IEEE UKRAINE SECTION

IASA

# *DYNAMIC PROCESSES FORECASTING AND RISK ESTIMATION UNDER UNCERTAINTY USING DECISION SUPPORT SYSTEMS*

Bidyuk Petro (Dr. of Science, Professor)
Prosyankina-Zharova Tatyana (PhD)
Terentiev Oleksandr (PhD)

*Institute for Applied System Analysis at the Igor Sikorsky Kyiv Polytechnic Institute*

**Kyiv-2017**

# Introduction. Definition of modern decision support systems

$$DSS = \{DKB, PDP, ST, MSE, MPE, RGP, \\ DQ, MQ, REQ, AQ\}$$

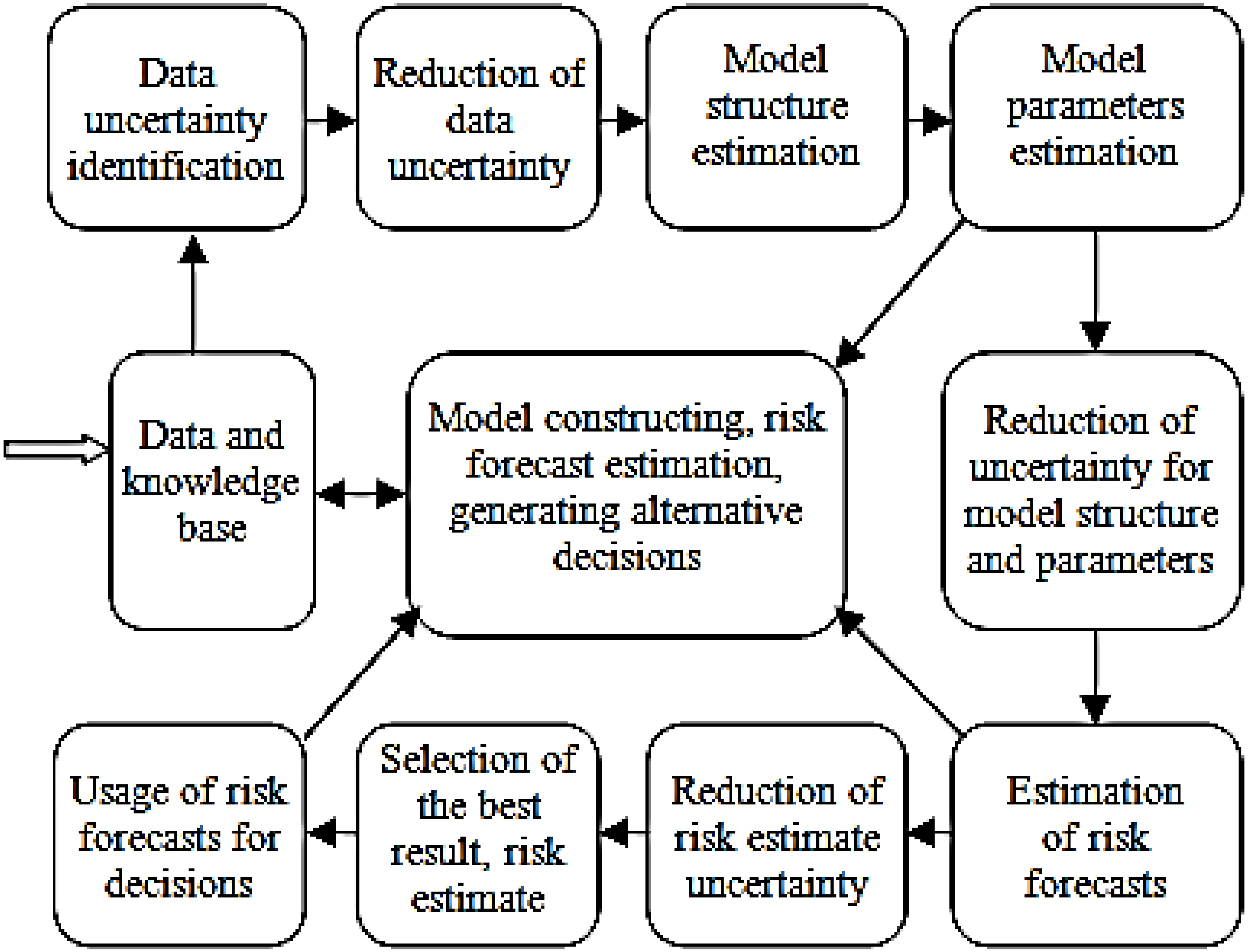
where

- *DKB* is data and knowledge base;
- *PDP* is a set of procedures for preliminary data processing;
- *ST* is a set of statistical tests for determining possible effects contained in data;
- *MSE* is a set of procedures for estimation of mathematical model structure;
- *MPE* is a set of procedures for estimation of mathematical model parameters;
- *RGP* are generating procedures for the risk estimates;
- *DQ, MQ, REQ*, *AQ* are the sets of statistical quality criteria for estimating quality of data, models, risk estimates, and decision alternatives, accordingly.

# Uncertainties in modeling and evaluation of forecasts

| № | Type of uncertainty | Causes | Methods |
|---|---|---|---|
| 1 | *Structural* | • *the impossibility of identification of causal relationships between variables;*<br>• *approximate values of the elements of the model structure;* | • *expert methods;*<br>• *description of revealed nonlinearities alternative using analytical processes with further assessment of the adequacy and quality model forecasts;*<br>• *application of statistical tests;*<br>• *applying the theory of hypothesis testing;* |
| 2 | *Statistical* | • *measurement error;*<br>• *stochastic external disturbance;*<br>• *multicollinearity;*<br>• *extreme values;*<br>• *missing measurements.* | • *digital and optimal filters (Kalman filter);*<br>• *clarifying the types of distributions;*<br>• *the principal components method;*<br>• *theory of extreme values;*<br>• *methods imputation missing values*<br>• *simple averaging* |
| 3 | *Parametric* | • *the incorrect method for assessment measurement;*<br>• *short samples.* | • *providing alternative methods for parameter estimation;*<br>• *the samples replication;*<br>• *Monte Carlo Markov chains* |
| 4 | *Probabilistic* | • *complex mechanisms of causality;*<br>• *lack of determinism;* | • *statical and dynamic Bayesian network;*<br>• *Markov model;*<br>• *probabilistic filters;*<br>• *conditional distributions;* |
| 5 | *Type amplitude* | • *availability of variables which is not measured;* | • *methods of processing fuzzy information.* |
| 6 | *Situational* | • *Failure to identify all causal relationships* | • *expert methods*<br>• *cognitive modeling*<br>• *SWOT, PEST, SPACE analysis* |

# Actions directed towards uncertainties identification, processing and taking them into consideration



Data uncertainty identification

Reduction of data uncertainty

Model structure estimation

Model parameters estimation

Data and knowledge base

Model constructing, risk forecast estimation, generating alternative decisions

Reduction of uncertainty for model structure and parameters

Usage of risk forecasts for decisions

Selection of the best result, risk estimate

Reduction of risk estimate uncertainty

Estimation of risk forecasts

# Adaptive estimation scheme for building “best” model

$$V_N(\theta, D_N) = e^{|1-R^2|} + e^{|2-DW|} + \alpha \ln(1 + \frac{SSE}{N}) + \beta \{\ln(1 + MSE) +$$

$$+ \ln(1 + MAPE)\}$$

where

- $\theta$ is a vector of model parameters;
- $D_N$ data in the form of time series (N is a power of time series used);
- R2 is a determination coefficient;
- DW is Durbin-Watson statistic;
- MSE is mean square error;
- MAPE is mean absolute percentage error for forecasts;
- $\alpha$, $\beta$ are adjustment coefficients that could be selected by a DSS user or searched for automatically by the decision support system itself.

# Estimation of system states via Kalman filter (1)

$$x(k) = A(k,k-1)x(k-1) + B(k,k-1)u(k-1) + w(k),$$

where

- x(k) is n-dimensional vector of system states;
- k=0,1,2, … is discrete time;
- u(k-1) is m-dimensional vector of deterministic control variables;
- w(k) is n-dimensional vector of external random disturbances;
- A(k, k-1) is (n×n) matrix of system dynamics;
- B(k, k-1) is (n×m) matrix of control coefficients.

# Estimation of system states via Kalman filter (2)

The measurement equation for vector *z*(k) of output variables is described by the following equation:

$$z(k) = H(k)x(k) + v(k)$$

where

- H(k) is (r×n) observation (coefficients) matrix;
- v(k) is r-dimensional vector of measurement noise with covariance matrix R.

# Estimation of system states via Kalman filter (3)

Another choice is in constructing separate algorithm for computing the values of  and . A convenient statistical algorithm for estimating the covariance matrices includes the following steps

$$\hat{R} = \frac{1}{2}\left[\hat{B}_1 + A^{-1}(\hat{B}_1 - \hat{B}_2)(A^{-1})^T\right]$$

$$\hat{Q} = \hat{B}_1 - \hat{R} - A\hat{R}A^T$$

$$\hat{B}_1 = E\left\{\left[z(k) - Az(k-1)\right]\left[z(k) - Az(k-1)\right]^T\right\};$$

$$\hat{B}_2 = E\left\{\left[z(k) - A^2 z(k-2)\right]\left[z(k) - A^2 z(k-2)\right]^T\right\};$$

# Estimation of system states via Kalman filter (4)

The matrices  and  are used in the optimal filtering procedure as follows:

$$S(k) = AP(k-1)A^T + \hat{Q}$$

$$\Delta(k) = S(k)\left\lfloor S(k) + \hat{R} \right\rfloor$$

$$P(k) = \left[I - \Delta(k)\right]S(k)$$

where S(k) and P(k) are prior and posterior covariance matrices of estimate errors respectively, the symbol “ # ” denotes pseudo-inverse; $A^T$ means matrix transposition; Δ(k) is a matrix of intermediate covariance results.

# Estimation of system states via Kalman filter (5)

The algorithm was successfully applied to the covariance estimating in many practical applications.

The computation experiments showed that the values of $\Delta(k)$ become stationary after about 20-25 periods of time (sampling periods) in a scalar case, though this figure is growing substantially with the growth of dimensionality of the system under study.

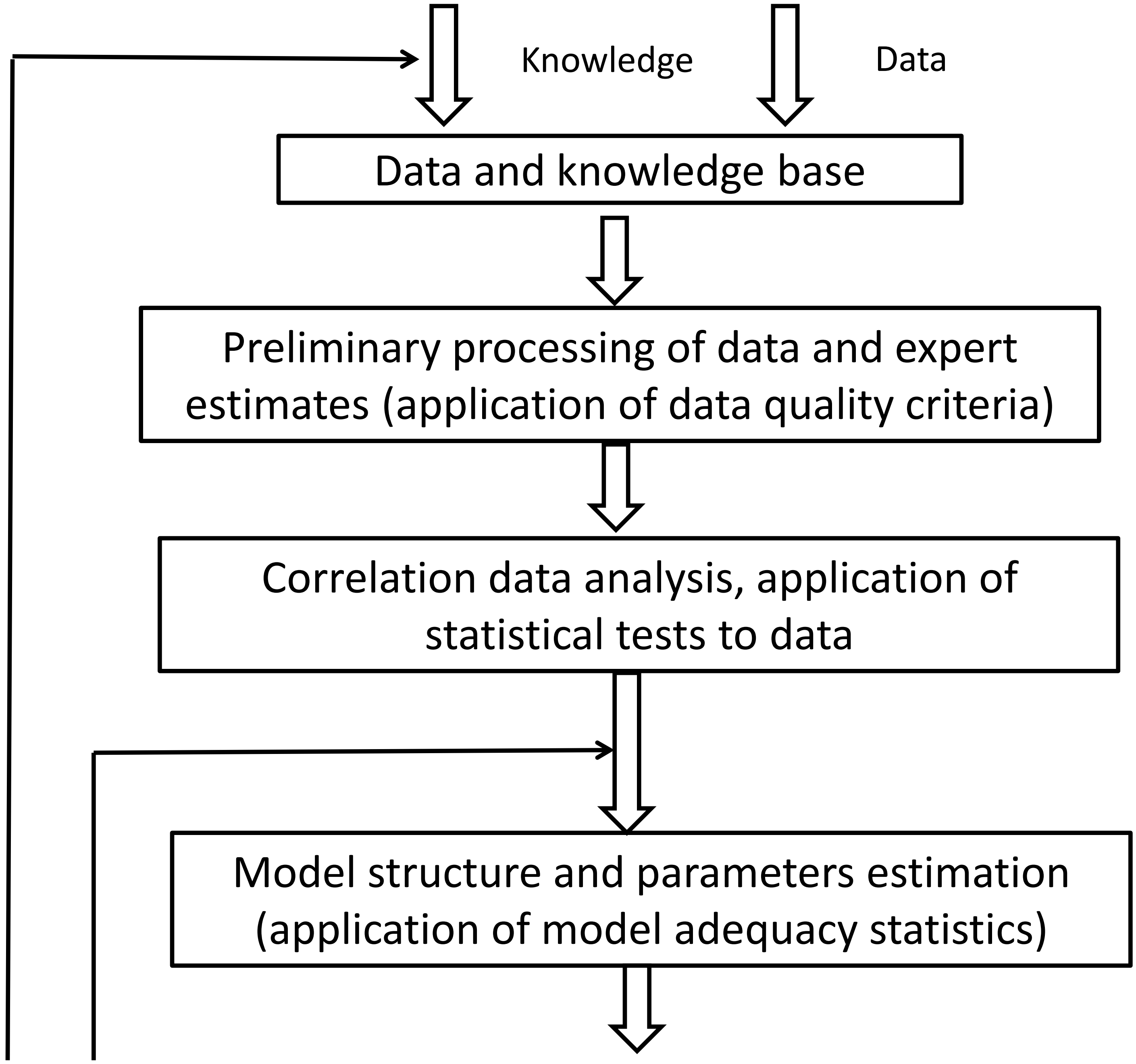
# Unified notion of a model structure

$$S = \{r,\ p, m, n, d, w, l\}$$

where

- r is model dimensionality (number of equations that constitute the model);
- p is model order (maximum order of differential or difference equation in a model);
- m is a number of independent variables in the right hand side of a model;
- n is a nonlinearity and its type;
- d is a lag or output reaction delay time;
- w is stochastic external disturbance and its type;
- l are possible restrictions for the variables and/or parameters.

# The functional layout of the DSS (1)



Knowledge

Data

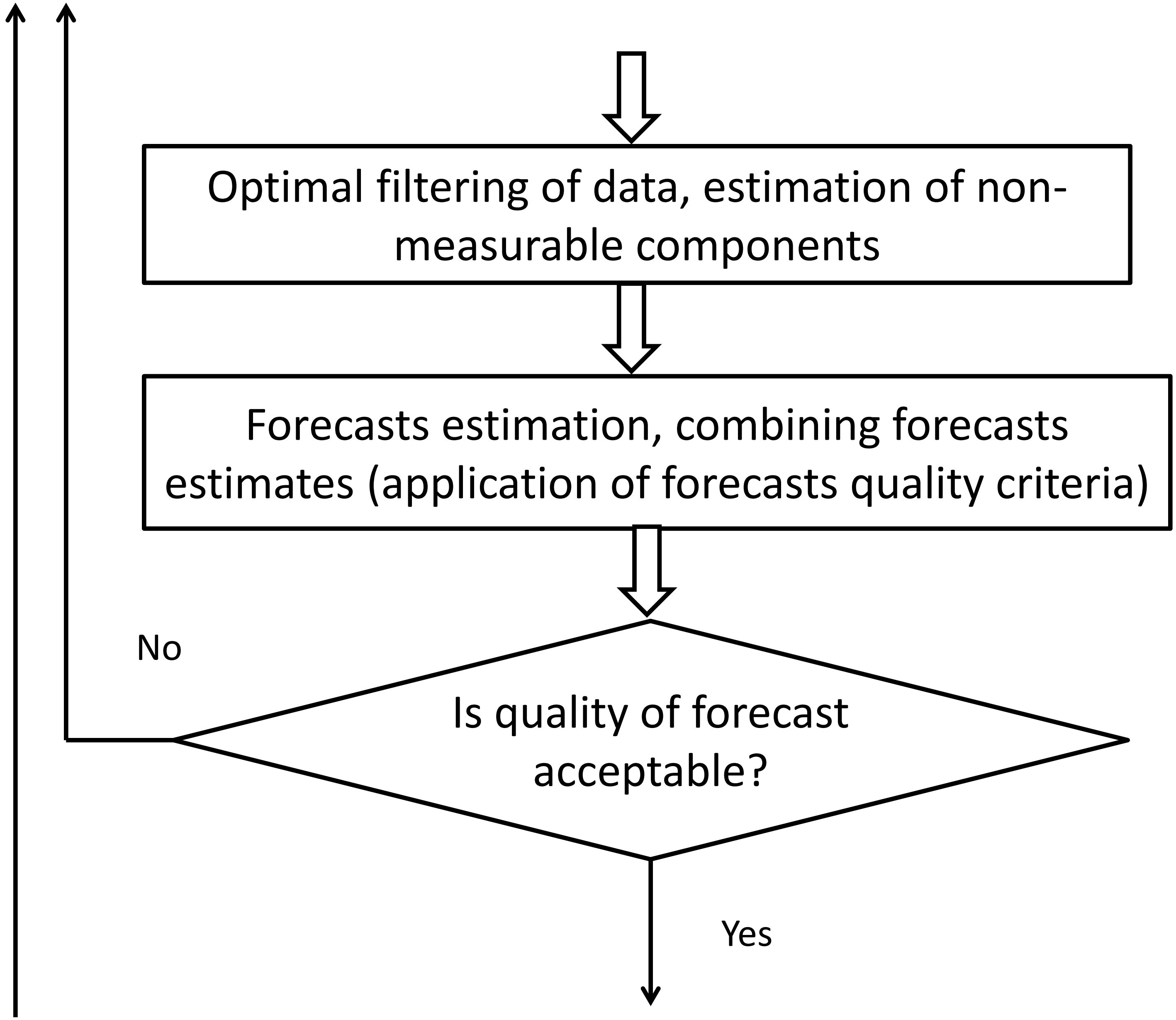
Data and knowledge base

Preliminary processing of data and expert estimates (application of data quality criteria)

Correlation data analysis, application of statistical tests to data

Model structure and parameters estimation (application of model adequacy statistics)

# The functional layout of the DSS (2)



Optimal filtering of data, estimation of non-measurable components

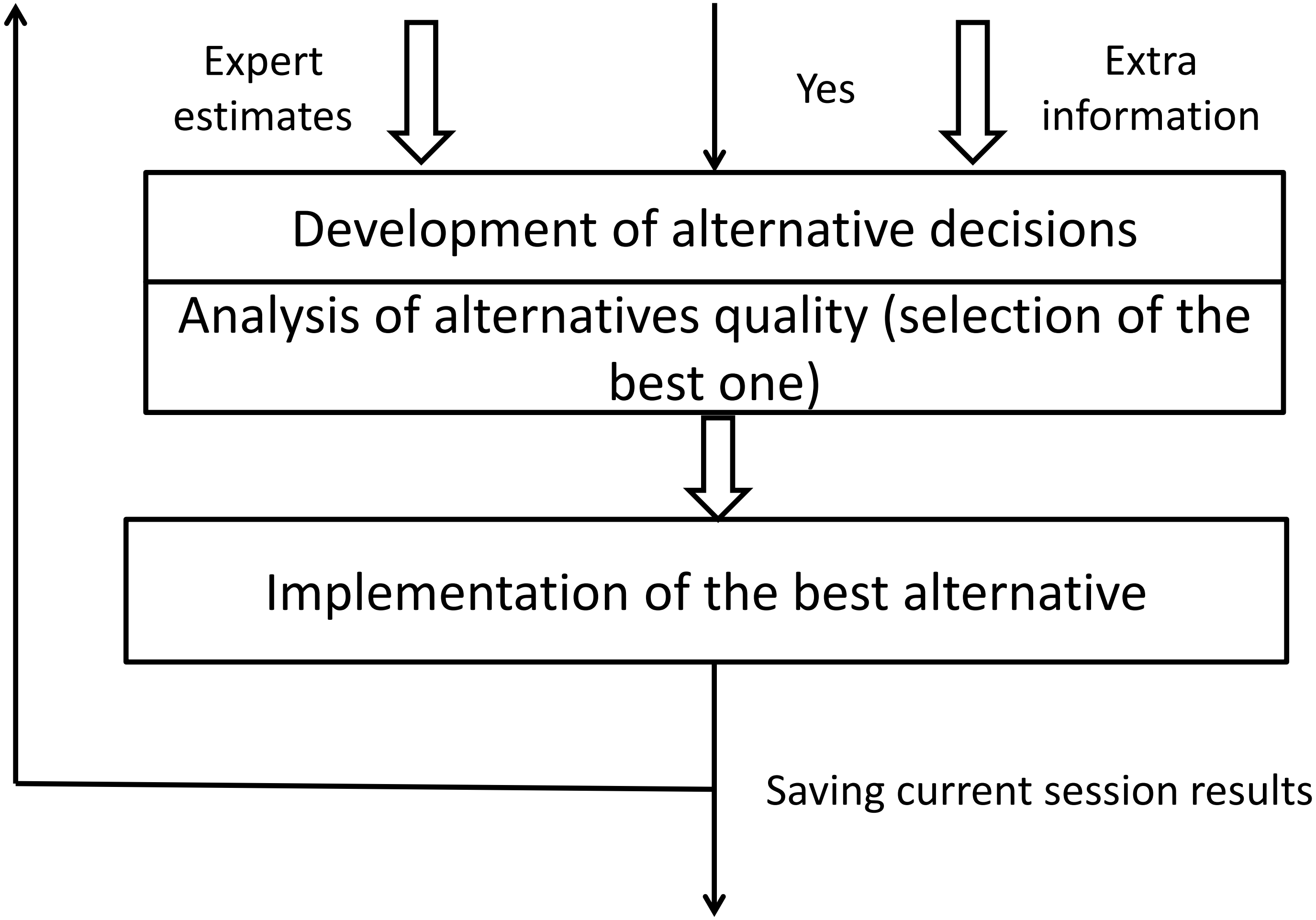
Forecasts estimation, combining forecasts estimates (application of forecasts quality criteria)

Is quality of forecast acceptable?

No

Yes

# The functional layout of the DSS  (3)



Expert estimates

Yes

Extra information

Development of alternative decisions

Analysis of alternatives quality (selection of the best one)

Implementation of the best alternative

Saving current session results

# Imputation techniques for replacing missing values

(1) Imputing constant value – user’s own constant, mean, median, mode, m-estimator value (Tukey's Biweight, Huber, Andrew's Wave) which are calculated from the active training predecessor data set.

(2) Estimate function approach uses for calculation of replacement values as estimation by analyzing each input-parameter as a target. In general sense all existing methods of predictive modeling could be used as an estimate function – decision trees, BN, neural networks, regression models, autoregressive models, exponential smoothing and so on.

(3) Special approaches: optimization techniques (EM-algorithm), cluster analysis (k-means), evolutionary algorithm (genetic algorithms), resampling.

# Imputation techniques for time series

For the data in the time series form the most suitable imputation techniques are:

– simple averaging when it is possible (when only a few values are missing);

– generation of forecast estimates with the model constructed using available measurements;

– generation of missing (lost) estimates from distributions the form and parameters of which are again determined using available part of data;

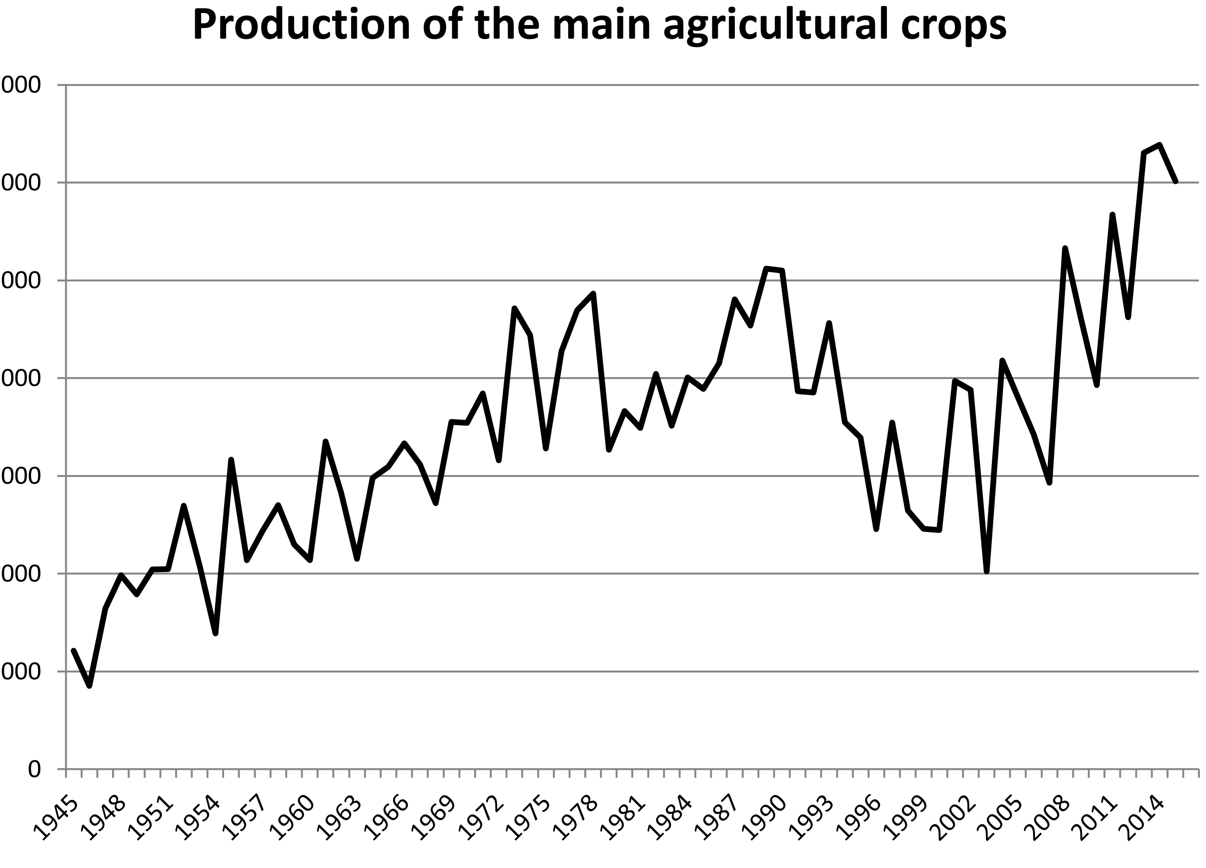
– the use of optimization techniques, say appropriate forms of EM-algorithms (expectation maximization);

– the exponential smoothing etc.

# The Universal Time Series Model

$$Y_t = f(T_t, S_t, X_t, E_t)$$

Where
$T_s$ – trend
$S_t$ – seasonal
$X_t$ – input
$E_t$ – error (irregular)



# Results of computational experiments (1)

The methodology proposed was applied for short term forecasting of production of the main agricultural crops in Ukraine.

The data used were indices of production of the main agricultural crops in the period within 1940-2014. The forecasting quality was tested with the data related to 2012-2014.

About 10% of input data is missing.

| Name of variable | Variable value in thousands of tons | | | RMSE |
|---|---|---|---|---|
| | 2012 | 2013 | 2014 | |
| Actual value | 2008.7 | 2295.3 | 1999.1 | |
| Forecasts: | | | | |
| - no imputation (as it is) | 1787.29 | 1888.14 | 1984.31 | 154.57 |
| - imputation of sample mean value | 1711.83 | 1777.28 | 1834.86 | 206.41 |
| - median value imputation | 1714.68 | 1783.74 | 1844.94 | 203.28 |
| - imputation of previously available data | 1751.87 | 1848.65 | 1938.20 | 172.94 |
| - imputation of the mean for neighboring values | 1741.19 | 1818.34 | 1888.30 | 185.99 |
| - imputation of the model generated values | 1719.67 | 1808.83 | 1889.51 | 192.12 |
| - imputation of averaged previous values | 1769.29 | 1877.80 | 1979.97 | **160.55** |

# Results of computational experiments (2)

For 11 agricultural time series were made following steps:

1. Generating 5, 10, 15, 20, 25, 30, 35, 40, 45, 50 % of missing data in the dataset.

2. for each variable were created ten samples each of which contained missing values with the appropriate portion.

3. Using various methods of imputation were filled missing values.

| | |
|---|---|
| x1 | Production of cereals |
| x2 | Sugar beet production |
| x3 | Sunflower production |
| x4 | Potato production |
| x5 | Vegetable production |
| x6 | Meat production |
| x7 | Milk production |
| x8 | Cattle stock |
| x9 | Cow production |
| x10 | Head of Pigs |
| x11 | Head of sheep and goats |

**Method 1.** Standard linear regression

$X_1$=$a_0$+$a_2$*$X_2$ + … + $a_{11}$*$X_{11}$

**Method 2.** Replacing using the average value by the sample

**Method 3.** For each missing value, the average between the first previous and the first subsequent values is calculated.

**Method 4.** For each missing value, based on the five previous values, the mean is calculated using weighting coefficients according to the principle of exponential smoothing.

For example, X1 contains a pass at time t, then the substitution value will be calculated using the formula

0,5*X1(t-1) + 0,25*X(t-2) + 0,125*X(t-3) + 0,1*X(t-4) + 0,025*X(t-5)

The weight coefficients were set based on the expert experience of using the method of exponential smoothing.

**Method 5.** Using the model of auto regression, based on the assumption that there are 3-year cycles.

| MAPE | Regression | Mean | Average between neighbor values | Exponential smoothing of 5 previous values | AR(p) |
|---|---|---|---|---|---|
| x1 | 15,05 | 32,18 | 13,54 | 19,64 | 20,94 |
| x2 | 19,09 | 69,17 | 15,26 | 23,90 | 30,37 |
| x3 | 32,14 | 78,30 | 11,27 | 19,24 | 22,74 |
| x4 | 11,14 | 13,53 | 8,37 | 12,74 | 12,73 |
| x5 | 10,12 | 28,78 | 6,50 | 10,93 | 12,64 |
| x6 | 8,41 | 46,92 | 6,32 | 14,72 | 12,01 |
| x7 | 5,54 | 35,38 | 4,76 | 9,17 | 9,08 |
| x8 | 6,85 | 66,22 | 5,55 | 12,78 | 10,08 |
| x9 | 4,48 | 42,59 | 3,74 | 8,38 | 7,11 |
| x10 | 11,03 | 54,89 | 10,19 | 16,37 | 21,44 |
| x11 | 22,82 | 97,53 | 7,34 | 17,13 | 15,21 |
| Average MAPE | 12,53 | 45,90 | **8,37** | 14,61 | 15,30 |

The best results show the average between neighbor values.

# General recommendations

| Method | Advantages and disadvantages. |
|---|---|
| Regression | Can be used for any percent of missing data. It is necessary to have regressors with high correlation towards the simulated variable. |
| Mean | Can be used for any percent of missing data. There is no need for explanatory variables. |
| Average between neighbor values | The method cannot be used when the number of missing data is more than 20%, because in this case appear long sequences of omissions. |
| Exponential smoothing of 5 previous values | The method cannot be used when the number of missing data is more than 25%, because in this case appear long sequences of omissions. |
| AR(p) | The method cannot be used when the number of missing data is more than 15%, because in this case appear long sequences of omissions. |

# Results of computational experiments (3)

- The methods proposed for eliminating information and situational uncertainties were used for solving the problem of decreasing the number of bank clients.
- During the process of constructing the scoring models quite acceptable results of forecasting were achieved with expanding the data sets by the values generated for restoring the client applications that were declined at the stage information verification and credit decision making using fuzzy logic and logistic regression. Such approach provides a possibility for increasing the quality of the final scoring model by 2-5% according to the Gini criterion.

# Conclusions

The general methodology was proposed for constructing DSS for mathematical modeling and forecasting of economic and financial processes, and different kinds of risks estimation that is based on the following system analysis principles: hierarchical system structure, taking into consideration of probabilistic and statistical uncertainties, availability of adaptation features, generating of multiple decision alternatives, and tracking of computational processes at all the stages of data processing and model constructing with appropriate sets of statistical quality criteria.

The DSS proposed could be used for support of decision making in various areas of human activities including strategy development for state government, financial institutions, industrial enterprises, agriculture, investment companies etc.

# THE END

Bidyuk Petro
pbidyuke_00@ukr.net

Prosyankina-Zharova Tatyana
t.pruman@gmail.com

Terentiev Oleksandr
o.terentiev@gmail.com

Institute for Applied System Analysis
National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»
Kyiv, Ukraine