MINISTRY OF EDUCATION AND SCIENCE OF UKRAINE

SIMON KUZNETS KHARKIV NATIONAL UNIVERSITY OF
ECONOMICS

STATISTICAL THINKING FOR SCIENCE ABOUT DATA

Guidelines
for independent work
of Master's (second) degree
students of speciality 122 "Computer Science"

Kharkiv
S. Kuznets KhNUE
2021

UDC 004.422.6(07.034)
S81

Compiled by: O. Rayevnyeva
V. Derykhovska

Затверджено на засіданні кафедри статистики і економічного прогнозування.

Протокол № 9 Б від 20.03.2021 р.

Самостійне електронне текстове мережеве видання

S81 **Statistical** Thinking for Science about Data [Electronic resource] : guidelines for independent work of Master's (second) degree students of speciality 122 "Computer Science" / compiled by O. Rayevnyeva, V. Derykhovska. – Kharkiv : S. Kuznets KhNUE, 2021. – 108 p. (English)

Tasks for independent work on the academic discipline and guidelines for carrying out the tasks are given to help the students gain practical skills in the use of the tools of economic and mathematical modeling in the study of complex socioeconomic processes and systems.

For Master's (second) degree students of speciality 122 "Computer Science".

UDC 004.422.6(07.034)

Introduction

The rapid development and wide application of the newest packages of applied programs and computer technology tools necessitate the formation of a specialist in business intelligence and information systems with new competences aimed at acquiring knowledge and skills in the use of econometric and mathematical modeling for the analysis of complex, mass socioeconomic phenomena and processes in various spheres of activity.

Statistical Thinking for Science about Data is one of the basic academic disciplines of the Master's program "Business Analysis and Information Systems in Entrepreneurship". The academic discipline is a response to the contemporary needs of the community. It provides students with an in-depth understanding of the business context of any socioeconomic processes and will enable them to solve the problems associated with analytical work in the IT-industry.

While studying educational material of this discipline students are involved in theoretical, practical, and independent training. In the credit-modular system of organization of the educational process, independent training is essential. The main purpose of independent training is the creation of conditions for the fullest realization of the creative potential of students through people-centred development of their abilities for research and individual activity. Moreover, the organization of the students' independent work based on the competence approach provides opportunities for personal involvement of students in the development of professional activity and the formation of their professionally significant qualities.

Carrying out tasks of independent works is aimed at the development of students' skills in the extension and deepening of theoretical knowledge and acquisition of professional competences in forecasting socioeconomic processes and modeling complex systems.

Studying this academic discipline enables students to get the ability:

to acquire practical knowledge of statistical modeling and forecasting;

to receive skills in the formation of the information space research;

to model relationships between economic processes and phenomena;

to model and predict time series in the study of the dynamics of the development of socioeconomic systems;

to identify and simulate the behavior of homogeneous complex socio-economic systems.

Content module 1. The methodological bases of statistical modeling and forecasting

Topic 1. The categorical basis of statistical modeling and forecasting

Task 1. Check the law of distribution of the variation series (Table 1).

Table 1

The dynamics of the value of fixed assets

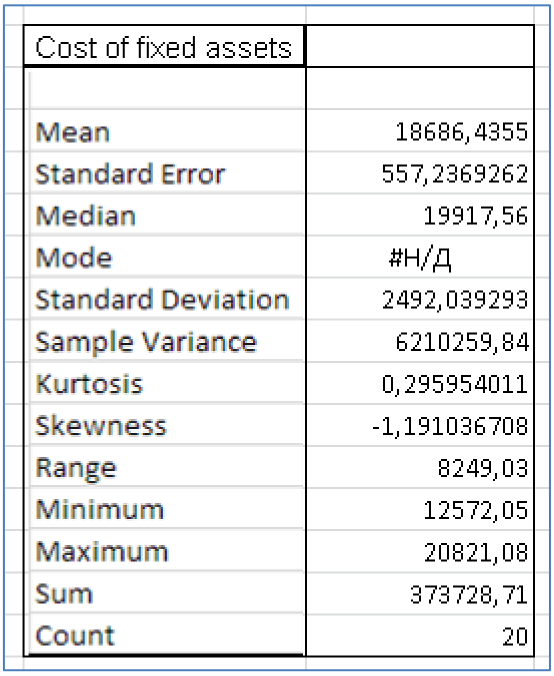
Period	Cost of fixed assets, thousand UAH
1st quarter of 2015	12 572.05
2nd quarter of 2015	14 405.39
3rd quarter of 2015	15 424.83
4th quarter of 2015	16 201.19
1st quarter of 2016	16 390.52
2nd quarter of 2016	16 598.67
3rd quarter of 2016	18 241.53
4th quarter of 2016	19 635.99
1st quarter of 2017	19 900.40
2nd quarter of 2017	19 900.10
3rd quarter of 2017	19 934.72
4th quarter of 2017	19 989.89
1st quarter of 2018	20 321.21
2nd quarter of 2018	20 333.16
3rd quarter of 2018	20 466.96
4th quarter of 2018	20 534.77
1st quarter of 2019	20 606.19
2nd quarter of 2019	20 657.29
3rd quarter of 2019	20 792.77
4th quarter of 2019	20 821.08

Guidelines

1.1. It is expedient to check statistical homogeneity of factorial features using the coefficient of variation with a limiting value of 33 %.

We can simplify verification of the law of distribution of the time series using special built-in functions of MS Excel or Statistica 10.0 applications. Let's use the add-in tool Data analysis "Descriptive statistics" MS Excel.

The calculation of statistical indicators is shown in Fig. 1.



Cost of fixed assets	
Mean	18686,4355
Standard Error	557,2369262
Median	19917,56
Mode	#Н/Д
Standard Deviation	2492,039293
Sample Variance	6210259,84
Kurtosis	0,295954011
Skewness	-1,191036708
Range	8249,03
Minimum	12572,05
Maximum	20821,08
Sum	373728,71
Count	20

Fig. 1. **The results of the calculation**

Using the calculations performed in Fig. 1, we obtain the value of the coefficient of variation for the studied variable:

$$V_x = \frac{\sigma_x}{\overline{x}} \times 100\,\% = \frac{2\,492.04}{18\,686.43} \times 100\,\% = 13.3\,\%.$$

The coefficient of variation is used to characterize the homogeneity of the studied population. The statistical population is considered to be quantitatively homogeneous if the coefficient of variation does not exceed 33 %.

The calculations confirm the hypothesis of homogeneity of the variation series.

1.2. The criterion of proximity to the normal distribution law is "the three-sigma rule" which expresses a conventional heuristic that nearly all values are taken to lie within three standard deviations of the mean. According to "the three-sigma rule", x_{min} and x_{max} should belong to the area $[\overline{x} \pm 3\sigma_x]$.

So, let's check:

$$\left[\bar{x} \pm 3\sigma_x\right] = 18\,686.43 \pm 3 \times 2\,492.04 = [11\,210.31; 21\,162.55]$$

$$\left.\begin{matrix} x_{min} = 12\,572.05 \\ x_{max} = 20\,821.08 \end{matrix}\right\} \rightarrow [11\,210.31; 26\,162.55].$$

Based on the calculations, we can conclude that the distribution of the value of fixed assets is close to the normal law of distribution.

Task 2. Using the *Basic statistics* menu of the program Statistica 10.0, calculate the main characteristics of the number of distributions (Table 2).

Table 2

The dynamics of the number of the registered unemployed in 2019

Month	The number of the registered unemployed, thousand people
January	364.3
February	367.0
March	340.7
April	311.4
May	300.9
June	287.1
July	280.8
August	275.0
September	268.2
October	259.3
November	288.9
December	338.2

Guidelines

After starting the program Statistica 10.0 and creating a new sheet with the original data (you need to create a spreadsheet with one variable (1 column) and 12 observations (rows)), in the *Statistics* menu, select *Basic statistics / Tables*. In the opened window, select *Descriptive statistics*, which will open a window for calculating the complex descriptive statistics (Fig. 2).

Next, you need to go to the *Advanced* tab and select the parameters you want to calculate – groups of indicators of the distribution center (mode, median, mean), the uniformity of distribution (range, minimum and maximum, variance, coefficient of variance, standard deviation), the distribution form (skewness, kurtosis) as shown in Fig. 3.

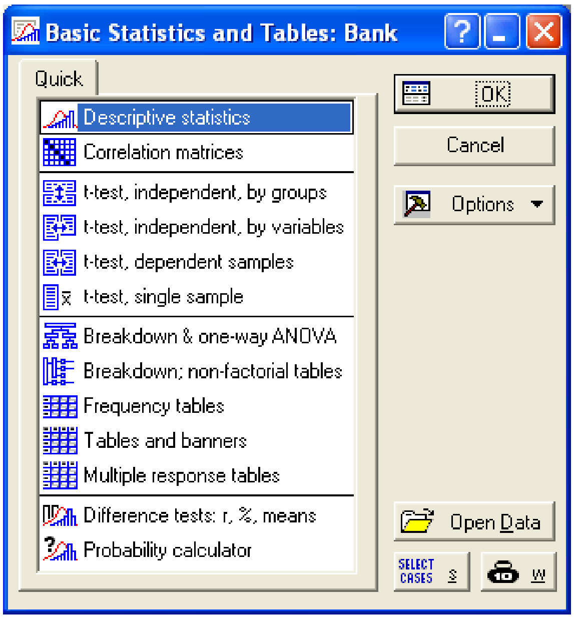
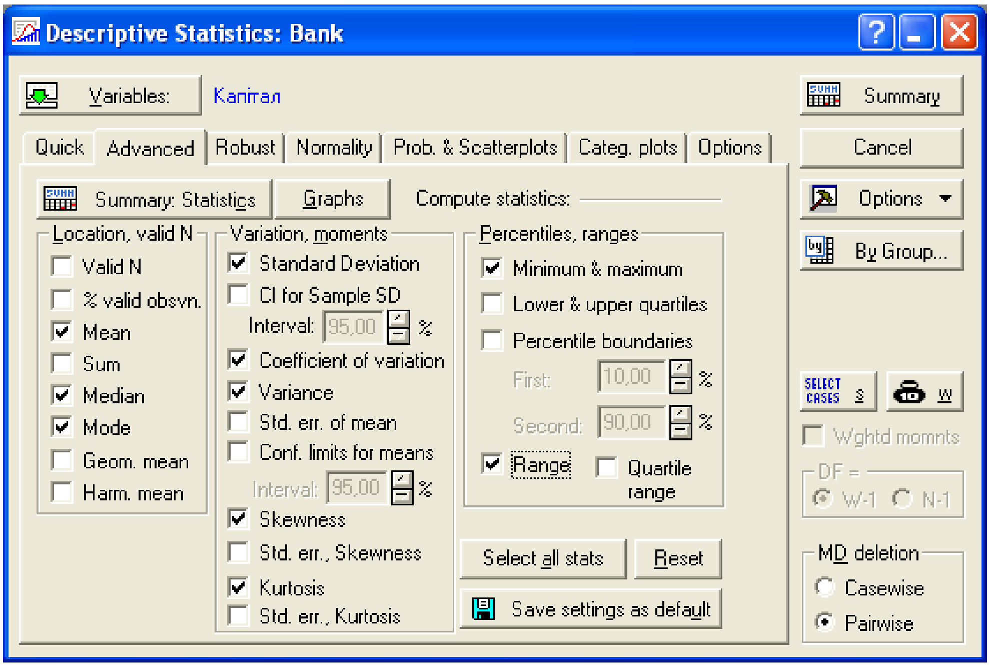

Basic Statistics and Tables: Bank
Quick
Descriptive statistics
Correlation matrices
t-test, independent, by groups
t-test, independent, by variables
t-test, dependent samples
t-test, single sample
Breakdown & one-way ANOVA
Breakdown; non-factorial tables
Frequency tables
Tables and banners
Multiple response tables
Difference tests: r, %, means
Probability calculator
OK
Cancel
Options
Open Data
SELECT CASES
S
W

Fig. 2. **The *Basic statistics* window**



Variables: Kanıran

Location, valid N
- ☐ Valid N
- ☐ % valid obsvn.
- ☑ Mean
- ☐ Sum
- ☑ Median
- ☑ Mode
- ☐ Geom. mean
- ☐ Harm. mean

Variation, moments
- ☑ Standard Deviation
- ☐ CI for Sample SD
- Interval: 95,00 %
- ☑ Coefficient of variation
- ☑ Variance
- ☐ Std. err. of mean
- ☐ Conf. limits for means
- Interval: 95,00 %
- ☑ Skewness
- ☐ Std. err., Skewness
- ☑ Kurtosis
- ☐ Std. err., Kurtosis

Percentiles, ranges
- ☑ Minimum & maximum
- ☐ Lower & upper quartiles
- ☐ Percentile boundaries
- First: 10,00 %
- Second: 90,00 %
- ☑ Range
- ☐ Quartile range

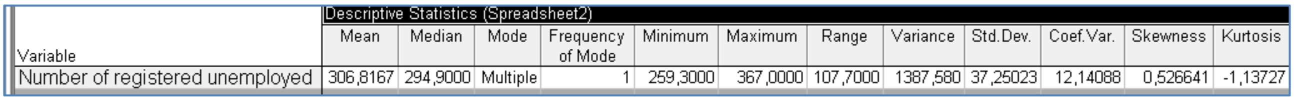
☐ Wghtd momnts

DF = ● W-1 ○ N-1

MD deletion
- ○ Casewise
- ● Pairwise

Fig. 3. **Choosing the parameters of *Descriptive statistics***

The system will calculate these indicators and present the results in the form of a table (Fig. 4) after the button is pressed Summary.



Variable	Mean	Median	Mode	Frequency of Mode	Minimum	Maximum	Range	Variance	Std.Dev.	Coef.Var.	Skewness	Kurtosis
Number of registered unemployed	306,8167	294,9000	Multiple	1	259,3000	367,0000	107,7000	1387,580	37,25023	12,14088	0,526641	-1,13727

Descriptive Statistics (Spreadsheet2)

Fig. 4. **The results of the function *Descriptive statistics***

So, on average, there were 306 thousand unemployed people in 2019, with the maximum number being 367 thousand people and the minimum being 259 thousand. In general, this time series is symmetrical (with inherent positive right-side asymmetry (Skewness > 0)), and insignificant sharpness of the distribution peak (Kurtosis < 0) is observed. The variance indicates a significant measure of the variation in the values of a random variable relative to its mathematical expectation and the average value, and the coefficient of variation is less than 33 %, which indicates that the number of unemployed corresponds to the normal distribution law (Coef.Var = 12.14 %).

Task 3. For a preliminary analysis of the information space of the study, it is necessary to verify the law of the distribution of indicators characterizing the retail trade in Ukraine (Table 3).

Table 3

Basic retail indicators

Years	Retail trade turnover of enterprises (legal entities), mln	Presence of objects of retail trade of enterprises (legal entities) by the end of the year, thousand	The number of markets for consumer goods at the end of the year, units
1	2	3	4
1990	78	145.7	1 576
1991	132	143.1	1 506
1992	1 456	138.0	1 482
1993	43 824	141.2	1418
1994	336 968	138.3	1 377
1995	11 964	133.7	1 282
1996	17 344	132.0	1 231
1997	18 933	127.5	1 551
1998	19 317	121.0	2 120
1999	22 151	111.6	2 320
2000	28 757	103.2	2 514
2001	34 417	96.4	2 715
2002	39 691	89.3	2 863
2003	49 994	83.8	2 891
2004	67 556	78.5	2 869
2005	94 332	75.2	2 886
2006	129 952	73.6	2 890
2007	178 233	71.9	2 834
2008	246 903	69.2	2 785

Table 3 (the end)

1	2	3	4
2009	230 955	65.3	2 761
2010	280 890	64.8	2 758
2011	350 059	64.2	2 698
2012	405 114	62.2	2 647
2013	433 081	59.8	2 609
2014	438 343	49.6	2 177
2015	487 558	49.6	2 134

Task 4. Find two time series that characterize the development of the Ukrainian industry in 2007 – 2019 and check them for normality of the distribution law.

Task 5. Using the statistical data of the website of the State Statistics Service of Ukraine [19], find information on the amount of capital investment (on a quarterly basis) and analyze this series using the module "Descriptive Statistics" of the program Statistica 10.0 and MS Excel. Compare the results.

Task 6. Using a powerful graphical toolkit of Statistica 10.0 (the Graphs menu) and your own information space of research, build a 2D histogram, point and line graphs.

Task 7. Find spatial one-dimensional data (at least 30 observations) and perform statistical analysis using descriptive statistics and graphic procedures, check your ranks for normal distribution law and draw conclusions (give an economic interpretation of the results).

The list of questions for independent work

1. What is the need for using economic and mathematical models in the process of analysis of socioeconomic systems?

2. What is the difference between strictly determined models and models that take into account uncertainty?

3. What are the basic principles of constructing models?

4. What is the adequacy of economic and mathematical models?

5. What is the place occupied by models in society?

6. What are the main classifications of economic and mathematical models?

7. What are the approaches to the system category interpretation?

8. What is the definition of a socioeconomic system?

9. What are the stages of the model construction process?

10. What are the physical models different from analog ones?

Topic 2. Regression models as a means of researching economic processes

Task 1. Using the program Statistica 10.0, find the forecast value of GDP in the next period with the following values of indicators: the consumer price index of 100.7 million UAH, export of goods and services in the amount of 3 908 266.6 million UAH, import of goods and services of 4 778 308.8 million UAH (Table 4).

Table 4

The input data

Period	GDP, mln UAH (Y)	Consumer price index (X1)	Export, mln UAH (X2)	Import, mln UAH (X3)
1	2	3	4	5
1	1937.3	102.6	2 339 402.7	2 712 852.43
2	1 820.2	101.2	2 518 534.3	3 171 353.67
3	2 054.4	101.3	3 129139.2	3 872 083.7
4	2 029.8	100.4	2 953 856.7	3 283 634.6
5	2 108	100.5	3 104 624.5	3 632 466
6	2 221.8	100.3	3 317 653.5	3 610 201.2
7	2 433.1	100.7	3 344 490.8	3 686 831.1
8	2 030	100.2	3 511 110.9	3 839 422.4
9	2 836.1	100.1	3 676 317.9	4 147 061.3
10	2 567.6	100.3	3 437 449.5	4 021 637.6
11	2 467.7	100.6	3 344 121.8	3 922 054
12	2 358	100.8	3 691 002.6	5 134 893.1
13	2 249.9	101.7	3208 484.9	3 700 647.9
14	2 049.5	101.7	3 409 760.2	4 297 569.8
15	2 487.4	100.8	4 108 205.4	4 953 434.5
16	2 488.3	100.6	4 067 299.9	4 820 236.3
17	2 576	100.6	4 083 445.1	4 852 211.6
18	2 692.1	101	4 235 294.1	4 682 039.7
19	2 768.2	100.9	4 258 746.5	5 315 298.1
20	2 978.5	100.1	4167 498.9	4 872 915.4
21	3 168.2	100.8	4 114 710.7	4 851 865.6
22	3 230.8	101.6	4 345 291.8	5 872 680.3
23	3 447.3	101.2	4 450 168.3	5 822 204
24	3 027.9	101.1	4 799 157.8	6 628 819.8
25	2 963.7	102.3	3 663 214.9	4 627 526.5
26	2 854	101.2	4 682 418.3	6 465 057.5
27	3 087.4	101.2	5 444 491.8	7 712 994.2
28	3 298.8	101.4	5 571 314.2	7 936 247.4

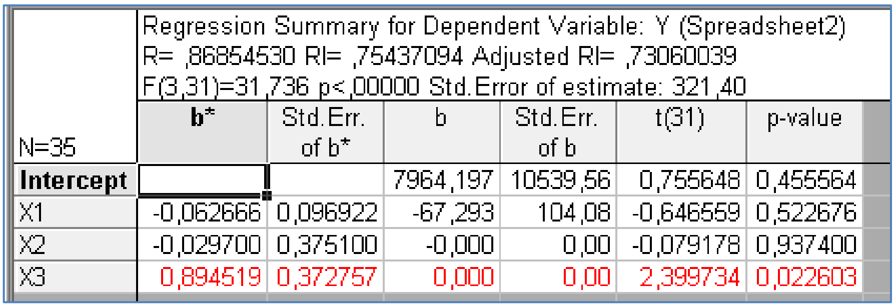
Table 4 (the end)

1	2	3	4	5
29	3 397.5	101.4	6 284 581.1	7 710 564
30	3 455	101	6 896 161.2	7 934 961.4
31	3 679.8	100.5	7 616 856.6	8 822 925.9
32	3 866.6	100.4	6 718 135.9	8 155 969.3
33	3 997.5	100.8	6 685 131.4	8 479 144.1
34	3 858.1	100.9	5 861 332.6	7 647 194.6
35	3 649.8	100.8	3 622 525.2	5 264 462.7

Guidelines

1. The construction of a forecast is possible only under the condition of a qualitative and adequate regression model. Therefore, first it is necessary to build a multifactor regression model of the dependence of GDP (Y) on the studied indicators (X1 – X3).

To do this, let's use the menu *Statistics* and the module *Multiple regression*. The characteristics of the model and the degree of their adequacy can be obtained by clicking the *Summary: Regression results* button. The results of building a multifactor econometric model are shown in Fig. 5.



Regression Summary for Dependent Variable: Y (Spreadsheet2)
R= ,86854530 RI= ,75437094 Adjusted RI= ,73060039
F(3,31)=31,736 p<,00000 Std.Error of estimate: 321,40

N=35	b*	Std.Err. of b*	b	Std.Err. of b	t(31)	p-value
Intercept			7964,197	10539,56	0,755648	0,455564
X1	-0,062666	0,096922	-67,293	104,08	-0,646559	0,522676
X2	-0,029700	0,375100	-0,000	0,00	-0,079178	0,937400
X3	0,894519	0,372757	0,000	0,00	2,399734	0,022603

Fig. 5. **The regression results**

The obtained results indicate the following:

the coefficient of multiple correlation (R) is 0.8685. The measured coefficient is from 0 to +1 (if the size of R is close to 1, the obtained model is adequate and can be used for the analysis and prediction of economic processes);

the model's determination coefficient (RI) is 0.7544 (if the size of R^2 is close to 1, the obtained model is adequate and can be used for the analysis and prediction of economic processes);

the adjusted determination coefficient based on the number of observations and the number of parameters is 0.7306 (*Adjusted RI*);

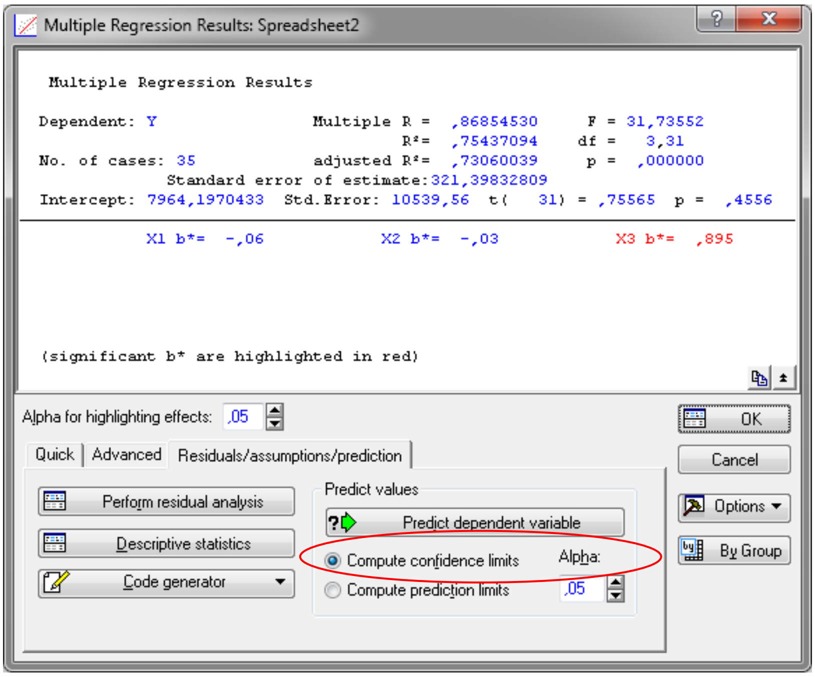
Fisher's eligibility criterion F (3. 31) = 31.736 (the Fisher's criterion is used to assess the statistical significance of the coefficient of determination. If the values obtained are more than the tabular ones, R^2 is significant, and the model is adequate);

B (a0, a1, a2, a3) = (7 964.197; –67.293; –0,00; 0.00) are the parameters of the model;

the mean square deviation of the model parameters is (10 539.56; 104.08; 0,00; 0.00);

t (31) = (0.7556; –0.6465; –0.0791; 2.3997) is the significance of the parameters according to the Student's criterion (the Student's criterion is used to assess the statistical significance of the coefficient of correlation if the values obtained are more than the tabular ones, R is significant, and the model is adequate).

The analysis of the above results shows that the model is adequate, and its parameters are significant, then the model can be used to build a forecast. To calculate the predictive values of a dependent variable, the *Predict dependent variable* option is at the bottom of the *Regression analysis results* window (Fig. 6).



```
 Multiple Regression Results

Dependent: Y               Multiple R =  ,86854530     F = 31,73552
                                  R²=  ,75437094    df =   3,31
No. of cases: 35           adjusted R²=  ,73060039     p =  ,000000
             Standard error of estimate:321,39832809
Intercept: 7964,1970433  Std.Error: 10539,56  t(   31) = ,75565  p =  ,4556

           X1 b*=  -,06             X2 b*=  -,03             X3 b*=  ,895



(significant b* are highlighted in red)
```

Alpha for highlighting effects: ,05

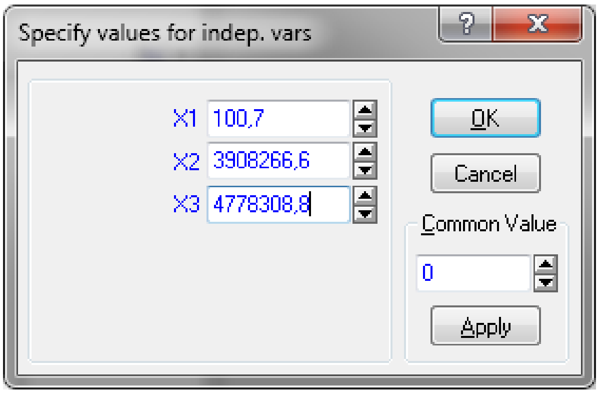
Multiple Regression Results: Spreadsheet2
OK
Cancel
Quick
Advanced
Residuals/assumptions/prediction
Perform residual analysis
Descriptive statistics
Code generator
Predict values
Predict dependent variable
Options
By Group

● Compute confidence limits
○ Compute prediction limits
Alpha: ,05

Fig. 6. **Choosing the option *Predict dependent variable***

Initiating the appropriate option, you need to specify the values of the factor characteristics (Fig. 7).

The forecasting results are presented in the form of a table, which additionally presents the parameters of the model and the limits of the forecast interval (Fig. 8).



Specify values for indep. vars

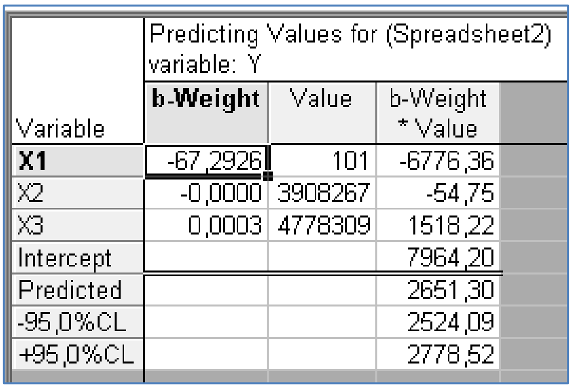
X1 100,7

X2 3908266,6

X3 4778308,8

Common Value 0

Fig. 7. **Predictive values of the factor characteristics**



Variable	Predicting Values for (Spreadsheet2) variable: Y b-Weight	Value	b-Weight * Value
X1	-67,2926	101	-6776,36
X2	-0,0000	3908267	-54,75
X3	0,0003	4778309	1518,22
Intercept			7964,20
Predicted			2651,30
-95,0%CL			2524,09
+95,0%CL			2778,52

Fig. 8. **The GDP forecast results**

The predicted GDP = 16 313.77 billion UAH; the confidence interval of the forecast values is 4 691.18 < y < 27 936.36.

Task 2. Based on Table 5, construct a single-factor regression model.

Table 5

The input data

Years	The number of agricultural animals (cattle number)	Milk production, thousand tons	Years	The number of agricultural animals (cattle number)	Milk production, thousand tons
1	2	3	4	5	6
1990	25 194.8	24 508.3	2005	6 902.9	13 714.4
1991	24 623.4	22 408.6	2006	6 514.1	13 286.9
1992	23 727.6	19 113.7	2007	6 175.4	12 262.1
1993	22 456.8	18 376.5	2008	5 490.9	11 761.3
1994	21 607.3	18 137.5	2009	5 079.0	11 609.6
1995	19 624.3	17 274.3	2010	4 826.7	11 248.5
1996	17 557.3	15 821.2	2011	4 494.4	11 086.0

Table 5 (the end)

1	2	3	4	5	6
1997	15 313.2	13 767.6	2012	4 425.8	11 377.6
1998	12 758.5	13 752.7	2013	4 645.9	11 488.2
1999	11 721.6	13 362.2	2014	4 534.0	11 132.8
2000	10 626.5	12 657.9	2015	3 884.0	10 615.4
2001	9 423.7	13 444.2	2016	3 750.3	10 381.5
2002	9 421.1	14 142.4	2017	3 682.3	10 280.5
2003	9 108.4	13 661.4	2018	3 530.8	10 064.0
2004	7 712.1	13 709.5	2019	3 332.9	96 63.2

Do the following:

2.1. Construct a linear econometric model and determine all its characteristics.

2.2. Check the statistical significance of the model parameters and the adequacy of the built model.

2.3. Calculate the theoretical values of the dependent variable and the model error. Construct a linear function graph with confidence intervals. Construct a histogram and the error distribution schedule on normally probabilistic paper.

2.4. Draw conclusions about the adequacy of the built model, give an economic interpretation of results.

Task 3. Using a single-factor regression model from task 2 and the program Statistica 10.0, build a forecast of milk production for 2020.

Task 4. Using statistical information from Table 6, construct a multifactor regression model of the dependence of GDP (Y) on the studied indicators (X1 – X3). Draw conclusions about the adequacy of the built model, give an economic interpretation of this dependence and the possibility of using it.

Table 6

The input data

Period	GDP per capital, UAH (Y)	Consumer price index (X1)	Indices of agricultural production (X2)	The number of registered unemployed, thousand people (X3)
1	2	3	4	5
1	6 792	112.104	88.715	983
2	7 273	112.3	92.5	981.8
3	7 782	112.014	94.142	946
4	8 309	111.436	94.119	9420

Table 6 (the end)

1	2	3	4	5
5	8 842	110.79	93.313	895
6	9 372	110.3	92.6	881.5
7	9 893	110.147	92.652	860
8	10 430	110.338	93.302	794
9	11 000	110.835	94.176	783
10	11 630	111.6	94.9	759.5
11	12 350	112.596	95.099	748
12	13 190	113.787	94.398	722
13	14 180	115.134	92.423	678
14	15 330	116.6	88.8	642.3
15	16 800	118.683	91.405	732
16	18 330	120.686	95.713	796
17	19 660	122.071	100.364	825
18	20 530	122.3	104	844.9
19	20 800	121.015	105.59	812
20	20 610	118.567	105.404	782
21	20 210	115.486	104.042	656
22	19 860	112.3	102.1	531.6
23	19 800	109.539	100.177	536
24	20 290	107.733	98.871	540
25	21 560	107.41	98.779	541
26	23 860	109.1	100.5	544.9
27	23 986	108	101	543.7
28	24 567	107	102.8	541.2
29	25 322	106.9	103.1	540
30	25 689	106.3	103.6	538

Task 5. Using the input data from Task 4 (Table 6) and the Farrar – Glauber test, test the econometric model for multicollinearity and, if it is necessary, eliminate it. Perform all calculations in the MS Excel and Statistica 10.0 programs.

Task 6. Check the existence of a linear multiplicity of the relationship between GDP and socioeconomic indicators (Table 7 shows the values for Ukraine in 2005 – 2018).

Construct a multiple regression model and determine all its characteristics. Check the statistical significance of the model parameters, the coefficient of multiple correlation. Calculate the theoretical values of the dependent variable and the model error. Construct a linear function graph with confidence intervals.

Calculate the predictive value of the dependent variable and confidence intervals of the change if the value of the independent indicator is known. Check the adequacy of the model according to Fisher's criterion. Draw conclusions about the adequacy of the constructed multifactorial model, and give an economic interpretation of the model as a whole. The input data are presented in Table 7.

Table 7

The input data for constructing a multifactor econometric model

Years	Production output thou UAH (X1)	Volume of retail turnover of enterprises (legal entities), mln (X2)	Average cost per month per household, UAH (X3)	Direct investments (million dollars) (X4)	The GDP (billion UAH) (У)
2005	226 358	19 317	395.6	2 063.6	186.5
2006	356 842	22 151	426.5	2 810.7	192.5
2007	373 893	28 757	541.3	3 281.8	198.9
2008	460 520	34 417	607	3 875	221.6
2009	504 008	39 691	658.3	4 555.3	225.8
2010	603 704	49 994	736.8	5 471.8	267.3
2011	809 988	67 556	903.5	6 794.4	345.1
2012	995 630	94 332	1229.4	9 047	441.5
2013	1 182 179	129 952	1442.8	16 890	544.2
2014	1 565 055	178 233	1722	21 607.3	720.7
2015	2 072 172	246 903	2590.4	29 542.7	948.1
2016	1 955 685	230 955	2754.1	35 616.4	913.3
2017	2 388 289	280 890	3072.7	40 053	1 082.6
2018	2 496 365	350 059	3456	44 806	1 316.6

Task 7. Using the input data from Task 6 (Table 7) and the Farrar – Glauber test, test the econometric model for multicollinearity and, if it is necessary, eliminate it. Perform all calculations in the MS Excel and Statistica 10.0 programs.

The list of questions for independent work

1. What problem does the econometric model solve?
2. Give the basic preconditions of the classical linear econometric model.
3. List the main stages of the use of MNCs for a single-factor model.
4. List the properties of a model parameter estimates.
5. What is the effectiveness of the model parameters?

6. What is the peculiarity of constructing a forecast using a regression model?

7. What is multicollinearity in a model?

8. List the steps of which the Farrar – Glauber algorithm is composed.

9. What are the main methods for elimination of multicollinearity?

10. List the stages of implementation of the method of main components.

Topic 3. Modeling and forecasting the development of trends

Task 1. Predict the level of income of the population of Ukraine in 2020. Build a decomposition model: to determine the form of the decomposition model, identify all the components, make a forecast of the trend component, carry out a spectral analysis of the cyclic constituent and check its quality. The input data are presented in Table 8.

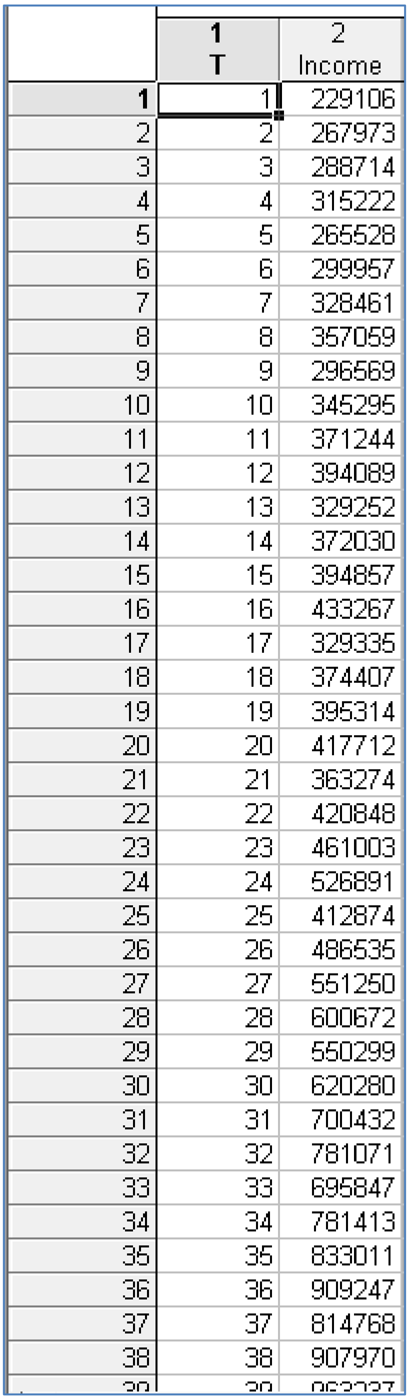
Table 8

Total income of the population of Ukraine on a quarterly basis

Period (T)	Incomes of the population, UAH million	Period (T)	Incomes of the population, UAH million
2010, quarter 1	229 106	2015, quarter 1	363 274
quarter 2	267 973	quarter 2	420 848
quarter 3	288 714	quarter 3	461 003
quarter 4	315 222	quarter 4	526 891
2011, quarter 1	265 528	2016, quarter 1	412 874
quarter 2	299 957	quarter 2	486 535
quarter 3	328 461	quarter 3	551 250
quarter 4	357 059	quarter 4	600 672
2012, quarter 1	296 569	2017, quarter 1	550 299
quarter 2	345 295	quarter 2	620 280
quarter 3	371 244	quarter 3	700 432
quarter 4	394 089	quarter 4	781 071
2013, quarter 1	329 252	2018, quarter 1	695 847
quarter 2	372 030	quarter 2	781 413
quarter 3	394 857	quarter 3	833 011
quarter 4	433 267	quarter 4	909 247
2014, quarter 1	329 335	2019, quarter 1	814 768
quarter 2	374 407	quarter 2	907 970
quarter 3	395 314	quarter 3	963 237
quarter 4	417 712	quarter 4	1 013 371

Guidelines

Form a time series and present it as a file in the package Statistica 10.0 (Fig. 9).



	1 T	2 Income
1	1	229106
2	2	267973
3	3	288714
4	4	315222
5	5	265528
6	6	299957
7	7	328461
8	8	357059
9	9	296569
10	10	345295
11	11	371244
12	12	394089
13	13	329252
14	14	372030
15	15	394857
16	16	433267
17	17	329335
18	18	374407
19	19	395314
20	20	417712
21	21	363274
22	22	420848
23	23	461003
24	24	526891
25	25	412874
26	26	486535
27	27	551250
28	28	600672
29	29	550299
30	30	620280
31	31	700432
32	32	781071
33	33	695847
34	34	781413
35	35	833011
36	36	909247
37	37	814768
38	38	907970
39	39	[illegible]

Fig. 9. **A fragment of initial data in Statistica 10.0**

In order to determine the model of decomposition of the time series components (additive or multiplicative), we present the initial data in the form of a graph (Fig. 10).

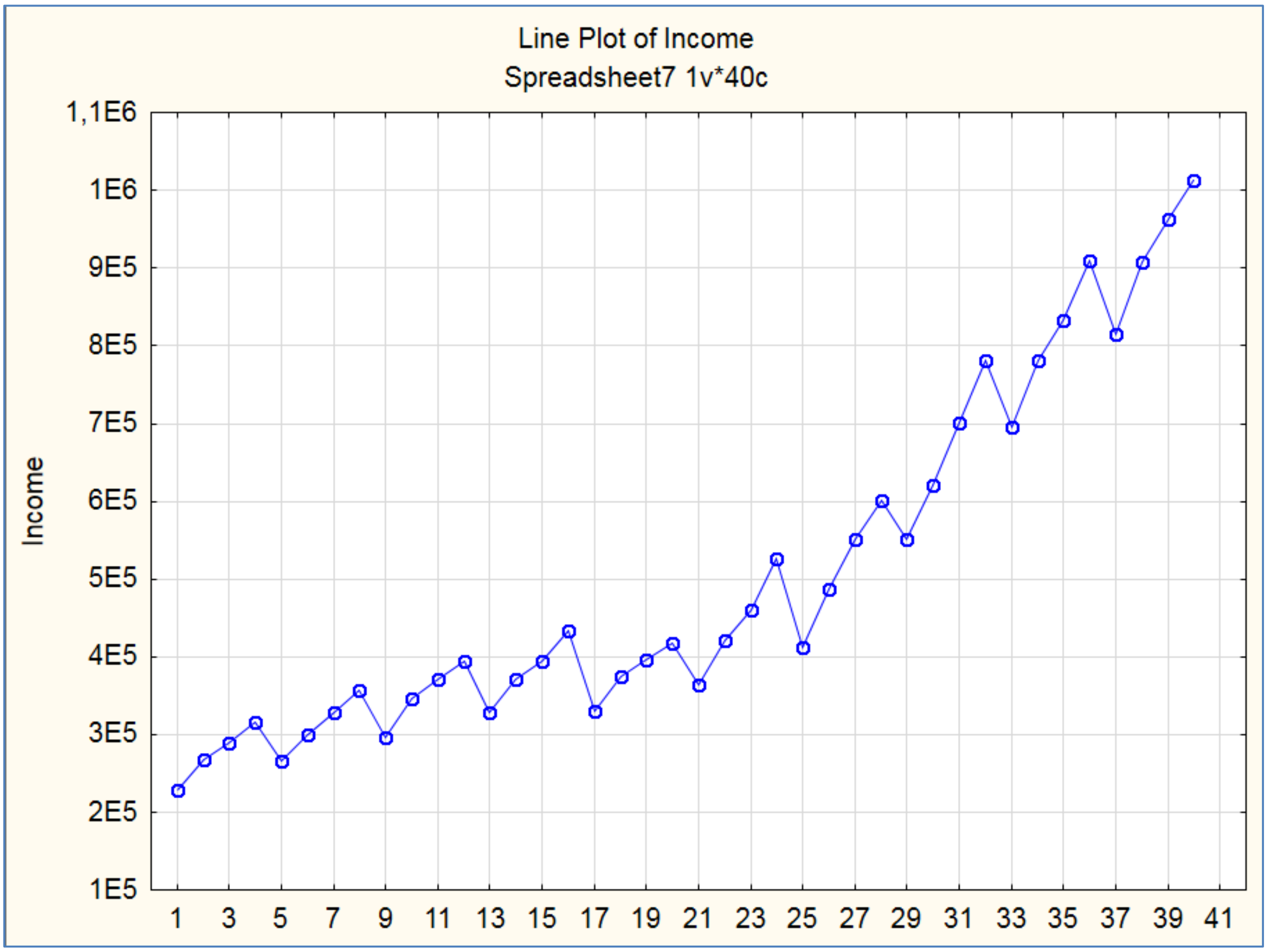


Fig. 10. **The line plot of income of the population of Ukraine**

Visual analysis shows the presence of a positive trend and some seasonality. Since the size of the population's income does not have a constant pronounced tendency to increase or decrease the amplitude of values, the multiplicative time series model should be used, which is presented by formula 1:

$$Y = T_t \times S_t \times C_t \times I, \quad (1)$$

where T is a trend component;

C is a cyclic component;

S is a seasonal component;

I is a random component;

T is the analyzed period.

In the case of a constant amplitude of changes in the values of the time series, it is expedient to use the additive model ($Y = T_t + S_t + C_t + I$).

To determine the presence of seasonality and the values of the seasonal lag, we use the *Time Series Analysis* module (Fig. 11).

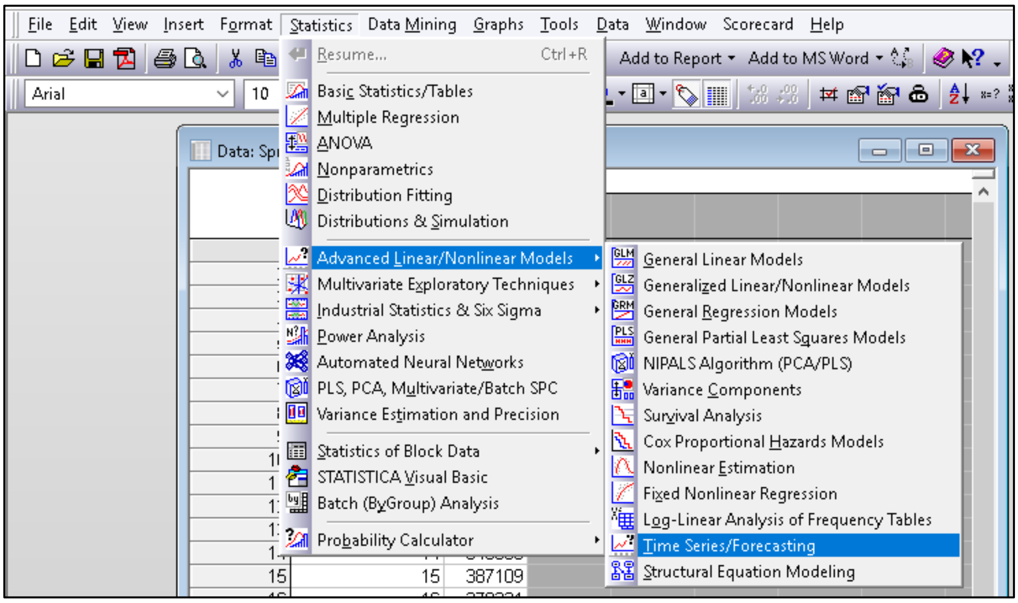
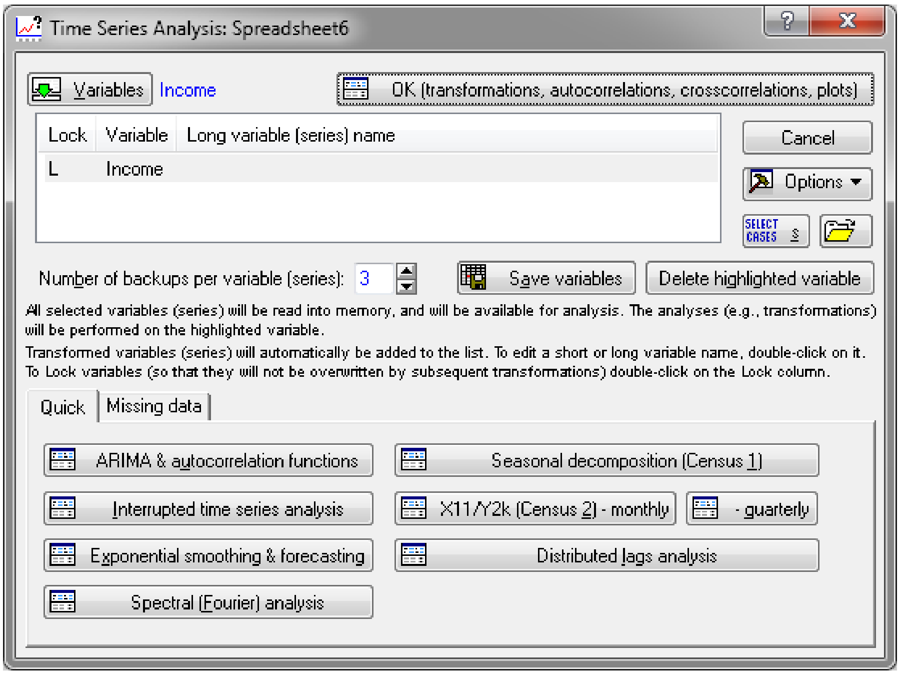

File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Scorecard Help
Add to Report
Add to MS Word
Arial
10
Data: Spr
Resume... Ctrl+R
Basic Statistics/Tables
Multiple Regression
ANOVA
Nonparametrics
Distribution Fitting
Distributions & Simulation
Advanced Linear/Nonlinear Models
Multivariate Exploratory Techniques
Industrial Statistics & Six Sigma
Power Analysis
Automated Neural Networks
PLS, PCA, Multivariate/Batch SPC
Variance Estimation and Precision
Statistics of Block Data
STATISTICA Visual Basic
Batch (ByGroup) Analysis
Probability Calculator
General Linear Models
Generalized Linear/Nonlinear Models
General Regression Models
General Partial Least Squares Models
NIPALS Algorithm (PCA/PLS)
Variance Components
Survival Analysis
Cox Proportional Hazards Models
Nonlinear Estimation
Fixed Nonlinear Regression
Log-Linear Analysis of Frequency Tables
Time Series/Forecasting
Structural Equation Modeling
15 15 387109

Fig. 11. **The *Time Series Analysis* module in Statistica 10.0**

Then, we need to choose a variable for analysis – the income of the population of Ukraine in the period of 2010 – 2019 (Fig. 12).


Time Series Analysis: Spreadsheet6

Variables: Income

Lock	Variable	Long variable (series) name
L	Income	

Number of backups per variable (series): 3

All selected variables (series) will be read into memory, and will be available for analysis. The analyses (e.g., transformations) will be performed on the highlighted variable.

Transformed variables (series) will automatically be added to the list. To edit a short or long variable name, double-click on it. To Lock variables (so that they will not be overwritten by subsequent transformations) double-click on the Lock column.

OK (transformations, autocorrelations, crosscorrelations, plots)
Cancel
Options
SELECT CASES
Save variables
Delete highlighted variable
Quick
Missing data
ARIMA & autocorrelation functions
Seasonal decomposition (Census 1)
Interrupted time series analysis
X11/Y2k (Census 2) - monthly
- quarterly
Exponential smoothing & forecasting
Distributed lags analysis
Spectral (Fourier) analysis

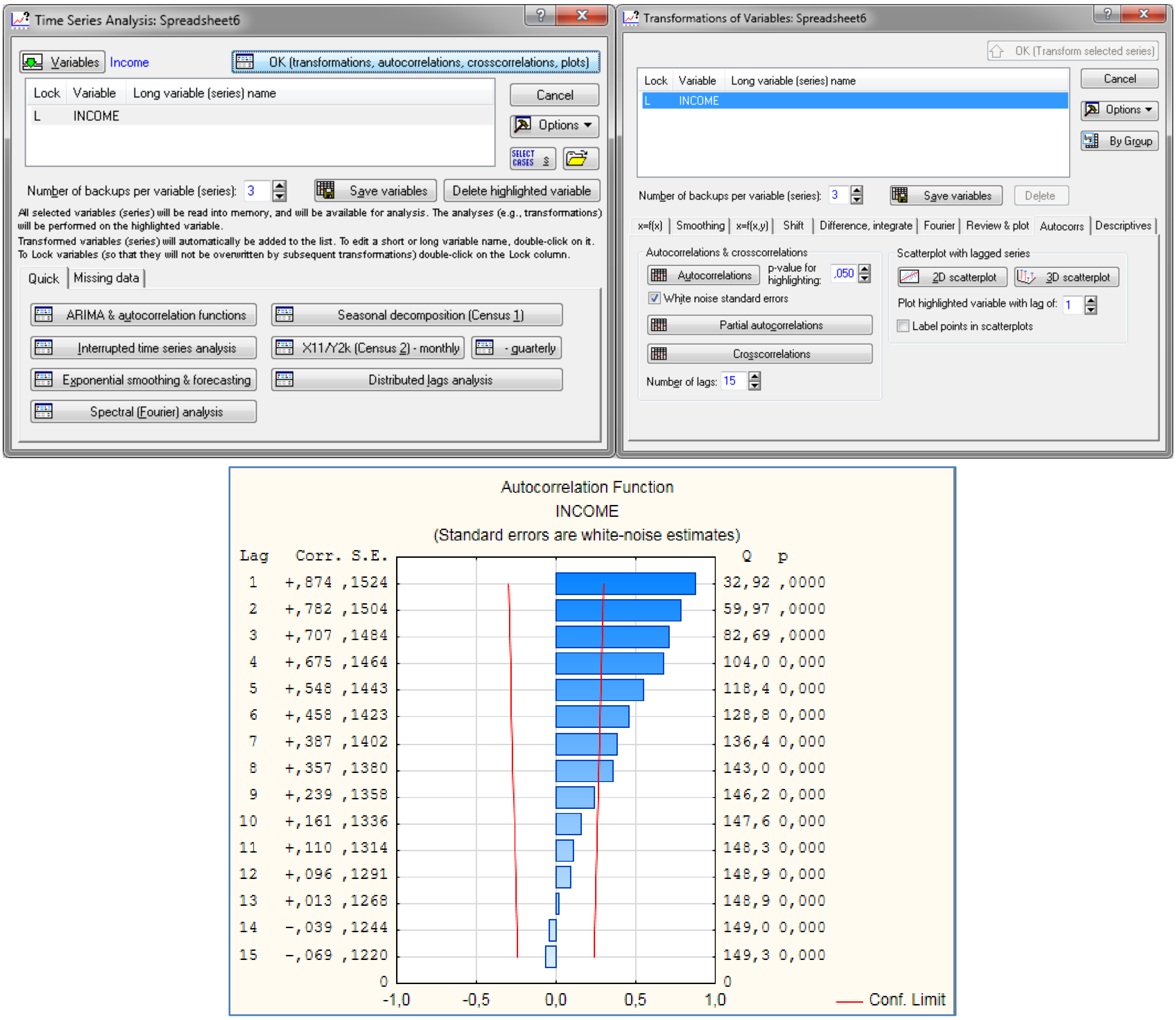
Fig. 12. **Choosing the parameters in the *Time Series Analysis* module**

The dialog box for this analysis contains the area where the original and converted time series are stored. The number of copies per row can be set by the user alone (a minimum of 3 is recommended). Your source row is denoted by the locked variable *L* (*Lock*). This means that it will always be saved and will not be deleted after all the manipulations.

Confirmations of visual analysis will be performed analytically, namely we will use:

1) autocorrelation analysis;
2) Fourier method.

Autocorrelation analysis helps to identify seasonality and determine seasonal lags of the time series. To do this, click on the button *OK* (*transformation, autocorrelations, crosscorrelations, plots*) and select the *Autocorrelation* tab. Then click on the button *Autocorrelations* (Fig. 13).


Time Series Analysis: Spreadsheet6
Variables
Income
OK (transformation, autocorrelations, crosscorrelations, plots)
Cancel
Options
SELECT CASES
Save variables
Delete highlighted variable
Quick
Missing data
ARIMA & autocorrelation functions
Seasonal decomposition (Census 1)
Interrupted time series analysis
X11/Y2k (Census 2) - monthly
- quarterly
Exponential smoothing & forecasting
Distributed lags analysis
Spectral (Fourier) analysis
Transformations of Variables: Spreadsheet6
OK (Transform selected series)
Cancel
Options
By Group
Save variables
Delete
x=f(x)
Smoothing
x=f(x,y)
Shift
Difference, integrate
Fourier
Review & plot
Autocorrs
Descriptives
Autocorrelations & crosscorrelations
Autocorrelations
Partial autocorrelations
Crosscorrelations
Scatterplot with lagged series
2D scatterplot
3D scatterplot
Autocorrelation Function
INCOME
(Standard errors are white-noise estimates)
-1,0
-0,5
0,0
0,5
1,0
Conf. Limit

Lock	Variable	Long variable (series) name
L	INCOME	

Number of backups per variable (series): 3

All selected variables (series) will be read into memory, and will be available for analysis. The analyses (e.g., transformations) will be performed on the highlighted variable.

Transformed variables (series) will automatically be added to the list. To edit a short or long variable name, double-click on it. To Lock variables (so that they will not be overwritten by subsequent transformations) double-click on the Lock column.

Lock	Variable	Long variable (series) name
L	INCOME	

Number of backups per variable (series): 3

p-value for highlighting: ,050

☑ White noise standard errors

Number of lags: 15

Plot highlighted variable with lag of: 1

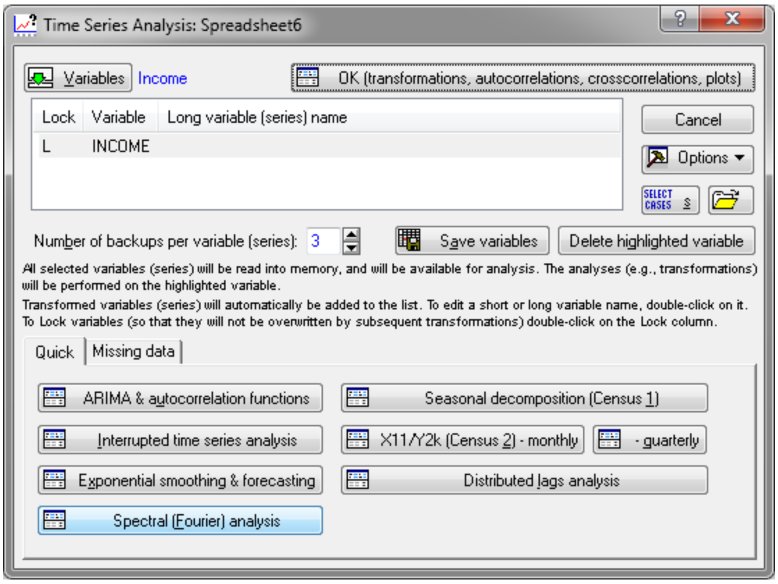
☐ Label points in scatterplots

```
Lag  Corr.  S.E.      Q     p
 1   +,874 ,1524   32,92 ,0000
 2   +,782 ,1504   59,97 ,0000
 3   +,707 ,1484   82,69 ,0000
 4   +,675 ,1464   104,0 0,000
 5   +,548 ,1443   118,4 0,000
 6   +,458 ,1423   128,8 0,000
 7   +,387 ,1402   136,4 0,000
 8   +,357 ,1380   143,0 0,000
 9   +,239 ,1358   146,2 0,000
10   +,161 ,1336   147,6 0,000
11   +,110 ,1314   148,3 0,000
12   +,096 ,1291   148,9 0,000
13   +,013 ,1268   148,9 0,000
14   -,039 ,1244   149,0 0,000
15   -,069 ,1220   149,3 0,000
   0                  0
```

Fig. 13. **The steps of autocorrelation analysis**

As we can see in Fig. 13, the greatest correlation values fall on the 1st lag, then there is a decline and its maximum autocorrelation coefficient is already achieved on the 4st lag. So in our data, there is a trend and seasonality equal to 4.

To confirm the presence of seasonality and to show the presence of hidden seasonalities, which could not be determined using the autocorrelation function, we use the Fourier method. To do this, go back to the previous level and click the button tab *Spectral (Fourier) analysis* (Fig. 14).



Time Series Analysis: Spreadsheet6

Variables: Income

Lock	Variable	Long variable (series) name
L	INCOME	

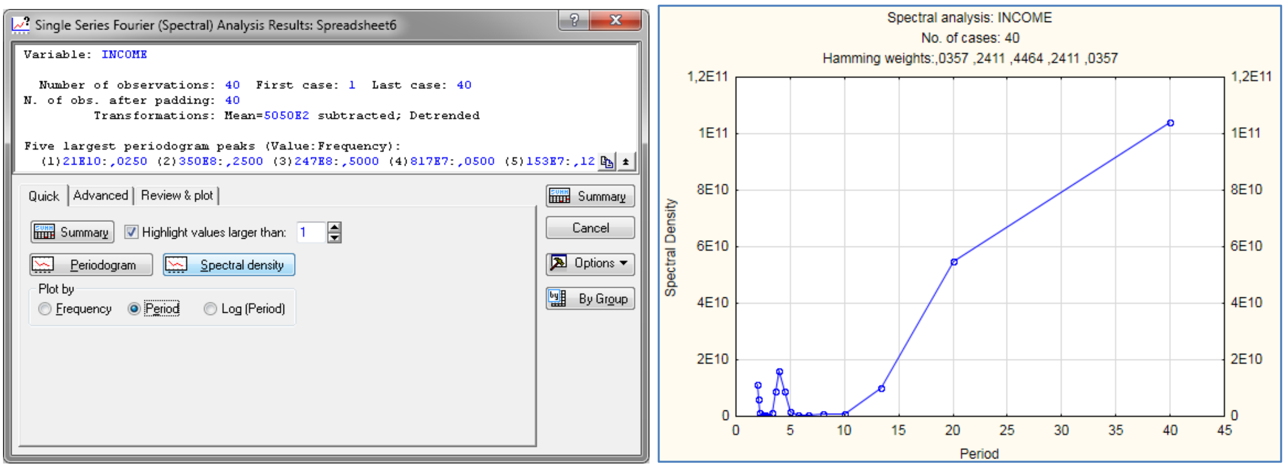
Number of backups per variable (series): 3

All selected variables (series) will be read into memory, and will be available for analysis. The analyses (e.g., transformations) will be performed on the highlighted variable.
Transformed variables (series) will automatically be added to the list. To edit a short or long variable name, double-click on it. To Lock variables (so that they will not be overwritten by subsequent transformations) double-click on the Lock column.

OK (transformations, autocorrelations, crosscorrelations, plots)
Cancel
Options
SELECT CASES
Save variables
Delete highlighted variable
Quick
Missing data
ARIMA & autocorrelation functions
Seasonal decomposition (Census 1)
Interrupted time series analysis
X11/Y2k (Census 2) - monthly
- quarterly
Exponential smoothing & forecasting
Distributed lags analysis
Spectral (Fourier) analysis

Fig. 14. **The Fourier method**

Here we are interested in one-dimensional (single series) analysis. We select the period to be the X-axis, and present it on the spectral plane (Fig. 15).



Single Series Fourier (Spectral) Analysis Results: Spreadsheet6

```
Variable: INCOME

  Number of observations: 40  First case: 1  Last case: 40
N. of obs. after padding: 40
         Transformations: Mean=5050E2 subtracted; Detrended

Five largest periodogram peaks (Value:Frequency):
  (1)21E10:,0250 (2)350E8:,2500 (3)247E8:,5000 (4)817E7:,0500 (5)153E7:,12
```

☑ Highlight values larger than: 1

Plot by: ○ Frequency ● Period ○ Log (Period)

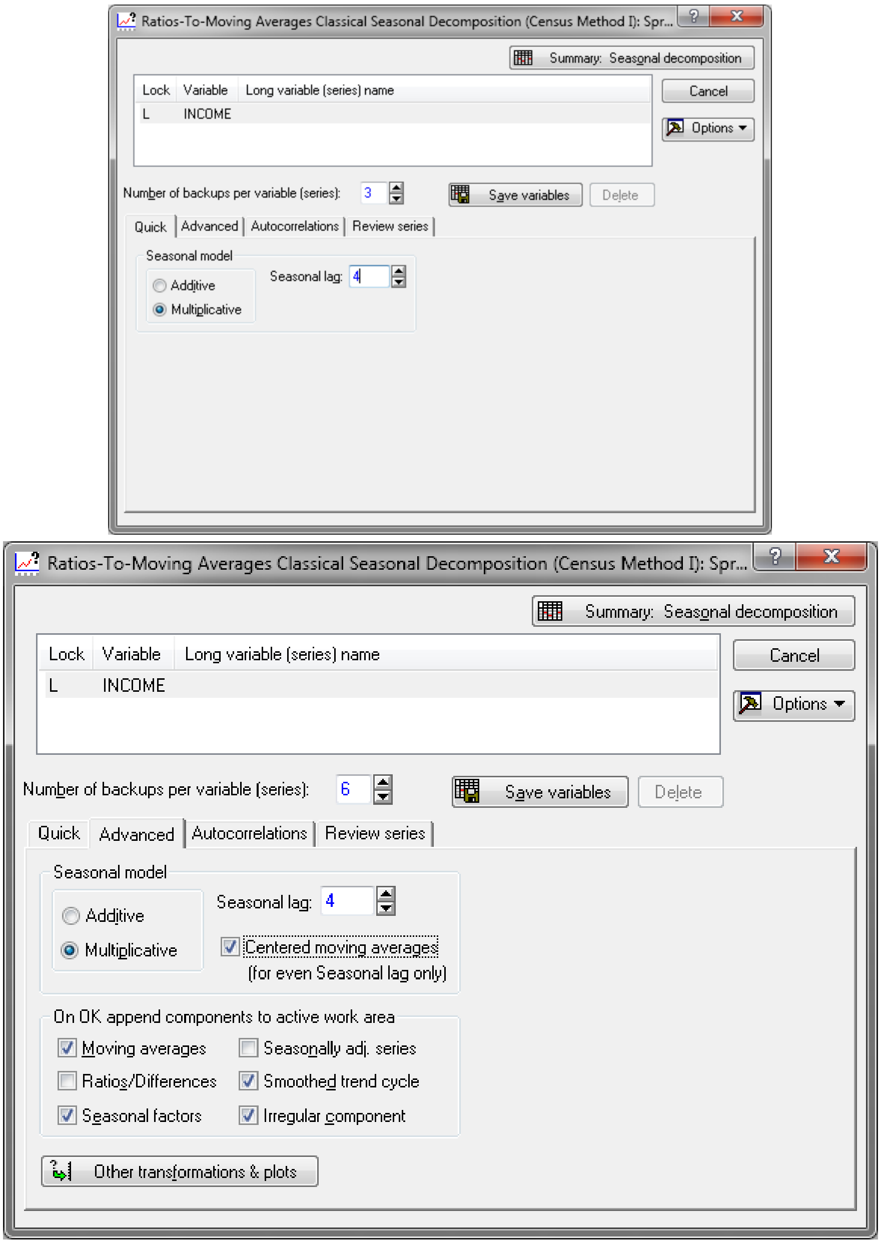
Fig. 15. **The results of the Fourier method**

We've obtained a plot of the spectral density over the period. Obviously, the absolute maximum is reached at the point with lag 4, and there is also a seasonality equal to half a quarter (lag 2). But since the value of the spectral density at this point is smaller, seasonality with lag 4 affects the variability of the data to a greater extent than seasonality with lag 2.

So, using graphical and analytical methods, we got convinced of the presence of a trend-cyclic and seasonal component in the model.

Decomposition of the time series is carried out on the following components: trend-cyclic, seasonal and random.

To do this, select the *Seasonal Decomposition* tab in the start-up panel of the *Advanced Linear / Nonlinear Models / Time Series / Forecasting* module and set the seasonal decomposition parameters (Fig. 16).



Ratios-To-Moving Averages Classical Seasonal Decomposition (Census Method I): Spr...

Lock	Variable	Long variable (series) name
L	INCOME	

Number of backups per variable (series): 3

Quick | Advanced | Autocorrelations | Review series

Seasonal model

○ Additive

● Multiplicative

Seasonal lag: 4

Ratios-To-Moving Averages Classical Seasonal Decomposition (Census Method I): Spr...

Lock	Variable	Long variable (series) name
L	INCOME	

Number of backups per variable (series): 6

Quick | Advanced | Autocorrelations | Review series

Seasonal model

○ Additive

● Multiplicative

Seasonal lag: 4

☑ Centered moving averages (for even Seasonal lag only)

On OK append components to active work area

☑ Moving averages

☐ Seasonally adj. series

☐ Ratios/Differences

☑ Smoothed trend cycle

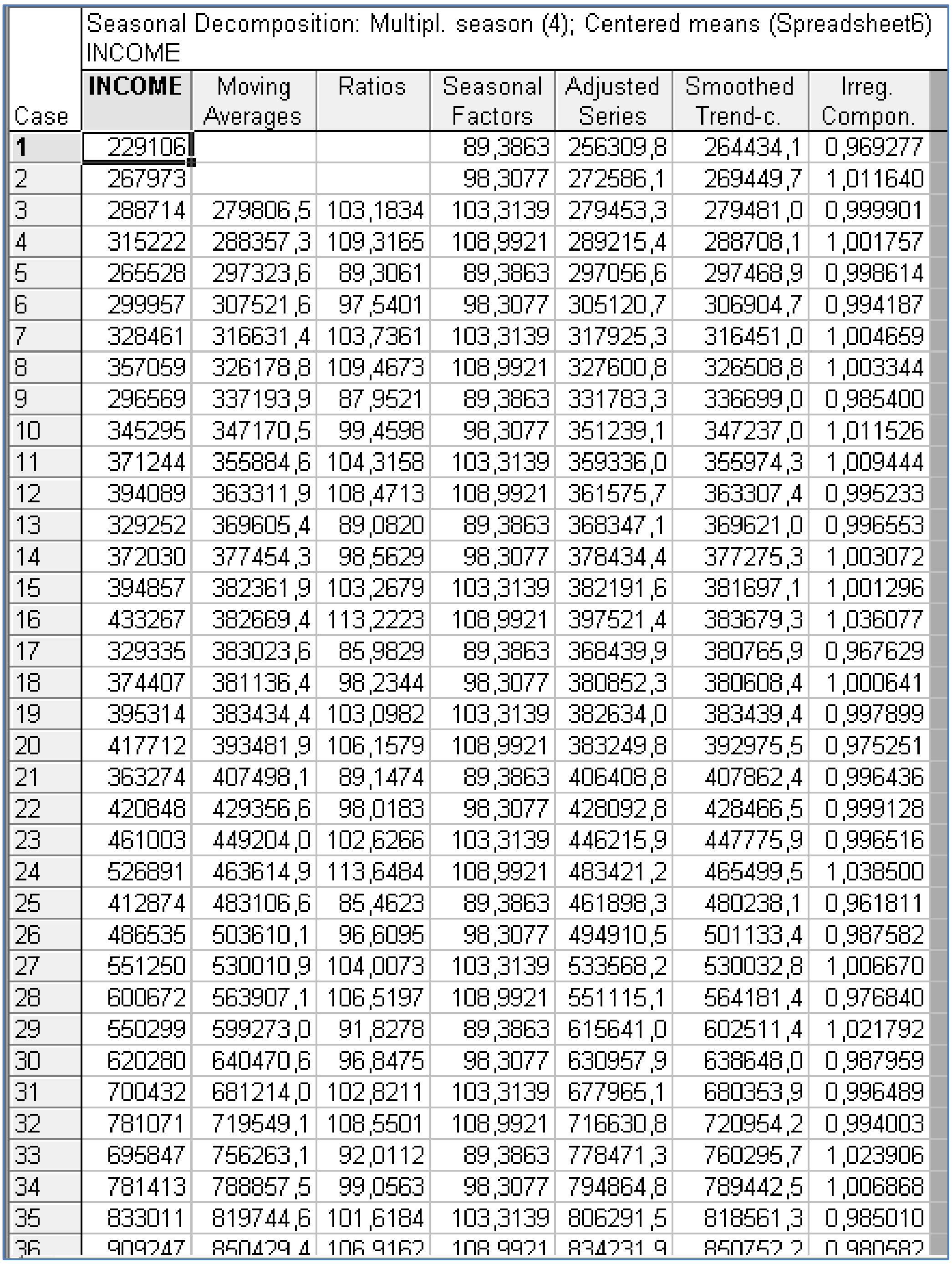
☑ Seasonal factors

☑ Irregular component

Summary: Seasonal decomposition
Cancel
Options
Save variables
Delete
Other transformations & plots

Fig. 16. **Choosing the *Time series analysis (TSA)***

Specifying the parameters of the seasonal decomposition model, we obtain the following result after clicking on the button *Summary: Seasonal decomposition* (Fig. 17).



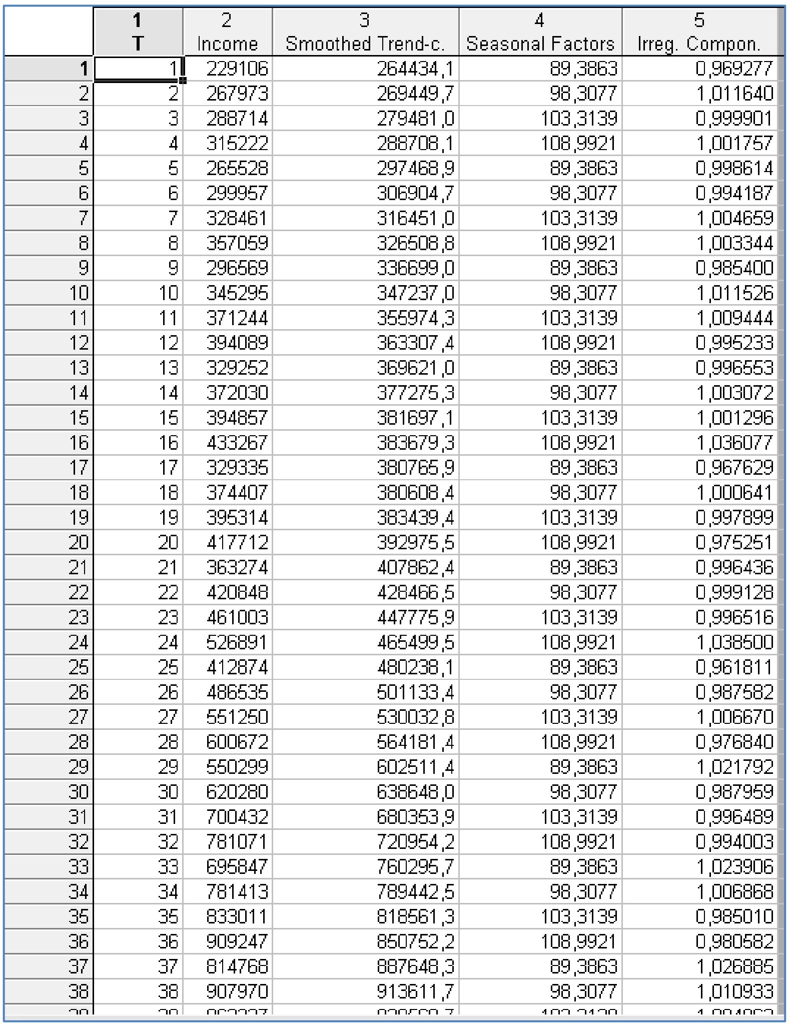
Seasonal Decomposition: Multipl. season (4); Centered means (Spreadsheet6)
INCOME

Case	INCOME	Moving Averages	Ratios	Seasonal Factors	Adjusted Series	Smoothed Trend-c.	Irreg. Compon.
1	229106			89,3863	256309,8	264434,1	0,969277
2	267973			98,3077	272586,1	269449,7	1,011640
3	288714	279806,5	103,1834	103,3139	279453,3	279481,0	0,999901
4	315222	288357,3	109,3165	108,9921	289215,4	288708,1	1,001757
5	265528	297323,6	89,3061	89,3863	297056,6	297468,9	0,998614
6	299957	307521,6	97,5401	98,3077	305120,7	306904,7	0,994187
7	328461	316631,4	103,7361	103,3139	317925,3	316451,0	1,004659
8	357059	326178,8	109,4673	108,9921	327600,8	326508,8	1,003344
9	296569	337193,9	87,9521	89,3863	331783,3	336699,0	0,985400
10	345295	347170,5	99,4598	98,3077	351239,1	347237,0	1,011526
11	371244	355884,6	104,3158	103,3139	359336,0	355974,3	1,009444
12	394089	363311,9	108,4713	108,9921	361575,7	363307,4	0,995233
13	329252	369605,4	89,0820	89,3863	368347,1	369621,0	0,996553
14	372030	377454,3	98,5629	98,3077	378434,4	377275,3	1,003072
15	394857	382361,9	103,2679	103,3139	382191,6	381697,1	1,001296
16	433267	382669,4	113,2223	108,9921	397521,4	383679,3	1,036077
17	329335	383023,6	85,9829	89,3863	368439,9	380765,9	0,967629
18	374407	381136,4	98,2344	98,3077	380852,3	380608,4	1,000641
19	395314	383434,4	103,0982	103,3139	382634,0	383439,4	0,997899
20	417712	393481,9	106,1579	108,9921	383249,8	392975,5	0,975251
21	363274	407498,1	89,1474	89,3863	406408,8	407862,4	0,996436
22	420848	429356,6	98,0183	98,3077	428092,8	428466,5	0,999128
23	461003	449204,0	102,6266	103,3139	446215,9	447775,9	0,996516
24	526891	463614,9	113,6484	108,9921	483421,2	465499,5	1,038500
25	412874	483106,6	85,4623	89,3863	461898,3	480238,1	0,961811
26	486535	503610,1	96,6095	98,3077	494910,5	501133,4	0,987582
27	551250	530010,9	104,0073	103,3139	533568,2	530032,8	1,006670
28	600672	563907,1	106,5197	108,9921	551115,1	564181,4	0,976840
29	550299	599273,0	91,8278	89,3863	615641,0	602511,4	1,021792
30	620280	640470,6	96,8475	98,3077	630957,9	638648,0	0,987959
31	700432	681214,0	102,8211	103,3139	677965,1	680353,9	0,996489
32	781071	719549,1	108,5501	108,9921	716630,8	720954,2	0,994003
33	695847	756263,1	92,0112	89,3863	778471,3	760295,7	1,023906
34	781413	788857,5	99,0563	98,3077	794864,8	789442,5	1,006868
35	833011	819744,6	101,6184	103,3139	806291,5	818561,3	0,985010
36	909247	850429,4	106,9162	108,9921	834231,9	850752,2	0,980582

Fig. 17. **A fragment of the results of TSA in Statistica 10.0**

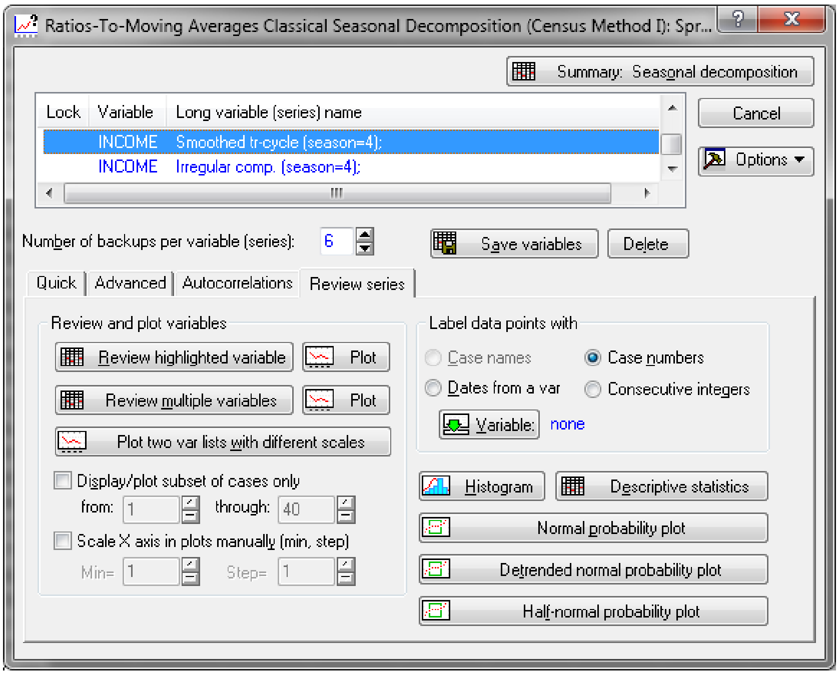
At the next stage it is necessary to copy the decomposition results, namely the trend-cyclic, seasonal and random components into the window with the initial data (Fig. 18).

Next, we need to visualize the components of the composition model. To do this, we need to return to the analysis tab, go to the charts and alternately choose the variable we need and press the *Plot* button (Fig. 19).



	1 T	2 Income	3 Smoothed Trend-c.	4 Seasonal Factors	5 Irreg. Compon.
1	1	229106	264434,1	89,3863	0,969277
2	2	267973	269449,7	98,3077	1,011640
3	3	288714	279481,0	103,3139	0,999901
4	4	315222	288708,1	108,9921	1,001757
5	5	265528	297468,9	89,3863	0,998614
6	6	299957	306904,7	98,3077	0,994187
7	7	328461	316451,0	103,3139	1,004659
8	8	357059	326508,8	108,9921	1,003344
9	9	296569	336699,0	89,3863	0,985400
10	10	345295	347237,0	98,3077	1,011526
11	11	371244	355974,3	103,3139	1,009444
12	12	394089	363307,4	108,9921	0,995233
13	13	329252	369621,0	89,3863	0,996553
14	14	372030	377275,3	98,3077	1,003072
15	15	394857	381697,1	103,3139	1,001296
16	16	433267	383679,3	108,9921	1,036077
17	17	329335	380765,9	89,3863	0,967629
18	18	374407	380608,4	98,3077	1,000641
19	19	395314	383439,4	103,3139	0,997899
20	20	417712	392975,5	108,9921	0,975251
21	21	363274	407862,4	89,3863	0,996436
22	22	420848	428466,5	98,3077	0,999128
23	23	461003	447775,9	103,3139	0,996516
24	24	526891	465499,5	108,9921	1,038500
25	25	412874	480238,1	89,3863	0,961811
26	26	486535	501133,4	98,3077	0,987582
27	27	551250	530032,8	103,3139	1,006670
28	28	600672	564181,4	108,9921	0,976840
29	29	550299	602511,4	89,3863	1,021792
30	30	620280	638648,0	98,3077	0,987959
31	31	700432	680353,9	103,3139	0,996489
32	32	781071	720954,2	108,9921	0,994003
33	33	695847	760295,7	89,3863	1,023906
34	34	781413	789442,5	98,3077	1,006868
35	35	833011	818561,3	103,3139	0,985010
36	36	909247	850752,2	108,9921	0,980582
37	37	814768	887648,3	89,3863	1,026885
38	38	907970	913611,7	98,3077	1,010933

Fig. 18. **Adding the decomposition components to the file**



Ratios-To-Moving Averages Classical Seasonal Decomposition (Census Method I): Spr...

Lock	Variable	Long variable (series) name
	INCOME	Smoothed tr-cycle (season=4);
	INCOME	Irregular comp. (season=4);

Number of backups per variable (series): 6

Review and plot variables

□ Display/plot subset of cases only

from: 1 through: 40

□ Scale X axis in plots manually (min, step)

Min= 1 Step= 1

Label data points with

○ Case names ● Case numbers

○ Dates from a var ○ Consecutive integers

Variable: none

Summary: Seasonal decomposition
Cancel
Options
Save variables
Delete
Quick
Advanced
Autocorrelations
Review series
Review highlighted variable
Plot
Review multiple variables
Plot
Plot two var lists with different scales
Histogram
Descriptive statistics
Normal probability plot
Detrended normal probability plot
Half-normal probability plot

Fig. 19. **Choosing the parameters**

Then we need to build a smoothed trend-cycle component (Fig. 20).

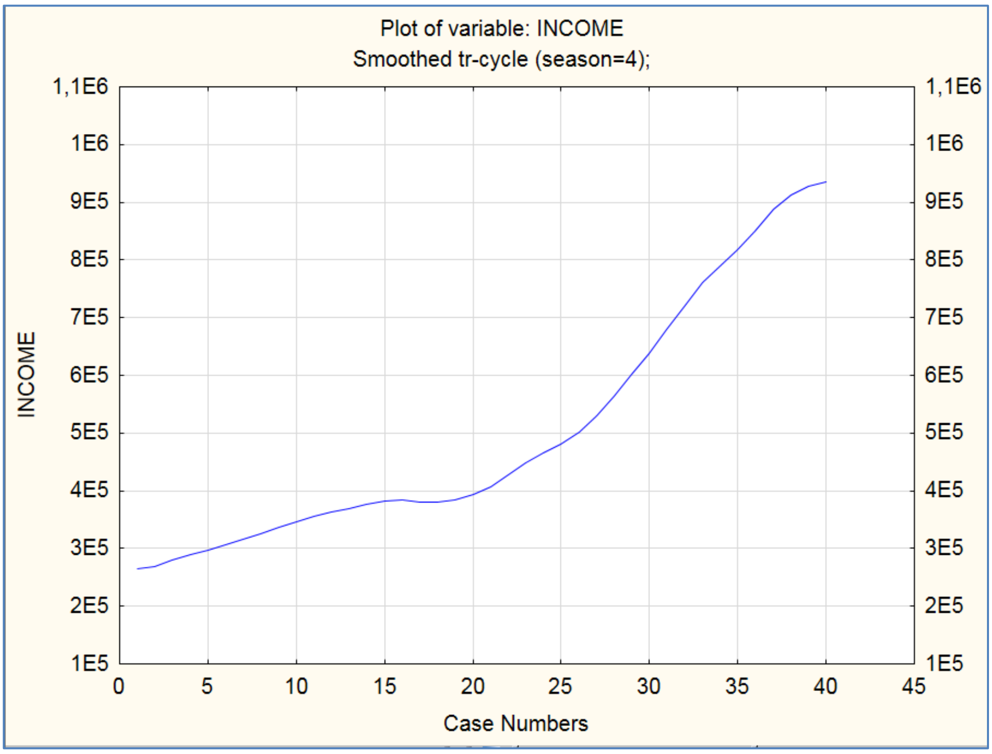


Fig. 20. **The trend-cycle component**

Then we need to build a seasonal component (Fig. 21).

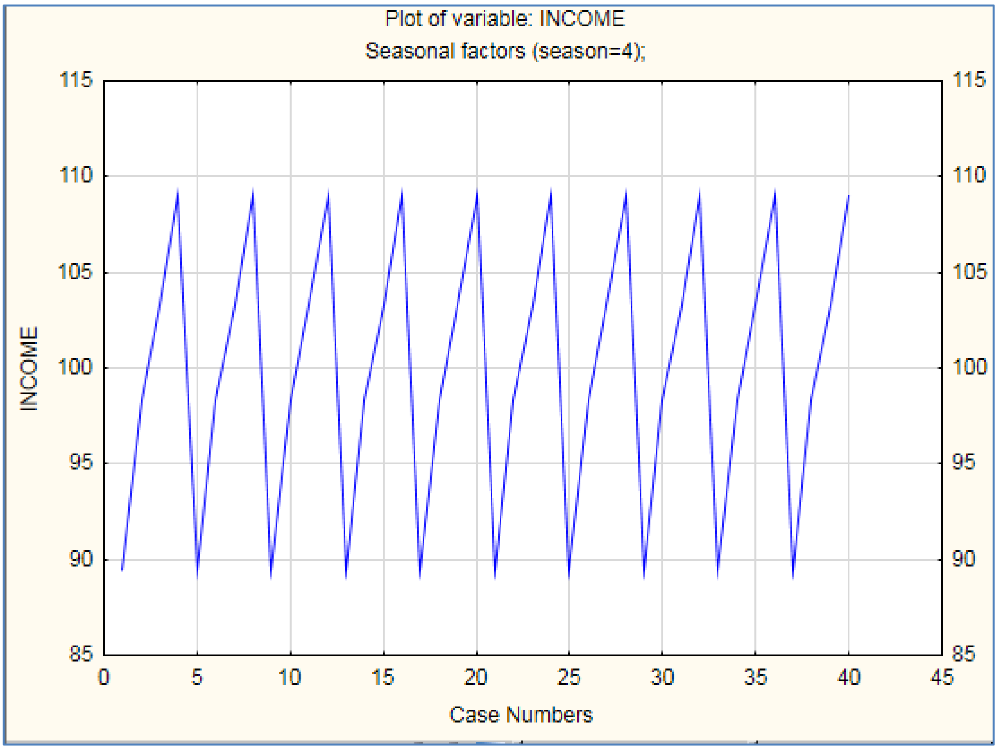


Fig. 21. **The seasonal component**

Then we need to build a random (irregular) component graph (Fig. 22).

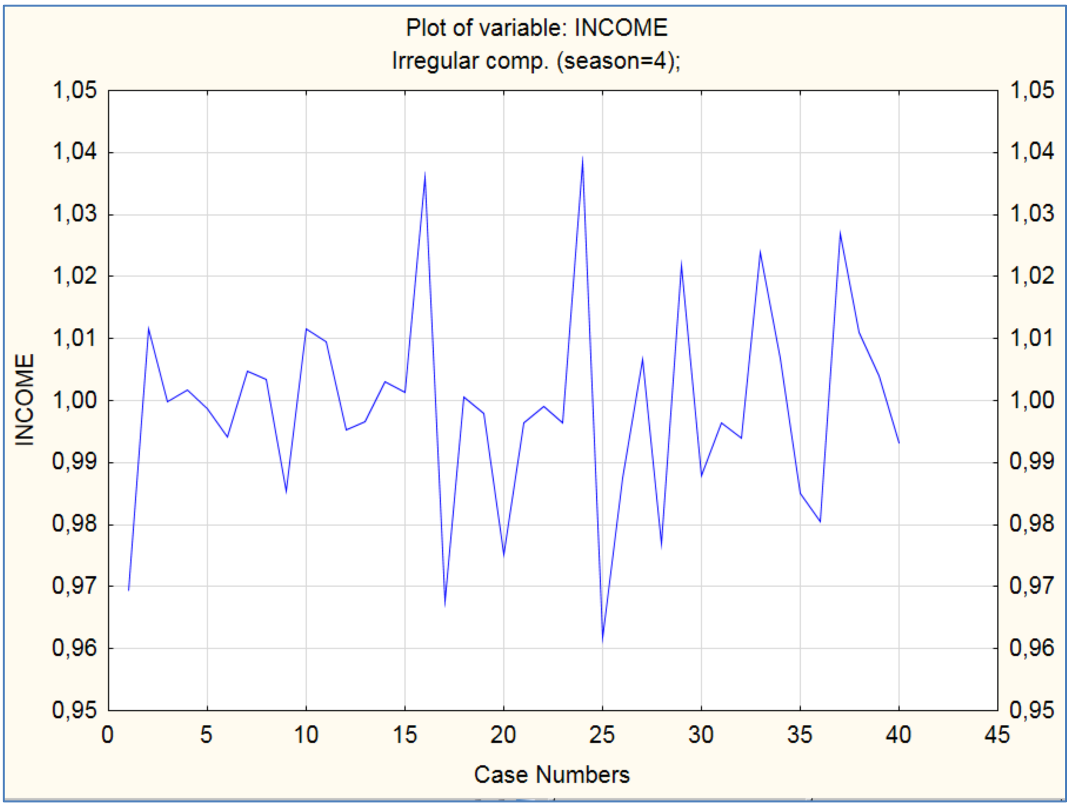
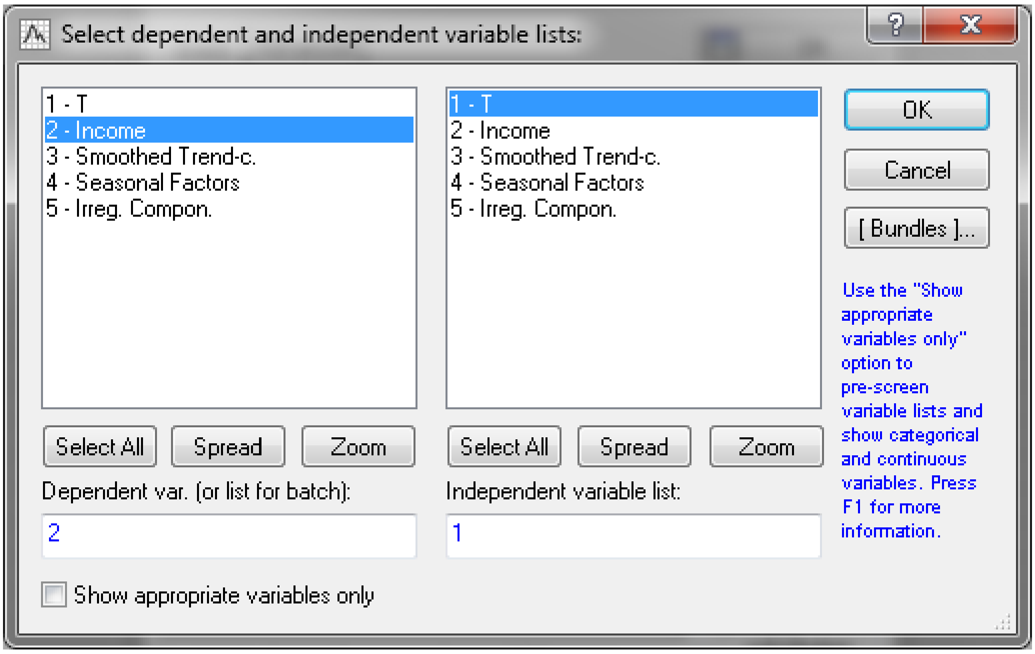


Fig. 22. **The random component graph**

The next step is to construct a regression model in which the independent variable is time (*T*) (Fig. 23).



Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

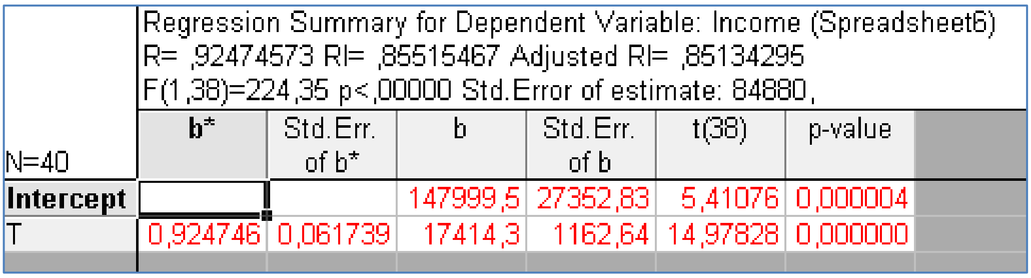
Dependent var. (or list for batch): 2

Independent variable list: 1

□ Show appropriate variables only

Fig. 23. **Choosing the parameters of the single-factor regression model**

The simulation results are given in Fig. 24.



Regression Summary for Dependent Variable: Income (Spreadsheet6)
R= ,92474573 RI= ,85515467 Adjusted RI= ,85134295
F(1,38)=224,35 p<,00000 Std.Error of estimate: 84880,

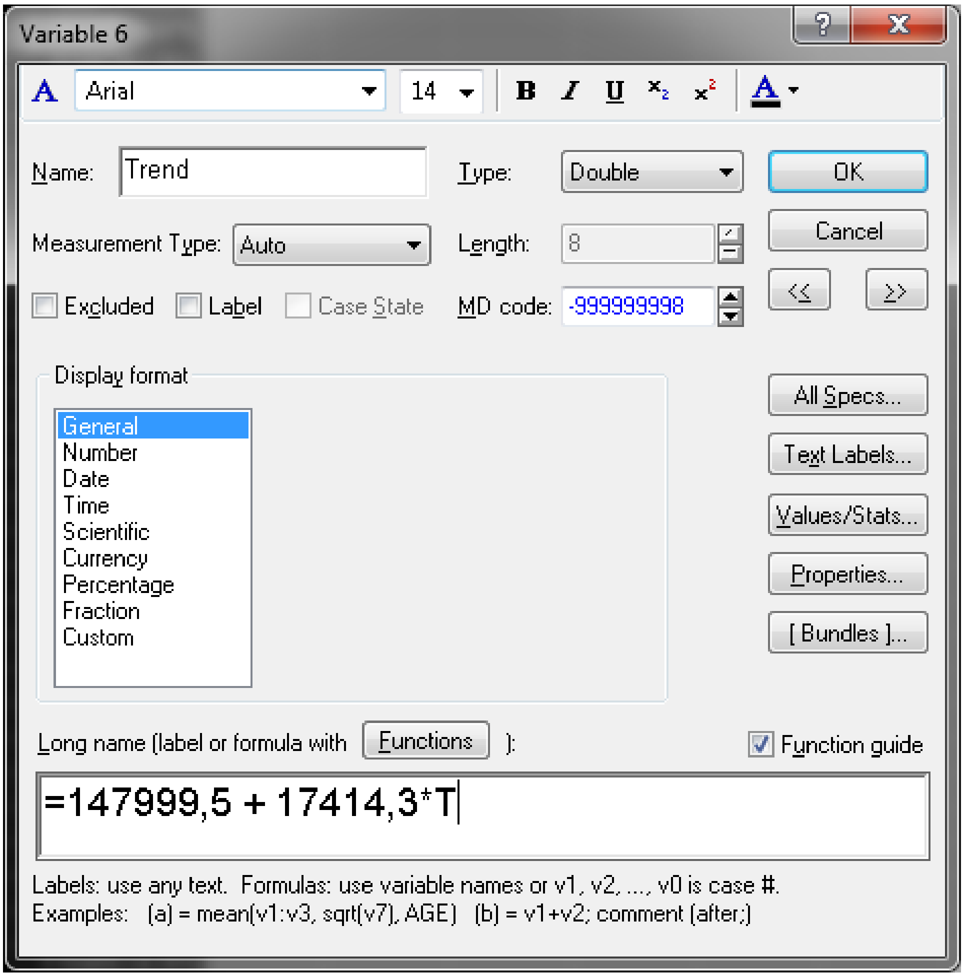
N=40	b*	Std.Err. of b*	b	Std.Err. of b	t(38)	p-value
Intercept			147999,5	27352,83	5,41076	0,000004
T	0,924746	0,061739	17414,3	1162,64	14,97828	0,000000

Fig. 24. **The regression model results**

So, we got an adequate and quality model that will look like:

$$У = 147\,999.5 + 17\,414.3 \times T.$$

Next, it is necessary to isolate the trend from the trend-cycle component. To do this, add a new variable with a calculation formula (Fig. 25).



Name: Trend

Type: Double

Measurement Type: Auto

Length: 8

□ Excluded □ Label □ Case State

MD code: -999999998

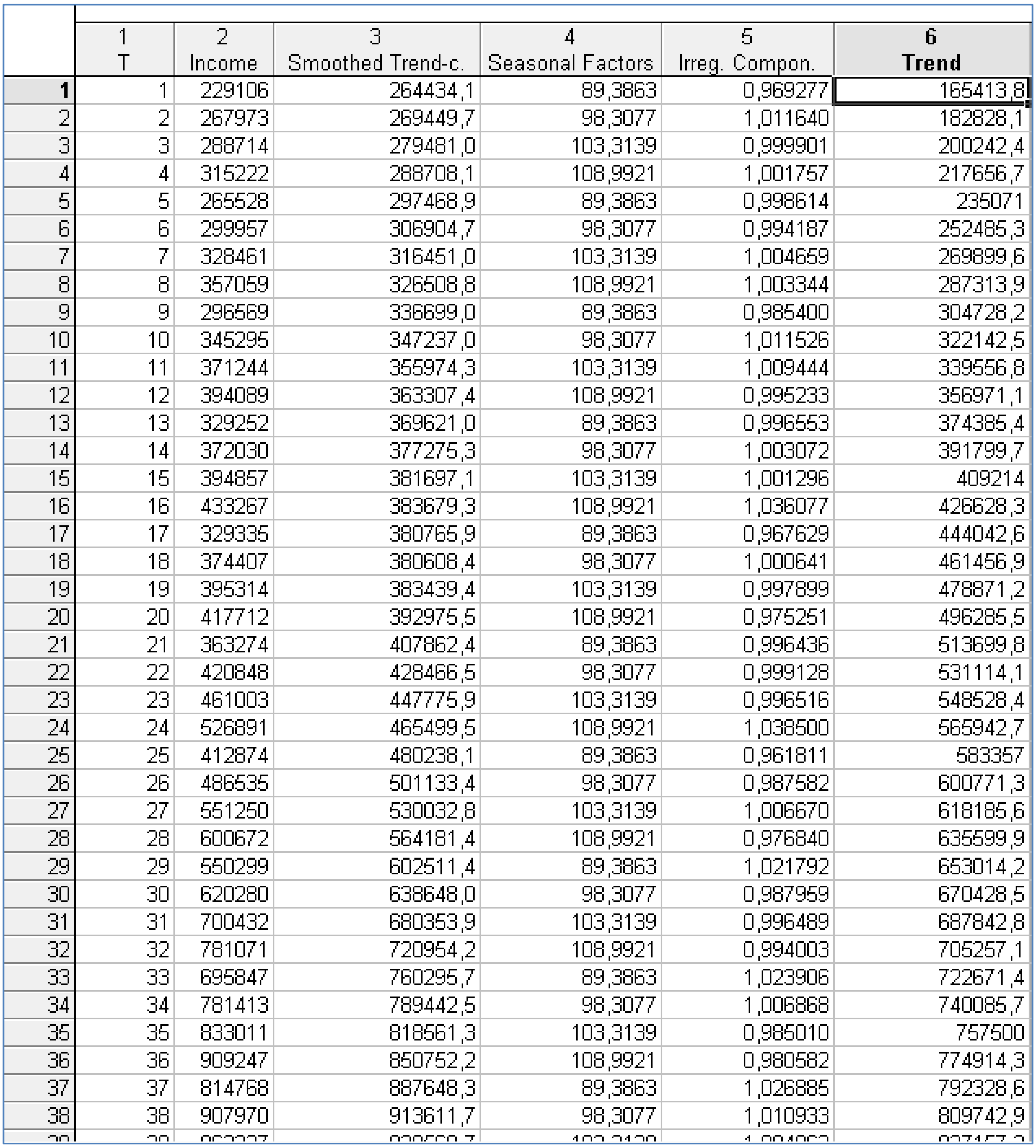
Display format: General (selected), Number, Date, Time, Scientific, Currency, Percentage, Fraction, Custom

Long name (label or formula with Functions): ☑ Function guide

=147999,5 + 17414,3*T

Labels: use any text. Formulas: use variable names or v1, v2, ..., v0 is case #.
Examples: (a) = mean(v1:v3, sqrt(v7), AGE) (b) = v1+v2; comment (after;)

Fig. 25. **Adding the trend component**



	1 T	2 Income	3 Smoothed Trend-c.	4 Seasonal Factors	5 Irreg. Compon.	6 **Trend**
1	1	229106	264434,1	89,3863	0,969277	165413,8
2	2	267973	269449,7	98,3077	1,011640	182828,1
3	3	288714	279481,0	103,3139	0,999901	200242,4
4	4	315222	288708,1	108,9921	1,001757	217656,7
5	5	265528	297468,9	89,3863	0,998614	235071
6	6	299957	306904,7	98,3077	0,994187	252485,3
7	7	328461	316451,0	103,3139	1,004659	269899,6
8	8	357059	326508,8	108,9921	1,003344	287313,9
9	9	296569	336699,0	89,3863	0,985400	304728,2
10	10	345295	347237,0	98,3077	1,011526	322142,5
11	11	371244	355974,3	103,3139	1,009444	339556,8
12	12	394089	363307,4	108,9921	0,995233	356971,1
13	13	329252	369621,0	89,3863	0,996553	374385,4
14	14	372030	377275,3	98,3077	1,003072	391799,7
15	15	394857	381697,1	103,3139	1,001296	409214
16	16	433267	383679,3	108,9921	1,036077	426628,3
17	17	329335	380765,9	89,3863	0,967629	444042,6
18	18	374407	380608,4	98,3077	1,000641	461456,9
19	19	395314	383439,4	103,3139	0,997899	478871,2
20	20	417712	392975,5	108,9921	0,975251	496285,5
21	21	363274	407862,4	89,3863	0,996436	513699,8
22	22	420848	428466,5	98,3077	0,999128	531114,1
23	23	461003	447775,9	103,3139	0,996516	548528,4
24	24	526891	465499,5	108,9921	1,038500	565942,7
25	25	412874	480238,1	89,3863	0,961811	583357
26	26	486535	501133,4	98,3077	0,987582	600771,3
27	27	551250	530032,8	103,3139	1,006670	618185,6
28	28	600672	564181,4	108,9921	0,976840	635599,9
29	29	550299	602511,4	89,3863	1,021792	653014,2
30	30	620280	638648,0	98,3077	0,987959	670428,5
31	31	700432	680353,9	103,3139	0,996489	687842,8
32	32	781071	720954,2	108,9921	0,994003	705257,1
33	33	695847	760295,7	89,3863	1,023906	722671,4
34	34	781413	789442,5	98,3077	1,006868	740085,7
35	35	833011	818561,3	103,3139	0,985010	757500
36	36	909247	850752,2	108,9921	0,980582	774914,3
37	37	814768	887648,3	89,3863	1,026885	792328,6
38	38	907970	913611,7	98,3077	1,010933	809742,9

Fig. 25. (the end)

It is necessary to build (visualize) a trend component (Fig. 26).
The values of the cycle component are then calculated as follows:
Cycle = Smoothed Trend-C / Trend (Fig. 27).
The window for entering a new variable is given in Fig. 28.
The graph of the cyclic component is presented in Fig. 29.

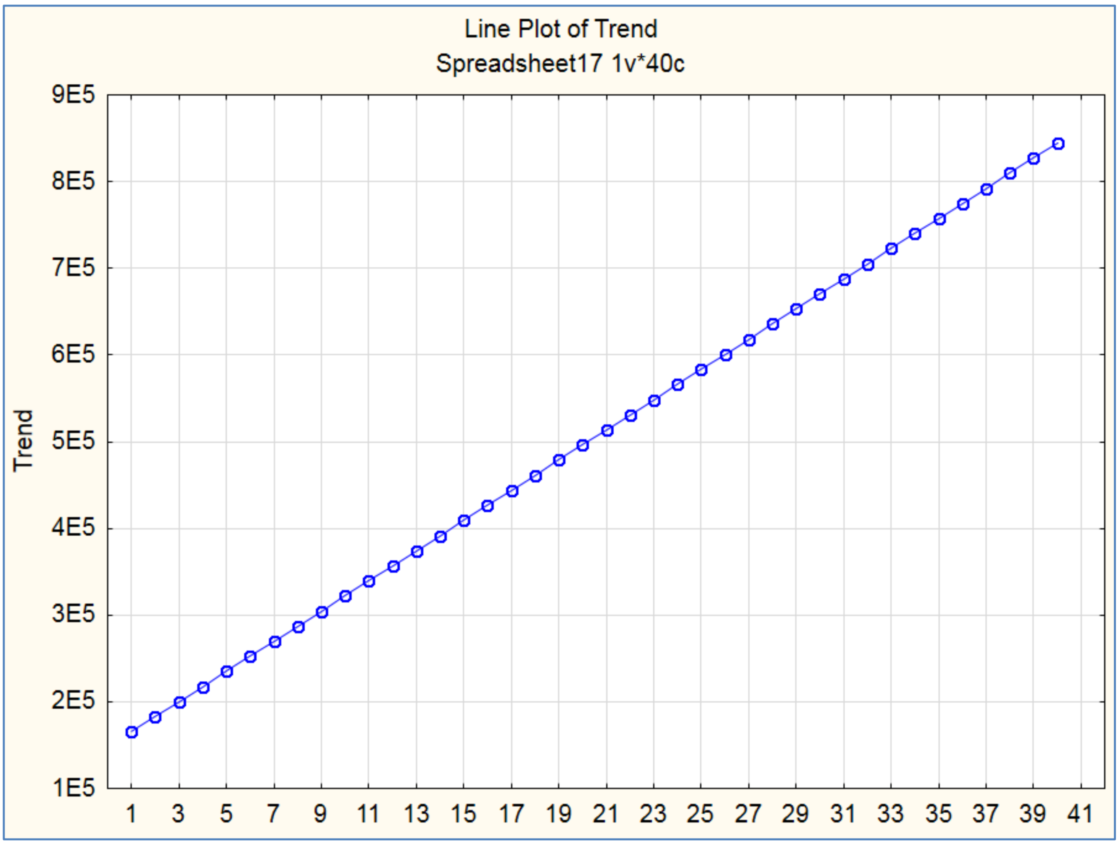
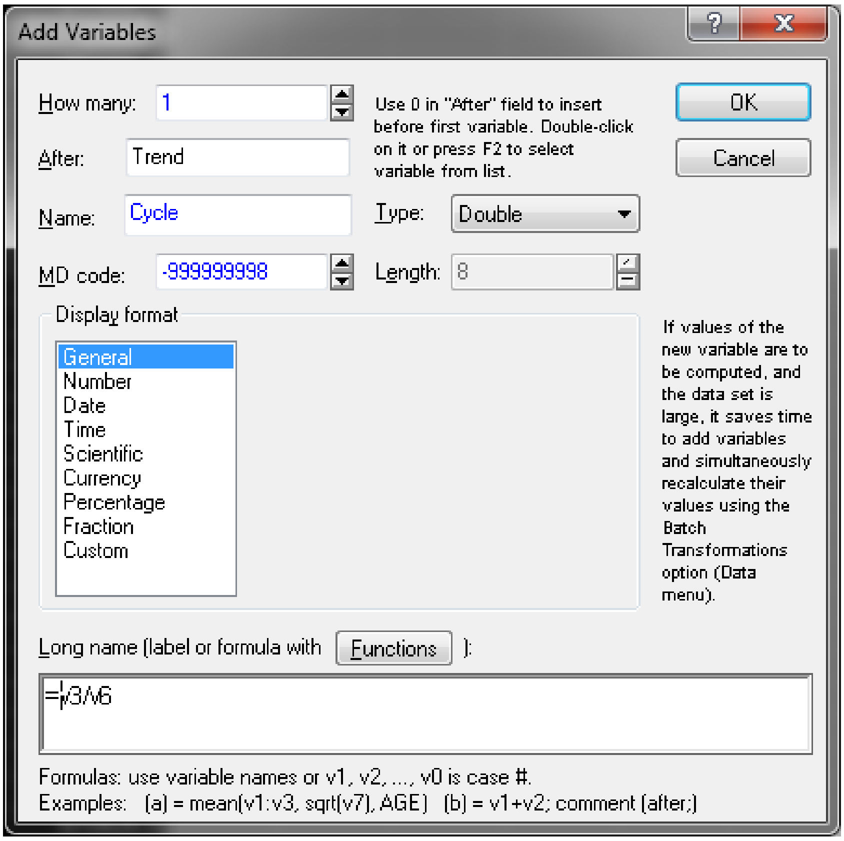


Fig. 26. **Visualization of the trend component**



How many: 1

After: Trend

Name: Cycle

MD code: -999999998

Use 0 in "After" field to insert before first variable. Double-click on it or press F2 to select variable from list.

Type: Double

Length: 8

Display format

- General
- Number
- Date
- Time
- Scientific
- Currency
- Percentage
- Fraction
- Custom

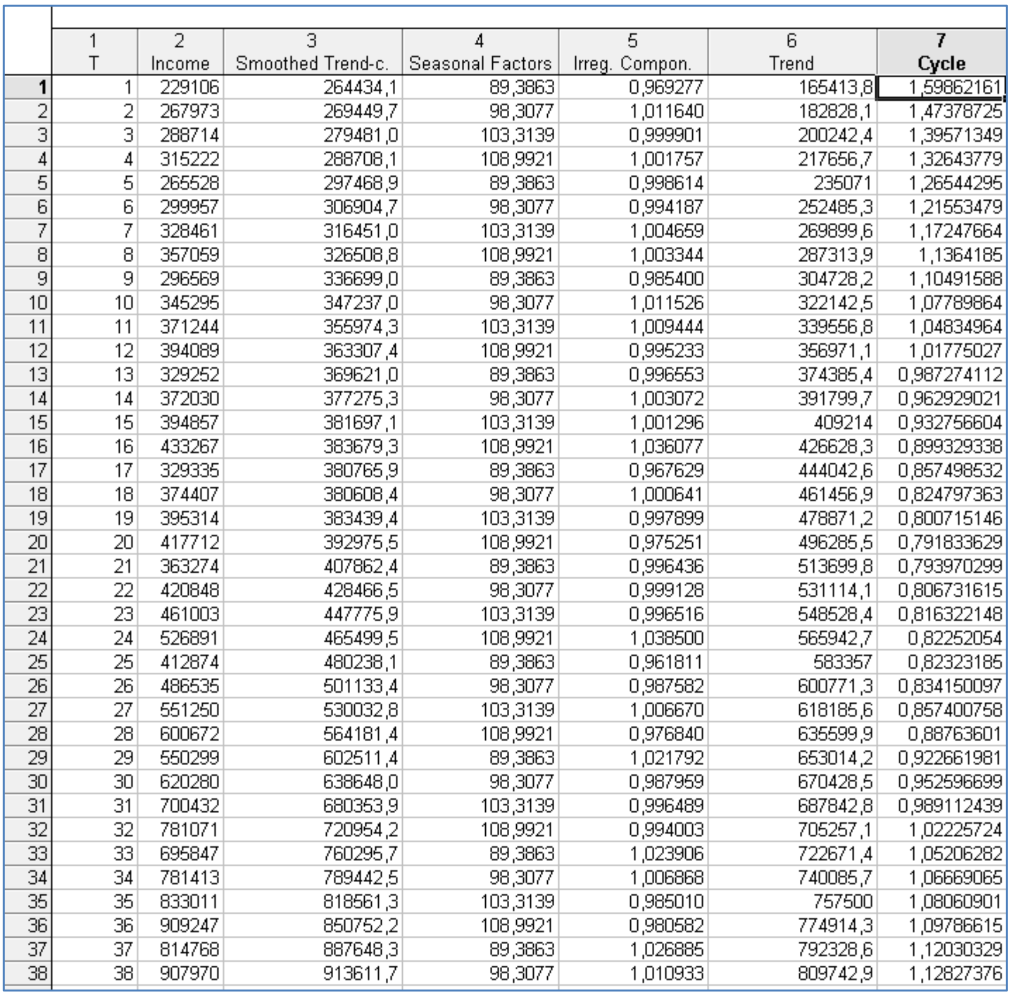
If values of the new variable are to be computed, and the data set is large, it saves time to add variables and simultaneously recalculate their values using the Batch Transformations option (Data menu).

Long name (label or formula with Functions):

```
=v3/v6
```

Formulas: use variable names or v1, v2, ..., v0 is case #.
Examples: (a) = mean(v1:v3, sqrt(v7), AGE) (b) = v1+v2; comment (after;)

Fig. 27. **Adding the cycle component**



	1	2	3	4	5	6	7
	T	Income	Smoothed Trend-c.	Seasonal Factors	Irreg. Compon.	Trend	**Cycle**
1	1	229106	264434,1	89,3863	0,969277	165413,8	1,59862161
2	2	267973	269449,7	98,3077	1,011640	182828,1	1,47378725
3	3	288714	279481,0	103,3139	0,999901	200242,4	1,39571349
4	4	315222	288708,1	108,9921	1,001757	217656,7	1,32643779
5	5	265528	297468,9	89,3863	0,998614	235071	1,26544295
6	6	299957	306904,7	98,3077	0,994187	252485,3	1,21553479
7	7	328461	316451,0	103,3139	1,004659	269899,6	1,17247664
8	8	357059	326508,8	108,9921	1,003344	287313,9	1,1364185
9	9	296569	336699,0	89,3863	0,985400	304728,2	1,10491588
10	10	345295	347237,0	98,3077	1,011526	322142,5	1,07789864
11	11	371244	355974,3	103,3139	1,009444	339556,8	1,04834964
12	12	394089	363307,4	108,9921	0,995233	356971,1	1,01775027
13	13	329252	369621,0	89,3863	0,996553	374385,4	0,987274112
14	14	372030	377275,3	98,3077	1,003072	391799,7	0,962929021
15	15	394857	381697,1	103,3139	1,001296	409214	0,932756604
16	16	433267	383679,3	108,9921	1,036077	426628,3	0,899329338
17	17	329335	380765,9	89,3863	0,967629	444042,6	0,857498532
18	18	374407	380608,4	98,3077	1,000641	461456,9	0,824797363
19	19	395314	383439,4	103,3139	0,997899	478871,2	0,800715146
20	20	417712	392975,5	108,9921	0,975251	496285,5	0,791833629
21	21	363274	407862,4	89,3863	0,996436	513699,8	0,793970299
22	22	420848	428466,5	98,3077	0,999128	531114,1	0,806731615
23	23	461003	447775,9	103,3139	0,996516	548528,4	0,816322148
24	24	526891	465499,5	108,9921	1,038500	565942,7	0,82252054
25	25	412874	480238,1	89,3863	0,961811	583357	0,82323185
26	26	486535	501133,4	98,3077	0,987582	600771,3	0,834150097
27	27	551250	530032,8	103,3139	1,006670	618185,6	0,857400758
28	28	600672	564181,4	108,9921	0,976840	635599,9	0,88763601
29	29	550299	602511,4	89,3863	1,021792	653014,2	0,922661981
30	30	620280	638648,0	98,3077	0,987959	670428,5	0,952596699
31	31	700432	680353,9	103,3139	0,996489	687842,8	0,989112439
32	32	781071	720954,2	108,9921	0,994003	705257,1	1,02225724
33	33	695847	760295,7	89,3863	1,023906	722671,4	1,05206282
34	34	781413	789442,5	98,3077	1,006868	740085,7	1,06669065
35	35	833011	818561,3	103,3139	0,985010	757500	1,08060901
36	36	909247	850752,2	108,9921	0,980582	774914,3	1,09786615
37	37	814768	887648,3	89,3863	1,026885	792328,6	1,12030329
38	38	907970	913611,7	98,3077	1,010933	809742,9	1,12827376

Fig. 28. **The results of adding the cycle component**

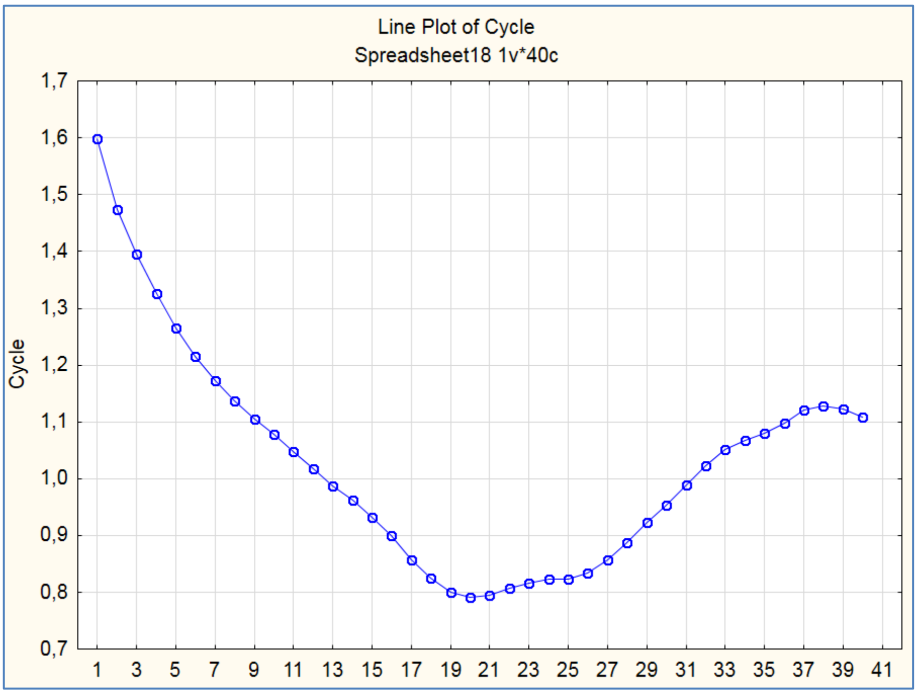
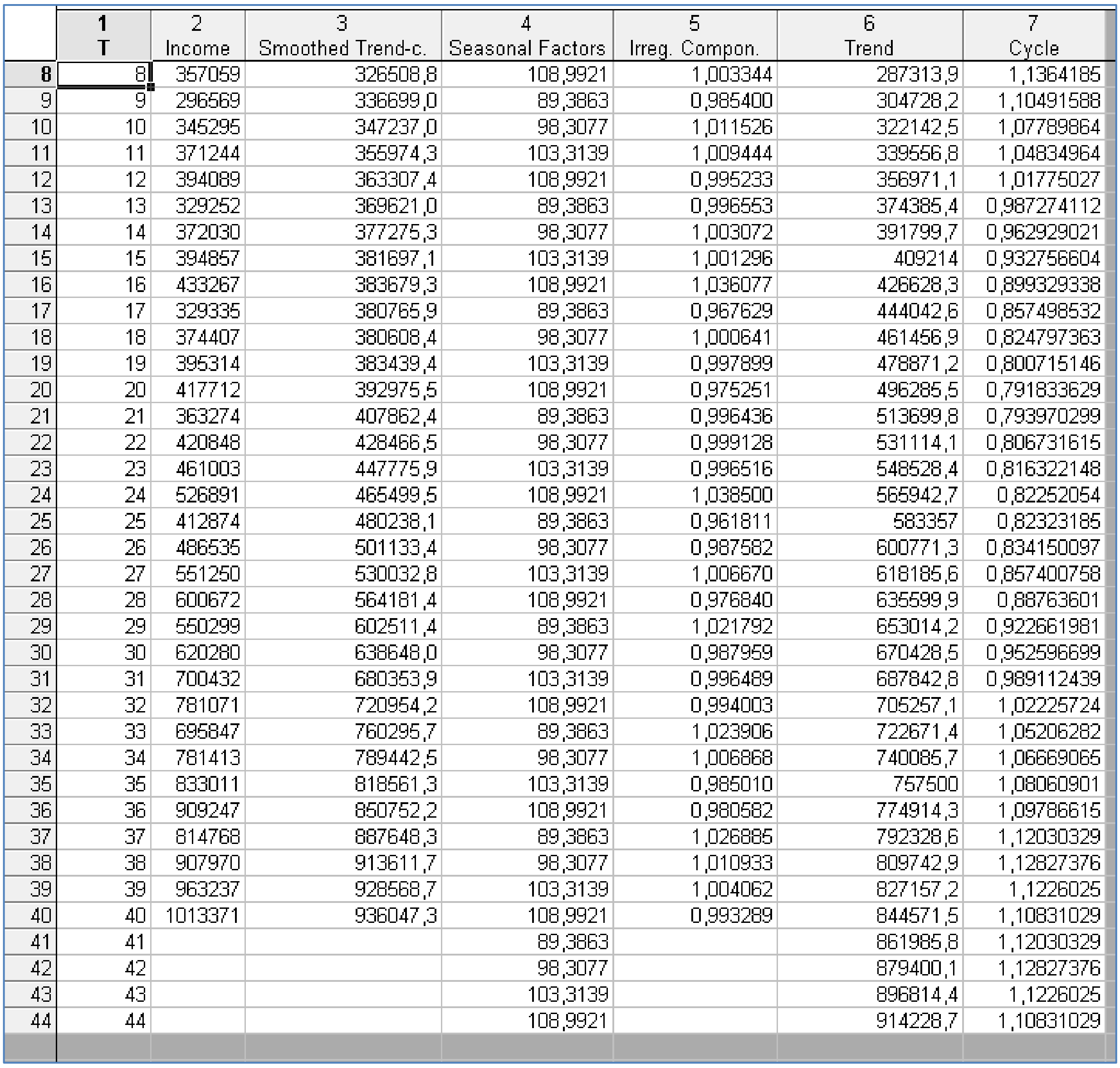


Fig. 29. **Visualization of the cycle component**

Before proceeding to forecasting the income for 1 year (4 quartiles = = 4 periods) ahead using the time series decomposition model, it is necessary to perform a number of actions:

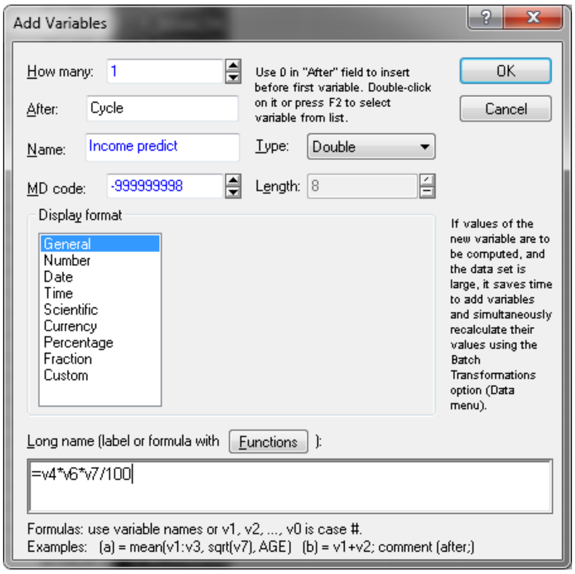
- add 4 observations after the last one available in the series;
- in the data column *T* (time period), enter the corresponding ordinal numbers, continuing the series;
- in the column *Seasonal Factors*, enter the corresponding values of the seasonal components;
- in the *Cycle* column, enter the corresponding cyclic component values, taking into account the cycle period;
- in the *Trend* column set the data recalculation (Fig. 30).



	1	2	3	4	5	6	7
	T	Income	Smoothed Trend-c.	Seasonal Factors	Irreg. Compon.	Trend	Cycle
8	8	357059	326508,8	108,9921	1,003344	287313,9	1,1364185
9	9	296569	336699,0	89,3863	0,985400	304728,2	1,10491588
10	10	345295	347237,0	98,3077	1,011526	322142,5	1,07789864
11	11	371244	355974,3	103,3139	1,009444	339556,8	1,04834964
12	12	394089	363307,4	108,9921	0,995233	356971,1	1,01775027
13	13	329252	369621,0	89,3863	0,996553	374385,4	0,987274112
14	14	372030	377275,3	98,3077	1,003072	391799,7	0,962929021
15	15	394857	381697,1	103,3139	1,001296	409214	0,932756604
16	16	433267	383679,3	108,9921	1,036077	426628,3	0,899329338
17	17	329335	380765,9	89,3863	0,967629	444042,6	0,857498532
18	18	374407	380608,4	98,3077	1,000641	461456,9	0,824797363
19	19	395314	383439,4	103,3139	0,997899	478871,2	0,800715146
20	20	417712	392975,5	108,9921	0,975251	496285,5	0,791833629
21	21	363274	407862,4	89,3863	0,996436	513699,8	0,793970299
22	22	420848	428466,5	98,3077	0,999128	531114,1	0,806731615
23	23	461003	447775,9	103,3139	0,996516	548528,4	0,816322148
24	24	526891	465499,5	108,9921	1,038500	565942,7	0,82252054
25	25	412874	480238,1	89,3863	0,961811	583357	0,82323185
26	26	486535	501133,4	98,3077	0,987582	600771,3	0,834150097
27	27	551250	530032,8	103,3139	1,006670	618185,6	0,857400758
28	28	600672	564181,4	108,9921	0,976840	635599,9	0,88763601
29	29	550299	602511,4	89,3863	1,021792	653014,2	0,922661981
30	30	620280	638648,0	98,3077	0,987959	670428,5	0,952596699
31	31	700432	680353,9	103,3139	0,996489	687842,8	0,989112439
32	32	781071	720954,2	108,9921	0,994003	705257,1	1,02225724
33	33	695847	760295,7	89,3863	1,023906	722671,4	1,05206282
34	34	781413	789442,5	98,3077	1,006868	740085,7	1,06669065
35	35	833011	818561,3	103,3139	0,985010	757500	1,08060901
36	36	909247	850752,2	108,9921	0,980582	774914,3	1,09786615
37	37	814768	887648,3	89,3863	1,026885	792328,6	1,12030329
38	38	907970	913611,7	98,3077	1,010933	809742,9	1,12827376
39	39	963237	928568,7	103,3139	1,004062	827157,2	1,1226025
40	40	1013371	936047,3	108,9921	0,993289	844571,5	1,10831029
41	41			89,3863		861985,8	1,12030329
42	42			98,3077		879400,1	1,12827376
43	43			103,3139		896814,4	1,1226025
44	44			108,9921		914228,7	1,10831029

Fig. 30. **The steps of the analysis**

Then we need to add a new variable – the *Income predict* (Fig. 31, 32).



How many: 1

After: Cycle

Use 0 in "After" field to insert before first variable. Double-click on it or press F2 to select variable from list.

Name: Income predict

Type: Double

MD code: -999999998

Length: 8

Display format

- General
- Number
- Date
- Time
- Scientific
- Currency
- Percentage
- Fraction
- Custom

If values of the new variable are to be computed, and the data set is large, it saves time to add variables and simultaneously recalculate their values using the Batch Transformations option (Data menu).

Long name (label or formula with Functions):

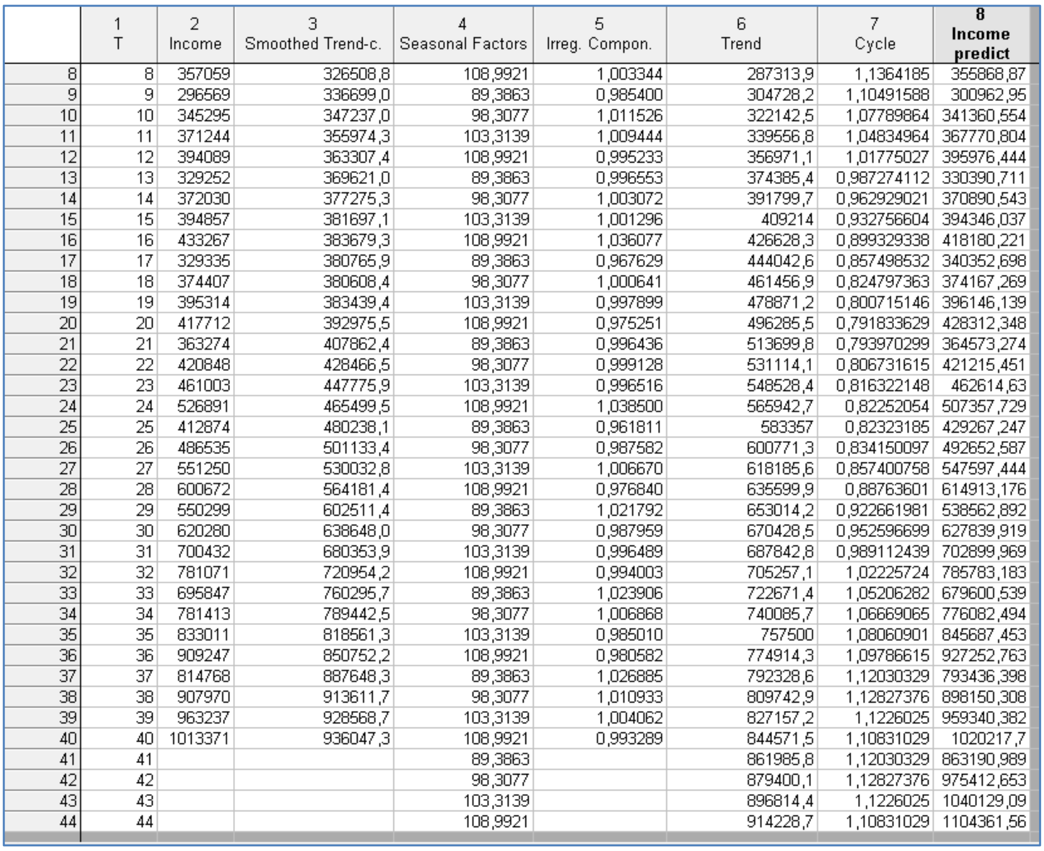
=v4*v6*v7/100

Formulas: use variable names or v1, v2, ..., v0 is case #.
Examples: (a) = mean(v1:v3, sqrt(v7), AGE) (b) = v1+v2; comment (after;)

Fig. 31. **Adding the *Income predict***

Then you can calculate the forecast values of the income indicator in 4 steps forward by specifying a model of the form:

Income predict = Trend × Cycle × Seasonal Factors / 100.



	1 T	2 Income	3 Smoothed Trend-c.	4 Seasonal Factors	5 Irreg. Compon.	6 Trend	7 Cycle	**8 Income predict**
8	8	357059	326508,8	108,9921	1,003344	287313,9	1,1364185	355868,87
9	9	296569	336699,0	89,3863	0,985400	304728,2	1,10491588	300962,95
10	10	345295	347237,0	98,3077	1,011526	322142,5	1,07789864	341360,554
11	11	371244	355974,3	103,3139	1,009444	339556,8	1,04834964	367770,804
12	12	394089	363307,4	108,9921	0,995233	356971,1	1,01775027	395976,444
13	13	329252	369621,0	89,3863	0,996553	374385,4	0,987274112	330390,711
14	14	372030	377275,3	98,3077	1,003072	391799,7	0,962929021	370890,543
15	15	394857	381697,1	103,3139	1,001296	409214	0,932756604	394346,037
16	16	433267	383679,3	108,9921	1,036077	426628,3	0,899329338	418180,221
17	17	329335	380765,9	89,3863	0,967629	444042,6	0,857498532	340352,698
18	18	374407	380608,4	98,3077	1,000641	461456,9	0,824797363	374167,269
19	19	395314	383439,4	103,3139	0,997899	478871,2	0,800715146	396146,139
20	20	417712	392975,5	108,9921	0,975251	496285,5	0,791833629	428312,348
21	21	363274	407862,4	89,3863	0,996436	513699,8	0,793970299	364573,274
22	22	420848	428466,5	98,3077	0,999128	531114,1	0,806731615	421215,451
23	23	461003	447775,9	103,3139	0,996516	548528,4	0,816322148	462614,63
24	24	526891	465499,5	108,9921	1,038500	565942,7	0,82252054	507357,729
25	25	412874	480238,1	89,3863	0,961811	583357	0,82323185	429267,247
26	26	486535	501133,4	98,3077	0,987582	600771,3	0,834150097	492652,587
27	27	551250	530032,8	103,3139	1,006670	618185,6	0,857400758	547597,444
28	28	600672	564181,4	108,9921	0,976840	635599,9	0,88763601	614913,176
29	29	550299	602511,4	89,3863	1,021792	653014,2	0,922661981	538562,892
30	30	620280	638648,0	98,3077	0,987959	670428,5	0,952596699	627839,919
31	31	700432	680353,9	103,3139	0,996489	687842,8	0,989112439	702899,969
32	32	781071	720954,2	108,9921	0,994003	705257,1	1,02225724	785783,183
33	33	695847	760295,7	89,3863	1,023906	722671,4	1,05206282	679600,539
34	34	781413	789442,5	98,3077	1,006868	740085,7	1,06669065	776082,494
35	35	833011	818561,3	103,3139	0,985010	757500	1,08060901	845687,453
36	36	909247	850752,2	108,9921	0,980582	774914,3	1,09786615	927252,763
37	37	814768	887648,3	89,3863	1,026885	792328,6	1,12030329	793436,398
38	38	907970	913611,7	98,3077	1,010933	809742,9	1,12827376	898150,308
39	39	963237	928568,7	103,3139	1,004062	827157,2	1,1226025	959340,382
40	40	1013371	936047,3	108,9921	0,993289	844571,5	1,10831029	1020217,7
41	41			89,3863		861985,8	1,12030329	863190,989
42	42			98,3077		879400,1	1,12827376	975412,653
43	43			103,3139		896814,4	1,1226025	1040129,09
44	44			108,9921		914228,7	1,10831029	1104361,56

Fig. 32. **The results of adding the *Income predict***

Let's build a graph of the income predict (Fig. 33).

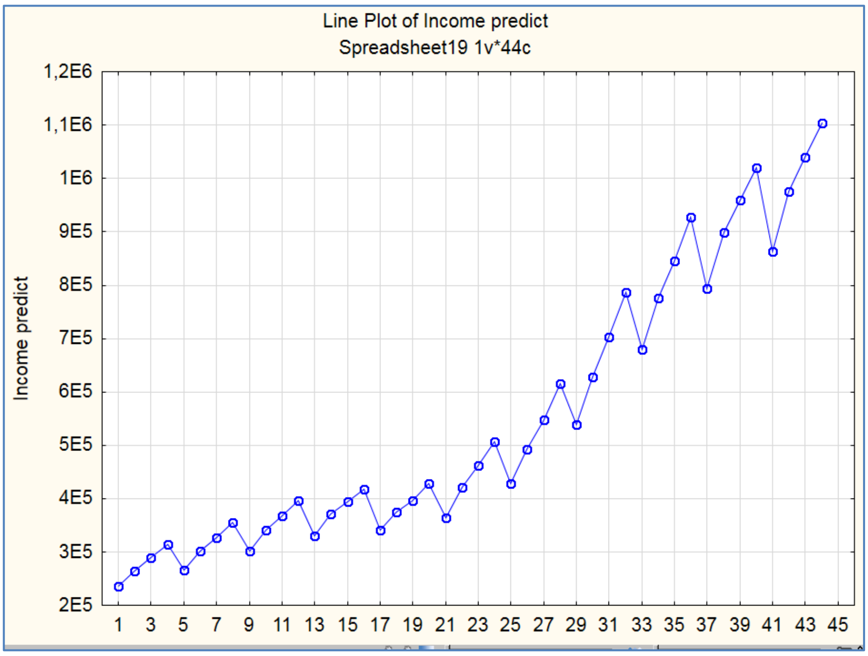


Fig. 33. **Visualization of the income predict**

Fig. 34 shows the histogram of error distribution (visualization of the irregular component with the help of the menu *Graphs of Block Data* / *Histogram: Block Columns*). The fact that this distribution is close to the normal law is a confirmation of the adequacy of the model and the accuracy of the forecast:

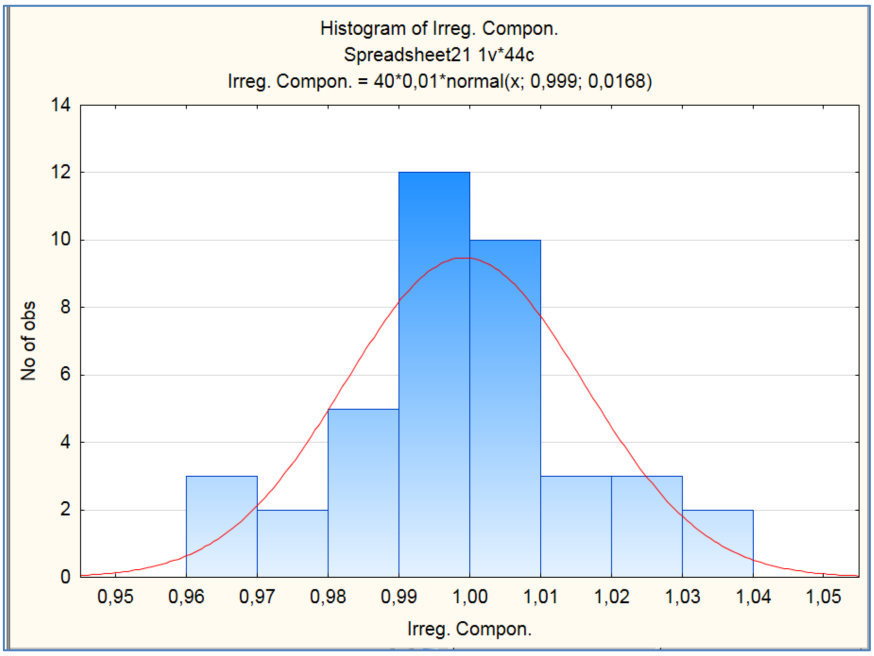


Fig. 34. **The histogram of the income predict with the curve of normal distribution**

To confirm the quality of the forecast, we calculate the average absolute percentage error by formula 2:

$$MAPE = \frac{1}{n} \times \sum \frac{|y - \bar{y}|}{y} \times 100\ \%, \qquad (2)$$

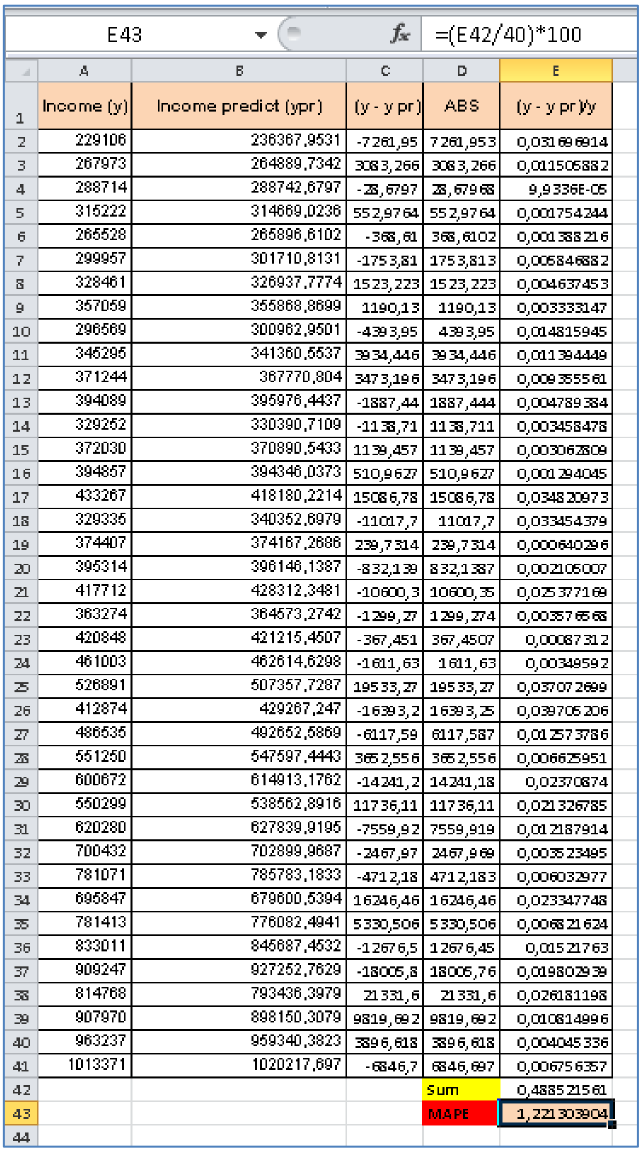
where n is the number of cases (periods);

y is the analyzed value;

$\bar{y}$ is the predicted value.

MAPE is a measure of the bias of the forecast (predicted errors of the time series). If MAPE < 10 %, this will indicate a high accuracy of the constructed forecast.

Calculations can be performed in MS Excel (Fig. 35).



E43 =(E42/40)*100

	A	B	C	D	E
1	Income (y)	Income predict (ypr)	(y - y pr)	ABS	(y - y pr)/y
2	229106	236367,9531	-7261,95	7261,953	0,031696914
3	267973	264889,7342	3083,266	3083,266	0,011505882
4	288714	288742,6797	-28,6797	28,67968	9,9336E-05
5	315222	314669,0236	552,9764	552,9764	0,001754244
6	265528	265896,6102	-368,61	368,6102	0,001388216
7	299957	301710,8131	-1753,81	1753,813	0,005846882
8	328461	326937,7774	1523,223	1523,223	0,004637453
9	357059	355868,8699	1190,13	1190,13	0,003333147
10	296569	300962,9501	-4393,95	4393,95	0,014815945
11	345295	341360,5537	3934,446	3934,446	0,011394449
12	371244	367770,804	3473,196	3473,196	0,009355561
13	394089	395976,4437	-1887,44	1887,444	0,004789384
14	329252	330390,7109	-1138,71	1138,711	0,003458478
15	372030	370890,5433	1139,457	1139,457	0,003062809
16	394857	394346,0373	510,9627	510,9627	0,001294045
17	433267	418180,2214	15086,78	15086,78	0,034820973
18	329335	340352,6979	-11017,7	11017,7	0,033454379
19	374407	374167,2686	239,7314	239,7314	0,000640296
20	395314	396146,1387	-832,139	832,1387	0,002105007
21	417712	428312,3481	-10600,3	10600,35	0,025377169
22	363274	364573,2742	-1299,27	1299,274	0,003576568
23	420848	421215,4507	-367,451	367,4507	0,00087312
24	461003	462614,6298	-1611,63	1611,63	0,00349592
25	526891	507357,7287	19533,27	19533,27	0,037072699
26	412874	429267,247	-16393,2	16393,25	0,039705206
27	486535	492652,5869	-6117,59	6117,587	0,012573786
28	551250	547597,4443	3652,556	3652,556	0,006625951
29	600672	614913,1762	-14241,2	14241,18	0,02370874
30	550299	538562,8916	11736,11	11736,11	0,021326785
31	620280	627839,9195	-7559,92	7559,919	0,012187914
32	700432	702899,9687	-2467,97	2467,969	0,003523495
33	781071	785783,1833	-4712,18	4712,183	0,006032977
34	695847	679600,5394	16246,46	16246,46	0,023347748
35	781413	776082,4941	5330,506	5330,506	0,006821624
36	833011	845687,4532	-12676,5	12676,45	0,01521763
37	909247	927252,7629	-18005,8	18005,76	0,019802939
38	814768	793436,3979	21331,6	21331,6	0,026181198
39	907970	898150,3079	9819,692	9819,692	0,010814996
40	963237	959340,3823	3896,618	3896,618	0,004045336
41	1013371	1020217,697	-6846,7	6846,697	0,006756357
42				Sum	0,488521561
43				MAPE	1,221303904
44					

Fig. 35. **Calculations of the MAPE**

The average absolute percentage error is 1.221 %, which is confirmed by the accuracy of the forecast for this multiplicative time series model.

As an output, it is necessary to give an economic interpretation of the results of forecasting, i.e. say what will happen with the population income in 2020.

Task 2. Using your own research information space and analytical functions of the program Statistica 10.0, build a forecast for 2 – 3 periods ahead and provide an economic interpretation of the results.

Task 3. Build a forecast of the country's export volume (Table 9) for 4 periods ahead:

1. Determine the form of the decomposition model.
2. Identify all components of the decomposition model.
3. Forecast the trend component.
4. Carry out a spectral analysis of the cyclic constituent.
5. Check the quality of the decomposition model.
6. Formulate conclusions.

Table 9

The input data

Period	Export, mln UAH	Import, mln UAH	Period	Export, mln UAH	Import, mln UAH
1	2 339 402.7	2 712 852.43	19	4 258 746.5	5 315 298.1
2	2 518 534.3	3 171 353.67	20	4 167 498.9	4 872 915.4
3	3 129 139.2	3 872 083.7	21	4 114 710.7	4 851 865.6
4	2 953 856.7	3 283 634.6	22	4 345 291.8	5 872 680.3
5	3 104624.5	3 63 2466	23	4 450 168.3	5 822 204
6	3 317 653.5	3 610 201.2	24	4 799 157.8	6 628 819.8
7	3 344 490.8	3 686 831.1	25	3 663 214.9	4 627 526.5
8	3 511 110.9	3 839 422.4	26	4 682 418.3	6 46 557.5
9	3 676 317.9	4 147 061.3	27	5 444 491.8	7 712 994.2
10	3 437 449.5	4 021 637.6	28	5 571 314.2	7 936 247.4
11	3 344 121.8	3 92 2054	29	6 284 581.1	7 710 564
12	3 691 002.6	5 134 893.1	30	6 896 161.2	7 934 9 61.4
13	3 208 484.9	3 700 647.9	31	7 616 856.6	8 822 925.9
14	3 409 760.2	4 297 569.8	32	6 718 135.9	8 155 969.3
15	4 108 205.4	4 953 434.5	33	6 685 131.4	8 479 144.1
16	4 067 299.9	4 820 236.3	34	5 861 332.6	7 647 194.6
17	4 083 445.1	4 852 211.6	35	3 622 525.2	5 264 462.7
18	4 235 294.1	4 682 039.7	36	4 287 044.3	5 908 471.9

Task 4. Build a forecast of the country's import volume (Table 9) for 4 periods ahead:

1. Determine the form of the decomposition model.
2. Identify all components of the decomposition model.
3. Forecast the trend component.
4. Carry out a spectral analysis of the cyclic constituent.
5. Check the quality of the decomposition model.
6. Formulate conclusions.

Task 5. Compare the results of the forecasting of the country's import and export volume. Explain the changes. How does the government foreign and domestic economic policy affect these areas?

Task 6. Using the statistical data of the website of the State Statistics Service of Ukraine [19], find information on the amount of capital investment (on a quarterly basis) and build a forecast for 1 year ahead (4 periods) using the module "Analysis of time series module" of the program Statistica 10.0.

Task 7. Using the statistical data of the website of the State Statistics Service of Ukraine [19], find information on the amount of the GDP (on a quarterly basis) and build a forecast for 5 years ahead using the module "Analysis of time series module" of the program Statistica 10.0.

The list of questions for independent work

1. What is the time series of spatial differences?
2. What types of time series are there?
3. How is the Durbin – Watson method used to study autocorrelation?
4. What is the definition of a stationary time series in the broadest sense?
5. What are the methods of stationary research?
6. In what case are additive and multiplicative models used?
7. List the antialiasing models.
8. What are the main stages of the Foster – Stewart method?
9. What are the main steps for constructing a decomposition model?
10. What methods are used to check the forecast quality?

Topic 4. Models of adaptive forecasting and an integrated model of autoregression

Task 1. Carry out smoothing of the time series by the method of exponential smoothing with different values of parameters. Give graphs of the smoothed data and the corresponding forecast values of the indicator. Build a forecast of changes in investment in the USA IT sector for 1 period ahead (the initial data are presented in Table 10). Evaluate the quality of the time series models (mean error, mean absolute error, standard deviation of errors, mean percentage error, mean absolute percentage error). Perform a comparative analysis of the models and determine the most adequate of them.

Provide an economic interpretation of the results.

Table 10

The input data

Period	Investments in the USA IT sector, billion dollars	Period	Investments in the USA IT sector, billion dollars
2014	380	2017	470
	420		430
	400		450
	480		480
2015	350	2018	490
	400		470
	450		490
	480		478
2016	370	2019	520
	450		490
	400		510
	420		410

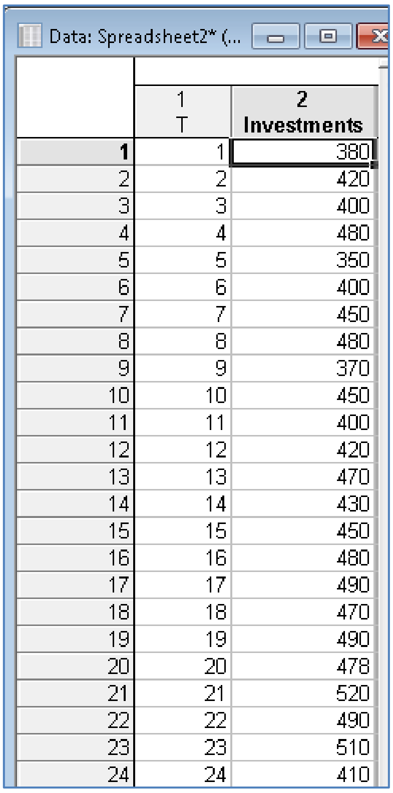
Guidelines

1. As always, we start by creating a source file in Statistica 10.0. To do this, specify the number of variables (2) and the objects of study (24 cases) (Fig. 36).

Let's build a graph of the original data.

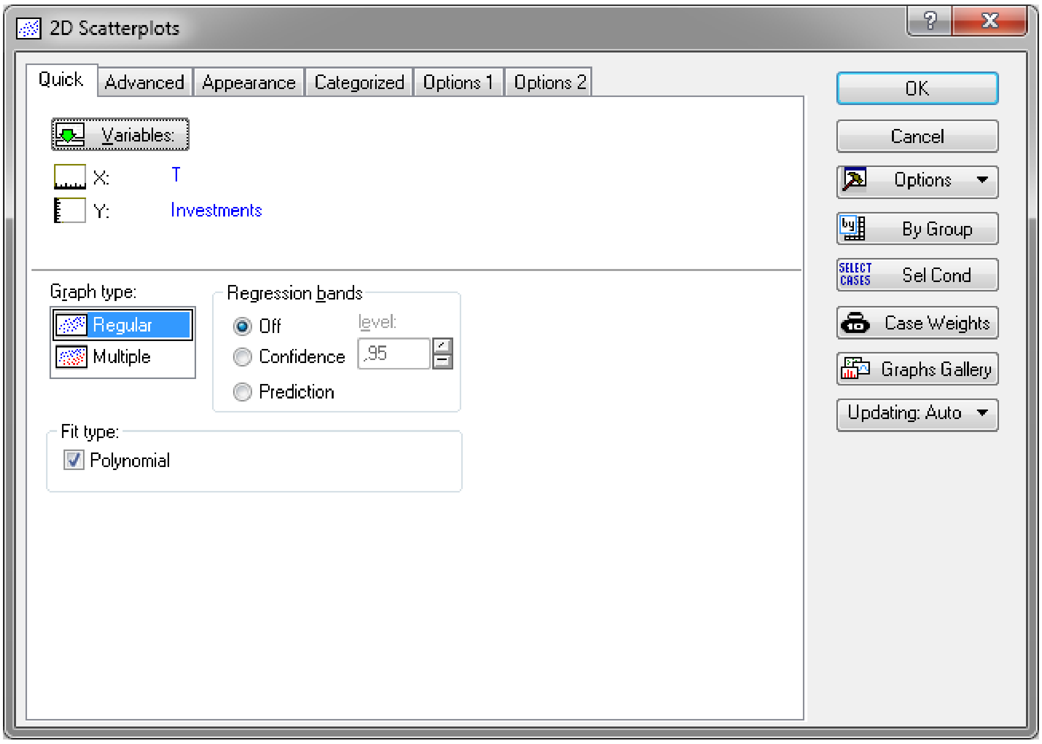
To plot the source data, select the *Scatterplots* sub-item in the *Graph items* menu. When setting the characteristics of the graph, select *Graph type* – *Regular*, go to the *Advanced* tab, and select the trend type – *Polynomial.* As variables displayed on the graph, select the period of time T for the X-axis and the investments for the Y-axis. The result is shown in Fig. 37, 38.

The choice out of the two main models of the time series decomposition – additive or multiplicative – is carried out using graphical analysis. The following rule is used for choosing a particular model: if the initial data have a constantly increasing or decreasing amplitude of oscillations of values, it is advisable to use a multiplicative model of decomposition; in the case of a constant amplitude of changes, it is advisable to use an additive model.



	1 T	2 Investments
1	1	380
2	2	420
3	3	400
4	4	480
5	5	350
6	6	400
7	7	450
8	8	480
9	9	370
10	10	450
11	11	400
12	12	420
13	13	470
14	14	430
15	15	450
16	16	480
17	17	490
18	18	470
19	19	490
20	20	478
21	21	520
22	22	490
23	23	510
24	24	410

Fig. 36. **The input data in Statistica 10.0**



X: T

Y: Investments

Graph type:
Regular
Multiple

Regression bands
● Off
○ Confidence
○ Prediction
level: ,95

Fit type:
☑ Polynomial

Fig. 37. **Choosing the parameters of the graph**

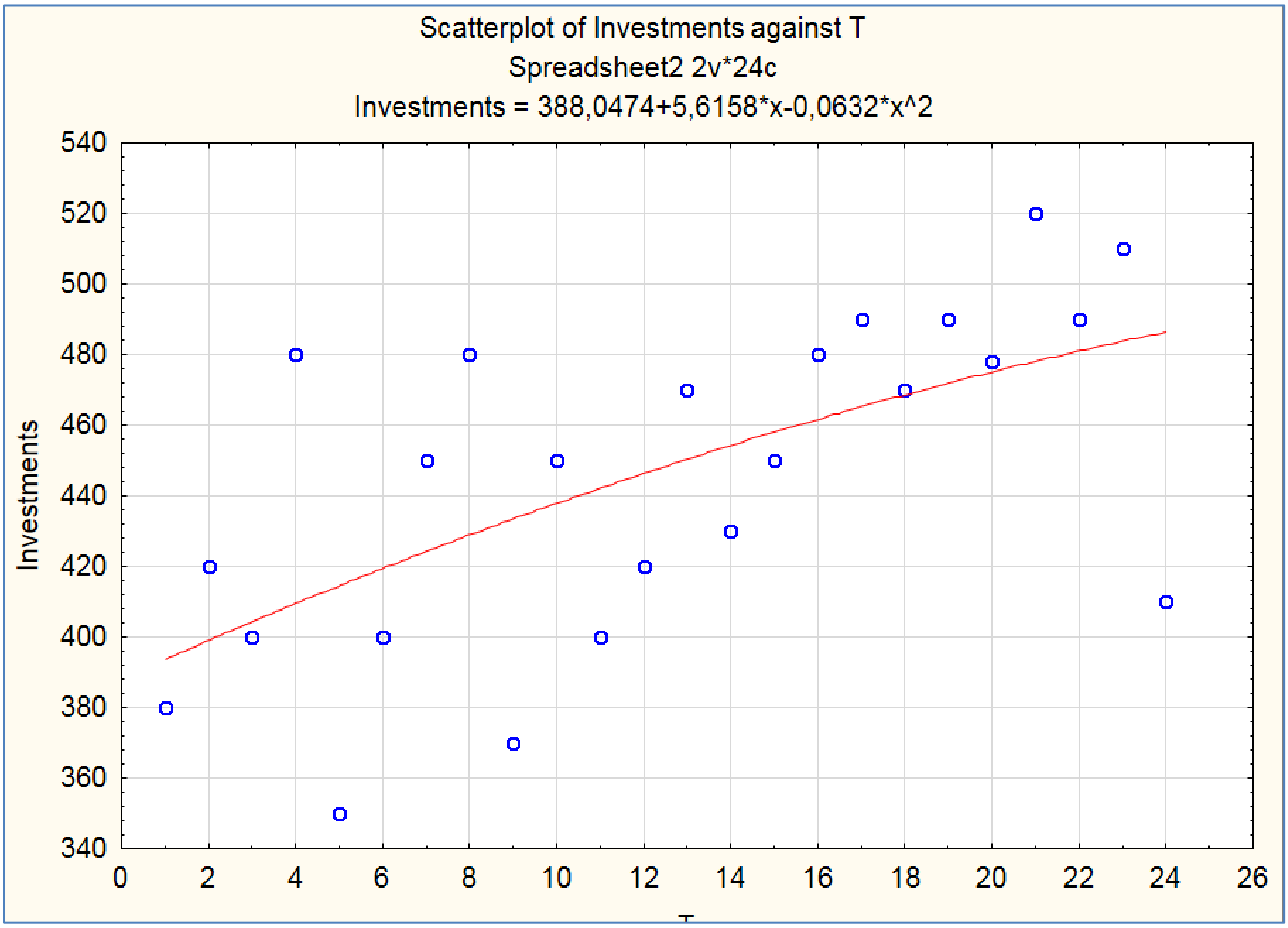


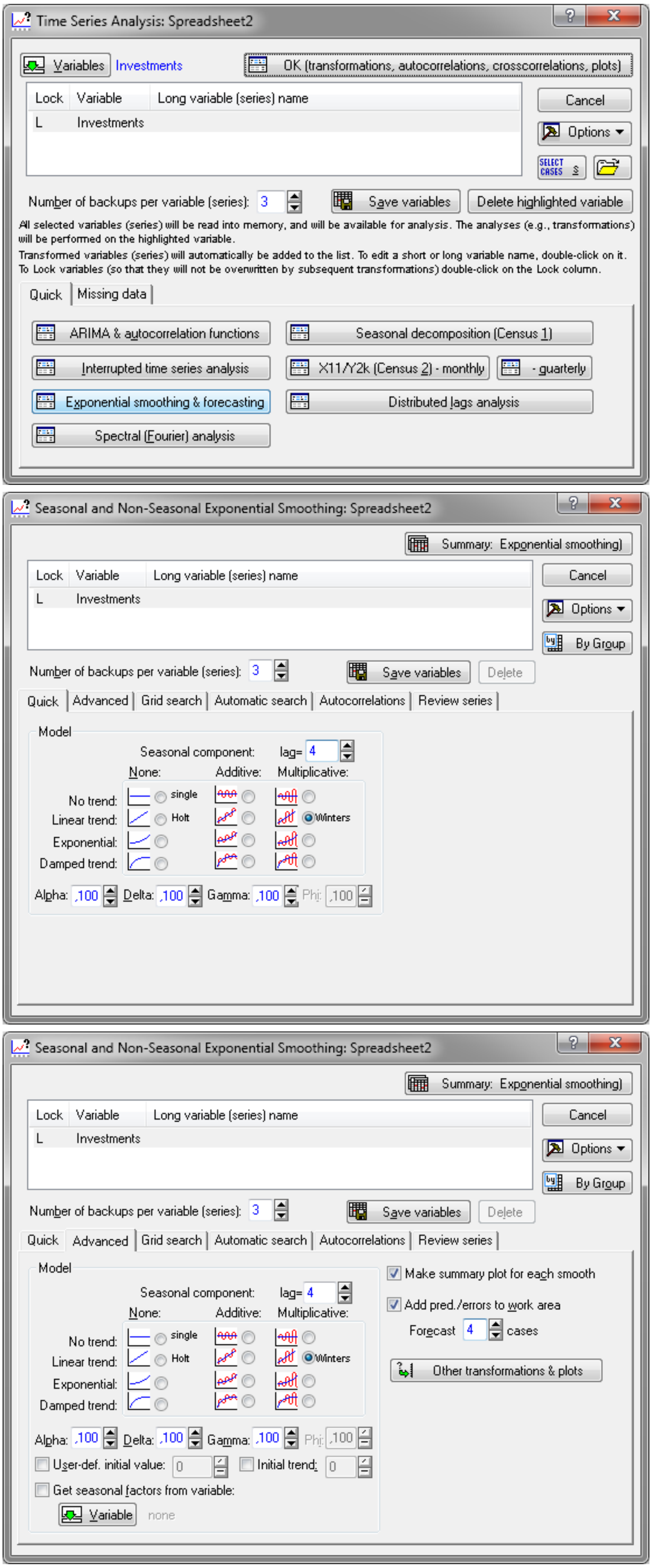
Fig. 38. **The graph of the initial data**

In our case (see Fig. 38), the graphical analysis allows us to conclude about the multiplicative nature of the relationship between the components.

2. Time series analysis is performed in the module *Statistics / Advanced Linear / Nonlinear Models / Time Series / Forecasting (Time series / Forecasting)*.

First, it is necessary to perform exponential smoothing of the original data. To do this, select the tab *Exponential smoothing and forecasting.*

In the next window, you need to set the parameters of exponential smoothing. Thus, the variable for analysis is investments, seasonal component lag (*Seasonal component 4*) since the data are presented on a quarterly basis). The choice of the model type is based on a preliminary visual analysis of the graph of the original data (see Fig. 38). We also specify the forecasting period – 4 (build a forecast for 1 year (4 quarters) ahead) (Fig. 39).



Time Series Analysis: Spreadsheet2
Variables
OK (transformations, autocorrelations, crosscorrelations, plots)
Cancel
Options
SELECT CASES
Save variables
Delete highlighted variable
Quick
Missing data
ARIMA & autocorrelation functions
Seasonal decomposition (Census 1)
Interrupted time series analysis
X11/Y2k (Census 2) - monthly
- quarterly
Exponential smoothing & forecasting
Distributed lags analysis
Spectral (Fourier) analysis

Variables: Investments

Lock	Variable	Long variable (series) name
L	Investments	

Number of backups per variable (series): 3

All selected variables (series) will be read into memory, and will be available for analysis. The analyses (e.g., transformations) will be performed on the highlighted variable.

Transformed variables (series) will automatically be added to the list. To edit a short or long variable name, double-click on it. To Lock variables (so that they will not be overwritten by subsequent transformations) double-click on the Lock column.

Lock	Variable	Long variable (series) name
L	Investments	

Number of backups per variable (series): 3

Seasonal component: lag= 4

● Winters

Alpha: ,100 Delta: ,100 Gamma: ,100 Phi: ,100

Lock	Variable	Long variable (series) name
L	Investments	

Number of backups per variable (series): 3

Seasonal component: lag= 4

● Winters

Alpha: ,100 Delta: ,100 Gamma: ,100 Phi: ,100

□ User-def. initial value: 0 □ Initial trend: 0

□ Get seasonal factors from variable:

Variable none

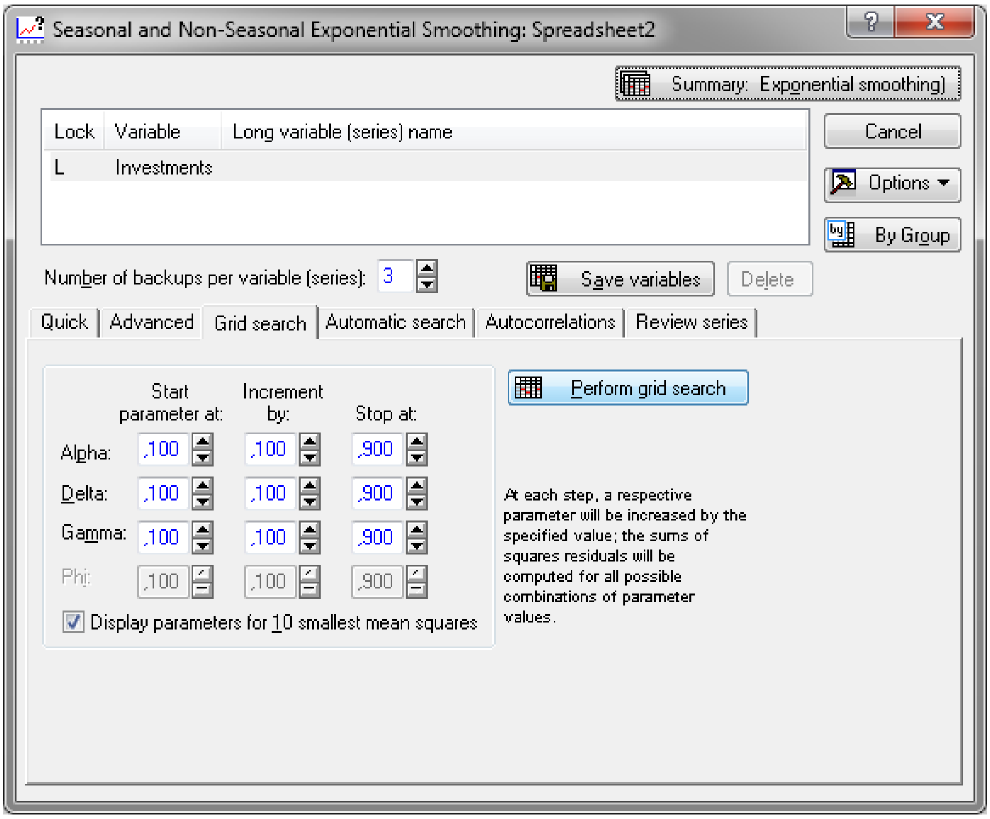
☑ Make summary plot for each smooth

☑ Add pred./errors to work area

Forecast 4 cases

Fig. 39. **Choosing the exponential smoothing parameters**

In the *Grid search* tab, we perform the process of finding and adjusting antialiasing parameters to determine the optimal values for this model. This is important because the performance of the entire model is based on the set values of the smoothing parameters. To do this, click the *Perform grid search* button (Fig. 40).



Seasonal and Non-Seasonal Exponential Smoothing: Spreadsheet2

Lock	Variable	Long variable (series) name
L	Investments	

Number of backups per variable (series): 3

Grid search tab:

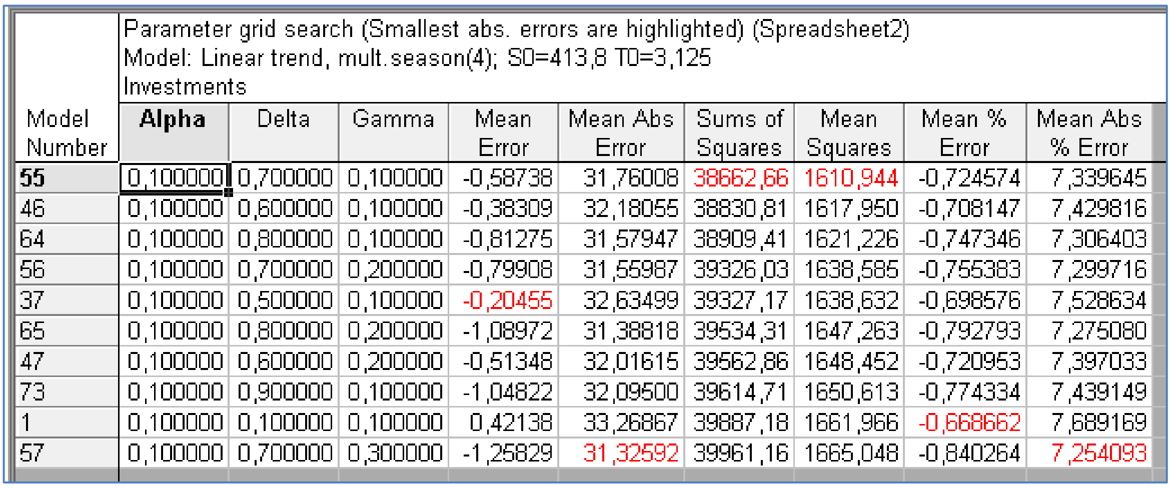
	Start parameter at:	Increment by:	Stop at:
Alpha:	,100	,100	,900
Delta:	,100	,100	,900
Gamma:	,100	,100	,900
Phi:	,100	,100	,900

☑ Display parameters for 10 smallest mean squares

At each step, a respective parameter will be increased by the specified value; the sums of squares residuals will be computed for all possible combinations of parameter values.

Fig. 40. **Launching the *Grid Search* function**

The result is a window of smoothing parameters for a given model (Fig. 41).



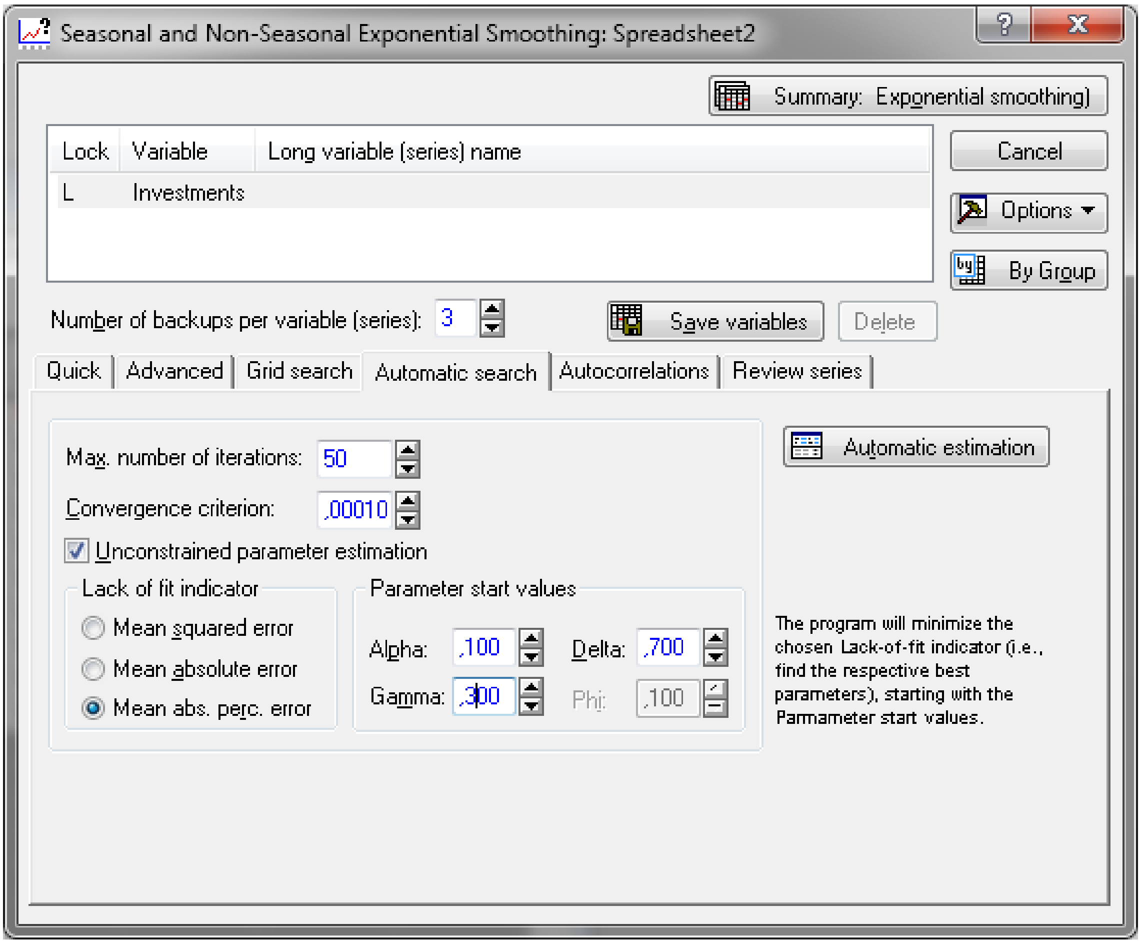
Parameter grid search (Smallest abs. errors are highlighted) (Spreadsheet2)
Model: Linear trend, mult.season(4); S0=413,8 T0=3,125
Investments

Model Number	Alpha	Delta	Gamma	Mean Error	Mean Abs Error	Sums of Squares	Mean Squares	Mean % Error	Mean Abs % Error
55	0,100000	0,700000	0,100000	-0,58738	31,76008	38662,66	1610,944	-0,724574	7,339645
46	0,100000	0,600000	0,100000	-0,38309	32,18055	38830,81	1617,950	-0,708147	7,429816
64	0,100000	0,800000	0,100000	-0,81275	31,57947	38909,41	1621,226	-0,747346	7,306403
56	0,100000	0,700000	0,200000	-0,79908	31,55987	39326,03	1638,585	-0,755383	7,299716
37	0,100000	0,500000	0,100000	-0,20455	32,63499	39327,17	1638,632	-0,698576	7,528634
65	0,100000	0,800000	0,200000	-1,08972	31,38818	39534,31	1647,263	-0,792793	7,275080
47	0,100000	0,600000	0,200000	-0,51348	32,01615	39562,86	1648,452	-0,720953	7,397033
73	0,100000	0,900000	0,100000	-1,04822	32,09500	39614,71	1650,613	-0,774334	7,439149
1	0,100000	0,100000	0,100000	0,42138	33,26867	39887,18	1661,966	-0,668662	7,689169
57	0,100000	0,700000	0,300000	-1,25829	31,32592	39961,16	1665,048	-0,840264	7,254093

Fig. 41. **The smoothing parameters**

The researcher chooses the combination of exponential smoothing parameters that has the lowest value of the mean absolute percentage error (MAPE < 15 %). Thus, in our case, the best forecast will be in the combination of parameters (coefficients) of the model No. 57, namely Alpha = 0.1; Delta = 0.7; Phi = 0.3.

In the next step, set these parameters in the *Automatic search* tab, set one of the criteria for selecting the best results (*Lack of fit indicator*) – mean squared error, mean absolute error, MAPE. Choose the MAPE (Fig. 42).



Lock	Variable	Long variable (series) name
L	Investments	

Number of backups per variable (series): 3

Max. number of iterations: 50

Convergence criterion: ,00010

☑ Unconstrained parameter estimation

Lack of fit indicator

○ Mean squared error
○ Mean absolute error
● Mean abs. perc. error

Parameter start values

Alpha: ,100 Delta: ,700
Gamma: ,300 Phi: ,100

The program will minimize the chosen Lack-of-fit indicator (i.e., find the respective best parameters), starting with the Parmameter start values.

Fig. 42. **Setting the parameters of exponential smoothing**

After clicking the *Automatic estimation* button, we get 3 groups of results at the same time:

graphs of initial data, smoothed and forecast, and model residues (Fig. 43);

a table with the original data, smoothed data (*Smoothed Series*), balances (*Resids*) and seasonal components (*Seasonal Factors*) (Fig. 44);

a model quality assessment table (Fig. 45).

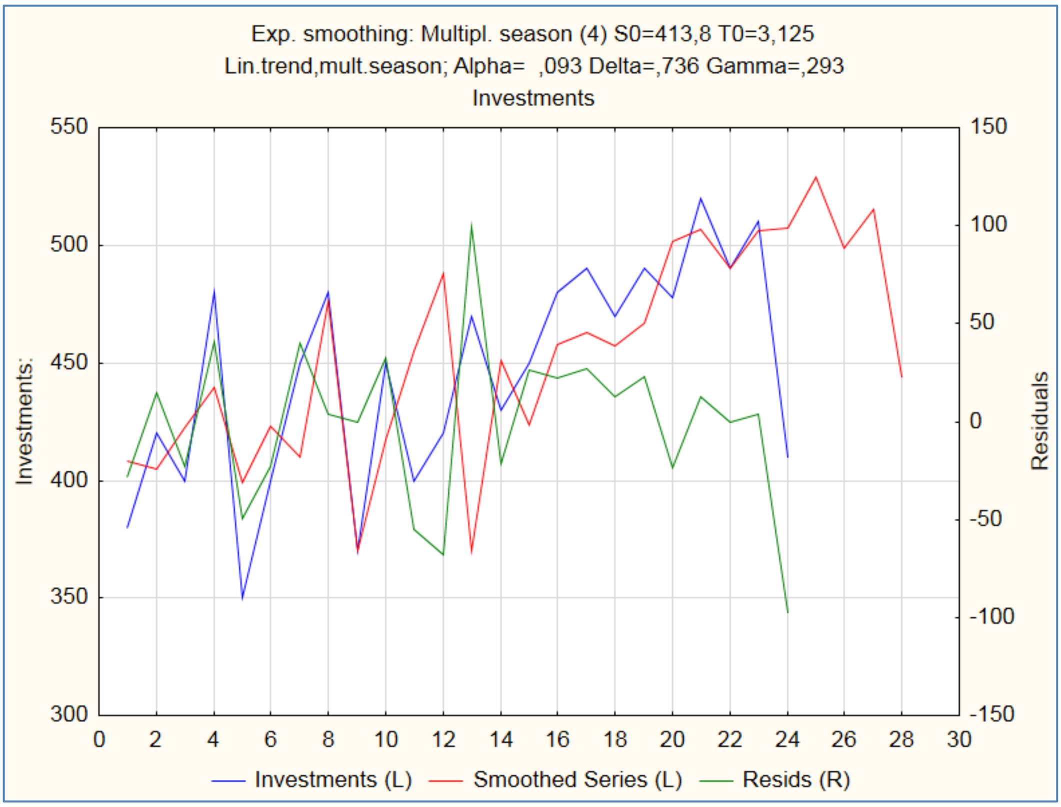
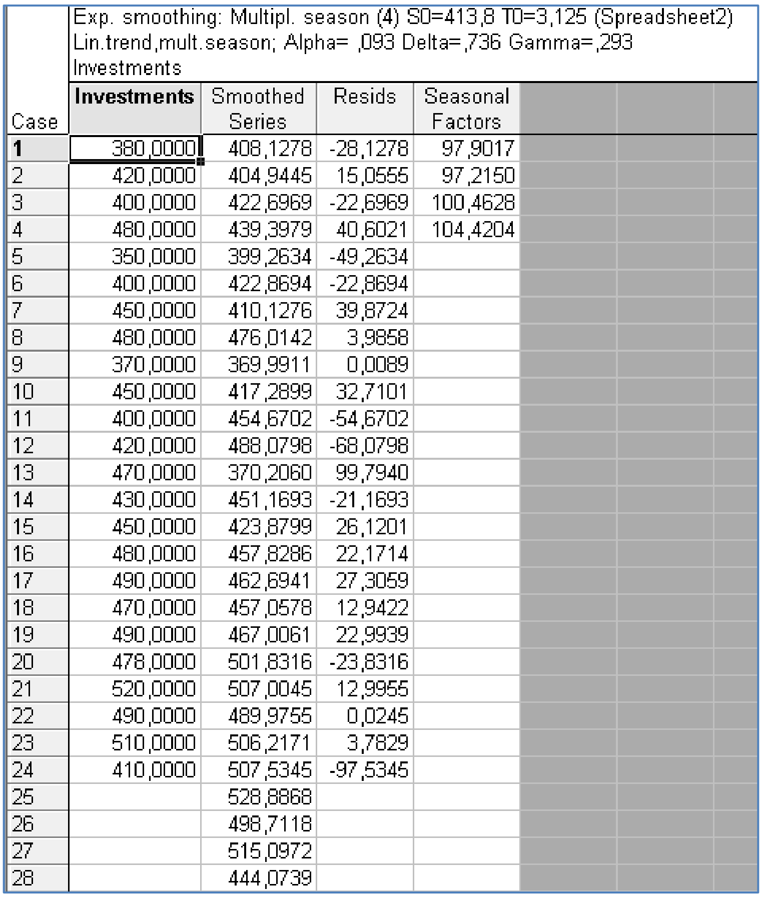


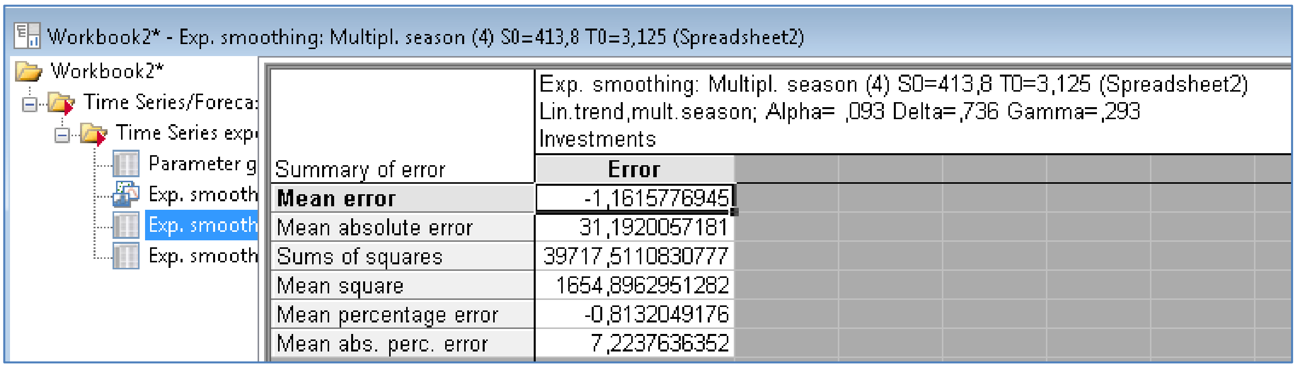
Fig. 43. **The graphs of exponential smoothing results**



Exp. smoothing: Multipl. season (4) S0=413,8 T0=3,125 (Spreadsheet2)
Lin.trend,mult.season; Alpha= ,093 Delta=,736 Gamma=,293
Investments

Case	Investments	Smoothed Series	Resids	Seasonal Factors
1	380,0000	408,1278	-28,1278	97,9017
2	420,0000	404,9445	15,0555	97,2150
3	400,0000	422,6969	-22,6969	100,4628
4	480,0000	439,3979	40,6021	104,4204
5	350,0000	399,2634	-49,2634	
6	400,0000	422,8694	-22,8694	
7	450,0000	410,1276	39,8724	
8	480,0000	476,0142	3,9858	
9	370,0000	369,9911	0,0089	
10	450,0000	417,2899	32,7101	
11	400,0000	454,6702	-54,6702	
12	420,0000	488,0798	-68,0798	
13	470,0000	370,2060	99,7940	
14	430,0000	451,1693	-21,1693	
15	450,0000	423,8799	26,1201	
16	480,0000	457,8286	22,1714	
17	490,0000	462,6941	27,3059	
18	470,0000	457,0578	12,9422	
19	490,0000	467,0061	22,9939	
20	478,0000	501,8316	-23,8316	
21	520,0000	507,0045	12,9955	
22	490,0000	489,9755	0,0245	
23	510,0000	506,2171	3,7829	
24	410,0000	507,5345	-97,5345	
25		528,8868		
26		498,7118		
27		515,0972		
28		444,0739		

Fig. 44. **The results of the forecast**



Exp. smoothing: Multipl. season (4) S0=413,8 T0=3,125 (Spreadsheet2)
Lin.trend,mult.season; Alpha= ,093 Delta=,736 Gamma=,293
Investments

Summary of error	Error
Mean error	-1,1615776945
Mean absolute error	31,1920057181
Sums of squares	39717,5110830777
Mean square	1654,8962951282
Mean percentage error	-0,8132049176
Mean abs. perc. error	7,2237636352

Fig. 45. **The table of quality estimates of the model of exponential smoothing**

Thus, according to the results of the forward forecast, the forecast values of the volume of investments in the USA IT sector for the next year have been obtained. The obtained values indicate an unstable trend of changes in investment between quarters. The quality of the prognosis is confirmed by low values of model errors (see Fig. 45).

Task 2. Build a model of the impact of socioeconomic indicators on the country's GDP (Table 11) and check the model for the presence of autocorrelation.

Table 11

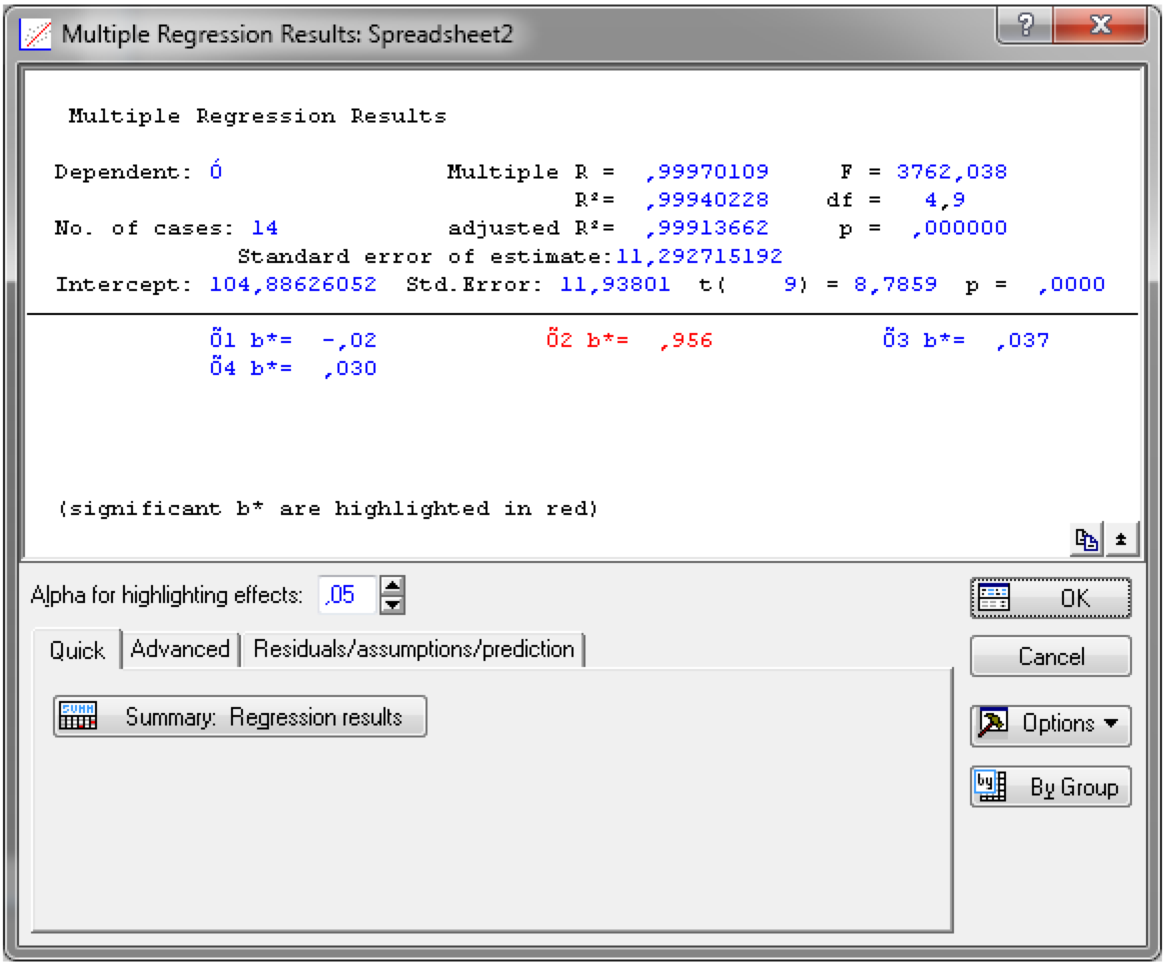
The initial data

Years	Production output, thousand UAH (X1)	Volume of retail trade turnover of enterprises, mln UAH (X2)	Total expenditures on average per month per household, UAH (X3)	Direct investment, million dollars (X4)	GDP, billion UAH (Y)
1998	226 358	19 317	395.6	2 063.6	186.5
1999	356 842	22 151	426.5	2 810.7	192.5
2000	373 893	28 757	541.3	3 281.8	198.9
2001	460 520	34 417	607	3 875	221.6
2002	504 008	39 691	658.3	4 555.3	225.8
2003	603 704	49 994	736.8	5 471.8	267.3
2004	809 988	67 556	903.5	6 794.4	345.1
2005	995 630	94 332	1 229.4	9 047	441.5
2006	1 182 179	129 952	1 442.8	16 890	544.2
2007	1 565 055	178 233	1 722	21 607.3	720.7
2008	2 072 172	246 903	2 590.4	29 542.7	948.1
2009	1 955 685	230 955	2 754.1	35 616.4	913.3
2010	2 388 289	280 890	3 072.7	40 053	1082.6
2011	2 496 365	350 059	3 456	44 806	1316.6

Guidelines

2.1. Construction of a multifactor econometric model.

The *Multiple Regression* module is provided in the Statistica package for the construction and comprehensive analysis of multiple linear econometric models. To start the calculation procedures, you must enter the menu item *Statistics / Multiple Regression*. In the start panel of this module, it is necessary to set variables for the analysis. The results of constructing a linear econometric model are presented in the dialog box (Fig. 46). At the top of the window, basic information about the model is contained, at the bottom, there are the function buttons that allow you to comprehensively view the results of the analysis.



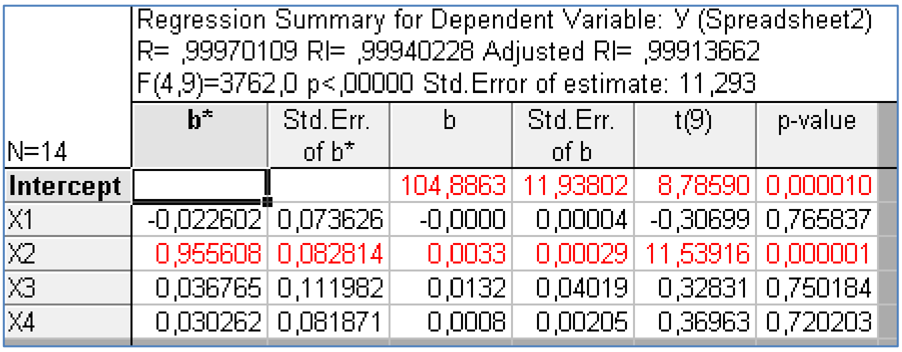
```
 Multiple Regression Results

Dependent: Ó                  Multiple R =  ,99970109     F = 3762,038
                                       R²=  ,99940228    df =    4,9
No. of cases: 14              adjusted R²=  ,99913662     p =  ,000000
             Standard error of estimate:11,292715192
Intercept: 104,88626052  Std.Error: 11,93801  t(    9) = 8,7859  p =  ,0000

            Õ1 b*=  -,02            Õ2 b*=  ,956            Õ3 b*=  ,037
            Õ4 b*=  ,030

(significant b* are highlighted in red)
```

Alpha for highlighting effects: ,05

Multiple Regression Results: Spreadsheet2
OK
Cancel
Quick
Advanced
Residuals/assumptions/prediction
Summary: Regression results
Options
By Group

Fig. 46. **The window of the results of regression analysis**

By initiating the *Summary: Regression results* button, we will determine the most significant characteristics of the model and the degree of its adequacy (Fig. 47).



Regression Summary for Dependent Variable: Y (Spreadsheet2)
R= ,99970109 RI= ,99940228 Adjusted RI= ,99913662
F(4,9)=3762,0 p<,00000 Std.Error of estimate: 11,293

N=14	b*	Std.Err. of b*	b	Std.Err. of b	t(9)	p-value
Intercept			104,8863	11,93802	8,78590	0,000010
X1	-0,022602	0,073626	-0,0000	0,00004	-0,30699	0,765837
X2	0,955608	0,082814	0,0033	0,00029	11,53916	0,000001
X3	0,036765	0,111982	0,0132	0,04019	0,32831	0,750184
X4	0,030262	0,081871	0,0008	0,00205	0,36963	0,720203

Fig. 47. **The results of regression analysis**

Based on the analysis of the obtained results, we note that this model is generally adequate and of high quality, but the parameters for X1, X3 and X4 are not significant.

2.2. Check the model for autocorrelation.

An important prerequisite for constructing a qualitative regression model using the least squares method is the independence of the values of random deviations ε_i from the values of deviations in all other observations. The absence of dependence guarantees the absence of correlation between any deviations ($\sigma(\varepsilon_i, \varepsilon_j) = \text{cov}\ (\varepsilon_i, \varepsilon_j = 0$ at $i \neq j$) and, in particular, between adjacent deviations ($\sigma(\varepsilon_{i-1}, \varepsilon_j) = 0$), $i = 2, 3, ..., n$.

Autocorrelation (sequential correlation) is defined as the correlation between the observed indicators, ordered in time (time series) or in space (cross-data). Autocorrelation of residues (deviations) is usually found in regression analysis using time series data. If cross-data are used, the presence of autocorrelation (spatial correlation) is extremely rare.

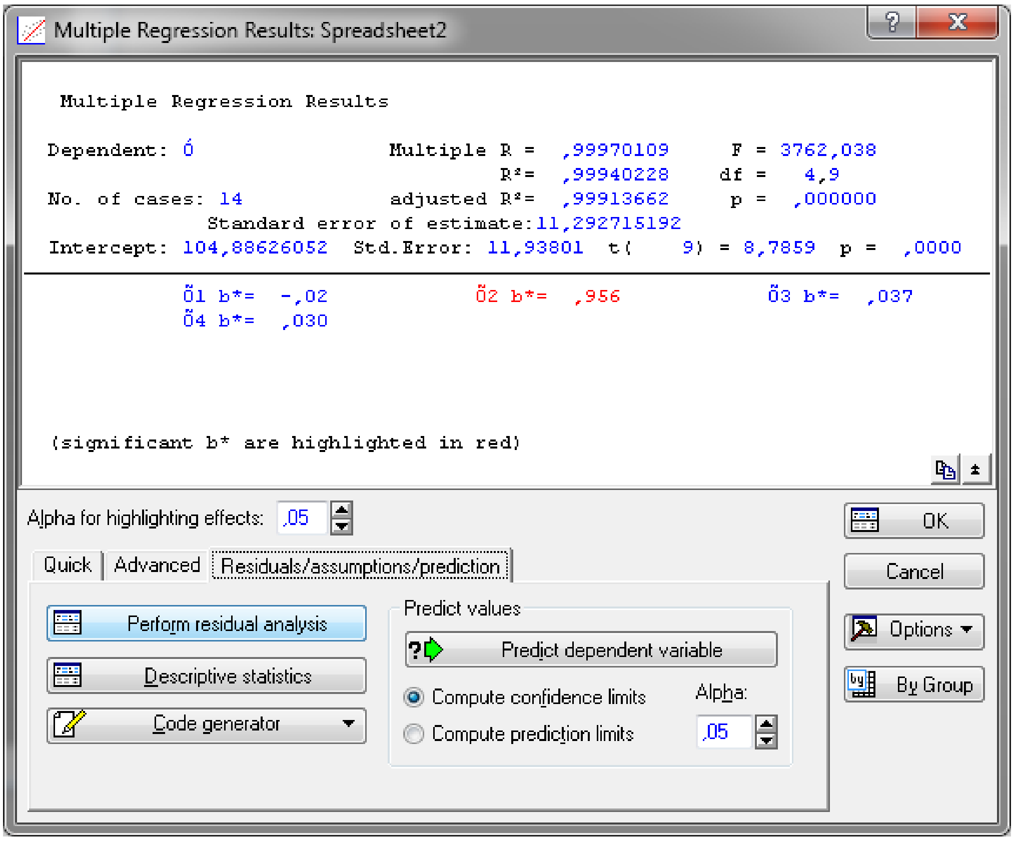
Among the main reasons that cause the appearance of autocorrelation, we can highlight specification errors; inertia (cyclicity); cobweb effect (presence of a time lag); data smoothing.

The presence of autocorrelation is determined using the following criteria:

- the Durbin – Watson criterion;
- the Von Neumann criteria;
- the non-cyclic autocorrelation criterion;
- the cyclic autocorrelation criterion.

2.3. Calculation of the Durbin – Watson criterion.

To determine the Durbin – Watson coefficient (the Durbin – Watson statistics), to assess the presence of autocorrelation in the model residues in the error analysis menu, initiate the *Perform residual analysis* button (Fig. 48).



Multiple Regression Results: Spreadsheet2

```
 Multiple Regression Results

Dependent: ó              Multiple R =  ,99970109     F = 3762,038
                                 R²=  ,99940228    df =    4,9
No. of cases: 14          adjusted R²=  ,99913662     p =  ,000000
             Standard error of estimate:11,292715192
Intercept: 104,88626052  Std.Error: 11,93801  t(    9) = 8,7859  p =  ,0000

           õ1 b*=  -,02            õ2 b*=  ,956            õ3 b*=  ,037
           õ4 b*=  ,030



(significant b* are highlighted in red)
```

Alpha for highlighting effects: ,05

Quick
Advanced
Residuals/assumptions/prediction
Perform residual analysis
Descriptive statistics
Code generator
Predict values
Predict dependent variable
OK
Cancel
Options
By Group

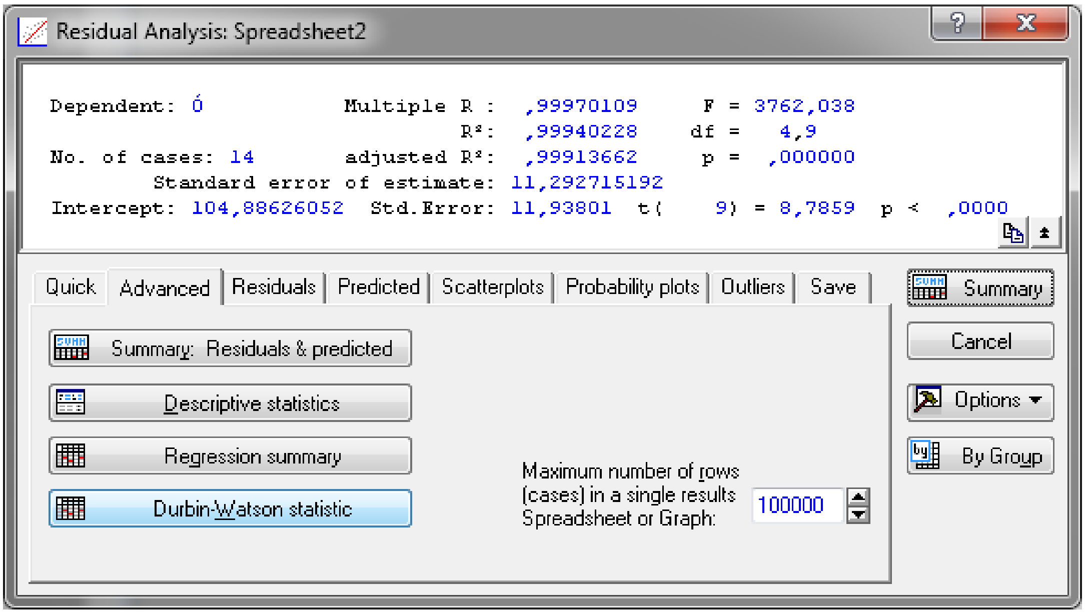
● Compute confidence limits

○ Compute prediction limits

Alpha: ,05

Fig. 48. **The error analysis menu**

In the error analysis menu, initiating the Durbin – Watson statistic button (Fig. 49), we get the value of autocorrelation of model errors by the Durbin – Watson criterion and the value of noncyclic autocorrelation coefficient (it expresses the degree of the interrelation of series).

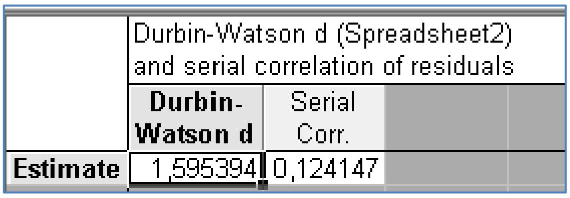


```
Dependent: ó           Multiple R :  ,99970109     F = 3762,038
                              R²:  ,99940228    df =    4,9
No. of cases: 14       adjusted R²:  ,99913662     p =  ,000000
        Standard error of estimate: 11,292715192
Intercept: 104,88626052  Std.Error: 11,93801  t(    9) = 8,7859  p <  ,0000
```

Maximum number of rows (cases) in a single results Spreadsheet or Graph: 100000

Fig. 49. **The error analysis module**

The result of the calculation of the Durbin – Watson coefficient is shown in Fig. 50.



Durbin-Watson d (Spreadsheet2)
and serial correlation of residuals

	Durbin-Watson d	Serial Corr.
Estimate	1,595394	0,124147

Fig. 50. **The result of the calculation of the Durbin – Watson test using the Statistica 10.0 package**

The critical values of the criterion for the number of observations n = 14, degrees of freedom (number of regressors) k = 4 and significance levels $\alpha = 0.05$ are equal to dl = 0.69, du = 1.97.

Conclusions are formed according to the following scheme:

1) if 0 < DW <dl, there is a positive autocorrelation of residues;

2) if dl ≤ DW ≤ du, a conclusion about the presence of autocorrelation is not determined (uncertainty zone);

3) if du < DW < 4 – du, there is no autocorrelation;

4) if 4 – du ≤ DW ≤ 4 – dl, a conclusion about the presence of autocorrelation is not determined;

5) if 4 – dl < DW < 4, there is a negative autocorrelation of residues.

That is 0.69 < 1.595 < 1.97 (dl < d < du), which indicates that the calculated value of the criterion is in the zone of uncertainty, and the use of one criterion is not enough.

The cyclic correlation coefficient (Serial Corr.) is 0.124. In fact, the calculated value of the cyclic autocorrelation coefficient is compared with the tabular value for the selected level of significance α and the length of the series n. If $| r | < | r_{5\ \%\ table} |$, we accept the hypothesis of non-autocorrelation of residues ε_t; if $| r | > | r_{1\ \%\ table} |$, we reject the hypothesis of their non-autocorrelation.

Therefore, it is difficult to draw an unambiguous conclusion about the presence of autocorrelation of the model residues.

2.4. Elimination of autocorrelation by the Aitken's method.

In case of autocorrelation of residues, the model parameters can be determined by the Aitken method.

Aitken's method differs from the usual least squares method in that the matrix of parameters is used to estimate the parameters of the model, which reflects the adjustment of the original data for the variability of the variance. The parameters of the model are determined by formulas 3 and 4:

$$a = \left(X^T \times \Omega^{-1} \times X\right)^{-1} X^T \times \Omega^{-1} \times Y; \tag{3}$$

$$\Omega^{-1} = \frac{1}{1-p^2}\begin{pmatrix} 1 & -p & 0 & 0 & 0 & \dots & 0 \\ -\rho & 1+\rho^2 & -\rho & 0 & 0 & \dots & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \tag{4}$$

In practice, to calculate ρ, the following relation is used to determine the cyclic correlation coefficient (formula 5):

$$\rho \approx r^0 \approx \frac{\sum\limits_{t=1}^{n} u_t \times u_{t-1}}{\sum\limits_{t=1}^{n} (u_t)^2}. \tag{5}$$

As can be seen from the calculations in Fig. 50, the value of the cyclic correlation coefficient is 0.124, so the parameter ρ = 0.124.

To eliminate autocorrelation in the model, it is proposed to use the Aitken method. Further calculations are performed in MS Excel, using built-in functions.

We present the following calculation algorithm:

1. Define the matrix Ω.

2. Calculate the inverse matrix Ω^{-1}.

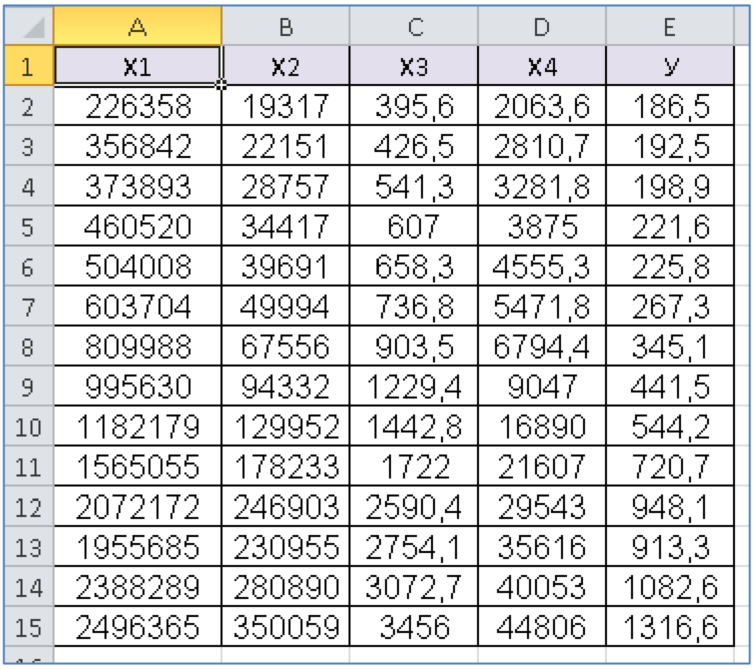
3. Multiply the matrix X^T by Ω^{-1}, where X^T is the matrix transposed to the matrix of independent variables X.

4. Find the product of $X^T\ \Omega^{-1}\ X$.

5. Calculate the inverse matrix $(X^T\Omega^{-1}X)^{-1}$ and the matrix $X^T\ \Omega^{-1}$ У.

6. Find the matrix A = $(X^T\ \Omega^{-1}\ X)^{-1}\ X^T\ \Omega^{-1}$ У whose elements will be the coefficients of the linear equation (Y is the vector of the dependent variable).

We begin calculations with the input of the initial data in a working sheet of the MS Excel package (Fig. 51).

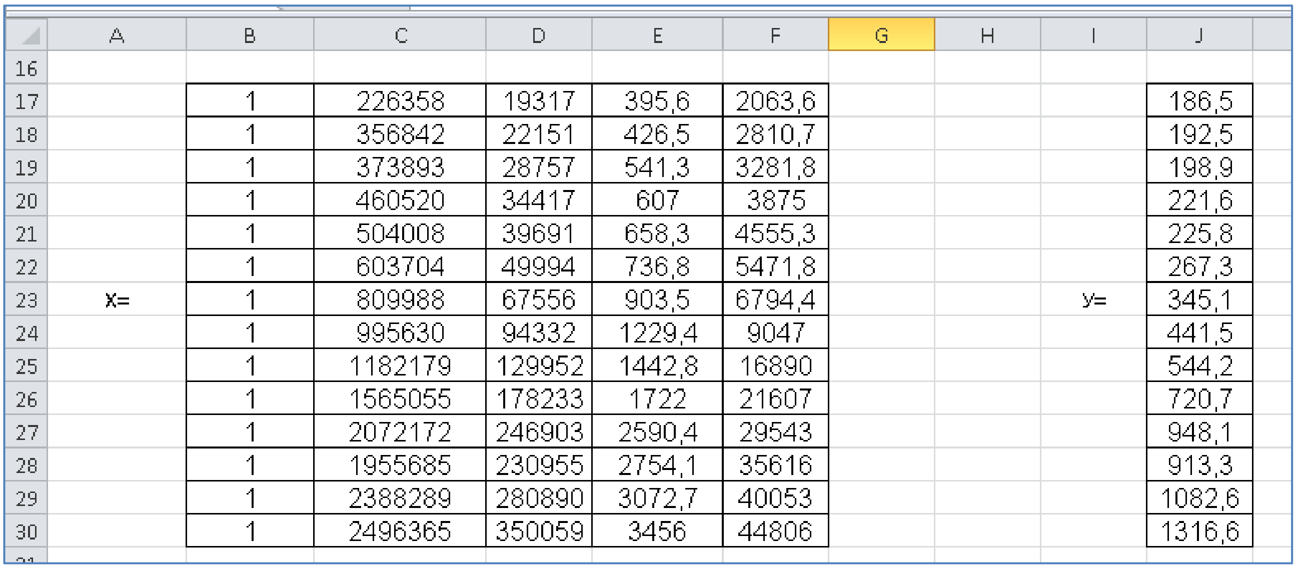


	A	B	C	D	E
1	X1	X2	X3	X4	y
2	226358	19317	395,6	2063,6	186,5
3	356842	22151	426,5	2810,7	192,5
4	373893	28757	541,3	3281,8	198,9
5	460520	34417	607	3875	221,6
6	504008	39691	658,3	4555,3	225,8
7	603704	49994	736,8	5471,8	267,3
8	809988	67556	903,5	6794,4	345,1
9	995630	94332	1229,4	9047	441,5
10	1182179	129952	1442,8	16890	544,2
11	1565055	178233	1722	21607	720,7
12	2072172	246903	2590,4	29543	948,1
13	1955685	230955	2754,1	35616	913,3
14	2388289	280890	3072,7	40053	1082,6
15	2496365	350059	3456	44806	1316,6

Fig. 51. **The initial data for building a multifactor model**

2.5. Carrying out intermediate calculations.

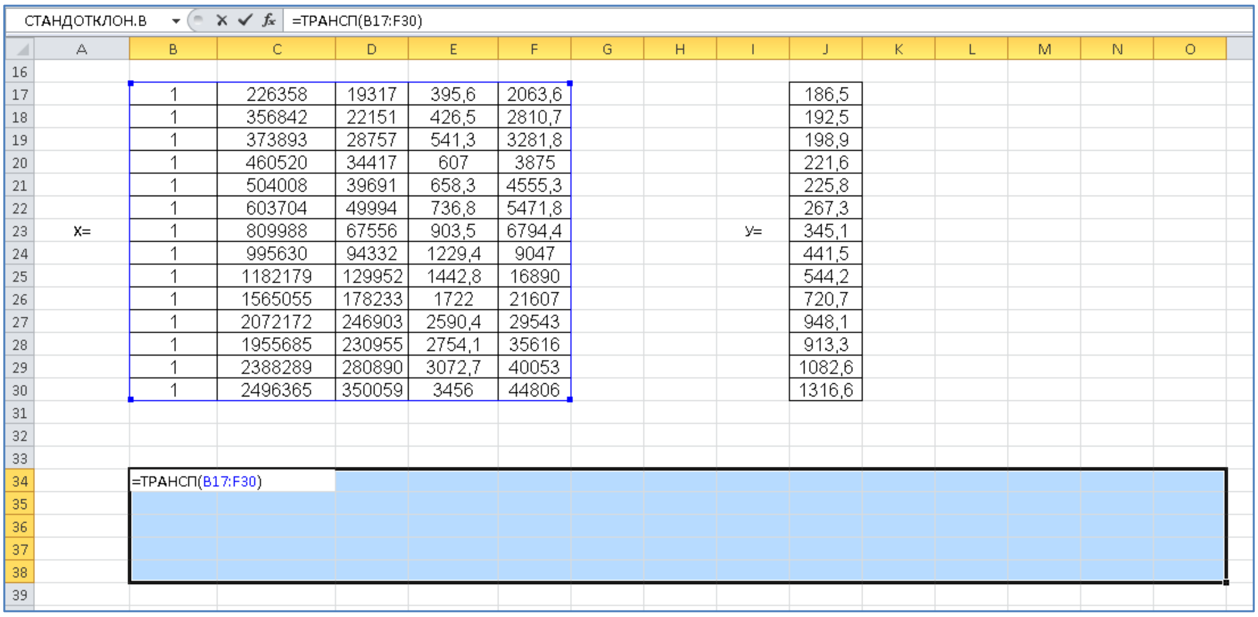
We have to construct a matrix of the independent variables X1 – X4 and the dependent variable Y (Fig. 52).



	A	B	C	D	E	F	G	H	I	J
16										
17		1	226358	19317	395,6	2063,6				186,5
18		1	356842	22151	426,5	2810,7				192,5
19		1	373893	28757	541,3	3281,8				198,9
20		1	460520	34417	607	3875				221,6
21		1	504008	39691	658,3	4555,3				225,8
22		1	603704	49994	736,8	5471,8				267,3
23	x=	1	809988	67556	903,5	6794,4			y=	345,1
24		1	995630	94332	1229,4	9047				441,5
25		1	1182179	129952	1442,8	16890				544,2
26		1	1565055	178233	1722	21607				720,7
27		1	2072172	246903	2590,4	29543				948,1
28		1	1955685	230955	2754,1	35616				913,3
29		1	2388289	280890	3072,7	40053				1082,6
30		1	2496365	350059	3456	44806				1316,6

Fig. 52. **Construction of matrixes**

Transposition in the MS Excel package is performed using the *ТРАНСП* (array) function (Fig. 53).

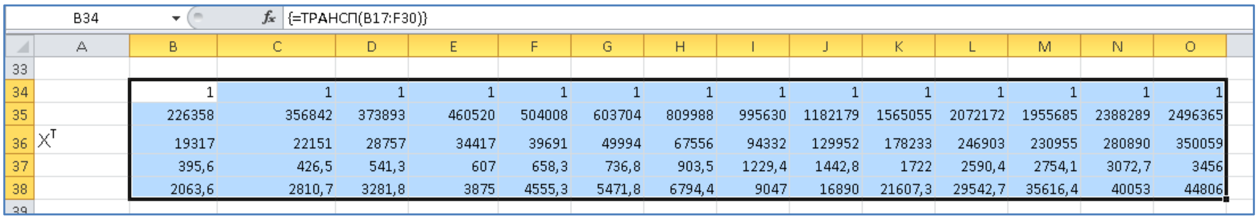


СТАНДОТКЛОН.В =ТРАНСП(B17:F30)

	A	B	C	D	E	F	G	H	I	J
16										
17		1	226358	19317	395,6	2063,6				186,5
18		1	356842	22151	426,5	2810,7				192,5
19		1	373893	28757	541,3	3281,8				198,9
20		1	460520	34417	607	3875				221,6
21		1	504008	39691	658,3	4555,3				225,8
22		1	603704	49994	736,8	5471,8				267,3
23	X=	1	809988	67556	903,5	6794,4			y=	345,1
24		1	995630	94332	1229,4	9047				441,5
25		1	1182179	129952	1442,8	16890				544,2
26		1	1565055	178233	1722	21607				720,7
27		1	2072172	246903	2590,4	29543				948,1
28		1	1955685	230955	2754,1	35616				913,3
29		1	2388289	280890	3072,7	40053				1082,6
30		1	2496365	350059	3456	44806				1316,6
31										
32										
33										
34		=ТРАНСП(B17:F30)								
35										
36										
37										
38										
39										

Fig. 53. **The transposition function window**

First, select the appropriate array of free cells and drive the formula, then press Ctrl + Shift + Enter (Fig. 54).

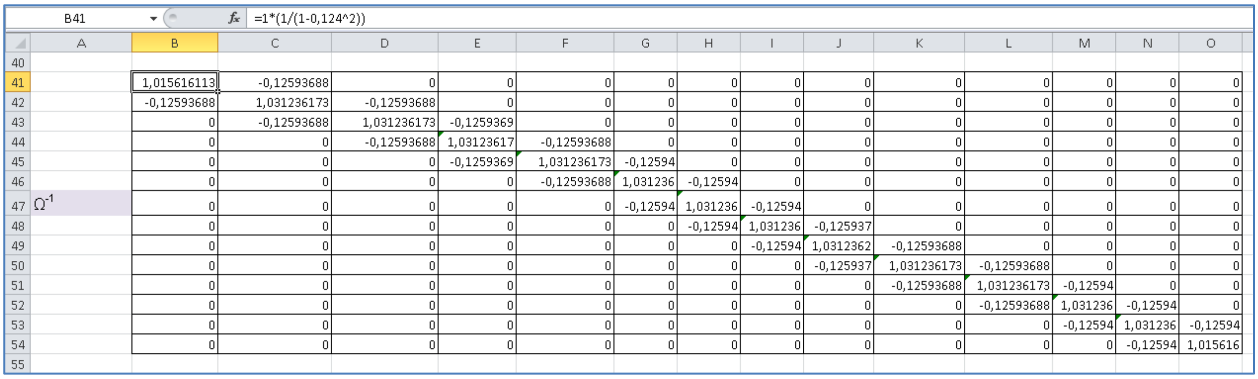


B34 {=ТРАНСП(B17:F30)}

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
33															
34		1	1	1	1	1	1	1	1	1	1	1	1	1	1
35		226358	356842	373893	460520	504008	603704	809988	995630	1182179	1565055	2072172	1955685	2388289	2496365
36	X^T	19317	22151	28757	34417	39691	49994	67556	94332	129952	178233	246903	230955	280890	350059
37		395,6	426,5	541,3	607	658,3	736,8	903,5	1229,4	1442,8	1722	2590,4	2754,1	3072,7	3456
38		2063,6	2810,7	3281,8	3875	4555,3	5471,8	6794,4	9047	16890	21607,3	29542,7	35616,4	40053	44806

Fig. 54. **The value of the matrix X^T**

Taking into account the value of the cyclic autocorrelation coefficient and formula 4, a matrix Ω – 1 with dimension 14 x 14 is formed (Fig. 55).

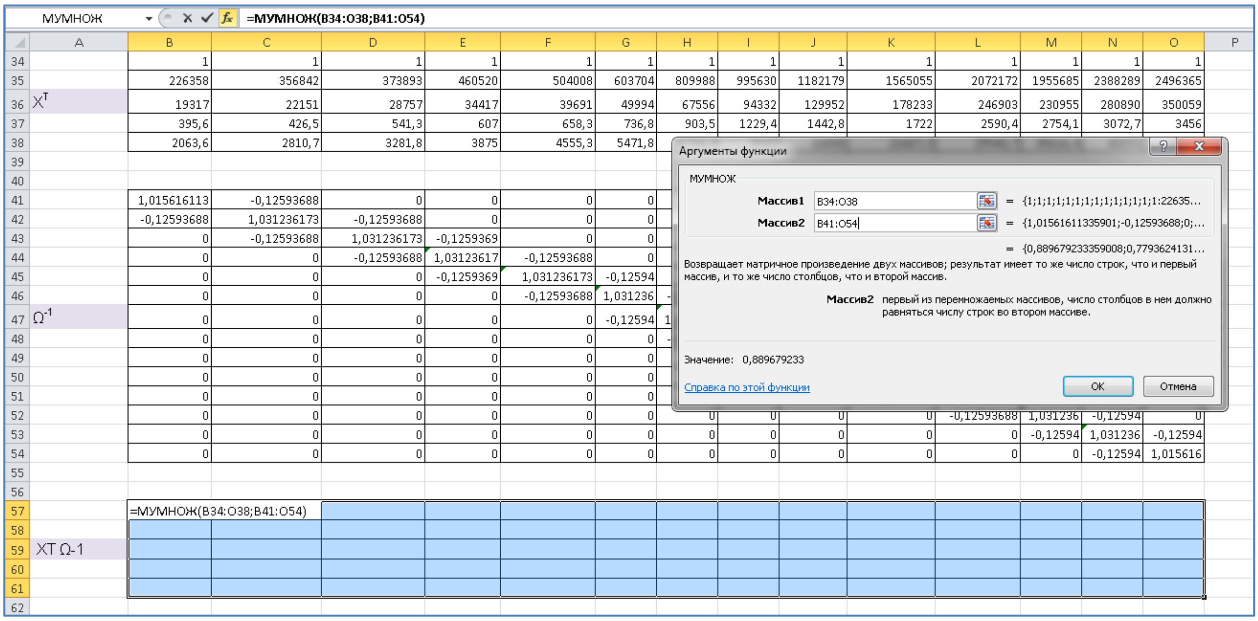


B41 =1*(1/(1-0,124^2))

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
40															
41		1,015616113	-0,12593688	0	0	0	0	0	0	0	0	0	0	0	0
42		-0,12593688	1,031236173	-0,12593688	0	0	0	0	0	0	0	0	0	0	0
43		0	-0,12593688	1,031236173	-0,1259369	0	0	0	0	0	0	0	0	0	0
44		0	0	-0,12593688	1,03123617	-0,12593688	0	0	0	0	0	0	0	0	0
45		0	0	0	-0,1259369	1,031236173	-0,12594	0	0	0	0	0	0	0	0
46		0	0	0	0	-0,12593688	1,031236	-0,12594	0	0	0	0	0	0	0
47	Ω^{-1}	0	0	0	0	0	-0,12594	1,031236	-0,12594	0	0	0	0	0	0
48		0	0	0	0	0	0	-0,12594	1,031236	-0,125937	0	0	0	0	0
49		0	0	0	0	0	0	0	-0,12594	1,0312362	-0,12593688	0	0	0	0
50		0	0	0	0	0	0	0	0	-0,125937	1,031236173	-0,12593688	0	0	0
51		0	0	0	0	0	0	0	0	0	-0,12593688	1,031236173	-0,12594	0	0
52		0	0	0	0	0	0	0	0	0	0	-0,12593688	1,031236	-0,12594	0
53		0	0	0	0	0	0	0	0	0	0	0	-0,12594	1,031236	-0,12594
54		0	0	0	0	0	0	0	0	0	0	0	0	-0,12594	1,015616
55															

Fig. 55. **The result of calculating the matrix Ω – 1**

Multiplication of two matrices in the MS Excel package is performed using the *МУМНОЖ* function (array 1; array 2) (Fig. 56, 57).

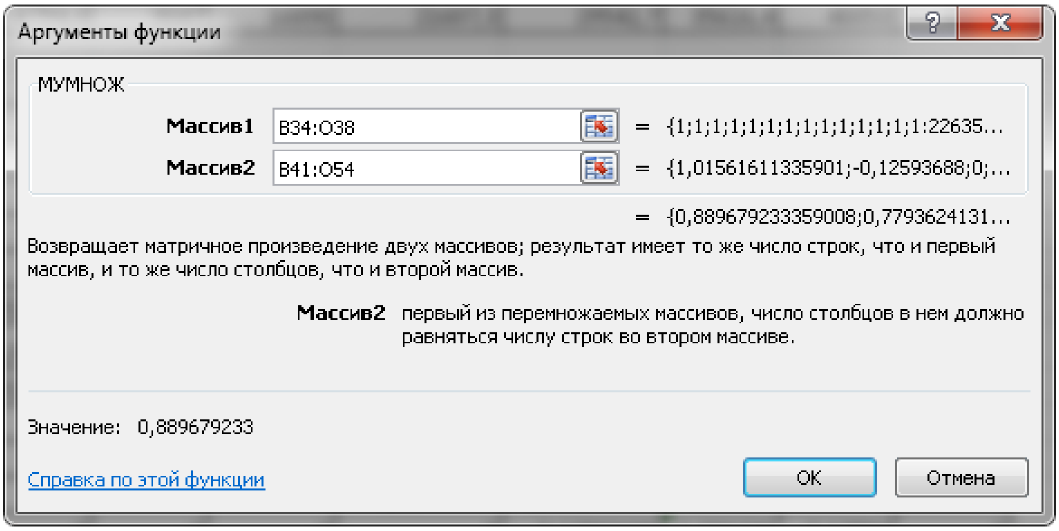


МУМНОЖ

=МУМНОЖ(B34:O38;B41:O54)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
34		1	1	1	1	1	1	1	1	1	1	1	1	1	1	
35		226358	356842	373893	460520	504008	603704	809988	995630	1182179	1565055	2072172	1955685	2388289	2496365	
36	X^T	19317	22151	28757	34417	39691	49994	67556	94332	129952	178233	246903	230955	280890	350059	
37		395,6	426,5	541,3	607	658,3	736,8	903,5	1229,4	1442,8	1722	2590,4	2754,1	3072,7	3456	
38		2063,6	2810,7	3281,8	3875	4555,3	5471,8									
39																
40																
41		1,015616113	-0,12593688	0	0	0	0									
42		-0,12593688	1,031236173	-0,12593688	0	0	0									
43		0	-0,12593688	1,031236173	-0,1259369	0	0									
44		0	0	-0,12593688	1,03123617	-0,12593688	0									
45		0	0	0	-0,1259369	1,031236173	-0,12594									
46		0	0	0	0	-0,12593688	1,031236	-								
47	Ω^{-1}	0	0	0	0	0	-0,12594	1								
48		0	0	0	0	0	0	-								
49		0	0	0	0	0	0									
50		0	0	0	0	0	0									
51		0	0	0	0	0	0									
52		0	0	0	0	0	0	0	0	0	0	-0,12593688	1,031236	-0,12594	0	
53		0	0	0	0	0	0	0	0	0	0	0	-0,12594	1,031236	-0,12594	
54		0	0	0	0	0	0	0	0	0	0	0	0	-0,12594	1,015616	
55																
56																
57		=МУМНОЖ(B34:O38;B41:O54)														
58																
59	XT Ω-1															
60																
61																
62																

Fig. 56. **The window for launching the function *МУМНОЖ***



Аргументы функции

МУМНОЖ

Массив1 B34:O38 = {1;1;1;1;1;1;1;1;1;1;1;1;1;1:22635...

Массив2 B41:O54 = {1,01561611335901;-0,12593688;0;...

= {0,889679233359008;0,7793624131...

Возвращает матричное произведение двух массивов; результат имеет то же число строк, что и первый массив, и то же число столбцов, что и второй массив.

Массив2 первый из перемножаемых массивов, число столбцов в нем должно равняться числу строк во втором массиве.

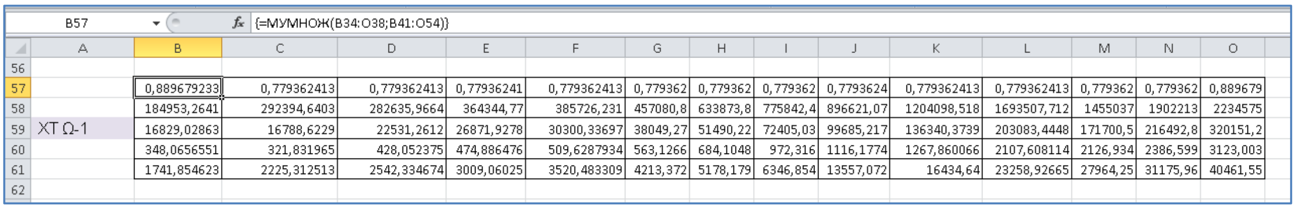
Значение: 0,889679233

Fig. 57. **The *МУМНОЖ* function implementation window**

Recall the rule of multiplication of matrices: two matrices can be multiplied if the number of rows of the second matrix is equal to the number of columns of the first matrix. When multiplying matrices, a matrix is obtained with the number of rows equal to the number of rows of the first matrix, and the number of columns equal to the number of columns of the second matrix. That is, in our case, when multiplying the X^T matrix of dimension of 5 x 14 by the matrix Ω – 1 of dimension of 14 x 14, we obtain a matrix of dimension of 5 x 14.

But first select the range of free cells, where the results of multiplication of two matrices will be placed, and at the same time press Ctrl + Shift + Enter.

As a result, we obtain a new matrix (Fig. 58).

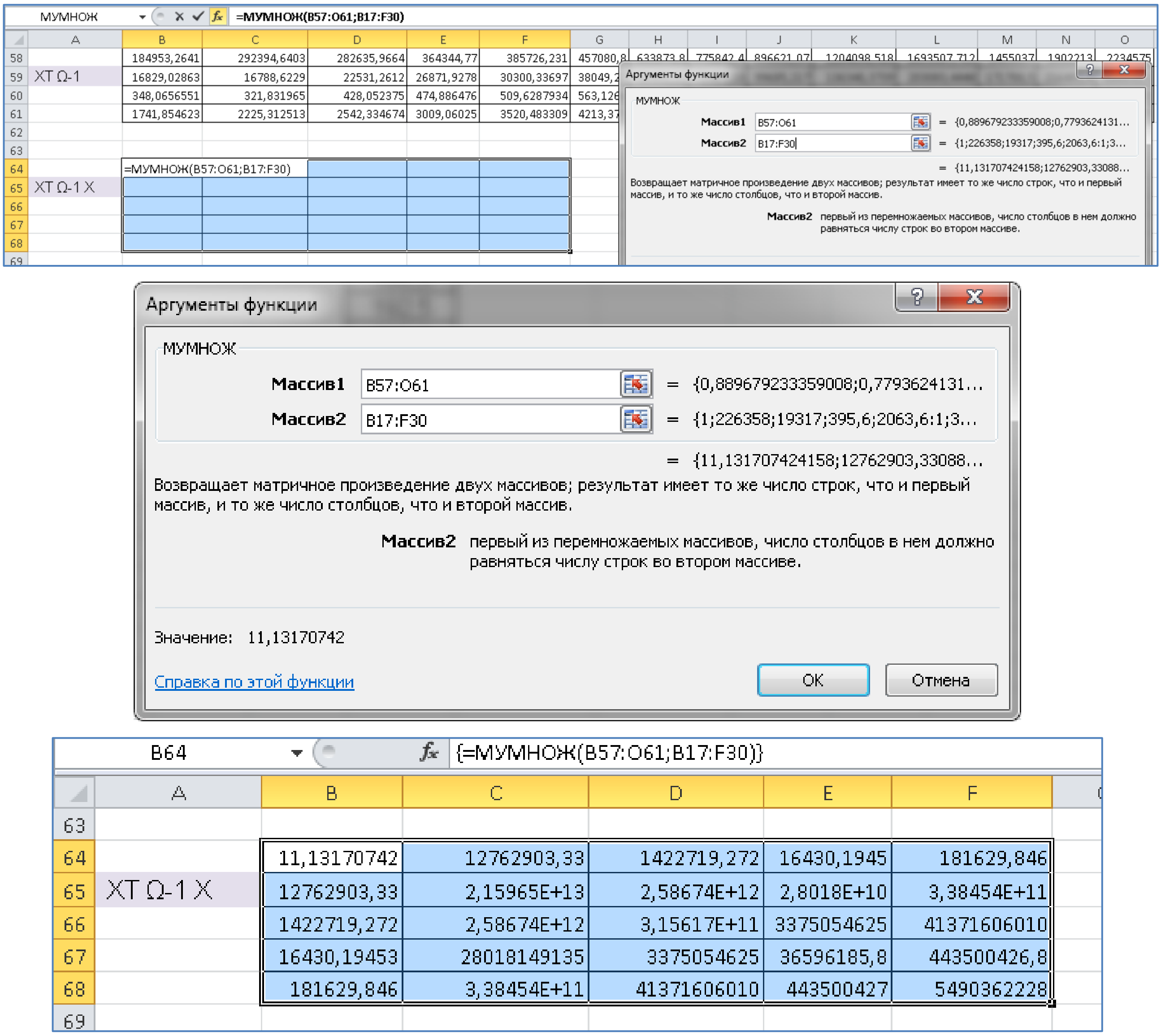


B57 | {=МУМНОЖ(B34:O38;B41:O54)}

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
56															
57		0,889679233	0,779362413	0,779362413	0,77936241	0,779362413	0,779362	0,779362	0,779362	0,7793624	0,779362413	0,779362413	0,779362	0,779362	0,889679
58		184953,2641	292394,6403	282635,9664	364344,77	385726,231	457080,8	633873,8	775842,4	896621,07	1204098,518	1693507,712	1455037	1902213	2234575
59	XT Ω-1	16829,02863	16788,6229	22531,2612	26871,9278	30300,33697	38049,27	51490,22	72405,03	99685,217	136340,3739	203083,4448	171700,5	216492,8	320151,2
60		348,0656551	321,831965	428,052375	474,886476	509,6287934	563,1266	684,1048	972,316	1116,1774	1267,860066	2107,608114	2126,934	2386,599	3123,003
61		1741,854623	2225,312513	2542,334674	3009,06025	3520,483309	4213,372	5178,179	6346,854	13557,072	16434,64	23258,92665	27964,25	31175,96	40461,55
62															

Fig. 58. **The results of using the *МУМНОЖ* function**

At the next stage, we find the product of the matrices X^T $\Omega^{-1}$and X using the function *МУМНОЖ* and get a matrix of dimension 5 x 5 (Fig. 59).



МУМНОЖ | =МУМНОЖ(B57:O61;B17:F30)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
58		184953,2641	292394,6403	282635,9664	364344,77	385726,231	457080,8	633873,8	775842,4	896621,07	1204098,518	1693507,712	1455037	1902213	2234575
59	XT Ω-1	16829,02863	16788,6229	22531,2612	26871,9278	30300,33697	38049,2								
60		348,0656551	321,831965	428,052375	474,886476	509,6287934	563,126								
61		1741,854623	2225,312513	2542,334674	3009,06025	3520,483309	4213,37								
62															
63															
64		=МУМНОЖ(B57:O61;B17:F30)													
65	XT Ω-1 X														
66															
67															
68															

Аргументы функции

МУМНОЖ

Массив1 B57:O61 = {0,889679233359008;0,7793624131…

Массив2 B17:F30 = {1;226358;19317;395,6;2063,6:1;3…

= {11,131707424158;12762903,33088…

Возвращает матричное произведение двух массивов; результат имеет то же число строк, что и первый массив, и то же число столбцов, что и второй массив.

Массив2 первый из перемножаемых массивов, число столбцов в нем должно равняться числу строк во втором массиве.

Аргументы функции

МУМНОЖ

Массив1 B57:O61 = {0,889679233359008;0,7793624131…

Массив2 B17:F30 = {1;226358;19317;395,6;2063,6:1;3…

= {11,131707424158;12762903,33088…

Возвращает матричное произведение двух массивов; результат имеет то же число строк, что и первый массив, и то же число столбцов, что и второй массив.

Массив2 первый из перемножаемых массивов, число столбцов в нем должно равняться числу строк во втором массиве.

Значение: 11,13170742

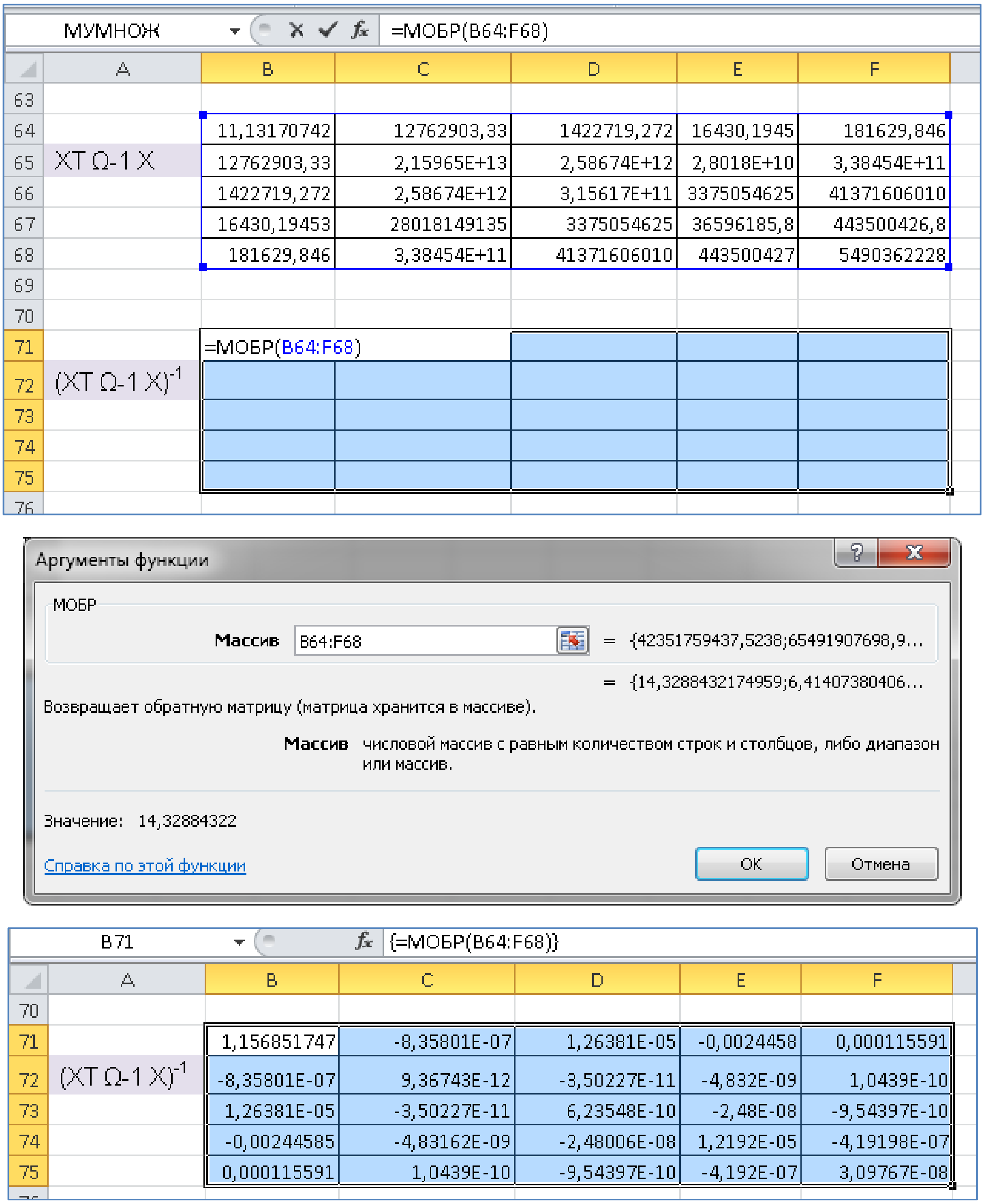
Справка по этой функции

B64 | {=МУМНОЖ(B57:O61;B17:F30)}

	A	B	C	D	E	F
63						
64		11,13170742	12762903,33	1422719,272	16430,1945	181629,846
65	XT Ω-1 X	12762903,33	2,15965E+13	2,58674E+12	2,8018E+10	3,38454E+11
66		1422719,272	2,58674E+12	3,15617E+11	3375054625	41371606010
67		16430,19453	28018149135	3375054625	36596185,8	443500426,8
68		181629,846	3,38454E+11	41371606010	443500427	5490362228
69						

Fig. 59. **The results of using the *МУМНОЖ* function**

To obtain the given inverse matrix, the function *MOBP* (array) is used (Fig. 60). The dimension of the inverse matrix will also be 5 x 5. But first select the range of free cells where the results of the inverse matrices will be placed, and simultaneously press Ctrl + Shift + Enter.



МУМНОЖ =МОБР(B64:F68)

	A	B	C	D	E	F
63						
64		11,13170742	12762903,33	1422719,272	16430,1945	181629,846
65	XT Ω-1 X	12762903,33	2,15965E+13	2,58674E+12	2,8018E+10	3,38454E+11
66		1422719,272	2,58674E+12	3,15617E+11	3375054625	41371606010
67		16430,19453	28018149135	3375054625	36596185,8	443500426,8
68		181629,846	3,38454E+11	41371606010	443500427	5490362228
69						
70						
71		=МОБР(B64:F68)				
72	(XT Ω-1 X)$^{-1}$					
73						
74						
75						

Аргументы функции

МОБР

Массив B64:F68 = {42351759437,5238;65491907698,9...

= {14,3288432174959;6,41407380406...

Возвращает обратную матрицу (матрица хранится в массиве).

Массив числовой массив с равным количеством строк и столбцов, либо диапазон или массив.

Значение: 14,32884322

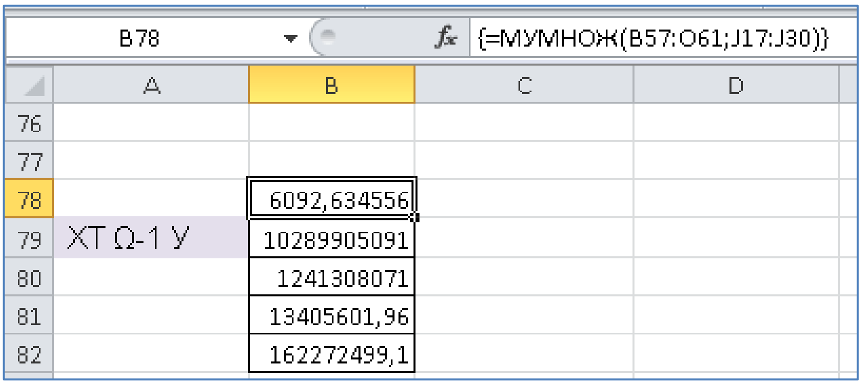
Справка по этой функции

B71 {=МОБР(B64:F68)}

	A	B	C	D	E	F
70						
71		1,156851747	-8,35801E-07	1,26381E-05	-0,0024458	0,000115591
72	(XT Ω-1 X)$^{-1}$	-8,35801E-07	9,36743E-12	-3,50227E-11	-4,832E-09	1,0439E-10
73		1,26381E-05	-3,50227E-11	6,23548E-10	-2,48E-08	-9,54397E-10
74		-0,00244585	-4,83162E-09	-2,48006E-08	1,2192E-05	-4,19198E-07
75		0,000115591	1,0439E-10	-9,54397E-10	-4,192E-07	3,09767E-08

Fig. 60. **The window for using the *MOBP* function**

According to the appropriate procedure, we determine the matrix $X^T\ \Omega^{-1}$ on the vector У (Fig. 61).

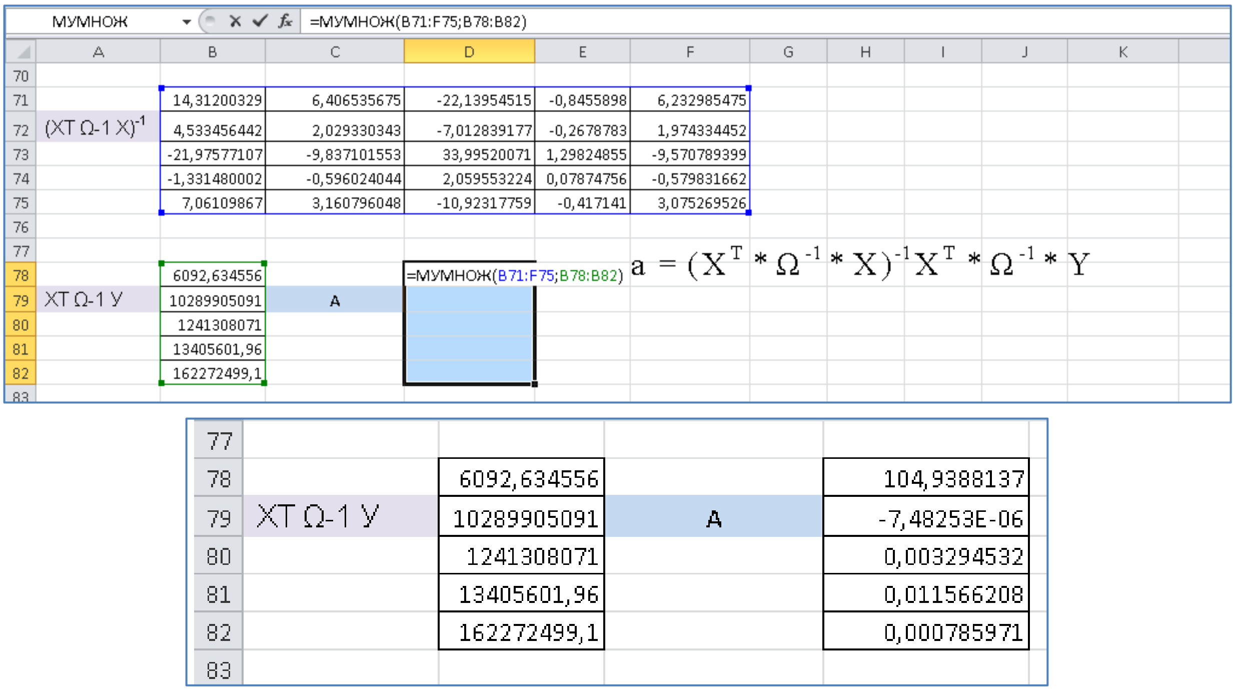


B78 {=МУМНОЖ(B57:O61;J17:J30)}

	A	B	C	D
76				
77				
78		6092,634556		
79	XT Ω-1 У	10289905091		
80		1241308071		
81		13405601,96		
82		162272499,1		

Fig. 61. **The matrix $X^T\ \Omega^{-1}$ У**

Determine the parameters of the model. To do this, use the above functions and perform calculations in accordance with formula 3. The results of the calculations are shown in Fig. 62.



МУМНОЖ =МУМНОЖ(B71:F75;B78:B82)

	A	B	C	D	E	F
70						
71		14,31200329	6,406535675	-22,13954515	-0,8455898	6,232985475
72	$(XT\ \Omega\text{-}1\ X)^{-1}$	4,533456442	2,029330343	-7,012839177	-0,2678783	1,974334452
73		-21,97577107	-9,837101553	33,99520071	1,29824855	-9,570789399
74		-1,331480002	-0,596024044	2,059553224	0,07874756	-0,579831662
75		7,06109867	3,160796048	-10,92317759	-0,417141	3,075269526
76						
77						
78		6092,634556		=МУМНОЖ(B71:F75;B78:B82)		
79	XT Ω-1 У	10289905091	A			
80		1241308071				
81		13405601,96				
82		162272499,1				

$$a = (X^T * \Omega^{-1} * X)^{-1} X^T * \Omega^{-1} * Y$$

77				
78		6092,634556		104,9388137
79	XT Ω-1 У	10289905091	A	-7,48253E-06
80		1241308071		0,003294532
81		13405601,96		0,011566208
82		162272499,1		0,000785971
83				

Fig. 62. **Calculation of parameters of a multifactor regression model by the Aitken method**

2.6. Formation of a general form of a multifactor regression model.

The general view of the model is as follows:

Y = 104.945 – 7.48X1 + 0.0033X2 + 0.0116X3 + 0.0008X4.

2.7. Checking the obtained model for the presence of an auto-residual relation.

In order to verify the elimination of autocorrelation of residues in the model, it is advisable to calculate the Durbin – Watson coefficient as:

$$DW = \frac{\sum_{t=2}^{T} (u_t - u_{t-1})^2}{\sum_{t=1}^{T} u_t^2}. \qquad (6)$$

where u_t is regression residues (the difference between the theoretical and predicted values of the indicator);

u_{t-1} is the value of the previous level of regression residues.

Find the calculated values based on the constructed econometric model and determine the residuals (Table 12).

Table 12

Calculation of model residues

Years	У	Уt	Ut	Ut^2	(Ut - Ut-1)^2
1998	186.5	176.4705435	10.02945653	100.59	
1999	192.5	187.7281931	4.77180687	22.77014	27.64287995
2000	198.9	211.3175286	–12.41752856	154.195	295.4732527
2001	221.6	231.8389076	–10.23890755	104.8352	4.746389515
2002	225.8	250.6677129	–24.86771294	618.4031	214.0019471
2003	267.3	286.9855456	–19.68554563	387.5207	26.85485801
2004	345.1	349.3552574	–4.255257357	18.10722	238.0937963
2005	441.5	444.4986271	–2.998627063	8.991764	1.579119697
2006	544.2	571.8783215	–27.67832146	766.0895	609.0873157
2007	720.7	740.7434552	–20.04345516	401.7401	58.29118354
2008	948.1	987.0545852	–38.95458522	1517.46	357.6308403
2009	913.3	940.3089128	–27.00891282	729.4814	142.699089
2010	1082.6	1115.23038	–32.63038033	1064.742	31.60089694
2011	1316.6	1352.087602	–35.487602	1259.37	8.163715681
			Sum	7154.295	2015.865284

$$DW = \frac{2\ 015.8652842}{7\ 154.295} = 0.282.$$

The critical values of the criterion are equal to dl = 0.69, du = 1.97.

Assuming that du < DW < 4 – du, we see that this indicates the absence of autocorrelation of the model. In our case: 0.69 > 0.28 < 4 – 1.97. That is 0 < DW < dl, so there is a positive autocorrelation of residues.

Thus, the model constructed using the Aitken method has autocorrelations in the residuals. This is one of the proofs of the insufficient quality of the model.

There are several ways to eliminate or reduce autocorrelation in time series. The most effective one is to exclude the trend from the numerical series and move to a random component. The following methods are used to eliminate autocorrelation:

1) the method based on the inclusion of time in the multiple regression equation as an argument – the Frisch – Waugh method;

2) the finite difference method, where the least squares method processes the levels of the original series and their successive differences between them;

3) the method of deviations of empirical values from those aligned with the trend;

4) the Cochrane – Orcutt method;

5) the Hildreth – Lu method.

Conclusion: to eliminate the autocorrelation in the model residues, the calculations of the model parameters have been performed using the Aitken method. According to the obtained results, in further calculations, it is necessary to use one of the methods of elimination of autocorrelation in the model residues.

Task 3. Construct a model of the influence of indicators of development of Ukraine's regions on the level of foreign investments (Table 13) and check the model for the presence of autocorrelation.

Table 13

The indicators of development of the regions of Ukraine

Regions of Ukraine	X1 Share of enterprises of collective ownership (in % to the total number in the region)	X2 Industry index (%)	X3 Labor productivity index (%)	Y Foreign investments (million US dollars)
1	2	3	4	5
Autonomous Republic of Crimea	0.220	0.180	0.170	0.200
Vinnytsia	0.820	0.540	0.450	0.530
Volyn	0.910	0.460	0.470	0.510

Table 13 (the end)

1	2	3	4	5
Dnipro	0.960	0.870	0.770	0.850
Donetsk	3.490	4.490	4.810	4.050
Zhytomyr	0.310	0.840	1.210	0.960
Transcarpathia	0.550	2.360	2.230	1.570
Zaporizhzhia	0.130	0.450	0.690	0.880
Ivano-Frankivsk	0.770	0.810	0.830	0.800
Kyiv	0.250	0.460	0.550	0.510
Kropyvnytskyi	0.500	1.080	1.040	0.800
Luhansk	0.120	0.390	0.530	0.750
Lviv	0.220	0.180	0.170	0.200
Mykolaiv	0.820	0.540	0.450	0.560
Odesa	0.910	0.460	0.470	0.570
Poltava	0.960	0.870	0.770	0.850
Rivne	0.160	0.710	0.880	0.860
Sumy	0.110	0.070	0.270	0.140
Ternopil	0.020	0.140	0.250	0.220
Kharkiv	0.010	0.030	0.070	0.090
Kherson	0.043	0.136	0.154	0.175
Khmelnytskyi	0.969	0.073	0.182	0.201
Cherkasy	0.172	0.105	0.178	0.289
Chernivtsi	0.108	0.150	0.181	0.352
Chernihiv	0.064	0.189	0.354	0.979

Task 4. Using the input data from Table 14, build a forecast for 1 year ahead (4 periods) using the module "Analysis of the time series module" of the program Statistica 10.0.

Table 14

The input data

Period	Air traffic load (%)	Period	Air traffic load (%)	Period	Air traffic load (%)
2014	78	2017	81	2018	77
	74		75		77
	77		74		77
	76		75		78
2015	75	2016	77	2019	78
	78		75		78
	74		75		78
	76		79		79

Task 5. Using the statistical data of the website of the State Statistics Service of Ukraine [19], find information about the tourist activity in Ukraine (on a yearly basis) and build a forecast for one year ahead using the models of adaptive forecasting.

Task 6. Using your own information space of research, build a multifactor regression model and check it for autocorrelation.

Task 7. Using the statistical data of the website of the State Statistics Service of Ukraine [19], find information about production and sale of industrial products according to the type (on a yearly basis) and build a forecast for 2 years ahead using the module "Analysis of the time series module" of the program Statistica 10.0.

The list of questions for independent work

1. What is the essence of smoothing methods?
2. What are the iterative smoothing methods?
3. What is the difference between exponential and simple smoothing?
4. What are the possibilities of using the methods of Brown, Holt, Winters?
5. What is the autoregression model used for?
6. What is the essence of the Granger test?
7. What is the classification of the vector autoregression models?
8. What are the ways of combining time series components into a model?
9. What is pulsed spectral analysis used for?
10. What is the difference between ARMA and ARIMA models?

Content module 2. Modeling and forecasting of multidimensional processes

Topic 5. Factor analysis of data

Task 1. In order to conduct a detailed analysis of the demographic situation in a country, it is necessary to reduce the information space (Table 15) using the methods of factor analysis.

Table 15

The input data

Years	X1 Employment to population ratio	X2 Unemployment (% of total labor force)	X3 Population	X4 Rate of natural increase	X5 GDP per capita	X6 Wage and salary
2005	58.26	7.26	8 391 850	89 939	1 578.40	123.6
2006	58.72	6.62	8 484 550	96 698	2 473.08	154.81
2007	58.91	6.33	8 581 300	98 308	3 851.44	199.67
2008	59.4	5.86	8 763 400	99 376	5 574.60	246.34
2009	59.83	5.74	8 947 243	99 625	4 950.29	291.29
2010	60.17	5.63	9 054 332	112 063	5 842.81	331.5
2011	60.52	5.42	9 173 082	122 310	7 189.69	351.86
2012	60.94	5.19	9 295 784	119 452	7 496.29	389.94
2013	61.36	4.97	9 416 801	118 288	7 875.76	418.25
2014	61.86	4.91	9 535 079	114 855	7 891.31	444.5
2015	62.27	4.96	9 649 341	111 513	5 500.32	466.9
2016	62.95	5	9 757 812	102 816	3 880.74	499.8
2017	63.26	5	9 854033	86 932	4 147.09	528.5
2018	63.56	4.9	9 939 800	81 732	4 722.38	544.6

Guidelines

The *Factor Analysis* module contains a wide range of methods for selection of factors, thus reducing the input information space.

Let's consider the main stages of conducting factor analysis in the system (package) Statistica 10.0 using the following example (see Table 15).

To reduce the initial information space we need the *Factor Analysis* module (*Statistics* / *Multivariate Exploratory Techniques / Factor Analysis* or

Multidimensional Methods / Factor Analysis). The *Factor Analysis* dialog box is shown in Fig. 63.

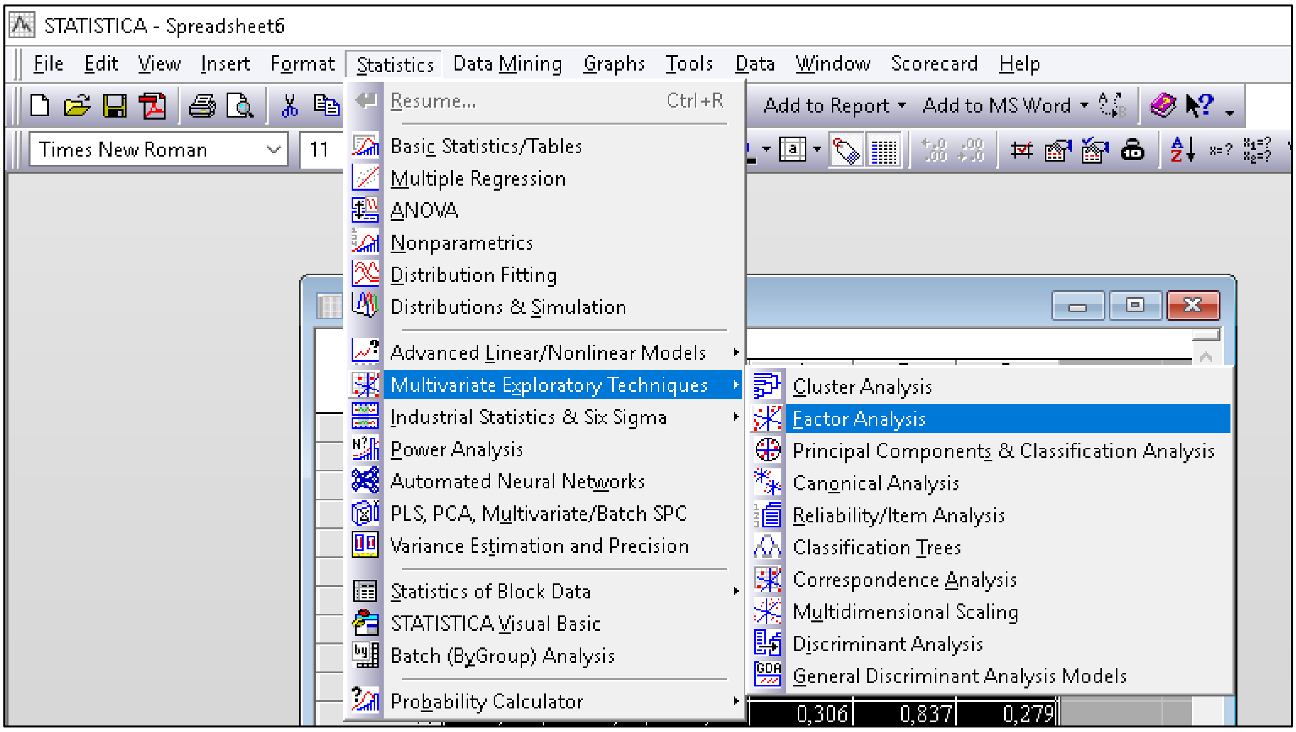
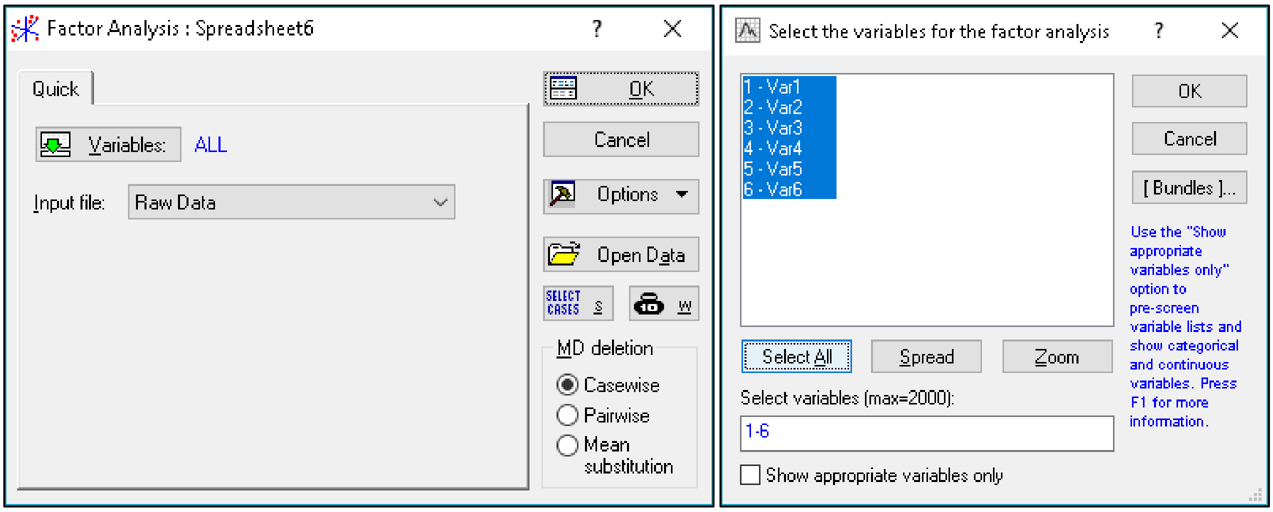

STATISTICA - Spreadsheet6
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Scorecard Help
Add to Report Add to MS Word
Times New Roman 11
Resume... Ctrl+R
Basic Statistics/Tables
Multiple Regression
ANOVA
Nonparametrics
Distribution Fitting
Distributions & Simulation
Advanced Linear/Nonlinear Models
Multivariate Exploratory Techniques
Industrial Statistics & Six Sigma
Power Analysis
Automated Neural Networks
PLS, PCA, Multivariate/Batch SPC
Variance Estimation and Precision
Statistics of Block Data
STATISTICA Visual Basic
Batch (ByGroup) Analysis
Probability Calculator
Cluster Analysis
Factor Analysis
Principal Components & Classification Analysis
Canonical Analysis
Reliability/Item Analysis
Classification Trees
Correspondence Analysis
Multidimensional Scaling
Discriminant Analysis
General Discriminant Analysis Models
0,306 0,837 0,279

Fig. 63. **The *Factor Analysis* module**

The *Variables* button allows you to select all variables from the data file that must be included in the factor analysis. If all variables are used for analysis, you can use the *Select All* button (Fig. 64).



Variables: ALL

Input file: Raw Data

MD deletion

● Casewise

○ Pairwise

○ Mean substitution

1 - Var1
2 - Var2
3 - Var3
4 - Var4
5 - Var5
6 - Var6

Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

Select variables (max=2000): 1-6

□ Show appropriate variables only

Fig. 64. **Choosing the variables**

The module includes the following output types: *Correlation Matrix* and *Raw Data*. Let's choose *Raw Data*. This is a regular data file, where the values of the variables are written in rows:

MD deletion (replace missed variables). A method for processing missed values;

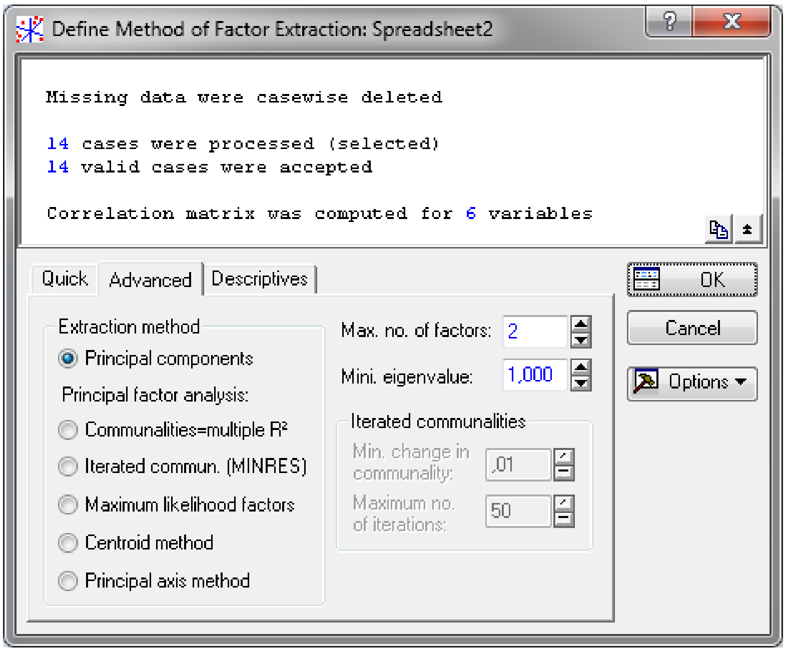
Casewise (a way to exclude missed cases) is a method where in the spreadsheet containing the data, all the rows (cases) that have at least one missing value are ignored. This applies to all variables. In the table, there are only cases in which there is no skip;

Pairwise (a duplicate way to exclude missed values). Missed cases are ignored for all variables, but only for the selected pair. All cases in which there are no spaces are used in processing, for example, with elemental calculation of the correlation matrix, when all pairs of variables are sequentially considered. Obviously, the Pairwise method has more observations for processing than the Casewise method;

Mean Substitution (substitution of the average for the missed values).

By clicking on the OK button in the startup window of the module, the analysis of the selected variables begins. The Statistica system will process the missed values in the way indicated, calculate the correlation matrix and offer a choice of several methods for factor analysis. The calculation of the correlation matrix (if not specified immediately) is the first stage of factor analysis. After clicking the Ok button, you can go to the next dialog.

The window *Define Method of Factor Extraction* is presented in Fig. 65.



Define Method of Factor Extraction: Spreadsheet2

```
Missing data were casewise deleted

14 cases were processed (selected)
14 valid cases were accepted

Correlation matrix was computed for 6 variables
```

Extraction method
● Principal components
Principal factor analysis:
○ Communalities=multiple R²
○ Iterated commun. (MINRES)
○ Maximum likelihood factors
○ Centroid method
○ Principal axis method

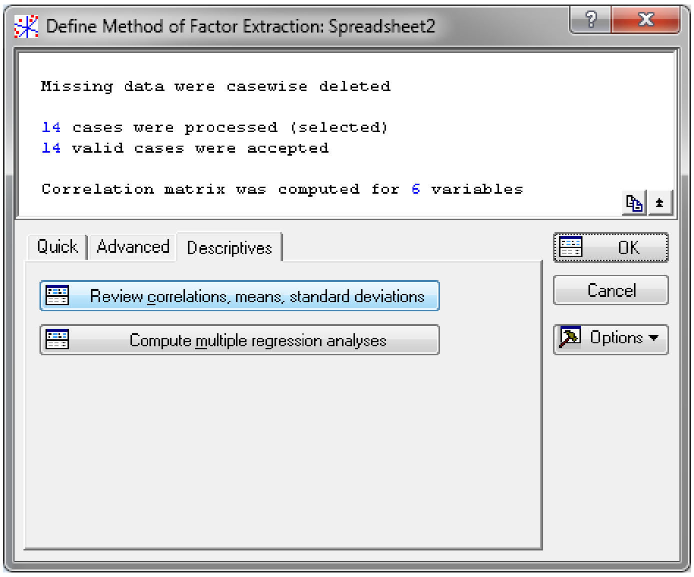
Max. no. of factors: 2
Mini. eigenvalue: 1,000

Iterated communalities
Min. change in communality: ,01
Maximum no. of iterations: 50

Fig. 65. **The *Define Method of Factor Extraction* window**

This window has the following structure. The upper part of the window is informational: it is reported here that the missing values are processed by the Casewise method. 14 cases were treated and 14 cases were taken for further calculations. The correlation matrix is calculated for 6 variables. The group of options merged under the heading *Extraction method* allows you to choose a method of processing.

To continue the analysis in the *Define Method of Factor Extraction* window (Fig. 66), we need to click on the *Review correlations*, *means, standard deviations* button (View correlations / average / standard deviations).

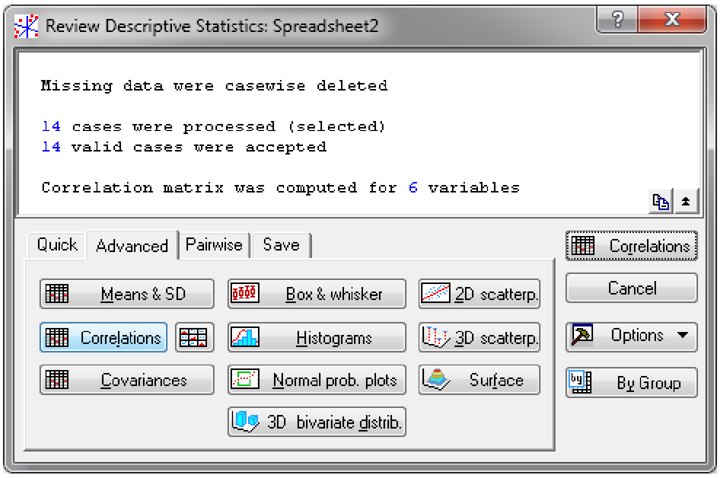


```
Missing data were casewise deleted

14 cases were processed (selected)
14 valid cases were accepted

Correlation matrix was computed for 6 variables
```

Define Method of Factor Extraction: Spreadsheet2
Quick
Advanced
Descriptives
Review correlations, means, standard deviations
Compute multiple regression analyses
OK
Cancel
Options

Fig. 66. **The *Review correlations, means, standard deviations* window**

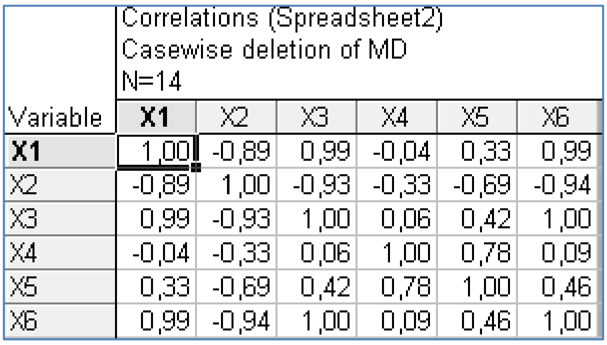
After that a window for viewing descriptive statistics for the analyzed data appears, where you can see the average, standard deviations, correlations, covariations, build different graphs (Fig. 67).



```
Missing data were casewise deleted

14 cases were processed (selected)
14 valid cases were accepted

Correlation matrix was computed for 6 variables
```

Fig. 67. **Additional analysis**

Here one can carry out additional analysis of the current data, verify the conformity of the sample variables to the normal distribution law and the existence of a linear correlation between the variables. Clicking the *Correlations* button will display the correlation matrix of the variables selected earlier (Fig. 68).

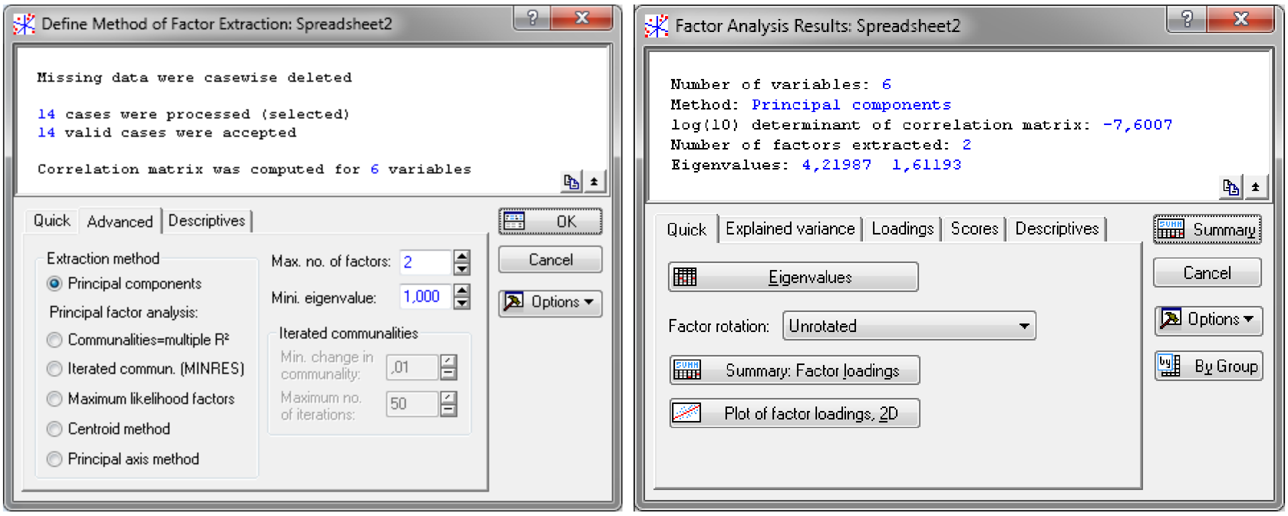


Correlations (Spreadsheet2)
Casewise deletion of MD
N=14

Variable	X1	X2	X3	X4	X5	X6
X1	1,00	-0,89	0,99	-0,04	0,33	0,99
X2	-0,89	1,00	-0,93	-0,33	-0,69	-0,94
X3	0,99	-0,93	1,00	0,06	0,42	1,00
X4	-0,04	-0,33	0,06	1,00	0,78	0,09
X5	0,33	-0,69	0,42	0,78	1,00	0,46
X6	0,99	-0,94	1,00	0,09	0,46	1,00

Fig. 68. **The correlation matrix**

So, at the next stage, we choose the method of allocation of factors – the method of the principal components (taking as a basis the correlation matrix of the initial data, this method allows you to reduce the dimension of the data and minimize the loss of information) and specify the maximum number of factors (6 in our case) and the minimum actual value of the Kaiser criterion (not less than 1). Statistica 10.0 automatically performs calculations and publishes the results of the factor analysis (Fig. 69).



Define Method of Factor Extraction: Spreadsheet2

```
Missing data were casewise deleted

14 cases were processed (selected)
14 valid cases were accepted

Correlation matrix was computed for 6 variables
```

Extraction method
● Principal components
Principal factor analysis:
○ Communalities=multiple R²
○ Iterated commun. (MINRES)
○ Maximum likelihood factors
○ Centroid method
○ Principal axis method

Max. no. of factors: 2
Mini. eigenvalue: 1,000
Iterated communalities
Min. change in communality: ,01
Maximum no. of iterations: 50

Factor Analysis Results: Spreadsheet2

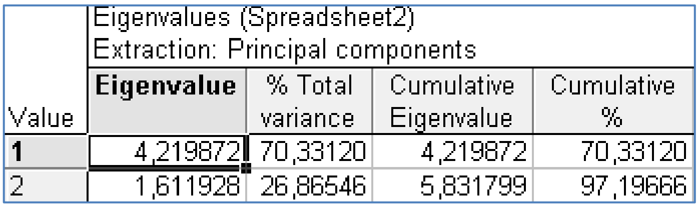
```
Number of variables: 6
Method: Principal components
log(10) determinant of correlation matrix: -7,6007
Number of factors extracted: 2
Eigenvalues: 4,21987  1,61193
```

Factor rotation: Unrotated

Fig. 69. **The *Define Method* window of factor analysis**

In the upper part of the window of the results of the factor analysis, an informative message is given: *Number of variables* (the number of analyzed variables): 6; *Method* (the method of analysis): *Principal components*; *log (10)* *determination of correlation matrix* (a decimal logarithm of the determinant of the correlation matrix): –7.6007; *Number of factors extracted* (the number of selected factors): 2; *Eigenvalues:* 4.21987; 1.61193.

As a result, two main factors that correspond to the highest eigenvalues of the correlation matrix were identified: λ1 = 4.21987 and λ2 = 1.61193, so two of these factors account for the largest part (97.2 %) of the variance explanation. Namely, the first factor explains 70 % (70.33 %) of the total dispersion, while the share of the second factor accounts for almost 27 % (26.86 %) of the dispersion explanation. Together, they describe approximately 97.2 % of the dispersion, that is, almost the entire array of data (Fig. 70).

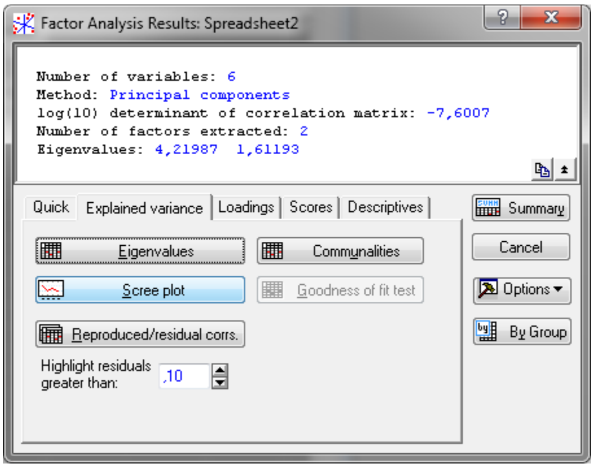


Eigenvalues (Spreadsheet2)
Extraction: Principal components

Value	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	4,219872	70,33120	4,219872	70,33120
2	1,611928	26,86546	5,831799	97,19666

Fig. 70. **The eigenvalues**

Thus, factorization is almost complete, although there are other, less significant factors. In order to ensure that the correct number of factors is obtained, it is expedient to use the Kettle criterion or Kettle Test (stony maturity criterion), which makes it possible to show graphically in descending order the eigenvalues of each selected factor and find a place on the graph where the reduction of these values from the left to the right as much as possible slows down. In accordance with this criterion, at points with coordinates 1, 2 the ash is slowed down most significantly, therefore, theoretically, one can restrict two factors (Fig. 71, 72).



```
Number of variables: 6
Method: Principal components
log(10) determinant of correlation matrix: -7,6007
Number of factors extracted: 2
Eigenvalues: 4,21987  1,61193
```

Highlight residuals greater than: ,10

Fig. 71. **Choosing the visualization of eigenvalues**

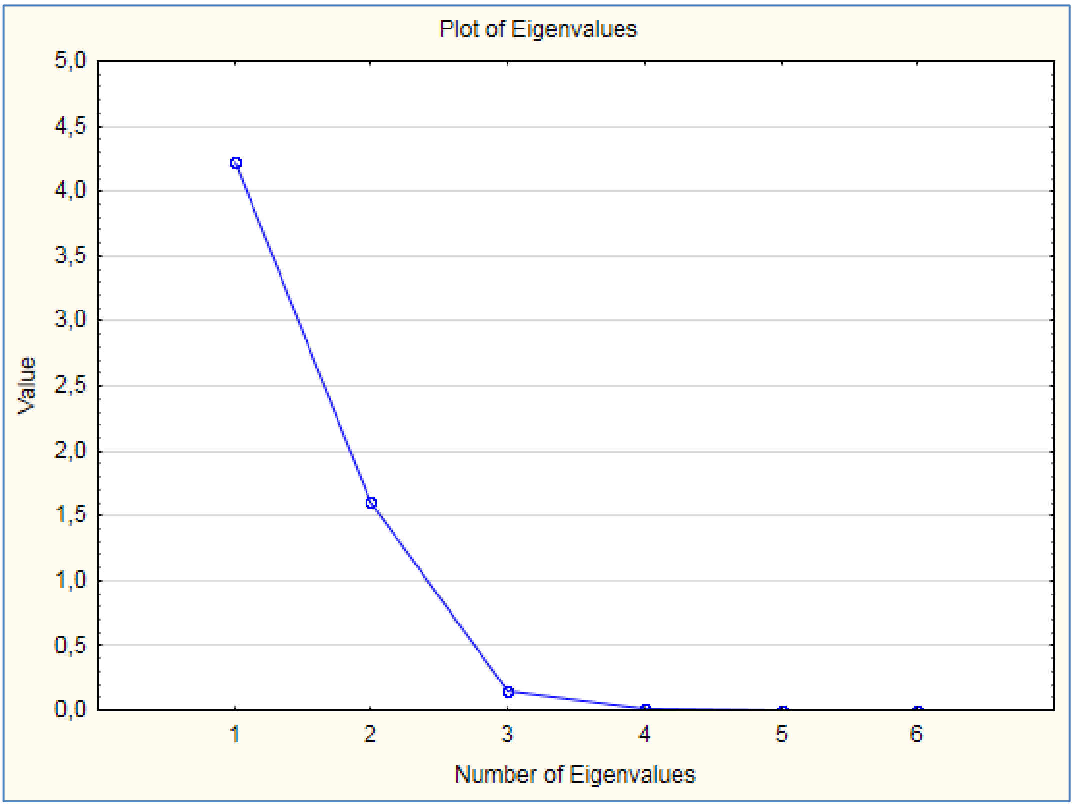
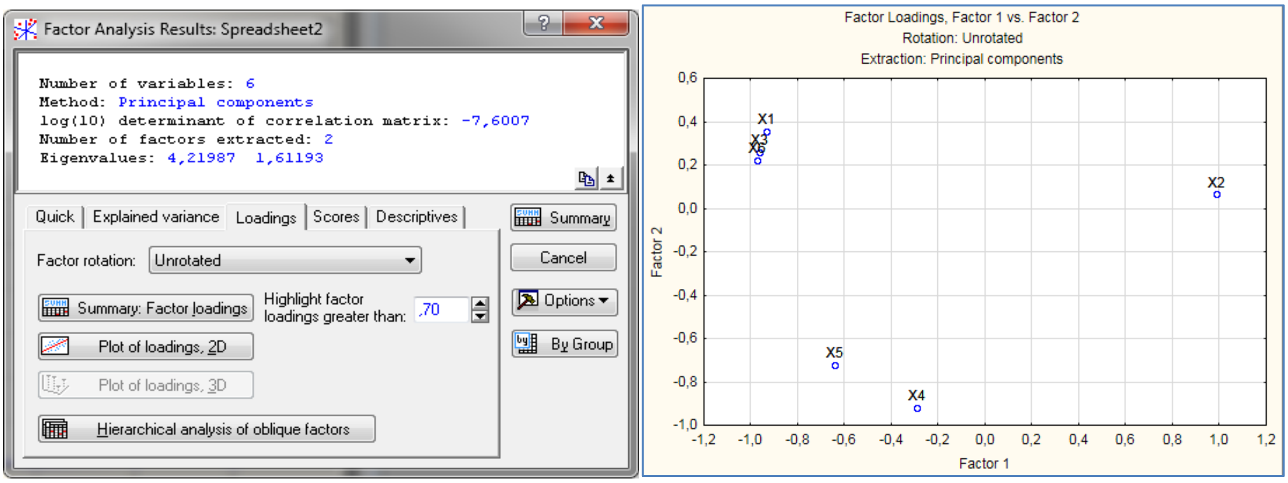


Fig. 72. **The plot of eigenvalues**

In the bottom of the window there are subdivisions that allow you to comprehensively get acquainted with the results of the analysis numerically and graphically. *Plot of loadings, 2D* and *Plot of loadings, 3D* (Load Charts) are options that will construct factor loading schedules in the projection onto the plane of any two selected factors and in projection into the space of the three selected factors (for which the presence of at least three selected factors is required) (Fig. 73).



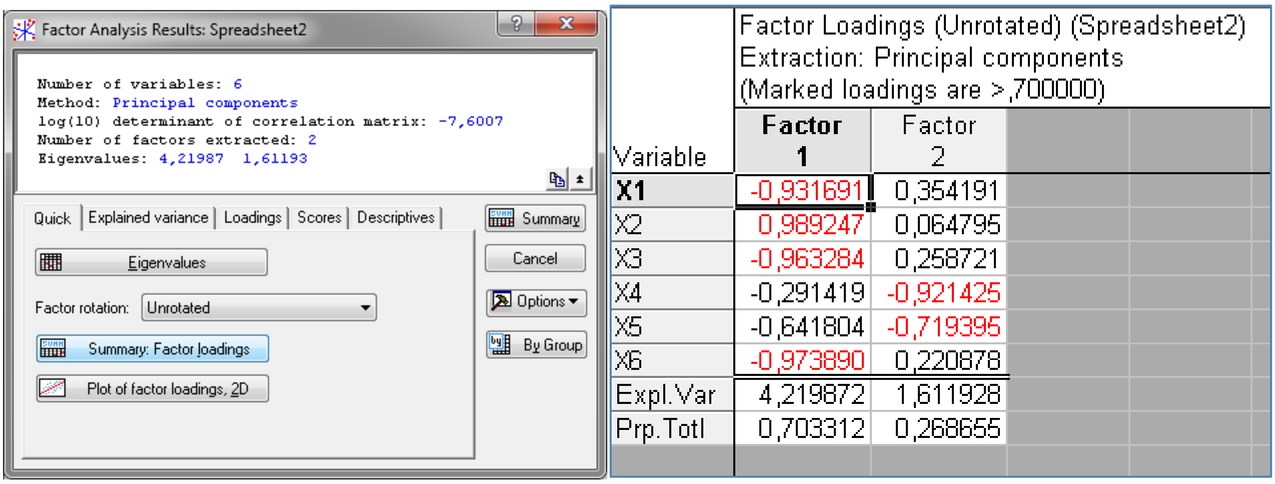
```
Number of variables: 6
Method: Principal components
log(10) determinant of correlation matrix: -7,6007
Number of factors extracted: 2
Eigenvalues: 4,21987  1,61193
```

Factor rotation: Unrotated

Highlight factor loadings greater than: ,70

Fig. 73. **The polygon of factor loadings**

Summary. Factor loadings. This option shows a table with current factor loadings, that is, those calculated for this method of rotation of factors, which is indicated to the right of the corresponding button. In this table, the factors correspond to the columns, and the variables are lines, and for each factor the loading of each output variable is given which shows the relative magnitude of the projection of the variable to the factor coordinate axis. Factor loadings can be interpreted as correlations between the corresponding variables and factors – the higher the loading modulus, the greater the proximity of the factor to the initial variable; and they represent the most important information for interpreting the resulting factors. In a generated table, for facilitation purposes, factor loadings will be allocated in absolute values greater than 0.7 (Fig. 74).



```
Number of variables: 6
Method: Principal components
log(10) determinant of correlation matrix: -7,6007
Number of factors extracted: 2
Eigenvalues: 4,21987  1,61193
```

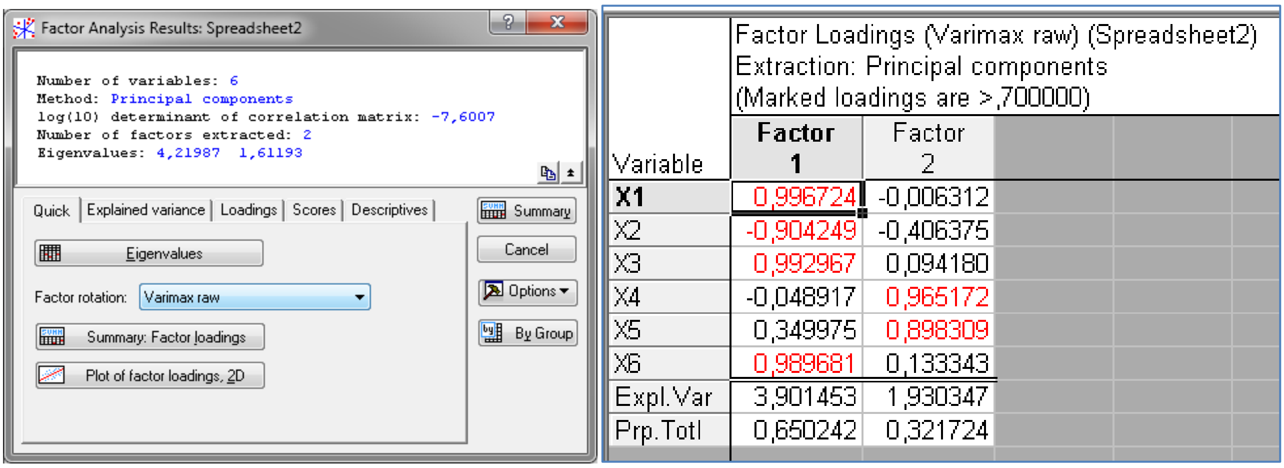
Factor rotation: Unrotated

Factor Loadings (Unrotated) (Spreadsheet2)
Extraction: Principal components
(Marked loadings are >,700000)

Variable	**Factor 1**	Factor 2
X1	-0,931691	0,354191
X2	0,989247	0,064795
X3	-0,963284	0,258721
X4	-0,291419	-0,921425
X5	-0,641804	-0,719395
X6	-0,973890	0,220878
Expl.Var	4,219872	1,611928
Prp.Totl	0,703312	0,268655

Fig. 74. **Factor loadings**

The results presented in Fig. 74 show that the first factor is more correlated with variables than the second one. Since the correlation of other factors is insignificant, in this case, it is advisable to resort to the rotation of the axes, hoping to obtain a solution that can be interpreted in the subject area. Now it is expedient to determine which indicators were included in the first and second factors. To do this, you should review the factor loading (correlations between the corresponding variables and factors – the higher the loading modulus, the greater the proximity of the factor to the output variable). However, one should immediately turn to the axes in order to obtain a simple structure in which most observations are located near the axes of coordinates. The results of changes in the composition of factors and factor loadings after their rotation using the normalized version are given in Fig. 75.



```
Number of variables: 6
Method: Principal components
log(10) determinant of correlation matrix: -7,6007
Number of factors extracted: 2
Eigenvalues: 4,21987  1,61193
```

Factor rotation: Varimax raw

Factor Loadings (Varimax raw) (Spreadsheet2)
Extraction: Principal components
(Marked loadings are >,700000)

Variable	Factor 1	Factor 2
X1	0,996724	-0,006312
X2	-0,904249	-0,406375
X3	0,992967	0,094180
X4	-0,048917	0,965172
X5	0,349975	0,898309
X6	0,989681	0,133343
Expl.Var	3,901453	1,930347
Prp.Totl	0,650242	0,321724

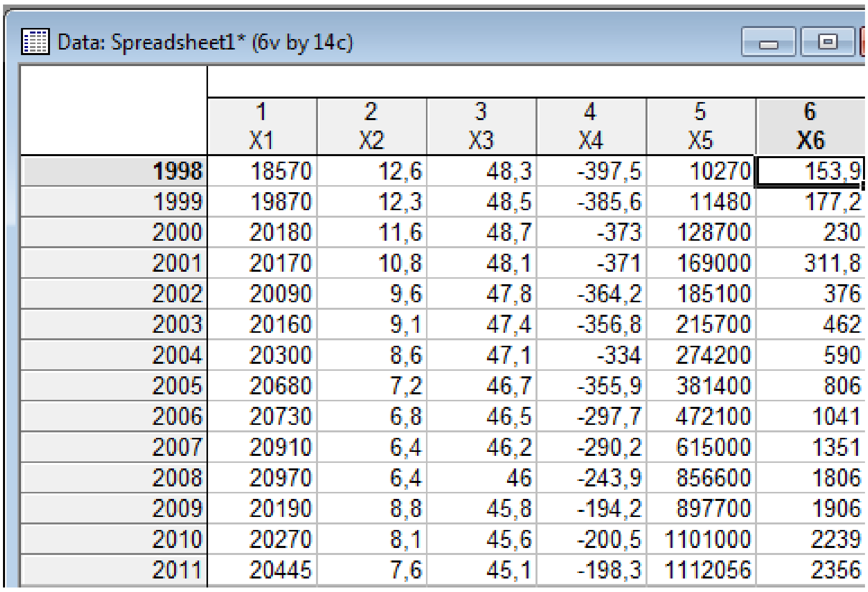
Fig. 75. **The results of changes in the composition of factors and factor loadings after their rotation**

Thus, according to the results, we have the following equations, where X1 – X6 are indicators characterizing the activity of a private enterprise, and f1, f2 are factor loadings:

X1 = 0.997 f1 – 0.006 f2;	X4 = -0.049 f1 + 0.965 f2;
X2 = –0.904 f1 – 0.406 f2;	X5 = 0.35 f1 + 0.898 f2;
X3 = 0.993 f1 + 0.094 f2;	X6 = 0.989 f1 + 0.133 f2.

The criterion for selecting quantitative indicators in the integral was applied factor loadings whose value was determined by the correlation coefficient of Pearson (R) and is within 0.7 ≥ R > 0.9, which, according to the Chaddock scale, indicates a strong correlation between the investigated parameters, which means that for the analyzed period 97 % of changes in the demographic situation is explained by the influence of the following indicators: X1, X2, X3 and X6. So, for detailed analysis of the demographic situation in the country it is advisable to leave such indicators as employment to population ratio, unemployment (% of total labor force), population and wage and salary.

Task 2. For the analysis of the demographic situation in Ukraine, the following indicators have been selected (Fig. 76): X1 – the number of employed; X2 – the unemployment rate; X3 – the number of permanent population; X4 – natural population increase (reduction); X5 – income of the population, UAH million; X6 – the average monthly nominal wage.



	1 X1	2 X2	3 X3	4 X4	5 X5	6 X6
1998	18570	12,6	48,3	-397,5	10270	153,9
1999	19870	12,3	48,5	-385,6	11480	177,2
2000	20180	11,6	48,7	-373	128700	230
2001	20170	10,8	48,1	-371	169000	311,8
2002	20090	9,6	47,8	-364,2	185100	376
2003	20160	9,1	47,4	-356,8	215700	462
2004	20300	8,6	47,1	-334	274200	590
2005	20680	7,2	46,7	-355,9	381400	806
2006	20730	6,8	46,5	-297,7	472100	1041
2007	20910	6,4	46,2	-290,2	615000	1351
2008	20970	6,4	46	-243,9	856600	1806
2009	20190	8,8	45,8	-194,2	897700	1906
2010	20270	8,1	45,6	-200,5	1101000	2239
2011	20445	7,6	45,1	-198,3	1112056	2356

Fig. 76. **The initial data in Statistica 10.0**

Make an aggregation of factors and provide an economic interpretation of the results of factor analysis.

Task 3. Reduce the information space (Fig. 77) using factor analysis methods. The initial indicators were the medical characteristics of the countries:

X1, population (thousand people);

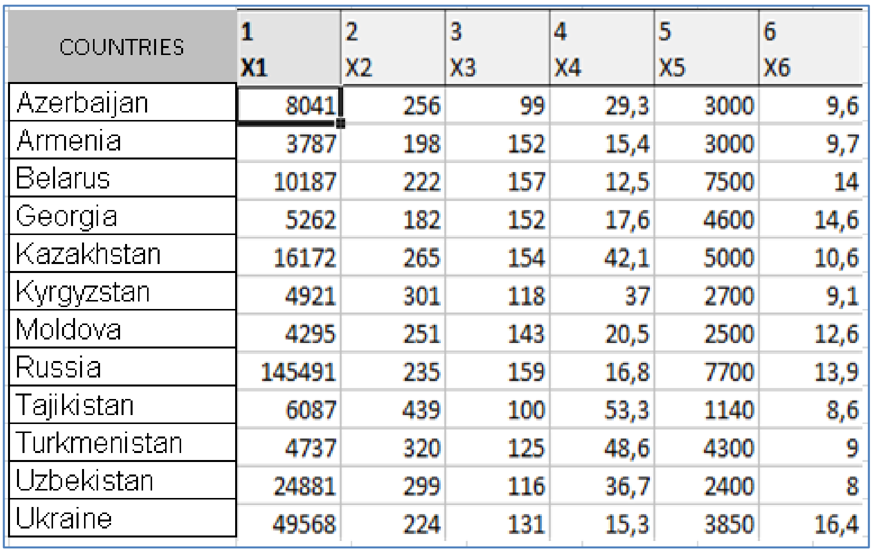
X2, the number of people per doctor;

X3, per capita health expenditures ($);

X4, infant mortality rate;

X5, GDP calculated at purchasing power parity per capita (million $);

X6, mortality per 1,000 people.



COUNTRIES	1 X1	2 X2	3 X3	4 X4	5 X5	6 X6
Azerbaijan	8041	256	99	29,3	3000	9,6
Armenia	3787	198	152	15,4	3000	9,7
Belarus	10187	222	157	12,5	7500	14
Georgia	5262	182	152	17,6	4600	14,6
Kazakhstan	16172	265	154	42,1	5000	10,6
Kyrgyzstan	4921	301	118	37	2700	9,1
Moldova	4295	251	143	20,5	2500	12,6
Russia	145491	235	159	16,8	7700	13,9
Tajikistan	6087	439	100	53,3	1140	8,6
Turkmenistan	4737	320	125	48,6	4300	9
Uzbekistan	24881	299	116	36,7	2400	8
Ukraine	49568	224	131	15,3	3850	16,4

Fig. 77. **The initial data in Statistica 10.0**

Task 4. Using the statistical data of the website of the State Statistics Service of Ukraine [19], find information about main indicators that characterize the level of country's economic development and reduce the information space using the methods of factor analysis using of the program Statistica 10.0.

Task 5. For analysis of the activities of a private enterprise (over the past 18 years), the following indicators were selected (Fig. 78):

X1, the proportion of losses from marriage;

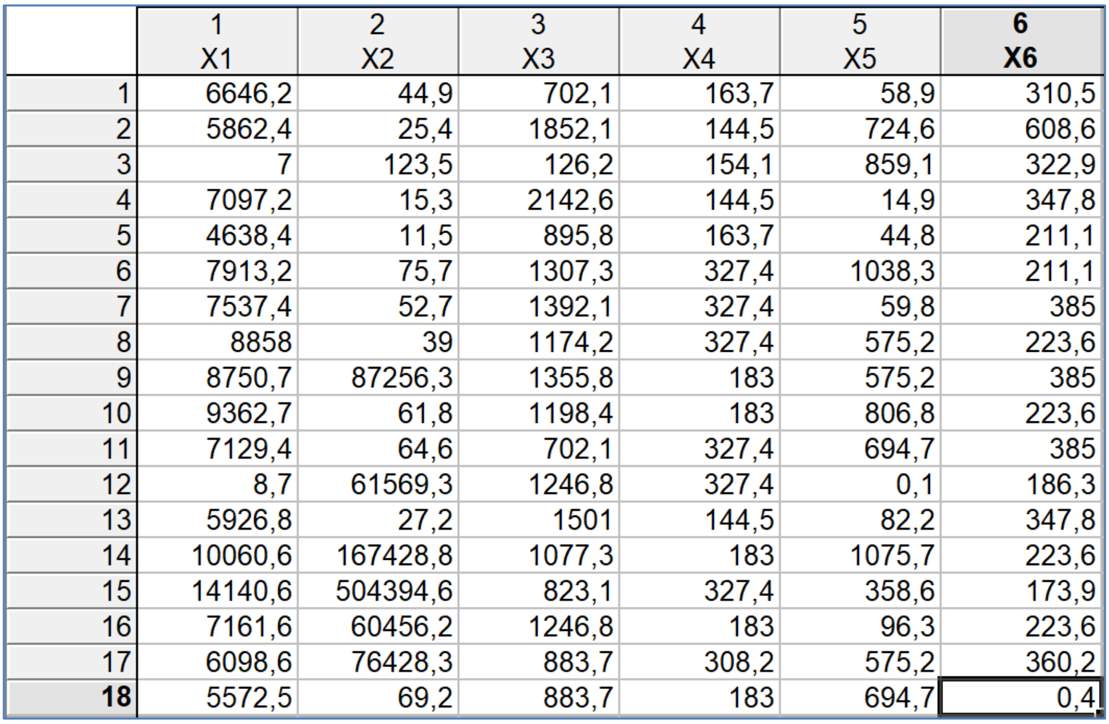
X2, the index of reducing production costs;

X3, return on capital;

X4, the equipment variability coefficient;

X5, labor productivity;

X6, the share of material costs.



	1	2	3	4	5	**6**
	X1	X2	X3	X4	X5	**X6**
1	6646,2	44,9	702,1	163,7	58,9	310,5
2	5862,4	25,4	1852,1	144,5	724,6	608,6
3	7	123,5	126,2	154,1	859,1	322,9
4	7097,2	15,3	2142,6	144,5	14,9	347,8
5	4638,4	11,5	895,8	163,7	44,8	211,1
6	7913,2	75,7	1307,3	327,4	1038,3	211,1
7	7537,4	52,7	1392,1	327,4	59,8	385
8	8858	39	1174,2	327,4	575,2	223,6
9	8750,7	87256,3	1355,8	183	575,2	385
10	9362,7	61,8	1198,4	183	806,8	223,6
11	7129,4	64,6	702,1	327,4	694,7	385
12	8,7	61569,3	1246,8	327,4	0,1	186,3
13	5926,8	27,2	1501	144,5	82,2	347,8
14	10060,6	167428,8	1077,3	183	1075,7	223,6
15	14140,6	504394,6	823,1	327,4	358,6	173,9
16	7161,6	60456,2	1246,8	183	96,3	223,6
17	6098,6	76428,3	883,7	308,2	575,2	360,2
18	5572,5	69,2	883,7	183	694,7	0,4

Fig. 78. **The initial data in Statistica 10.0**

Reduce the information space (see Fig. 78) using the factor analysis methods.

Task 6. Using your own information space of research make an aggregation of factors and provide an economic interpretation of the results of factor analysis.

Task 7. Using the statistical data of the website of the State Statistics Service of Ukraine [19], find information about the main indicators that characterize the level of socioeconomic growth of regions and reduce the information space using the methods of factor analysis in the program Statistica 10.0.

The list of questions for independent work

1. What is the essence of the method of factor analysis?
2. What is the basis for constructing a factor matrix?
3. What are the stages of the construction of the main components?
4. What is the essence of the procedure varimax?
5. What is matrix transformation and what does it depend on?
6. What are the methods of factor analysis?
7. What is the essence of factor loadings?
8. Expand on the essence of the previous quotient analysis procedures.
9. What is a group of factors?
10. What is the factor rotation method used for?

Topic 6. Cluster analysis as a means of forming homogeneous data groups

Task 1. Based on the analysis of the dendrogram shown in Fig. 79, draw a conclusion about the number of clusters in the population under study. Justify the answer.

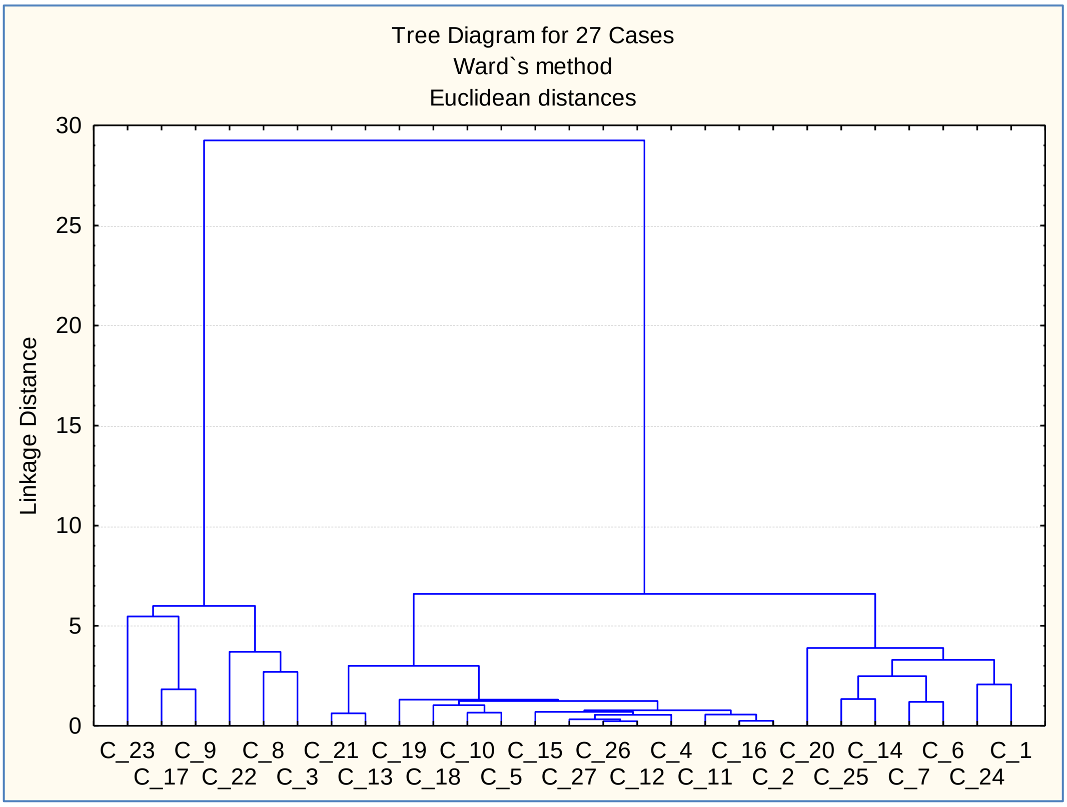


Fig. 79. **The dendrogram built by means of the program Statistica 10.0**

Guidelines

Using the constructed dendrite, you can determine the number of clusters into which it is advisable to divide the set of objects using two approaches.

The first approach is based on the visual analysis of the dendrites shown in Fig. 79. It can be concluded that it is necessary to divide the set of objects into 3 clusters.

The second approach is based on natural ways of allocating the number of clusters into which the set is divided. This approach consists of the following steps:

1. The bonds of the dendrites built on the units of the set are arranged in descending order of their length (formula 7):

$$i_2 = \frac{d_1}{d_2},\ i_3 = \frac{d_2}{d_3}, ..., i_{\varpi-1} = \frac{d_{\varpi-2}}{d_{\varpi-1}}, \tag{7}$$

where d_1, d_2,… $d_{\varpi-1}$ are the ordered bond lengths;

i_1, i_2,… $i_{\varpi-1}$ is the ratio of bond lengths.

2. A search is made for the following value of i_k for which the following relation holds:

$$i_k < i_{k+1} \text{ for } k = 2,3,...,\overline{\varpi} - 1.$$

If this relation is satisfied, it can be argued that it is preferable to divide the population into k parts:

$$i_2 = \frac{23}{22} = 1.04,\ i_3 = \frac{22}{6} = 3.6,\ i_4 = \frac{6}{4} = 1.5.$$

That is, for k = 2, the condition 1.04 < 3.6 is fulfilled.

Based on the analysis, we can conclude that in a further study it is necessary to build two and three clusters by artificial clustering and compare the quality of the obtained clusters.

Task 2. Classify (group) cars and their owners into insurance risk groups, which is assessed based on the following indicators:

1) the cost of the car, million UAH (X1);

2) the age of the driver, years (X2);

3) the experience of the driver, years (X3);

4) the age of the car, years (X4).

The initial values of the indicators are given in Table 16.

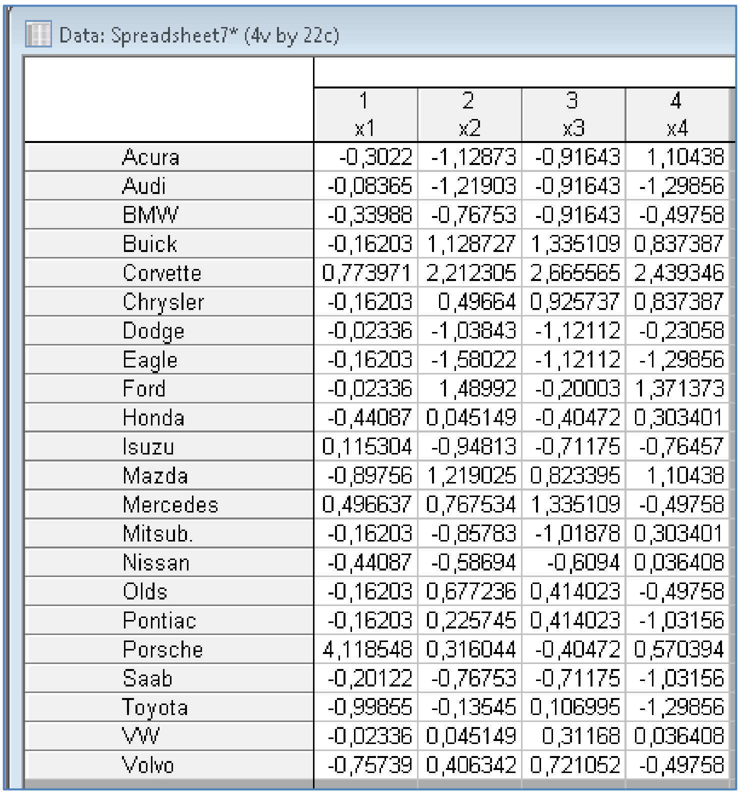
Table 16

The initial data

No.	Car model	X1	X2	X3	X4
1	Acura	0.521	25	3	10
2	Audi	0.666	24	3	1
3	BMW	0.496	29	3	4
4	Buick	0.614	50	25	9
5	Corvette	1.235	62	38	15
6	Chrysler	0.614	43	21	9
7	Dodge	0.706	26	1	5
8	Eagle	0.614	20	1	1
9	Ford	0.706	54	10	11
10	Honda	0.429	38	8	7
11	Isuzu	0.798	27	5	3
12	Mazda	0.126	51	20	10
13	Mercedes	1.051	46	25	4
14	Mitsubishi	0.614	28	2	7
15	Nissan	0.429	31	6	6
16	Olds	0.614	45	16	4
17	Pontiac	0.614	40	16	2
18	Porsche	3.454	41	8	8
19	Saab	0.588	29	5	2
20	Toyota	0.059	36	13	1
21	VW	0.706	38	15	6
22	Volvo	0.219	42	19	4

Guidelines

1. To construct cluster groups, we assess the values of the indicators. For this purpose, in the context menu, we need to standardize the initial data. So, select *Fill / Standardize Block / Standardize Columns* as shown in Fig. 80.



Data: Spreadsheet7* (4v by 22c)

	1 x1	2 x2	3 x3	4 x4
Acura	-0,3022	-1,12873	-0,91643	1,10438
Audi	-0,08365	-1,21903	-0,91643	-1,29856
BMW	-0,33988	-0,76753	-0,91643	-0,49758
Buick	-0,16203	1,128727	1,335109	0,837387
Corvette	0,773971	2,212305	2,665565	2,439346
Chrysler	-0,16203	0,49664	0,925737	0,837387
Dodge	-0,02336	-1,03843	-1,12112	-0,23058
Eagle	-0,16203	-1,58022	-1,12112	-1,29856
Ford	-0,02336	1,48992	-0,20003	1,371373
Honda	-0,44087	0,045149	-0,40472	0,303401
Isuzu	0,115304	-0,94813	-0,71175	-0,76457
Mazda	-0,89756	1,219025	0,823395	1,10438
Mercedes	0,496637	0,767534	1,335109	-0,49758
Mitsub.	-0,16203	-0,85783	-1,01878	0,303401
Nissan	-0,44087	-0,58694	-0,6094	0,036408
Olds	-0,16203	0,677236	0,414023	-0,49758
Pontiac	-0,16203	0,225745	0,414023	-1,03156
Porsche	4,118548	0,316044	-0,40472	0,570394
Saab	-0,20122	-0,76753	-0,71175	-1,03156
Toyota	-0,99855	-0,13545	0,106995	-1,29856
VW	-0,02336	0,045149	0,31168	0,036408
Volvo	-0,75739	0,406342	0,721052	-0,49758

Fig. 80. **The normative values of energy security indices**

2. To perform a cluster analysis, we need to log into the cluster analysis module; so, we need to use the menu *Statistics / Multivariate Exploratory / Cluster Analysis* menu (Fig. 81).

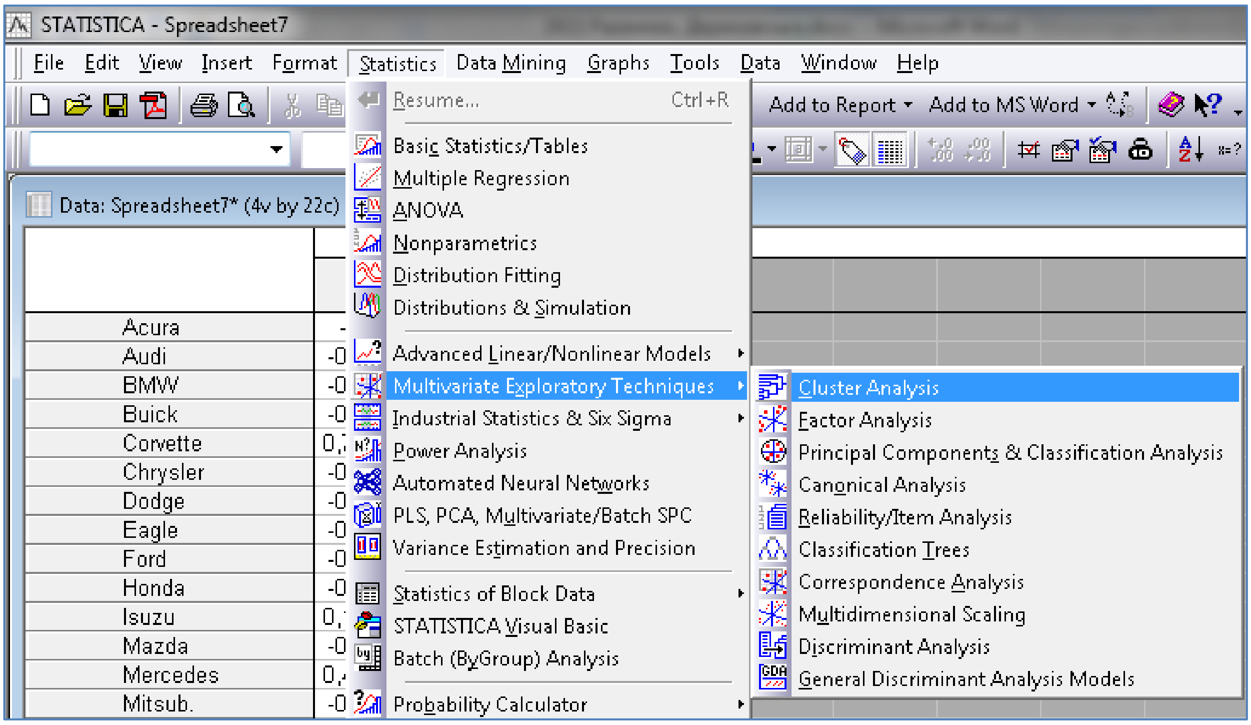


Fig. 81. **The *Cluster Analysis* module**

The resulting dialog allows you to use one of the methods of clustering:

1) Joining (tree clustering);

2) K-means clustering;

3) Two-way joining.

Let's start cluster analysis with methods of natural hierarchical clustering – Single Linkage (Fig. 82).

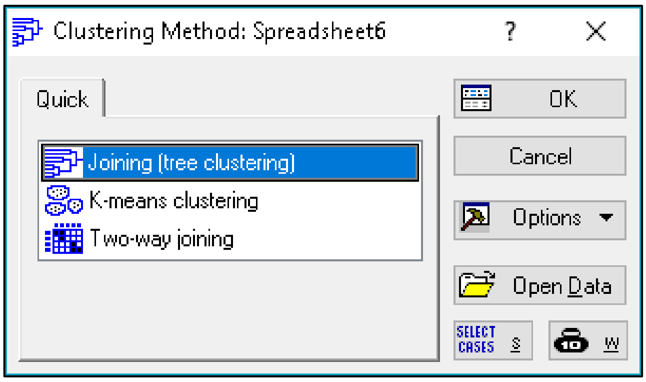
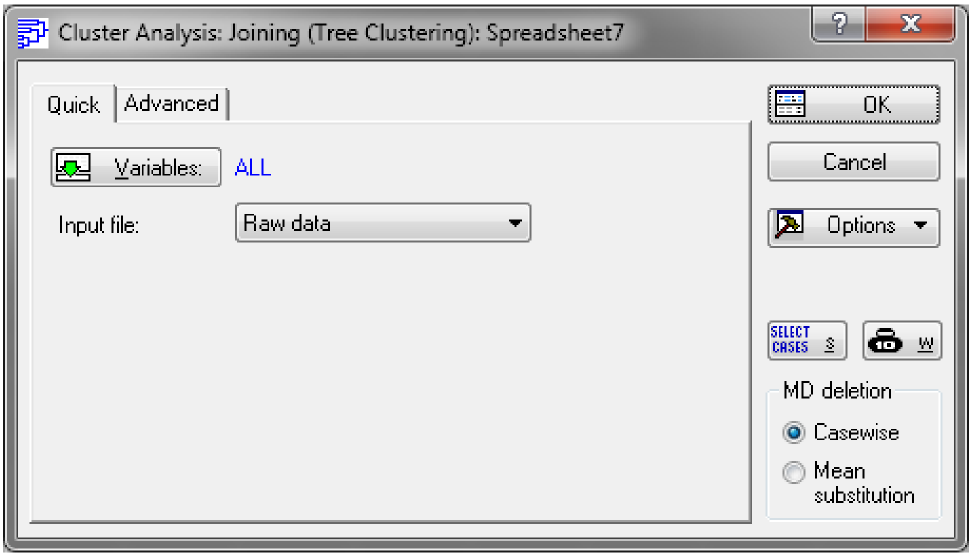


Fig. 82. **Choosing the Ward method**

3. In order to determine the number of clusters it is expedient to initially conduct natural (tree-like) clustering. In the Statistica 10.0 package. This type of clustering involves the implementation of several stages.

3.1. Selection of indicators for which clustering is carried out (Fig. 83).


Variables: ALL

Input file: Raw data

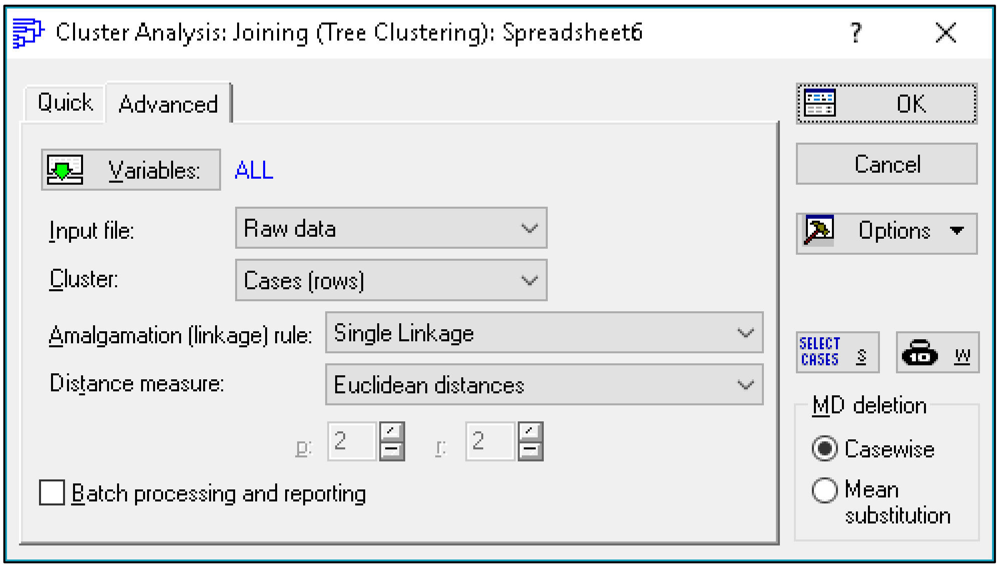
MD deletion

● Casewise

○ Mean substitution

Fig. 83. **Selection of the indicators for the analysis**

3.2. Selection of objects of classification in the *Cluster* field. When clustering the variables themselves, they are labeled *Variables* [*columns*], in this task *Cases* [*rows*] (Fig. 84).



Variables: ALL

Input file: Raw data

Cluster: Cases (rows)

Amalgamation (linkage) rule: Single Linkage

Distance measure: Euclidean distances

p: 2 r: 2

□ Batch processing and reporting

MD deletion

● Casewise

○ Mean substitution

Fig. 84. **Choosing the parameters of cluster analysis**

3.3. The choice of rules for grouping objects. To do this, use the *Amalgamation [linkage] rule* menu, which allows you to choose one of the following rules:

- ➢ Single Linkage (one-way method "Closest neighbor's principle").
- ➢ Complete Linkage ("full-length" method).
- ➢ Unweighted pair-group average (unweighted pair average).
- ➢ Weighted pair-group average (weighted pairwise average).
- ➢ Unweighted pair-group centroid (unweighted centroid method).
- ➢ Weighted pair-group centroid (Weighted centroid method).
- ➢ Ward's method.

According to the work purpose, let's use the single-linkage method.

3.4. Choosing the distance type to be used in the clustering process. For this purpose, in the *Distance measure* window, you must select one of the distance types used in the package:

• Squared Euclidean distances (square of the Euclidean distance);

• Euclidean distances;

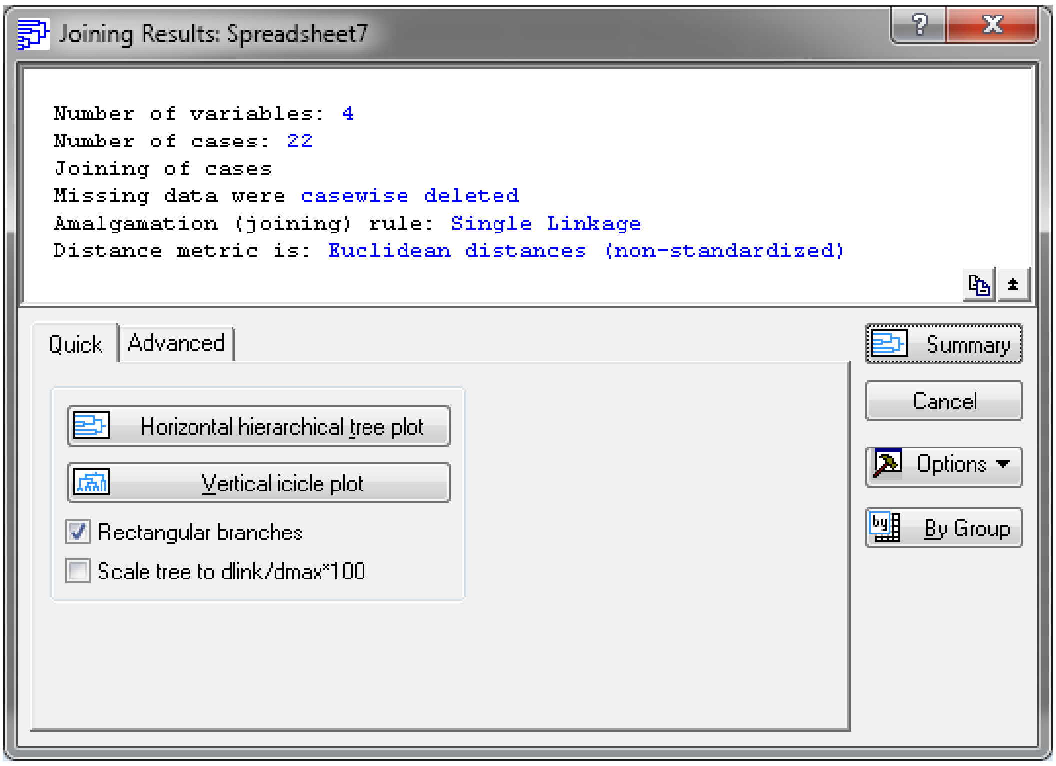
• City-block (Manhattan) distance (distance from city districts (Manhattan distance));

• Chebyshev distance metric (Chebyshev distance);

• Percent disagreement.

According to the work purpose, let's use the Euclidean distance.

After setting all clustering parameters, we go to the window of its results (Fig. 85).



Joining Results: Spreadsheet7
Quick
Advanced
Horizontal hierarchical tree plot
Vertical icicle plot
Summary
Cancel
Options
By Group

```
Number of variables: 4
Number of cases: 22
Joining of cases
Missing data were casewise deleted
Amalgamation (joining) rule: Single Linkage
Distance metric is: Euclidean distances (non-standardized)
```

☑ Rectangular branches

□ Scale tree to dlink/dmax*100

Fig. 85. **Choosing the parameters of cluster analysis**

Using the *Vertical (Horizontal) icicle plot* button, we build a vertical dendrogram (Fig. 86) and a horizontal hierarchical dendrogram (Fig. 87).

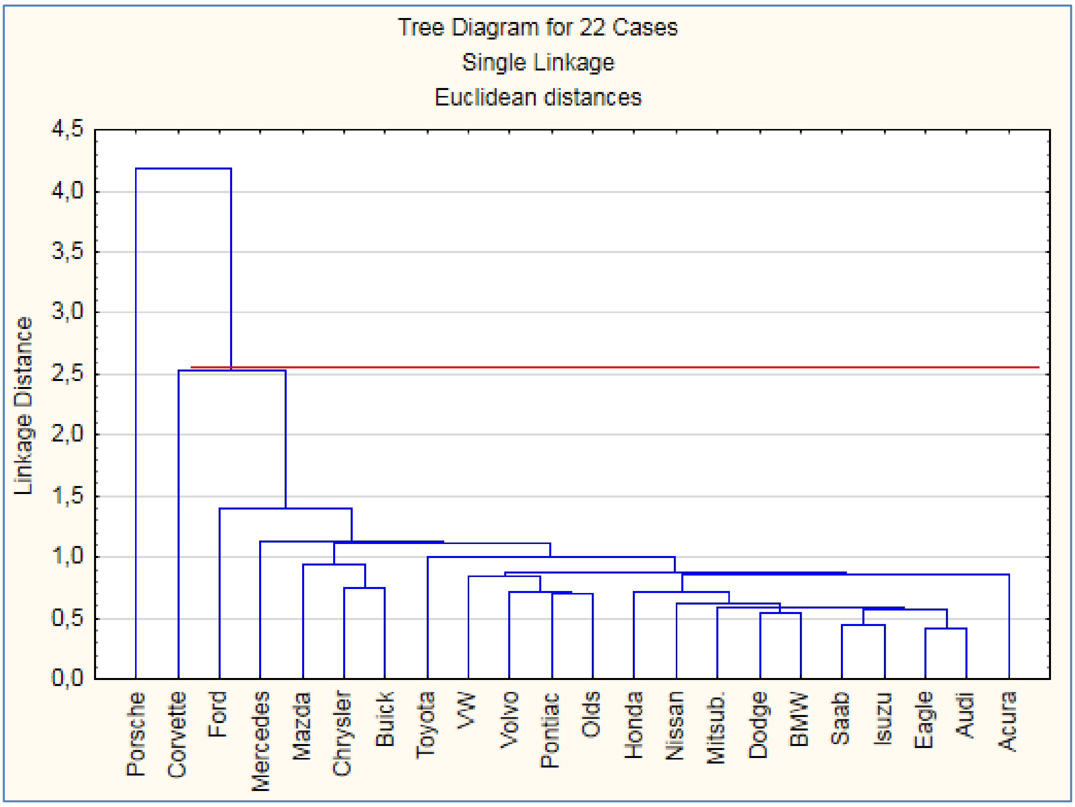


Fig. 86. **The vertical dendrogram**

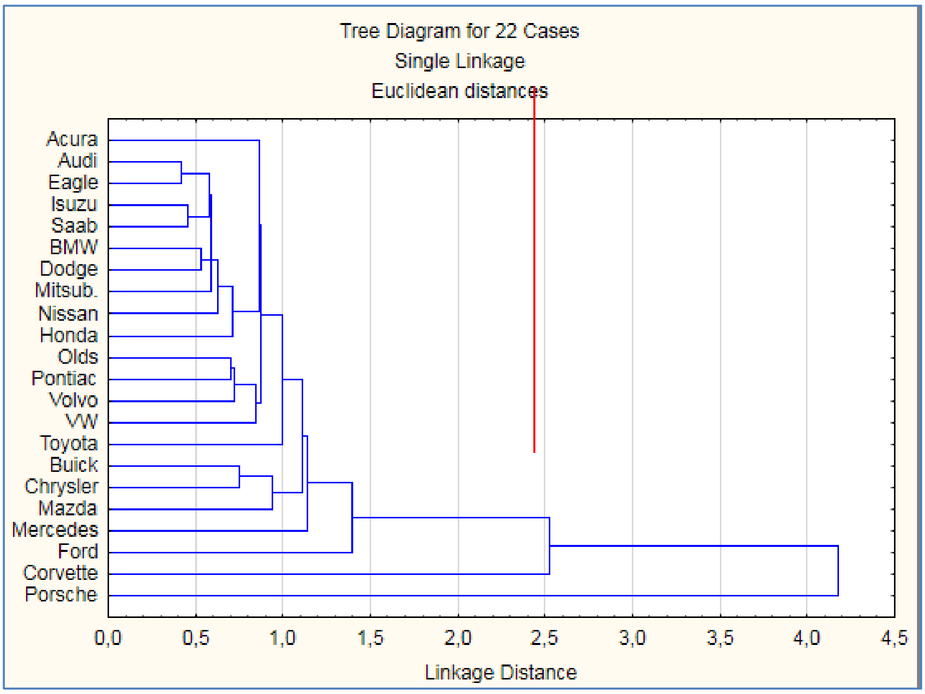
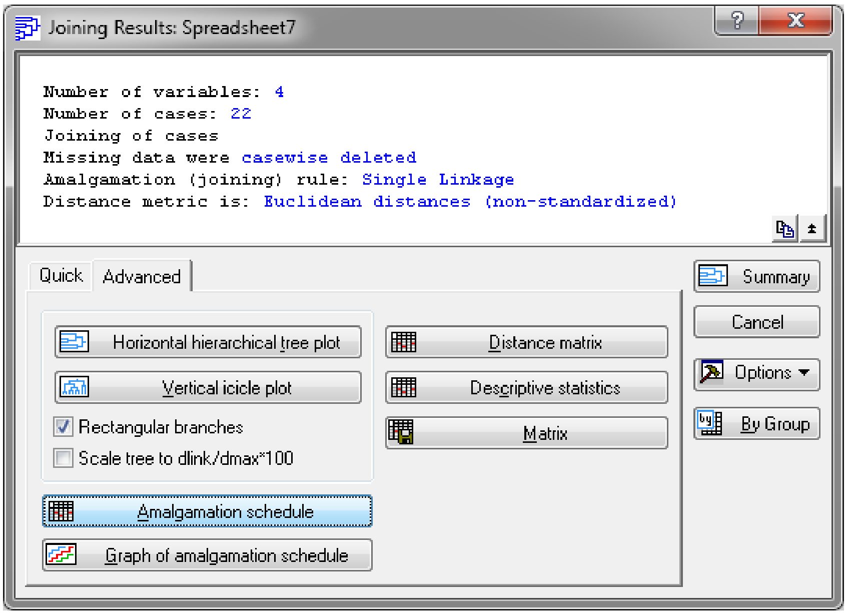


Fig. 87. **The horizontal hierarchical dendrogram**

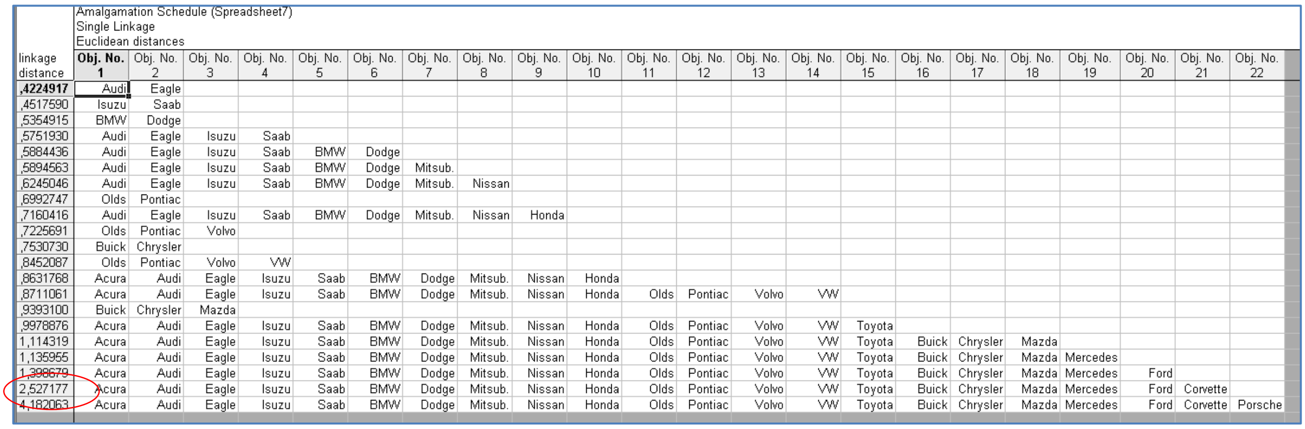
Graphical analysis of the number of cluster groups based on the use of dendrograms in Fig. 86, 87, suggests that it is most appropriate to divide the set of the car models into 3 clusters, as the number of border crossings is 3 times.

Also, one of the tools available in Statistica for selecting the number of clusters is the graph of the merging process (the button *Graph of Amalgamation schedule*) and the table of the merging objects (the button *Amalgamation schedule*), presented in Fig. 88, 89.



```
Number of variables: 4
Number of cases: 22
Joining of cases
Missing data were casewise deleted
Amalgamation (joining) rule: Single Linkage
Distance metric is: Euclidean distances (non-standardized)
```

Fig. 88. **The table of the merging objects**



Amalgamation Schedule (Spreadsheet7)
Single Linkage
Euclidean distances

linkage distance	**Obj. No. 1**	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5	Obj. No. 6	Obj. No. 7	Obj. No. 8	Obj. No. 9	Obj. No. 10	Obj. No. 11	Obj. No. 12	Obj. No. 13	Obj. No. 14	Obj. No. 15	Obj. No. 16	Obj. No. 17	Obj. No. 18	Obj. No. 19	Obj. No. 20	Obj. No. 21	Obj. No. 22
,4224917	Audi	Eagle																				
,4517590	Isuzu	Saab																				
,5354915	BMW	Dodge																				
,5751930	Audi	Eagle	Isuzu	Saab																		
,5884436	Audi	Eagle	Isuzu	Saab	BMW	Dodge																
,5894563	Audi	Eagle	Isuzu	Saab	BMW	Dodge	Mitsub.															
,6245046	Audi	Eagle	Isuzu	Saab	BMW	Dodge	Mitsub.	Nissan														
,6992747	Olds	Pontiac																				
,7160416	Audi	Eagle	Isuzu	Saab	BMW	Dodge	Mitsub.	Nissan	Honda													
,7225691	Olds	Pontiac	Volvo																			
,7530730	Buick	Chrysler																				
,8452087	Olds	Pontiac	Volvo	VW																		
,8631768	Acura	Audi	Eagle	Isuzu	Saab	BMW	Dodge	Mitsub.	Nissan	Honda												
,8711061	Acura	Audi	Eagle	Isuzu	Saab	BMW	Dodge	Mitsub.	Nissan	Honda	Olds	Pontiac	Volvo	VW								
,9393100	Buick	Chrysler	Mazda																			
,9978876	Acura	Audi	Eagle	Isuzu	Saab	BMW	Dodge	Mitsub.	Nissan	Honda	Olds	Pontiac	Volvo	VW	Toyota							
1,114319	Acura	Audi	Eagle	Isuzu	Saab	BMW	Dodge	Mitsub.	Nissan	Honda	Olds	Pontiac	Volvo	VW	Toyota	Buick	Chrysler	Mazda				
1,135955	Acura	Audi	Eagle	Isuzu	Saab	BMW	Dodge	Mitsub.	Nissan	Honda	Olds	Pontiac	Volvo	VW	Toyota	Buick	Chrysler	Mazda	Mercedes			
1,398679	Acura	Audi	Eagle	Isuzu	Saab	BMW	Dodge	Mitsub.	Nissan	Honda	Olds	Pontiac	Volvo	VW	Toyota	Buick	Chrysler	Mazda	Mercedes	Ford		
2,527177	Acura	Audi	Eagle	Isuzu	Saab	BMW	Dodge	Mitsub.	Nissan	Honda	Olds	Pontiac	Volvo	VW	Toyota	Buick	Chrysler	Mazda	Mercedes	Ford	Corvette	
4,182063	Acura	Audi	Eagle	Isuzu	Saab	BMW	Dodge	Mitsub.	Nissan	Honda	Olds	Pontiac	Volvo	VW	Toyota	Buick	Chrysler	Mazda	Mercedes	Ford	Corvette	Porsche

Fig. 88. (the end)

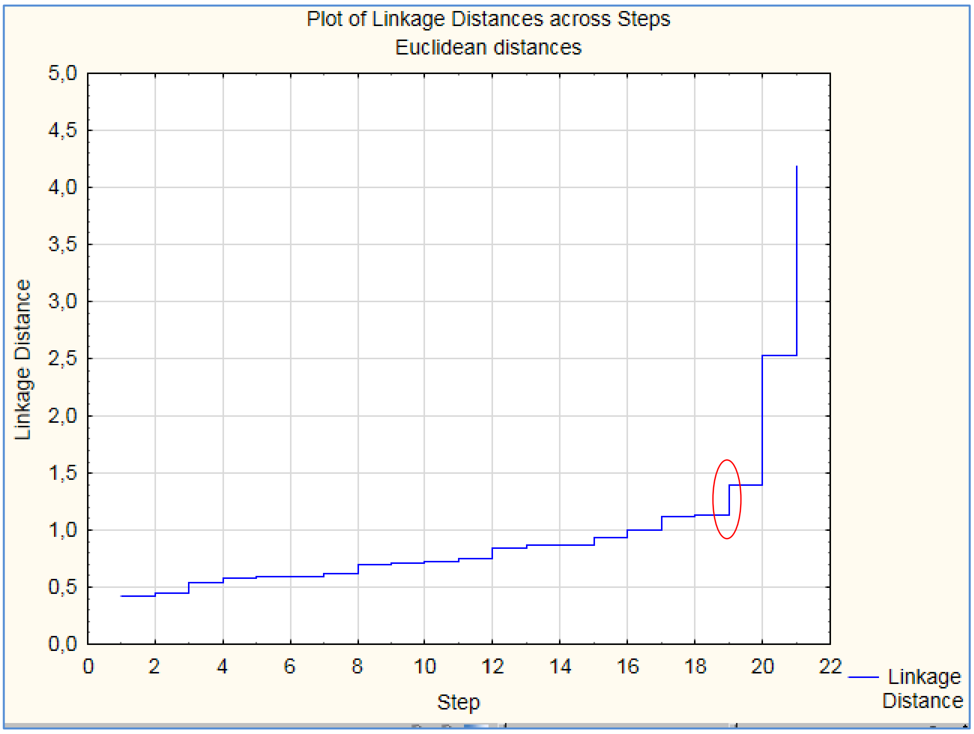


Fig. 89. **The graph of the merging process**

How to use these tools to determine the number of clusters? There are some practical recommendations:

1) on the graph, there is a point of fracture and a step number m, where this fracture occurred; then the number of clusters is n – m, where n is the number of objects in the sample;

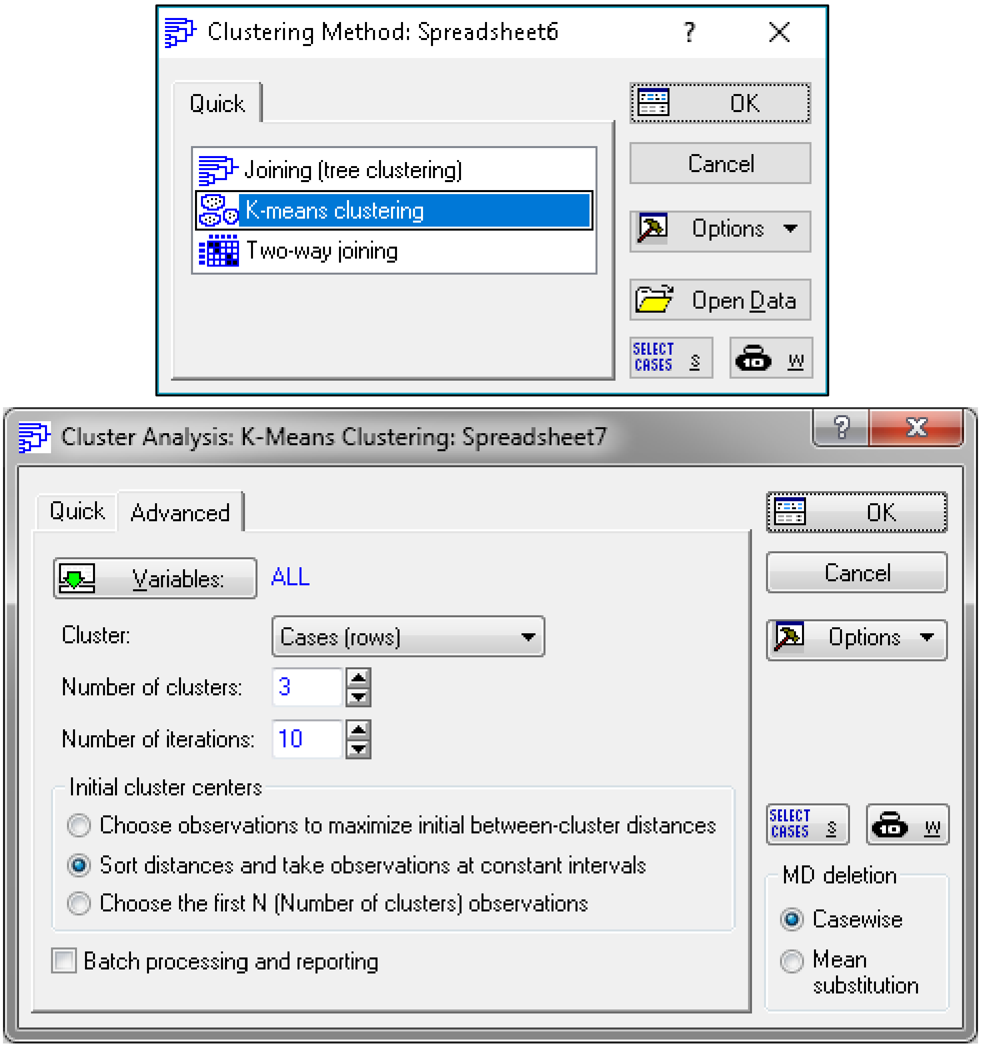
2) in the column *linkage distance* of the table of merging objects, there is a step number m, where merging of objects took place at a significantly greater distance than in the step m – 1; then the number of clusters is n – m, where n is the number of objects in the sample.

In our case, we can consider step number 19, as a turning point, where we get 22 – 19 = 3 clusters. Also, the analysis of the table of merging objects shows that in the 19th step, there was a distance jump of almost 1.13 units, while in the previous steps, the jumps did not exceed 0.2 units.

Thus, according to the level of risk, the studied car models should be divided into 3 clusters.

4. Construction of clusters using the method of k-means (non-hierarchical clustering) is carried out in the following stages:

4.1. Setting the basic clustering parameters. Similar to the method of tree clustering, the indicators based on which the clustering is carried out are selected. Taking into account the results of the hierarchical method, the number of clusters equal to 3 is indicated (Fig. 89).


Clustering Method: Spreadsheet6

Quick

- Joining (tree clustering)
- K-means clustering
- Two-way joining

Cluster Analysis: K-Means Clustering: Spreadsheet7

Quick | Advanced

Variables: ALL

Cluster: Cases (rows)

Number of clusters: 3

Number of iterations: 10

Initial cluster centers

○ Choose observations to maximize initial between-cluster distances

● Sort distances and take observations at constant intervals

○ Choose the first N (Number of clusters) observations

□ Batch processing and reporting

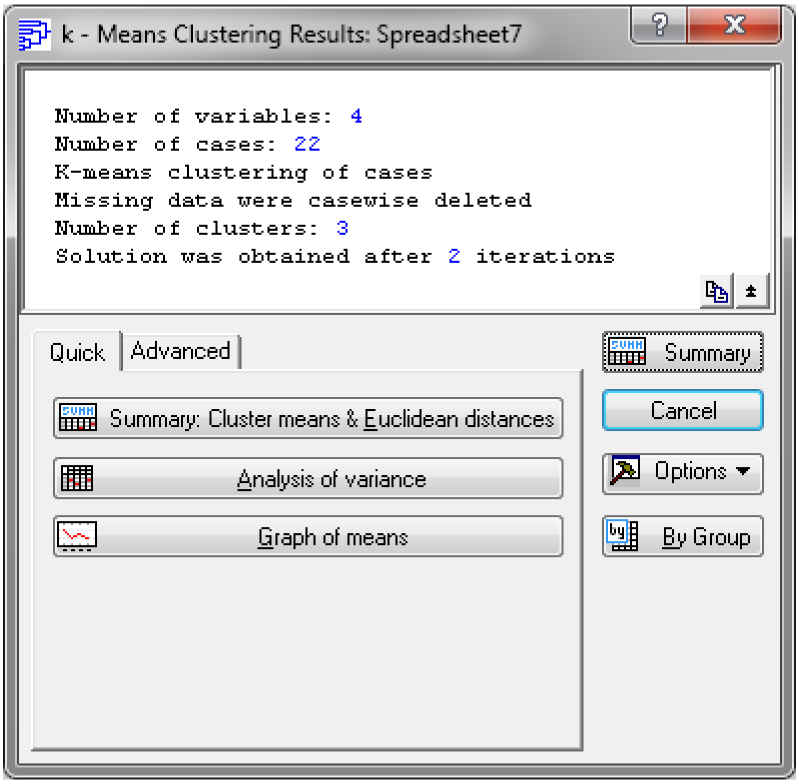
MD deletion

● Casewise

○ Mean substitution

Fig. 89. **The stages of the k-means method**

4.2. In the clustering results window, you can select those calculations and reports for the cluster analysis that the user needs (Fig. 90).



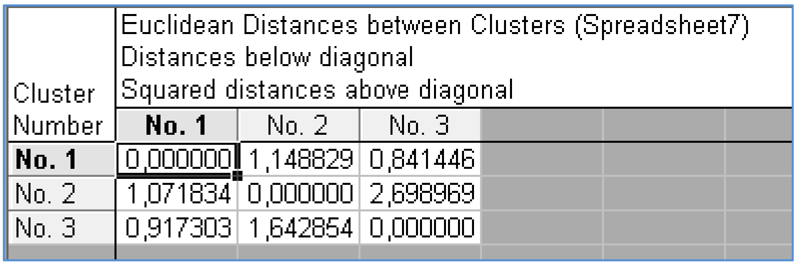
k - Means Clustering Results: Spreadsheet7

```
Number of variables: 4
Number of cases: 22
K-means clustering of cases
Missing data were casewise deleted
Number of clusters: 3
Solution was obtained after 2 iterations
```

Fig. 90. **Selection of the parameters of the k-means method**

4.3. The results of cluster analysis.

4.3.1. The *Cluster Means & Euclidean Distances* button (mean values in the clusters and Euclidean distances) is presented in Fig. 91.



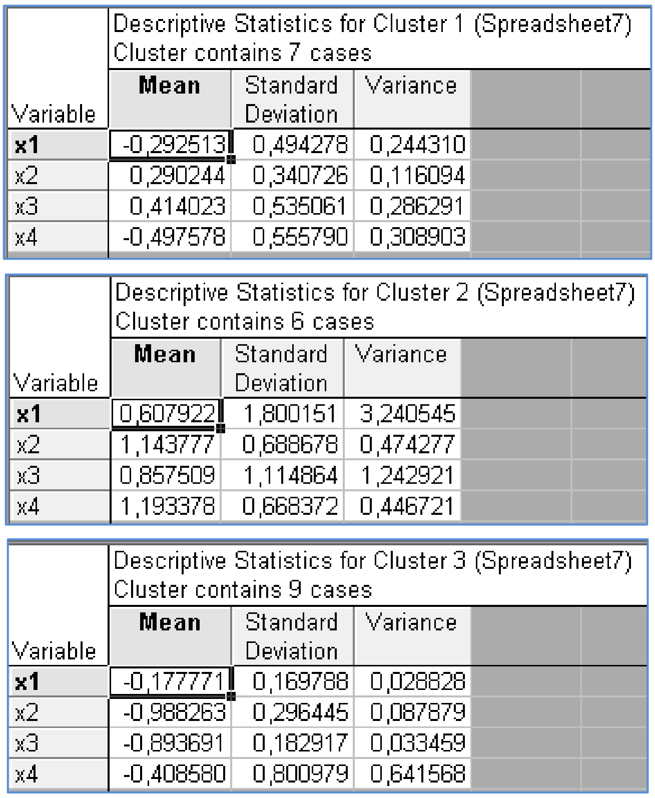
Euclidean Distances between Clusters (Spreadsheet7)
Distances below diagonal
Squared distances above diagonal

Cluster Number	No. 1	No. 2	No. 3
No. 1	0,000000	1,148829	0,841446
No. 2	1,071834	0,000000	2,698969
No. 3	0,917303	1,642854	0,000000

Fig. 91. **The Euclidean distances**

By the matrix of distances between clusters, one can determine the quality of the clusterization carried out. The greater the distance between the clusters and the less the distance between the elements of the clusters, the more qualitative clustering is carried out.

4.3.2. The *Descriptive Statistics* button for each cluster allows you to define descriptive statistics for each cluster (Fig. 92).



Descriptive Statistics for Cluster 1 (Spreadsheet7)
Cluster contains 7 cases

Variable	Mean	Standard Deviation	Variance
x1	-0,292513	0,494278	0,244310
x2	0,290244	0,340726	0,116094
x3	0,414023	0,535061	0,286291
x4	-0,497578	0,555790	0,308903

Descriptive Statistics for Cluster 2 (Spreadsheet7)
Cluster contains 6 cases

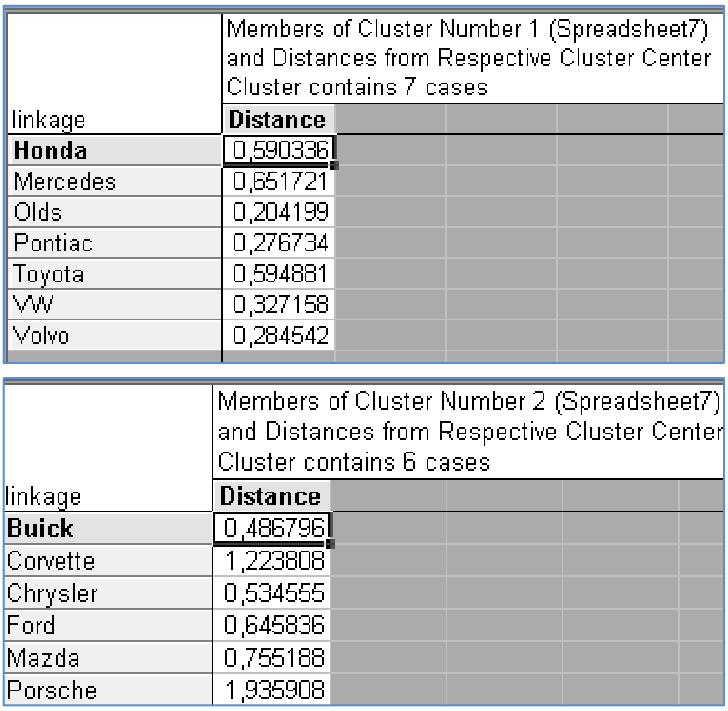
Variable	Mean	Standard Deviation	Variance
x1	0,607922	1,800151	3,240545
x2	1,143777	0,688678	0,474277
x3	0,857509	1,114864	1,242921
x4	1,193378	0,668372	0,446721

Descriptive Statistics for Cluster 3 (Spreadsheet7)
Cluster contains 9 cases

Variable	Mean	Standard Deviation	Variance
x1	-0,177771	0,169788	0,028828
x2	-0,988263	0,296445	0,087879
x3	-0,893691	0,182917	0,033459
x4	-0,408580	0,800979	0,641568

Fig. 92. **The descriptive statistics for each cluster**

4.3.3. The list of cars included in each cluster can be obtained using the *Members for each cluster & distances* button (group members and distances) (Fig. 93).



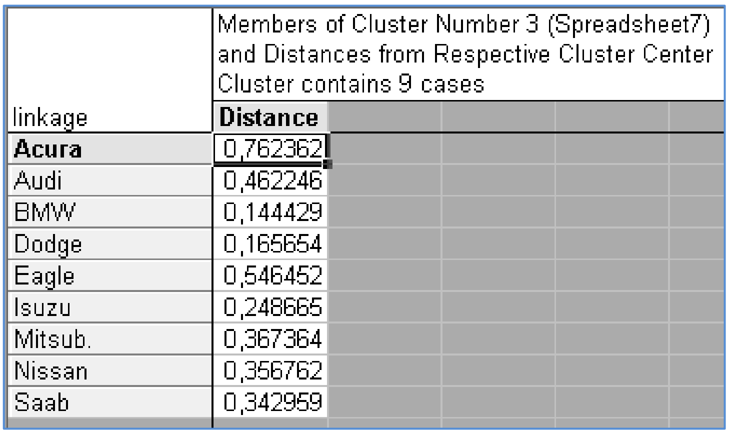
Members of Cluster Number 1 (Spreadsheet7)
and Distances from Respective Cluster Center
Cluster contains 7 cases

linkage	Distance
Honda	0,590336
Mercedes	0,651721
Olds	0,204199
Pontiac	0,276734
Toyota	0,594881
VW	0,327158
Volvo	0,284542

Members of Cluster Number 2 (Spreadsheet7)
and Distances from Respective Cluster Center
Cluster contains 6 cases

linkage	Distance
Buick	0,486796
Corvette	1,223808
Chrysler	0,534555
Ford	0,645836
Mazda	0,755188
Porsche	1,935908

Fig. 93. **The members of each cluster**



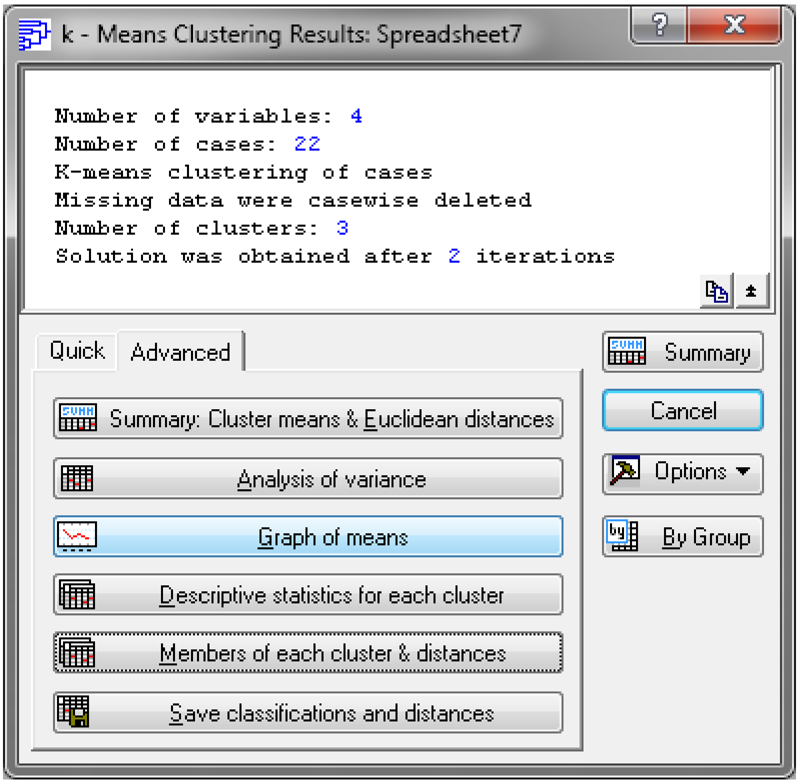
Members of Cluster Number 3 (Spreadsheet7)
and Distances from Respective Cluster Center
Cluster contains 9 cases

linkage	Distance
Acura	0,762362
Audi	0,462246
BMW	0,144429
Dodge	0,165654
Eagle	0,546452
Isuzu	0,248665
Mitsub.	0,367364
Nissan	0,356762
Saab	0,342959

Fig. 93. (the end)

Fig. 93 shows that the representative for the first cluster is the car model Olds, for the second once it is Buick, for the third cluster it is BMW. A comparative analysis of the Euclidean distances allowed us to conclude that the built-up clusterization is qualitative, as evidenced by a significant excess of distance between groups and within them.

4.3.4. To construct a graph showing the character of the breakdown of cars into clusters according to the level of insurance risk, the *Graph of means* button is used (Fig. 94).



k - Means Clustering Results: Spreadsheet7

```
Number of variables: 4
Number of cases: 22
K-means clustering of cases
Missing data were casewise deleted
Number of clusters: 3
Solution was obtained after 2 iterations
```

Quick
Advanced
Summary: Cluster means & Euclidean distances
Analysis of variance
Graph of means
Descriptive statistics for each cluster
Members of each cluster & distances
Save classifications and distances
Summary
Cancel
Options
By Group

Fig. 94. **The graphs of average values of indicators for 3 clusters**

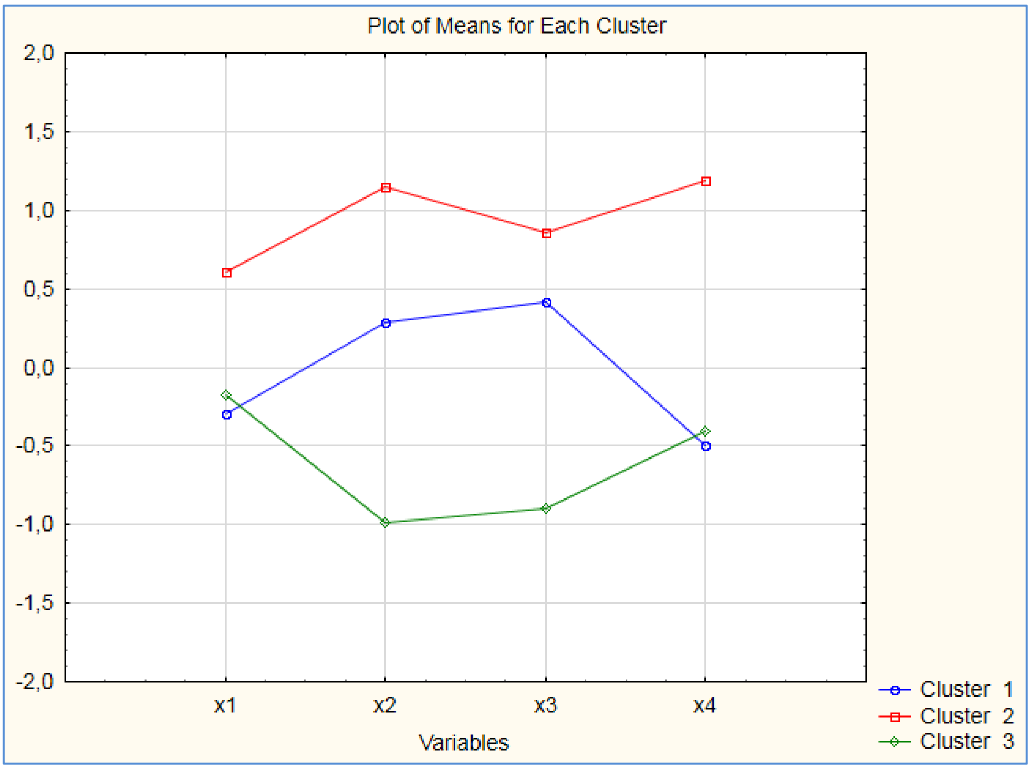


Fig. 94. (the end)

Analyzing the results we can give the recommendations presented in Table 17.

Table 17

The general characteristics of the insurance risk clusters

Cluster number	List of car models included in the cluster	Key characteristics of the class	Recommendation
The first cluster	Honda, Olds, Mercedes, VW, Pontiac, Toyota, Volvo	This group is characterized by the presence of budget models of cars owned by middle-aged drivers with sufficient driving experience. Most cars have low mileage	Maintain all indicators at the same level
The second cluster	Buick, Corvette, Chrysler, Ford, Mazda, Porsche	This group is characterized by the presence of expensive cars owned by mature drivers with significant driving experience	Maintain indicators X1 – X3 at the same level and try to decrease X4 – the car age
The third cluster	Acura, Audi, BMW, Dodge, Eagle, Isuzu, Mitsubishi, Nissan, Saab	This group is characterized by the presence of models of middle-class cars, which belong to younger drivers with little driving experience. Most cars have an average age of use	Try to increase indicator X3 and decrease indicators X1, X4.

Thus, the cars and their owners were divided into classes, each of which corresponds to a certain risk group. Observations in the same group are characterized by the same probability of an insured event, which may later become a reminder during the insurance assessment of the car.

Task 3. Based on the analysis of the dendrogram shown in Fig. 95, draw a conclusion about the number of clusters in the population under study. Justify your answer.

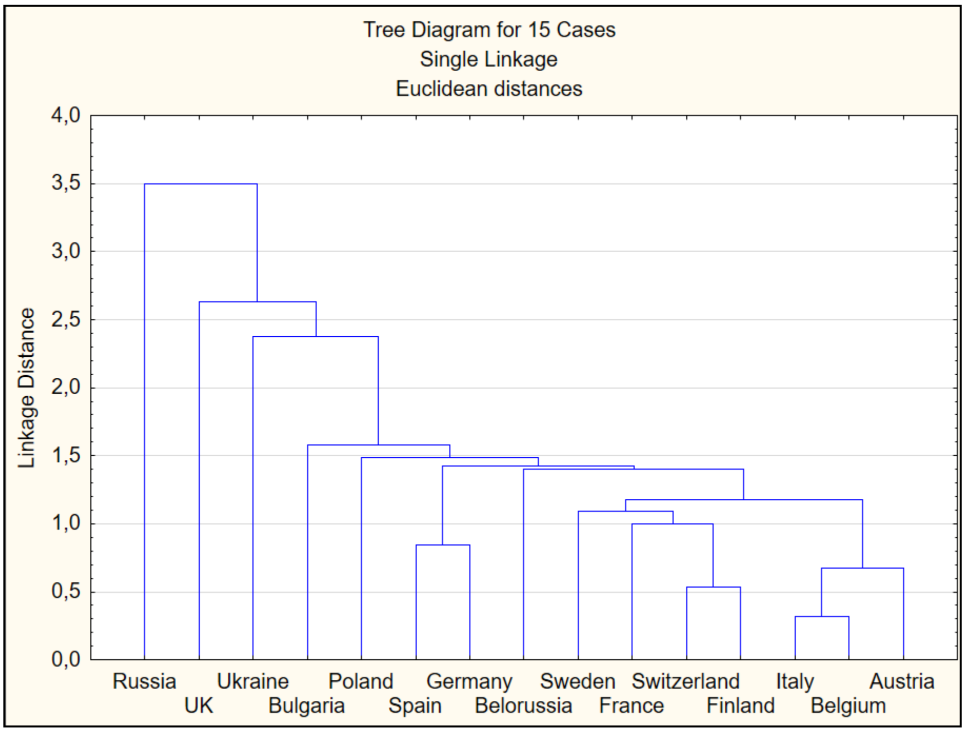


Fig. 95. **The vertical dendrogram**

Task 4. Using the data in Table 18 and the methods of hierarchical clustering, classify the countries of the world according to the level of education of the population.

Table 18

The indicators of the quality of education of the population

Countries	Level of employment	Unemployment rate	The number of people who graduated from HEI, thousand people	The number of permanent residents, people
1	2	3	4	5
Austria	71.5	6.0	422	8 662 588
Belgium	62.3	7.8	488	11 291 746
Greece	52.0	23.5	659	10 846 979

Table 18 (the end)

1	2	3	4	5
Denmark	74.9	6.2	291	5 699 220
France	64.6	10.1	2338	66 539 000
Germany	74.7	4.1	2780	81 292 400
Italy	57.2	11.7	1 872	60 685 487
Poland	64.5	6.2	1 902	38 484 000
Spain	60.5	19.6	1 959	46 423 064
Sweden	76.2	7.0	436	9 838 480
Slovakia	64.9	6.6	209	9 838 480
United Kingdom	75.3	3.9	2 380	65 572 409
Czech Republic	72	3.4	427	10 541 466
Ireland	64.7	5.4	199	4 635 400
Switzerland	79.6	3.3	279	8 306 200
Finland	69.2	7.3	309	5 496 591

Task 5. Based on the results of the research, the data were obtained on the following indicators: population (X1), GDP per capita (X2), exports of goods and services (X3), the ratio of students to teachers (number of students per teacher) (X4) and the number of doctors per 1000 people (X5) (Table 19). Classify 22 countries according to their level of development, using the methods of hierarchical and non-hierarchical clustering.

Table 19

The initial data

Counties	X1	X2	X3	X4	X5
1	2	3	4	5	6
Algeria	35 978 000	4 480.7245	61 975 371 709.730	16.8079	1.207
Angola	18 992 708	3 587.8838	51 572 818 660.867	27.2341	0.1311
Azerbaijan	8 997 586	5 842.8058	28 728 665 753.083	8.0237	3.6629
Canada	34 108 752	47 450.3185	471 736 717 163.276	18.1009	2.0384
China	1 339 724 852	4 550.4536	1 654 815 752 520.770	15.46166	1.4627
France	62 791 013	40 638.3340	707 910 169 575.170	12.68018	3.0134
Georgia	4 436 391	3 233.2959	4 034 786 511.811	7.5016	4.3073
Germany	81 802 257	41 531.9342	1 445 674 190 819.170	12.90596	3.7536
Hungary	10 014 324	13 113.5260	107 203 863 216.488	10.19559	2.8894
Iceland	317 630	43 024.9238	7 113 368 298.680	11.0971	3.5654
India	1 182 105 564	1 357.5637	375 353 472 834.938	25.32821	0.6634

Table 19 (the end)

1	2	3	4	5	6
Japan	128 056 000	44 507.6764	857 109 901 329.889	11.8603	2.2471
Kazakhstan	16 442 000	9 070.4883	65 502 334 498.320	7.5036	3.4867
Moldova	3 563 695	1 958.1337	1 941 104 508.743	10.4993	2.439
New Zealand	4 367 800	33 692.0108	44 356 174 905.218	14.50074	2.6114
Russia	142 849 472	10 674.9972	445 513 189 914.350	8.5003	2.3984
Slovakia	5 424 925	16 727.2913	69 727 530 607.620	12.03266	3.305
Switzerland	7 785 806	74 605.7745	373 420 687 857.230	9.3983	3.8053
Turkey	73 722 988	10 672.3892	157 844 709 209.476	18.1024	1.7068
Ukraine	45 782 592	2 965.1424	63 998 815 464.489	6.5301	3.483
United States	309 349 689	48 466.8234	1 846 280 000 000.000	14.3967	2.4383
Uzbekistan	28 001 400	1 634.3121	13 030 252 051.819	13.00627	2.5352

Task 6. Using the data in Table 18 and methods of non-hierarchical clustering, classify the countries of the world according to the level of education of the population.

Task 7. Based on the analysis of the dendrogram shown in Fig. 96, draw a conclusion about the number of clusters in the population under study. Justify you answer.

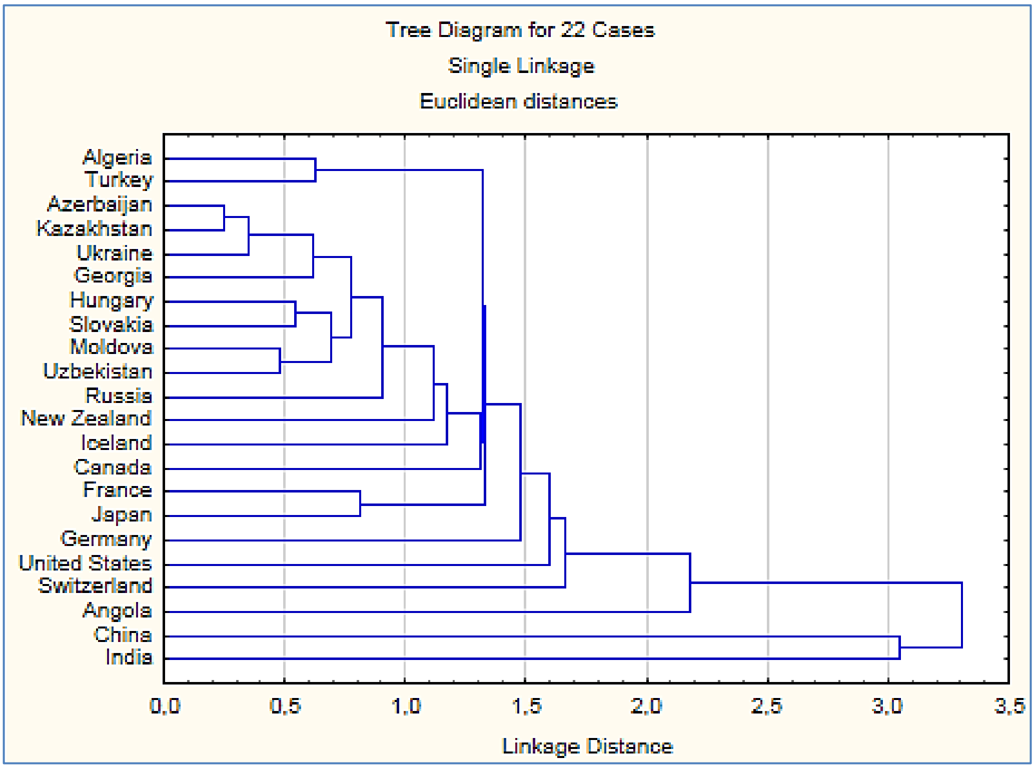


Fig. 96. **The horizontal hierarchical dendrogram**

The list of questions for independent work

1. What is the difference between clustering and classification?
2. Name the main properties of the cluster.
3. List the main uses of cluster analysis.
4. List the stages of the formation of the matrix of observations.
5. List the distances that are most often used in the multidimensional analysis.
6. What is the matrix of distances?
7. What is the difference between calculation of the Chebyshev distance and the calculation of the average absolute difference in values of signs?
8. Name the attributes of the matrix of distances.
9. List the cluster analysis methods.
10. Provide the implementation stages of the k-medium method.

Topic 7. Data recognition and discriminatory analysis

Task 1. Check the quality of clustering of the cars and their owners into insurance risk groups (see Table 17). The initial data are presented in Table 16.

Guidelines

Normalized initial data about 22 car models, which were distributed in tree groups by the method of cluster analysis (in terms of insurance risk), are shown in Fig. 97.

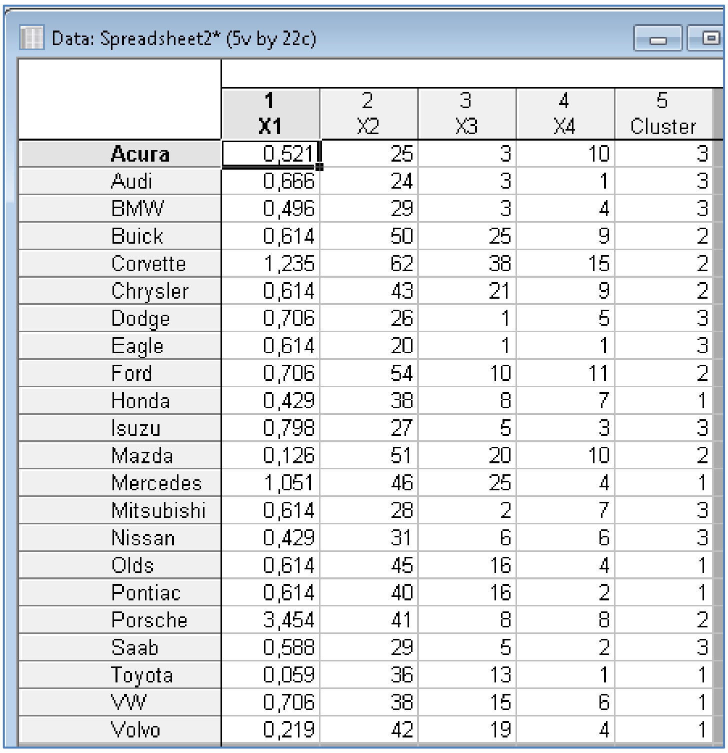
Discriminant analysis is a multidimensional statistical method that allows you to study the differences between two or more groups of objects in several variables at a time. The main task of discriminant analysis is to study group differences, that is, to discriminate objects based on certain attributes.

With the help of discriminant analysis, two types of problems are solved:

1. Search for a function according to which the object belongs to one of the known classes.
2. Classification of new objects according to the found rules.

Let's consider an example of using discriminant analysis to solve the problem of object recognition.

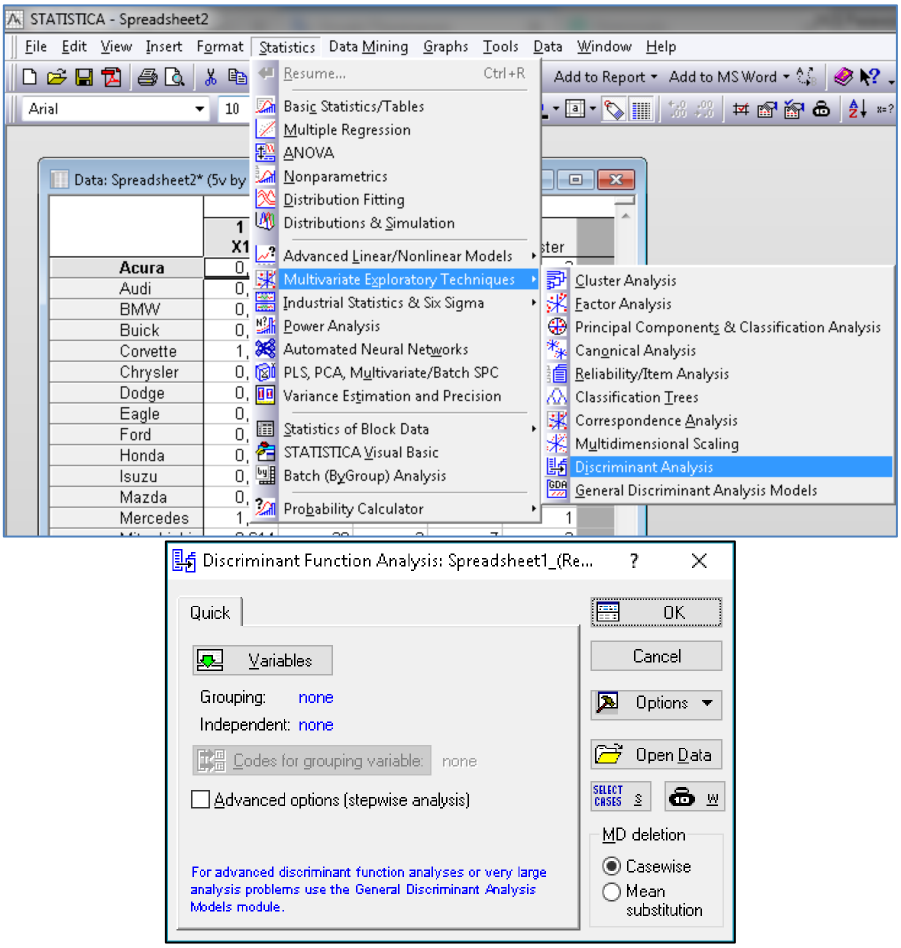
The choice of the module *Discriminant analysis* is possible through the *Statistics / Multivariate Exploratory Techniques / Discriminant analysis* menu (Fig. 98).



Data: Spreadsheet2* (5v by 22c)

	1 X1	2 X2	3 X3	4 X4	5 Cluster
Acura	0,521	25	3	10	3
Audi	0,666	24	3	1	3
BMW	0,496	29	3	4	3
Buick	0,614	50	25	9	2
Corvette	1,235	62	38	15	2
Chrysler	0,614	43	21	9	2
Dodge	0,706	26	1	5	3
Eagle	0,614	20	1	1	3
Ford	0,706	54	10	11	2
Honda	0,429	38	8	7	1
Isuzu	0,798	27	5	3	3
Mazda	0,126	51	20	10	2
Mercedes	1,051	46	25	4	1
Mitsubishi	0,614	28	2	7	3
Nissan	0,429	31	6	6	3
Olds	0,614	45	16	4	1
Pontiac	0,614	40	16	2	1
Porsche	3,454	41	8	8	2
Saab	0,588	29	5	2	3
Toyota	0,059	36	13	1	1
VW	0,706	38	15	6	1
Volvo	0,219	42	19	4	1

Fig. 97. **The initial data with cluster distribution**


STATISTICA - Spreadsheet2
File Edit View Insert Format Statistics Data Mining Graphs Tools Data Window Help
Resume... Ctrl+R
Basic Statistics/Tables
Multiple Regression
ANOVA
Nonparametrics
Distribution Fitting
Distributions & Simulation
Advanced Linear/Nonlinear Models
Multivariate Exploratory Techniques
Industrial Statistics & Six Sigma
Power Analysis
Automated Neural Networks
PLS, PCA, Multivariate/Batch SPC
Variance Estimation and Precision
Statistics of Block Data
STATISTICA Visual Basic
Batch (ByGroup) Analysis
Probability Calculator
Cluster Analysis
Factor Analysis
Principal Components & Classification Analysis
Canonical Analysis
Reliability/Item Analysis
Classification Trees
Correspondence Analysis
Multidimensional Scaling
Discriminant Analysis
General Discriminant Analysis Models
Add to Report Add to MS Word
Arial 10
Data: Spreadsheet2* (5v by
Discriminant Function Analysis: Spreadsheet1_(Re...
Quick
OK
Cancel
Options
Open Data
SELECT CASES S
W
Variables

Grouping: none
Independent: none
Codes for grouping variable: none
☐ Advanced options (stepwise analysis)

MD deletion
● Casewise
○ Mean substitution

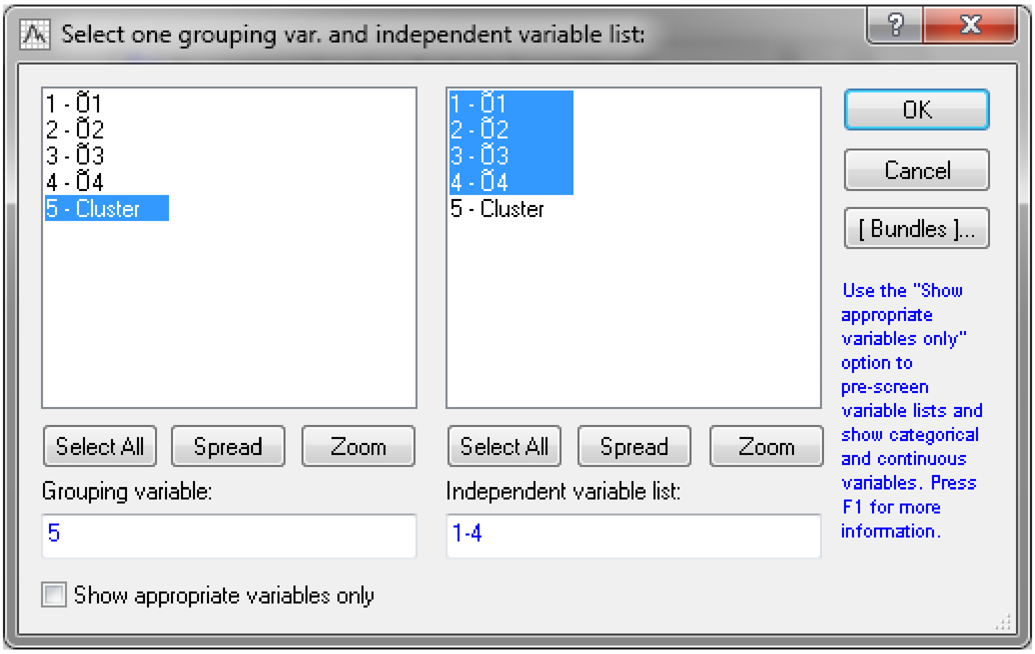
For advanced discriminant function analyses or very large analysis problems use the General Discriminant Analysis Models module.

Fig. 98. **Launching the discriminant analysis in Statistica 10.0**

The screen for the *Discriminant Function Analysis* module will appear, with the help of which you can perform the following functions:

- open the data file using the *Open Data* button;
- select a variable – *Variables*;
- determine the number of groups of objects being analyzed – *Codes for grouping variable*;
- permanently remove variables from the *Casewise* list or replace them with the average *Mean substitution*;
- specify the conditions for selecting observations from the database – *Select Cases*;
- make weight of variables by selecting them from the list – *W*.

You can use the *Variables* button to select a *Grouping* and an *Independent variable* (Fig. 99).



Select one grouping var. and independent variable list:

1 - Ŏ1
2 - Ŏ2
3 - Ŏ3
4 - Ŏ4
5 - Cluster

1 - Ŏ1
2 - Ŏ2
3 - Ŏ3
4 - Ŏ4
5 - Cluster

Grouping variable: 5

Independent variable list: 1-4

□ Show appropriate variables only

Use the "Show appropriate variables only" option to pre-screen variable lists and show categorical and continuous variables. Press F1 for more information.

Fig. 99. **Selection of variables**

Using the buttons located in the *Variables* selection panel it is possible:

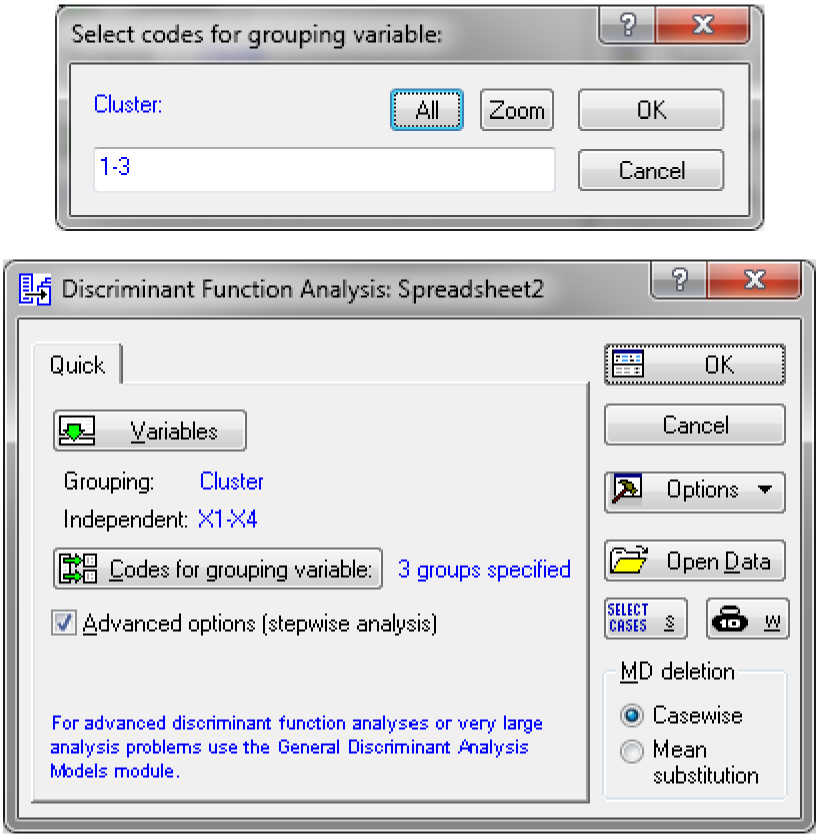
1) to select all variables – *Select All*;

2) to view the type of name – *Spread*;

3) to see additional information on the *Zoom variable*.

4) to define the model by clicking the *OK* button.

The *Model Definition* dialog box that is used to select a model is shown in Fig. 100.



Select codes for grouping variable:

Cluster: 1-3

Discriminant Function Analysis: Spreadsheet2

Grouping: Cluster

Independent: X1-X4

Codes for grouping variable: 3 groups specified

☑ Advanced options (stepwise analysis)

For advanced discriminant function analyses or very large analysis problems use the General Discriminant Analysis Models module.

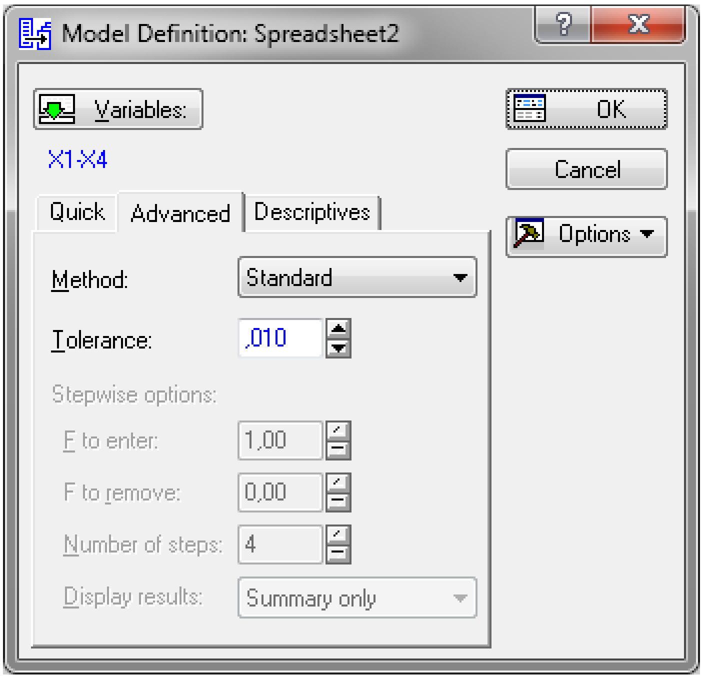
MD deletion

● Casewise

○ Mean substitution

Fig. 100. **Choosing the parameters of the discriminant analysis**

On the *Advanced* tab, you can specify the method that will be used to select meaningful variables (Fig. 101).



Model Definition: Spreadsheet2

Variables: X1-X4

Method: Standard

Tolerance: ,010

Stepwise options:

F to enter: 1,00

F to remove: 0,00

Number of steps: 4

Display results: Summary only

Fig. 101. **Choosing the method of discriminant analysis**

The following methods may be used:

- *Standard*. All variables are included in the model at the same time;
- *Forward stepwise*. At each step in the model, a variable with a maximum F value is selected. The procedure ends when all variables whose values F are greater than the values specified in the *F to enter* field are included in the model;
- *Backward stepwise* (step-by-step). At each step, all variables are selected in the model, which are then deleted depending on the value of F. The steps end when there are no variables with F values less than those specified by the user in the *F to remove* field.

The *Number of steps* field determines the maximum number of analysis steps that the procedure ends in.

The *Tolerance* field allows you to exclude non-informative variables from the model. If the tolerance is less than the value of 0.01, the variable is considered non-informative and not included in the model.

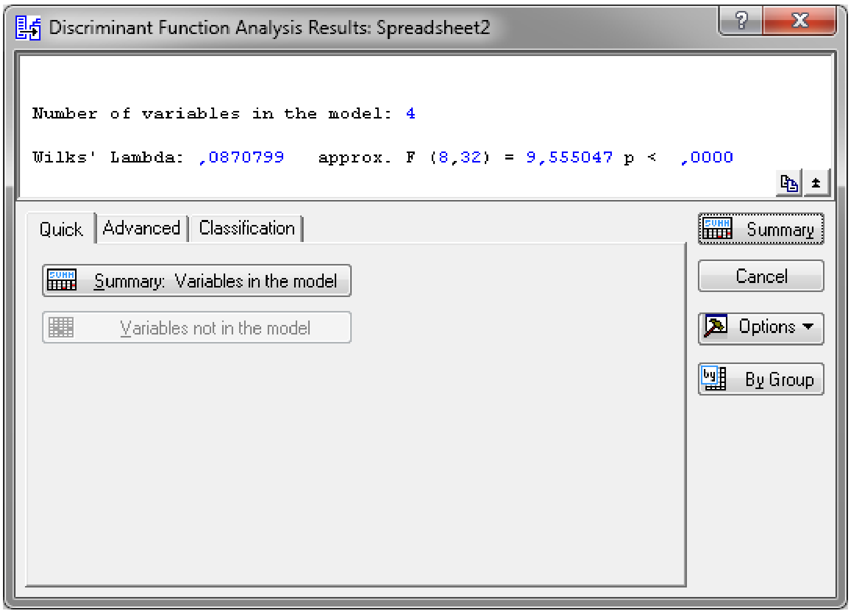
As a method of analysis, choose *Standard*. According to the results obtained during the calculations presented in the window *Discriminant Function Analysis Results* (Fig. 102), it is possible to obtain the following information:

the number of variables in the model: 4;

the value of Wilks' lambda: 0.0870799;

the approximate value of F-statistics, which is related to Wilks' lambda (Approx. F (8; 32)) = 9.555047;

the significance level of F-criterion p < 0.0000 for the obtained value of 9.555047.



Discriminant Function Analysis Results: Spreadsheet2
Quick
Advanced
Classification
Summary: Variables in the model
Variables not in the model
Summary
Cancel
Options
By Group

```
Number of variables in the model: 4

Wilks' Lambda: ,0870799   approx. F (8,32) = 9,555047 p <  ,0000
```

Fig. 102. **The *Discriminant function analysis results* window**

The criteria for assessing the quality of classification are based on the calculation of possible transitions of the analyzed objects from one class to another. The criteria are the following:

1) λ-statistics of Wilks (Wilks' lambda);

2) the significance of F-statistics ($F_{table} < F$);

3) the Mahalanobis distance;

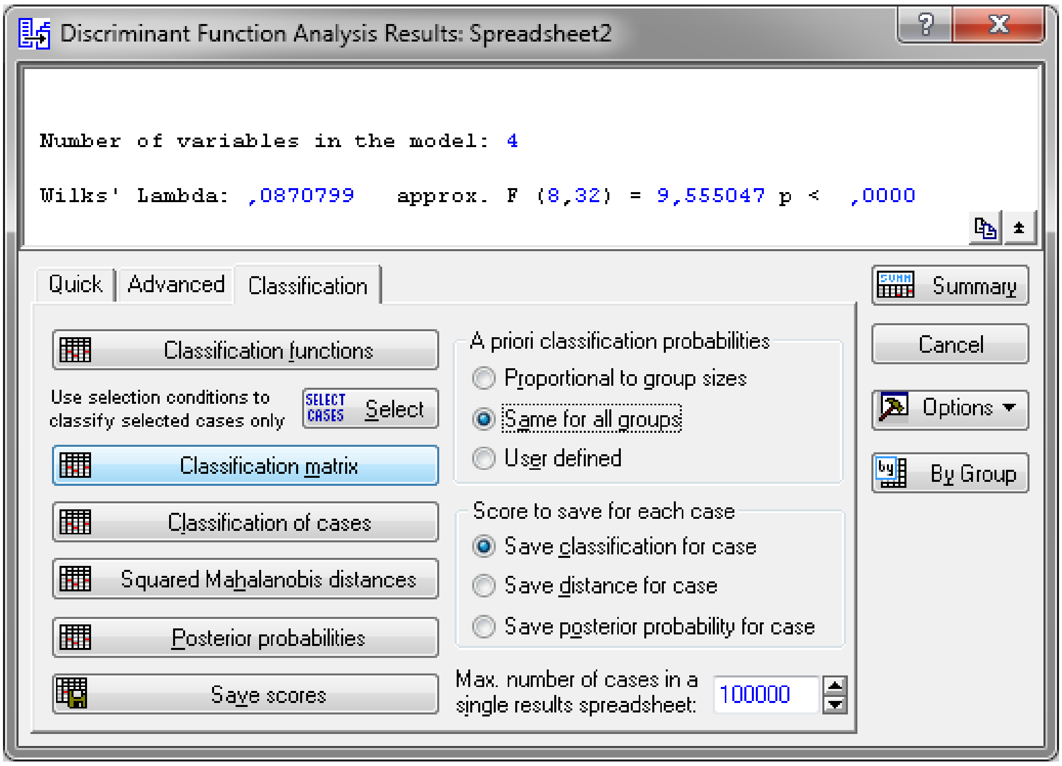
4) criterion x^2.

Values of Wilks' statistics (Wilks' lambda) are in the range [0; 1]. If Wilks' statistics is close to 0, this indicates good discrimination, while values close to 1 indicate bad discrimination of the studied objects.

Thus, according to Wilks' lambda which is 0.087701 it is possible to conclude that the classification is correct.

Also, the significance of F-statistics confirms the existence of differences between groups. In our case, F_{table} ($\alpha = 0.05$, $k1 = 4$, $k2 = 22$) = 2.82. So, $F_{table} < F$ ($2.82 < 9.555047$). Thus, the classification of the car models according to the insurance risk level is correct.

As a validation check, let's see the results of the classification matrix by clicking the *Classification matrix* button (Fig. 103), pre-selecting *Same for all groups* in the right-hand window of *Discriminant Function Analysis Results* (Fig. 104).



Discriminant Function Analysis Results: Spreadsheet2

```
Number of variables in the model: 4

Wilks' Lambda: ,0870799   approx. F (8,32) = 9,555047 p <  ,0000
```

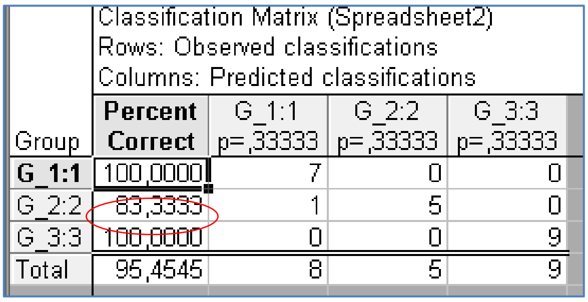
A priori classification probabilities
○ Proportional to group sizes
● Same for all groups
○ User defined

Score to save for each case
● Save classification for case
○ Save distance for case
○ Save posterior probability for case

Use selection conditions to classify selected cases only

Max. number of cases in a single results spreadsheet: 100000

Fig. 103. **The *Classification matrix* button**

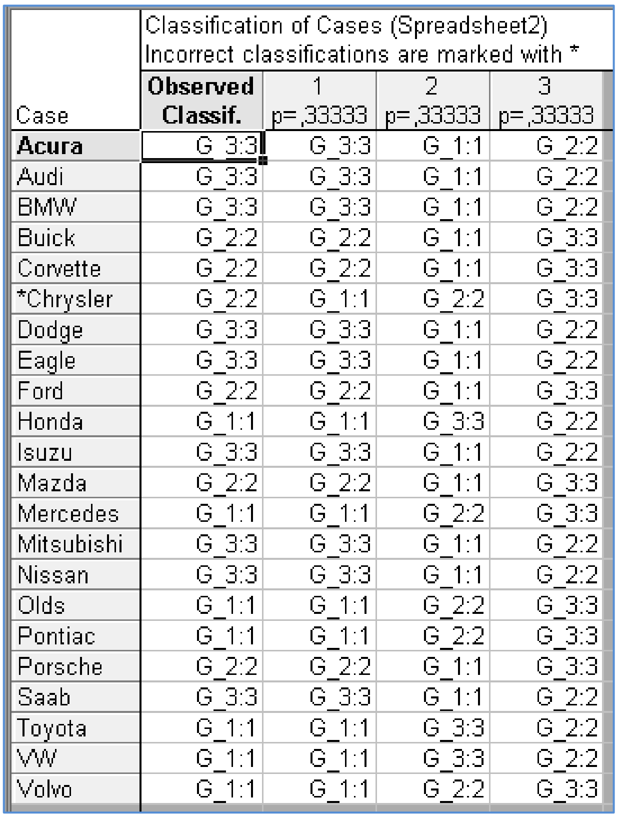


Classification Matrix (Spreadsheet2)
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	G_1:1 p=,33333	G_2:2 p=,33333	G_3:3 p=,33333
G_1:1	100,0000	7	0	0
G_2:2	83,3333	1	5	0
G_3:3	100,0000	0	0	9
Total	95,4545	8	5	9

Fig. 104. **The classification matrix results**

By the results of the classification matrix, we can conclude that not all objects are correctly broken down into four groups by cluster analysis. If there are car models that are incorrectly assigned to the appropriate groups, see the classification of cases (Fig. 105).



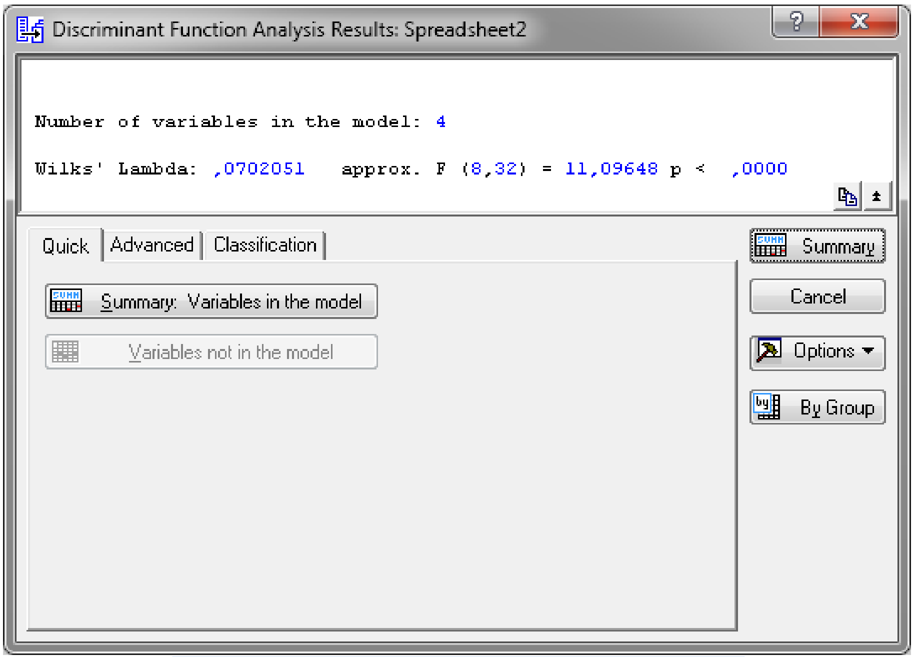
Classification of Cases (Spreadsheet2)
Incorrect classifications are marked with *

Case	Observed Classif.	1 p=,33333	2 p=,33333	3 p=,33333
Acura	G_3:3	G_3:3	G_1:1	G_2:2
Audi	G_3:3	G_3:3	G_1:1	G_2:2
BMW	G_3:3	G_3:3	G_1:1	G_2:2
Buick	G_2:2	G_2:2	G_1:1	G_3:3
Corvette	G_2:2	G_2:2	G_1:1	G_3:3
*Chrysler	G_2:2	G_1:1	G_2:2	G_3:3
Dodge	G_3:3	G_3:3	G_1:1	G_2:2
Eagle	G_3:3	G_3:3	G_1:1	G_2:2
Ford	G_2:2	G_2:2	G_1:1	G_3:3
Honda	G_1:1	G_1:1	G_3:3	G_2:2
Isuzu	G_3:3	G_3:3	G_1:1	G_2:2
Mazda	G_2:2	G_2:2	G_1:1	G_3:3
Mercedes	G_1:1	G_1:1	G_2:2	G_3:3
Mitsubishi	G_3:3	G_3:3	G_1:1	G_2:2
Nissan	G_3:3	G_3:3	G_1:1	G_2:2
Olds	G_1:1	G_1:1	G_2:2	G_3:3
Pontiac	G_1:1	G_1:1	G_2:2	G_3:3
Porsche	G_2:2	G_2:2	G_1:1	G_3:3
Saab	G_3:3	G_3:3	G_1:1	G_2:2
Toyota	G_1:1	G_1:1	G_3:3	G_2:2
VW	G_1:1	G_1:1	G_3:3	G_2:2
Volvo	G_1:1	G_1:1	G_2:2	G_3:3

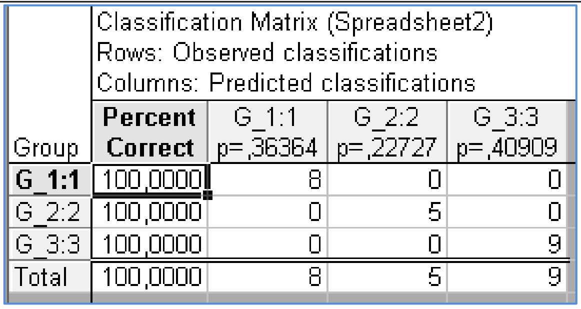
Fig. 105. **The classification of cases**

In Fig. 105, incorrectly assigned objects are marked with an asterisk (*). So, in our case only one object is not in its cluster (group) – Chrysler. According to the obtained results (see Figs. 104, 105), this car model must be transferred to the 1st cluster.

After transferring Chrysler to the 1st cluster and re-performing the discriminant analysis, we obtain the following result (Fig. 106).

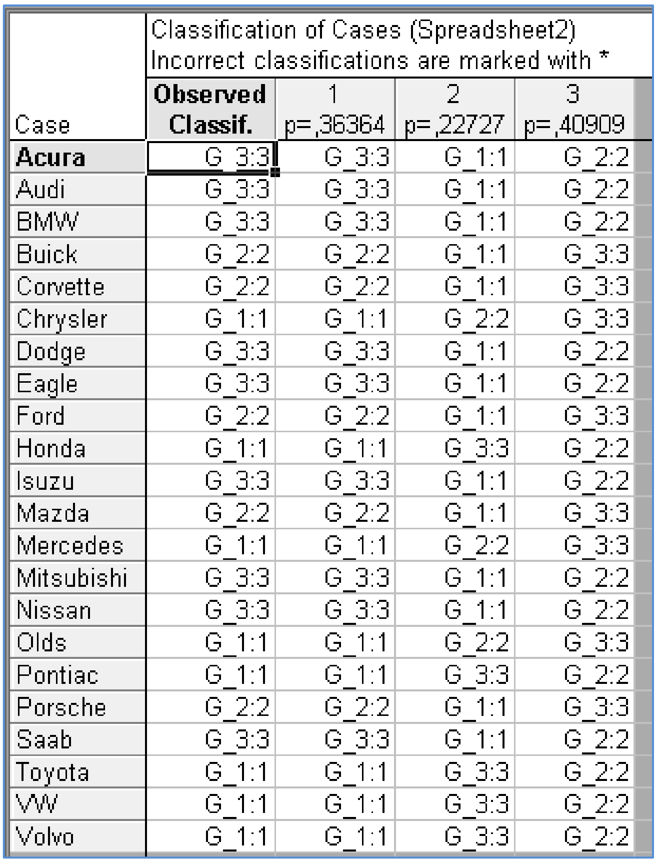

Discriminant Function Analysis Results: Spreadsheet2
Quick
Advanced
Classification
Summary: Variables in the model
Variables not in the model
Summary
Cancel
Options
By Group

```
Number of variables in the model: 4
Wilks' Lambda: ,0702051   approx. F (8,32) = 11,09648 p < ,0000
```



Classification Matrix (Spreadsheet2)
Rows: Observed classifications
Columns: Predicted classifications

Group	Percent Correct	G_1:1 p=,36364	G_2:2 p=,22727	G_3:3 p=,40909
G_1:1	100,0000	8	0	0
G_2:2	100,0000	0	5	0
G_3:3	100,0000	0	0	9
Total	100,0000	8	5	9



Classification of Cases (Spreadsheet2)
Incorrect classifications are marked with *

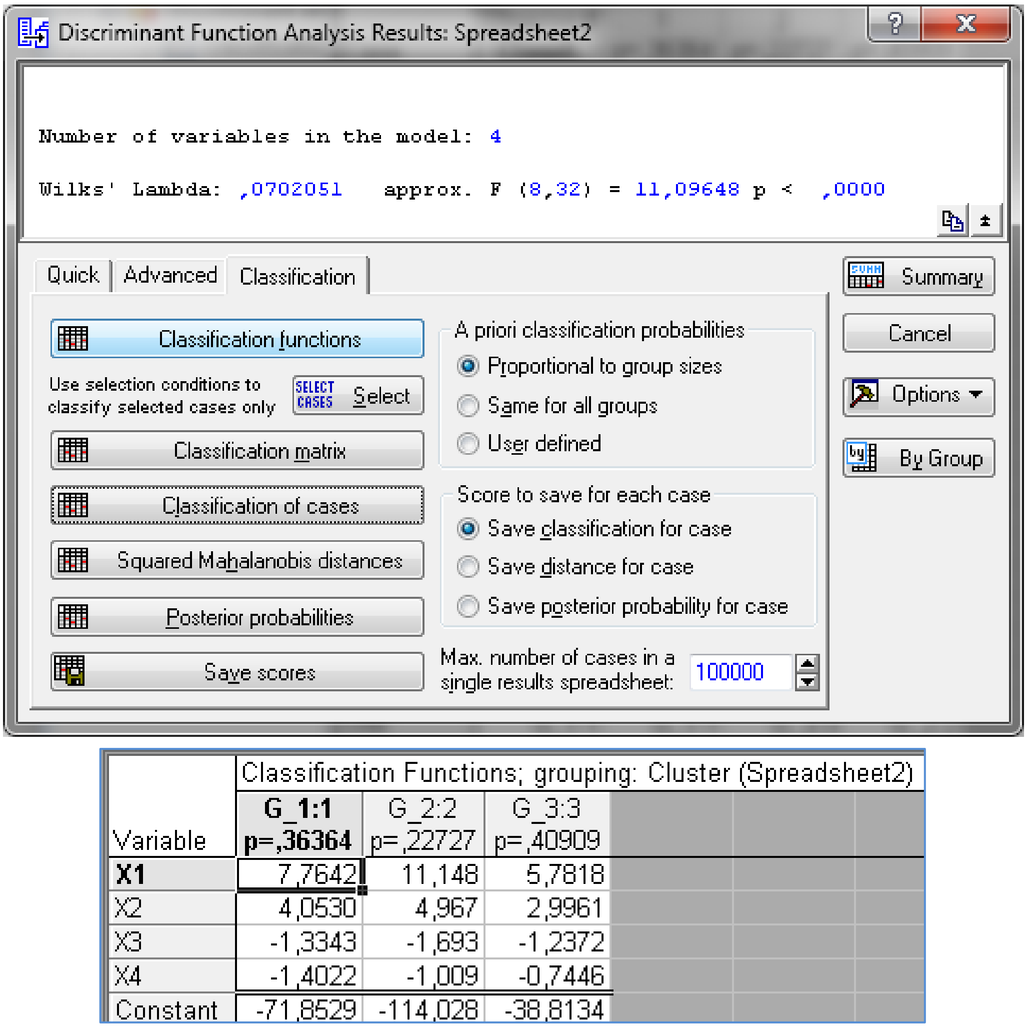
Case	Observed Classif.	1 p=,36364	2 p=,22727	3 p=,40909
Acura	G_3:3	G_3:3	G_1:1	G_2:2
Audi	G_3:3	G_3:3	G_1:1	G_2:2
BMW	G_3:3	G_3:3	G_1:1	G_2:2
Buick	G_2:2	G_2:2	G_1:1	G_3:3
Corvette	G_2:2	G_2:2	G_1:1	G_3:3
Chrysler	G_1:1	G_1:1	G_2:2	G_3:3
Dodge	G_3:3	G_3:3	G_1:1	G_2:2
Eagle	G_3:3	G_3:3	G_1:1	G_2:2
Ford	G_2:2	G_2:2	G_1:1	G_3:3
Honda	G_1:1	G_1:1	G_3:3	G_2:2
Isuzu	G_3:3	G_3:3	G_1:1	G_2:2
Mazda	G_2:2	G_2:2	G_1:1	G_3:3
Mercedes	G_1:1	G_1:1	G_2:2	G_3:3
Mitsubishi	G_3:3	G_3:3	G_1:1	G_2:2
Nissan	G_3:3	G_3:3	G_1:1	G_2:2
Olds	G_1:1	G_1:1	G_2:2	G_3:3
Pontiac	G_1:1	G_1:1	G_3:3	G_2:2
Porsche	G_2:2	G_2:2	G_1:1	G_3:3
Saab	G_3:3	G_3:3	G_1:1	G_2:2
Toyota	G_1:1	G_1:1	G_3:3	G_2:2
VW	G_1:1	G_1:1	G_3:3	G_2:2
Volvo	G_1:1	G_1:1	G_3:3	G_2:2

Fig. 106. **New results of the discriminant analysis**

Thus, the task of getting the correct groups is complete.

A discriminant function is a linear combination of a certain set of features which are called classification features and on the basis of which classes of objects that are homogeneous in some properties are identified.

To do this, in the *Discriminant Function Analysis Results* window, click *Classification functions* (Fig. 107); a window will appear, from which it is possible to write the classification functions for each class.



Discriminant Function Analysis Results: Spreadsheet2

```
Number of variables in the model: 4

Wilks' Lambda: ,0702051    approx. F (8,32) = 11,09648 p <  ,0000
```

Use selection conditions to classify selected cases only

A priori classification probabilities
- ● Proportional to group sizes
- ○ Same for all groups
- ○ User defined

Score to save for each case
- ● Save classification for case
- ○ Save distance for case
- ○ Save posterior probability for case

Max. number of cases in a single results spreadsheet: 100000

Classification Functions; grouping: Cluster (Spreadsheet2)

Variable	G_1:1 p=,36364	G_2:2 p=,22727	G_3:3 p=,40909
X1	7,7642	11,148	5,7818
X2	4,0530	4,967	2,9961
X3	-1,3343	-1,693	-1,2372
X4	-1,4022	-1,009	-0,7446
Constant	-71,8529	-114,028	-38,8134

Fig. 107. **The classification functions**

Based on the results in Fig. 107, let's build classification functions:

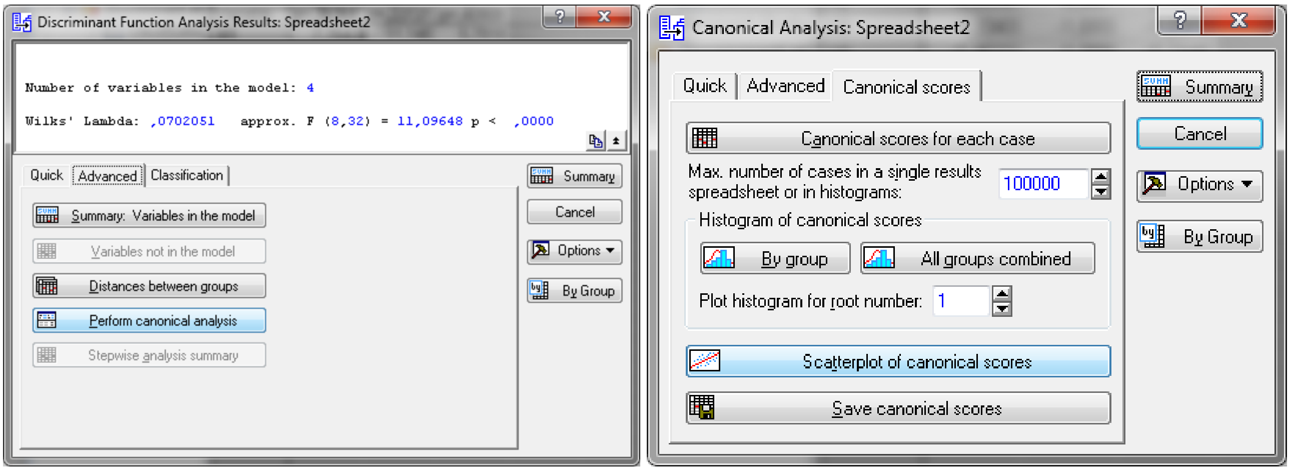
car models with low level of the insurance risk = 7.7642X1 + 4.0530X2 – – 1.3343X3 – 1.4022X4;

car models with high level of the insurance risk = 11.148X1 + 4.967X2 – – 1.693X3 – 1.009X4;

car models with middle level of the insurance risk = 5.7818X1 + + 2.9961X2 – 1.2373X3 – 0.7446X4.

You have obtained the coefficients for each variable and each discriminant function. They can also be interpreted in the usual way: the higher the standardized coefficient, the greater the contribution of the corresponding variable to the discrimination of the population.

For more detailed information, it is possible to review the results of a canonical analysis that can be performed if at least three groups have been selected and at least two variables in the model are selected by clicking the *Perform canonical analysis* button (Fig. 108).


Discriminant Function Analysis Results: Spreadsheet2
Quick
Advanced
Classification
Summary
Cancel
Options
By Group
Summary: Variables in the model
Variables not in the model
Distances between groups
Perform canonical analysis
Stepwise analysis summary
Canonical Analysis: Spreadsheet2
Quick
Advanced
Canonical scores
Summary
Cancel
Options
By Group
Canonical scores for each case
Histogram of canonical scores
By group
All groups combined
Scatterplot of canonical scores
Save canonical scores

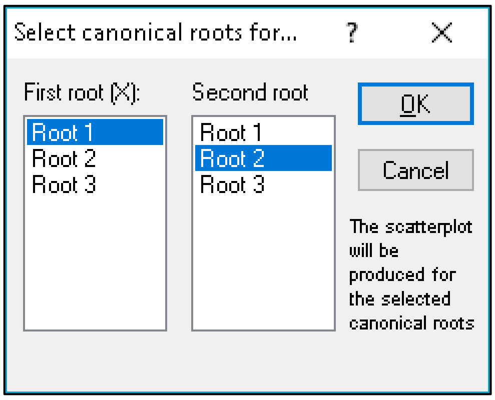
```
Number of variables in the model: 4
Wilks' Lambda: ,0702051   approx. F (8,32) = 11,09648 p < ,0000
```

Max. number of cases in a single results spreadsheet or in histograms: 100000

Plot histogram for root number: 1

Fig. 108. **Choosing the canonical analysis**

A canonical analysis window appears in which the *Scatterplot of canonical scores* option is possible to construct the next scatter plot for values. With this diagram, it is possible to determine the contribution that each discriminating function makes to the distribution between the groups (Fig. 109).



The scatterplot will be produced for the selected canonical roots

Fig. 109. **Selection the parameters of the canonical analysis**

The graph of scattering of canonical values for canonical roots is given in Fig. 110.

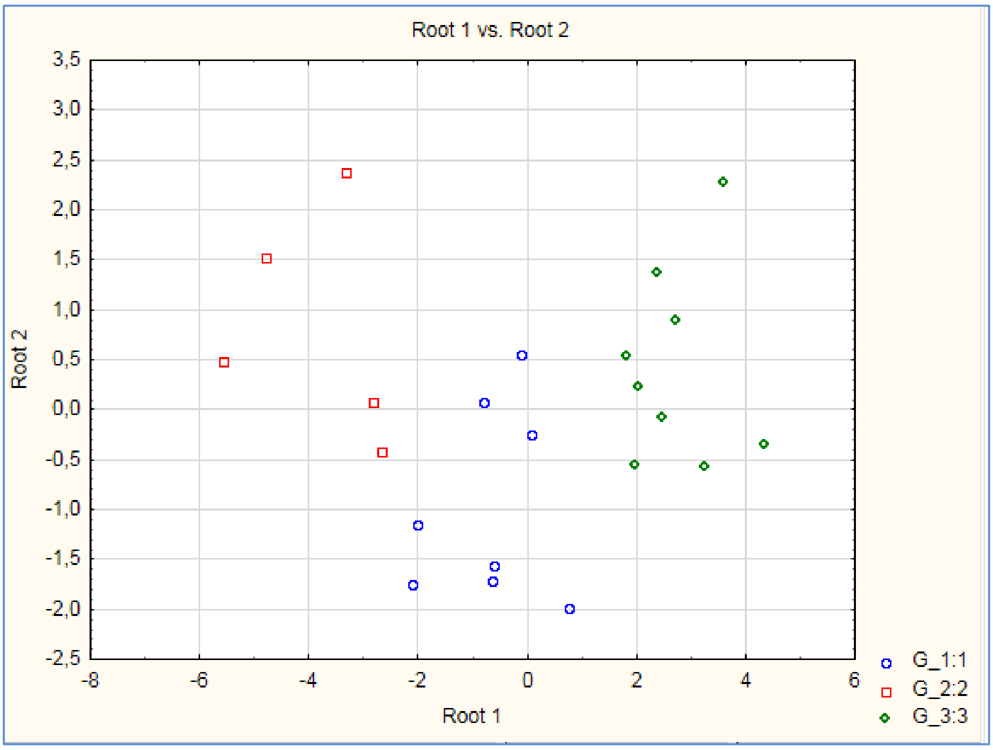
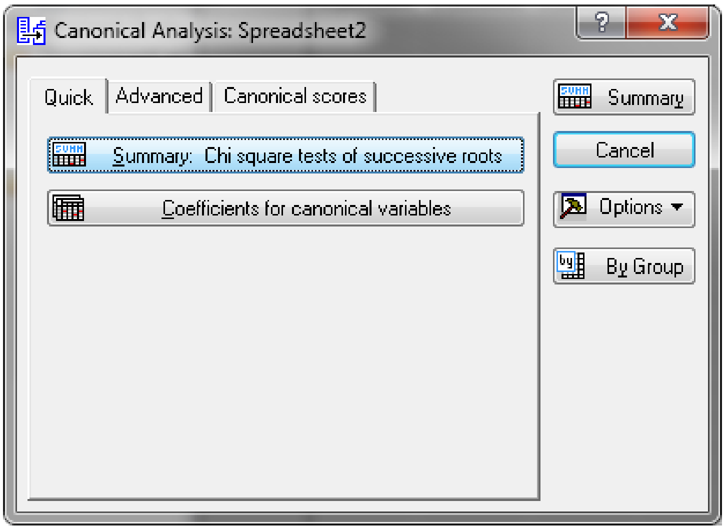
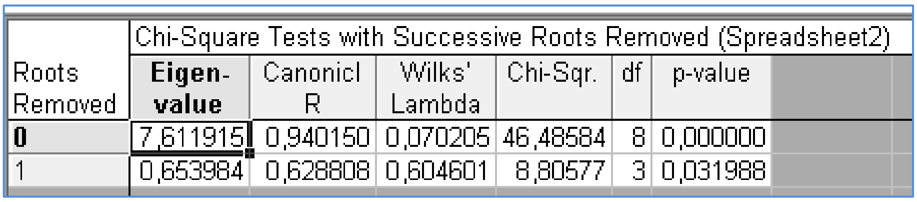


Fig. 110. **Visualization of the canonical analysis**

Also, determine whether the constructed discriminant functions are statistically significant. To do this, click on the *Chi-square criteria* for remote roots from the results window (Fig. 111).


Canonical Analysis: Spreadsheet2
Quick
Advanced
Canonical scores
Summary: Chi square tests of successive roots
Coefficients for canonical variables
Summary
Cancel
Options
By Group



	Chi-Square Tests with Successive Roots Removed (Spreadsheet2)					
Roots Removed	**Eigen-value**	Canonicl R	Wilks' Lambda	Chi-Sqr.	df	p-value
0	7,611915	0,940150	0,070205	46,48584	8	0,000000
1	0,653984	0,628808	0,604601	8,80577	3	0,031988

Fig. 111. **Checking the statistical significance of the discriminant function**

In this example, the discriminant functions are statistically significant.

Conclusion. The classification of car models according to the level of insurance risk by the method of cluster analysis is adequate and correct. In the course of discriminant analysis, some functions were obtained that can be used in the future to assign a certain (new or not participating in the analysis) model of car to one of the obtained classes (clusters, groups).

Task 2. According to the results of the cluster analysis, 15 countries of the world were divided according to the level of energy security into 4 clusters (groups) based on the following indicators:

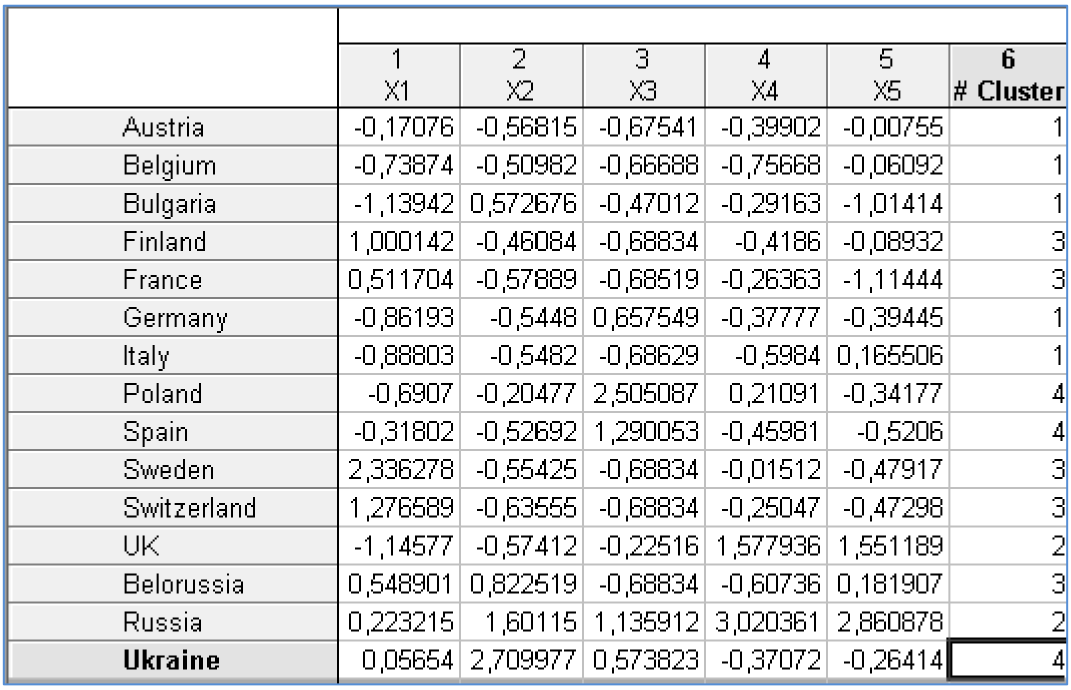
1) the share of their own sources in the balance of fuel and energy resources of the state, % (X1);

2) the share of the dominant fuel resource in the consumption of fuel and energy resources, % (X2);

3) energy intensity of GDP, kg of conventional fuel / UAH (X3);

4) the volume of coal production, million tons (X4);

5) the degree of provision of fuel and energy resources (X5). The initial data are presented in Fig. 112.



	1 X1	2 X2	3 X3	4 X4	5 X5	**6** **# Cluster**
Austria	-0,17076	-0,56815	-0,67541	-0,39902	-0,00755	1
Belgium	-0,73874	-0,50982	-0,66688	-0,75668	-0,06092	1
Bulgaria	-1,13942	0,572676	-0,47012	-0,29163	-1,01414	1
Finland	1,000142	-0,46084	-0,68834	-0,4186	-0,08932	3
France	0,511704	-0,57889	-0,68519	-0,26363	-1,11444	3
Germany	-0,86193	-0,5448	0,657549	-0,37777	-0,39445	1
Italy	-0,88803	-0,5482	-0,68629	-0,5984	0,165506	1
Poland	-0,6907	-0,20477	2,505087	0,21091	-0,34177	4
Spain	-0,31802	-0,52692	1,290053	-0,45981	-0,5206	4
Sweden	2,336278	-0,55425	-0,68834	-0,01512	-0,47917	3
Switzerland	1,276589	-0,63555	-0,68834	-0,25047	-0,47298	3
UK	-1,14577	-0,57412	-0,22516	1,577936	1,551189	2
Belorussia	0,548901	0,822519	-0,68834	-0,60736	0,181907	3
Russia	0,223215	1,60115	1,135912	3,020361	2,860878	2
Ukraine	0,05654	2,709977	0,573823	-0,37072	-0,26414	4

Fig. 112. **The results of cluster analysis**

Under the condition of the task, it is necessary to assign two countries: Moldova and the Czech Republic to already known groups received in the previous task. The data for classification are given in Table 20.

Table 20

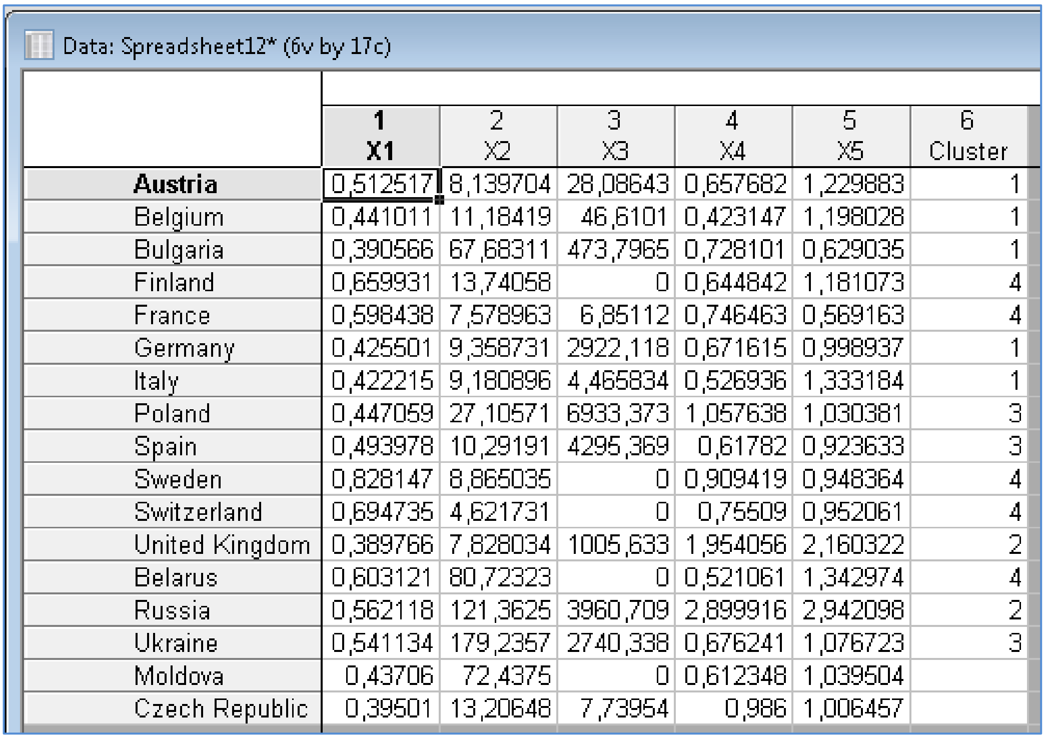
The input data

Countries	X1	X2	X3	X4	X5
Moldova	0.43706	72.4375	0	0.612348	1.039504
Czech Republic	0.39501	13.20648	7.73954	0.98600	1.006457

Guidelines

Now, let's look at an example of the use of discriminant analysis to solve the problem of classification of new objects according to the found rules.

First, we need to enter new data into the already created table in Statistica 10.0 (Fig. 113).



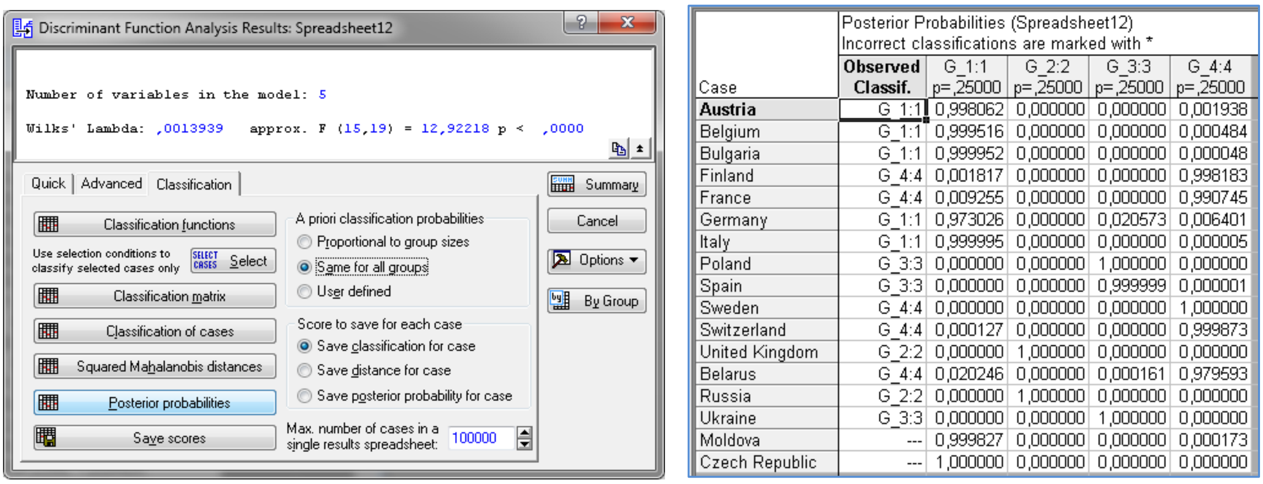
Data: Spreadsheet12* (6v by 17c)

	1 X1	2 X2	3 X3	4 X4	5 X5	6 Cluster
Austria	0,512517	8,139704	28,08643	0,657682	1,229883	1
Belgium	0,441011	11,18419	46,6101	0,423147	1,198028	1
Bulgaria	0,390566	67,68311	473,7965	0,728101	0,629035	1
Finland	0,659931	13,74058	0	0,644842	1,181073	4
France	0,598438	7,578963	6,85112	0,746463	0,569163	4
Germany	0,425501	9,358731	2922,118	0,671615	0,998937	1
Italy	0,422215	9,180896	4,465834	0,526936	1,333184	1
Poland	0,447059	27,10571	6933,373	1,057638	1,030381	3
Spain	0,493978	10,29191	4295,369	0,61782	0,923633	3
Sweden	0,828147	8,865035	0	0,909419	0,948364	4
Switzerland	0,694735	4,621731	0	0,75509	0,952061	4
United Kingdom	0,389766	7,828034	1005,633	1,954056	2,160322	2
Belarus	0,603121	80,72323	0	0,521061	1,342974	4
Russia	0,562118	121,3625	3960,709	2,899916	2,942098	2
Ukraine	0,541134	179,2357	2740,338	0,676241	1,076723	3
Moldova	0,43706	72,4375	0	0,612348	1,039504	
Czech Republic	0,39501	13,20648	7,73954	0,986	1,006457	

Fig. 113. **The initial data**

Now, we first perform a discriminant analysis for all countries, and in the results window, calculate the a posteriori probabilities, the values of which are given in Fig. 114.

The next step is to calculate the distances from the new cases to the group centers. The results of the calculations are shown in Fig. 115. To do this, select the *Squared Mahalanobis distances* button (the distance of Mahalanobis determines the affiliation of a variable to a particular class).



Discriminant Function Analysis Results: Spreadsheet12

```
Number of variables in the model: 5

Wilks' Lambda: ,0013939   approx. F (15,19) = 12,92218 p < ,0000
```

A priori classification probabilities
○ Proportional to group sizes
● Same for all groups
○ User defined

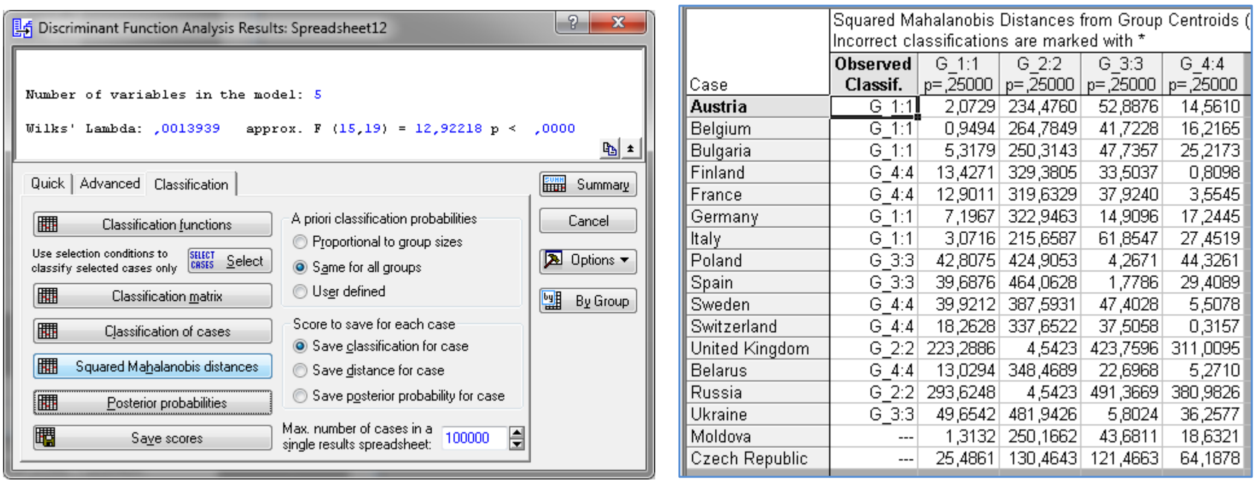
Score to save for each case
● Save classification for case
○ Save distance for case
○ Save posterior probability for case

Max. number of cases in a single results spreadsheet: 100000

Posterior Probabilities (Spreadsheet12)
Incorrect classifications are marked with *

Case	Observed Classif.	G_1:1 p=,25000	G_2:2 p=,25000	G_3:3 p=,25000	G_4:4 p=,25000
Austria	G_1:1	0,998062	0,000000	0,000000	0,001938
Belgium	G_1:1	0,999516	0,000000	0,000000	0,000484
Bulgaria	G_1:1	0,999952	0,000000	0,000000	0,000048
Finland	G_4:4	0,001817	0,000000	0,000000	0,998183
France	G_4:4	0,009255	0,000000	0,000000	0,990745
Germany	G_1:1	0,973026	0,000000	0,020573	0,006401
Italy	G_1:1	0,999995	0,000000	0,000000	0,000005
Poland	G_3:3	0,000000	0,000000	1,000000	0,000000
Spain	G_3:3	0,000000	0,000000	0,999999	0,000001
Sweden	G_4:4	0,000000	0,000000	0,000000	1,000000
Switzerland	G_4:4	0,000127	0,000000	0,000000	0,999873
United Kingdom	G_2:2	0,000000	1,000000	0,000000	0,000000
Belarus	G_4:4	0,020246	0,000000	0,000161	0,979593
Russia	G_2:2	0,000000	1,000000	0,000000	0,000000
Ukraine	G_3:3	0,000000	0,000000	1,000000	0,000000
Moldova	---	0,999827	0,000000	0,000000	0,000173
Czech Republic	---	1,000000	0,000000	0,000000	0,000000

Fig. 114. **The table of a posteriori probabilities**



Discriminant Function Analysis Results: Spreadsheet12

```
Number of variables in the model: 5

Wilks' Lambda: ,0013939   approx. F (15,19) = 12,92218 p < ,0000
```

A priori classification probabilities
○ Proportional to group sizes
● Same for all groups
○ User defined

Score to save for each case
● Save classification for case
○ Save distance for case
○ Save posterior probability for case

Max. number of cases in a single results spreadsheet: 100000

Squared Mahalanobis Distances from Group Centroids (
Incorrect classifications are marked with *

Case	Observed Classif.	G_1:1 p=,25000	G_2:2 p=,25000	G_3:3 p=,25000	G_4:4 p=,25000
Austria	G_1:1	2,0729	234,4760	52,8876	14,5610
Belgium	G_1:1	0,9494	264,7849	41,7228	16,2165
Bulgaria	G_1:1	5,3179	250,3143	47,7357	25,2173
Finland	G_4:4	13,4271	329,3805	33,5037	0,8098
France	G_4:4	12,9011	319,6329	37,9240	3,5545
Germany	G_1:1	7,1967	322,9463	14,9096	17,2445
Italy	G_1:1	3,0716	215,6587	61,8547	27,4519
Poland	G_3:3	42,8075	424,9053	4,2671	44,3261
Spain	G_3:3	39,6876	464,0628	1,7786	29,4089
Sweden	G_4:4	39,9212	387,5931	47,4028	5,5078
Switzerland	G_4:4	18,2628	337,6522	37,5058	0,3157
United Kingdom	G_2:2	223,2886	4,5423	423,7596	311,0095
Belarus	G_4:4	13,0294	348,4689	22,6968	5,2710
Russia	G_2:2	293,6248	4,5423	491,3669	380,9826
Ukraine	G_3:3	49,6542	481,9426	5,8024	36,2577
Moldova	---	1,3132	250,1662	43,6811	18,6321
Czech Republic	---	25,4861	130,4643	121,4663	64,1878

Fig. 115. **The table of distances from the new case to the centers of the groups**

The maximum value of a posteriori probabilities (see Fig. 114) and the minimum distance from the new case to the centroids of groups (see Fig. 115) correspond to cluster No. 1. Therefore, the studied countries Moldova and the Czech Republic should be assigned to the first cluster – countries with an average level of energy supply and a high level of use of energy-saving technologies.

Task 3. Using the methods of discriminant analysis of the program Statistica 10.0 and the statistical data of the website of the State Statistics Service of Ukraine [19], find information about main indicators that characterize the level of socioeconomic growths of regions and search for a function according to which the object (regions) belongs to one of the known classes.

Task 4. Using your own information space of research make discriminant analysis and provide an economic interpretation of the results of the factor analysis.

Task 5. Using the data in Table 21 and methods of discriminant analysis, check the quality of clustering and learn to classify objects based on the discriminant function.

Table 21

The indicators of the quality of education of the population

Countries	Level of employment	Unemployment rate	The number of people who graduated from HEI, thousand people	The number of permanent residents, people	Cluster No.
Austria	71.5	6.0	422	8 662 588	2
Belgium	62.3	7.8	488	11 291 746	2
Greece	52.0	23.5	659	10 846 979	1
Denmark	74.9	6.2	291	5 699 220	2
France	64.6	10.1	2338	66 539 000	1
Germany	74.7	4.1	2780	81 292 400	1
Italy	57.2	11.7	1872	60 685 487	1
Poland	64.5	6.2	1902	38 484 000	1
Spain	60.5	19.6	1959	46 423 064	1
Sweden	76.2	7.0	436	9 838 480	2
Slovakia	64.9	6.6	209	9 838 480	2
United Kingdom	75.3	3.9	2380	65 572 409	1
Czech Republic	72	3.4	427	10 541 466	2
Ireland	64.7	5.4	199	4 635 400	2
Switzerland	79.6	3.3	279	8 306 200	2
Finland	69.2	7.3	309	5 496 591	2

Task 6. Check the quality of clustering by the methods of discriminant analysis and learn to classify objects by discriminant function. The initial data is presented in Table 22.

Table 22

The distribution of American states according to the level of development of Agile IT Project Management

States of America	Main indicators that characterize the level of agile development					Cluster No.
	X1	X2	X3	X4	X5	
1	2	3	4	5	6	7
Arizona	70.53	57.8	2 900	33.26	48.7	2
Arkansas	80.4	74.5	1 421	10.12	49.2	2

Table 22 (the end)

1	2	3	4	5	6	7
California	92.3	40.1	5 075	10.84	75.4	3
Colorado	75.21	70	2 111	15.63	51.3	2
Connecticut	63.2	71.2	1 100	10.97	49.5	1
Delaware	64.79	65.3	700	11.99	48.8	1
District of Columbia	65.83	34.8	500	13.4	43.44	1
Florida	64.51	57.9	1 200	13.99	46.3	1
Georgia	76.31	55.9	400	11.4	58.56	3
Hawaii	64.38	22.77	600	11.39	39.19	1
Idaho	78.37	84	508	19.75	40.2	2
Illinois	76.4	63.7	865	14.57	48.9	2
Indiana	68.72	81.5	658	16.96	46.3	2
Iowa	68.99	88.7	583	9.46	47.7	2
Kansas	60.33	78.2	862	16.21	47.3	1
Kentucky	70.36	86.3	913	12.7	52.8	2
Louisiana	63.48	60.3	979	15.74	43.2	1

Task 7. Using the Internet resources, find information about the level of development of Agile IT Project Management in other US states and divide them into already known groups received in the previous task (Table 22).

The list of questions for independent work

1. Formulate the definition of discriminant analysis.
2. What approaches are used for conducting discriminant analysis?
3. What is the general view of discrimination?
4. What is the concept of centroid?
5. What characterizes the interclass variation?
6. What are the main tasks of discriminant analysis?
7. What are the criteria for checking the quality of discrimination?
8. Give a definition of the discriminant function.
9. In what limits is Wilks' lambda measured?
10. What is the difference between standardized and structural coefficients of the discriminant function?

References

Main

1. Антохонова И. В. Методы прогнозирования социально-экономических процессов : учебное пособие / И. В. Антохонова. – Улан-Удэ : Изд-во ВСГТУ, 2004. – 212 с.

2. Бабешко Л. О. Основы эконометрического моделирования : учебное пособие / Л. О. Бабешко. – Изд. 3-е. – Москва : КомКнига, 2007. – 432 с.

3. Вітлинський В. В. Моделювання економіки : навч. посіб. / В. В. Вітлинський. – Київ : КНЕУ, 2003. – 408 с.

4. Геєць В. М. Моделі і методи соціально-економічного прогнозування : підручник / В. М. Геєць, Т. С. Клебанова, О. І. Черняк та ін. – 2-ге вид., виправ. – Харків : ВД ″ІНЖЕК″, 2008. – 396 с.

5. Єріна А. М. Статистичне моделювання та прогнозування : навч. посіб. / А. М. Єріна. – Київ : КНЕУ, 2001. – 170 с.

6. Клебанова Т. С. Эконометрия : учебно-методическое пособие для самостоятельного изучения дисциплины / Т. С. Клебанова, Н. А. Дубовина, Е. В. Раевнева. – Харьков : Изд. Дом ″ИНЖЭК″, 2003. – 132 с.

7. Магнус Я. Р. Эконометрика: начальный курс : учеб. / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. – 8-е изд., испр. – Москва : Дело, 2007. – 504 с.

8. Многомерный статистический анализ в экономике : учеб. пособие для вузов / Л. А. Сошникова, В. Н. Тамашевич, Г. Уебе и др. ; под ред. проф. В. Н. Тамашевича. – Москва : ЮНИТИ-ДАНА, 1999. – 598 с.

9. Присенко Г. В. Прогнозування соціально-економічних процесів : навч. посіб. / Г. В. Присенко, Є. І. Равікович. – Київ : КНЕУ, 2005. – 378 с.

10. Статистика : навчальний посібник / під ред. д-ра екон. наук, професора О. В. Раєвнєвої. – Харків : ХНЕУ, 2010. – 520 с.

Additional

11. Андрієнко В. Ю. Статистичні індекси в економічних дослідженнях / В. Ю. Андрієнко. – Київ : Академперіодика, 2004. – 536 с.

12. Дуброва Т. А. Факторный анализ с использованием пакета "STATISTICA" : учебное пособие / Т. А. Дуброва, Д. Э. Павлов, Н. П. Осипова. – МГУ экономики, статистики и информатики. – Москва : Финансы и статистика, 2002. – 702 с.

13. Моделирование экономики : учебное пособие / Т. С. Клебанова, В. А. Забродский, О. Ю. Полякова и др. – Харьков : Изд-во ХГЭУ, 2001. – 140 с.

14. Орлов А. И. Организационно-экономическое моделирование : учебник : в 3 ч. / А. И. Орлов. – Москва : Изд-во МГТУ им. Н. Э. Баумана. – 2009. – 254 с.

15. Халафян А. А. STATISTICA 6. Статистический анализ данных / А. А. Халафян. – Москва : ООО ″Бином-Пресс″, 2008. – 512 с.

16. Dickey D. A. Distribution of the estimators for autoregressive time series with a unit root / D. A. Dickey, W. A. Fuller // Journal of the American Statistical Association. – 1979. – Vol. 74. – P. 427–431.

17. Granger C. W. J. Forecasting economic time series / C. W. J. Granger, P. Newbold. – 2nd ed. – New York : Academic Press, 1986. – 324 p.

18. Lachenbruch P. A. Discriminant Analysis / P. A. Lachenbruch. – New York : Hafner, 1974 – 234 p.

Information resources

19. Офіційний сайт державної служби статистики України [Електронний ресурс]. – Режим доступу : http://www.ukrstat.gov.ua.

20. Электронный учебник StatSoft [Электронный ресурс]. – Режим доступа : http://www.statsoft.ru.

Content

НАВЧАЛЬНЕ ВИДАННЯ

СТАТИСТИЧНЕ МИСЛЕННЯ ДЛЯ НАУКИ ПРО ДАНІ

Методичні рекомендації
до самостійної роботи студентів
спеціальності 122 "Комп'ютерні науки"
другого (магістерського) рівня
(англ. мовою)

Самостійне електронне текстове мережеве видання

Укладачі: **Раєвнєва** Олена Валентинівна
Дериховська Вікторія Ігорівна

Відповідальний за видання *О. В. Раєвнєва*

Редактор *З. В. Зобова*

Коректор *З. В. Зобова*

Подано завдання для самостійної роботи з навчальної дисципліни та методичні рекомендації до їх виконання, що допоможуть студентам набути практичних навичок використання інструментів економіко-математичного моделювання під час вивчення складних соціально-економічних процесів та систем.

Рекомендовано для студентів спеціальності 122 "Комп'ютерні науки" другого (магістерського) рівня.

План 2021 р. Поз. № 138 ЕВ. Обсяг 108 с.

Видавець і виготовлювач – ХНЕУ ім. С. Кузнеця, 61166, м. Харків, просп. Науки, 9-А

Свідоцтво про внесення суб'єкта видавничої справи до Державного реєстру
Дк № 4853 від 20.02.2015 р.