Bayesian Data Analysis in Modeling and Forecasting Nonlinear Nonstationary Financial and Economic Processes

Oleksandr Trofymchuk, Petro Bidyuk,
Tatyana Prosiankina-Zharova and Oleksandr Terentiev

December 3, 2023

UDC 004.89:519.22](043.3)

O. Trofymchuk, P. Bidyuk, T. Prosyankina-Zharova, O. Terentiev

BAYESIAN DATA ANALYSIS IN MODELING AND FORECASTING NONLINEAR NONSTATIONARY FINANCIAL AND ECONOMIC PROCESSES

Oleksandr Trofymchuk

Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine, Kyiv,
orcid: 0000-0003-3358-6274

Trofymchuk@nas.gov.ua

Petro Bidyuk

National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»,
orcid: 0000-0002-7421-3565

pbidyuke_00@ukr.net

Tetyana Prosyankina-Zharova

Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine, Kyiv,
orcid: 0000-0002-9623-8771

t.pruman@gmail.com

Oleksandr Terentiev

Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine, Kyiv,
orcid: 0000-0002-4288-1753

o.terentiev@gmail.com

The study focuses on some aspects of modeling and forecasting the nonlinear nonstationary processes (NNP) of applying the modern Bayesian methods of data, in particular, generalized linear model (GLM) that are popular in analysis of NNP. All Bayesian techniques of data analysis are very popular today thanks to their flexibility, high quality of results, availability of possibilities for structural and parametric optimization and adaptation to new data and conditions of functioning. The structural and parametric adaptation of Bayesian generalized linear models supposes taking into consideration the following elements: number of equations that are necessary for adequate formal description of the processes under study; availability of nonlinearity and nonstationarity; type of random disturbance — its probability distribution and corresponding parameters; order of model equations, and some other structural elements. Such approach to modeling improves model adequacy and quality of final result of their application. Parameter estimation of the models can be performed by making use of rather wide set of methods, more precisely the following: ordinary LS (OLS), nonlinear LS (NLS), maximum likelihood (ML), the method of additional variable (MAV), and Monte Carlo for Markov Chain (MCMC). The last method is distinguished by universality of application to estimation of linear and nonlinear models. Besides, each of Bayesian approaches to data analysis is well supported by appropriate sets of statistical criteria that make it possible thorough quality analysis of intermediate and final results of computations. Illustrative examples are presented the usage of the Bayesian approach for analysis and forecasting of NNP, in particular, in specialized intellectual decision support system.

Keywords: nonlinear nonstationary processes, Bayesian methods, modeling, forecasting, generalized linear models.

Introduction

Many studies today are related to modeling and forecasting evolution of processes in various areas; they are mostly touching upon widely spread nonlinear nonstationary processes (NNP). To be more exact, definition of NNP means that such processes exhibit at least one type on nonstationarity (regarding trend or integration, and variance or heteroscedasticity) as well as nonlinearity regarding variables or model parameters. Such processes create majority in ecology, economy, finances, industrial technologies, engineering systems, hydrology, climate studies, in the problems of technical, medical and economic diagnostics, physical experiments of various type etc. For example, many processes in economy show availability of low (first or second) order trend, but transition to second order of integration automatically shifts the process from the class of linear to the nonlinear ones because quadratic and higher order dependence indicates relation to the class of nonlinear characteristics.

When we talk about process analysis we mean, as a rule, solving first of all the two most often met problems such as constructing the selected process model and forecasting. The widely used models of NNPs include, at least, the following types: differential equations, nonlinear regression, combination of linear and nonlinear regression, nonlinear autoregression, semi- and nonparametric methods, kernel based models, vector parametric models and methods, vector semi- and nonparametric approaches, state space and frequency-domain models, generalized linear models (GLM), and some others. Today high popularity acquired the methods of intellectual data analysis such as neural networks, Bayesian networks, decision trees and forests, immune, genetic and molecular algorithms. A separate subclass of nonlinear models belongs to models that are nonlinear in parameters such as logit and probit (logistic regression), ecological function, irrational and hyperbolic functions, Tornquist functions, and many others [1, 2]. The models nonlinear in parameters require application of special nonlinear estimation techniques for parameter estimation such as nonlinear LS, maximum likelihood (ML), Monte Carlo for Markov Chains (MCMC) and others.

One of the most widely met in practice subclass of nonlinear processes (and their models) is created by heteroscedastic processes, for example, they are very often considered in financial analysis. This direction of modeling and forecasting is widely known and practically used due to the fact that variance is considered in this case as dynamic variable which is described by dynamic model and its value is predicted to solving many problems. For example, predicted standard deviation (volatility) and variance itself are used for financial risk estimation and short-term forecasting of process volatility. Existing today variety of variance models and their estimation techniques is very high due to wide possibilities for their practical applications [3, 4].

Very popular approach to modeling NNP today is based upon Bayesian data analysis. Besides well-known static and dynamic Bayesian networks includes Bayesian regression, structural equation models, probabilistic filtering techniques, complex distribution analysis etc. The Bayesian approach to analysis of data and expert estimates has such positive features as structural flexibility of the models, possibility for taking into consideration uncertainties that are available practically in every area and case of studies, availability of model parameter estimation procedures for the cases of linear and nonlinear modeling, forecasting of complex distributions etc. [5–8].

This paper consider uncertainties as the factors of negative influence to the computational procedures used for model structure and parameter estimation, forecasts and control actions computing etc. Influence of the factors results in lower quality of intermediate and final results of data analysis, i.e. model adequacy, forecast estimates, control actions and decision alternatives.

The main goals of the study are as follows:

— to provide a review of Bayesian data processing and model constructing methods for their further use in intellectual decision support system for modeling and forecasting nonlinear nonstationary processes in economy and finances;

— to present illustrative examples of Bayesian techniques application to solve the problems mentioned;

— to stress the necessity of development intellectual decision support system (IDSS) for high quality solving the problems.

The review of Bayesian methods for modeling NNP illustrates some applications, and highlight the methods that could be used for reaching high quality results of data analysis. For example, it is often useful to stress development and application of a specialized IDSS and apply it to solving specific complicated problem. Then we applied this to building the GLM model of actuarial process and to building the Bayesian network to model the socio-economic processes taking place in regions of Ukraine.

Methodologies.

Bayesian methods for modeling and forecasting

Today there exists a wide set of Bayesian methods that are often used for preliminary data processing, model constructing, forecasting of future processes evolution, risk estimation, control in various spheres, decision support, classification and solving some other practical problems. Among others, the following Bayesian methods and techniques are actively used in practice, and should be mentioned [7–13]:

— generalized linear models; the set of exponential distribution laws used in the case of GLM application are as follows: normal, Poisson, binomial, Gamma, inverse Gaussian; such approach enhances substantially number of process that can be formally described with GLM;

— structural equation models (SEM); the models of this type make it possible constructing mathematical models for another class of random process that exhibit specific structure that can be adequately described by Bayesian techniques;

— static and dynamic Bayesian networks (BN); BN represent powerful probabilistic instrumentation capable formally describe sophisticated stochastic process and expert estimates, and generate probabilistic inference;

— Bayesian filtering of data and recursive parameter estimation; filtering touches upon the problem of data processing, i.e. reducing influence of noise components spoiling collected observation; parameter estimation relates to parameters of distributions and various models, first of all regression models we have to construct in multiple applications;

— Bayesian maps; trajectory synthesis and control of robotic systems;

— Markov localization models; the problem of robot localization and control is considered;

— multivariate distribution constructing and analysis; forecasting multivariate distributions, parameter estimation;

— decision trees and forests;

— combining Bayesian and statistical techniques into single model; model combining provides a possibility for modeling and forecasting complicated NN processes.

The methods listed above are distinguished with their high flexibility and possibilities for taking into consideration possible data uncertainties and generating alternative final results directed to support of decisions according to specific problem statement [14–19]. Consider details of some methods.

Generalized linear models

Generalized linear models is a class of regression models that allow for taking into consideration interaction between model factors, specific distribution law of dependent variable and possible nonlinearity [12, 13]. GLM consists of the three basic components: systematic, stochastic, link function and can be formally represented as follows:

$$\mu_i = E[y_i] = g^{-1}\left(\sum_j X_{ij}\beta_i + \xi_i\right), \tag{1}$$

$$Var[y_i] = \frac{\phi V(\mu_i)}{\omega_i},$$

where y_i is a vector of observations for dependent variable; $g(x)$ is a link function X_{ij}; is a matrix of observations for a model factors; β_j is a vector of parameters estimated on factor observations; ξ_i is a vector of stochastic residuals; ϕ is a vector of scale parameters for the function of $V(x)$; ω_i — prior weights of confidence level.

Thus, GLM is characterized by the following elements: distribution law for dependent variable, Y; parameters and specific features of a link function $g(\cdot)$; features of the linear predictor, $\eta = X\beta$.

Usually it is reasonable to make the following suggestions regarding GLM:

- all the components of dependent variable Y are independent and their distribution law belongs to the family of exponential distributions;
- the suggestion regarding systematic feature of a model is treated as follows: p predictors are combined into single «linear predictor» η;
- the suggestion regarding link function: mutual dependence between suggestions of stochastic and systematic features is expressed by the link function that is supposed to be differentiable and monotonic and has an inverse:

$$E[y] = \mu = \mathrm{g}^{-1}(\eta). \tag{2}$$

The set of possible distribution laws for GLM and parameters of the distributions for dependent variable are presented in Table 1 [17].

Table 1

Component	Normal distribution	Poisson	Binomial	Gamma distribution	Inverse Gaussian
	$N(\mu, \sigma^2)$	$P(\mu)$		$G(\mu, v)$	$\mathrm{IG}(\mu, \sigma^2)$
Variance param., ϕ	$\phi = \sigma^2$	1	$\frac{1}{n}$	$\phi = \mathrm{v}^{-1}$	$\phi = \sigma^2$
Cumulant function, $b(\theta)$	$\frac{\theta^2}{2}$	$\exp(\theta)$	$\log(1+e\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$c(y, \phi)$	$-\frac{1}{2}\left\{\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right\}$	$-\log(y!)$	$\log\frac{n}{y}$	$v\log(vy)$ – $-\log(y)$ – $-\log(\Gamma(\nu))$	$-\frac{1}{2}\left\{\log(2\pi\phi y^3) + \frac{1}{\phi y}\right\}$
$\mu(\theta) = E(Y, \theta)$	θ	$\exp(\theta)$	$\frac{e^\theta}{1+e^\theta}$	$-\frac{1}{\theta}$	$(-2\theta)^{1/2}$
Canonic link, $\theta(\mu)$	identity	log	logit	reciprocal	$\frac{1}{\mu^2}$
Variance func, $V(\mu)$	1	M	$\mu(1-\mu)$	μ^2	μ^3
$Var(\mu)$	σ^2	$n\mu(1-\mu)$	μ	$\frac{\mu^2}{\nu}$	$\frac{\sigma^2}{\mu^3}$

The link function links linear predictor η to the estimate μ related to *Y*. In a classic linear model mean the value of dependent variable and linear predictor are identical, and identity link (η and μ) are selected arbitrarily but from the set of real numbers. GLM is distinguished from linear model of general form (special case of such model is multiple regression) by the following:

— distribution of dependent variable (system reaction) can be non-Gaussian and not necessarily continuous, for example, binomial;

— dependent variable predictions are computed as linear combination of predictors that are linked to the dependent variable with a link function.

Dependently on the distribution law of dependent variable and type of the link function there exist types of GLM [17, 18] presented in Table 2.

Table 2

Model type	Link function	Dependent variable distribution
Linear model of general type	$g(\mu) = \mu$	Normal distribution
Log-linear model	$g(\mu) = \ln(\mu)$	Poisson distribution
Logit-model	$g(\mu) = \ln(\mu/(1-\mu)$	Binomial distribution
Probit-model	$g(\mu) = \Phi - 1\mu$	Binomial distribution
«Survival» analysis	$g(\mu) = \mu - 1$	Gamma distribution, exponential distribution

Thus, GLM represents quite general class of statistical models that includes linear regression, variance and covariance analysis, Log-linear models for analysis of random tables, logit/probit models, Poisson regression and many others.

Each distribution has its specific link function for each exists substantiated equality statistic regarding parameter vector β of linear predictor, $\eta = \sum x_i \beta_i$. Such canonic link comes to being when, θ = η, where θ is canonic parameter which is determined when likelihood function is introduced. Specific distribution type corresponding to each link function is given in Table 2, and generalized form of distribution is given below:

$$f(y, \theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{\alpha(\phi)} + c(y, \phi)\right], \qquad (3)$$

where *a*, *b*, *c* — are functions that correspond to specific distribution law; *y* is dependent variable; θ is canonic parameter or a function of some parameter of a specific distribution; ϕ is variance parameter.

Function *b*(·) has a special meaning in generalized linear models because it described relation between mean value and variance in a distribution. When ϕ is known then we have exponential model with canonic parameter θ. When ϕ is unknown, then exponential distribution can be of two-parameter type. Thus, canonic link for a set of exponential distributions has the form presented in Table 3.

Table 3

Distribution type	Canonic link
Normal distribution	$\eta = \mu$
Poisson	$\eta = \log(\mu)$
Binomial distribution	$n = \log\left\{\frac{\pi}{1-\pi}\right\}$
Gamma distribution	$\eta = \mu - 1$
Inverse Gaussian distribution	$\eta = \mu - 2$

A substantiated statistic for canonical forms is the vector $X^{\mathrm{T}}y$:

$$\textstyle\sum_i x_{ij} y_i,\ j = 1 \ldots p. \qquad (4)$$

It should be noted that canonical link results in desirable features of a model, especially for short samples, for example existence of systematic effects. And here are approaching the problem of selecting statistic criteria for model estimating. As a result of analysis of various criteria for model selection the conclusion was made that Akaike information criterion fits better to solving the problem together with the maximum likelihood approach to estimation.

Quality estimation for generalized linear models

Usually in the process of analyzing quality (adequacy) of a model several different criteria are applied. In some cases these criteria coincide with the same statistical criteria used in constructing linear and/or nonlinear regression. When we use the models on the purpose of classification, among them there are the following: common accuracy of a model; I-st and II-nd type errors; ROC-curve, and Gini index.

Common accuracy (CA) statistic is determined by the expression:

$$CA = \frac{Correct\_Forecast}{N}, \tag{5}$$

where, *Correct Forecast* is a number of correctly forecasted cases (examples); *N* is a total number of cases under investigation.

This criterion is somewhat subjective measure of a model quality because it depends on a number of defaults in the dataset, and the level of the cut-off threshold. Different levels of the threshold result in different values of common accuracy.

ROC-curve (receiver operation characteristic) shows dependence of a number of correctly classified positive examples on the number of incorrectly classified negative examples. The first set of examples is called true positive, and the second set are referred to as false positive ones. Here it is suggested that the classifier has some parameter that can be varied to reach necessary division into classes. This parameter is called cut-off point or threshold. Depending on the parameter value different values of the I-st and II-nd type errors will be achieved.

Most often the following statistics (in percentages) are used for determining quality of a model:

— a part of true positive examples (true positives rate) [16]:

$$TPR = \frac{TP}{TP + FN};$$

— a part of false positive examples (false positives rate):

$$FPR = \frac{FP}{TN + FP}.$$

Usually for completeness of the analysis two more characteristics are applied: sensitivity and specificity.

Model sensitivity is defined as a part of true positive cases [16, 17]:

$$Se = TPR = \frac{TPR}{TP + FN}. \tag{6}$$

Model specificity is defined as a part of true negative cases that were correctly classified by a model [16, 17]:

$$Sp = \frac{TN}{TN + FP}. \tag{7}$$

Now, it is obvious that [16, 17]:

$$Sp = \frac{TN - FP + FP}{TN + FP} = 1 - \frac{FP}{TN + FP} = 1 - FPR. \tag{8}$$

The model that exhibits high sensitivity provides for a true result when number of positive cases is also high (i.e., it reveals well positive cases). And the model that exhibits high specificity usually provides for better (true) results when the number of negative cases is high (i.e., the model discovers better negative cases). The ROC curve (or Lorentz curve) graph uses the axes *Y* for sensitivity *Se*, and axis *X* for a part of false positive cases *FPR* or (1-Sp).

The graph of ROC-curve for an ideal classifier tends to the left upper corner where part of true positive cases tends to 1 (the case of ideal sensitivity), and the part of false positive samples tends to zero. Thus, the closer approaches ROC-curve to the left upper corner the better is the model regarding prediction of a true value. As a consequence, straight diagonal line corresponds to the classifier that is unable to distinguish between the two classes. As far as visual comparison of different ROC-curves not always allows selection of better model, there is a numerical criterion in the form of area under curve (AUC) that makes the model selection easier. It is computed using the trapeze method as follows [18]:

$$AUC = \int f(x)dx = \sum_i \left[\frac{x_{i+1} + x_i}{2} \right] * (Y_{i+1} - Y_i). \qquad (9)$$

As alternative model quality measure is used Gini index that is computed as an area between diagonal and Lorentz curve divided by the whole area under the diagonal. The index is widely used for resolution analysis of a system developed for credit risk estimation. In this case the model is used for dividing the clients into two groups: those who are inclined to default, and those who are not. Usually, probability is hired as a measure of inclination to default.

Application of Bayesian approach to estimation

In most cases of solving the problems of process modeling, forecasting and decision support based on statistical data we meet uncertainties of various types. As an example can serve structural, parametric and statistical uncertainties available in development and practical application of identification procedures for analysis of processes of various nature. Structural uncertainties are linked to the uncertainties of developed model structure, and parametric are referred to the model parameter estimates. Statistical uncertainties are mostly related to observations, for example to the difficulties of determining true data distribution, influence of external random factors corrupting the measurements, missing observations, extreme values etc. Theoretical studies of such problems are mostly related to analysis of reasons for emergence, classification and influence estimation of the uncertainties as well as level and probabilities of respective risks.

Very often we lack statistical data to solve the problems of risk analysis and decision making. Such problems cannot be solved with traditional statistical frequency approach because the data available cannot provide necessary information. Moreover, the situations related to decision making may change substantially and lack the results of preliminary analysis. Such particular features lead to complications with decision making and may influence negatively final results. That is why the Bayesian approach becomes in such cases useful and highly effective instrument of modeling, forecasting and decision support.

A distinctive feature of Bayesian approach to data and expert estimates processing is that researcher considers the level of his belief to possible models and forecasts before receiving data and represents his view in the form of a probability. As soon as the necessary data is received the Bayes theorem provides a possibility for computing another set of probabilities representing with them refined beliefs to the candidate models.

The key advantage of Bayesian approach to data expert estimate analysis is in the possibility of using any prior information related to system under study state and its model (for example, model structure and its parameters). Such information is presented in the form of prior probabilities and/or density distribution. The prior probabilities are recalculated (improved) further on using the data that are used to determine posterior distribution of parameter estimates or respective output variables.

Consider as example application of conjugate distributions what means that prior and posterior distributions are of the same type.

Let experimental data $X_1,\ldots, X_n$ is normal random sample with mean value μ and variance σ^2. Suggest that prior distribution is also normal with mean μ_0 and variance σ_0^2. Then posterior distribution for μ with known sample $X_1,\ldots, X_n$ and known prior distribution will also be normal with mean μ_* and variance σ_*^2 that are computed as follows [8, 10]:

$$\mu_* = \frac{\sigma^2\mu_0 + n\sigma_0^2\bar{X}}{\sigma^2 + n\sigma_0^2}, \quad \sigma_*^2 = \frac{\sigma^2 + \sigma_0^2}{\sigma^2 + n\sigma_0^2},$$

where $\bar{X} = \frac{\sum_{i-1}^{n} X_i}{n}$ is sample mean.

Bayesian analysis often uses parameter characterizing quality of results, and determined as inverse to variance: $\eta = 1/\sigma^2$. Thus, we have for the prior distribution, $\eta_0 = = 1/\sigma^2$ and for posterior distribution $\eta_* = \frac{1}{\sigma_*^2}$. Now the expressions for posterior parameter will take the form [8, 10]:

$$\eta_* = \eta_0 + n\eta, \quad \mu_* = \frac{\eta_0}{\eta_*}\mu_0 + \frac{n\eta}{\eta_*}\bar{X}.$$

Thus, for the case of normal sample information about μ is contained in the sample mean $\bar{X}$ that is complete statistic for μ. The quality of distribution identification is determined by the relation: X: $\frac{\eta}{\sigma^2} = n\eta$. Quality of posterior distribution description is determined by the quality of two following components: prior distribution representation, and statistical characteristics of sample data. The posterior mean is weighted prior mean and sample mean with weighting coefficient that is proportional to the quality parameter. It means that influence of prior distribution is decreasing with growing size of data sample *n*.

Consider the problem of forecasting random value x on the basis of historical observations y using Bayesian approach. That is, it is necessary to determine type of distribution for the future values of x given values of y. From the probabilistic point of view the problem is in determining forecast density $\pi(x|y)$ that describes possible changes with known values of y.

Most often, the model for which the forecast density is needed does not exist. But known is the probabilistic model for x that is expressed in terms of distribution $g(x|\theta)$ depending on unknown parameter θ that is based on the model describing observations y. If we denote posterior density of θ given y as $P(\theta|y)$ then forecasting density for x can be written as follows [8, 10]:

$$\pi(x|y)) = \int_{\theta} g(x|\theta)P(\theta|y)d\theta. \qquad (10)$$

If a point forecast is required, it is enough to use the point estimate of $(x|y)$. To find an interval forecast it is enough to compute interval estimates for $(x|y)$.

Results & discussion.

The application of GLM for modeling the actuarial processes

As an example of actuarial process modeled by GLM the problem of estimating (forecasting) financial loss in car insurance had been selected. The experimental data includes one dependent variable, «Loss», reflecting paid volume of insurance among the cars of the following brands: VAZ, Mitsubishi and Toyota. The regions of selling (distribution) of the car policies include the cities of Kyiv, Donetsk and Odesa. The data relates to the years starting from 2006 with the sample size of 9546 examples. The results of GLM constructing for different distribution laws are presented in Tables 4–6.

Table 4

Model	
Distribution of dependent variable	Link function
Gamma	LOG
Normal	LOG
Poisson	LOG
Normal	Identity

Table 5

Total loss, bln. UAH	Mean	Std. deviance	Min	Max	Mean of standard error	Variance, %
102,01	11805,69	15358,118	6273,867	18549,819	0,075	130,091
18,11	1897,46	939,910	4010,98	634,054	0,120	49,535
17,92	1877,53	1027,57	4234,95	558,35	0,176	54,73
17,92	1877,53	999,30	3535,4	118,0	224,35	53,22

Table 6

Total loss	Log-likelihood	Difference	Risk
102008320,905	– 15742,754	84087288,32	1,301
18111231,380	– 98700,167	190198,799	0,495
17921032,574	– 42173677,24	0,007	0,547
17921032,589	– 98700,167	0,009	0,532
102008320,905	– 15742,754	84087288,32	1,301

Table 6 shows that the risk of financial loss for the models constructed varies approximately between 40–60% what is marginally acceptable but requires undertaking of some measures regarding its minimization.

Usage of Bayesian methods for modeling the socio-economic processes

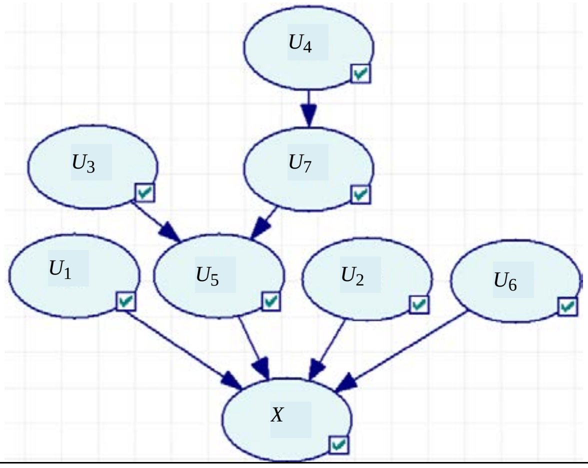
As a part of the study, computational experiments using various socio-economic indicators were performed the quality of results obtained are quite acceptable. In particular, the paper presents the use of a continuous Bayesian network to study the effectiveness of the decentralization reform [20, 22].

The time series of indices that are relevant to the reform period: 2015-2021 [20–22], were considered, which characterize the socio-economic component of the reform — the growth of local budget revenues per resident of the community due to the increase in tax revenues of local budgets as a result of the activation of the community economy. These processes are sophisticated for studying: they are nonlinear and nonstationary. Table 7 shows the values of the estimates of mathematical expectation and standard deviation, according to the normal distribution of the regressors and the target variable.

Table 7

Variable	Indicator, %	Mathematical expectation of the variable	Standard deviation of the variable
U_1	Index of the volume of industrial products sold	– 2,414	5,845
U_2	Agricultural production volume index	1,428	8,625
U_3	Index of construction products	5,314	12,232
U_4	Increase in revenues to the general fund	– 30,214	43,204
U_5	Growth rate of land fee revenues	19,457	21,464
U_6	Basic grant	– 0,771	0,969
U_7	Index of physical change of gross regional product	0,128	5,055
X	Dynamics of local budget income per inhabitant of community	3,891	12,474

A continuous time Bayesian network was developed and used to model the studied processes. Continuous Time Bayesian networks are used to model stochastic processes in continuous time state space. The Gaussian distribution is used in this problem what was proved by appropriate statistics. The topology of the Bayesian network developed is presented in figure.



The target variable X has multiple parents $U = \{U_1, U_2, U_5, U_6\}$.

$$f(X|U_i) = N(x; \mu_x + b_i * \mu_i, \sigma_x),$$

$$N(x; \mu_x + b_i * \mu_i, \sigma_x) = \frac{1}{\sqrt{2\pi\sigma_x}} \exp\left(\frac{1}{2} * \frac{(x - (\mu_x + b_i * \mu_i, \sigma_x))^2}{\sigma_x}\right),$$

where μ — is the mathematical expectation, σ — is the variance, and is the coefficient $\frac{1}{\sqrt{2\pi\sigma_x}}$ normalization constant ensuring that $\int_x N(x; \mu_x + b_i * \mu_i, \sigma_x) = 1$, b_i — coefficient characterizing the relationship between X and its i-th parent (it is also called a weight coefficient) [23].

The relationship between specific variable X and its parents (U_1, U_2, ..., U_n) is formally described by a linear regression model:

$$X = b_1 U_1 + b_2 U_2 + \ldots + b_n U_n + Q_x,$$

where Q_x is a noise component that can be written in the form of a Gaussian distribution with zero mathematical expectation, and b_1, b_2,..., b_n are regression coefficients showing the relationship between the variable X and U_1, U_2,..., U_n — its parents.

Regression coefficients of the model were estimated by the method of least squares. The structure of the regression model is as follows:

$$X = -1,23U_1 + 3,34U_2 - 0,39U_5 + 9,15U_n + Q_x.$$

The greatest impact on the target variable X is exerted by U_6 (base subsidy, % of the schedule), which indicates insufficient independence of local budgets.

The predictive characteristics of the model are characterized by the following statistics: RMSE = 3,4, MAPE = 13,5 % — they are quite acceptable.

Conclusion

Thus, it is presented short review of modern data processing Bayesian techniques. An example of modeling and forecasting of financial actuarial process is presented. It was shown that generalized linear models can serve as effective instrument analyzing financial and economic processes that helps to take into consideration actual complicated factor interactions and their influence on dependent variable. There also exists a possibility for model and forecasts quality analysis (of the results achieved) using a set of appropriate statistical criteria.

Further research in this direction should be focused on the problems touching upon the following issues: refinement of the forecasting model; more profound analysis of factors influencing dependent variable; more active use of Bayesian and neural networks and other methods of intellectual data analysis to modeling and forecasting actuarial processes; development and use of commercial intellectual decision support system. The DSS will help to construct and study more sophisticated combined model consisting of linear and nonlinear parts, to reach higher quality forecasts of dependent variable, and consequently improve estimates of possible financial loss. This approach to financial processes analysis will help to minimize financial risks in insurance as well as in many other spheres of human activities. Finally such studies will positively influence macro economy as a whole.

О.М. Трофимчук, П.І. Бідюк,
Т.І. Просянкіна-Жарова, О.М. Терентьєв

БАЙЄСІВСЬКИЙ АНАЛІЗ ДАНИХ У МОДЕЛЮВАННІ ТА ПРОГНОЗУВАННІ НЕЛІНІЙНИХ НЕСТАЦІОНАРНИХ ФІНАНСОВО-ЕКОНОМІЧНИХ ПРОЦЕСІВ

Трофимчук Олександр Миколайович

Інститут телекомунікацій і глобального інформаційного простору НАН України, м. Київ,

Trofymchuk@nas.gov.ua

Бідюк Петро Іванович

Національний технічний університет «Київський політехнічний інститут імені Ігоря Сікорського»,

pbidyuke_00@ukr.net

Просянкіна-Жарова Тетяна Іванівна

Інститут телекомунікацій і глобального інформаційного простору НАН України, м. Київ,

t.pruman@gmail.com

Терентьєв Олександр Миколайович

Інститут телекомунікацій і глобального інформаційного простору НАН України, м. Київ,

o.terentiev@gmail.com

У статті представлено короткий огляд сучасних байєсівських методів аналізу даних, наведено особливості застосування узагальнених лінійних моделей (УЛМ) в аналізі нелінійних нестаціонарних процесів, підкреслено їхні можливості та особливості застосування до процесів різної природи. Усі байєсівські методи аналізу даних сьогодні дуже популярні завдяки своїй гнучкості, високій якості результатів, можливості структурно-параметричної оптимізації та адаптації до нових даних і умов функціонування. Структурно-параметрична адаптація байєсівських УЛМ передбачає врахування кількості рівнянь, які необхідні для адекватного опису досліджуваних процесів; наявності нелінійності та нестаціонарності; типу випадкового збурення — його розподілу ймовірностей; порядку використаних рівнянь та деяких інших елементів структури. Все це сприяє підвищенню адекватності моделей, що будуються, і якості остаточного результату їх застосування. Для оцінювання параметрів цих моделей можна примінити досить широку множину методів, зокрема таких: звичайний метод найменших квадратів (МНК), нелінійний МНК (НМНК), метод максимальної правдоподібності (ММП), метод допоміжної змінної (МДП) і метод Монте–Карло для марковських ланцюгів (МКМЛ), який відрізняється універсальністю застосування до оцінювання параметрів лінійних та нелінійних моделей. Крім того, кожен з байєсівських методів аналізу даних добре підтримується відповідними множинами статистичних критеріїв, які роблять можливим ретельний якісний аналіз проміжних і кінцевих результатів. У статті наведено приклад застосування GLM для прогнозування фінансових втрат у страхуванні, запропоновано використання байєсівського аналізу даних у спеціалізованій інтелектуальній системі підтримки прийняття рішень, що дозволило підвищити якість результатів обчислень.

Ключові слова: нелінійні нестаціонарні процеси, байєсівські методи, моделювання, прогнозування, узагальнені лінійні моделі.

REFERENCES

1. De Gooijer J.G. Elements of nonlinear time series analysis and forecasting. Cham : Springer, 2017. 618 p. https://doi.org/10.1007/978-3-319-43252-6
2. Bidyuk P.I., Romanenko V.D., Tymoshchuk O.L. Time series analysis. Kyiv : Polytechnika Publisher at the National Technical University of Ukraine «Igor Sikorsky KPI», 2013. 600 p.
3. Tsay R.S. Analysis of financial time series. Hoboken Wiley & Sons, Inc., 2010. 720 p. https://doi.org/10.1002/9780470644560
4. Hansen B.E. Econometrics. Madison : University of Wisconsin, 2021. 1026 p.
5. Rossi P.E., Allenby G.M., McCulloch R. Bayesian statistics and marketing. Hoboken : John Wiley & Sons, Ltd., 2005. 368 p. https://doi.org/:10.1111/j.1467-985X.2006.00446_13.x
6. Press S.J. Subjective and objective Bayesian statistics: principles, models, applications. Hoboken : John Wiley & Sons, Inc., 2003. 600 p.
7. Bidyuk P.I., Terentyev O.M., Konovalyuk M.M. Bayesian networks in technologies of intellectual data analysis. *Artificial intelligence*. 2010. Vol. 2. P. 104–113.
8. Zgurovsky M.Z., Bidyuk P.I., Terentiev O.M., Prosyankina-Zharova T.I. Bayesian networks in decision support systems. Kyiv : Edelweys, 2015. 300 p.
9. Jensen F.V. Bayesian networks and decision graphs. New York : Springer-Verlag, 2001. 268 p. https://doi.org/10.1007/978-1-4757-3502-4
10. Lee Sik-Yum. Structural equation modeling: a bayesian approach. Chichester : John Wiley & Sons, Ltd., 2007. 432 p. https://doi.org/10.1002/9780470024737

11. Chen Ming-Hui, Qi-Man Shao, Ibrahim J.G. Monte Carlo methods in Bayesian computation. New York : Springer-Verlag, 2000. 387 p. https://doi.org/10.1007/978-1-4612-1276-8
12. De Jong P., Heller G.Z. Generalized linear models for insurance data. New York : Cambridge University Press, 2008. 196 p.
13. Mc Cullagh P., Nelder J.A. Generalized linear models. New York : Chapman & Hall, 1989. 526 p.
14. Sugumaran V. Intelligent support systems technology: Knowledge management. London : IRM Press, 2002. 318 p. https://doi.org/10.4018/978-1-931777-00-1
15. SAS Institute Inc. SAS/OR 14.2 User's Guide: Mathematical Programming. http://support.sas.com/thirdpartylicenses.
16. Kuznietsova N., Bidyuk P. Intelligence information technologies for financial data processing in risk management. *Data Stream Mining & Processing. Communications in Computer and Information Science*. Cham : Springer, 2020. Vol. 1158. https://doi.org/10.1007/978-3-030-61656-4_36
17. Trofymchuk O., et al. Decision support systems for modelling, forecasting and risk estimation. Riga : Lambert Academic Publishing, 2019. 176 p.
18. Davidson R., MacKinnon J.G. Econometric theory and methods. Oxford : Oxford University Press, 2004. 652 p. https://doi.org/10.1017/S0266466605000356
19. Castle J.L., Doornik J.A., Hendry D.F. Modelling nonstationary 'Big Data'. *International Journal of Forecasting*. 2021. Vol. 37, N 4. P. 1556–1575. pttps://doi.org/10.1016/j.ijforecast.2020.08.002
20. State Statistics Service of Ukraine. https://www.ukrstat.gov.ua
21. Center-for-sociological-research-decentralization-and-regional-development. https://kse.ua/kse-impact/center-for-sociological-research-decentralization-and-regional-development/
22. Decentralization. https://decentralization.gov.ua/en/
23. Hautaniemi Sampsa K., Petri T. Korpisaari, Jukka P.P. Saarinen. Target identification with dynamic hybrid Bayesian networks. Target identification with dynamic hybrid Bayesian networks. *VI Image and Signal Processing for Remote Sensing*. 2001. Vol. 4170. P. 92–102. https://doi.org/10.1117/12.413885

Submitted 21.07.2023