**МІНІСТЕРСТВО ОСВІТИ І НАУКИ,
МОЛОДІ ТА СПОРТУ УКРАЇНИ**

ХАРКІВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ

**Лабораторний практикум
з навчальної дисципліни**

"БІЗНЕС-СТАТИСТИКА"

**для студентів спеціальності
8.03050601 "Прикладна статистика"
денної форми навчання**

Харків. Вид. ХНЕУ, 2013

Затверджено на засіданні кафедри статистики та економічного прогнозування.

Протокол № 1 від 29.08.2012 р.

Укладачі: Раєвнєва О. В.
Чанкіна І. В.
Гольтяєва Л. А.

Л12 Лабораторний практикум з навчальної дисципліни "Бізнес-статистика" для студентів спеціальності 8.03050601 "Прикладна статистика" денної форми навчання / укл. О. В. Раєвнєва, І. В. Чанкіна, Л. А. Гольтяєва. – Х. : Вид. ХНЕУ, 2013. – 68 с. (Укр. мов.)

Подано лабораторні роботи, метою яких є закріплення й поглиблення знань теоретичного і практичного матеріалу, набуття навичок аналізу різних типів даних за допомогою пакета STATISTICA.

Рекомендовано для студентів економічних спеціальностей.

Вступ

Ведення бізнесу – справа не для всіх. Лише частина тих, хто наважується відкрити власну справу, досягають успіху. Лише десята частина з тих, хто зважився, отримує прибуток, надприбуток, решта ж або працюють на малих обертах, або змушені піти з ринку, зазнавши невдачі.

Бізнес-статистика є дисципліною, яка вивчає сукупність кількісних відомостей, що характеризують стан явищ та процесів. Тобто завдання, які можуть бути вирішені за допомогою методів бізнес-статистики є пошук статистичних даних про кількість знову відкритих фірм, які є потенційними конкурентами, вибір правильної стратегії поведінки, для того, щоб утримати свої позиції в даному сегменті ринку та ін.

Уже давно визнано той факт, що найбільш цінним ресурсом є своєчасна та достовірна інформація, а тому бізнес-статистика, будучи похідною математики, статистики та аналізу, здатна надати таку інформацію.

Лабораторний практикум з даної дисципліни призначений для вирішення сучасних бізнес-проблем та закріплення теоретичного та практичного матеріалу, вироблення навичок роботи з пакетами прикладних програм, що забезпечують аналіз статистичних даних. Для виконання лабораторних робіт пропонується використовувати пакет Statistica 8.0.

Лабораторні роботи розроблені за основними темами дисципліни й ґрунтуються на теоретичному матеріалі відповідної теми а також попередніх тем. Кожна робота містить мету, завдання й методичні рекомендації до виконання. Лабораторні роботи рекомендується виконувати послідовно, оскільки дії й прийоми, загальні для всіх робіт, будуть вказуватися один раз. Крім того, послідовне виконання дозволяє краще засвоїти й закріпити матеріал дисципліни.

Для кожної лабораторної роботи оформлюється звіт. Оцінка за виконання роботи ставиться за результатами виконання та захисту лабораторної роботи. Особлива увага приділяється правильності висновків та повноті економічної інтерпретації результатів.

Знання та вміння, отримані в процесі навчання, формують такі компетентності (табл. 1).

Таблиця 1

Компетентності, які повинен здобути студент, що вивчає навчальну дисципліну

Компетентності фахівця	Зміст компетентності	Уміння фахівця освітньо-кваліфікаційного рівня "магістр"
1. Інформаційна	1.1. Здатність використовувати сучасні методи інтелектуального аналізу даних для моделювання розвитку соціально-економічних систем та процесів	1.1.1. Використовувати різноманітні можливості пакету прикладної програми Statistica 8.0 при обґрунтуванні прийняття рішень у бізнесі. 1.1.2. Використовувати інформаційно-аналітичні пошукові системи щодо отримання необхідної інформації
2. Аналітична	2.1. Здатність щодо проведення статистичного аналізу різноманітних економічних процесів	2.1.1. Володіти сучасними економіко-математичними методами. 2.1.2. Уміти визначати існуючі методи, які застосовуються в аналізі тенденцій основних характеристик сегментів бізнесу. 2.1.3. Здійснювати формування інформаційного простору відповідно до наукових методів первинної обробки інформації. 2.1.4. Здійснювати моделювання бізнес процесів за допомогою новітніх методів. 2.1.5. Прогнозувати перебіг бізнес-подій на підставі розроблених моделей
3. Обліково-статистична	3.1. Здатність проводити розрахунки основних показників, що характеризують певні процеси бізнес-середовища	3.1.1. Визначати потреби сучасних підприємств, використовуючи індивідуальний підхід, впроваджувати сучасні методи оцінки. 3.1.2. Досліджувати взаємозв’язки соціально-економічних показників. 3.1.3. Володіти сучасними інформаційними технологіями збору, обробки та поширення даних

Модуль 1. Методи аналізу даних для прийняття рішень у бізнесі

Лабораторне заняття на тему "Первинний аналіз даних у системі Statistica (розрахунок та використання показників рівня, розсіяння та асиметрії)"

Мета – набуття навичок пошуку статистичних даних у середовищі Internet, освоїти способи класифікації типів наборів даних.

Завдання – необхідно знайти багатовимірні просторові кількісні дані та провести їх аналіз за допомогою описових статистик.

Методичні рекомендації

Для того, щоб почати пошук статистичної інформації, необхідно запустити програму перегляду (браузера), наприклад, Microsoft Internet Explorer. У полі адреси введіть назву сайта зі статистичною інформацією.

Пошук статистичних даних для аналізу можливо здійснити на офіційних державних сайтах : http://www.ukrstat.gov.ua, http://me.kmu.gov.ua, http://www.nbuv.gov.ua, сайтах міжнародних організацій http://www.cisstat.com/, http://laborsta.ilo.org, http://www.oecd.org/statistics, http://apps.fao.org, http://www.imf.org, http://unescostat.unesco.org, http://www.eclac.cl/estadisticas, http://unstats.un.org/unsd/, http://www.un.org/russian/, http://www.unescap.org/stat/, http://www.unido.org/doc/3474, http://www3.who.int/whosis/menu.cfm, http://www.world-bank.org, http://www.wto.org тощо, користуючись пошуковими системами.

Для аналізу статистичних даних у пакеті Statistica необхідно в меню програм выбрати ярлик програми Statistica. Для збереження файла з початковими даними необхідно вибрати в меню пункт File / New Data. З'явиться діалогове вікно New Data: Specify file name (Нові дані: Визначте ім'я файла), в якому необхідно вказати ім'я файла і його розміщення. Після введення імені файла натисніть кнопку зберегти у вікні, що з’явилось.

У результаті з’являється порожнє поле даних, яке є таблицею розміром 10 × 10. Стовпці таблиці називаються Variables (змінні), а рядки - Cases (випадки, спостереження) (рис 1).

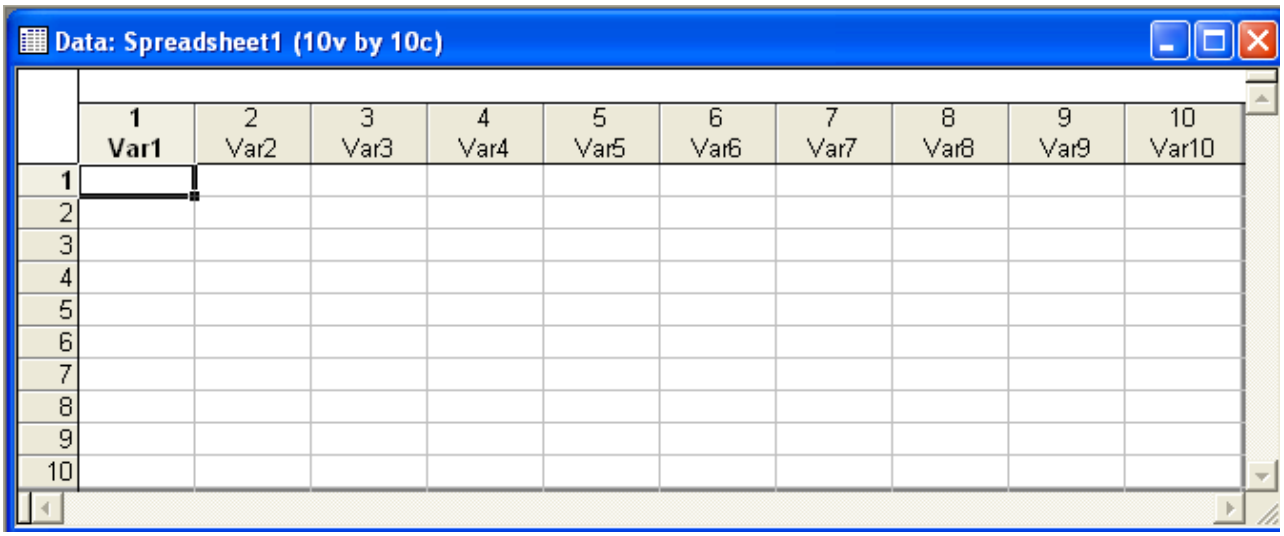
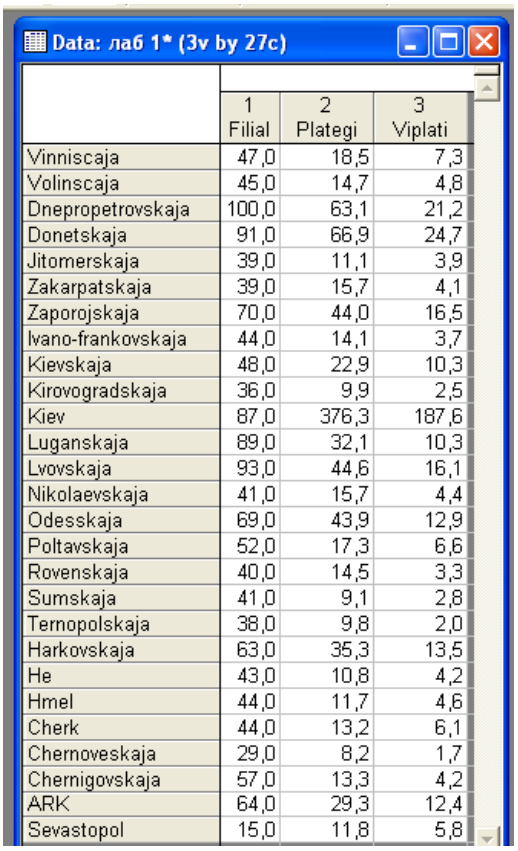


Рис. 1. **Порожнє поле даних**

Кожна змінна має своє ім'я, формат та інші атрибути (які називаються специфікацією змінної), що задаються користувачем. Для зручності роботи з даними необхідно спочатку встановити кількість змінних і спостережень. Для даного прикладу поле даних повинно містити 6 змінних (Variables) і 27 спостережень (Cases). Операції над змінними Vars і спостереженнями Cases доступні або в меню Data, вибравши відповідну кнопку Vars (Cases) або через контекстне меню, натиснувши правою кнопкою миші на імені змінної (спостереження). За допомогою команд Add (додати), Move (перемістити), Copy (копіювати), Delete (видалити) можливе проведення дій як над змінними, так і над спостереженнями. Таблиця початкових статистичних даних матиме вигляд зображений на рис. 2.

Далі необхідно провести графічне представлення даних, що можливо за допомогою графічного аналізу. Для побудови гістограм для графічного аналізу на панелі інструментів необхідно вибрати Graphs – 2DGraphs – Histograms. 2DGraphs – це візуальний аналіз даних на площині, який здійснюється за допомогою різноманітних гістограм, діаграм розсіювання, імовірнісних графіків, лінійних графіків, діаграм діапазонів, діаграм розмахів, кругових діаграм, стовпчастих графіків, графіків послідовних значень і т. д.



Data: лаб 1* (3v by 27c)

	1 Filial	2 Plategi	3 Viplati
Vinniscaja	47,0	18,5	7,3
Volinscaja	45,0	14,7	4,8
Dnepropetrovskaja	100,0	63,1	21,2
Donetskaja	91,0	66,9	24,7
Jitomerskaja	39,0	11,1	3,9
Zakarpatskaja	39,0	15,7	4,1
Zaporojskaja	70,0	44,0	16,5
Ivano-frankovskaja	44,0	14,1	3,7
Kievskaja	48,0	22,9	10,3
Kirovogradskaja	36,0	9,9	2,5
Kiev	87,0	376,3	187,6
Luganskaja	89,0	32,1	10,3
Lvovskaja	93,0	44,6	16,1
Nikolaevskaja	41,0	15,7	4,4
Odesskaja	69,0	43,9	12,9
Poltavskaja	52,0	17,3	6,6
Rovenskaja	40,0	14,5	3,3
Sumskaja	41,0	9,1	2,8
Ternopolskaja	38,0	9,8	2,0
Harkovskaja	63,0	35,3	13,5
He	43,0	10,8	4,2
Hmel	44,0	11,7	4,6
Cherk	44,0	13,2	6,1
Chernoveskaja	29,0	8,2	1,7
Chernigovskaja	57,0	13,3	4,2
ARK	64,0	29,3	12,4
Sevastopol	15,0	11,8	5,8

Рис 2. **Таблиця даних для аналізу**

Для обрання параметрів візуального аналізу необхідно виділити вкладку Advanced, натиснувши кнопку Variables:, вибрати змінну, для якої будується гістограма. Гістограму для інших змінних можливо побудувати, натиснувши кнопку Change Variable. У результаті з'явиться таке вікно (рис.3).

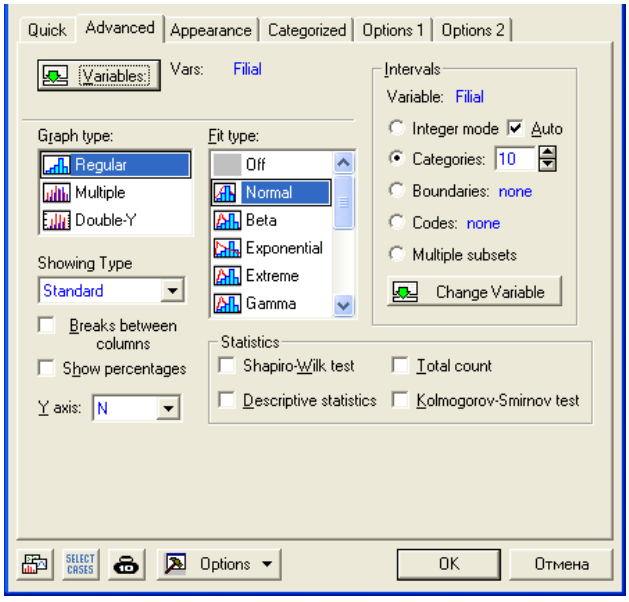


Рис. 3. **Вікно побудови гістограм**

Вибір всіх параметрів підтверджується кнопкою OK. У результаті була отримана гістограма для змінної "філія" (рис. 4), аналогічним чином для інших змінних. Наочне уявлення гістограм зберігається у файлі Woorkbook для можливості використання в подальшому.

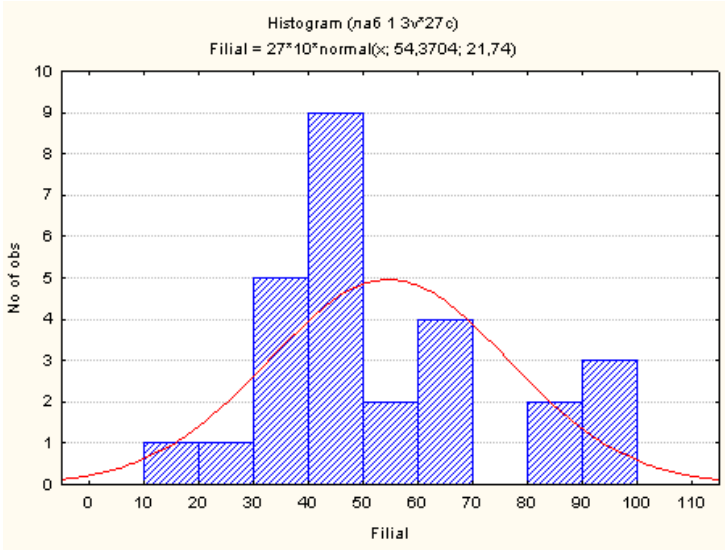


Рис. 4. **Гістограма розміщення філій страхових компаній України за областями**

Для подальшого аналізу даних використовується модуль Descriptive Statistics (описових статистик) (рис.5).

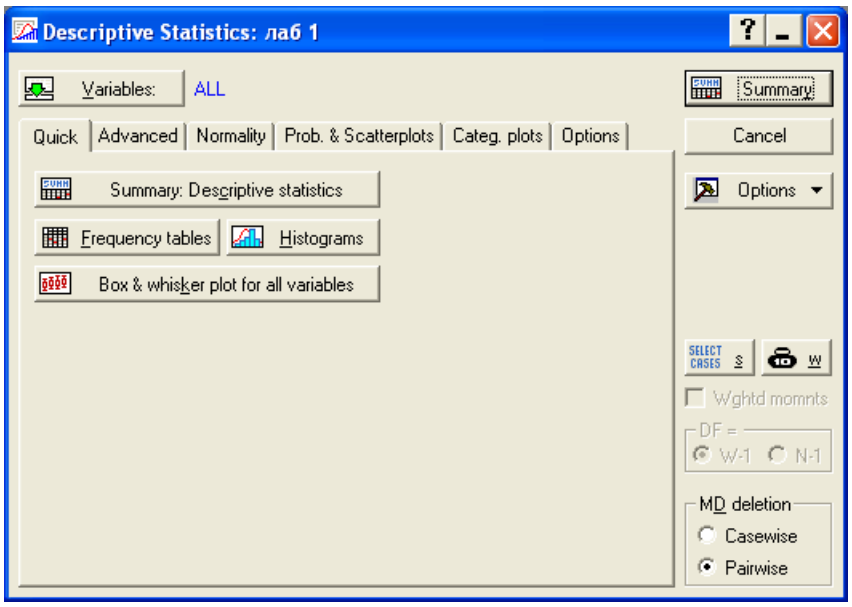


Рис. 5. **Діалогове вікно описових статистик**

Вибір змінних для аналізу виконується у такому вікні (рис.6)

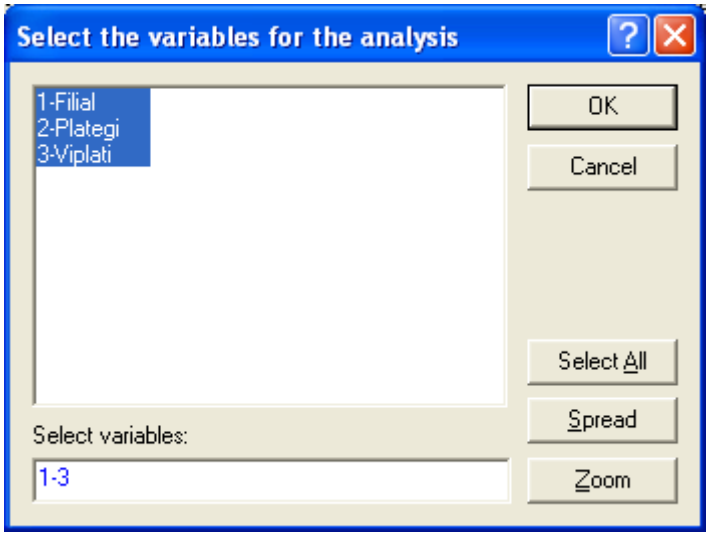


Рис. 6. **Вибір змінних**

У результаті натискання кнопки Summary розраховуються описові статистики (рис.7.)

Descriptive Statistics (па6 1)

Variable	Valid N	Mean	Confidence -95,000%	Confidence +95,000%	Median	Mode	Frequency of Mode	Percentile 10,00000	Percentile 90,00000	Range	Quartile Range	Variance	Std.Dev.	Skewness	Std.Err. Skewness	Kurtosis
Filial	27	54,37037	45,77032	62,97042	45,00000	44,00000	3	36,00000	91,00000	85,0000	29,00000	472,627	21,73998	0,716793	0,447852	-0,26939
Plategi	27	36,21481	8,55302	63,87661	15,70000	15,70000	2	9,80000	63,10000	368,1000	23,60000	4889,643	69,92598	4,756960	0,447852	23,74804
Viplati	27	14,72222	0,84599	28,59846	5,80000	Multiple		2,50000	21,20000	185,9000	9,00000	1230,438	35,07760	4,953540	0,447852	25,21913

Рис. 7. **Вікно розрахованих описових статистик**

Розраховані значення основних показників можливо інтерпретувати таким чином:

Valid N (число спостережень) – обсяг вибірки. У даному випадку обсяг вибірки становить 27 одиниць для кожного показника. Mean (середнє) – це узагальнюючий показник, що характеризує типовий рівень явища. Показує центральне становище змінної і розглядається спільно з довірчим інтервалом. У даному випадку для змінної філіал середнє дорівнює 54,37.

Conf. limits for mean (довірчий інтервал для середнього) – інтервал значень навколо оцінки, де з певною ймовірністю знаходиться "істинне" середнє генеральної сукупності. Для змінної "філіал" цей інтервал – 45,77; 62,97.

Median (медіана) – вимір центральної тенденції, значення, яке розбиває вибірку на дві рівні частини так, що 50 % значень лежить нижче значення медіани, а інші 50 % – вище. (Медіана для змінної, що аналізується дорівнює 45).

Mode (мода) – значення, відповідне найбільшої частоті появи змінної у вибірці. (Для змінною філія мода дорівнює 44).

Standart Deviation (середньоквадратичне відхилення) – показує абсолютне відхилення виміряних значень від середньоарифметичного (21, 73 значення середньоквадратичне відхилення досліджуваної змінної).

Variance (дисперсія) – один з показників варіації кількісної змінної, дорівнює відношенню суми квадратів відхилень від середнього арифметичного до числа ступенів свободи даної суми квадратів (n - 1) (472,62 для змінної філія).

Skewness (асиметрія) – міра симетричності розподілу. Якщо розподіл симетричний, то вона дорівнює 0; якщо асиметрія істотно відрізняється від 0 – розподіл не симетричний. Асиметрія з довгим правим хвостом позитивна, з лівим – негативна. Розподіл несиметричний.

Kurtosis (ексцес) – міра гостроти піку розподілу, при нормальному розподілі ексцес дорівнює 0. Якщо ексцес позитивний - розподіл має загострений пік, якщо від'ємний - плоский пік.

Range (розмах вибірки) вимірює різницю між максимальним і мінімальним значеннями ознаки, що варіює (85 для змінної, що аналізується).

Лабораторна робота на тему "Перевірка гіпотез за допомогою статистичних критеріїв (Z-та t-тестів)"

Мета роботи – засвоєння методів перевірки статистичних гіпотез.

Завдання – з використанням пакету Statistica перевірити гіпотезу за допомогою t-критерію двох залежних вибірок.

Методичні рекомендації

Розглянемо такі дані про обіг фірми за 15 днів до і після публікації реклами (рис. 8). Необхідно визначити чи збільшився обіг фірми після проведення рекламної компанії, тобто, чи є значиміше ефект від реклами.

	1 До	2 После
1	100,744	115,842
2	102,497	99,655
3	81,305	98,71
4	105,532	120,96
5	96,715	101,85
6	87,659	122,346
7	109,51	117,175
8	102,047	113,978
9	97,978	101,172
10	89,96	115,694
11	120,777	110,822
12	113,219	95,506
13	77,832	121,507
14	97,831	90,734
15	97,105	114,223

Рис. 8. **Початкові дані**

Аналіз необхідно розпочинати відкриттям у меню Statistics модуля Basic Statistics / Tabels. У запропонованому меню необхідно вибрати рядок t-test for dependent samples (t-критерій для залежних вибірок) і натиснути кнопку 

. На екрані з'являється діалогове вікно t-критерію для залежних вибірок (рис. 9)

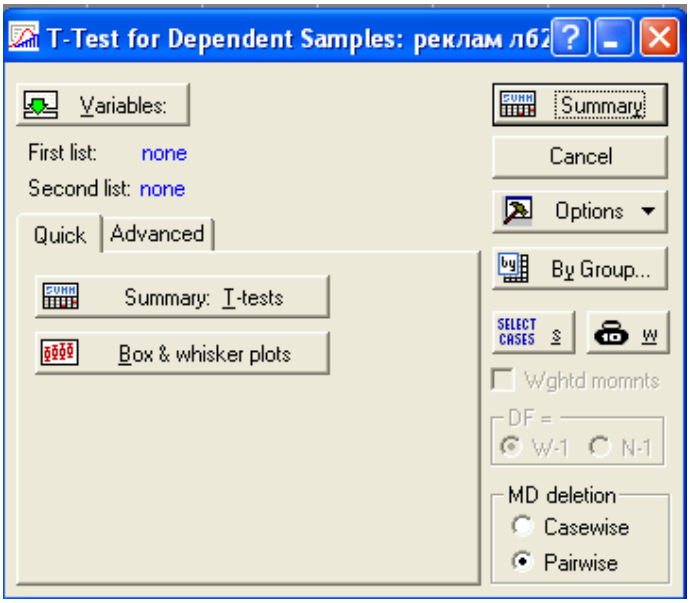


Рис. 9. **Вікно t-критерію для залежних вибірок**

У представленому вікні необхідно вибрати змінні для аналізу натисненням кнопки Variables:. Після чого відкриється вікно Select one or two variable lists (вибрати один або два списки змінних). У лівому списку вибираємо змінну до, в правому – після (рис. 10)

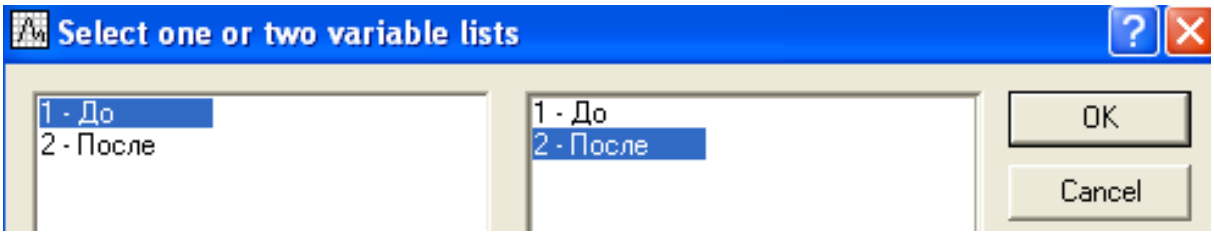


Рис. 10. **Вибір змінних для аналізу**

Далі повертаємось у діалогове вікно t-test for dependent samples натисненням кнопки OK. Для візуалізації даних натискаємо кнопку Box & whisker plots і на екрані з'являється діаграма розмаху ("ящики і вуса") (рис. 11).

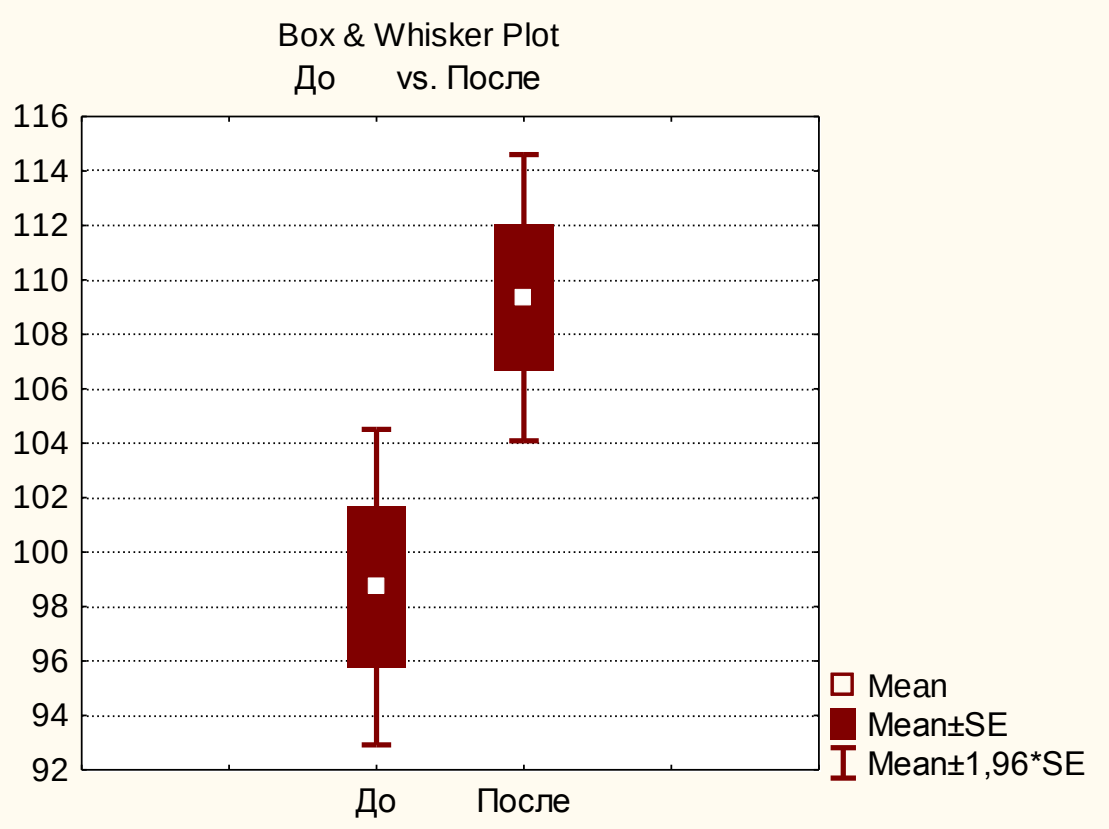


Рис. 11. **Діаграма розмаху для змінних**

Точки в центрі прямокутників відповідають середнім значенням змінних. Від цих значень береться позитивне і негативне стандартне відхилення ("вуса") і позитивна і негативна стандартна помилка ("ящики").

З графіка добре видно, що середній оборот після реклами збільшився, інтервали стандартних помилок не перекриваються. Натисненням кнопки 

на екран виводиться докладна таблиця результатів тесту (рис. 12)

Variable	T-test for Dependent Samples (реклам л62) Marked differences are significant at p < ,05000 **Mean**	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
До	98,7141	11,45062						
После	109,3449	10,38879	15	-10,6309	16,58011	-2,48329	14	0,026302

Рис. 12. **Електронна таблиця результатів**

Оскільки ρ = 0,02 < 0,05, то можна стверджувати, що різниця в середніх значеннях доходу до і після проведення реклами є значущою. Тобто результати статистичного аналізу показали значуще збільшення обороту фірми після проведення реклами.

Лабораторне заняття на тему "Прогнозування значень на основі багатофакторної регресійної моделі"

Мета роботи – набуття навичок побудови та аналізу нелінійних виробничих функцій.

Завдання – побудувати багатофакторну регресійну модель, використовуючи виробничу функцію Коба – Дугласа.

Методичні рекомендації

Необхідно перевірити наявність лінійного і нелінійного зв'язку між величиною факторів виробництва і обсягом продукції, що випускається. Початкові дані подані на рис.13.

	1 **Вартість основних фондів, тис. грн.**	2 Середньо-спискова чисельність робітників, чол.	3 Випуск валової продукції, тис. грн.
1	66	60	54
2	174	159	162
3	99	153	138
4	111	255	261
5	1059	1215	1347
6	576	288	252
7	1149	471	708
8	1887	264	135
9	231	240	243
10	333	153	138
11	381	240	324
12	693	225	117
13	66	60	54
14	174	168	162
15	99	171	138
16	111	255	261
17	279	258	213
18	408	42	21
19	66	66	42
20	105	27	9
21	62	53	41
22	195	60	3
23	30	87	32
24	51	98	45
25	23	294	57

Рис. 13. **Початкові дані**

Для початку роботи запустимо програму Statistica. Сформуємо таблицю (файл) вихідних даних. Побудова і вивчення багатофакторної моделі проводиться в модулі Multiple Regression.

Далі, на рис. 14 наводяться найбільш важливі параметри отриманої регресійної моделі.

Multiple R – коефіцієнт множинної кореляції, характеризує тісноту лінійного зв'язку між залежною і всіма незалежними змінними. Може приймати значення від 0 до 1.

R2 – коефіцієнт детермінації. Чисельно виражає частку варіації залежної змінної, поясненою за допомогою регресійного рівняння. Чим більше R2, тим більшу частку варіації пояснюють змінні, включені в модель.

Adjusted R – скоригований коефіцієнт множинної кореляції. Цей коефіцієнт позбавлений недоліків коефіцієнта множинної кореляції.

F – F-критерій; **df** - число ступенів свободи для F-критерію; **p** – імовірність нульової гіпотези для F-критерію; **Standard error of estimate** – стандартна помилка оцінки (рівняння); **Intercept** - вільний член

рівняння; **Std.Error** – стандартна помилка вільного члена рівняння; **t** – t-критерій для вільного члена рівняння; **p** – імовірність нульової гіпотези для вільного члена рівняння; **Beta** – параметр рівняння.

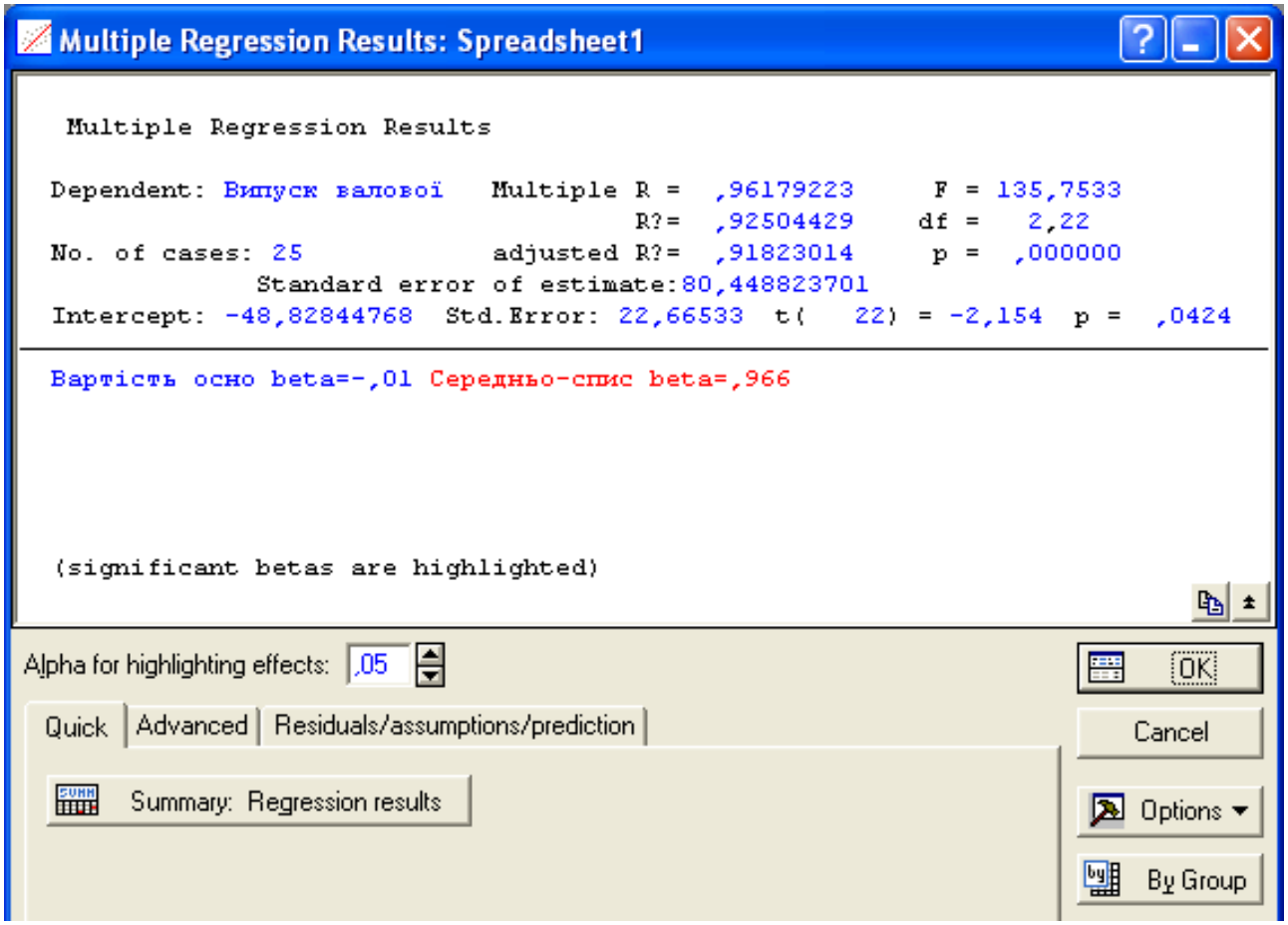


Рис. 14. **Результати модуля Multiple Regression**

Розраховані параметри моделі наведені на рис. 15.

N=25	Regression Summary for Dependent Variable: Випуск валової п R= ,96179223 R?= ,92504429 Adjusted R?= ,91823014 F(2,22)=135,75 p<,00000 Std.Error of estimate: 80,449					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(22)	p-level
Intercept			-48,8284	22,66533	-2,15432	0,042429
Вартість основних фондів, тис. грн.	-0,008397	0,068802	-0,0053	0,04373	-0,12205	0,903970
Середньо-спискова чисельність робітників, чол.	0,966211	0,068802	1,1605	0,08264	14,04335	0,000000

Рис. 15. **Параметри моделі**

Наступним важливим кроком у побудові моделі є оцінка залишків, яку можна провести за допомогою аналізу гістограми залишків (рис. 16).

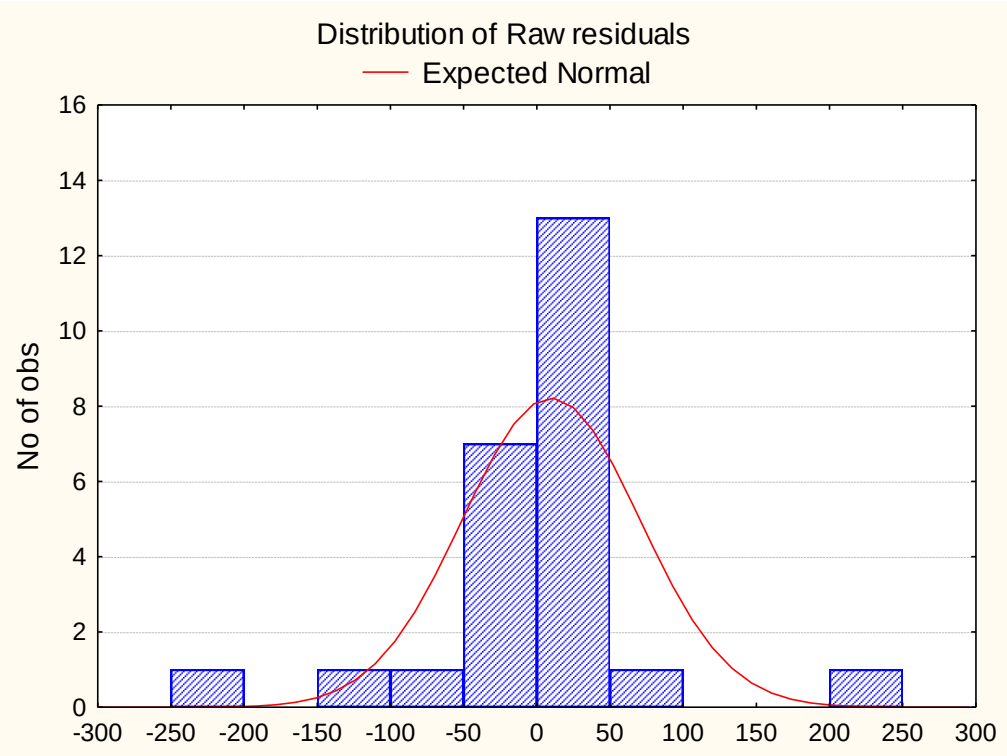


Рис. 16. **Гістограма залишків**

Та графіку залишків на нормальному ймовірнісному папері (рис. 17).

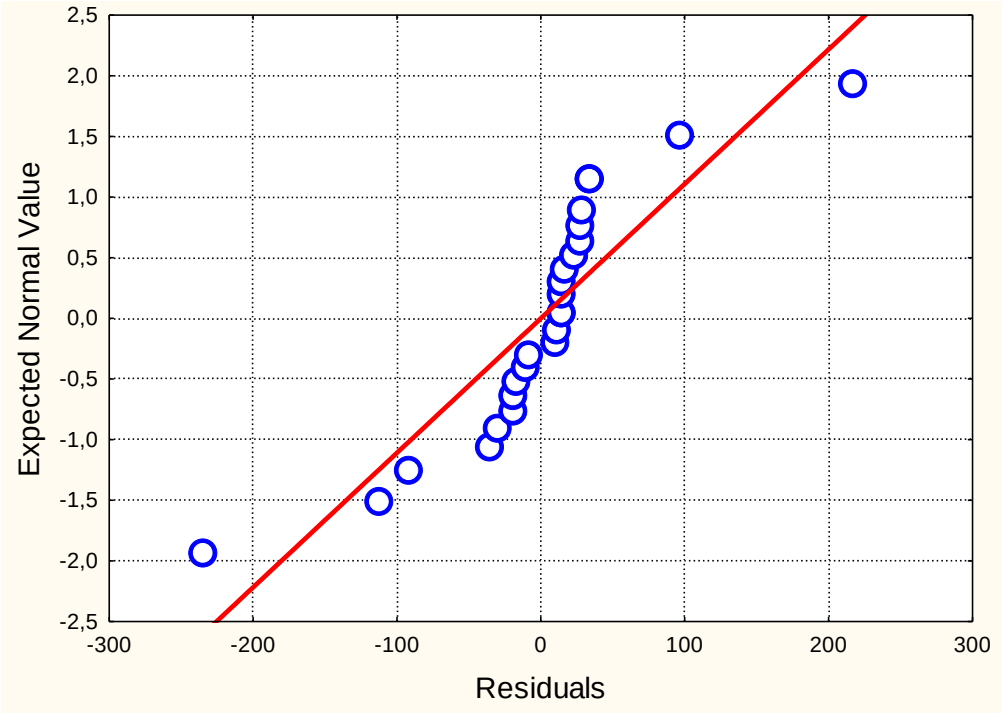


Рис. 17. **Графік залишків на нормальному імовірнісному папері**

За рис. 16 та рис. 17 видно, що розкид залишків є досить великим. Тому перевіримо дані на наявність нелінійної залежності.

Першим етапом вирішення задач на побудову нелінійних моделей є встановлення виду залежності між незалежними і залежною величинами. Це можна зробити за допомогою графічних та аналітичних методів статистики. У даній задачі залежність випуску валової продукції від вартості основних виробничих фондів та середньоспискової чисельності робітників виражається функцією Коба – Дугласа. Вона є нелінійною і має такий вигляд: $Y = a_0 \times OF^{a1} \times L.^{a2}$ Отже, задача побудови моделі зводиться до оцінки параметрів вказаної функції. Для побудови нелінійних багатофакторних моделей у системі Statistica використовується модуль Нелінійного оцінювання (рис. 18).

На панелі інструментів Statistics або в меню Statistics виберемо функцію **Nonlinear Estimation** – Нелінійне оцінювання.

У стартовому вікні Нелінійного оцінювання необхідно обрати вид нелінійної моделі. В даному випадку обираємо **User-specified regression, custom loss function** – Задана користувачем регресія та функція залишків.

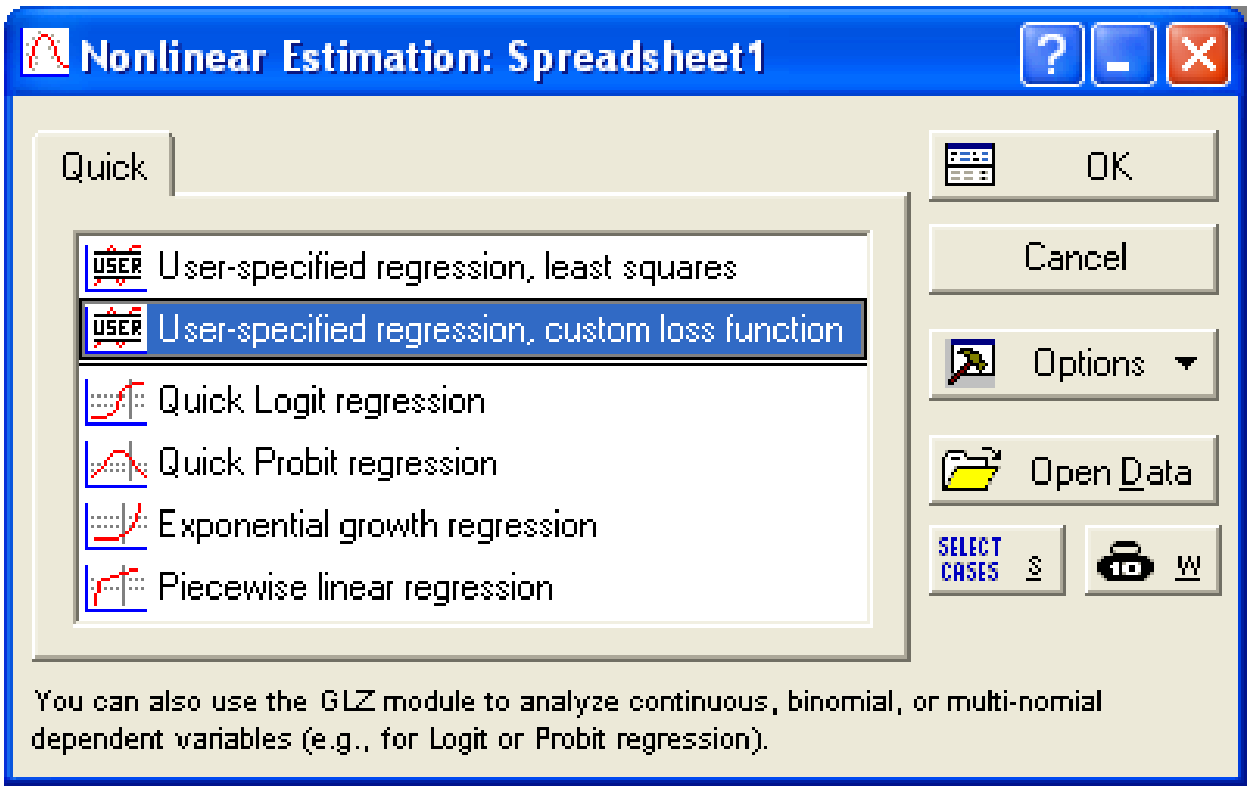


Рис. 18. **Стартове вікно нелінійного оцінювання**

У наступному вікні, ініціювавши кнопку **Function to be estimated & loss function** – Функція для оцінки параметрів і функція залишків маємо задати функцію, параметри якої потрібно оцінити за вихідними даними, та функцію залишків.

Функція залишків за замовчуванням – мінімізація суми квадратів відхилень модельних значень від спостережуваних. Зверніть увагу на запис функцій. Підказки щодо символьного позначення математичних операцій знаходяться в нижній частині вікна (рис.19).

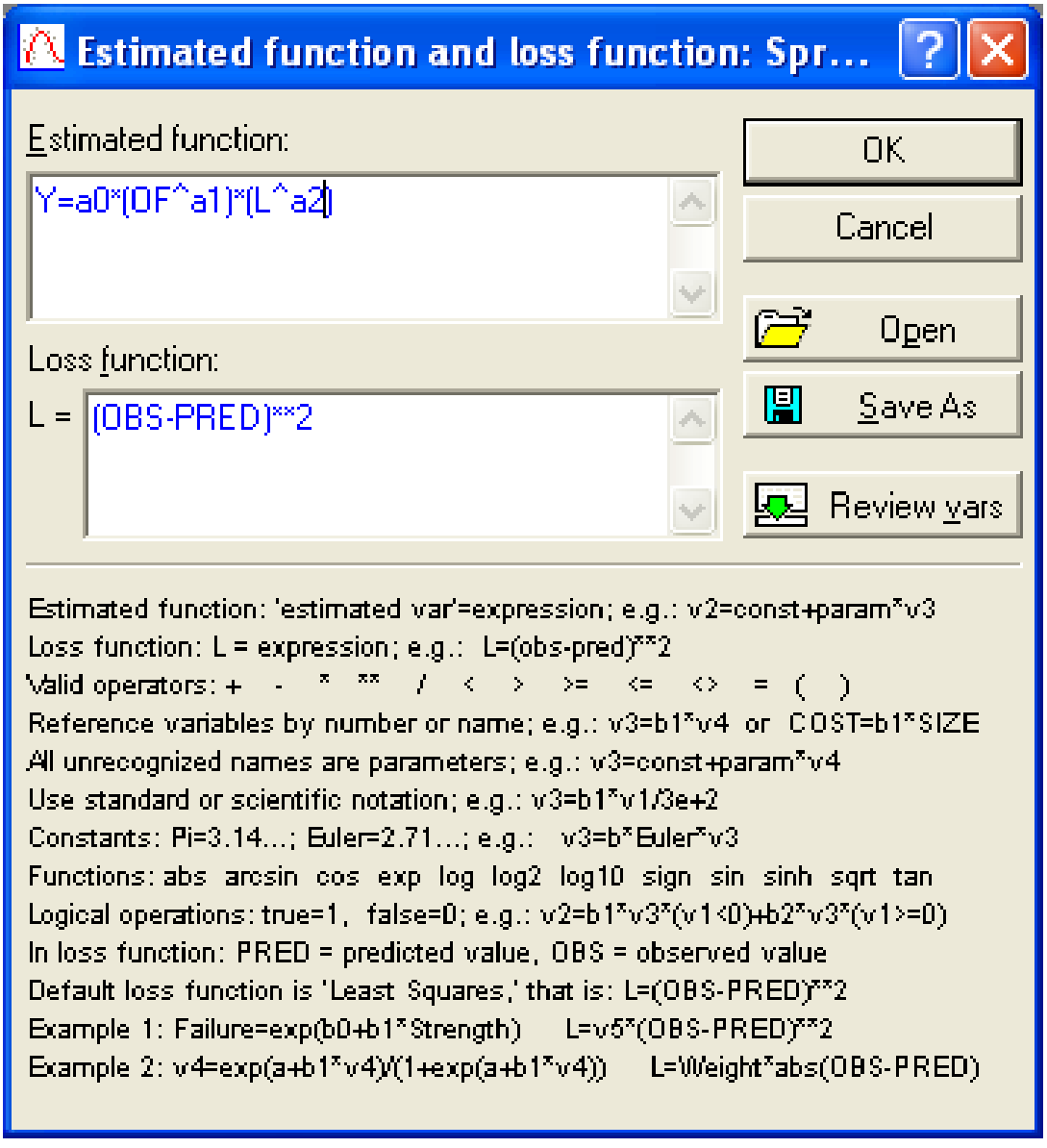


Рис.19. **Функція для оцінки параметрів і функція залишків**

Задавши вид функцій і натиснувши двічі кнопку ОК, переходимо до вікна **Model Estimation** – Оцінка моделі (рис. 20).

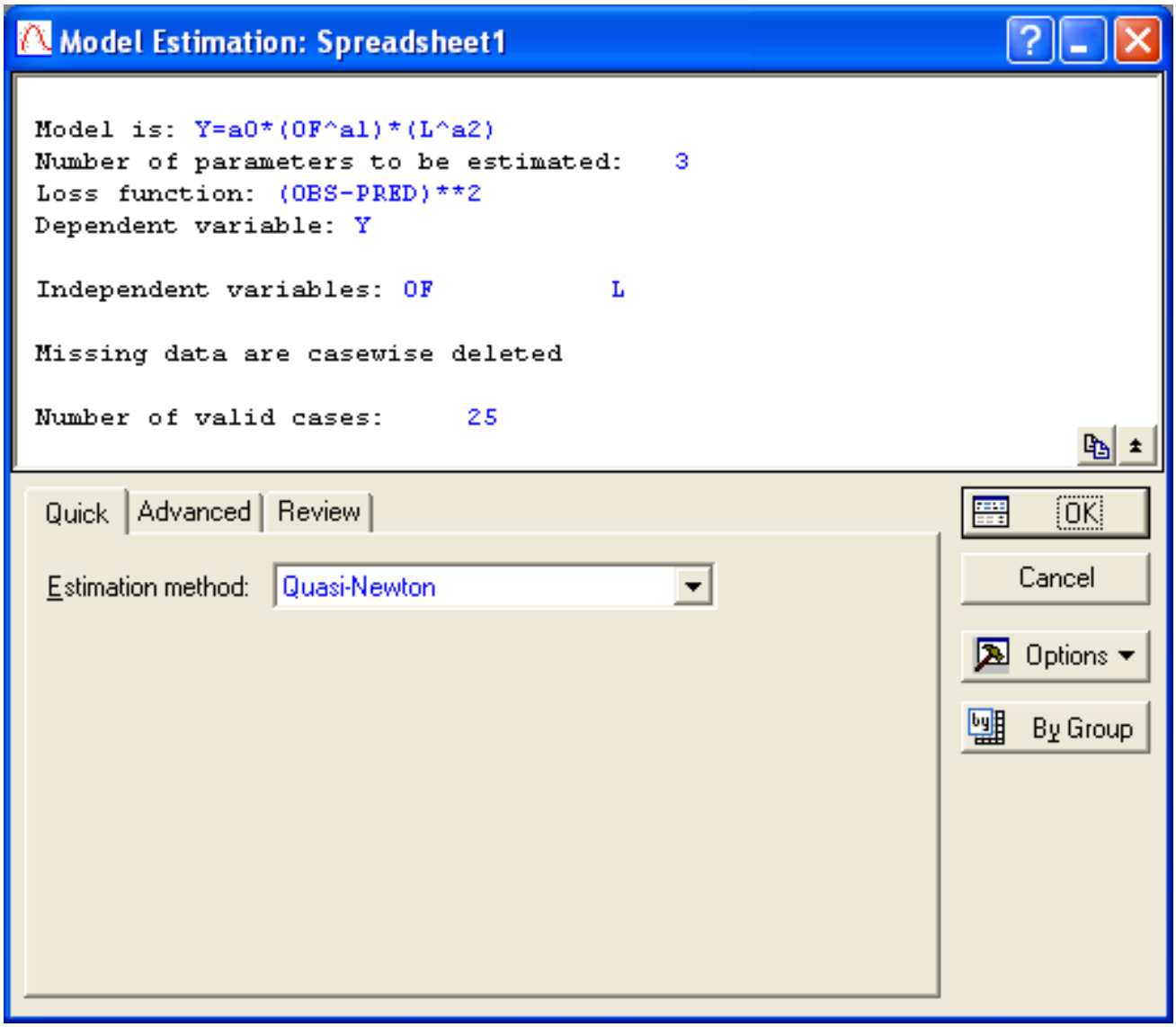


Рис 20. **Оцінка моделі**

Вікно Оцінки моделі складається з двох частин – інформаційної та функціональної. У верхній частині вікна (інформаційній) міститься інформація про вид моделі, кількість оцінюваних параметрів, функцію залишків, залежну і незалежну змінні, метод обробки пропущених значень і кількість точок спостереження. У функціональній частині вікна необхідно перейти на закладку **Advanced** – Додатково і задати метод оцінки параметрів (Estimation method). За замовчуванням це буде метод **Quasi-Newton**. Окрім того, в цьому вікні можна вибрати опцію **Asymptotic standart errors** (Асимптотична стандартна помилка), задати кількість ітерацій для оцінки параметрів (Maximum number of iterations – за замовчуванням 50), точність оцінювання параметрів та інші параметри. Основні характеристики моделі представлені на рис. 21.

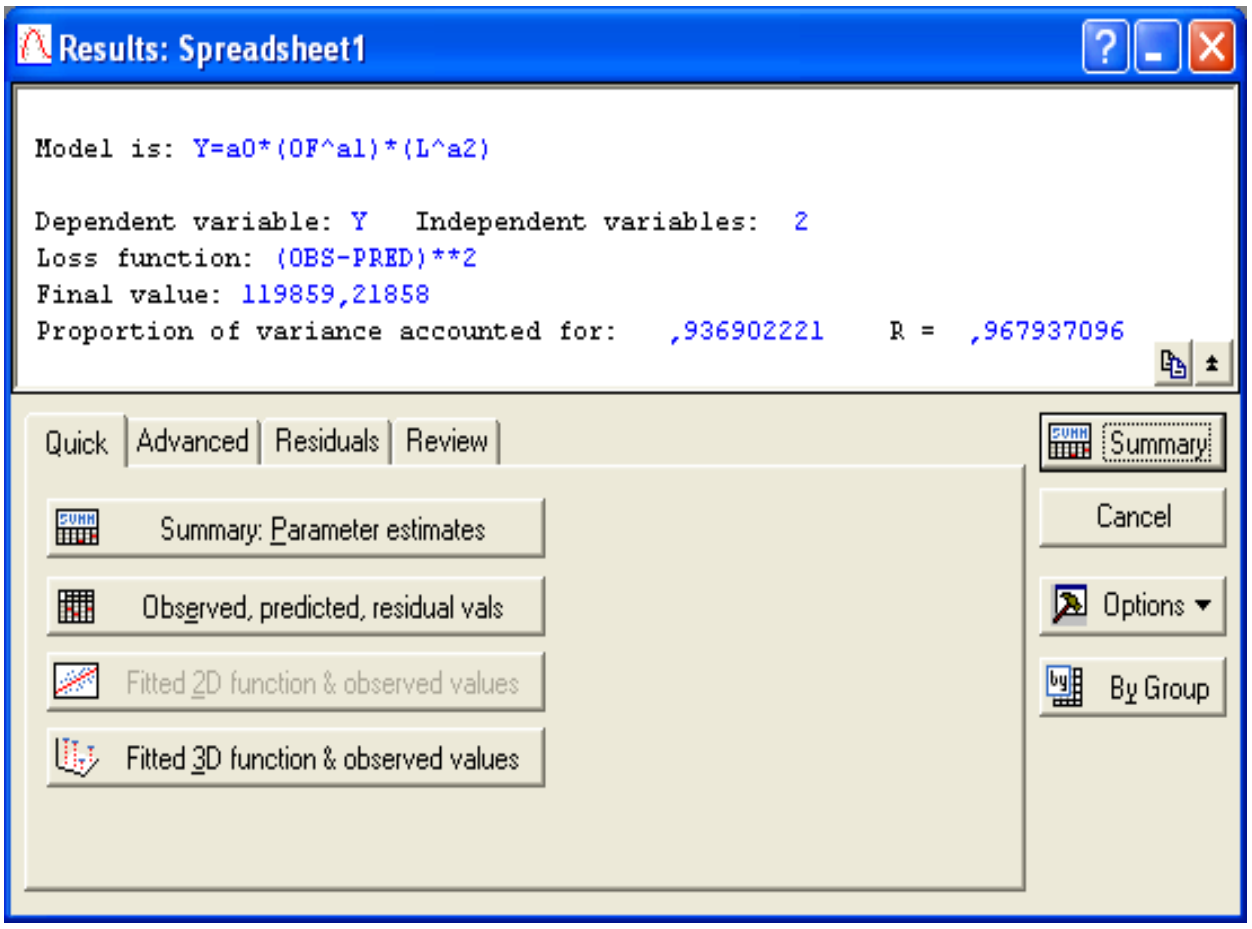


Рис. 21. **Основні характеристики моделі**

Параметри моделі наведено на рис. 22.

Таким чином функція Кобба – Дугласа має наступний вигляд:

$$Y = 0{,}354414 \times OF^{0{,}136895} \times L^{1{,}029019}.$$

	Model: Y=a0*(OF^a1)*(L^a2) (Spreadsheet1) Dep. var: Y Loss: (OBS-PRED)**2 Final loss: 119859,21858 R= ,96794 Variance explained: 93,690%						
N=25	a0	a1	a2				
Estimate	0,354414	0,136895	1,029019				

Рис. 22. **Параметри моделі**

Аналізуючи адекватність моделі необхідно провести аналіз залишків (закладка Residuals – Залишки). Тут маємо ряд кнопок, ініціювавши які можна всебічно проаналізувати залишки (рис. 23).

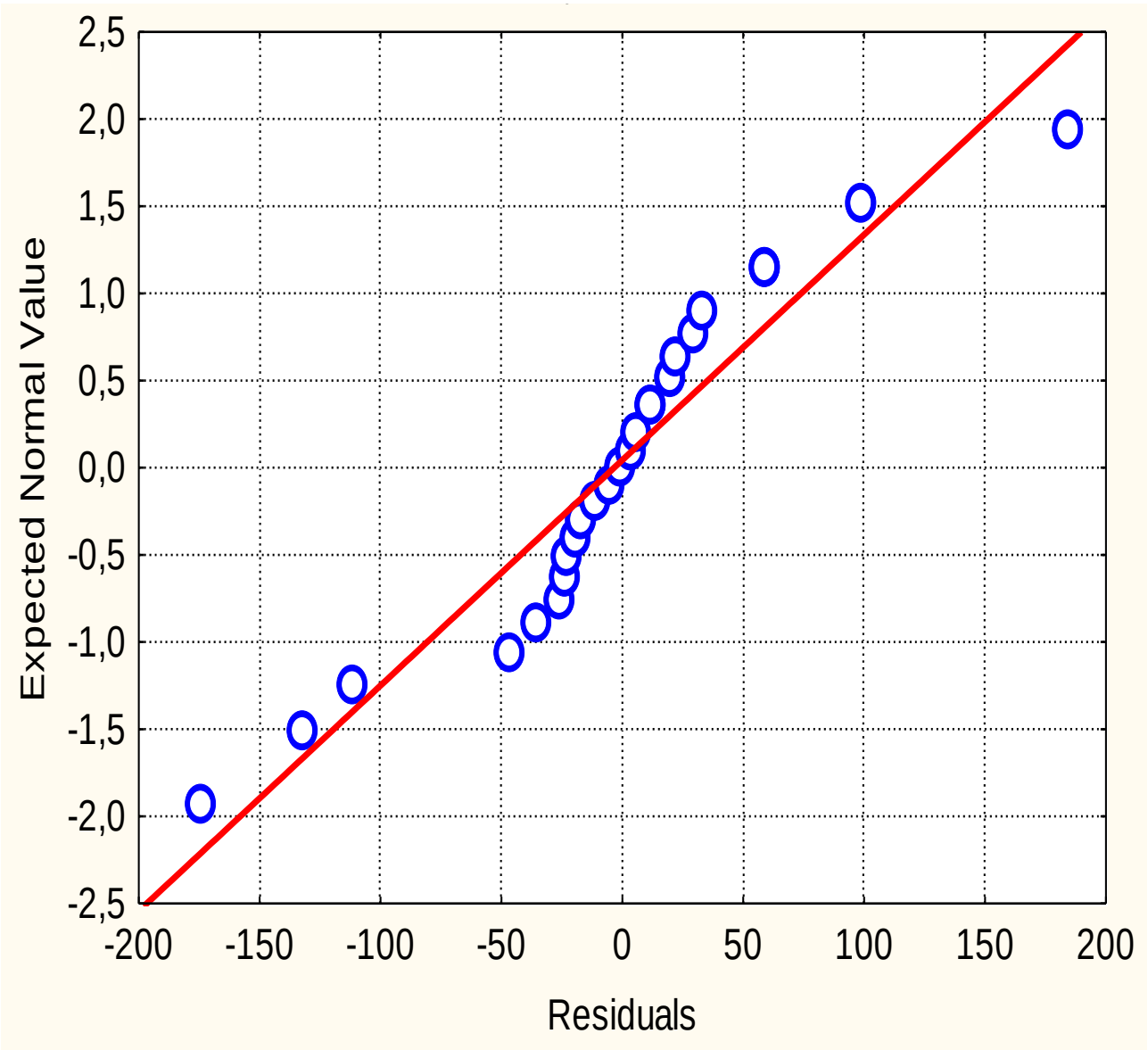


Рис. 23. **Залишки на імовірнісному папері**

Основними кнопками, на які варто звернути увагу, є кнопка Normal probability plot of residuals – Графік залишків на нормальному імовірнісному папері та кнопка Distribution of residuals – Гістограма розподілу залишків (рис. 23 та 24).

Даний графік будується у системі координат, де по осям відкладаються отримані залишки та очікувані значення залишків для кожної точки спостереження. Якщо залишки (точки на графіку) добре лягають на пряму, то це свідчить про адекватність побудованої моделі.

Якщо залишки будуть розподілені за нормальним законом розподілу, то модель вважається адекватною.

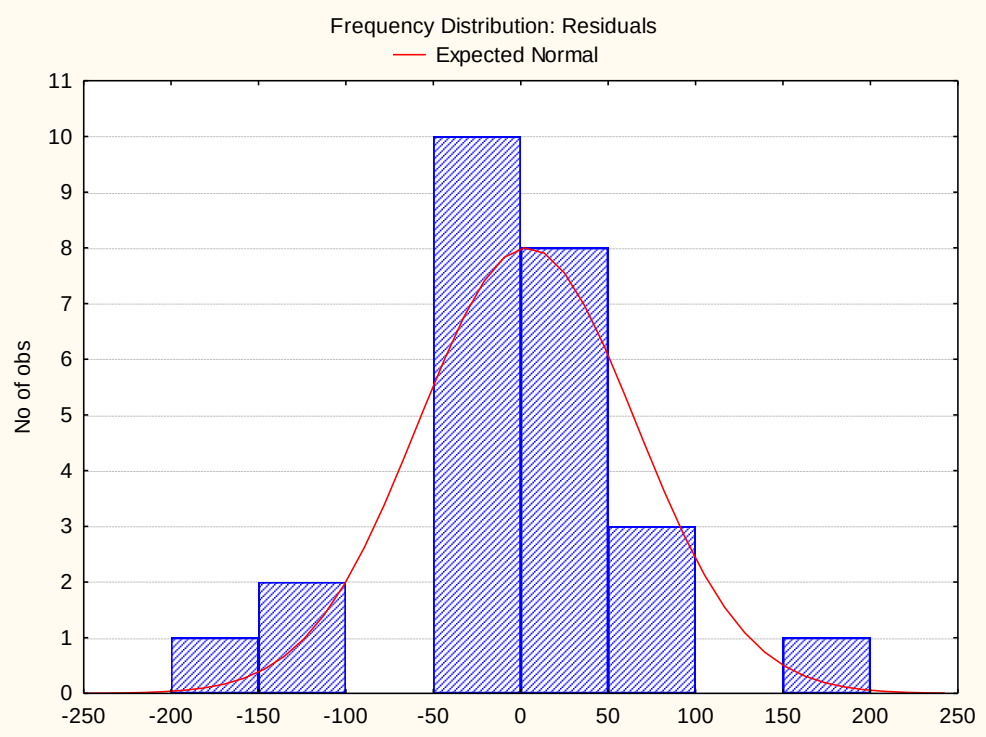


Рис. 24. **Гістограма залишків**

Таким чином, побудована модель достатньо добре відображає залежність випуску продукції від чисельності робітників та вартості основних виробничих фондів і може бути використана для прогнозу. Прогнозні значення випуску продукції обчислюємо, підставивши у модель значення вартості основних фондів та чисельності робітників.

Лабораторне заняття на тему "Використання апарату кластерного аналізу для класифікації даних"

Мета – отримання навиків використання кластерного аналізу в пакеті Statistica.

Завдання – отримати однорідні групи об’єктів з використанням методів кластерного аналізу.

Методичні рекомендації

Розглянемо основні етапи проведення кластерного аналізу в системі STATISTICA на такому прикладі.

Завдання полягає в тому, що приватному підприємцю для прийняття рішення про інвестування капіталу необхідно розподілити

банки за однорідними групами, використовуючи всі відомі методи кластерного аналізу. Вхідні дані за 28 банками за 2011 р. за такими показниками, поданими на рис. 25: X1 – кредитно-інвестиційний портфель; X2 – міжбанківські кредити; X3 – резерв під заборгованість банків; X4 – кредити юридичним особам; X5 – кредити фізичним особам; X6 – резерв під кредити та заборгованість клієнтів; X7 – цінні папери.

	1 bank	2 KIP	3 Mezhb krediti	4 Zadolzh bankov	5 Krediti Ur licam	6 Krediti fiz licam	7 Rezervi klientam	8 Cennie bumagi
1	ІНГ БАНК УКРАЇНА	8546,34	234,19	17,04	7720,5	95,52	334,13	847,3
2	АКТИВ-БАНК	2916,65	364,94	0,38	1800,91	842,3	99,21	8,1
3	АЛЬФА-БАНК	19595,91	2019,04	16,34	16856,73	4880,14	5484,73	1341,07
4	БАНК КРЕДИТ-ДНІПРО	5569,11	402,32	0,29	5028,75	466,24	482,93	155,02
5	БАНК ТАВРИКА	4151,16	1064,12	1,35	2923,92	256,87	97,56	5,17
6	БРОКБІЗНЕСБАНК	12689,63	1466,98	2,46	9224,82	2302,17	728,17	426,3
7	ВІЕЙБІ Банк	6252,38	1508,12	5,51	4098,07	1699,09	1452,7	405,31
8	ВТБ БАНК	28552,02	716,57	3,24	28021,71	3333,64	5344,58	1827,91
9	ДЕЛЬТА БАНК	15558,31	1456,12	40,26	9028,93	5937,46	4033,84	3209,91
10	ЗЛАТОБАНК	2873,32	694,58	7,37	2331,77	172,17	317,89	0,06
11	КІБ Креді Агріколь	2886,81	70	1,5	3041,47	24,34	251,5	4
12	КИЇВСЬКА РУСЬ	4277,19	505,93	0,97	3408,63	399,74	226,49	190,34
13	КРЕДІ АГРІКОЛЬ БАНК	5204,68	193	1,26	3240,6	2165,96	458,11	64,5
14	КРЕДИТПРОМБАНК	9833,73	586,71	3,74	8688,12	2169,93	2323,31	716,01
15	МАРФІН БАНК	3293,71	462,52	2,64	2156,65	905,71	664,49	435,97
16	МЕГАБАНК	3568,92	322,78	0,29	2799,14	522,81	183,24	107,71
17	ОТП БАНК	19516,59	269,41	31,87	12920,34	6862,48	3104,15	2600,38
18	ПІВДЕНКОМБАНК	3306,78	239,59	9,98	2670,44	64,47	147,27	489,53
19	ПІВДЕННИЙ	8358,6	484,66	1,93	7833,81	619,44	593,89	16,5
20	ПЛАТИНУМ БАНК	2990,49	922,54	34,3	31,57	2406,95	337,45	1,17
21	ПРАВЕКС-БАНК	4036,57	137,99	1,38	1677,4	3126,8	905,31	1,07
22	ПРОМІНВЕСТБАНК	30466,77	1275,58	25,87	28022,21	792,63	2191,17	2593,4
23	ПУМБ	23409,44	4781,8	32,03	13853,77	4604,9	4121,16	4322,17
24	СІТІБАНК УКРАЇНА	4404,05	3,89	0,04	2112,58	56,94	63,67	2294,35
25	СБЕРБАНК РОСІЇ	13853	155,45	9,75	13707,79	1228,61	2667,47	1438,37
26	СОЮЗ	3763,37	604,46	0,67	3324,15	101,89	372,85	106,4
27	УКРІНБАНК	3185,2	678,96	2,39	2384,71	315,72	199,55	7,74

Рис. 25. **Вхідні дані**

Для побудови кластерних утворень нормуємо значення показників. З цією метою в контекстному меню необхідно обрати Edit/Fill/Standardize Block/Standardize Columns. Нормовані значення наведені на рис. 26.

Spreadsheet12* (8v by 27c)

	1 bank	2 KIP	3 Mezhb krediti	4 Zadolzh bankov	5 Krediti Ur licam	6 Krediti fiz licam	7 Rezervi klientam	8 Cennie bumagi
1	ІНГ БАНК УКРАЇНА	-0,09985	-0,59825055	0,613155395	0,047606959	-0,8394993	-0,62241669	-0,02312681
2	АКТИВ-БАНК	-0,78017	-0,46020509	-0,7307493	-0,74964381	-0,452828	-0,7625852	-0,73254172
3	АЛЬФА-БАНК	1,235438	1,28618871	0,556688811	1,27807501	1,63790366	2,45076534	0,394280014
4	БАНК КРЕДИТ-ДНІПРО	-0,45964	-0,4207394	-0,73800928	-0,31491811	-0,6475461	-0,53363296	-0,60834339
5	БАНК ТАВРИКА	-0,63099	0,277987046	-0,65250274	-0,59839675	-0,7559547	-0,7635697	-0,73501859
6	БРОКБІЗНЕСБАНК	0,400846	0,703325421	-0,56296288	0,250208874	0,3030703	-0,38730687	-0,37901775
7	ВІЕЙБІ Банк	-0,37707	0,74676091	-0,3169299	-0,44026215	-0,0091953	0,044994718	-0,39676158
8	ВТБ БАНК	2,317743	-0,0889552	-0,50004297	2,78177524	0,83714966	2,36714276	0,805828581
9	ДЕЛЬТА БАНК	0,747512	0,691859466	2,48623264	0,223826396	2,18536775	1,58507026	1,97409766
10	ЗЛАТОБАНК	-0,78541	-0,11217217	-0,16689012	-0,67814755	-0,7998111	-0,63210653	-0,73933831
11	КІБ Креді Агріколь	-0,78378	-0,77160186	-0,64040276	-0,58256511	-0,8763552	-0,67171911	-0,73600764
12	КИЇВСЬКА РУСЬ	-0,61576	-0,31134828	-0,68315603	-0,53311597	-0,6819788	-0,6866417	-0,57848575
13	КРЕДІ АГРІКОЛЬ БАНК	-0,50368	-0,64173883	-0,65976273	-0,55574626	0,23254285	-0,54844219	-0,68486417
14	КРЕДИТПРОМБАНК	0,055723	-0,22606099	-0,45970969	0,177926083	0,23459846	0,564457112	-0,13411237
15	МАРФІН БАНК	-0,73461	-0,35718042	-0,5484429	-0,70173272	-0,4199953	-0,42530249	-0,37084325
16	МЕГАБАНК	-0,70135	-0,50471749	-0,73800928	-0,61520213	-0,618255	-0,71244746	-0,64833674
17	ОТП БАНК	1,225852	-0,56106538	1,8094403	0,747921749	2,66432893	1,0303569	1,45883352
18	ПІВДЕНКОМБАНК	-0,73303	-0,59254925	0,043649565	-0,63253545	-0,8555765	-0,73390949	-0,32556648
19	ПІВДЕННИЙ	-0,12254	-0,33380508	-0,60571615	0,062867557	-0,5682215	-0,46742703	-0,72544081
20	ПЛАТИНУМ БАНК	-0,77125	0,128507308	2,00546001	-0,98793863	0,35732378	-0,62043577	-0,73839997
21	ПРАВЕКС-БАНК	-0,64484	-0,69981822	-0,65008275	-0,76627814	0,73005108	-0,28161366	-0,73848451
22	ПРОМІНВЕСТБАНК	2,549132	0,501245878	1,32544101	2,78184258	-0,4785464	0,485613818	1,452933
23	ПУМБ	1,696285	4,2031024	1,82234695	0,873636162	1,49538861	1,63717103	2,91434294
24	СІТІБАНК УКРАЇНА	-0,60043	-0,8414006	-0,75817592	-0,70766807	-0,8594754	-0,78379067	1,20013208
25	СБЕРБАНК РОСІЇ	0,541433	-0,681384	0,025096259	0,853975566	-0,2528026	0,769805291	0,476532244
26	СОЮЗ	-0,67785	-0,2073206	-0,707356	-0,54449375	-0,836201	-0,59931383	-0,64944415
27	УКРІНБАНК	-0,74772	-0,12866372	-0,56860953	-0,67101759	-0,7254831	-0,70271585	-0,73284605

Рис. 26. **Нормовані значення показників кредитно-інвестиційного портфеля**

Для проведення кластерного аналізу необхідно скористатися меню Statistics/ Multivariate Exploratory/Cluster analysis після чого з'явиться діалогове вікно (рис. 27), що дозволяє вибрати один з методів кластеризації.

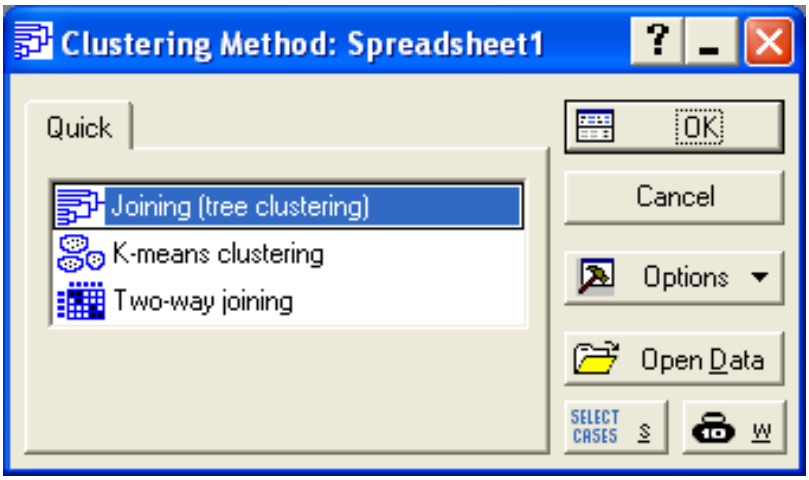


Рис. 27. **Діалогове вікно модуля кластерний аналіз**

Діалогове вікно, що з’явилося дозволяє використати один з методів кластеризації: **Joining (tree clustering)** – Об'єднання (деревоподібна кластеризація); **K – means clustering** – Кластеризація методом К-середніх; **Two-way joining** – Двовхідне об'єднання.

Розглянемо Об'єднання (деревоподібна кластеризація) – **Joining (tree clustering).** У вікна, в якому обираються параметри кластеризації, вибираємо Variables (Змінні), рис. 28.

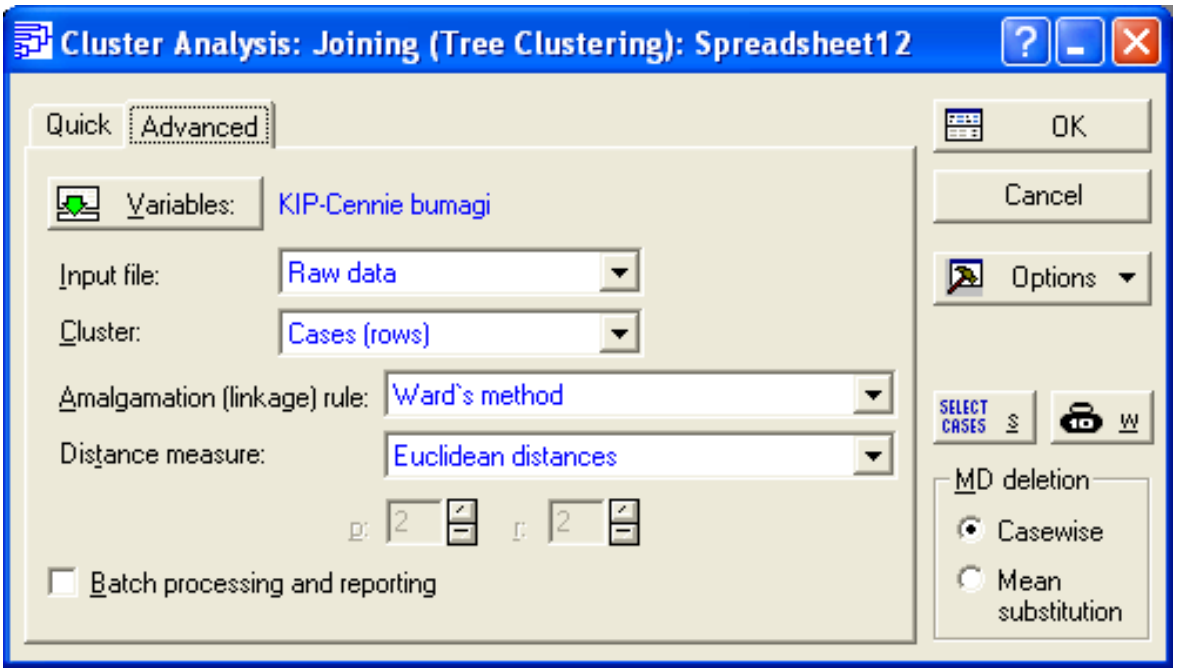


Рис. 28. **Діалогове вікно параметрів кластеризації**

Input file (Вихідні дані) (рис.28) становить меню, в якому обираємо Raw data (Вихідні дані). Distance matrix (Матриця відстаней) передбачена на той випадок, якщо вхідна інформація представлена у вигляді матриці подібності.

У полі **Cluster** (Кластер) обирається напрям класифікації. При кластеризації самих змінних обирається Variables [Columns] (Змінні [Стовпці]), в даній задачі обирається класифікація за спостереженнями Cases [rows] (Спостереження [рядки]).

Рядок **Amalgamation [linkage] rule** (Правило об’єднання [зв’язки]) містить установки для вибору таких мір подібності: **Single Linkage** (Метод одиночного зв'язку "принцип найближчого сусіда"). **Complete Linkage** (Метод повного зв'язку "принцип далекого сусіда”). **Unweighted pair-group average** (Незважене попарне середнє). **Weighted pair-group average** (Зважене попарне середнє). **Unweighted pair-group centroid**

(Незважений центроїдний метод). **Weighted pair-group centroid** (Зважений центроїдний метод). **Ward's method** (Метод Варда).

Для вирішення даної задачі вибираємо метод Варда.

У полі **Distance measure** (Міра відстані) (рис. 28) пропонуються різні види відстаней: **Squared Euclidean distances** (квадрат Евклідової відстані); **Euclidean distances** (Евклідова відстань); **City-block (Manhattan) distance** (Відстань міських кварталів (Манхеттенська відстань)); **Chebychev distance metric** (Відстань Чебишева); **Power: SUM(ABS(x-y)**p)**1/r** (Степенева відстань); **Percent disagreement** (Відсоток незгоди).

Для вирішення поставленого завдання оберемо **City-block (Manhattan) distance** (Відстань міських кварталів (Манхеттенська відстань)).

Після встановлення всіх параметрів кластеризації переходимо до вікна її результатів (рис. 29).

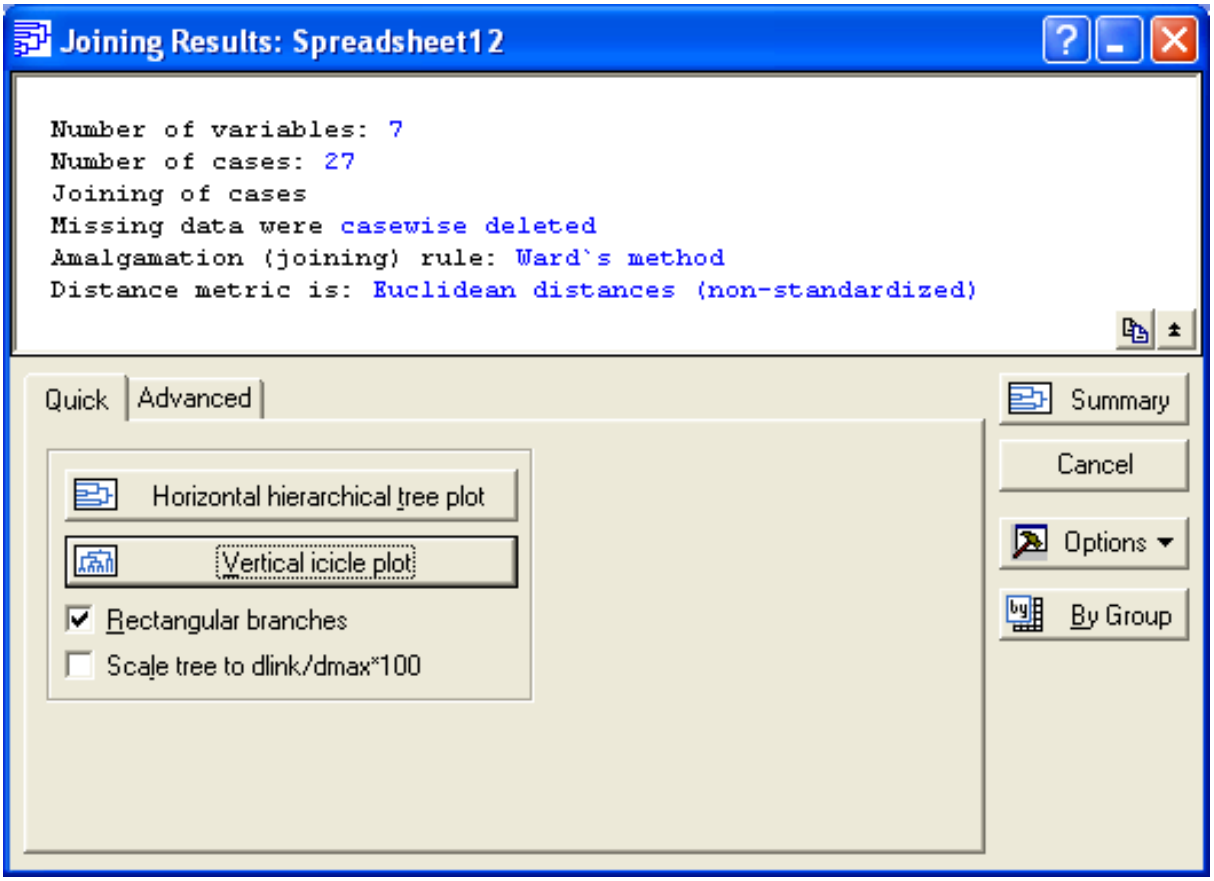


Рис. 29. **Вікно результатів кластеризації**

Далі за допомогою кнопки **Vertical icicle plot** будуємо вертикальну дендрограму (рис. 30).

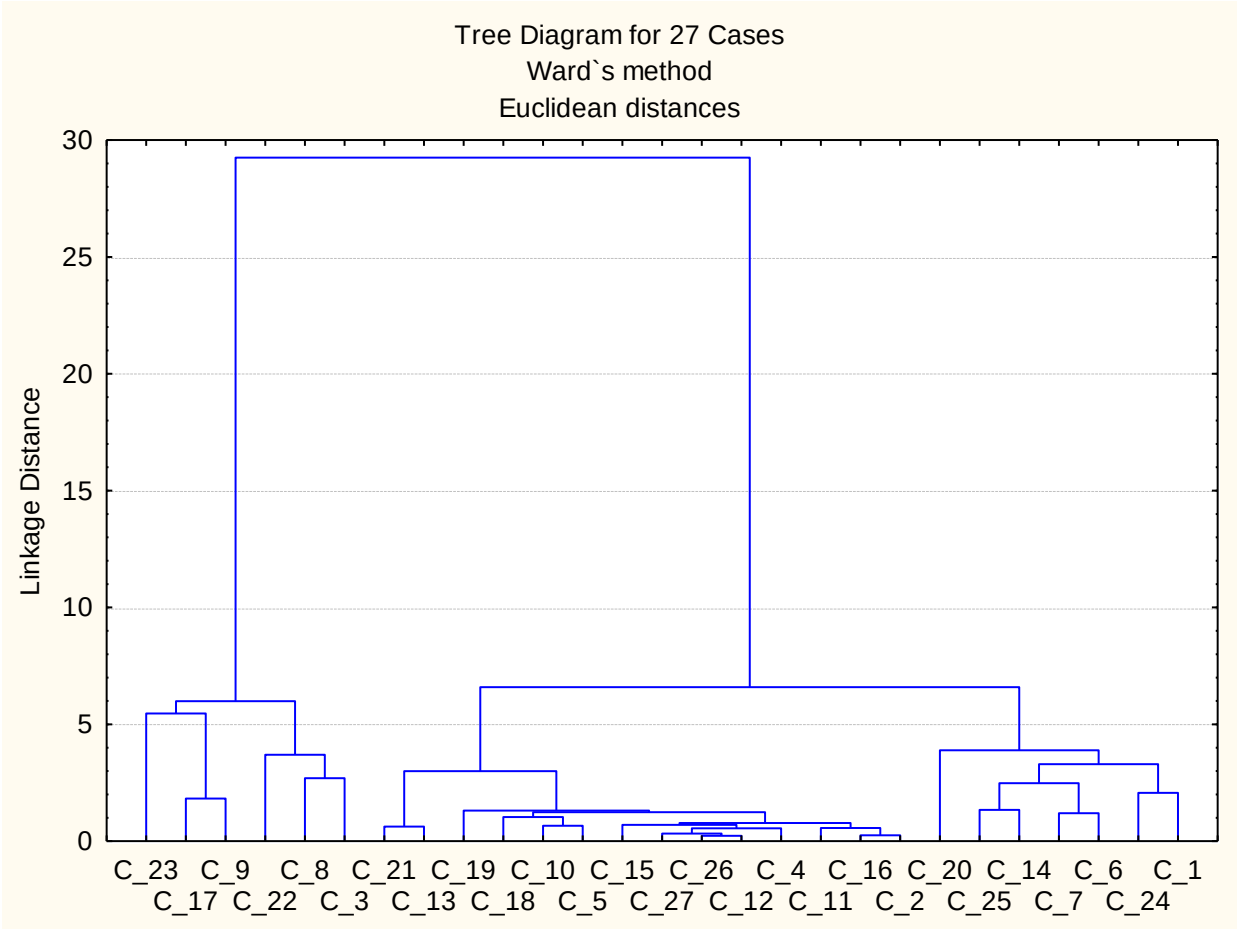


Рис. 30. **Вертикальна дендрограма**

На дендрограмі видно, що всю сукупність банків доцільно розбити на 3 групи.

Іншим методом кластеризації є метод **K – means clustering** (К-середніх). Відмінністю даного методу від попереднього є те, що цей метод використовується коли користувач має уяву про кількість кластерів. Діалогове вікно даного методу викликається вибором на рис. 27.

За результатами дендрограми розіб'ємо сукупність на 3 кластери, що відзначимо в полі **Number of clusters** (Число кластерів) (рис. 31).

За допомогою кнопки **Variables** (Змінні) оберем показники, за якими проводитиметься кластеризація. У рядку **Cluster** (Кластер) зазначимо об'єкти для класифікації Cases [rows] (Спостереження [строки]).

Оскількі метод k-середніх є ітераційним, для зазначення максимального числа ітерацій, у результаті яких на кожній ітерації об'єкти розміщуються в різні кластери, призначено поле **Number of iterations** (Число ітерацій).

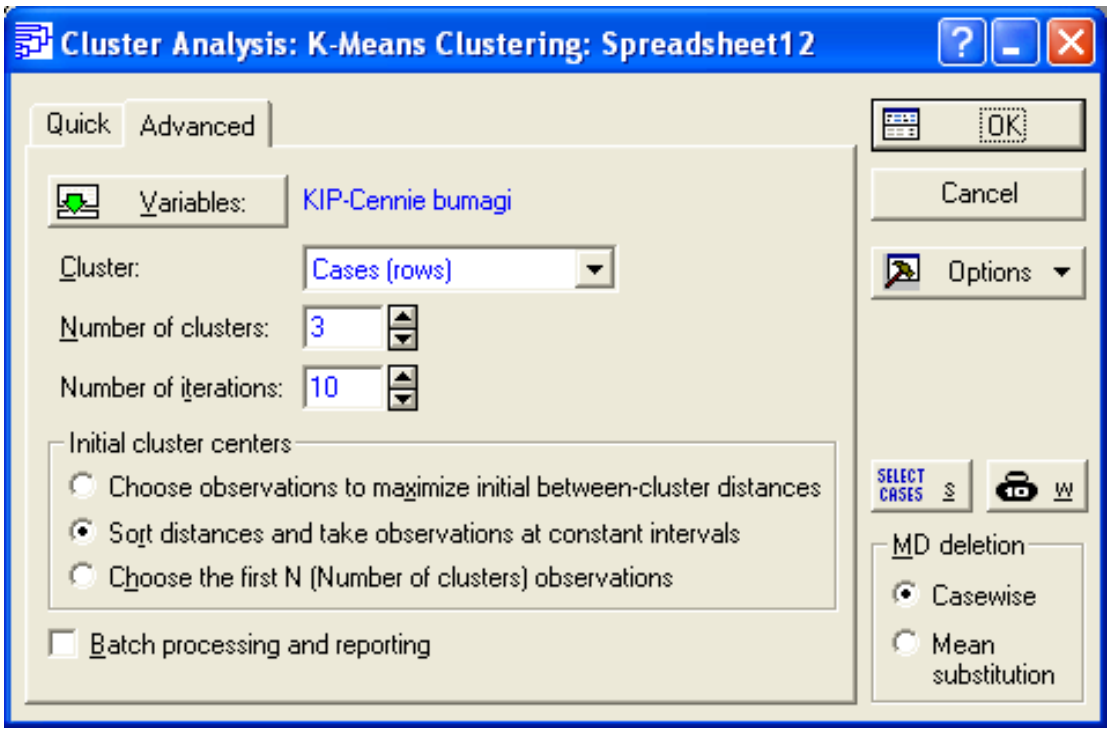


Рис. 31. **Діалогове вікно модуля**
Кластерний аналіз: Метод k-середніх

Опція **Choose observations to maximize initial between-cluster distances** (Обрати спостереження, максимізуючи початкові відстані між кластерами) обирає перші k спостережень відповідно до кількості кластерів, спостережень, які є центрами кластерів.

Подальші спостереження замінюють раніше обрані центри в тому випадку, якщо найменша відстань до будь-якого з них більше, ніж найменша відстань між кластерами. В результаті цієї процедури початкові відстані між кластерами максимізують.

У даному випадку обрана опція **Sort distances and take observations at constant intervals** (Сортувати відстані і вибрати спостереження на постійних інтервалах), що значить спочатку сортування відстані між усіма об'єктами, а потім в якості початкових центрів кластерів вибір спостережень на постійних інтервалах.

Наступна опція **Choose the first N** (Number of cluster) (Обрати перші N [кількість кластерів] спостережень). Ця опція бере перші N (кількість кластерів) спостережень в якості початкових центрів кластерів.

Після натиснення кнопки OK Statistica здійснить обчислення і з'явиться нове вікно: "K-Means Clustering Results" (рис. 32).

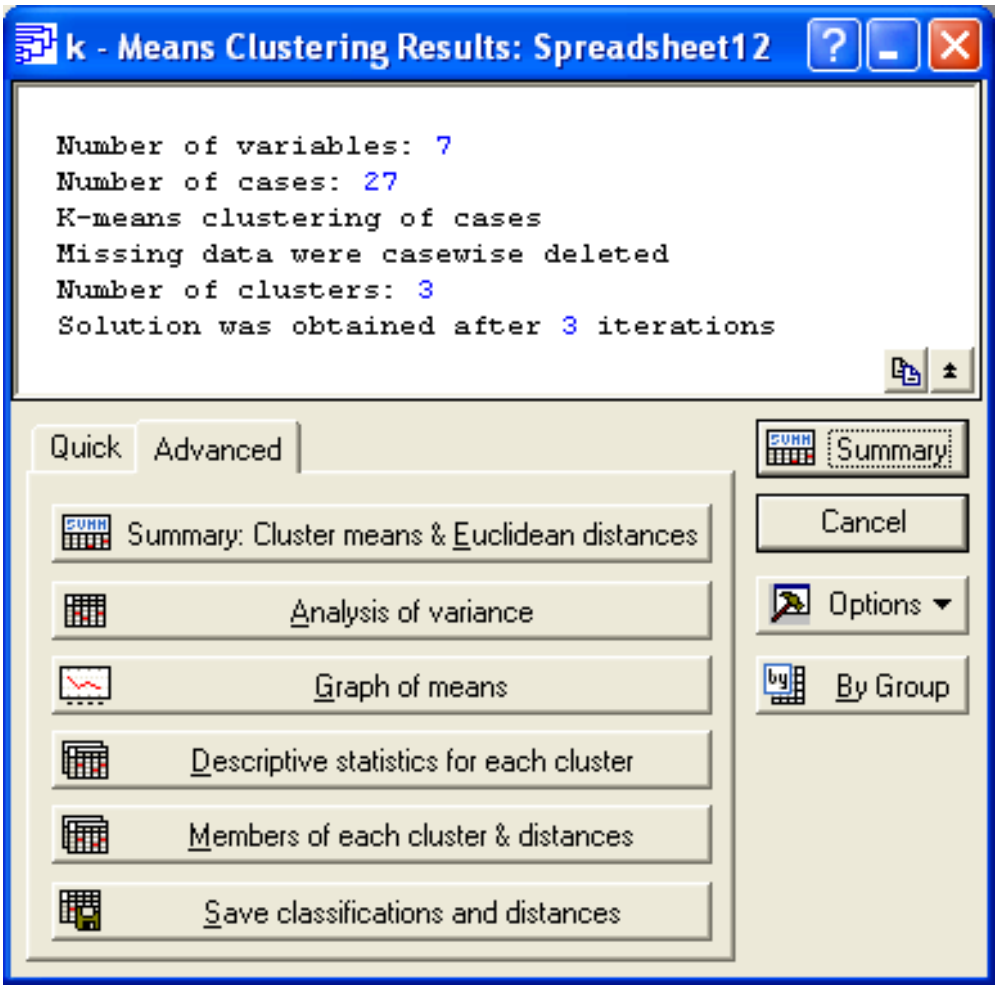


Рис. 32. **Вікно результатів кластеризації**

Вікно з результатами класифікації умовно можливо розділити на дві частини. У верхній частині містяться значення параметрів, за якими проводиться аналіз, а в нижній – кнопки для виведення результатів.

У верхній частині вікна: кількість змінних – 7; кількість спостережень - 27; класифікація спостережень (або змінних, залежить від установки в попередньому вікні в рядку Cluster) методом K – середніх; спостереження з пропущеними даними видаляються або змінюються середніми значеннями; кількість кластерів – 3; рішення досягнуто після 3 ітерацій.

За допомогою кнопки **Cluster Means & Euclidean Distances** (середні значення в кластерах та евклідові відстані) на екран виводяться дві таблиці. У першій таблиці (рис. 33) наведено відстані між класами. Над діагональними елементами, які дорівнюють нулю, вказані квадрати, а нижче – евклідові відстані.

У другій таблиці (рис. 34) вказані середні величини кластера по всім змінним (спостереженням). По вертикалі вказані номери кластерів, а по горизонталі змінні.

Cluster Number	Euclidean Distances between C Distances below diagonal Squared distances above diagor		
	No. 1	No. 2	No. 3
No. 1	0,000000	3,898457	2,082241
No. 2	1,974451	0,000000	0,311362
No. 3	1,442997	0,557998	0,000000

Рис. 33. **Матриця відстаней між кластерами**

Variable	Cluster Means (Spreadsheet12)		
	Cluster No. 1	Cluster No. 2	Cluster No. 3
KIP	1,628660	-0,634786	-0,041695
Mezhb krediti	1,005396	-0,407018	0,012150
Zadolzh bankov	1,250018	-0,586948	0,217352
Krediti Ur licam	1,447846	-0,572573	-0,016414
Krediti fiz licam	1,390265	-0,542339	-0,034417
Rezervi klientam	1,592687	-0,620348	-0,041817
Cennie bumagi	1,500053	-0,520362	-0,199148

Рис. 34. **Значення середніх величин кластерів**

Натиснувши кнопку **Graph of means** (Графік середніх), можна отримати графічне зображення середніх значень змінних для кожного кластера (рис. 35).

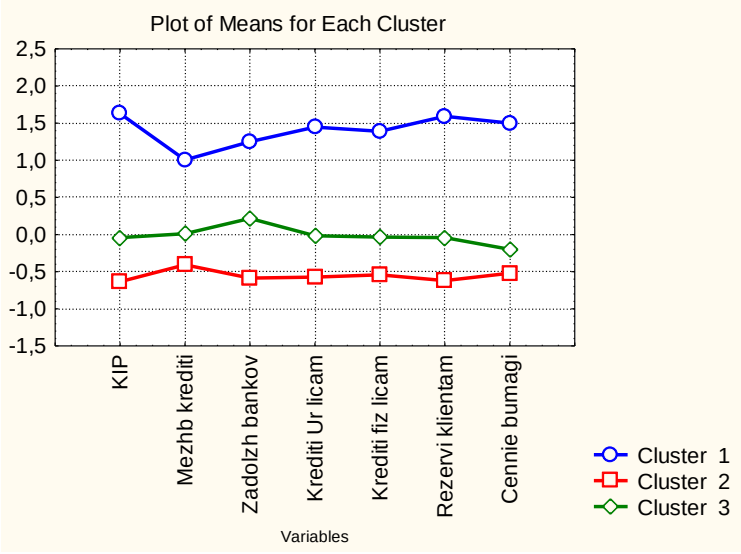


Рис. 35. **Графік середніх значень змінних для кожного кластера**

Перелік спостережень, що входять до кожного з кластерів, можна отримати з використанням кнопки **Members for each cluster & distances** (члени групи і відстані) (рис. 36).

Так, до першого кластера увійшли Альфа-банк, ВТБ Банк, Дельта Банк, ОТП Банк, Промінвестбанк та ПУМБ.

До другого кластера увійшли такі банки: Актив-банк, банк Кредит-Дніпро, банк Таврика, Злотобанк, КІБ Креді Агріколь, Марфін банк, Мегабанк, Південкомбанк, Південний, Правекс-банк, Сітібанк Україна, СОЮЗ та Укрінбанк.

А ті, що залишились, об'єднано у третій кластер: ІНГ банк Україна, Брокбізнесбанк, Віейбі Банк, Кредитпромбанк, Платинум Банк та Сбербанк Росії.

Кнопка **Descriptive Statistics for each cluster** дозволяє визначити описові статистики для кожного з кластерів (рис. 37).

Case No.	Members and Distances Cluster co **Distance**
C_3	0,628303
C_8	1,062489
C_9	0,824460
C_17	0,875055
C_22	1,042453
C_23	1,357541

Case No.	Members and Distances Cluster cor **Distance**
C_2	0,146025
C_4	0,144618
C_5	0,289441
C_10	0,242486
C_11	0,213403
C_12	0,082462
C_13	0,318600
C_15	0,123045
C_16	0,099453
C_18	0,291371
C_19	0,324699
C_21	0,522171
C_24	0,689184
C_26	0,151412
C_27	0,162893

Case No.	Members and Distances Cluster cor **Distance**
C_1	0,471012
C_6	0,480945
C_7	0,407861
C_14	0,379206
C_20	0,883664
C_25	0,630027

Рис.36. **Спостереження, що входять до 1, 2, та 3 кластерів відповідно**

Variable	Descriptive Statistics for Cluster 1 Cluster contains 6 cases		
	Mean	Standard Deviation	Variance
KIP	1,628660	0,695697	0,483995
Mezhb krediti	1,005396	1,691629	2,861610
Zadolzh bankov	1,250018	1,069378	1,143570
Krediti Ur licam	1,447846	1,086791	1,181114
Krediti fiz licam	1,390265	1,106905	1,225239
Rezervi klientam	1,592687	0,758939	0,575989
Cennie bumagi	1,500053	0,886965	0,786707

Variable	Descriptive Statistics for Cluster 2 (Cluster contains 15 cases		
	Mean	Standard Deviation	Variance
KIP	-0,634786	0,172503	0,029757
Mezhb krediti	-0,407018	0,292763	0,085710
Zadolzh bankov	-0,586948	0,225532	0,050865
Krediti Ur licam	-0,572573	0,207644	0,043116
Krediti fiz licam	-0,542339	0,448926	0,201534
Rezervi klientam	-0,620348	0,144443	0,020864
Cennie bumagi	-0,520362	0,493671	0,243711

Variable	Descriptive Statistics for Cluster 3 (Cluster contains 6 cases		
	Mean	Standard Deviation	Variance
KIP	-0,041695	0,488565	0,238696
Mezhb krediti	0,012150	0,623095	0,388248
Zadolzh bankov	0,217352	0,974034	0,948743
Krediti Ur licam	-0,016414	0,631513	0,398808
Krediti fiz licam	-0,034417	0,455239	0,207243
Rezervi klientam	-0,041817	0,604168	0,365019
Cennie bumagi	-0,199148	0,413268	0,170791

Рис. 37. **Описові статистики для кожного кластера**

Ще один метод, який використовується для кластеризації є **Two-way joining (Двовхідного об'єднання).** Основна відмінність даного методу полягає в тому, щоб одночасно класифікувати як спостереження, так і змінні. Але недоліком є те, що кластери, що отримуються є досить часто неоднорідними за своєю природою. Діалогове вікно даного методу викликається вибором на рис. 37 і наведено на рис.38.

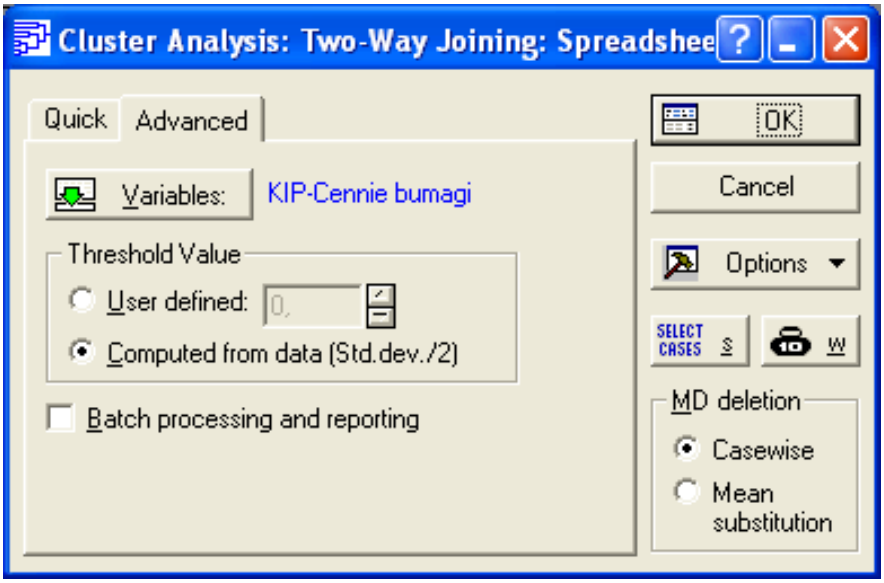


Рис. 38. **Діалогове вікно методу Двовхідного об'єднання**

Група операцій **Threshold Value (Значення порогу)** містить два режими: **User defined (Заданий користувачем) і Computed from data (Std. Dev./2) (Обчислене за даними).**

Пороговий параметр визначає належність елементів матриці даних до кластерів, що формуються. Якщо ця величина дуже велика порівняно зі значеннями елементів у матриці, то формується тільки один кластер; якщо дуже мала, то кластером буде кожна точка даних.

Для більшості випадків беруть порогове значення, рівне половині величини загального стандартного відхилення (режим **Computed from data (Std. Dev./2) (Обчислене за даними))**. Після завдання всіх параметрів натиснемо Ok.

Вікно з результатами обчислень має такий вигляд (рис.39).

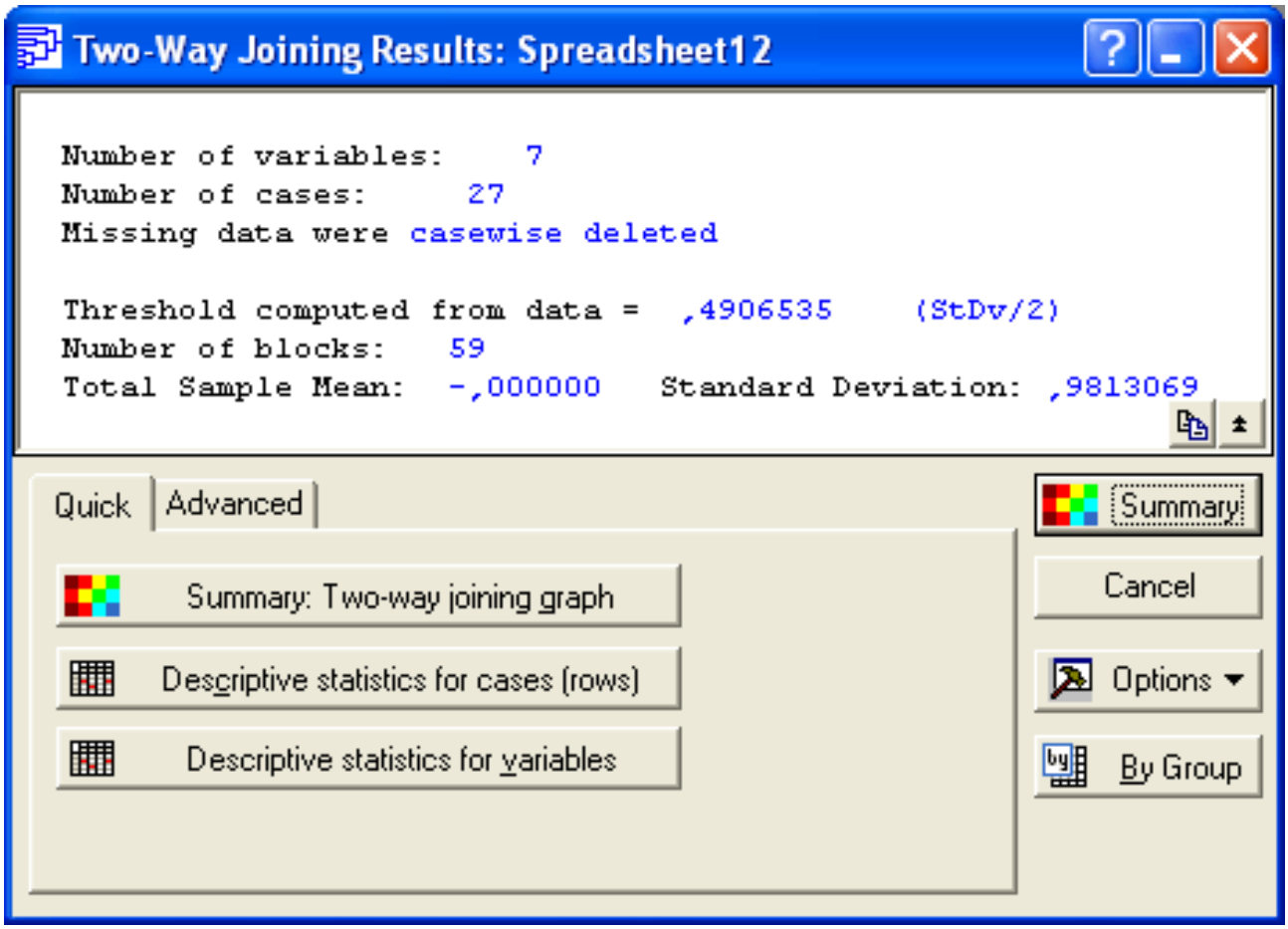


Рис. 39. **Вікно результатів Двовхідного методу**

У даному вікні кнопка **Descriptive statistics for cases (rows)** (Описові статистики [рядків]) і **Descriptive statistics for variables (Описові статистики для змінних)** – виводять на екран таблиці зі

статистичними характеристиками даних: середніми значеннями і стандартними відхиленнями (рис. 40).

Case ID	Means and Standard D	
	Mean	Std.Dev.
C_1	-0,217483	0,501282
C_2	-0,666961	0,144781
C_3	1,262763	0,683912
C_4	-0,531832	0,145121
C_5	-0,551206	0,371309
C_6	0,046880	0,483908
C_7	-0,106923	0,423007
C_8	1,217234	1,287769
C_9	1,413424	0,863121
C_10	-0,559125	0,292893
C_11	-0,723205	0,099093
C_12	-0,584355	0,134123
C_13	-0,480241	0,321233
C_14	0,030403	0,337235
C_15	-0,508301	0,156344
C_16	-0,648331	0,079153
C_17	1,196524	0,993123
C_18	-0,547074	0,308775
C_19	-0,394326	0,281799
C_20	-0,089533	1,049681
C_21	-0,435866	0,539001
C_22	1,231094	1,169861
C_23	2,091753	1,111270
C_24	-0,478687	0,745369
C_25	0,247522	0,568926
C_26	-0,603140	0,197075
C_27	-0,611008	0,221022

Variable	Means and Standard D	
	Mean	Std.Dev.
KIP	0,000000	1,000000
Mezhb krediti	0,000000	1,000000
Zadolzh bankov	0,000000	1,000000
Krediti Ur licam	-0,000000	1,000000
Krediti fiz licam	0,000000	1,000000
Rezervi klientam	-0,000000	1,000000
Cennie bumagi	-0,000000	1,000000

Рис. 40. **Статистичні характеристики отриманих кластерів**

Опція **Reordered statistics for variables (Перевпорядкована матриця даних)** формує таблицю зі спостереженнями відповідно до результатів двовхідного об'єднання (рис. 41).

Кнопка **Summary: Two-way joining graph** на рис. 39 виводить графічне зображення результатів двовхідного об'єднання. В даному випадку перевпорядкована матриця даних відображається у вигляді карти ліній рівня (рис. 42).

Case ID	Reordered Data Matrix (Spreadsheet12)						
	KIP	Krediti Ur licam	Rezervi klientam	Zadolzh bankov	Cennie bumagi	Krediti fiz licam	Mezhb krediti
C_1	-0,099852	0,047607	-0,622417	0,613155	-0,023127	-0,839499	-0,598251
C_19	-0,122540	0,062868	-0,467427	-0,605716	-0,725441	-0,568222	-0,333805
C_4	-0,459637	-0,314918	-0,533633	-0,738009	-0,608343	-0,647546	-0,420739
C_12	-0,615759	-0,533116	-0,686642	-0,683156	-0,578486	-0,681979	-0,311348
C_26	-0,677852	-0,544494	-0,599314	-0,707356	-0,649444	-0,836201	-0,207321
C_2	-0,780174	-0,749644	-0,762585	-0,730749	-0,732542	-0,452828	-0,460205
C_16	-0,701350	-0,615202	-0,712448	-0,738009	-0,648337	-0,618255	-0,504718
C_15	-0,734608	-0,701733	-0,425303	-0,548443	-0,370843	-0,419995	-0,357180
C_13	-0,503676	-0,555746	-0,548442	-0,659763	-0,684864	0,232543	-0,641739
C_21	-0,644837	-0,766278	-0,281614	-0,650083	-0,738484	0,730051	-0,699818
C_18	-0,733029	-0,632536	-0,733909	0,043650	-0,325566	-0,855576	-0,592549
C_25	0,541434	0,853976	0,769805	0,025096	0,476532	-0,252803	-0,681384
C_3	1,235438	1,278075	2,450765	0,556689	0,394280	1,637904	1,286189
C_5	-0,630989	-0,598397	-0,763570	-0,652503	-0,735019	-0,755955	0,277987
C_11	-0,783780	-0,582565	-0,671719	-0,640403	-0,736008	-0,876355	-0,771602
C_24	-0,600429	-0,707668	-0,783791	-0,758176	1,200132	-0,859475	-0,841401
C_10	-0,785411	-0,678148	-0,632107	-0,166890	-0,739338	-0,799811	-0,112172
C_27	-0,747721	-0,671018	-0,702716	-0,568609	-0,732846	-0,725483	-0,128664
C_20	-0,771251	-0,987939	-0,620436	2,005460	-0,738400	0,357324	0,128507
C_6	0,400846	0,250209	-0,387307	-0,562963	-0,379018	0,303070	0,703326
C_14	0,055723	0,177926	0,564457	-0,459710	-0,134112	0,234598	-0,226061
C_7	-0,377067	-0,440262	0,044995	-0,316930	-0,396762	-0,009195	0,746761
C_9	0,747512	0,223826	1,585070	2,486233	1,974098	2,185368	0,691859
C_23	1,696285	0,873636	1,637171	1,822347	2,914343	1,495389	4,203103
C_8	2,317743	2,781775	2,367143	-0,500043	0,805829	0,837150	-0,088955
C_22	2,549131	2,781842	0,485614	1,325441	1,452933	-0,478546	0,501246
C_17	1,225852	0,747922	1,030357	1,809440	1,458833	2,664329	-0,561065

Рис. 41. **Перевпорядкована матриця даних**

На графіку по горизонталі відкладені беруть участь у класифікації змінні, а по вертикалі – спостереження. Кольори клітинок, що знаходяться на перетині, вказують на належність елементів матриці до певного кластера. У даному прикладі видно, що, використовуючи цей метод, усю сукупність можливо розподілити на п’ять кластерів. Досить складна інтерпретація результатів і спірна їх практична цінність робить розглянутий метод не дуже привабливим.

Спираючись на дендрограму (рис. 30) та результати методу k-середніх робимо висновок про те, що всю сукупність банків можливо розподілити на 3 кластери.

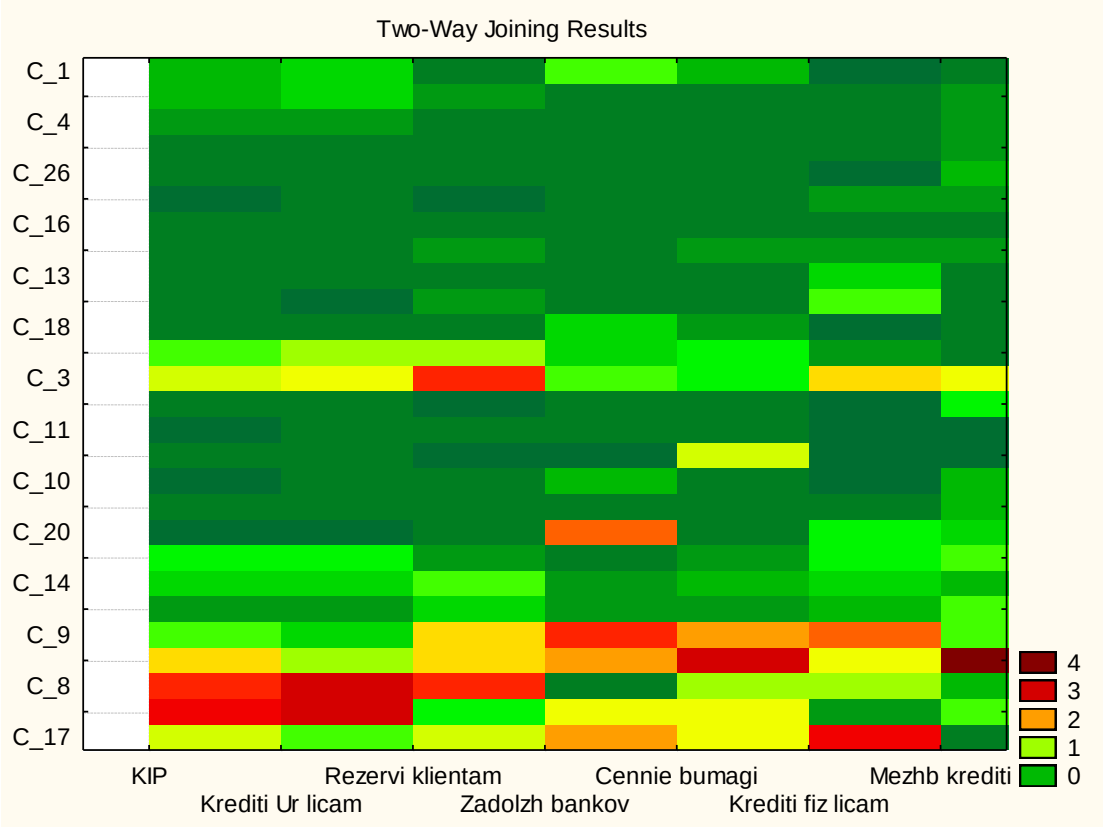


Рис. 42. **Графічний результат методу двовхідного об'єднання**

Отриманим кластерам можливо надати таку інтерпретацію (табл. 2).

Таблиця 2

Інтерпретація результатів кластерного аналізу

Номер кластера	Розмір кредитно-інвестиційного портфеля	Назва банку
Кластер № 1	великий	Альфа-банк, ВТБ Банк, Дельта Банк, ОТП банк, Промінвестбанк та ПУМБ
Кластер № 2	низький	Актив-банк, банк Кредит-Дніпро, банк Таврика, Златобанк, КІБ Креді агріколь, Марфін банк, Мегабанк, Південкомбанк, Південний, Правекс-банк, Сітібанк Україна, СОЮЗ та Укрінбанк
Кластер № 3	середній	ІНГ банк Україна, Брокбізнесбанк, Віейбі Банк, Кредитпромбанк, Платинум банк та Сбербанк Росії

Таким чином, можна зробити висновок, що інвестувати капітал найкраще в банки, які увійшли до кластера № 1, який характеризується великим розміром кредитно-інвестиційного портфеля.

Модуль 2. Сучасні методи аналізу бізнес-процесів

Лабораторне заняття на тему "Застосування методів дискримінантного аналізу перевірки класифікації, отриманої при використанні апарату кластерного аналізу"

Мета роботи – отримання навичок використання дискримінантного аналізу в пакеті Statistica.

Завдання – перевірити якість кластеризації методами дискримінантного аналізу.

Методичні рекомендації

У результаті попередньої лабораторної роботи методами кластерного аналізу розподілено 27 банків України за трьома групами (за розміром кредитно-інвестиційного портфеля) є вихідними даними для виконання лабораторної роботи, які подано на рис. 43. Необхідно перевірити якість класифікації методами дискримінантного аналізу. Та після отримання підтвердження класифікації віднести такі банки до класифікаційних груп.

Таблиця 3

Дані для класифікації

Назва банку	Міжбанківські кредити	Заборгованість банків	Кредити юридичним особам	Кредити фізичним особам	Резерви клієнтам	Цінні папери
ГРАНТ	9,94	0,01	362,08	94,37	7,41	4,41
КРЕДОБАНК	105,21	6,77	1338,77	615,17	248,31	631,88

Вибір даного модуля можливо здійснити через меню **Statistics/ Multivariate Exploratory Techniques/Discriminant analysis** (рис. 44). Або можливо використовувати перемикач **Module Switcher**, який містить перелік усіх доступних модулів та натиснути **Discriminant Analysis,** а потім кнопку Switch to.

На екрані з'явиться стартова панель модуля **Discriminant Function Analysis** (рис. 44), яка містить такі функціональні кнопки:

Variables дозволяє вибрати **Grouping** (Групувальні змінні) і **Independent** (Незалежні змінні).

	1 bank	2 Mezhb krediti	3 Zadolzh bankov	4 Krediti Ur licam	5 Krediti fiz licam	6 Rezervi klientam	7 Cennie bumagi	8 Nomer klastera
1	ІНГ БАНК УКРАЇНА	234,19	17,04	7720,5	95,52	334,13	847,3	3
2	АКТИВ-БАНК	364,94	0,38	1800,91	842,3	99,21	8,1	2
3	АЛЬФА-БАНК	2019,04	16,34	16856,73	4880,14	5484,73	1341,07	1
4	БАНК КРЕДИТ-ДНІПРО	402,32	0,29	5028,75	466,24	482,93	155,02	2
5	БАНК ТАВРИКА	1064,12	1,35	2923,92	256,87	97,56	5,17	3
6	БРОКБІЗНЕСБАНК	1466,98	2,46	9224,82	2302,17	728,17	426,3	3
7	ВІЕЙБІ Банк	1508,12	5,51	4098,07	1699,09	1452,7	405,31	3
8	ВТБ БАНК	716,57	3,24	28021,71	3333,64	5344,58	1827,91	1
9	ДЕЛЬТА БАНК	1456,12	40,26	9028,93	5937,46	4033,84	3209,91	1
10	ЗЛАТОБАНК	694,58	7,37	2331,77	172,17	317,89	0,06	2
11	КІБ Креді Агріколь	70	1,5	3041,47	24,34	251,5	4	2
12	КИЇВСЬКА РУСЬ	505,93	0,97	3408,63	399,74	226,49	190,34	3
13	КРЕДІ АГРІКОЛЬ БАНК	193	1,26	3240,6	2165,96	458,11	64,5	3
14	КРЕДИТПРОМБАНК	586,71	3,74	8688,12	2169,93	2323,31	716,01	3
15	МАРФІН БАНК	462,52	2,64	2156,65	905,71	664,49	435,97	2
16	МЕГАБАНК	322,78	0,29	2799,14	522,81	183,24	107,71	2
17	ОТП БАНК	269,41	31,87	12920,34	6862,48	3104,15	2600,38	1
18	ПІВДЕНКОМБАНК	239,59	9,98	2670,44	64,47	147,27	489,53	2
19	ПІВДЕННИЙ	484,66	1,93	7833,81	619,44	593,89	16,5	2
20	ПЛАТИНУМ БАНК	922,54	34,3	31,57	2406,95	337,45	1,17	3
21	ПРАВЕКС-БАНК	137,99	1,38	1677,4	3126,8	905,31	1,07	2
22	ПРОМІНВЕСТБАНК	1275,58	25,87	28022,21	792,63	2191,17	2593,4	1
23	ПУМБ	4781,8	32,03	13853,77	4604,9	4121,16	4322,17	1
24	СІТІБАНК УКРАЇНА	3,89	0,04	2112,58	56,94	63,67	2294,35	2
25	СБЕРБАНК РОСІЇ	155,45	9,75	13707,79	1228,61	2667,47	1438,37	3
26	СОЮЗ	604,46	0,67	3324,15	101,89	372,85	106,4	2
27	УКРІНБАНК	678,96	2,39	2384,71	315,72	199,55	7,74	2
28	ГРАНТ	9,94	0,01	362,08	94,37	7,41	4,41	
29	КРЕДОБАНК	105,21	6,77	1338,77	615,17	248,31	631,88	

Рис. 43. **Результати кластеризації банків за розміром кредитно-інвестиційного портфеля**

Codes for grouping variable (Коди для груп змінних) вказують кількість аналізованих груп об'єктів.

Missing data (пропущені змінні) дозволяє вибрати **Case wise** (построкове видалення змінних зі списку), або **Mean substitution** (змінити їх на середні значення).

Open Data – відкриває файл з даними.

Кнопка **Select Cases** – задає умови вибору спостережень з бази даних.

Кнопка **W** – задає ваги змінних, обравши їх зі списку.

Після натискання кнопки **Variables** можливо обрати (рис. 45):

Grouping – групувальну змінну;

Independent - незалежну змінну.

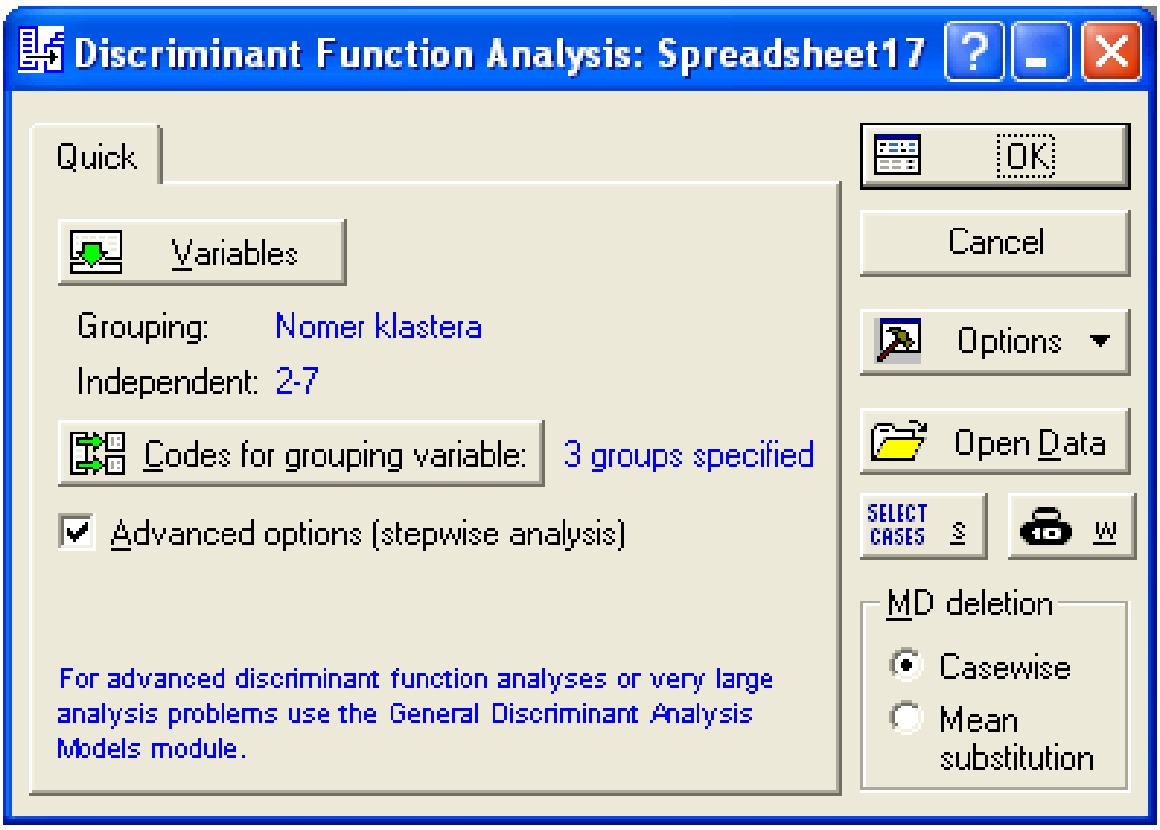


Рис. 44. **Діалогове вікно модуля дискримінантного аналізу**

Категорії групувальної змінної обиратимуться у вікні (рис. 46).

Рис. 45. **Вікно для вибору змінних**

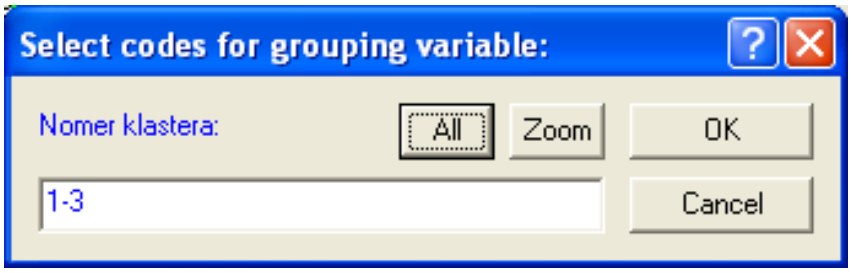


Рис. 46. **Вибір групувальної змінної**

Вигляд діалогового вікна **Model Difinition**, яке призначене для вибору моделі, подано на рис. 47. У вкладці Advanced можливо задати метод (Method), який буде використовуватись для вибору значимих змінних.

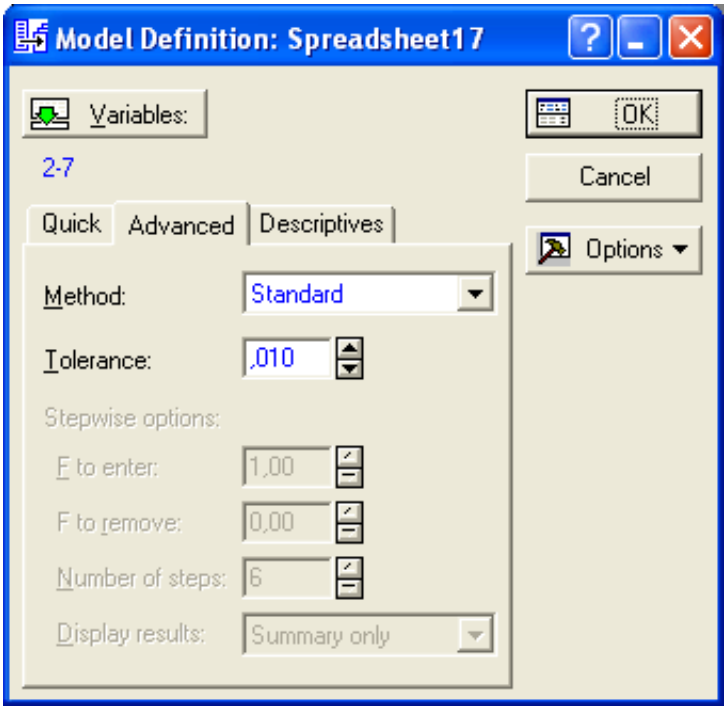


Рис. 47. **Діалогове вікно визначення вибору моделі**

Можливе використання таких методів:

Standart (Стандартний). Усі змінні одночасно включені в модель.

Forward stepwise (Покроковий з включенням). На кожному кроці в модель відбирається змінна з максимальним F-значенням. Процедура закінчується, коли всі змінні, значення F яких більше значення вказаного в полі F to enter, увійшли до моделі.

Backward stepwise (Покроковий з виключенням). На кожному кроці в модель відбираються всі змінні, які потім видаляються залежно від величини F-значення. Кроки закінчуються, коли немає змінних, F-значення яких менше певного, визначеного в полі F to remove.

Поле **Number of steps** (число кроків) визначає максимальну кількість кроків аналізу, по досягненню яких процедура закінчується.

Поле **Tolerance** (толерантність) дозволяє виключити з моделі неінформативні змінні. Якщо толерантність має значення менше, ніж значення 0,01, то змінна визнається не інформативною та не включається до моделі.

В якості методу аналізу виберемо **Standard**. За результатами, отриманими в ході обчислень, представленими у вікні **Discriminant Function Analysis Results**, рис. 48, можливо отримати таку інформацію: число змінних в моделі (Number of variables in the model) – 6;
значення лямбди Уилкса (Wilks' Lambda) – 0,022; приблизне значення *F*-статистики, яке пов'язане з лямбдою Уилкса (Approx. *F* (12, 38)) – 10,32641; рівень значимості *F*-критерію $p < 0,0000$ для значення 10,32641.

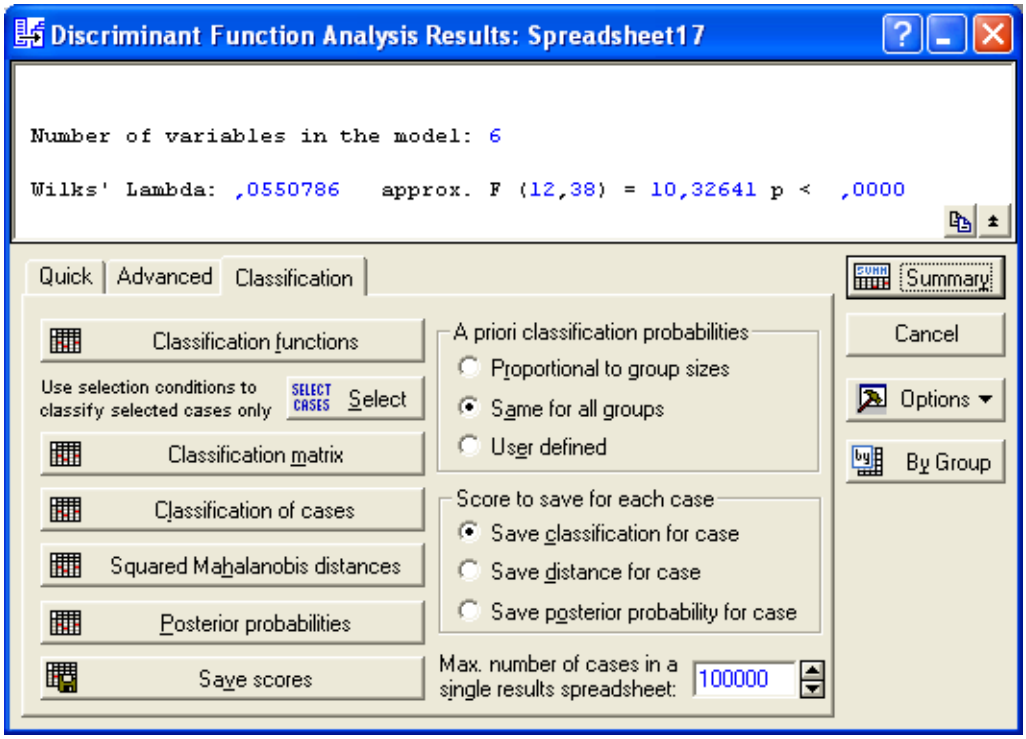


Рис. 48. **Вікно результатів дискримінантного аналізу**

Значення статистики Уілкса лежить в інтервалі [0,1]. Значення статистики Уілкса, які наближуються до 0, свідчать про гарну дискримінацію, а значення, які наближуються до 1, свідчать про погану дискримінацію. Таким чином, за даними показника Wilks' Lambda, який дорівнює 0,055, можливо зробити висновок, що класифікація є коректною.

В якості перевірки коректності навчальних вибірок подивимося результати класифікаційної матриці, натиснувши кнопку Classification matrix, попередньо обравши **Same for all groups** у правій частині вікна **Discriminant Function Analysis Results** (рис. 49).

Group	Classification Matrix (Spreadsheet17) Rows: Observed classifications Columns: Predicted classifications **Percent Correct**	G_1:1 p=,33333	G_2:2 p=,33333	G_3:3 p=,33333
G_1:1	100,0000	6	0	0
G_2:2	83,3333	0	10	2
G_3:3	66,6667	0	3	6
Total	81,4815	6	13	8

Рис. 49. **Класифікаційна матриця**

За результатами класифікаційної матриці (рис. 49) можна зробити висновок, що об'єкти розбито правильно на три групи за допомогою кластерного аналізу. Якщо є підприємства, неправильно віднесені до відповідних груп, можна подивитися **Classification of cases** (класифікація випадків). У таблиці класифікації (рис. 50) випадків некоректно віднесені об'єкти позначаються зірочкою (*).

Case	Classification of Cases (Spreadsheet17) Incorrect classifications are marked with * **Observed Classif.**	1 p=,33333	2 p=,33333	3 p=,33333
1	G_3:3	G_3:3	G_2:2	G_1:1
2	G_2:2	G_2:2	G_3:3	G_1:1
3	G_1:1	G_1:1	G_3:3	G_2:2
4	G_2:2	G_2:2	G_3:3	G_1:1
5	G_3:3	G_2:2	G_3:3	G_1:1
6	G_3:3	G_3:3	G_2:2	G_1:1
7	G_3:3	G_3:3	G_2:2	G_1:1
8	G_1:1	G_1:1	G_3:3	G_2:2
9	G_1:1	G_1:1	G_3:3	G_2:2
10	G_2:2	G_2:2	G_3:3	G_1:1
11	G_2:2	G_2:2	G_3:3	G_1:1
12	G_3:3	G_2:2	G_3:3	G_1:1
13	G_3:3	G_2:2	G_3:3	G_1:1
14	G_3:3	G_3:3	G_2:2	G_1:1
15	G_2:2	G_2:2	G_3:3	G_1:1
16	G_2:2	G_2:2	G_3:3	G_1:1
17	G_1:1	G_1:1	G_3:3	G_2:2
18	G_2:2	G_2:2	G_3:3	G_1:1
19	G_2:2	G_3:3	G_2:2	G_1:1
20	G_3:3	G_3:3	G_2:2	G_1:1
21	G_2:2	G_3:3	G_2:2	G_1:1
22	G_1:1	G_1:1	G_3:3	G_2:2
23	G_1:1	G_1:1	G_3:3	G_2:2
24	G_2:2	G_2:2	G_3:3	G_1:1
25	G_3:3	G_3:3	G_2:2	G_1:1
26	G_2:2	G_2:2	G_3:3	G_1:1
27	G_2:2	G_2:2	G_3:3	G_1:1

Рис. 50. **Класифікація випадків**

Класифікаційні функції для кожного класу можливо отримати у вікні Discriminant Function Analysis Results натиснувши кнопку Classification functions (рис. 51).

Variable	Classification Functions; groupin		
	G_1:1 p=,33333	G_2:2 p=,33333	G_3:3 p=,33333
Mezhb krediti	0,0049	0,00120	0,00235
Zadolzh bankov	0,7008	0,07434	0,23979
Krediti Ur licam	0,0030	0,00056	0,00112
Krediti fiz licam	0,0077	0,00156	0,00295
Rezervi klientam	0,0006	-0,00077	-0,00078
Cennie bumagi	0,0030	0,00027	-0,00007
Constant	-63,3806	-2,64771	-7,97837

Рис. 51. **Параметри класифікаційних функцій**

Банки з великим розміром кредитно-інвестиційного портфеля = 0,0049 • Міжбанківські кредити + 0,7008 • Заборгованість банків + + 0,003 • Кредити юридичним особам + 0,0077 • Кредити фізичним особам + 0,0006 • Резерви клієнтам + 0,003 • Цінні папери – 63,38;

Банки з низьким розміром кредитно-інвестиційного портфеля = 0,0012 • Міжбанківські кредити + 0,0743 • Заборгованість банків + + 0,00056 • Кредити юридичним особам + 0,00156 • Кредити фізичним особам – 0,00077 • Резерви клієнтам + 0,00027 • Цінні папери – 2,65;

Банки з середнім розміром кредитно-інвестиційного портфеля = 0,00235 • Міжбанківські кредити + 0,24 • Заборгованість банків + + 0,00112 • Кредити юридичним особам + 0,00295 • Кредити фізичним особам – 0,00078 • Резерви клієнтам - 0,00007 • Цінні папери – 7,98.

Для отримання більш детальної інформації можна переглянути результати канонічного аналізу, який можливо проводити, якщо були вибрані, принаймні, три групи і є хоча б дві змінні в моделі, натиснувши кнопку **Perform canonical analysis** (рис. 48). З'являється вікно канонічного аналізу (рис. 52), в якому за допомогою опції Scatterplot of canonical scores можливо побудувати таку діаграму розсіювання для значень (рис. 53).

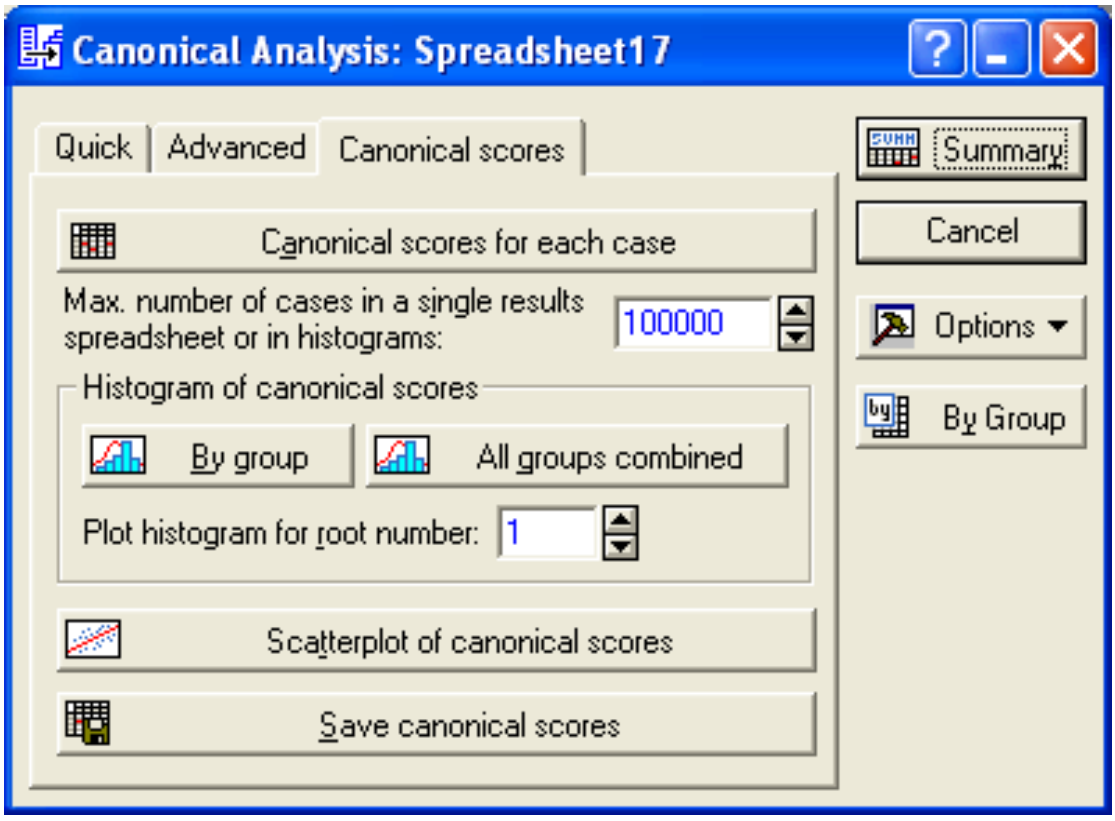


Рис. 52. **Діалогове вікно канонічного аналізу**

За допомогою цієї діаграми можливо визначити вклад, який вносить кожна дискримінантна функція в розподіл між групами.

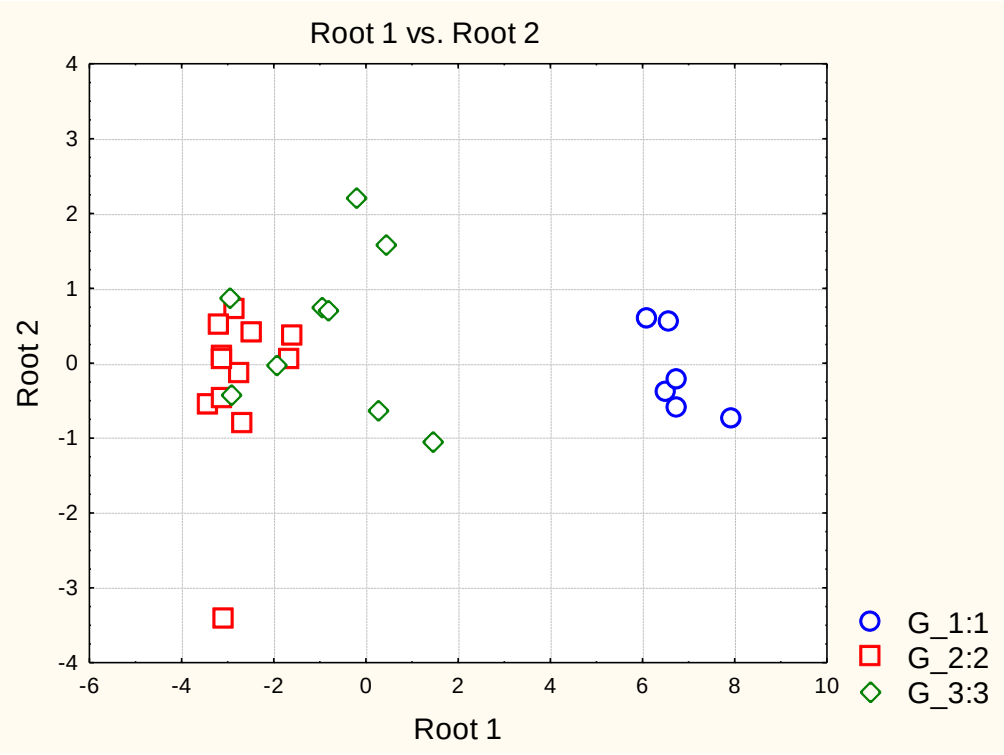


Рис. 53. **Графік розсіювання канонічних значень**

Таким чином, проведена класифікація банків за розміром кредитно-інвестиційного портфеля методом кластерного аналізу є адекватною. В ході проведення дискримінантного аналізу, побудовано функції, які можна використовувати в подальшому для віднесення певного банку в один з класів. Для визначення класу кредитно-інвестиційного портфеля банку Грант та Кредобанк внесемо змінні у вихідні дані (рис. 54).

	1 bank	2 Mezhb krediti	3 Zadolzh bankov	4 Krediti Ur licam	5 Krediti fiz licam	6 Rezervi klientam	7 Cennie bumagi	8 Nomer klastera
1	ІНГ БАНК УКРАЇНА	234,19	17,04	7720,5	95,52	334,13	847,3	3
2	АКТИВ-БАНК	364,94	0,38	1800,91	842,3	99,21	8,1	2
3	АЛЬФА-БАНК	2019,04	16,34	16856,73	4880,14	5484,73	1341,07	1
4	БАНК КРЕДИТ-ДНІПРО	402,32	0,29	5028,75	466,24	482,93	155,02	2
5	БАНК ТАВРИКА	1064,12	1,35	2923,92	256,87	97,56	5,17	3
6	БРОКБІЗНЕСБАНК	1466,98	2,46	9224,82	2302,17	728,17	426,3	3
7	ВІЕЙБІ Банк	1508,12	5,51	4098,07	1699,09	1452,7	405,31	3
8	ВТБ БАНК	716,57	3,24	28021,71	3333,64	5344,58	1827,91	1
9	ДЕЛЬТА БАНК	1456,12	40,26	9028,93	5937,46	4033,84	3209,91	1
10	ЗЛАТОБАНК	694,58	7,37	2331,77	172,17	317,89	0,06	2
11	КІБ Креді Агріколь	70	1,5	3041,47	24,34	251,5	4	2
12	КИЇВСЬКА РУСЬ	505,93	0,97	3408,63	399,74	226,49	190,34	3
13	КРЕДІ АГРІКОЛЬ БАНК	193	1,26	3240,6	2165,96	458,11	64,5	3
14	КРЕДИТПРОМБАНК	586,71	3,74	8688,12	2169,93	2323,31	716,01	3
15	МАРФІН БАНК	462,52	2,64	2156,65	905,71	664,49	435,97	2
16	МЕГАБАНК	322,78	0,29	2799,14	522,81	183,24	107,71	2
17	ОТП БАНК	269,41	31,87	12920,34	6862,48	3104,15	2600,38	1
18	ПІВДЕНКОМБАНК	239,59	9,98	2670,44	64,47	147,27	489,53	2
19	ПІВДЕННИЙ	484,66	1,93	7833,81	619,44	593,89	16,5	2
20	ПЛАТИНУМ БАНК	922,54	34,3	31,57	2406,95	337,45	1,17	3
21	ПРАВЕКС-БАНК	137,99	1,38	1677,4	3126,8	905,31	1,07	2
22	ПРОМІНВЕСТБАНК	1275,58	25,87	28022,21	792,63	2191,17	2593,4	1
23	ПУМБ	4781,8	32,03	13853,77	4604,9	4121,16	4322,17	1
24	СІТІБАНК УКРАЇНА	3,89	0,04	2112,58	56,94	63,67	2294,35	2
25	СБЕРБАНК РОСІЇ	155,45	9,75	13707,79	1228,61	2667,47	1438,37	3
26	СОЮЗ	604,46	0,67	3324,15	101,89	372,85	106,4	2
27	УКРІНБАНК	678,96	2,39	2384,71	315,72	199,55	7,74	2
28	ГРАНТ	9,94	0,01	362,08	94,37	7,41	4,41	
29	КРЕДОБАНК	105,21	6,77	1338,77	615,17	248,31	631,88	

Рис. 54. **Вихідні дані з новими спостереженнями**

Case	Observed Classif.	G_1:1 p=,22222	G_2:2 p=,44444	**G_3:3 p=,33333**
28	---	0,000000	0,994894	0,005106
29	---	0,000000	0,952121	0,047879

Рис. 55. **Фрагмент таблиці апостеріорних імовірностей**

Case	Observed Classif.	G_1:1 p=,22222	G_2:2 p=,44444	G_3:3 p=,33333
28	---	120,8471	2,4331	12,40223
29	---	94,9461	0,9161	6,32077

Рис. 56. **Фрагмент таблиці відстаней від нового випадку до центрів груп**

Максимальне значення апостеріорних імовірностей та мінімальна відстань від нового випадку до центроїдів груп відповідають кластеру № 2 (рис. 55, 56). Тому банки Грант та Кредобанк відносяться до другого кластера та мають низький розмір кредитного портфеля.

Лабораторне заняття на тему "Проведення редукції даних методами факторного аналізу"

Мета роботи – набуття навичок обробки даних за допомогою методів факторного аналізу.

Завдання – скоротити інформаційний простір, використовуючи методи факторного аналізу.

Методичні рекомендації

Модуль **Factor Analysis** (Факторний аналіз) містить широкий набір методів, за допомогою яких можливо проводити виділення факторів, тим самим скорочуючи вхідний інформаційний простір.

Розглянемо основні етапи проведення факторного аналізу в системі Statistica на такому прикладі.

Для аналізу діяльності приватного підприємства було відібрано такі показники (рис. 57):

X1 – питома вага втрат від браку;

X2 – індекс зниження собівартості продукції;

X3 – фондовіддача;

X4 – коефіцієнт змінності устаткування;

X5 – продуктивність праці;

X6 – питома вага виробів, що купуються.

	1 X1	2 X2	3 X3	4 X4	5 X5	6 X6
1	5,571	47,88	0,522	0,153	1,071	0,225
2	4,914	27,09	1,377	0,135	0,873	0,441
3	5,85	131,76	0,63	0,144	1,035	0,234
4	5,949	16,29	1,593	0,135	0,018	0,252
5	3,888	12,24	0,666	0,153	0,054	0,153
6	6,633	80,82	0,972	0,306	1,251	0,153
7	6,318	56,25	1,035	0,306	0,072	0,279
8	7,425	41,67	0,873	0,306	0,693	0,162
9	7,335	93,123	1,008	0,171	0,693	0,279
10	7,848	65,97	0,891	0,171	0,972	0,162
11	5,976	68,94	0,522	0,306	0,837	0,279
12	7,29	65,709	0,927	0,306	0,09	0,135
13	4,968	29,07	1,116	0,135	0,099	0,252
14	8,433	178,686	0,801	0,171	1,296	0,162
15	11,853	538,308	0,612	0,306	0,432	0,126
16	6,003	64,521	0,927	0,171	1,116	0,162
17	5,112	81,567	0,657	0,288	0,693	0,261
18	4,671	73,89	0,657	0,171	0,837	0,27
19	9,018	68,58	0,765	0,297	0,117	0,243
20	7,344	107,523	0,927	0,306	1,557	0,261
21	4,671	73,89	0,657	0,171	0,837	0,27

Рис. 57. **Початкові дані**

Для виклику модуля факторного аналізу можна використовувати **Statistics/Multivariate Exploratory Techniques/Factor Ananlysis** (Багатовимірні методи/факторний аналіз). На екрані з'явиться діалогове вікно (рис. 58) **Factor Analysis**.

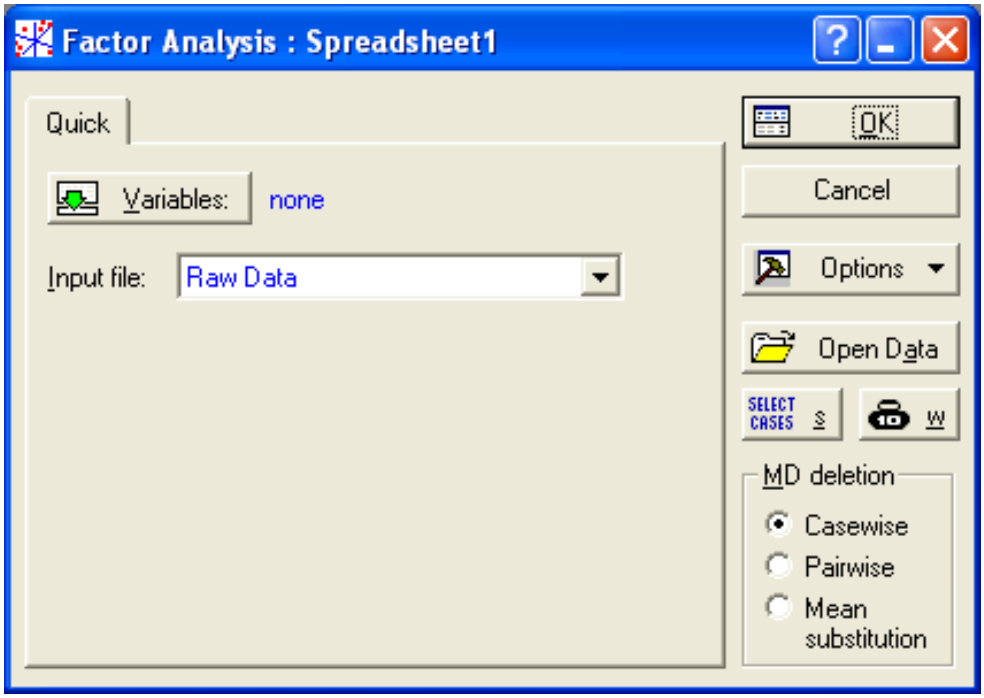


Рис. 58. **Діалогове вікно факторного аналізу**

Кнопка **Variables** (Змінні) дозволяє відібрати всі змінні з файла даних, які повинні бути включені в факторний аналіз (рис. 59). Якщо при аналізі використовуватимуться не всі змінні, то можна скористатися кнопкою **Select All** (Виділити все).

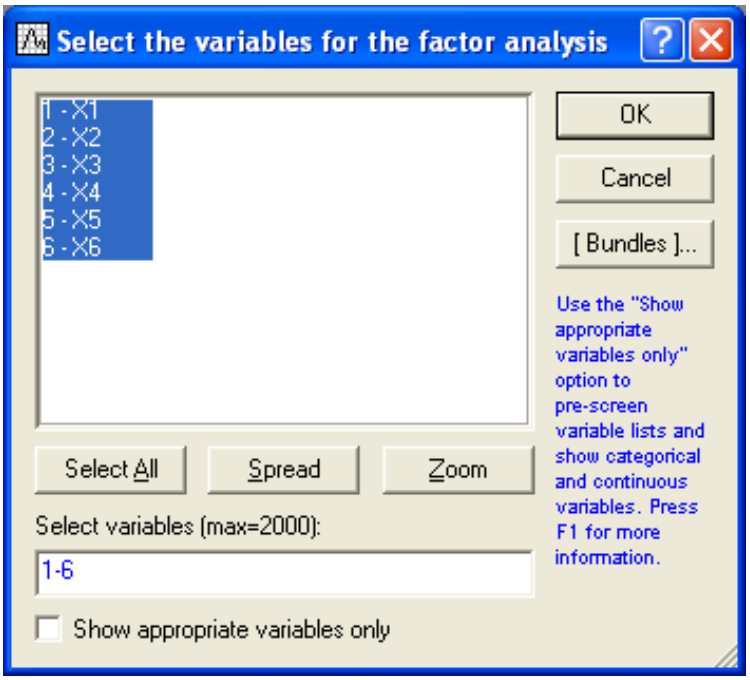


Рис 59. **Вікно вибору змінних**

У модулі можливі такі типи вихідних даних: **Correlation Matrix** (Кореляційна матриця) та **Raw Data** (Вихідні дані).

Оберемо, наприклад, **Raw Data**. Це звичайний файл даних, де по рядках записані значення змінних.

MD deletion (Заміна пропущених змінних). Спосіб обробки пропущених значень.

Casewise (спосіб виключення пропущених випадків) – полягає в тому, що в електронній таблиці, що містить дані, ігноруються всі рядки (випадки), у яких є хоча б одне пропущене значення. Це стосується всіх змінних. У таблиці залишаються тільки випадки, в яких немає жодного пропуску.

Pairwise (Парний спосіб виключення пропущених значень) – ігноруються пропущені випадки не для всіх змінних, а лише для вибраної пари. Всі випадки, в яких немає пропусків, використовуються в обробці, наприклад, при поелементному обчисленні кореляційної матриці, коли послідовно розглядаються всі пари змінних. Очевидно, в способі **Pairwise** залишається більше спостережень для обробки, ніж у способі **Casewise.**

Mean Substitution (підстановка середнього замість пропущених значень).

Натиснувши в стартовому вікні модуля на кнопку **OK**, розпочинається аналіз обраних змінних. Система Statistica обробить пропущені значення тим способом, який вказано, обчислить кореляційну матрицю і запропонує на вибір кілька методів факторного аналізу. Обчислення кореляційної матриці (якщо вона не задається відразу) – перший етап факторного аналізу. Після натиснення кнопки **Ok** можна перейти до наступного діалогового вікна.

Define Method of Factor Extraction (Визначити метод виділення факторів) (рис. 60).

Дане вікно має таку структуру. Верхня частина вікна є інформаційною: тут повідомляється, що пропущені значення оброблені методом **Casewise**. Опрацьовано 21 випадок та 21 випадок прийнятий для подальших обчислень. Кореляційна матриця обчислена для 6 змінних. Група опцій, об'єднаних під заголовком **Extraction method** (Методи виділення факторів), дозволяє вибрати метод обробки.

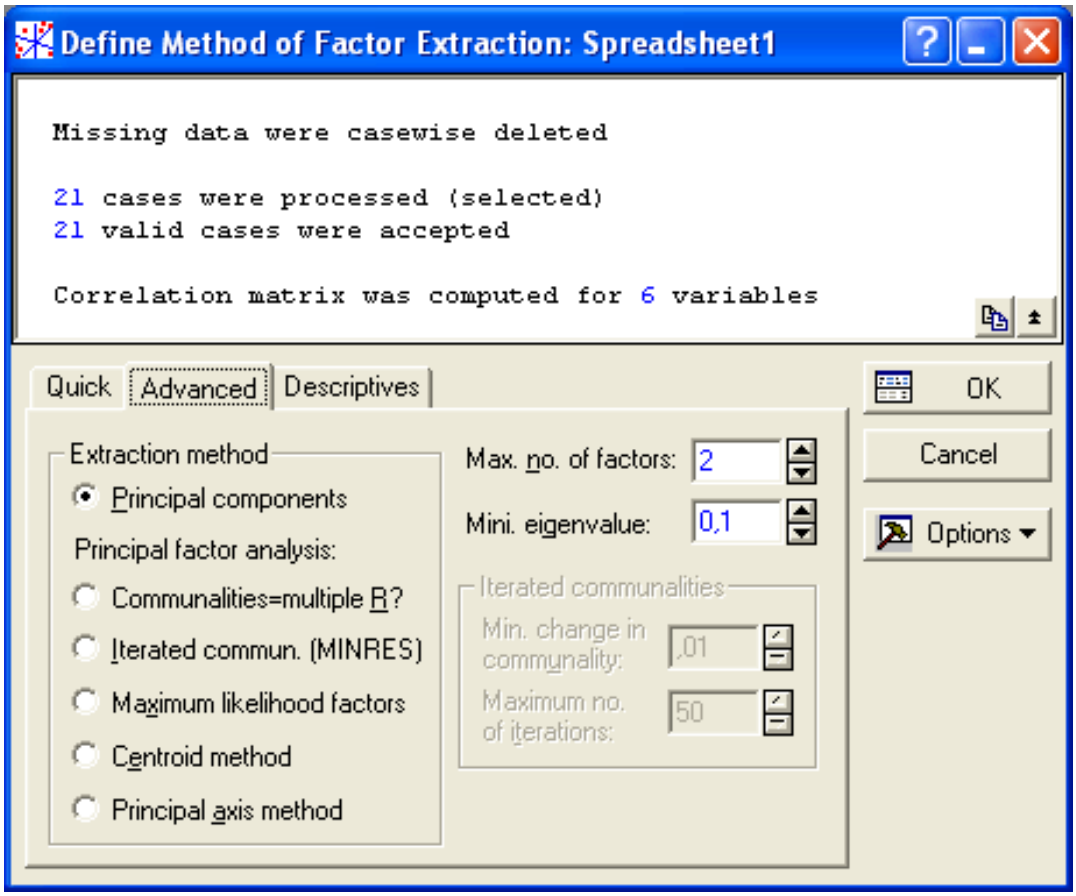


Рис. 60. **Вікно вибору методу виділення факторів**

Для продовження аналізу у вікні Define Method of Factor Extraction (Визначити метод виділення факторів) (рис. 61) необхідно натиснути на кнопку Review correlations, means, standard deviations (Проглянути кореляції/середні/стандартні відхилення).

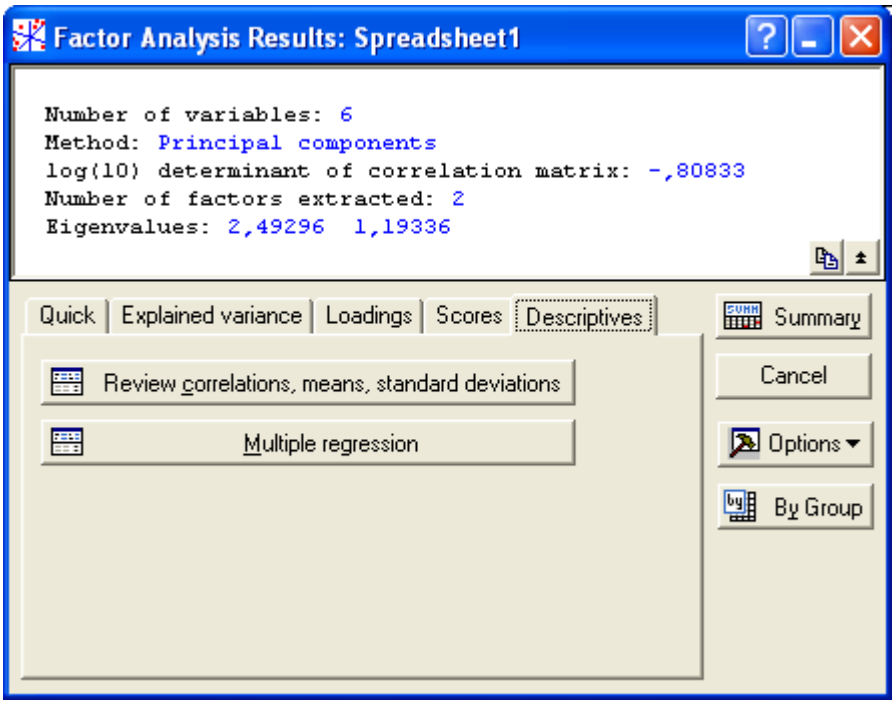


Рис. 61. **Вкладки вікна вибору методу виділення факторів**

Після чого вам з'явилося вікно перегляду описових статистик для аналізованих даних (рис. 62), де можна подивитися середні, стандартні відхилення, кореляції, коваріації, побудувати різні графіки.

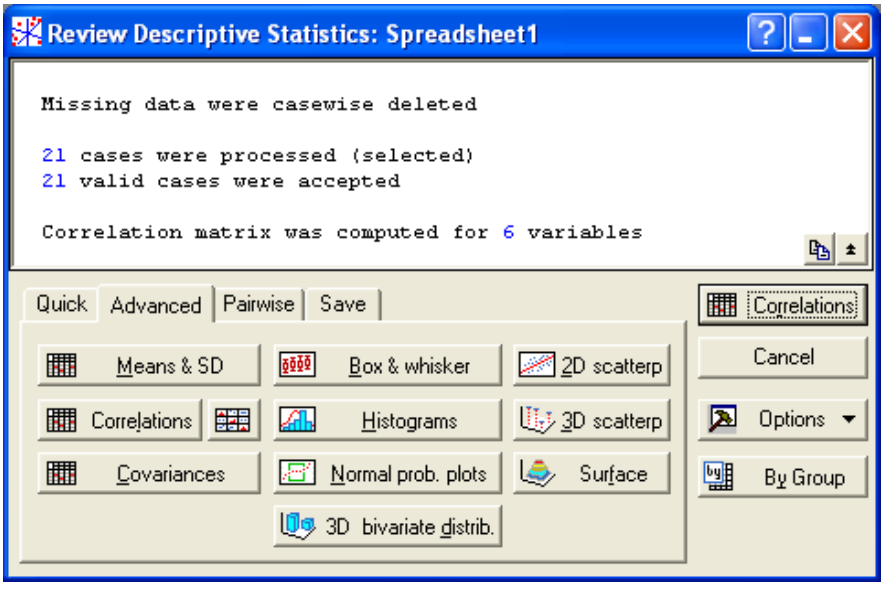
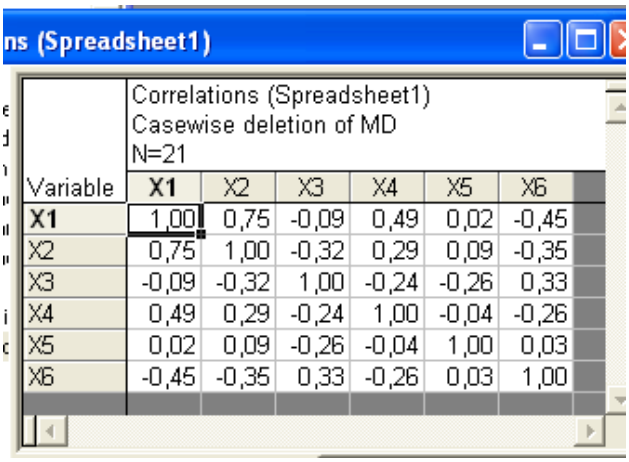


Рис. 62. **Вікно перегляду описових статистик**

Тут можна провести додатковий аналіз поточних даних, перевірити відповідність вибіркових змінних нормальному закону розподілу і існування лінійної кореляції між змінними.

Натиснувши кнопку Correlations (Кореляції) (рис. 63), на екрані з'явиться кореляційна матриця обраних раніше змінних.



Variable	X1	X2	X3	X4	X5	X6
X1	1,00	0,75	-0,09	0,49	0,02	-0,45
X2	0,75	1,00	-0,32	0,29	0,09	-0,35
X3	-0,09	-0,32	1,00	-0,24	-0,26	0,33
X4	0,49	0,29	-0,24	1,00	-0,04	-0,26
X5	0,02	0,09	-0,26	-0,04	1,00	0,03
X6	-0,45	-0,35	0,33	-0,26	0,03	1,00

Рис. 63. **Кореляційна матриця**

Далі оберемо опцію **Principal components** (Головні компоненти) і натиснемо кнопку OK. Система швидко зробить обчислення, і на екрані (рис.64) з'явиться вікно Factor Analysis Results (Результати факторного аналізу).

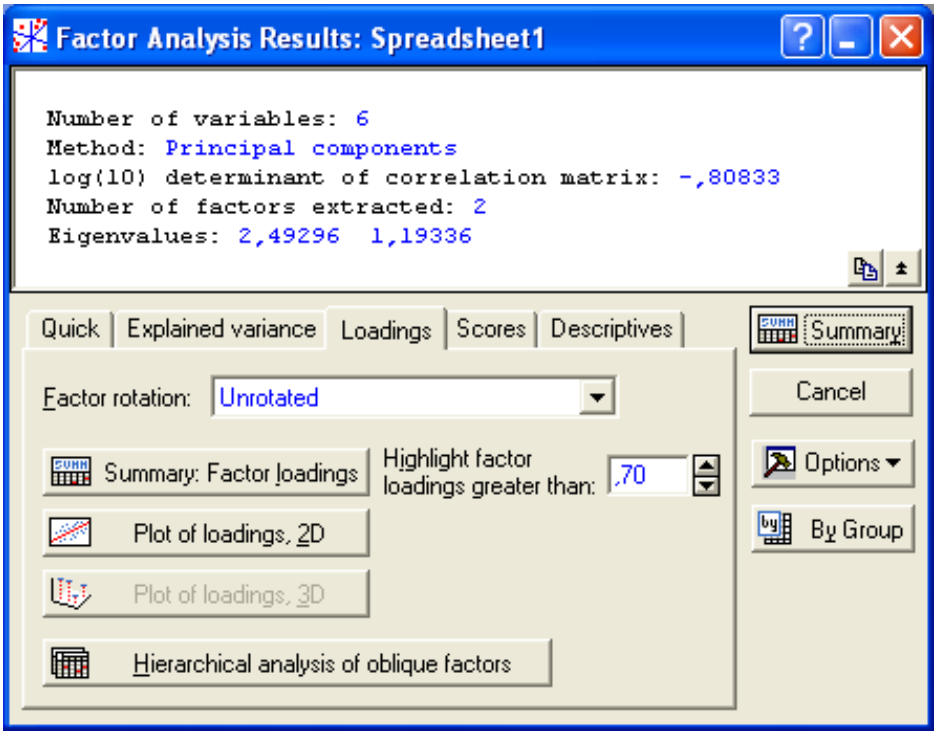


Рис 64. **Вікно результатів факторного аналізу**

У верхній частині вікна результатів факторного аналізу дається інформаційне повідомлення: **Number of variables** (число аналізованих змінних) – 6; **Method** (метод аналізу) – головні компоненти; **log (10) determination of correlation matrix** (десятковий логарифм детермінанта кореляційної матриці) – 0,80833; **Number of Factor extraction** (число виділених факторів) – 2; **Eigenvalues** (власні значення) – 2,49296; 1,19336.

У нижній частині вікна знаходяться підрозділи, що дозволяють всебічно ознайомитись з результатами аналізу чисельно та графічно.

Plot of loadings, 2D і Plot of loadings, 3D (Графіки навантажень) – ці опції побудують графіки факторних навантажень у проекції на площину будь-яких двох обраних факторів (рис. 65) і в проекції в простір трьох обраних факторів (для чого необхідна наявність як мінімум трьох виділених факторів).

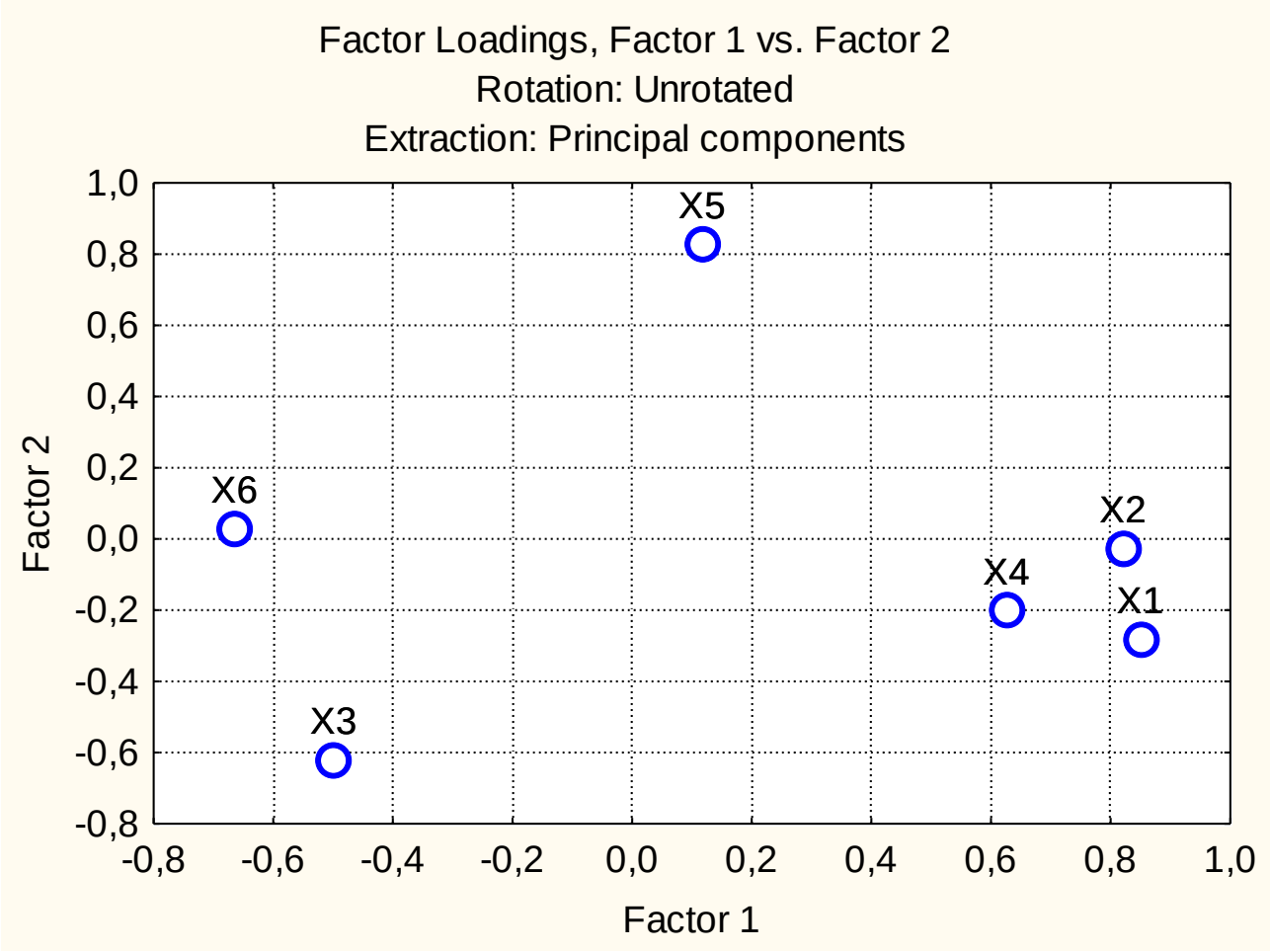


Рис. 65. **Графік факторних навантажень**

Summary. Factor loadings (факторні навантаження). Ця опція викликає таблицю з поточними факторними навантаженнями (рис.66), тобто обчисленими для даного методу обертання факторів, який вказаний праворуч від відповідної кнопки. У цій таблиці факторам відповідають стовпці, а змінним – рядки і для кожного фактора вказується навантаження кожної вихідної змінної, яка показує відносну величину проекції змінної на факторну координатну вісь. Факторні навантаження можуть інтерпретуватися як кореляції між відповідними змінними і чинниками – чим вище навантаження по модулю, тим більше близькість фактора до початкової змінної; та вони представляють найбільш важливу інформацію для інтерпретації отриманих факторів. У згенерованої таблиці для полегшення трактування будуть виділені факторні навантаження по абсолютній величині більше 0,7.

Variable	Factor Loadings (Unrota Extraction: Principal cor (Marked loadings are >, **Factor 1**	Factor 2
X1	0,850059	-0,282785
X2	0,821316	-0,026843
X3	-0,498412	-0,621588
X4	0,626536	-0,202623
X5	0,117676	0,827365
X6	-0,664075	0,026784
Expl.Var	2,492965	1,193365
Prp.Totl	0,415494	0,198894

Рис. 66. **Таблиця факторних навантажень**

За результатами таблиці видно, що перший фактор більше корелює зі змінними, ніж другий. Оскільки кореляція інших факторів незначна, у цьому випадку доцільно вдатися до повороту осей, сподіваючись отримати рішення, яке можна інтерпретувати в предметній області.

Мета обертання – отримання простої структури, при якій більшість спостережень знаходиться поблизу осей координат. При випадковій конфігурації спостережень неможливо отримати просту структуру. Далі необхідно натиснути Factor rotation (Обертання факторів) (рис. 67).

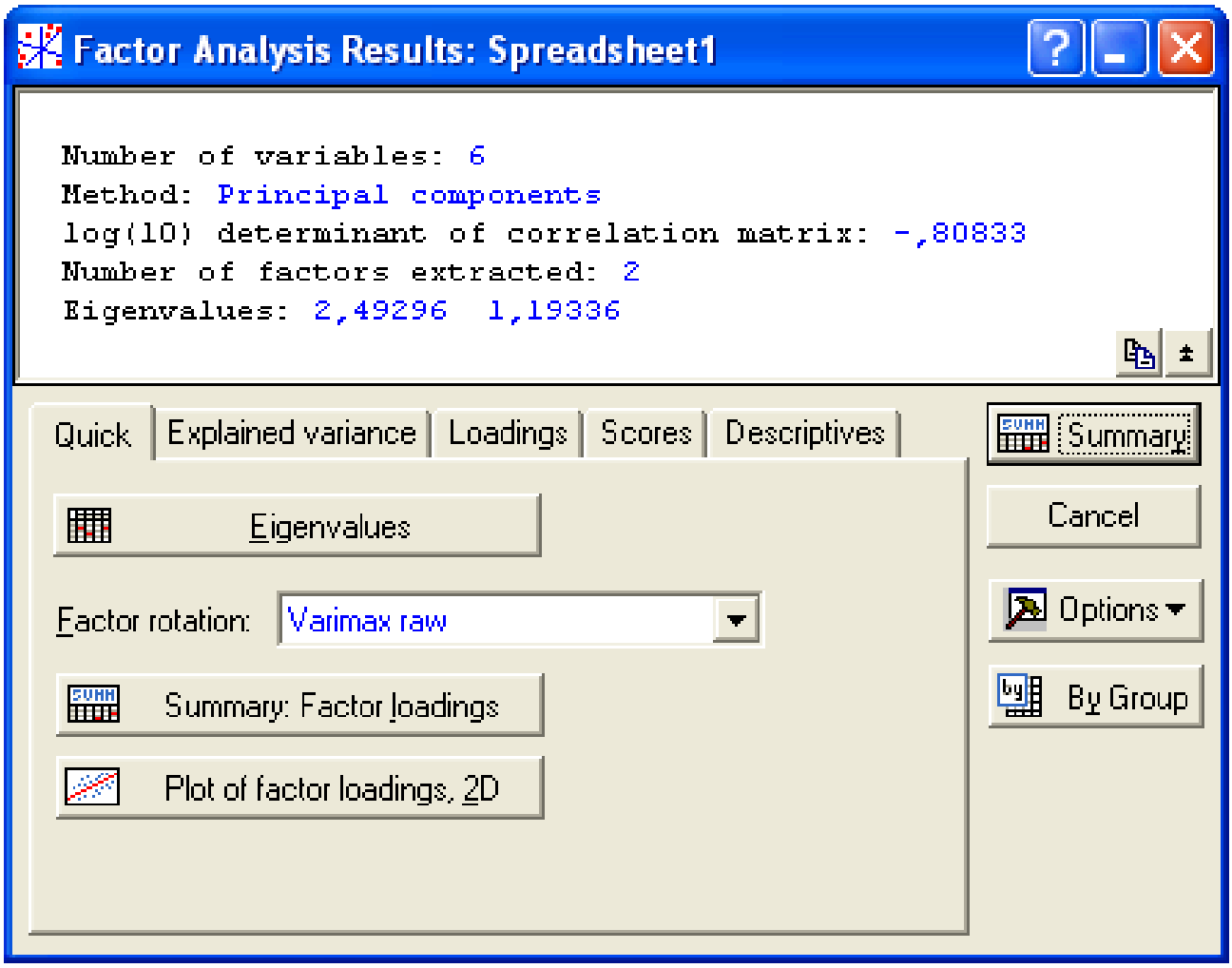


Рис. 67. **Вибір методу повороту осі**

У даному меню можливо обирати різні повороти осі. Вікно пропонує кілька можливостей оцінити й знайти потрібний поворот наступними методами: **Varimax** – варімакс; **Biquartimax** – Біквартімакс; **Quartimax** – Квартімакс; **Equamax** – Еквімакс.

Термін – **normalized** (нормалізовані) – вказує на те, що факторні навантаження в процедурі нормалізуються, тобто діляться на корінь квадратний з відповідною дисперсією.

Термін **raw** (вихідні) показує, що обертаються навантаження не нормалізовані.

Ініціювавши кнопку **Varimax raw** (варімакс). Система справить обертання факторів методом варімакс, і вікно **Factor Analysis Results** (Результати факторного аналізу) знову з'являться на моніторі. Знову ініціювавши в цьому вікні кнопку **Plot of Loadings 2D** (Двовимірний графік навантажень) з’явиться графік навантажень (рис. 68).

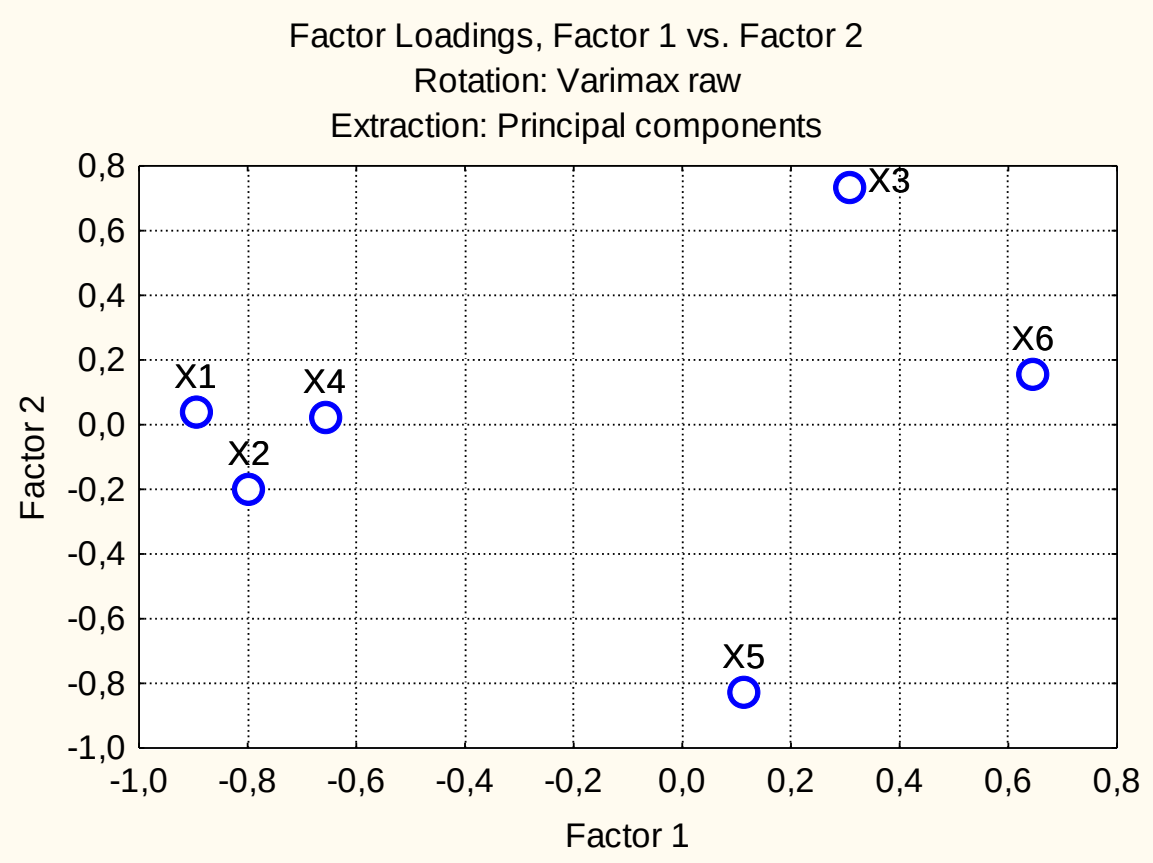


Рис. 68. **Графік навантажень**

Цей графік відрізняється від попереднього. Навантаження ще можна подивитися чисельно, ініціювавши кнопку Факторні навантаження (Factor loadings). Натиснувши на кнопку Summary. Factor loadings, відкриється вікно (рис.69).

Variable	Factor Loadings (Varim Extraction: Principal co (Marked loadings are >, **Factor 1**	Factor 2
X1	-0,895031	0,038570
X2	-0,797132	-0,199649
X3	0,308632	0,734529
X4	-0,758089	0,022847
X5	0,113966	-0,827884
X6	0,745916	0,156541
Expl.Var	2,395031	1,291299
Prp.Totl	0,399172	0,215216

Рис. 69. **Матриця факторних навантажень**

Тепер знайдене рішення вже можна інтерпретувати. Фактори частіше інтерпретують по навантаженнях. Перший фактор найтісніше пов'язаний з X1, X2, X4, X6. Другий фактор – X3 і X5. Таким чином, початковий простір скорочено й отримано два нових фактори.

Для підтвердження цього в програмному пакеті Statistica існує критерій Scree plot (Критерій кам'янистого осипу, рис. 70). У вікні Factor Analysis Results натиснувши кнопку Scree plot отримується наступний графік власних значень.

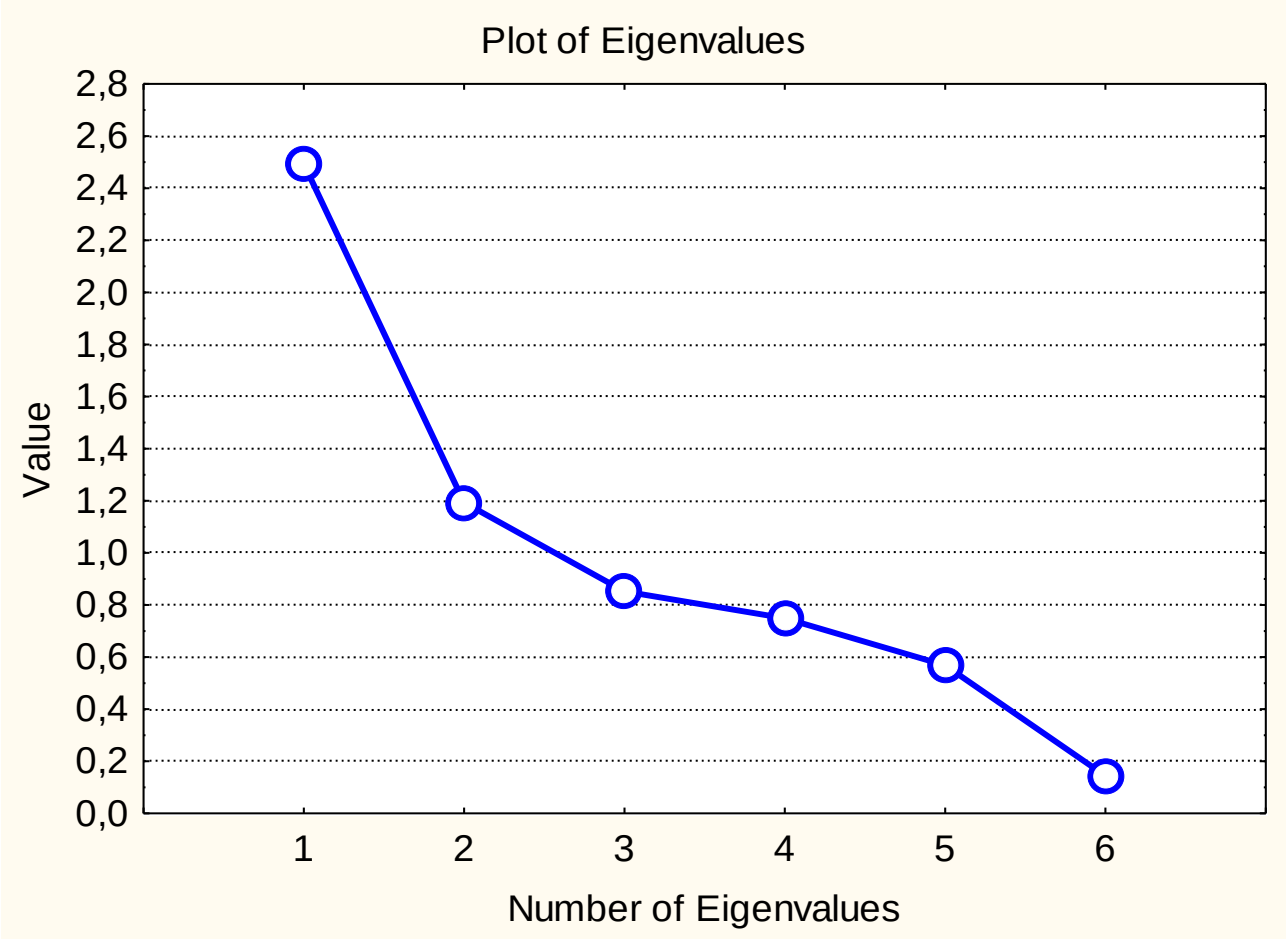


Рис. 70. **Результати критерію кам'янистого осипу.**

У точках з координатами 1, 2 осипання сповільнюється найбільш істотно, отже, теоретично можна обмежуватися двома факторами.

Отримані результати можливо інтерпретувати таким чином:

F1 = 0,746 • X6 – 0,895 • X1 – 0,797 • X2 - 0,758 • X4; – виробничі фактори;

F2 = X3 • 0,735 – 0,828 • X5 – фактори, пов’язані з трудовими ресурсами.

Лабораторне заняття на тему "Основні принципи роботи в системі Statistica Data Mining. Побудова моделі Data Mining"

Мета роботи – набути навичок роботи в модулі Statistica Data Mining.

Завдання – необхідно знайти багатовимірні просторові кількісні дані та провести їх аналіз за допомогою описових статистик.

Методичні рекомендації

Система Statistica Data Miner (розробник – компанія StatSoft) спроектована і реалізована як універсальний і всебічний засіб аналізу даних – від взаємодії з різними базами даних до створення готових звітів, реалізує так званий графічно-орієнтований підхід.

Серцем Statistica Data Miner є браузер процедур Data Mining (рис. 71), який містить більше 300 основних процедур, спеціально оптимізованих під завдання Data Mining, засоби логічного зв'язку між ними та управління потоками даних, що дозволяє конструювати власні аналітичні методи.

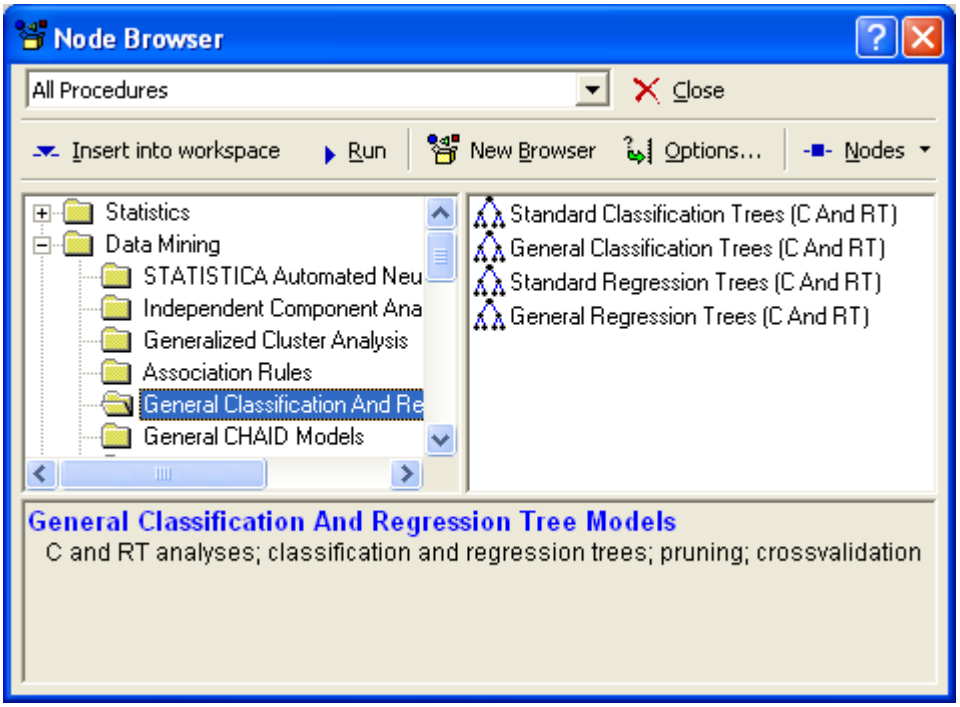


Рис. 71. **Браузер процедур Data Mining**

Робочий простір Statistica Data Miner складається з чотирьох основних частин (рис. 72).

Рис. 72. **Робочий простір Statistica Data Miner**

Data Acquisition – збір даних. У даній частині користувач ідентифікує джерело даних для аналізу, будь то файл даних або запит з бази даних.

Data Preparation, Cleaning, Transformation – підготовка, перетворення й очищення даних. Тут дані перетворюються, фільтруються, групуються і т. д.

Data Analysis, Modeling, Classification, Forecasting – аналіз даних, моделювання, класифікація, прогнозування. Тут користувач може за допомогою браузера або готових моделей задати необхідні види аналізу даних, таких, як: прогнозування, класифікація, моделювання і т. д.

Reports – результати. У даній частині користувач може переглянути, задати вигляд і налаштувати результати аналізу (наприклад, робоча книга, звіт або електронна таблиця).

Також в меню Data Mining є безліч процедур, які дозволяють проводити попередній аналіз даних ("буріння даних"), за допомогою чого можливо виявити приховані тенденції та закономірності, які непомітні на перший погляд, рис. 73.

Система Statistica включає величезний набір різних аналітичних процедур, і це робить його недоступним для звичайних користувачів, які слабо розбираються в методах аналізу даних. Але також запропонований варіант роботи для звичайних користувачів, що володіють невеликим досвідом і знаннями в аналізі даних і математичній статистиці.

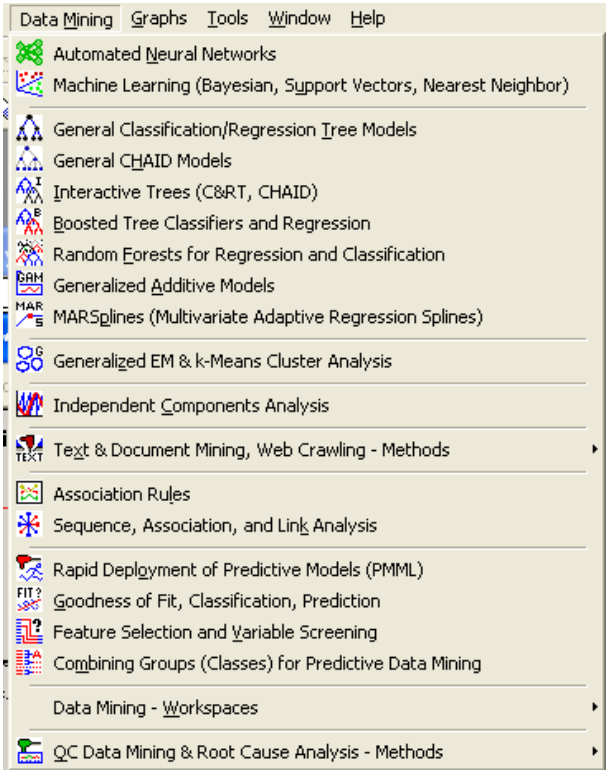


Рис. 73. **Пункт "Видобувник даних"**

Для цього, крім загальних методів аналізу, були вбудовані готові закінчені (сконструйовані) модулі аналізу даних, призначені для вирішення найбільш важливих і популярних завдань: прогнозування, класифікації, створення правил асоціації і т. д.

Роботу сконструйованих модулів аналізу даних розглянемо на прикладі використання методів багатовимірного аналізу даних (кластерного, дискримінантного та факторного). Для чого скористаємося вхідними даними попередніх лабораторних робіт за темами "використання апарату кластерного аналізу для класифікації даних", "застосування методів дискримінантного аналізу перевірки класифікації, отриманої при використанні апарату кластерного аналізу", "проведення редукції даних методами факторного аналізу" та зіставимо результати.

Для пошуку виду аналізу запускаємо "Диспетчер вузлів" (натискаємо на кнопку  у вікні Data Miner). У даному діалоговому вікні (рис. 74), можливо вибрати вид аналізу або задати операцію перетворення даних.

Диспетчер вузлів включає в себе всі доступні процедури для видобутку даних. Усього доступно близько 260 методів фільтрації й очищення даних, методів аналізу. За замовчуванням, процедури поміщені в папки і відсортовані відповідно до типу аналізу, який вони виконують. Однак користувач має можливість створити власну конфігурацію сортування методів.

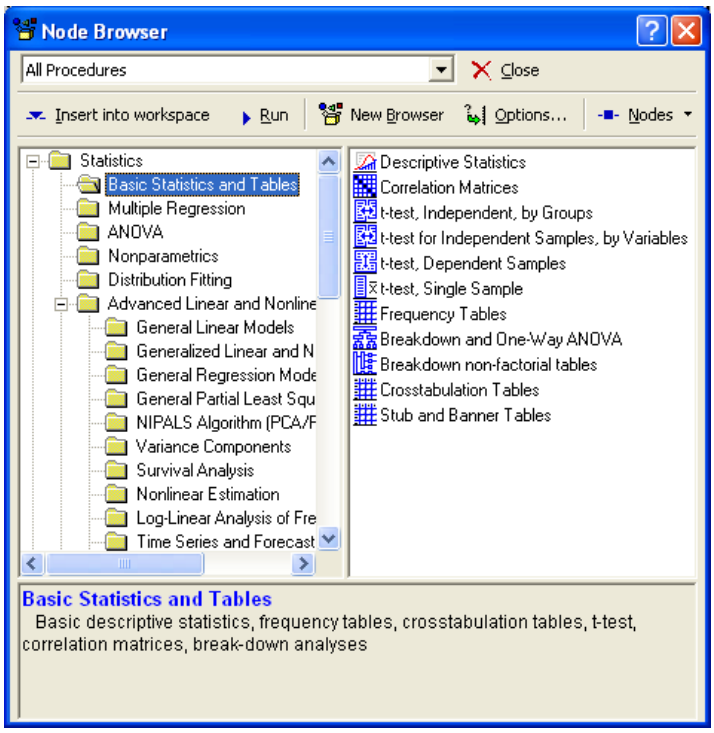


Рис. 74. **Діалогове вікно диспетчеру вузлів**

Далі необхідно обрати вхідні дані для проекту, натиснувши кнопку **Data Source**. Натиснувши двічі на ярлик, який з'явиться після обирання або створення файла з початковими даними у частині **Data Acquisition,** з'явиться вікно для вибору змінних (рис. 75). Принцип відбору змінних (безперервних і категоріальних) та предикторів (безперервних і категоріальних),здійснюється виходячи зі знань про структуру даних, який описаний у методичних рекомендаціях до попередніх лабораторних робіт. Так, для вибору змінних для проведення кластерного аналізу у вікні обираємо такі (рис. 76) та натискаємо OK.

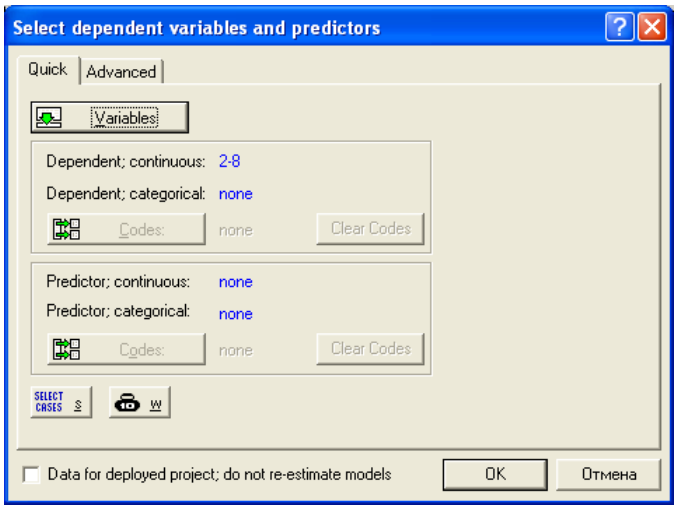


Рис. 75. **Вікно вибору залежних змінних і предикторів**

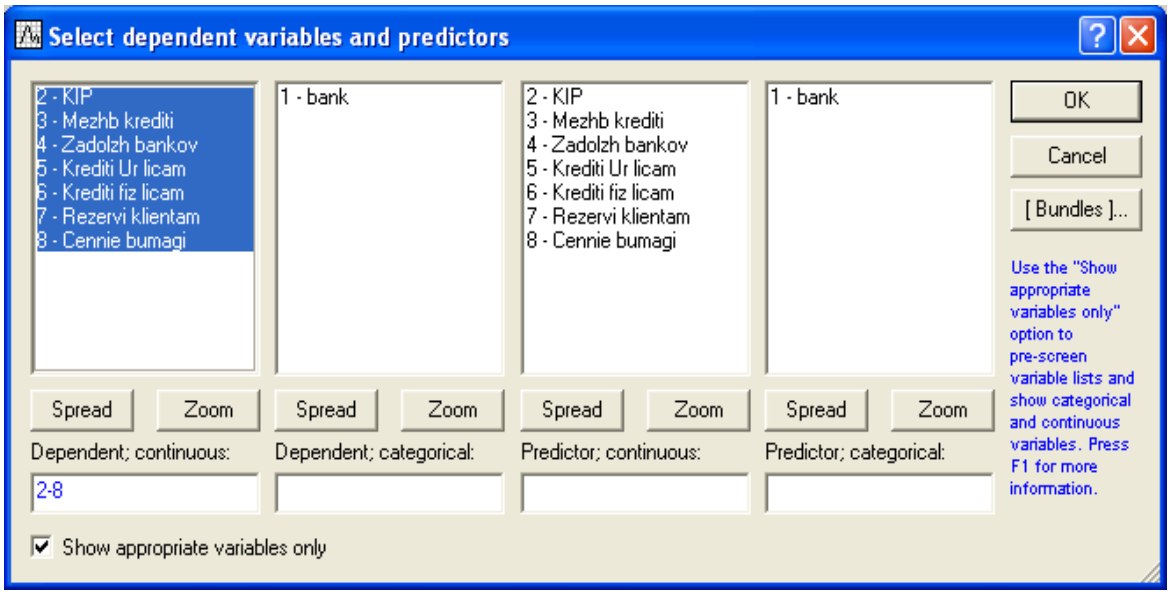


Рис. 76. **Вибір залежних змінних і предикторів для проведення кластерного аналізу**

Потім обираємо метод для проведення аналізу за допомогою діалогового вікна **Node Browser**. Після вибору кластерного аналізу обираємо одразу три методи його здійснення (рис. 77).

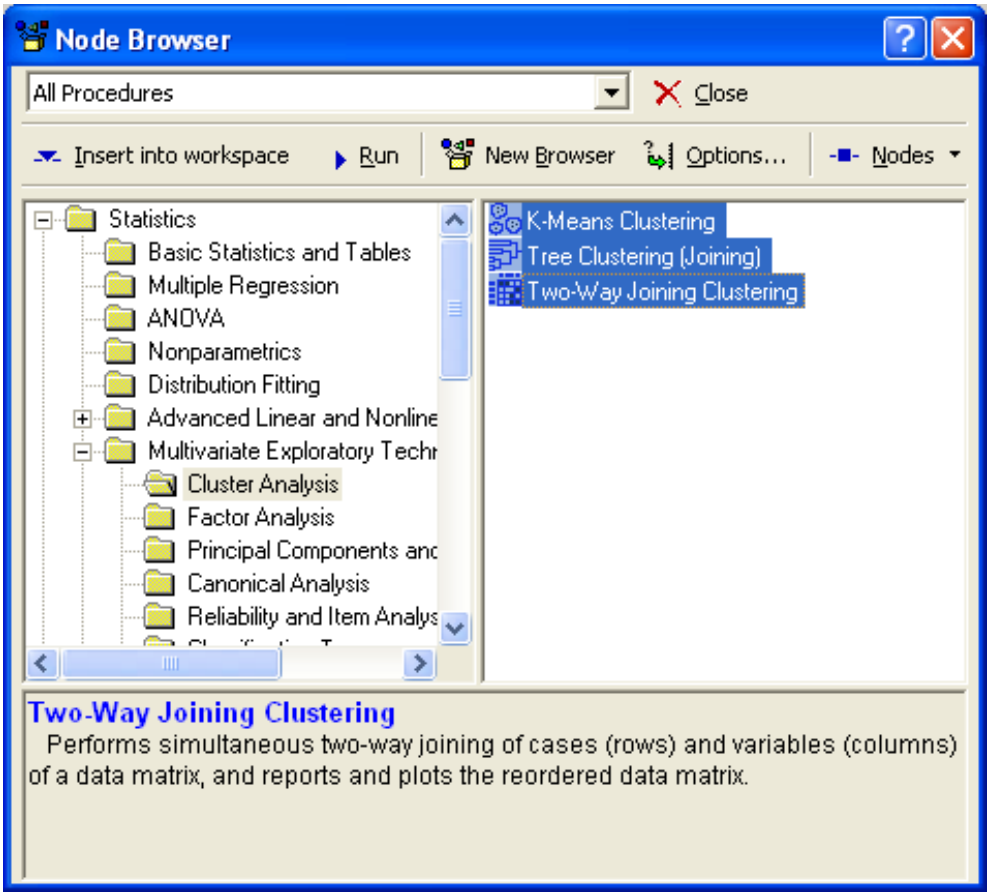


Рис. 77. **Діалогове вікно вибору методів кластерного аналізу**

Джерело даних у робочій області Data Miner автоматично з'єднується з вузлами обраних аналізів. Усі вузли, з'єднані з джерелами даних активними стрілками будуть проведені (рис. 78).

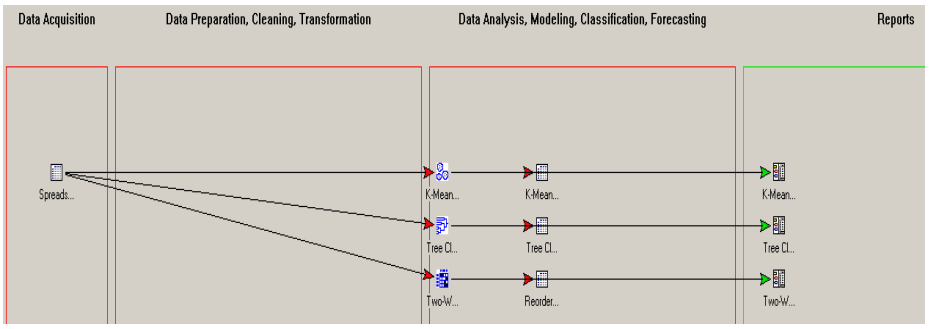


Рис. 78. **Вікно Data Miner після виконання проекту**

Операції створення (видалення) зв'язків можна виробляти і вручну. Також можна переглянути результати (у стовпці звітів). Докладні звіти створюються за замовчуванням для кожного виду аналізу. Для робочих книг результатів доступна повна функціональність системи STATISTICA, за якими можливо переглянути результати, відредагувати параметри аналізу. Після проведення дискримінантного та факторного аналізу вікно Data Miner виглядатиме таким чином (рис. 79).

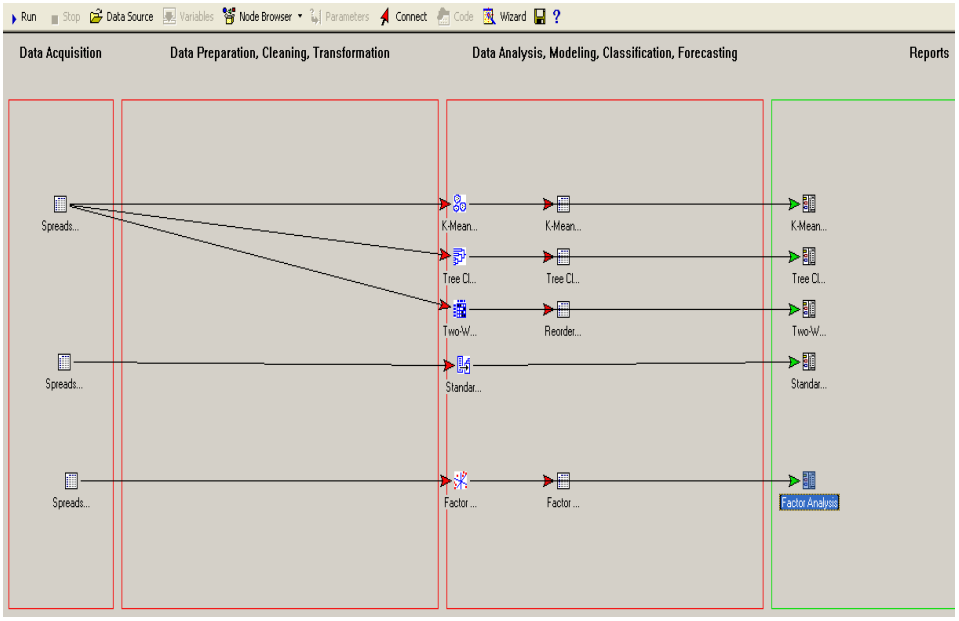


Рис. 79. **Вікно Data Miner після проведення багатовимірного аналізу**

Результати проведеного багатовимірного аналізу даних можливо переглянути у частині вікна Riports.

Крім того, в диспетчері вузлів STATISTICA Data Miner містяться різноманітні процедури розрахунків описових статистик, кореляційних матриць, дисперсійного аналізу, регресійного аналізу, дерев класифікації, а також методи узагальнення часових рядів і прогнозування. Всі ці інструменти можна використовувати для проведення складного аналізу в автоматичному режимі, а також для оцінювання якості моделей.

Лабораторне заняття на тему "Застосування дерев класифікації у вирішенні завдань інтелектуального аналізу даних"

Мета роботи – ознайомитись з алгоритмом побудови дерев класифікації даних з використанням STATISTICA Data Miner.

Завдання – провести класифікацію даних за допомогою дерев класифікації.

Методичні рекомендації

Досліджуються страхові компанії України. Вихідні дані для проведення класифікації подані на рис. 80. У дослідження включені дані по 30 страховим компаніям України. Зареєстровані такі показники ефективності роботи підприємства за 2011 р.: прибуток, фонд заробітної плати; інтегральний показник страхових виплат; рівень надійності (низький – 3; середній – 2; високий – 1); плинність кадрів (низька – 1; середня – 2; висока – 3).

	1 Назва страхової компанії	2 Прибуток	3 Фонд заробітньої плати	4 Інтегральний показник страхових виплат	5 Рівень надійності	6 Плинність кадрів
1	КРАЇНА	47,5	5,34	0,57	2	1
2	АРСЕНАЛ СТРАХУВАННЯ	43,1	2,58	0,57	2	1
3	ЄВРОПЕЙСЬКА	80,8	4,3	0,31	1	3
4	ЮНІВЕС	84,8	2,46	0,4	2	2
5	ТЕКОМ	61,9	0,75	0,6	2	1
6	УКРАЇНА	59,9	4,45	0,38	2	2
7	УКРАЇНСКА ЕКОЛОГІЧНА	80,1	0,77	0,33	1	3
8	СТРАХОВА КОМПАНІЯ "ПЕРША"	62,6	4,63	0,31	1	3
9	ОВЕ УКРАЇНА	48,1	4,19	0,54	3	1
10	АРМА	34	0,85	0,41	2	2
11	НОВА	60	4,72	0,82	3	1
12	ВІЙСЬКОВО-СТРАХОВА КОМПАНІЯ	39,1	2,65	0,57	2	1
13	СТРОЙПОЛІС	39,9	5,35	0,31	1	3
14	НАФТОГАЗСТРАХ	24,9	5,54	0,42	2	2
15	ІНВЕСТСЕРВІС	50,6	0,72	0,81	3	1
16	ГЛОБУС	50,6	0,74	0,43	2	2
17	ЕКСПРЕС СТРАХУВАННЯ	54,4	2,43	0,79	3	1
18	ГРАВЕ УКРАЇНА	53,4	2,63	0,58	2	1
19	ГОРОДСЬКА СТРАХОВА КОМПАНІЯ	42,5	2	0,57	2	1
20	АСКО-ДОНБАС	49,7	4,35	0,79	3	1
21	УКРАЇНСЬКИЙ СТРАХОВИЙ ДІМ	49,6	0,6	0,62	2	1
22	ІНДІГО	45,8	3	0,81	3	1
23	ДОБРОБУТ ТА ЗАХИСТ	81,9	5,57	0,31	1	3
24	НАСТА	84,3	2,65	0,34	1	3
25	ЕТАЛОН	55,8	2,86	0,3	1	3
26	РАРИТЕТ	61,5	4,45	0,38	2	2
27	СТАТУС	82,2	5,43	0,56	2	2
28	ЕНЕРГОПОЛІС	58,9	4,34	0,41	2	1
29	ГАРАНТІЯ СО	43,6	0,71	0,4	2	1
30	ЮТІКО (UTICO)	35,4	4,36	0,55	2	1

Рис. 80. **Вихідні дані**

Після вводу даних виберемо змінні (рис. 81).

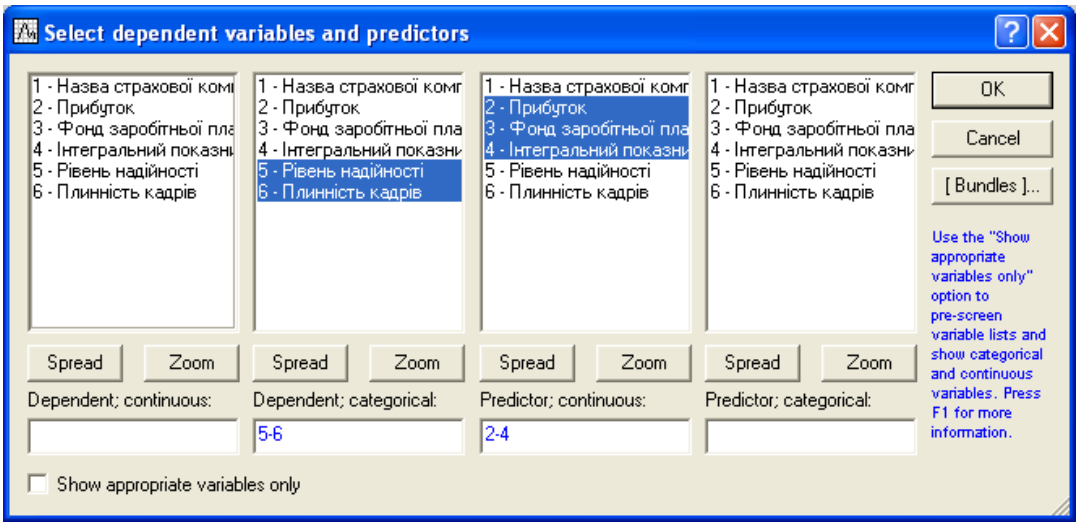


Рис. 81. **Вікно вибору змінних**

Далі запускаємо проект **Data Mining** та за допомогою процедури **Node Browser** обираємо **Classification Trees/Classification from Categorical and Ordered Predictors.** Результати класифікації представлено на рис. 82.

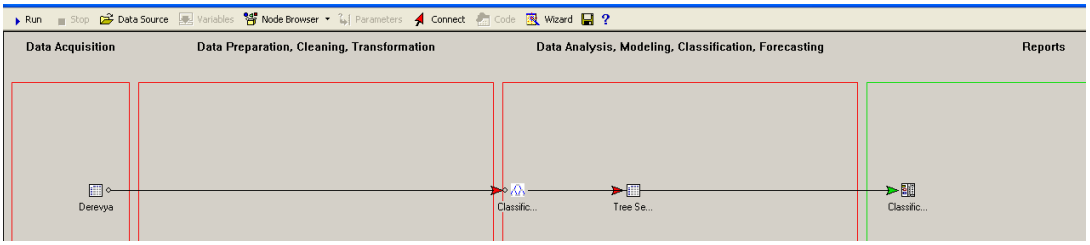


Рис. 82. **Вікно проекту Data Mining**

Результати аналізу представлені у вигляді структури дерева (рис. 83), яка поділяється на ліву та праву гілку, що містять по два вузла.

Tree Structure (Derevya)
Child nodes, observed class n's,
predicted class, and split condition for each node

Node	Left branch	Right branch	n in cls 1	n in cls 2	n in cls 3	Predict. class	Split constant	Split variable
1	2	3	16	7	7	1	-0,476875	Інтегральний показник страхових виплат
2	4	5	2	6	7	3	-0,354865	Інтегральний показник страхових виплат
3			14	1	0	1		
4			0	0	7	3		
5			2	6	0	2		

Рис. 83. Номера вузлових вершин

Так, ліва вершина містить два вузла: 2-й та 4-й, права – 3-й та 5-й. Далі, з першого рядка таблиці, зображеної на рис. 83, випливає, що в першій вершині кількість страхових компаній, які мають рівень плинності

кадрів низький – 16; середній – 7; високий – 7, класифіковані як такі, що мають низький рівень плинності кадрів. З вершини 1 виходять дві гілки (права та ліва) з відповідними вершинами 2 та 3.

Умова розділення страхових компаній за вершинами 2 та 3 така: якщо значення інтегрального показника страхових виплат менше або дорівнює 0,477, то плинність кадрів висока. 15 страхових компаній, які мають рівень плинності кадрів низький – 2; середній – 6; високий – 7 класифіковані як такі, що мають високий рівень плинності кадрів. Решта страхових компаній, які мають рівень плинності кадрів низький – 14; середній – 1, класифіковані як такі, що мають низький рівень плинності кадрів. З рядків 4 та 5 випливає, що якщо значення інтегрального показника страхових виплат менше або дорівнює 0,355, то 7 страхових компаній мають високий рівень плинності кадрів, 8 – середній. Інтерпретація результатів значно спрощується за допомогою графа дерева класифікації (рис. 84).

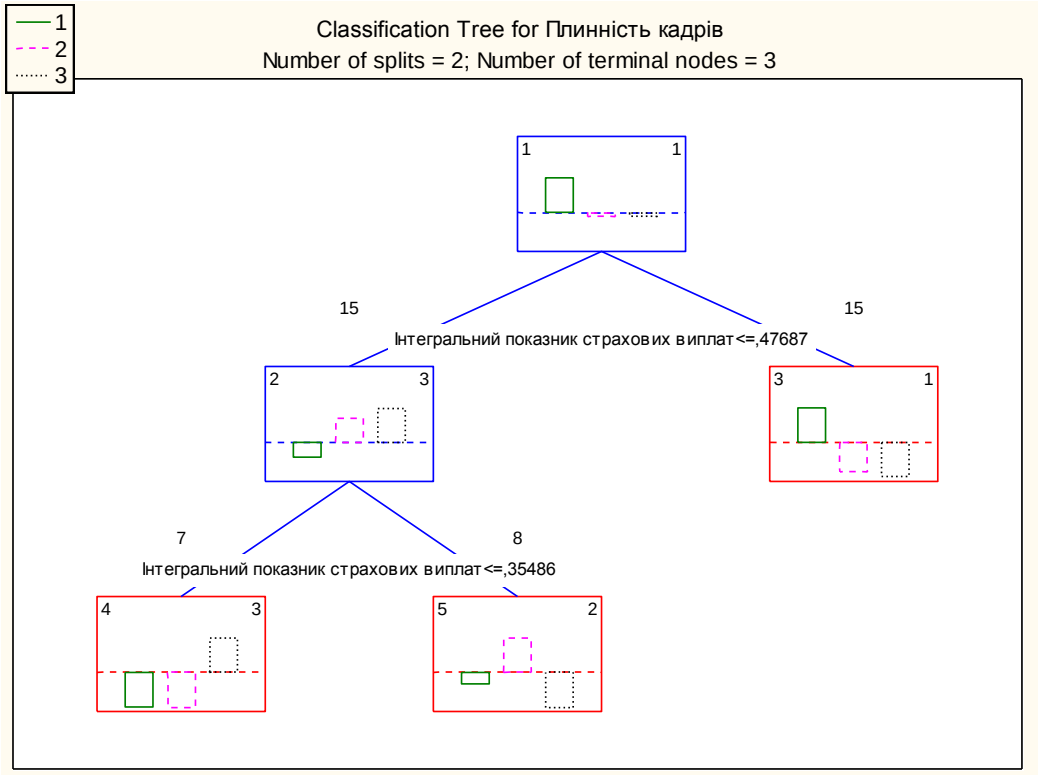


Рис. 84. **Дерево класифікації за рівнем плинності кадрів**

У таблиці результатів класифікації (рис. 85) виводиться інформація про те, скільки спостережень кожного з класів віднесено за результатами кластеризації до того чи іншого класу (за строками матриці), склад початкових класів (за стовпцями матриці) та об’єм вибірки.

Class	Predicted Class x Obser Predicted (row) x observe Learning sample N = 30 Class 1	Class 2	Class 3
1	14	1	0
2	2	6	0
3	0	0	7

Рис. 85. **Вікно результатів класифікації**

Також страхові компанії класифіковано за рівнем надійності, результати представимо графічно (рис. 86).

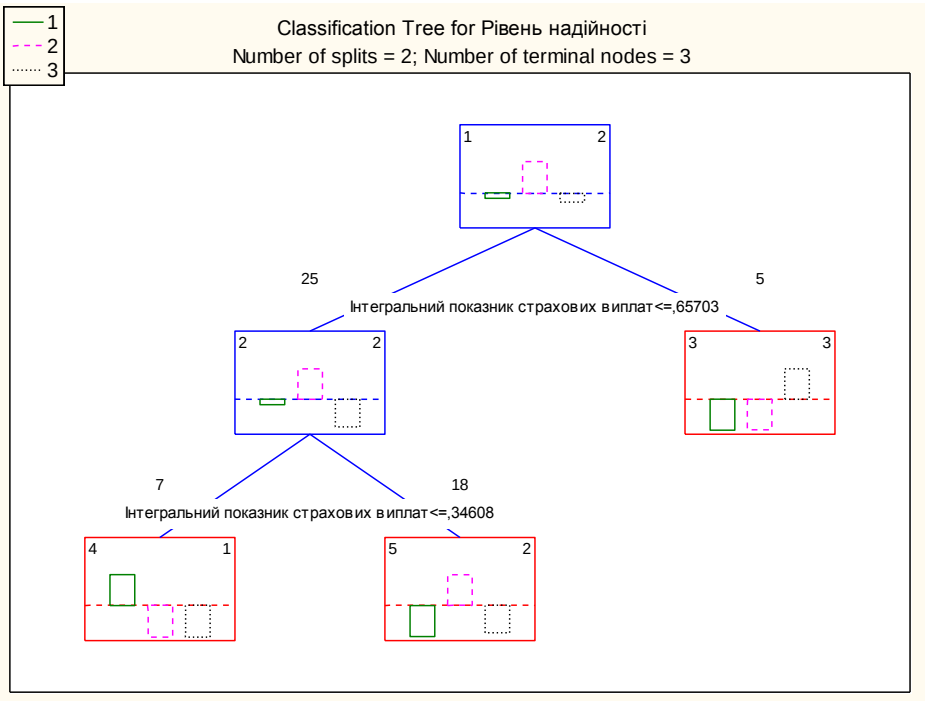


Рис. 86. Дерево класифікації за рівнем надійності

Метод дерев класифікації є гнучким засобом для передбачення користувачем приналежності спостережень до відповідних класів. Застосування цього методу в STATISTICA Data Mining надає можливість класифікувати одразу за декількома змінними та різноманітними способами, що полегшує аналіз та підвищує достовірність результатів.

Рекомендована література

1. Барсегян А. А. Технология данных Data Mining, Visual Mining, Text Mining, OLAP / А. А. Барсегян. – СПб. : БХВ-Петербург, 2007. – 384 с.

2. Боровиков В. П. Популярное введение в программу STATISTICA. / В. П. Боровиков – М. : Издательский дом "Вильямс", 2008 – 267 с.

3. Буреева Н. Н. Многомерный статистический анализ с использованием ППП “STATISTICA” / Н. Н. Буреева – Н. Новгород, Статистика, 2007. – 112 с.

4. Захарченко Н. И. Бизнес-статистика и прогнозирование в MS Exel / Н. И. Захарченко: – М. : Издательский дом "Вильямс", 2004. – 208 с.

5. Куприенко Н. В. Статистика. Методы анализа распределений. Выборочное наблюдение / Н. В. Куприенко. – СПб. : Изд. Политехн. ун-та, 2009. – 138 с.

6. Халафян А. А. STATISTICA 6. Статистический анализ данных / А. А.Халафян – М. : ООО "Бином-Пресс", 2008. – 512 с.

7. Статистичний збірник "Регіони України" – 2010 р, [Електронний ресурс]. – Режим доступу : http://www.ukrstat.gov.ua.

8. Статистичний щорічник України – 2010 р., [Електронний ресурс]. – Режим доступу : http://www.ukrstat.gov.ua.

НАВЧАЛЬНЕ ВИДАННЯ

Лабораторний практикум з навчальної дисципліни "БІЗНЕС-СТАТИСТИКА"

для студентів спеціальності 8.03050601 "Прикладна статистика" денної форми навчання

Укладачі: **Раєвнєва** Олена Валентинівна
Чанкіна Ірина Володимирівна
Гольтяєва Людмила Анатоліївна

Відповідальний за випуск **Раєвнєва О. В.**

Редактор **Пушкар І. П.**

Коректор **Бриль В. О.**

План 2013 р. Поз. № 117.
Підп. до друку Формат 60 × 90 1/16. Папір MultiCopy. Друк Riso.
Ум.-друк. арк. 4,25. Обл.-вид. арк. 5,31. Тираж прим. Зам. №

Видавець і виготівник – видавництво ХНЕУ, 61166, м. Харків, пр. Леніна, 9а

Свідоцтво про внесення до Державного реєстру суб'єктів видавничої справи ***Дк № 481 від 13.06.2001 р.***