ISSN 2277-3061

# Forecasting based on Bayesian type models

Peter Bidyuk [1], Alexander Gozhyi [2], Alexandr Trofymchuk [3]
Dr.of eng.sci., professor at the Institute for Applied System Analyssis, NTUU "KPI",Kiev, Ukraine [1]
pbidyke@gmail.com
Ph.D., associate professor, Department of Information Technology and Program Systems, Black Sea State University named after Petro Mohyla, Nikolaev, Ukraine [2]
alex_daos@mail.ru
Dr.of eng.sci., professor at the Institute of Telecommunication and Global Information Sphere at NASof Ukraine, Kiev, Ukraine [3]
itelua@kv.ukrtel.net

## ABSTRACT

A review of some Bayesian data analysis models is proposed, namely the models with one and several parameters. A methodology is developed for probabilistic models construction in the form of Bayesian networks using statistical data and expert estimates. The methodology provides a possibility for constructing high adequacy probabilistic models for solving the problems of classification and forecasting. An integrated dynamic network model is proposed that is based on combination of probabilistic and regression approaches; the model is distinguished with a possibility for multistep forecasts estimation. The forecast estimates computed with the dynamic model are compared with the results achieved with logistic regression combined with multiple regression. The best results were achieved in this case with the combined dynamic net model.

## Indexing terms/Keywords

Bayesian network, Bayesian models, statistical data, forecasting, dynamic Bayesian network, multistep forecasts estimation, logistic regression, multiple regression.

Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol. 15, No. 3
www.ijctonline.com, editorijctonline@gmail.com

ISSN 2277-3061

## INTRODUCTION

The modern information technologies and statistical analysis of experimental data is based on usage of a wide variety of methods that are characterized by the in-depth and comprehensive analysis of available measurements and development of mathematical and statistical models of high degree of adequacy. The models are suitable for solving practical problems of prediction and control and studying the objects selected. In analysis of data of different nature, we often meet so called uncertainties in the form of incomplete information about the process under study, significant influence of random disturbances and measurements noise (errors), nonstationary process structure and its parameters [1, 2]. That stresses availability of structural, statistical and parametric uncertainties which create difficulties with analyzing the data and reduces quality of the final result: estimate forecasts and/or quality of automatic control. All these problems require development of new approaches and methods for modeling, forecasting and control in the presence of mentioned uncertainties.

Quite effective tool for evaluation and consideration of statistical nature uncertainties is adaptive Kalman filter (KF), which allows assessing and prediction the state of dynamic processes [2 – 5] in real time. Adaptation of a model to the characteristics of random disturbances and measurement noise is achieved in this case by using computed in real-time estimates of covariance matrices for these random processes. The advantages of optimal filtration procedures are the opportunity to take into account explicitly statistical characteristics of disturbances and measurement noise, possibility for calculating optimal estimates of state variables, evaluating non-measurable components of a state vector and simultaneous evaluation states and some model parameters. The disadvantages of KF are as follows: significant reduction in the quality of forecasts, when the number of predicting steps is more than one, sensitivity of the state estimation procedure to adequacy of models and significant complication of computational algorithms in the case of non-linear system analysis.

Significant progress has been achieved towards modeling and forecasting processes with uncertainty using fuzzy logic and neuro-fuzzy networks [6, 7]. When using the methods of this class the uncertainty is referred to as "fuzzy" linguistic variables that require creating rules for decision making. Modeling procedures based on fuzzy logic are characterized by relative simplicity and the possibility of adapting to the processes of special class. Disadvantages include the need for generating a large number of rules when studying multidimensional processes and the impossibility of tracking their use by a decision making person (DMP) while forming the inference.

Another broad class of modeling methods for prediction and management, which are also aimed at treating uncertainty, is based on the Bayesian approach to data analysis [8 – 12]. The methodology of Bayesian data analysis and expert estimates is consistent with the logic of DMP actions while analyzing the processes of arbitrary nature, alternatives generating and decision making. Prior information about the process under study is complemented by experimental data, additional information of qualitative or quantitative nature that may be obtained from various sources. The integrated knowledge and data are necessary for estimating posterior probability for variables, parameters, conditions, situations and so on. All Bayesian methods are successfully used in all stages of data analysis in modeling, forecasting and decision making. At the stage of preliminary data processing are applied probabilistic filtering, filling possible data gaps; at the stage of modeling the formation of model structures and parameters estimates; and at the stage of alternatives generating: calculation of probabilistic inferences (decisions) using previously constructed models. Bayesian methods have the following advantages: the possibility for taking into consideration uncertainties of statistical, structural and parametric nature (e.g. using Bayesian networks (BN)); include a large number of heterogeneous variables in one model; availability of sufficiently flexible parameter estimation procedures as well as a presence of broad range methods for generating accurate and approximate inference. The disadvantages include difficulties of obtaining prior information and relative complexity of some computational procedures relating to numerical integration, parameter estimation and forming probabilistic inference. Regarding the shortcomings we can say that in some cases they do exist and create difficulties for the researcher, but with the acquisition of experience with these methods and quality of relevant computational data analysis procedures and knowledge level and scope of these challenges is significantly reduced [8].

In view of the needs to improve the methods of fighting uncertainties of various types and the availability of a wide range of methods for Bayesian data analysis, it is necessary to know their application possibilities, advantages and disadvantages. This is particularly important in the context of decision support system (DSS) development [13 – 15] as far as probabilistic methods and models make it possible to obtain important additional alternative methods for decision making based on regression analysis, fuzzy logic, neural networks, and so on. To some extent this problem is partially solved in this paper through examination of some popular methods of Bayesian data analysis and probabilistic models, and an attempt is made for association with other types of models, including regression.

## FORMULATION OF THE PROBLEM

The objectives of the study are as follows: to review some types of Bayesian models and to determine their possible use for mathematical modeling of studied processes dynamics and short-term forecasting; to offer an integrated model, that is based on a combination of regression and probabilistic approaches; to perform comparative analysis of predictions using the proposed combination of probabilistic model with regression models.

ISSN 2277-3061

## SOME BAYESIAN METHODS FOR DATA ANALYSIS

Bayesian analysis (BA) means creating methods and models to form opinions on the probability values for selected variables or parameters based on statistical (experimental) data and expert assessments. The essential difference between Bayesian methods of regression analysis and other methods is the explicit use of probabilities for quantitative description of uncertainties in the inference formed on their basis of [9]. The ease of probability application for quantitative description of uncertainties can be explained as follows: (1) by analogy, as far as uncertainty is the result of physical randomness, it is logical to describe this ambiguity in the language of random events; (2) axiomatic approach: a probabilistic approach leads to statistical inference context of decision making in the form of losses or winnings; in this case using the axioms of ordering and transitivity uncertainty should be represented in terms of probability; (3) the existence of the principle of probability coherence: according to this principle assigning the probabilities to all possible events must be such that a particular person could not get quantified gain as a result of taking part in the game and other events. Thus, the probability can play a measure of uncertainty in applied statistics, but definitive proof of this is the possibility of successful practical application of relevant models and methods in each case.

The model development procedure can be divided into three stages:

– Building a full probabilistic model as a joint probability distribution for all observed and not observed variables that are used to describe the state of the investigated object or process; this model should be consistent with the laws of the process and available data.

– Using existing data and knowledge about the process under study, calculate and interpret appropriate posterior distribution, that is to find the conditional probability distribution for specified target variables.

– Assess the quality of the model constructed and interpret the posterior distribution; i.e. assess the quality of data description model, identify the sensitivity of the result to the assumptions made at the first stage, and check the correctness of the conclusions; if necessary, the model can be modified or expanded, i.e. all three phases of data analysis can be repeated.

It is obvious that considerable difficulties may arise during the first stage in evaluating a priori distributions for building probabilistic models. For solving this problem a significant role plays in-depth knowledge about the functioning, structure and parameters of the facility and the existing experience of analyzing the processes of specified class. However, Bayesian approach to data analysis has the advantage that it is able to present a clear logical interpretation of probabilistic and statistical findings, in particular to calculate the probability of hitting unknown quantity at a certain interval in contrast to the formation of the confidence interval at a frequency approach. This strict interpretation of the final result of the frequency approach is possible only in respect of a certain number of implementations of similar events that repeat in practice. In addition, the Bayesian approach provides an opportunity to express uncertainty directly in quantitative (probabilistic) form without using multivariable models and complex (hierarchical and often iterative) procedures for their construction. Bayesian approach provides researchers conceptually simple methods of forming structured multiparameter data models.

Bayesian statistical inference on the evaluation parameter (in general this is vector of parameters) or unobservable values formulated in terms of probability statements. These statements of conditional probability using the observed values $y$, leads to the inference formulated in the form of conditional probabilities $p(\theta \mid y)$ or $p(\tilde{y} \mid y)$. Obviously, this condition may be extended by the observed explanatory variables or known parameters.

For the possibility of forming opinions on the probability $\theta$ subject to availability of observations $y$ it is necessary to build a model for joint probability distribution regarding $\theta$ and $y$. The joint distribution (joint distribution function) could be written as the product of two densities:

$$p(\theta, y) = p(\theta)\, p(y \mid \theta),$$

where $p(\theta)$ is prior distribution for parameter $\theta$; $p(y \mid \theta)$ is data distribution. According to Bayes theorem posterior conditional density for the parameter can be calculated as follows:

$$p(\theta \mid y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)\, p(y \mid \theta)}{p(y)}, \qquad (1)$$

where $p(y) = \sum_{\theta} p(\theta)\, p(y \mid \theta)$, that is the sum is calculated for all values of $\theta$ (if parameter is discrete) or $p(y) = \int p(\theta)\, p(y \mid \theta)\, d\theta$ (in the case of continuous parameter). Equivalent to form (1) is approximate posterior density

$$p(\theta, y) \propto p(\theta)\, p(y \mid \theta), \qquad (2)$$

where the symbol "$\propto$" denotes proportionality.

In order to form an inference on the unknown value of the measured variable $y$ we need to know distribution of this variable:

$$p(y) = \int p(y,\theta)\,d\theta = \int p(\theta)\,p(y\,|\,\theta)\,d\theta. \qquad (3)$$

This probability is often referred to as marginal $y$ but informative title for it is *a priori distribution forecast*. Prior since there are no previous observaions, and forecast because the distribution is built for the observed values. After receiving the observaions $y$ we can calculate the estimate of the unknown variable value $\tilde{y}$ based on previous observations. For example, suppose $y=[y_1, y_2, \ldots, y_n]$ is a vector of weight measurements for an object; $\theta = [\mu, \sigma^2]$ is unknown true weight of the object and variance of the measurements; $\tilde{y}$ is expected value of the next measurement. In this case, the distribution of values $\tilde{y}$ is called predictor of posterior distributions because it is determined subject to availability of observations $y$:

$$p(\tilde{y}\,|\,y) = \int p(\tilde{y},\theta\,|\,y)\,d\theta = \int p(\tilde{y}\,|\,\theta,y)\,p(\theta\,|\,y)\,d\theta =$$
$$= \int p(\tilde{y}\,|\,\theta)\,p(\theta\,|\,y)\,d\theta. \qquad (4)$$

The ultimate expression in (4) reflects a posteriori distribution forecast as averaged forecasts for using posterior distribution parameter vector $\theta$. This last expression can be written because $y$, $\tilde{y}$ in this model is relatively independent on a certain $\theta$.

All statistical methods that use probability, contain elements of subjectivity in the sense that they are based on a mathematical idealization of the world, i.e. the processes and facilities are described mathematically. Sometimes they say that Bayesian methods are particularly subjective because they require a priori knowledge of distributions. However, in most cases scientific data analysis statements (assumptions) are necessary to determine "likelihood" and "a priori" component model. For example, in linear regression analysis must be done regarding a priori assumptions about distribution of the regression parameters. Here works known general principle: if it is possible to repeat the experiment in order to obtain additional observations, the distribution characteristics can be estimated using the data what increases objectivity of the analysis. Thus, if it is possible to repeat the experiment, the parameters of prior distribution can be completely objectively determined by experimental data. Of course, some elements of analysis that require the use of scientific judgment, remain. For example, the choice of methods for experiment planning, selection of data for analysis and methods of preliminary processing, the definition of parametric shapes to describe the distributions, the creation of methods validation procedures for implementing analysis and selection and use of appropriate quality criteria.

Consider models with one parameter. A simple example of a statistical model based on Bayesian inference is a model with one parameter. Widely used models of distribution of this type are: binomial, normal, Poisson and exponential. Binomial distribution is a statistical model which describes the experiment with a sample in which each member of the set of possible results takes one of two possible values. This data regarding the permutations condition, that the results do not bind to the future, can be interchanged in the analysis. This model is used to evaluate the order parameter, which represents the proportion of successful results in the sample, or the probability of success in each experiment; the model is as follows:

$$p(y\,|\,\theta) = Bin(y\,|\,n,\theta) = C_n^y\ \theta^y (1-\theta)^{n-y}, \qquad (5)$$

and all probabilities, that are considered in the context of this model are determined by the condition $n$.

**Example 1**. Consider an example of this model application to evaluation of the sex ratio among the newborns. It is known that about two hundred years ago the share of women among infants in Europe was less than 0.5 [9]. Today the proportion of accepted value is 0.485. For setting $\theta$ the binomial models choose the proportion of female infants; an alternative parameter may be related proportion of male to female: $\phi = (1-\theta)/\theta$. Let $y$ is the number of newborn girls in $n$ cases of birth; Assume also that birth cases are conditionally independent for a given value $\theta$, and the probability of the birth of girls is $\theta$ for all cases.

In order to form a conclusion on Bayesian binomial model it is necessary to determine a priori distribution for $\theta$. A simple version of this assumption is a little information even for distribution in the range $[0, 1]$. Approximate posterior distribution find as a result of Bayes Theorem application to (5):

$$p(\theta \mid y) \propto \theta^{y} (1-\theta)^{n-y}. \quad (6)$$

If fixed values $n$ and $y$ multiplier $C_n^y$ does not depend on unknown parameter $\theta$, so it can be considered as a constant when calculating a posteriori distribution. This form of approximate posterior density is called beta distribution:

$$\{\theta \mid y\} \sim \text{Beta}(y+1,\ n-y+1). \quad (7)$$

The posterior distribution is considered as a compromise between a priori distribution and data displayed following expressions [8, 9]:

$$E(\theta) = E[E(\theta \mid y)], \quad (8)$$

$$\text{var}(\theta) = E[\text{var}(\theta \mid y)] + \text{var}[E(\theta \mid y)], \quad (9)$$

resulting from the substitution of $(\theta, y)$ instead $(u, v)$ of the expression:

$$E(u) = E[E(u \mid v)]. \quad (10)$$

The average value of $u$ can be obtained by averaging the equivalent average marginal distribution for $v$. In (10) internal expectation is averaging $u$ given $v$. The posterior distribution contains all the current (existing) information on $\theta$, ideally on a complete posterior distribution for $p(\theta \mid y)$. The advantage of the Bayesian approach is in flexible procedures forming a posteriori conclusions provided by the simulation.

**Models with multiple options.** In most practical problems there is a need of evaluation of several parameters. In such cases, the ultimate goal is to perform Bayesian analysis of marginal distribution for selected parameters. The sequence of actions to achieve this may be as follows: first, we need to get the joint distribution of all unknowns, what is followed by integration of the distribution of the unknowns that are not of direct interest to the goal in order to get the desired marginal distribution. Another way of finding a solution may be the use of simulation to generate joint distribution of values to calculate estimates of unknown parameters. In solving many problems there is no need to assess all the unknown parameters of the model, as far as part of the parameters can not be of interest. A classic example is the scale of random errors in the measurement of variables. Consider the case of the presence of such generalized settings.

Assume that the parameter vector $\theta$ consists of two parts: $\theta = [\theta_1, \theta_2]$, each of which could also be a vector. If the conclusion is applicable, $\theta_1$ and $\theta_2$ can be seen as related options; let the following information is available:

$$\{y \mid \mu, \sigma^2\} \sim N(\mu, \sigma^2).$$

Both parameters $\mu = \theta_1$ and $\sigma^2 = \theta_2$ are unknown, but we are interested in $\mu$. It is necessary to find through observations the conditional distribution $p(\theta_1 \mid y)$. For the joint posterior density distribution it can be written:

$$p(\theta_1, \theta_2 \mid y) \propto p(y \mid \theta_1, \theta_2)\ p(\theta_1, \theta_2),$$

and after averaging over $\theta_2$:

$$p(\theta_1 \mid y) = \int p(\theta_1, \theta_2 \mid y)\, d\theta_2.$$

An alternative representation is as follows:

$$p(\theta_1 \mid y) = \int p(\theta_1 \mid \theta_2, y)\ p(\theta_2 \mid y)\ d\theta_2, \quad (11)$$

which shows that a posteriori distribution $p(\theta_1 \mid y)$ is a mixture of conventional posterior distributions $\theta_2$, $p(\theta_2 \mid y)$. The last one acts as a weighting function with different possible meanings of $\theta_2$. The result depends on the weighting coefficients for a posteriori density $\theta_2$ and thus we get combination of the information contained in the data

and a priori model. Integral (11) is rarely evaluated explicitly, but it reflects an important practical strategy for building models with multiple options. A posteriori distribution can be calculated by marginal and conditional simulation. For example, initially measured $\theta_2$ at its marginal posterior distribution, and then $\theta_1$ – for its conditional posterior distribution at a certain value $\theta_2$.

**Hierarchical models.** In constructing probabilistic models here are many options, among which there are relationships that challenge the necessity for building such joint probability distributions for these parameters that reflect current relationship between them. For example, in the the study of the effectiveness of treatment of specific type of disease and the hospital probability of survival it is logical to expect that the estimates for a sample of hospitals must be in some way related. The model that will ensure the existence of such relationships can be created using a priori distribution in which the values are considered as elements of common sampling distribution. The main point of such applied problems is that observations (the number is variable in the group) make it possible to estimate the distribution parameters values, even if these options are not observed. It is naturally to create models of processes hierarchically, i.e. so that the observations were described by conditions with respect to certain parameters. In turn, these parameters are described in the form of probability in terms of other parameters, called hyperparameters. Such hierarchical representation enables better understand (and description) of the problem with multiple parameters, and it also plays an important role in development of the solution required for computing procedures.

When solving practical problems it is important that models that do not have a hierarchical structure, usually are unsuitable for describing hierarchical data. If the number of parameters in the model is insufficient, these models adequately describe the large amounts of data, and an excessive increase in the number of parameters results in "over-learning" (good mathematical description of the existing data but poor predictive characteristics). In models of hierarchical type we can adjust the number of options to improve the adequacy and, at the same time, the introduction of the relationships between parameters makes it possible to avoid retraining [9].

Consider a simple example of building a hierarchical model with parameter estimation using data of the experiment and a priori distribution obtained on the basis of previous (historical) experiments. We assume that the results of current and past experiments form a random sample of the population. As the practice of modeling shows some settings in different experiments may have the same value. For example, let several vectors of observations are related to various experiments with normal distribution with different mean values but the same variance. In cases when for the values of the parameters for the separation there is no other information other than observations we must assume the symmetry parameter in their prior distribution. In a probabilistic sense the symmetry means interchangeability parameters in their joint distribution, that has to be invariant to permutations of indices. In general, it can be argued that the less information is available about the process, there is more ground for introduction of the concept of mutual exchange and options.

In the simplest form of distribution interchangeable with each setting parameters serves as an independent value of the a priori distribution, characterized by the vector of parameters that can be written:

$$p(\theta|\phi) = \prod_{j=1}^{J} p(\theta_j|\phi)$$

As generally known, vector $\phi$ in the distribution for $\theta$ has to take into account the uncertainties in $\phi$:

$$p(\theta) = \int \left[ \prod_{j=1}^{J} p(\theta_j|\phi) \right] p(\phi)\, d\phi.$$

The key "hierarchical" part of these models is that vector $\phi$ is known, and has its a priori distribution $p(\phi)$. Thus, proper Bayesian posterior distribution is $p(\phi,\theta)$:

$$p(\phi,\theta) = p(\phi)\, p(\theta|\phi),$$

and common posteriori distribution is as follows:

$$p(\phi,\theta|y) \propto p(\phi,\theta)\, p(y|\phi,\theta) = \\ = p(\phi,\theta)\, p(y|\theta).$$

Other simplification is possible because hyperparameters $\phi$ affect only $y$ indirectly through options $\theta$. In order to build joint probability distribution for $(\phi,\theta)$ we need to assign a priori distribution for $\phi$. If there is no information

about these options, then choose the uniform distribution, but you need to be sure that a posteriori distribution can be assessed properly. In any case, you must know the domain of these parameters in order to limit the value estimates.

Computational procedures necessary to build hierarchical models are similar to procedures used in assessing multi models, but they are more complex due to the increased number of parameters. In cases where the distribution $p(\theta|\phi)$ is conjugate to the likelihood $p(y|\theta)$, evaluating a posteriori distribution $p(\theta,\phi|y)$ can be done by combining the analytical and computational methods. However, in practice it is often necessary to assess not conjugated hierarchical models, what leads to some complications with computational procedures. The overall analytical process for determining conditional and marginal distributions can be presented in three steps: (1) add joint posterior density $p(\theta,\phi|y)$ of irregular form as the product of a priori distribution hiperparametrs $p(\theta|\phi)$ conditional distribution parameters $p(\theta|\phi)$ and credibility $p(y|\theta)$; (2) get analytically conditional posterior density for $\theta$ given known hiperparametrs $\phi$; for fixed observation $y$ it will function of $\phi$, i.e. $p(\theta|\phi,y)$; (3) compute estimates of $\phi$ using Bayesian approach, i.e. find marginal posterior distribution $p(\phi|y)$. The first step is done by definition; the second step is quite simply implemented using conjugated models as far as conditional posterior density is a product of conjugated posteriori densities for the component $\theta_j$. The third step can be performed via direct integration of posteriori joint distribution:

$$p(\phi|y)=\int p(\theta,\phi|y)\,d\theta.$$

And in the case of standard models, including normal distribution, marginal a posteriori distribution $\phi$ will be calculated by the formula of conditional probability:

$$p(\phi|y)=\frac{p(\theta,\phi|y)}{p(\theta|\phi,y)}.$$

**Bayesian network.** A popular type of Bayesian model is static and dynamic Bayesian network (BN), the first method of their construction was proposed in the eighties of the last century (another name: Bayesian belief network). They are successfully used to create mathematical description of causal relationships in simple and complex systems then use this description to a probability of forming an inference on chosen variables (states) and / or parameters of the process under study. BN is successfully used in systems of technology and medical diagnostics for pattern recognition, classification and forecasting. The spectrum of possible applications is continuously growing [11, 12, 15].

Formally BN is a probabilistic model in the form of directed acyclic graph (DAG), which arcs reflect explicitly the connections (links) between process variables. The DAG can be constructed using expert information or based on data on the evolution data for the relevant variables. The data is used for the analysis of conditional independence process variables and build tables of conditional probabilities (CPT) for each node variable. Today there is a wide range of methods for constructing the network structure. All BN structure estimation algorithms can be divided into two groups: (1) heuristic search algorithms for estimating the structure probabilistic model using a scoring method for its evaluation; (2) the second group of algorithms is based on analyzing patterns of mutual dependency between variables (or nodes) of a model. In applying the first group of algorithms the search process for the structure estimation continues as long as the scoring function does not change or it changes very slightly from one iteration to the next. In the first case, for the quality of BN structure estimation the following criteria are used: Bayesian scoring function [16], entropy based criterion [17], the functional based on the minimum description length [18]. The algorithms in this group require less computational cost, but the result of their application can be not the best model structure due to the nature of the heuristic search procedures. In the second case, the mutual dependence of nodes is evaluated using the tests for conditional independence [19, 20]. The advantage of search algorithms of this group is the possibility of getting asymptotically correct result, but tests for conditional independence are sometimes unreliable, especially in cases of small volumes of data. The general procedure for constructing BN can be represented as the following steps:

**Step 1.** Reduction of the modeling problem dimension that is performed by the known methods. In general, with increasing of the number of variables and parameters the number of sessions for estimating these variables and parameters increases exponentially; so reducing the number of variables and parameters makes it possible to simplify the problem solving. In addition, it is known that the reduction of model dimension improves the accuracy of estimates of the parameters as finite data sample contains a limited amount of information. To solve the problem of reduction one could use the following methods: the method of principal components (PCM); factor analysis; multidimensional scaling (MDS); teaching methods for nonlinear structures, such as linear local dive and others. Factor analysis, PCM and MDS are based on the use of their own vectors. Thus, PCM calculate the maximum variance linear projection, determined by eigenvectors of covariance matrix dimensions. Factor analysis is based on identifying and modeling the correlation structure of the data,

excluding from consideration of random variation data. PCM is more commonly used for the reduction of measurements (number of variables), and factor analysis is used for identifying structural relationships between variables. MDS calculation method provides projections of small dimensions that best preserve the pairwise distances between the values of measurements. Teaching methods in nonlinear structures (configurations) apply to certain types of data of high dimensionality (e.g., pattern recognition), which can form explicit material nonlinearity. Generally, the use of the PCM, factor analysis or multidimensional scaling to such structures gives a positive result regarding the transformation of data in order to bring them to the form required for correct parameter estimation models.

**Step 2.** Most of the known structure and parameters estimation algorithms of probabilistic models as well as the opinions based on them are based on discrete data. Therefore, this step can be performed zooming distribution of data for the purpose of bringing it to easy to use forms and sampling of continuous variables. The "custom" distributions, such as distributions with a strong asymmetry, are scaled by logarithm conversion or by the square root approximation to the distribution of known forms. It is obvious that this lost original scope of data that must be considered in interpreting the results of further evaluation the structures and parameters of models being constructed. For sample data there are developed several effective schemes that ensure sustainable compliance in the sampling intervals (e.g., intervals of equal width or identical frequencies falling values).

The size of data may impose limits on the number of intervals and the number of parameters to be estimated. Obviously, the sample must be available in each interval. Number of intervals is preferably limited because it enables to reduce the number of estimated parameters. Thus, if the network consists of ten binary variables and each variable has on average three parent nodes, then the network parameters necessary to evaluate about $10 \cdot 2^3 = 80$ parameters. If the network consists of ten ternary variables (and each variable has three states), then it is necessary to evaluate $10 \cdot 3^3 = 270$ parameters.

On the other hand, the reduction of the data dimension leads to reducing of their resolution (precision of measurements and presentation of expert estimates), what reduces the accuracy of input and output data related to their assessments of probabilities. The granularity (the depth) of any analysis is determined by the number of options for available events and completeness of relevant data. Hence, for in-depth situational analysis of processes and objects of arbitrary nature we must have large data sets that provide high precision measurements of inputs and outputs. Most algorithms for evaluating of the structure and parameters of network models give better results in the absence of missing values, i.e. there are no intervals with missing values. The lack of gaps makes it possible to apply a relatively simple parameter estimation method of maximum likelihood, and it is not difficult to implement the method of expectation maximizing.

**Step 3.** Formulation of semantic constraints. In the process of finding the best network structure we should fomulate context-focused semantic restrictions that limit the nuber of search structures. This recommendation applies to procedures for any type (i.e., full and partial search). Since the dimension of search space grows exponentially as the number of model variables is growing, the exhaustive search is actually impossible. Thus, for the three variables space the search of network structures is limited by 25 directed acyclic graphs (11 Markov equivalent classes), and for 10 variables that number increases to $3 \cdot 10^{17}$ graphs (or $1 \cdot 10^{17}$ Markov equivalent classes). The semantic restrictions make it possible to reduce the search space to only those network structures that are consistent with the precedence of time requirements or other requirements of dependence between the variables. The limited search space automatically reduces the time required for data processing.

An acceptable basis for formulating semantic constraints provides the basic causal theory of structures and assessments of experienced experts. In the process of the model building we should at least consider the precedents of time interactions between the variables and not make any significant shift in the process of finding the model structure. The semantic restrictions contribute to reducing the number of structures that can be implemented, and increase the probability of constructing a rational structure. The rational use of knowledge of specific subject area (especially when simulating large-scale systems) allows to reduce significantly the number of possible combinations of units without reducing the quality of the inference that is based on the constructed model [21].

**Step 4.** Search of the structures for candidate models. In this step, among the set of possible model structures it is necessary to choose the best models out of several candidate using appropriate quality criteria and assess their options. The search is performed with heuristic algorithms using scoring functions (SF), which enable evaluation of the constructed graphic models. The result of using each combination in the form of a scoring function, structure, search algorithm and corresponding data sample is a candidate model that specifies the network structure. Thus, the task of evaluating the optimal structure is due to the combination of a scoring function and heuristic algorithm. The aim of solving this optimization problem is in evaluation of directed acyclic graph $\mathbf{G}$ structure in the space of acceptable structures $\mathbf{\Omega}^G$, which minimizes the value of scoring functions and meets the characteristics of training data $\mathbf{D}$.

As the scoring function the posteriori probability distribution can be hired in the form:

$$P(\mathbf{G}|\mathbf{D},\Theta) \propto P(\mathbf{D}|\mathbf{G})\cdot P(\mathbf{G}),$$

but the exact calculation of this function even for networks of small dimension requires significant computational cost. Therefore, when evaluating distribution $P(\mathbf{G}|\mathbf{D})$ it is recommended make simplifications, for example, on the type of distribution. Thus, in [22] the algorithm K2 for the implementation of which is considered a priori distribution for uniform $P(\mathbf{G})$ and marginal plausibility $P(\mathbf{D}|\mathbf{G})$ are calculated using conjugate Dirichlet distribution for the network settings. The procedure of K2 algorithm is based on "greedy" search of local extremum and reordering of the network structure such that each variable $X_i$ is added to the parent node that most affects the increase in value of a scoring function. This procedure is repeated for each variable $X_i$ as long as the increase in value will not stop scoring function or variable number of parameters exceeds preset threshold.

A simple approximation of posteriori probability distribution networks provide other scoring functions, including Bayesian information criterion (BIC) that is a marginal likelihood estimation model for large samples. It should be noted that to obtain acceptable approximation quality does not require large sample; in addition, in this case we do not need to specify a priori distribution of the parameters.

The scoring info-geo function is modification of Bayesian information criterion, which is as follows [22]:

$$Ig = -\log P(\mathbf{D}|\hat{\Theta}) + \frac{|\Theta|}{2}\log\frac{N}{2\pi} + \log\int(\det\mathbf{I}(\Theta))^{1/2}\,d\Theta,$$

where the first term represents the likelihood logarithm using estimates $\hat{\Theta}$ obtained by the method of maximum likelihood; the second term is a measure of the complexity of the model, which is determined by the number of parameters; the last term, which includes Fisher information matrix $\mathbf{I}(\Theta)$ determinants is interpreted as a measure of "geometric" complexity. The first two members of the info-geo function correspond to BIC with negative sign.

A well-known criterion of learning is minimum description length (MDL). According to Shannon coding theory, if distribution $P(X)$ of random variable $X$ is known, then optimal length of the code for transferring the value of $x$ through the communication channel is determined by the expression: $L(x) = -\log P(x)$. The source entropy $S(P) = -\sum_x P(x)\cdot\log P(x)$ is a minimum expected length of encoded message. Any other code that is based on the incorrect representation of source messages leads to greater message length. In other words, the better is source model, the more compact can be encoded data.

In the problem of training network by a source of information is some unknown distribution function $P(D|h_0)$, where $D = \{d_1,...,d_N\}$ is training data; $h$ is probabilistic hypothesis concerning the nature of the data. If we consider an empirical risk function, $L(D|h) = -\log P(D|h)$ which is proportional to the empirical evaluation of the distribution error, the difference between the model $P(D|h_0)$ and the distribution is definded by the Kullback-Leibler measure $P(D|h)$ as follows:

$$\left|P(D|h) - P(D|h_0)\right| = \sum_D P(D|h_0)\cdot\log\frac{P(D|h_0)}{P(D|h)} = \sum_D P(D|h_0)\cdot\left|L(D|h) - L(D|h_0)\right| \geq 0.$$

Thus, the measure is the difference between the expected length of encoding (by hypothesis) and the minimum possible. This difference is always non-negative and it equals to zero only when full convergence of the two distributions is observed. The MDL principle generally means that from plurality of models we have to choose the one that makes it possible to describe data with maximum compactness and without loss of information.

To find the global optimum we could apply genetic algorithm or search methods of the Markov chain Monte Carlo type [8, 10]. Thus, when using each annealing algorithm the simulating network structure is interpreted as a state of Markov chain. At each step of the search the algorithm makes a disturbance for network to move from one state of Markov chain to another. This disturbance for a network is implemented using three following operations: adding of an arc,

withdrawal of arc or changing arc direction for the opposite. These operations make it possible to create a set of potential network structures from which one is randomly selected for further study by selected scoring function (SF). Thus, the search algorithm selects the network with improved values of the scoring functions for further processing and removes from further consideration the networks with smaller values of SF some using some finite probabilities. This narrows the search space by removing acyclic structures and the use of semantic constraints. This algorithm requires more computational cost than the "greedy" search algorithm, but it is characterized by a high probability of convergence to global maximum.

**Step 5.** At this stage comparing is performed of probabilistic candidate models characteristics to select the best one for description the process under study. To evaluate the quality of this type of models the criteria of forecasting accuracy is applied uusing existing data for testing. If the model is built to solve the classification problem, for assessment of their quality it is possible to calculate an average weighted utility (or value) obtained through their probabilistic forecasts. This approach is used in cases where we can get information on the cost of possible losses due to incorrect classification or utility achieved through proper processing. Quite reasonable metric for comparing the true joint probability distribution (it is always unknown) with its assessment is the Kullback-Leibler distance that can be seen as some standardized quality assessment for constructed models, including Bayesian networks.

It is clear that the main criterion for a model quality is the final result of its application to solve relevant practical problems that prompted its construction. Bayesian networks provide the opportunity to generate probabilistic inference using several different methods and to get, therefore, alternative results, from which one can select the best one. As far as the quality of a model operation is affected by the quality of relevant statistical data and expert estimates, it is necessary to properly prepare the data according to requirements of estimation theory.

BN construction in the presence of hidden nodes (variables). The method of calculating the parameters of BN was proposed that is based on the expectation maximization algorithm (EM algorithm) in conditions of incomplete input information and known network topology. The resulting method describes the whole process of finding unknown parameters and consists of the following steps: (1) BN building using training data or expert estimates; (2) generate data samples with given network structure (it is used when training data is not available); (3) add hidden nodes to the network structure; (4) initialization of unknown parameters of the network; (5) calculation of the network parameters based on data generated using EM algorithm. Schematic representation of the proposed method is shown in Fig. 1.

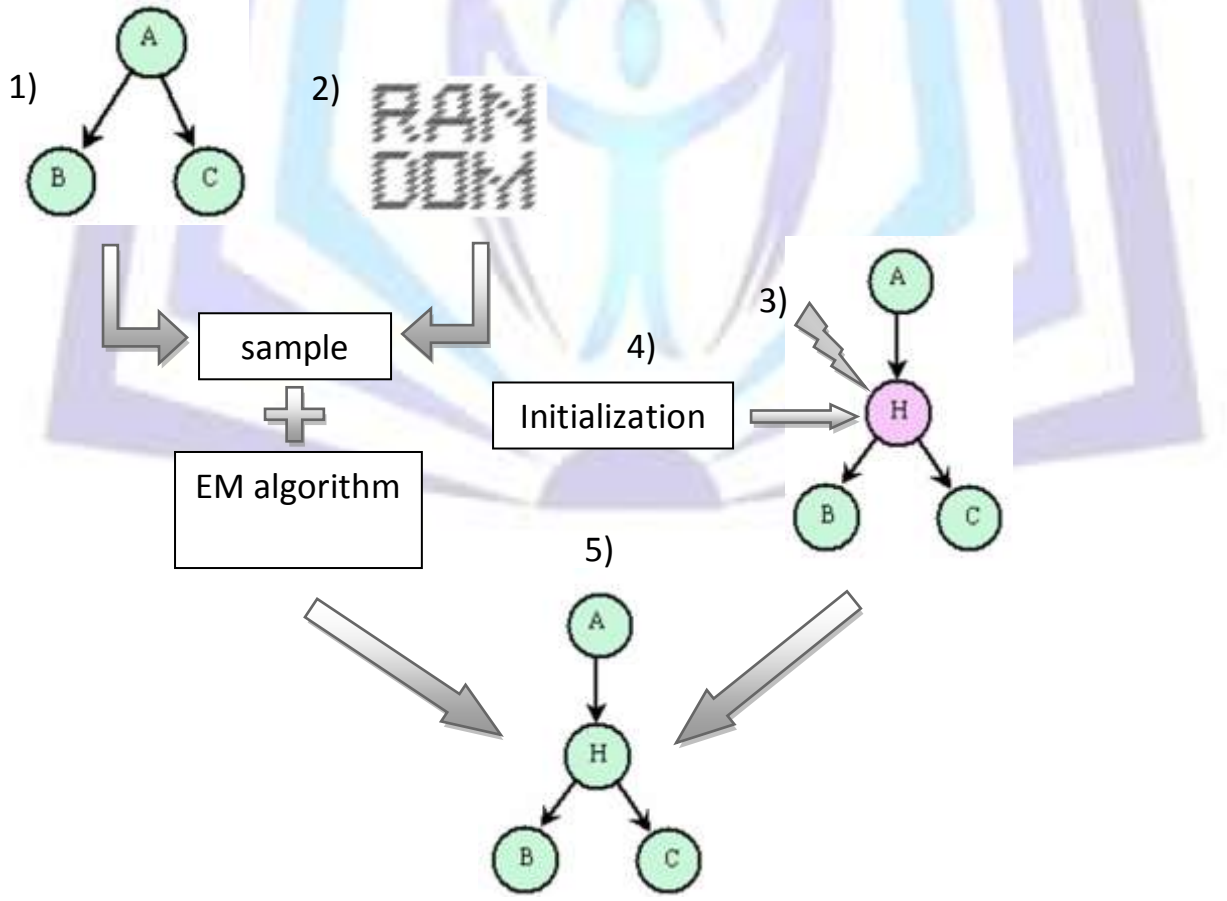


**Fig. 1 Scheme of the technique for BN parameters estimation with hidden nodes**

In the first stage topology of BN is built and the entire network parameters are calculated. At this point there are two options: if we know the network structure it should be moved in the software environment and then the CPTs are to be computed, or when there is only training data. In the second case the BN structure estimation is carried out in two steps: the first step is to built network topology using, say, heuristic algorithm, and the second step is to find the network parameters that are most suitable for the training data.

In the second stage when the training data is not given or it is not sufficient it is necessary to generate pseudorandom samples using the network structure constructed during the first phase. The generating is performed as follows. First, a probabilistic inference is calculated using the network without instantiated nodes $P(S_{ij}),\ i=1,\ldots,N;\ j=1,\ldots,S$.

Then we select the node, $N_{i^*}$, and instantiate selected one of its states, $S_{i^*j^*}$, with probability of this state $P(S_{i^*j^*})$. Then the node probabilities are computed after instantiation: $P(S_{ij} \mid S_{i^*} = S_{i^*j^*})$. The next step is selection of the next node. This operation is repeated until there is no non-instatiated nodes. The instantiated states are stored as a record in the sample: $(S_{i_1}\ldots S_{i_N}),\ i_k \in (1,\ldots,S)$. The algorithm is repeated until we create a required number of records in the sample.

The third step is adding of a hidden node to the network. If the node is inserted between several existing nodes, then previously existing arcs between them are removed and new ones are created that are related to the hidden node. To add hidden parent node, it is created and respective arcs are added. At the next stage the hidden nodes are initialized with initial values. These can be randomly generated values or values provided by experts. At the final stage an iterative process is implemented by the EM algorithm, which uses pre-generated sample of unknown parameters estimates for the hidden nodes of Bayesian network.

**Prediction using BN.** To achieve the possibility of solving the problem of forecasting with BN the static structure of probabilistic graphical model should be modified. One such opportunity is to build the so-called additive BN, which enables to reduce the size of conditional probability tables in the case of simulation of dynamic systems [23]. Additive BN form a basis for building dynamic Bayesian network models (DBN), that enable to calculate and update the values of forecasts with the receiving of new evidences (measurements). The probabilistic inference that is based on DBM is represented in the form of forecast probability distribution based on time series of current observations. Generally BN is represented by the complete joint probability distribution:

$$p(X_1,\ldots,X_n) = \prod_{i=1}^{n} p(X_i \mid \pi(X_i)),$$

where $X_1,\ldots,X_n$ are node variables; $\pi(X_i)$ is a set of parent nodes for variable $X_i$. The probabilistic inference regarding the network is represented by conditional probability of acceptance by selected variables some specific values due to availability of some evidence (information) on the state of the network: $p(\mathbf{X}=\mathbf{x} \mid \mathbf{E}=\mathbf{e})$ That is, $\mathbf{X}$ is any set of nodes that takes the values $\mathbf{x}$ provided that the observed components are taking values $\mathbf{e}$. The degree of probabilistic network complexity (number of nodes and the links between them) can affect the quality of probabilistic inference. Depending on the complexity of a problem the algorithms for exact or approximate inference are used, although problems may arise in both cases. For many practical problems the algorithms are available that provide acceptable accuracy in a reasonable time.

The additive probabilistic network models are generally regarded as a class of separable type models. The idea of separability is that the overall impact of a basic set of variables $X_1,\ldots,X_m$ on dependent variable can be expressed through the influence of individual variables. It is assumed that each independent variable $X_i$ can be in some state (states) $s_i^*$ such that it will not create an impact on $Y$. Thus, conditional probabilities $p(Y \mid X_i, X_{j\neq i} = x_j^*)\ i=1,\ldots,m$, are describing individual effect of each variable $X_i$ on $Y$. In general, the whole set of separable variables $\{X_1,\ldots,X_m\}$ is divided into subsets $\mathbf{X}_i,\ i=1,\ldots,l$, and influence of each on dependent varable is determined separately through the conditional probabilities $p(Y \mid \mathbf{X}_i, \mathbf{X}_{j\neq i} = \mathbf{x}_j^*)$. Now the additive network model can be represented as follows:

$$p(Y=y \mid X_1,...,X_m)=\begin{cases}\sum_{i=1}^{l}\alpha_i\, p(Y=y \mid \mathbf{X}_i, \mathbf{X}_{j\neq i}=x_j^*), \ if \quad y\neq y^*; \\ 1-\sum_{y'\neq y^*} p(Y=y' \mid \mathbf{X}_1,...,\mathbf{X}_l), \ if \quad y=y^*, \end{cases} \tag{12}$$

where $y^*$ are dependent variable states that are evoked by the individual influences of variables $\mathbf{X}_i,\ i=1,...,l$; $\alpha_i \geq 0,\ i=1,...,l$ are model parameters that should satisfy the condition:

$$\sum_{i=1}^{k}\alpha_i\, p(Y \mid X_{i,}\, X_{j\neq i}=x_j^*) \leq 1.$$

Other constraints on these parameters can be determined for specific applications of the models. Similarly to other separable models to determine the conditional probabilities $p(Y \mid X_1,...,X_m)$ for additive models we have to determine only conditional probability due to certain influences. Thus, for BM with binary variables the size of CPTs can be reduced to $2^{m+1}$ to $\sum_{i=1}^{m} 2^{|X_i|+1}$.

The properties of additive models enable us to use for solving the prediction problems BN in combination with regression models. If measurements of independent variables can be represented by the vector $\mathbf{X}(k-i)=\{x_1(k-i),...,x_m(k-i)\}$, then the additive model can be represented as follows:

$$E(y(k) \mid \mathbf{X}(k),...,\mathbf{X}(k-l)) = \sum_{i=0}^{l} f_i(\mathbf{X}(k-i)), \tag{13}$$

where $f_i(\cdot)$ is an arbitrary function. Thus, the equation (13), which reflects the structure of the additive model, is directly related to the equation (12):

$$E(Y \mid X_1,...,X_m) = \sum_{i=1}^{l} f_i(\mathbf{X}_i),$$

where $f_i(\mathbf{X}_i) = \phi_i\, E(Y \mid \mathbf{X}_i,\ \mathbf{X}_{j\neq i}=x_j^*)$.

Representation of BN as additive model makes it possible to move to dynamic network model, which will be used for calculating forecasts. To compute conditional probabilities for this model the additive decomposition of the type described above is used. The main difference of DBN is that the decomposition parameters are computed again after receiving new measurements. In DBN the variable $Y(k)$ depends on the set of variables $\mathbf{X}(k-i)=\{\mathbf{X}_1(k-i),...,\mathbf{X}_m(k-i)\}$, i.e. the vector of measurements independent variables over time. The conditional probability for the variable $Y(k)$ is determined using the additive decomposition of probabilistic models:

$$p(Y(k)=y \mid \mathbf{X}(k),...,\mathbf{X}(k-l)) =$$
$$=\begin{cases}\sum_{i=1}^{l}\alpha_i(k)\, p(Y(k)=y \mid \mathbf{X}(k-i), \mathbf{X}_{j\neq i}(k-j)=x^*(k-j)), \ if \quad y\neq y^*; \\ 1-\sum_{y'\neq y^*} p(Y(k)=y' \mid \mathbf{X}(k),...,\mathbf{X}(k-l), \quad if \quad y=y^*. \end{cases} \tag{14}$$

Equation (14) is similar in structure to the equation (12). In addition to consideration of new measurements in the calculations, it gives an opportunity to update conditional probability by recursively updating values of parameters (weighting coefficients) $\alpha_1(k),...,\alpha_l(k)$.

The final result (inference) based on the model of this type is performed by the generalized probabilistic method of forming inferences presented in [23]. According to this algorithm first an additive decomposition of the total BN on

individual components is performed. An inference for certain subsets of nodes of the basic model is performed by L-S algorithm [21]. For each subset (clicks) units $C$ is calculated joint probability distribution. For this purpose is calculated probability $\prod_{X_i \in C} |X_i|$, where $|X_i|$ is the number of categorical variable values $X_i$. Generally the problem is reduced to generating of a set of sub-networks with weights $\alpha_i$. This way $i-th$ subnet is formed by setting $X_{j \neq i}$ equal to the values $x^*_{j \neq i}$ which were considered above. The algorithm operates recursively as long as the largest dimension subsets for each subnet will not be less than the selected threshold. The resulting tree contains leaf subnets, which inference is formed using probabilistic weights $\alpha_i$.

The method considered was applied to forecasting asset price related to a pre-specified level, and the results were compared with logistic regression. The problems of this type arise in stock assets trading. The characteristics of forecast quality obtained by the methods used are shown in Table. 1 (the bottom three lines are describing the results of dynamic network models).

**Table 1 The characteristics of forecast quality obtained by the methods used**

| The value of threshold *c* | The best model | The threshold probability | Number of correspondences for trend forecast (with probability *p*) |
|---|---|---|---|
| 0,0075 | LR (BS) + MR | 0,47 | 0,869 |
| 0,0065 | LR (FS) + MR | 0,5 | 0,861 |
| 0,0060 | LR (BS) + MR | 0,5 | 0,846 |
| 0,0055 | LR(CHAID) | 0,45 | 0,832 |
| 0,0050 | LR (FS) + MR | 0,52 | 0,831 |
| 0,0045 | LR (BS) + MR | 0,52 | 0,828 |
| 0,0040 | LR (BS) + MR | 0,43 | 0,826 |
| 0,0035 | LR (BS) + MR | 0,49 | 0,822 |
| 0,0010 | LR (FS) + MR | 0,34 | 0,732 |
| 0,0005 | LR (FS) + MR | 0,4 | 0,710 |
| -0,0020 | LR (BS) + MR | 0,43 | 0,677 |
| -0,0025 | LR (BS) + MR | 0,47 | 0,699 |
| 0,0075 | DNM-3 | 0,52 | 0,729 |
| 0,0075 | DNM-3 + FK | 0,52 | 0,837 |
| 0,0075 | DNM-5 +FK | 0,52 | 0,871 |

Abbreviations in the table: LR – logistic regression; MR – multiple regression; TS – tree solutions; DBN – dynamic network model; KF – Kalman filter; FS – forward selection for regression model; BS – backward selection; CHAID – CHi-squared Automatic Interaction Detector.

The forecasting results obtained by the dynamic network model, are compared to the results obtained with logistic regression in combination with multiple regression:

$$g_{\min}(x_1) = \frac{e^{x_1(k)}}{1+e^{x_1(k)}},$$

$$x_1(k) = -0{,}626 - 0{,}424 \cdot \hat{S}2(k) - 0{,}616 \cdot \hat{P}(k) - 0{,}81 \cdot \hat{R}2(k) + \\ + 0{,}773 \cdot \hat{R}3(k) + 1{,}739 \cdot yf(k)'$$

Where $\hat{S}2(k), \hat{P}(k), \hat{R}2(k), \hat{R}3(k)$ are technical analysis indicators; $yf(k)$ is output variable of multiple regression model that takes the value 1 if the forecasted price increases and 0 if the forecasted price decreases. So the best forecasting result was achieved with the logistic regression and the method of independent variables selection by the Backward Selection procedure. The variables are in the right hand side together with the values of the stock indicators,

and the forecast was made by multiple regression ( $p=0,869$ ). The best result of forecasting when using dynamic network model was obtained by the depth of memory 5 and using linear Kalman filter to smooth the data ( $p=0,871$ ). It should be noted that the computational cost in the latter case was significantly higher than in the case of logistic regression. Also it was found that quality of forecasting depends on the threshold $C$ against which the prediction is performed.

## CONCLUSIONS

An overview of some Bayesian models to analyze the data aiming to determine the possibility of their use for predicting development of processes of arbitrary nature was presented. The methodology is proposed for constructing probabilistic models in the form of Bayesian networks that is based based on statistical data and expert estimates. The technique provides correct building of probabilistic models as directed graphs that displays existing causal relationships in data. An integrated probabilistic/regression approach to forecasting was proposed, which is based on a combination of regression equations and probabilistic model. It is distinguished from existing models by the possibility of performing the multistep forecasting. The forecasting results obtained by the dynamic network model are compared with the results obtained with the logistic regression in combination with multiple regression model.

In predicting the price of the selected stock asset the best results were obtained with logistic regression and the method of selection of independent variables for multiple regression by the Backward Selection procedure ( $p=0,869$ ). The best results of forecasting using dynamic network model obtained by using the depth of memory equal to five sampling periods and the linear Kalman filter to smooth the data ( $p=0,871$ ). It should be noted that the computational costs in the latter case were significantly higher than in the case of logistic regression. Also it was found that quality of this forecasting depends on the threshold $C$ value against which the prediction was performed.

In the future work it is necessary to investigate the possibility of obtaining high-quality forecasting results using multivariate hierarchical models and hybrid dynamic network models combined in the frames of specialized decision support system.

## REFERENCES


1. Zgurowskyj M.Z., Pankratova N.D. System Analysis. – Kyiv: Naukova Dumka, 2011. – 900 p.
2. Zgurowskyj M.Z., Podladchikov V.M. Analytical Methods in Kalman Filtering. – Kyiv: Naukova Dumka, 1995. – 285 p.
3. Haykin S. Kalman filtering and neural networks. – New York: John Wiley & Sons, Inc., 2001. – 284 p.
4. Rao M.J.M. Filtering and control of macroeconomic systems. – Amsterdam: North-Holland, 1987. – 280 p.
5. Zgurowskyj M.Z., Bidyuk P.I. Analysis and Control of Large Space Structures. – Kyiv: Naukova Dumka, 1997. – 451 p.
6. Rassel S., Norvig P. Artificial Intelligence. – Moscow: Williams, 2006. – 1407 p.
7. Zaichenko Yu.P. Fuzzy Models and Methods in Intelligent Systems. – Kyiv: Slovo, 2008. – 344 p.
8. Bernardo J.M., Smith A.F.M. Bayesian theory. – New York: John Wiley & Sons, Inc., 2001. – 586 p.
9. Gelman A., Carlin J.B., Stern H.S., Rubin D.B. Bayesian data analysis. – New York: Chapman and Hall/CRC, 2004. – 670 p.
10. Rossi P.E., Allenby G.M., McCulloch R. Bayesian statistics and marketing. – New Jersey: John Wiley & Sons, Ltd, 2005. – 348 p.
11. Cowell R.G., Dawid A.P., Lauritzen S.L., Spiegelhalter D.J. Probabilistic networks and expert systems. – New York: Springer, 1999. – 323 p.
12. Zgurovsky M.Z., Bidyuk P.I., Terentyev O.M. Method of constructing Bayesian networks based on scoring functions // Cybernetics and System Analysis, 2008, Vol. 44, No.2, pp. 219-224.
13. Holsapple C.W., Winston A.B. Decision Support Systems. – Saint Paul (Minnesota): West Publishing Company, 1996. – 850 p.
14. Turban E., Aronson J.E. Decision Support Systems. – New Jersey: Prentice Hall, 2001. – 865 p.
15. Bidyuk P.I., Gozhyj O.P., Korshevnyuk L.O. Computer based DSS. – Mykolaiv: The Black Sea State University, 2011. – 380 c.
16. Cooper G.F., Herskovits E.H. A Bayesian method for the induction probabilistic networks from data // Machine Learning, 1992, № 9, pp. 309 – 347.
17. Herskovits E.H. Computer-based probabilistic network construction / Doctoral dissertation, Stanford University, Stanford, CA, 1991. – 225 p.
18. Bouckaert R.R. Probabilistic network construction using the MDL principle / Technical report TR-RUU-CS-94-27, Utrecht University, 1994. – 26 p.

19. Wermuth N., Lauritzen S. Graphical and recursive models for contingency tables // Biometrica, 1983, vol. 72, pp. 537 – 552.
20. Fung R.M., Crawford S.L. Constructor: a system for the induction of probabilistic models / Boston, MA: MIT Press, Proc. AAAI, pp. 762 – 769.
21. Lauritzen S., Spiegelhalter D. Local computations with probabilities on graphical structures and their application to expert systems // Journal of the Royal Statistical Society, Series B, 1988, vol. 50, No. 2, pp. 157 – 224.
22. Rissanen J. Fisher information and stochastic complexity // IEEE transactions on Information Theory, 1996, vol. 42, pp. 40 – 47.
23. Cooper G. The computational complexity of probabilistic inference using Bayesian belief networks / Artificial Intelligence, 1990, vol. 42, pp. 393 – 405.