## CONTENTS

# PREFACE

Recently modeling is the most effective means of knowledge of the laws and patterns of the world. For modern mathematics is characteristic the intense penetration into other disciplines, particularly in the economic sciences. Economics as a science of the objective laws of social development, constantly uses various quantitative features, and as resulting was accumulated a lot of kinds of mathematical methods and models for scientific researches. Today in economics the mathematical model comes at a leading position as an effective tool for the study and forecasting of economic processes and phenomena

This book is an outgrowth of work under course the Mathematical Methods and Models for Economists (MMME) at Odessa National Economic University. The MMME course has been offered since more than 15 years, and addressed to bachelors, masters, and scientists who are interested to researches of Economics through the modern Mathematical Methods and Models with using IT technology.

Most of the world's university curricula to prepare masters in economics include discipline Micro-, Macroeconomics, Econometrics I, II and III that widely use economic modeling. But all curricula that offered are very complicated (by our opinions so as we have had many years of experience lecture under these courses) and although there are a lot of good books on Mathematical Methods of Modeling and related topics, we felt that many of them have either too high-level or too advanced. Our goal was to describe the main methods that focus on the fundamental algorithms of Mathematical Methods and Models, Data Mining and Forecasting in Economics.

It lays the mathematical foundations for the core Data Mining methods, with key concepts explained when first encountered; the book also tries to build the intuition behind the formulas to aid understanding.

The main parts of the book include Exploratory Data Analysis, Factor Analysis, Clustering, and Classification, Dynamic Optimization Models, Neural Networks. Simulation Modeling.

The book lays the basic foundations of these tasks, and it also covers cutting-edge topics such as Projection on Latent Structure, High-dimensional Data Analysis, and Complex Graphs and Networks. It integrates concepts from related disciplines such as machine learning and Statistics and is also ideal for a course on Data Analysis. Most of the material in the text is grounded on using Linear Algebra, and Probability and Statistics.

The book includes many examples to illustrate the main technical concepts. All offered examples have been used and realized by students in class and some examples have been realized by the doctoral students for their scientific works.

All of the algorithms in the book have been implemented by the authors. We suggest that readers can use their favorite data analysis and mining software to work through our examples and to implement the algorithms we describe in text. We recommend use the package STATISICA software or other packages that you know.

**Chapter 1. Fundamentals of mathematical modeling and forecasting**

1.1. The logic applied economics and mathematical modeling. Economic Data (information base of a model)

Applied economic modeling is a cyclical process that can be divided into the following stages:

1) Description of the purpose and object modeling.

2) Computer exploratory analysis.

3) Mathematical formalization of models according to economic theory.

4) Estimation of model parameters (using specialized software).

5) Verification of model adequacy.

6) Analysis and interpretation of results (if necessary, execute the clarify to mathematical model and go to step 3)).

In the first stage, researcher defines purpose and object modeling. Depending on the purpose of the study the same economic unit or process can be described in various models. The final method of research should be kept in mind throughout the process of simulation!

1) Description of object modeling includes the following points:

- selection of a unit element in the aggregate as a carrier of characteristics the properties of the object; (in terms of explorer)
- Determining spatial and temporal boundaries of the object modeling;
- formation the feature space of model.

2) As the basis of modeling we use certain statistical totality (set of objects, values of characteristics etc). Formally, any set can be represented as an ordered set of data parameters – *N*, where *N*(***n*** = 1, 2, ..., *N*) is an amount of elements in totality; *I*(***i*** = 1, 2, ..., *I*) is the registered number indicators of the ***n***-th element ; *T*(***t*** = 1,2, ..., *T*) is time calendar periods (for economic data this is usually a year, quarter, month, day, etc.).Thus, the unit of information for the object of modeling is the value of the ***i***-th feature of ***n***-th element into totality for ***t***-th period.

If the totality is studied in statics, we have that the information (data) is represented by a matrix of dimension $N \times I$ (spatial data) in case of dynamics, then by a matrix of dimension $N \times I \times T$ (panel data).

Estimations of significance and informativeness of the individual features play the crucial role in forming a feature space but we also need to take into account the belonging to certain types (continuous, discrete, categorial etc.) and the range of variation, laboriousness of collecting the information.

**Note that, the absolutely accurate measurement results do not exist in nature. The actual data always is regarded as the implementation of the unpredictable stochastic process.**

This gives a ground for the probabilistic estimate of the results modeling. The aim of this estimation to establish as far the identified a regularity is deprived of the random influences, and as far it is typical for all complex of conditions in which the functioning object of modeling and this procedure includes the following steps:

- statistical description of the object ( determination of averages, standard deviations, and other characteristics of the distribution);
- unification of types features (often as standardization of data), convert of them to the same species;
- testing of totality on uniformity, identification of abnormal (atypical) observations (so-called emissions) and, if necessary, correcting them;
- restoration (if it is possible) of missing data;
- estimation of the interconnection between the features

3) Building a model is based on the certain rules and algorithms that determine the order of mathematical calculations and actions necessary for processing of information.

At this stage, we find the algebraic form of calculations, the relationship between the properties of the process that are described by symbols and signs, order of calculations through the flowchart

4) Estimation of model parameters it's a stage of computer processing of data. Currently, there is a number of developed software products that provide to users unique opportunities for the experimentation, exploration, graphic mapping and in-depth analysis of data carried out by the modern methods of modeling and forecasting implemented using the latest computer technology (sometimes even in automatic mode).

5) Verification of model adequacy means evaluating of the degree of correspondence of parameters the model to the factual characteristics of the object. At this stage, we use different procedures for comparison of model conclusions testing statistical hypotheses using statistical criteria. **Verification of model adequacy makes sense only for the aim of research and cannot be abstract**.

6) The final stage of modeling is one of the most difficult and the most critical.

The difficulty lies for the interpretation of the results and do not exist ready-made algorithms or recipes. There is a common requirement for all models that interpretation is consistent with the original hypothesis. The main conclusions are stated in terms of content such as the content of the model parameters, accuracy tested hypotheses, evaluating their degree of reliability.

Thus, we can formulate the two basic principles of economic and mathematical modeling:

**1) subordination to the goal of all steps of the modeling;**
**2) the adequacy of the model**.

Again, that only correct, "ideal" model does not exist. The choice of the type of model depends on the purpose of research, the specific process (phenomenon), scale object modeling, the available information, hardware and software.

1.2. Data Analysis with using computer technology. Homogeneity and Typology. Multivarious Ranking. The Methods of expert estimations

The statistical data processing implemented through any statistical package of applied programs (even into the spreadsheets of *MS Excel*). Below we consider examples of exploratory data analysis with software package *STATISTICA*. Its modules generate a large amount of initial information as the tables of data (*Scrollsheet*) and graphs. Table *Scrollsheet* supports all standard operations with blocks of values (copy, move, insert, extrapolation, standardization, etc.) and you can edit them, save as results file (with the .scr file extension) or convert the raw data (with the *.sta* file extension) can be exported to other *Windows* applications. Consider the following analysis of data on the ***n*** = ***15*** sugar mills in the following indicators (features - *Variables*):

*Table 1.1. Data on the work of 15 sugar mills*

| *Variables* | *Name* | *LongName* |
|---|---|---|
| *VAR1* | *Quality* | *Sugar content of beet%* |
| *VAR2* | *Wastage* | *Wastage the feedstock during transportation and storage,%* |
| *VAR3* | *Molasses* | *The sugar content in molasses,%* |
| *VAR4* | Effect | *The output of sugar from 1 tonne of sugar beet%* |

File the primary data of sugar mills RM.sta 4v * 15

*Table 1.2. A data file in the system Statistica*

| | *VAR1 Quality* | *VAR2 Wastage* | *VAR3 Molasses* | *VAR4 Effect* |
|---|---|---|---|---|
| *1* | *15,1* | *0,99* | *2,5* | *9,78* |
| *2* | *15,41* | *1,06* | *2,68* | *9,13* |
| *3* | *15,22* | *0,98* | *2,19* | *10,46* |
| *4* | *15,16* | *0,95* | *2,06* | *10,69* |
| *5* | *15,43* | *1* | *2,05* | *10,58* |
| *6* | *15,41* | *1* | *2,06* | *10,84* |
| *7* | *15,15* | *0,97* | *2,34* | *10,87* |
| *8* | *16,06* | *0,9* | *2,24* | *12,24* |
| *9* | *15,95* | *0,92* | *2,27* | *11,94* |
| *10* | *15,59* | *0,95* | *2,13* | *11,26* |
| *11* | *15,52* | *0,93* | *2,26* | *11,01* |
| *12* | *15,33* | *0,97* | *2* | *11,88* |
| *13* | *15,48* | *0,91* | *2,2* | *11,53* |
| *14* | *15,18* | *0,98* | *2,23* | *11,03* |
| *15* | *15,17* | *0,98* | *2,18* | *10,37* |

For example, we consider the procedure for creating a table *Scrollsheet* by module *Basic Statistics / Tables* as main statistics and tables, which combine methods of exploratory data analysis. We should open source data file (for example, *Table 1.2. A data file in the STATISTICA* ( a totality of the sugar mills )) into the starting panel of the module . The first phase of data analysis procedures *Descriptive Statistics* that use the descriptive statistics. The choice of indicators for the analysis carried out in windows *Select the variables for the analysis*. You can select just four variables: VAR1 - VAR4. In order to analyze the distribution of the full set of these features, we use the option *More Statistics* for advanced set of descriptive statistics. Among them choose: *Mean* as the average value, *Median* as the median, *Lower and Upper Quartiles for* the calculating of lower and upper quartiles, *Standard Deviation* for standard deviation, and for other coefficients (*Skewness* , *Kurtosis)*. After the team's analysis of the procedure, the system creates a spreadsheet *Scrollsheet* with the results of the calculation:

*Table 1.3. Descriptive statistics*

| *DescriptiveStatistics (RM.sta)* | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Continue …* | *Mean* | *Median* | *LowerQuartile* | *UpperQuartile* | *Std. Dev.* | *Skewness* | *Kurtosis* |
| *VAR1* | *15,37* | *15,33* | *15,16* | *15,52* | *0,331* | *0,603* | *0,755* |
| *VAR2* | *0,97* | *0,97* | *0,93* | *0,99* | *0,041* | *0,384* | *0,817* |
| *VAR3* | *2,23* | *2,2* | *2,06* | *2,27* | *0,179* | *1,275* | *1,970* |
| *VAR4* | *10,91* | *10,87* | *10,46* | *11,53* | *0,820* | *–0,398* | *0,391* |

As the data show, the aggregate distribution of the sugar mills is characterized by moderate variation of indicators (the ratio of standard deviation to average is less than 10%) and a notable asymmetry, especially in terms of loss of sugar in the processing of raw materials (VAR3); to exit 1 ton of sugar from raw material is typical left-sided asymmetry. We continue to the exploratory analysis of data to assess the relationship between features and it select on the starting panel module procedure *Correlation matrices* . Further, choosing in the window *Pearson Product* the tab *Moment Correlation* than the type of matrix such as *One variable list* (square matrix) , and for the selection of signs press the button *Select all*. By the command to the procedure we obtain a new table whose elements are coefficients of the pair correlation:

*Table 1.4. Coefficients of pair correlation*

| *Variable* | *VAR1* | *VAR2* | *VAR3* | *VAR4* |
|---|---|---|---|---|
| *VAR1* | *1* | *– 0,265* | *0,238* | *0,377* |
| *VAR2* | *– 0,265* | *1* | *0,492* | *– 0,726* |
| *VAR3* | *0,238* | *0,492* | *1* | *– 0,573* |
| *VAR4* | *0,378* | *– 0,726* | *– 0,573* | *1* |

To visualize the interconnections you must select the chart type and specify the indicators. For example, the connexion between sugar content of beets (VAR1) and sugar yield 1 ton of raw material (VAR 4) we represent as a two-dimensional scattering diagram. Command *Continue* is addressing to the procedure window of *Correlation matrices*. Choose the option *2D scatter plot*, and the corresponding schedule the correlation field appears on the screen:

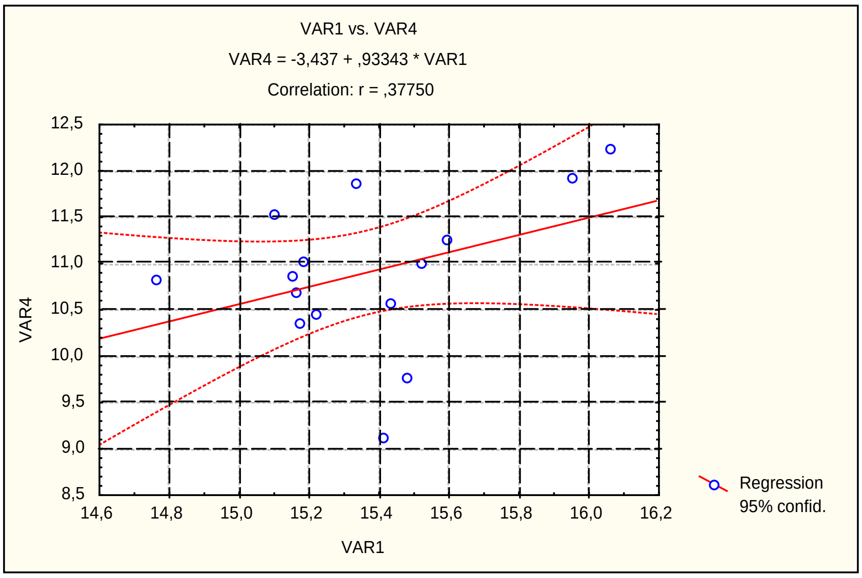



*Figure 1.1. Correlation field (scatter plot)*

Consequently, there is successively displayed three documents the on the screen by the results of exploratory analysis of data:

- table with the statistical characteristics of the distribution totality;
- correlation matrix;
- scatter diagram (correlation field).

These documents have standard headers and require editing for sense interpretation of the results.

Specification of data in table *Scroll Sheet* is given by the command of a context menu: for columns it is *ColumnSpecs*, for rows it is *ReName* and for the table name by the command *Titles*. The graphics can edit by options in the *Graphic Text Editor*

Edited numeric, text, and graphical information can be exported to other *Windows* applications such as a text file, for a printer or a special file so-called a *Report*. The output parameters are showed in the dialog box *Page /Output Setup*. For numerical and text data is selected by the command *Text / Scrollsheets Spreadsheets* , for graphic it is *Graphs* (with further refinement of the format and other parameters of the graphic). In the frame, *Output Header* we can specify the title, date, time and more.

The report is one of the documents of *Statistica*, that is created by option from *Windows* (frame *Output*) in expanded text format (RTF). RTF files can be edited directly in *STATISTICA* or any other word processor, such as *MS Word*. The system has the power to automatically prepare a report with *Auto-report*.

One of the conditions of statistical modeling is the homogeneity of totality. The identified regularities are stable only in a homogeneous totality and those can be applied to all units of totality.

The notion of homogeneity is associated with the presence for of all units the totality such common characteristics and traits that define isomorphism or the belonging to the same type. Evaluation of the degree of homogeneity uses the criteria of mathematical statistics, the most of them are focused on the analysis of the unimodal distributions.The set is considered the homogeneous if it has the symmetric normal distribution by the inherently. Of course, the social and economic phenomena do not have of normal distribution in pure form.But it is close to the other the unimodal distributions, it is often used as a first approximation in modeling. Some unimodal distributions are reduced to normal type through the transformation of the features to their logarithms. Number of asymmetric distributions, primarily such as the right asymmetry description can log-normal curve homogeneity.A number of asymmetric distributions, primarily with a Right asymmetry could be described through the log-normal curves.

The main properties of the normal distribution:

- distribution curve is symmetrical about the maximum ordinate, which corresponds to the "mean" or average arithmetic $\bar{x}$;
- 68,3 % values lie within the interval $\bar{x} \pm \sigma$; 95,4 % values lie within the interval $\bar{x} \pm 2\sigma$; 99,7 % values lie within the interval $\bar{x} \pm 3\sigma$;
- the ratio of standard deviation σ to average absolute deviation $\bar{l}$ is $\frac{\bar{l}}{\sigma} = \sqrt{\frac{2}{\pi}} = 0{,}8$ or $\frac{\sigma}{\bar{l}}$ = 1,25. Its value depends on atypical, abnormal observations which are contained in the totality and can serve as an indicator of the "contamination";
- the third central moments is $m_3 = 0$, and fourth central moments is $m_4 = 3m_2^2$. Hence, we have a skewness is $a_3 = m_3/\sigma^3 = 0$ and a kurtosis is $a_4 = m_4/m_2^2 = 3$.

Thanks to these properties the normal curve is used as *Standard* and it performs a significant role for the use of methods the sampling, regression and factor analysis.

Regularities of the one-dimensional distribution one can analyze by the procedures *Distribution* in the *Startup Panel - Descriptive Statistics* (*Basic Statistics and Tables* module) in *STATISTICA*.

Distribution of totality under variations of the features one can execute in the table *Frequency Table*, where are specified the intervals of groupings, frequency, relative frequency, cumulative frequency, and cumulative relative frequency.

Option *Normal expected frequencies* adds to the table the theoretical frequencies (group and cumulative). Verification the assumption of normal distribution is carried out by option *Kolmogorov-Smirnov & Lilliefors test for normality*. The following table shows the distribution of 120 firms that participated in the international exhibition by the level of advertising costs ( % of total expenditure) - VAR1.

The main characteristics of distribution:

*Table 1.5. Characteristics of distribution*

| DescriptiveStatistics (s______. sta) | | | | | | |
|---|---|---|---|---|---|---|
| | Mean | Minimum | Maximum | Std. Dev. | Skewness | Kurtosis |
| VAR1 | 2,295 | 0,7 | 3,9 | 0,6812 | 0,1361 | –0,473 |

On the data of the grouping:

*Table 1.6. The results of the grouping*

| K – S d = ,06713, p > .20; Lilliefors p > .20 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Continue… | Count | Cumul. Count | Percent ofValid | Cumul. % ofValid | Expecte dCount | Cumul. Expected | PercentExpe cted | Cumul. % Expected |
| ,0 < x < = ,5 | 0 | 0 | 0 | 0 | 0,5 | 0,5 | 0,42 | 0,42 |
| ,5 < x < = 1,0 | 2 | 2 | 1,67 | 1,67 | 2,9 | 3,4 | 2,44 | 2,86 |
| 1,0 < x < = 1,5 | 15 | 17 | 12,50 | 14,17 | 11, 2 | 14,6 | 9,29 | 12,15 |
| 1,5 < x < = 2,0 | 26 | 43 | 21,67 | 35,83 | 25,3 | 39, 9 | 21,11 | 33,26 |
| 2,0 < x < = 2,5 | 38 | 81 | 31,67 | 67,50 | 34,3 | 74,2 | 28,57 | 61,83 |
| 2,5 < x < = 3,0 | 22 | 103 | 18,33 | 85,83 | 27,8 | 102,0 | 23,14 | 84,97 |
| 3,0 < x < = 3,5 | 13 | 116 | 10,83 | 96,67 | 13,4 | 115,4 | 11,19 | 96,16 |
| 3,5 < x < = 4,0 | 4 | 120 | 3,33 | 100 | 3,9 | 119,3 | 3,23 | 99,39 |

maximum of the deviation of the cumulative relative frequencies lies within a fifth interval ***:|67,5–61,83|= 5,67%; | 67,5-61,83 | = 5,67***, ie ***d = 0,0567***. The same result was obtained with using cumulative frequencies: ***d = | 81,0 - 74,2 |:120 = 0.0567***. As you can see, there is slightly different criterion value ***d = 0,06713*** at the left top of the table. The differences between them are caused by the fact that the procedure *C-S* and *Lillieforstest for normality* for the ***d*** calculation is carried out for not grouped data, both of ***d*** is considerably less than critical $\left(1{,}22:\sqrt{120}\right)=0{,}11$ at ***α = 0,10***. Hence, the hypothesis of the normal distribution of firms by the relative frequencies of advertising costs is not rejected. The concordance of the empirical distribution for the firms by the level of advertising costs with a normal distribution we can see in the figure1.2.:

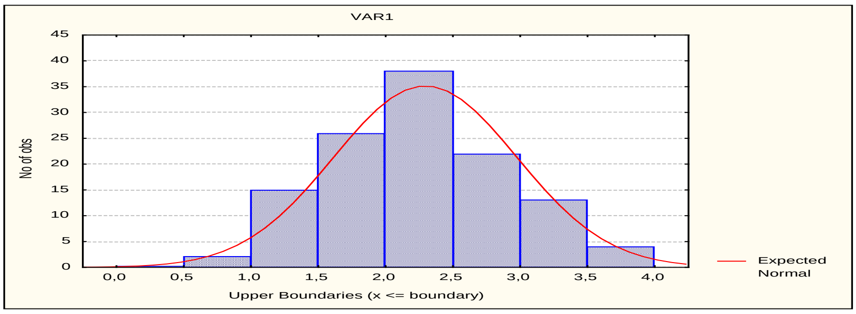



*Figure 1.2. The empirical and theoretical distribution of firms by the level of advertising costs*

The conclusion about the normal distribution the totality of firms due the advertising costs with the probability of ***0.95*** along $\chi^2$ - criterion is confirmed.

If you need to verify the hypothesis of concordance of data for other distributions (lognormal, exponential, etc.), one can use the module *Nonparametrics / Distribution.*

There are some components from the totality for which values of the features can lie far removed from the center of distribution, so called the atypical( with *abnormal* values). That can be maximum $x_n$ or minimum $x_1$ are values from the ordered row of observations $x_1 \le x_2 \le x_3 \le \ \dots \ \le x_n$. The origin of anomalous observations (the outliers) can be different. We need to identify the source and then to take an objective decision on withdrawal of these observations for further analysis.

Grubbs' test for outliers $T_n$ with statistical characteristics for which are the standardized marginal deviations of the anomalous values $x_n$ ($x_1$) from the mean $\bar{x}$, we can assess materiality deviations from the main mass.In the case of the maximum $x_n$.

Test detects one outlier at a time. This outlier is expunged from the dataset and the test is iterated until no outliers are detected. However, multiple iterations change the probabilities of detection, and the test should not be used for sample sizes of six or fewer since it frequently tags most of the points as outliers. then

$$T_n = \frac{x_n - \bar{x}}{\sigma},$$

were $\bar{x}$ and $\sigma$ are determined to totality as a whole.

If the actual value $T_n$ is less than critical, we have the deviation with probability ***(1 - α)*** is recognized random, insignificant, and if it is exceeding critical and the deviation is considered essential and therefore value is abnormal, unusual for the totality of whole.

In this case, the value is removed and the criterion uses to $x_{n-1}$ and etc. and so on, till we recognize that do not stay the outliers, and therefore the totality is homogeneous.

For example, we identify the challenges associated with anomaly detection of the maximum of the sugar losses under the processing of raw materials for sugar factories in the totality (see above). the asymmetry coefficient (coefficient of skewness) is the biggest by this feature (VAR3) that is 1,275. Maximum value of the feature is $x_n$ = 2,68, the mean is $\bar{x}$= 2,23, and σ = 0,179. From here

$$T_n = \frac{2{,}68 - 2{,}23}{0{,}179} = 2{,}51,$$

it is less than the critical value $\boldsymbol{T_{0,95}(15) = 2,705}$. Thus, the maximum the value features different from all set of observations, and we accept the homogeneity of totality.

In the description of the object modeling, it is important to organize the elements of totality according to some properties (qualities, value) and to determine the belonging of them to a certain type. If the property is characterized by one feature then the ordering units of totality we can do by the replacement the values of feature by the corresponding ranks. We can do it in *STATISTICA* by procedure *Rank Variables* in menu *Vars*.

One can choose the features, the scheme of arrangement (increase and decrease of values), processing conditions of the related ranks, rank type: regular (from1 to *n*) or factious (from 0 to 1) at dialog box *Rank Order Values of STATISTICA.*

Since the properties of socio-economic phenomena are characterized set of features ($m \geq 2$), it is necessary to aggregate all the features of set xi to *integral assessment* $G_j$ for ordering the units of totality.

This assessment interprets by the geometrical means as a point from multidimensional space, the coordinates of which indicate the scale or position of the *j* -th unit.

With the algebraic point of view, a value of features for the *j*-th element of the totality is represented by vector $x_j = (x_1, x_2, ..., x_m)^T$ and aggregate of them means that the vector converts to the scalar.

The aggregation of features is based on so-called the additive value theory, according to which the value of the whole is the sum of the values of its components.

This approach is implemented for the determining of the ratings that are based on methods of using the expert estimations (of ranks or scores). If the features of set ***X*** have the different units of measurement we should bring them to one basis, i.e. to do the pre-standardization. The methods use expert estimations of objects quality and criteria weights. This expert estimations are changed during the computation. The expert estimation are supposed to be measured in linear and ordinal scales. Each object is described by the set of linear, ordinal or nominal criteria. The constructed object estimations must not contradict both the measured criteria and the expert estimations Vector of the original values features $x_j = (x_1, x_2, ..., x_m)^T$ replaced by a vector standardized values $z_j = (z_1, z_2, ..., z_m)^T$.

More often, integrated assessment $G_j$ is defined as the arithmetic mean of the standardized values of the features $z_{ij}$. For the *j*-th unit of totality:

$$G_j = \frac{1}{m}\sum_{1}^{m} z_{ij} .$$

If the features of set have different weights then to each of them confer the especial weight $\omega_i$,, i.e. integrated assessment is a weighted arithmetic mean:

$$G_j = \sum_{1}^{m} z_{ij}\omega_i \text{, were } \Sigma\, \omega_i = 1.$$

Construction of integrated assessment involves four stages:

- organize the feature space;

- choice the method of standardization of features;

- ground the function of weight coefficients;

- define the procedure to aggregation of the features.

At the stage of the organization of the feature space X a priori qualitative analysis of the essence of the phenomenon plays a crucial role. Thus, for the characterisation of demographic situation, one can use indicators such as life expectancy, total fertility rate, infant mortality rate, the demographic burden of the working population, gross Migration and so on.

It is important to clarify what the importance of indicators of average income or provision of housing is not included in the demographic feature space because they are inherent characteristics of living standards. As for choosing of the weights coefficients, it is also based on the theoretical analysis of the essence of the phenomenon and for any one study determined by the expert statistical method. In forming the feature space it is important to ensure the unidirectional presentation of information $x_i$.

Demographics ceteris paribus be better that greater life expectancy and lower infant mortality. In this manner, life expectancy and infant mortality are the multidirectional information, and this should be considered when aggregating them into a united assessment.

To provide the unidirectional information of features one should divide them on *stimulants* and *de stimulants*.

The link between the assessment $G$ and $x_{st}$ (feature - stimulant) is a direct or between the assessment G and $x_{dst}$ (feature de stimulant) is inverse. Within the aggregation of the de stimulants are converted to stimulants, for example

$$x_{st} = 1 - x_{dst} \quad \text{or} \quad x_{st} = 1/x_{dst}.$$

In practice, one can use various methods of standardization. They are based on a comparison of empirical values of $x_{ij}$ and a certain quantity $a$. This is quantity can be a maximum $x_{\max}$, minimum $x_{\min}$, mean $\bar{x}$ or $x_0$ as an reference values of the feature. A result of comparison can be represented by the ratio $\frac{x_{ij}}{a}$ or as the deviation $\frac{x_{ij} - a}{q}$, where $q$ is a unit of standardization.

For example, determine the ratings of investment attractiveness of supplying companies the electronic equipment. The feature space is representing (%):
$x_1$-profitability of production, $x_2$-liquidity assets, $x_3$-share of expenditures on research. The easiest way the standardization is a relation $\frac{x_{ij}}{a}$, and because all of the features are stimulants, it is advisable to take $a = x_{\min}$ thus, $z_{ij} = \frac{x_{ij}}{x_{\min}}$.

The calculated ratings for the three companies show that among them the most attractive for investors are Kyivstar:

*Table 1.7. Calculation of integrated assessment (ranking) of companies*

| Компанія | $x_{1j}$ | $x_{2j}$ | $x_{3j}$ | $Z_{1j}$ | $Z_{2j}$ | $Z_{3j}$ | $G_j$ |
|---|---|---|---|---|---|---|---|
| Kyivstar | 21 | 31 | 34 | 1,3125 | 1,1481 | 1,2143 | 1,225 |
| Lifecell | 16 | 32 | 28 | 1 | 1,1852 | 1 | 1,0617 |
| WodafoneR | 19 | 27 | 32 | 1,1875 | 1 | 1,1428 | 1,1101 |

If there are standards, regulations, or any other the reference values of features $x_{i0}$, then aggregating of relation $\frac{x_{ij}}{x_{i0}}$, we can estimate the level of deviation from the "standard".

The values of integrated assessment $G_j$ should change from 0 to 1, and we do the calculation in which is aggregated the positive and negative deviations:

$$G_j = \frac{1}{m}\sum_{1}^{m}\left|\frac{x_{ij}}{x_{i0}} - 1\right|,$$

The depending on the specific aim of the study we can aggregate only positive or only negative deviation.

Sometimes we need to find the average of the squares of deviations (average quadratic).

Comparative analysis within a totality where each indicator has a typical average level implemented on the base the aggregation of relations $x_{ij}$ to the average level (mean) $\bar{x}$:

$$G_j = \frac{\mathbf{1}}{\boldsymbol{m}}\sum_{\mathbf{1}}^{\boldsymbol{m}}\frac{\boldsymbol{x_{ij}}}{\bar{\boldsymbol{x}}_{\boldsymbol{i}}}.$$

Obviously, if $G_j > 1$ then level the growth of phenomena in the $j$ -th unit above average on aggregate, while $G_j < 1$ it is lower.

Such a generalized assessment called the *multidimensional average* (or in this case as the normalized *mean vector*) and its significance can carry out a typology units of totality, say for the transport companies in terms of efficiency of fleet vehicles, for agricultural enterprises study in terms of the availability of resources and more..

If the features of set have different weights then multivariate mean is calculated as the weighted arithmetic mean: $G_j = \sum_{1}^{m}\frac{x_{ij}}{\bar{x}}\omega_i$ , where $\omega_i$ is weight of $i$-th feature and $\sum_{1}^{m}\omega_i = 1$.

For example, there are an average 13.5 of bodied, 0.85 standard units of the tractors and 23.8 cows per 100 hectares of farmland in agricultural farms of the region. The average rating of soil quality score is 51. In assessing of the resource potential of farms for indices softness the soil and availability of tractors, we offer the weights equal 0.3, for indicators of human resources and density of the number of cows is 0.2.

If for *j*-th farm the values of these indicators are 13.8; 0.8; 29.9 and 62.7 respectively then the multivariate mean of the availability of resources is

$$G_j = \frac{13,8}{13,5}0,2 + \frac{0,8}{0,85}0,3 + \frac{29,9}{23,8}0,2 + \frac{62,7}{51,0}0,3 = 1,106.$$

Hence, the extent of the availability of resources is higher than the average for the region. A similar content integrated assessment and can be calculated based on the percentage:

$$d_{ij} = \frac{x_{ij}}{\sum_{1}^{n} x_{ij}},$$

where $\sum_{1}^{n} x_{ij}$ is the total scope of the values *i*-th indicator of the totality on the whole;

$x_{ij}$ is the scope of the *j* -th component by this indicator. It is obvious that $\sum_{1}^{n} d_{ij} = 1$, or it is 100%. The formula of integral assessment is as follows:

$$G_j = \frac{n}{m}\sum_{1}^{m} d_{ij}.$$

For example, on *j* -th region (*j* =1, 2, 3) is accounted for 25% of exports of steel products, 48% - engineering products, 24% - chemicals and 32% - products of other industries. The rating of export potential of the industry in the region is

$$G_j = \frac{3}{4}(0,25 + 0,48 + 0,24 + 0,32) = 0,9675,$$

that is lower than the average level in three regions.

The integral assessments are widely used for socio-economic research, they are calculated based on the deviation ($x_{ij} - a$), that are standardized by the range ($x_{max} - x_{min}$).

Thus for stimulant it is $a = x_{min}$ and for de stimulant and where $a = x_{max}$.

$$z_{ij} = \frac{x_{ij} - x_{min}}{x_{max} - x_{min}}; \quad z_{ij} = \frac{x_{max} - x_{ij}}{x_{max} - x_{min}}.$$

Thus, $z_{ij}$ shows the relative position of the *j* -th unit of totality within a variation by the *i* -th feature. The $z_{ij}$ is approaching to 1 at high value of the *i* -th feature and it is approaching to 0 at low one. The integral assessment has the same property

$$G_j = \frac{1}{m}\sum_{1}^{m} z_{ij}.$$

The higher the level of development properties, the more value of $G_j$ deviates from 0.

A classic example of this type of evaluation is the integral Human Development Index (HDI) by the United Nations Development Programme. HDI is a composite statistic of life expectancy ($x_1$), education ($x_2$), and per capita income indicators ($x_3$), which are used to rank countries into four tiers of human development. A country scores higher HDI when the lifespan is higher, the education level is higher, the GDP per capita is higher, the fertility rate is lower, and the inflation rate is lower. Unit of standardization is theoretically possible range: for life (years) it is the interval (85 - 25), for education (100 - 0%), and for GDP of US per capita prior to 1997 (5120 - 100). If the actual per capita income indicators have exceeded average 5120 USD, the value was discounted under the particular technique. We'll define the Human Development Index (HDI) in a country where life expectancy is 69.4 years, education is 87% GDP per capita is 5010 USD:

$$G_j = \frac{1}{3}\left[\frac{69{,}4-25}{85-25} + 0{,}87 + \frac{5010-100}{5120-100}\right] = 0{,}863.$$

The using of theoretically possible range allows comparative analysis in space and in time.

If the analysis of the dynamic is not expected, then the actual range can be taken as the unit of standardization. The relative position of the $j$ -th unit of totality in a multidimensional space also characterizes the level of a taxonomic indicator of the development level and it is calculated according to the following rule of the standardization for deviations from the mean:

$$z_{ij} = \frac{x_{ij} - \bar{x}_i}{\sigma_i} \qquad \text{for the stimulant,}$$

$$z_{ij} = \frac{\bar{x}_i - x_{ij}}{\sigma_i} \qquad \text{for the de stimulant.}$$

This standardization allows you to get rid the measurement units, but at the same time it is happening the leveling of mean and the variance, where: for each feature $\bar{z} = 0$, the variance $\sigma_z^2 = 1$, and the range of variation $z_{ij}$ under Three Sigma Rule: from -3 to +3.

When calculating the integral assessment we are using the normal range of variation for all signs on one and the same level. For example, at two standard deviations (from -2 to +2). The distance between the upper (2) and lower (-2) range points is $|C| = 2/z_0/\sqrt{m}$, in multidimensional space, where $z_0$ is a point taken as a basis of comparison. If $z_0 = –2$, then for five feature we have

$$|C| = 2/–2/\sqrt{5} = 8{,}94.$$

The position of the $j$ -th unit concerning the base of comparison $z_0$ is defined as the Euclidean distance

$$G_{j0} = \left[ \sum_{1}^{m} \left( z_{ij} - z_0 \right)^2 \right]^{1/2},$$

and the ratio of the distance $G_{j0}$ to the standard range of variation |C| is called a taxonomic measure of development (TMD) for *jth* country

$$G_j = \frac{C_{j0;}}{|C|}$$

$$G_j \in [0; 1] .$$

The higher the level of the phenomenon development, then the value Gj is more. If the conditional object coordinates were determined at $z_0$ = +2 (in a top range of variation), its interpretation is true for the deviation ($1 - G_j$).
One can define the taxonomic measure of development for countries according to the following indicators: $x_1$- GDP per capita, ( thous. USD.); $x_2$- external debt, % of GDP; $x_3$- degree of energy self-sufficiency, %.

There are the standardized and absolute values of indicators represented in the table below

*Table 1.8. Calculation of the taxonomic measure of development*

| Country, j | Primary values of indicators , $x_{ij}$ | | | Standardized values of indicators , $z_{ij}$ | | | $C_{j0}$ | $G_j$ |
|---|---|---|---|---|---|---|---|---|
| | $x_{1j}$ | $x_{2j}$ | $x_{3j}$ | $z_{1j}$ | $z_{2j}$ | $z_{3j}$ | | |
| 1 | 5,8 | 14 | 48 | 0,524 | –1,259 | 0,481 | 4,811 | 0,538 |
| 2 | 4,7 | 22 | 55 | –0,543 | 0,140 | 1,182 | 3,963 | 0,443 |
| 3 | 3,9 | 28 | 29 | –1,319 | 1,189 | –1,422 | 1,206 | 0,135 |

For the calculating $C_{j0}$, the standardized value $z_{2j}$ should multiply on -1, since $x_2$ is de stimulant. For example, the Euclidean distance for the first country is

$$C_{10} = \left[(0{,}524 + 2)^2 + (1{,}259 + 2)^2 + (0{,}481 + 2)^2\right]^{1/2} = 4{,}811,$$

and TMD is $G_j = \frac{4{,}811}{8{,}94} = 0{,}538$.

The calculation of integral assessments or so-called Multi-Dimensional Assessment can be implemented in Data Management software module of STATISTICA by any of the considered methods above.

The main purpose of integral assessments is a ranking and typology of objects. However, like any other statistic indicator, Gj has the specific socio-economic contents, the variation of its values submits to certain laws of distribution, thus we can use such assessments for study the patterns of distribution, relationships, and trends of development.

A characteristic feature of the modeling and forecasting social and economic processes is multiversion so, one can use the different methods, models, information support, criteria of estimation, verification of model adequacy and other.

The choice between competing options is based on a certain system of rules which ensure the well-grounded estimates for each variant.

It is believed that the expert (lat. Expertus-experienced) has a system of rules and can compare options and it gives to each of them the number.

Often, the relative significance or the advantage set by the methods of ranking, pairwise comparisons or direct assessment. During the ranking, the expert must compare the variations (factors, models, objects, etc.) in order that deems reasonable and attributed to each of them the natural numbers or Ranks 1, 2, ..., n.

The number of ranks is equal to the number of variants. If the expert gives two or more variants of the same rank, then each of these options is attributed to the average grade calculated from the respective natural numbers 1,2, . , . , N. For the justification complex management decisions under uncertainty, the long-term forecasting of science, technology, economics one can use the method of expert evaluations.

The reliability of the group estimations depends on the consistency of expert opinion, which requires appropriate statistical processing of information. For the expert evaluations (*n* experts) for each *i*- th variant is determined a sum of ranks $\Sigma R_i$ by which the variants are ordered.

For example, we assign the first rank (the highest) to the variant which has the least total sum of ranks and the latter rank for the variant with the greatest total sum of ranks.

The results the survey of experts For example, we have the data of ranking for three variants under five experts:

*Table 1.9. The results of ranking*

| *Variant* | Expert | | | | | *Sum of ranks* | *d* | *$d^2$* |
|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | | | |
| A | 2 | 1 | 1 | 1 | 1 | 6 | -4 | 16 |
| B | 1 | 2 | 3 | 2 | 2 | 10 | 0 | 0 |
| C | 3 | 3 | 2 | 3 | 3 | 14 | 4 | 16 |
| Amount | X | X | X | X | X | 30 | X | 32 |

Variant A will get the first rank, for which $\Sigma R_i = 6$, variant B will get the second rank and variant C will get the third rank.

It should be noted that the ranks define only place among other options, not taking into account the existing distance between them.

Statistical analysis of the results of ranking involves evaluating the degree of concordance the expert opinions. The measure of consistency serves *concordance coefficient W*. The formula for the *W* statistic is:

$$W = \frac{12S}{n^2\left(m^3 - m\right)}$$

Where: S is the sum of squared deviations $S = \Sigma d^2$, *n* is the number of judges (raters or expert), *m* is the total number of objects being ranked. It is based on the deviation *d* from $\Sigma R_i$ - sum of ranks of detached variations from the average sum of ranks $S_{max} = n^2(m^3 - m) / 12$.

In the case of inconsistencies expert opinions W = 0 and vice versa the higher the degree of concordance, the more value W is close to 1.

By data of the table:

*Table 1.10. Estimation of the concordance of expert opinions*

| Variant | A | B | C | Amount | $W_i$ |
|---|---|---|---|---|---|
| A | 0 | 4 | 5 | 9 | 0,60 |
| B | 1 | 0 | 4 | 5 | 0,33 |
| C | 0 | 1 | 0 | 1 | 0,07 |
| Amount | 1 | 5 | 9 | 15 | 1,00 |

The average sum of ranks is 30: 3 = 10, the sum of squared deviations $S$ = 32, and the coefficient of concordance $W$ = (12 • 32) / 52 (33 - 3) = 0.64, which indicates to some differences in the assessments of experts regarding the importance of options.

Verification of significance the concordance coefficient $W$ by using the criterion of $\chi^2$ with ($m$ –1) *number of degrees of freedom*. Statistical characterization of criterion calculated as $\chi^2 = Wn(m-1)$. For this example $\chi^2$=0.64 ×5(3 – 1) = 6.4, it is exceeding the critical $\chi^2$ (2) = 5.99. This allows us to assert with a probability of 0.95 that the value of W = 0,64 is not accidental and the views of experts are agreed.

There are two estimates of experts such as 0 and 1 for the *pairwise comparisons*.

If the variant is more weighty, we give to it the rating 1, and if the variant is less weighty - 0. Results of pairwise comparisons are formed as a matrix $a_{ij}$ with a quantity of provided advantages as elements of it.

The diagonal elements of this matrix are equal 0. It is one of the properties of the matrix $a_{ij} + a_{ji} = n$, where $n$ is a number of experts. A *weightiness* of variant (or an option) is characterized by the ratio of the advantages for appropriate variant(or an option) to aggregate sum of matrix elements. According to the table, the option A is the most significant and for it ω= 9/15 =0,60.

Often expert can only directly assess the levels of individual properties or particular phenomenon. For example quality of output, competitiveness of firms and so on. In such situations, the scale (range) of estimates where the expert should estimate the subject (characteristic or properties) or phenomenon and give to it certain score $Z_{ij}$, where $i$ is number of characteristic, $j$ is number of element in totality.

For the certain set m of the properties of the phenomenon, one should determine the average score $G_j = \frac{\sum z_{ij}}{m}$.

Most rating systems are based on these methodological principles.

There is the world-famous rating system CAMEL. The rating system is designed to take into account and reflect all significant financial and operational factors examiners assess in their evaluation of an institutions performance. Institutions are rated using a combination of specific financial ratios and examiner qualitative judgments.

For each bank estimated capital adequacy, asset quality, management effectiveness, profitability and liquidity balance.

The average score is a rating of the financial condition of j-th bank and its value depends on the degree of interference of banking supervision and a set of measures to address the shortcomings.

The rating system is designed to take into account and reflect all significant financial and operational factors examiners assess in their evaluation of an institutions performance. Institutions are rated using a combination of specific financial ratios and examiner qualitative judgments.

The following describes some details of the CAMEL system in the context of examining a credit union.

Credit unions that maintain a level of capital fully commensurate with their current and expected risk profiles and can absorb any present or anticipated losses are accorded a rating of 1 for capital. Such credit unions generally maintain capital levels at least at the statutory net worth requirements to be classified as "well capitalized" and meet their risk-based net worth requirement. Further, there should be no significant asset quality problems, earnings deficiencies, or exposure to credit or interest-rate risk that could negatively affect capital.

A capital adequacy rating of 2 is accorded to a credit union that also maintains a level of capital fully commensurate with its risk profile both now and in the future and can absorb any present or anticipated losses. However, its capital position will not be as strong overall as those of 1 rated credit unions. Also, there should be no significant asset quality problems, earnings deficiencies, or exposure to interest-rate risk that could affect the credit union's ability to maintain capital levels at least at the "adequately capitalized" net worth category. Credit unions in this category should meet their risk-based net worth requirements.

A capital adequacy rating of 3 reflects a level of capital that is at least at the "undercapitalized" net worth category. Such credit unions normally exhibit more than ordinary levels of risk in some significant segments of their operation. There may be asset quality problems, earnings deficiencies, or exposure to credit or interest-rate risk that could affect the credit union's ability to maintain the minimum capital levels. Credit unions in this category may fail to meet their risk-based net worth requirements.

A capital adequacy rating of 4 is appropriate if the credit union is "significantly undercapitalized" but asset quality, earnings, credit or interest-rate problems will not cause the credit union to become critically undercapitalized in the next 12 months. A 4 rating may be appropriate for a credit union that does not have sufficient capital based on its capital level compared with the risks present in its operations.

If the properties $z_{ij}$ have a different weight, the rating is defined as the weighted arithmetic mean $G_j = \sum d_i z_{ij}$ where $d_i$ is the weight of the *i-th* properties. The commercial, political and other risks are estimated in this way.

For example, the commercial risk associated with internationalization of banking is estimated by BERI index (Business Environment Risk Information Index). The feature space of this index includes 15 parameters with the different weight that characterize the political and economic situation in the partner countries.

In particular, political stability (weight is 12%), the balance of payments (weight is 6%), the pace of economic development (weight is 10%) and others.
The amount of all weights equals 100 %.

One of the popular methods of forming the group examination is *Delphi method*. The Delphi method is a structured communication technique or method, originally developed as a systematic, interactive forecasting method which relies on a panel of experts. The experts answer questionnaires in two or more rounds. After each round, a facilitator or change agent provides an anonymous summary of the experts' forecasts from the previous round as well as the reasons they provided for their judgments. Thus, experts are encouraged to revise their earlier answers in light of the replies of other members of their panel. It is believed that during this process the range of the answers will decrease and the group will converge towards the "correct" answer. Finally, the process is stopped after a predefined stop criterion (e.g. number of rounds, achievement of consensus, stability of results) and the mean or median scores of the final rounds determine the results.

Delphi is based on the principle that forecasts (or decisions) from a structured group of individuals are more accurate than those from unstructured groups. The technique can also be adapted for use in face-to-face meetings, and is then called mini-Delphi or Estimate-Talk-Estimate (ETE). Delphi has been widely used for business forecasting and has certain advantages over another structured forecasting approach, prediction markets.

Regarding software, it should be noted that there are different software packages for the statistical modeling: *STATISTICA, SPSS, E-Wiues, Statgraphics, SAS, MathLab* ets.

## 1.3. Fundamentals and Tapes of Forecasting methods

*Forecast***:** A prediction, projection, or estimate of some future activity, event, or occurrence. One of the most difficult problems of management is to predict the future and find effective decisions under uncertainty. Prediction serves as a tool to minimize uncertainties and is scientifically justified of conclusion about future events, prospects the development of processes and the consequences of management decisions.

Types of Forecasts.

*Economic forecasts*: predict a variety of economic indicators, like money supply, inflation rates, interest rates, etc.

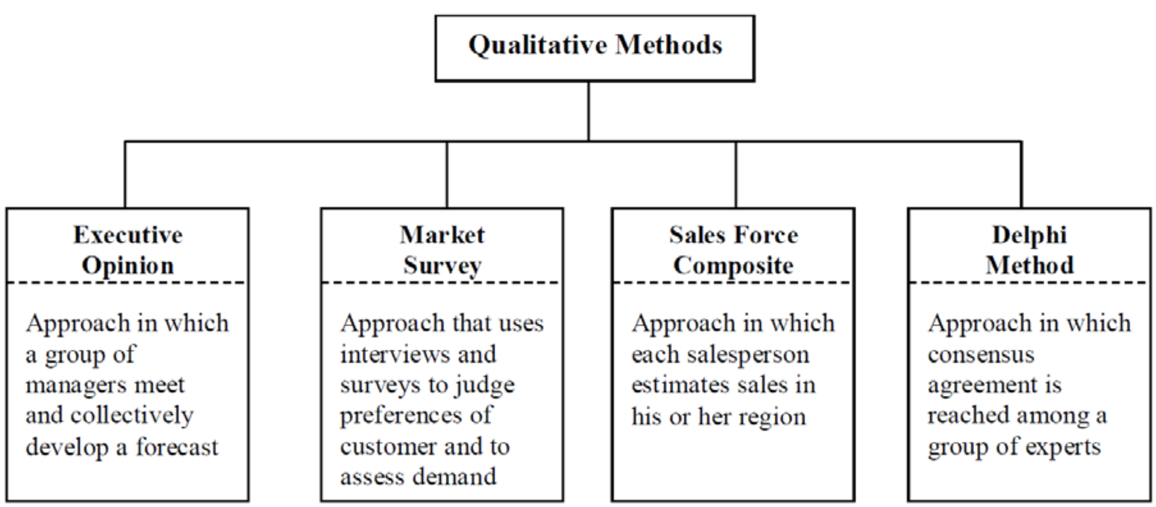
*Technological forecasts*: predict rates of technological progress and innovation.

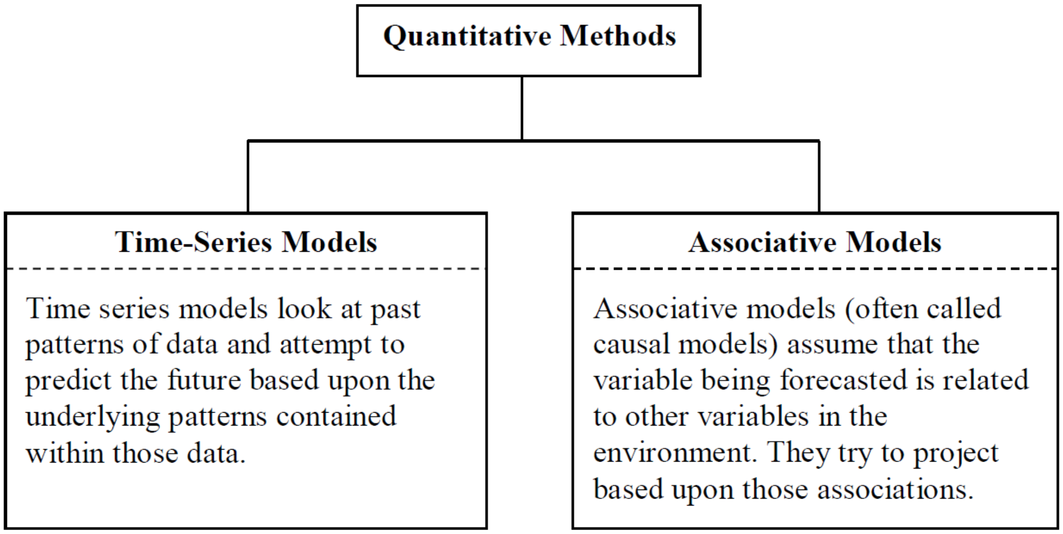
*Demand forecasts*: predict the future demand for a company's products or services.

Since virtually all the operations management decisions (in both the strategic category and the tactical category) require as input a good estimate of future demand, this is the type of forecasting that is emphasized in our textbook and in this course.

*Qualitative methods***:** These types of forecasting methods are based on judgments, opinions, intuition, emotions, or personal experiences and are subjective in nature. They do not rely on any rigorous mathematical computations.

*Quantitative methods***:** These types of forecasting methods are based on mathematical (quantitative) models, and are objective in nature. They rely heavily on mathematical computations.






Time series Models

| *Model* | *Description* |
|---|---|
| Naïve | Uses last period's actual value as a forecast |
| Simple Mean (Average) | Uses an average of all past data as a forecast |
| Simple Moving Average | Uses an average of a specified number of the most recent observations, with each observation receiving the same emphasis (weight) |
| Weighted Moving Average | Uses an average of a specified number of the most recent observations, with each observation receiving a different emphasis (weight) |
| Exponential Smoothing | A weighted average procedure with weights declining exponentially as data become older |
| Trend Projection | Technique that uses the least squares method to fit a straight line to the data |
| Seasonal Indexes | A mechanism for adjusting the forecast to accommodate any seasonal patterns inherent in the data |

*Decomposition of time series,* the patterns that may be present in a time series:

*Trend*: Data exhibit a steady growth or decline over time.

*Seasonality*: Data exhibit upward and downward swings in a short to intermediate time frame (most notably during a year).

*Cycles:* Data exhibit upward and downward swings in over a very long time frame.

*Random variations*: Erratic and unpredictable variation in the data over time with no discernable pattern.

Data set to demonstrate forecasting methods. The following data set represents a set of into yearly totals.

For various illustrations that follow, we may make slightly different assumptions about starting points to get the process started for different models. In most cases we will assume that each year a forecast has been made for the subsequent year. Then, after a year has transpired we will have observed what the actual demand turned out to be (and we will surely see differences between what we had forecasted and what actually occurred, for, after all, the forecasts are merely educated guesses).

Finally, to keep the numbers at a manageable size, several zeros have been dropped off the numbers (i.e., these numbers represent demands in thousands of units).

Hypothetical demands that have occurred over several consecutive years. The data have been collected on a quarterly basis, and these quarterly values have been amalgamated

| Year | Quarter 1 | Quarter 2 | Quarter 3 | Quarter 4 | Total Annual |
|---|---|---|---|---|---|
| 1 | 62 | 94 | 11 | 41 | 31 |
| 2 | 73 | 11 | 13 | 52 | 36 |
| 3 | 79 | 11 | 14 | 58 | 39 |
| 4 | 83 | 12 | 14 | 62 | 41 |
| 5 | 89 | 13 | 16 | 65 | 45 |
| 6 | 94 | 13 | 16 | 70 | 46 |

Another particularity of a statistical prediction is definiteness in time. All the forecast period is called prediction time. We distinguish the forecasts by the duration of these periods: short-term (1 year), medium term (5 years) and long-term (5 to 20 years or more). Length of the prediction time depends on the specific object forecasting of dynamics intensity, the duration of the identified patterns and trends. Projected results for the period prejudice can be represented by a single number (Forecast Point) or range of values, which belongs to a certain probability of predictive value (Prediction interval). Statistical forecasts are based on hypotheses about the stability of quantity values projected; law of distribution; relationships with other variables like. The main forecasting tool is an extrapolation. In mathematics, extrapolation is the process of estimating, beyond the original observation range, the value of a variable on the basis of its relationship with another variable. It is similar to interpolation, which produces estimates between known observations, but extrapolation is subject to greater uncertainty and a higher risk of producing meaningless results. Extrapolation may also mean extension of a method, assuming similar methods will be applicable. In general, to extrapolate is a process of projecting, expanding or extending a given data or observation towards an area that is not experienced so far, in order to predict a conjectural knowledge about the unknown or untouched area. The extrapolation reflects trends that construct a future image. It is a way of predicting by observing known data or past experiences. Abstracting from the causes of the process, the patterns of development seen as a function of time. The univariate time series serve as an information base of the forecasting.

In multivariate prediction, the process considered as a function of a set of factors whose impact is analyzed simultaneously or with some delay. The system of interrelated time series serves as the information base.

So, the factors included in the model explicitly, it is particularly important prior, theoretical analysis of the structure of relationships.

An important step is verification of statistical forecasting example, that the assessment of their accuracy and validity. During the verification, one can use a set of criteria, methods, and procedures that make it possible to assess the quality of the forecast.

A hindcast of prognosis estimation or forecast for the last time (ex-post forecast) is the most common.

A hindcast is a way of testing a mathematical model that is run in past periods for which actual demand history is also available. The system calculates the forecast accuracy measurements by comparing the differences between the actual values and the ex-post values. Hindcasting is also known as backtesting.

The verification procedure is as follows. Dynamic range is divided into two parts, the first : for $t$ =1, 2, 3, …, $p$ is called retrospection (prehistory) or hindcast, the second for $t = p + 1, p + 2, p + 3, \dots, p + (n - p)$ is the forecast period.
According to retrospection modeled dynamics and pattern based on the calculated prediction model $Y_{p+v}$, where $v$ - prediction time. Retrospection consistently changing and accordingly the forecast period changes too.

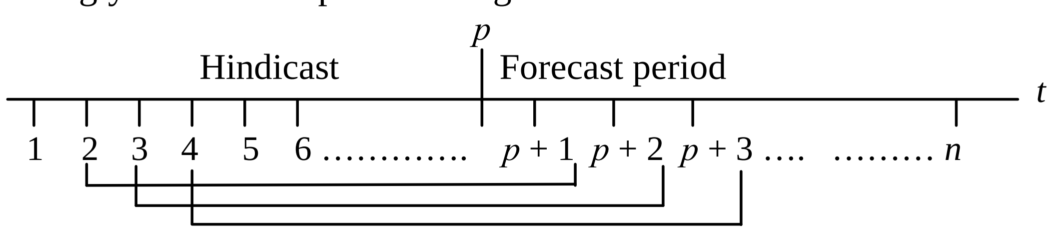



*Fig.1.3. Scheme of verification the accuracy of the retrospective forecasting for v=1*

Since the actual value of the forecast period are known, we can determine the prediction error as the difference between actual $y_t$ and forecast $Y_t$ levels: $e_t = y_t - Y_t$. There will be $n - p$ errors.

An univariate forecasting allows you to measure the forecast error in next ways:

- Mean absolute deviation (MAD) or Mean absolute error (MAE) $\bar{e} = \frac{\sum |e_t|}{n - p}$
- Error total (ET) $e_t = y_t - Y_t$
- Mean square error (MSE) $= \frac{\sum e_t^2}{n - p}$ and root mean squared error (RMSE) $s = \sqrt{\frac{\sum e_t^2}{n - p}}$

- Mean percentage error (MPE) $\widehat{A} = 100 \frac{\sum \left| \frac{e_t}{y_t} \right|}{n - p}$.

If the result of evaluating forecast accuracy meets the criteria for the accuracy of, say, 10%, the predictive model being considered and recommended for practical use. Obviously, the prediction error depends on the length of retrospection and forecasting horizon. The optimum ratio between them is 3: 1.

In evaluating and comparing the accuracy of forecasts used as *Theil inequality coefficient*, which is zero in the absence of forecast errors and has no upper limit:

$$T = \frac{\sqrt{\sum (y_t - Y_t)^2}}{\sqrt{\sum y_t^2}}.$$

In most cases, existing methods of verification forecasts are based on statistical procedures that lead to building confidence limits of the forecast, ie to the creation of interval forecasts.

When forecasting processes, the development of which is wholly or partly is formalized (e.g., science and technology, socio-economic and political consequences of the adoption of certain management decisions), using methods of expert estimates or method Delfi. They are based on the mobilization of professional experience and intuition of experts (judges or raters) who are selected on the basis their competencies.

# Chapter 2. Methods and models of Multiple Factor Analysis

## 2.1. Cluster procedures for classification. Discriminant Analysis

Diversity of explored objects is a source of one of the most important problems for researcher, because every next object to be explored is not same as previous one. Hence researcher have to find some similarity existing between the objects. There is a very constructive approach to do that. It's so called cluster analysis. Cluster analysis is a generic name for a variety of mathematical methods, numbering in the hundreds, that can be used to find out the objects in a set are similar. The best known of these research goals is the making of classifications. One reason that cluster analysis is so useful is that researchers in all fields need to make and revise classifications continually. Classification is a general process related to categorization, the process in which objects are recognized, differentiated, and understood. Classification as approach is based on two points:

- Classes, are conceptually meaningful groups of objects that share common characteristics, play an important role in how people analyze and describe the world.
- Level of resemblance is higher between objects included in a same class, than between objects which we include in different ones.

One can estimate resemblance using single or several features forming an "image of a class". Traditional method of classification is to put features in order by importance. At every step of the way a researcher pays attention only to a single feature of the sequence, so he makes consecutive description of classes. But possibilities of the method are often restrained.

Another method is to use simultaneously the whole set of features. Every element of a population is described by the whole set of features and it is interpreted as a point in a multidimensional feature space. In this case every member of the set is represented as point in a multidimensional space or so called feature space, and if two points presented corresponding elements have a small distance from one another are not far one from another then we can say that two objects are very similar. There are two ways to build classification:

- Classification is based on constructing multidimensional estimates (indexes, ratings and so on);
- Multidimensional classification is based on the idea that level of similarity depicted by some metrics.

One can notice that any approach is to some extent subjective, since the classification depend on choice of a feature space and applied metrics.

Further we consider multidimensional classification using methods of the cluster analysis.

The notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms. But mostly a ***cluster*** is a set of one or more objects that we are willing to call similar to each other. However, different researchers employ different cluster models, and for each of these cluster jby different algorithms, varies significantly in its properties. Understanding these "cluster models" is key to understanding the differences between the various algorithms. Similarity inside a cluster and dissimilarity outside the cluster is defined by some metric measuring resemblance between elements. Using the metric we can calculate so called resemblance coefficient.

A ***resemblance coefficient*** measures the overall resemblance — the degree of similarity between each pair of objects. We denote $d_{ij}$ (distance) or $r_{ij}$ as a resemblance coefficient between $i$-th and $j$-th objects. If a sample contain $n$ elements then the $\frac{n(n-1)}{2}$ resemblance coefficients form a ***resemblance matrix.*** Obviously the diagonal elements of the matrix are equal to zero. In order to obtain the resemblance coefficients, we make use of the idea. Let a data matrix has $m$ attributes that are subscripted $i = 1, 2,. .. ,m$; and $n$ objects subscripted $j = 1, 2,.. ,n$ . In the data matrix, the value of data for any $i$-th attribute and $j$-th object is denoted as $x_{ij}$. The corresponding value in the standardized data matrix is denoted as $z_{ij} = \frac{x_{ij} - \overline{x_i}}{\sigma_i}$ . Researchers use more often than others Euclidean distance metric (the symbol $d_{jk}$ stand for the Euclidean distance coefficient for objects $j$ and $k$)

$$d_{jk} = \sqrt{\sum_{i=1}^{m} \left(z_{ij} - z_{ik}\right)^2} .$$

If attributes makes unequal contributions towards the resemblance between objects, then more appropriate is so called weighted Euclidean distance coefficient

$$d_{jk} = \sqrt{\sum_{i=1}^{m} w_i \left(z_{ij} - z_{ik}\right)^2} .$$

The factors $w_i$ are represent as weights. Also one can find very useful Manhattan distance which is defined by

$$d_{jk} = \sum_{i=1}^{m} \left|z_{ij} - z_{ik}\right|$$

and so on.

There are two categories approaches of clustering methods, namely an iterative clustering, and an hierarchical one. There are two basic approaches for generating a hierarchical clustering:

Agglomerative: Start with the points as individual clusters and, at each step, merge the closest pair of clusters. This requires defining a notion of cluster proximity.

Divisive: Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide which cluster to split at each step and how to do the splitting. Sometimes researcher can restrain distance between objects of the same cluster putting so called threshold, g. e. maximal possible distance. If the distance between objects exceed some value $c_0$ then the objects belong to different clusters. Many agglomerative hierarchical clustering techniques are variations on a single approach: starting with individual points as clusters, successively merge the two closest clusters until only one cluster remains. Basic agglomerative hierarchical clustering algorithm.

1. Compute the proximity matrix, if it is necessary – repeat;

2. Merge the closest two $j$-th and $k$-th clusters and setting for the new cluster new number $q$;

3. Update the proximity matrix to reflect the proximity between the new cluster and the original clusters by the equation: $d_{qs} = a_1 d_{js} + a_2 d_{ks} + a_3 d_{jk} + a_4 (d_{js} - d_{ks})$ until only one cluster remains.

Values of the parameters $a_1$, $a_2$, $a_3$, $a_4$ depends on a selected agglomerative hierarchical clustering techniques, such as MIN, MAX, and Group Average, come from a graph-based view of clusters. MIN defines a cluster proximity as the proximity between the closest two points that are in different clusters, or using graph terms, the shortest edge between two nodes in different subsets of nodes. Alternatively, MAX takes the proximity between the farthest two points in different clusters to be the cluster proximity, or using graph terms, the longest edge between two nodes in different subsets of nodes. Another graph-based approach, the group average technique, defines cluster proximity to be the average pairwise proximities (average length of edges) of all pairs of points from different clusters.

A hierarchical clustering is often displayed graphically using a tree-like diagram called a dendrogram, which displays both the cluster-subcluster relationships and the order in which the clusters were merged.

Analytics software Statistica includes module *Cluster Analysis* which also performs actions of the agglomerative hierarchical clustering algorithm.

If attributes takes only binary values g. e. “0” or “1”, then each resemblance coefficient $r_{ij}$ is calculated as we show below. We will pretend that the two objects $j$ and $k$ are samples of products, and the eight attributes are the qualities the products must be complied. The code “1” means a quality is present in the sample and “0” otherwise.

*Table 2.1. Attributes*

| **Objects** | **Attributes** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *A* | *B* | *C* | *D* | *E* | *F* | *G* | *H* |
| *j* | *0* | *1* | *1* | *0* | *1* | *0* | *0* | *1* |
| *k* | *0* | *0* | *1* | *1* | *1* | *1* | *0* | *1* |

The two-by-two Table 2.2. summarizes the four types of matches among these data. There are a = 3 cases of 1-1 matches (attributes for which both objects are coded “1”); b = 1 cases of 1- 0 matches (attributes for which object j is coded “1” and k is coded “0”); c = 2 cases of 0 -1 matches (for which j is coded “O” and k is coded “1”); and d = 2 cases of 0 — 0 matches (for which both objects are coded “0”).

*Table 2.2. Contingency table*

| *Value of Attribute* | *1* | *0* |
|---|---|---|
| *1* | *a* | *b* |
| *0* | *c* | *d* |

There are various resemblance coefficients, for example *Simple Matching Coefficient*

$$r_{jk} = \frac{a+d}{a+b+c+d}.$$

If we are more interested in the cases of 1-1 then the ones of 0 — 0 matches, we use the *Russell and Rao Coefficient*

$$r_{jk} = \frac{a}{a+b+c+d}$$

or the *Jaccard Coefficient*

$$r_{jk} = \frac{a}{a+b+c}.$$

There are also the *Sorenson Coefficient*

$$r_{jk} = \frac{2a}{2a+b+c}.$$

It ignores the 0—0 matches, *d*, but gives twice the weight to the 1-1 matches, *a*. In essence it makes the weight of a equal to the combined weights of *b* and *c*. There are other coefficients, but values all of them are in the segment $0 \le r_{jk} \le 1$. The choice of the coefficient is very subjective and depend on objectives of a researcher.

In addition to the *Hierarchical Clustering* there is another approach to the clustering procedure. That is so called *Iteration Clustering*. The typical example of the latter is *K-means Clustering*. We first choose K initial points called *centroids*, where K is a user-specified parameter, namely, the number of clusters desired. Each point is then assigned to the closest centroid, and each collection of points assigned to a centroid is a cluster. The centroid of each cluster is then updated based on the points assigned to the cluster. We repeat the assignment and update steps until no point changes clusters, or equivalently, until the centroids remain the same. The K-means algorithm may be described as follows.

1: Select K points as initial centroids.
2: Repeat.
3: Form K clusters by assigning each point to its closest centroid.
4: Recompute the centroid of each cluster.
5: Until Centroids do not change.

But there are also another very different approach. It is so called *Discriminant Analysis*. Discriminant analysis is a statistical analysis to predict a categorical dependent variable (called a grouping variable) by one or more continuous or binary independent variables (called predictor variables). Discriminant analysis is used when groups are known *a priori* (unlike in cluster analysis). For example, we might attempt to classify loan applicants into three credit risk categories (X): good, moderate, or bad. You might use continuous variables such as current salary, years in current job, age, and debt burden, (Ys) to predict an individual's credit risk category. You could build a predictive model to classify an individual into a credit risk category using discriminant analysis. Generally speaking, discriminant analysis is based on minimizing the classification error.

Discriminant analysis works by creating one or more linear combinations of predictors, creating a new latent variable for each function. These functions are called *discriminant functions*. The number of functions possible is either p-1 where p is a number of groups, or m (the number of predictors), whichever is smaller. The first function created maximizes the differences between groups on that function. The second function maximizes differences on that function, but also must not be correlated with the previous function. This continues with subsequent functions with the requirement that the new function not be correlated with any of the previous functions. A linear discriminant function is a linear combination of numerical characteristics of features formed some set. It is defined for every *j*-th group ($j$ = 1, 2, …, p):

$$f_j = a_{0j} + a_{1j}\bar{x}_{1j} + a_{2j}\bar{x}_{2j} + ... + a_{mj}\bar{x}_{mj}$$

where $a_{ij}$ are coefficients of the function; $\bar{x}_{ij}$ are means of *i*-th predictor in *j*-th group.

Coefficients $a_{ij}$ may be calculated as $a_{ij} = (n-p)\sum_{k=1}^{m} b_{ik}\bar{x}_{kj}$,

where $b_{ik}$ is element of the matrix which is inverse matrix of the

$$W_{ik} = \sum_{j=1}^{p}\sum_{h=1}^{n_j}(x_{ijh} - \bar{x}_{ij})(x_{kjh} - \bar{x}_{kj}).$$

The discriminant function coefficients are partial coefficients, reflecting the unique contribution of each variable to the classification of the criterion variable. Also a constant is defined as $a_{0j} = -0{,}5\sum_{k=1}^{m} a_{kj}\bar{x}_{kj}$ .

Geometrically speaking, is an imaginary point in an m-dimensional Euclidian space. Coordinates of the point are means of the features of a j-th group. Values of $f_j$ for *p* groups are called *centroids* because one can describe them as centers of mass. The Procedure is based on the measuring proximity of an h-th element to the centroids of distinguished by a researcher classes. To determine whether the element is contained in a group one can use the *Mahalanobis distance* which is a measure of the distance between a point P and a centroid.

$$D^2 = (n-p)\sum_{i=1}^{m}\sum_{k=1}^{m} b_{ik}(x_{ijh} - \bar{x}_{ij})(x_{kjh} - \bar{x}_{kj}).$$

A discriminant function maximizes differences between the groups and minimizes variance within the groups. One can say that the population divided optimally in to groups if the ratio of the variance between groups to the variance within groups is minimal. The variance between groups is characterized by square of difference between centroids $\left(f_j - f_s\right)^2$. The variance within groups may be calculated as

$$\sigma_f^2 = \frac{\sum_{j=1}^{p}\sum_{h=1}^{n_j} a_{ij}(x_{ijh} - \overline{x}_{ij})^2}{\sum_{j=1}^{p} n_j - p}.$$

Hence the approach is based on minimizing the square of $D$:

$$D^2 = \frac{\left(f_j - f_k\right)^2}{\sigma_f^2}.$$

The value is called *generalized Mahalanobis distance*. To estimate whether a discriminant function is appropriate one g.e. it can effectively distinguish groups one from another we can use Wilks' lambda

$$\lambda = \prod_{j=1}^{p} \frac{1}{1+\lambda_j},$$

where $\lambda j$ are eigenvalues of the covariance matrix. The $\lambda$-test takes into account as well differences between the classes as homogeneity within every group. Since $\lambda$ is calculated as inverse value hence the more difference between centroids the less value of the lambda and vice-versa. Also we can instead use F-test or chi-squared test.

A software package *STATISTICA* include a module *Discriminant Analysis* where the procedures are contained. For example consider a procedure for testing the suitability of applicants for a some job. The procedure provides for the dividing a set of claimants into two groups: qualified claimants (group C) un qualified ones (group NC) for the position. The features of our interest are random access memory (VAR2) and attention span (VAR3). Values of the variables are given in the Table 2.3

*Table 2.3. Values of the variables*

| | VAR1 | VAR2 | VAR3 |
|---|---|---|---|
| 1 | C | 72 | 75 |
| 2 | C | 57 | 70 |
| 3 | C | 59 | 62 |
| 4 | C | 67 | 72 |
| 5 | C | 75 | 59 |
| 6 | C | 62 | 73 |
| 7 | NC | 67 | 50 |
| 8 | NC | 56 | 59 |
| 9 | NC | 58 | 54 |
| 10 | NC | 47 | 60 |

Using commands on the start panel of the module we can select features: *independent variable list* - VAR2 and VAR3; *grouping variable* - VAR1; method of analysis – *Standard*. According to the results of analysis we can find number of the features, values of lambda test and F-test

*Discriminant Function Analysis Results*
*Number of variables in the model: 2*
*Wilks'Lambda:,270128 approx.F(2,7)=9,45681 p < ,01024.*

Selected options *Distances between groups* and *Squared Mahalanobis distances* determine distances between the groups and within the groups. So generalized Mahalanobis distance between the groups is equal to 11,258. Also we can find the distances from the elements to the centroids. The Share of correctly identified elements is 90%.

*Table 2.5. Table of squared Mahalanobis distance and classifications*

| *Squared Mahalanobis Distances from Group Centroids (new.sta)* | | | |
|---|---|---|---|
| *Incorrect classifications are marked with ** | | | |
| *Continue…* | *Observed Classif.* | *C p =,600* | *NC p =,400* |
| *1* | *C* | *3,180* | *22,786* |
| *2* | *C* | *1,227* | *6,898* |
| **3* | *C* | *3,051* | *1,653* |
| *4* | *C* | *0,586* | *14,174* |
| *5* | *C* | *3,140* | *8,023* |
| *6* | *C* | *0,616* | *12,305* |
| *7* | *NC* | *11,076* | *2,004* |
| *8* | *NC* | *6,560* | *0,315* |
| *9* | *NC* | *10,313* | *0,090* |
| *10* | *NC* | *12,275* | *1,789* |

New object will be included in a group for which the value of the discriminant function will be the most. So in the example new claimant get 65 points in the random access memory test and 65 points in the attention span test. The value of the discriminant function for the group C is equal to 163,957 and for the group NC is equal to 159,39. Since the first value is more then the another one then claimant is included into the group C.

Described procedure may be used also in cases when number of groups is more then two. It's important that number of elements within every group should be not least then two. Sometimes we are not interested to include an object into certain group, but instead to determine probabilities for the object to be included into one or another group. For that we have to chose on option *Posterior Probabilities*. We'll next consider specificities using the methods solving som interesting problems.

*Example 2.1 Hotel business in Odessa.*

Effectiveness of an enterprise is characterized by many indicators which are often interrelated, then it's necessary to use methods and algorithms of multi-dimensional statistical analysis.

The main task of our study is to explore population of enterprises (hotels) and to divide into groups taking into account level of functioning effectiveness. We choose such indicators: *x1* - average workload for one bed; *x2* - average cost of a bed per day; *x3* - average rate of return per one lodger; *x4* - labour consumption of services rendered; *x5* - material consumption of services rendered.

It is worth noting that the indicator *x1* was calculated as the ratio of a number of a night in hotel to a number of beds in an one. Data about indicator *x2* was provided by State Statistics Service of Odessa region. The indicator *x3* was calculated as ratio of net income of an enterprise to a number of lodgers. The indicator *x4* was calculated as a ratio of labour costs to net income. And finally, the indicator *x5* was calculated as a ratio of material costs to net income.

As result we obtain the table of the main indicators

*Table 2.6. The main indicators of economic activities*

| № | x1 | x2 | x3 | x4 | x5 |
|---|---|---|---|---|---|
| 1 | 54,12 | 109,0 | 276,06 | 63,04 | 18,53 |
| 2 | 60,59 | 1001 | 2561,70 | 10,37 | 48,04 |
| 3 | 41,88 | 420 | 985,01 | 25,37 | 37,84 |
| 4 | 11,92 | 380 | 861,02 | 31,55 | 56,70 |
| 5 | 92,64 | 227 | 479,44 | 61,76 | 13,14 |
| 6 | 27,50 | 140 | 636,36 | 22,20 | 70,00 |
| 7 | 51,28 | 2126 | 7145,54 | 13,51 | 14,28 |
| 8 | 9,68 | 551 | 1184,85 | 45,18 | 78,43 |
| 9 | 88,69 | 1297 | 2923,79 | 23,28 | 41,76 |
| 10 | 32,13 | 139 | 139,00 | 73,88 | 35,82 |
| 11 | 7,44 | 2085 | 4037,42 | 6,84 | 27,35 |
| 12 | 106,77 | 608 | 1285,78 | 11,51 | 16,68 |
| 13 | 69,72 | 900 | 1266,16 | 7,34 | 29,20 |
| 14 | 35,04 | 800 | 1616,73 | 39,43 | 32,16 |
| 15 | 71,36 | 1701 | 4636,11 | 9,55 | 96,32 |
| 16 | 78,79 | 1167 | 3526,30 | 15,62 | 36,98 |
| 17 | 2,89 | 522 | 3627,49 | 19,71 | 48,75 |
| 18 | 109,10 | 2363 | 2611,05 | 20,36 | 14,69 |
| 19 | 106,07 | 323 | 825,55 | 38,45 | 16,95 |
| 20 | 110,49 | 213 | 628,24 | 46,74 | 32,29 |
| 21 | 76,38 | 185 | 857,45 | 29,45 | 7,72 |
| 22 | 132,44 | 550 | 558,00 | 19,53 | 14,40 |
| 23 | 86,30 | 275 | 696,75 | 20,30 | 14,72 |
| 24 | 37,00 | 411 | 4156,46 | 1,33 | 0,75 |
| 25 | 172,69 | 70 | 238,11 | 8,42 | 86,04 |
| 26 | 42,17 | 150 | 1028,58 | 51,03 | 5,97 |
| 27 | 139,16 | 390 | 322,50 | 13,12 | 17,42 |
| 28 | 52,13 | 495 | 714,67 | 9,97 | 72,92 |
| 29 | 142,94 | 726 | 963,58 | 11,99 | 44,18 |
| 30 | 172,68 | 580 | 665,13 | 12,63 | 68,91 |
| 31 | 141,80 | 602 | 1390,52 | 7,72 | 75,41 |
| 32 | 120,75 | 239 | 320,14 | 2,75 | 10,85 |
| 33 | 10,14 | 1540 | 4037,42 | 6,84 | 23,35 |
| 34 | 27,14 | 300 | 367,87 | 21,34 | 29,81 |
| 35 | 64,17 | 394 | 1170,98 | 12,10 | 37,01 |
| 36 | 62,31 | 750 | 1312,22 | 21,49 | 32,95 |
| 37 | 31,00 | 350 | 7510,63 | 9,54 | 30,39 |
| 38 | 103,46 | 1457 | 3698,69 | 15,96 | 13,64 |
| 39 | 142,61 | 912 | 2218,65 | 15,94 | 13,69 |
| mean | 75,01 | 703,70 | 1884,15 | 22,49 | 35,03 |

In the first stage a set of variables must be divided into two subsets:

- stimulant variables,
- destimulant variables.

Such a variable is recognized as a stimulant variable for which the greater the value the higher the position of the object under consideration. Destimulants are transformed into stimulants by the formulas:

$$\hat{x}_i = -x_i \,,$$

or

$$\hat{x}_i = \frac{1}{x_i}$$

where $x_i$ is a variable recognized as a destimulant. In the case the variables *x1* and *x3* are recognized as stimulants, and the variables *x2* , *x4* and *x5* are recognized as destimulants. We transformed all destimulants into stimulants replacing formers by inverted values of them. Then we replaced all variables by normalized ones, using the formula

$$z_{ij} = \frac{x_{ij} - \overline{x_i}}{\sigma_i}$$

where $x_{ik}$ is a value of *i*-th feature of *j*-th element ( $j = \overline{1,39}$ ), $\overline{x_i}$ is an average value of the feature, $\sigma_i$ is standard deviation of it calculated for all elements of population.

The main task of the clustering analysis is to create groups containing elements which seemed so similar. The similarity is determined by a rule of calculating some metric $d_{jk}$ describing it between j-th and k-th elements of population. As a metric we can choose the widely known Euclidian one

$$d_{jk} = \left[ \sum_{i=1}^{5} (z_{ij} - z_{ik} \right]^{1/2}$$

where $z_{ij} = \frac{x_{ij} - \overline{x_i}}{\sigma_i}$ and $z_{ik} = \frac{x_{ik} - \overline{x_i}}{\sigma_i}$ are normalized values of i-th feature ($i = \overline{1,5}$) for j-th and k-th elements ( $j,k = \overline{1,39}$ ) respectively. The clustering procedure will be executed using the *STATISTICA*, and it based on using obtained matrix of normalized values. First we use agglomerative hierarchical procedure “Joining (Tree clustering)” to consolidate elements joining them by “Single Linkage (nearest neighbor)” in a sense of an Euclidean metric. We was provided by a diagram called a tree, or dendrogram, that shows the hierarchy of similarities among all pairs of objects. From the tree the clusters can be read off. Also we get the graph which demonstrates how the elements were joined (Fig 2.1 and Fig 2.1 respectively). We come to conclusion that the population may be grouped into three clusters.

To clarify the results we use iteration clustering namely K-means method. The method groups using the principle “the nearest neighbor” to form the clusters near centroids (points whose coordinates are the group means of features). At the same time we stated that number of the clusters have to be equal to 3 (as it was shown by the previous hierarchical analysis). As it’s known the iterations of the K-means method minimize variances within the groups, maximizing by that the similarity within the corresponding clusters. As a result the objects(hotels) were grouped into three clusters(see below the results table) where the first and the seventh columns are the hotel numbers, the columns from second to sixth describe standardized data, the second-to-last one contains cluster numbers, and the last one contains “distances” from objects to the centroids of the clusters.

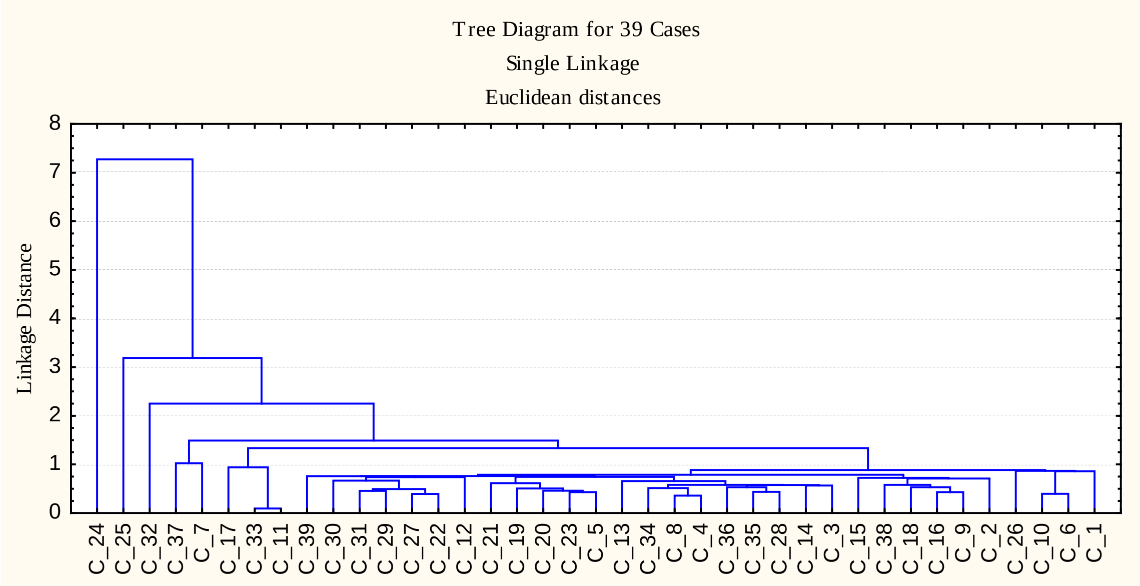



Fig. 2.1. Hierarchical dendrogram classificating the hotels according to a feature “efficiency of enterprises”.

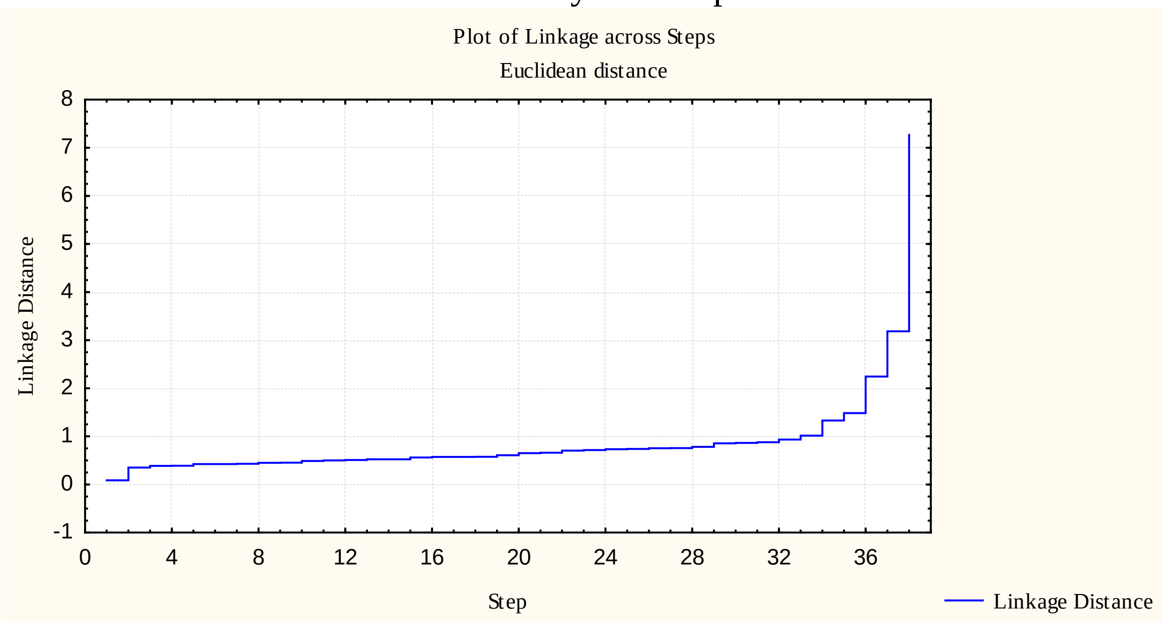



Fig. 2.2. Step-by-step graphics of including the objects into the clusters

*Table 2.7. Distribution of object for the clusters*

*Table.*sta

| | **x1** | **x2** | **x3** | **x4** | **x5** | **observ_num** | **Cluster** | **Dist.** |
|---|---|---|---|---|---|---|---|---|
| **C_1** | -0,44 | 2,23 | -0,88 | -0,60 | -0,11 | 1 | 1 | 0,91 |
| **C_2** | -0,30 | -0,68 | 0,37 | 0,05 | -0,27 | 2 | 2 | 0,20 |
| **C_3** | -0,70 | -0,19 | -0,49 | -0,41 | -0,24 | 3 | 2 | 0,58 |
| **C_4** | -1,33 | -0,10 | -0,56 | -0,47 | -0,29 | 4 | 2 | 0,71 |
| **C_5** | 0,37 | 0,53 | -0,77 | -0,59 | -0,01 | 5 | 1 | 0,22 |
| **C_6** | -1,00 | 1,51 | -0,68 | -0,36 | -0,30 | 6 | 1 | 0,81 |
| **C_7** | -0,50 | -0,86 | 2,88 | -0,13 | -0,04 | 7 | 2 | 0,99 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| ***C_8*** | *-1,37* | *-0,39* | *-0,38* | *-0,55* | *-0,31* | *8* | *2* | *0,65* |
| ***C_9*** | *0,29* | *-0,76* | *0,57* | *-0,38* | *-0,26* | *9* | *2* | *0,39* |
| ***C_10*** | *-0,90* | *1,52* | *-0,95* | *-0,62* | *-0,24* | *10* | *1* | *0,81* |
| ***C_11*** | *-1,42* | *-0,86* | *1,18* | *0,45* | *-0,20* | *11* | *2* | *0,54* |
| ***C_12*** | *0,67* | *-0,45* | *-0,33* | *-0,03* | *-0,08* | *12* | *1* | *0,45* |
| ***C_13*** | *-0,11* | *-0,64* | *-0,34* | *0,37* | *-0,21* | *13* | *2* | *0,55* |
| ***C_14*** | *-0,84* | *-0,59* | *-0,15* | *-0,52* | *-0,22* | *14* | *2* | *0,43* |
| ***C_15*** | *-0,08* | *-0,82* | *1,50* | *0,12* | *-0,32* | *15* | *2* | *0,45* |
| ***C_16*** | *0,08* | *-0,73* | *0,90* | *-0,21* | *-0,24* | *16* | *2* | *0,30* |
| ***C_17*** | *-1,52* | *-0,35* | *0,95* | *-0,32* | *-0,27* | *17* | *2* | *0,48* |
| ***C_18*** | *0,72* | *-0,88* | *0,40* | *-0,33* | *-0,05* | *18* | *2* | *0,60* |
| ***C_19*** | *0,65* | *0,07* | *-0,58* | *-0,52* | *-0,09* | *19* | *1* | *0,26* |
| ***C_20*** | *0,75* | *0,64* | *-0,69* | *-0,55* | *-0,22* | *20* | *1* | *0,24* |
| ***C_21*** | *0,03* | *0,89* | *-0,56* | *-0,45* | *0,25* | *21* | *1* | *0,34* |
| ***C_22*** | *1,21* | *-0,39* | *-0,72* | *-0,31* | *-0,04* | *22* | *1* | *0,52* |
| ***C_23*** | *0,24* | *0,26* | *-0,65* | *-0,33* | *-0,05* | *23* | *1* | *0,17* |
| ***C_24*** | *-0,80* | *-0,17* | *1,24* | *5,31* | *6,01* | *24* | *3* | *0,00* |
| ***C_25*** | *2,05* | *4,02* | *-0,90* | *0,23* | *-0,32* | *25* | *1* | *1,75* |
| ***C_26*** | *-0,69* | *1,34* | *-0,47* | *-0,57* | *0,43* | *26* | *1* | *0,72* |
| ***C_27*** | *1,35* | *-0,12* | *-0,85* | *-0,11* | *-0,10* | *27* | *1* | *0,49* |
| ***C_28*** | *-0,48* | *-0,31* | *-0,64* | *0,08* | *-0,31* | *28* | *1* | *0,57* |
| ***C_29*** | *1,43* | *-0,54* | *-0,50* | *-0,05* | *-0,26* | *29* | *1* | *0,64* |
| ***C_30*** | *2,05* | *-0,42* | *-0,67* | *-0,09* | *-0,30* | *30* | *1* | *0,82* |
| ***C_31*** | *1,40* | *-0,44* | *-0,27* | *0,32* | *-0,31* | *31* | *1* | *0,65* |
| ***C_32*** | *0,96* | *0,46* | *-0,85* | *2,20* | *0,07* | *32* | *1* | *1,08* |
| ***C_33*** | *-1,36* | *-0,80* | *1,18* | *0,45* | *-0,17* | *33* | *2* | *0,51* |
| ***C_34*** | *-1,01* | *0,15* | *-0,83* | *-0,35* | *-0,21* | *34* | *1* | *0,69* |
| ***C_35*** | *-0,23* | *-0,13* | *-0,39* | *-0,06* | *-0,24* | *35* | *1* | *0,43* |
| ***C_36*** | *-0,27* | *-0,56* | *-0,31* | *-0,35* | *-0,23* | *36* | *2* | *0,47* |
| ***C_37*** | *-0,93* | *-0,02* | *3,08* | *0,12* | *-0,21* | *37* | *2* | *1,12* |
| ***C_38*** | *0,60* | *-0,79* | *0,99* | *-0,22* | *-0,02* | *38* | *2* | *0,54* |
| ***C_39*** | *1,42* | *-0,64* | *0,18* | *-0,22* | *-0,02* | *39* | *1* | *0,75* |

The table reflects as every element was assigned to an appropriate cluster taking into account the all five features. Hence we get the preliminary results:
- The first cluster contains the hotels 1, 5, 6, 10, 12, 18, 19-23, 25-32, 34, 35, 39;
- The second cluster contains the hotels 2-4, 7-9, 11, 13-17, 33, 36-38;
- The third cluster (so called abnormal cluster) contains only the hotel 24.

To ensure that our clustering procedure is robust we use one of the techniques of the *Discriminant Analysis*, namely by a table showing so called posterior probabilities. Using the data from the table we deduce that the clustering procedure by the latent feature “efficiency of enterprises” was made correctly.

Also we have obtained another result of the *Discriminant Analysis*, namely the *Discriminant Functions*

*Table 2.8. Using the Discriminant Functions for classification*

| Variable | Classification functions; grouping: x6 (Table.sta) | |
|---|---|---|
| | G_1:1 p=,55263 | G_2:2 p=,44737 |
| x1 | 0,81639 | -0,8857 |
| x2 | 0,64565 | -0,8247 |
| x3 | -0,81039 | 1,5696 |
| x4 | -0,56665 | -0,7703 |
| x5 | -3,85831 | -10,0819 |
| Constant | -1,45822 | -2,9284 |

The functions let us include new elements into the clusters by substituting standardized values of features into the discriminant functions and then comparing the values:

$$f_1 = -1,46 + 0,82z_1 + 0,65z_2 - 0,81z_3 - 0,57z_4 - 3,86z_5;$$

$$f_2 = -2,93 - 0,89z_1 - 0,82z_2 + 1,57z_3 - 0,77z_4 - 10,08z_5.$$

It is worth for noting that the three former features and two latter ones tend to be separated (see below the dendrogram). The features may be included into two different groups of features. For example, the "operational productive factor" and "resource factor".

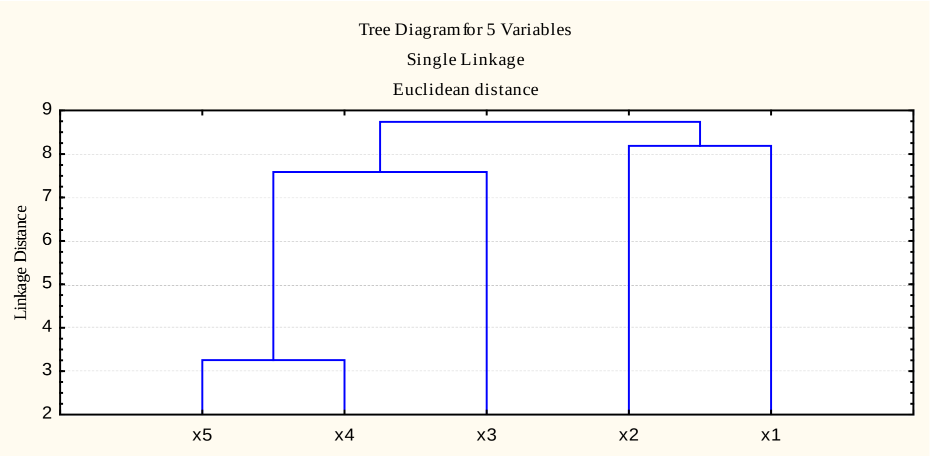



*Fig. 2.3. Hierarchical dendrogram for the clustering procedure*

In the next phase of our work we explore the impact one or another factor for the decision to include an element into a certain group. For this purpose we use another clustering method called *Double Clustering*. According to the method we have to choose a value for the threshold to obtain three clusters. Results obtained by using the method (Fig. 2.4) and rearranged matrix (Table 2.9) of the Double Clustering allow us to estimate the impact of the groped features on forming clusters.


*Fig. 2.3. Variations of the impact on the clustering procedure*

*Table 2.9. Rearranged matrix of the Double Clustering*

*Rearranged matrix* (table.sta)

| | x1 | x2 | x3 | x4 | x5 |
|---|---|---|---|---|---|
| **C_1** | -0,43944 | 2,228370 | -0,878933 | -0,596801 | -0,113389 |
| **C_6** | -0,99929 | 1,506411 | -0,682003 | -0,362402 | -0,302646 |
| **C_10** | -0,90183 | 1,524674 | -0,953844 | -0,615493 | -0,237627 |
| **C_21** | 0,02874 | 0,888938 | -0,561162 | -0,451500 | 0,247009 |
| **C_26** | -0,69073 | 1,337178 | -0,467627 | -0,566795 | 0,428459 |
| **C_3** | -0,69693 | -0,185923 | -0,491445 | -0,407591 | -0,244716 |
| **C_14** | -0,84066 | -0,587852 | -0,146164 | -0,520495 | -0,222462 |
| **C_28** | -0,48122 | -0,314130 | -0,639201 | 0,081387 | -0,305374 |
| **C_35** | -0,22806 | -0,129168 | -0,389796 | -0,060646 | -0,241891 |
| **C_36** | -0,26716 | -0,558236 | -0,312602 | -0,350394 | -0,226009 |
| **C_4** | -1,32697 | -0,096853 | -0,559211 | -0,469626 | -0,286660 |
| **C_8** | -1,37407 | -0,386747 | -0,382218 | -0,546442 | -0,309972 |
| **C_34** | -1,00680 | 0,152544 | -0,828753 | -0,347881 | -0,210798 |
| **C_25** | 2,05448 | 4,016065 | -0,899677 | 0,229321 | -0,315350 |
| **C_5** | 0,37089 | 0,534886 | -0,767775 | -0,594154 | -0,007893 |
| **C_20** | 0,74636 | 0,637977 | -0,686444 | -0,552371 | -0,223087 |
| **C_19** | 0,65326 | 0,069895 | -0,578600 | -0,515315 | -0,089439 |
| **C_23** | 0,23743 | 0,259299 | -0,648999 | -0,328504 | -0,046734 |
| **C_32** | 0,96209 | 0,457391 | -0,854842 | 2,199071 | 0,069005 |
| **C_2** | -0,30330 | -0,677055 | 0,370328 | 0,050089 | -0,271500 |
| **C_13** | -0,11116 | -0,637212 | -0,337776 | 0,369808 | -0,207451 |
| **C_15** | -0,07676 | -0,823184 | 1,504135 | 0,116496 | -0,321269 |
| **C_9** | 0,28773 | -0,758017 | 0,568232 | -0,379119 | -0,256562 |
| **C_16** | 0,07943 | -0,727557 | 0,897547 | -0,209873 | -0,241809 |
| **C_18** | 0,71697 | -0,881692 | 0,397300 | -0,329696 | -0,046052 |
| **C_38** | 0,59832 | -0,788087 | 0,991771 | -0,220936 | -0,021173 |
| **C_12** | 0,66800 | -0,447278 | -0,327050 | -0,026121 | -0,084805 |
| **C_39** | 1,42192 | -0,642194 | 0,182827 | -0,220462 | -0,022312 |
| **C_22** | 1,20801 | -0,385926 | -0,724836 | -0,313019 | -0,039583 |
| **C_27** | 1,34935 | -0,120833 | -0,853551 | -0,111978 | -0,096945 |
| **C_29** | 1,42884 | -0,542437 | -0,503159 | -0,054427 | -0,262832 |
| **C_31** | 1,40484 | -0,441741 | -0,269803 | 0,315568 | -0,307533 |
| **C_30** | 2,05443 | -0,419348 | -0,666278 | -0,088300 | -0,301571 |

| | | | | | |
|---|---|---|---|---|---|
| **C_7** | -0,49906 | -0,864910 | 2,875711 | -0,129475 | -0,036664 |
| **C_37** | -0,92567 | -0,016689 | 3,075255 | 0,117691 | -0,213813 |
| **C_11** | -1,42123 | -0,861639 | 1,176909 | 0,449595 | -0,196394 |
| **C_33** | -1,36442 | -0,801317 | 1,176909 | 0,449595 | -0,166493 |
| **C_17** | -1,51695 | -0,351266 | 0,952853 | -0,316786 | -0,272944 |
| **C_24** | -0,79938 | -0,166340 | 1,241971 | 5,307980 | 6,007281 |

For example, the values of forth and fifth features had the most impact on to form an "abnormal" third cluster concluding only element namely a 24th hotel.

Completing the economic-mathematical analysis for every cluster we obtain average values of the features according to the results which allow us to describe the outcome of the clustering procedure.

*Table 2.10. Average values of the features for the clusters.*

| *Features* / *Clusters* | *X1* | *X2* | *X3* | *X4* | *X5* |
|---|---|---|---|---|---|
| *1* | *97,83* | *454,00* | *836,11* | *26,10* | *32,60* |
| *2* | *46,01* | *1065,37* | *3183,19* | *18,85* | *40,51* |
| *3* | *37,00* | *410,50* | *4156,46* | *1,33* | *0,75* |

So for the first cluster (hotels 1, 5, 6, 10, 12, 18, 19-23, 25-32, 34, 35, 39) are typical the average values of the features: average workload for one bed is equal to 97,83 persons, average cost of a bed per day is 454UAH, average rate of return per one lodger is 836,11UAH, labour consumption of services rendered is 0,261 UAH. and material consumption of services rendered is 0,326 UAH.

For the second cluster (hotels 2-4, 7-9, 11, 13-17, 33, 36-38) are typical the average values of the features: average workload for one bed is equal to 46,01 persons, average cost of a bed per day is 1065,37UAH, average rate of return per one lodger is 3183,19UAH, labour consumption of services rendered is 0,1885 UAH. and material consumption of services rendered is 0,3051 UAH.

As for third cluster which include only 24th element are typical the average values of the features: average workload for one bed is equal to 37,0 persons, average cost of a bed per day is 410,5 UAH, average rate of return per one lodger is 4156,46UAH, labour consumption of services rendered is 0,0133 UAH. and material consumption of services rendered is 0,0075 UAH.

Taking into account that third cluster which include only one element, let's consider it's activities as abnormal (abnormal cluster to be separately explored) and exclude it from the analysis. Then using let us analyze the clusters we built, using found maximal, minimal and average values of the features.

The first cluster is characterized by high average workload for one bed, high average cost of a bed per day, high average rate of return per one lodger, high labour consumption of services rendered and by low material consumption of services rendered.

As for the elements included into the second cluster we can say that the hotels are characterized by low average workload for one bed, but by high average cost of a bed per day and by high average rate of return per one lodger, also take place low labour consumption of services rendered by high material consumption of services rendered.

Hence we can infer that the enterprises included into both clusters are focused on customer service of different categories of lodgers. Enterprises included into the first cluster are more preferable for the lodgers which are less wealthy, and enterprises of the second cluster are preferable for the more wealthy ones.

Taking into account that the explored enterprises share the same business, we have to consider the correlations reflecting activities of enterprises to find more profitable behavior.

The main criteria for estimating of efficiency of enterprises included into the clusters is the overall impact of activities of an enterprise, resource efficiency, reduce the whole operating cost which did not worse services and so on, to maximize profit and to improve the financial position of the enterprises.

Using the above obtained results, we can infer that

- The researched population was divided into three clusters. The grouping was based on the latent feature "Efficiency of enterprise"
- Guided by the economic indicators of enterprise activities and by the results of classification, we get two cluster need to be explored further.
- After grouping the enterprises into clusters we should explore dynamics of the main indicators of enterprise activities.
- To increase the efficiency of the enterprises included into the clusters we should provide further exploration of correlations that reflect the efficiency and can explain to some extent whose position is more profitable.

*Example 2.2. Cluster analysis for Regions of Ukraine (not to be confused with a party of traitors) by the latent feature "attractiveness for business of a manufacturing baby plant based food"*

Using the data by region for the years 2009-2011 we provide the Clustering for regions of Ukraine (see below the analysis for 2011).

*Table 2.10. Data by regions*

| № | Name of the Region | Average per capita Income, (UAH) | Number of children under 3 years old | Annual demand for the food (tons) | Land suitability (for growing the plants) | restriction on the level of profitability (%) | restriction on the level of trade margin (%) |
|---|---|---|---|---|---|---|---|
| c1 | Crimea | 1549,13 | 108699 | 1542,32 | 1 | 25 | 25 |
| c2 | Vinnitsa | 1556,66 | 69996 | 992,93 | 2 | 25 | 20 |
| c3 | Volyn | 1365,9 | 59207 | 838,99 | 5 | 100 | 100 |
| c4 | Dnipro | 1977 | 143989 | 2040,21 | 1 | 100 | 100 |
| c5 | Donetsk | 2051,9 | 168048 | 2379,24 | 2 | 25 | 20 |
| c6 | Zhytomyr | 1535,73 | 58429 | 829,46 | 2 | 15 | 15 |
| c7 | Zakarpattya | 1221,91 | 71430 | 1014,62 | 1 | 15 | 25 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| c8 | Zaporizhia | 1928,57 | 72364 | 1025,68 | 5 | 100 | 25 |
| c9 | Ivano-Frankivsk | 1436,53 | 66618 | 943,46 | 2 | 15 | 10 |
| c10 | Kyiv | 2676,05 | 206632 | 2931,35 | 2 | 20 | 20 |
| c11 | Kirovohrad | 1476,02 | 41677 | 591,18 | 2 | 20 | 25 |
| c12 | Lugansk | 1715,75 | 84544 | 1198,29 | 5 | 20 | 30 |
| c13 | Lviv | 1603,35 | 114710 | 1626,26 | 6 | 100 | 25 |
| c14 | Mykolaiv | 1636,46 | 51483 | 730 | 4 | 100 | 25 |
| c15 | Odessa | 1573,2 | 113734 | 1614,59 | 2 | 25 | 25 |
| c16 | Poltava | 1707,08 | 57137 | 809,69 | 2 | 25 | 25 |
| c17 | Rivne | 1420,01 | 67740 | 962,18 | 5 | 15 | 20 |
| c18 | Sums | 1591,75 | 41704 | 591,11 | 4 | 25 | 25 |
| c19 | Ternopil | 1334,28 | 47982 | 679,9 | 5 | 15 | 15 |
| c20 | Kharkov | 1785,09 | 105247 | 1491,68 | 5 | 25 | 20 |
| c21 | Herson | 1434,2 | 48463 | 687,41 | 6 | 20 | 20 |
| c22 | Khmelnytsky | 1532,18 | 57418 | 814,32 | 2 | 100 | 15 |
| c23 | Cherkassy | 1483,19 | 49129 | 697,12 | 2 | 10 | 25 |
| c24 | Chernivtsi | 1302,87 | 43383 | 616,29 | 3 | 100 | 15 |
| c25 | Chernihiv | 1559,51 | 40097 | 568,72 | 2 | 10 | 20 |

(*Note: Unfortunately, the regions was arranged by the Ukrainian alphabet*)

As above we for the clustering use for the measuring a "distance" the Euclidian metric and the feature space is 6-dimensional one. We choose such indicators:

$x_{1j}$ is the average per capita income in j-th region (j=1,2,…,25) for the corresponding year ;

$x_{2j}$ is the number of children under 3 years old (as potential consumers of the production) in j-th region (j=1,2,…,25) for the corresponding year ;

$x_{3j}$ is the annual demand for the food. We mean the annual demand as a scientifically based volumes of the consumed food in j-th region (j=1,2,…,25) ;

$x_{4j}$ is the land suitability in j-th region (j=1,2,…,25) for cultivation of the vegetables and fruits that are essential to produce the food. Unfortunately, that national statistics on this indicator were unavailable. Hence, using by expert estimates, we introduce such values for this indicator: 1 – unusable lands; 2 – marginal lands; 3 – marginal or usable lands; 4 – lands of a mixed type; 5 – usable lands; $x_{5j}$ is the restriction on the level of profitability (%) for food producers. The parameter was established for j-th region (j=1,2,…,25) by its body of the local self-government. If the parameter for a region was not was established then we state that it is equal to 100%;

$x_{6j}$ is the restriction on the level of trade margin (%);The parameter was too established for j-th region (j=1,2,…,25) by its body of the local self-government. If the parameter for a region was not was established then we also state that it is equal to 100%.

It' worth for noting that for most regions of Ukraine except Kyiv city, Odessa, Mykolaiv and Crimea regions, according to the statistical data, actual consuming of the production is lesser then a scientifically based one.

As it's known a necessary condition to provide the clustering is the indicators to be pointed in a same direction (fortunately, in the case the indicators are stimulants) and the indicators to be standardized. Hence we calculate

$$z_{mj} = \frac{x_{mj} - \overline{x_m}}{\sigma_m}, \quad (m = 1, 2, ..., 6;\ j = 1, 2, ..., 25)$$

where $\overline{x_m}$ are the average values of the features and $\sigma_m$ are their standard deviations. The values of the standardized indicators is represented by a matrix $(25 \times 6)$. The rows of the latter correspond to the regions (c1, c2,...,c25), the columns, respectively correspond to the features.

The modeling was based on the matrices of standardized values provided by the program *STATISTIA* using its module prescribed to execute the clustering procedure. We was provided by dendrogram, that shows the hierarchy of similarities among all pairs of objects. Also we get the graph which demonstrates how the elements were joined (for example see Fig. 2.5 shown below).

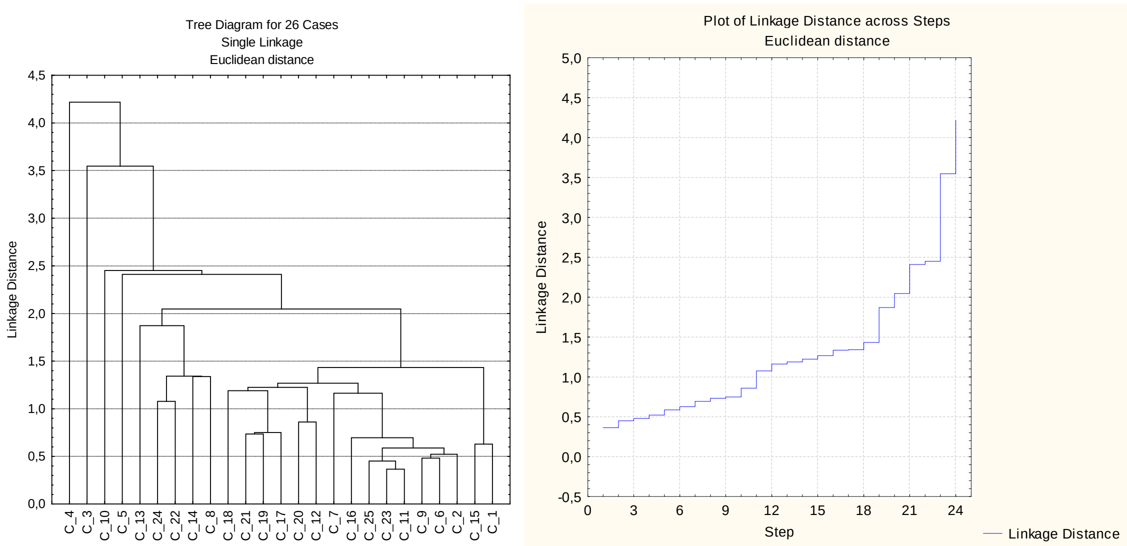



*Fig. 2.5. Hierarchical dendrogram Step-by-step graphics of including the objects into the clusters.*

Analyzing the results acquired in the first step lead to the conclusion that regions of Ukraine may be included into three clusters (groups): 1) "leaders", which are the most favorable for development of the industry; 2) "averages" are moderately favorable; 3) "outsiders" are unfavorable regions.

Then, in the next phase of our research we use iteration clustering namely K-means method. The method groups using the principle "the nearest neighbor" to form the clusters near centroids (points whose coordinates are the group means of features). At the same time we stated that number of the clusters have to be equal to 3(as it was shown by the previous hierarchical analysis). As it's known the iterations of the K-means method minimizes variances within the groups, maximizing by that the similarity within the corresponding clusters. . As a result the objects(regions) were grouped into three clusters(see below the results table) where rows corresponds to the regions, the first and second columns are region numbers, the second-to-last one contains cluster numbers, and the last one contains "distances" from objects to the centroids of the clusters.

*Table 2.12. Distribution of objects for the clusters*

| | Table | | |
|---|---|---|---|
| | 1 Case_N | 2 Cluster | 3 Distance |
| C_1 | 1 | 3 | 0,75 |
| C_2 | 2 | 3 | 0,26 |
| C_3 | 3 | 2 | 1,26 |
| C_4 | 4 | 1 | 1,26 |
| C_5 | 5 | 1 | 0,61 |
| C_6 | 6 | 3 | 0,30 |
| C_7 | 7 | 3 | 0,65 |
| C_8 | 8 | 2 | 0,57 |
| C_9 | 9 | 3 | 0,35 |
| C_10 | 10 | 1 | 0,96 |
| C_11 | 11 | 3 | 0,43 |
| C_12 | 12 | 3 | 0,62 |
| C_13 | 13 | 2 | 0,81 |
| C_14 | 14 | 2 | 0,28 |
| C_15 | 15 | 3 | 0,69 |
| C_16 | 16 | 3 | 0,39 |
| C_17 | 17 | 3 | 0,51 |
| C_18 | 18 | 3 | 0,44 |
| C_19 | 19 | 3 | 0,62 |
| C_20 | 20 | 3 | 0,80 |
| C_21 | 21 | 3 | 0,79 |
| C_22 | 22 | 2 | 0,65 |
| C_23 | 23 | 3 | 0,37 |
| C_24 | 24 | 2 | 0,65 |
| C_25 | 25 | 3 | 0,45 |

To ensure that our clustering procedure is robust we use one of the techniques of the Discriminant Analysis, namely by a table showing so called posterior probabilities. Using the data from the table shown below we deduce that the clustering procedure by the latent feature "favorability for business" was made correctly.

*Table 2.13. Discriminant Analysis results*

| Case | Posterior Probabilities (Table)<br>Incorrect classifications marked with * | | | |
|---|---|---|---|---|
| | Observed Classif. | G_1:1 p=,12000 | G_2:2 p=,24000 | G_3:3 p=,64000 |
| 1 | G_3:3 | 0,000000 | 0,000044 | 0,999956 |
| 2 | G_3:3 | 0,000071 | 0,000000 | 0,999929 |
| 3 | G_2:2 | 0,000000 | 0,999999 | 0,000001 |
| 4 | G_1:1 | 1,000000 | 0,000000 | 0,000000 |
| 5 | G_1:1 | 0,999536 | 0,000000 | 0,000464 |
| 6 | G_3:3 | 0,000043 | 0,000000 | 0,999957 |
| 7 | G_3:3 | 0,000000 | 0,000000 | 1,000000 |
| 8 | G_2:2 | 0,000000 | 1,000000 | 0,000000 |
| 9 | G_3:3 | 0,000060 | 0,000000 | 0,999940 |
| 10 | G_1:1 | 1,000000 | 0,000000 | 0,000000 |
| 11 | G_3:3 | 0,000071 | 0,000000 | 0,999929 |
| 12 | G_3:3 | 0,000324 | 0,000000 | 0,999676 |
| 13 | G_2:2 | 0,000000 | 1,000000 | 0,000000 |
| 14 | G_2:2 | 0,000000 | 0,999998 | 0,000002 |
| 15 | G_3:3 | 0,000013 | 0,000034 | 0,999954 |
| 16 | G_3:3 | 0,000035 | 0,000000 | 0,999965 |
| 17 | G_3:3 | 0,000002 | 0,000000 | 0,999998 |
| 18 | G_3:3 | 0,004659 | 0,000000 | 0,995341 |
| 19 | G_3:3 | 0,000006 | 0,000000 | 0,999994 |
| 20 | G_3:3 | 0,035335 | 0,000000 | 0,964665 |
| 21 | G_3:3 | 0,000000 | 0,000000 | 1,000000 |
| 22 | G_2:2 | 0,000000 | 0,993147 | 0,006853 |
| 23 | G_3:3 | 0,000014 | 0,000000 | 0,999986 |
| 24 | G_2:2 | 0,000000 | 0,999926 | 0,000074 |
| 25 | G_3:3 | 0,000057 | 0,000000 | 0,999943 |

To provide a fuller description of the obtained clusters we use another procedure for multidimensional clustering called *Two-way joining*. The method is that we group not only the observations(regions) but also the variables(features). The essential in the method is the value for the threshold which simultaneously takes into account proximity between the objects and between the features. Choosing a value of the parameter we try to achieve the grouping of the population into three blocks. Consequently we obtain the *Reordered date matrix*, shown below. The classification allows us to get additional information respect to the features about cluster structure, namely, we can characterize every region not only as included in one or another cluster but by its features.

*Table 2.14. Reordered date matrix*

| Case | (Table) | | | | | |
|---|---|---|---|---|---|---|
| | z1 | z2 | z3 | z4 | z5 | z6 |
| C_1 | -0,23055 | 0,680908 | 0,682747 | -1,27225 | -0,457073 | -0,116734 |
| C_15 | -0,15018 | 0,798702 | 0,802034 | -0,67213 | -0,457073 | -0,116734 |
| C_7 | -1,32324 | -0,191008 | -0,188265 | -1,27225 | -0,725939 | -0,116734 |
| C_2 | -0,20541 | -0,224556 | -0,224066 | -0,67213 | -0,457073 | -0,341221 |
| C_16 | 0,29689 | -0,525395 | -0,526519 | -0,67213 | -0,457073 | -0,116734 |
| C_18 | -0,08823 | -0,886453 | -0,887303 | 0,52811 | -0,457073 | -0,116734 |
| C_6 | -0,27530 | -0,495169 | -0,493887 | -0,67213 | -0,725939 | -0,565709 |
| C_9 | -0,60656 | -0,303585 | -0,305721 | -0,67213 | -0,725939 | -0,790197 |
| C_11 | -0,47469 | -0,887085 | -0,887188 | -0,67213 | -0,591506 | -0,116734 |
| C_23 | -0,45075 | -0,712744 | -0,712325 | -0,67213 | -0,860372 | -0,116734 |
| C_25 | -0,19589 | -0,924049 | -0,924260 | -0,67213 | -0,860372 | -0,341221 |
| C_22 | -0,28716 | -0,518821 | -0,518877 | -0,67213 | 1,559425 | -0,565709 |
| C_24 | -1,05289 | -0,847172 | -0,845741 | -0,07201 | 1,559425 | -0,565709 |
| C_8 | 1,03651 | -0,169157 | -0,170010 | 1,12823 | 1,559425 | -0,116734 |
| C_14 | 0,06107 | -0,657671 | -0,658054 | 0,52811 | 1,559425 | -0,116734 |
| C_13 | -0,04950 | 0,821536 | 0,821296 | 1,72835 | 1,559425 | -0,116734 |
| C_12 | 0,32584 | 0,115797 | 0,114897 | 1,12823 | -0,591506 | 0,107754 |
| C_20 | 0,55739 | 0,600147 | 0,599161 | 1,12823 | -0,457073 | -0,341221 |
| C_17 | -0,66173 | -0,277336 | -0,274822 | 1,12823 | -0,725939 | -0,341221 |
| C_19 | -0,94800 | -0,739578 | -0,740748 | 1,12823 | -0,725939 | -0,565709 |
| C_21 | -0,61434 | -0,728325 | -0,728352 | 1,72835 | -0,591506 | -0,341221 |
| C_3 | -0,84242 | -0,476967 | -0,478157 | 1,12823 | 1,559425 | 3,250583 |
| C_4 | 1,19823 | 1,506524 | 1,504555 | -1,27225 | 1,559425 | 3,250583 |
| C_5 | 1,44835 | 2,069388 | 2,064151 | -0,67213 | -0,457073 | -0,341221 |
| C_10 | 3,53257 | 2,972068 | 2,975454 | -0,67213 | -0,591506 | -0,341221 |

Let us make the economic-mathematical analysis for every obtained cluster.
1) the “first cluster” include the most favorable for the business regions: Kyiv region, Donetsk region, Dnipro region. It’s worth for noting that for including elements into the cluster values of the first three features are crucial(as it is shown by the diagram of isolines for *Two-way joining*). But the regions are unfavorable for cultivation of the necessary plants. Hence the circumstances shall be taken into account.
2) the “second cluster” include favorably for the business regions:Volyn, Zaporizhia, Lviv, Mykolayiv, Khmelnytsky and Chernivtsi regions. The lands of the regions are very suitable for the cultivation. Also the regions are attractive by absence of restrictions on the level of profit.
3) the “third cluster” include the regions: Crimea, Vinnytsia, Zhytomyr, Ivano-Frankivsk, Kirovohrad, Luhansk, Odessa, Poltava, Rivne, Sumy, Ternopil, Kharkiv, Kherson, Cherkasy and Chernihiv regions. For the elements is essential more or less even influence of oll features. The cluster show that most of Ukrainian regions are favorable for the business.

The multidimensional modeling we made allows us to plan our business within formed clusters not only to minimize costs but also to run the business more effectively.

*Example 2.3. Cluster analysis of Districts of the Mykolayiv region of revenue collection from local property taxes*

Now, Ukrainian politicians and scientists consider property taxes as one of the important issues of money to local budgets. Hence we conduct cluster analysis of the districts of the Mykolayiv region of revenue collection from local property taxes (a land tax, vehicle tax) to a local budgets(general and special funds respectively) for 2006-2009 years.

The Feature Space is 4-dimensional one. We choose such indicators:

$x_{1j}$ is the land tax revenue (thousands of UAH) into a local budget of the j-th district ( j=1,2,…,24) for a corresponding year.
$x_{2j}$ is the vehicle tax revenue (thousands of UAH) into a local budget of the j-th district ( j=1,2,…,24) for a corresponding year.
$x_{3j}$ is the income to a general fund of a local budget of the j-th district ( j=1,2,…,24) for a corresponding year.
$x_{4j}$ is the income to a special fund of a local budget of the j-th district ( j=1,2,…,24) for a corresponding year.

It's worth for noting that we don't take into account immediate income to the budget of the region, although the income is a significant part of the region income as whole (for a land tax is near 25%, for a vehicle tax is more than 40%, to a general fund is more than 23% and to a special fund average 25%).
We make standardization putting

$$z_{mj} = \frac{x_{mj}}{x_m} \quad (m = 1, 2, 3, 4; \; j = 1, 2, ..., 24)$$

Where $x_m$ is a land tax revenue and an income to the funds to the local budgets in the region as whole for a corresponding year. The values of the standardized indicators is represented by a matrix $(24 \times 4)$. The rows of the latter correspond to the districts (c1, c2,...,c24), the columns, respectively correspond to the features.
The modeling was based on the matrices of standardized values provided by the program *STATISTICA* using its module prescribed to execute the clustering procedure. We was provided by dendrogram, that shows the hierarchy of similarities among all pairs of objects. Also we get the graph which demonstrates how the elements were joined (for example see Fig 2.6 which show hierarchical dendrogram classification the districts and step-by-step graphics of including the objects into the clusters according to the data for 2009 year.).

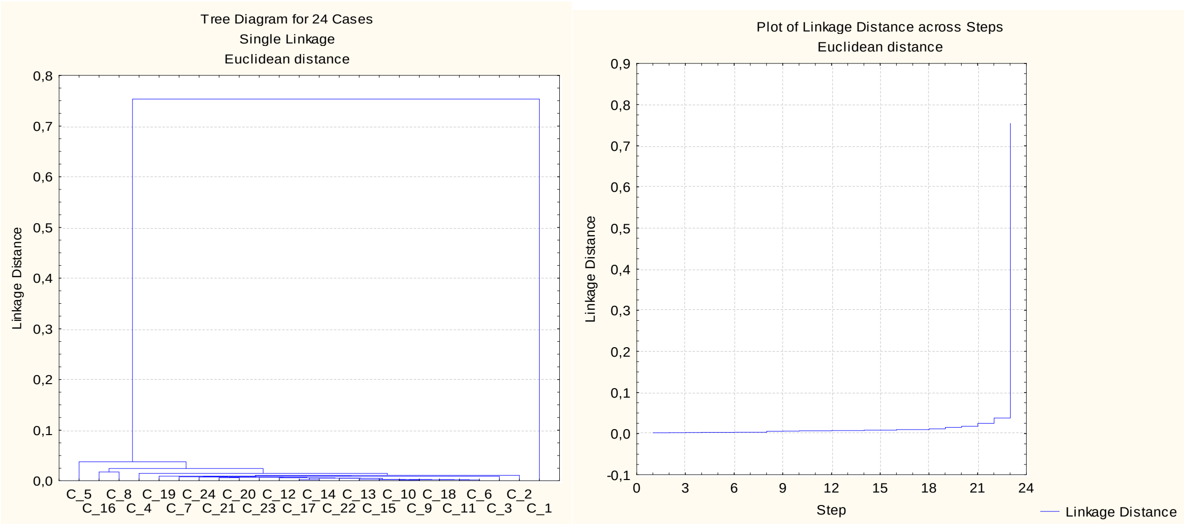



*Fig. 2.6. Hierarchical dendrogram and step-by-step graphics of including the objects into the clusters*

Analyzing the results acquired in the first step lead to the conclusion that the distrcts of the Mykolayiv region may be included into the four clusters (groups)
1) the Mykolayiv city 2) "leaders; 3) "averages"; 4) "outsiders".
To clarify the results we use iteration clustering namely K-means method. The method groups using the principle "the nearest neighbor" to form the clusters near centroids (points whose coordinates are the group means of features). At the same time we stated that number of the clusters have to be equal to 4(as it was shown by the previous hierarchical analysis). As it's known the iterations of the K-means method minimizes the variances within the groups, maximizing by that the similarity within the corresponding clusters. As a result the objects(districts) were grouped into four clusters(see below the results table) where the columns from first to fourth describe standardized data, the fifth column contain district numbers, the second-to-last one contains cluster number, and the last one contains "distances" from objects to the centroids of the clusters.

*Table 2.15. Distribution of object for the clusters for 2009 year.*

| | Table.sta | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1<br>mzem09 | 2<br>mtran09 | 3<br>mzf09 | 4<br>msf09 | 5<br>Case_N | 6<br>Cluster | 7<br>Dist. |
| C_1 | 0,408995 | 0,340098 | 0,434206 | 0,445691 | 1 | 1 | 0,00 |
| C_2 | 0,017839 | 0,022778 | 0,022628 | 0,022713 | 2 | 2 | 0,01 |
| C_3 | 0,014482 | 0,005559 | 0,009686 | 0,024009 | 3 | 3 | 0,01 |
| C_4 | 0,026343 | 0,032351 | 0,026356 | 0,01684 | 4 | 2 | 0,01 |
| C_5 | 0,043439 | 0,021557 | 0,056919 | 0,009458 | 5 | 2 | 0,02 |
| C_6 | 0,006481 | 0,00745 | 0,009188 | 0,003604 | 6 | 4 | 0,00 |
| C_7 | 0,023889 | 0,013968 | 0,018086 | 0,006294 | 7 | 3 | 0,01 |
| C_8 | 0,025698 | 0,008949 | 0,013791 | 0,044944 | 8 | 2 | 0,01 |
| C_9 | 0,006861 | 0,005672 | 0,007379 | 0,002953 | 9 | 4 | 0,00 |
| C_10 | 0,007034 | 0,004464 | 0,007132 | 0,002208 | 10 | 4 | 0,00 |
| C_11 | 0,00685 | 0,008088 | 0,007602 | 0,003372 | 11 | 4 | 0,00 |
| C_12 | 0,013454 | 0,007911 | 0,009661 | 0,01172 | 12 | 3 | 0,00 |
| C_13 | 0,004783 | 0,005465 | 0,005755 | 0,003534 | 13 | 4 | 0,00 |
| C_14 | 0,013673 | 0,008417 | 0,008613 | 0,003673 | 14 | 4 | 0,00 |
| C_15 | 0,00898 | 0,004916 | 0,005828 | 0,003256 | 15 | 4 | 0,00 |
| C_16 | 0,022842 | 0,017076 | 0,028796 | 0,046547 | 16 | 2 | 0,01 |
| C_17 | 0,012489 | 0,007053 | 0,009325 | 0,002834 | 17 | 4 | 0,00 |
| C_18 | 0,004682 | 0,008318 | 0,008458 | 0,002881 | 18 | 4 | 0,00 |
| C_19 | 0,01834 | 0,0151 | 0,015863 | 0,019336 | 19 | 3 | 0,00 |
| C_20 | 0,010347 | 0,012709 | 0,010831 | 0,007533 | 20 | 4 | 0,00 |
| C_21 | 0,014118 | 0,014115 | 0,012615 | 0,011899 | 21 | 3 | 0,00 |
| C_22 | 0,011018 | 0,005265 | 0,006709 | 0,007884 | 22 | 4 | 0,00 |
| C_23 | 0,010565 | 0,009239 | 0,012121 | 0,01717 | 23 | 3 | 0,00 |
| C_24 | 0,016798 | 0,011134 | 0,01623 | 0,006504 | 24 | 3 | 0,00 |

To provide a fuller description of the obtained clusters we use another procedure for multidimensional clustering called *Two-way joining*. The method is that we group not only the observations(regions) but also the variables(features). The essential in the method is the value for the threshold which simultaneously takes into account proximity between the objects and between the features. Choosing a value of the parameter we try to achieve the grouping of the population into three(we exclude the “abnormal” cluster Mikolaiv city) blocks. Consequently we obtain the *Reordered date matrix,* shown below. The classification allows us to get additional information respect to the features about cluster structure, namely, we can characterize every district not only as included in one or another cluster but by its features.

*Table 2.14. Reordered date matrix*

| Case | Reordered matrix (Table.sta) | | | |
|---|---|---|---|---|
| | mzem09 | mzf09 | mtran09 | msf09 |
| C_2 | 0,017839 | 0,022628 | 0,022778 | 0,022713 |
| C_19 | 0,018340 | 0,015863 | 0,015100 | 0,019336 |
| C_4 | 0,026343 | 0,026356 | 0,032351 | 0,016840 |
| C_3 | 0,014482 | 0,009686 | 0,005559 | 0,024009 |
| C_23 | 0,010565 | 0,012121 | 0,009239 | 0,017170 |
| C_6 | 0,006481 | 0,009188 | 0,007450 | 0,003604 |
| C_11 | 0,006850 | 0,007602 | 0,008088 | 0,003372 |
| C_18 | 0,004682 | 0,008458 | 0,008318 | 0,002881 |
| C_14 | 0,013673 | 0,008613 | 0,008417 | 0,003673 |
| C_17 | 0,012489 | 0,009325 | 0,007053 | 0,002834 |
| C_9 | 0,006861 | 0,007379 | 0,005672 | 0,002953 |
| C_10 | 0,007034 | 0,007132 | 0,004464 | 0,002208 |
| C_13 | 0,004783 | 0,005755 | 0,005465 | 0,003534 |
| C_15 | 0,008980 | 0,005828 | 0,004916 | 0,003256 |
| C_12 | 0,013454 | 0,009661 | 0,007911 | 0,011720 |
| C_21 | 0,014118 | 0,012615 | 0,014115 | 0,011899 |
| C_20 | 0,010347 | 0,010831 | 0,012709 | 0,007533 |
| C_22 | 0,011018 | 0,006709 | 0,005265 | 0,007884 |
| C_7 | 0,023889 | 0,018086 | 0,013968 | 0,006294 |
| C_24 | 0,016798 | 0,016230 | 0,011134 | 0,006504 |
| C_8 | 0,025698 | 0,013791 | 0,008949 | 0,044944 |
| C_16 | 0,022842 | 0,028796 | 0,017076 | 0,046547 |
| C_5 | 0,043439 | 0,056919 | 0,021557 | 0,009458 |

Let us perform economic and mathematical analysis of each of the obtained clusters:

1) Mykolaiv city, for which the proceeds of the land tax is almost 40.9% of the total (or 68689.191 thousand UAH in absolute terms) vehicle tax revenue amounts 34% of the region collected (or 11798.764 thousand UAH ) the the general fund revenues is 43.4% of regional one (or 553,454.88 thousand UAH), and the income to the special fund is 44.5% of regional (or 51082.157 thousand).

Note that the share of income among land tax revenues of the general fund ranges from 6.7% in 2007 to 12.4% in 2009, while the share of revenues among transport fees revenues of the special fund ranges from 15.6% in 2008 to 25 5% in 2006;

2) The "leaders" are Voznesensk, Pervomaisk and Yuzhnoukrainsk towns, Berezansky and Oktyabrsky districts, for which, receipts of land tax are on average 2.7% of the total and range from 1.8% regional (or 2995.936 thousand UAH) Voznesensk town to 4.3% of regional (or 7295.504 thousand UAH) for Yuzhnoukrainsk town; vehicle tax revenue averaged 2% of regional and range from 0.9% regional (or 310.464 thousand) for Berezan district to 3.2% of regional (or 1122.343 thousand ) for Pervomaysk; the volume of general fund revenues averaged 2.9% of the total and range from 1.4% regional (or 17578.836 thousand UAH) for Berezan district to 5.7% regional (or 72551.353 thousand UAH) for Yuzhnoukrainsk; amounts of special fund revenue account for an average of 2.8% of the total and range from 0.9% of regional (or 1084.061 thousand UAH account for an average) for the town Juzhnoukrainsk to 4.7% of regional (or 5334.928 thousand UAH account for an average) for the October district. Note that this cluster share the proceeds of the land tax among others incomes to general fund revenues in

2009 ranged from 10.4% (Voznesensk town) to 24.6% (Berezan district) and the share of transportation fees revenue r among incomes of special funds ranges from 6% (Berezan district) to 69% (Yuzhnoukrainsk);

3) The "averages" are Ochakiv town, Bashtanka, Voznesensk, Mykolayiv, Novoodeska, Pervomajskiy and Snigirevskaya districts, for which:

receipt of land tax are on average nearly 1.6% of the total and range from 1.1% regional (or 1774,357 thousand UAH) for Pervomaisky district to 2.4% regional (or 4012.102 thousand UAH) for Bashtansky district; vehicle tax revenue average at 1.1% of regional and ranged from 0.6% regional (or 192.842 thousand) for the Ochakiv town to 1.5% regional (or 523.851 thousand. USD) for the Mykolayiv region; the volume of general fund revenues are on average 1.3% of the total and range from 1% regional (or 12314.647 thousand UAH) for the Ascension district to 1.7% regional (or 23053.611 thousand) for Bashtansky district; amounts of special fund revenue account for an average of almost 1.4% of regional and range from 0.6% regional (or 721.407 thousand) for the Bashtansky district to 2.4% of regional (or 2751.724 thousand UAH) for the Ochakiv town. For this cluster share the proceeds of the land tax of general fund revenues in 2009 ranged from 11.5% (the Pervomaisky district) to 19.7% (the Ochakiv town) and the share of revenue among transportation fees revenue of special funds ranges from 7% (the Ochakiv town) to 67.2% in the Bashtanka district;

4) The "outsiders" are Arbuzinsky, Berezneguvate, Bratsk, Veselinovsky, Vradiyivskyy, Domanivsky, Yelanetsky, Kazankivskyi, Kryvoozerskyy, Novobugsky and Ochakiv areas, for which:

receipt of land tax are on average 0.8% of the total and range from 0.5 % regional (or 786.379 thousand UAH) for the Kryve Ozero district to 1.4% regional (or 2296.256 thousand) for the Domanivsky district;

vehicle tax revenue average at 0.7% of regional and ranged from 0.4% regional (or 154.872 thousand UAH) for the Bratsky district to 1.1% regional (or 440.912 thousand UAH) for the Novobugsky distrct; the volume of general fund revenues are on average nearly 0.8% of the total and range from 0.6% regional (or 7335.501 thousand) for the Vradiivka district to 1.1% regional (or 13805.937 thousand ) for Novobugsky area; amounts of special fund revenue account for an average of almost 0.4% of the total and range from 0.2% regional (or 223.274 thousand UAH) for the Bratsky district to 0.8% regional (or 1655.863 thousand) for Ochakiv district.

Note that this cluster share the proceeds of the land tax general fund revenues for 2009 range from 7.3% (the Kryve Ozero district) to 21.6% (the Ochakiv district) and the share of revenue among transportation fees revenue of special funds ranges from 20.2% % (the Ochakiv district) to 87.4 (the Kryve Ozero district).

Conducted multidimensional classification leads to the following conclusions:

- the structure of revenues from land tax (increase in the share of rent) is somewhat "distorted", and this leads to the fact that the largest revenue go to local budgets of cities. On average revenues of the districts from land tax up 14.6% of total assets, and, for most "leaders" the figure is below average, while the other is usually higher than the average;

- We have to take into account identified above clusters for planning, forecasting, monitoring and controlling of revenues from property taxes. For example, for the "leaders" we have to carry out a risk-based unused capacity (g.e., the optimistic forecasts), concentrating in the current work focus on control functions, and for the "averages" and "outsiders" we should give priority to preventive work, expansion of the tax base, improving administration etc.

- vehicle tax are on average 44.6% of the revenues of special funds, and for most "average" and "outsiders" form the lion's share (to 87.4%) of all revenues special fund ;

- It is necessary to carry out further economic modeling of property taxes on the designated clusters that will not only reduce costs but also to obtain more precise results (to be continued Dynamics revenue property taxes in clusters).

*Example 2.4. Cluster analysis of the competitiveness of the Odessa Region food industry enterprises*

It is known that the activity any enterprises is characterized by a lot of indicators (indications) that influence various aspects of this activity. Some of these indications cannot be measured directly (so-called *latent indications*) but there are manifested as a result of acquisition of specific numerical values for several interrelated indication-symptoms. One of these latent indications is level of competitiveness of economic entities: regions, enterprises and others. The concept of competitiveness is obviously relative integral characteristic. It envisages the comparison of the enterprise with its competitors through a lot of indicators their activities.

Therefore, necessary to apply methods and models of multivariate statistical analysis to quantitative investigation these of integral indicators.

We consider the data on the activities of the 24 Odessa region food industry enterprises for 2008-2012 on the following parameters:

*Table 2.17. Indicators of enterprises.*

| | |
|---|---|
| *x1 - materials consumption,%* | *x23 - coefficient of financial stability margin* |
| *x2 - coefficient of the manufacturing cost* | *x24 - turnover ratio of current assets* |
| *x3 - gross income ratio* | *x25 - inventory turnover ratio* |
| *x4 - capital-labor ratio* | *x26 - asset turnover ratio* |
| *x5 - capital productivity ratio* | *x27 - turnover ratio of receivables* |
| *x6 - coefficient of mobility* | *x28 - days sales outstanding* |
| *x7 - the share of productive capacity in assets ,%* | *x29 - coefficient of accounting payable turnover ratio* |
| *x8 - coefficient of wear and tear* | *x30 - payable turnover in day,* |
| *x9 - coefficient of suitability* | *y31 - profitability of production,%* |
| *x10 - coefficient of renewal* | *y32 - profitability ratio,%* |
| *x11 - coefficient of withdraws* | *y33 - operation return on working capital,%* |
| *x12 - growth rate of fixed assets* | *y34 - return on equity,%* |
| *x13 - critical point of sales,* | *y35 - return on fixed assets,%* |
| *x14 - operation leverage,* | *y36 - return on assets,%* |
| *x15 - specific wage,* | *x37 - maneuverability equity ratio* |
| *x16 - glut of finished product,* | *x38 - financial stability coefficient* |
| *x17 - share of long-term assets in total assets,%* | *x39 - debt to equity of equity ratio* |
| *x18 - share of fixed assets in total assets,%* | *x40 - ratio of receivables and payables,* |
| *x19 - share of fixed assets in long-term assets,%* | *x41 - absolute liquidity ratio,* |
| *x20 - coefficient of labor productivity* | *x42 - quick ratio* |
| *x21 - coefficient of internal level of competitiveness* | *x43 - current ratio* |
| *x22 - coefficient of competitiveness margin* | *x44 - equity ratio* |

Note that some indicators (namely $x_j$) act as predictors (regressors, controlled variables, manipulated variables etc.), and the rest (g.e. $y_j$) are explained variables (regressands, predicted variables etc.).

Let us draw classification (dividing into classes or groups - clusters) of the enterprises on the set of regressands $y_j,\ j=\overline{31,36}$. First, we find the matrix of correlation coefficients between the pairs of regressands for all 24 enterprises for 2008-2012:

*Table 2.18. Matrix paired correlation coefficients*

| Variable | Correlations (Table.sta) Market correlations witch significance p <,0500( N=120 (Casewise deleted) | | | | | |
|---|---|---|---|---|---|---|
| | y31 | y32 | y33 | y34 | y35 | y36 |
| y31 | 1,00 | 0,93 | 0,73 | 0,03 | 0,50 | 0,70 |
| y32 | 0,93 | 1,00 | 0,82 | 0,02 | 0,52 | 0,69 |
| y33 | 0,73 | 0,82 | 1,00 | -0,05 | 0,64 | 0,79 |
| y34 | 0,03 | 0,02 | -0,05 | 1,00 | 0,11 | 0,04 |
| y35 | 0,50 | 0,52 | 0,64 | 0,11 | 1,00 | 0,90 |
| y36 | 0,70 | 0,69 | 0,79 | 0,04 | 0,90 | 1,00 |

Analysis of the table shows the presence of a sufficiently close relationship between factors $y_j,\ j=\overline{31,36}$ (besides $y_{34}$), and some of their grouping. This fact must be considered in further studies. Next we consider the dynamic clusters (taking into account changes that have occurred in the five years) into which a set of 24 enterprises may be divided. With clustering, given the relationship between the detected signs of a measure of similarity (distance $d_{ps}$) between $p$-th and $s$-th enterprises we use the Mahalanobis metric (or so-called *1-r Pearson*):

$$d_{ps}=\left(z_p-z_s\right)r^{-1}\left(z_p-z_s\right)^T$$

where $z_p, z_s$ are vectors-rows corresponding to the $p$-th and $s$-th objects in the standardized feature space

$$\left(z_k=\frac{y_k-\overline{y_j}}{\sigma_j},\ j=\overline{31,36}\right)$$

$\overline{y_j}$ is an average value, $\sigma_j$ is a standard deviation of j-th feature, and $r^{-1}$ is the inverse matrix of the pair correlation matrix. Let us perform the distribution of our set of 24 companies on group-dynamic clusters for years, using hierarchical, agglomerative clustering procedure according to the criterion Ward (Ward, Ward's method) which based on an analysis of the increasing variations of the features for all possible association of clusters (which leads to formation of clusters of approximately equal size). In the calculations we use the module "Hierarchical cluster analysis" of the *STATISTICA*. We present the results of multidimensional modeling for the years 2008-2011 on figure below :

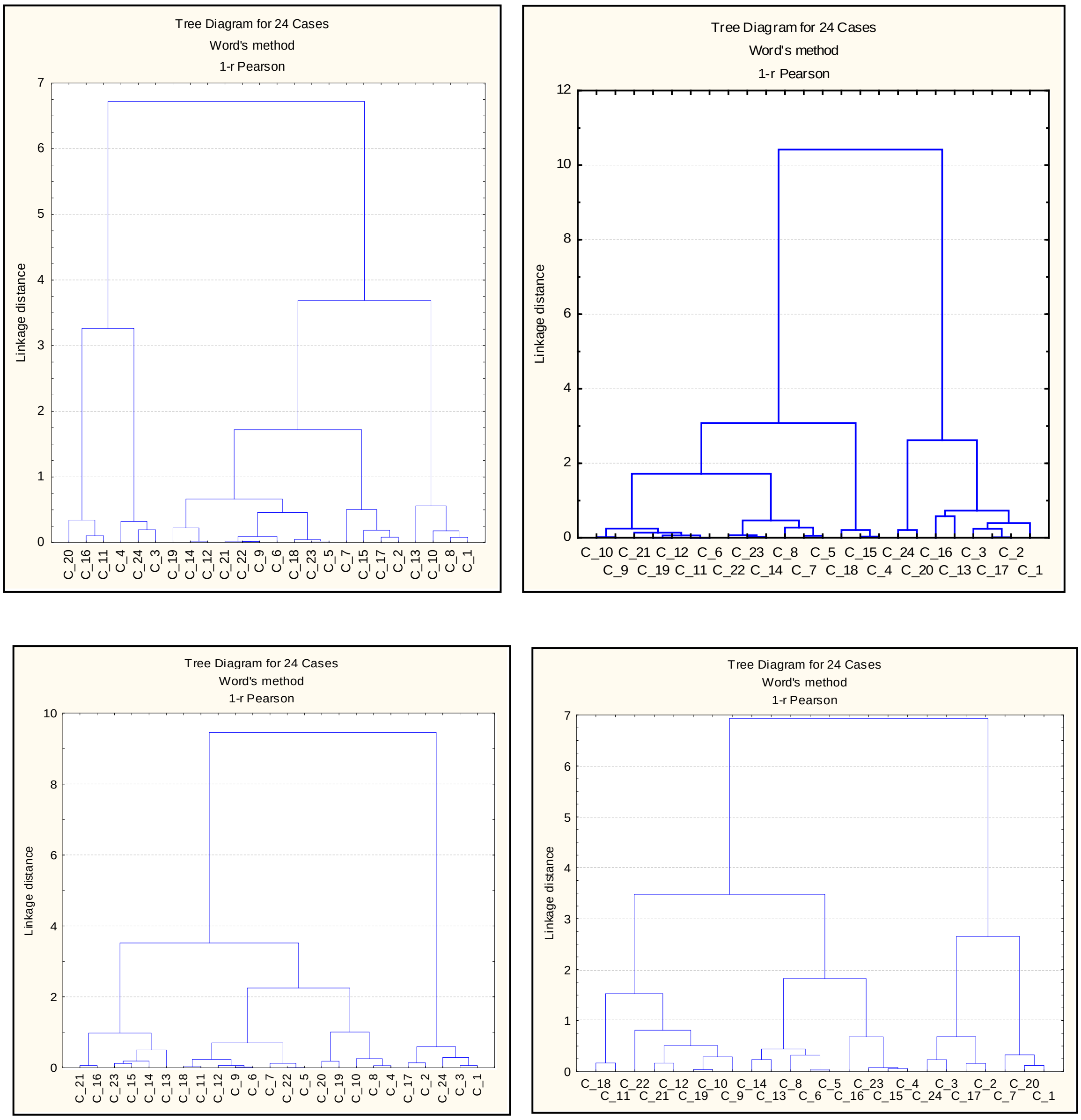



*Fig. 2.7. Dendrograms describe the including companies into clusters respect to their activities (performance indicators* $y_j,\ j=\overline{31,36}$*) for 2008-2011.*

Analysis of the method of hierarchical cluster analysis shows that the proposed set of 24 enterprises (objects) can be divided into 4 groups (clusters) for the indicators ($y_j,\ j=\overline{31,36}$) of their activities. In addition, let us provide the clustering within the years taking into account all features (see figure 2.8):

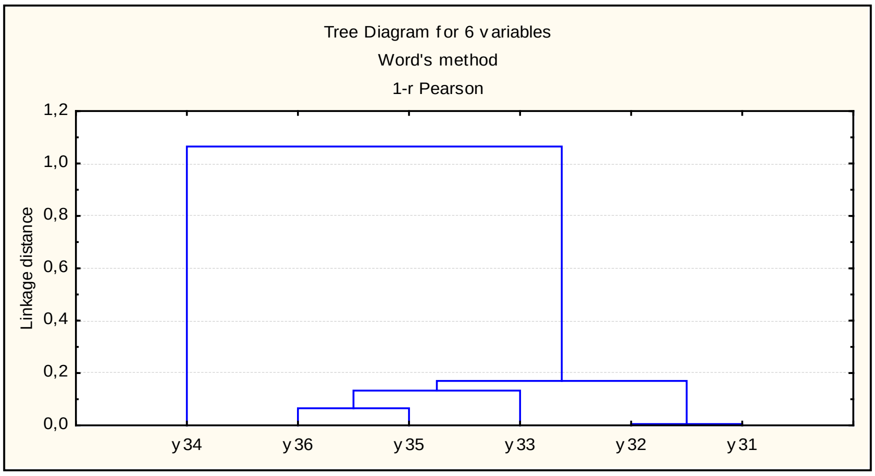



*Fig. 2.8. The dendrogram describing distribution for 2008-2012.*

Analysis of these results shows that the population can be divided into three groups ( "hidden" or so-called latent factors) the first group which contains only $y_{34}$ which is efficiency (profitability) of equity; the second one includes $y_{33}, y_{35}, y_{36}$ The variables describe the operating efficiency of (financial) activities; the third group includes $y_{31}, y_{32}$. The features reflect the efficiency of business (economic activity).

Then we consider the indicators of enterprises according to the results of 2012. First let us apply the procedure hierarchical cluster analysis:

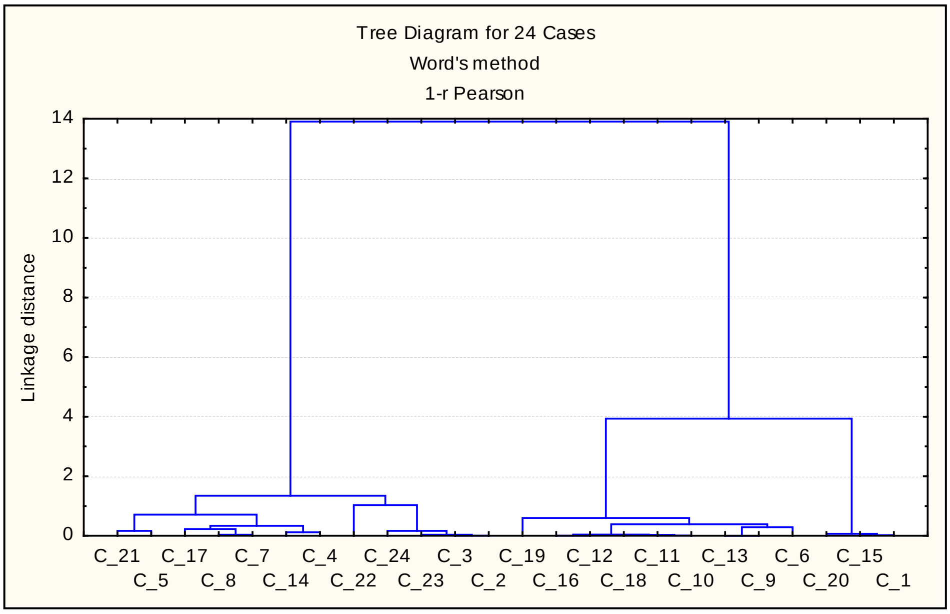



*Fig. 2.9. The dendrogram describes as companies grouped into clusters respect to the results of 2012.*

Then, using the method of K-means, we divide population of 24 enterprises into 4 clusters:

*Table 2.19. Distribution companies together into clusters according to the data of 2012*

| group (cluster) | Number of companies | Industry | name |
|---|---|---|---|
| (S_12){1} | 7 | Wine | PJSC "Dolinka" (S_11) |
| | | | JSC "Odessa factory of sparkling wines" (S_9) |
| | | | PJSC "Southern" group |
| | | Canning | PJSC "Artsyzkyy plant of foodstuffs" (S_16) |
| | | | JSC "Plant" Illichivsk "(S_18) |
| | | | JSC "Odessa baby food cannery" (S_19) |
| | | Dairy and meat | PJSC «Plant Shkirsyrovynnyy" (S_20) |
| (S_10){2} | 5 | Bakery | JSC "Odessa loaf" (S_1) |
| | | vodka | PJSC "Dniester" group |
| | | | PJSC "tenants" (S_13) |
| | | | JSC "Odesavinprom" (S_15) |
| | | Flour | JSC "Belgorod-Dniester bakery" (S_6) |
| (S_21){3} | 7 | Bakery | JSC "Lyubashivka elevator" (S_2) |
| | | Canning | JSC "VO" Odessa cannery "(S_17) |
| | | Dairy and meat | PJSC "Reni meat" group |
| | | | JSC "Odessa meat" (S_22) |
| | | | JSC "Combine" (S_23) |
| | | Flour | PJSC "Razdel'nyanskiy elevator" (S_5) |
| | | | PJSC "Zaplazke Cereal" (S_7) |
| (S_8){4} | 5 | Bakery | JSC "Balta Cereal" (S_3) |
| | | Wine | JSC "Odessa Cognac Factory" group |
| | | | PJSC "Harchovyk" (S_14) |
| | | Dairy and meat | JSC "Yantar" (S_24) |
| | | Flour | PJSC "Aliyahske Cereal" (S_4) |

So, using the latent indicator "level of competitiveness", the investigated population of enterprises is divided into 4 clusters. "leaders" is group {2} "above average" is group {4} "below average" is group {1} and "outsiders" is group {3}. Next we will provide a deeper study of the set of objects (so-called batch methods of principal components analysis).

## 2.2. Factor analysis. The Principal Component Analysis (PCA). The Principal Component Method (batch PLS / PCA-R, SPC)

Factor Analysis is a method for modeling observed variables, and their covariance structure, in terms of a smaller number of underlying unobservable (latent) “factors.” The factors typically are viewed as broad concepts or ideas that may describe an observed phenomenon. For example, a basic desire of obtaining a certain social level might explain most of the consumption behavior. These unobserved factors are more interesting to the social scientist than the observed quantitative measurements.

Factor analysis is generally an exploratory/descriptive method that requires many subjective judgments by the user. It is a widely used tool but can be controversial because the models, methods, and subjectivity are so flexible that debates about interpretations can occur.

The method is similar to principal components although, as the textbook points out, factor analysis is more elaborate. In one sense, factor analysis is an inversion of principal components. In factor analysis, we model the observed variables as linear functions of the “factors.” In principal components, we create new variables that are linear combinations of the observed variables. But in both PCA and FA dimension of the data are reduced. Recall that in PCA interpretation of principal components is often not very clean. A particular variable may, on occasion, contribute significantly to more than one of the components. Ideally, we like each variable to contribute significantly to only one component. A technique called factor rotation is employed towards that goal. Examples of fields in which factor analysis is involved include economics, health, intelligence, sociology, and sometimes ecology and others.

Quite often, when we are modeling complex causal complexes (factor analysis), a so-called "curse of dimension" arises, that is, a problem when exogenous (independent, externally specified) chi variables (predictors) that form the feature space of models with highly correlated (multi-collinear). Factor analysis, as a set of mathematical methods and models with latent indicators, is designed to solve the following problems:

1) Reduce the number of variables describing the described objects.

2) Indirectly quantify latent indicators.

3) Classify variables together with the introduction of more general variables (principal component) by aggregating primary features.

4) The creation or confirmation of the structure of the investigated array of information that is, conducting the search or confirmatory structural analysis.

5) Convert the original data to a form convenient for use or interpretation (e.g., ‘orthogonalization’ of variables for further correlation-regression analysis).

To address these problems and ensure the adequacy of the model we can use the transition from the initial set of signs to the set of uncorrelated variables (*principal components*) substantially smaller dimension that would retain all information concerning the causal mechanism of formation phenomenon or process and does not affect the accuracy of the analysis. The instrument of the change is the *method of principal components (Principal Components Analysis -PCA*). The purpose of the method of principal component - reveal the hidden (latent) root causes that explain the correlation between signs and meaningfully interpreted. It is assumed that $x_i$ signs are indicators of latent properties that are not directly measured. The root cause correlation characteristics *j*-th group called $G_j$ *component*. Features that belong to different groups, uncorrelated, and therefore components $G_j$ independent (orthogonal). The method of principal components use to moves from a large number of the features $x_i$ to minimum number of the most informative $G_j$ component:

$$\underset{i=1,\,2,\,...,\,m}{x_i} \Rightarrow \underset{j=1,\,2,\,...,\,p}{G_j}\text{, where } p << m .$$

Principal components method solves the following problems:

- The identification the component that is providing them with specific content. This task depends on the Feature space *X*. As a rule, it is based on theoretically grounded hypotheses about the nature of the latent properties of the phenomenon. If this hypothesis is not confirmed, then we used to the maximum number of features, relying on the ability of a method to identify such properties. But in this case, the interpretation of the components will be complicated. Thus components are hypothetical quantities, it can only measure them indirectly by means of specially designed models. The model principal component relationship between the primary features and components described as a linear combination $z_i = \sum_1^m a_{ij} G_j$ ,

where $z_i$ is standardized values of *i*-th feature with the single variances;

$m$ is the aggregate variance equal to the number of features;

$a_{ij}$ is the factor weights or *loadings* of *j*-th component to *i*-th feature.

Loadings of $a_{ij}$ describes the relation between density of *i*-th feature and *j*-th component and, as any measure of the density of communication varies within the range [-1, 1].

There are not residues (characteristics) in the model of principal components, that is, a priori assumed that all m components fully explain the aggregate variation (variance) of the feature space.

The square of loading factor $a_{ij}^2$ characterize the contribution of the *j*-th component to the variation of *i*-th feature through the orthogonality conditions the component. Total contribution of *j*-th component to total variance of *m* features are $\lambda_j = \sum_1^m a_{ij}^2$ . That total variance of feature space X can be represented as the sum of the variances of the component $\sum_1^m \lambda_j$ or through factor *loadings*

$$m = \sum_1^m \lambda_j = \sum_1^m \sum_1^m a_{ij}^2 .$$

The scheme of decomposition of the total variance feature space X can be represented as a matrix:

*Table 2.20. Decomposition of variance*

| $G_j$ / $Z_i$ | $G_1$ | $G_2$ | … | $G_m$ | Variance of $z_i$ |
|---|---|---|---|---|---|
| $z_1$ | $a_{11}^2$ | $a_{12}^2$ | … | $a_{1m}^2$ | *1* |
| $z_2$ | $a_{21}^2$ | $a_{22}^2$ | … | $a_{2m}^2$ | *1* |
| $z_3$ | $a_{31}^2$ | $a_{32}^2$ | … | $a_{3m}^2$ | *1* |
| … | … | … | … | … | … |
| $z_m$ | $a_{m1}^2$ | $a_{m2}^2$ | … | $a_{mm}^2$ | *1* |
| Variance *of* $G_j$ | $\lambda_1$ | $\lambda_2$ | … | $\lambda_m$ | $m$ |

Analysis of the matrix by rows shows which components and their weights are forming the variation of *i*-th feature. Each *factor structure* inherent to each feature. The component easier loads the features, the factor structure is simpler. Analysis of the columns of the matrix shows which features are indicators of the j-th component. The components ordered by values dispersions (characteristic (own) numbers matrix of pair correlation coefficients $r_{ik}$ ):

$$\lambda_1 > \lambda_2 > \lambda_3 > ... > \lambda_m .$$

The largest part of the total of variation falls on several first components. Experience has shown that the number of such weighty components is 10-15% of the number of primary signs. They are called the *principal components* and are subject to meaningful interpretation. Thus, the principal component model transforms *m*-dimensional feature space to the *p*-dimensional feature- space of components ($p << m$). The total variance of the principal components significantly less than the total variance of feature space, and the ratio $\frac{\sum_1^p \lambda_j}{m}$ describes the completeness of the factorization. That is, the exploration of the components is the classic problem of characteristic numbers λ and V vectors that is corresponded to correlation matrix R. The components are deemed as principal, for which:

- $\lambda_j > 1$ by the *Kaiser Criteriaa* ;
- completeness of factorization not less 70%.

To satisfy the condition $\lambda_j = \sum_1^m a_{ij}^2$ , the characteristic vector is normalized

$$a_{ij} = V_{ij}\sqrt{\frac{\lambda_j}{\sum_1^m V_{ij}^2}}\,.$$

Therefore, the factor loading of *j*-th component is not anything but a normalized characteristic (own) vector of matrix R.

Procedures of principal components method (*Principal components)* are presented in module *Factor Analysis*. The information base component analysis can be as *Raw data*, and *Correlation matrix*.

So, the building a model of principal components is carried out in three stages:

- Calculation of the correlation matrix R;
- Finding and calculation of principal components factor loadings;
- Identification of the principal components.

Summarize, we can say that the technically, a *principal component* can be defined as a linear combination of optimally-weighted observed variables. In order to understand the meaning of this definition, it is necessary to first describe how subject scores on a principal component are computed. In the course of performing a principal component analysis, it is possible to calculate a score for each subject on a given principal component. For example, in the preceding study, each subject would have scores on two components: one score on the satisfaction with supervision component, and one score on the satisfaction with pay component. The subject's actual scores on the seven questionnaire items would be optimally weighted and then summed to compute their scores on a given component.

*Example 2.5. This example ( by STATISTICA (Examples, Factor.sta data file ) is based on a fictitious data set describing a study of life satisfaction. Suppose that a questionnaire is administered to a random sample of 100 adults. The questionnaire contains 10 items that are designed to measure satisfaction at work, satisfaction with hobbies, satisfaction at home, and general satisfaction in other areas of life. Responses to all questions are recorded via computer and scaled so that the mean for all items is approximately 100.*

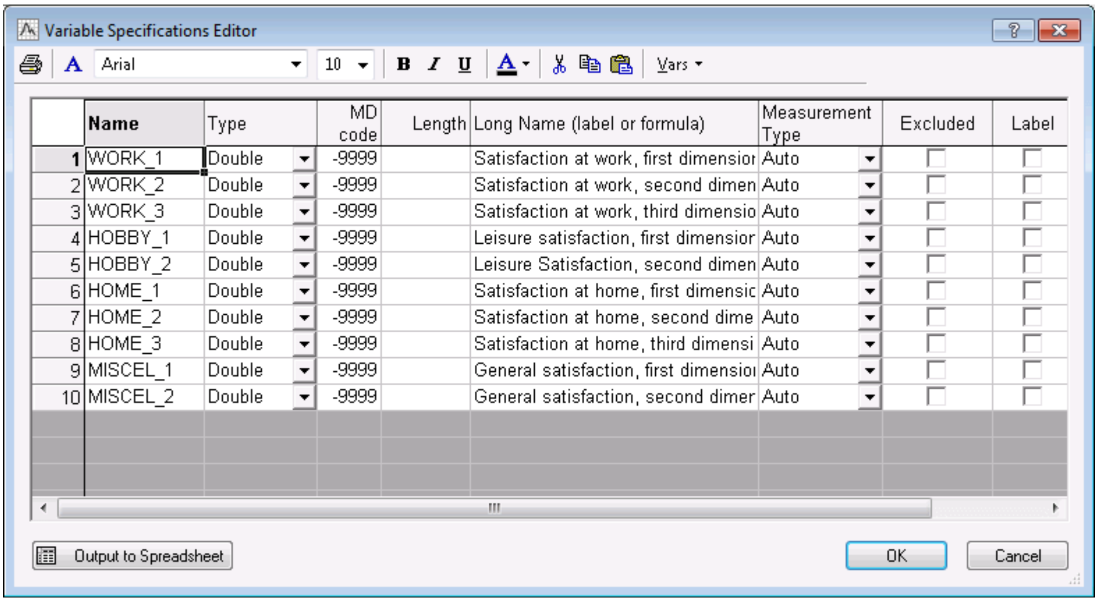
When we open data file, then Ribbon bar. Select the *Home* tab. In the *File* group, click the *Open* arrow and on the menu, select *Open Examples* to display the *Open a STATISTICA Data File* dialog box. Open the *Factor.sta* data file, which is located in the *Datasets* folder.

Classic menus. On the *File* menu, select *Open Examples* to display the *Open a STATISTICA Data File* dialog box. Open the *Factor.sta* data file, which is located in the *Datasets* folder.

Following is a listing of the variables in the data file. To display the *Variable Specifications Editor*:

Ribbon bar. Select the *Data* tab. In the *Variables* group, click *All Specs*.

Classic menus. On the *Data* menu, select *All Variables Specs*.

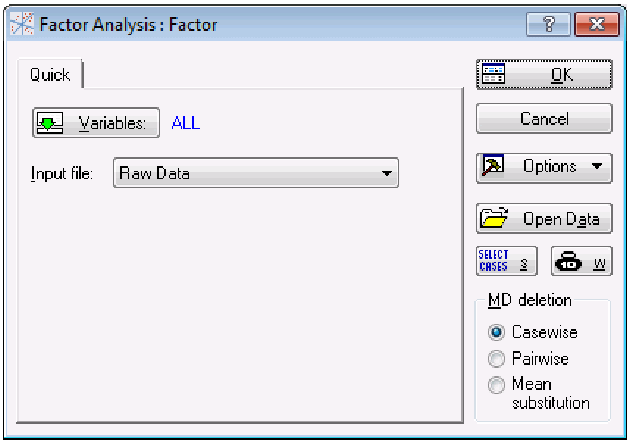



**Purpose of the analysis.** The goal is to learn more about the relationships between satisfaction in the different domains. Specifically, it is desired to learn about the number of factors "behind" these different domains of satisfaction, and their meaning.

**Specifying the analysis.** Ribbon bar. Select the *Statistics* tab. In the *Advanced/Multivariate* group, click *Mult/Exploratory* and on the menu, select *Factor* to display the *Factor Analysis* Startup Panel.

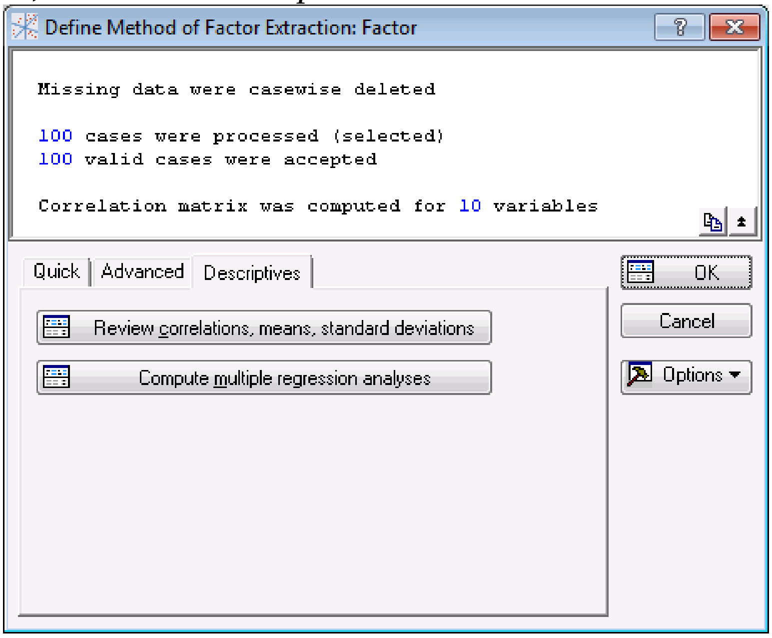
Classic menus. On the *Statistics - Multivariate Exploratory Analysis* submenu, select *Factor Analysis* to display the *Factor Analysis* Startup Panel.

Click the *Variables* button, select all 10 variables, and click the *OK* button. The Startup Panel will look as shown below:

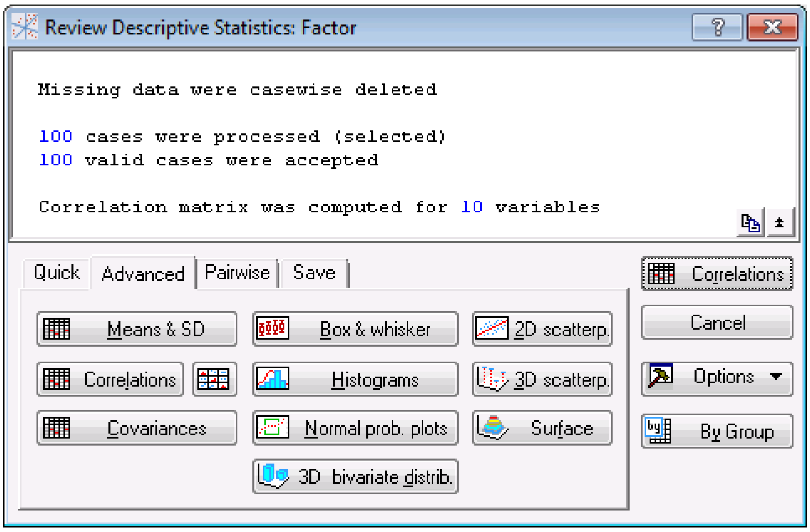



**Other options.** In order to perform a standard factor analysis, this is all that you need to specify in this dialog box. Note that you could also choose either *Casewise* or *Pairwise* deletion, or *Mean substitution* of missing data (via the *MD deletion* group box) or a *Correlation Matrix* data file (via the *Input file* option).

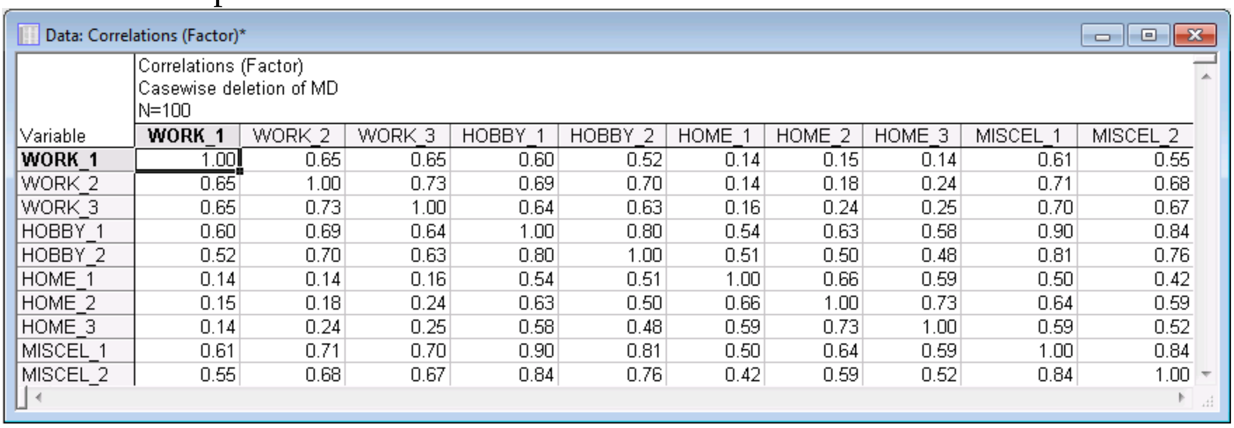
**Define method of factor extraction.** Click the *OK* button to display the *Define Method of Factor Extraction* dialog box. In this dialog box, you can review descriptive statistics, perform a multiple regression analysis, select the extraction method for the factor analysis, select the maximum number of factors and the minimum eigenvalue, and select other options related to specific extraction methods. For now, select the *Descriptives* tab.




**Review descriptive statistics.** Click the *Review correlations, means, standard deviations* button to display the *Review Descriptive Statistics* dialog box. Select the *Advanced* tab.




Here, you can review the descriptive statistics graphically or through spreadsheets.
**Computing correlation matrix.** Click the *Correlations* button to produce the *Correlations* spreadsheet.


Data: Correlations (Factor)*

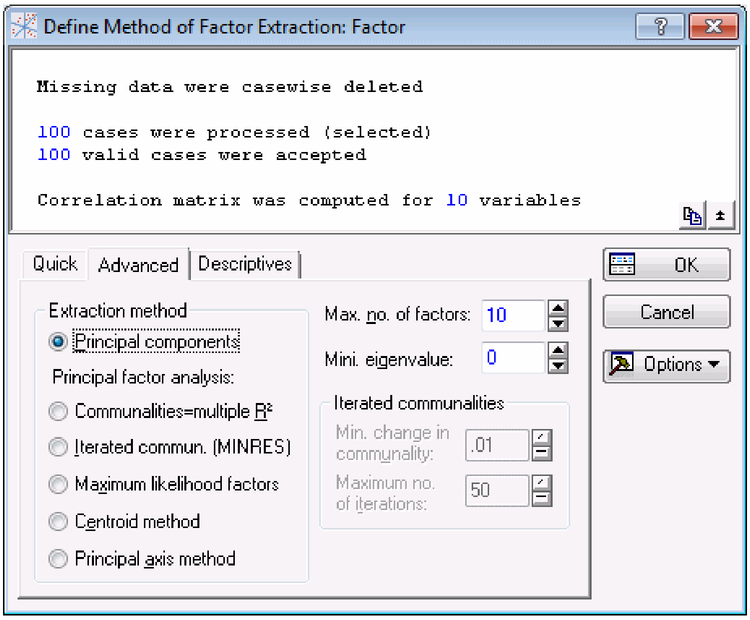
Correlations (Factor)
Casewise deletion of MD
N=100

| Variable | WORK_1 | WORK_2 | WORK_3 | HOBBY_1 | HOBBY_2 | HOME_1 | HOME_2 | HOME_3 | MISCEL_1 | MISCEL_2 |
|---|---|---|---|---|---|---|---|---|---|---|
| WORK_1 | 1.00 | 0.65 | 0.65 | 0.60 | 0.52 | 0.14 | 0.15 | 0.14 | 0.61 | 0.55 |
| WORK_2 | 0.65 | 1.00 | 0.73 | 0.69 | 0.70 | 0.14 | 0.18 | 0.24 | 0.71 | 0.68 |
| WORK_3 | 0.65 | 0.73 | 1.00 | 0.64 | 0.63 | 0.16 | 0.24 | 0.25 | 0.70 | 0.67 |
| HOBBY_1 | 0.60 | 0.69 | 0.64 | 1.00 | 0.80 | 0.54 | 0.63 | 0.58 | 0.90 | 0.84 |
| HOBBY_2 | 0.52 | 0.70 | 0.63 | 0.80 | 1.00 | 0.51 | 0.50 | 0.48 | 0.81 | 0.76 |
| HOME_1 | 0.14 | 0.14 | 0.16 | 0.54 | 0.51 | 1.00 | 0.66 | 0.59 | 0.50 | 0.42 |
| HOME_2 | 0.15 | 0.18 | 0.24 | 0.63 | 0.50 | 0.66 | 1.00 | 0.73 | 0.64 | 0.59 |
| HOME_3 | 0.14 | 0.24 | 0.25 | 0.58 | 0.48 | 0.59 | 0.73 | 1.00 | 0.59 | 0.52 |
| MISCEL_1 | 0.61 | 0.71 | 0.70 | 0.90 | 0.81 | 0.50 | 0.64 | 0.59 | 1.00 | 0.84 |
| MISCEL_2 | 0.55 | 0.68 | 0.67 | 0.84 | 0.76 | 0.42 | 0.59 | 0.52 | 0.84 | 1.00 |

All correlations in this spreadsheet are positive; some correlations are of substantial magnitude. For example, variables *Hobby_1* and *Miscel_1* are correlated at the level of *.90*. Some correlations (for example the ones between work satisfaction and home satisfaction) seem comparatively small. So, it looks like there is some clear structure in this matrix.

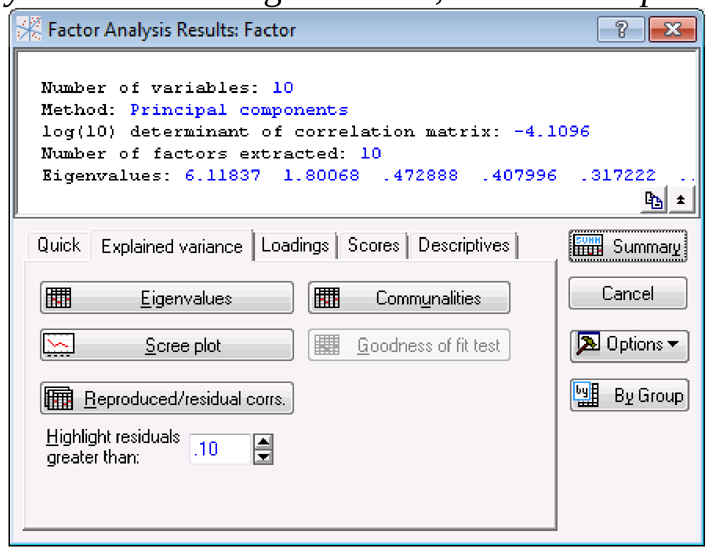
**Extraction method.** In the Review Descriptive Statistics dialog box, click the *Cancel* button to return to the *Define Method of Factor Extraction* dialog box. You can choose from several extraction methods on the *Advanced* tab (see *the Define Method of Factor Extraction - Advanced* tab topic for a description of each method, and the Introductory Overviews for a description of *Principal Components* and *Principal Factors*).

For this example, accept the default extraction method of *Principal components* and change the *Max. no. of factors* to *10* (the maximum number of factors in this example) and the *Mini. eigenvalue* to *0* (the minimum value for this option).




Click the *OK* button to continue the analysis.

**Reviewing results.** You can interactively review the results of the factor analysis in the *Factor Analysis Results* dialog box. First, select the *Explained Variance* tab.



**Reviewing the eigenvalues.** Click the *Eigenvalues* button to produce the spreadsheet of eigenvalues, percent of total variance, cumulative eigenvalues, and cumulative percent.

Data: Eigenvalues (Factor)*

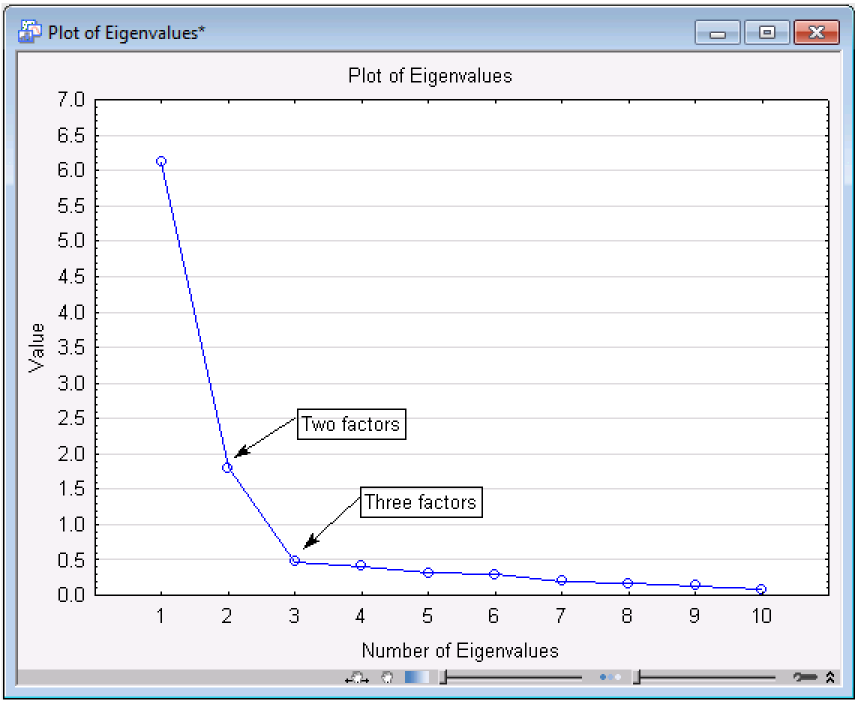
Eigenvalues (Factor)
Extraction: Principal components

| Value | Eigenvalue | % Total variance | Cumulative Eigenvalue | Cumulative % |
|---|---|---|---|---|
| 1 | 6.118369 | 61.18369 | 6.11837 | 61.1837 |
| 2 | 1.800682 | 18.00682 | 7.91905 | 79.1905 |
| 3 | 0.472888 | 4.72888 | 8.39194 | 83.9194 |
| 4 | 0.407996 | 4.07996 | 8.79993 | 87.9993 |
| 5 | 0.317222 | 3.17222 | 9.11716 | 91.1716 |
| 6 | 0.293300 | 2.93300 | 9.41046 | 94.1046 |
| 7 | 0.195808 | 1.95808 | 9.60626 | 96.0626 |
| 8 | 0.170431 | 1.70431 | 9.77670 | 97.7670 |
| 9 | 0.137970 | 1.37970 | 9.91467 | 99.1467 |
| 10 | 0.085334 | 0.85334 | 10.00000 | 100.0000 |

As you can see, the eigenvalue for the first factor is equal to *6.118369*; the proportion of variance accounted for by the first factor is approximately *61.2%*. Note that these values happen to be easily comparable here because there are 10 variables in the analysis, and thus the sum of all eigenvalues is equal to 10. The second factor accounts for about *18%* of the variance. The remaining eigenvalues each account for less than *5%* of the total variance.

**Deciding on the number of factors.** We have described how these eigenvalues can be used to decide how many factors to retain, that is, to interpret. According to the *Kaiser criterion* (Kaiser, 1960), you would retain factors with an eigenvalue greater than 1. Based on the eigenvalues in the *Eigenvalues* spreadsheet shown above, that criterion would suggest you choose 2 factors.

**Scree test.** Now, to produce a line graph of the eigenvalues in order to perform Cattell's scree test (Cattell, 1966), click the *Scree plot* button. The *Plot of Eigenvalues* graph shown below has been "enhanced" to clarify the test. Based on Monte Carlo studies, Cattell suggests that the point where the continuous drop in eigenvalues levels off suggests the cutoff, where only random "noise" is being extracted by additional factors. In our example, that point could be at factor 2 or factor 3 (as indicated by the arrows). Therefore, you should try both solutions and see which one will yield the most interpretable factor pattern.




Now, examine the factor loadings.

**Factor loadings.** Factor loadings can be interpreted as the correlations between the factors and the variables. Thus, they represent the most important information on which the interpretation of factors is based.

First look at the (unrotated) factor loadings for all 10 factors. In the *Factor Analysis Results* dialog box, select the *Loadings* tab.

On the *Factor rotation* drop-down list, select *Unrotated*, and then click the *Summary: Factor Loadings* button to produce the *Factor Loadings* spreadsheet of loadings.

Data: Factor Loadings (Unrotated) (Factor)*

Factor Loadings (Unrotated) (Factor)
Extraction: Principal components
(Marked loadings are >.700000)

| Variable | Factor 1 | Factor 2 | Factor 3 | Factor 4 | Factor 5 | Factor 6 | Factor 7 | Factor 8 | Factor 9 | Factor 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| **WORK_1** | -0.652601 | 0.514217 | 0.301687 | 0.439108 | -0.013701 | 0.127061 | 0.051149 | -0.022306 | 0.080008 | 0.003894 |
| WORK_2 | -0.756976 | 0.494770 | -0.078826 | -0.211795 | -0.090859 | 0.172261 | -0.236057 | 0.163030 | 0.103633 | 0.012210 |
| WORK_3 | -0.745706 | 0.456680 | -0.104749 | 0.030826 | -0.204913 | -0.422007 | 0.033818 | 0.019468 | -0.017932 | 0.038980 |
| HOBBY_1 | -0.941630 | -0.021835 | 0.012653 | 0.001861 | 0.120655 | 0.093740 | -0.023837 | 0.001553 | -0.243305 | 0.171990 |
| HOBBY_2 | -0.875615 | 0.051643 | 0.099675 | -0.324541 | -0.015852 | 0.091249 | 0.312973 | -0.025504 | 0.088684 | 0.017996 |
| HOME_1 | -0.576062 | -0.604977 | 0.490999 | -0.114927 | -0.112513 | -0.114838 | -0.145100 | -0.023820 | 0.004027 | -0.019576 |
| HOME_2 | -0.671289 | -0.617962 | -0.125776 | 0.159963 | 0.225012 | -0.103741 | 0.028392 | 0.201046 | 0.145372 | 0.048318 |
| HOME_3 | -0.641532 | -0.573925 | -0.268572 | 0.152709 | -0.362524 | 0.159987 | 0.011420 | -0.079979 | 0.006890 | 0.000902 |
| MISCEL_1 | -0.951516 | 0.013513 | -0.050164 | 0.026706 | 0.076795 | 0.012644 | 0.035938 | 0.095864 | -0.156713 | -0.223847 |
| MISCEL_2 | -0.900333 | 0.048154 | -0.151805 | -0.034832 | 0.226647 | -0.050720 | -0.120513 | -0.292831 | 0.087690 | -0.030324 |
| Expl.Var | 6.118369 | 1.800682 | 0.472888 | 0.407996 | 0.317222 | 0.293300 | 0.195808 | 0.170431 | 0.137970 | 0.085334 |
| Prp.Totl | 0.611837 | 0.180068 | 0.047289 | 0.040800 | 0.031722 | 0.029330 | 0.019581 | 0.017043 | 0.013797 | 0.008533 |

Remember that factors are extracted so that successive factors account for less and less variance . Therefore, it is not surprising to see that the first factor shows most of the highest loadings. Also note that the sign of the factor loadings only counts insofar as variables with opposite loadings on the same factor relate to that factor in opposite ways. However, you could multiply all loadings in a column by -1 (i.e., reverse all signs), and the results would not be affected in any way.

***Note:*** Computation details for Explained Variance and Proportion of Total Variance: Explained Variance for a given factor is the square of the loadings across the variables for the given factor. Proportion of Total Variance is Explained Variance divided by the Total Variance in the data set.

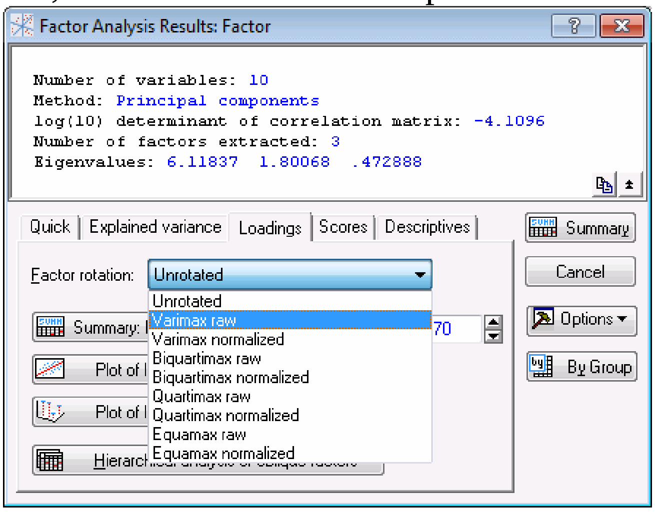
In *Factor Analysis*, the analysis works on the correlation matrix or equivalently the standardized variables so that each variable is a variance of 1, thus, you can divide the explained variance by the total number of variables to get the proportion of total variance.

**Rotating the factor solution.** As described in the Introductory Overviews, the actual orientation of the factors in the factorial space is arbitrary, and all rotations of factors will reproduce the correlations equally well. This being the case, it seems natural to rotate the factor solution to yield a factor structure that is simplest to interpret; in fact, the formal term *simple structure* was coined and defined by Thurstone (1947) to basically describe the condition when factors are marked by high loadings for some variables, low loadings for others, and when there are few high cross-loadings, that is, few variables with substantial loadings on more than one factor. The most standard computational method of rotation to bring about simple structure is the *varimax* rotation (Kaiser, 1958); others that have been proposed are *quartimax*, *biquartimax*, and *equamax* (see Harman, 1967) and are implemented in Statistica.

**Specifying a rotation.** First, consider the number of factors that you want to rotate, that is, retain and interpret. It was previously decided that two is most likely the appropriate number of factors; however, based on the results of the Scree plot, it was also decided to look at the three-factor solution. We will start with three factors.

In the *Factor Analysis Results* dialog box, click the *Cancel* button to return to the *Define Method of Factor Extraction* dialog box, and change the *Maximum no. of factors* on the *Quick* tab from *10* to *3*. Then click the *OK* button to continue with the analysis.

On the *Factor Analysis Results* dialog box - *Loadings* tab, on the *Factor rotation* drop-down list, select *Varimax Raw* to perform a varimax rotation.




Click the *Summary: Factor loadings* button to produce the *Factor loadings* spreadsheet.

Data: Factor Loadings (Varimax raw) (F...

Factor Loadings (Varimax raw) (Factor)
Extraction: Principal components
(Marked loadings are >.700000)

| Variable | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| WORK_1 | 0.839579 | -0.157384 | 0.227287 |
| WORK_2 | 0.898615 | 0.118837 | -0.048899 |
| WORK_3 | 0.865608 | 0.151923 | -0.057003 |
| HOBBY_1 | 0.731037 | 0.501711 | 0.318078 |
| HOBBY_2 | 0.726495 | 0.371499 | 0.336895 |
| HOME_1 | 0.099696 | 0.426704 | 0.864238 |
| HOME_2 | 0.148303 | 0.823540 | 0.384856 |
| HOME_3 | 0.147422 | 0.857607 | 0.236350 |
| MISCEL_1 | 0.758585 | 0.518368 | 0.252834 |
| MISCEL_2 | 0.736270 | 0.524719 | 0.136161 |
| Expl.Var | 4.495100 | 2.591518 | 1.305320 |
| Prp.Totl | 0.449510 | 0.259152 | 0.130532 |

**Reviewing the three-factor rotated solution.** In the *Factor loadings* spreadsheet, substantial loadings on the first factor appear for all but the home-related items. *Factor 2* shows fairly substantial factor loadings for all but the work-related satisfaction items. *Factor 3* only has one substantial loading for variable *Home_1*. The fact that only one variable shows a high loading on the third factor makes us wonder whether we cannot do just as well without it (the third factor).

**Reviewing the two-factor rotated solution.** Once again, click the *Cancel* button in the *Factor Analysis Results* dialog box to return to the *Define Method of Factor Extraction* dialog box. Change the *Maximum no. of factors* on the *Quick* tab from *3* to *2* and click the *OK* button to continue to the *Factor Analysis Results* dialog box. Again select the *Loadings* tab, and on the *Factor rotation* drop-down list, select *Varimax raw*. Click the *Summary: Factor loadings* button.

Data: Factor Loadings (...

Factor Loadings (Varimax
Extraction: Principal com
(Marked loadings are >.7(

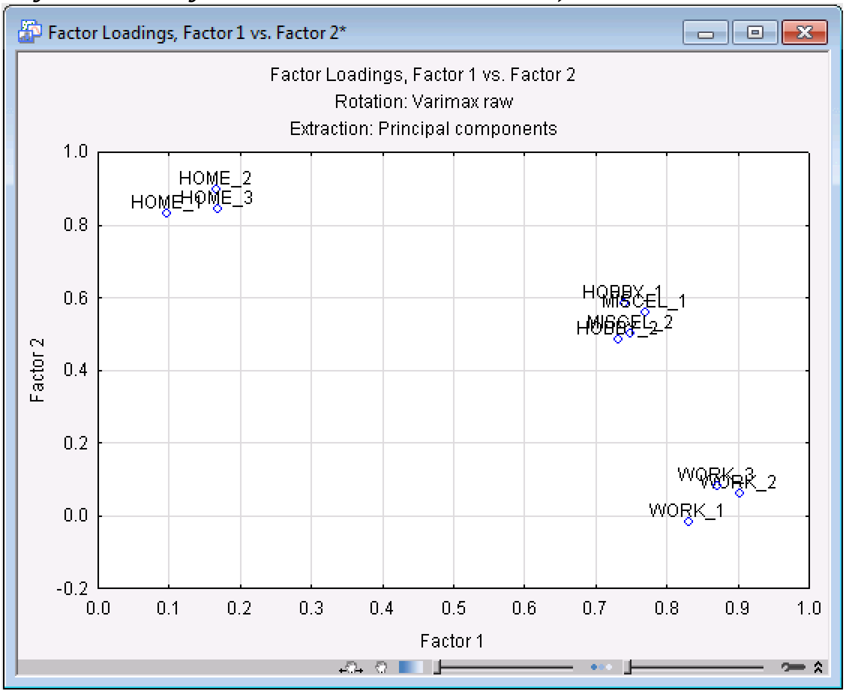
| Variable | Factor 1 | Factor 2 |
|---|---|---|
| WORK_1 | 0.830623 | -0.019320 |
| WORK_2 | 0.902408 | 0.058905 |
| WORK_3 | 0.870524 | 0.082595 |
| HOBBY_1 | 0.739857 | 0.582885 |
| HOBBY_2 | 0.731191 | 0.484489 |
| HOME_1 | 0.097371 | 0.829676 |
| HOME_2 | 0.165722 | 0.897242 |
| HOME_3 | 0.168370 | 0.844159 |
| MISCEL_1 | 0.768988 | 0.560555 |
| MISCEL_2 | 0.748861 | 0.502121 |
| Expl.Var | 4.561544 | 3.357507 |
| Prp.Totl | 0.456154 | 0.335751 |

*Factor 1* shows the highest loadings for the items pertaining to work-related satisfaction. The smallest loadings on that factor are for home-related satisfaction items. The other loadings fall in-between. *Factor 2* shows the highest loadings for the home-related satisfaction items, lowest loadings for work-related satisfaction items, and loadings in-between for the other items.

**Interpreting the two-factor rotated solution.** Does this pattern lend itself to an easy interpretation? It looks like the two factors are best identified as the work satisfaction factor (*Factor 1*) and the home satisfaction factor (*Factor 2*).

Satisfaction with hobbies and miscellaneous other aspects of life seem to be related to both factors. This pattern makes some sense in that satisfaction at work and at home may be independent from each other in this sample, but both contribute to leisure time (hobby) satisfaction and satisfaction with other aspects of life.

**Plot of the two-factor rotated solution.** Click the *Plot of loadings, 2D* button on the *Factor Analysis Results* dialog box - *Loadings* tab to produce a scatterplot of the two factors. The graph simply shows the two loadings for each variable. Note that this scatterplot nicely illustrates the two independent factors and the 4 variables (*Hobby_1*, *Hobby_2*, *Miscel_1*, *Miscel_2*) with the cross-loadings.




Now we will see how well we can reproduce the observed correlation matrix from the two-factor solution.

**Reproduced and residual correlation matrix.** Select the *Explained variance* tab, and click the *Reproduced/residual corrs.* button to produce two spreadsheets with the reproduced correlation matrix and the residual correlations (observed minus reproduced correlations).

Data: Residual Correlations (Factor)*

Residual Correlations (Factor)
Extraction: Principal components
(Marked residuals are > .100000)

| Variable | WORK_1 | WORK_2 | WORK_3 | HOBBY_1 | HOBBY_2 | HOME_1 | HOME_2 | HOME_3 | MISCEL_1 | MISCEL_2 |
|---|---|---|---|---|---|---|---|---|---|---|
| WORK_1 | 0.31 | -0.10 | -0.07 | -0.01 | -0.08 | 0.08 | 0.02 | 0.01 | -0.02 | -0.06 |
| WORK_2 | -0.10 | 0.18 | -0.06 | -0.01 | 0.01 | 0.01 | -0.02 | 0.03 | -0.02 | -0.02 |
| WORK_3 | -0.07 | -0.06 | 0.24 | -0.06 | -0.05 | 0.01 | 0.02 | 0.04 | -0.02 | -0.02 |
| HOBBY_1 | -0.01 | -0.01 | -0.06 | 0.11 | -0.02 | -0.02 | -0.01 | -0.03 | 0.01 | -0.00 |
| HOBBY_2 | -0.08 | 0.01 | -0.05 | -0.02 | 0.23 | 0.03 | -0.06 | -0.05 | -0.02 | -0.04 |
| HOME_1 | 0.08 | 0.01 | 0.01 | -0.02 | 0.03 | 0.30 | -0.10 | -0.13 | -0.04 | -0.06 |
| HOME_2 | 0.02 | -0.02 | 0.02 | -0.01 | -0.06 | -0.10 | 0.17 | -0.05 | 0.01 | 0.02 |
| HOME_3 | 0.01 | 0.03 | 0.04 | -0.03 | -0.05 | -0.13 | -0.05 | 0.26 | -0.02 | -0.03 |
| MISCEL_1 | -0.02 | -0.02 | -0.02 | 0.01 | -0.02 | -0.04 | 0.01 | -0.02 | 0.09 | -0.02 |
| MISCEL_2 | -0.06 | -0.02 | -0.02 | -0.00 | -0.04 | -0.06 | 0.02 | -0.03 | -0.02 | 0.19 |

The entries in the *Residual Correlations* spreadsheet can be interpreted as the "amount" of correlation that cannot be accounted for with the two factor solution. Of course, the diagonal elements in the matrix contain the standard deviation that cannot be accounted for, which is equal to the square root of one minus the respective communalities for two factors (remember that the communality of a variable is the variance that can be explained by the respective number of factors). If you review this matrix carefully you will see that there are virtually no residual correlations left that are greater than *0.1* or less than *-0.1* (actually, a few are about of that magnitude). Added to that is the fact that the first two factors accounted for *79%* of the total variance (see *Cumulative %* eigenvalues displayed in the *Eigenvalues* spreadsheet).

**The "secret" to the perfect example.** The example you have reviewed does indeed provide a nearly perfect two-factor solution. It accounts for most of the variance, allows for ready interpretation, and reproduces the correlation matrix with only minor disturbances (remaining residual correlations). Of course, nature rarely affords one such simplicity, and, indeed, this fictitious data set was generated via the normal random number generator accessible in the spreadsheet formulas. Specifically, two orthogonal (independent) factors were "planted" into the data, from which the correlations between variables were generated. The factor analysis example retrieved those two factors as intended (i.e., the work satisfaction factor and the home satisfaction factor); thus, had nature planted the two factors, you would have learned something about the underlying or latent structure of nature.

**Miscellaneous other results.** Before concluding this example, brief comments on some other results will be made.

**Communalities.** To view the communalities for the current solution, that is, current numbers of factors, click the *Communalities* button on the *Factor Analysis Results* dialog box - *Explained Variance* tab. Remember that the communality of a variable is the portion that can be reproduced from the respective number of factors; the rotation of the factor space has no bearing on the communalities. Very low communalities for one or two variables (out of many in the analysis) may indicate that those variables are not well accounted for by the respective factor model.

**Factor score coefficients.** The factor score coefficients can be used to compute factor scores. These coefficients represent the weights that are used when computing factor scores from the variables. The coefficient matrix itself is usually of little interest; however, factor scores are useful if one wants to perform further analyses on the factors. To view these coefficients, click the *Factor score coefficients* button on the *Factor Analysis Results* dialog box - *Scores* tab.

**Factor scores.** Factor scores (values) can be thought of as the actual values for each respondent on the underlying factors that you discovered. Click the *Factor scores* button on the *Factor Analysis Results* dialog box - *Scores* tab to compute factor scores. These scores can be saved via the *Save factor scores* button and used later in other data analyses.

**Final Comment.** Factor analysis is a not a simple procedure. Anyone who is routinely using factor analysis with many (e.g., 50 or more) variables has seen a wide variety of "pathological behaviors" such as negative eigenvalues, uninterruptable solutions, ill-conditioned matrices, and such adverse conditions. If you are interested in using factor analysis in order to detect structure or meaningful factors in large numbers of variables, it is recommended that you carefully study a textbook on the subject (such as Harman, 1968). Also, because many crucial decisions in factor analysis are by nature subjective (number of factors, rotational method, interpreting loadings), be prepared for the fact that experience is required before you feel comfortable making those judgments. The *Factor Analysis* module of *STATISTICA* was specifically designed to make it easy for you to switch interactively between different numbers of factors, rotations, etc., so that different solutions can be tried and compared.

# Chapter 3. Modeling and forecasting the dynamics of economic processes, a time series analysis

## 3.1. Fundamentals of time series analysis (modeling of the dynamics)

The goal of empirical economic analysis is to highlight economic mechanisms and decision making: a number of observations on the relevant variables are thus required to study the existing links among them. The impossibility of controlled experiments in most economic fields has, as a consequence, which observations can be obtained only through surveys or resorting to existing databases.

Many statistical methods relate to data which are independent, or at least uncorrelated. There are many practical situations where data might be correlated. This is particularly so where repeated observations on a given system are made sequentially in time. Data gathered sequentially in time are called a *time series.* Time series (TS) econometrics is a rapidly evolving field. In particular, the cointegration revolution has had a substantial impact on applied analysis. As a consequence of the fast pace of development, there are no textbooks that cover the full range of methods in current use and explain how to proceed in applied domains. This gap in the literature motivates the present volume. The methods are sketched out briefly to remind the reader of the ideas underlying them and to give sufficient background for empirical work. The volume can be used as a textbook for a course on applied time series econometrics. The coverage of topics follows recent methodological developments. Unit root and *cointegration* analysis play a central part. Other topics include structural *vector autoregressions*, *conditional heteroskedasticity*, and nonlinear and *nonparametric time series models*. A crucial component in empirical work is the software that is available for analysis. New methodology is typically only gradually incorporated into the existing software packages.

A time series consists of a set of observations ordered in time, on a given phenomenon (target variable). Usually, the measurements are equally spaced, e.g. by year, quarter, month, week, day. The most important property of a time series is that the ordered observations are dependent on time, and the nature of this dependence is of interest in itself. Examples of time series are the gross national product, the unemployment rate, or the daily closing value of the Dow Jones index.

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. The simplest form of data is a longish series of continuous measurements at equally spaced time points. That is observations are made at distinct points in time, these time points being equally spaced and, the observations may take values from a continuous distribution. The above setup could be easily generalized: for example, the times of observation need not be equally spaced in time, the observations may only take values from a discrete distribution, If we repeatedly observe a given system at regular time intervals, it is very likely that the observations we make will be correlated. So we cannot assume that the data constitute a random sample. The time-order in which the observations are made is vital. Objectives of time series analysis:

- description - summary statistics, graphs;

- analysis and interpretation - find a model to describe the time dependence in the data, can we interpret the model?

- forecasting or prediction - given a sample from the series, forecast the next value, or the next few values;

- control - adjust various control parameters to make the series fit closer to a target

- adjustment - in a linear model the errors could form a time series of correlated observations, and we might want to adjust estimated variances to allow for this.

Time series data have a natural temporal ordering. This makes time series analysis distinct from other common data analysis problems, in which there is no natural ordering of the observations (e.g. explaining people's income relative to their education level, where the individuals' data could be entered in any order).

Time series analysis is also distinct from spatial data analysis where the observations typically relate to geographical locations (e.g. house prices). A time series model will generally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time series models will often make use of the natural one-way ordering of time so that values for a given period will be expressed as deriving in some way from past values, rather than from future values.

Thus, the time series is a sequence of data indexed by time, often comprising uniformly spaced observations. It is formed by collecting data over a long range of time at a regular time interval (data points should be at the same interval on the time axis). Consecutive points are then linked by means of straight lines to form the series.

The model accounts for patterns in the past movements of a variable and uses that information to predict its future movements. In a sense, a time series is a sophisticated method of extrapolation.

In time series analysis, we have no structural knowledge about the real world causal relationships which affect the variable we are trying to forecast. For example, an economic variable may be influenced by external factors which we cannot explain like the weather, changes in taste or seasonal cycles in spending. We could have used a regression model to forecast but the standard errors may become so large that the coefficients will become insignificant and the standard error of forecast unreasonably large.

We have to observe the evolution of the time series and draw some conclusions about its past behaviour that would allow us to infer something about its probable future. We should not attempt to construct a model which offers a structural explanation for its behaviour in terms of other variables (like in regression) but rather obtain one which replicates its past behaviour in a way that will help us forecast its future behaviour.

Besides the levels fluctuations, the dynamic series are characterized as random fluctuations associated with a mass random process.

Rows, where levels fluctuate around a mean, called stationary. The economic numbers are usually transient. For most of them characterized by systematic changes in the level of irregular fluctuations as alternate ups and dips of varying intensity. For example, economic cycles (industrial, construction, the stock market, etc.) are repeated with different length and different amplitude fluctuations.

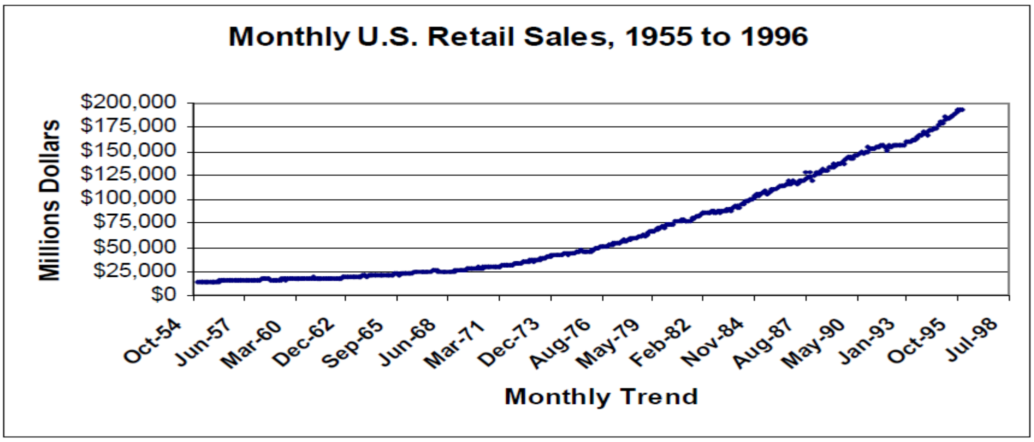



*Fig. 3.1. Time Series Plot of Monthly U.S. Retail Sales, 1955 to 1996*

*Components of a time series*

A time series is essentially composed of the following four components:

1. Trend $f(t)$;
2. Seasonality $S_t$
3. Cycle $C_t$;
4. Residuals $e_t$

*Trend*. The trend can usually be detected by inspection of the time series. It can be upward, downward or constant, depending on the slope of the trend-line. The trend-line equation of the line is actually the equation of the regression line of y(t) on t.

*Seasonality***.** The seasonal factor can easily be detected from the graph of the time series. It is usually represented by peaks and troughs occurring at regular time intervals, suggesting that the variable attains maxima and minima. The time interval between any two successive peaks or troughs is known as the period.

*Cycle.* A cycle resembles a season except that its period is usually much longer. Cycles occur as a result of changes of qualitative nature, that is, changes in taste, fashion and trend for example. A cycle is very hard to detect visually from a time series graph and is thus very often assumed to be negligible, especially for short-term data.

*Residuals.* Residuals are also known as errors which are put on the account of unpredictable external factors commonly known as freaks of nature. They are the differences between the expected and observed values of the variable. Theoretical values are the combination (addition or multiplication) of trend, seasonality and cycle. It is assumed that residuals are normally distributed and that, over a long range of time, they cancel one another in such a way that their sum is zero.

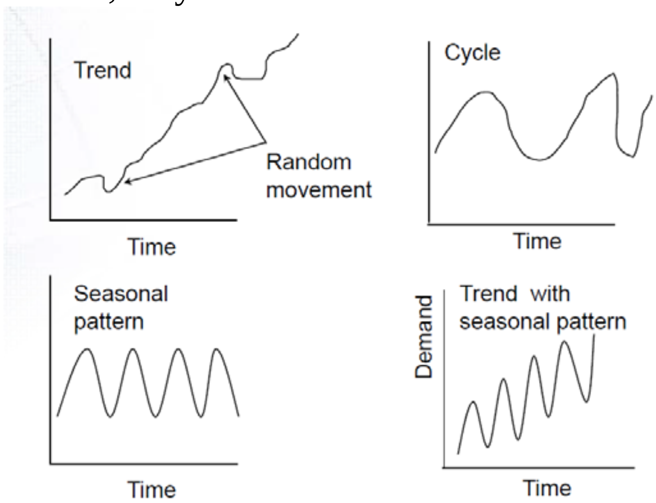



*Fig. 3.2. Components of a time series*

*A Time series models*, **t**here are two types of time series models – *additive* and *multiplicative*. In the additive model, the components are added and, in the multiplicative model, they are multiplied! Using $\boldsymbol{f(t)}$ for trend, $\boldsymbol{C_t}$ for cycle, $\boldsymbol{S_t}$ for season and $R$ for residuals, we can represent these models as follows:

*Additive* $y_t = f(t) + C_t + S_t + e_t$,

*Multiplicative* $y_t = f(t)\, C_t\, S_t\, e_t$.

The main difference between these two models is that, for the multiplicative, it is assumed that the factors influencing the components are dependent. This is a much more sensible and practical model for real-life situations since, in most situations, factors are interdependent. Whenever the model to be used is not specified, use the multiplicative.

It should be clear that, for an additive model, the components are expressed in the same units. For the multiplicative model, the trend has the same units as the $\boldsymbol{y_t}$ (or $\boldsymbol{y(t)}$) values and the three other components are considered to be unitless, thus acting as indices.

Depending on the purpose of the study, this conditional model allows studying the trend by eliminating (removing) the oscillation or studying the fluctuations by eliminating of the trend. During the forecasting, the different forecasts are reduced to that is final.

*Problems on time series mainly involve***:**

1. Detrending;
2. Deseasonalisation;
3. Forecasting.

Isolation of components:

*Trend* . Given a set of data, it is relative easy to determine the trend-line equation and hence the trend values by using the method of regression. The equation will be of the form

$$y_t = a + bt,$$

where $a$ and $b$ are known as regression coefficients. It is obvious that $a$ is the

$y$-intercept and $b$ is the slope of the regression line. It is noted that, since the values of $t$ occur at a regular interval, we can determine the trend values by simply adding $b$ continuously to the first trend-value.

*Seasonality.* The seasonal component is rather strangely calculated! It is first removed, together with the residuals, from the original data by using the method of *moving averages* and then obtained back by dividing the original data by the remaining components (multiplicative model) or by subtracting the remaining components from the original data (additive model). The residuals are then eliminated by averaging as will be seen in an example further on.

*Cycle.* Cycles are expected to sum up to zero in an additive model and negligible in a multiplicative model. Hence, they do not take an active part in the process of forecasting.

*Residuals.* This unpredictable component is also removed together with the season when applying the method of moving averages. It makes no sense to isolate it but it may be extracted from the seasonality by some form of averaging.

***Forecasting.*** Forecasting is the ultimate objective of time series analysis. This delicate procedure involves only trend and seasonality. We will look at two different aspects of forecasting – using tables or graphs.

The dependence of levels are a characteristic feature of any dynamic series: value $y_t$ on some extent depends on previous values $y_{t-1}$, $y_{t-2}$ etc.

To assess the degree of dependence the levels time series, we use *autocorrelation coefficient* $r_p$ with time lag $p = 1, 2, \dots, m$.

Coefficient $r_p$ describes the density of connection between primary time series and their shifting on $p$ moments:

*Table 3.1. The shifts of time series*

| *Changing the time t* | *Level of $y_t$* | *p = 1* | *p = 2* | *p = 3* |
|---|---|---|---|---|
| *1* | $y_1$ | — | — | — |
| *2* | $y_2$ | $y_1$ | — | — |
| *3* | $y_3$ | $y_2$ | $y_1$ | — |
| … | … | … | … | … |
| *n – 2* | $y_{n-2}$ | $y_{n-3}$ | $y_{n-4}$ | $y_{n-5}$ |
| *n – 1* | $y_{n-1}$ | $y_{n-2}$ | $y_{n-3}$ | $y_{n-4}$ |
| *n* | $y_n$ | $y_{n-1}$ | $y_{n-2}$ | $y_{n-3}$ |

The shifts of time series on lags $p$=1,2,3 are shown in the *Table 3.1.* As we see, with increasing values for lag $p$, the number of the correlated pairs levels decreases. Thus, if $p$=1 then a length of correlated a time series is less by one unit than the primary one. If p $p$=2 we have shift by two units, etc. Because of this, on practice to determine the autocorrelation function should follow the rules by which the number of lags are $\boldsymbol{m} \leq \frac{\boldsymbol{n}}{\mathbf{2}}$.

For time series, the errors may not be independent. Often errors are autocorrelated; that is, each error is correlated with the error immediately before it. Autocorrelation is also a symptom of systematic lack of fit.

The coefficient of autocorrelation $-1 \leq r_p \leq 1$ determine by the value of lag $p$:

$$r_p = \frac{c_p}{c_0}, \text{ where } c_p = \frac{1}{n}\sum_{t=1}^{n-p}(y_t - \bar{y})(y_{t+p} - \bar{y}); \quad c_0 = \frac{1}{n}\sum_{t=1}^{n}(y_t - \bar{y})^2 .$$

The sequence of coefficients $r_p$ is called *the autocorrelation function* ***(****ACF****)*** and a plot of $r_p$ against $p$ is known as a *correlogram* (not to be confused with *Scatterplot*). The correlogram, also known as an *autocorrelation plot*, is a plot of the sample autocorrelations $r_p$ versus $p$ (the time lags). The coefficients $r_p$ are represented by rectangles, the lines that define 95% confidence limits of significant for $r_p$ are designated by the dotted line.

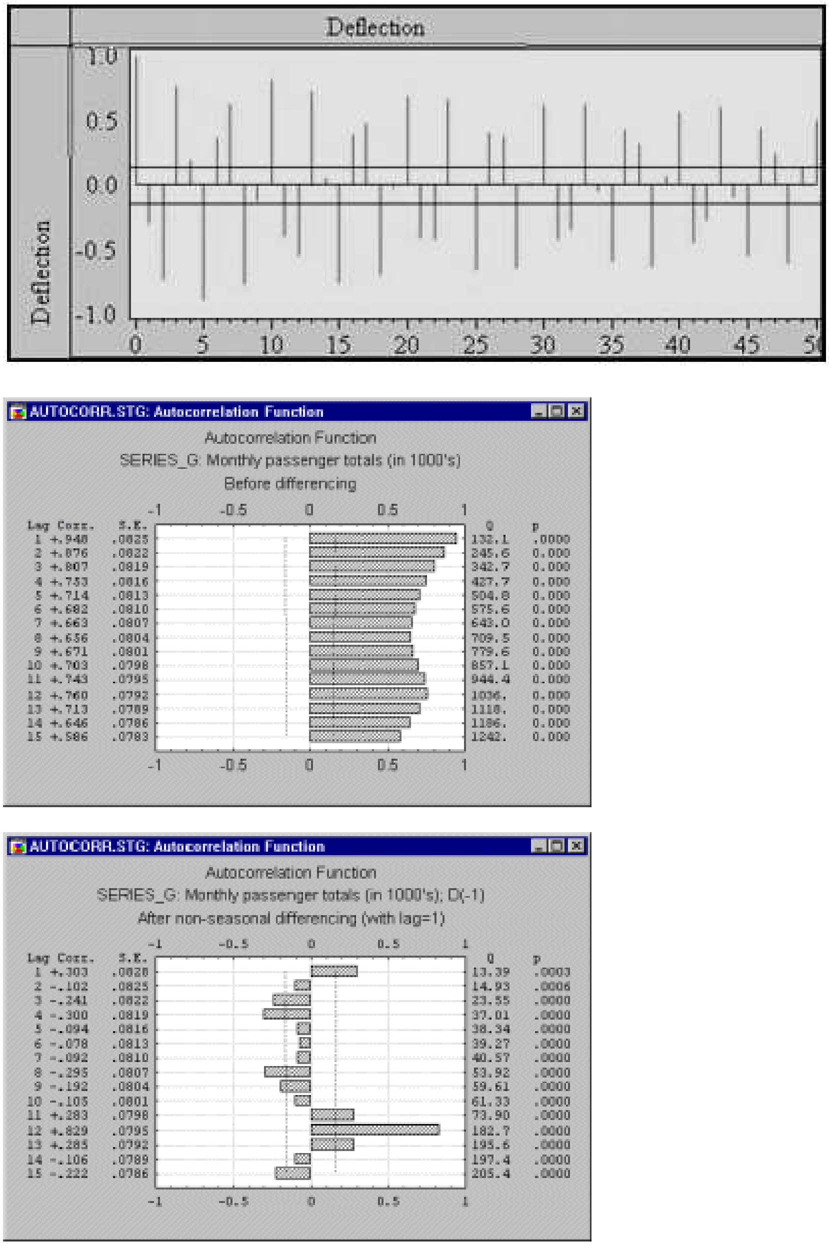



*Fig. 3.3. Examples of Autocorrelation plot and Autocorrelation function*

A time series considered stationary when $r_p$ doesn't leave confidence interval. It can be concluded about the nature of the dynamics by speed fading of the autocorrelation function.

The values $r_1$ is most commonly used. Describing the degree of dependence of two successive members of the series, autocorrelation coefficient is a measure of continuity of the series. If $r_1 \to 1$, then a time series has a tendency to development, if $r_1 \to 0$, then the levels of series are independent. Relatively high autocorrelation coefficient at $p=k$, $2k$, $3k$… indicate to the regular fluctuations.

Random component, unlike deterministic component, is not connected with the change of time. Analysis of this component is the basis for testing the hypotheses about the adequacy of a model to real process. Provided that the model is correct, the random component is a stationary process with expectation $M(e)$=0 and dispersion is

$$s_e^2 = \frac{\sum_1^n (y_t - f_t)^2}{n-m},$$

where $m$ is number the parameters of function $f_t$.

To estimate stationarity of random component use a cyclic autocorrelation coefficient of the first order $r_1$. Series of residual values correlated: $e_1$, $e_2$, $e_3$, …, $e_n$ and $e_2$, $e_3$, $e_4$, …, $e_1$.

If $\sum e_e = \sum e_{e+1} = 0$, the formula is simplified:

$$r_1 = \frac{\sum_1^n e_t e_{t+1}}{\sum_1^n e_t^2}.$$

There are tables of critical values of cyclic autocorrelation coefficients for positive and negative values. If the actual value $r_1$ is less than critical, autocorrelation is considered negligible and the random component is stationary process (so-called "*white noise*"). In case the actual value exceeds the critical $r_1$ can be concluded about the inadequacy of deterministic real part of the process.

Modeling and prediction of dynamic processes can be carried out according to the procedures in *STATISTICA* by using the modules: *Multiple Regression*, *Time Series / Forecasting*, *Nonlinear Estimation*. Modeling trends and trend extrapolation procedures performed by module *Multiple Regression* and *Nonlinear Estimation*; comprehensive analysis of dynamic processes, identification models, adaptive prediction with use the procedures module Time Series / Forecasting, which includes data processing methods, such as

- *Arima & autocorrelation functions* for the autoregression model and integration moving average;

- *Interrupted time series analysis* for analysis a broken time series (models for intervention ARIMA);

- *Exponential smoothing & forecasting* for exponential smoothing and forecasting;

- *Seasonal decomposition* (1, 2) for seasonal decomposition 1 and 2 (monthly and quarterly);

- *Distributed lags analysis* for the analysis of distributed lags (regression model for two time series);

- *Spectral (Fourier) analysis* for spectral (Fourier) analysis.

If necessary, selected for the study of dynamic range can be transformed using a sequence of commands: ***OK*** (*transformations, autocorrelations, crosscorrelations, plots*) → ***OK*** (*Transform highlighted variable*) → *Transformations of variables*. A wide range the transformations for dynamic series are offered in the box *Time series transformations*, including:

- *Add a constant* – add a constant to value range;

- *Power* – raise to a power;

- *Inverse power* – to obtain the root;

- *Natural log* – to obtain the natural logarithm, it is based of the mathematical constant ***e***;

- *Exponent* – raise constant ***e*** to a power;

- *Mean subtract* –deviations from the mean;

- *Standardize* – feature standardization.

All variants of transformation used consecutively only for the selected time series. In the latter two options, you can specify the average and standard deviation or identify them automatically using the option *Estimate mean & std.dev.from data*.

Module transformation involves the removing linear trend, use *Trend subtract*, introducing his parametry manually or use automatic calculation *Estimate a / b from data*. Similarly, carried out the removal of autocorrelation lag appropriate option for *Autocorr*.

Under options *Shift relative starting point of series* offered a number of options nudged forward or backward on a log. Option *Differencing* makes it possible to determine the difference between the current $y_t$ and the levels ($y_t - y_{t+p}$) shift alongside on lag *p*.

Equally important is the group of options *Smoothing* which when repeated application allows us to determine the type of trend. All information displayed in the transformation of the starting panel, the maximum number is nine. If one or another transformation in the further analysis is not used, it will remove by command *Delete highlighted variable*. The command *Save*, on the contrary, retains certain features in a separate file.

## 3.2. Trend Analysis. Simple forecasting and smoothing methods.

The simple forecasting and smoothing methods model components in a series that is usually easy to observe in a time series plot of the data. This approach decomposes the data into its component parts and then extends the estimates of the components into the future to provide forecasts. You can choose from the static methods of trend analysis and decomposition, or the dynamic methods of moving average, single and double exponential smoothing, and Winters method. Static methods have patterns that do not change over time; dynamic methods have patterns that do change over time and estimates are updated using neighboring values.

You can use two methods in combination. That is, you can choose a static method to model one component and a dynamic method to model a different component. For example, you can fit a static trend using trend analysis and dynamically model the seasonal component in the residuals using Winters method. Or, you can fit a static seasonal model using decomposition and dynamically model the trend component in the residuals using double exponential smoothing. You can also apply a trend analysis and decomposition together so that you can use the wider selection of trend models offered by trend analysis. A disadvantage of combining methods is that the confidence intervals for forecasts are not valid.

In this chapter, we will refer to one from three types of time series patterns.

A trend exists when there is a long-term increase or decrease in the data. It does not have to be linear. Sometimes we will refer to a trend "changing direction" when it might go from an increasing trend to a decreasing trend.

In a time series econometrics, there is often interest in the dynamic behavior of a variable over time. A dynamic conditional mean model specifies the expected value of $y_t$ as a function of historical information. The studying some of this models is an important part of the dynamic processes. In the analysis of time series, a trend is represented as a smooth path and described of the function, called the trend $Y_t=f(t)$, where $t=1,2,\ \dots,\ n$ is time variables.

In practice, we have mostly used functions for which the parameters have specific interpretation. That are depended on character the dynamics of trends. there are the most common curves: the polynomials, various exhibitors, and logistic curves etc. A polynomial of degree 1, i.e. linear trend $Y_t=a+bt$ to describe the processes that change evenly over time and have stable gain levels. Polynomial 2nd degree (parabola) $Y_t=a+bt+ct^2$ is able to describe the process, which is characteristic of accelerated growth or decrease. Parabola shape is determined by the c: when $c>0$ it means the branches of the parabola are directed to upward and has a minimum vice versa, if parabola at $c<0$ has branches are directed to down, in this case, the parabola has a maximum. An increment of sign the level a time series can change one or two times if it is 3rd-degree polynomial $Y_t=a+bt+ct^2+dt^3$.

If stable relative velocity (growth rate) is a characteristic feature of the process then this process is described by the exponent, which can take different equivalent forms. For example, the main form of exponential equation

(dependence) is $Y_t = ab^t$ ,

where $b$ — average rate of change the levels: if $b > 1$ then levels are increased with a constant rate, if $b < 1$, on the contrary, they are decreased. Absolute growth is proportional to the level of achieved,

Exponential trend can be represented in the form $Y_t = ae^{\lambda t}$ or $Y_t = e^{a + bt}$, where $\lambda = \ln b$, $e = 2{,}718$ is the *base* of the *natural logarithm*, $\ln e = 1$.

Evaluation the parameters of equations trend often do by the method of *least squares (OLS)*, the method of least squares is a standard approach in regression analysis to the approximate solution of overdetermined systems, i.e., sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes sum of the square residuals, that is to say $\sum_1^n (y_t - Y_t)^2 \rightarrow \min$ or $\sum_1^n e_t^2 \rightarrow \min$, where $y_t$ are actual values , $Y_t$ are theoretical values and residuals $e_t = y_t - Y_t$

Construction of trend by *OLS* model can easily be implemented by modules *Multiple Regression* in *spreadsheets of MS Excel* and one can use the option *Add trend line* in the *Chart Wizard*.

The tendency can be extended beyond the dynamic range. This procedure is called *extrapolation* of trend. The theoretical possibility of extrapolation based on the assumption that the conditions that determine the trend in the past, do not undergo significant changes in the future. Formally extrapolation operation can be represented as a definition of function

$$Y_{t+v} = f(Y_t^*, v),$$

where $Y_{t+v}$ is a predicted value on *v –ahead* ; $Y_t^*$ is a basis of extrapolation, it often takes the last level of series that is defined by the trend.

One can make extrapolation by option *Predict dependent var* under the module *Multiple Regression*. It should specify the value $v$ in the box *Specify dependent for index vars.* Extrapolation of the trend is a point forecast. Obviously, the '*hit to point'* is unlikely. Indeed, the trend has a property of uncertainties, primarily because of error parameters. The source of these errors is a limited set of observations $y_t$ , each of which contains a random component $e_t$.

If observation period is shifted just in one step, it leads to the parameter estimates shifting. The random component will be present also outside the dynamic range, and therefore it should be taken into account. To do this, define a confidence interval that would be with a certain probability limits outlined the possible values of $Y_{t+v}$.. A point forecast is converted to the prediction interval. The width of the interval is depended on the level of variation around the trend of a time series and probability of inference (1–α): $Y_{t+v} \pm t_{1-\alpha}\, s_p$, where $s_p$ is mean squared prediction error, the value of which depends on the variance of trend $s_Y^2$ and variance of the deviations from trend $s_e^2$. The forecast error depends on the residual variance $s_e^2$, a length of a time series (prehistory) and period of shift *v*. The longer the period of prehistory, the error is less on the contrary increase in the period leads to an increase in forecast error.

As a first step in moving beyond mean models, random walk models, and linear trend models, nonseasonal patterns and trends can be extrapolated using a moving-average or smoothing model. The basic assumption behind averaging and smoothing models is that the time series is locally stationary with a slowly varying mean. Hence, we take a moving (local) average to estimate the current value of the mean and then use that as the forecast for the near future. This can be considered as a compromise between the mean model and the random-walk-without-drift-model. The same strategy can be used to estimate and extrapolate a local trend. A moving average is often called a "smoothed" version of the original series because short-term averaging has the effect of smoothing out the bumps in the original series. By adjusting the degree of smoothing (the width of the moving average), we can hope to strike some kind of optimal balance between the performance of the mean and random walk models.

The procedure of *smoothing* is called *filtering*, and operators of those procedures are named the *filters*. This technique, when properly applied, reveals more clearly the underlying trend, seasonal and cyclic components. There are two distinct groups of smoothing methods: *Averaging Methods* and *Exponential Smoothing Methods*, the simplest kind of averaging model is a simple moving average (SMA) with smoothing interval $m < n$. Where $n$ is number of observation. We define the mean for each of intervals $\bar{y}_t$, that are a mid-interval. If $m$ is an odd number, that $m=2p+1$ and weight of members of the series within the same interval $a_r = \frac{1}{(2p+1)}$, then $\bar{y}_t = \frac{1}{2p+1}\sum_{t-p}^{t+p} y_i$, where $y_i$ is actual value and level, at a time *i*; and *i* is the serial number of level in the interval. For even number *m*, the mid-interval lies between two points of the time and then we perform additional procedures centering (averaging of each pair of values). Moving average with equal weights $a_r$ at smoothing time series are repaying not just incidental, but inherent in a particular process the periodic fluctuations.

Assuming the existence of such fluctuations, we use the weighted moving average. That is, we gave the concrete weight to each level within the of the interval.

There are various ways of forming a weight function. In some cases, the weights correspond to members of binomial expansion $\left(\frac{1}{2}+\frac{1}{2}\right)^{2p}$, at $m = 3$ it be $a_r = 1/4,\ 1/2,\ 1/4$. In other cases, we are looking for the data smoothing interval as a polynomial function such how, for example, the parabola such how $\bar{y}_t = a + b_i + c_i^2$, where $i=-p,\dots,p$. Then the weight function :

for $m = 5$ $a_r = \frac{1}{35}(-3, 12, 17, 12, -3)$;

for $m = 7$ $a_r = \frac{1}{21}(-2, 3, 6, 7, 6, 3, -2)$ etc.

As seen from formulas, the weights are symmetric with respect to interval center of the smoothing, and a sum of all multipliers is equal $\sum a_r = 1$.

The main advantage of moving average is visibility and ease the interpretation of trends. But we should not forget that a number of moving average series is shorter on *2p* levels than an original series. Therefore, information on the extreme members of the series is lost. The losses are more tangible, especially for new information, if the interval of smoothing is wider. In addition, moving averages always are dependent, though they have a common base of calculation.

This may indicate a cyclical process (Slutsky effect), under smoothing the significant fluctuations, even if we have the absence of cycles in the primary series.

Old and new information is equally weighty. One can use asymmetric filters if a new information is more important for predict. The simplest of them is a moving average that replaces not central, and the last member of the series (*adaptive average*): $\bar{y}_t = \bar{y}_{t-1} + \frac{y_t - y_{t-m}}{m}$.

Intuitively, past data should be discounted in a more gradual fashion, for example, the most recent observation should get a little more weight than 2nd most recent, and the 2nd most recent should get a little more weight than the 3rd most recent, and so on. The simple exponential smoothing (SES) model accomplishes this.

Let *a* denote a *smoothing constant* (a number between 0 and 1). One can write the model is to define a series $Y_t$ that represents the current level (i.e., local mean value) of the series as estimated from data up to the present. The value of $Y_t$ at time t is computed recursively from its own previous value like this:

$$Y_t = a y_t + (1-\alpha) Y_{t-1}.$$

Thus, the current smoothed value is an interpolation between the previous smoothed value and the current observation, where *a* controls the closeness of the interpolated value to the most recent observation. The *forecast* for the *next* period is simply the current smoothed value:

Equivalently, we can express the next forecast directly in terms of previous forecasts and previous observations, in any of the following equivalent versions. In the first version, the forecast is an interpolation between previous *forecast* and previous *observation:*

$$Y_{t+1} = a y_t + (1-\alpha) Y_t.$$

In the second version, the next forecast is obtained by adjusting the previous *forecast* in the direction of the previous *error* by a fractional amount *a*:

$$Y_{t+1} = Y_t + \alpha e_t,$$

where $e_t = y_t - Y_t$ is the error made at time t. In the third version, the forecast is an *exponentially weighted (i.e. discounted) moving average* with discount factor 1-α:

$$Y_{t+1} = a\,[Y_t + (1-a)\,Y_{t-1} + (1-a)^2\,Y_{t-1} + (1-a)^3\,Y_{t-2} + ... + +(1-a)^r\,Y_{t-r+1}]$$

or

$$Y_t = \sum_{r=0}^{t} a(1-a)^r y_{t-r}.$$

The interpolation version of the forecasting formula is the simplest to use if you are implementing the model on a spreadsheet: it fits in a single cell and contains cell references pointing to the previous forecast, the previous observation, and the cell where the value of *a* is stored. Note that if *a=1*, the SES model is equivalent to a random walk model (without growth). If *a=0*, the SES model is equivalent to the mean model, assuming that the first smoothed value is set equal to the mean.

Exponential average adapts to new conditions by the giving more weight to new information, and in this case method of *short-term forecasting* is very effective and reliable. To calculate the average exponential $Y_t$ it is necessary to determine the initial conditions for initial value $Y_0$ and parameter *a*. The average level of the past period (to dynamic series) or in the absence of such data, we can use the first level of series, that $Y_0 = y_1$. As for parameter *a*, in practice the value in the range of 0.1 to 0.3 often use. The sum of weight coefficients $\sum a_r$ in certain time interval *m* depends on parameter *a*. Then one can determine value *a* by the predetermined values: $a = 1 - \sqrt[m]{1 - \sum_{1}^{m} a_r}$. For example, if the time interval is *m* = 10 months, and the sum of the weights equal $\sum a_r = 0{,}90$, then $a = 1 - \sqrt[10]{1 - 0{,}9} \approx 0{,}2$.

That is, ten members of the dynamic series determine 90% of the values of the exponential average when *a = 0,2.*

The 12-day and 26-day exponential averages of exchange closing prices with smoothing parameters of 0.15 and 0.075 respectively use in monitoring the foreign exchange market

We examine the trend lines as fast and slow (or lines of support and resistance). A significant deviation between of this averages indicates a power of trend and intersection signalize that their changes are possible. If the fast average crosses slow top, it signals the birth of a new downtrend, and if it crosses low that is the birth of a growing trend. The exponential average second order $Y_t^*$ is determined by the same recurrence formula based on the smoothed series $Y_t$:

$$Y_t^* = aY_t + (1-a)Y_{t-1}^*.$$

Assuming the presence of a linear trend, forecast level $Y_{t+1}$ can be calculated by the formula $Y_{t+1} = \frac{(2-a)Y_t - Y_t^*}{1-a}$.

## 3.3. Estimating the seasonal component. Holt-Winter's and X-11/Y2k Monthly Seasonal Adjustment (Census Method II)

Many time series display seasonality. By seasonality, we mean periodic fluctuations. For example, retail sales tend to peak for the Christmas season and then decline after the holidays. So time series of retail sales will typically show increasing sales from September through December and declining sales in January and February.

Seasonality is quite common in economic time series. It is less common in engineering and scientific data.

If seasonality is present, it must be incorporated into the time series model Seasonal fluctuations are forming influenced by not only climatic but also socio-economic factors. The strength and direction of the individual factors create a different configuration of seasonal waves.

The nature of the seasonal component can be additive or multiplicative.

For the typical constant additive components are characterized the oscillation around the average level or trend and for multiplicative growth as amplitude attenuation over time (visually resembles a cone). Similarly, we can suppress (remove) certain features in a time series, such as seasonality, in order to model the trend and/or cycle. Once we have built a suitable model for the smoothed series, we can add back the appropriate seasonal component in order to produce predictions.

If we have monthly data, our first moving average value is calculated on observations 1 to 12, and the second moving average value is calculated on observations 2 to 13. We then average these two values to get our first moving average value which then replaces observation 7 in our original series. Similarly, at the end of our series, there are six observations that we have no moving average values for. The method used is to first smooth the trend and cycle using a lowess smoother (fitting a local regression to a window of points and using the point on the fitted regression line as the value of the smooth for the time value in the middle of the window). The regression that is used is “weighted”, in that observations near the edge of the window are given less weight than observations near the centre of the window when determining the local regression line. Then a separate lowess smoother is used on each seasonal sub-series (i.e. all the January observations, all the February observations, …). The *trend and cycle* smoothed value and the appropriate *seasonal* smoothed value can be subtracted from the original observation to yield the remainder or random component for that observation.

*Holt-Winters Model*

This model, often referred to as a procedure, was first proposed in the early 1960s. It uses a process known as exponential smoothing. All data values in a series contribute to the calculation of the prediction model.

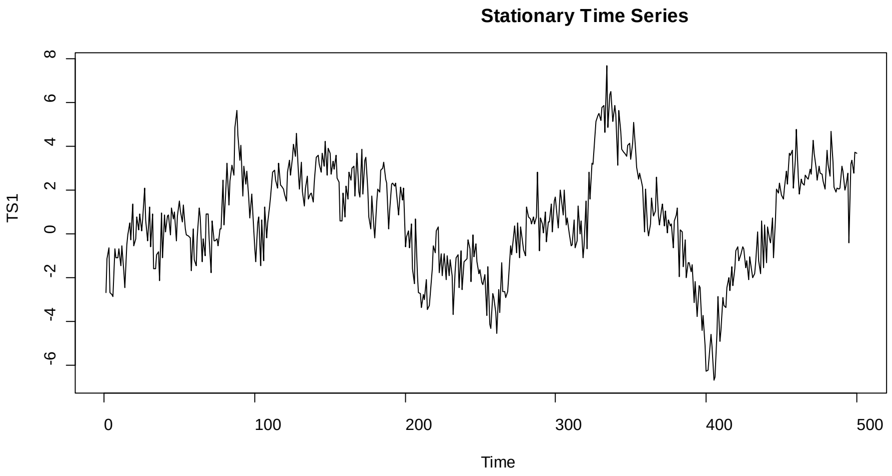



*Fig. 3.4. Examples of stasionary time series*

*Exponential smoothing* in its simplest form should only be used for non-seasonal time series exhibiting a constant trend (or what is known as a stationary time series). It seems a reasonable assumption to give more weight to the more recent data values and less weight to the data values from further in the past. An intuitive set of weights is the set of weights that decrease each time by a constant ratio. Strictly speaking this implies an infinite number of past observations but in practice there will be a finite number. Such a procedure is known as *exponential smoothing* since the weights lie on an exponential curve.

If the smoothed series is denoted by $S_t$ , $\alpha$ denotes the smoothing parameter, the exponential smoothing constant, $0 < \alpha < 1$.

The smoothed series is given by: $S_t = \alpha y_t + (1 - \alpha)S_{t-1}$, where $S_1 = y_1$. The smaller the value of $\alpha$, the smoother the resulting series.

It can be shown that: $S_t = \alpha y_t + \alpha(1 - \alpha)y_{t-1} + \alpha(1 - \alpha)^2 y_{t-2} + \ldots + (1 - \alpha)^{t-1} y_1$.

Consider the following Time Series: 14 24 5 18 10 17 23 17 23 …

Using the formulae above, with an exponential smoothing constant, $\alpha = 0.1$:

$S_1 = y_1 = 14$,

$S_2 = \alpha y_2 + (1 - \alpha)S_1 = 0.1(24) + 0.9(14) = 15$,

$S_3 = \alpha y_3 + (1 - \alpha)S_2 = 0.1(5) + 0.9(15) = 14$,

$S_4 = \alpha y_4 + (1 - \alpha)S_3 = 0.1(18) + 0.9(14) = 14.4$ etc.

Thus the smoothed series depends on all previous values, with the most weight given to the most recent values.

Exponential smoothing requires a large number of observations. Exponential smoothing is not appropriate for data that has a seasonal component, cycle or trend. However, modified methods of exponential smoothing are available to deal with data containing these components.

The Holt-Winters model uses a modified form of exponential smoothing. It applies three exponential smoothing formulae to the series. Firstly, the level (or mean) is smoothed to give a local average value for the series. Secondly, the trend is smoothed and lastly each seasonal sub-series ( ie all the January values, all the February values….. for monthly data) is smoothed separately to give a seasonal estimate for each of the seasons.

The exponential smoothing formulae applied to a series with a trend and constant seasonal component using the Holt-Winters additive technique are:

$$\mathrm{a}_t = \alpha(Y_t - \mathrm{s}_{t-p}) + (1-\alpha)(\mathrm{a}_{t-1} + \mathrm{b}_{t-1}),$$

$$\mathrm{b}_t = \beta(\mathrm{a}_t - \mathrm{a}_{t-1}) + (1-\beta)\mathrm{b}_{t-1},$$

$$\mathrm{s}_t = \gamma(Y_t - \mathrm{a}_t) + (1-\gamma)\mathrm{s}_{t-p},$$

where: $\alpha$, $\beta$ and $\gamma$ are the smoothing parameters; $\mathrm{a}_t$ is the smoothed level at time $t$

$\mathrm{b}_t$ is the change in the trend at time $t$

$\mathrm{s}_t$ is the seasonal smooth at time $t$

$p$ is the number of seasons per year

The Holt-Winters algorithm requires starting (or initial) values. Most commonly:

$$\mathrm{a}_p = \frac{1}{p}(Y_1 + Y_2 + \ldots + Y_p)$$

$$\mathrm{b}_p = \frac{1}{p}\left[\frac{Y_{p+1} - Y_1}{p} + \frac{Y_{p+2} - Y_2}{p} + \ldots + \frac{Y_{p+p} - Y_p}{p}\right]$$

$$\mathrm{s}_1 = Y_1 - \mathrm{a}_p, \quad s_2 = Y_2 - \mathrm{a}_p, \quad \ldots, \quad s_p = Y_p - \mathrm{a}_p.$$

The Holt-Winters forecasts are then calculated using the latest estimates from the appropriate exponential smooths that have been applied to the series.

So we have our forecast for time period $T+\tau$:

$$\hat{y}_{T+\tau} = \mathrm{a}_T + \tau\ \mathrm{b}_T + \mathrm{s}_T$$

where: $\mathrm{a}_T$ is the smoothed estimate of the level at time *T*;

$\mathrm{b}_T$ is the smoothed estimate of the change in the trend value at time *T*;

$\mathrm{s}_T$ is the smoothed estimate of the appropriate seasonal component at *T*.

As mentioned earlier the Holt-Winters model assumes that the seasonal pattern is relatively constant over the time period. Students would be expected to notice changes in the seasonal pattern and identify this as a potential problem with the model, particularly if long–term predictions are made. In practice this is dealt with by transforming the original data and modelling the transformed series or using a multiplicative model. Students are not expected to know this, but are required to identify a variable seasonal pattern as a potential problem. The exponential smoothing formulae applied to a series using Holt-Winters Multiplicative models are:

$$\mathrm{a}_t = \alpha \frac{Y_t}{\mathrm{s}_{t-p}} + (1-\alpha)(\mathrm{a}_{t-1} + \mathrm{b}_{t-1}),$$

$$\mathrm{b}_t = \beta(\mathrm{a}_t - \mathrm{a}_{t-1}) + (1-\beta)\mathrm{b}_{t-1},$$

$$\mathrm{s}_t = \gamma \frac{Y_t}{\mathrm{a}_t} + (1-\gamma)\mathrm{s}_{t-p}$$

The initial values are as for the additive model, except:

$$\mathrm{s}_1 = \frac{Y_1}{\mathrm{a}_p}, \quad \mathrm{s}_2 = \frac{Y_2}{\mathrm{a}_p}, \quad \ldots, \quad s_p = \frac{Y_p}{\mathrm{a}_p}.$$

So we have our prediction for time period $T+\tau$:

$$\hat{y}_{T+\tau} = (\mathrm{a}_T + \tau\ \mathrm{b}_T)\mathrm{s}_T.$$

There are many situations where it is important to give interval predictions, rather than point predictions, as a means of assessing future uncertainty. An interval prediction associated with a prescribed probability a confidence interval, but that the term prediction interval used in the context of time series analysis. This is interval estimates of model parameters. It is relatively common to see predictions made without any reference to prediction intervals. This may be because there are a number of different ways that prediction intervals. The paper above provides not only details of how the prediction intervals for Holt-Winters are produced but also compares the authors' preferred method with several alternative methods. It also compares the prediction intervals calculated for the same data set by a variety of different models.

In one example, a monthly index of employment in manufacturing in Canada, a prediction for three years after the end of the actual data is provided of 115.9. A prediction interval of [113.95,117.85] is also calculated. A suggested interpretation of this prediction interval (P.I.) is ' There is a 95% chance that the true index value of employment in manufacturing in Canada in three years time will be between

113.95 and 117.85.' The details given in this paper apply to an additive Holt-Winters model only.

*Assessing non-stationary model forecasts*

The test of any prediction model is how well does it predict when compared to actual data values. To do this either remove the last few given observations or find the next few actual observations. Different prediction models can then be compared using a statistic known as the Root Mean Squared Error of Prediction (RMSEP). The formula for calculating this statistic is given below

$$RMSEP = \sqrt{\frac{1}{\tau}\sum_{t=1}^{\tau}(y_t - \hat{y}_t)^2}$$

where $\tau$ is the number of predictions we are using in the calculation of *RMSEP*.

Students are not expected to calculate RMSEP.

Summary of what students need to know

- Holt –Winters Additive model assumes seasonal pattern is reasonably constant
- Holt-Winters Model uses a technique of exponential smoothing, which is a weighted sum of previous values in a series. More weight is given to more recent values and less weight is given to values from the distant past.
- Holt-Winters Additive model exponentially smooths three series in order to produce predictions – the level, the trend and the seasonal sub-series.
- Students should be able to identify cyclical components and inconsistent seasonal patterns. They should note that such features are incompatible with assumptions underlying Holt-Winters Additive model and suggest a multiplicative model be considered instead. Such a comment would be expected at Excellence level only. Students are NOT expected to calculate a multiplicative model.

The parameters of smoothing (*Alpha*), (*Delta*) and (*Gamma*) in the default system *STATISTICA* determined at 0.1 (and we have the opportunity to choose their optimal value option *Search on the net*)

Often in modeling time series with different seasonal components is used the methods of seasonal adjustment (*decomposition*) X-II (or so-called *CENSUS X-II methods* according to the abbreviation of the Central Bureau of Statistics USA). These methods are based on numerous practical applications and one takes into account various properties of dynamics: outliers, a different number of days in months, working and days off, examine different assessments for final clarification of the receiving component-cyclical trend, seasonality, an irregular component of a time series and seasonal amended.

If periodic fluctuations of monthly dynamics series exist we can use a seasonal wave model that is based on a harmonic analysis. Its main features are amplitude, phase, and frequency of fluctuations period.

*The amplitude A* describes the distance of the average maximum (minimum) seasonal wave, cycle frequency is *oscillation period T* (the number of units of time necessary to complete one full cycle), *f* is the number of cycles per unit time (where each observation is treated as one unit of time). Thus, the frequency corresponds to a period (the number of units of time necessary to complete one full cycle), $f = 1/T$.

If $T = 12$ months, then $f = 1/12$ of cycle a month. The distance between the timing point *t = 0* and nearest a peak is called *a phase* $\Theta$. Seasonal wave at period *T* can be described by a function: $Y = a + b\cos\omega t + d\sin\omega t$, where $\omega$ is harmonics angular frequency; it is measured in radians per unit time $\omega = 2\pi f = 2\pi/T$ and varies in the interval $0 \le \omega \le 2\pi$; *b*, *d* is harmonics coefficient, functionally related to the amplitude: $A = \sqrt{b^2 + d^2}$. Harmonics coefficients determined by the least squares method. A system of normal equations reduced to the identities through the properties of orthogonal sine and cosine functions:

$\sum y = an$;

$\sum y\cos\omega t = \frac{1}{2}nb$;

$\sum y\sin\omega t = \frac{1}{2}nd$.

Hence, if $n = 12$ then we obtain : $a = \frac{\sum y}{12}$; $b = \frac{\sum y\cos\omega t}{6}$; $d = \frac{\sum y\sin\omega t}{6}$.

Therefore, *a* is nothing like the average monthly level of series. Coefficients *b* and *d* determine the amplitude of fluctuations around the average level.

Obviously, it is larger the amplitude, we have the more significant contribution of harmonic to the total variance of the process. Variance ratio serves as an estimate of this contribution $R^2 = \frac{\delta^2}{\sigma^2}$, where $\delta^2 = 0{,}5A^2$ is variance of harmonic.

One can involve into the model of harmonic analysis a lot of harmonics with different periods of oscillations. For instance, the first harmonic has the period 12, the second has the period 6, the third - 4 and etc.

A time series is decomposed to regular periodic waves the sine wave through the harmonic function. The adequacy of it to real process depends as far as frequency and amplitude have a sustainable character. A relatively sustainable character of the intra-year dynamics is inherent for a market of seasonal items.

*Example 3.1. Consider the dynamics of the monthly average price the fresh cucumbers (hrn.):*

*Table 3.2. Dynamics of the monthly average price the fresh cucumbers (UAH)*

| *Month* | *Price,UAH.* | *Month* | *Price,UAH* | *Month* | *Price,UAH.* |
|---|---|---|---|---|---|
| *1* | *5,56* | *5* | *2,60* | *9* | *0,76* |
| *2* | *5,70* | *6* | *1,38* | *10* | *1,43* |
| *3* | *4,72* | *7* | *0,70* | *11* | *4,36* |
| *4* | *3,68* | *8* | *0,57* | *12* | *5,89* |

To build a harmonic function in *STATISTICA* we should click the *Spectral (Fourier) analysis* button in the *Time Series Analysis* Startup Panel to display the *Fourier (Spectral) Analysis* dialog box, which contains four tabs: *Quick*, *Advanced*, *Autocorrelations*, and *Review series*. Both the *Quick tab* and the *Advanced tab* of the *Single Series Fourier (Spectral) Analysis Results* dialog have an option for producing a summary spreadsheet with the frequencies, periods, cosine and sine coefficients, periodogram values, spectral density estimates (computed according to the selection in the *Data* windows for spectral density estimates group box on the *Single Series Fourier (Spectral) Analysis Results* dialog - *Advanced* tab), and the weights used to produce the spectral density estimates. Note that as the default graph for this spreadsheet, you may select to plot the sine/cosine coefficients, the periodogram values (or log-periodogram), or the spectral density estimates (or the log-densities) against the frequencies or period.

We have obtained the coefficients of harmonics which were determined by the option *Summary:*

*Table 3.3. The coefficients of harmonics (by STSTISTICA)*

| *Spectral analysis: VAR3 (_______.sta)* | | | | |
|---|---|---|---|---|
| *No. Of cases: 12* | | | | |
| *Continue …* | *Frequency* | *Period* | *Cosine Coeffs* | *Sine Coeffs* |
| *0* | *0* | | *–3,7E-17* | *–0* |
| *1* | *0,083* | *12* | *2,493* | *0,055* |
| *2* | *0,167* | *6* | *0,098* | *–0,678* |
| *3* | *0,25* | *4* | *–0,328* | *–0,456* |
| *4* | *0,333* | *3* | *–0,455* | *–0,037* |
| *5* | *0,417* | *2,4* | *–0,289* | *0,060* |
| *6* | *0,5* | *2* | *–0,176* | *–0* |

The first harmonic is the most significant, with an amplitude $A = \sqrt{2{,}493^2 + 0{,}055^2} = 2{,}494$ that explains 76.5% the variation of a series. The second harmonic explains 5.8% the variation of a series. The contribution of the other harmonics is equal 17.7%.

Obtained amplitude of oscillation can be use in predicting the seasonal process.

3.4. Autoregression models AR, ARIMA*ARIMAS, ARSH, GARSH. Dynamic factor models and DFM DFMS (with Markov switching)

This section presents the Box-Jenkins Approach, its different models and their basic properties in a rather elementary and heuristic way. These models have become an indispensable tool for short-run forecasts. We first present the most important approaches for statistical modeling of time series. These are autoregressive (AR) processes and moving average (MA) processes, as well as a combination of both types, the so-called ARMA processes and some other. We also show how this class of models can be used for predicting the future development of a time series in an optimal way.

As noted above, the autocorrelation of the levels time series is a characteristic feature for a lot of them. The internal structure of the dynamic range and level of dependence of $y_t$ from its previous values $_{-1}, y_{t-2}, \ldots, y_{t-p}$ described by an *autoregression function* $y_t = a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_p y_{t-p} + e_t$ , where $p$ is a order of autoregression; $a_p$ is autoregression coefficient.

Autoregressive process of order *p (AR (p))* functionally linked to the autocorrelation function $r_p = a_1 r_{p-1} + a_2 r_{p-2} + \ldots + a_p$ , where $p = 1, 2, \ldots, m$ is the autocorrelation lag (displacement $y_t$ to $p$ values ago) $r_0 = 1$.

According to this relation, a unified first-order autoregression coefficient $y_t = a_1 y_{t-1} + e_t$ is equal to the coefficient of the autocorrelation of the first order, that is $a_1 = r_1$. For the second order autoregression $y_t = a_1 y_{t-1} + a_2 y_{t-2} + e_t$ we have a system of equations

$$r_1 = a_1 + a_2 r_1$$

$$r_2 = a_1 r_1 + a_2.$$

Therefore

$$a_1 = -\frac{r_1(1-r_2)}{1-r_1^2},\ a_2 = -\frac{r_2 - r_1^2}{1-r_1^2}.$$

In the simulation, transient in nature of economic processes autoregression function is combined with other methods of analyzing the dynamics: the moving (exponential) average, trend, seasonal wave. It is combining the different models into a single unit greatly expands the scope of their practical use. In addition, the combined model formed based on the same statistical characteristics such as the autocorrelation functions, it is developed an algorithm for calculating the parameters of the model and it is developed one algorithm of model parameters, determined forecasts. Together with the *moving-average (MA)* model, it is a special case and key component of the more *general ARMA* and *ARIMA* (*Autoregressive integrated moving average*) models of time series, which have a more complicated stochastic structure; it is also a special case of the *vector autoregressive model (VAR),* which consists of a system of more than one stochastic difference equation.

*ARIMA* models are applied in some cases where data show evidence of non-stationarity, where an initial differencing step (corresponding to the integrated part of the model) can be applied to reduce the non-stationarity.

The *AR* part of *ARIMA* indicates that the evolving variable of interest is regressed on its own lagged (*i.e., prior*) values. The *MA* part indicates that the regression error is actually a linear combination of error terms whose values occurred contemporaneously and at various times in the past. The I (*integrated*) indicates that the data values have been replaced with the difference between their values and the previous values (and this differencing process may have been performed more than once). The purpose of each of these features is to make the model fit the data as well as possible.

*Non-seasonal ARIMA* models are generally denoted *ARIMA(p,d,q)* where parameters *p*, *d*, and *q* are non-negative integers, *p* is the order (number of time lags) of the autoregressive model, *d* is the degree of differencing (the number of times the data have had past values subtracted), and *q* is the order of the moving-average model. Seasonal ARIMA models are usually denoted $ARIMA(p,d,q)(P,D,Q)_m$, where *m* refers to the number of periods in each season, and the uppercase *P,D,Q* refer to the autoregressive, differencing, and moving average terms for the seasonal part of the ARIMA model.

When two out of the three terms are zeros, the model may be referred to based on the non-zero parameter, dropping *"AR", "I" or "MA"* from the acronym describing the model. For example, ARIMA (1,0,0) is AR(1), ARIMA(0,1,0) is I(1), and ARIMA(0,0,1) is MA(1).

ARIMA models can be estimated following the Box–Jenkins approach.

Practical implementation of the models is possible only if the time series contains at least the 50-60 observations. One can illustrate the multiplicative model ARIMA * ARIMA by an example of the program *STATISTICA.*

*Example 3.1. Monthly US dynamics simulation of carriage passengers (1000's), 1949 to1960 (n = 144).*

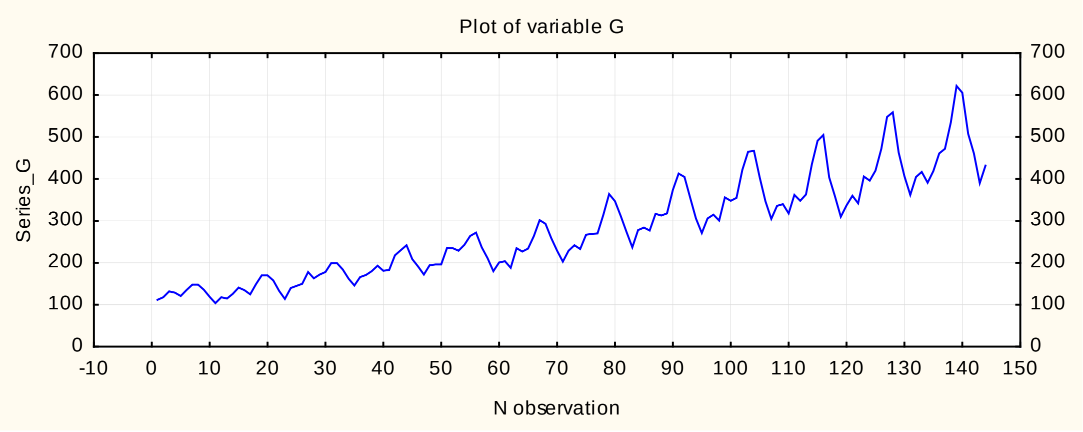



*Fig. 3.4. Monthly US dynamics simulation of carriage passengers (1000's), 1949-1960 (n = 144)*

Visual analysis graph the time series lets you see a clear yearly seasonality (on 12 months lag ), which grows "in the cone" and has the basic trend towards sustainable growth the linear trend. If we use a number of transformation such as *Natural Log*, and take *Difference* at first order Lag 1 to the primary series then we obtain a stationary process. After the identification process of model *ARIMA*ARIMAS*, we get the following options: $(0, 1, 1) \cdot (0, 1, 1)_{12}$

The model parameters are estimated by maximum likelihood, it is only necessary to specify the computational procedure: *Approximate* for approximation or *Exact* - precise in the lower left corner of the dialog box *Single series ARIMA*, iterative procedure of determining the parameters of the model in case of its acceptability makes by the command *Begin parameter estimation* and the evaluating of the results opens in window by *OK* command.

***Single Series ARIMA Results***

*Variable: SERIES_G: Monthly passenger totals (in 1000's)*

*Transformations: ln(x),D(1),D(12)*

*Model: (0,1,1)(0,1,1) Seasonal lag: 12*

*No.of obs.: 131 Initial SS= .273 Final SS= .183(66,95%) MS = .0014*

*Parameters (p/Ps-Autoregressive, q/Qs-Moving aver.); p < .05*

| | $q(1)$ | $Q_s(1)$ |
|---|---|---|
| *Estimate:* | *.40182* | *.55694* |
| *Std. Err.:* | *.09069* | *.07395* |

*Example 3.2. Canadian Fish Price Data*

Statistics Canada reports a number of price indices at each stage in the fish supply chain. For the purposes of this research, data for the period January 1981 to March 2010 have been collected on the price of processed fish and the ex-vessel price of fish. The processed index represents a monthly processed price (Process) of filleted fresh/frozen fish. And the price in the first-hand market is a monthly index for the ex-vessel price (Ex-vessel) of ground fish. The data represent Laspeyres price indices and are transformed to real values using a monthly all commodity consumer price index (CPI) for Canada.

*Price Graphs.* To provide a visual inspection of the data, the two real price series are graphed out in Figure 1. For the real processed price index for filleted fresh/frozen fish the interesting points are the wide swings in the price data up the early 2000s and then after this time the price follows a general downward trend to the current period. Within the data set the highest processed price of fillets is observed in March 1982 and lowest in May 2008.

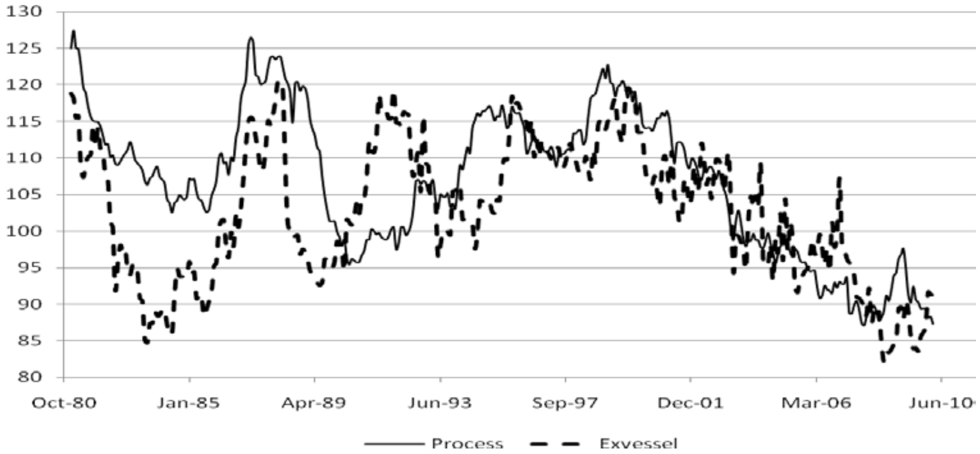



*Fig. 3.5. Processing and Ex-vessel Price of Fish, January 1981 to March 2010*

In the first-hand market for fish we observe an upward or positive trend in ex-vessel prices over the period 1981 to 1999. Notice the wide swings in prices similar to the processed price over the period. After 1999 the price trend is very much negative. In fact, a simple trend line over the two periods shows a positive trend of 0.066 (p-value 0.00) over the first period and a negative trend of -0.221 (p-value 0.00) over the latter period. It also appears as if the variation in price is much wider in the former period compared to the latter. The main point here is that over the last ten years the price received by fishermen for their catch has decreased substantially from a real price index of 119.6 in November 1999 to 91.3 in March 2010. This has implications for both revenue and profit generation. On the revenue side, the price fall tells us that without increases in individual vessel harvest (or constant fleet harvest with a decline in the overall number of fishermen) real revenue to individual fishermen must decline. More so, under this scenario unless the real cost of harvesting has fallen, real income to fishermen must also decline.

It is important to keep in mind that each segment in the supply chain represents a well defined and functioning market. Price determination within each segment is subject to market specific supply and demand characteristics and, of course, supply links between the two segments. Visually it appears that price development in the processing and ex-vessel sectors follows a somewhat similar path. The pair wise price correlation between ex-vessel and processing is measured at just over 60%. The large price shocks early in the data are certainly visible in both sectors and both sectors suffer a general decline in real price over the last ten years.

*Summary Statistics.*Table 1 provides summary statistics for the two price series. The table shows the mean, standard deviation and coefficient of variation (CV). As the prices are indices the mean and standard deviation are useful primarily in calculating the CV. The CV measures the ratio of the standard deviation to the mean. For presentation, the CV has been multiplied by 100. The CV is a unit less measure and allows a comparison of dispersion across the variables of interest; the larger the CV the greater the dispersion in the variable.

*Table 3.4. Summary Statistics of Processed and Ex-vessel Price of Fish*

*January 1981 – March 2010*

| Variable | Mean | Standard Deviation | Coefficient of Variation |
|---|---|---|---|
| Process [a] | 106.84 | 9.82 | 9.19 |
| Ex-vessel [b] | 102.63 | 9.45 | 9.21 |
| Obs. | 352 | | |

[a] Industrial price fresh/frozen

[b] Ex-vessel price

It is interesting that Figure 1 appears to show a difference in variation between the two series but in fact the CV measures very similar variation for the two prices over the period. Given what appears to be a structural break in the ex-vessel price in November 1999 (Figure 3.5), the calculated CV varies somewhat from 8.67 prior to the break to 9.06 after the break. Consequently, whatever the cause, the break not only changed the direction of the price trend but also increased the variation in ex-vessel prices.

The final summary table will describe some time series or data generating properties of the price variables. If the data generating process is stable this indicates that the mean, variance and pairwise correlations of the realizations are stable or stationary over time. If on the other hand this is not true, then econometric modelling of such non-stationary variables tends to measure common trends in the data and the underlying economic relationship of interest is obscured. A number of statistics are available for testing stationarity and here the augmented Dickey-Fuller approach is used with constant, trend and six lags for testing (Gordon 1995). In the level form of the variables, the null hypothesis is that the price series is characterized as nonstationary with an alternative hypothesis of stationary in first-differenced values of the variable. For each of the price variables the results of the test are reported in column 2 of Table 3.5. In both cases we cannot reject the null hypothesis at p-values less than 5%. Next, we take the first differences of the variables and reapply the test. The null hypothesis is that the series is stationary in second-differences against as alternative hypothesis of stationary in first-differences. The results are reported in column 3 and now for all price variables we can easily reject the null and accept the alternative hypothesis of stability/stationarity in the first-difference values of the variables.

*Table 3.5. Tests for Stationarity[a)]*

| | Dickey-Fuller Levels | Dickey-Fuller First-differences |
|---|---|---|
| Process [b)] | -1.92 | -6.08 |
| | (0.64)* | (0.00) |
| Ex-vessel [c)] | -2.50 | -8.79 |
| | (0.33) | (0.00) |

[a)] All statistics include constant, trend and 6 lags.

[b)] Processed price

[c)] Ex-vessel price

* Mackinnon approximate p-value

What these stationarity results mean is that, in most case, we must approach modelling using the first differences of the prices rather than the levels. It is argued that using first differences rather than levels removes long-run price information in the data but this can be recovered within an error-correction model as we will see below. First, we will concentrate on a univariate short-run model using first differenced price data. *An ARMAX Model of Ex-vessel Prices: T*he initial modelling will be to fit an ARIMA model to the ex-vessel price data. This is a univariate modelling technique based on the maintained assumption that current realizations of price can be explained by lagged values of the price (dynamic shocks) and current and lagged values of the stochastic error term (stochastic shocks). The ARIMA can be considered a reduced form price model for the purpose of short-run forecasting. It is possible to augment the ARIMA price model by including exogenous variables in specification for the purpose of improving forecasting possibilities and to reduce forecast error. These extensions are defined as ARMAX or transfer function models and for the case at hand the US/Canada exchange rate and seasonal dummy variables may serve this purpose well. Of course, as our data are nonstationary the price equation will be modelled in first differences rather than levels.

*Univariate model :* the specification of the univariate price model is defined as:

$$\Delta Exvessel_t = \delta_o + \alpha_p \Delta Ex_t + \sum_{s=1}^{12} \beta_s D_s + \sum_{i=1}^{p} \gamma_{Vi} \Delta Exvessel_{t-i} + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t \quad (1)$$

where $\Delta$ represents first differences, $Exvessel_t$ is the ex-vessel price index for fish in period *t*, $Ex_t$ is the US/Canada exchange rate, $D_S$ are seasonal monthly dummies, $\sum_{i=1}^{p}\gamma_{Vi}\Delta Exvessel_{t-i}$ represents the autoregressive (AR) component (dynamic shocks), $\sum_{j=1}^{q}\theta_j\varepsilon_{t-j}$ represents the moving average (MA) component (stochastic shocks) and $\varepsilon_t$ is an *id* random error term. Estimation of Equation (1) is routine using maximum likelihood procedures.

Selecting the correct lag specification for Equation (1) is critical for generating an estimated equation with good forecasting potential. Our research strategy is to evaluate alternative AR and MA lag structures based on review of the autocorrelation and partial autocorrelation functions with possible candidate specifications defined on testing *id* conditions in the stochastic error term using a Box-Lung Q-statistic. Among those candidate specifications the preferred model is identified by measured RMSE and AIC statistics. For comparison and evaluation purposes both the estimated ARIMA and ARMAX models will be reported.

Initial evaluation of the ex-vessel price showed very strong monthly effects with prices slightly but significantly higher in the months August through January. Testing showed that we are able to control seasonal effects using a single dummy variable with the value 1 in the months mentioned and zero otherwise. Following this procedure the final specification of the ARIMA model resulted in an AR specification of one and twelve lags and a MA specification of twelve lags. The results for both the ARIMA and ARMAX are reported in Table

3.

*Table 3.6. ARIMA and ARMAX Equations for the Ex-vessel Price*

| Variables | ARIMA | ARIMAs | ARMAX |
|---|---|---|---|
| Exchange[a)] | - | - | 0.129 |
| | | | (0.111) |
| Seasonal[c)] | - | 0.016 | 0.017 |
| | | (0.000) | (0.000) |
| $Exvessel_{t-1}$[d)] | -0.212[b)] | -0.240 | -0.244 |
| | (0.000) | (0.000) | (0.000) |
| $Exvessel_{t-12}$[e)] | -0.352 | -0.406 | -0.425 |
| | (0.037) | (0.008) | (0.003) |
| $\varepsilon_{t-12}$[f)] | 0.577 | 0.585 | 0.607 |
| | (0.000) | (0.000) | (0.000) |

| | | | |
|---|---|---|---|
| Intercept | -0.001 | -0.009 | -0.009 |
| | (0.624) | (0.000) | (0.000) |
| | | | |
| RMSE | 0.0276 | 0.0264 | 0.0263 |
| AIC | -1507.03 | -1537.52 | -1538.02 |
| Q-statistic | 60.97 | 36.02 | 37.26 |
| | (0.018) | (0.649) | (0.594) |
| Observations | 350 | 350 | 350 |

[a)] US/Canada exchange rate

[b)] p-value in parentheses, robust standard errors

[c)] Seasonal dummy variable

[d)] AR lag 1

[e)] AR lag 12

[f)] MA lag 12

The second column in Table 3.6. reports the basic ARIMA model with all ARMA components statistically significant but the Q-statistic, testing the null hypothesis of no correlation in the estimated errors is rejected with a p-value of less than 2%. We were unable to find any reasonable combination of ARMA structure that generated statistically significant parameters and non-correlated predicted errors. This lead to a search for seasonality in the price series and resulted in the estimates reported in column three under the heading ARIMAs. Here we observe that the seasonal dummy and the ARMA components are statistically important with a small Q-statistic. Also, note the reduction in RMSE and the AIC statistic relative to ARIMA. Clearly ARIMAs is statistically preferred to ARIMA.

We extend the ARMAX structure one step further by including the US/Canada exchange rate in specification. These results are reported in column four under the heading ARMAX. The inclusion of the exchange rate variable does in fact reduce the RMSE of the specification and lowers somewhat the AIC criteria relative to the ARIMAs model. However, the exchange rate although correctly signed is itself only significant with a p-value of 11%. Nevertheless, international trade is very important for Canadian fisheries and short-run price forecasting will be carried out using the ARMAX model.

*Price Forecasts:* the ARMAX model will be used to forecast both in-sample and dynamic forecasting. For in-sample forecasting of the ex-vessel price the actual values of the seasonal dummy, exchange rate and ARMA components are used in making the one step ahead forecast. Whereas, for dynamic forecasting the actual values of the seasonal dummy and exchange rate are combined with the predicted value of the ARMA components for forecasting. For purposes of presentation in-sample forecasting is carried out for the period January 1981 to October 2007 with dynamic forecasts over the period November 2007 to March 2010. Figure 2a graphs out the in-sample forecasting of ex-vessel price whereas Figure 2b provides a closer look at the dynamic forecast.

Figure 3.6a. shows that one step ahead in-sample forecasting of the ex-vessel price of fish is very good but this is what is expected in a well specified equation. But it does provide further evidence that the ARMAX model is a reasonable candidate for short-run ex-vessel price forecasting.

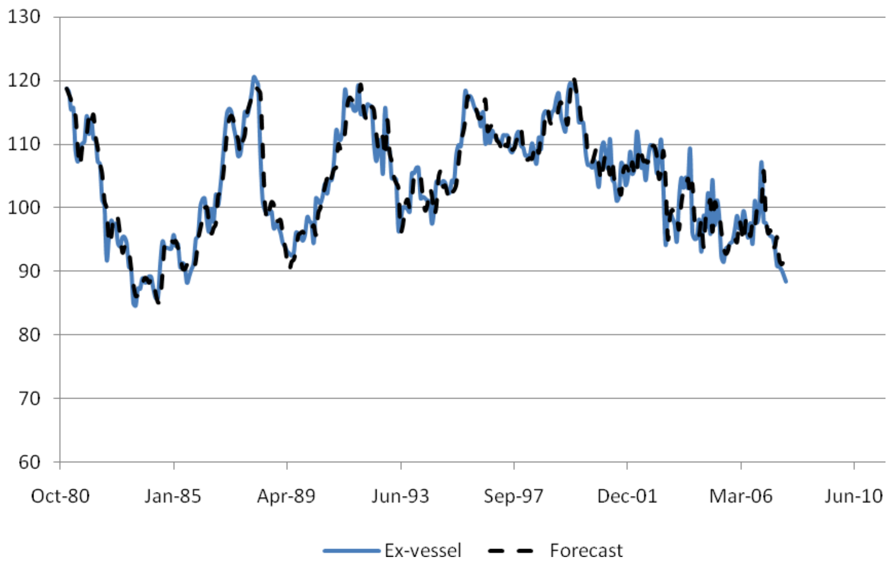



*Figure 3.6a: In-Sample Ex-vessel Price Forecasts: January 1981-October 2007*

Figure 3.6b. shows the dynamic forecast over the last months of the data period. The merit of ARIMA modelling is in dynamic forecasting. The dynamic forecast does a good job of picking up all four turning points in the series but does not capture the full magnitude of the variation in actual ex-vessel prices. Nevertheless, this model has much to recommend it and we will move on to out-of-sample forecasting under alternative exchange rate scenarios.

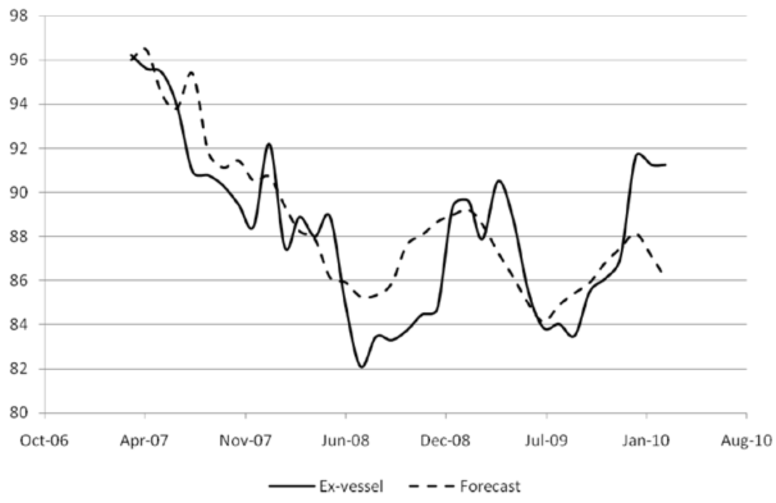



*Figure 3.6b. Dynamic Ex-vessel Price Forecasts: November 2007-March* 2010

Over the period of our data the US/Canada exchange rate reached a high in January 2002 (1US=1.61Can) and low in November 2007 (1US=0.967Can) and for simulation purposes these rates are set as the upper and lower bound on the exchange. Dynamic forecasting will follow ex-vessel price realizations as we simulate the exchange rising from par to the highest bound and repeat the simulation from par to the lowest bound. The purpose is to compare the difference in forecasted ex-vessel prices from the different exchange rate scenarios. In interpretation we must be mindful that such out-of-sample forecasting depends on starting values or in other words the forecast is path dependent. The starting point is set as October 2007 when the actual exchange rate was par. The results of the two exchange rate dynamic forecasting scenarios are shown in Figure 3.7.

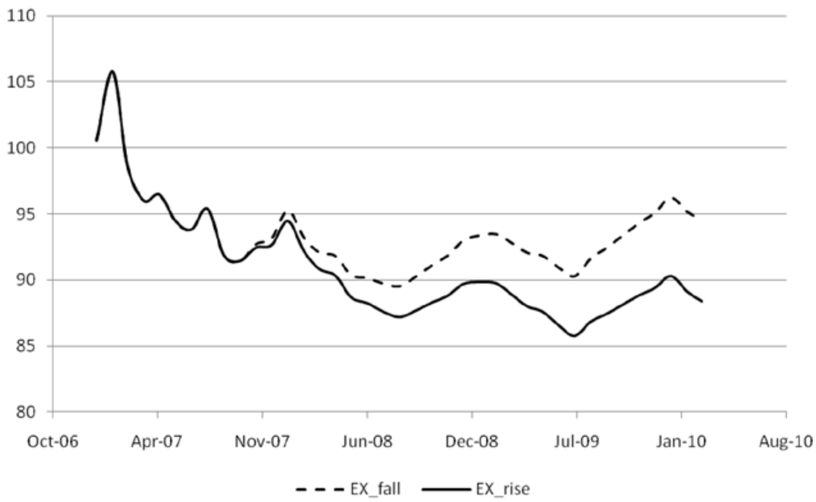



*Figure 3.7. Ex-vessel Price Forecasts, Two Exchange rate Scenarios*

The upper dashed line in Figure 3.7. show ex-vessel price predictions as the value of the Canadian dollar falls relative to the US dollar and the solid bottom line shows price predictions as the value of the Canadian dollar rises relative to the US dollar. The results are as expected with ex-vessel price increasing as a cheaper Canadian dollar increases US demand for Canadian fish.

At the upper and lower bounds of the simulation, predicted values reflect a 6.3% difference in the high to low ex-vessel price. This is an interesting but serious complication for fisheries managers: Interesting because it statistically shows the importance of international factors impact the welfare/income of Canadian fishermen and complicating because regardless of the success of fisheries management on stocks and harvest levels the ultimate impact on fishermen is influenced by external/exogenous factors.

To provide an alternative but realistic exchange rate scenario consider the current world situation where the US dollar is falling against the major world currencies. In this scenario we will simulate the ex-vessel price impact of the Canadian dollar on par with the US. For dynamic forecasting we pick as a starting point March 2009 when the actual exchange rate was 1US=1.265Can and allow the Canadian dollar to gradually rise each month until par with the US dollar. Under this scenario the movement of the exchange rate to par causes the real ex-vessel price to Canadian fishermen to fall by 5.3%.

Dynamic price forecasts are certainly useful in providing probable ex-vessel price paths in the short run and can be used to predict the consequence of exogenous shocks in the system. We turn now to multi-variable price modelling of the links in the Canadian fish supply chain.

*Error-Correction Modelling of Ex-vessel Prices in Canadian Fisheries: r*esearch in applied time series econometrics is concerned with the issue of short-run dynamics and long-run equilibrium (Engle and Granger 1987). The problem is that short-run dynamics, as represented by vector autoregressive models are subject to omitted variable bias if long-run parameters are neglected in specification. Error-correction modelling is an attempt to combine both short- and long-run parameters in a single equation. What is interesting is that long-run parameters are derived from prices in level (i.e., non-stationary) form and short-run parameters are derived from prices in first difference (i.e., stationary) form. The idea is that there may exist a vector of parameters that in combination with the level prices forces stationarity. Estimation of the error-correction model for the first-hand/processing segment will provide a statistical picture of the short- and long-run price parameters that link the markets.

*The EC Model:* in a pairwise comparison of prices in the two market segments of the fish supply chain, it is not unreasonable to think of an economic equilibrium describing the long-run relationship and written as:

$$P_t^v = \beta_o + \beta_l \cdot P_t^p + \varepsilon_t \qquad (2)$$

where $P_t$ is a spesific market price in log level form, and $v\ and\ p$ represent the vessel and processing market segments, respectively. If equilibrium exists then $(1, -\beta_0, -\beta_l)$ represents a cointegrating vector that produces a random error term

$\varepsilon_t$ that is stationary. Moreover, $\beta_l$ defines the equilibrium link (or long-run parameter) between prices in the different market segments. Rearranging equation (2) and lagging one period the equation can be written in error form as:

$$\varepsilon_{t-1} = P_{t-1}^{v} - \beta_o - \beta_l \cdot P_{t-1}^{p} \tag{3}$$

In equilibrium the error term takes the value zero, but if process prices are above the equilibrium the error term is negative and below the equilibrium the error term is positive. The error equation in combination with the short-run model ensures that prices above the equilibrium are adjusted downward and prices below the equilibrium are adjusted upward. By including equation (3) in a short-run distributed lag representation the error-correction model can be written as:

$$\Delta P_t^{v} = \sum_{i=1}^{I} \alpha_i \Delta P_{t-i}^{v} + \sum_{j=0}^{J} \alpha_j \Delta P_{t-j}^{p} + \gamma(P_{t-1}^{v} - \beta_o - \beta_l \cdot P_{t-1}^{p}) + \vartheta_t \tag{4}$$

where short-run parameters are defined by $\alpha_i \, and \, \alpha_j$, long-run parameters by $\beta_l$, and $\gamma$ is the speed of adjustment to regain the equilibrium. The speed of adjustment to recover the equilibrium is an important characteristic of the market. The first point of interest is that all arguments in Equation (4) are stationary ($\sim I(0)$) and with correct specification of lag structure on the distributed lag (short run) variables the error term $\vartheta_t$ is non-autocorrelated and asymptotically normal. Second, the speed of adjustment to a short-run price shock can be approximated with the simple expression $\tau = ((1 - |\gamma|)/|\gamma|)$. Finally, a possible serious econometric problem in Equation (4) may result if there is correlation between the current valued process price and the error term i.e. $E\left(\vartheta_t P_t^{p}\right) \neq 0$ and between the change in actual vessel price and the error term $E(\vartheta_t \Delta P_t^{p} \neq 0$. This problem can arise by feedback links between the market segments and will cause inconsistent parameter estimates in both Equations (3) and (4) (Hahn, 1990). To avoid this problem it is necessary that current valued process price and the change in current valued process price be weakly exogenous to the error term both in Equations (3) and (4), respectively. What this means is that in an economic sense price leadership both in the short run and long run, moves from the process sector to the vessel sector without feedback effects. Boswijk and Urbain (1997) suggest a modified application of the Hausman test or variable inclusion test to test weak exogeneity. Consider a distributed lag model defined for processed price

$$\Delta P_t^{p} = \delta_o + \sum_{k=1}^{K} \delta_k P_{t-k}^{v} + \sum_{s=1}^{S} \delta_s P_{t-s}^{p} + \epsilon_t \tag{5}$$

Von Cramon-Traubadel (1998) refers to Equation (5) as a marginal model. Long-run exogeneity can be tested using Equation (5) by including the error-correction term, $\varepsilon_{t-1}$ of Equation (3) as an additional regressor. The equation is estimated under the assumption of weak exogeneity in processed price and an F-test is appropriate in testing.

Short-run exogeneity can be tested by estimating first Equation (5), calculating the fitted residuals $\widehat{\epsilon}_t$ and including this as an additional variable in

Equation (4). Equation (4) is re-estimated under the null hypothesis of exogeneity in processed price. An F-test is appropriate in testing.

*Validation and Hypothesis Testing: a*s with the univariate model, selecting the correct lag specification in Equation (4) and (5) is critical in model specification and selection. The autocorrelation function is used to provide possible candidate specifications and this is augmented with both t-tests of individual parameters and tests to measure for correlation in the stochastic error term. The final specification is based on *id* error terms defined using a Box-Lung Q-statistic and comparisons of measured RMSE and AIC statistics.

Following Halicioglu (2008) cointegration is tested in Equation (4) from a null hypothesis of no cointegration or $H_o\colon \gamma = 0\ and\ \gamma \cdot \beta = 0$ against the alternative of cointegration. Critical values for the test are provided by Pesaran *et al.* (2001).

*Model Estimates:* table 5 reports the results from estimating the error-correction and testing equations. Column 2 reports the test for short-run weak exogeneity based on Equation (4) by including the predicted errors from Equation (5). Column 3 reports the test for long-run weak exogeneity based on Equation (5) including the error-correction term from Equation (3). The F-statistics used in testing are reported at the bottom of the table. Both F-statistics provide support for the conclusion that process price in terms of vessel price can be considered weakly exogenous in both the short-run and long-run equations. This is a strong result and indicates that process price is the driving factor or price leader in determining ex-vessel price. Although a strong result it is consistent with published research in agricultural economics from producer to processor (Von Cramon-Taubadel, 1998).

The error-correction results are reported in column 4 of Table 5. Final specification of the equation relies on the current change in process price plus the first and third lags of the process price variable. The second lag on process price was insignificant and was dropped with no statistical consequence from the equation. The first lag of vessel price is also included as is the error-correction term from Equation (3). It is interesting that this specification fails to produce white noise error terms based on the Box-Lung test. From experience with the ARIMA selection process, we search for omitted seasonality in vessel price data. We found that by including a 12$^{th}$ lag on vessel price the Box-Lung test showed white noise error terms. Note the speed of adjustment coefficient is measure at 0.072 (absolute value) and statistically significant implying a somewhat long time to adjustment of approximately 13 months. It is the long lag on vessel price that causes the lengthy adjustment process.

Prior to evaluation of the error-correction model, we test to ensure a long-run cointegratred model exits. This test will also generate the long run parameter of interest. The empirical process is to re-estimate Equation (4) but include the full specification of the long run model (Equation 3). The hypothesis of no cointegration is tested from a null of $H_o\colon \gamma = 0\ and\ \gamma \cdot \beta = 0$ against a general alternative. The testing generated a calculated $F_{(2, 330)}$ -statistic of 8.45 and falls outside of the bounds test rejecting the null hypothesis and indicates a cointegrated system. The long-run estimated equilibrium is written as (p-value in parentheses):

$$e_{t-1} = Vessel_{t-1} - \underset{(0..00)}{0.605} Process_{t-1} \quad (6)$$

This equation supports a long-run price elasticity where just over 60% of a price change at the processing level is passed on to the vessel level.

Returning to Table 3.6, the short run parameters show the response elasticities of the changes in vessel price to a shock in the system. Short run vessel price is impacted by current and lagged values of change in process price and past change in vessel price. Note that short-run elasticities with respect to process prices are less than half the magnitude of the long run value and declining in lagged response.

*Table 3.7. Marginal Equation Error Correction, Processing and Error Correction Equation*

| Variables | Marginal Error Correction | Marginal Processing | Error Correction |
|---|---|---|---|
| ΔProcess | -0.250<br>(0.530) | - | 0.321<br>(0.002) |
| ΔProcess_1 | 0.325<br>(0.009) | 0.127<br>(0.040) | 0.244<br>(0.023) |
| ΔProcess_2 | - | 0.008<br>(0.890) | - |
| ΔProcess_3 | 0.196<br>(0.094) | -0.023<br>(0.689) | 0.212<br>(0.070) |
| ΔProcess_4 | - | 0.105<br>(0.023) | - |
| ΔProcess_12 | - | 0.137<br>(0.030) | - |
| ΔProcess_13 | - | 0.134<br>(0.017) | - |

| | | | |
|---|---|---|---|
| ΔProcess_14 | - | -0.123<br>(0.015) | - |
| ΔVessel_1 | -0.161<br>(0.022) | 0.017<br>(0.503) | -0.149<br>(0.031) |
| ΔVessel_2 | - | 0.020<br>(0.484) | - |
| ΔVessel_3 | - | 0.015<br>(0.645) | - |
| ΔVessel_4 | - | -0.052<br>(0.095) | - |
| ΔVessel_12 | 0.279<br>(0.000) | - | 0.224<br>(0.000) |
| EC_v | -0.067<br>(0.001) | 0.008<br>(0.403) | -0.072<br>(0.000) |
| r_v | 0.618<br>(0.36) | - | - |
| RMSE | 0.0272 | 0.0128 | 0.0272 |
| AIC | 1460.97 | -1959.76 | -1471.14 |
| Q-statistic | 46.93<br>(0210) | 50.86<br>(0.117) | 50.15<br>(0.131) |
| Obs. | 338 | 336 | 338 |
| Short-run F-statistic | 2.23<br>(0.136) | - | - |
| Long-run F-statistic | - | 0.70<br>(0.403) | - |

A useful way to evaluate the equilibrium and the error-correction model is to graph out the dynamic response to a shock in process price. This dynamic response is shown in Figure 3.7. and 3.8. for two alternative scenarios. In Figure 3.7. we allow for a one time price shock to the process price and then return to pre-shock levels in the next period. Figure 3.7a graphs out the dynamic response of vessel price to the shock to process price (i.e., the cointegrated equation) and Figure 3.7b. graph out short run changes in prices (i.e., the error-correction equation). In Figure 3.8a., we allow for a persistent price shock to process price and report price dynamics in Figure 5a and short run changes in price in Figure 3.8b.

In Figure 3.7a, equilibrium vessel price at first jumps up in response to a one time shock to process price. However, in the next period process price returns to pre-shock levels and now the response process reduces vessel price but overshoots the pre-shock vessel level. There are two reasons for this; the return to pre-shock process price in the second period causes a negative adjustment in vessel price, and at the same time in the second period the error-correction forces an additional correction on vessel price caused by the first period positive price shock. The oscillations in vessel price decrease overtime but note we observe the seasonal impact in the market at long lag length. Figure 4b drafts out the change in vessel price from the change in process price. Notice the short run vessel prices both increase and decrease in response to the shock but are not long lasting.

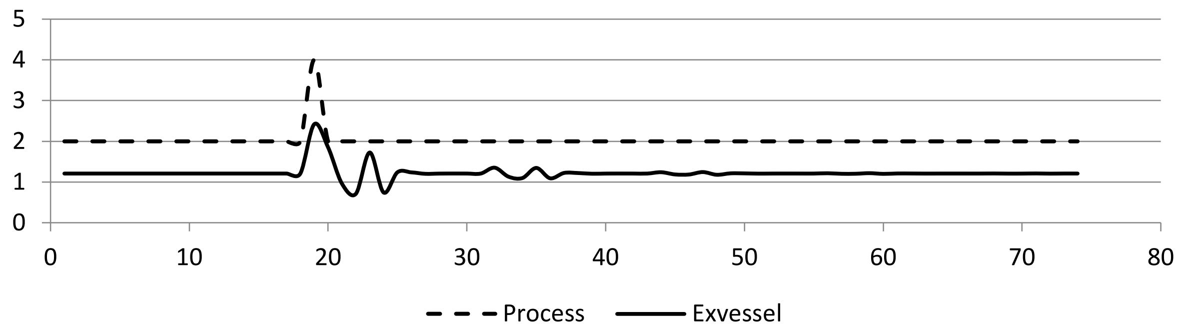



*Figure 3.7a. Price Simulation: One time price shock*

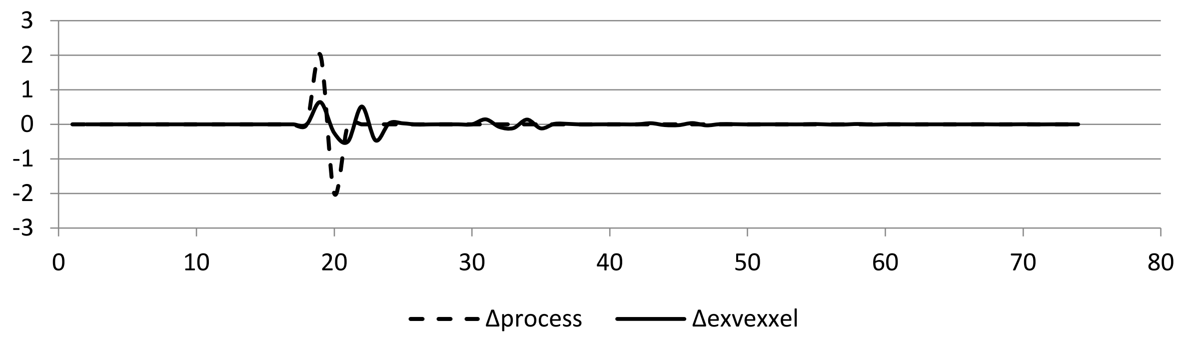



*Figure 3.7.b Change in Price Simulation: One time price shock*

Figure 3.8a shows the long run effect of a positive and persistent price shock to process price. In response vessel price initially overshoots the new equilibrium price level but stability in the system forces a return to the new equilibrium price. Notice in Figure 5b that short run price adjustment shows a more moderate adjustment to a persistent price shock relative to Figure 3.8b.

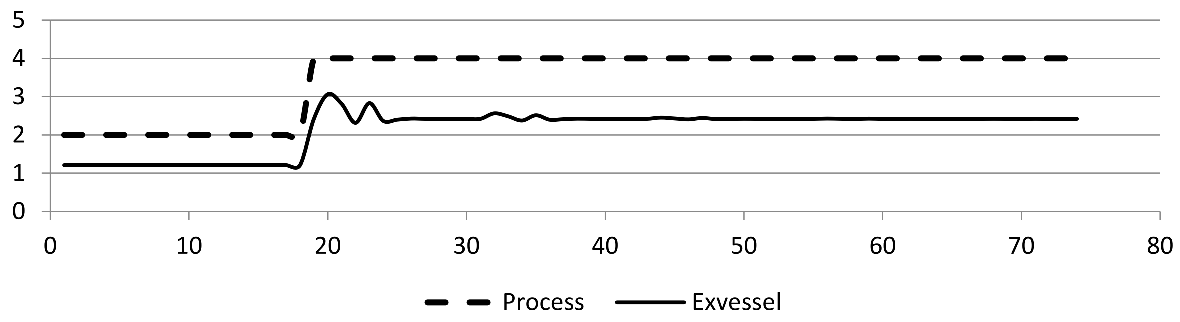



*Figure 3.8.a Price Simulation: Persistent price shock*

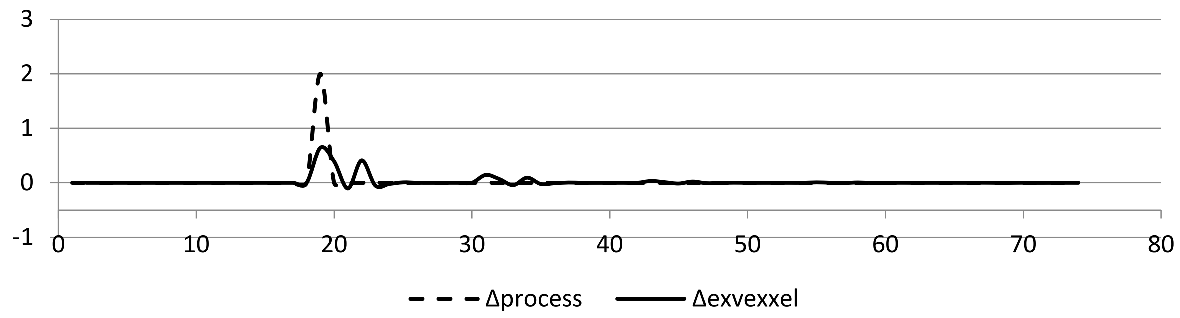



*Figure 3.8b. Change in Price Simulation Persistent price shock*

The ARMAX model fit the data well providing dynamic forecasts and policy evaluation. Interestingly we show that changes in the US/Canada exchange rate impact the vessel price with rising Canadian dollar value causing downward pressure on ex-vessel prices. This is an important fisheries management issue to the extent that it is an exogenous shock impacting income levels of fishermen. The ARMAX model is statically straightforward and can be updated by fisheries managers for up-to-date short-run price forecasts.

The error-correction model provided a statistical characterization of the links between the process and vessel markets. Importantly testing showed that there exists a long run-equilibrium relationship across process and vessel prices. What this means is that there exists natural market forces that allow for price deviations in the short run but eventually force price to return to the equilibrium. Second, we show that process price can be considered weakly exogenous with respect to both the short and long run. In other words, process price acts like a price leader in determining vessel price. We measure that a one time shock can cause serious variation in vessel price but the shock is stable and prices do return to pre-shock levels Future research in this area could approach the price problem in the spirit of structural modelling and concentrate on introducing a retail demand shift variable and a marketing cost index to augment model specification and improve forecasting. In addition, the issues of asymmetric price response to upside and downside process price shocks is a natural extension to this work.

If the nature of the dynamics of rapidly is changing under the influence of external factors, we can use the model *Interrupted ARIMA* - (with intervention).

Modeling of dynamic processes with the effect of saturation when the rate of increase (decrease) the level of slowing down and approaching a certain limit (specific consumption of resources, food consumption per capita, etc.) has its own characteristics. One can use a class of curves with horizontal asymptote to describe it using $K \neq 0$.

The simplest of these is a modified exponential function: $Y_t = K + ab^t$, where parameter $a$ is a difference between ordinate $Y_t$ at t $t = 0$ and asymptote K. If $a < 0$, the asymptote is above the curve, if $a > 0$ it is below the asymptote of the curve. The parameter $b$ describes the relation of successive increments ordinates.

If coordinates on the time axis have a normal distribution then these ratios are constant: $b = \frac{Y_{t+1} - Y_t}{Y_t - Y_{t-1}} = const$ . The modified exponential function describes processes to which operates a limiting factor, and the impact of this factor increases with increasing $Y_t$.

If the limiting factor affects only after a certain point, to which the process developed by the exponential law, then the process is best approximated by S-shaped function inflection at point P. The accelerated growth is changed to the slow down at this point. For example, the demand for a new product at first is insignificant; then, if consumers accept, it is growing rapidly, but as we have the market saturation, the growth rate is slowing down, it is extinguished. Demand stabilized at a certain level. There is a similar phase of development with process innovations and inventions, resource efficiency and more. Pearl-Reed function or *logistic curve*: $Y_t = \frac{K}{1 + be^{-at}}$ . is the most common among the *S-shaped* curves that describe the full cycle of development.

If the indicator of a process is the proportion that varies from 0 to 1, the form of logistic function is simplified $Y_t = \frac{1}{e^{a+bt} + 1}$ or $\frac{1}{Y_t} = 1 + e^{a+bt}$ .

One can use another S-shaped function such as a *Gompertz curve* $Y_t = Ka^{b^t}$ : (or by logarithms $\lg Y_t = \lg K + b^t \lg a$ ) in insurance and demographic statistics. That curve is adjusted to the modified Gompertzian exhibitors, whose constant is the ratio in increments of ordinates logarithms.

Thus, the Gompertz curve is converted to the modified exponential function, in which the ratio of ordinate growth (in logarithm ) are constant.

Estimation of parameters the functions with asymptotes is much more difficult compared with polynomials and exponential functions. We can use two options.

By the first option, asymptote as a norm or a standard is defined a priori $K$*. Then the modified exponential function can be written as $(Y_t - K^*) = ab^t$ .

If $z$ substitute for $(Y_t - K^*)$ then we take logarithm of equation, and obtain a linear function of the logarithms $lgz = lga + tlgb$.

Similarly, logistic functions $1/Y_t = K^* + ab^t$ can be linearized if we $z$ substitute for $(Y_t - K^*)$ obtain lgz = lg$a$ + $t$lg$b$.

The transformed linear form the parameters of functions and polynomials parameters can be estimated by least squares using procedures *Multiple Regression*. Forecast and its confidence limits can be defined traditionally , although the confidence limits of the forecast for the full cycle curves are arbitrary.

By the second option, the asymptote is unknown, therefore, necessary to determine all three parameters: *K, a, b*.

There are various procedures that implement through *OLS* method. Since the logistic curve and the curve Gomperz are transformed to the modified exponential function, it is advisable to consider the method is described by Briant (the so-called three points) for all of three functions.

By this method, parameter *b* is determined at first, then *a* and *C*. For example, the formulas for calculating the parameters modified exponential function are:

$$b = \frac{(n-1)\sum_{1}^{n-1} y_t y_{t+1} - \sum_{1}^{n-1} y_t \sum_{1}^{n-1} y_{t+1}}{(n-1)\sum_{1}^{n-1} y_t^2 - \left(\sum_{1}^{n-1} y_t\right)^2};$$

$$a = \frac{n\sum_{1}^{n} b^t y_t - \sum_{1}^{n} b^t \sum_{1}^{n} y_t}{n\sum_{1}^{n} b^{2t} - \left(\sum_{1}^{n} b^t\right)^2};$$

$$K = \frac{\sum_{1}^{n} y_t - a\sum_{1}^{n} b^t}{n}.$$

Calculation of the S-shaped curves (or other non-linear regressions) can be implemented in *Nonlinear Estimation* module. Users should choose the type of function on their own by using the *User-specified regression (STATISTICA)*

*Example 3.3. For example, apply the logistic curve to the data series of population dynamics of the metropolis*:

*Table 3.8. Annual population dynamics metropolis*

| *Year* | *Million. people..* | *Year* | *Million. people.* | *Year* | *Million. people..* |
|---|---|---|---|---|---|
| *1950* | *3,48* | *1970* | *4,78* | *1990* | *6,14* |
| *1955* | *3,86* | *1975* | *5,13* | *1995* | *6,37* |
| *1960* | *4,17* | *1980* | *5,52* | *2000* | *7,04* |
| | *4,56* | *1985* | *5,90* | | |

In the dialog box *Estimated function & loss function*, we should choose kind of functional type curve. $v2 = b1/(1 + b2 \cdot \exp(-b3 \cdot v1))$, where $b1 = K$, $b2 = a$, $b3 = b$

Regarding the loss function, you can use residual deviations, which is determined by the system default. We identify feature for dynamics which is modeled (in this example - *v2*), and the method of parameter estimation (*Quasi-Newton*) by button Variables.

After we click *OK* to open box Result. The correlation index R = 0,997 indicates the high approximating quality of a model. One can estimate of parameters through Parameter estimates are as follows:

*Table 3.9. Estimation of model parameters*

| *Model: v2 = b1 / (1 + b2*exp(–b3*v1)) (_______.sta)* | | | |
|---|---|---|---|
| *Final loss: ,0723 R=,997 Variance explained: 99,428%* | | | |
| | *B1* | *B2* | *B3* |
| *Estimate* | *12,37* | *2,79* | *0,114* |

According to the data, population growth of metropolis for five years averages 11.4%, approaching the limit - 12.37 m. per.

Consequently, there is wide enough class of dynamics models, and they describe different processes of development.

Selecting the type of model in a particular study is based primarily on the theoretical analysis of the specific process, its internal structure, relationships with other processes. A character of dynamics (steady, accelerated, with saturation, etc.) is defined on the basis of this analysis and it is clarifying a range of functions that can approximate this process. Formal methods serve as a major help in choosing a particular model.

For example, we use analysis of successive differences for polynomials.

Equality of differences of the *p*- order is considered as a sign that the process was described by polynomial p-order.

For the difference is 1st order $\Delta'_t = y_t - y_{t-1}$, we use approximately the same linear trend, if there is the difference is 2nd order $\Delta''_t = \Delta_t - \Delta_{t-1}$ , we use parabola and so on.

Some difficulties in selecting a model can arise for the exponential function.

S-curve is described as the exponential function before the inflection point but the inflection point can be outside the dynamic series.

So if saturation is theoretically possible in the future and the process can fade or there are some limitations to the process (legal, financial resources, capacity utilization, etc.), we should prefer S-curve.

The significant variation of levels Y is characteristic for a primary dynamics series; thus, analysis of differences should conduct on base the series of moving averages.

The following table summarizes the main characteristics (differences) of this analysis (a priori tests), which determine the specific type of model complete cycle:

*Table 3.10. Chain pace (growth)*

| *Characterization* | *Property of Characterization* | *Type of trend* |
|---|---|---|
| $\Delta'_t$ | *About the same* | *Polynomial of degree 1* |
| $\Delta'_t$ | *It is changed as linear* | *Polynomial of degree 2* |
| $\Delta'_t / y_{t-1}$ | *About the same* | *Exponential function* |
| $\lg \Delta'_t$ | *It is changed as linear* | *Modified exponential function* |
| $\lg \Delta'_t / (y_{t-1})^2$ | *It is changed as linear* | *Logistic curve* |
| $\lg \Delta'_t / y_{t-1}$ | *It is changed as linear* | *Gompertz curve* |

If we study the reverse direction of the trend, the difference is calculated from the end. So as the differences are negative we cannot use logarithm, so you need to increase the interval smoothing moving average.

*Example 3.4. Modeling and forecasting VAT revenues at the micro and macro levels.*

Tax forecasting carried out to assess the potential tax state as a whole and for individual regions, moreover, we determine the likelihood of unforeseen events and social nature unpredictable financial, in particular, tax situations by forecasting.

In terms of the unstable nature of economic processes in practice, in fact, one can be realized only two types of forecasts such as an operational (a one month) and short-term (a one year).

In forecasting tax revenue, we should take into account the actual time series of income taxes for prior periods. Also, it is necessary to consider that the figures for previous periods need to adjust in many cases, which requires the study of specific economic features at a given period by considering the value of tax rates, changes in tax base and so on.

The goal of the example is to build models of the dynamics of monthly VAT revenues at the micro level (according to State Tax Inspectorate in Prymorskyi District for the period from 2005 to 2011) and at the macro level (according to State Statistics Committee of Ukraine and other sources for the period from 2007 to 2011) with amendments "emissions" and with variable damping or trend. The example is based on the adequate models of forecasting performance.

I. First, we consider the micro level, i.e. VAT revenues of the district STI. In the simulation, we use the following data tax revenue of Primorsky district. Odesa:

*Table 3.11. The dynamics of monthly VAT revenues to Primorsky district STI. Odesa for the years 2005-2011 (thousand UAH)*

| Years | *2005* | *2006* | *2007* | *2008* | *2009* | *2010* | *2011* |
|---|---|---|---|---|---|---|---|
| Months | VAT | VAT | VAT | VAT | VAT | VAT | VAT |
| *1* | *9949,278* | *16722,58* | *20829,38* | *42686,98* | *21587,9* | *17634,8* | *60078,4* |
| *2* | *6828,165* | *10289,37* | *14043,34* | *15902,04* | *12054,5* | *14515,8* | *45384,5* |
| *3* | *10554,97* | *18059,67* | *12967,92* | *17969,49* | *22632,4* | *16492* | *63664,1* |
| *4* | *9687,427* | *13955,77* | *17624,91* | *23458,78* | *19021,4* | *17352,8* | *66547,9* |
| *5* | *11115,32* | *12869,36* | *18007,45* | *20621,95* | *16379,8* | *18471,2* | *53162,1* |
| *6* | *10858,56* | *13757,19* | *16871,61* | *20637,98* | *18298,4* | *19627,6* | *76498,3* |
| *7* | *10731,88* | *14440,41* | *18850,29* | *21330,15* | *22959* | *21850,2* | *85979,3* |
| *8* | *11947,37* | *14660,06* | *24756,63* | *21722,23* | *18948,4* | *22804,4* | *83397,7* |
| *9* | *13666,76* | *16533,46* | *18642,41* | *25945,56* | *24279,2* | *32101,8* | *94064,1* |
| *10* | *14272,88* | *16879,82* | *18539,1* | *25706,97* | *19213,1* | *46736,2* | *109826,8* |
| *11* | *13957,57* | *14843,69* | *20204,11* | *21033,23* | *20100* | *61695,4* | *109536,3* |
| *12* | *14764,04* | *15695,59* | *18433,3* | *21706,06* | *22396,9* | *56211,6* | *114792,4* |

For clarity, we present data in a chart:

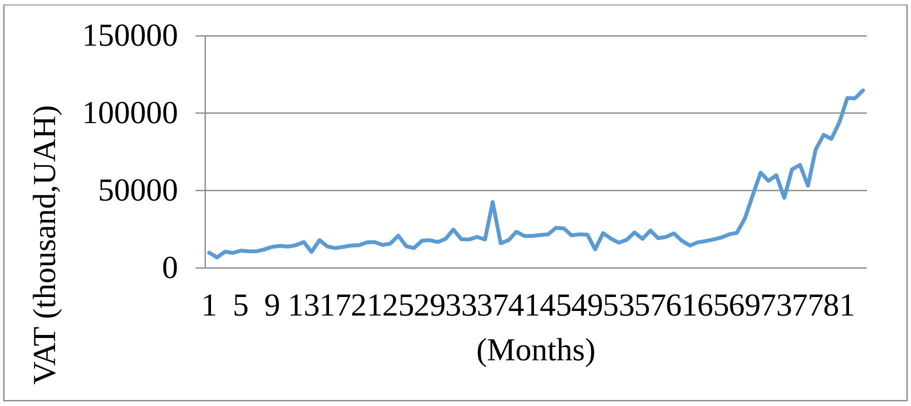



*Figure 3.9. The dynamics of monthly VAT revenues to STI Primorsky district of Odessa (thousand)*

Note that there is an anomalous value of 42,686.98 thousand. UAH. That is VAT receipts in January 2008 (outlier or a disambiguation), which is not statistically significant (as the government has forced the businesses to pay VAT in advance). Therefore, in future modeling, we will replace it for monthly average value is 2008 - 23,226.8 (thousand. UAH).

After visual analysis of a dynamics series, we conclude that it has certain characteristics:

a)trend is traced the long-term and evolution is deterministic in time, which is described by a function of time $f(t)$;

b) seasonal fluctuations $S_t$;

c) random fluctuations $e_t$.

The relationship between these components of time series presented in this case as the additive model (a summing):

$$y_t = f(t) + S_t + e_t,$$

where $y_t$ is a level of time series at time $t = 1, 2, ..., 84$.

This design allows conditional, depending on the purpose of research, we can study the trend with eliminating of oscillations or study the fluctuations with eliminating of the trend. We carry out the forecasting of different elements into to final forecast

Note that in this case, there is so-called "broken" linear trend (in October of 2010 revenues rose sharply). This is due to the growth of the tax base (from the great companies as taxpayers (VAT)) and also there is a significant improvement in the administration of taxes. So to find these trend equations we should use the module Piecewise Regression Models of nonlinear estimation in STATISTICA:

*Table 3.12. Coefficients of model by the results of the STATISTICA*

| Piecewise Regression Models<br>*N=84, R=0,98019* | | | | |
|---|---|---|---|---|
| *B₀* | *t* | *B₀* | *t* | *break point* |
| *12160,91* | *151,64* | *-300818* | *4879,71* | *27885,35* |

As a result, with a probability of 98%, we found that the equation of a broken line of trend can look as follows:

$f(t) = 12160,91 + 151,64t$ , $t = 1, 2, ..., 38$;

and

$f(t) = -300818 + 4879,71t$ , $t = 39, 40, ..., 84$.

After the analysis of the dynamics, we use additive model *ARIMA*ARIMAS* with intervention into the trend on the 39th observation and seasonal lag 12: $y_t = p_{t-1}y_{t-1} + a + bt - q_{t-1}e_{t-1} + S_t$ ,

where

$y_t$ is the levels of time series at time $t = 1, 2, ..., 84$;

$f(t) = a + bt$ is a trend component (with intervention);

$p_{t-1}$ is an autoregression coefficient of the first order;

$q_{t-1}$ is a coefficient the moving average model;

$e_{t-1}$ is an irregular component (random deviations or so-called *White noise*.

We apply the Smoothed Moving Average (SMMA) of first order $S_t = c + Q_{t-1}S_{t-1}$ to find seasonal coefficients. Parameters of the model one should estimate at 95% confidence level (or 5% risk) with the condition of minimizing the MSE (mean square error) $MSE = \sum_{i=1}^{n} \frac{(y_i - \tilde{y}_i)^2}{n}$ , where $y_i$ is the actual value, and $\tilde{y}_i$ estimates of indicator $y$ at $i$ -th period.

*Table 3.13. The results of modeling and forecasting*

| *Observations №* | *Forecast. Model: (1,1,1)(0,0,1);* Seasonal.*: 12; 1* Intervention*; Start SS=3212E6; End SS=2753E6 (85,69%); MS=3484E4; p<0,05* | | | |
|---|---|---|---|---|
| | *Forecasting* | *Lower 90%* | *Upper 90%* | *Standard deviation* |
| *85* | *115714,9* | *105890,5* | *125539,2* | *5902,7* |
| *86* | *107810,0* | *95881,1* | *119739,0* | *7167,2* |
| *87* | *118452,6* | *104573,9* | *132331,3* | *8338,7* |
| *88* | *119024,4* | *103453,9* | *134594,9* | *9355,2* |
| *89* | *111308,5* | *94211,2* | *128405,8* | *10272,5* |
| *90* | *123875,8* | *105377,5* | *142374,2* | *11114,3* |
| *91* | *129134,2* | *109333,7* | *148934,7* | *11896,6* |

| | | | | |
|---|---|---|---|---|
| *92* | *127062,8* | *106040,6* | *148084,9* | *12630,6* |
| *93* | *132886,4* | *110709,8* | *155063,0* | *13324,3* |
| *94* | *136332,8* | *113058,9* | *159606,7* | *13983,5* |
| *95* | *132433,6* | *108111,9* | *156755,3* | *14613,1* |
| *96* | *137125,3* | *111799,1* | *162451,5* | *15216,6* |

High adequacy of the model is confirmed comparing average predicted by this model values the VAT in January 2012 115,714.9 thousand corresponds to the actual, which is a UAH 116,056.1 thousand. There is a small relative error of prediction $\delta = \frac{|116056,1-115714,9|}{116056,1} = 0,0029$ or about 0.3%.

Note that, for the medium term of forecasting (for 2013-2015) we should use the trend model is based on actual annual revenues over the past years.

II. Perform modeling and forecasting dynamics of monthly revenues (billion) VAT at the macro level (according to State Statistics Committee of Ukraine for the period from 2007 to 2011):

*Table 3.14. The dynamics of monthly VAT revenues for the years 2007-2011 (billion)*

| *Years* | *2007* | *2008* | *2009* | *2010* | *2011* |
|---|---|---|---|---|---|
| *Months* | VAT | VAT | VAT | VAT | VAT |
| *1* | *4,7218* | *6,1866* | *6,3259* | *8,2303* | *9,5469* |
| *2* | *3,9682* | *7,1212* | *6,1968* | *4,7877* | *9,4425* |
| *3* | *5,0313* | *6,6084* | *9,6554* | *8,2598* | *10,3564* |
| *4* | *4,7422* | *9,3798* | *7,4880* | *10,054* | *11,2816* |
| *5* | *4,6714* | *8,7213* | *5,4998* | *7,1548* | *9,6094* |
| *6* | *5,2116* | *7,4539* | *4,8555* | *8,0634* | *9,2842* |
| *7* | *4,9334* | *8,3256* | *6,4424* | *9,5333* | *11,2136* |
| *8* | *4,8521* | *9,3311* | *6,0222* | *8,1873* | *12,9507* |
| *9* | *4,5546* | *9,3647* | *3,0412* | *7,8726* | *11,5065* |
| *10* | *5,672* | *8,7470* | *8,0639* | *9,2611* | *10,8824* |
| *11* | *5,4232* | *4,6886* | *8,1171* | *10,7776* | *11,0774* |
| *12* | *5,601* | *6,1544* | *12,9882* | *10,534* | *12,9422* |
| *Year* | *59,3828* | *92,0826* | *84,6964* | *102,72* | *130,09* |

For value added tax (VAT), the analysis of dynamics monthly income and dynamics for years allow to detect the presence of a pronounced linear trend with a seasonal lag of 12:

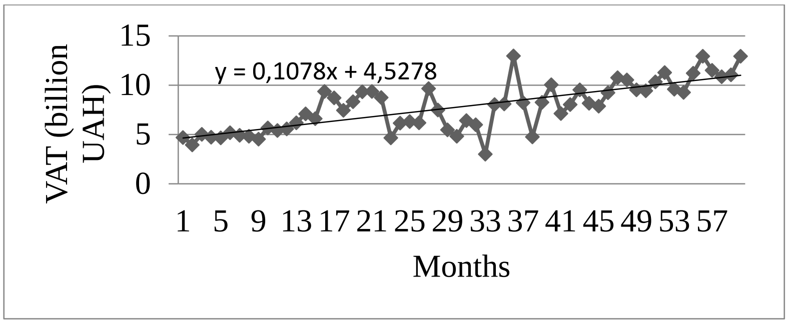



*Figure 3.10. The dynamics of monthly VAT revenues for the years 2007-2011 (billion)*

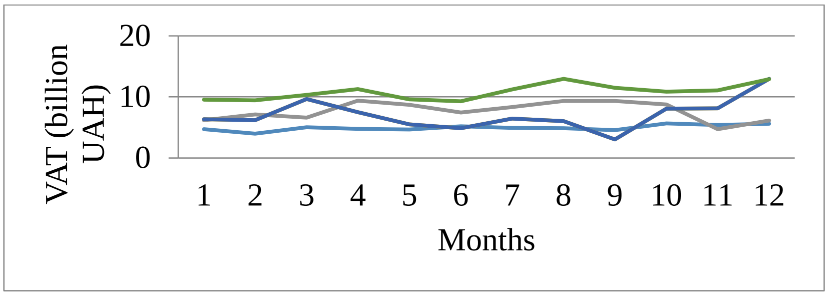



*Figure 3.11. The dynamics of monthly VAT revenues for the months (billion)*

Therefore, we should use the modeling by the additive Holt model (exponential smoothing model) with a linear trend and seasonality 12 $y_t = f(t) + S_t + e_t$, where $y_t$ are the levels of dynamics series at time $t = 1, 2, ..., 70$;

$f(t)$ is a trend-cycle component;

$e_t$ is the irregular component (random deviations or so-called *White noise*).

To find the trend-cycle component one can use exponential smoothing: $f_t = \alpha f_t + (1-\alpha) f_{t-1}$, where the smoothing parameter is $\alpha = 0{,}1$. It is chosen in terms of the best approximation to empirical data (with a minimizing the sum of squared errors SSE), and according to cyclical "damping" of the trend with parameter $\gamma = 0{,}1$.

When this formula is applied recursively, each new smoothed value (which is also forecast) is calculated as a weighted average of the current observation and smoothed series.

To find seasonal factors we use exponential smoothing $S_t = S_{t-p} + \delta(1-\alpha)e_t$ with parameter $\delta = 0{,}1$ .

Note that, the assessment of seasonal ingredients at the time *t* is calculated as the corresponding component in the last seasonal cycle plus an error $e_t$ (it is observation minus predicted value at the same time). The standard error of the additive trend–cyclical seasonal model $y_t = f(t) + S_t + e_t$ is $\sigma = 0,075$. Results of modeling methods are presented in the following table:

*Table 3.15. The results of modeling and forecasting*

| *Observations №* | *Exponential smoothing : Additive. Seasonal.(12)* *$S_0$=4,497; $T_0$=0,0752; Alfa=0,100; Delta=0,100; Gama=0,100* | | | |
|---|---|---|---|---|
| | *VAT* | *Smoothing* | Errors | Seasonal component |
| *1* | *4,72180* | *4,39545* | *0,32635* | *-0,176962* |
| *2* | *3,96820* | *3,70289* | *0,26531* | *-0,980652* |
| *3* | *5,03130* | *5,35457* | *-0,32327* | *0,563352* |
| *4* | *4,74220* | *5,89115* | *-1,14895* | *1,054347* |
| *5* | *4,67140* | *4,29806* | *0,37334* | *-0,490273* |
| *6* | *5,21160* | *4,31274* | *0,89886* | *-0,583093* |
| *7* | *4,93340* | *5,23880* | *-0,30540* | *0,17333* |
| *8* | *4,85210* | *5,01927* | *-0,16717* | *-0,091145* |
| *9* | *4,55460* | *4,21432* | *0,34028* | *-0,953803* |
| *10* | *5,67200* | *5,84308* | *-0,17108* | *0,563105* |
| *11* | *5,42320* | *5,19584* | *0,22736* | *-0,143145* |
| *12* | *5,60100* | *6,50444* | *-0,90344* | *1,064335* |
| *66* | *9,28420* | *9,46645* | *-0,18225* | |
| *67* | *11,21360* | *10,31627* | *0,89733* | |
| *68* | *12,95070* | *10,26001* | *2,69069* | |
| *69* | *11,50650* | *9,82415* | *1,68235* | |
| *70* | *10,88240* | *11,73559* | *-0,85319* | |
| *71* | | *11,16136* | | |
| *72* | | *12,56185* | | |
| *73* | | *11,45848* | | |
| *74* | | *10,81256* | | |
| *75* | | *12,54480* | | |
| *76* | | *13,18181* | | |
| *77* | | *11,85066* | | |
| *78* | | *11,91138* | | |
| *79* | | *12,92076* | | |
| *80* | | *12,97137* | | |

The point forecasting of VAT revenues in November 2011 ($t = 71$), in December 2011 ($t = 72$) and the next in 2012 has performed through the collecting of forecasts by each component to the one. For prediction interval with 95% reliability for our model $y_{71} = 11{,}1613$ (billion), and $y_{72} = 12{,}5618$ (billion) we find the error $\Delta = 1{,}96\sigma = 0{,}145$.

Thus, with 95% reliability could be argued that VAT revenue in November 2011 can range from 11.018 billion UAH (pessimistic forecast) to the 11.3063 billion UAH (optimistic forecast) and in December 2011 from 12,4068 (pessimistic) to 12.7068 billion (optimistic forecast). Note that the actual VAT receipts in November 2011 have amounted the 11.0774 billion UAH, that is approaching to the average expected by our forecasting model.

The relative error of forecasting VAT revenues for November 2011 is $\delta = \frac{|11{,}0774 - 11{,}1613|}{11{,}0774} = 0{,}0075$ (or 0.75%) and for December 2011 $\delta = \frac{|12{,}9422 - 12{,}7068|}{12{,}9422} = 0{,}018$ (or 1.8%).

Economic-mathematical analysis of time series of tax revenues allowed to build adequate mathematical models and make forecasting the volume of revenues. The results confirmed a prediction of actual data.

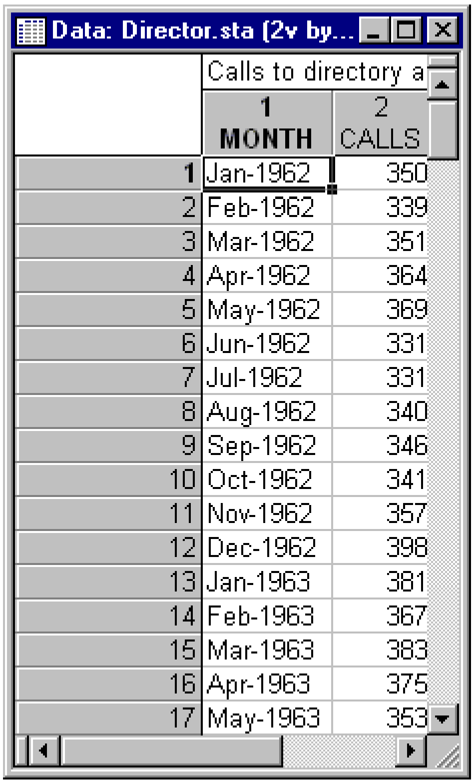
As we see, the modeling of the VAT under Odesa State Tax Inspectorate in Prymorskyi District is showed that a significant sharp increase in revenues arise after the tax base increase and it is improving tax administration.

Note that, the economic crisis Ukraine significantly affects the amounts of revenues to the state budget (including the VAT).

Thus, we can conclude on the need for continuous monitoring of revenues to the revenue side of the state budget and operational analysis, forecasting, risk-based. The research results can be used in practice of the financial, services, tax, and in making scientifically grounded management decisions.

*Example 3.5. Modeling of decrease in the number of requests for assistance ( on months) in Cincinnati by Interrupted ARIMA( from Statistica Help)*

This example is based on data presented in McDowall, McCleary, Meidinger, and Hay (1980), who provide an excellent applications-oriented introduction to interrupted time series analysis. The data in the example file Director.sta will be used. This file contains the number of monthly calls (in 100's) to Cincinnati Directory Assistance over the period from January, 1962, through December, 1976. In March 1974 (the 147'th month in the series), Cincinnati Bell initiated a new 20 cent charge for calls to Directory Assistance. This caused a marked drop-off in the number of requests for assistance, and the purpose of this analysis is to fit a model to the series that takes this abrupt change into account.

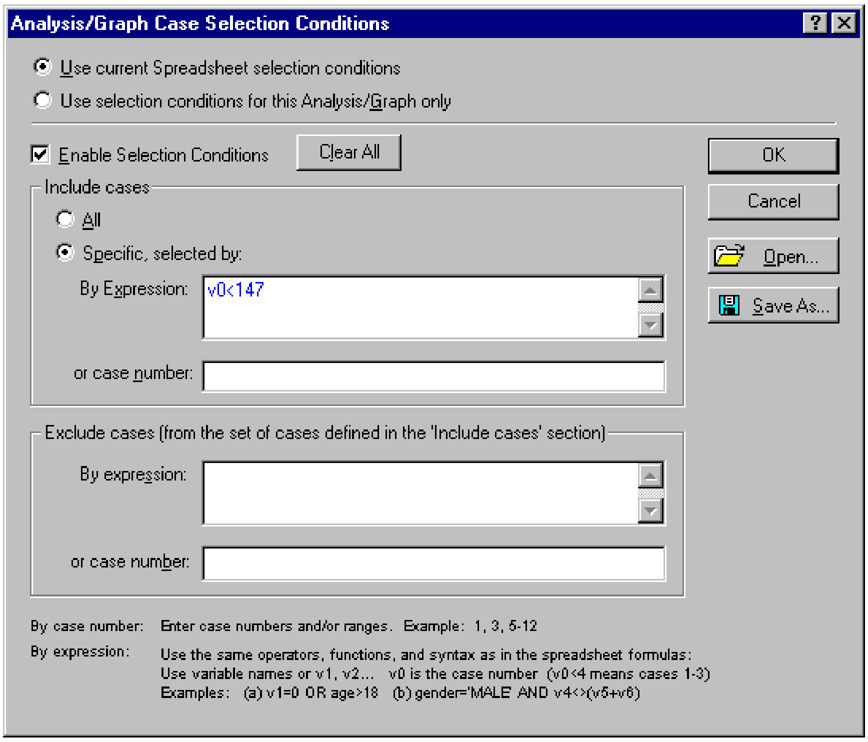


Data: Director.sta (2v by...

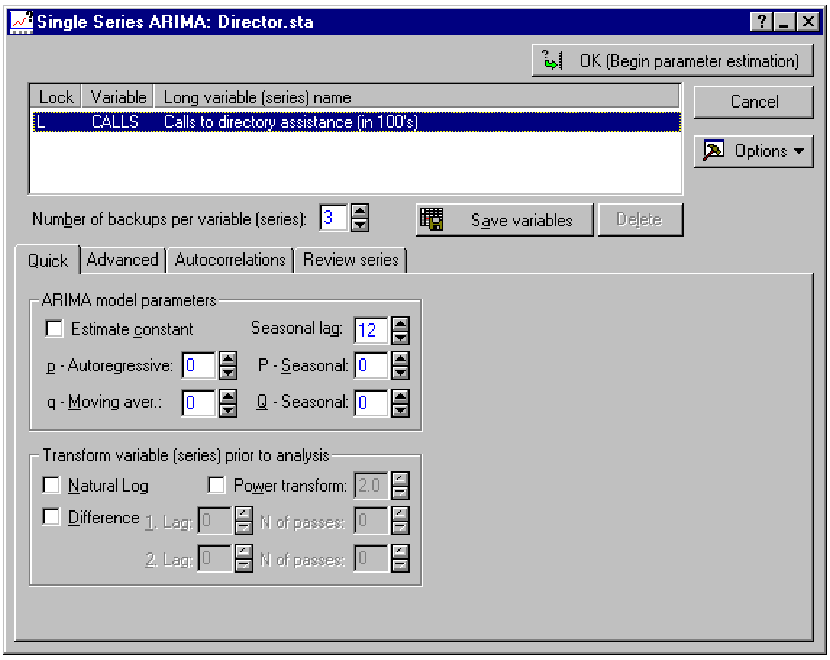
| | Calls to directory a | |
|---|---|---|
| | 1 MONTH | 2 CALLS |
| 1 | Jan-1962 | 350 |
| 2 | Feb-1962 | 339 |
| 3 | Mar-1962 | 351 |
| 4 | Apr-1962 | 364 |
| 5 | May-1962 | 369 |
| 6 | Jun-1962 | 331 |
| 7 | Jul-1962 | 331 |
| 8 | Aug-1962 | 340 |
| 9 | Sep-1962 | 346 |
| 10 | Oct-1962 | 341 |
| 11 | Nov-1962 | 357 |
| 12 | Dec-1962 | 398 |
| 13 | Jan-1963 | 381 |
| 14 | Feb-1963 | 367 |
| 15 | Mar-1963 | 383 |
| 16 | Apr-1963 | 375 |
| 17 | May-1963 | 353 |

*Model Identification*. The first step is to identify an appropriate model for the series. In this case, a regular ARIMA model will first be fit to the observations prior to the introduction of the service charge; then, an intervention component will be added to analyze the complete series.

To start the analysis, select T*ime Series/Forecasting* from the *Statistics - Advanced Linear/Nonlinear Models* menu to display the *Time Series Analysis Startup Panel*. There are 146 months of observations in the series prior to the introduction of the service charge. To open only those observations into the active work area, click the *Select Cases* button (below the Options button on the *Startup Panel*) and specify the condition Include cases v0<147.

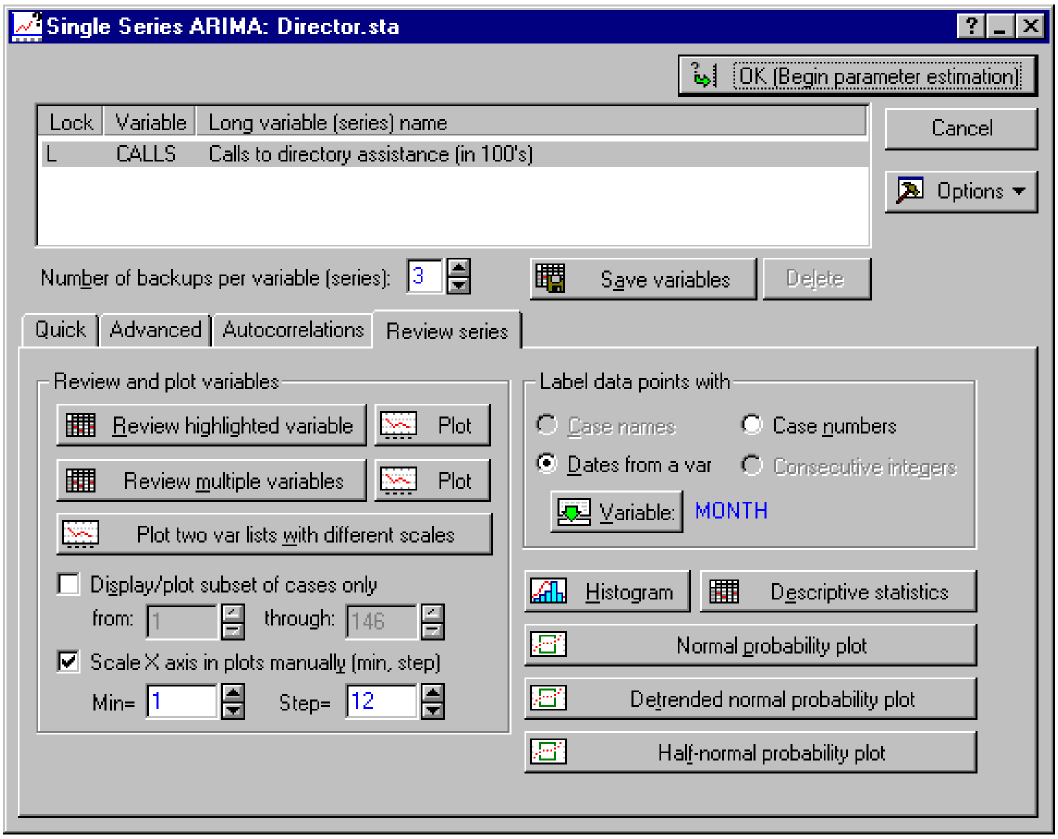



V0 in the selection condition shown above refers to the case numbers, so only cases numbers less than 147 will be included in the analysis. Now click the OK button, then click the *Variables* button to display the standard variable selection dialog. Here, select the variable *Calls* and click the *OK* button. Then click the *ARIMA & autocorrelation* functions button to display the *Single Series ARIMA* specification dialog.

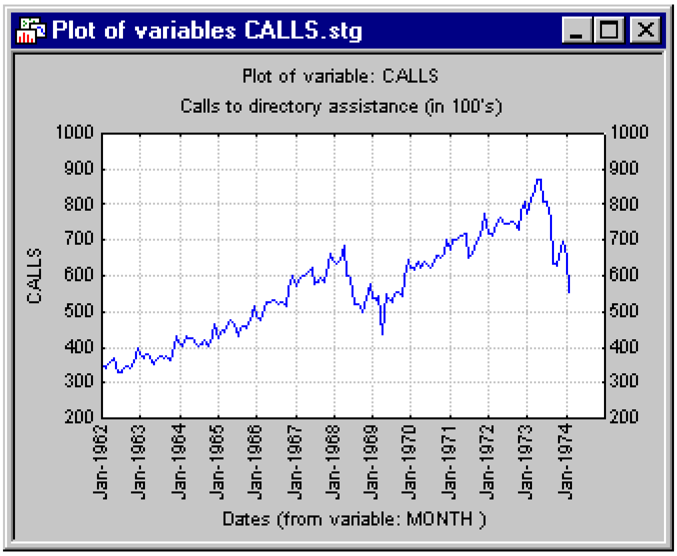



First plot the series; file Director.sta contains a variable with dates that can be used to label the horizontal x-axis of line plots. Click on the *Review series* tab, then select the *Dates* from a var option button to display a standard variable selection dialog. Here, select variable Month as the date variable and click the OK button.

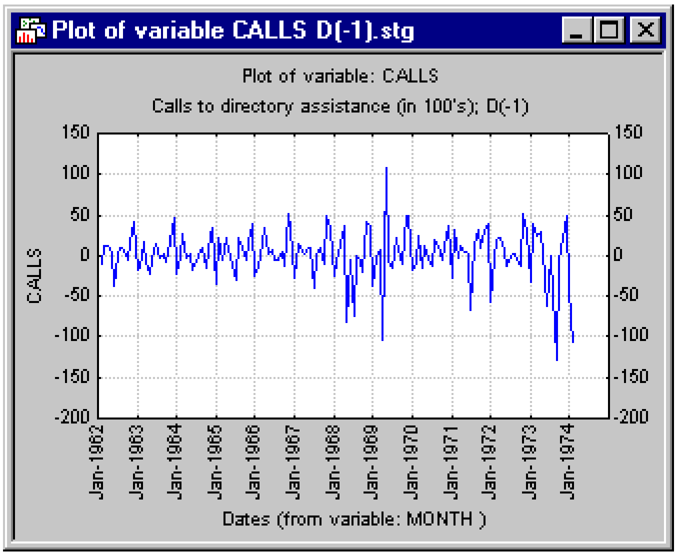
Since each year consists of 12 observations, also select the *Scale X* axis in plots manually check box and enter 12 in the Step box.



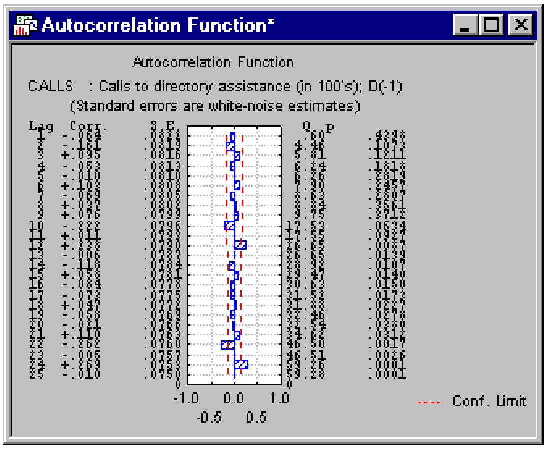

Now click the *Plot* button next to the Review highlighted variable button.



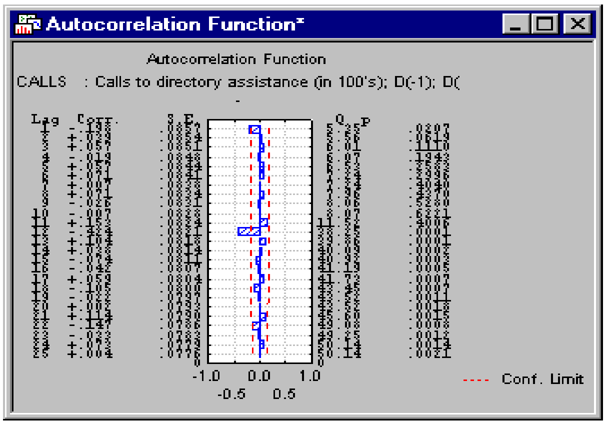

The series shows a linear upward trend, so some differencing will probably be necessary. Click on the *Advanced* tab, and then click the *Other transformations & plots* button to display the Transformations of Variables dialog. To perform simple (non-seasonal) differencing, click on the *Difference*, integrate tab and select the *Differencing* (*x=x-x(lag)*) option button. Be sure that the lag parameter is set to 1, and then click the OK (*Transform selected series*) button. After all cases have been transformed, the differenced series will be plotted (unless you previously cleared the Plot variable (series) after each transformation check box on the *Review & plot* tab).

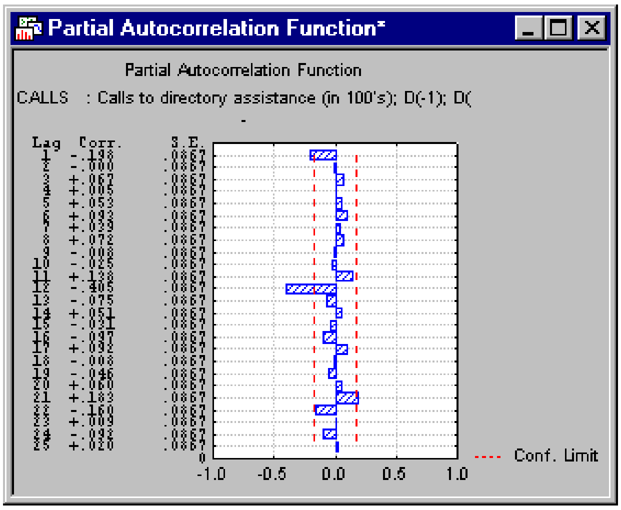



*Autocorrelation*. As described in the previous example and the Introductory Overview, the autocorrelation and partial autocorrelation functions are the most important tools for identifying an appropriate ARIMA model. Because the series consists of monthly observations, you may suspect that any seasonality will occur with a seasonal lag of 12. If seasonal differencing is necessary you would expect substantial autocorrelations at multiples of the seasonal lag (by contrast, a stationary series would be characterized by an autocorrelation function that dies out, the longer the lag); therefore, on the Autocorrs tab, set the Number of lags parameter under Autocorrelations & crosscorrelations to 25 so that you may detect such a pattern. Then click the Autocorrelations button.




Indeed, the autocorrelation function shows no sign of decaying but suggests that further (seasonal) differencing is necessary. Bring up the Difference, integrate tab again and perform *Differencing (x=x-x(lag))* with lag=12. Shown below are the autocorrelation and partial autocorrelation functions for the resulting series. (Click the *Autocorrelations and Partial autocorrelations* buttons on the *Autocorrs* tab.)

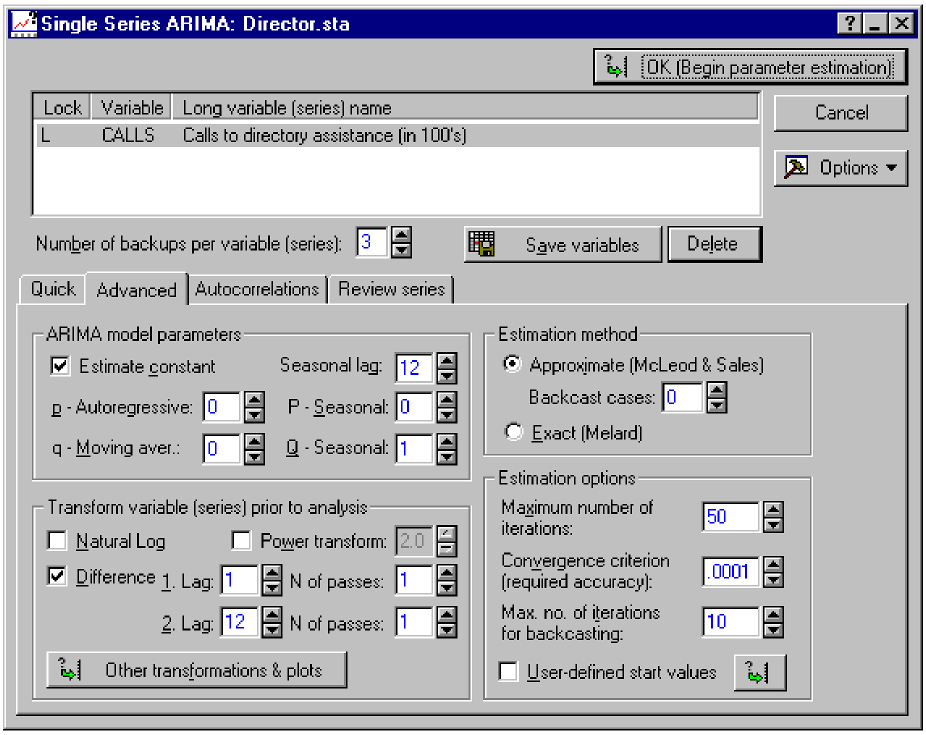


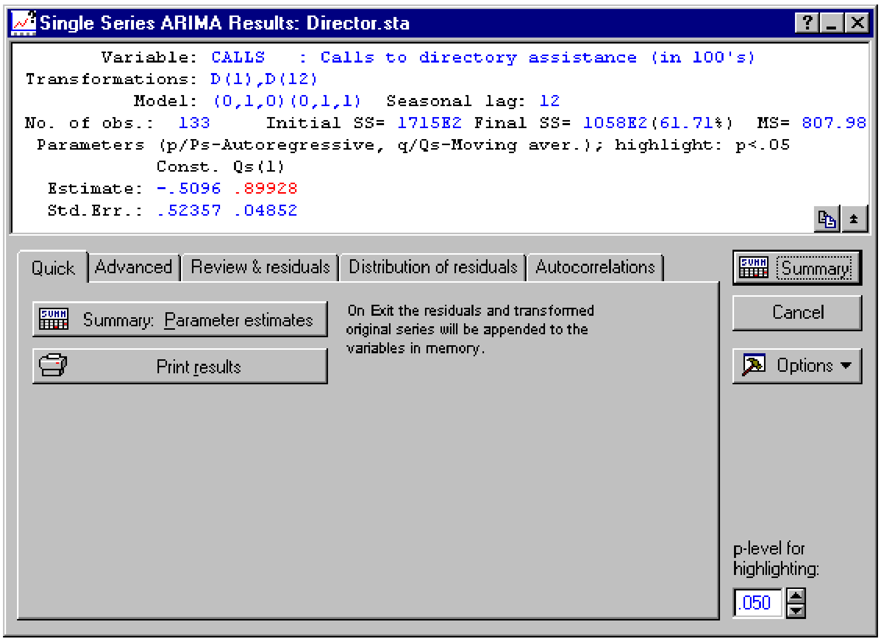

Both plots show a small spike at lag 1 and another stronger spike at 12. The two spikes show no sign of gradual decay in the autocorrelation function, and so McDowall et al. (1980, page 69) suggest a moving-average process rather than an autoregressive process. Specifically, they initially fit a model to the first 146 observations of the series that includes non-seasonal and seasonal differencing, one seasonal autoregressive parameter, and a constant.

In standard terminology, the model (0, 1, 0) (0, 1, 1) will be fit with seasonal lag 12 and a constant.

Specifying the ARIMA Model. Now return to the Single Series ARIMA dialog (click the Cancel button on the *Transformations* of *Variables* dialog). Delete the differenced series from the active work area (by selecting each differenced series and clicking the *Delete* button), and select (highlight) the original untransformed series *Calls*. Next, specify *1Seasonal moving average parameter* (Q) and a Seasonal lag of 12, and select the Estimate constant check box on the Advanced tab. Then, select the Difference check box under Transform variable (series) prior to analysis and specify differencing with 1. Lag: 1 (1 N. of passes) and 2. Lag: 12 (1 N. of passes). Make sure that the Approximate (McLeod & Sales) option button is selected under Estimation method.




You are now ready to proceed; click the OK (Begin parameter estimation) button. After the parameter estimation finishes, the *Single Series ARIMA Results* dialog is displayed.



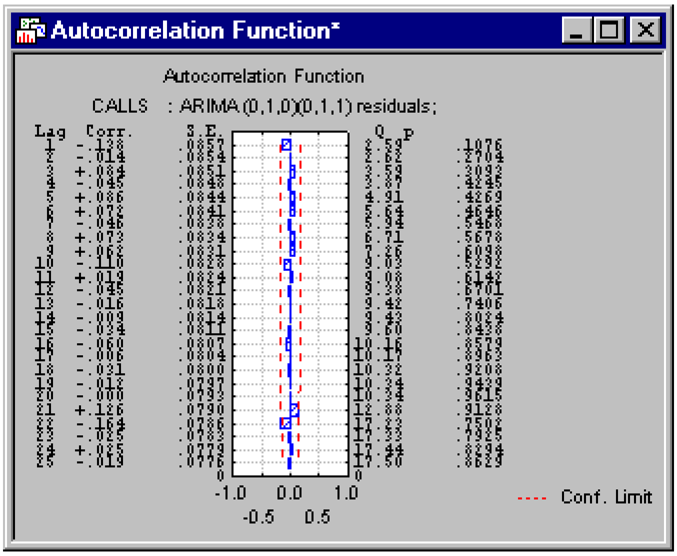

Preliminary Results. Click the *Summary*: Parameter estimates button to display a spreadsheet with the parameter estimates.

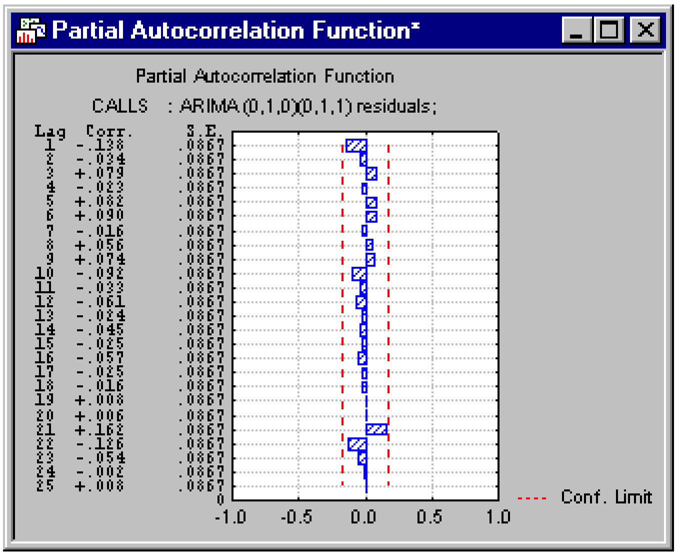
Data: Input: CALLS : Calls to directory assistance (in 100's) (Direc...

Input: CALLS : Calls to directory assistance (in 100's) (Director.st
Transformations: D(1),D(12)
Model:(0,1,0)(0,1,1) Seasonal lag: 12 MS Residual=807.98

| Paramet. | Param. | Asympt. Std.Err. | Asympt. t( 131) | p | Lower 95% Conf | Upper 95% Conf |
|---|---|---|---|---|---|---|
| Constant | -0.509620 | 0.523565 | -0.97337 | 0.332165 | -1.54536 | 0.526116 |
| Qs(1) | 0.899279 | 0.048515 | 18.53609 | 0.000000 | 0.80330 | 0.995253 |

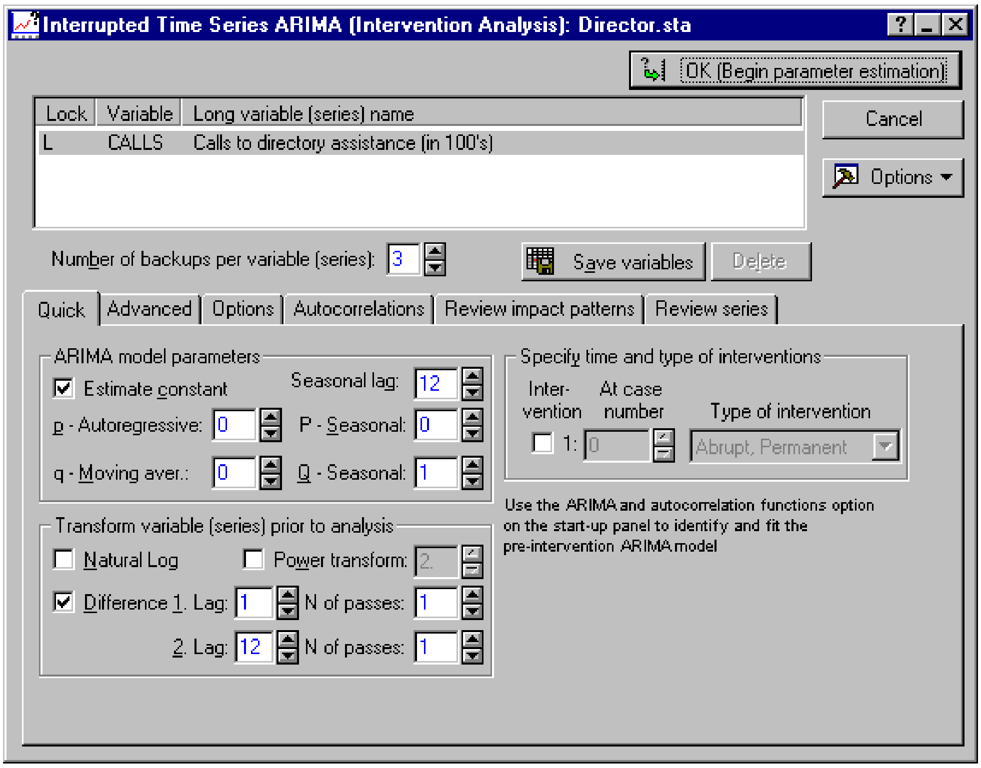
The constant is not statistically significant; therefore, McDowall et al. (1980) suggest to drop it from further analyses. Now, quickly look at the *Autocorrelations* and *Partial autocorrelations* of residuals on the *Autocorrelations* tab.



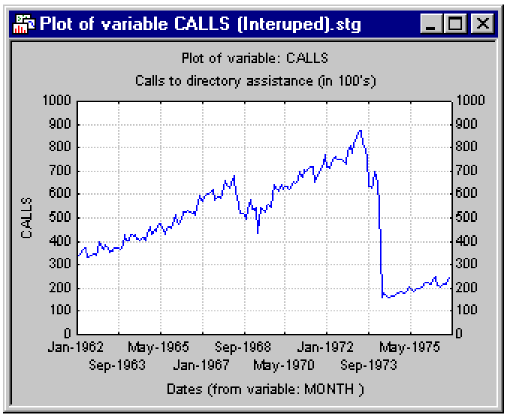




There is no evidence of any significant residual serial dependency in the data. Therefore, you can be satisfied that you have identified a reasonable model for the series.

Specifying Interrupted ARIMA. Now, exit the Single Series ARIMA Results dialog by clicking the Cancel button to specify an ARIMA model for the entire series, accounting for the introduction of the 20 cents user-fee for calls to Directory Assistance. Now, click the Cancel button on Single Series ARIMA dialog to return to the Time Series Analysis Startup Panel and click the Select Cases button. Note that, after exiting from the Single Series ARIMA Results dialog, the differenced series as well as the ARIMA residuals were automatically appended to the active work area. When you change the selection conditions, the active work area will be cleared, and a respective message box will appear to warn you that this is about to happen; when you subsequently select another type of time series analysis, the selected original variables will be read again, using the new selection conditions. Simply click Yes in response to the warning, and then turn off the case selection conditions (click the Clear All button and then the OK button). Next click the *Interrupted time series analysis* button to display the Interrupted Time Series ARIMA dialog.

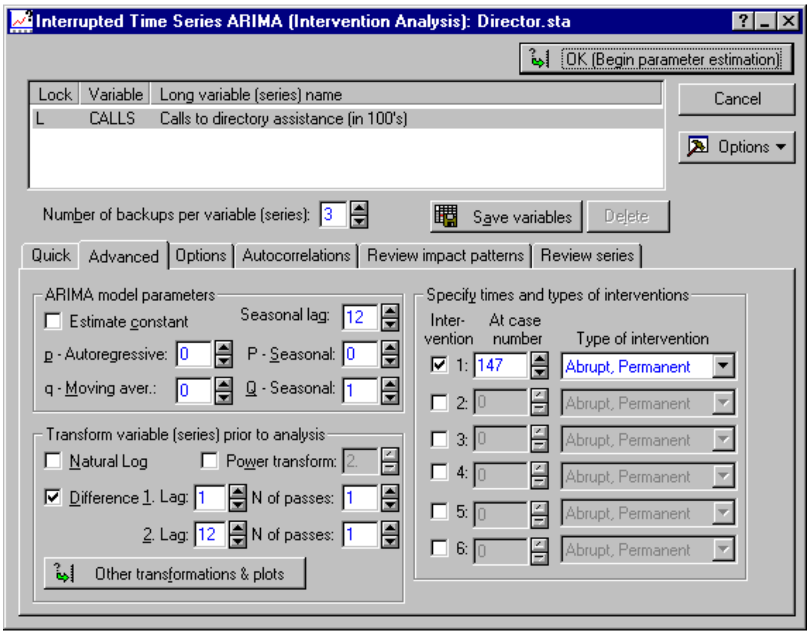



As you can see, all prior specifications are still in place. However, before proceeding, look at the entire series, that is, including the observations following the introduction of the service charge. Click the *Other transformations & plots* button on the *Advanced* tab and then plot the series (use the Plot button beside the *Review highlighted variable* button on the *Review & plot* tab as before).

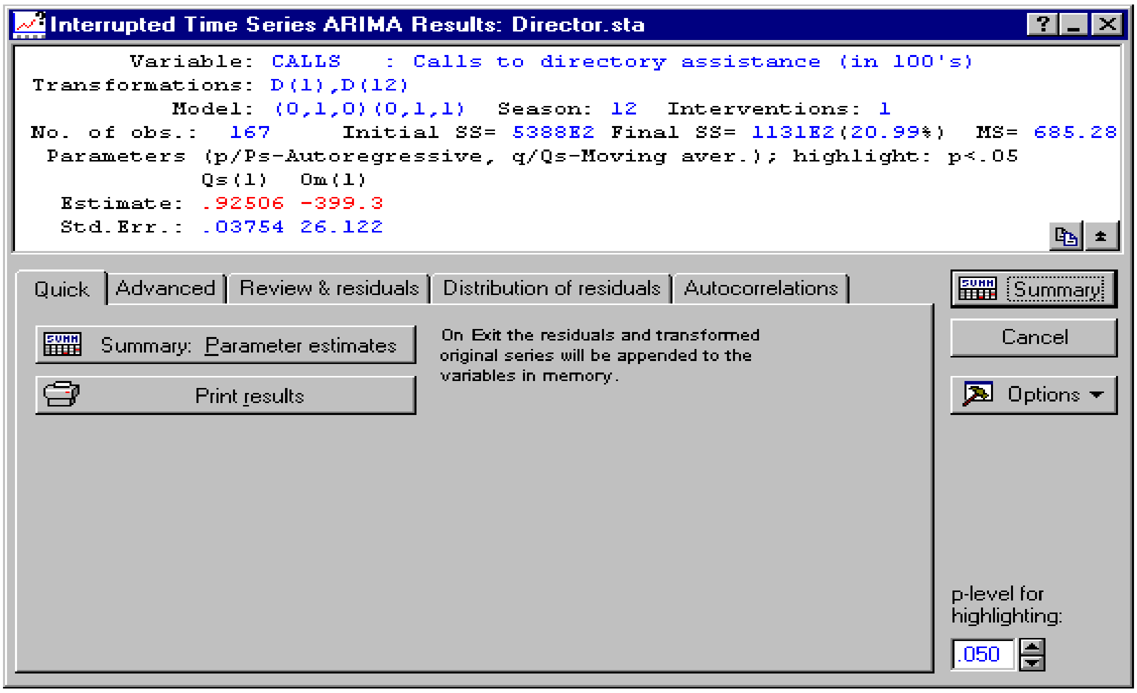



Evidently, the service charge caused a sharp abrupt decline in the number of calls to Directory Assistance. Now click the Cancel button to exit the Transformations of Variables dialog and go back to the Interrupted Time Series ARIMA dialog.

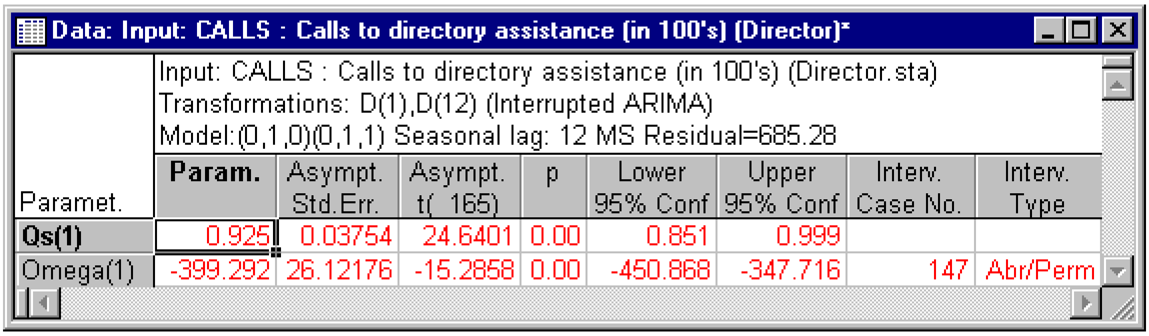
Specifying the intervention. The Time Series module allows you to model three types of interventions: Abrupt-Permanent, Gradual- Permanent, and Abrupt-Temporary. The Introductory Overview discusses these types of interventions in detail. In this case, looking back at the plot of the entire series, it certainly looks like the introduction of the service charge caused an abrupt and permanent change in the number of calls to Directory Assistance. However, in other cases the result of an intervention may not be as apparent; McDowall et al., (1980, pages 83-85) discuss how one might proceed when the researcher has no a priori hypotheses concerning the nature of the impact. Now, specify the intervention. Select the 1 check box in the Specify times and types of interventions group on the Advanced tab. The service charge was introduced at the 147'th month in the series; thus specify 147 in the At case number field. Finally, remember that you (and McDowall et al.) concluded that a constant is not necessary for these data (i.e., after differencing); therefore clear the Estimate constant check box.



You are now ready to proceed with the parameter estimation. Click the OK (Begin parameter estimation) button, and when the parameter estimation is done, the Interrupted Time Series ARIMA Results dialog is displayed.




Reviewing the *Final Results*. Click the *Summary*: Parameters estimates button to display a spreadsheet with the final parameter estimates.

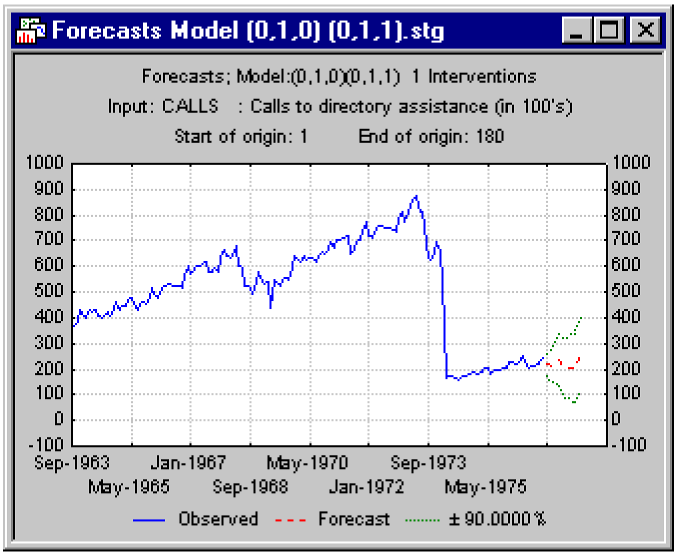

Data: Input: CALLS : Calls to directory assistance (in 100's) (Director)*

Input: CALLS : Calls to directory assistance (in 100's) (Director.sta)
Transformations: D(1),D(12) (Interrupted ARIMA)
Model:(0,1,0)(0,1,1) Seasonal lag: 12 MS Residual=685.28

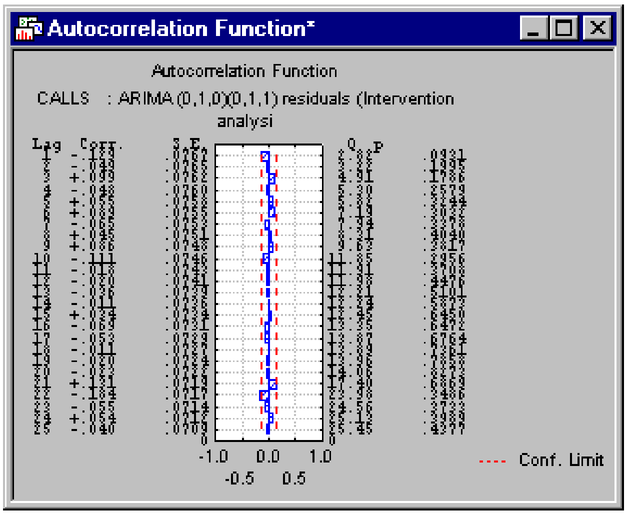
| Paramet. | Param. | Asympt. Std.Err. | Asympt. t( 165) | p | Lower 95% Conf | Upper 95% Conf | Interv. Case No. | Interv. Type |
|---|---|---|---|---|---|---|---|---|
| Qs(1) | 0.925 | 0.03754 | 24.6401 | 0.00 | 0.851 | 0.999 | | |
| Omega(1) | -399.292 | 26.12176 | -15.2858 | 0.00 | -450.868 | -347.716 | 147 | Abr/Perm |

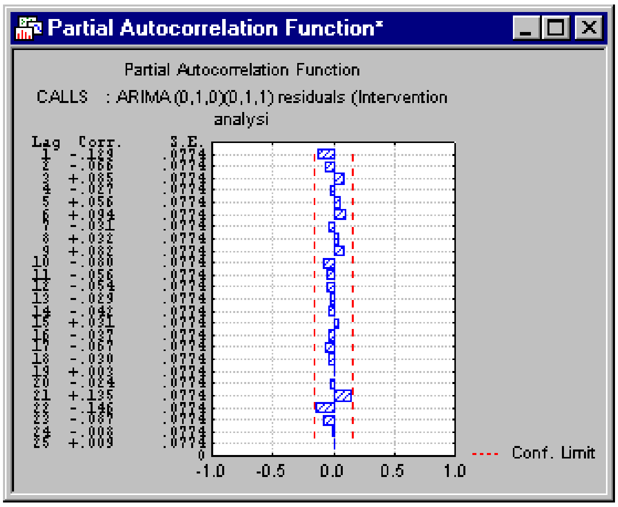
The seasonal moving average parameter is highly significant. Parameter Omega(1) is the intervention parameter, and it is also highly significant. For Abrupt, Permanent interventions, Omega can simply be interpreted as the amount of permanent change that occurred at the point of intervention. Thus, you may conclude that the introduction of the 20 cents service charge for Directory Assistance reduced the number of such calls by about 399 * 100 = 39,000 calls.

*Forecasts.* As you can see by reviewing the options available on the *Advanced* tab, you can compute forecasts, taking into account the interventions in the series. Click the *Plot series & forecasts* button to plot the series together with one seasonal cycle (one year) of forecasts.

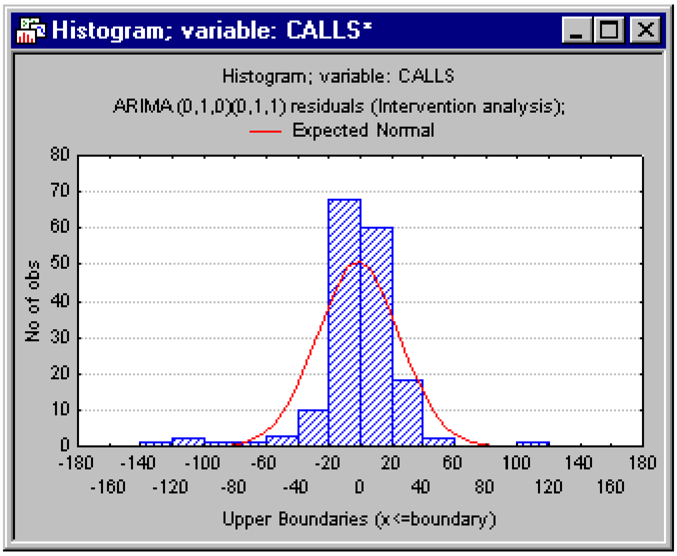



*Residual analysis*. Before concluding the analysis, perform some final model checks. Shown below are the autocorrelation and partial autocorrelation functions for the residuals. (Click the *Autocorrelations and Partial autocorrelations* buttons on the *Autocorrelations* tab.)

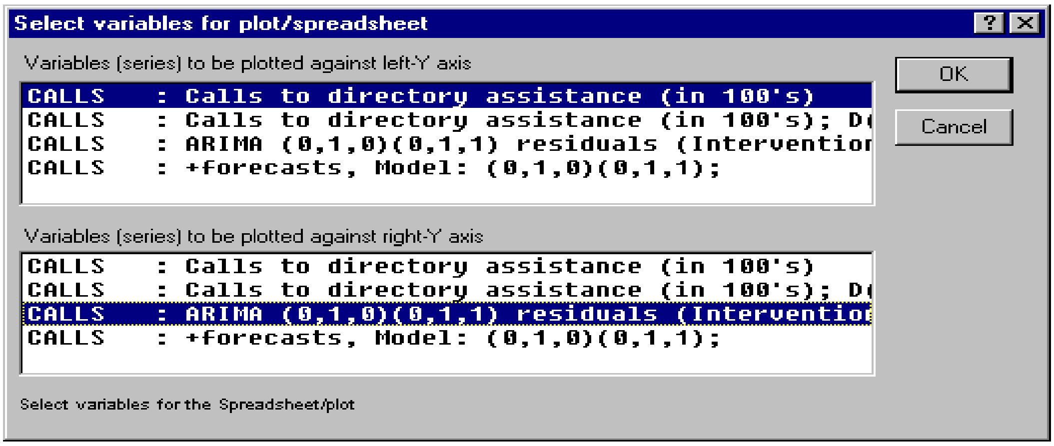


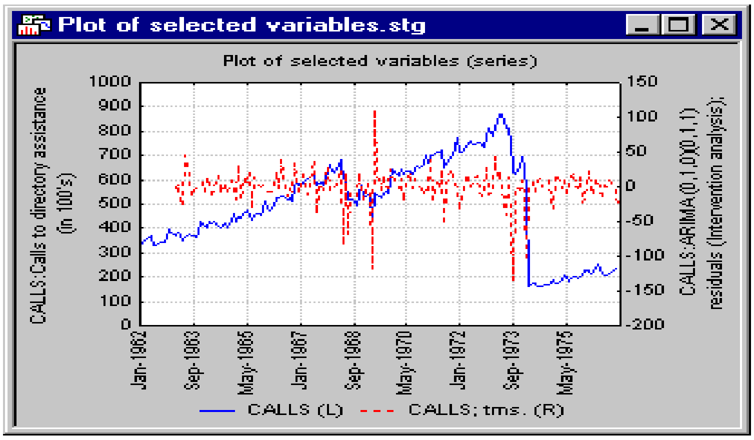
There is no evidence of any residual serial dependency in the data. Click the *Histogram* button on the *Distribution of residuals* tab to review the distribution of residuals.




Even though the distribution of the residuals seems leptokurtic (i.e., more "peaked" than the standard normal distribution), they approximate the normal distribution reasonably well. Now click the Cancel button to return to the *Interrupted Time Series ARIMA* dialog again; the ARIMA residuals as well as the differenced series and the ARIMA forecasts will be appended to the active work area. As in Example 1, look at a final summary graph that includes the original series as well as the residuals. Click the *Other transformations & plots* button on the Advanced tab, and then select the Plot two variables lists with different scales button on the Review & plot tab. In the Select variables for plot/spreadsheet dialog, select to plot the original series Calls and the ARIMA residuals.




Now, click the OK button to produce the plot.




It looks like the fit of the ARIMA model is as good or better following the intervention as it is prior to the intervention. In other words, there is no evidence that after the intervention the ARIMA residuals show some kind of non-random (white noise) pattern. If that were the case, one might suspect that the ARIMA model you identified from the series prior to the intervention (i.e., based on the first 146 observations only) does not fit the series after the intervention; however, in your case there is no evidence of this.

*Example 3.6. Modeling and Forecasting Ukrainian bond index (UB).*

This problem has arisen in connection with research of calculations the parameters of hybrid program-target (for the implementation of great structural projects) bonds. Financial engineering examines the hybrid financial tools for which the indicators are tied to various financial markets.

According to preliminary results of the study, we will define a floating coupon rate of the hybrid bonds by equation $q(t) = \max\{q_b; q_b + \Delta(t)\}$; where $q(t)$ is the coupon rate at the time $t$; $q_b$ is a reference rate and $\Delta(t)$ is a coupon bonus.

In our view, the parameters to adjust of yields before selling (buy back) of bonds (on issue and circulation on the secondary market) and purchase price $P_{buy}$ or sell price $P_{sell}$ of bonds it should reflect the general trends on the state of the securities market( in this case for the bonds). It is proposed to use the Ukrainian bond index (UB), located in the stock list.

The Ukrainian stock market has been operating for many stock exchanges.

Perspektiva Stock Exchange which occupied 80% in the structure of trading in 2013, had a leading role in trading government bonds. PFTS was ranked first (57% of trading) into the volume of trading in corporate bonds and Perspektiva Stock Exchange share in this segment of the stock market was 24.2% in 2013. Ukrainian bonds Performance Index (UB) can be used to adjust various parameters of hybrid-purpose bonds.

Of course, a more balanced approach for selecting an index is to use the average bond index, which takes into account changes in market prices of bonds on different trading platforms. Moreover, the weighting of indices for different exchanges can take place in terms of trading volume in the period, but unfortunately, these exchanges use different approaches to the bond index, making it impossible to use a weighted average to be used in the calculations.

We will study the stock market, namely the value of the index Ukrainian bonds (UB). As mentioned above, the Perspektiva Stock Exchange plays a leading in bonds trading . Therefore, we examine a daily dynamic from the 02.01.2013 to 30.05. 2014:

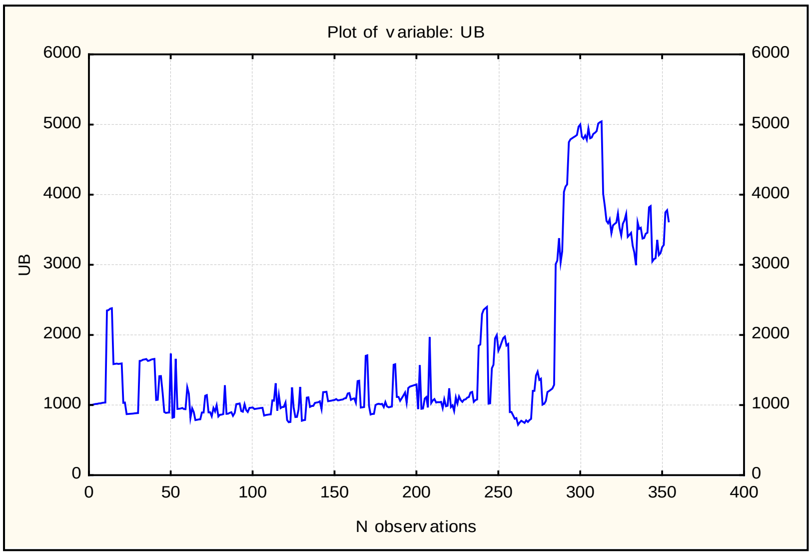



*Fig. 3.12. Daily dynamics of the Ukrainian bond index (UB) from 02.01.2013 till 30.05. 2014.*

After analyzing of the dynamics series we can draw a conclusion that it has the following characteristics:

a) fairly significant fluctuations is traced around some constants;

b) there is a leap the UB index in 18.02.2014 (almost a sharp increase on 1730 basis points);

c)there are the random fluctuations $e_t$.

According to techniques of Exchange "Perspective", the index $UB_t$ at time $t$ is calculated by the formula: $UB_t = UB_{t-1} \cdot \frac{\sum_{i=1}^{n} R_{i,t} V_{i,t}}{\sum_{i=1}^{n} R_{i,t-1} V_{i,t}} \cdot Z_t$ ,where $t$ – current period (index is calculated at this period); $t-1$ – previous period (at the close of the trading session of the previous trading day); $i$ – issuance of bonds on the List; $n$ – number issuances bonds in the List; $R_{i,t}$ – effective yield bonds *i*-th issue currently, %; $V_{i,t}$ – nominal capitalization (total nominal amount) *i*-th bond issue, hrn.; $Z_t$ – correction coefficient for the current period.

We perform modeling and prediction the dynamics of the index of UB by the so-called ARIMA (autoregressive integrated moving average) model.

As we know, a characteristic feature of any dynamic series is dependence levels: values $UB_t$ to some extent depends on previous values $UB_{t-1}, UB_{t-2}$ and etc.

To assess the degree of dependence the levels one should use a series of autocorrelation coefficients $r_p$ with time lag $p = 1, 2, ..., m$ that characterizes a density of connection between the primary number and the same number with shift on *p* points.

The sequence of coefficients $r_p$ is called autocorrelation function (ACF) and depicted graphically as autocorrelogram. We can conclude about the character of dynamics by the speed damping of ACF (Figure 3.13):

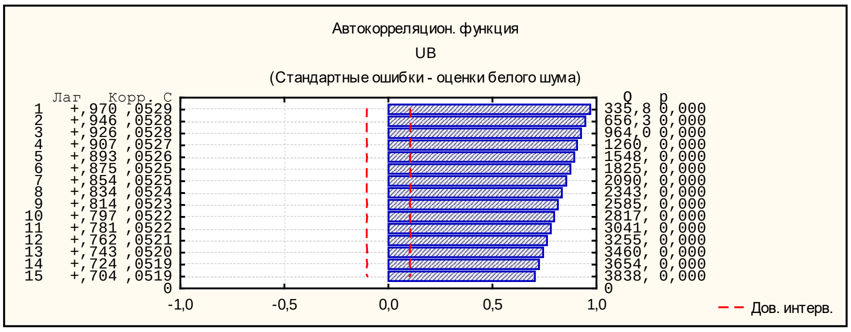



*Figure 3.13. Graph and values of ACF dynamics for* $UB_t$.

Graph of ACF indicates a presence of the first-order autocorrelation with lag 1. Partial autocorrelation function (PACF) (Figure 3.4) indicates a "pure" description (it is presence or absence of a higher order autocorrelation). Hence, We analyze the PACF or the correlation with shift one step ahead:

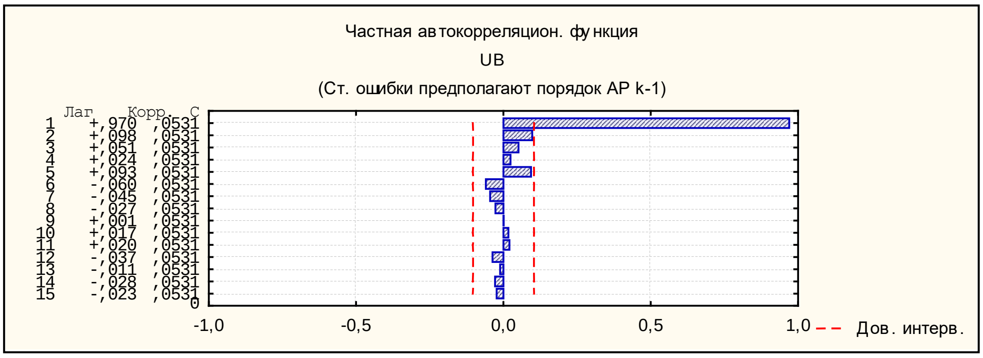



*Figure 3.14. Graph and values PACF of dynamics* $UB_t$.

Graph PACF confirms the presence only autocorrelation of the first order.

Difference operator of the first order one can apply to a time series $UB_t$ with lag 1 $d(1) = UB_t - UB_{t-1}$ to obtain the stationary form of process.

The ARIMA with intervention at the 285 level (18.02.2014 year) is used to modeling and predicting the dynamics

Note that to compare forecasting accuracy in modeling we choose only 335 levels (), then we take the next levels for the forecast.

So ARIMA model will be as follows: $UB_t = const + p(1)UB_{t-1} - q(1)e_t$, evaluation parameters which get using the module *Time series and forecasting* in program STATISTICA (see the table below).

*Table 3.16. Assessment the parameters of autoregression model a time series*

| Параметр | Исход.:UB1 (Таблица.sta)<br>Преобразования: D(1) (Прерванная АРПСС)<br>Модель(1,1,1) MS Остаток= 62513, | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Парам. | Асимпт. Ст.ошиб. | Асимпт. t( 330) | p | Нижняя 95% дов. | Верхняя 95% дов. | Интерв. Набл. N | Интерв. Тип |
| Конст. | 1,834 | 8,3897 | 0,218648 | 0,827059 | -14,670 | 18,338 | | |
| p(1) | 0,514 | 0,2374 | 2,165978 | 0,031029 | 0,047 | 0,981 | | |
| q(1) | 0,705 | 0,2030 | 3,472670 | 0,000584 | 0,306 | 1,104 | | |
| Омега(1) | 1838,109 | 261,2105 | 7,036887 | 0,000000 | 1324,261 | 2351,956 | 285 | ск/уст |

A random component $e_t$, unlike the deterministic component, is not connected with the change of time.

It is the basis for testing hypotheses about model adequacy to real process (see Figure below):

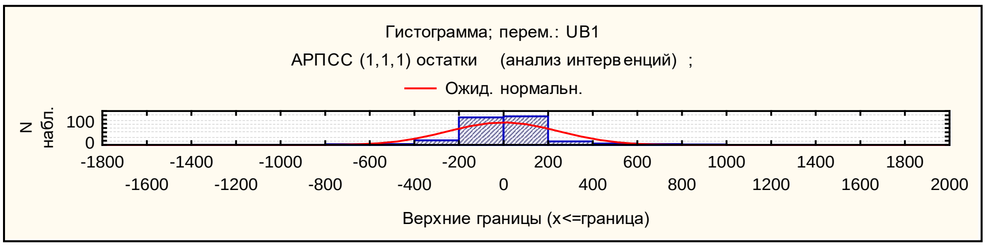



*Figure.3.15. Graph residues of the dynamics model and the comparing them with a normal distribution*

We conclude sufficiently adequacy of the model and perform the forecast on following periods:

*Table 3.17. Forecasts by ARIMA model*

| | Прогнозы; Модель:(1,1,1) 1 интервенции (Таблица.sta)<br>Исход.:UB1<br>Начало исходных: 1 Конец исходн.: 335 | | | |
|---|---|---|---|---|
| Набл. N | Прогноз | Нижний 95,0000% | Верхний 95,0000% | Ст.ошиб. |
| 336 | 3539,292 | 3047,445 | 4031,139 | 250,0266 |
| 337 | 3506,971 | 2874,260 | 4139,682 | 321,6338 |
| 338 | 3491,240 | 2768,297 | 4214,183 | 367,5024 |
| 339 | 3484,040 | 2691,435 | 4276,646 | 402,9150 |
| 340 | 3481,229 | 2629,367 | 4333,091 | 433,0375 |
| 341 | 3480,674 | 2575,655 | 4385,692 | 460,0593 |

We'll compare the results of prediction for May 2014 by the ARIMA model with actual values:

*Table 3.18. Forecasts by ARIMA model and actual values*

| *Forecasts* | *Actual data* |
|---|---|
| *3539,292* | *3511,1209* |
| *3506,971* | *3525,3314* |
| *3491,240* | *3473,5498* |
| *3484,040* | *3479,4659* |
| *3481,229* | *3440,7611* |
| *3480,674* | *3457,0099* |
| *3481,279* | *3487,6691* |
| *3482,482* | *3455,5293* |
| *3483,991* | *3456,2835* |
| *3485,658* | *3479,0848* |

As you can see, the maximum relative error of prediction:

$\delta_{\max} = \frac{\max\left|UB_t^{прогноз} \text{-} UB_t^{факт}\right|}{\min UB_t^{факт}} \le \frac{40}{3440} = 0,011$ that is, less than 1.1%, and this suggests the possibility of using ARIMA with intervention for modeling and predicting the dynamics of the index Ukrainian bonds $UB_t$ .

The basic version of the least squares model assumes that the expected value of all error terms, when squared, is the same at any given point. This assumption is called homoskedasticity, and it is this assumption that is the focus of ARCH/GARCH models. Data in which the variances of the error terms are not equal, in which the error terms may reasonably be expected to be larger for some points or ranges of the data than for others, are said to suffer from heteroskedasticity. The standard warning is that in the presence of heteroskedasticity, the regression coefficients for an ordinary least squares regression are still unbiased, but the standard errors and confidence intervals estimated by conventional procedures will be too narrow, giving a false sense of precision. Instead of considering this as a problem to be corrected, ARCH and GARCH models treat heteroskedasticity as a variance to be modeled. As a result, not only are the deficiencies of least squares corrected, but a prediction is computed for the variance of each error term. This prediction turns out often to be of interest, particularly in applications in finance.

Engle (1982, 1995), Bollerslev (1986), Bollerslev and Ghysels (1996) and others developed a class of models that address such concerns and allow for modeling both the level (the first moment) and the variance (the second moment) of a process. This class of models is referred to as conditional heteroscedastic models. In this chapter, the term volatility refers to a measure associated with either the conditional variance or the conditional standard deviation of a process.

Volatility is an important concept not just in theory, but also in practice in financial markets. It is one of the primary factors in the determination of option prices for stocks and stock indexes. Since the combined option markets is larger than the combined stock markets in the United States, it is easy to understand the interest and the importance of modeling volatility in financial markets. In addition to its use in option pricing, volatility is very important in financial risk management and asset allocation. Finally, modeling the volatility of a time series may improve the efficiency of the estimates of model parameters as well as the accuracy of interval forecast.

The conditional heteroscedastic models discussed in this chapter include the autoregressive conditional heteroscedastic (ARCH) model of Engle (1982), the generalized ARCH (GARCH) model of Bollerslev (1986), the GARCH-M model of Engle, Lilien, and Robins (1987), the exponential GARCH (EGARCH) model of Nelson (1991), and a variety of threshold GARCH models. Useful literature in this area of research can be found in the review articles by Bollerslev, Chou, and Kroner (1992), and Bollerslev, Engle, and Nelson (1999). We shall discuss strength and weaknesses of each model and show applications of the models.

For example, we can consider a model AR(1) (ARIMA(1)):

$y_t = \gamma y_{t-1} + e_t$, where $e_t$ is Gaussian random process with constant variance $\sigma^2$ (if we do a standardization then we have zero expectation and unit variance). The process $e_t$ is *Autoregressive Conditionally Heteroscedastic* (*ARCH)*, if the conditional expectation is zero, but the conditional variance is dependent on time, for example (Engle): $\sigma_t^2 = \omega + \alpha_1 e_{t-1}^2 + \cdots + \alpha_q e_{t-q}^2$, and with probability 1, the conditional variance has remained positive if $\omega > 0, \alpha_1, \cdots, \alpha_q \geq 0$ Thus, the model allowed the data to determine the best weights to use in forecasting the variance.

Values of the realization process in the immediate past, large in absolute value are pulling an increase in the conditional variance at the moment, and the conditional probability realization $e_t$ with is great in an absolute value is arising again.

On the contrary, relatively small values $e_{t-1}, \cdots, e_{t-q}$ result in a reduction its probability.

Thus the key issue is the variance of the error terms and what makes them large. This question often arises in financial applications where the dependent variable is the return on an asset or portfolio and the variance of the return represents the risk level of those returns. These are time series applications, but it is nonetheless likely that heteroskedasticity is an issue. Even a cursory look at financial data suggests that some time periods are riskier than others; that is, the expected value of the magnitude of error terms at sometimes is greater than at others. Moreover, these risky times are not scattered randomly across quarterly or annual data. Instead, there is a degree of autocorrelation in the riskiness of financial returns. Financial analysts are looking at plots of daily returns and can notice that the amplitude of the returns varies over time and describe this as *volatility clustering*.

The ARCH and GARCH models, which stand for autoregressive conditional heteroskedasticity and generalized autoregressive conditional heteroskedasticity, are designed to deal with just this set of issues. They have become widespread tools for dealing with time series heteroskedastic models. The goal of such models is to provide a volatility measure – like a standard deviation that can be used in financial decisions concerning risk analysis, portfolio selection and derivative pricing.

The simplest ARCH (1) model has the form: $y_t = \sigma_t e_t$, $\sigma^2{}_t = a_0 + a_1 y^2{}_{t-1}$, where $e_t \in N(0,1)$. $a_0, a_1 \geq 0$ are conditions of the non-negativity variance. We'll obtain the estimation of parameters by the maximum likelihood estimation (MLE).

The success of conditionally Gaussian models ARSH (q), which has given a number of explanations of phenomena in the behavior of financial indices (clustering, heavy tail distribution, density values $y_t$), spawned a number of different generalizations of this model, pursuing the goal to explain several other phenomena.

Historically one of the first generalizations models ARCH (q) became the generalized ARCH model or GARCH that been proposed by T.Bollerslevom (T.Bollerslev) in 1986.

This model is characterized by two parameters (p ; q) and determined as

GARCH (p, q). Regarding volatility form, we have $\sigma^2{}_t = a_0 + \sum_{i=1}^{p} \alpha_i y^2{}_{t-i} + \sum_{j=1}^{q} \beta_j \sigma^2{}_{t-j}$, where $a_0 > 0, \alpha_i \geq 0, \beta_j \geq 0$.

The main advantage of this model is that for fitting of a time series GARCH (p, q) we can use not big of *p* and *q*.

For the simplest variant of GARCH(1,1) we obtain $\sigma^2{}_t = a_0 + \alpha_1 y^2{}_{t-1} + \beta_1 \sigma^2{}_{t-1}$, where $a_0 > 0, \alpha_1 + \beta_1 < 1$ for limitation of variance.

Further analysis of financial time series found the following phenomenon: values $y_{t-1}$ and $\sigma_t$ have negative covariance $Cov(y_{t-1}, \sigma_t) < 0$. Owing to proposed economic explanation of this phenomenon, it was named *leverage effect*. The so-called "leverage effect" first noted by Black(1976) refers to the tendency for stock prices to be negatively correlated with changes in stock volatility. A firm with debt and equity outstanding typically becomes more highly leveraged when the value of the firm falls. This raises the equity return volatility.

In practice, prices under low volatility behave so that there increasing or decreasing continue as long as possible. Similarly, under the great volatility of prices, it looks like as growth slow or decline of them to try to return movement in the opposite direction. This asymmetry effect can not be explained within ARSH end GARSH models because they are not sensitive to the volatility of past time series level.

To explain the asymmetry effect, Nelson (Nelson D.B.) proposed a model EGARSH (*p, q*) (Exponential GARSH) in 1991 for which the logarithm of the conditional variance is determined using a function of standard errors: $\ln \sigma^2{}_t = a_0 + \sum_{i=1}^{p} \alpha_i g(e_{t-i}) + \sum_{j=1}^{q} \beta_j \ln \sigma^2{}_{t-j}$, where the leverage effect is taken into account by function $g(e_t) = \theta e_t + \gamma(|e_t| - E|e_t|)$. Thus, $\sigma^2{}_t$ from EGARSH model depends on the magnitude and the sign of $y_t$, that allow reacting asymmetrically to unexpected ups and downs of the market.

We can recommend some investigation of various stock indexes in the works of Bollerslev, Engle, Nelson in which the models and their generalizations are described.

The Capital Asset Pricing Model (CAPM) describes the relationship between risk and expected return, and it serves as a model for the pricing of risky securities.

CAPM says that the expected return of a security or a portfolio equals the rate on a risk-free security plus a risk premium. If this expected return does not meet or beat our required return, the investment should not be undertaken.

The traditional CAPM model and its dynamic modification by R. Merton, Arbitration pricing theory by S. Ross indicate a proportional relationship between the expected excess market portfolio yield and its conditional variance. Economic theory holds that investors should be rewarded for taking risks. The ARCH-M (ARCH in mean) model provides an explicit link between the risk (conditional volatility) and the best forecast of a time series. These models foresee clear functional dependence conditional average random variable $y_t$ from its own conditional variance.

These models foresee clear functional dependence conditional average random variable $E(y_t / y_i) = \tau + \delta\sigma_t$ from its own conditional variance.

Commonly, one could use linear relationship, which has the following interpretation: the expected return on the market portfolio is divided into two components, such as risk-free income $\tau$ and risk premium $\delta\sigma_t$.

Economic agents expect an increase in yield due to the growing uncertainty factor of sensitivity to changes in the expected return conditional standard deviation associated with the degree of *relative risk aversion* that is assumed positive.

ARSH-M model had used to identify the risk premium in the term structure of interest rates, due to the hypothesis of efficiency currency market. Under the term structure of interest rates means the ratio between income securities with different maturities. This structure is illustrated by the curve and shows less income to short-term securities and over for long-term securities. Growth rates when moving from short- to long-term securities may explain increased risk of investing. Engle, Lilien, Robins modeled difference in yield of 6 and 3-month Treasury bills, using quarterly data for the period 1960-1984.

The dynamics of excess holding yield has discovered a significant component associated with changes in the conditional variance, but the average value of this component is only 0.14% for the quarter. The hypothesis of the efficiency of the currency market says that the forward exchange rate is the best not shifted forecast of the future course cash. But practical observations call into question the effectiveness of the currency market in this sense.

The shifting of forward rate does not necessarily indicate irrationality of market participants and can be a manifestation of the risk premium.

Empirically, various approximations of risk premium are directly related to the conditional variance spot rate (Kendall, McDonald) and shown conflicting results regarding the adequacy ARSH-M model, as developed by other specifications ARCH models (IGARCH is the *integrated model* GARCH, CHARMA is ARMA model with *conditional heteroskedasticity*, RCA is the *autoregression model with random coefficients*, SV is the *stochastic volatility model*, etc.).

*Example 3.7. Modeling and predicting of British Petroleum share price* .

We used data (*www.bp.com*) for modeling of the British Petroleum shares price from 1 January to 31 December 2010:

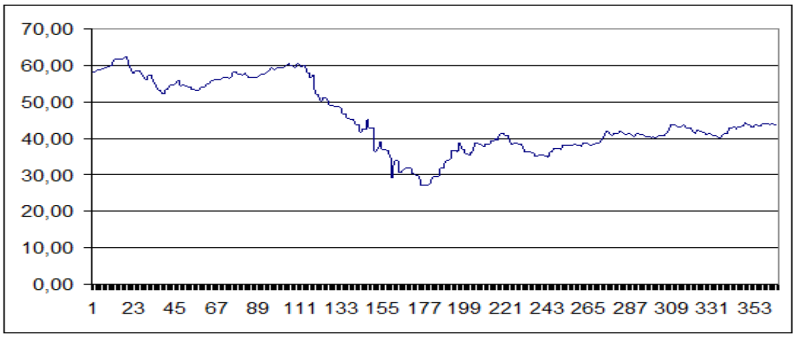



*Figure 3.16. Dynamic of British Petroleum share price movements in 2010*

We obtained by adding the trend line to the graph, that the choosing of parabolic trend (third order) would be best:

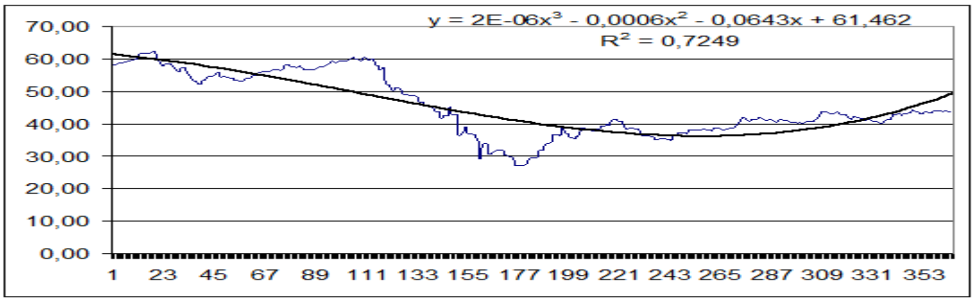



*Figure 3.17. Third order polynomial in comparison with the dynamics of the original time series*

So we transform the original time series by taking the third order finite differences on lag 3: $x_t = y_t - 3y_{t-1} + 3y_{t-2} - y_{t-3}$

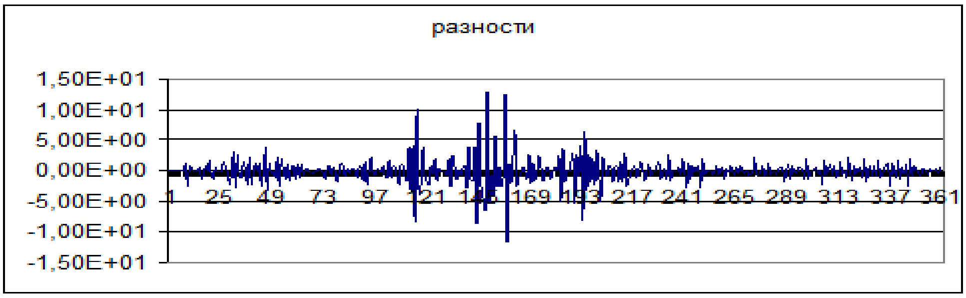



*Figure 3.18. Graph  finite difference third order*

We have got almost stationary a time series (Mathematical Expectation is zero, but we are observed clusters of variance). Indeed, we can conduct various tests (*Dickey-Fuller*, *Sidsel -Tukey*, *Akaike*, *Schwarz*), and find out that the error variance is approximated by model ARCH (5)

$$\sigma^2(\mathrm{e}_t) = 0{,}199459 + 0{,}186017 \cdot e_{t-1}^2 + 0{,}099732 \cdot e_{t-2}^2 + 0{,}027316 \cdot e_{t-3}^2 + 0{,}279598 \cdot e_{t-4}^2 +$$

$$+0{,}261907 \cdot e_{t-5}^2 ,$$

Moreover, all coefficients in the model of error variance satisfy the Engle conditions (they are positive and with a sum less than 1), ie, the model reduces the variance of the process.

We return to the original time series, finally have got a model AR (2) ARSH (5):

$$y_t = 3 \cdot y_{t-1} - 3 \cdot y_{t-2} + y_{t-3} + e_t - 1,958 e_{t-1} + 0,958 e_{t-2} .$$

According to the obtained model, we calculate data for the original time series. The following table presents the results for the last ten days of 2010:

*Table 3.19. Modeled and actual British Petroleum stock price on 22-31 December 2010*

| *Date* | *Actual data* | *Model* |
|---|---|---|
| *22.12.2010* | *43,61* | *43,61142431* |
| *23.12.2010* | *44,00* | *43,68235904* |
| *24.12.2010* | *44,00* | *44,08660048* |
| *25.12.2010* | *44,00* | *44,08395011* |
| *26.12.2010* | *44,00* | *44,08141926* |
| *27.12.2010* | *43,97* | *44,07900268* |
| *28.12.2010* | *44,11* | *44,04544419* |
| *29.12.2010* | *43,95* | *44,18913526* |
| *30.12.2010* | *43,89* | *44,02015424* |
| *31.12.2010* | *44,17* | *43,95574309* |

More clearly, one can see presented the actual and modeled time series level on the graph:

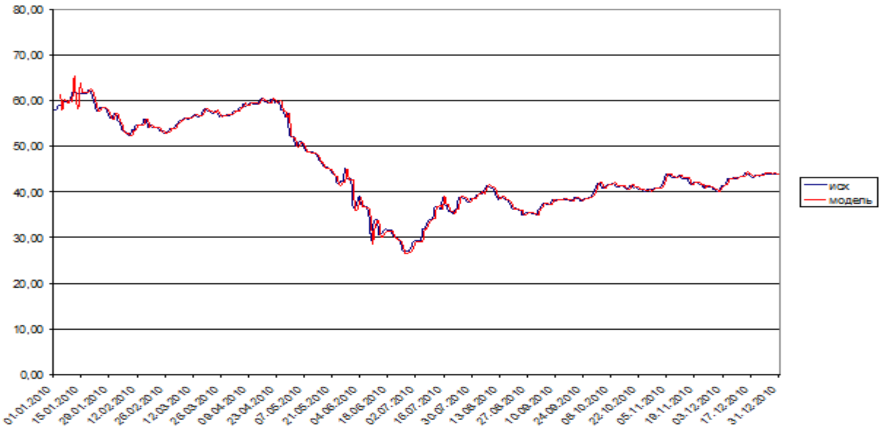



*Figure 3.19. Modeled and actual British Petroleum stock price on 22-31 December 2010*

According to information received we fulfill the forecasting model of stock price on 1.01.2011 and 2.01.2011:

*Table 3.20. Forecasts*

| Date | Forecast | Actual value |
|---|---|---|
| *01.01.2011* | *44,2457091* | *44,17* |
| *02.01.2011* | *44,3224263* | *45,15* |

As can be seen from the table, the forecast was quite accurate. The actual absolute error of the forecast for January 1 was 0.0757, and on January 2 it was 0.82757. The relative errors do not exceed 1.8%.

*Example 3.8. Analysis of time structure of interest rates based models ARSH (1) GARSH (1.1).*

We do the analysis of the term structure of interest rates on US government bonds with different maturities action on the subject of confirmation or refutation of the Expectancy theory. According to the Expectancy theory, the interest rate on long-term bonds must equal the arithmetic average of interest rates on short-term bonds for the corresponding period of possession. Checking the legality of the theory of expectations will be implemented through a simple model, which includes two periods (Mankiw, Miron) and has the form:

$$I_t = \theta + 0,5\left(i_t + E_t i_{t+1}\right),$$

where $I_t$ is the yield on long-term bonds in the time period $t$ ; $i_t$ , $i_{t+1}$ is yield on short-term bonds in the time period $t$ and $t+1$ respectively; $E_t$ are expectations in the time $t$ of return on short-term bonds in the time period $t+1$; $\theta$ is a bonus. According to the presented model, the expectancy theory can be presented as follows: the current yield is calculated for two periods must be equal to the arithmetic mean of return calculated for the current period and the expected return plus permanent time income. In other words, the yield of two-year government bond purchase should be the same as the acquisition successively one after another two-year government bonds. The simplest model can be represented as: $(i_{t+1}-i_t)=\alpha+\beta(I_t-i_t)+\varepsilon_{t+1}$, where $(i_{t+1}-i_t)$ is expected changes in short-term interest rates; $(I_t-i_t)$ reflects the slope of the yield curve; $\alpha=-2\theta$ and $\beta=2$ based on the definition of the theory of expectations; $\varepsilon_{t+1}$ is a forecast error. Checking the expectations theory is to establish the impact of market expectations to the slope of the yield curve. That is, we test how well the market can predict the movement of short-term interest rates are based on the estimated properties that are inherent to the spread between long-term and short-term interest rates on government bonds.

Monthly observations on US government bonds with a constant maturity of three (GS3m), of six months (GS6m), of one (GS1), of two (GS2),five month (GS5), ten month (GS10) and twenty month (GS20 ) at years from January 1982 to February 2003 are served as the information base. Calculations were made using a database of the Federal Reserve Bank of St. Louis. Originally, we should determine the order of the integrity data by the test of unit root *Dickey-Fuller* (*ADF test*). The optimal lag length was determined by using Akaike's Information Criterion:

$$\Delta Y_t = Y_t - Y_{t-1} = \alpha + \rho\sum_{j=1}^{p}\lambda_j \Delta Y_{t-j} + \varepsilon_t ,$$

where $\alpha, \rho$ are coefficients; $\varepsilon_t$ is a residual. To examine of the unit root we tested two hypotheses: $H_0 : \rho = 0$ (there is unit root); $H_A : \rho \neq 0$ (there is not unit root).

To identify the clustering of fluctuations the interest rates on government bonds we are used ARCH (1), GARCH (1.1) models.

It is known that the yield on long-term government bonds depends on factors such as inflation, fiscal and monetary policy and growth. These factors, providing a state of uncertainty may reduce the significance of the tests used. Therefore, further data are tested for the presence of structural shifts using the Chow test, which consists in the selection equation for each sub-sample to see whether there are significant differences in the constructed equations. Substantial differences indicate structural changes in existing relationships. The test is based on comparing the residual sum of squares is obtained by constructing separate equations for each sub-sample study.

To check the significance of Chow test we can use F-statistics and log-likelihood ratio (LR). F-statistic is based on a comparison of limited and unlimited sum squares residuals and in the simplest case it involves one structural shift that is tested on the promotion of the null hypothesis of no structural shifts:

$$F_{k,T-2k} = \frac{\left(ESS_R - ESS_{UR}\right)/k}{ESS_{UR}/\left(T-2k\right)},$$

where $\frac{m-1}{k}$ is limited sum of squares; $\frac{m-1}{k}$ is sum of the squared residuals of sample; $\frac{m-1}{k}$ are number observations; $\frac{m-1}{k}$ are the number of parameters in the equation. Log likelihood ratio is a maximum ratio of logarithmic likelihood Gaussian function. F-statistic has asymptotic $\chi^2$ – distribution with degrees of freedom $\frac{m-1}{k}$ under the null hypothesis of no structural break, where $m$ is the number of sub-samples.

The results of the unit root test presented in Table 3.21 . For interest rates on government bonds for all series, the null hypothesis can not be refuted even at the level of significance of 10%, which indicates the presence of a random walk data and, therefore, could lead to a distortion of the results in the construction of regression models. To resolve this issue, continue throughout the analysis, we should be using the spread between long-term and short-term interest rates and the first difference in short-term government bonds.

For short-term rates for all term, the null hypothesis is rejected. For the slope of yield curve is saved stationarity of interest rates on bonds of the term of three and six months and for interest rates on bonds of the term of six months and a year.

*Table 3.21. Unit root test (ADF)*

| *Governme* | | ***It*** | ***it*** | ***It– it*** | ***it+1*** |
|---|---|---|---|---|---|
| *GS3m–GS6m* | *Constant and trend (Lag)* | *–2.63 (0)* | *–2.44 (0)* | *–5.73* (0)* | *–11.22* (0)* |
| | *Constant (Lag)* | *–2.35 (0)* | *–2.11 (0)* | *–5.09* (0)* | *–11.18*(0)* |
| | *Without constant and trend (Lag)* | *–3.10* (0)* | *–2.83* (0)* | *–3.84* (0)* | *–10.99*(0)* |
| *GS6m–GS1* | *Constant and trend (Lag)* | *–2.08 (0)* | *–2.63 (0)* | *–4.18* (0)* | *–11.25*(0)* |
| | *Constant (Lag)* | *–1.57 (0)* | *–2.35 (0)* | *–3.84* (0)* | *–11.22* (0)* |
| | *Without constant* | *(0)* | *(0)* | *(0)* | |
| *GS1–GS2* | *Constant and trend (Lag)* | *–2.18 (0)* | *–2.08 (0)* | *–3.06 (0)* | *–10.56*(0)* |
| | *Constant (Lag)* | *–1.50 (0)* | *–1.57 (0)* | *–2.70***(0)* | *–10.54*(0)* |
| | *Without constant* | *(0)* | *(0)* | | *–10.27*(0)* |
| *GS5–GS10* | *Constant and trend (Lag)* | *–2.71 (0)* | *–2.53 (0)* | *–1.65 (0)* | *–10.13*(0)* |
| | *Constant (Lag)* | *–1.90 (0)* | *–1.72 (0)* | *–1.62 (0)* | *–10.12*(0)* |
| | *Without constant* | *(0)* | *(0)* | *(0)* | *–10.00*(0)* |
| *GS10–GS20[2] 1982:01–1986:12* | *Constant and trend (Lag)* | *–1.07 (0)* | *–1.15 (0)* | *–2.25 (1)* | *–3.46***(3)* |
| | *Constant (Lag)* | *–0.13 (0)* | *–0.22 (0)* | *–2.05 (1)* | *–3.48**(3)* |
| | *Without constant* | *(0)* | *(0)* | *(1)* | *–3.18** |
| *GS10–GS20 1993:10–2003:02* | *Constant and trend (Lag)* | *–2.95 (0)* | *–2.70 (0)* | *–2.25 (0)* | *–4.62* (3)* |
| | *Constant (Lag)* | *–0.73 (0)* | *–0.57 (0)* | *–1.23 (0)* | *–4.44* (3)* |
| | *Without constant and trend (Lag)* | *–0.65 (0)* | *–0.67 (0)* | *–0.21 (0)* | *–4.44* (3)* |

*Note. * - coefficients statistically significant at the 1% level, ** at the 5% level, *** at the 10% level.*

*2 - Sample of government bond with twenty-year maturity dates was the gap between December 1986 and October 1993 by the stop of bonds issue and the renewal of their registration only in 1993*

As we can be seen in Table 3.21, for the spread between interest rates on government bonds with two-year, five-year and ten-year, ten-year and twenty-year maturity dates(for both periods) we cannot refute the existence of a unit root hypothesis. Thus, the using of transformed data of the slope of the yield curve for the analysis of long-term government bonds is insufficient for it. The first difference of data are fixed at 1% significance level (see table 3.22 and 3.23):

*Table 3.22. The unit root test for first differences of yield curve slope*

| *Government Bonds* | ***ΔGS1–GS2*** | | | ***ΔGS5–GS10*** | | |
|---|---|---|---|---|---|---|
| | *Constant and trend (Lag)* | *Constant (Lag)* | *Without constant and trend (Lag)* | *Constant and trend (Lag)* | *Constant (Lag)* | *Without constant and trend (Lag)* |
| *It–it* | *–-11.90* (0)* | *–11.91* (0)* | *–11.93* (0)* | *–8.46* (2)* | *–8.47* (2)* | *–8.45* (2)* |

*Table 3.23. The unit root test for first differences of yield curve slope*

| *Government Bonds* | ***ΔGS10–GS20*** | | | | | |
|---|---|---|---|---|---|---|
| | ***1982:01–1986:12*** | | | ***1993:10–2003:02*** | | |
| | *Constant and trend (Lag)* | *Constant (Lag)* | *Without constant and trend (Lag)* | *Constant and trend (Lag)* | *Constant (Lag)* | *Without constant and trend (Lag)* |
| *It–it* | *–6.83* (0)* | *–6.90* (0)* | *–6.95* (0)* | *–8.24* (0)* | *–5.23* (2)* | *–5.24* (2)* |

*Note. * - coefficients statistically significant at the 1% level, ** at the 5% level, *** at the 10% level*

One can use the amendment to autoregression (model AR (1)) if we are convinced in the stationarity of data and build a regression model by the least squares method. However, standardized statistic $R^2$ for all kinds of bonds do not exceed value 0.213 it indicates that there is not property of prediction for the slope of the yield curve

Further analysis that is based on the identification of the time-dependent conditional of error variance through to heteroskedasticity of data, ie, on ARCH (1), GARCH (1,1) models. We can use Lagrange Multiplier (LM) as an assessment criterion, which for ARCH-first order model is equal $NR^2$, where $N$ is a sample size, and $R^2$ is the coefficient of determination calculated for the equation.

This statistic has $\chi^2$-distribution with one degree of freedom. We have obtained the following results:

*Table 3.24. Testing the Expectations Theory by ARCH (1) model*

| *Government Bonds* | *GS3m-GS6m* | *GS6m-GS1* | *GS1-GS2* | *GS5-GS10* | *GS10-GS20* | |
|---|---|---|---|---|---|---|
| | | | | | *1982:01-1986:12* | *1993:10-2003:02* |
| | *ARCH(1)[a]* | *ARCH(1)[b]* | *ARCH(1)[c]* | *ARCH(1)[d]* | *ARCH(1)[e]* | *ARCH(1)[f]* |
| *conditional mean* | | | | | | |
| *Constant* | *-0,065*** *(-2,272)* | *-0,093** *(-3,499)* | *-0,022* *(-0,987)* | *-0,035* *(-1,022)* | *-0,121**** *(-1,493)* | *-0,021* *(-0,572)* |
| $I_t$-$i_t$ | *0,230** *(3,064)* | *0,093** *(4,566)* | *0,395*** *(2,064)* | *0,261* *(0,972)* | *1,263*** *(2,272)* | *1,065** *(-2,272)* |
| *AR(1)* | *0,301** *(8,461)* | *0,223** *(7,429)* | *0,212** *(4,767)* | *0,425** *(6,819)* | *0,436** *(5,080)* | *0,432** *(2,906)* |
| $R^2$ | *0,106* | *0,090* | *0,122* | *0,176* | *0,235* | *0,104* |
| *N* | *252* | *252* | *251* | *251* | *58* | *110* |
| *LM* | *26,712* | *22,68* | *30,622* | *44,176* | *13,63* | *11,44* |
| Standard error | *0,330* | *0,318* | *0,314* | *0,295* | *0,377* | *0,226* |

*Note.1)* coefficients statistically significant at the 1% level, ** at the 5% level, *** at the 10% level ; there are t-statistics in parentheses.*

*2) Equations the time-dependent conditional error variance are:*

*a)* $\underset{(10,416)^{*}\ (5,291)^{*}}{\sigma_t^2 = 0,033 + 0,566\left(\varepsilon_{t-1}^2\right)}$ *b)* $\underset{(9,832)^{*}\ (4,868)^{*}}{\sigma_t^2 = 0,036 + 0,641\left(\varepsilon_{t-1}^2\right)}$ *; c)* $\underset{(9,185)^{*}\ (3,939)^{*}}{\sigma_t^2 = 0,050 + 0,512\left(\varepsilon_{t-1}^2\right)}$ *;*

*d)* $\underset{(9,206)^{*}\ (1,170)}{\sigma_t^2 = 0,077 + 0,093\left(\varepsilon_{t-1}^2\right)}$ *;e)* $\underset{(5,208)^{*}\ (-4,391)^{*}}{\sigma_t^2 = 0,148 - 0,151\left(\varepsilon_{t-1}^2\right)}$ *; f)* $\underset{(3,916)^{*}\ (0,971)}{\sigma_t^2 = 0,038 + 0,223\left(\varepsilon_{t-1}^2\right)}$ .

As one can be seen from the table, the slope of the yield curve under the conventional dispersion modeling has some, but it is a negligible information about the movement of short-term interest rates. Model GARCH (1,1) shows to preserve of the obtained effect(see the table bellow):

*Table 3.25. Testing the Expectations Theory by ARCH (1,1) model*

| *Government Bonds* | *GS3m-GS6m* | *GS6m-GS1* | *GS1-GS2* | *GS5-GS10* | *GS10-GS20* | |
|---|---|---|---|---|---|---|
| | | | | | *1982:01-1986:12* | *1993:10-2003:02* |
| | *ARCH(1)*[a] | *ARCH(1)*[b] | *ARCH(1)*[c] | *ARCH(1)*[d] | *ARCH(1)*[e] | *ARCH(1)*[f] |
| *Conditional mean* | | | | | | |
| *Constant* | *-0,076** *(-2,464)* | *-0,083** *(-3,499)* | *-0,029* *(-1,987)* | *-0,039**** *(-1,322)* | *-0,132**** *(-2,493)* | *-0,034* *(-1,572)* |
| $I_t$-$i_t$ | *0,303** *(2,525)* | *0,278** *(3,566)* | *0,139* *(0,764)* | *0,195* *(0,772)* | *1,363*** *(2,872)* | *0,565* *(1,272)* |
| *AR(1)* | *0,399** *(7,461)* | *0,387** *(6,429)* | *0,423** *(8,767)* | *0,439** *(7,819)* | *0,490** *(6,080)* | *0,332** *(3,906)* |
| $R^2$ | *0,106* | *0,128* | *0,168* | *0,177* | *0,237* | *0,106* |
| *N* | *252* | *252* | *251* | *251* | *58* | *110* |
| *LM* | *26,46* | *32,68* | *42,622* | *44,476* | *13,73* | *11,66* |
| Standard error | *0,311* | *0,312* | *0,307* | *0,296* | *0,380* | *0,227* |

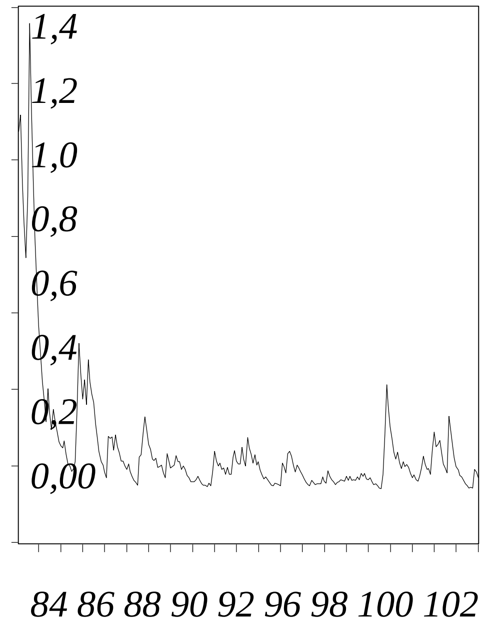
*Note.1)* coefficients statistically significant at the 1% level, ** at the 5% level, *** at the 10% level ; there are t-statistics in parentheses.*

*2) Equations the time-dependent conditional error variance are:*

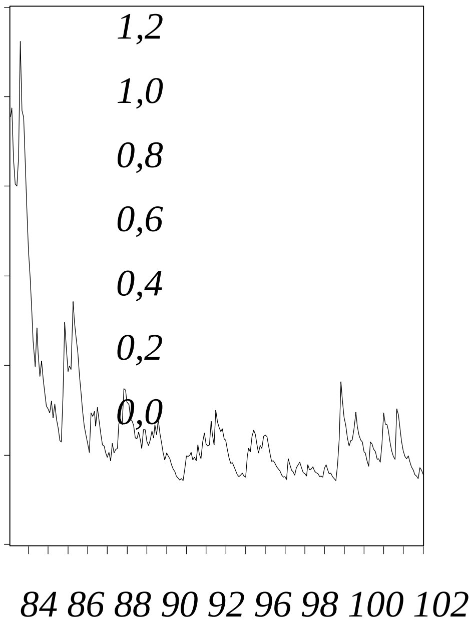
*a)* $\underset{(3,316)^{*}\ (4,291)^{*}\quad (14,720)^{*}}{\sigma_t^2 = 0,006 + 0,200\varepsilon_{t-1}^2 + 0,680\sigma_{t-1}^2}$ ; *b)* $\underset{(2,316)^{*}\ (3,291)^{*}\quad (15,720)^{*}}{\sigma_t^2 = 0,004 + 0,183\varepsilon_{t-1}^2 + 0,741\sigma_{t-1}^2}$ ; *c)* $\underset{(2,316)^{*}\ (3,091)^{*}\quad (12,720)^{*}}{\sigma_t^2 = 0,007 + 0,170\varepsilon_{t-1}^2 + 0,726\sigma_{t-1}^2}$ ;

*d)* $\underset{(7,316)^{*}\ (-2,091)^{**}\quad (50,720)^{*}}{\sigma_t^2 = 0,002 - 0,040\varepsilon_{t-1}^2 + 1,080\sigma_{t-1}^2}$ ; *e)* $\underset{(11,316)^{*}\ (-3,291)^{*}\quad (15,720)^{*}}{\sigma_t^2 = 0,015 - 0,151\varepsilon_{t-1}^2 + 1,025\sigma_{t-1}^2}$ ; *f)* $\underset{(4,816)^{*}\ (-4,291)^{*}\quad (30,720)^{*}}{\sigma_t^2 = 0,007 - 0,181\varepsilon_{t-1}^2 + 1,033\sigma_{t-1}^2}$ .

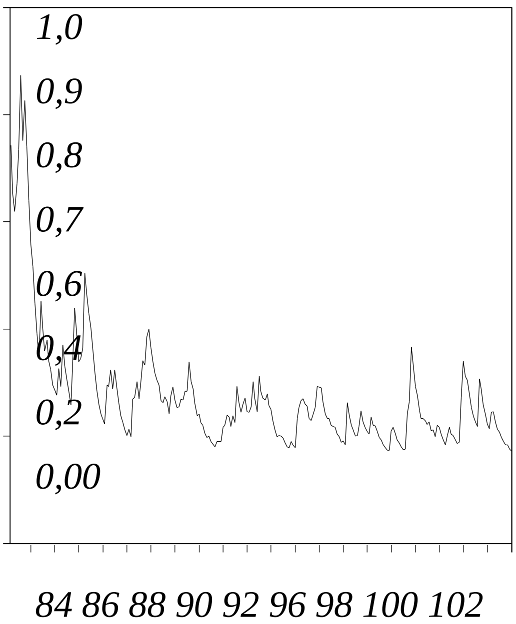
Studies show that, contrary to expectations theory, the slope of the yield curve does not have sufficient information that would allow to form the basis for the forecasts of the movement of short-term interest rates for government bonds in all the examined lifetime. Thus, the spread between short-term and long-term interest rates has the low predictive ability (it should be noted that the yield curve still has some ability to predict short-term interest rates on government bonds for a period of up to one year).There are graphs of conditional standard deviations on figure bellow:



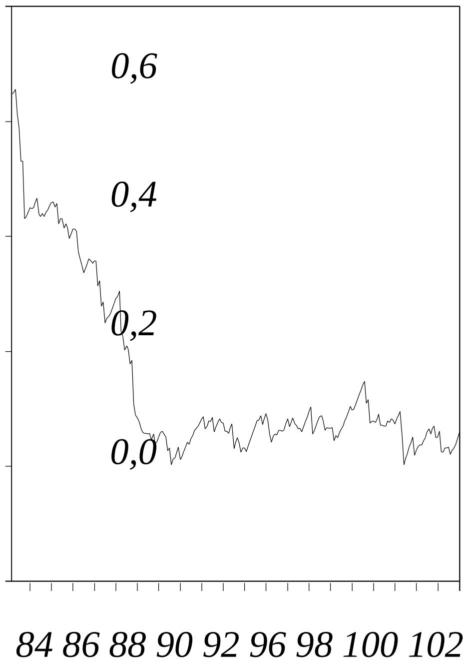

*(a) three and six months*



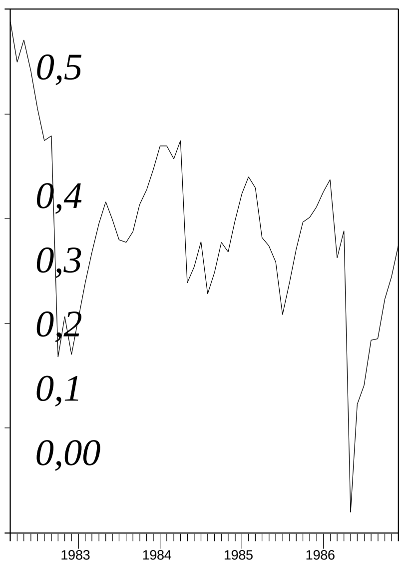

*(b) six months and one year*

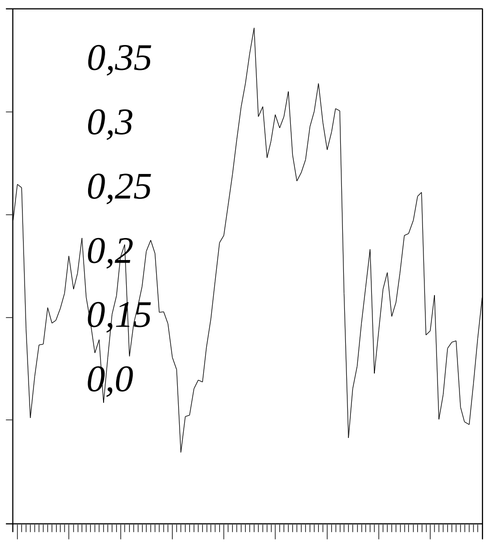



*(c) one and two years*




*(d) five and ten years*




*(e) ten and twenty years*
*January 1982 to December 1986*

*(f) ten and twenty years*
*October 1993 to February 2003.*

*Figure 3.17. Conditional standard deviation of short-term interest rate for government bonds with different dates of maturities*

Properties the constancy of indicators the model GARCH (1,1) of government bonds for all maturities action is significant. On the next step of the analysis, we should examine the structural breaks in the data that is tested by the test Chow. October of 1982 was chosen as a potential point gap( when the Federal Reserve Bank has changed the principles of monetary policy). The results are presented in the following table:

*Table 3.26. The test for the presence of structural shifts: Chow test*

| | GS6m-GS3m | | GS1-GS6m | | GS2-GS1 | | GS10-GS5 | | GS20-GS10 | |
|---|---|---|---|---|---|---|---|---|---|---|
| erm | Fstat. | LLR | Fstat. | LLR | Fstat. | LLR | F-stat. | LLR | F-stat. | LLR |
| 5.82 | .771 | 3.804 | 383 | 4.530 | | | | | | |
| 6.82 | 3.610 | 8.696 | 2.250 | 5.085 | .592 | 6.625 | .178 | .581 | .823 | 5. 801 |
| 7.82 | 3.664 | 3.893 | 8.943 | 2.375 | 1.063 | 1.887 | .446 | 3.306 | .158 | 6. 804 |
| 8.82 | 2.106 | 58.992 | 6.416 | 0.373 | .238 | 4.123 | .096 | .338 | .956 | 3. 113 |
| 9.82 | 9.785 | 8.083 | 0.692 | 6.705 | .593 | 2.315 | .320 | .030 | .367 | 1. 214 |
| 0.82 | 7.601 | 9.001 | 3.636 | 3.826 | 4.773 | 1.733 | .060 | 5.090 | .210 | 3. 914 |
| 1.82 | 3.344 | 7.996 | 8.997 | 2.510 | .295 | 4.280 | .908 | 1.734 | .217 | 3. 937 |
| 2.82 | 3.308 | 7.900 | 7.170 | 7.910 | .660 | 2.503 | .846 | .599 | .430 | 1. 421 |
| 1.83 | 2.811 | 6.582 | 6.696 | 6.701 | .355 | 1.645 | .772 | .388 | .453 | 1. 498 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2.83 | 2.193 | 4.935 | 4.219 | 0.296 | .692 | 6.911 | .238 | .777 | .494 | 1. 629 |
| 3.83 | 1.423 | 2.866 | 1.715 | 3.652 | .043 | 2.130 | .495 | .552 | .615 | 2. 024 |
| 4.83 | 1.688 | 3.580 | 1.383 | 2.758 | .730 | 1.210 | .211 | .694 | .556 | 1. 833 |

*Note. F-stat.is the F-statistics; LLR is a log likelihood ratio.*

For data series with the spread between government bonds with one and two-year, five and ten-year maturities confirmed the hypothesis about the existence of structural shift in October 1982. However, the short term of the securities in the three and six months, six months and year showed that the structural shift was in August 1982. There were no structural shifts (all F-statistics have been insignificant) for time series on ten and twenty years.

## Dynamic factor models DFM and DFMS or MS-DFM (with Markov switching)

We will treat a class of models, so-called dynamic factor models (DFMs), which has received considerable attention in the past decade because of their ability to model simultaneously and consistently data sets in which the number of series exceeds the number of time series observations. Dynamic factor models were originally proposed by Geweke (1977) as a time-series extension of factor models previously developed for cross-sectional data. In early influential work, Sargent and Sims (1977) showed that two dynamic factors could explain a large fraction of the variance of important U.S. quarterly macroeconomic variables, including output, employment, and prices. This central empirical finding that a few factors can explain a large fraction of the variance of many macroeconomic series has been confirmed by many studies; see for example Giannone, Reichlin, and Sala (2004) and Watson (2004).

Dynamic factor models have the twin appeals of being grounded in dynamic macroeconomic theory (Sargent (1989), Boivin and Giannoni (2006)) and providing a good first-order description of empirical macroeconomic data, in the sense that a small number of factors explain a large fraction of the variance of many macroeconomic series (Sargent and Sims (1977), Giannone, Reichlin, and Sala (2004), Watson (2004)). Theoretical econometric research on DFMs over the past decade has made a great deal of progress, and a variety of methods are now available for the estimation of the factors and of the number of factors. Theoretical work is ongoing in many related areas, including weak factor structures, tests for the number of factors, and factor models with instability and breaks.

The bulk of empirical work to date has focused on forecasting. A great deal is now known about the performance of factor forecasts and about best practices for forecasting using factors. Broadly speaking, this research has found that linear factor forecasts perform well to very well relative to competitors for many, but not all, macroeconomic series. For U.S. real activity series, reductions in pseudo out-of-sample mean squared forecast errors at the two- to four-quarter horizon are often in the range of 20%-40%, although smaller or no improvements are seen for other series, such as U.S. inflation post-1990. Parametric (third-generation) DFMs are also particularly well suited for now casting. Work on other empirical applications of DFMs, such as structural VARs and estimation of parameters of DSGEs, is newer and while there are relatively few such applications to date, these constitute promising directions for future empirical research.

The dynamic factor models DFM play an important role in modern models of short-term forecasting of various economic indicators. We'll consider a generalization of DFM that is dynamic factor model with Markov switching (MS-DFM) on the example of its application for calculating short-term forecast of social and economic development of Russia (within CES: Byelorus Republics, Kazakhstan and the Russian Federation).

*Example 3.9.The prediction model of short-term indicators of socio-economic development of Russia*

We consider computer model of short-term prediction for indicators of socio-economic development of the Russian Federation, which consists of two components: the first is a forecast model DFM; the second is DFMS.

DFMS model can be viewed as a combination of standard DFM model and model for time series with a Markov switching (Hamilton).

The basic assumption of DFM is that the dynamic behavior of all macroeconomic time series that is observed and determined by a relatively small number of common factors with those are not observed (so-called the latent factors or principal components).

Each variable (latent) is a linear combination of factors plus the idiosyncratic noise through the measurement error. Potentially, DFM may contain quite a number of observable factors, but it does not lead to the "curse of dimensionality", as in the case of vector autoregression (VAR).

The number of estimated parameters increases linearly, it is not quadratically as a case of the VAR.

Substantiated estimation of common unobservable factors can be obtained by the method of principal components (Stock and Watson), owing to the imposed factorial structure.

DFM with Markov switching (DFMS or MS-DFM) is based on an earlier generation of DFMs that is operated with a relatively small number of observed variables (Stock and Watson), which allow estimate them by maximum likelihood method (MMP). Hamilton was first, who proposed an idea of unobservable Markov switching between the modes, he has been using it in an econometric model for time series of GDP growth.

If we say the unobservable mode we mean the phase of the business cycle recovery or recession. In practice, one can talk about the unobserved condition of the economy as the regime(mode) and about the probability that it is in one of two modes.

Therefore, we would use as a variable the real GDP growth, then the simple model with Markov switching can be considered as autoregression first order for real GDP growth in which the free term - random variable that takes two values: it is high for recovery mode and it is low for mode recession (as average growth rate in the corresponding phase of the business cycle).All known specifications of DFMS (on published works: Chauvet, 1998; Kimand Nelson, 1999; Chauvet and Hamilton, 2006; Chauvet and Potter, forth coming 2013) contains only one factor, which is interpreted as the state of business activity. It is assumed that the dynamic behavior of unobservable (latent) factor obeys a simple Markov model with Markov switching between modes - recovery and recession. Generally to assess this latent factor "business activity" one should take four time series, for which there is a consensus that each of them varies in one phase of latent (coincident indicator), but not leading or lagging. Typically, that is such series as using real income, employment levels, sales in industry and trade, and the index of industrial production ( factor structure is superimposed on these variables). If the autoregression equation for the factor "business activity" and equations that link the observed numbers of these factors are known, directly the factor can be estimated by a Kalman filter (we assume the normality of errors), and the same parameters can be estimated by MLM. One can use short-term economic indicators and survey data and indicators of financial and commodity markets as the input data. Specifications of predictive models have examined with different numbers of factors (one to three). For each specification calculated the mean prediction error, which is the criterion for a model. Mean square error of prediction is calculated on historical data in the prediction outside the sample evaluation (out-of-sample forecasting), predicting that plays in real time. Forecasts are calculated on the basis of information available at the end of the first, second and third months of the relevant quarter. In this case, each time situation is reproduced where different data are available at different levels on the forecasting date (called an end to unequal sample). Forecasting Horizon is current quarter (nowcast), as well as one and two-quarters ahead. Used 92 series. The first month of observation is December 1995. Results forecasting by DFM model:

*Table 3.27. Mean square error of forecast real GDP growth of (by % to the previous quarter, with adjustment for seasonality) is based on the information at the end of the first month of the relevant quarter, percentage points*

| | *Current* | *One quarter* | *Two* |
|---|---|---|---|
| *One* | *1.9* | *1.8* | *3.2* |
| *Two* | *1.7* | *1.8* | *3.2* |
| *Three* | *1.4* | *1.6* | *3.1* |

*Table 3.28. Mean square error of forecast real GDP growth of (by % to the previous quarter, with adjustment for seasonality) based on information as of the end of the second month of the relevant quarter, percentage points*

| | *Current* | *One quarter* | *Two* |
|---|---|---|---|
| *One* | *1.7* | *1.7* | *3.1* |
| *Two* | *1.6* | *1.7* | *3.1* |
| *Three* | *1.6* | *1.6* | *3.1* |

*Table 3.29. Mean square error of forecast real GDP growth of (in% to the previous quarter, with adjustment for seasonality) based on the information at the end of the third month of the relevant quarter, percentage points*

| | *Current* | *One quarter* | *Two* |
|---|---|---|---|
| *One* | *1.4* | *1.6* | *3.1* |
| *Two* | *1.1* | *1.7* | *3.1* |
| *Three* | *1.1* | *1.6* | *3.0* |

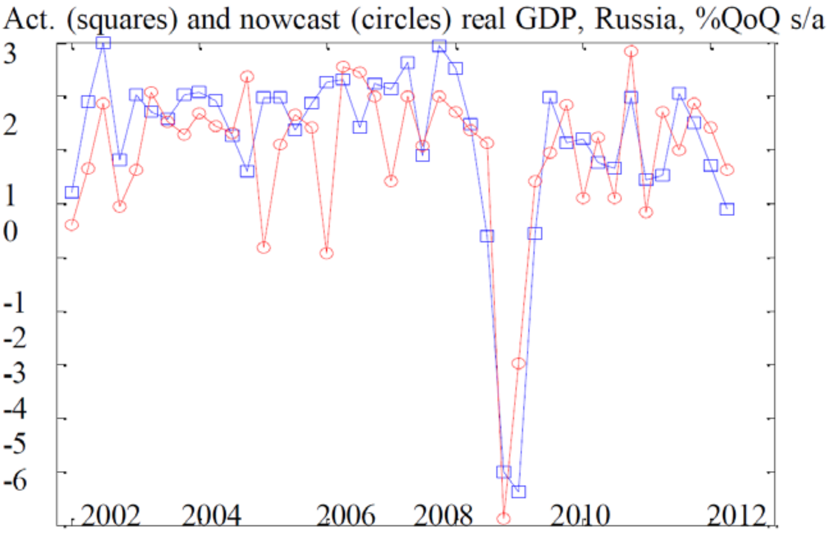



*Figure 3.19. Actual (squares) and projected (circles) values of real GDP growth, by % to the previous quarter, with adjustment for seasonality. Forecast based on information as of the end of the third month of the relevant quarter. The predictive model is DFM of three factors. The horizon of forecasting is the current quarter*

Forecasting by DFMS models.

The model is based on the Markov-switching model of Hamilton (1989), which endogenously estimates the timing of regime shifts in the parameters of a time series model, seems well suited for the task of modeling business cycle phase shifts. This feature of the business cycle is often captured using the dynamic common factor model of Stock and Watson (1989, 1991). Chauvet (1998) combined the dynamic-factor and Markov- switching frameworks to create a statistical model capturing both regime shifts and comovement.

We'll do modeling with using four indicators of business activity.

We suppose the existence of an aggregated cyclical factor $F_t$, which dynamics is described by the equation: $F_t = \alpha^{(S_t)} + \phi F_{t-1} + e_t, \quad S_t = 1,2$, where $e_t \sim N(0, \sigma_e^2)$ and $\alpha^{(S_t)} = \alpha^{(1)}$, when the economy is in a state of recovery ($S_t = 1$), i $\alpha^{(S_t)} = \alpha^{(2)}$, when the economy is in a state of decline ($S_t = 2$).

Also assume that the growth rate of each of the four monthly indicators $y_{rt}$ satisfy the relation $y_{rt} = \lambda_r F_t + \nu_{rt}, \quad r = 1,2,3,4.$ where the idiosyncratic component of monthly indicator $\nu_{rt}$ obeys to the first order autoregressive process: $\nu_{rt} = \theta_r \nu_{r,t-1} + \varepsilon_{rt}$.

Thus, the increase (decrease) in aggregate factor $F_t$ on $\Delta$ leads to an increase (decrease) in each variable monthly indicator $y_{rt}$ to value $\lambda_r \Delta$; the larger the ratio $\lambda_r$ (load factor), the more indicator $r$ is subjected to fluctuations of an aggregate factor $F_t$. The indicator $r$ also exposed to idiosyncratic shock $\nu_{rt}$ that does not affect other indicators.

DFMS model assumes that unobservable component $S_t$ that is responsible for the phase of the business cycle, subject to Markov chains, ie $P(S_t = j / S_{t-1} = i, S_{t-2} = k, ..., Y_{t-1}) = p_{ij}$, where $Y_{t-1}$ describes the history of implementations four indicators $y_{rt}$ until the $t-1$ inclusively. The described system of dynamic equations can be represented in a state space of a Markov switching. In this case, the state vector $\overline{f_t} = (F_t, \nu_{1t}, \nu_{2t}, \nu_{3t}, \nu_{4t})^T$ contains 5 component and its dynamics can be characterized by the equation (transition equation): $\overline{f_t} = \alpha^{(S_t)} \overline{e_5} + \Phi \overline{f_{t-1}} + \overline{a_t}$ , where $\overline{e_5} = (1,0,0,0,0)^T$, $\overline{a_t} = (e_t, \varepsilon_{1t}, \varepsilon_{2t}, \varepsilon_{3t}, \varepsilon_{4t})$, $\Phi = diag\{\phi, \theta_1, \theta_2, \theta_3, \theta_4\}$.

All values and model parameters are calculated (estimated) through recursive Kalman filter by method (quasi) maximum likelihood under standard limits (laws normality of distribution errors). Russia has used four series, the most similar to those used to identify NBER recession. This is the real retail sales (similar to sales in the Industry and Trade), the first major component of disposable real income and real wages (analog real disposable income), the first major component of the rate of unemployment and the needs of the workers (similar employment outside the agricultural sector) and the index of industrial production. The first month of observation is December 1995.

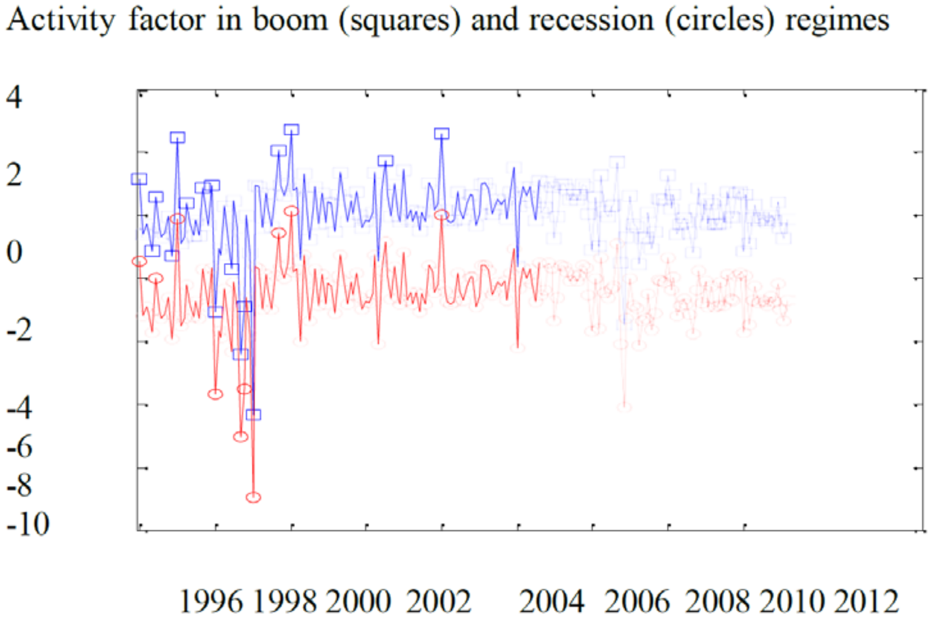



*Figure 3.19. The estimated factor activity in a state of boom (squares) and recession (circles) for Russia*

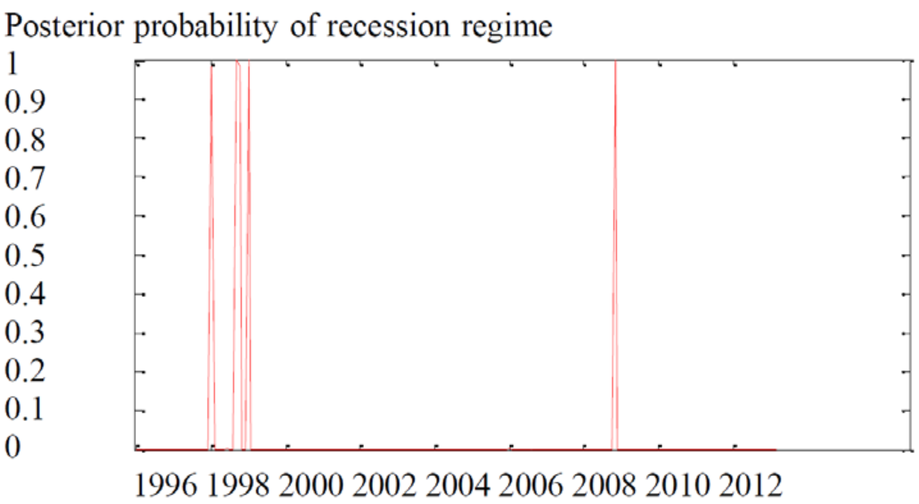



*Figure 3.20. A posteriori probability of recession for Russia*

Russia had only two cases of recession probability close to 1 (see Figure 3.20 for 1996 -2013 years). This is a crisis of the late 1990s and the latest financial and economic kryza2007-2009 years. Finally, the prediction for the current quarter has performed:

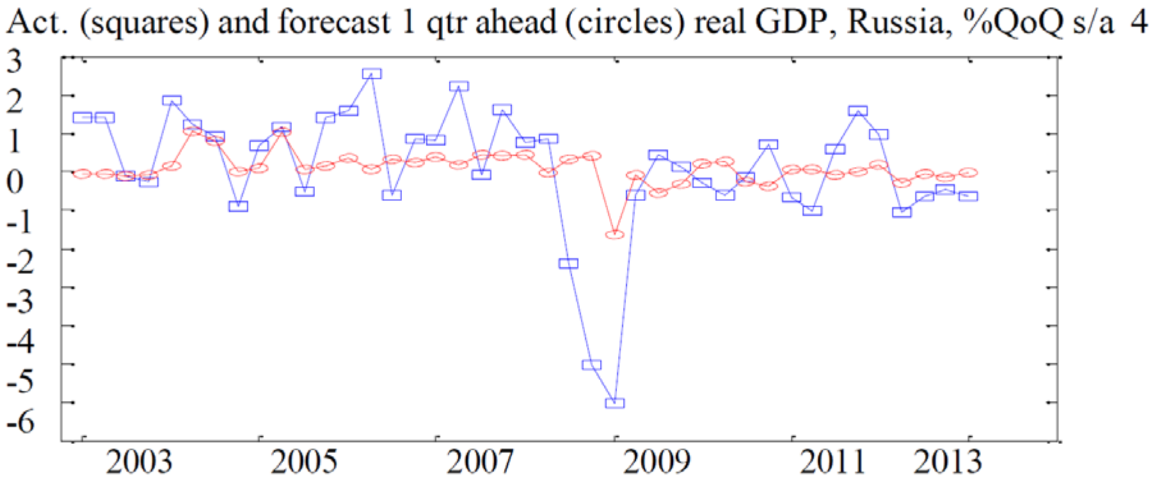



*Figure 21. Actual (squares) and projections (circles) values of real GDP growth of, by % to the previous quarter, with adjustment for seasonality. Forecast based on information as of the end of the third month of the relevant quarter. The predictive model is DFMS. The Horizon is forecasting the current quarter*

The results show the ability to use the predictive models DFMS as predictive indicators of business activity and ability of application to the economies of different countries.

## Chapter 4. Special cases of regression analysis

4.1. Nonlinear regression analysis, specifics of analyse of panel (space-time) data. Modeling of causal systems (Systems of structural regressions)

Often at a simulating of nonlinear relationships (especially on dynamic data series) relative quantities are widely used, in particular, various indexes (rates). Their behavior is characterized by greater constancy in time compared to the absolute values. In addition, the transition to the dimensionless variables is made, as well as the possibility to exclude multicolinearity and autocorrelation of the residuals appears. A model of the so-called production regression is prevalent on practice:

$$Y = f(x_1, x_2, ..., x_m, t),$$

where $Y$ is the result of the production, $x_1, x_2, ..., x_m$ are production factors , and $t$ is the time. The classic example of this type of nonlinear functions is the production function (regression) of the Cobb-Douglas-Tinbergen, which describes the relationship between the factors and the result of production at any level of economic activity (firm, branch, region, the economy in General):

$$Y = Ax_1^{b_1} x_2^{b_2} x_3^{b_3} ... x_m^{b_m} e^{\lambda t},$$

where $b_i$ is elasticity coefficient, which shows how much% at average will change $Y$ at 1% change of $x_i$ provided all other factors are constant (i.e., the coefficient of elasticity is the relative effect of influence of factor $i$ on $Y$), and the trend of the results, due to the influence of others, non-extensive (non-production) factors taken into account as a time variable $e^{\lambda t}$ in the model . For exponent function we find logarithm and bring it to linear, and therefore the parameters we define by the method of least squares.

For example, consider the macro function of the Cobb-Douglas-Tìnbergen for United States, Britain and Japan, $Q = AK^{\alpha}L^{\beta}e^{\lambda t}$ (here $Q$ is the gross national product; $K$ is fixed assets, $L$ - human resources; λ characterizes the growth rate function by nonextensive factors), constructed according to the period of 1950-1977. Applying logarithmic differentiation to this function, we obtain the model that describes the interconnection of rates of increase:

$$q = \alpha k + \beta l + \lambda,$$

where *q, k, l is the* growth in accordance with the result, capital and labour costs:

*Table 4.1. Estimation of parameters of regression rate*

| Country | α | β | λ |
|---|---|---|---|
| United States | 0.447 | 0.553 | 1.34 |
| United Kingdom | 0.506 | 0.494 | 0.53 |
| Japan | 0.397 | 0.603 | 4.66 |

Based on this model, you can define the contribution of intensive and intensive factors into the development processes of reproduction:

$$d_{ex} = (\alpha k + \beta l)/q;$$

$$d_{int} = \lambda / q.$$

Based on the values of function parameters it is possible to make conclusions about the features of economic development in each of these countries in the postwar years. So, technological progress was the most intensive implemented in the economy of Japan: the parameter λ is higher in comparison with the United States in 3.5 times, compared with the UK-9 times. Meanwhile, Japan's economy is characterized by the lowest capital intensity (α = 0,397) and a relatively high level of efficiency of labour resources (β = 0,603). For the U.S. economy it is typical the balance of elasticity ratio of capital and labor. Average annual growth in the United States in the postwar years were (in %): gross national product - 3.38, fixed assets - 2.79, manpower - 1.46. Hence the contribution of factors in shaping the dynamics of GDP is:

Intensive $d_{ex} = \frac{0,447 \cdot 2,79 + 0,553 \cdot 1,46}{3,38} = 0,625;$

Intensive $d_{int} = \frac{1,34}{3,38} = 0,375$ .

In practice, different modifications of the production function are applied. For example, dividing both parts by $L$ we obtain the function of productivity:

$$W = AF^{\alpha} e^{\lambda t},$$

where $W$ is the efficiency of labor; $F$ is the capital-labor ratio.

The rate of increase in this function is written as follows:

$$w = \alpha f + \lambda = \alpha (k - l) + \lambda.$$

The contribution of intensive and intensive factors in the dynamics of labour productivity is defined similarly:

$$d_{ex} = \alpha (k - l) / w;$$

$$d_{int} = \lambda / w.$$

The classical regression is applied to the rates of increase.

Power function describes also the relationship between demand $C$, average per capitaincome $D$ and the prices on product $P$. The trend of demand is driven by habits, fashions, etc., and is introduced into the model as a time variable $e^{\lambda t}$:

$$C = A D^{\alpha} P^{\beta} e^{\lambda t},$$

where α and β are the coefficients of elasticity of demand depending on income and prices.

*Example 4.1. Simulation of the influence of professional skill of workers improvement on the economic activity results.*

Lately many studies of economists around the world dedicated to the influence of the so-called "human capital" on the growth of the economy.In particular, the Nobel Laureate Jo Stiglitz writes that human capital is an important factor of economic growth. According to the estimations of a number of specialists, increasing the duration of education for one year leads to increase of GDP by 5-

15%. Now the theory of human capital is the new direction of research, and new problem is reproduction of the labor force, its development and use. Common economic characteristic of investment in human capital is that their performance far exceeds the profitability of investments in physical capital. This was the theoretical basis of unprecedented development of education, training and retraining in the developed countries. American scientist Denison E. quantitatively assessed the impact of the growth of various economic factors in the United States GDP growth for 1929-1982. He estimates that improving the quality of the labor force identified a 14% increase in GDP.

We will consider that GDP is modelled by a function of the Cobb-Douglas-Tìnbergena:

$$Q = AK^{\alpha} L^{\beta} e^{\lambda t},$$

where $Q$ is GDP (result); $K$ is the fixed capital; $L$ is labour costs (number of employees); and $e^{\lambda t}$ describes the so-called neutral technical progress.

Parameters $\alpha$ and $\beta$ are the coefficients of elasticity: $\alpha$ describes the relative GDP increment per unit of capital at $L = const$ and $\beta$ is the relative GDP increment per unit labour costs at $K = const$. Capital and labor costs are considered as factors of the extensive development (attracting new resources). Parameter $A$ results the scale (dimension) of factors to the scale of the result. When $A = 1$ and the GDP is trend of the result, caused by the influence of the other, nonextensive factors taken into account in the model as a time variable $e^{\lambda t}$ ($\lambda$ describes the rate of GDP growth by nonextensive factors: improvements in technology, the growth of qualification, etc.).

Applying logarithmic differentiation to the Cobb-Douglas-Tìnbergen function ( $(\ln f)' = \frac{f'}{f}$), we obtain a linear model, which describes the relationship of growth rates: $q = \alpha k + \beta l + \lambda$ where $q, k, l, \lambda$ are the growth rates of the GDP, capital, labor costs and level of qualification respectively. Everywhere in the future as the growth rate of the index we understand the ratio of the next level to the previous: $\frac{f_i}{f_{i+1}}$. This model allows you to define the contribution of intensive and intensive factors in GDP.It is obvious that a similar approach can be applied to the modeling of growth of gross product for separate economic sectors.

First let's investigatethe influence of total number of workers, who increased qualification (educ) during 2004-2014 on GDP (vvp) of Ukraine:

*Table 4.2. The number of people who get a qualification (thousand) and GDP (million USD) and the rate of their growth*

| *Year* | *The number of people who get a qualification, thsd. persons* | *GDP mln. UAH* | *Growth of qualification* | *GDP growth* |
|---|---|---|---|---|
| *2004* | *930.3* | *345113* | | |
| *2005* | *976* | *441452* | *1.049124* | *1.279152* |

| | | | | |
|---|---|---|---|---|
| *2006* | *994.5* | *544153* | *1.018955* | *1.232644* |
| *2007* | *994.5* | *720731* | *1* | *1.324501* |
| *2008* | *1071.2* | *948056* | *1.077124* | *1.315409* |
| *2009* | *1022.7* | *913345* | *0.954724* | *0.963387* |
| *2010* | *943.9* | *1082569* | *0.922949* | *1.185279* |
| *2011* | *978.4* | *1316600* | *1.03655* | *1.216181* |
| *2012* | *1016.5* | *1408889* | *1.038941* | *1.070096* |
| *2013* | *976.9* | *1454931* | *0.961043* | *1.03268* |
| *2014* | *804.1* | *1566728* | *0.823114* | *1.07684* |

For clarity let's construct the diagram of this dependence:

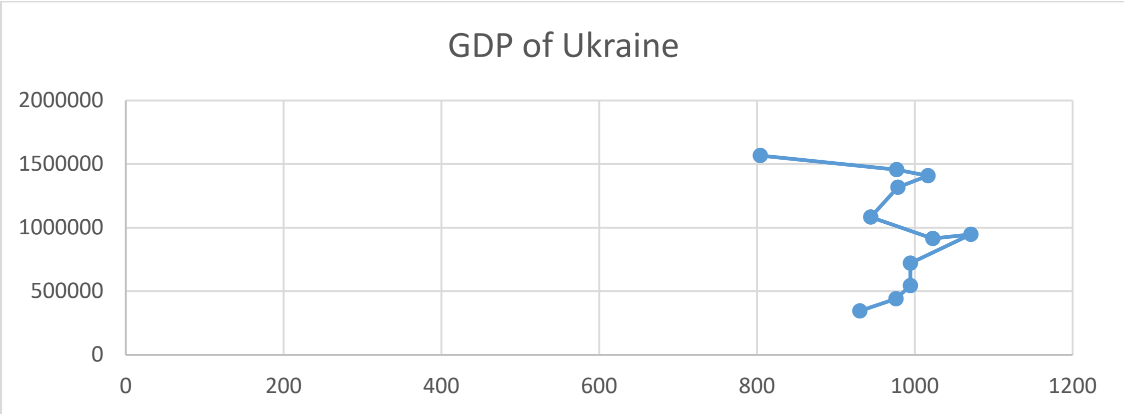



*Fig. 4.1. Dependence of Ukraine's GDP (million USD) from the number of people who get a qualification (thousand persons).*

Analyzing the diagram, we make the conclusion about the non-linearity of dependencies. So turn to growth (ted) and GDP (tvvp), the relationship between them is depicted in the following figure:


*Fig. 4.2. Dependenceof rates of Ukraine's GDP change on the rate of qualification change.*

The resulting dependence is close to linear, and more precisely to the piece-linear. Therefore, when modeling we use the module "piece-linear regression" of specialized programStatistica.

As a result, we get:

*Table 4.3. Results of modeling the impact of rate of qualification growth (ted) on the GDP growth (tvvp)*

| N=10 | Model: Piecewise-lin. with point (Table data1)<br>Depend. var.: twp loss: l.s.<br>Final loss: ,018007158 R= ,93512 explanatory div.: 87, | | | | |
|---|---|---|---|---|---|
| | B0 | ted | B0 | ted | P.gap |
| Estimate | 1,135363 | -0,105470 | 0,614574 | 0,633237 | 1,169617 |

So we get the following model:

tvvp=1,135363-0,10547ted for 2009, 2012-2014;

tvvp=0,614574+0,633237ted for 2005-2008, 2010, 2011.

Compare the results of simulation with actual data:

*Table 4.4. Comparison of modelling results influence the rate of qualification growth (ted) on the GDP growth (tvvp)*

| | Table data1 | | | |
|---|---|---|---|---|
| | 1<br>ted | 2<br>twp | 3<br>Predict. | 4<br>Residual |
| 1 | 1,049 | 1,279 | 1,28 | 0,00 |
| 2 | 1,019 | 1,233 | 1,26 | -0,03 |
| 3 | 1,000 | 1,325 | 1,25 | 0,08 |
| 4 | 1,077 | 1,315 | 1,30 | 0,02 |
| 5 | 0,955 | 0,963 | 1,03 | -0,07 |
| 6 | 0,923 | 1,185 | 1,20 | -0,01 |
| 7 | 1,037 | 1,216 | 1,27 | -0,05 |
| 8 | 1,039 | 1,070 | 1,03 | 0,04 |
| 9 | 0,961 | 1,033 | 1,03 | -0,00 |
| 10 | 0,823 | 1,077 | 1,05 | 0,03 |

Analyze the remains for autocorrelation and heteroscedastic absence (conditions of the Gauss-Markov theorem):

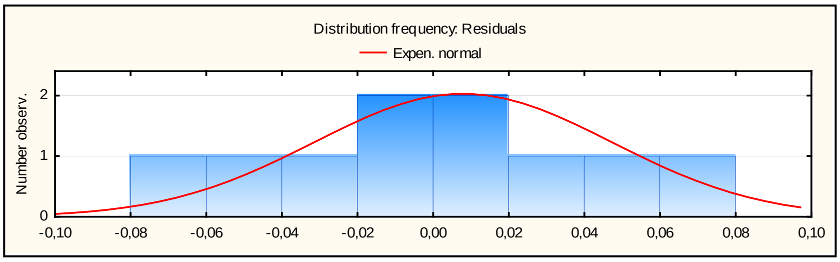



*Fig. 4.3. Diagram of the normal distribution of the residuals of the model.*

Often instead of this diagram it is applied the graphic of normal distribution of the residuals (the normality of the distribution indicates the location of the points of deviations, similar to normal stochastic line):

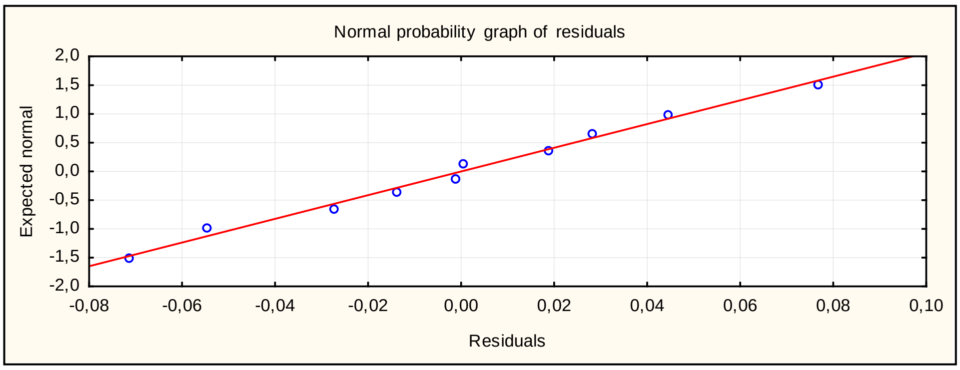



*Fig. 4.4. Graph of normality of the distribution of the residuals of the model.*

Based on the graph we make a conclusion about the absence of heteroscedastic and autocorrelation.The value of the coefficient of correlation R=0,93512 is close to 1, so we speak about high adequacy of the constructed model. Under this model, you can conduct the forecast: for example, if in 2015 the number of people who successfully qualify will decrease by 5% compared with the year 2014 (taking into account the trend of recent years), i.e., $ted_{2015}$= 0.95, then by constructed model $tvvp_{2015}$= 1,135363-0,10547 ($ted_{2015}$= 0.95) = 1,0351665, that is, Gdp (nominal) increase in 2015 compared with 2014 year by about 3.5%.

Quite often the information base of the regression model is represented by dynamic series that leads to some difficulties caused by the dependence of levels and their autocorrelation. The presence of the latter violates one of the requirements (independence of observations) of regression analysis and it can disfigure the simulation results. In practice, use different ways to eliminate autocorrelation. The simplest is a method of numerical transformation, when instead of initial levels of interconnected dynamic series$y_t$,$x_t$finite differences are used ( $\Delta_y = y_t - y_{t-1}$, $\Delta_x = x_t - x_{t-1}$and so on), or a method of deviations from the trend, when the initial levels $y_t$,$x_t$ are replaced by deviations from the trend $d_y = y_t - f(t)$; $d_x = x_t - f(t)$. Putting the time factor $t$in the regression equation $Y = f(x,t)$ also contributes to the elimination of autocorrelation. Load on the variable $t$ depends on the complex of factors, included in the model. The sense of the parameters of such a model we consider on the example of relationship of dynamics of oil imnport $y_t$ and the price for a barrel of oil $x_t$ at the world market. For example, the volume of imports of crude oil in the country systematically diminished, due to both changing prices, and internal factors. The relationship between these parameters, you can present as a linear function

$$Y = a + bx + ct,$$

where $b$ is an average increase of resulting index $Y$ per unit of increment of factor index $x$; $c$ is average annual increase of $Y$ under the influence of changes of unidentified factors that uniformly changing in time.

*Table 4.5. Analysis of the dynamics of indicators*

| Serial the number of the year | Imports of oil, $y_t$, million. barrels | The price for the 1 barrel, $x_t$,. | Model | $e_t = y_t - Y_t$ |
|---|---|---|---|---|
| 1 | 1749 | 13.48 | 1808 | – 59 |
| 2 | 1702 | 14.76 | 1743 | -41 |
| 3 | 1769 | 18.92 | 1653 | 116 |
| 4 | 1600 | 22.97 | 1562 | 38 |
| 5 | 1431 | 30.29 | 1442 | – 11 |
| 6 | 1325 | 34.66 | 1349 | 24 |
| 7 | 1302 | 30.77 | 1332 | -30 |
| 8 | 1341 | 29.36 | 1292 | 49 |
| 9 | 1232 | 28.07 | 1251 | -19 |
| 10 | 1180 | 26.40 | 1213 | – 33 |
| 11 | 1162 | 27.79 | 1147 | 15 |
| Total | 15793 | x | 15793 | 0 |

The model of imported oil (model is constructed in Statistica) is described by equation:

$$Y = 1984{,}340 - 2{,}497x - 52{,}986t$$

$$(27.97)\ (-2.50)\ (-6.99).$$

In the brackets there are values of $t$–criterion, and they are exceeding critical $t_{0,95}(8)=2.31$, so we can consider as significant the influence of each factor on the volume of imports with probability 0.95. In accordance with the values of the regression coefficients theincrease of the price of one barrel of oil on 1 dollar reduces the import of crude oil in the country by about 2.5 million barrels. Due to other factors (e.g., energy), the import of oil decreases annually by roughly 53 million barrels.The value of the determination coefficient $R^2$=0.951 and trajectory criterion $F(28)$ = 77.48 attest to the adequacy of the model.

So, if you have a linear trend in the series, then the variable of time is included into the model

$$Y = a_0 + \sum_{1}^{m} b_i x_i + ct,$$

where $b_i$ is a net effect of influence of factor $i$ on $Y$; $c$ is the effect of unidentified factors that form the trend of dynamic series.

Note that the dynamic model can display not only a trend but also more complex components of series, such as intermittent or seasonal fluctuations, the discontinuity of process (intervention), etc.

The peculiarity of the regression analysis of the dynamic series is the estimation of autocorrelation of residual values $e_t = y_t - Y_t$. If autocorrelation is significant, then factors included in a model not entirely represent the mechanism of the process, model is considered as inadequate. Checking the significance of autocorrelation can be carried out on the basis of the circular coefficient of the first order $r_1$. The software tools to test autocorrelation significanceoften use Durbin-Watson criterion, whose characteristic $D$ is functionally associated with $r_1$:

$$D = \frac{\sum_{2}^{n}(e_t - e_{t+1})}{\sum_{1}^{n} e_t^2} = 2\,(1 - r_1)\,.$$

If there is no autocorrelation between adjacent members of series than the value of $D$ is approximately 2, with high positive autocorrelation $D$ approaching to 0, and at high negative autocorrelation to -4. The critical limits of its values are defined as folloing: lower $D_L$ and upper $D_U$, based on it, the hypothesis of no autocorrelation (primary or the null hypothesis $H_0$: $r_1 = 0$) is accepted or rejected. When checking the hypothesis there are three possible conclusions:

- $D > D_U$–autocorrelation is absent;
- $D < D_L$-a hypothesis of absence of autocorrelation is rejected;
- $D_L \leq D \leq D_U$ -the conclusion is uncertain.

The critical limits of $D$ depend on the number $n$ of members in a series and the number $m$ of model parameters. The negative autocorrelation checking is carried out on the basis of values (4-$D$). In the *MultipleRegression* module to check the significance of autocorrelation of residual values in the *ResidualAnalysis*it is provided the option of *Durbin-Watson stat*. In our example, $D$=1.831 that lie in the interval of possible values of the hypotheses $h_0$: $r_1 = 0$, and therefore, the autocorrelation significance is not proven. A similar conclusion gives test of hypotheses using cyclic autocorrelation coefficient whose value $r_1 = 0.085$ significantly less than critical $r_{0.95}(11) = 0.353$. Absence of autocorrelation of the residuals confirms the adequacy of the model.

A characteristic feature of the formation mechanism of variation and the dynamics of socio-economic indicators is the delay of influence of factors when cause and effect ripped apart in time (e.g., investment). Time lags are conditioned by the duration of the production cycle, the inertia of the processes, feedback, etc. For the evaluation of the effects of delay of factor $i$ influence in the model it is introduced lag variable $x_{i.p.}$. The factors that have two or more lag (*timephased lag*) are introduced in the model as units of corresponding variables. General view of the model with distributed lags:

$$Y = a_{00} + \sum_{i=1}^{m} \sum_{p=0}^{k} b_{i,p} x_{i,p}\,,$$

where $p$ = 0, 1, ..., $k$ lags; $m$ is number of factors included in the model.

Theoretically you can generalize model with distributed lags to any number of factors, but the practical implementation of such a model stumbles on the difficulties caused by the limited nature of the dynamic series and the complexity of their inner structure. As a rule, in the model are included such lag variables, for which the lags are grounded theoretically and tested empirically. The tool for determining the lag is an intercorrelation function, which is a set of coefficients of correlation between the rows of $x_i$ and $y$, biased relative to each other on the lag *p*. With increasing of lag intercorrelation function extinct.

Because of the limited nature of dynamic series of socio-economic phenomena, all the features of the development process cannot be taken into account in the model. To expand the information base of the model, the combination of spatial and dynamic series is applied. For example, it is described the dependence of $Y = f(x_1, x_2, x_3)$ according to the data of 10 sites for five years. There are different options to use such mixed static-dynamic (panel) information. Look at the two of them.

1. *Dinamization of spatial models*. For example, for each *t*year from 5 it is determined the static model $Y_t = f(x_1, x_2, x_3)$. Regression coefficients of static models form dynamic series. If the effect of influence of the $i$ factor varies in time, such a change will be trend series $b_i$. By trend extrapolation method you can determine the expected effect of influence during the bias *v*. At the same time the forecasted level of the factor $x_{i,t+v}$ is determined. The combination of these forecasts gives the forecast function *Y*:

$$Y_{t+v} = a_{0,\,t+v} + b_{1,\,t+v}\,x_{1,\,t+v} + b_{2,\,t+v}\,x_{2,\,t+v} + b_{3,\,t+v}\,x_{3,\,t+v}.$$

Prediction of the effects of influence of factors and their levels can be done in any way if the functional view of the forecast model is grounded (of course, while maintaining the sufficiency of information for both dynamic and spatial series).

2. *Model of object-periods*. Often the spatial and dynamic rows are combined into one information set, the unit of this set is the object-period. For example, for 10 objects and five years, we have $10 \cdot 5 = 50$ object-periods. This approach to combining spatial and dynamic series greatly expanding the information base of the model, at the same time gives it special properties. The main feature of static-dynamic information (Panel data) is dependence of observations. In this model not only the levels of the dynamic series are dependent, but all the series as well (spatial and time) as belonging of the levels to this or that series is fixed. So, the relationship between dynamic series is the result of spatial variation, which because of the inertial processes is maintained for some time. The dependence of the spatial series reflects the synchronicity of the dynamics of indicators of separate objects, occasioned by common conditions of development. Ignoring these features of informational database modeling leads to false conclusions.

The peculiarities of the spatial variations are taken into account in the model using a structuralvariables of individual objects $u_j$. The trend that is inherent to all objects of function *Y* is filtered using a variable of time *t*. However, due to the uneven development of individual objects of set, along with the common trend individual trendscan be substantial. For filtering you can use variables in *dynamic interaction*: for factors - $x_{it}$, for objects - $u_{jt}$. Taking into account all these features, the regression model for the set of objects-periods can be written as:

$$Y = a_0 + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m} c_i x_i t + \sum_{j=1}^{n-1} a_j u_j + \sum_{j=1}^{n-1} d_j u_j t + ft,$$

where the parameters of the model $b_i$ is neteliminated from interactions within the model effect of the impact factor $x_i$; $c_i$ is change effects the impact of $b_i$ in time; $a_j$ - the difference between the functions of the *j*-th object and in the whole population; $d_j$ - change these differences in time; *f* - common to all objects of a population trend - the impact of unidentified factors in the model; $a_0$ - free member of the equation. For each *j*-th object the free member equation is equal to the sum of $(a_0 + a_j)$; unlike $a_0$, the sum has an economic meaning, it measures the impact of factors that determine the specificity of this object.

Therefore, model of objects-periods includes two options: one of them presents the assessment of the effects of influence factors and their change in time, the second is the features of the population, specifics of development of individual objects. It is possible to avoid the model overloading and save the maximum information to evaluate options we use, for example, the algorithm of step by step regression analysis. In recent years, more and more often in such cases, the methods and models of regression in latent structure (main components) or regression of the partial least squares (PLS-R, PCA-R, PLS-PM methods) are applied.

Simulation of causal systems (System structural regressions).

The complexity and versatility of interconnections, reverse effects imply a need for the use of models in the form of a system of interdependent (simultaneous) equations. There are two types of such systems. Some systems of equations describe a serial chain of causation, which makes it possible to solve them serially. Other systems have regenerative connections, when one variable at the same time acts and as a cause, and as a consequence. In this case, we have to solve the equations simultaneously. The construction scheme of any system of equations includes forming a logical framework models and specification of equations. Logical framework of model can be represented geometrically as *the connectivity graph* or as an *adjacency matrix*.

An important stage in the model specification is the division of variables that form the structure of the models into endogenous and exogenous. *Endogenous* (internal or related) variables due to the internal structure of the process and are the subject of analysis. Their number is the number of equations and identities of the model. The independent variables (external) that are included in the model are called *exogenous*. They cause changes in the system of relationships without the reverse influence. Classification of variables by the endogenous and exogenous rather conditional and depends on the nature and essence of the phenomenon, which is studied, and the purpose of the study. In dynamic models *lag* variables appear. For example, the interconnection of all types of variables we will present by the following system of equations:

$$Y_{1,t} = f_1(\, y_{3,t},\, y_{2,t-1},\, x_{1,t})$$

$$Y_{2,t} = f_2(\, y_{1,t},\, y_{3,t-1},\, x_{2,t})$$

$$Y_{3,t} = f_3(\, y_{2,t}, y_{1,t-1},\, x_{3,t}\,).$$

The number of equations of the systemis equal to the number of endogenous variables ($y_{1,t}$,$y_{2,t}$, $y_{3,t}$). They are interdependent, and each of them separately exposed by independent (exogenous) variables ($x_{1,\ t}$, $x_{2,\ t}$, $x_{3,\ t}$) and endogenous with delay ($y_{2,\ t-\ 1}$,$y_{3,\ t\text{-}1}$,$y_{1,\ t\text{-}1}$). Lag variables have the same properties as exogeneous so they are combined into one class *defined by predetermined* variables $z_j$. So, the separate $i$ equation of the system can be written as:

$$y_i = Y_i\, a_i + Z_j\, b_j + e_i,$$

where $Y_i$ is vector of endogenous variables of $i$-th equation ($i$ = 1, 2, ..., $k_i$); $a_i$ are coefficients of the endogenous variables in $i$-th equation; $Z_j$ is vector of exogenous and lag variables in $i$-th equation ($j$ =1, 2, ..., $m_i$); $b_j$ are the coefficients of the variables $z_j$ in $i$-th equation. The model reflects the structure of correlations between variables and so it is called *structural*. Since the same endogenous variables are included into various equations of the model, this leads to the dependence of the residuals of the endogenous variables, which complicates the estimation of parameters of the model by classic OLS. To exclude correlation, we need to transform the structural model and lead it to shorter, *normalized* form. In the present form all endogenous variables are expressed solely by predetermined (exogenous and lag) variables: $y_i = Z_j\ r_j + v_i$, where $r_j$ are coefficients of the reduced form in the variable $z_j$ that are estimated by OLS; $v_i$ is a residual.

The problem of estimation of parameters of the model and the possibility of its transformation related to the concept of *model identification*. A model is called identificated, if the equation of structural form clearly describes the interconnection. The condition of the identification is verified for each $i$-equation by criterion: $(k_i - 1) \leq (m - m_i)$. In the identified equation the difference between the total number of exogenous and lag variables in all system $m$ and the number of such variables in the $i$-th equation $m_i$ is one unit greater than the number of endogenous variables this equation $k_i$. Every equation of identified system reflects a system of mutual interconnections, not duplicates and cannot be substituted by any combination of other equations.

When $(m\text{-}m_i) > (k_i - 1)$, estimation of model parameters cannot be identified unambiguously, the system is considered to be *overidentificated*. If the $i$-th equation $(m\text{-}m_i) \leq (k_i - 1)$, such a system is *nonidentificated*, and it is impossible to define the parameters by statistical methods. Let’s check the identification of the above system of equations in which the total number of exogeneous and lag variables is equal to $m$ = 6. The first equation contains two endogeneous and two predetermined variables $z_j$, so $k_i - 1 = 2\text{-}1 = 1$, and $m\text{-}m_i = 6 - 2 = 4$, and so the equation is overidentified. Similarly, we can see that the second and third equation is overidentified.

Estimation of parameters of overidentificated equations is carried out by *two steps* OLS. In step 1 the parameters of the given system of equations and theoretical values of endogenous variables shall be determined; on the step 2 in the system of structural equations the values of endogenous variables shall be replaced with values calculated within the present system for the linear model:

1. $\hat{y}_i = \sum_{1}^{m_i} r_j z_j$ ;

2. $Y_i = \sum_{1}^{k_i} a_i \hat{y}_i + \sum_{1}^{m_i} b_j x_j$ .

The equations that are constructed in this way are considered as ordinary regression equations, their parameters are determined by the OLS and are used in forecasting. If the specification of the model does not meet the requirements of mathematical apparatus and empiric data, and the system of equations is nonidentificated, there are three possible ways of model modification:

- exclude from the model some endogenous variables;
- include into the model some additional exogenous or lag variables;
- replace some set of interrelated variables by multidimensionalestimations, for example, the main components.

In *Statistica* methods of structural simulation are implemented in modules *Sepath*. Powerful mathematics apparate of the module is effective in building simple econometric models. We consider the system of equations presented by ambiguous causal chains. These systems are called recurrent (recursive). *Recurrent* model is a systems ofequations, in which the dependent variable of *i*level acts as an independent in the equation $i + 1$. For example, the feature set of model contains $m$-1 endogenous variables $Y_i$ and exogenous variable $Z_i$. Then the system of equations looks like:

$Y_1 = b_{10} + a_1 Z_1$

$Y_2 = b_{20} + b_{21} Y_1 + a_2 Z_2$

$Y_3 = b_{30} + b_{31} Y_1 + b_{32} Y_2 + a_3 Z_3$

$Y_4 = b_{40} + b_{41} Y_1 + b_{42} Y_2 + b_{43} Y_3 + a_4 Z_4$

……………………………………………………………………….

$Y_m = b_{m0} + b_{m1} Y_1 + b_{m2} Y_2 + b_{m3} Y_3 + b_{m4} Y_4 + ... + b_{m-1} Y_{m-1} + a_m Z_m$.

The calculation of the coefficients of regression recurrent models is carried out consistently for each equation by classic OLS. The matrix of regression coefficients is triangular, i.e.

$$\begin{vmatrix} -1 & 0 & 0 & 0 & ... & 0 \\ b_{21} & -1 & 0 & 0 & ... & 0 \\ b_{31} & b_{32} & -1 & 0 & ... & 0 \\ b_{41} & b_{42} & b_{43} & -1 & ... & 0 \\ ... & ... & ... & ... & ... & ... \\ b_{m1} & b_{m2} & b_{m3} & b_{m4} & ... & -1 \end{vmatrix}.$$

Based on the principles of sequential decomposition process for reasons of its formation, the recurrent model allows the assessment of *the full influence effect* for each of them. The latter is the sum of *direct* and *indirect* influences.

## 4.2. The projection on the latent structure. PLS-PM/PCA-PM methods. Logit-regression of MakFadden

PLS-PM (Partial Least Squares Path Modeling or Projecton Latent Structures Path Modeling) methods started to be widely used in a variety of applied researches of the 70-s years of the twentieth century because of the work of Herman Vold and his son Sven Wold, in this works the basic principles of the technique of modeling PLS–PM were established. PLS-PM is a tool for modeling of interconnections between latent variables. The method PLS-PM OL is used for analyse of high dimensional data in the conditions of ill-structured environment. It is widely applied, in particular, in the economy to evaluate such latent indexes as a utility, level of economic development, etc.

The modeling problem (by means of PLS-PM) can be writted as follows: let X is a nxp data matrix:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix},$$

where n is the number of observations (objects), and p is the number of variables (characteristics).

Matrix X can be divided into J blocks: $X_1$, $X_2$, ... ,$X_j$, each of which is associated with the latent variable $LV_j$, whose evaluation is denoted by:

$$\widehat{LV_j} = Y_j .$$

All interconnections between variables can be divided into two types: interconnections between latent variables and blocks $X_j$ (these interconnections form the outer model) and interconnections between the latent variables with each other (these interconnections form the internal model).

Internal model can be written in the form of a system of linear equations:

$$LV_j = \beta_0 + \sum_{i\to j}\beta_{ji}LV_i + in\_error_j,$$

where $LV_i$ is all latent variables that affect the laten tvariable $LV_j$; coefficients $\beta_{ji}$ are structural (enroute) coefficients that characterize the intencity and the direction of connection between latent variables $LV_i$ and $LV_j$; $\beta_0$ – free member; $in\_error_j$ – a random deviation of the internal model. The internal model imposed by the following requirements:

1. The system of latent equations should be recursive so it has to correspond to loop-free graph.
2. Internal model is recursive, so the following requirement is true:

$$E(LV_j|LV_i) = \beta_{0i} + \sum_{i\to j}\beta_{ji}LV_i.$$

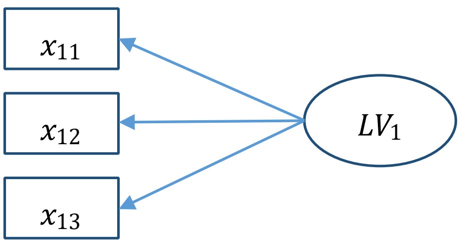
3. Random deviations are independent of the explanatory variables, that is, the following requirement is satisfied:

$$cov(LV_j, error_j) = 0.$$

Thus, at the simulation by using the method of projection on the latent structure the requirements for the statistical distributions of variables and random deviations are absent.

The external model describes two type of interconnection between latent and explicit variables: reflective and formative.

Reflective type is the most widespread type of external models in which the latent variable is the cause of implicit variables, that is, explicit variables reflect latent. The graph of reflective external models looks like:

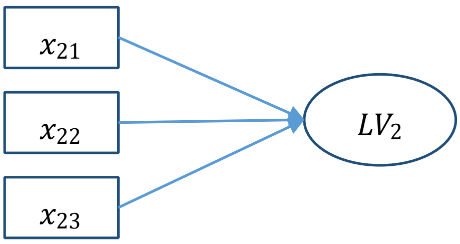



The external model can be written in the form of a system of linear equations:

$$X_{jk} = \lambda_{0jk} + \lambda_{jk}LV_j + out\_error_{jk},$$

where $\lambda_{jk}$ are load coefficients, $\lambda_{0jk}$ are free members, $out\_error_{jk}$ is random deviation of external model.

Formative type is a kind of external model when explicit variables are "cause” of latent variables, i.e. they «form» latent variables. The graph of formative external models type looks like:




External formative model can be represented in the form of a system of linear equations:

$$LV_j = \lambda_{0j} + \lambda_{jk} X_{jk} + out\_error_j,$$

where $\lambda_{jk}$ are lead coefficients, $\lambda_{0j}$ are free members, $out\_error_j$ is random deviation of external model.

Latent variables can not be measured directly. Therefore we enter the notion of evaluation of latent variable $\widehat{LV_J}$, which is a linear combination of the corresponding implicit variables:

$$\widehat{LV_j} = Y_j = \sum_k w_{jk} X_{jk},$$

where $w_{jk}$ – external weight (weight multipliers) of a model.

Thus, variables $LV_j$ and $Y_j$ characterize one and the same entity, but are used in different purposes: the first is for theoretical, the other is for practical.

PLS-PM modelling can be represented as a three-steps algorithm:

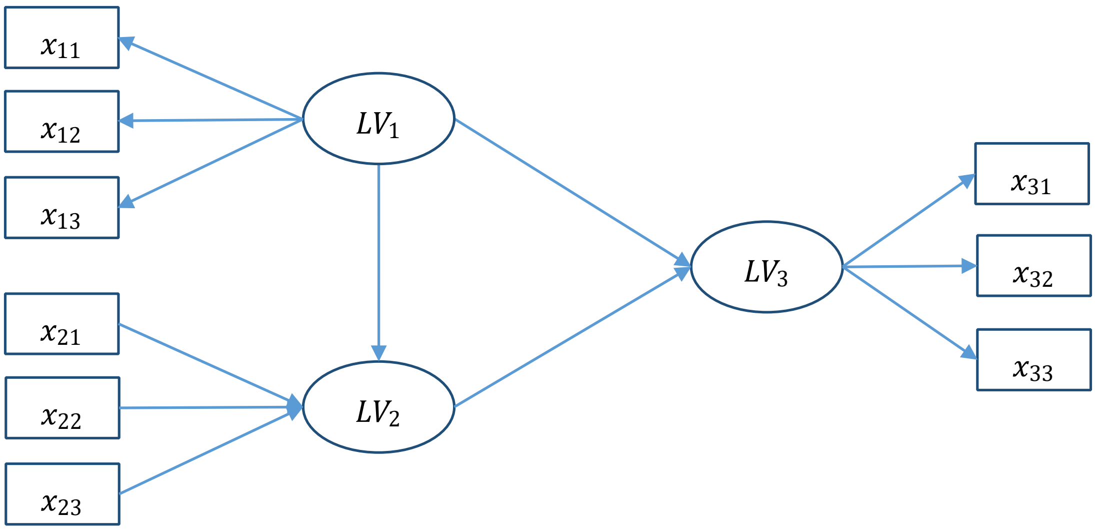
a) calculating of external weights $w_{jk}$ to obtain estimates of the latent variables;

b) computation of structural coefficients $\beta_{ji}$ of internal model;

c) calculation of loads of $\lambda_{jk}$ of external model.

The first step is the main and the most complicate step of PLS-PM metodology, that is the iteration, its purpose is to calculate the values of external measures and, as a result, ratings $Y_j$. The result of this step is that abstract latent variables materialize. The second stepis a solution of a system of linear equations by the method of least squares, the third is calculation of correlations.

Consider a simple model with three latent variables and corresponding blocks:




Each latent variable is associated with three explicit variables, so the external model can be represented as a system of equations:

$$X_{1k}= \lambda_{01k}+ \lambda_{1k}LV_1 + out\_error_{1k},\ k= 1, 2,3$$
$$X_{2k}= \lambda_{02k}+ \lambda_{2k}LV_2 + out\_error_{2k},\ k= 1, 2,3$$
$$LV_3= \lambda_{03}+ \sum_k \lambda_{jk} X_{3k}+ out\_error_{3k},\ k= 1, 2, 3$$

Internal model can be written as a system of equations:

$$\begin{cases} LV_2 = \beta_{02} + \beta_{21}LV_1 + in\_error_1 \\ LV_3 = \beta_{03} + \beta_{31}LV_1 + \beta_{32}LV_2 + in\_error_2 \end{cases}$$

The first phase of the simulation PLS-PM is an iteration process for obtaining the values of the assessments of the latent variables.

First the initial values of external weighting coefficients $w_{jk}$ are set.

Step1: Calculate the values of the latent variables $Y_j$ by means of external model.

Step2: Calculate the internal weights $e_{ij}$.

Step3: Calculate the values of the latent variables $Z_j$ based on internal models.

Step4: Calculate the new values of external weights $w_{jk}$.

Next, steps1-4 need to be repeated untill the necessary convergence of external weights.

Iterative process starts with setting some values of external weights. Suppose for simplicity that all the weights are equal to 1:

$$\tilde{w}_1 = (\tilde{w}_{11} = 1,\ \tilde{w}_{12} = 1,\ \tilde{w}_{13} = 1)$$
$$\tilde{w}_2 = (\tilde{w}_{21} = 1,\ \tilde{w}_{22} = 1,\ \tilde{w}_{23} = 1)$$
$$\tilde{w}_3 = (\tilde{w}_{31} = 1,\ \tilde{w}_{32} = 1,\ \tilde{w}_{33} = 1)$$

Here swung dash indicates that the values of weights are conditional. Now pass to the step 1 of simulation and compute the value of latent variables by the formula:

$$Y_k \propto X_k \widetilde{w}_k, \qquad k = 1, 2, 3$$

We will get the following system for the demonstrated model:

$$Y_1 \propto 1X_{11} + 1X_{12} + 1X_{13}$$
$$Y_2 \propto 1X_{21} + 1X_{22} + 1X_{23}$$
$$Y_3 \propto 1X_{31} + 1X_{32} + 1X_{33}$$

Sign $\propto$ we use instead of an equal sign (some authors indicate by this sign the normalization of variables), because in this case variables Y depend on the variables X, but the right and left parts of the expression are not equal, regardless of the approximate value of the external weights.

On step 2, we pas to the internal models with the purpose of calculation of values estimates of the latent variables, but in another way than on the first step: using internal, rather than external model. That is, instead of calculating ratings $Y_j$ as linear combinations of implicit variables, you must calculate the value of assessments of latent variables $Z_j$ as linear combinationsof assesements of other latent variables associated with $LV_j$ given by:

$$Z_j = \sum_{i \leftrightarrow j} e_{ij} Y_i,$$

where the double arrow indicates that assessments of $Y_i$ are summarized but only of latent variables $LV_i$ that are related to j-th latent variable. The values $e_{ij}$ are called internal weights and calculated by the formula:

$$e_{ji} = \begin{Bmatrix} cor(Y_j, Y_i), & i \leftrightarrow j \\ 0, & i \not\leftrightarrow j \end{Bmatrix}$$

Thus, the internal weights characterize the direction and strength of the interconnection between latent variables. Getting the value of the internal weights, you can calculate the estimation of latent variables based on internal model:

$$Z_1 = \sum_{i \leftrightarrow 1} e_{i1} Y_i = e_{21} Y_2 + e_{31} Y_3$$
$$Z_2 = \sum_{i \leftrightarrow 2} e_{i2} Y_i = e_{12} Y_1 + e_{32} Y_3$$
$$Z_3 = \sum_{i \leftrightarrow 3} e_{i3} Y_i = e_{13} Y_1 + e_{23} Y_2$$

Step 4 is the recalculation of the external weights values. For reflective type of external model we apply formula:

$$\widetilde{w}_{jk} = (Y_j' Y_j)^{-1} Y_j' X_{jk},$$

and for the formative type the following formula:

$$\widetilde{w}_j = (X_j' X_j)^{-1} X_j' Y_j.$$

One each iteration ($S$= 1, 2, 3 ...) the degree of convergence of external weights is evaluated, i.e. external weight on the iteration $S$ compares with external weights on the iteration $S-1$. The convergence degree is considered sufficient if:

$$|w_{jkS-1} - w_{jkS}| < 10^{-5}.$$

The second step of simulation is calculation of structural coefficients $\widehat{\beta_{ji}} = B_{ji}$ that are calculated using the OLS for multiple regression $Y_j$ on $Y_i$ by the formula:

$$Y_j = \sum_{i \to j} \widehat{\beta_{ji}}\, Y_i \quad .$$

Structural coefficients for OLS are:

$$B_{ji} = (Y_i'Y_i)^{-1} Y_i'Y_j \quad .$$

The third step is calculation of loads $\lambda_{jk}$. For simplification the loads are calculated as the ratios of correlations between latent and explicit variables:

$$\widehat{\lambda_{jk}} = cor(X_{jk}, Y_j) \quad .$$

The ultimate goal of PLS-PM modelling is to get assesements for latent variables to implement further forecasting procedures.

*Example 4.2. Building PLS-PM for the model of sustainable development of tourism.*

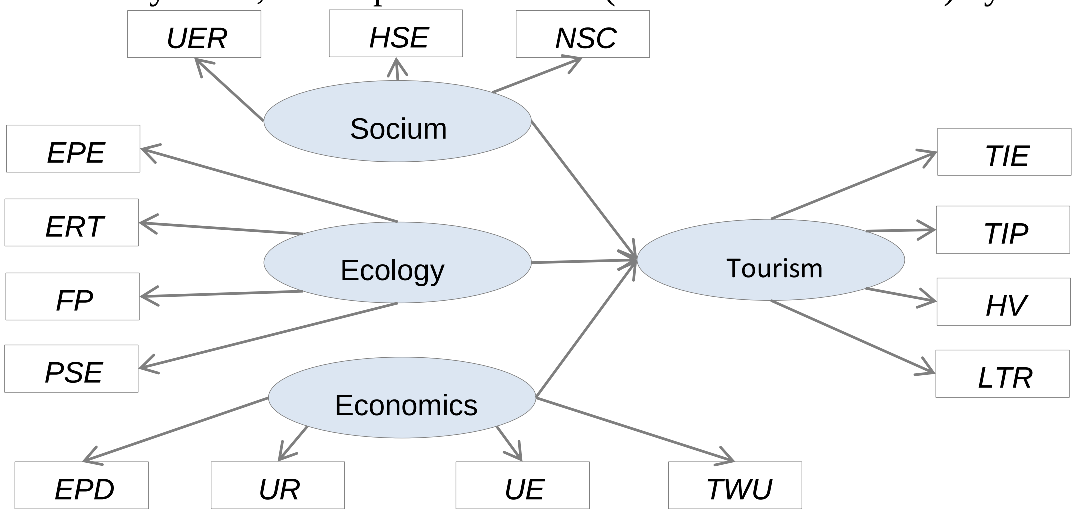
The following table presents indicators of sustainable tourism development, offered by Greenwood J.B. for the Plimsoll model for the region with developed tourism sector:

*Table4.6. Indicators of sustainable tourism development*

| *Indicator* | *Description of the indicator* | *The notation in the model* |
|---|---|---|
| *The number of jobs in the travel industry* | *The total number of all jobs in the travel industry for the calendar year* | *TIE* |
| *Stock of paid in the travel industry* | *Stock of paid in the travel industry by calendar year* | *TIP* |
| *Cost of housing* | *The average cost of real estate (area no more than 10 acre and in the Houseon one family)* | *HV* |
| *Tax revenues from the travel industry* | *Total tax revenues from the travel industry for kalendar year* | *LTR* |

| | | |
|---|---|---|
| *Number of childrens receive subsidies for treatment* | *Number of childrens receive subsidies for treatment, for the year* | *NSC* |
| *State cost on social services for the population* | *The sum of state cost for health and other social services for a year* | *HSE* |
| *The unemployment rate, %* | *The proportion of the total number of unemployed persons to the economically active population* | *UER* |
| *State cost on the protection of the environment* | *The sum of state cost for the protection of the environmentby kalendar year* | *EPE* |
| *The share of jobs in the travel industry, %* | *The share of jobs in the travel industry of the total volume of jobs* | *ERT* |
| *Number of poor families* | *The number of families with income less than subsistence operation* | *FP* |
| *State cost on the protection of public safety* | *The amount of the state cost on the protection of public safety (Paulandcand, service, rescue, etc.)* | *PSE* |
| *State cost on economic development* | *The amount of the state cost on planning and economic development (increase of employment of the population, and more)* | *EPD* |
| *State cost for utilities* | *The amount of the cash flows from the population for utility services* | *UR* |
| *Costs of komunal services* | *The amount of the state cost for Komunal servants populated* | *UE* |
| *The total volume of water consumed* | *With an average daily volume of consumed water of the population from all sources available for the calendar year* | *TWU* |

The original model is represented in the following figure (latent variables are denoted by ovals, and explicit variables (indicatorsfrom a table) by rectangulars.




The relationship of the latent variables is the internal model, and the relationships between latent and explicit variables represent an external model. Let's write these models in analytical form.

Internal model:

$$LV_{tourism} = \beta_0 + \beta_1 LV_{social} + \beta_2 LV_{ecologic} + \beta_3 LV_{economic} + error_{tourism},$$

where$LV_{tourism}$, $LV_{social}$, $LV_{ecologic}$, $LV_{economic}$ are latent variables $\beta_1, \beta_2, \beta_3$ are structural coefficients, which characterize the strength and direction of the interconnection between latent variables, $\beta_0$ is free member, $\mathrm{error}_{\mathrm{tourism}}$ is residual.

The external model looks like the following system:

$$\begin{cases} X_{TIE} = \lambda_{0TIE} + \lambda_{1TIE} LV_{tourism} + error_{TIE} \\ X_{TIP} = \lambda_{0TIP} + \lambda_{1TIP} LV_{tourism} + error_{TIP} \\ \dots \\ X_{UER} = \lambda_{UER} + \lambda_{1UER} LV_{social} + error_{UER} \\ X_{HSE} = \lambda_{0HSE} + \lambda_{1HSE} LV_{social} + error_{HSE} \\ \dots \\ X_{EPD} = \lambda_{0EPD} + \lambda_{1EPD} LV_{economic} + error_{EPD} \\ X_{UR} = \lambda_{0UR} + \lambda_{1UR} LV_{economic} + error_{UR} \\ \dots \\ X_{EPE} = \lambda_{0EPE} + \lambda_{1EPE} LV_{ecologic} + error_{EPE} \\ X_{ERT} = \lambda_{0ERT} + \lambda_{1ERT} LV_{ecologic} + error_{ERT} \end{cases},$$

where $\mathrm{X}_{\mathrm{TIE}}$, $\mathrm{X}_{\mathrm{TIP}}$, ...,$\mathrm{X}_{\mathrm{ERT}}$ are explicit variables, $\lambda_{\mathrm{1TIE}}$, $\lambda_{\mathrm{1TIP}}$, ...,$\lambda_{\mathrm{1ERT}}$ are coefficients of loads, $\lambda_{\mathrm{0TIE}}$, $\lambda_{\mathrm{0TIP}}$,...,$\lambda_{\mathrm{0ERT}}$ are free members, $\mathrm{error}_{\mathrm{TIE}}$ $\mathrm{error}_{\mathrm{TIP}}$,, ...,$\mathrm{error}_{\mathrm{ERT}}$ are residuals. Note that at structural coefficients simulating by means of the partial least squares (PLS) we use the estimations of latent variables:

$$\widehat{LV}_j = Y_j = \sum_k w_{jk} X_{jk} \quad .$$

For our model, these estimations can be presented in the form of the following system of equations:

$$\begin{cases} \widehat{LV}_{tourism} = Y_{tourism} = w_{TIE}X_{TIE} + w_{TIP}X_{TIP} + w_{HV}X_{HV} + w_{LTR}X_{LTR} \\ \widehat{LV}_{social} = Y_{social} = w_{UER}X_{UER} + w_{HSE}X_{HSE} + w_{NSC}X_{NSC} \\ \widehat{LV}_{ecologic} = Y_{ecologic} = w_{EPE}X_{EPE} + w_{ERT}X_{ERT} + w_{FP}X_{FP} + w_{PSE}X_{PSE} \\ \widehat{LV}_{economic} = Y_{economic} = w_{EPD}X_{EPD} + w_{UR}X_{UR} + w_{UE}X_{UE} + w_{TWU}X_{TWU}, \end{cases}$$

where $w_{TIE}$, $w_{TIP}$, ...,$w_{TWU}$ are external weights of the model.

Stages of PLS-PM simulation that arerelated to optimization model and calculation of all parameters, are implemented in the software environment R (for the analysis the following tools are used: package for component analyse - Plsdepot; package for PLS analysis – Plspm; package for Microsoft Excel files - Excel link.). For the model review it is performed a component analysis using the package Plsdepot. On the figure there arevisualized explicit variables correlations with two first principal components:

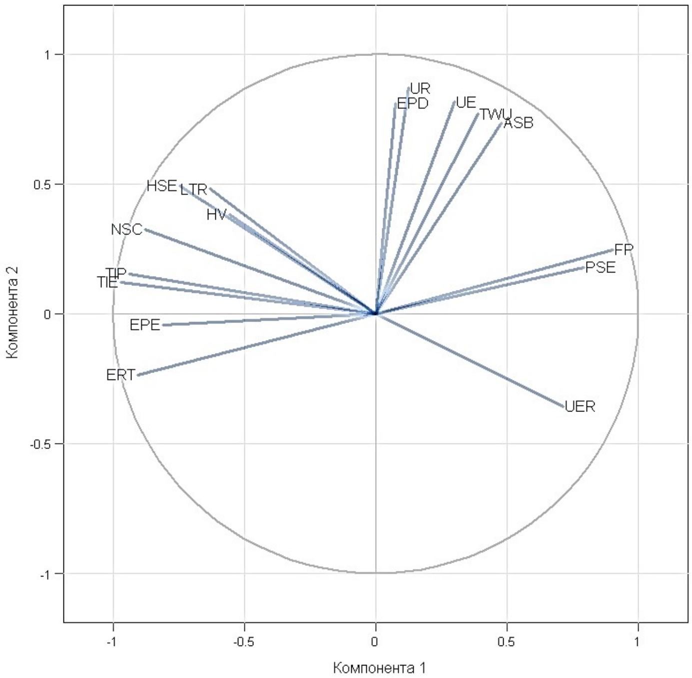



*Fig. 4.5. Correlation diagram in two principal componets space*

Component analysis confirms the correctness of the variables division on sectors: tourism-socium-ecology-economics).

Analyse of the model by Plspm includes the following items: a) verification of the internal consistency of the blocks; b) verification of significance of variables in the external model; c) verification of cross correlations of variables in block with latent variables with variables of other blocks; g) verification of the internal model; d) verification of the model qualification by index of conformity of model to data; e) optimization of a model.

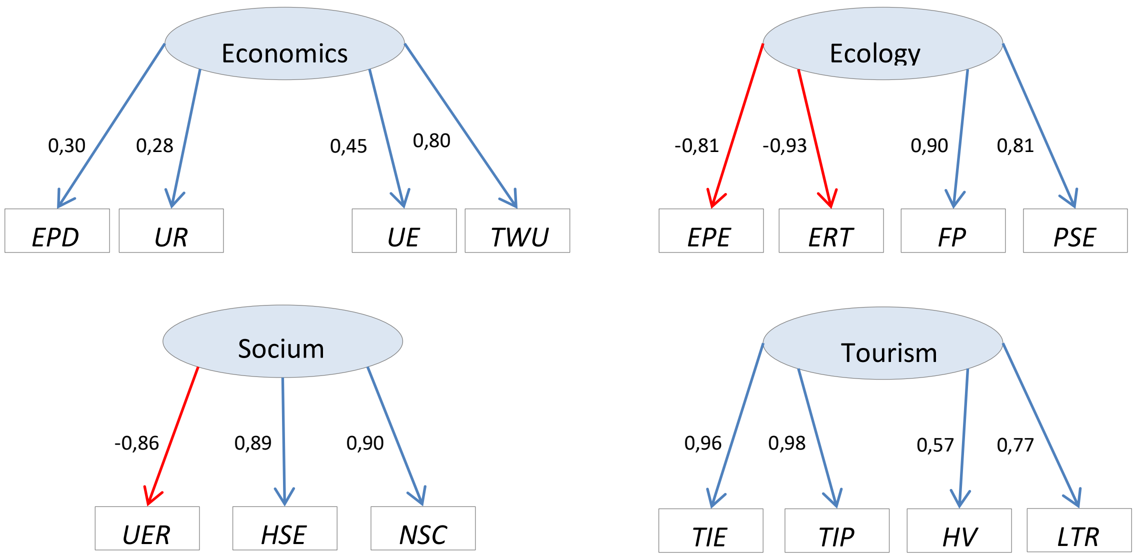
a) to check internal consistency in blocks in the Plspm the following criteria are used: Kronbah coefficient α; coefficient $\rho$ of Dillan-Goldstein; eigenvalue matrix of explicit variables correlation.

*Tables 4.7.Checking internal consistency in blocks*

| Block | Alpha Kronbah, $\alpha_K$ | Po Dillon-Goldstein, $\rho_{DG}$ | Eigenvalue, $\lambda_1$ | Eigenvalue, $\lambda_2$ |
|---|---|---|---|---|
| Economics | 0.88 | 0.91 | 3.41 | 0.72 |
| Ecology | 0.00 | 0.00 | 2.98 | 0.73 |
| Socium | 0.00 | 0.58 | 2.37 | 0.38 |
| Tourism | 0.85 | 0.90 | 2.84 | 0.70 |

The table shows that two unit ("Economics" and"Tourism") have very high values of coefficients $\alpha_K$ and $\rho_{DG}$ and blocks of "ecology" and "socium" characterized by poor internal consistency ($\alpha_K$< 0.7 and $\rho_{DG}$< 0).

On the figure there are correlation coefficients of explicit and latent variables (by blocks):




In the «Ecology" block two variables (*ERT* and *EPE*) have a negative correlation with latent variable. In a "Socium" also there is a variable with negative correlation (*UER*). This leads to internal inconsistencies in blocks. After modification of variables (*FP*, *PSE*, *UER*) we obtain the index of cooperation of external model, are presented in the following table:

*Table 4.8. Cooperation indexes of the external model*

| Block | Alpha Kronbah, $\alpha_K$ | Po Dillon-Goldstein, $\rho_{DG}$ | Eigenvalue, $\lambda_1$ | Eigenvalue, $\lambda_2$ |
|---|---|---|---|---|
| Economy | 0.88 | 0.91 | 3.41 | 0.72 |
| Ecology | 0.88 | 0.94 | 1.78 | 0.22 |
| Socium | 0.85 | 0.93 | 1.74 | 0.26 |
| Tourism | 0.85 | 0.90 | 2.84 | 0.70 |

All three of the cooperation criterions satisfy the necessary criteria ($\alpha_K > 0.7$, $\rho_{DG} > 0.7$, , $\lambda_1 > 1$, $\lambda_2 < 1$) after modifications. Figure 4.6 presents the correlation schemes:

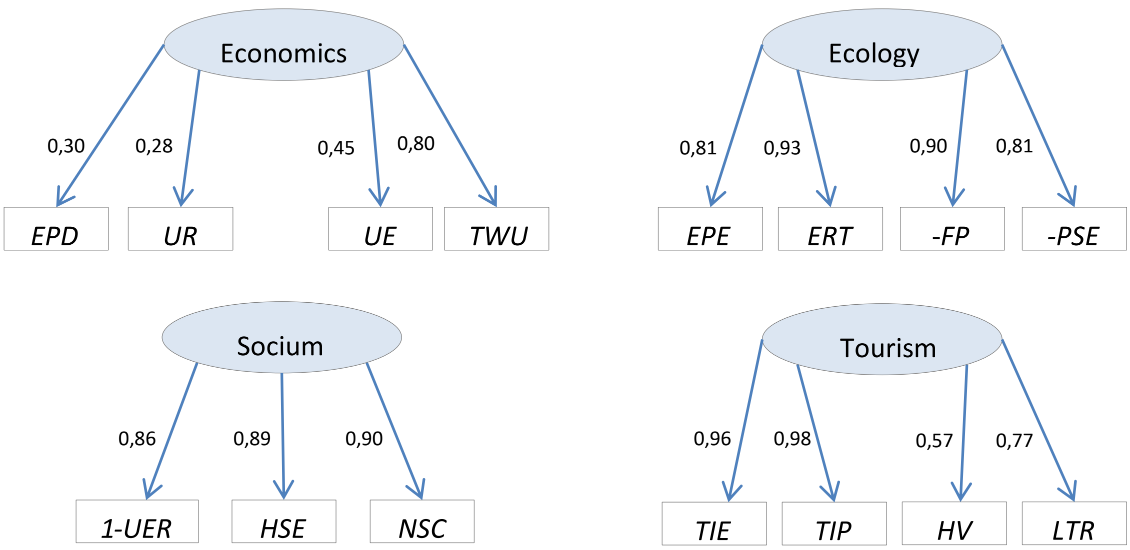



*Fig. 4.6. Correlation connections.*

The second phase is verification of the external model:

*Table 4.9. The value of the coefficients of the external model*

| Block | Variable | External weight $w_j$ | Load, $\lambda_{1j}$ |
|---|---|---|---|
| Economy | UR | 0.53 | 0.28 |
| | TWU | -0.34 | 0.84 |
| | UE | -0.45 | 0.45 |
| | EPD | 0.36 | 0.31 |
| Ecology | ERT | 0.34 | 0.93 |
| | EPE | 0.30 | 0.81 |
| | -FP | 0.29 | 0.90 |
| | -PSE | 0.23 | 0.81 |
| Socium | HSE | 0.34 | 0.90 |
| | NSC | 0.42 | 0.91 |
| | 1-UER | 0.36 | 0.86 |
| Tourism | TIP | 0.38 | 0.98 |
| | LTR | 0.23 | 0.78 |
| | HV | 0.12 | 0.57 |
| | TIE | 0.40 | 0.96 |

Variables are considered significant if the load factor $\lambda_{1j}$>0.7. By the results of the model validation it is necessary to eliminate variables *UR*, *UE*, *EPD*, *HV*. After exclusion of nonsignificant variables we get the following test results of external model:

*Table 4.10. The value of the coefficients of the external model after modifications*

| Block | Variable | External weight $w_j$ | Load, $\lambda_{1j}$ |
|---|---|---|---|
| Economy | TWU | 0.39 | 0.91 |
| Ecology | ERT | 0.33 | 0.93 |
| | EPE | 0.30 | 0.81 |
| | -FP | 0.29 | 0.90 |
| | -PSE | 0.23 | 0.81 |
| Socium | HSE | 0.34 | 0.90 |
| | NSC | 0.42 | 0.91 |
| | 1-UER | 0.36 | 0.86 |
| Tourism | TIP | 0.41 | 0.99 |
| | LTR | 0.22 | 0.77 |
| | TIE | 0.44 | 0.97 |

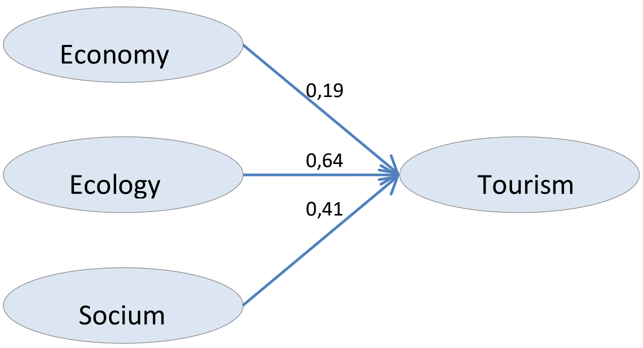
After checking the load factors $\lambda_{1j}$, that is, after checking the interconnection of explicit variables with latent variable of the corresponding block, you need to check the cross-loading, i.e. to determine the strength of the connection between explicit variables and latent variables of other blocks. This will exclude such indicators whose connection with latent variable of another block is greater than with latent variable of the corresponding block:

*Table 4.11. The value of the coefficients of cross load*

| Block | Variable | Block | | | |
|---|---|---|---|---|---|
| | | Economy | Ecology | Socium | Tourism |
| Economy | TWU | 0.91 | -0.50 | -0.09 | -0.13 |
| Ecology | ERT | -0,49 | 0.93 | 0.58 | 0.81 |
| | EPE | -0.33 | 0.81 | 0.58 | 0.73 |
| | -FP | -0.63 | 0.90 | 0.65 | 0.72 |
| | -PSE | -0.48 | 0.81 | 0.64 | 0.56 |
| Socium | HSE | -0.04 | 0.58 | 0.90 | 0.69 |
| | NSC | -0.11 | 0.75 | 0.91 | 0.84 |
| | 1-UER | -0.08 | 0.52 | 0.86 | 0.72 |
| Tourism | TIP | -0.24 | 0.84 | 0.80 | 0.99 |
| | LTR | 0.09 | 0.43 | 0.68 | 0.77 |
| | TIE | -0.29 | 0.89 | 0.85 | 0.97 |

The table shows that the strength of the connection of all explicit variables with latent variables of corresponding blocks is greater than with latent variables of other blocks, that is, all variables are "loyal" to their blocks.

The fourth step is to check the quality of the internal model. This figure is a graphic representation of the internal model with the specified values of structural coefficients:




In the table are estimates $\beta_i$ of the equation of structural models, as well as the results of the criteriont-statistic:

*Table 4.12. Estimation of the parameters of a structural model*

| | Estimator$\beta_i$ | Standarderror, $SE_{\beta i}$ | t-statistics | Pr(> \|$t$\|) |
|---|---|---|---|---|
| Freemember | 2,1628E-17 | 0.085354054 | 2,53403E-16 | 1 |
| Economy | 0.19028279 | 0.119057102 | 1.598248144 | 0.1249264 |
| Ecology | 0.64313833 | 0.167063135 | 3.849672335 | 0.00093012 |
| Socium | 0.41436382 | 0.139606488 | 2.968084329 | 0.00733725 |

Criterion of t-statistics is held for blocks “Ecology” and “Socium” (Pr (> |$t$|) < 0.05), but not to block “Economics” (Pr (> |$t$|) > 0,05). On the stage of optimization this block can be excluded from the model.

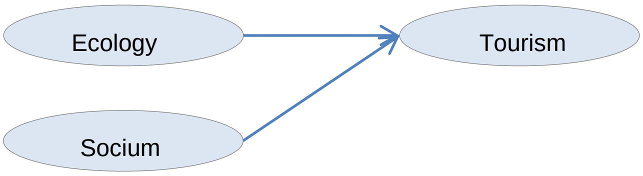
*Table 4.13.Total statistic of internal model*

| Block | $R2$ | Ratio explanation variation, BC | Ratio rejecting var., AVE |
|---|---|---|---|
| Economy | 0.00 | 0.89 | 0.89 |
| Ecology | 0.00 | 0.74 | 0.74 |
| Socium | 0.00 | 0.79 | 0.79 |
| Tourism | 0.85 | 0.84 | 0.84 |

Showing $R^2$ for target block "tourism" more then 80%. The proportion of variation characterizes share variations of a block, which reproduced latent variable this block. For the all blocks and in this figure far exceeds 50%, that positively characterises the model. In the last column of a given indicator, which characterizes the dispersion excluded share (average particle dispersion indicators unit that explains the latent variable dispersion in the block that contains the error of measurement). Fashion Nick AVE for the all blocks exceeds 50%, therefore, this criterion of the internal model is also considered satisfactory.

The fifth stage is the calculation of the single factor of quality conformity model data-GoF (Goodness-of-Fit). Coefficient characterizes the quality as an internal model of the system, and external, and serves as an indicator of the reliability of the forecast model (predictive reliability model is considered to be high, if the coefficient of GoF > 70%). Our model coefficient GoF=82%.

The sixth stage is the optimization of the model of all criteria as a model are performed, in addition to the criteriont-statistics for the block "Economics" the model therefore exclude this block and perform all stages of checking the updated model, internal (structural), part of which will be as follows:




The first stage – validation of internal consistency in blocks:

*Table 4.14. Checking internal consistency in blocks (optimization)*

| Block | AlphaKronbaha, $\alpha K$ | Ro Dandllona-Goldstein,$\rho DG$ | Their own value, $\lambda 1$ | Their own value, $\lambda 2$ |
|---|---|---|---|---|
| Environmentalist and I | 0.88 | 0.92 | 2.98 | 0.73 |
| Social and UM | 0.87 | 0.92 | 2.37 | 0.38 |
| Tourism | 0.90 | 0.94 | 2.53 | 0.45 |

On Fig. 4.7 presents correlation scheme after shutdown D block:

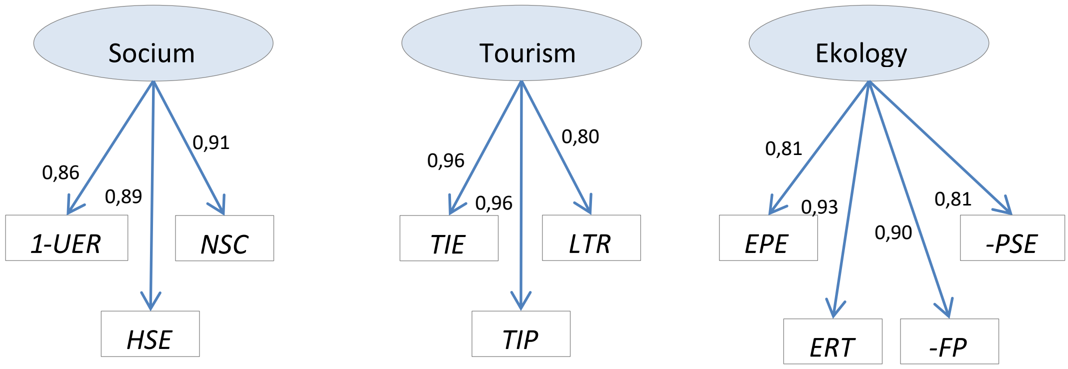



*Fig. 4.7. Coefficients of loads of variables in the model (after Optimization)*

Second stage – checking the value of variables in the model:

*Table 4.15. The importance declines sharply coefficients of external models (after Optimizition)*

| *Block* | *Variable* | *The outer weight, wj* | *Onloading, λ1j* |
|---|---|---|---|
| *Ekology and I* | *ERT* | *0.33* | *0.93* |
| | *EPE* | *0.30* | *0.81* |
| | *-FP* | *0.29* | *0.90* |
| | *-PSE* | *0.23* | *0.81* |
| *Social and UM* | *HSE* | *0.34* | *0.90* |
| | *NSC* | *0.42* | *0.91* |
| | *1-UER* | *0.37* | *0.86* |
| *Tourism* | *TIP* | *0.39* | *0.98* |
| | *LTR* | *0.27* | *0.80* |
| | *TIE* | *0.42* | *0.96* |

The third stage – checking the cross-correlations of variables with variable block other units:

*Table 4.16. The importance declines sharply coefficients of cross load (after Optimizition)*

| Block | Variable | Block | | |
|---|---|---|---|---|
| | | Ekology and I | Social and UM | Tourism |
| Ekology and I | ERT | 0.93 | 0.58 | 0.80 |
| | EPE | 0.81 | 0.58 | 0.73 |
| | -FP | 0.90 | 0.65 | 0.71 |
| | -PSE | 0.81 | 0.64 | 0.55 |
| Social and UM | HSE | 0.58 | 0.90 | 0.69 |
| | NSC | 0.75 | 0.91 | 0.84 |
| | 1-UER | 0.52 | 0.86 | 0.73 |
| Tourism | TIP | 0.84 | 0.80 | 0.98 |
| | LTR | 0.43 | 0.68 | 0.80 |
| | TIE | 0.89 | 0.85 | 0.96 |

IV etap –Validating internal models:

*Table 4.17. The statistics of the internal model (after Optimizition)*

| | DandNCA$\beta i$ | Standarderror, $SE\beta i$ | t-statistics | Pr(>\|t\|) |
|---|---|---|---|---|
| Freea member of the | 1,7726E-17 | 0.090808067 | 1,95203E-16 | 1 |
| «Ekology and I " | 0.433665664 | 0.127880991 | 3.391165947 | 0.002626084 |
| "Social and UM» | 0.54539304 | 0.127880991 | 4.264848383 | 0.000315956 |

*Table 4.18. The total statistics of internal model (after Optimizition)*

| Block | Type of unit | $R2$ | The proportion of variation, BC | Share remote dispers.,AVE |
|---|---|---|---|---|
| Environmentalist and I | Ekzogenous | 0 | 0.743174071 | 0.743174071 |
| Social and UM | Ekzogenous | 0 | 0.787916914 | 0.787916914 |
| Tourism | Endogenous | 0.818585688 | 0.84015712 | 0.84015712 |

V etap – check the quality of the model index matching model data: new model coefficient GoF=86%. Therefore, all the necessary conditions for the new model.

Analyze the results. Internal model has the form equation:

$LVtourism = 0{,}55 LVsocial + 0{,}43 LVecologic + errortourism.$

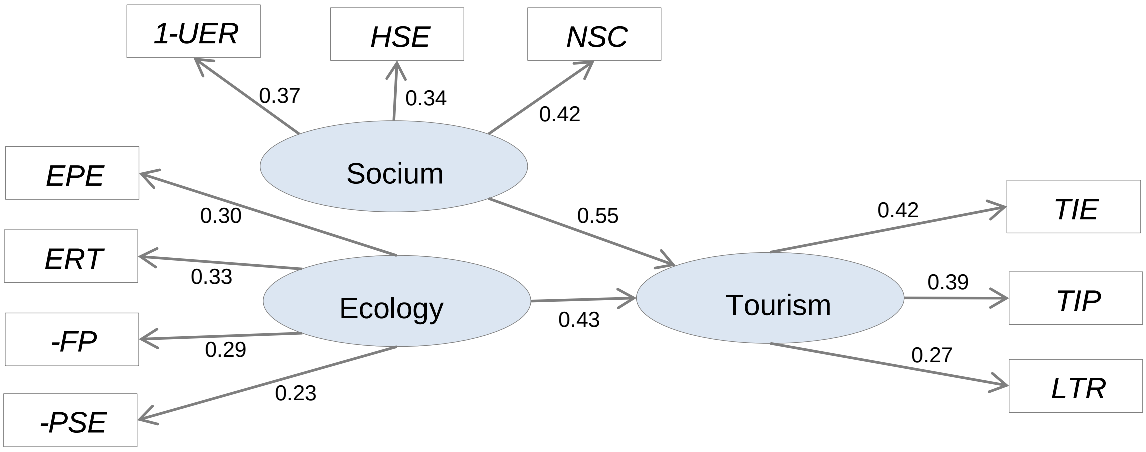
Estimations of latent variables represent in the form of a system of equations:

$L\check{V}tourism = 0{,}42 XTIE + 0{,}39 XTIP + 0{,}27 XLTR$

$L\check{V}social = 0{,}37(1 - XUER) + 0{,}34 XHSE + 0{,}42 XNSC$

$L\check{V}ecologic = 0{,}3 XEPE + 0{,}33 XERT - 0{,}29 XFP - 0{,}23 XPSE$

The obtained results can be represented as a graph, which indicated a structural coefficients for internal models and external weight for the external model:




Essential effect on the level of development of tourism in the region with developed tourist sector two characteristics: «The level of social comfortin» and «The level of ready stand community to engage the protection of the environment» the force of the impact 0,55 and 0,43 in accordance. In addition, research has revealed enough weak impact of the level of economic development, tourism development can suggests that sustainable tourism development is possible as in the economically developed region and is developing. Level of social comfort is possible by three factors:

- the level of employment of the population (the proportion of people who have a steady job, a total number of economically active population);

- the amount of thepublic spending on health care and other social services;

- the number of children receiving subsidies for treatment.

The level of the formation of the system of environmental values in society – the 4 factors:

- the amount of the Government spending on the protection of the environment;

- the share of jobs in the tourist industry in the total volume of population;

- the number of families with income below the subsistence minimum (reverse dependency);

- the amount of the Government spending on the protection of public safety (reverse dependency).

D and NCA the level of development of tourism depends on three factors:

- the number of jobs in the travel industry;

- Foundationsalary travel industry;

- the amount of the tax revenues from the travel industry.

On the model you can count latent indexes for each year and create a stimulant to predict future values. Sustainability of development can be assessed by comparing the values of estimates for different periods. For example, if the estimation of the latent variables/units (economy, ecology, Society) current period previous period ratings are smaller, then the corresponding block is considered to be sustainable. Or for every unit you can define a baseline with which to conduct the comparison.

Logìt-regression McFadden.

In 2000, the Nobel Prize in economic sciences were awarded achievement scholars James Heckman and Daniel McFadden (Daniel L. McFadden), which made a significant contribution to the development of theory and methods for analyzing discrete choice. Hekman has methodological problems of formation of statistical sampling, and McFadden has developed a model of discrete choice. The work of McFadden devoted to microeconomic analysis, demographic processes demand, methods of modelling, theory of choice among his publications Note: structural analysis of discrete models and econometric applications (Structural analysis of discrete data with econometric applications, including Monkey). Preferences, uncertainty and optimality (preferences, uncertainty and optimality). A guide to econometrics (Handbook of econometrics, together with Engle) and others. They created the econometric methods to assess production technologies and research the factors that underlie the demand of companies for capital and labour. His main achievement is considered the development of economic theory and econometric models methods for analyzing discrete choice. In 1974, has developed the so-called logìt model, which is applied to predict the effectiveness of construction of transport lines in California based on data analysis for choosing locals way of commuting. In 90th years of the extended methodology ìmìtacìjnimi models, allowing base research on more common principles. As a result of the simulation of discrete choice has greater realism and accuracy.

Often reflect discrete choice, that is, the choice of a finite set of alternatives (for example, information about classes of individuals, their permanent residence, a method of moving or travelling, etc.) in the economic theory of the traditional analysis of demand predict that individual choice is described by continuous variables, but this interpretation does not correspond to the practice of research povedìnidentification of discrete choice. Methodology of discrete choice McFadden comes from analysis of conditional logìt, the essence of which is that in the life of each person appear specific alternatives: X-characteristics associated with each alternative and Y-characteristics of individuals that the researcher can observe by using real data for the study, for example, the choice of ways to travel, where al′ternativami can be a car, bus or metro, X-characteristics can cover time and expenses, and Y-characteristics include age, income and education, but the differences between the individual and the alternatives other than between X and Y. "Despite the fact that they are not notable to the researcher, they determine individual maximum useful. These characteristics of the presented random vectors errors ".

The standard logìt model utility is a linear function of the properties of the alternatives: $u_{ij} = \sum_{i=1}^{k} x_{ij}\beta_{ij} + \varepsilon_{ij}$, $x_{ij} - k \times 1$ vector, which contains the characteristics of consumeriand alternative j, $\beta_{ij} - k \times 1$ the vector of parameters and variables $\varepsilon_{ij}$, $j = 1, 2, ..., J$, provides casual and such that are independent of the standard distributions of extreme values, cumulative distribution function which is equal to $F(\varepsilon_{ij}) = \exp\left(-\exp(-\varepsilon_{ij})\right)$ and distribution density function: $\mathrm{f}(\varepsilon_{ij}) = \exp(-\varepsilon_{ij}) \cdot \exp\left(-\exp(-\varepsilon_{ij})\right)$. McFadden suggested that random errors have a certain statistical distribution among the population, which he called "extreme importance”. In these conditions, he showed that the probability that a personiwill alternatives j, equal to $p_{ij} = \dfrac{\exp\left(\sum_{i=1}^{k} x_{ij}\beta_{ij}\right)}{\sum_{j=1}^{J}\left(\exp\left(\sum_{i=1}^{k} x_{ij}\beta_{ij}\right)\right)}$ - the probability that the consumeriselects product jall. Consequently, it is likely that the product jis purchased on the market. This value on the market with a large number of consumer product market equals j.

The logìt model is used CDF logistics distribution:
$p(x)=1-\exp(-x^{Tb})/\left(1+\exp(-x^{Tb})\right)=\dfrac{\exp(x^{Tb})}{1+\exp(x^{Tb})}$ . Also, when calculating the coefficient of the used determìnation McFadden:

$$McFadden\ R^2=1-\frac{\ln L_1}{\ln L_0}\ w_1=\frac{1}{n}\sum_i\left(y_i-\widehat{y_i}\right)^2,$$

$$\Pr edicten\ R^2=1-\frac{w_1}{w_0},\ \mathrm{w}_0=\begin{cases}1-\widehat{p_i},\ \widehat{p}>0,5\\ \widehat{p_i},\ \widehat{p}\le 0,5\end{cases},\ \widehat{p}=\frac{n_1}{n}.$$

Logìt model applied to transport in the demand for urban transport, to study the effectiveness of political action, social reform, changes in the environment, etc. For example, these models explain how changes in the prices of goods affect their availability, the demographic situation, the volume of traffic. They are used in studies of choice of housing, places of residence or education, demand for household energy, telephone services, provision of housing the elderly. McFadden concluded that the conditional logìt models have a particular property for the probability of choosing between two alternatives, for example, travel by bus or by train, regardless of price and quality of the other options.

# Chapter 5. Dynamical optimizations models

## 5.1. Arrow-Debreu's model (ADM). Applied models CGE (computable general equilibrium) and DCGE (dynamical CGE)

General equilibrium (GE) specify conditions tranquility, when each of economical agents operate with them utility function, and all markets are in equilibrium. The ADM is one of the most general model of competitive economy and is a crucial part of general equilibrium theory, as it can be used to prove the existence of general equilibrium (or Walrasian equilibrium) of an economy. Leon Walras proved the existence of general equilibrium when every agents in the economy operate with yours purposeful function and all markets are in the equilibrium. Walrasian equilibrium is the solution of the system of equations when number of independents variables is equal to number of the equations with condition perfect competition.

In 1954, McKenzie and the pair Arrow and Debreu (Nobel's laureates) proved the existence of general equilibria by invoking the Kakutani fixed-point theorem on the fixed points of a continuous function from a compact, convex set into itself. In the Arrow-Debreu approach, convexity is essential, because such fixed-point theorems are inapplicable to non-convex sets. Notes that the Kakutani theorem does not assert that there exists exactly one fixed point.

The assumption of precluded many applications, which were discussed in the Journal of Political Economy from 1959 to 1961 by Francis M.Bator, M.J.Farrel, Tjalling Koopmans and Thomas J.Rothenberg. In 1969, Ross M.Starr proved the existence of economic equilibria when some consumer preferences need not be convex. He proved (his proof used Shapley-Folkman theorem) that a "convexified" economy has general equilibria that are closely approximated by "quasi-equilibria" of the original economy. Compared to earlier models, the Arrow-Debreu model radically generalized the notion of a commodity, differentiating commodities by time and place of delivery. So, for example, "apples in Berlin in September" and "apples in Paris in June" are regarded as distinct commodities. The Arrow-Debreu model applies to economics with maximally complete markets, in which there exists a market for every time period and forward prices for every commodity at all time periods and in all places. This model specifies the conditions of perfectly competitive markets. In financial economics the term Arrow-Debreu is most commonly used with reference to an Arrow-Debreu security. A canonical Arrow-Debreu security is a security that pays one unit of numeraire if a particular state of the world is reached and zero otherwise (the price of such a security being a so-called "state price"). As such, any derivatives contract whose settlement value is a function on an underlying whose value is uncertain at contract date can be decomposed as linear combination of Arrow-Debreu securities. Since the work of Breeden and Lizenberger in 1978, a large number of researchers have used options to extract Arrow-Debreu prices for a variety of applications in financial economics.

A common economic equilibrium (CGE) is determined as the state of economy, at that every economic agent operates in accordance with the objective function, and all markets simultaneously are in equilibrium. Founder of CGE theory Leon Walras proved that a common equilibrium is consonant with the economic system in that at every market the condition of perfect competition is fair. In the Walras model a common equilibrium is determined as a result of solution of the system of equations on condition that market demand on goods is equal to the initial stock and producing (at equilibrium prices). Walras showed that such system of equations has a solution, if the number of independent equations coincides with the number of unknowns. But this condition is only necessary, but not sufficient for existence of solutions of the system. Nobel laureates in the economy Arrow, K.J., and of G. Debreu proved the fundamental possibility of existence of CGE, applying the mathematical theorem about fixed point.

The neoclassical CGE models are able to help to estimate effects from different meaningful changes in the parameters of economic politics or external conjuncture taking into account the real present structure of economy. Exactly the inertia of this structure allows to talk about fundamental possibility of application of CGE conception as an instrument of analysis of structural changes in the conditions of transitional economy.

Lets describe a simple mathematical model ADM. Let $N$ consumers and $S$ firms (fixed number but enough large number: that they cannot effect on the prices) effect on economy.

Suppose that our model economy has $K$ goods. Commodity bungles are now lists $x_i = \left(x_{i1}, ..., x_{ij}, ..., x_{ik}\right), x_{ij} \geq 0$ and a utility function assigns a number $U_i\left(X_i\right)$ to each such lists $x_i$. In the economy $K$ goods produces without public goods (really public goods easily can introduce in the model: must addition the Samuelson’s condition about equilibrium sum of the individual estimations utility of consumers from consumption public goods to these costs of the production). The consumer’s maximization problem can be stated in the following way

$$muximize\, U_i\left(x_i\right)$$

subject to the constraints

$$p_{i1}x_{i1} + ... + \mathrm{p}_{ij}\, x_{ij} + ... \mathrm{p}_{ik}\, x_{ik} = p \cdot x_i \leq M \in X.$$

Any consumer realize the problem of choice consumers lists $x_i = \left(x_{i1}, ..., x_{ij}, ..., x_{ik}\right), x_{ij} \geq 0$ , which quantities

Neoclassic models of CGE are able to estimate effects of significant changes in parameters of economic politics or external state of affairs taking into account the real structure of economics. Namely persistence of this structure allows to speak about fundamental possibility of application of CGE conception as an instrument for analyze of structure biases in conditions of transitional economics.

Now lets describe Arrow and Debreu CGE model. Suppose that in economics there are $N$ consumers and $s$ firms (fixed constant number), but quite enough number, such that firms and suppliers do not have significant influence on the prices. $K$ goods are produced , but no public goods (in fact it is easy to intake public goods into a model because only condition of Samuelson about the equality of sum of individual estimations of utility of consumers from consumption of public good of producers value need to be added).

Every consumer accomplish a mission of consumer basket selection $x_i = \left(x_{i1}, ..., x_{ij}, ..., x_{ik}\right), x_{ij} \geq 0$ from potential set $X_i$ which maximizes its utility function $U_i\left(X_i\right)$ that describes advantages of consumer ( its possible to make assumptions about the utility function on the ground of empirical data or ease of optimal solution finding).

The basic property of utility function in all cases is absence of external effects in consumption (i.e. when the choice of one consumer has no effect on another consumer choice). Besides, consumer has initial reserve of goods and production factors $w_i$. Consumers are also owners of all firms in economic, at that $a_{ij}$ is the part of firm's capital $j$ owned by consumer $i$.

It allows consumer to participate in distribution of firm's profit $\pi_j$ proportionally to its part in capital. The condition of full distribution of profit is equality $\sum_{i=1}^{N} \alpha_{ij} = 1$. Quite often Arrow and Debreu model is formulated with condition of existence of permanent return of scale in all economic sectors that finally leads to zero long-term profit of companies.

The problem of consumer in Arrow and Debreu model is formulated as following:

$$\begin{cases} U_i\left(x_i\right) \rightarrow \max\limits_{x_i} \\ x_i \in X_i \\ p \cdot x_i \leq p w_i + \sum_{j=1}^{S} \alpha_{ij} \pi_j \end{cases}, (*)$$

where $p$ is price of goods vector. The solution of this problem determines first of all the function of individual demand of consumer for each produced good.

The function of market requirement if the sum of individual demand functions of all consumers. It is often assumed that consumers are identical and it allows to describe marker supply and demand in more simple form by inclusion of representative agent. In the most simple Arrow and Debreu model the behavior of representative agent corresponds to process of joint decision making by consumers and state. It is suspected that firms solve a problem of choice of optimal vector $y_j \in Y_j$, where $Y_j$ is the potential production set, with the aim of maximization of their profits $\pi_j$. Elements of vector $y_j \in \left(y_{j1}, ..., y_{ji}, ..., y_{jK}\right)$ can be greater than zero and less than zero as well, because the meaning of $y_{ji}$ is the production output $i$ by firm $j$ with the deduction of its industrial consumption by the firm.

$$\begin{cases} \pi_j = p_j \cdot y_j \to \max\limits_{y_j} \\ y_j \in Y_j \end{cases} \quad (**)$$

Note that Arrow and Debreu model describes competitive behavior of economic agents that act based only on its own interest.

Back to Arrow and Debreu model lets formulate the condition of supply and demand balance of all goods in economics:

$$\sum_i w_i + \sum_j y_j \geq \sum_i x_i \quad \forall k \,. \; (***)$$

Usually supposed that the demand function is homogeneous of zero degree, and profit function is linearly homogeneous with respect to prices. As a result, same as in Walras model, the absolute level of prices has no influence on equilibrium of the model.

The set $\left(\overline{x}, \overline{y}\right)$ is called equilibrium state of economic system and prices $\overline{p}$ equilibrium prices if

1. At a price $\overline{p_i}$ the solution of the system (*) is $\overline{x_i} \quad \forall i$.
2. At prices vector $\overline{p}$ the solution of problem (**) is $\overline{y_j} \quad \forall j$.
3. Inequality (***) holds at $\overline{x_i}$ and $\overline{y_j}$. Or if (***) is leaded to well-known Walras law then $\left(\overline{p}, \sum_j \overline{y_j} + \sum_i w_i - \sum_i \overline{x_i}\right) = 0$.

The equilibrium in this model is optimal by Pareto if local non-saturation is expected.

Passing from theoretical formulation of a problem of CGE finding to applied analysis first of all was concerned with necessity to answer the question how the state have to make a practical choice of optimal economic policy. That's why two big classes of CGE models appear in practical applications.

1) *Analysis of tax*. This class of models is used for modeling of all taxes set that are the part of modern tax structure (income tax, corporate income tax, assessed tax, excise tax and others). As distinct from the real situation, often in models instead of specific tax rates equally equivalents are used. The main assignment of application of these models is the study of contorting effects that produce taxes and execution of optimal from the point of view of economic efficiency tax policy.
2) *International trade problem analyze.* As distinct from tax policy modeling this model class is less homogeneous by its structure because often it is based on different types of international trade theory. In some models

several countries or regions are considered in their commercial cooperation. Others represent itself models of one economics and the rest of the world is considered only as functions of demand on export of this country and supply of import. And some models are oriented only on analyze of international trade questions, for others the possibility of scenarist's modelling of fiscal policy is provided.

Applied CGE models allow to pass from theory to quantitative estimation of behavior of various variants of economic policy in concrete economics. Application of such detailed picture of economic as it is presented in national accounts system allows us not only to provide quantitative verification of theoretical hypothesis but also to analyze how the structure of concrete economic (usually quite different from neoclassic ideal) can have effect on consequences of economic policy.

Considerable progress in use of applied CGE models happen after appearance of programming language GAMS (General Algebraic Modeling System) which was worked out by the World Bank especially for economists that study quantitative analyze of economic policy. GAMS give possibility to researcher to choose the computational algorithm for the calculation of equilibrium state and aimed especially to economic problems solution.

Besides choice of reliable numerical methods of equilibrium economic states finding at change of parameters of economic policy, one more formal problem that appears at use of applied CGE models is problem of finding of parameters in models of functional forms. As a rule, calibration procedure is used in this case (consists in determination of estimations of parameters of model functions based on statistical data for reference period, usually one year).

Therefore these applied models are called Computable General Equilibrium or DCGE – Dynamic Computable General Equilibrium. By our site the process of applied CGE model construction has to consist of following steps:

1. *General conception.* Based on analyze of economic policy questions that are interested for investigator, and available statistic data to chose the necessary level of detalization of basic analyzed variables: regions, consumers producers, government.
2. *Behavior of economic agents*. It is necessary to determine the objective function for every economic agent. This function can be quite simple (as, for example, government that collect taxes and show demand on public goods and services), and quite complicate as well (as, for example consumer that distributes income among alternative goods in accordance with his preferences).
3. *Supply and demand functions*. We need to clarify the functional forms that describe the behavior of each economic agent in model (usually these forms are received in process of optimization problems solution from v.2).
4. *Determination of parameters*. On this stage the hypothesis about compatibility of observable data and behavior forecast is checked based on empirical data. The actual values of parameters of behavior functions are then determined.
5. *Programming of model and computer calculations.*

6. *Reproduction of outgoing equilibrium.* If parameters of production functions and utility functions were based on outgoing data then at outgoing values of exogenous variables the model has to represent outgoing equilibrium condition described by initial data.
7. *Model testing.* Tests consist of estimation of simple changes in economic policy (such as, for example, tax reform), and also may include diagnostic calculations (for example, determination of tax burden to provide comparative analysis of tax instruments efficiency).

In many countries of the world (USA, Denmark, Norway, Netherlands, France, Sweden, Australia, Brazil, Argentina, India, Thailand, Switzerland, Russian Federation and others) CGE models are applied by official government for estimation of consequences of changes in national economic policy. In Australia by means of ORANI model decisions about railway and power industry privatization, fall in level of state support, changes in labor-market control, taking measures in separate branches, in trade policy were made. Nowadays Center for Policy Studies (organization that support and develop CGE models ORANI, MONASH, MMRF, TERM) render services for developing of CGE models for some countries and conduct training courses for many developing countries. Center participated in CGE models developing for USA joint with International Trade Commission of USA (USAGE model). On the website of the center training model is presented (http://www.monash.edu.au/policy/minimal.htm).

CGE models were widely applied for estimation of consequences of creation of integration formations. For example, by means of KPMG, USITC the influence of NAFTA on USA economics were investigated (the main question was about influence of creation on employment because of influx of emigrants). In Europe CGE models are applied for estimation of consequences of changes in trade relationships between EU and other countries, consequences of expansion of EU, for modeling of some economic problems (for example, common agriculture politic) and etc. Lets mention about OECDTAX-model in which interconnection of countries were investigated through trade flows, flows of direct foreign investments and activity of transnational corporations.

Many CGE models (over recent years GLOBE model is considered as one of standards) realize within the framework of project GTAP (Global Trade Analysis Project), which is coordinated by Center of World Commerce Analysis at faculty of agriculture economic of Purdue University West-Lafayett city, Indiana, USA).

Lets describe the main principles of GLOBE model construction.

1. The model has to include blocks that describe economic processes in the investigated region and countries that are the main trade partners who have material effect on the world economy. All basic economic sectors have to be described in the model, such as households, business sector, state.
2. The behavior of each sector has to be described in explicit form. The sector of household for each region modeled based on the behavior of a representative consumer whose benefits are described by Stone-Geary utility

function that depends on the total consumption. Public consumption of households consists of consumption of domestic and imported goods, and is set in the model by using the constant elasticity of substitution (CES-function). The parameters of this function, as well as the level of existing income, determined by the matrix of social accounts, by volumes of trade, domestic and imported consumer goods, as well as by amounts of tax. The demand of households on domestic and imported goods is derived from solution of the problem. Correspondingly the demand for import for one country should be balanced by proposal of export of all other countries of the world. Government activity is modeled by means of the budget constraints of the State. The Government's demand on goods and services in the model is exogenous (fixed) percentage of total consumption. The public consumption is funded by various taxes: export tax (if it is in the country) and import tax (i.e., duties and tariffs), sales tax, value added tax, the tax on the factors of production, the tax on the household income, as well as a variety of indirect taxes. In the model total savings are defined as the sum of household savings (set as exogenous percentage of available household income), state savings and savings of other states (debts) that come to the state. Demand for investment is exogenous value defined from "capital" account of social accounts matrix (and in the model there is the mechanism of exogenous changes of this value). Manufacturing sector in the model consists of firms that operate in conditions of perfect competition, minimize costs and thereby maximize their profits. It is assumed that production function consists of several levels and has constant returns to scale. At the very top level comes CES-aggregating two "baskets": the basket of basic factors of production and the basket of intermediate goods. Each of these baskets in its turn is a unit. The basket of the main factors is the CES-unit from two types of labor (skilled and unskilled), as well as capital, land and natural resources. The parameters of the production function, related to the main factors of production, calibrated on data for each factor of social accounts matrix (with all taxes inclusive). Basket of intermediate goods is a unit consisting of energy, raw materials and other goods used in production. This unit is performed by applying the Leontiev function and based on a matrix of social accounts. For closing of the model the following steps need to be performed: 1) to fix the balance of capital (the exchange rate is considered to be flexible); 2) to fix saving of households; 3) to fix taxes (in accordance with the costs of the Government also fixed); 4) assume full employment and full mobility of all factors of production, except for unskilled labor for which the possibility of unemployment is anticipated; 5) offer factors of production are fixed (but the prices of factors are flexible and may change); 6) production technology is defined and fixed, productivity of factors remains the same.

3. For model of each country the main sectors (activities) should be allocated, for example: agriculture; coal production; oil production; gas production; mining operation; food industry; textile industry; chemical industry; metallurgy; other manufacturing industry; the production of electricity; trade; service industry, etc. Accounting industry structure is necessary because it allows you to track the impact of different events on certain industries/activities that, in the future, allows to develop and conduct the necessary corrective measures from the Government.

4. The model should be based on official data. As for the implementation of the model large arrays of data are necessary, the GTAP database may be used. This database is the most complete collection of data of the world economy and world trade. Its advantage is also that it contains data that is typically used in the models. The GTAP database in the last 8-th version covers 57 sectors 129 regions in the world and contains data on trade flows between the regions of economies (matrices of social accounts, taxes, duties, subsidies, tariffs, etc.).

5. It is better to construct the model by means of GAMS.

## 5.2. Application of CGE, DSGE models

*Example 5.1.* Static calculation model CGE for the Russian Federation (A.V. Alekseev A Diversified Computable General Equilibrium Model. / Working paper # WP/2007/217. – Moscow, CEMI Russian Academy of Sciences, 2007).

This model is a static CGE- model. It presents the 5 regions, where 3 agents and 15 branches are modeled in every region. Thus, the model is defined by the 15 economic agents. The first representative consumer has a utility function with a constant elasticity of substitution and chooses optimal consumer set provided the fulfillment of budget constraints.

The second type of agent-producer is modeled using a production function with constant feedback from the scale. It is assumed that all markets are markets with perfect competition, because the manufacturer made a condition of zero profits.

In the third type of agent the state is modeled implicitly. Namely, the only function performed by the State it is the collection of taxes and tariffs on exogenous rates and redirection of taxes and tariff revenues in the form of transfers from producers to consumers. Thus, the state in this model can be termed as the agent that has a specific objective function that provides the complete redistribution of tax and tariff revenue that maximizes its objective function. An example of such a function can be any smooth convex function of the transfers' value. It should be emphasized that more detailed modeling of the state is not necessary as the tools of influence on the economy, such as tax, tariff and non-tariff means of regulation, are exogenous parameters of this model.

This is the inter-regional model that is built for estimation of trade scenarios. So it is necessary to use the Armington's offer which means that the same goods produced in different regions, are treated as different products. Regarding to the consumer utility these products are aggregated into a single generalized product applying the appropriate elasticity replacement (Armington's elasticity). In the model the elasticity levels of Armington are applied for Russia, same as in the papers of Zemninskiy, Ballard, Faini, Piggot. The Armington's elasticity for all regions, except for the RF, is set at the level of 0.9. This value is widespread in the literature, but in order to understand how this affects the results, it was performed a test for model resistance in relation to the selection of Armington's elasticity levels, which showed that the results of the simulation are resistant towards the Armington's elasticity changes. Then the optimization problem for representative consumers is solved, the relative rates of consumers and producers are determined, the equilibrium conditions of the model are formulated and the existence and uniqueness of model equilibrium is proved.

*Endogenous and exogenous variables of the model.*

The most important exogenous variables in this model are the elasticity of substitution and elasticity of Armington. Other aspects of the consumption and production functions are endogenous and are calibrated in balance. In addition tariffs equivalent rates of non-tariff regulations, and tax rates are set as exogenous.

It is also necessary to note that at the stage of calibration the following parameters are exogenous: consumption, production, trade flows. Once settings of production and consumption functions are found, i.e. at the stage of scenarios testing, the optimal consumption, production and trade flows are calculated within the model.

*Demand of representative consumer.*

Consumers are represented by two levels of utility with a constant elasticity of substitution on each of the levels. The structure of demand in each region can be described as follows:

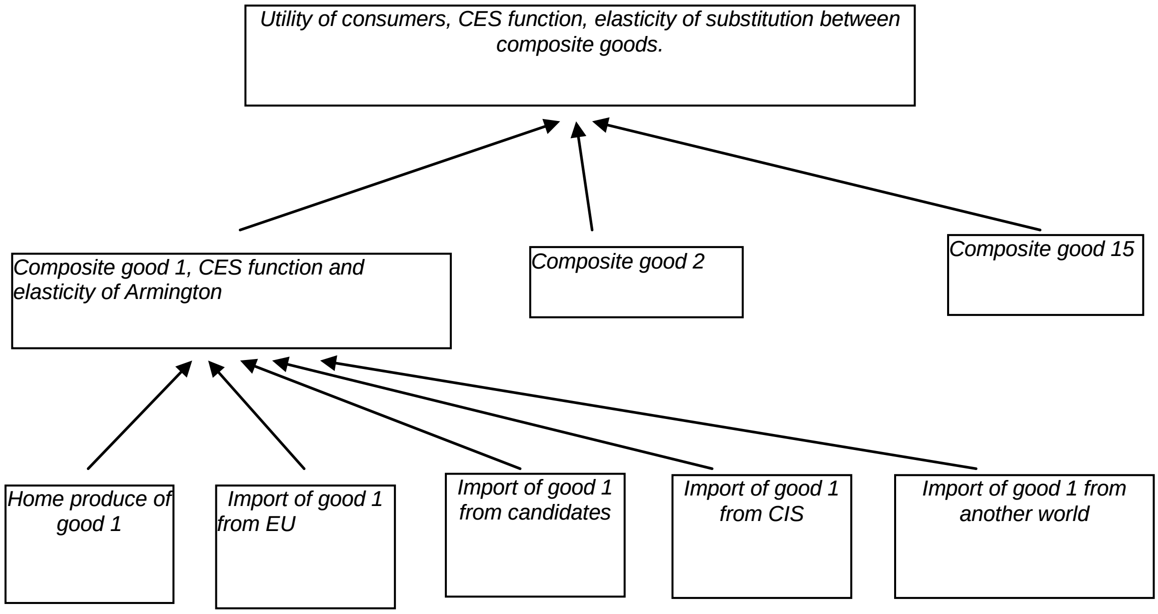



*Figure 5.1.* Demand of consumers.

On the second level of consumers utility the following more detailed function is applied:

$$U_j = \left[ \sum_{i=1}^{15} \gamma_{ij} C_{ij}^{\frac{\sigma_{ij}-1}{\sigma_{ij}}} \right]^{\frac{\sigma_{ij}}{\sigma_{ij}-1}}. \quad (1)$$

In this formula $U_j$ is the utility in the region $j$, that consists of composite goods $C_{ij}$, and these composite goods are weighted with parameters $\gamma_{ij} \left( \sum_{i=1}^{15} \gamma_{ij} = 1 \right)$, which are also calibrated in the model, and $\sigma_{ij}$ is elasticity of substitution between composite goods.

On the first level these composite goods are calculated on the basis of data on consumption of relevant goods inside the specified region. I.e., composite goods consist of identical goods, but produced in different regions, and then imported into a single region, and for each region the volume of consumption of some composite goods is calculated. Composite product is designated by:

$$C_{ij} = \left[ \sum_{k=1}^{5} \beta_{ijk} C_{ij}^{k \frac{\sigma^{arm}-1}{\sigma^{arm}}} \right]^{\frac{\sigma^{arm}}{\sigma^{arm}-1}}. \qquad (2)$$

$\sigma^{arm}$ - is the elasticity of Armington; $C_{ij}^{k}$ - consumption of good, which is produced in the region $k$ and consumed in the region $j$; coefficients $\beta_{ijk} \left( \sum_{k=1}^{5} \beta_{ijk} = 1 \right)$ are parameters of the model and calibrated in equilibrium. Thus, $\gamma_{ij}$ and $\beta_{ijk}$ are calibrated in equilibrium, while elasticity and consumption are set in the initial state of economics.

Consumer solves the following problem:

$$\begin{cases} U_j \to \max \\ s.t.\, I_j = \sum_i P_{ji} C_{ji} \end{cases}, \qquad (3)$$

where $P_{rj}$ is the price of composite good $i$ in the region $j$; $I_j$ is the income of consumer. Lets shortly describe the solution of consumer's problem:

$$\frac{\partial L}{\partial C_{ij}} = \left[ \sum_{j=1}^{15} \gamma_{ij} C_{ij}^{\frac{\sigma^{arm}-1}{\sigma^{arm}}} \right]^{\frac{\sigma^{arm}}{\sigma^{arm}-1}} \cdot \gamma_{ij} \cdot C_{ij}^{-1/\sigma^{arm}} - \lambda_j \cdot P_{ij} = 0, \qquad (4)$$

Thus, the demand for composite good $i$ in the region $j$:

$$C_{ij} = \frac{\gamma_{ij}^{\sigma^{arm}} \cdot I_j}{P_{ij}^{\sigma^{arm}} \cdot \sum_j \gamma_{ij}^{\sigma^{arm}} P_{ij}^{1-\sigma^{arm}}} \quad . \qquad (5)$$

For finding the consumption demand on the good $i$ in the region $j$, that is delivered from region $k$, the following consumer problem need to be solved

$$C_{ij} = \left[ \sum_{k=1}^{5} \alpha_{ijk} \cdot C_{ijk}^{\frac{\sigma^{arm}-1}{\sigma^{arm}}} \right]^{\frac{\sigma^{arm}}{\sigma^{arm}-1}} \to \max$$

$$s.t.\, P_{ji} \cdot C_{ji} = \sum_{k=1}^{5} PC_{jik} \cdot C_{jik} = I_j - \sum_{m \neq j} P_{mj} \cdot C_{mj}, \qquad (6)$$

where $PC_{jik}$ is the price of the good $i$ in the region $j$, which is delivered from region $k$. The solution of this problem is the demand on good $i$ in the region $j$ which is delivered from region $k$:

$$C_{jik} = \frac{\gamma_{ijk}^{\sigma^{arm}} \cdot I_j - \sum_{m \neq j} P_{mj} \cdot C_{mj}}{PC_{jik}^{\sigma^{arm}} \cdot \sum_{k=1}^{5} \gamma_{ijk}^{\sigma^{arm}} \cdot PC_{jik}^{1-\sigma^{arm}}}. \quad (7)$$

*Offer of representative producer.*

The function of the constant elasticity of production substitution in sectors is used as a production function:

$$Y_j = \left[ \sum_{i=1}^{15} \beta_{ij} S_{ij}^{\frac{\sigma_{ij}^t - 1}{\sigma_{ij}^t}} \right]^{\frac{\sigma_{ij}^t}{\sigma_{ij}^t - 1}}. \quad (8)$$

In this formula $i$ and $j$ represent goods and regions correspondingly, $i$ may change from 1 to 15, and $j$ - from 1 to 5. $Y_j$ is the general production in the region $j$, which is constructed as following: the production ($S_{ij}$) of 15 goods, produced in the given region $j$ is weighed, ($i$ may change from 1 to 15), together with weights $\beta_{ij} \left( \sum_{i=1}^{15} \beta_{ij} = 1 \right)$, where $\sigma_{ij}^t$ is corresponding elasticity of substitution. By means of parameters $\beta_{ij}$ the production sectors are weighted, and these parameters are calibrated in a model, elasticities are fixed, and the production is given in the initial balance.

The producer's problem may be written as follows:

$$PP_{ij} \cdot S_{ij} \rightarrow \max,$$

$$s.t.: Y_j = \left[ \sum_{i=1}^{15} \beta_{ij} S_{ij}^{\frac{\tau_j^t - 1}{\tau_j^t}} \right]^{\frac{\tau_j^t}{\tau_j^t - 1}}. \quad (9)$$

Here $Y_j$ is the price of good $i$ in the region $j$. Shortely the solution of this problem:

$$\frac{\partial L}{\partial S_{ij}} = PP_{ij} - \lambda_j \cdot \left[ \sum_{j=1}^{15} \beta_{ij} \cdot S_{ij}^{\frac{\tau_j^t - 1}{\tau_j^t}} \right]^{\frac{\tau_j^t}{\tau_j^t - 1}} \cdot \beta_{ij} \cdot S_{ij}^{-1/\tau_j^t} = 0, \quad (10)$$

The solution of the problem is the supply of the good $i$ in the region $j$:

$$S_{ij} = \frac{\beta_{ij}^{\tau_j^t} \cdot Y_j}{PP_{ij}^{\tau_j^t} \cdot \left[ \sum_{i=1}^{15} \beta_{ij}^{\tau_j^t} \cdot PP_{ij}^{1-\tau_j^t} \right]^{\tau_j^t / \tau_j^t - 1}}. \quad (11)$$

*Taxes and tariffs*

Taking into account commodity taxes on the production of good $i$ in a region $j$ ($\tau_{ij}$)

and tariff $tar_{ijk}$ on a product $i$ imported to a region $j$ from a region $k$; the price of good $i$ in the region $j$, imported from region $k$, looks like following:

$$PC_{jik} = PP_{ij} \cdot 1 + tar_{ijk} \cdot 1 + t_{ij}. \qquad (12)$$

Thus, an income from tax and tariffs collection in a region $j$ is following:

$$REV_j = \sum_{i=1}^{15} \sum_{j=1}^{5} tar_{ijk} + t_{ij} + tar_{ijk} \cdot t_{ij} \cdot PP_{ij} \cdot C_{jik} \qquad (13)$$

This income is redistributed as a lumpsum transfer to consumer. Thus, the consumer's income looks like:

$$I_j = \sum_{i=1}^{15} PP_{ij} \cdot S_{ji} + REV_j + AID_j \cdot PCINDEX_j, \qquad (14)$$

where $AID_j = \sum_i \sum_j C^B_{ijk} - \sum_i S^B_{ij}$ is trade disbalance in the region $j$ in the basic year (index $B$ means "the basic year"), $PCINDEX_j = \dfrac{\sum_i \sum_k PC_{ijk} \cdot C^B_{jik}}{\sum_i \sum_k PC^B_{jki} \cdot C^B_{jik}}$ prices index in the region.

*Equilibrium.*

An equilibrium is determined by the set of equilibrium costs of producers (and as tariffs and taxes are set exogenous, the costs of consumers are unequivocal determined in an equilibrium). Except that, an optimal production and optimal consumption of all commodities on all markets are set in equilibrium.

Equilibrium terms:

1) equality of demand and supply on all 15 markets, that is described by equation

$$\sum_{k=1}^{5} C_{ijk} = S_{ij}; \qquad (15)$$

2) strength of balance of trade is in each of regions

$$\sum_{i=1}^{15} \left( PP_{ik} \cdot S_{ik} - C_{ikk} \right) - \sum_{i=1}^{15} \sum_{m \neq k} PP_{im} \cdot C_{imk} - AID_k \cdot PCINDEX_k = 0, \qquad (16)$$

what testifies that a disbalance in the found equilibrium must equal to a trade disbalance in a base equilibrium.

*Model calibration*

At first it is necessary to define weight coefficients $\beta_{ij}, \gamma_{ij}, \gamma_{ijk}$, what would correspond to an equilibrium in a basic year. From (11) we get expressions for $\beta_{ij}$ :

$$\beta_{ij} = \frac{PP_{ij} \cdot S_{ij}^{1/\tau} \cdot}{\left[\frac{Y_j}{\sum_k \beta_{jk}^{\tau} \cdot PP_{jk}^{1-\tau\,\tau/\tau-1}}\right]^{1/\tau}} . \tag{17}$$

Applying the condition $\left(\sum_{j=1}^{15} \beta_{ij} = 1\right)$ we find:

$$\beta_{ij} = \frac{PP_{ij}^{B} \cdot S^{B}_{ij}{}^{1/\tau} \cdot}{\sum_{i=1}^{15}\left(PP_{jk}^{B} \cdot S_{ij}^{B\,1/\tau}\right)} . \tag{18}$$

From (5) we get the expression for $\gamma_{ij}$ :

$$\gamma_{ij} = \frac{P_{ij} \cdot C_{ij}^{1/\sigma}}{\left(\frac{I_j}{\sum_i \gamma_{ij}^{\sigma} P_{ij}^{1-\sigma}}\right)^{1/\sigma}} . \tag{19}$$

Applying the condition $\sum_{i=1}^{15} \gamma_{ij} = 1$ we obtain:

$$\gamma_{ij} = \frac{P_{ij}^{B} \cdot C^{B}_{ij}{}^{1/\sigma} \cdot}{\sum_i \left(P_{ij}^{B} \cdot C_{ij}^{B\,1/\sigma}\right)} . \tag{20}$$

From (7) we get the expression for $\gamma_{ijk}$ :

$$\gamma_{ijk} = \frac{PC_{ijk} \cdot C_{ijk}^{1/\sigma}}{\left(\frac{I_j - \sum_{m \neq i} P_{mj} \cdot C_{mj}}{\sum_k \alpha_{ijk}^{\sigma} \cdot CP_{ijk}^{1-\sigma}}\right)^{1/\sigma}} . \tag{21}$$

Applying the condition $\sum_{i=1}^{15} \gamma_{ijk} = 1$ we obtain:

$$\gamma_{ijk} = \frac{PC_{ijk}^{B} \cdot C_{ijk}^{B\,1/\sigma} \cdot}{\sum_i PC_{ijk}^{B} \cdot C_{ijk}^{B\,1/\sigma}} . \tag{22}$$

By substitution of formulas (5) and (7) into tthe constraint of problem (6), we obtain:

$$\gamma_{ijk} = \frac{P_{ijk}^{B} \cdot C^{B}_{ijk}{}^{1/\sigma} \cdot}{\sum_i \left(PC_{ijk}^{B} \cdot C_{ijk}^{B\,1/\sigma}\right)} . \tag{23}$$

The expression in brackets is equal to $\gamma_{ij}$ in accord to (19). So, we get a formula for price of the composite good $i$ in the region $j$:

$$P_{ij} = \left[ \sum_i \left( PC_{ijk}^{1-\sigma} \cdot \gamma_{ijk}^{\sigma} \right) \right]^{1/_{1-\sigma}} . \quad (24)$$

Formula (24) is included in the system of equations of the model. This formula is applied for the calibration of the model. The last parameter, which must be defined for the base year is the consumption of composite product $i$ in the region $j$:

$$C_{ij}^{B} = \left[ \sum_{i=1}^{5} \left( \gamma_{ijk} \cdot C_{ijk}^{B\ \sigma-1/_{\sigma}} \right) \right]^{\sigma/_{\sigma-1}} . \quad (25)$$

As can be seen from the obtained ratio, equilibrium prices are not available in explicitly and should be calculated in balance by means of computational processes. Calculations in the model were carried out using the package with the optimization algorithm. This algorithm is a generalization of the composite gradient Wolf 's algorithm. The method is based on reducing the dimension by using the representation of all the variables through a set of independent variables (method is similar to the standard gradient method).

*Existence and uniqueness of solution*

The existence and uniqueness of solution follows from the formulation of an optimization problem. Namely, in the optimization problem consumers optimize convex functions of utility with convex constraints (3). From the other side, manufacturers present a convex production function and also are limited by the convex budget restrictions (9) in their actions. Thus, the maximization of convex functions in convex sets occurs in this optimization problem. So, the solution exists and is unique. Note that from the very beginning the functionы in the model were chosed such that there was no difficulty with existence and uniqueness of solution.

*Industries and regions*

This model were calibrated on data for year 2000 and includes 5 geographic regions, each of which is modeled as a separate economy: 1) Russian Federation; 2) 15 EU countries (EU-15: Belgium, Denmark, Germany, Greece, Spain, France, Ireland, Italy, Luxembourg, Netherlands, Austria, Portugal, Finland, Sweden and the United Kingdom); 3) 10 countries that are joining the EU (SK-10: Cyprus, the Czech Republic, Estonia, Latvia, Hungary, Lithuania, Malta, Poland, Slovakia and Slovenia); 4) 11 countries of the Commonwealth of independent States (CIS-11: Armenia, Azerbaijan, Belarus, Georgia, Kazakhstan, Kyrgyzstan, Moldova, Tajikistan, Turkmenistan, Ukraine and Uzbekistan); 5) the rest of the world.

In each of the regions 15 sectors were modeled: energy, oil and gas industry, the other fuel industry, ferrous metallurgy, non-ferrous metallurgy, chemical and petrochemical industry, mechanical engineering and metal processing, light industry, food industry, other industries, agriculture and forestry, construction, transportation and communication, utilities, finance, banking and insurance.

*Calculation Experiments*

Model was used to estimate the effects on the economy of the Russian Federation from different scenarios of liberalization of the Russian trade, in particular, the influence of impact of EU enlargement on the Russian economy was estimated, as well as the economy of a number of other countries. Non-tariff regulation measures were not included in the scripts and expansion processes were simulation, anticipating that the countries that come in the EU will adjust tariff rates that apply to the old EU Member States. First the base model was calibrated, which is the equilibrium of economy before the EU expansion. Then tariff rates, that are applied in countries that are joining EU, change, setting them on a level that applies the old EU members. Note that there is a big difference between these tariff rates. So for Russia the most powerful export sector is «oil and gas»: the EU applies a tariff rate close to zero, and the countries that are joining the EU (by our estimation), apply a tariff rate of about 2.08.

Yet we note significant differences in tariff rates in such sectors as food, chemical industry and oil processing industry: 17.69% - the old EU members and 6.8% of the country that join the EU. Model scenarios evaluation show that trade flows will presumably change as following: trade between old and new members of the EU have significantly accelerate of 4,73%, but the trade with countries of other regions, which includes Russia, CIS countries and the rest of the world, is deteriorating. That is, the effect of the creation of the trade, and the effect of the substitution of trade may be observed (see following diagram):

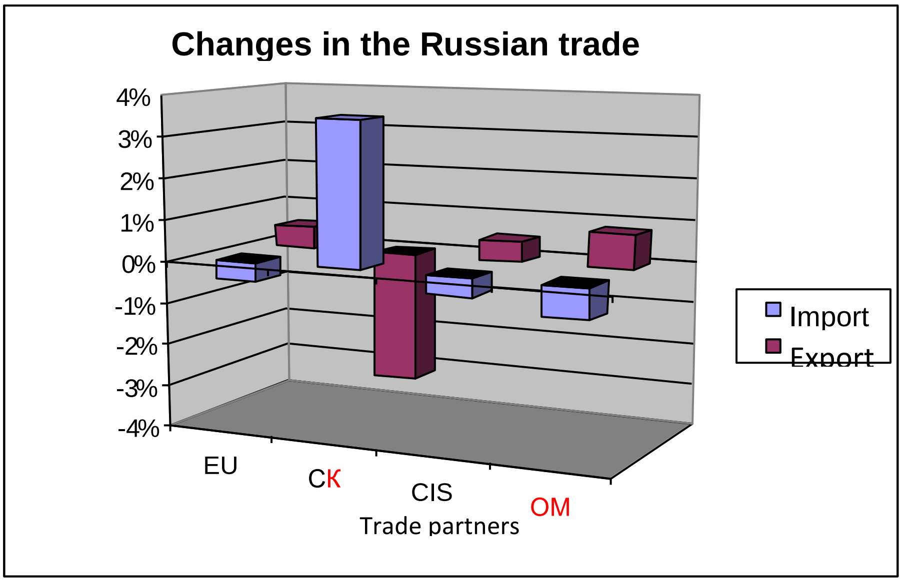



*Figure 5.2. Changes in the Russian trade (%)*

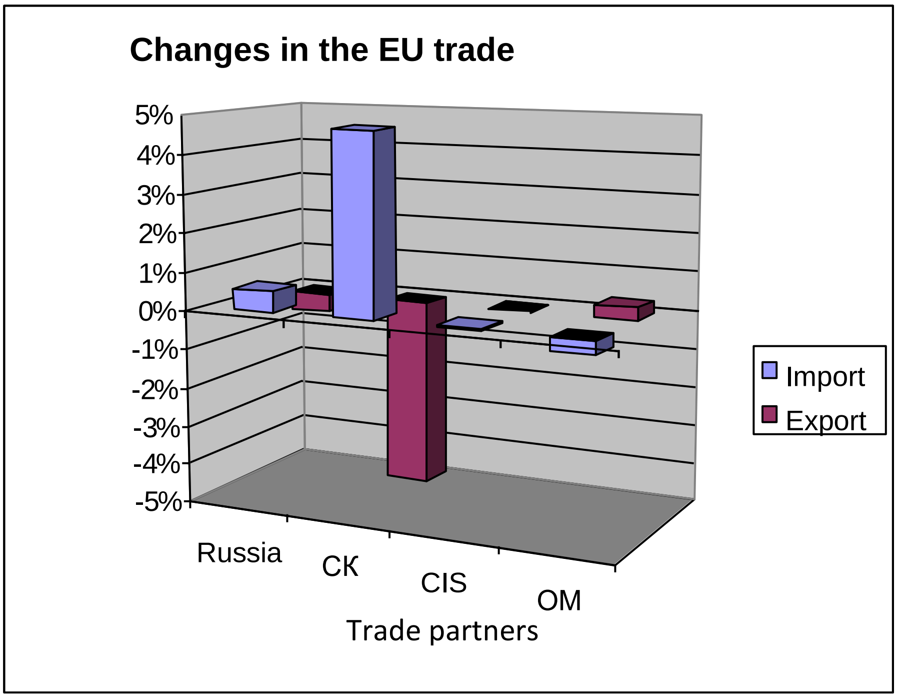



*Figure 5.3. The changes in the trade of the EU (%)*

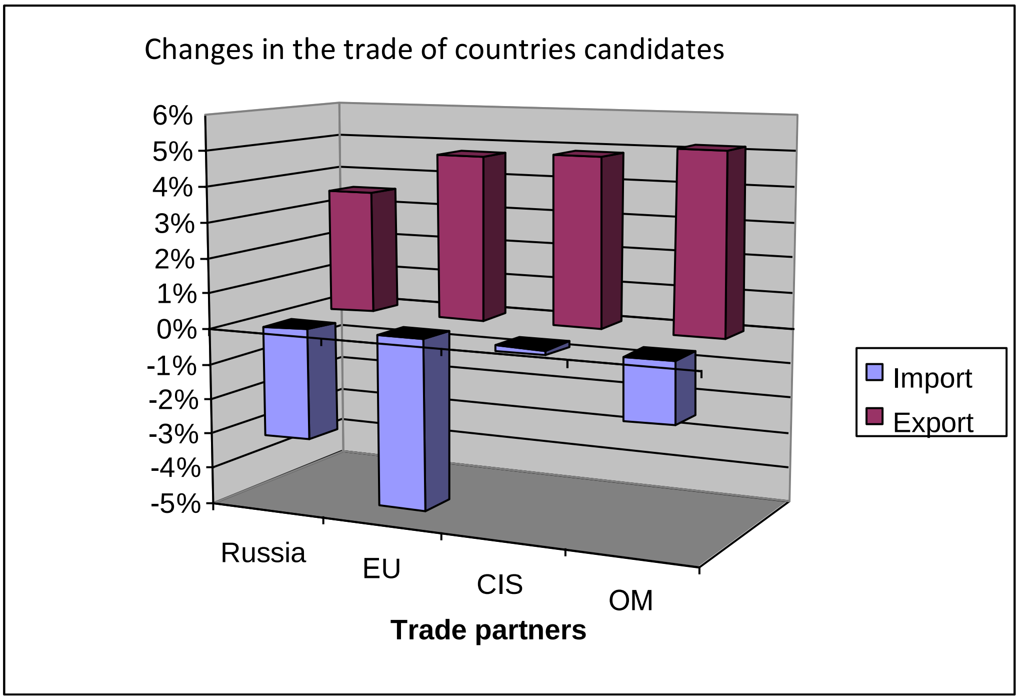



*Figure 5.4. Changes in the trade of the countries candidates (%)*

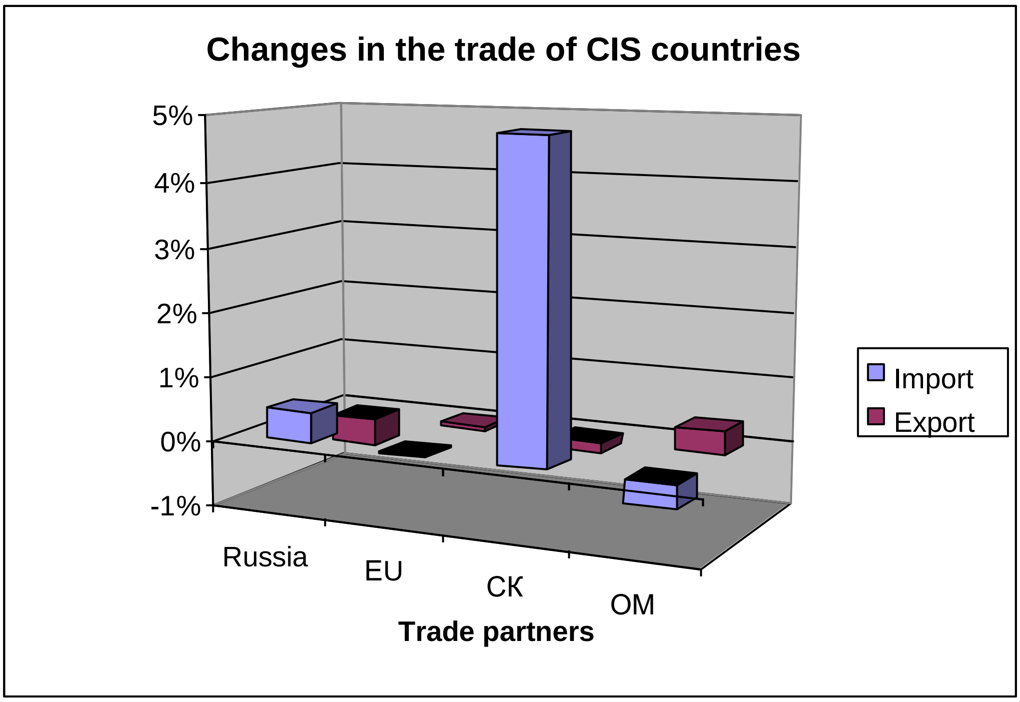



*Figure 5.5. Changes in the trade of CIS countries (%)*

Obvious fact is that trade between the countries, that join the EU and the Russian Federation, as well as CIS countries increased after the enlargement of the eurozone. This is the result of the fares discount that is applicable by countries join the EU to the level of the tariffs applied to older members of the EU. RF also experienced a significant reduction in the level of exports to countries that join the EU, and the existing exports to the EU countries is partially replaced by exports from countries of the former Union. On the other hand, Russia begins to export more to other regions, which means a greater degree of diversification of trade to Russia, since the expansion of the most part of RF exports was to countries that join the EU. Import, in turn, varies in the opposite direction. Countries that join the EU increase their export into Russia, while other regions decrease.

In order to have an overall picture of the changes of trade flows, it is necessary to examine the so-called terms of trade:

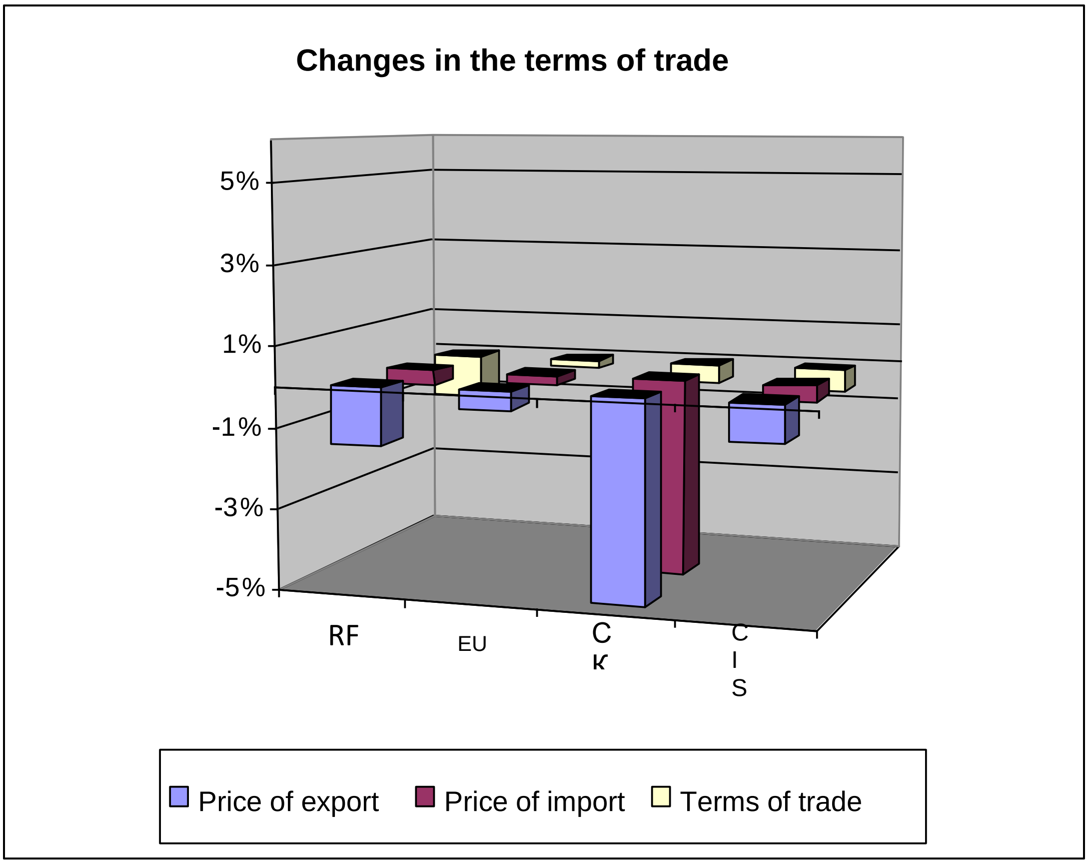



*Figure 5.6. Terms of trade by region*

Figure 5.6 shows how changes in tariffs that will occur as a result of EU expansion will affect on terms of trade changes. Russia's terms of trade worsened in average on 1%. It is important to note that these changes are relative. For example, dropping to 0.2% in the terms of trade for countries that are joining the EU, must be considered as a dropping of much smaller size than the deterioration of terms of trade for the RF and the CIS countries. The cause of the deteriorating of terms of trade in the EU and joining countries of the EU, is the fact that after the expansion of the EU the prices of manufacturers dropped, and this drop outweighs the price increase due to the improved trade opportunities. The graph shows that the country that most loses in terms of terms of trade, actually is RF.

In this scenario, the well-being in the countries that are joining the EU is at 2.3% after entry into the European Union. The only factor that could explain such effect is a change in the terms of trade. Welfare of the RF also drops, but only at 0.07%. Because the model is fixed by GDP, namely the terms of trade lead to this effect.

The following Figures 5.7 and 5.8 show the production of the Russian Federation, export and import in the basic balance, as well as illustrate the interest change of production at the extension scripts.

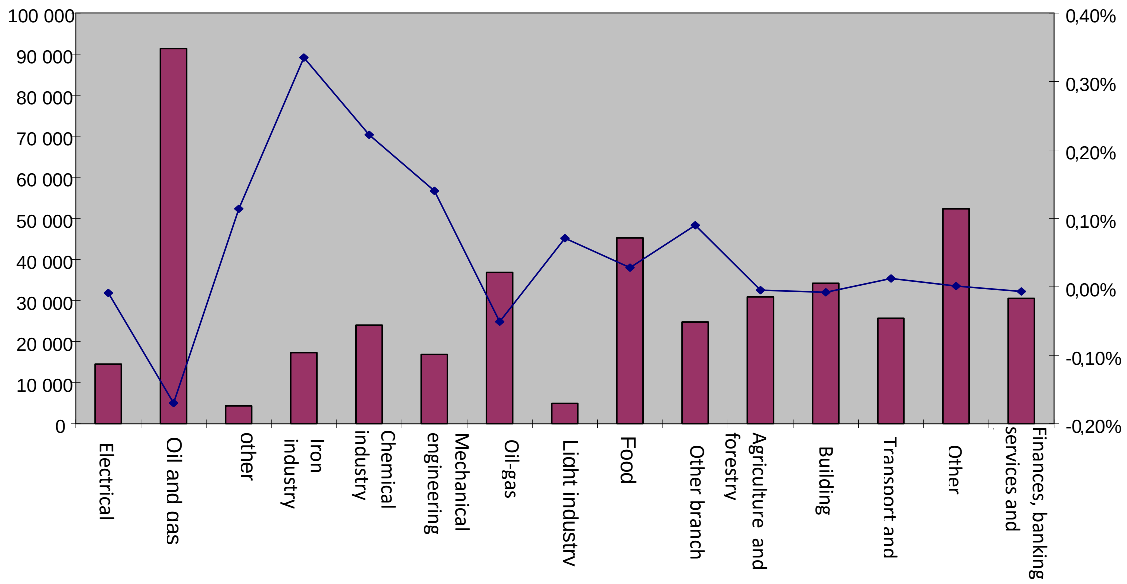



*Figure 5.7. Changes of Russian production*

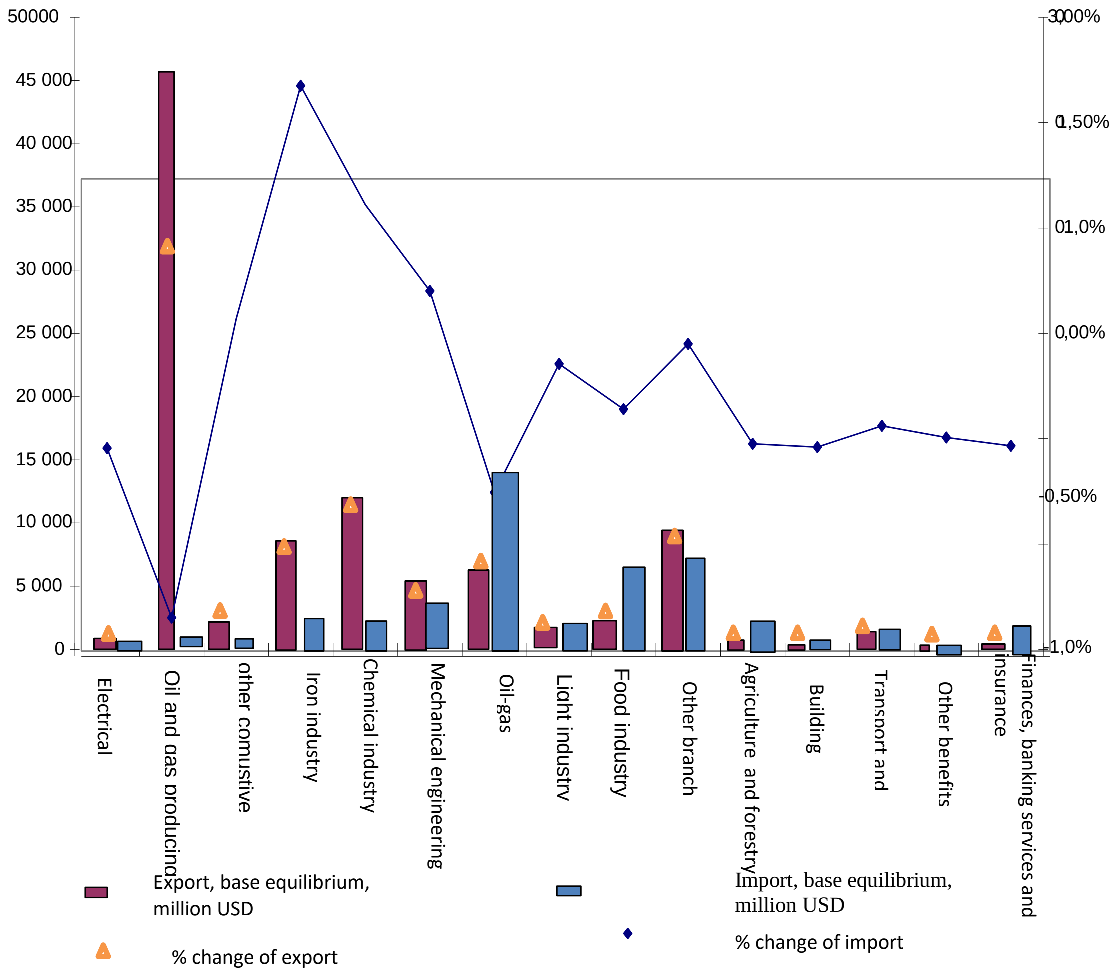



*Figure 5.8. Changes of Russian trade*

Structure changes of trade flows is quite obvious. Exports mostly increases, and import decreases. The only sector, where import increases is the “electricity and heat” sector, but the nominal volumes are small, because RF almost does not import electricity. Most likely it can be errors of model. Food industry and Agriculture and the forest industry undergo the biggest changes in the export. As has been said, it is precisely those sectors where countries have relatively high levels of tariff rates. So in these sectors RF faces a higher competition from countries that are joining the EU, and this is the reason that explains why in these areas RF had experienced the greatest losses.

In this scenario also a slight diversification of Russian production happen, namely: in the greatest industry «oil and gas» level of production drops to 0.17%, while in other areas it is on average growing (so you can talk about diversification, but insignificant).

*Conclusions*

Constructed model of general equilibrium calculation (using the Global v. 1) allows to assess the impact of trade policy changes on the economy of individual States, in particular, explore industry changes and is very flexible tool for obtaining numerical experiments.

Unlike models of partial equilibrium, such as autoregression analysis (with the exception of vector regressions), this model allows you to explore not only direct changes of any specific tariffs or taxes, but also is able to assess indirect effects. In addition, this model can not have endogeneouse problems (unlike the models of partial equilibrium).

By means of the model it is obviouse that EU expansion is disadvantageous for RF but Russia wins in terms of well-being.

The disadvantages of the model include assumptions about markets with perfect competition, that is actually incorrect for a number of highly monopolized industries (oil and gas). One would include a field with the growing impact of the scale but you need to know the industry ratios, and this is very problematic for Russian reality.

*Example 5.2. Dynamic stochastic model of the general economic equilibrium of Russia («*DYNAMIC STOCHASTIC MODEL OF RUSSIA'S GENERAL ECONOMIC EQUILIBRIUM» *D.L. Andrianov, D.N. Shultz, I.A. Oshchepkov )*

Recently, an important place in the analysis, modeling and forecasting of macroeconomic processes belongs to dynamical model of the general economic equilibrium (DSGE models). This approach to modeling comes in the framework of the so-called New Neoclassical Synthesis – a new direction economic thought, which became a product of the synthesis of neoclassic and neo-Keynesian. In DSGE models is the ability to take into account the effect of technological factors and such imperfection of real markets as tight prices and salaries, imperfect information and imperfect competition. DSGE modeling is used in the practice of the State regulation of the United States, Canada, Britain, Sweden, Chile, New Zealand and others. It is considered the DSGE model of the Russian economy taking into account the current realities.

Any DSGE model includes 3 required components: 1) the dynamic version of the equation IS to national income; 2) New Keynesian Phillips curve for inflation; 3) Taylor rule for interest rate modelling. The last equation is important in terms of the transition of the Bank of Russia to the policy of inflation targeting. Deemed necessary to include in the system the equation for exchange rate based on uncovered interest rate parity, and also separately note the equation for inflation expectations. In addition to the specified key equations of the model, will also need an auxiliary equation for trend-cyclical component separation, the calculation of the short-term deviations from equilibrium, calculation of deflation variables, etc.

Thus, in the opinion of the authors, small DSGE model of Russia should include 4 endogenous variables: GDP, inflation, interest rate and exchange rate. Exogenous variables traditionally are the variables that determine the dynamics of the Russian economy: world GDP, inflation and interest rate. The relationship between these indicators will show in the picture:

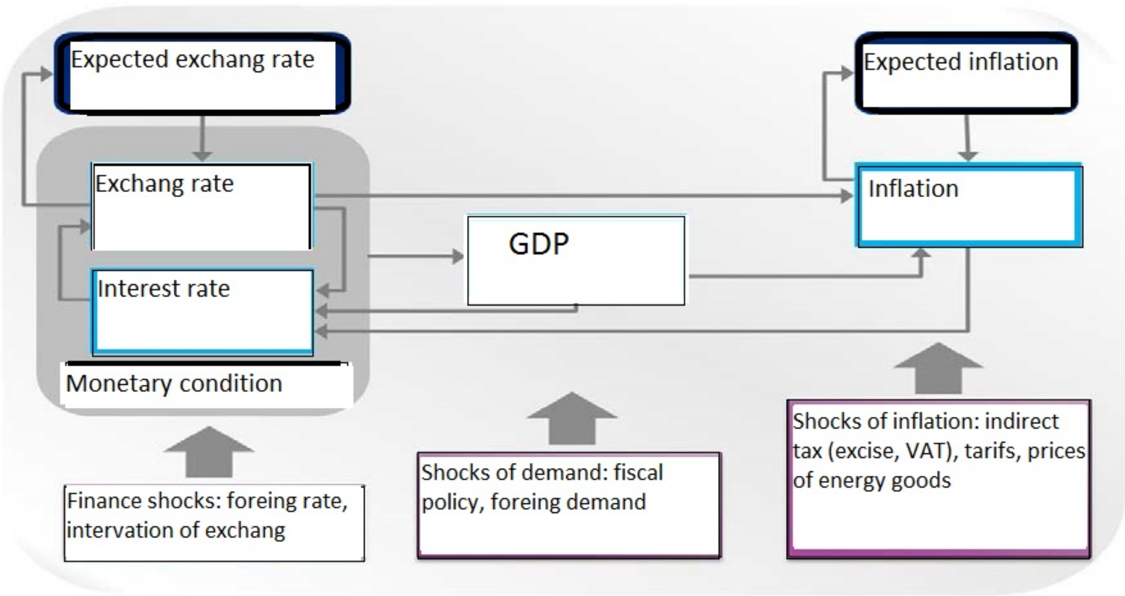



*Figure 5.9. Graph of interaction of activities of DSGE model for Russia*

Let's consider in more detail the key equations of the system. As mentioned above, for the simulation of GDP the dynamic version of the curve IS is applied:

$\hat{y}_t = a_1\hat{y}_{t-1} - a_2 mci_t + a_3\hat{y}_t^* + \varepsilon_t^y$ , where $\hat{y}_t$ is the deviation (gap) of output from equilibrium (long-term) level in time $t$; $mci_t$ is index of monetary conditions at the moment of time $t$, calculated by formula $mci_t = a_4\hat{r}_t + (1-a_4)(-\hat{z}_t)$; $\hat{y}_t^*$ - the deviation of the world release from the equilibrium level in time $t$; $\varepsilon_t^y$ is shock of demand in time $t$; $\hat{r}_t$ are deviations of the real rate of interest from the equilibrium values in a given time $t$; $a_1$ is inertia coefficient of deviations release; $a_2$ is coefficient of influence of monetary conditions on the real economy; $a_3$ is coefficient of influence of foreign demand on release; $a_4$ is the coefficient of significance of interest rate value in monetary policy.

These equations allow us to consider the following features of the economy: - inertial change of release; - the impact of monetary policy on the real economy, especially interest rates, and non-monetary factors; - the impact of external demand in the domestic economy; - the choice of the consumers between present and future consumption; - the choice of the consumers between domestic and imported goods.

The peculiarity of the DSGE models is the use of filters for the separation of the trend-cycle component. While the results of the filter depends on the selected method, parameter smoothing, etc. So when the new management of the Bank of Russia announce that in forming monetary policy's DSGE-approach, the economists it caused a certain wariness. It is shown that different procedures of smoothing can produce enough conflicting signals about the phases of the economic cycle. This model for the selection of trend-cycle component uses the filter Hodric-Prescott on the period from the 1st quarter of 2000. 1 quarter of 2014. with the standard parameter smoothing 1600 for quarterly dynamics. Results filtering are presented here:


*Figure 5.10. Validity filter Hodric-Prescott to temp of real GDP RF.*

This specification reflects the following characteristics of the economy: - the presence of inflationary expectations; - the impact of Economics on monetary processes; - rigidity (inertial) nominal values; - nonmonetary facet of inflation, including the impact of tariffs of natural monopolies; - the effect of imported inflation.

Theoretical basis of DSGE models is the concept of rational expectations. According to this concept: a) the economic agents available complete information, as well as the State; b) information is used as efficiently as possible; c) as a result, in the economic agents have perfect foresight, that is able to accurately predict actual future inflation, exchange rate, etc.

Today, the hypothesis of rational expectations questioned by many economists as unrealistic. For Russia, perhaps more realistic is the hypothesis of

adaptive expectations. Under the hypothesis of adaptive expectations of economic agents is not required perfect foresight, it is enough to consider the previous errors.
Taylor's idea was this: to develop the rule of conduct of the Central Bank, which would count the economy when it is overheated (that is, when inflation exceeds the target level and when national income exceeds the potential level) and stimulating in the recession. But for oil-firms, export-oriented countries rule Taylor turns out to be not quite adequate. Central banks such countries are forced to follow not only for inflation, and perhaps even to a greater extent, due to exchange rate. Taylor rule for Russia was identified as economists the center of macroeconomic research as follows:
In addition, the complex has the question of choice of interest rate. In some of the DSGE models used interbank crediting rate, but in the process of transition to inflation targeting CBR use interest politics had everything to a single "key rates", and was chosen.

This specification reflects the following characteristics of the economy: - target and the expected level of exchange rates; - the impact of currency interventions on currency rate; - the impact of domestic and foreign interest rate on exchange rate; - flexibility of exchange rates; - the impact of domestic and foreign inflation on the exchange rate.

Developed a DSGE model was implemented in the Prognoz Platform. In comparison with popular Dynare, Prognoz Platform has several advantages: the integrity of the databases of the international and Russian socio-economic statistics; availability of convenient interface for working with the DSGE models; no need for knowledge of the syntax of the programming language.
The Prognoz Platform was conducted calibration of model parameters. While largely ignored ranges of values of the parameters, featured in previous research.

On the basis of calibrated DSGE models were calculated projections on key economic indicators for the 3 years to come while maintaining economic curremt (Figure 5.11 and aggregate representation for years – Table 5.2):

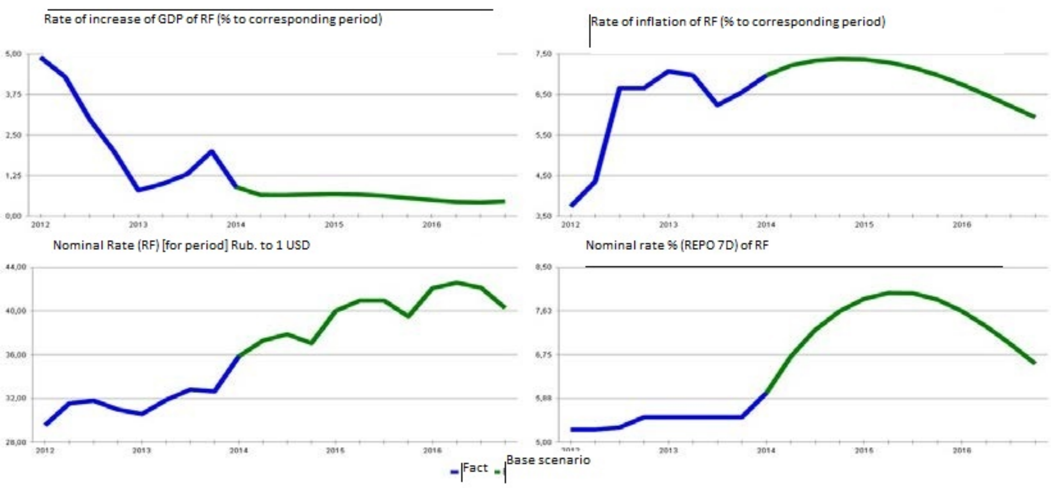



*Figure 5.11. Prediction of key economic indicators RF.*

*Table 5.2. Prediction of key economic indicators RF*

| *Variable* | *2014* | *2015* | *2016* |
|---|---|---|---|
| *Temp increment of real GDP, %* | *0.72* | *0.64* | *0.46* |
| *Temp increment of inflation, %* | *7.23* | *7.21* | *6.35* |
| *Exchange rate (RUR/US $)* | *37.02* | *40.36* | *41.78* |
| *Rate REPO 7D, %* | *6.89* | *7.92* | *7.12* |

Note that in 2013 were large-scale events that it was impossible to predict at the time of the prediction at the beginning of the year and who have made a significant contribution to the total value of the macroeconomic indicators in 2014:
- Western sanctions imposed on the main sectors of the Russian economy, which do not allow the use in the production of modern foreign technologies and restricting access to financial institutions in overseas capital markets; - reverse countersanctions on imports of agricultural products; - the rapid fall of oil prices.

Western sanctions have caused an additional reduction of GDP, countersanctions get in an accelerated growth in prices and the fall in oil prices has caused a major devaluation of the national currency. But despite all the mark, the model allowed exactly catch the trends of economic development of Russia on the qualitative level. For example, was able to correctly predict the slowing of economic growth against the devaluation of the ruble and inflation (stagnation); natural for similar conditions of the reaction of the Central Bank in the form of increasing the rate of compound interest. Also in the following periods of time on a background of weakening inflationary expectations and a slowing inflation rate expected took interest discount. Thus, the model demonstrates that RUSSIAN economic problems are internal in nature, and external shocks just upped their scale.

So increasingly, having a discussion about carrying out structural reforms and the search for a new model of development of the Russian economy. There are many suggestions from economists to accelerate their medium-term economic growth: freezing of tariffs of natural monopolies, increase government spending (including through the Fund of national wellbeing), interest rate cuts, the stabilization or further devaluation of the exchange rate, etc.

To assess the effects of various instruments of State economic policy, from external shocks DSGE models contain all the equations the variables $\varepsilon_t$ (shocks, or innovation). Using the constructed model estimated impact of 1% changes in exogenous variables in the model variables: GDP, inflation, exchange rate and interest rate.

Analyze the results. One-off negative shock inflation (for example caused by freezing tariffs) will conduct a stimulatory effect on economic growth, including through interest rate cuts (Keynes). This is especially noteworthy is that according to the model, lowering of inflation induces is not strengthening, but weakening the national currency due to declining rates of interest. Indeed, in the current context of the RUSSIAN capital flows have a more meaningful impact on the exchange rate than trade balance.

Analysis of time lag allows you to assert that the effect of freezing the tariffs will not in the short and medium term. And here's a one-off positive fiscal shock (for example, the increase in public spending) just has a short-term – virtually the entire rise of GDP is in the first year. An increase in government spending will lead to a slight increase in inflation, and as a result, the growth rate of the Central Bank.

Single growth rate (devaluation of the ruble) increases and the GDP and inflation. This acceleration of the economy going without even looking at the growth of nominal interest rates associated with an acceleration of inflation. That is, in this case positive external economic effect outweighs the negative effect of rising interest rates.

Finally, the single reduced interest rate will lead to accelerating economic growth in spite of accelerating inflation. Again the maximum GDP increase is expected in the medium term. The effect of reducing interest rates is due to the guarantee of the national currency.

Noteworthy that when lowering interest rates going accelerated inflation. As a result, the Central Bank, which pursues the goal of monetary stability, and sooner or later will be forced to again raise the rate of interest that will lead to a slowdown in economic activity. Judging from this, it is this line of reasoning, as well as the obvious devaluation effect in holding back the regulator from the reduction of key rates in terms of stagnation.

# Chapter 6. Modeling using Neural Network

Introduction to neural networks

*What is a Neural Network?* An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurones) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process. Learning in biological systems involves adjustments to the synaptic connections that exist between the neurones. This is true of ANNs as well.

*Historical background.* Neural network simulations appear to be a recent development. However, this field was established before the advent of computers, and has survived at least one major setback and several eras.

Many import and advances have been boosted by the use of inexpensive computer emulations. Following an initial period of enthusiasm, the field survived a period of frustration and disrepute. During this period when funding and professional support was minimal, important advances were made by relatively few researchers. These pioneers were able to develop convincing technology which surpassed the limitations identified by Minsky and Papert. Minsky and Papert, published a book (in 1969) in which they summed up a general feeling of frustration (against neural networks) among researchers, and was thus accepted by most without further analysis. Currently, the neural network field enjoys a resurgence of interest and a corresponding increase in funding. The history of neural networks that was described above can be divided into several periods:

1. **First Attempts:** There were some initial simulations using formal logic. McCulloch and Pitts (1943) developed models of neural networks based on their understanding of neurology. These models made several assumptions about how neurons worked. Their networks were based on simple neurons which were considered to be binary devices with fixed thresholds. The results of their model were simple logic functions such as "a or b" and "a and b". Another attempt was by using computer simulations. Two groups (Farley and Clark, 1954; Rochester, Holland, Haibit and Duda, 1956). The first group (IBM reserchers) maintained closed contact with neuroscientists at McGill University. So whenever their models did not work, they consulted the neuroscientists. This interaction established a multidisciplinary trend which continues to the present day.
2. **Promising & Emerging Technology:** Not only was neroscience influential in the development of neural networks, but psychologists and engineers also contributed to the progress of neural network simulations. Rosenblatt (1958)

stirred considerable interest and activity in the field when he designed and developed the Perceptron. The Perceptron had three layers with the middle layer known as the association layer. This system could learn to connect or associate a given input to a random output unit. Another system was the ADALINE (ADAptive LInear Element) which was developed in 1960 by Widrow and Hoff (of Stanford University). The ADALINE was an analogue electronic device made from simple components. The method used for learning was different to that of the Perceptron, it employed the Least-Mean-Squares (LMS) learning rule.

3. **Period of Frustration & Disrepute:** In 1969 Minsky and Papert wrote a book in which they generalised the limitations of single layer Perceptrons to multilayered systems. In the book they said: "...our intuitive judgment that the extension (to multilayer systems) is sterile". The significant result of their book was to eliminate funding for research with neural network simulations. The conclusions supported the disenchantment of researchers in the field. As a result, considerable prejudice against this field was activated.
4. **Innovation:** Although public interest and available funding were minimal, several researchers continued working to develop Neuromorphically based computational methods for problems such as pattern recognition. During this period several paradigms were generated which modern work continues to enhance. Grossberg's (Steve Grossberg and Gail Carpenter in 1988) influence founded a school of thought which explores resonating algorithms. They developed the ART (Adaptive Resonance Theory) networks based on biologically plausible models. Anderson and Kohonen developed associative techniques independent of each other. Klopf (A. Henry Klopf) in 1972, developed a basis for learning in artificial neurons based on a biological principle for neuronal learning called heterostasis. Werbos (Paul Werbos 1974) developed and used the back-propagation learning method, however several years passed before this approach was popularized. Back-propagation nets are probably the most well known and widely applied of the neural networks today. In essence, the back-propagation net. is a Perceptron with multiple layers, a different threshold function in the artificial neuron, and a more robust and capable learning rule. Amari (A. Shun-Ichi 1967) was involved with theoretical developments: he published a paper which established a mathematical theory for a learning basis (error-correction method) dealing with adaptive pattern classification. While Fukushima (F. Kunihiko) developed a step wise trained multilayered neural network for interpretation of handwritten characters. The original network was published in 1975 and was called the Cognitron.
5. **Re-Emergence:** Progress during the late 1970s and early 1980s was important to the re-emergence on interest in the neural network field. Several factors influenced this movement. For example, comprehensive books and conferences provided a forum for people in diverse fields with specialized technical languages, and the response to conferences and publications was quite positive. The news media picked up on the increased

activity and tutorials helped disseminate the technology. Academic programs appeared and courses were introduced at most major Universities (in US and Europe). Attention is now focused on funding levels throughout Europe, Japan and the US and as this funding becomes available, several new commercial with applications in industry and financial institutions are emerging.

6. **Today:** Significant progress has been made in the field of neural networks- enough to attract a great deal of attention and fund further research. Advancement beyond current commercial applications appears to be possible, and research is advancing the field on many fronts. Neurally based chips are emerging and applications to complex problems developing. Clearly, today is a period of transition for neural network technology.

The first artificial neuron was produced in 1943 by the neurophysiologist Warren McCulloch and the logician Walter Pits. But the technology available at that time did not allow them to do too much.

*Why use neural networks?* Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "*expert*" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "*what if*" questions. Other advantages include:

1. Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.
2. Self-Organization: An ANN can create its own organization or representation of the information it receives during learning time.
3. Real Time Operation: ANN computations may be carried out in parallel, and special hardware devices are being designed and manufactured which take advantage of this capability.
4. Fault Tolerance via Redundant Information Coding: Partial destruction of a network leads to the corresponding degradation of performance. However, some network capabilities may be retained even with major network damage.

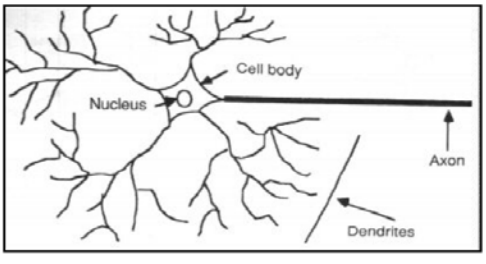
*Neural networks versus conventional computers.* Neural networks take a different approach to problem solving than that of conventional computers. Conventional computers use an algorithmic approach i.e. the computer follows a set of instructions in order to solve a problem. Unless the specific steps that the computer needs to follow are known the computer cannot solve the problem. That restricts the problem solving capability of conventional computers to problems that we already understand and know how to solve. But computers would be so much more useful if they could do things that we don't exactly know how to do.

Neural networks process information in a similar way the human brain does. The network is composed of a large number of highly interconnected processing elements (neurones) working in parallel to solve a specific problem. Neural networks learn by example. They cannot be programmed to perform a specific task. The examples must be selected carefully otherwise useful time is wasted or even worse the network might be functioning incorrectly. The disadvantage is that because the network finds out how to solve the problem by itself, its operation can be unpredictable.

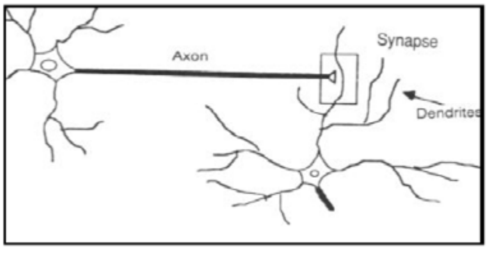
On the other hand, conventional computers use a cognitive approach to problem solving; the way the problem is to solved must be known and stated in small unambiguous instructions. These instructions are then converted to a high level language program and then into machine code that the computer can understand. These machines are totally predictable; if anything goes wrong is due to a software or hardware fault.

Neural networks and conventional algorithmic computers are not in competition but complement each other. There are tasks are more suited to an algorithmic approach like arithmetic operations and tasks that are more suited to neural networks. Even more, a large number of tasks, require systems that use a combination of the two approaches (normally a conventional computer is used to supervise the neural network) in order to perform at maximum efficiency.

*Human and Artificial Neurones - investigating the similarities. How the Human Brain Learns?* Much is still unknown about how the brain trains itself to process information, so theories abound. In the human brain, a typical neuron collects signals from others through a host of fine structures called *dendrites*. The neuron sends out spikes of electrical activity through a long, thin stand known as an *axon*, which splits into thousands of branches. At the end of each branch, a structure called a *synapse* converts the activity from the axon into electrical effects that inhibit or excite activity from the axon into electrical effects that inhibit or excite activity in the connected neurones. When a neuron receives excitatory input that is sufficiently large compared with its inhibitory input, it sends a spike of electrical activity down its axon. Learning occurs by changing the effectiveness of the synapses so that the influence of one neuron on another changes.

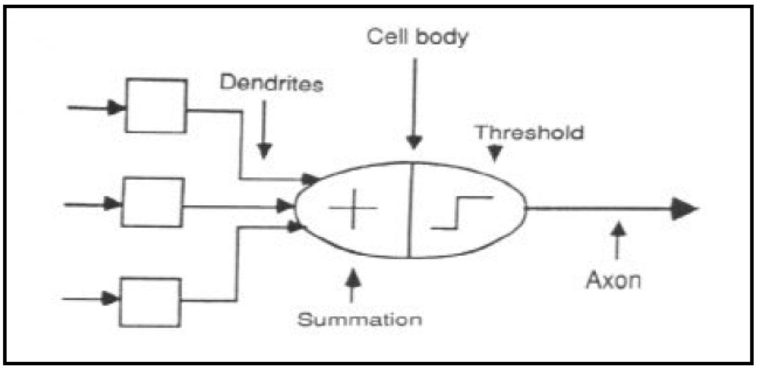


Components of a neuron



The synapse

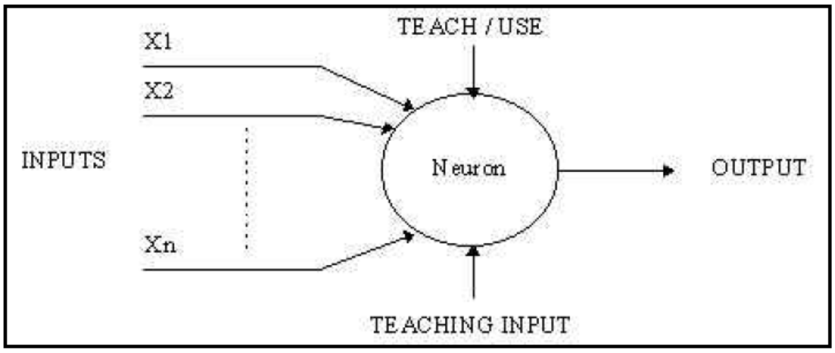
*From Human Neurones to Artificial Neurones*. We conduct these neural networks by first trying to deduce the essential features of neurons and their interconnections. We then typically program a computer to simulate these features. However because our knowledge of neurons is incomplete and our computing power is limited, our models are necessarily gross idealizations of real networks of neurons.




The neuron model

*An engineering approach:*

*a) A simple neuron.* An artificial neuron is a device with many inputs and one output. The neuron has two modes of operation; the training mode and the using mode. In the training mode, the neuron can be trained to fire (or not), for particular input patterns. In the using mode, when a taught input pattern is detected at the input, its associated output becomes the current output. If the input pattern does not belong in the taught list of input patterns, the firing rule is used to determine whether to fire or not.




A simple neuron

*b) Firing rules.* The firing rule is an important concept in neural networks and accounts for their high flexibility. A firing rule determines how one calculates whether a neuron should fire for any input pattern. It relates to all the input patterns, not only the ones on which the node was trained.

A simple firing rule can be implemented by using Hamming distance technique. The rule goes as follows: *take a collection of training patterns for a node, some of which cause it to fire (the 1-taught set of patterns) and others which prevent it from doing so (the 0-taught set). Then the patterns not in the collection cause the node to fire if, on comparison , they have more input elements in common with the 'nearest' pattern in the 1-taught set than with the 'nearest' pattern in the 0-taught set. If there is a tie, then the pattern remains in the undefined state.*

*Example 6.1.* a 3-input neuron is taught to output 1 when the input (X1,X2 and X3) is 111 or 101 and to output 0 when the input is 000 or 001. Then, before applying the firing rule, the truth table is;

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| X1: | | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| X2: | | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| X3: | | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| OUT: | | 0 | 0 | 0/1 | 0/1 | 0/1 | 1 | 0/1 | 1 |

As an example of the way the firing rule is applied, take the pattern 010. It differs from 000 in 1 element, from 001 in 2 elements, from 101 in 3 elements and from 111 in 2 elements. Therefore, the 'nearest' pattern is 000 which belongs in the 0-taught set. Thus the firing rule requires that the neuron should not fire when the input is 001. On the other hand, 011 is equally distant from two taught patterns that have different outputs and thus the output stays undefined (0/1).

By applying the firing in every column the following truth table is obtained;

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| X1: | | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| X2: | | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| X3: | | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| OUT: | | 0 | 0 | 0 | 0/1 | 0/1 | 1 | 1 | 1 |

The difference between the two truth tables is called the *generalization of the neuron*. Therefore the firing rule gives the neuron a sense of similarity and enables it to respond 'sensibly' to patterns not seen during training.

*c) Pattern Recognition - an example.* An important application of neural networks is pattern recognition. Pattern recognition can be implemented by using a feed-forward (figure 1) neural network that has been trained accordingly. During training, the network is trained to associate outputs with input patterns. When the network is used, it identifies the input pattern and tries to output the associated output pattern. The power of neural networks comes to life when a pattern that has no output associated with it, is given as an input. In this case, the network gives the output that corresponds to a taught input pattern that is least different from the given pattern.

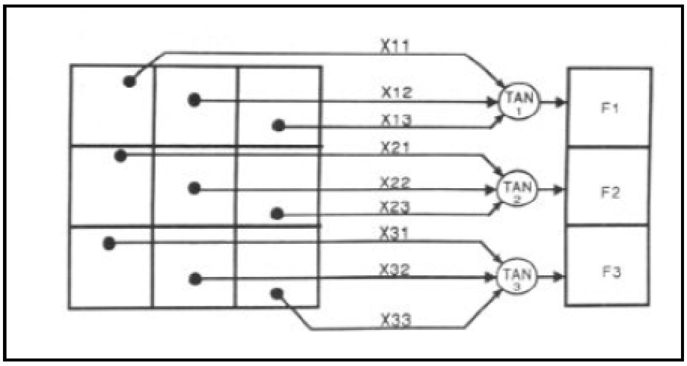



*Figure 6.1*

*Example 6.2:* The network of figure 6.1 is trained to recognize the patterns T and H. The associated patterns are all black and all white respectively as shown below.


If we represent black squares with 0 and white squares with 1 then the truth tables for the 3 neurones after generalization are;

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| X11: | | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| X12: | | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| X13: | | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| OUT: | | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |

*Top neuron*

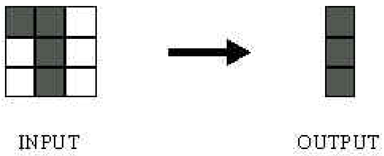
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| X21: | | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| X22: | | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| X23: | | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| OUT: | | 1 | 0/1 | 1 | 0/1 | 0/1 | 0 | 0/1 | 0 |

*Middle neuron*

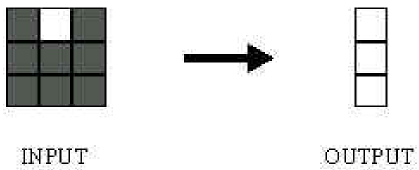
| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| X21: | | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 |
| X22: | | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 |
| X23: | | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| OUT: | | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 |

*Bottom neuron*

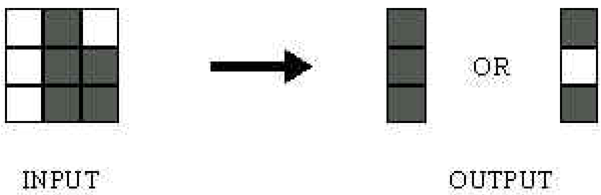
From the tables it can be seen the following associations can be extracted:




In this case, it is obvious that the output should be all blacks since the input pattern is almost the same as the 'T' pattern.




Here also, it is obvious that the output should be all whites since the input pattern is almost the same as the 'H' pattern.




Here, the top row is 2 errors away from the a T and 3 from an H. So the top output is black. The middle row is 1 error away from both T and H so the output is random. The bottom row is 1 error away from T and 2 away from H. Therefore the output is black. The total output of the network is still in favour of the T shape.

*d) A more complicated neuron.* The previous neuron doesn't do anything that conventional computers don't do already. A more sophisticated neuron (figure 6. 2) is the McCulloch and Pitts model (MCP). The difference from the previous model is that the inputs are 'weighted', the effect that each input has at decision making is dependent on the weight of the particular input. The weight of an input is a number which when multiplied with the input gives the weighted input. These weighted inputs are then added together and if they exceed a pre-set threshold value, the neuron fires. In any other case the neuron does not fire.

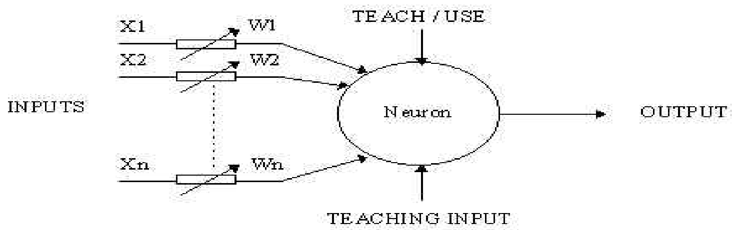



*Figure 6.2. An MCP neuron*

In mathematical terms, the neuron fires if and only if;

$$X_1W_1 + X_2W_2 + X_3W_3 + ... > T.$$

The addition of input weights and of the threshold makes this neuron a very flexible and powerful one. The MCP neuron has the ability to adapt to a particular situation by changing its weights and/or threshold. Various algorithms exist that cause the neuron to 'adapt'; the most used ones are the Delta rule and the back error propagation. The former is used in feed-forward networks and the latter in feedback networks.

*Architecture of neural networks: a) Feed-forward networks.* Feed-forward ANNs (figure 6.3) allow signals to travel one way only; from input to output. There is no feedback (loops) i.e. the output of any layer does not affect that same layer. Feed-forward ANNs tend to be straight forward networks that associate inputs with outputs. They are extensively used in pattern recognition. This type of organization is also referred to as bottom-up or top-down.

*b) Feedback networks.* Feedback networks can have signals travelling in both directions by introducing loops in the network. Feedback networks are very powerful and can get extremely complicated. Feedback networks are dynamic; their 'state' is changing continuously until they reach an equilibrium point. They remain at the equilibrium point until the input changes and a new equilibrium needs to be found. Feedback architectures are also referred to as interactive or recurrent, although the latter term is often used to denote feedback connections in single-layer organizations.

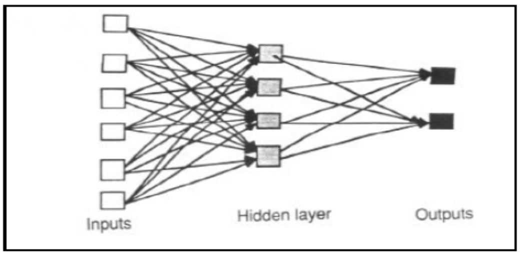



*Figure 6.3 An example of a simple feedforward network*

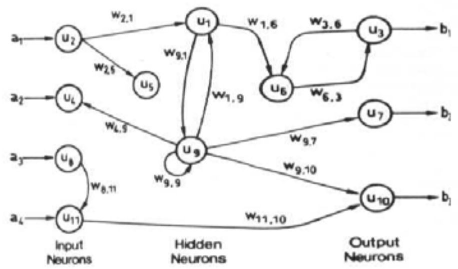



*Figure 6.4 An example of a complicated network*

*c) Network layers.* The commonest type of artificial neural network consists of three groups, or layers, of units: a layer of "**input**" units is connected to a layer of "**hidden**" units, which is connected to a layer of **"output**" units (see Figure 6.4).

- The activity of the input units represents the raw information that is fed into the network.
- The activity of each hidden unit is determined by the activities of the input units and the weights on the connections between the input and the hidden units.
- The behavior of the output units depends on the activity of the hidden units and the weights between the hidden and output units.

This simple type of network is interesting because the hidden units are free to construct their own representations of the input. The weights between the input and hidden units determine when each hidden unit is active, and so by modifying these weights, a hidden unit can choose what it represents.

We also distinguish single-layer and multi-layer architectures. The single-layer organization, in which all units are connected to one another, constitutes the most general case and is of more potential computational power than hierarchically structured multi-layer organizations. In multi-layer networks, units are often numbered by layer, instead of following a global numbering.

*d) Perceptrons.* The most influential work on neural nets in the 60's went under the heading of '*perceptrons*' a term coined by Frank Rosenblatt. The perception (figure 6.5.) turns out to be an MCP model ( neuron with weighted inputs ) with some additional, fixed, pre--processing. Units labeled A1, A2, Aj , Ap are called association units and their task is to extract specific, localized featured from the input images. Perceptrons mimic the basic idea behind the mammalian visual system. They were mainly used in pattern recognition even though their capabilities extended a lot more.

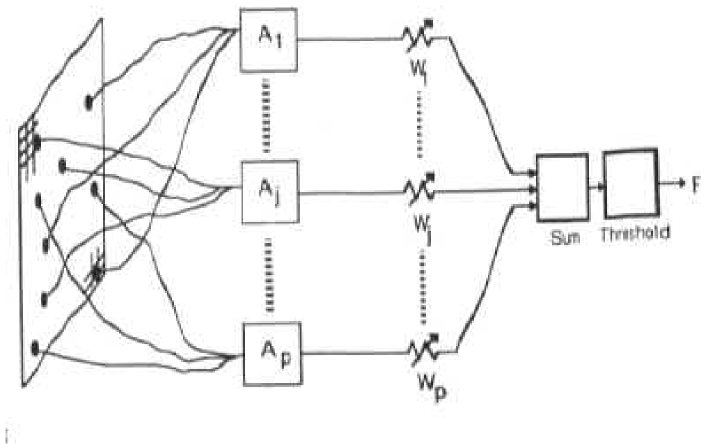



*Figure 6.5. The MCP model*

In 1969, Minsky and Papert wrote a book in which they described the limitations of single layer Perceptrons. The impact that the book had was tremendous and caused a lot of neural network researchers to loose their interest. The book was very well written and showed mathematically that *single layer* perceptrons could not do some basic pattern recognition operations like determining the parity of a shape or determining whether a shape is connected or not. What they did not realized, until the 80's, is that given the appropriate training, multilevel perceptrons can do these operations.

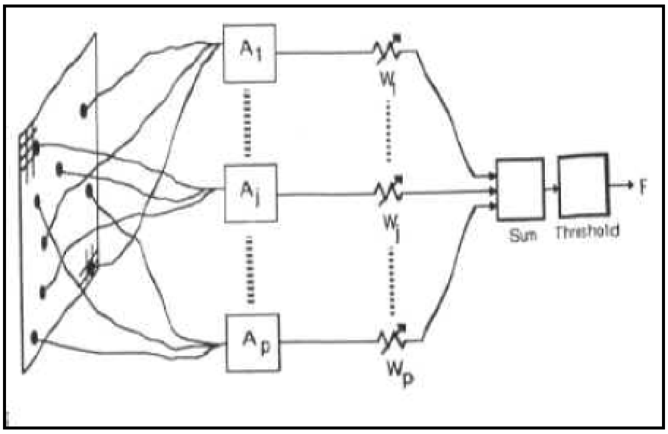



*Figure 6.6. The Multilevel Perceptron*

The memorization of patterns and the subsequent response of the network can be categorized into two general paradigms:

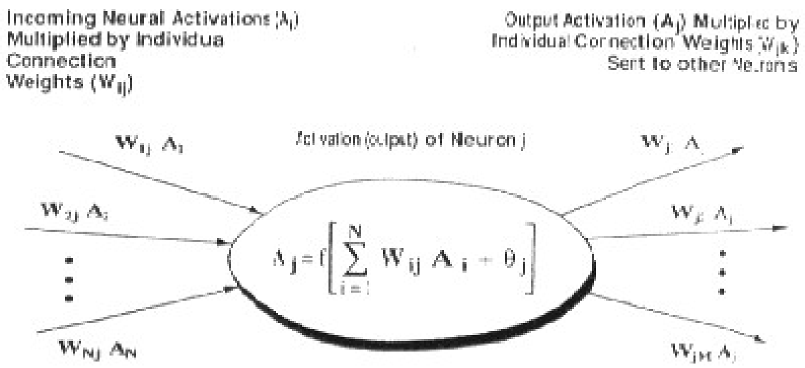
- **associative mapping** in which the network learns to produce a particular pattern on the set of input units whenever another particular pattern is applied on the set of input units. The associative mapping can generally be broken down into two mechanisms;

- *auto-association*: an input pattern is associated with itself and the states of input and output units coincide. This is used to provide pattern completition, ie to produce a pattern whenever a portion of it or a distorted pattern is presented. In the second case, the network actually stores pairs of patterns building an association between two sets of patterns;
- *hetero-association*: is related to two recall mechanisms;

✓ *nearest-neighbour* recall, where the output pattern produced corresponds to the input pattern stored, which is closest to the pattern presented, and
✓ *interpolative* recall, where the output pattern is a similarity dependent interpolation of the patterns stored corresponding to the pattern presented. Yet another paradigm, which is a variant associative mapping is classification, ie when there is a fixed set of categories into which the input patterns are to be classified.

- **regularity detection** in which units learn to respond to particular properties of the input patterns. Whereas in associative mapping the network stores the relationships among patterns, in regularity detection the response of each unit has a particular 'meaning'. This type of learning mechanism is essential for feature discovery and knowledge representation.

Every neural network possesses knowledge which is contained in the values of the connections weights. Modifying the knowledge stored in the network as a function of experience implies a learning rule for changing the values of the weights



Information is stored in the weight matrix W of a neural network. Learning is the determination of the weights. Following the way learning is performed, we can distinguish two major categories of neural networks:

- **fixed networks** in which the weights cannot be changed, i.e. ***dW/dt=0***. In such networks, the weights are fixed a priori according to the problem to solve.
- **adaptive networks** which are able to change their weights, i.e. ***dW/dt ≠ 0.***

All learning methods used for adaptive neural networks can be classified into two major categories:

- **Supervised learning** which incorporates an external teacher, so that each output unit is told what its desired response to input signals ought to be. During the learning process global information may be required. Paradigms of supervised learning include error-correction learning, reinforcement learning and stochastic learning.
  An important issue concerning supervised learning is the problem of error convergence, i.e. the minimization of error between the desired and computed unit values. The aim is to determine a set of weights which minimizes the error. One well-known method, which is common to many learning paradigms is the least mean square (LMS) convergence.
- **Unsupervised learning** uses no external teacher and is based upon only local information. It is also referred to as self-organization, in the sense that it self-organizes data presented to the network and detects their emergent collective properties. Paradigms of unsupervised learning are Hebbian learning and competitive learning .
  Ano2.2 From Human Neurones to Artificial Neuronesther aspect of learning concerns the distinction or not of a separate phase, during which the network is trained, and a subsequent operation phase. We say that a neural network learns off-line if the learning phase and the operation phase are distinct. A neural network learns on-line if it learns and operates at the same time. Usually, supervised learning is performed off-line, whereas unsupervised learning is performed on-line.

*Transfer Function.* The behavior of an ANN (Artificial Neural Network) depends on both the weights and the input-output function (transfer function) that is specified for the units. This function typically falls into one of three categories:

- *linear (or ramp)*
- *threshold*
- *sigmoid*

For ***linear units***, the output activity is proportional to the total weighted output.

For ***threshold units***, the output is set at one of two levels, depending on whether the total input is greater than or less than some threshold value.

For ***sigmoid units***, the output varies continuously but not linearly as the input changes. Sigmoid units bear a greater resemblance to real neurones than do linear or threshold units, but all three must be considered rough approximations.

To make a neural network that performs some specific task, we must choose how the units are connected to one another (see figure 6.1), and we must set the weights on the connections appropriately. The connections determine whether it is possible for one unit to influence another. The weights specify the strength of the influence. We can teach a three-layer network to perform a particular task by using the following procedure:

1. We present the network with training examples, which consist of a pattern of activities for the input units together with the desired pattern of activities for the output units.
2. We determine how closely the actual output of the network matches the desired output.
3. We change the weight of each connection so that the network produces a better approximation of the desired output.

*An Example to illustrate the above teaching procedure:*

Assume that we want a network to recognize hand-written digits. We might use an array of, say, 256 sensors, each recording the presence or absence of ink in a small area of a single digit. The network would therefore need 256 input units (one for each sensor), 10 output units (one for each kind of digit) and a number of hidden units.

For each kind of digit recorded by the sensors, the network should produce high activity in the appropriate output unit and low activity in the other output units.

To train the network, we present an image of a digit and compare the actual activity of the 10 output units with the desired activity. We then calculate the error, which is defined as the square of the difference between the actual and the desired activities. Next we change the weight of each connection so as to reduce the error. We repeat this training process for many different images of each different images of each kind of digit until the network classifies every image correctly.

To implement this procedure we need to calculate the error derivative for the weight (EW) in order to change the weight by an amount that is proportional to the rate at which the error changes as the weight is changed. One way to calculate the EW is to perturb a weight slightly and observe how the error changes. But that method is inefficient because it requires a separate perturbation for each of the many weights.

Another way to calculate the EW is to use the Back-propagation algorithm which is described below, and has become nowadays one of the most important tools for training neural networks. It was developed independently by two teams, one (Fogelman-Soulie, Gallinari and Le Cun) in France, the other (Rumelhart, Hinton and Williams) in U.S.

*The Back-Propagation Algorithm.* In order to train a neural network to perform some task, we must adjust the weights of each unit in such a way that the error between the desired output and the actual output is reduced. This process requires that the neural network compute the error derivative of the weights (**EW**). In other words, it must calculate how the error changes as each weight is increased or decreased slightly. The back propagation algorithm is the most widely used method for determining the **EW**.

The back-propagation algorithm is easiest to understand if all the units in the network are linear. The algorithm computes each **EW** by first computing the **EA**, the rate at which the error changes as the activity level of a unit is changed. For output units, the **EA** is simply the difference between the actual and the desired output. To compute the **EA** for a hidden unit in the layer just before the output layer, we first identify all the weights between that hidden unit and the output units to which it is connected. We then multiply those weights by the **EA**s of those output units and add the products. This sum equals the **EA** for the chosen hidden unit. After calculating all the **EA**s in the hidden layer just before the output layer, we can compute in like fashion the **EA**s for other layers, moving from layer to layer in a direction opposite to the way activities propagate through the network. This is what gives back propagation its name. Once the **EA** has been computed for a unit, it is straight forward to compute the **EW** for each incoming connection of the unit. The **EW** is the product of the EA and the activity through the incoming connection.

Note that for non-linear units, (see Appendix C) the back-propagation algorithm includes an extra step. Before back-propagating, the **EA** must be converted into the **EI**, the rate at which the error changes as the total input received by a unit is changed.

*Applications of neural networks. Neural Networks in Practice.*

Given this description of neural networks and how they work, what real world applications are they suited for? Neural networks have broad applicability to real world business problems. In fact, they have already been successfully applied in many industries.

Since neural networks are best at identifying patterns or trends in data, they are well suited for prediction or forecasting needs including:

- *sales forecasting*
- *industrial process control*
- *customer research*
- *data validation*
- *risk management*
- *target marketing*

But to give you some more specific examples; ANN are also used in the following specific paradigms: recognition of speakers in communications; diagnosis of hepatitis; recovery of telecommunications from faulty software; interpretation of multimeaning Chinese words; undersea mine detection; texture analysis; three-dimensional object recognition; hand-written word recognition; and facial recognition.

*Electronic noses.* ANNs are used experimentally to implement electronic noses. Electronic noses have several potential applications in telemedicine. Telemedicine is the practice of medicine over long distances via a communication link. The electronic nose would identify odours in the remote surgical environment. These identified odours would then be electronically transmitted to another site where an door generation system would recreate them. Because the sense of smell can be an important sense to the surgeon, telesmell would enhance telepresent surgery.

*Instant Physician.* An application developed in the mid-1980s called the "instant physician" trained an autoassociative memory neural network to store a large number of medical records, each of which includes information on symptoms, diagnosis, and treatment for a particular case. After training, the net can be presented with input consisting of a set of symptoms; it will then find the full stored pattern that represents the "best" diagnosis and treatment.

*Neural Networks in business.* Business is a diverted field with several general areas of specialization such as accounting or financial analysis. Almost any neural network application would fit into one business area or financial analysis.

There is some potential for using neural networks for business purposes, including resource allocation and scheduling. There is also a strong potential for using neural networks for database mining, that is, searching for patterns implicit within the explicitly stored information in databases. Most of the funded work in this area is classified as proprietary. Thus, it is not possible to report on the full extent of the work going on. Most work is applying neural networks, such as the Hopfield-Tank network for optimization and scheduling.

*Marketing.* There is a marketing application which has been integrated with a neural network system. The Airline Marketing Tactician (a trademark abbreviated as AMT) is a computer system made of various intelligent technologies including expert systems. A feed forward neural network is integrated with the AMT and was trained using back-propagation to assist the marketing control of airline seat allocations. The adaptive neural approach was amenable to rule expression. Additionally, the application's environment changed rapidly and constantly, which required a continuously adaptive solution. The system is used to monitor and recommend booking advice for each departure. Such information has a direct impact on the profitability of an airline and can provide a technological advantage for users of the system. [Hutchison & Stephens, 1987]. While it is significant that neural networks have been applied to this problem, it is also important to see that this intelligent technology can be integrated with expert systems and other approaches to make a functional system. Neural networks were used to discover the influence of undefined interactions by the various variables. While these interactions were not defined, they were used by the neural system to develop useful conclusions. It is also noteworthy to see that neural networks can influence the bottom line.

*Credit Evaluation.* The HNC company, founded by Robert Hecht-Nielsen, has developed several neural network applications. One of them is the Credit Scoring system which increase the profitability of the existing model up to 27%. The HNC neural systems were also applied to mortgage screening. A neural network automated mortgage insurance underwriting system was developed by the Nestor Company. This system was trained with 5048 applications of which 2597 were certified. The data related to property and borrower qualifications. In a conservative mode the system agreed on the underwriters on 97% of the cases. In the liberal model the system agreed 84% of the cases. This is system run on an Apollo DN3000 and used 250K memory while processing a case file in approximately 1 sec.

*The back-propagation Algorithm - a mathematical approach.*

Units are connected to one another. Connections correspond to the edges of the underlying directed graph. There is a real number associated with each connection, which is called the weight of the connection. We denote by $W_{ij}$ the weight of the connection from unit $u_i$ to unit $u_j$. It is then convenient to represent the pattern of connectivity in the network by a weight matrix W whose elements are the weights $W_{ij}$. Two types of connection are usually distinguished: excitatory and inhibitory. A positive weight represents an excitatory connection whereas a negative weight represents an inhibitory connection. The pattern of connectivity characterizes the architecture of the network.

The computing world has a lot to gain from neural networks. Their ability to learn by example makes them very flexible and powerful. Furthermore there is no need to devise an algorithm in order to perform a specific task; i.e. there is no need to understand the internal mechanisms of that task. They are also very well suited for real time systems because of their fast response and computational times which are due to their parallel architecture. Neural networks also contribute to other areas of research such as neurology and psychology. They are regularly used to model parts of living organisms and to investigate the internal mechanisms of the brain. Perhaps the most exciting aspect of neural networks is the possibility that some day '*conscious*' networks might be produced. There is a number of scientists arguing that consciousness is a 'mechanical' property and that '*conscious*' neural networks are a realistic possibility. Finally, I would like to state that even though neural networks have a huge potential we will only get the best of them when they are integrated with computing, AI, fuzzy logic and related subjects.

Among the many neural networks software, we should note the package STATISTICA – *Neural Networks*, which has a large arsenal of statistical methods, implemented the powerful genetic algorithm and all kinds of neural networks.

The mathematical apparatus of neural networks software enables not only to obtain the desired parameters in discrete form, but also have a functional representation of the output parameter input provided immutability other, that is an expression of qualitative relations between them. As we have noted above, due to its versatility, the network can be used to determine the optimal structure portfolios and investment, it can forecast the bankruptcies of financial and industrial structures, define the credit risk, creditworthiness of borrowers, forecasting inflation, exchange rates, liquidity of commercial banks, tax revenues budget and other economic indicators. For example, if a bank loan has a certain base of knowledge about the person who applied for a loan. This may be her age, education, occupation, property and more. The analyst can determine the essential characteristics of the client, make the "learn" of neural network on this basis, and assign to the client a category of credit risk. To quantify the alternatives we can also use discriminant or cluster analysis.

To optimize the investment process, we should have the available simulation model, which allow the determine market demand in line with market conditions and the economy of the region.

"Identify the function of demand," it is quite a difficult task because it has the fractal properties. However, we can determine the econometric dependence between the most important macro- and microeconomic indicators for modeling the economic situation of the region. We can use the neural network approach for the solution of problems of this type.

The indicators which describe and simultaneously determine the regional market for a given type of products or services come to the input of the neural network. Outlined input information can be divided into several blocks: Block micro and macro economic and social indicators in the region; data block advertising costs; block of data that characterize the price and quality of services; block data on market conditions and block statistics inner nature. For practical implementation of this approach using a neural network we need to determine: input and output variables, principle of operation and appearance of the network.

Performing this procedure will allow by changing prices and advertising costs to achieve the necessary indicators of demand in accordance with the accepted strategy. In addition, the neural network makes it possible to determine quantitative values for the input parameters.

The above method is very effective for using in certain marketing activities in specific markets. At the same time, it help minimize the costs of advertising and a maximum external effect of the magnitude of demand.

Neural networks are a good software product to solve the problem of forecasting the investment potential of the region. It can be defined as the ability to obtain the maximum possible amount of the investment component of a gross regional product, which is implemented through the use of the investment factor of economic growth.

This problem has the peculiar special features: a large amount of input data; incompleteness or excess data, their noisy and partial contradiction; lack of clear formal prediction algorithm. To solve this problem it is necessary to build a model that analyze the information on admission quantify the potential to detect patterns in it, take into account the heterogeneity and uncertainty of data and perform forecast.

For practical implementation of the developed model to solve the problem: -

create an information base and on its basis to allocate the training and test sample - hold sample preprocessing, normalization, data encryption, network design (choice of network topology, activation function of neurons, learning algorithm) ; - make the learning network based on retrospective data; - estimate the functioning of the model and the quality of learning network.

Tools neural networks can be used for the procedures evaluation of performance the liquidity and reliability of commercial banks based on the integral index (rule-making). To find this figure in neural networks performed following selection procedure: - coding system input values or set them - topology or network architecture, ie the number and structure relationships (inputs, layers, outputs); - Activation function; - Learning algorithm network. For rule-making that defines the category of bank ( "reliable" or "problematic"), using the procedure "classification with a teacher" are several scenarios depending on the goal: - the prediction of the criterion of troubled banks National Bank of Ukraine; - Predicting bank failure; - The choice of reliable agents in lending counter. Here the main role for the study sample, which is formed for each of the above cases, by its own rules based on regulatory and statistical base.

Another application of neural networks we can use to the tax and customs system. For example, to select of taxpayers as the candidates for tax checks we should file on the input of learning network a sequence of vectors, components of which are the parameters of these taxpayers. The neural network selects those taxpayers that have the same characteristics as a training sample. The algorithm is constructed so that it will split the tax returns in respect of checks in two classes: 1) *can give the great additional charges*; 2) *Additional charges are unlikely*.

To train neural network of recognition classes declarations, one can use training file that contains information on the results of previous inspections and data from tax returns audited taxpayers. After the procedure, network learning, it can be used for the classification of tax returns, for which these checks are not carried out. Tax returns selected network as a potentially productive in the future be considered an expert analyst, who makes the final decision: whether to conduct or not and if necessary, on what issues need priority attention. Outlines the method can be used as a method to identify links between the values that appear in specific locations declarations, tax evasion probabilistic, ie, as a preliminary selection procedure for econometric modeling processes through taxation regression and discriminant analysis. The same procedure simulation using neural networks can be used for quantitative analysis and forecasting economic performance in customs.

*Example 6.3. Modeling by using the neural networks the Ukrainian bond index (UB) (according to Example 3.6)*

We examine a daily dynamic from the 02.01.2013 to 30.05. 2014 (Figure 6.7)

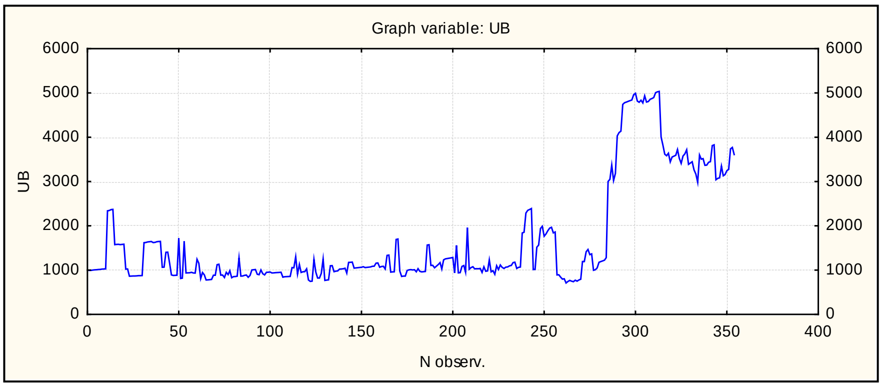



*Figure 6.7. Daily dynamics of the Ukrainian bond index (UB) from 02.01.2013 to 30.05. 2014.*

To set up the neural network models ( to form  the training , reference and test samples), we choose the dynamics series $UB1$, i.e., the first level 334 dynamics series of *index* $UB_t$ . With the "*master making*" module of "Neural Networks" from STATISTICA, it searches the most optimal model of the dynamic range. Here is an excerpt of this process (the results of the top 5 models) as the following table in which the first column is a level modeled time series, and the second is error of models):

*Table 6.1. The search results optimal models (fragment)*

| | Residuals (1-5 ) (Table.sta) | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | ub1.1 | Res.ub1.1 | ub1.2 | Res.ub1.2 | ub1.3 | Res.ub1.3 | ub1.4 | Re |
| 1 | 1033,535 | 33,535 | 1028,040 | 28,040 | 1000,004 | 0,00359 | 1000,003 | 0 |
| 2 | 1034,768 | 32,637 | 1029,344 | 27,213 | 1002,135 | 0,00418 | 1002,133 | 0 |
| 3 | 1037,783 | 30,453 | 1032,536 | 25,206 | 1007,335 | 0,00520 | 1007,330 | -0 |
| 4 | 1040,896 | 28,217 | 1035,831 | 23,153 | 1012,684 | 0,00565 | 1012,676 | -0 |
| 5 | 1042,422 | 27,128 | 1037,448 | 22,153 | 1015,300 | 0,00565 | 1015,292 | -0 |
| 6 | 1045,776 | 24,751 | 1040,999 | 19,974 | 1021,030 | 0,00519 | 1021,023 | -0 |
| 7 | 1046,781 | 24,043 | 1042,064 | 19,326 | 1022,743 | 0,00493 | 1022,737 | -0 |
| 8 | 1050,808 | 21,227 | 1046,328 | 16,748 | 1029,584 | 0,00334 | 1029,582 | 0 |
| 9 | 1052,324 | 20,175 | 1047,935 | 15,786 | 1032,151 | 0,00252 | 1032,152 | 0 |
| 10 | 1053,852 | 19,119 | 1049,553 | 14,821 | 1034,734 | 0,00157 | 1034,736 | 0 |
| 11 | 2204,845 | -142,139 | 2268,063 | -78,921 | 2346,602 | -0,38247 | 2346,808 | -0 |
| 12 | 2211,985 | -141,218 | 2275,470 | -77,733 | 2352,991 | -0,21203 | 2353,054 | -0 |
| 13 | 2234,202 | -138,285 | 2298,499 | -73,988 | 2372,785 | 0,29812 | 2372,428 | -0 |
| 14 | 2241,561 | -137,292 | 2306,121 | -72,732 | 2379,312 | 0,45943 | 2378,825 | -0 |
| 15 | 1441,908 | -140,562 | 1462,924 | -119,546 | 1581,918 | -0,55258 | 1582,375 | -0 |
| 16 | 1445,503 | -141,290 | 1466,759 | -120,034 | 1586,215 | -0,57745 | 1586,699 | -0 |
| 17 | 1448,973 | -141,984 | 1470,461 | -120,495 | 1590,360 | -0,59645 | 1590,865 | -0 |
| 18 | 1443,120 | -140,809 | 1464,217 | -119,712 | 1583,368 | -0,56155 | 1583,834 | -0 |
| 19 | 1446,625 | -141,516 | 1467,957 | -120,184 | 1587,557 | -0,58414 | 1588,048 | -0 |
| 20 | 1450,162 | -142,219 | 1471,730 | -120,651 | 1591,780 | -0,60181 | 1592,291 | -0 |
| 21 | 1051,722 | 20,592 | 1047,297 | 16,167 | 1031,134 | 0,00286 | 1031,133 | 0 |
| 22 | 1053,271 | 19,520 | 1048,937 | 15,187 | 1033,752 | 0,00194 | 1033,753 | 0 |
| 23 | 961,369 | 91,896 | 951,835 | 82,363 | 869,465 | -0,00713 | 869,475 | 0 |
| 24 | 962,453 | 90,929 | 952,977 | 81,454 | 871,514 | -0,00975 | 871,524 | 0 |

As a result of runs (eras) through  iterative the Backpropagation method**,** we have found the optimal model that showed the least error in training, control and test samples. That model has been a three-layer perceptron with one hidden layer, consisting of five neurons and radial basis function activated neurons $\psi(s) = \exp(-ks^2)$, where $s \in R$, and $k > 0$ is signal amplification ratio (compression ratio-stretching)

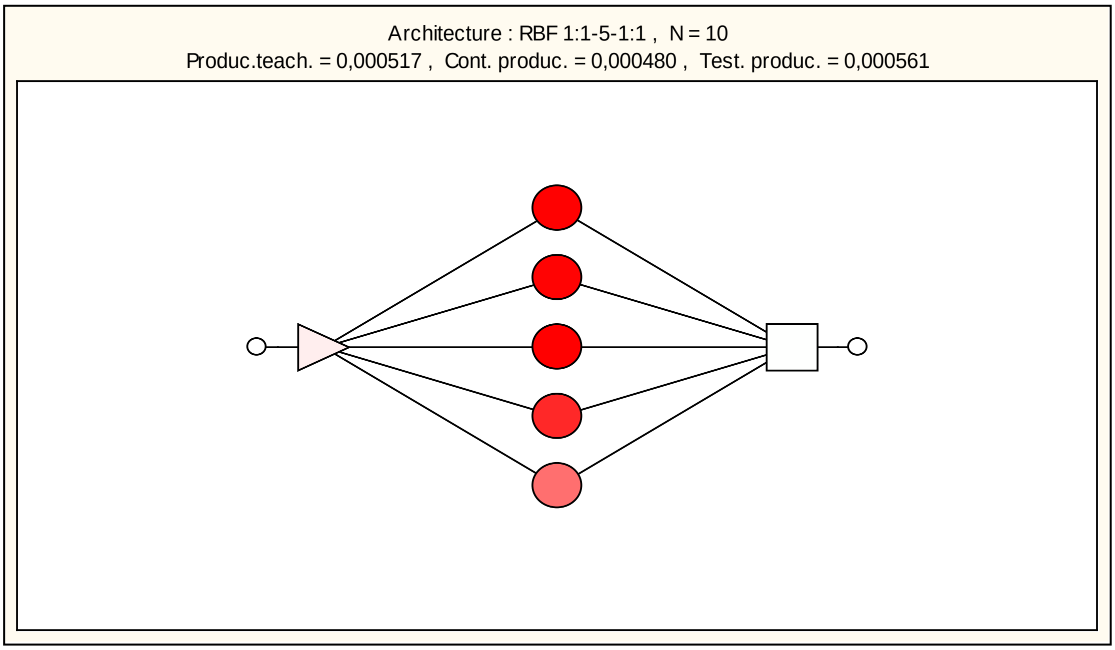



*Figure 6.8. The architecture of neural network model time series $UB_t$*

For clarity, we have plotted the output of time series and the results of modeling onto one chart:

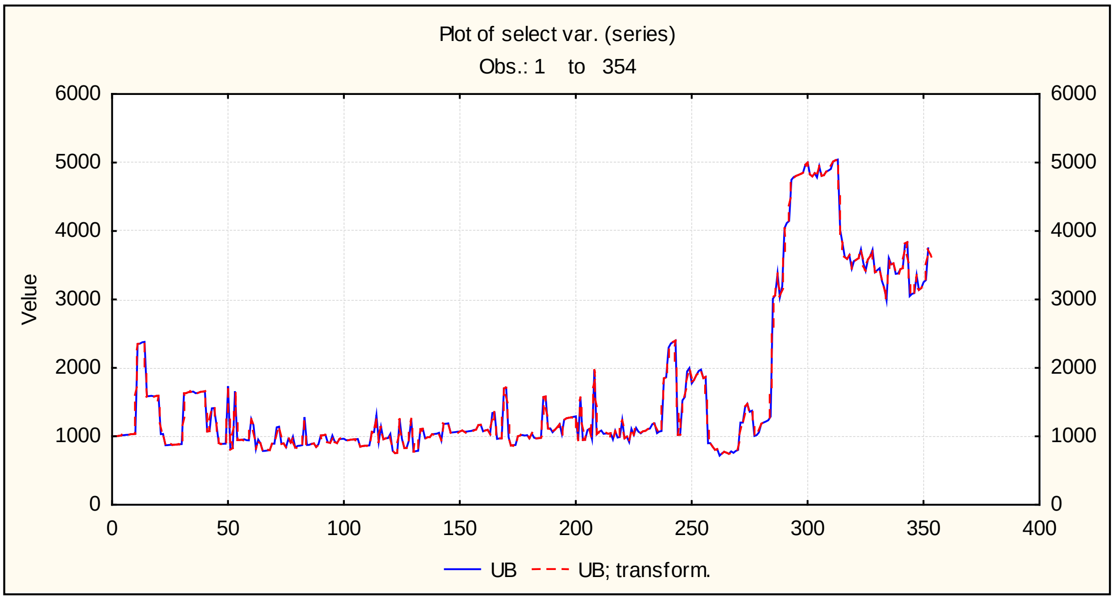



*Figure 6.9. Diagrams of time series (original and optimal model)*

Visually, we can observe a high enough precision built models, and prediction of results for this model better than the model ARIMA (1,1,1) with the intervention:

*Table 6.2. Forecasts for ARIMA model, neural network and actual*

| *Forecasts (ARIMA)* | *Forecasts (neural network)* | *Actual* |
|---|---|---|
| *3539,292* | *3509,562* | *3511,1209* |
| *3506,971* | *3526,361* | *3525,3314* |
| *3491,240* | ***3481,346*** | ***3473,5498*** |
| *3484,040* | *3480,745* | *3479,4659* |
| *3481,229* | *3441,269* | *3440,7611* |
| *3480,674* | *3460,074* | *3457,0099* |
| *3481,279* | *3485,379* | *3487,6691* |
| *3482,482* | *3452,982* | *3455,5293* |
| *3483,991* | *3453,591* | *3456,2835* |
| *3485,658* | *3480,058* | *3479,0848* |

As we can see, the relative error of ARIMA forecasting with the intervention is:

$$\delta_{\max} = \frac{\max\left|UB_t^{pred} - UB_t^{fact}\right|}{\min UB_t^{fact}} \le \frac{40}{3440,7611} = 0,011,$$

i.e., it is less than 1.1%, then for the neural network:

$$\delta_{\max} = \frac{\max\left|UB_t^{pred} - UB_t^{fact}\right|}{\min UB_t^{fact}} \le \frac{8}{3440,7611} = 0,002,$$

i.e., it is less than 0,2%.

Again, if we have a relatively small amount of sample data for modeling by neural network, it can lead to so-called "retraining effect".

*Example 6.4. Forecasting revenues by using the multivariate models and neural network technology*

For an effective sustainable socio-economic development of Ukraine is important the increased revenue to the budgets of all levels to facilitate the development and stability of the financial system and increase economic security.
And therefore, the question of improving the efficiency of the revenue side of the budget, which is directly related to revenue forecasting and allocation of budget allocations for different types of taxes in the regions with taking into account sectoral characteristics have the particular relevance.

Note that the main feature of macroeconomic forecasting (which include the revenues for certain types of taxes) is that any macro variable is defined as set of components and tends to lower of the deviation from the average than each of its components. Accordingly, even in the presence of a representative sample of sustainable development of the economy and the principles of the tax system to reduce the total error calculations should not attempt to determine potential tax capacity of individual taxpayers and simulate of revenues for each type of tax in the regions or Ukraine as a whole.

Given the nonstationarity and lack of statistical homogeneity for corresponding time series in the Ukrainian economy, it has ruled out the possibility of adequate forecasting based on econometric models. In particular, when forecasting revenues were decided to abandon the use of ARIMA or other factor models. we made the choice to study the many of factor forecasting. And in order to increase the statistical sample, we analyze simultaneously many different time series of income tax for different regions of Ukraine due to a small number of statistics. Accordingly, in this case, exercise time series prediction by extrapolating them will not succeed.

To carry out research, primarily we build an econometric model of forecasting tax revenues regression type. It is appropriate to have the ability to compare the quality prediction of more complex models (although above it was proved problematic use of econometric models).

Then we will choose the nonlinear forecasting models of tax revenues that do not require for random variables and explanatory variables the hypothesis about normal distribution and it will be based on the methods of the theory of neural networks.

We will construct appropriate economic and mathematical models and conduct a comparative analysis of their effectiveness on the example the statistical data of VAT of regions. Originally, we have built a significant number of econometric models forecasting VAT revenues, which were based on different sets of explanatory variables and, given the lack of statistical information, tuned by the data on these variables for different years. That is, for some models, we have taken only the statistics for 2006 (those explanatory variables for which statistics were available only for 2006). Therefore we have recreated the VAT revenues in the corresponding quarter of 2007.

Some models were built on statistics for 2004-2005., made the forecasting of VAT revenues in 2005-2006., Respectively, we also built economic and mathematical models for other time intervals. All built econometric models showed determination coefficient at level 0.15 and low, and corresponding values of F-test (which is considerably lower than the corresponding tabular value). It indicates to lack of suitability for forecasting revenues of VAT.

A normalized mean square error of forecast for these models was in the range of 0.9 to 1.0 (the weather was not much better than the normal mean).

Comparative analysis the accuracy of VAT revenues based on the several explanatory variables prior periods has shown that the economic and mathematical models are built based on statistical data from 2002 to 2005 years for indicators ' "VAT revenue " ," Loss of tax benefits "," VAT Refund " "The volume of exports", "imports", "Consumer price Index", "Collection of VAT" and "Overpayment" are more adequate.

In Figure 6.10 shows the actual number of relative changes in VAT revenues by region (solid line) and playback (dashed line) for the first quarter of next year (2003 to 2006) using the specified econometric model.

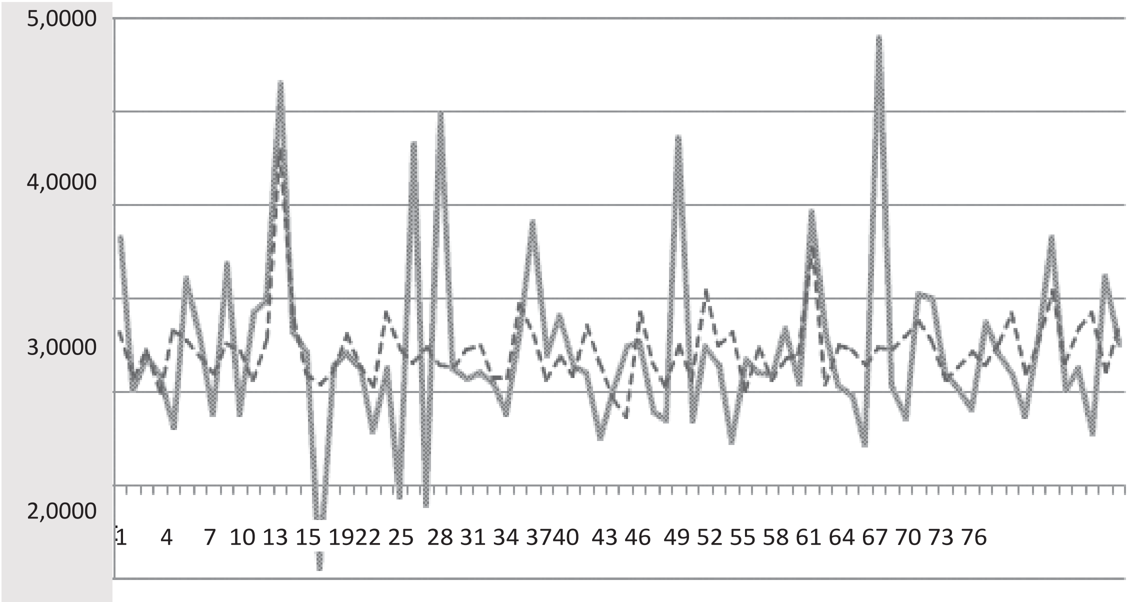



*Figure 6.10. Reproduction of relative changes in VAT revenues by region in the first quarter using a linear model*

On Figure 6.10, we can see how the forecast of VAT revenues received through regression models, at odds with the real data.

A similar situation with the prediction, as shown in figure 6.10, we can observe for other quarters. Almost all built econometric models were unable to reproduce VAT revenues through values of a number of different indicators of the previous period. This can be explained by the low dependence of the output variable "VAT revenue" of change of informative factors. High correlation coefficients between input and output variables appear only in absolute terms, due to the general trend of growth of all financial and economic indicators related to the overall development of the economy and underlying inflation. Accordingly, the possibility of tax revenues modeling using classical econometric tools has considerable doubt.

We will make reproduction the VAT revenues using economic and mathematical models built using neural networks *perceptron* type on the same statistics and regression models to selected explanatory variables.

Forming patterns on neural networks comes down to choosing the optimal configuration of neural networks (determining the number of inner layers networks and neurons in these layers), the type of activation functions for different neurons and substantiation of recommendations for pre-processing of data. For the design of neural network models, we select explanatory variables by which built the latest econometric models, as they have the complete statistics for all regions. Thus the amount of training set is particularly important for the correct settings neural network to avoid the effect of retraining. To build a neural network model for forecasting the VAT revenues we select such a neural network for each quarter, that most accurately reflects the statistics, but it has a simple structure and avoids the effect of retraining.

As a result of experiments with the simulation proceeds of VAT using the tools neural network model, we have defined the most adequate model with the following configuration: fully connected neural network of multilayer perceptron type with one inner layer is consisting of six neurons; the first layer consists of eight neurons by the number of model input variables and one output neuron layer. Exactly the first nonlinear model based on neural networks tool showed the high fidelity reproduction of output variable based on the set of input parameters. This is confirmed by the values of normalized mean forecast error $\sigma_{norm}$=0,095.

We can see the result the reproduction of indicator " VAT revenues" for the first quarter using this model in Figure 6.11:

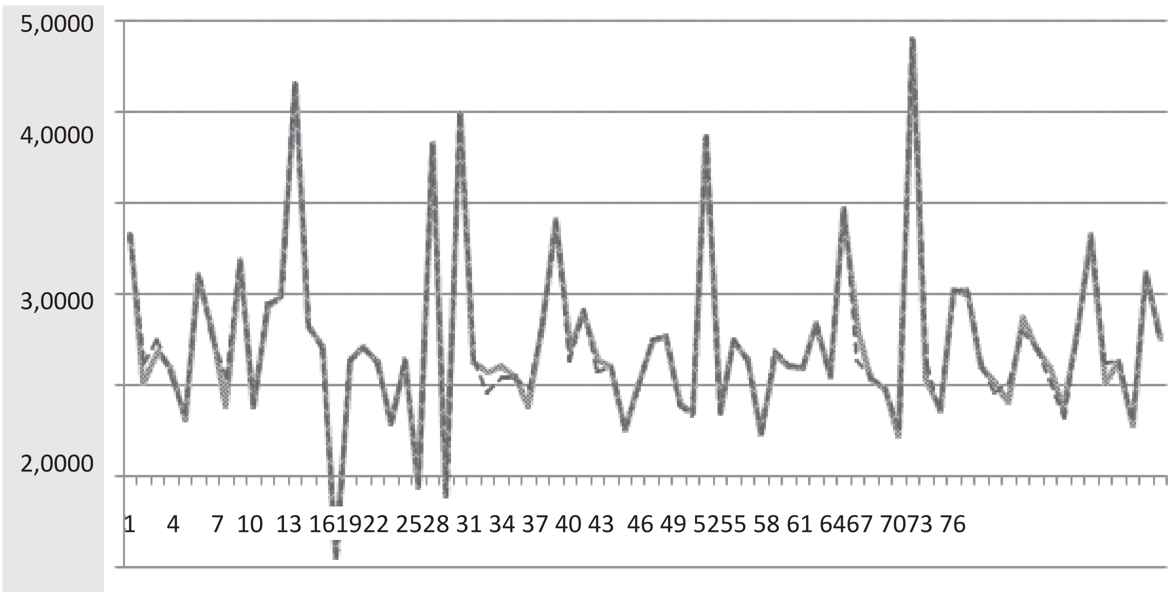



*Figure 6.11. The reproduction of relative changes in VAT revenues in the first quarter with the use of neural network models*

The results of reproduction the VAT revenues with using neural networks methods confirmed the high efficiency of the instruments and the feasibility of its use for solving the problem of forecasting tax revenues.

*Example 6.5. The prediction of dynamics of stock index S & P 500 by using neural networks and regression models.*

The price quotations are the most effective way of allocation of scarce resources, including capital. That is why pricing or price quotation in financial markets is actual topic to study by scientists and traders.

We put the task for comparative analysis of two modeling methods and forecasting Stock Exchange dynamics through regression and neural network on the example of index S & P 500, it is an American stock market index based on the market capitalizations of 500 large companies having common stock listed on the NYSE or NASDAQ. The S&P 500 index components and their weightings are determined by S&P Dow Jones Indices.

To build models, one should use the monthly data on the value of the index from the 1965 year, To build models, one should use the monthly data on the value of the index from the 1965 year, because since that time it began widespread use CAPM model of interpretation for market dynamics, which greatly affected the pricing of financial instruments. Testing was conducted at different periods of study neural networks and building regressions.

We have selected the following variables as exogenous (explanatory) factors:

*Table 6.3. The exogenous variables*

| *Variable* | *Variable description* |
|---|---|
| *D12,$* | *The dividends of the companies included in the S & P500 index at the past 12 months* |
| *E12,$* | *The profit of the companies included in the S & P500 index at the past 12 months* |
| *b/m* | *Value of book and market value of the company* |
| *Tbl* | *T-billrates - US government bond quotes* |
| *AAA* | *Yield bonds of companies with different credit rating from AAA to BAA* |
| *BAA* | |
| *Lty* | *Return on long-term US government bonds* |
| *Ntis* | *Net Equity Expansion (12-month, moving), which shows the ratio of new shares of the total market capitalization* |
| *Rfree* | *Risk-free interest rate* |
| *Infl* | *The value of inflation* |
| *Ltr* | *Return on long-term bonds (private)* |
| *corpr* | *Yield on Corporate Bonds* |
| *svar* | *Variance of index* |

The results forecasting options are shown in Table 6.4, which contains the real value of the index and its assessment:

*Table 6.4. The real value of the index and assessment of models*

| *Date* | *The actual values* | *Price assessment / yield assessment* | | | |
|---|---|---|---|---|---|
| | | The neural network (fact). | Autoregression | Factor regression | Factor regression(sta.) |
| *2007.01* | *1438,24* | *1418,7516* | *1222,023417* | *1548,752* | *0,218564452* |
| *2007.02* | *1406,82* | *1376,5382* | *1224,818591* | *1560,65* | *-5,627317311* |
| *2007.03* | *1420,86* | *1364,4893* | *1236,287678* | *1614,723* | *-0,803467346* |
| *2007.04* | *1482,37* | *1405,404* | *1253,613319* | *1618,68* | *-0,605499206* |
| *2007.05* | *1530,62* | *1414,0254* | *1301,996449* | *1636,744* | *-1,269921051* |
| *2007.06* | *1503,35* | *1391,996* | *1344,256816* | *1656,208* | *0,831851476* |
| *2007.07* | *1455,27* | *1386,0798* | *1310,259959* | *1604,523* | *0,213324461* |
| *2007.08* | *1473,99* | *1296,1634* | *1335,028306* | *1631,025* | *-4,745558672* |
| *2007.09* | *1526,75* | *1377,2197* | *1309,505706* | *1633,429* | *0,89041493* |
| *2007.10* | *1549,38* | *1411,1852* | *1283,173393* | *1561,379* | *0,648988383* |
| *2007.11* | *1481,14* | *1322,1271* | *1321,607034* | *1548,035* | *0,902918155* |
| *2007.12* | *1468,36* | *1419,9702* | *1321,418471* | *1500,951* | *0,354602651* |

The following series of experiments were conducted with neural networks and the multilayer structure of neurons with different activation functions based on factor dependence. Neural network 25_10_10 (3 hidden layers with 25, 10 and 10 neurons and sigmoid activation function) have shown the best results. This allows the neural network to predict trend changes in the index and provides quarterly forecast a slight error.

Building a regression, we should into account that the original data should be fixed. Therefore, we use first differences and then we built regression. Also, we should say that the dependent variable is not the value of an index, only its yield. Forecasting results presented in figures below:

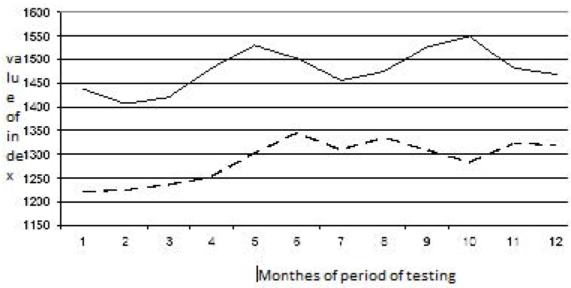



*Figure 6.12. The forecast neural network*

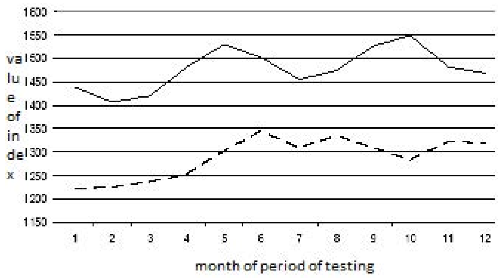



*Figure 6.13. The forecast autoregression*

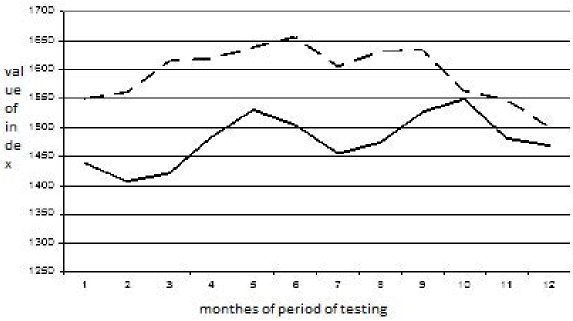



*Figure 6.14. The forecast factor regression*

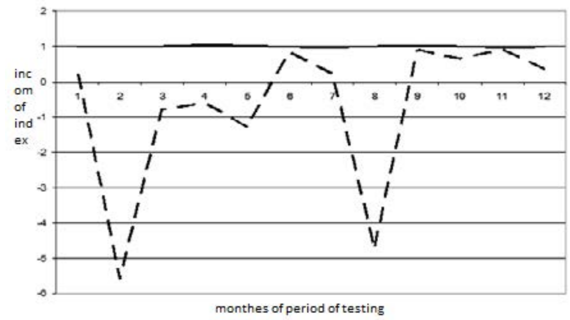



*Figure 6.15. Forecast factor regression with stationary data*

We can see are the results of comparing models based on the coefficient of determination and quality the forecast of trend in the following table:

*Table 6.5. Comparison of forecasting techniques based on the coefficient of determination and expert assessment of the quality trend prediction*

| *Model* | *$R^2$ on 12 months* | *$R^2$ on 3 months* | *trend prediction (0 - 5)* |
|---|---|---|---|
| *Neural network 25_10_10* | *-5,61* | *0,99* | *4* |
| *Autoregression* | *-19,94* | *0,97* | *2* |
| *Factor regression* | *-7,89* | *0,98* | *3* |
| *Factor regression(sta.)* | *-10448,78* | *-27,13* | *0* |

Thus, the neural network on base the factor dependence, at the time period 2003-2006, has shown better indications. It coped with the prediction trend with sufficient accuracy.

Compare the model on trading criteria. The evaluation procedure is as follows: if the model is showed a price increase from $ 1 to $ 3, then we open a long position (to buy); if the price has shown a real change with $ 1 to $ 2, then the profit has been 2-1 = 1 dollar.

*Drawdown* - the difference between the maximum and the minimum of the capital curve, expressed in absolute terms has been elected as a risk indicator. In a comparative analysis also evaluated the predictive accuracy of the trend.

*Table 6.6. The comparison of regression and neural network models based on training criteria*

| *Comparison criteria* | *Neural network* | *Factor regression* | *Autoregression* |
|---|---|---|---|
| *on 12 months* | | | |
| *Trend* | *63,64%* | *72,73%* | *54,55%* |
| *Drawdown* | *231,24* | *274,48* | *163,33* |
| *Standard deviation.* | *71,11* | *89,01* | *52,47* |
| *Earnings* | *218,46* | *243,06* | *1,06* |
| Return | *218,46%* | *243,06%* | *0,01* |
| *Economic earnings* | *0,94* | *0,89* | *0,0065* |
| *on 3 months* | | | |
| *Trend* | *66,67%* | *66,67%* | *66,67%* |
| *Drawdown* | *78,89* | *75,55* | *75,55* |
| *Standard deviation.* | *29,31* | *28,42* | *28,42* |
| *Earnings* | *78,89* | *44,13* | *44,13* |
| Return | *78,89%* | *44,13%* | *0,44* |
| *Economic earnings* | *1,00* | *0,58* | *0,584* |

As can be seen from Table 6.6, factor regression showed better earnings on long period (for 12 months) but the neural network has won in terms of economic earnings. When considering a short period (3 months), the neural network outstripped the autoregression factor by both indicators.

At the same time, we should take into consideration Efficient-market hypothesis (EMH). We note: the neural network is a free-for-all instrument and cannot afford to get stable enormous profits. We can be found the only temporary inefficiencies with them, which, however, are quickly corrected actions by arbitrageurs with having the same information and tools for forecasting. We can say that they create the efficient-market.

In conclusion, we note that the regression models have a greater explanatory ability. They allow you to identify patterns in an explicit form.

Neural networks do not provide an answer to the question "why?". But our study shows that they are well suited to assess the dynamics of the stock due to its flexibility and ability to find nonlinear patterns. The apparent advantage of neural networks is that they do not require a large sample size to correctly forecast.

## Bibliography


1. Box G. E., Jenkins G. M., Reinsel G. C., Ljung G. M. Time Series Analysis: Forecasting and Control.-N.Y.: John Wiley & Sons, 2015.

2. Balisteri E.J., Rutherford T.F. Computing General Equilibrium theories of Monopolistic Competition and Heterogeneous Firms // Chapter prepared for the Handbook of Computable General Equilibrium Modeling edited by Dixin P.B. and Jorgenson D.W., 2011.

3. Callan R. The Essence of Neural Networks.- London: Prentice Hall Europe, 1999.

4. Diamantaras K. I. Neural Networks and Principal Component Analysis.- CRC Press, 2002.

5. Draper N.R., Smith H. Applied Regression Analysis. 3rd ed.- New York: John Wiley&Sons,1998.

6. Engle R.F. Autoregressive conditional Heteroskedasticity with estimates of the variance of the United Kingdom inflation / R.F. Engle // Econometrics. – 1982. – V. 50, – P. 987– 1007.

7. Engle R.F. and Granger G.W.J. Cointegration and error-correction: Representation, estimation and testing / *Econometrica* 55: 1987, P.251-276.

8. Franke J., Hardle W. and Hafner C. Statistics of Financial Markets, 2nd Edition.- Berlin, Heidelberg, New York: Springer, 2008.

9. Franke, J., Hardle, W. and Hafner, C. Statistics of Financial Markets, 2nd Edition.-, Berlin, Heidelberg, New York: Springer, 2008.

10. Factor Analysis: Statnotes, from North Carolina State University, http://www2.chass.ncsu.edu/garson/pa765/factor.htm
Principal Components and Factor Analysis, http://www.uta.edu /faculty/ sawasthi/ Statistics/ stfacan.htm

11. Geladi, P., Kowlaski B Partial least square regression: A tutorial// Analytica Chemica Acta.- 1986. – V. 35, - P.1–17.

12. Gordon D. V. Price modelling in the Canadian fish supply chain with forecasts and simulations of the ex-vessel price of fish / Rome, FAO, 2010.

13. Gordon D.V. Optimal Lag Length in Estimating Dickey-Fuller Statistics: An Empirical Note/ *Applied Economics Letters*, 2: 1995, P.188-190.

14. Hazova D.S. Mathematcal Modeling of sustainable tourism development: dys.kand. ekon. Sciences: 08.00.13, Moscow, 2014.- 126 p.

15. Intelligent simulation technology in the information-analytical system of the State Tax Service: Monograph // com.red. L.L.Taranhul.- K .: Alerts, 2010.

16. Klein, L. R. A Textbook of Econometrics.-N.Y.: Greenwood Press, 1977.

17. Kovalyova E.M. Venture funding based on public-private partnership: candidate thesis. Econ. Sciences: 08.00.08, Odessa, 2015.- 196 p.

18. Litterman, R., Scheinkman, J. Common factors affecting bond returns // Journal of Fixed Income.- 1991, Vol. 1, - P. 54–61.

19. Mardia K. V., Kent J. T. and Bibby J. M. (1979). Multivariate Analysis.- London: Academic Press, Duluth, 1979.

20. Matskul V. M., Kirillova L. A. Modeling and forecasting of VAT on micro and macro levels // V.M.Matskul / Bulletin socioeconomic doslidzhen.- ONEU, 2014, №54, P.27-34.

21. Matskul V.M. Podvalna N.E. Economic-mathematical analysis of hotel enterprises Odessa // V.M.Matskul / - Economy: Realities time. Scientific zhurnal.- 2014.-№4 (14) .- P.53-62.

22. Nelson D.B. Conditional heteroskedasticity in asset pricing: a new approach / D.B. Nelson // Econometrics. – 1990. – V. 59. – P. 347–370.

23. Pearl J. CAUSALITY: models, reasoning, and inference. – Cambridge Univ. Press, 2000.

24. Podvalna N.E. Analytical justification efficiency of hotel enterprises: Dis candidate. Econ. Sciences: 08.00.09, Odessa, 2014.- 164p.

25. Piazalo D. Poland's Membership in the European Union: An Analysis with a Dynamic Computable General Equilibrium (CGE) Model// LICOS Discussion Papers, №8900, 2000.

26. Statistica. A systematic approach to data analysis [electronic resource] .- StatSoft Russia.[rus]- Access: http: //www.statsoft.ru.

27. Soto, G. M. Using principal component analysis to explain term structure movements: performance and stability/ Progress in Economics Research, Vol. 8, Nova Science Publishers: 2004, P. 203–225.

28. Sanchez, G. PLS Path Modeling with R / Gaston Sanchez. – Berkeley, California, 2013. – 210 p.

29. Stock, J.H. and M.W. Watson Combination forecasts of output growth in a seven-country data set // Journal of forecasting.- 2004, Vol. 23, P.405-430.

30. Tenenhaus, M. La régression PLS.- Paris: Technip.,1998.

31. The electronic textbook on statistics StatSoft, Inc. [Electron resource]. – Access to the tutorial:http://www.statsoft.ru/home/textbook/default.htm.

32. Vukolov E.A. Basic statistic analysis. Workshop on statistical methods and operations research using packages Statistica and Excel.- FORUM: INFRA-M, Moscow, 2004

33. Yerina A. M., Motoryn R. M., Holovach A. V. Statistic.- K.: KNEU, 2001.

34. Wold H. Estimation of principal components and related models by iterative least squares / H. Wold // In P.R. Krishnaiaah (Ed.), Multivariate Analysis II. . – New York: Academic Press, 1966. – P. 391-420.

35. Wold H.// In P.R. Krishnaiah (Ed.), Multivariate Analysis II.- New York: Academic Press, 1973. – P. 383-407.