Теми практичних занять

№ з/п	Назва теми	Кількість годин	
		ДФН	ЗФН
1	Статистичні методи в суспільній географії: основні поняття і терміни. Усвідомлення суспільно-географічної та математичної сутності поставленого завдання	8	
2	Статистичні показники: поняття, форми, види. Матриця даних	8	4
3	Побудова одновимірних статистичних моделей векторів змінних	8	2
4	Побудова двовимірних статистичних моделей на матриці вихідних даних	8	2
5	Кластер-аналіз вихідних даних. Інтерпретація отриманих результатів	6	2
6	Факторний аналіз вихідних даних. Інтерпретація отриманих результатів	6	2
7	Просторові змінні як випадкові величини	4	
	Разом	**48**	**12**

Практичні роботи виконуються за індивідуальними завданнями для кожного студента. Індивідуальність завдань забезпечується тим, що студенти працюють з власною інформаційною базою даних з обраної теми дослідження.

МЕТОДИЧНІ РЕКОМЕНДАЦІЇ ДО ПРАКТИЧНИХ ЗАНЯТЬ

Практичне заняття 1-4
СТАТИСТИЧНІ МЕТОДИ В СУСПІЛЬНІЙ ГЕОГРАФІЇ: ОСНОВНІ ПОНЯТТЯ І ТЕРМІНИ. УСВІДОМЛЕННЯ СУСПІЛЬНО-ГЕОГРАФІЧНОЇ ТА МАТЕМАТИЧНОЇ СУТНОСТІ ПОСТАВЛЕНОГО ЗАВДАННЯ

Питання для обговорення:

1. Сутність і співвідношення понять:
   методи, методика і методологія;
   статистика і геостатистика;
   статистичні методи, математичні методи і математико-статистичні методи;
   модель, математична модель, статистична модель;
   моделювання, географічне моделювання, математичне моделювання, статистичне моделювання;
   географічна інформація (дані) і статистична інформація (дані);
   статистичне дослідження і статистичне спостереження.
2. Способи отримання статистичної інформації.
3. Класифікація статистичних досліджень за різними критеріями.
4. Статистичні таблиці, їх види.
5. Етапи проведення статистичного дослідження.

Завдання до практичної роботи:

Завдання 1. Побудуйте схеми, на яких відобразіть, як співвідносяться поняття:
методи, методика і методологія;
статистика і геостатистика;
статистичні методи, математичні методи і математико-статистичні методи;
модель, математична модель, статистична модель;
моделювання, географічне моделювання, математичне моделювання, статистичне моделювання;
географічна інформація (дані) і статистична інформація (дані);
статистичне дослідження і статистичне спостереження.

На схемах (або нижче як висновок) зробіть відповідні пояснення, які б відображали, сутність понять та чим відрізняються і чим схожі ці поняття.

Завдання 2. Користуючись сайтом Державної служби статистики, побудуйте прості, групові та комбіновані статистичні таблиці. Визначте підмет і присудок побудованих статистичних таблиць.

Завдання 3. Користуючись сайтом Державної служби статистики,

розробіть етапи невеликого статистичного дослідження (тематика – на вибір студента):

I етап: 1.1. Сформулювати мету і завдання, об’єкт і предмет дослідження.

1.2. Розробити план і програму дослідження.

1.3. Скласти макети статистичних таблиць.

II етап: 2.1. Зібрати статистичний матеріал (5-10 показників).

III етап: 3.1. Звести статистичний матеріал в розроблені макети статистичних таблиць.

3.2. Статистично обробити отриманий матеріал відповідно до програми аналізу даного дослідження (*за можливості*).

3.3. Графічно відобразити отримані статистичні результати (*для прикладу*).

IV етап: 4.1. На підставі аналізу отриманих даних зробити відповідні висновки, розробити необхідні практичні рекомендації та провести літературне оформлення.

Практичне заняття 5-8
СТАТИСТИЧНІ ПОКАЗНИКИ:
ПОНЯТТЯ, ФОРМИ, ВИДИ.
МАТРИЦЯ ДАНИХ

Питання для обговорення:

1. Поняття «статистичний показник».
2. Класифікація статистичних показників за різними критеріями.
3. Абсолютні показники, їх види та форми.
4. Відносні показники, їх види та форми.
5. Індекси та їх використання в географічних дослідженнях.
6. Поняття «матриці даних». Вимоги до складання матриці даних

Завдання до практичної роботи:

Завдання 1. Розгляньте класифікацію статистичних показників за різними критеріями – способом одержання, часовою ознакою, суттю досліджуваних явищ, ступенем агрегування досліджуваних явищ, об'єктом дослідження, взаємозалежністю. Користуючись сайтом Державної служби статистики – розділом «Статистична інформація» або розділом «Публікації» (наприклад, Україна у цифрах 2022: статистичний збірник / Відп. за випуск О. А. Вишневська. – К., 2023. – 34 с. або Статистичний щорічник України 2021 / За ред. І. Є. Вернера. – К., 2022. – 447 с.), наведіть по 3-4 приклади *конкретних статистичних показників* кожного виду. Відповідь оформіть у вигляді таблиці.

Таблиця

Критерій класифікації статистичних показників	**Види статистичних показників за вказаним критерієм**	**Приклади статистичних показників**
часова ознака	моментні	а) чисельність наявного населення України на 01. 01. 2022 р. = 41,2 млн. осіб б) середній розмір місячної пенсії пенсіонерів України на кінець 2022 року = 4622,6 грн. ……
	інтервальні	

Завдання 2. Розрахуйте відносні статистичні показники, якщо відомо наступне:

а) розрахуйте показники динаміки (темпи росту і приросту ланцюговим і базисним способом) за варіантами (*на вибір студента*). Побудуйте графіки.

Роки	***Варіант 1.***	***Варіант 2.***	***Варіант 3.***	**Темпи росту**	**Темпи приросту**

	Обсяг іноземних інвестицій в Україну, на початок року, млн. дол. США	Обсяг іноземних інвестицій з України, на початок року, млн. дол. США	Обсяг експорту послуг України, млн. дол. США	(зменшення), %		(зменшення), %	
				ланцюговим способом	базисним способом	ланцюговим способом	базисним способом
2000	3281,8	98,5	3655,1				
2001	3875,0	170,3	3731,9				
2002	4555,3	155,7	4303,8				
2003	5471,8	144,3	4524,9				
2004	6794,4	166,0	5612,7				
2005	9047,0	198,6	6443,2				
2006	16890,0	219,5	7791,8				
2007	21607,3	243,3	9435,1				
2008	29542,7	6196,6	12260,1				
2009	35616,4	6203,1	10129,7				
2010[1]	38992,9	5760,5	11936,3				
2011[1]	45370,0	6402,8	14180,3				
2012[1]	48197,6	6435,4	14096,2				
2013[1]	51705,3	6568,1	14233,2				
2014[1]	53704,0	6702,9	11520,8				
2015[1]	40725,4	6456,2	9736,6				
2016[1]	32122,5	6315,2	9868,0				
2017[1]	31230,3	6346,3	10790,3				
2018[1]	31606,4	6322,0	11679,9				
2019[1]	32905,1	6294,4	15660,9				
2020[1]	35809,6	6272,7	11547,2				

[1] Без урахування тимчасово окупованої території Автономної Республіки Крим, м. Севастополя та частини тимчасово окупованих територій у Донецькій та Луганській областях.

2) розрахуйте показники структури і координації за варіантами (*на вибір студента*). Графічно візуалізуйте отримані результати.

Варіант 1

Гендерний склад ВРУ протягом років незалежності країни

Номер скликання ВРУ	Кількість народних депутатів		
	всього	з них жінок	з них чоловіків
I (1992-1993 рр.)	475	12	463

II (1994-1997 рр.)	436	18	418
III (1998-2001 рр.)	477	38	439
IV (2002-2005 рр.)	509	28	481
V (2006-2007 рр.)	483	42	441
VI (2008-2012 рр.)	541	42	499
VII (2013-2014 рр.)	478	46	432
VIII (2015-2019 рр.)	468	56	412
IX (2020-2021 рр.)	440	92	348

Варіант 2
Вікова структура населення України на 01. 01. 2022 р.

	Усього, осіб	**з них у віці**		
		0–14 років	**15–64 роки**	**65 років і старше**
Україна[1]	**40997698**	**6119886**	**27646706**	**7231106**
Вінницька	1502430	228250	1004375	269805
Волинська	1018628	196450	685096	137082
Дніпропетровська	3093176	467034	2081587	544555
Донецька	4046487	423519	2674678	948290
Житомирська	1179801	190552	790179	199070
Закарпатська	1241643	242204	846345	153094
Запорізька	1637673	232973	1102443	302257
Івано-Франківська	1349096	224384	927622	197090
Київська	1789300	316611	1201941	270748
Кіровоградська	897297	131792	597961	167544
Луганська	2098324	189316	1396732	512276
Львівська	2459763	396357	1688299	375107
Миколаївська	1091106	164901	737530	188675
Одеська	2340332	393810	1569062	377460
Полтавська	1344445	188739	912376	243330
Рівненська	1140724	229140	764199	147385
Сумська	1033580	132578	702848	198154
Тернопільська	1018462	156857	699887	161718
Харківська	2583325	349863	1782295	451167
Херсонська	1000166	158536	673166	168464
Хмельницька	1225666	190583	819510	215573
Черкаська	1157115	157088	776993	223034
Чернівецька	887392	151873	606264	129255
Чернігівська	950773	125440	631852	193481
м.Київ	2910994	481036	1973466	456492

[1] Без урахування тимчасово окупованої території Автономної Республіки Крим, м. Севастополя та частини тимчасово окупованих територій у Донецькій та Луганській областях.

Варіант 3
Робоча сила за статтю, типом місцевості та віковими групами у 2021 році[1]

	всього	жінки	чоловіки	міська місцевість	сільська місцевість
населення віком 15 років і старше	**17405,0**	**8293,0**	**9112,0**	**11948,7**	**5456,3**
з нього					
15-70 років	17321,6	8248,2	9073,4	11906,7	5414,9
20-64 роки	17018,3	8091,9	8926,4	11716,4	5301,9
працездатного віку	16666,8	7911,6	8755,2	11483,3	5183,5
за віковими групами					
15-24 роки	1128,9	502,8	626,1	671,0	457,9
25-29 років	1839,0	748,9	1090,1	1217,1	621,9
30-34 роки	2459,9	1056,6	1403,3	1695,6	764,3
35-39 років	2664,2	1206,4	1457,8	1978,8	685,4
40-49 років	4723,3	2353,8	2369,5	3332,5	1390,8
50-59 років	3851,5	2043,1	1808,4	2588,3	1263,2
60-70 років	654,8	336,6	318,2	423,4	231,4
71 рік і старше	83,4	44,8	38,6	42,0	41,4

[1] Без урахування тимчасово окупованої території Автономної Республіки Крим, м. Севастополя та частини тимчасово окупованих територій у Донецькій та Луганській областях.

3) розрахуйте показники порівняння (просторового порівняння, порівняння зі стандартом) для споживання населенням України якогось продукту харчування (*на вибір студента*) у 2020 р. Проранжируйте регіони України за отриманими показниками. Побудуйте гістограми.

Регіони	**Споживання населенням, 2020 р.**[1]									
	м'яса та м'ясо-продуктів, кг	**молока та молоко-продуктів, кг**	**яєць, шт.**	**хлібних продуктів, кг**	**картоплі, кг**	**овоче-баштанних культур, кг**	**плодів, ягід, винограду, кг**	**риби та рибо-продуктів, кг**	**цукру, кг**	**олії, кг**
Раціональні норми споживання	**83**	**380**	**290**	**101**	**124**	**161**	**90**	**20**	**38**	**13**
Мінімальні норми споживання	**52**	**341**	**231**	**94**	**96**	**105**	**68**	**12**	**32**	**8**
Україна	**53,8**	**201,9**	**228**	**96,6**	**134**	**164**	**56,5**	**12,4**	**27,8**	**12,3**
Вінницька	57,5	200,9	299	111,5	168	174,5	63,6	14,7	30,1	13,3
Волинська	55,1	209,2	285	110,1	178,5	164,5	52,3	13,5	31,4	12,5

Дніпропетровська	65,6	196,4	303	88,8	115,7	168,1	63,7	13	27,3	13,1
Донецька	48,7	171,2	251	93,3	92,5	140,4	42,3	12,5	26,8	10,8
Житомирська	54,2	202,8	318	103,7	183,5	169,2	49,8	14,4	26,6	11,5
Закарпатська	52,2	232,8	281	116,7	153,5	170,1	52,2	8	28,4	12,6
Запорізька	56,1	180,4	275	93,2	96,2	172,6	52,5	13,2	27,4	12,2
Івано-Франківська	49,8	300,6	276	112,8	187,3	154,7	54,4	9,4	31,5	12,5
Київська	60,6	209,4	301	75,9	118,9	173,3	77,5	16,1	23,5	11,9
Кіровоградська	57,8	229,2	327	100,4	144,8	186,6	49,4	13,4	32,6	11,9
Луганська	43,5	150,6	217	90,1	110,9	106,5	43	9,2	27,9	11,6
Львівська	53,1	225,9	278	93,6	180,1	182,8	58,5	9,6	29,2	13,6
Миколаївська	57,2	205,4	266	105,8	123,5	172,4	60	13,4	28,3	13,6
Одеська	50,1	180,2	263	98,2	96,7	165,3	59,5	15,1	26,3	14
Полтавська	50,9	195,5	284	97,2	137,7	179	58,3	12,2	27,9	11,2
Рівненська	48,5	190,5	278	90,5	158,8	145,5	44,7	11	27,6	10,6
Сумська	51,4	180,4	275	98,5	168	163,4	43,3	10,6	29,3	10,7
Тернопільська	47,7	237,8	280	93,6	156,5	162,8	52,1	9,4	24,8	13,1
Харківська	53,5	202,6	272	91,9	103,5	166,4	50,7	9,5	23,6	11,2
Херсонська	55	195,8	255	99,6	135,6	181,5	50,6	14,5	30,6	12,9
Хмельницька	53,8	204,2	289	115,4	169,2	162,6	62,8	11,1	28,9	12,7
Черкаська	52,1	222,2	275	113,2	164,4	168,3	60,9	14,2	33,8	12,5
Чернівецька	44,4	220,1	281	109,2	143,7	175,7	65,7	10,3	29,9	12,9
Чернігівська	53	207,5	269	114,1	164,1	177,5	55,7	12,8	38,2	13,4

[1] Без урахування тимчасово окупованої території Автономної Республіки Крим, м. Севастополя та частини тимчасово окупованих територій у Донецькій та Луганській областях.

4) розрахуйте показники планового завдання і виконання плану за варіантами (*на вибір студента*):

Варіант 1

Номер супермаркетів	**Реалізація продукції, млн. грн.**		
	фактично за минулий рік	**планове завдання на звітний рік**	**фактично у звітному році**
1	22,1	22,4	23,5
2	32,4	32,5	35,7
3	41,6	42,7	40,2

Варіант 2

	Урожайність сільськогосподарських культур, ц/га

	фактично за 2021 р.	планове завдання на 2022 р.	фактично у 2022 р.
зернових та зернобобових	53,9	54,6	45,8
картоплі	166,4	170,0	173,5
соняшнику	24,6	24,0	21,6

Варіант 3

	Виробництво чавуну на металургійному комбінаті, тис. т		
	фактично за минулий рік	планове завдання на звітний рік	фактично у звітному році
Передільний	42	45	48
Ливарний	40	42	43
Хромонікелевий	37	37	37
Феромарганцевий	30	28	25
Ферофосфорний	22	25	24

5) розрахуйте показники інтенсивності за варіантами (*на вибір студента*), використовуючі наступні дані:

Регіони	Середня чисельність наявного населення у 2021 р., тис. осіб	Площа, тис. км^2	Використання електроенергії, тис. кВт*год.	Обсяг валового регіонального продукту України в 2021 р., млн. грн.
Україна	**41377,8**[1]	**603,5**	**83888551**[1]	**5450849**[1]
Вінницька	1519,3	26,5	1728791	173531
Волинська	1024,4	20,1	683202	92535
Дніпропетровська	3119,3	31,9	22012015	582363
Донецька	4079,8	26,5	8901484	283326
Житомирська	1187,3	29,8	1268438	113919
Закарпатська	1247,3	12,8	443351	75626
Запорізька	1652,5	27,2	7454972	228906
Івано-Франківська	1356,5	13,9	2577862	119680
Київська	1791,8	28,1	3014648	291519
Кіровоградська	911,9	24,6	2284442	99564
Луганська	2112,1	26,7	1161036	52135
Львівська	2487,9	21,8	2302941	296182
Миколаївська	1100,1	24,6	3299320	124162
Одеська	2359,7	33,3	2320087	271669
Полтавська	1361,9	28,8	3822701	266694
Рівненська	1145,1	20,1	2678332	88859
Сумська	1044,6	23,8	1196932	105254
Тернопільська	1026,1	13,8	419851	81485

Харківська	2616,4	31,4	3596405	319796
Херсонська	1009,2	28,5	997488	88182
Хмельницька	1236,3	20,6	1710960	119876
Черкаська	1169,5	20,9	1773689	131154
Чернівецька	893,5	8,1	1900689	54582
Чернігівська	968,0	31,9	786969	113474
м. Київ	2957,2	0,8	5551947	1276376

[1] Без урахування тимчасово окупованої території Автономної Республіки Крим, м. Севастополя та частини тимчасово окупованих територій у Донецькій та Луганській областях.

6) за допомогою всіх можливих видів відносних показників проаналізувати нижче наведені статистичні дані щодо наукової діяльності в Україні в 2017-2020 рр. (*за варіантами на вибір студента*):

Наукова діяльність України в 2017-2020 рр.

	всього				**з них жінки**			
	2017	**2018**	**2019**	**2020**	**2017**	**2018**	**2019**	**2020**
ВВП, млрд. грн.	**2981,2**	**3560,3**	**3977,2**	**4222,0**				
Витрати на виконання наукових досліджень і розробок, млн. грн.	**13379,3**	**16773,7**	**17254,6**	**17022,4**				
Кількість працівників, задіяних у виконанні наукових досліджень і розробок	**94274**	**88128**	**79262**	**78860**	**44173**	**41323**	**36989**	**37501**
Варіант 1 **з них за рівнем освіти**								
мають вищу освіту	81783	76455	68744	68060	38671	36202	32364	32277
доктори наук	6942	7043	6526	7060	1883	1884	1766	2053
доктори філософії (кандидати наук)	19219	18806	16929	17949	9030	8837	7900	8530
магістри (спеціалісти)	46612	43291	39148	37406	22940	21586	19489	18670
бакалаври (молодші бакалаври, молодші спеціалісти)	9010	7315	6141	5645	4818	3895	3209	3024
Інші рівні освіти	12491	11673	10518	10800	5502	5121	4625	5224
Варіант 2 **з них за галузями наук**								
природничі	22140	21805	21305	21106	10611	10138	10039	10064

інженерія та технології	48985	43423	39033	36837	18769	16625	14982	14286
медичні науки та науки про здоров'я	5228	5461	4192	4914	3618	3789	2993	3384
сільськогосподарські та ветеринарні науки	7451	7428	6508	6212	4352	4208	3711	3611
суспільні	7239	6968	6096	7187	4787	4655	3994	4565
гуманітарні науки та мистецтво	3231	3043	2128	2604	2036	1908	1270	1591
Варіант 3 **з них за категоріями персоналу**								
дослідники	59392	57630	51121	51427	26533	25780	22649	23338
техніки	9144	8553	7470	7117	5368	4994	4511	4137
допоміжний персонал	25738	21945	20671	20316	12272	10549	9829	10026

Завдання 3. Користуючись сайтом Державної служби статистики або сайтами Головних управлінь статистики регіонів, проведіть вибірковий збір статистичних показників по Україні або будь-якому регіону України в територіальному розрізі (виходячи з обраного предмета дослідження). Необхідно зібрати до 30 статистичних показників. Побудуйте інформаційну базу даних по статистичній сукупності. База даних повинна включати як абсолютні, так і відносні показники.

Практичне заняття 9-12
ПОБУДОВА ОДНОВИМІРНИХ СТАТИСТИЧНИХ МОДЕЛЕЙ ВЕКТОРІВ ЗМІННИХ

Питання для обговорення

1. Сутність одновимірного аналізу
2. Нормальний розподіл випадкових величин
3. Мода
4. Медіана
5. Цетральний момент першого, другого, третього порядків
6. Види середнього значення

Рекомендації до підготовки та проведення заняття

Наступним завданням практичної роботи є виявлення залежності динаміки певних показників від часу, включає методи згладжування і аналітичного виміру. Згладжування базується на процедурі визначення усередненої траєкторії розвитку показників в минулому і її продовження на майбутнє. Аналітичне вимірювання передбачає підбір математичної функції, яка найкращим чином відображає тенденцію динаміки показника в часі, і розрахунок на його основі прогнозних значень даного показника. Виконується за допомогою одновимірного аналізу, здійснюється в програмі Statistica 8.0. Одновимірний аналіз відноситься до базових, виконується шляхом «Statistics → Basic Statistics/Tables → Descriptive Statistics».

Одновимірний аналіз містить одну випадкову величину (аналіз однієї випадкової величини). В основі даного аналізу є дослідження властивостей і характеристик рядів випадкових величин, встановлення приналежності цих рядів до певного теоретичного розподілу, визначення подібності або відмінності даного ряду в порівнянні з іншими рядами.

Основними вихідними складовими одновимірного аналізу є поняття: простий статистичний ряд (сукупність), упорядкований статистичний ряд (ранжируваних), варіаційний ряд, генеральна сукупність і вибірка, обсяг вибірки, частота і ймовірність, закон розподілу, критерії однорідності та інші.

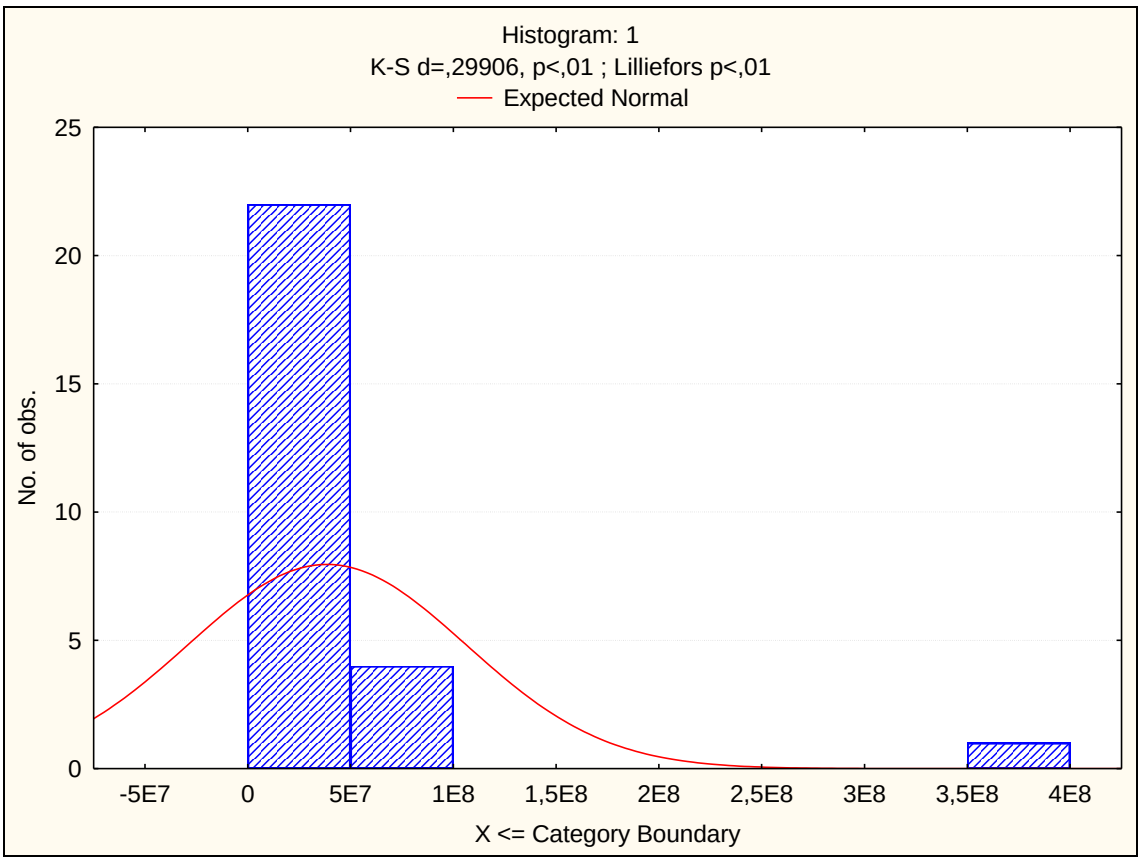

Рис. 1. Нормальний розподіл статистичного показника

Сукупність випадкових величин, отриманих за однакових умов проведення спостереження, дослідів, експериментів називається простою статистичною сукупністю або статистичним рядом. Статистичний ряд – це первинна форма запису статистичного матеріалу у вигляді таблиці. Для наочного представлення вихідного матеріалу за даними статистичного ряду будується графік змін значень даної величини в часі або в просторі або в хронологічній послідовності.

Генеральна сукупність – це нескінченне або кінцеве число елементів або компонентів, що складаються з якісно однорідних показників. Будь-яку частину генеральної сукупності, відібрану за певними правилами, яка характеризує генеральну сукупність, називають статистичної вибіркою. Вибіркову сукупність (вибірку) створюють зазвичай для полегшення обробки інформації.

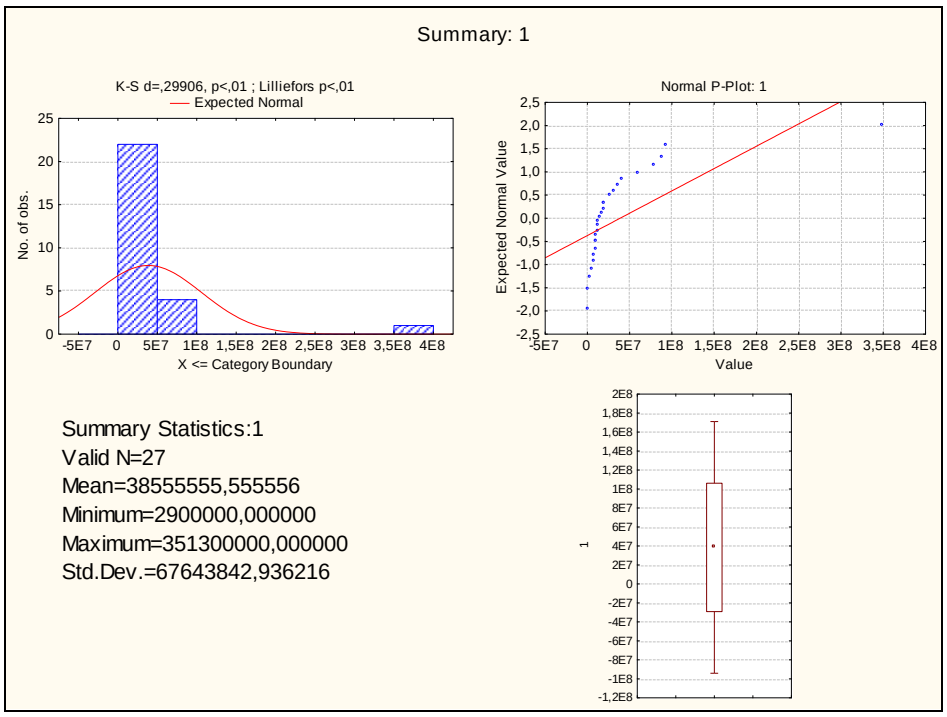

Рис. 2. Приклад одновимірного аналізу статистичних показників

Одним з найбільш складних і важливих питань одновимірного статистичного аналізу є визначення кількості спостережень в дослідженнях для отримання надійного уявлення про характер мінливості ознаки, що вивчається, в генеральній сукупності. Звичайно оптимальний обсяг вибірки пропорційний ступеню мінливості ознаки. Якщо ознака сильно змінюється, то кількість вимірювань слід збільшити. Величину вибіркової сукупності при виконанні географічних досліджень можна визначити двома способами: по таблиці досить великих чисел; розрахунковим способом. В обох випадках кількість спостережень (чисельність або обсяг вибірки) визначається, виходячи з довірчої ймовірності.

Систематизація та впорядкування даних, що представляють статистичну сукупність. Приведення їх у певну систему і характеристика цієї системи. Будь-яка статистична сукупність характеризується обсягом ряду і абсолютної частотою повторення однакового ознаки. Частота – це число, яке показує скільки разів зустрічається дана ознака в досліджуваної сукупності. Закон зміни частоти – це і є закон розподілу. Існує три способи подання закону розподілу: табличний, графічний і аналітичний. Результати обробки вихідних даних, як правило, спочатку завжди оформляються у вигляді статистичних таблиць. Така форма дозволяє надати матеріалу зручність, компактність і раціональність.

Спосіб, який дозволяє в наочній формі отримати уявлення про закономірності розподілу, називається графічним зображенням варіаційного ряду. Існує кілька способів графічного зображення рядів розподілу. При цьому зазвичай використовують дані таблиці емпіричного розподілу. При графічному зображенні рядів розподілу на горизонтальній осі відкладаються значення інтервалів або спостережень значення випадкової величини, а на

вертикальній осі – частоти. Для наочного уявлення варіаційного ряду частіше використовують графічні зображення у вигляді гістограми, полігону, кумуляти, кривою концентрації Лоренса і інші. Гістограма наочно показує розподіл досліджуваних величин, через що подібний спосіб вже часто використовується для ілюстрації особливостей статистичного розподілу. Варіаційний ряд при цьому зображується у вигляді стовпчиків, кордони між якими проходять по координатах, відповідних кордонів між класами. При цьому підстава стовпчиків по ширині дорівнює величині інтервалу ознаки, а висота пропорційна частоті окремих класів. Розглянемо основні статистичні характеристики варіаційних рядів. Одним з основних параметрів статистичного ряду є середнє значення ознаки або центр, щодо якого розподіляються члени сукупності. Значення середніх при вивченні різного роду закономірностей географічних явищ, процесів і об'єктів дуже велике. Вони дозволяють: визначити загальну тенденцію розвитку явищ; оцінювати значення окремої величини шляхом порівняння її з середньою, визначати наявність зв'язку між явищами за допомогою аналізу середніх двох або декількох ознак по територіях або часових проміжків.

При обробці даних географічних досліджень як найважливіших характеристик варіаційного ряду застосовуються різні середні значення. Середні величини прийнято розділяти на прості і зважені. Якщо середні значення обчислюються по безпосередньому переліку значень ознаки в кожній одиниці сукупності, то такі середні називаються простими (невиваженими). Якщо середні обчислюються по вариационному ряду з урахуванням статистичної ваги кожного варіанту, то їх називають зваженими.

Середні величини бувають різного роду. Якщо вид середньої невідомий, то мається на увазі середня арифметична величина.

Середні величини частіше використовують статечні і структурні (порядкові) середні. Статечні середні, в свою чергу, поділяються на середні арифметичні, середні гармонійні, середні квадратичні, середні кубічні та інші. Структурні середні – це мода, медіана, квартили, децили і ін.

Медіаною називається середнє (серединне) значення ознаки ранжированного варіаційного ряду, тобто значення, рівновіддалене від початку і кінця, перебудованого в зростаючому або спадному порядку.

Модою називається ймовірна (що частіше зустрічається) в даному статистичному ряду величина. Мода є найбільшою ординатою кривої розподілу при одновершинная розподілі. У загальному випадку крива розподілу може мати кілька вершин і, відповідно, вона буде мати кілька мод.

Вважається, що нормальний розподіл є однією з емпірично перевірених істин і його положення розглядаються як один з фундаментальних законів природи. Точна форма нормального розподілу «колоколоподібна крива» визначається тільки двома параметрами: середнім і стандартним відхиленням. Для нормального розподілу характерно, що 68% всіх значень знаходяться в межах ± 1 стандартне відхилення від середнього, а діапазон ± 2 стандартних відхилення включає 95% значень. Про прийняття за основу від

нормального розподілу свідчать коефіцієнти асиметрії та ексцесу. Коефіцієнт асиметрії показує відхилення розподілу від симетричного (нормальний розподіл абсолютно симетричне, відповідно коефіцієнт асиметрії дорівнює нулю). Позитивне значення коефіцієнта асиметрії свідчить про наявність розподілу з «довгим правим хвостом», негативне з «довгим лівим хвостом». Коефіцієнт ексцесу показує «гостроту піку» розподілу (для нормального розподілу коефіцієнт ексцесу також дорівнює нулю). Якщо коефіцієнт ексцесу є позитивним, пік є загостреним, якщо є негативним – пік закруглений.

План виконання практичної роботи:

1. Відкриваємо програму Statistica 8.0
2. Натискаємо кнопку New ( ) або Ctrl+N. Натискаємо ОК.
3. З'являється таблицю, в яку необхідно додати дані з бази даних.
4. Для цього копіємо в таблиці Excel базу даних і вставляємо її у таблицю програми Statistica.
5. Дана операція роться наступним чином. У першій ячійці правою кнопкою миші натискуємо у вікні, що відкривається, натискуємо «Paste With Headers → Paste With Both». Таким чином, ми копіюємо у таблицю програми Statistica наші показники, а також назви показників та назви районів/регіонів (рис. 3).

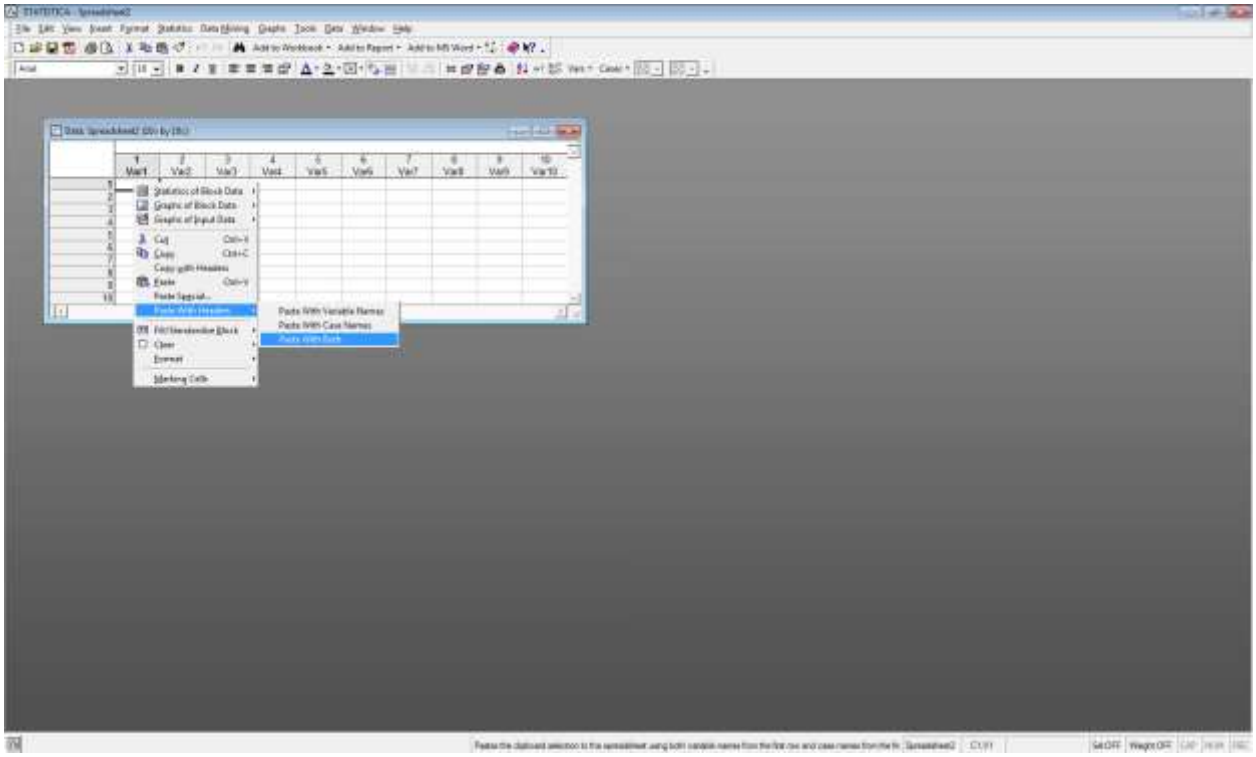

Рис. 3. Додавання бази даних у таблицю програми Statistica

6. Після цього на панелі задач натискаємо «Statistics → Basic

Statistics/Tables → Descriptive Statistics → OK».

7. Відкривається діалогове вікно (рис. 4).

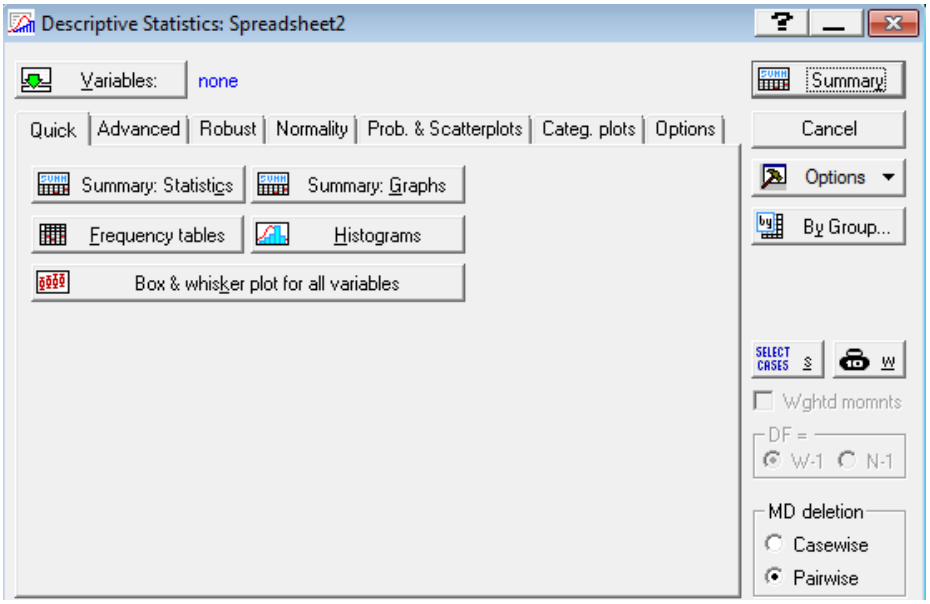

Рис. 4. Діалогове вікно описової статистики

8. Натискаємо кнопку «Variables» та обираємо показники для аналізу.
9. Обираємо опцію «Advanced». Галочки ставимо навпроти «Valid N» (кількість значень), «Mean» (середнє значення), «Median» (медіана), «Mode» (мода), «Standard Deviation» (стандартне відхилення), «CI for Sample SD» (довірчий інтервал), «Variance» (дисперсія), «Skewness», «Kutosis», «Minimum & maximum» (мінімальне та максимальне значення), «Lower & upper quartiles» (нижній та верхній квартилі) (рис. 5).

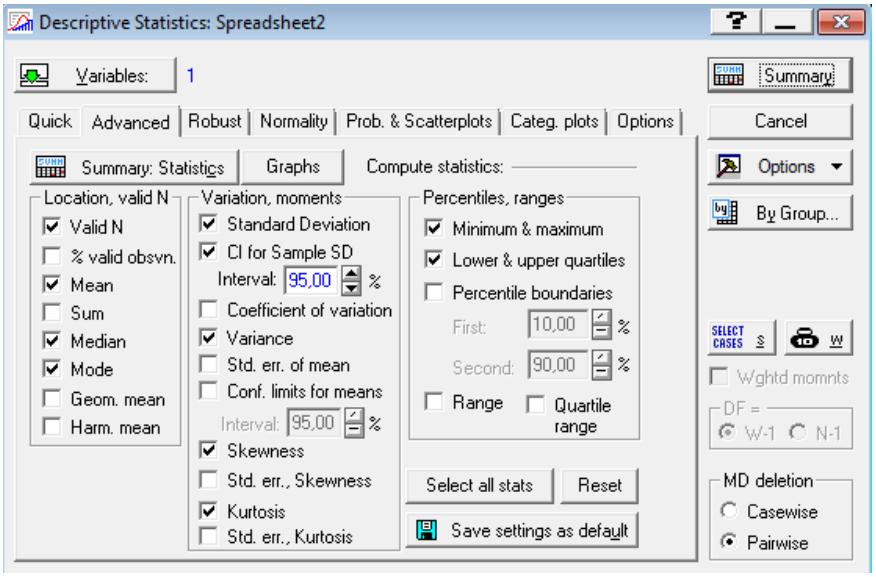

Рис. 5. Опції для описової статистики

10. В опції «Options» галочку ставимо навпроти «Median/Quartiles/Range». Тобто в графіку Box and Whiskers Plot буде відображатися медіана, квартилі, максимальне та мінімальне значення вибірки.
11. Далі ми повертаємося у вкладку «Quick» та натискаємо кнопку «Summary: Graph» (рис. 6).

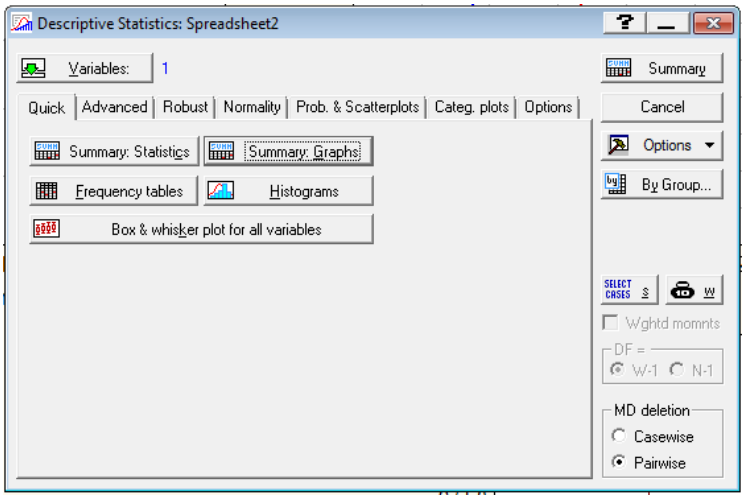

Рис. 6. Натискання кнопки «Summary: Graph»

12. Як результат, з'являється наступний результуючий графік (рис. 7). За цим графіком можна зрозуміти, чи відповідає даний розподіл нормальному.

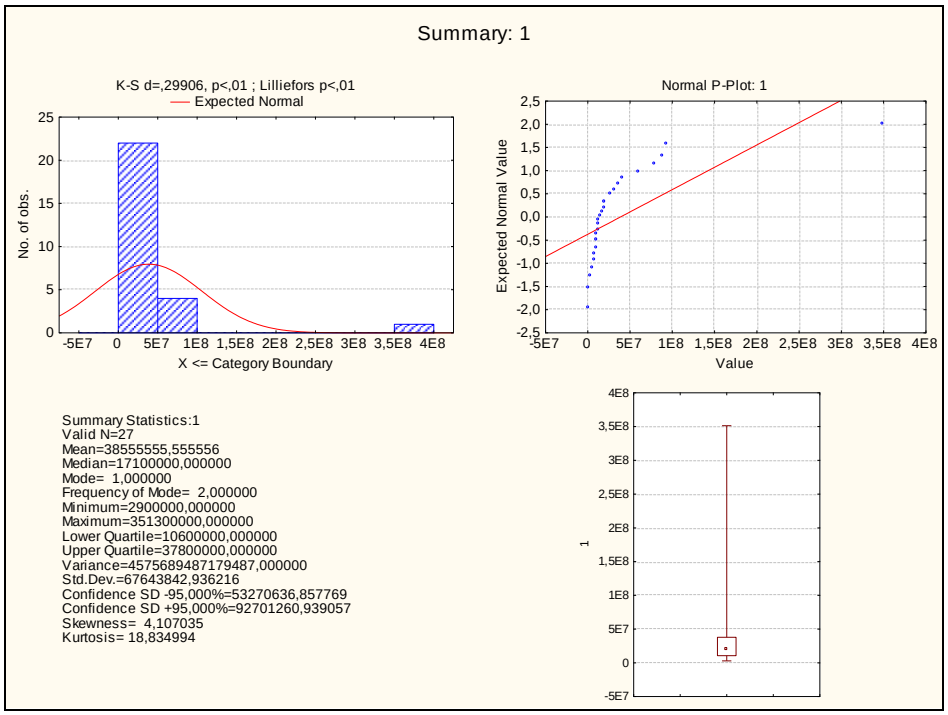

Рис. 7. Результуючий графік одновимірного статистичного аналізу

13. Студенти мають за отриманими параметрами описової статистики проаналізувати 5 показників зі свої баз даних та на зробити висновки

стосовно ступеню територіальної диференціації досліджуваних явищ в межах країни/регіону.

Практичне заняття 13-16
ПОБУДОВА ДВОВИМІРНИХ СТАТИСТИЧНИХ МОДЕЛЕЙ НА МАТРИЦІ ВИХІДНИХ ДАНИХ

Рекомендації до підготовки та проведення заняття

Завданням цієї практичної роботи є визначення пар показників, пов'язаних між собою, визначення кількісної оцінки сили зв'язку між ознаками. Двовимірний аналіз містить дві випадкові величини, утворює двовимірну модель і аналізує дві величини. В ході практичної роботи вивчаються такі двовимірні моделі як: кореляційний і коваріаційний аналізи, виконуваної в програмі Statistica 8.0 шляхом «Statistics → Basic Statistics/Tables → Correation Matrix».

Кореляційний аналіз метод обробки статистичних даних, що полягає у вивченні тісноти зв'язку між змінними, при цьому порівнюються коефіцієнти кореляції між однією парою або великою кількістю пар ознак для встановлення між ними статистичної взаємодії, дозволяє виявити найбільш зв'язані змінні і побудувати поверхні їх залежності.

Мета кореляційного аналізу – забезпечити отримання деякої інформації про однієї змінної за допомогою іншої змінної. У випадках, коли можливе досягнення мети, кажуть, що змінні корелюють. У найзагальнішому вигляді сприйняття гіпотези про наявність кореляції означає, що зміна значення змінної А станеться одночасно з пропорційною зміною значення В. Мірою залежності між експериментальними наборами даних є числа – коефіцієнти зв'язку.

Головні завдання кореляційного аналізу:

1) оцінка за вибірковими даними коефіцієнтів кореляції;

2) перевірка значущості вибіркових коефіцієнтів кореляції або кореляційного відношення;

3) оцінка близькості виявленої зв'язку до лінійної;

4) побудова довірчого інтервалу для коефіцієнтів кореляції.

Визначення сили й напрямку взаємозв'язку між змінними є однією з важливих проблем аналізу даних. У загальному випадку для цього застосовують поняття кореляції. Кореляція є залежністю двох випадкових величин. При цьому, зміна однієї або декількох цих величин приводить до систематичної зміни іншої або інших величин.

Математичної мірою кореляції двох випадкових величин служить коефіцієнт кореляції.

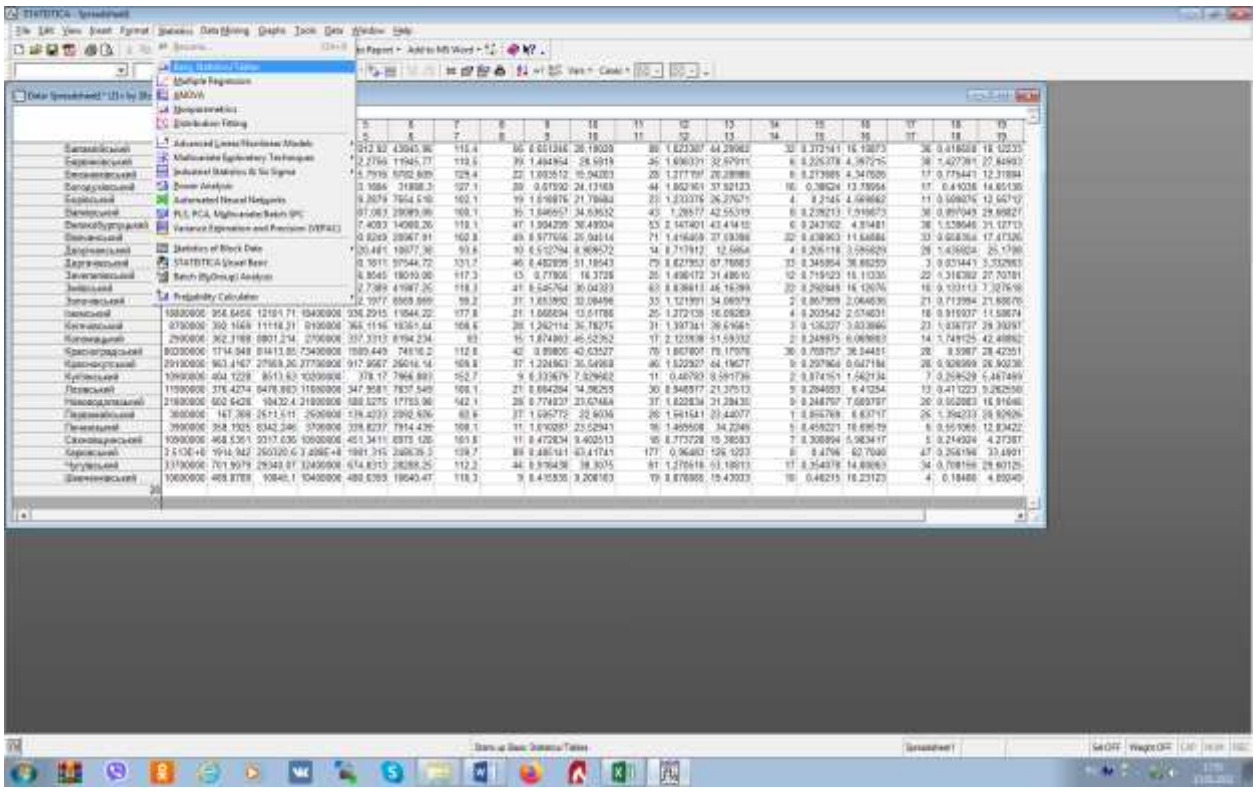

Рис. 1. Алгоритм вибору кореляційного аналізу в програмі Statistica 8.0.

Коефіцієнт кореляції приймає значення в межах від -1 до +1. Знак «-» означає зворотну залежність, «+» – пряму. Якщо коефіцієнт дорівнює нулю, то лінійний зв'язок між динамічними рядами є відсутнім, а якщо одиниці – є функціональна залежність. Досить часто помилково вважається, що значення коефіцієнта кореляції менше 0.2 свідчить про відсутність зв'язку. Насправді він свідчить про наявність дуже слабкого зв'язку, але зв'язок між показниками є.

Коеффіціент корреляції	**Тіснота зв'язку**
> ± 0.91	Дуже сильна
± 0.71-0.90	Сильна
± 0.51-0.70	Відносна
± 0.21-0.50	Слабка
<± 0.20	Дуже слабка

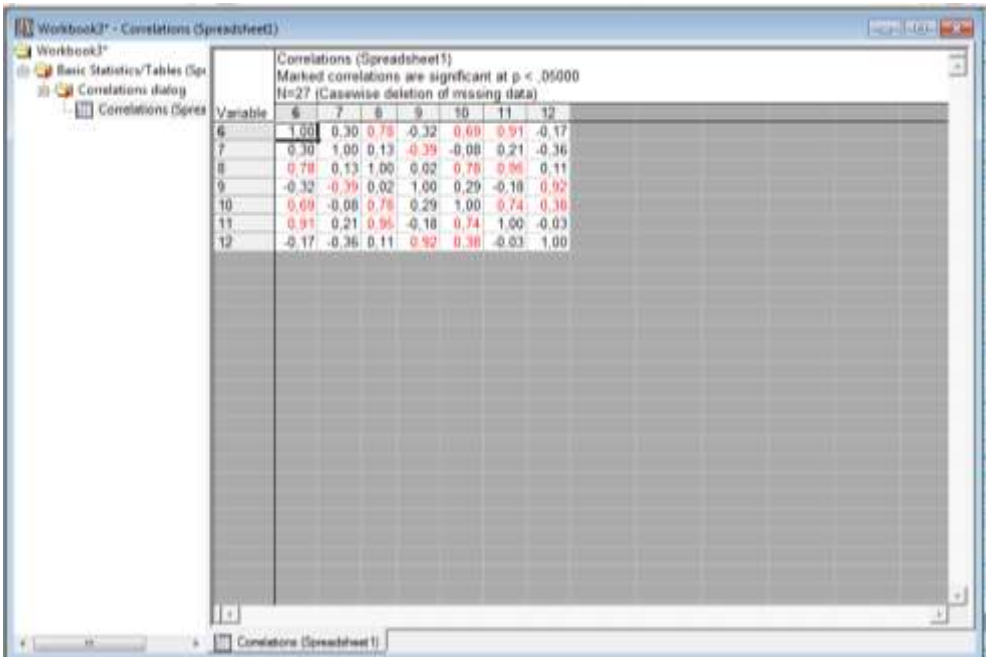

Рис. 2. Кореляційна матриця

Коефіцієнт кореляції, а в загальному випадку кореляційна функція, дозволяє встановити ступінь взаємозв'язку між змінними. Кореляція може бути лінійної або нелінійної залежно від типу залежності, яка фактично існує між змінними. Досить часто на практиці розглядають тільки лінійну кореляцію (взаємозв'язок), але більш глибокий аналіз вимагає використання для дослідження процесів нелінійних залежностей. Складну нелінійну залежність можна спростити, але знати про її існування необхідно для того, щоб побудувати адекватну модель процесу.

У разі максимальної тісноти зв'язку між показниками на діаграмі розсіювання їх залежність буде представлена прямою лінією. Іноді зустрічається таке явище, як псевдокорреляція, тобто кореляція, обумовлена впливом інших показників, які залишилися поза увагою дослідника. Значимість коефіцієнта кореляції залежить від довжини динамічних рядів. У великих динамічних рядах навіть слабкі залежності будуть значущими, в той час, як в незначних навіть дуже сильні залежності не є статистично надійними. Тому говорять про надійність кореляційної залежності, пов'язаної з репрезентативністю вихідних динамічних рядів і свідчить про те, наскільки ймовірно, що виявлена залежність, знову буде виявлена при збільшенні періоду ретроспекції або екстраполяції на майбутнє. Надійність виявлених залежностей оцінюється за допомогою стандартної статистичної міри – p-рівня – статистичного рівня значимості. Даний показник знаходиться в зворотній залежності до надійності результату: чим вище p-рівень, тим нижча статистична надійність виявленої залежності, і навпаки.

План виконання практичної роботи:

1. Відкриваємо програму Statistica 8.0

2. Натискаємо кнопку New ( ) або Ctrl+N. Натискаємо ОК.
3. З’являється таблицю, в яку необхідно додати дані з бази даних.
4. Для цього копіємо в таблиці Excel базу даних і вставляємо її у таблицю програми Statistica.
5. Дана операція роться наступним чином. У першій ячійці правою кнопкою миші натискуємо у вікні, що відкривається, натискуємо «Paste With Headers → Paste With Both». Таким чином, ми копіюємо у таблицю програми Statistica наші показники, а також назви показників та назви районів/регіонів (рис. 3).

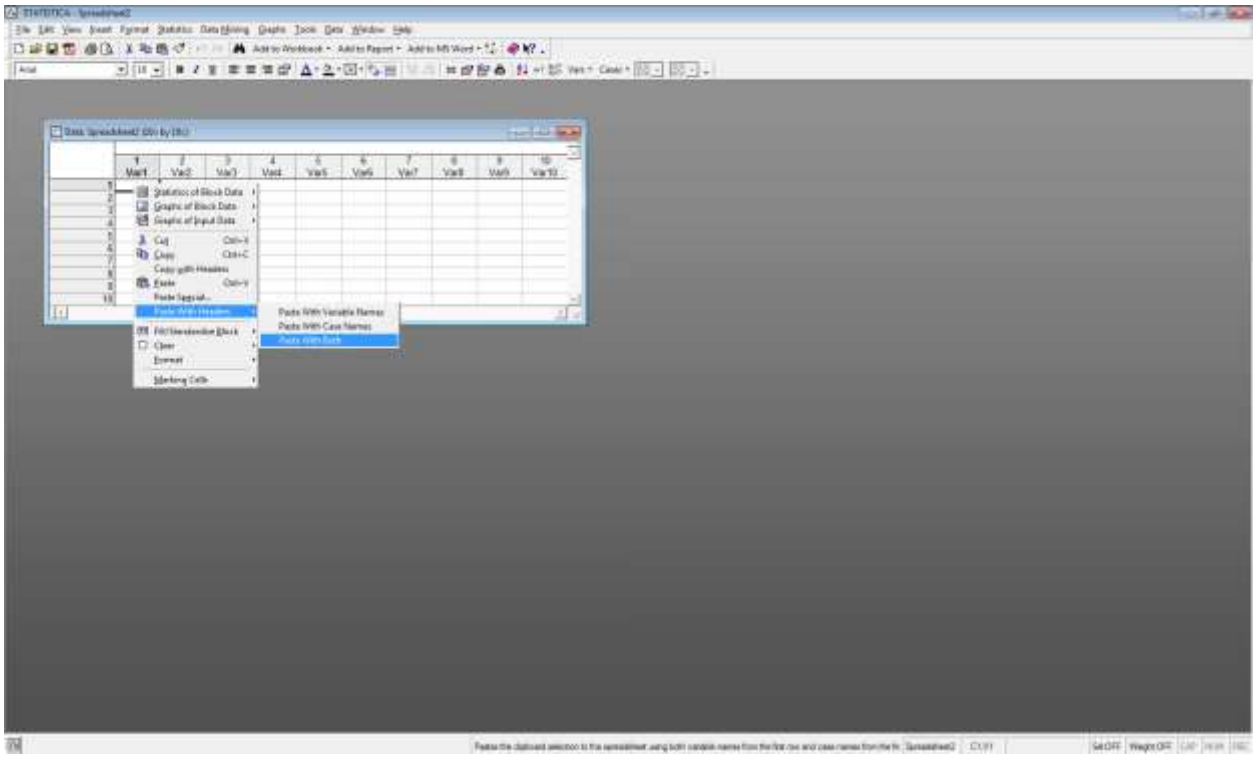

Рис. 3. Додавання бази даних у таблицю програми Statistica

6. Після цього на панелі задач натискаємо «Statistics → Basic Statistics/Tables → Correation Matrix → OK».

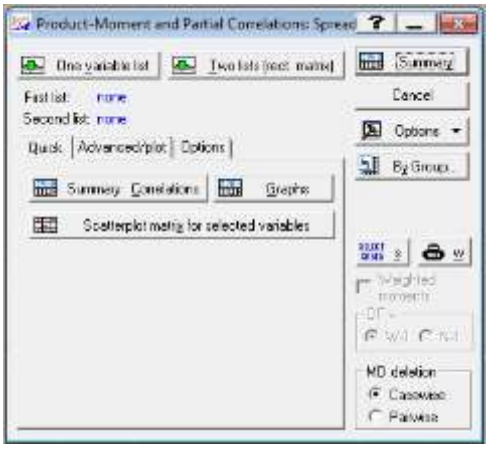

Рис. 4. Діалогове вікно виконання кореляційного аналізу

7. Натискаємо кнопку «Two lists (rect. matrix)». У вікні, що відкрилося, обираємо необхідні показники. Натискаємо «ОК» (рис. 5).

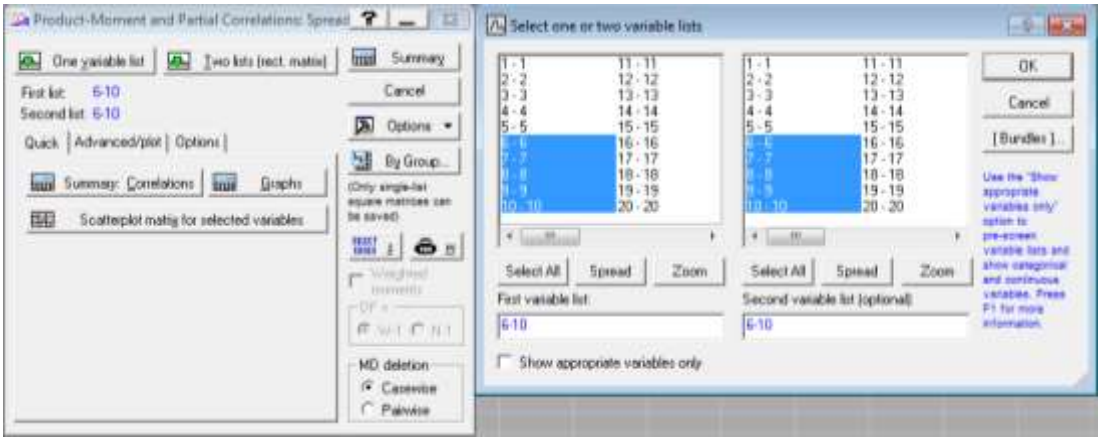

Рис. 5. Вибір показників для аналізу

8. Обираємо вкладку «Option» та перевіряємо, щоб у рядку «p-level for highlighting» було «,05». Цу і є p-рівень – статистичний рівень значимості (рис. 6).

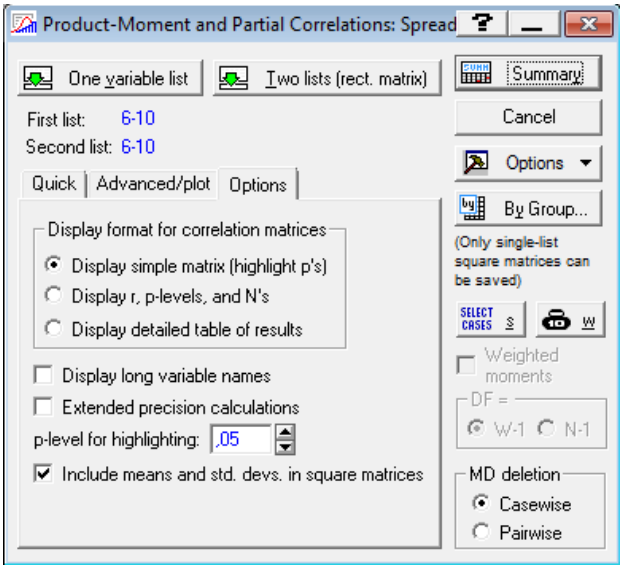

Рис. 6. Перевірка статистичного рівня значимості (p-рівень)

Після цього переходимо на вкладку «Quick» та настискаємо на кнопку «Summary: Correlations». В результаті отримуємо кореляційну матрицю. Червоним кольором підсвічуються ті коефіцієнти кореляції, для яких $p < 0{,}05$. Ці коефіцієнти кореляції значимі та можуть враховуватися при аналізі (рис. 7).

Студенти мають продемонструвати отиману кореляційну матрицю, визначити ступінь сили зв'язку між окремими показниками та зробити висновок стосовно того чи іншого значення коефіцієнта кореляції між показниками. Бажано для аналізу обирати показники, які взаємозалежні між

собою (наприклад, кількість обєктів торгівлі у районі та чисельність населення тощо). Також потрібно взяти показники, між якими гіпотетично не може бути взаємозявзу. Коефіцієнти кореляції це підтвердять. Необхідно проаналізувати прийнамні 5 пар показників.

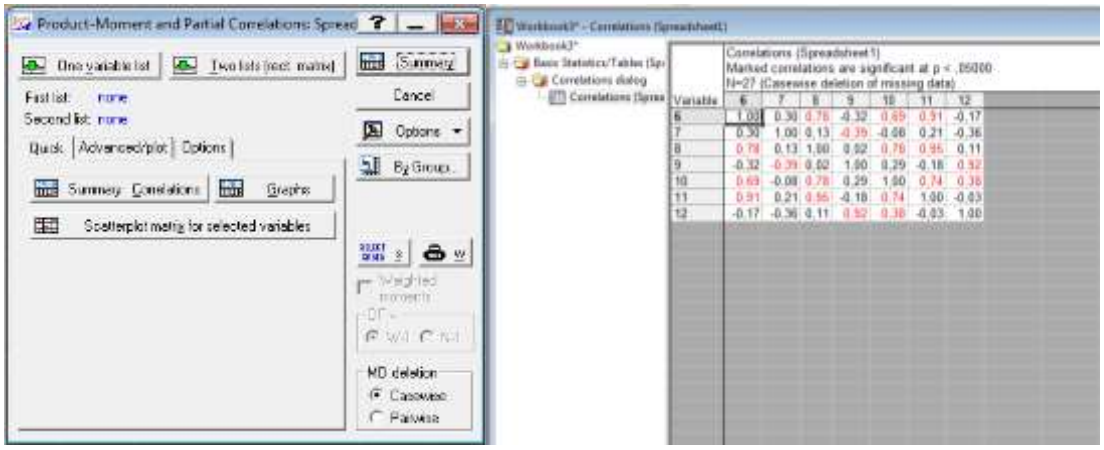

Рис. 7. Кореляційна матриця

Практичне заняття 17-19
КЛАСТЕР-АНАЛІЗ ВИХІДНИХ ДАНИХ. ІНТЕРПРЕТАЦІЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Рекомендації до підготовки та проведення заняття

Завданням цієї практичної роботи буде визначення територіальної диференціації адміністративно-територіальних одиниць і виявлення значущих чинників суспільно-географічного процесу.

Багатовимірним аналізом називають сукупність різних методів, призначених для вивчення багатовимірних явищ. У багатовимірному просторі досліджувані об'єкти розташовуються, як правило, не рівномірно, а утворюють певні скупчення. Ці скупчення можна розглядати як класи об'єктів. Причому в різних просторах виділяються різні класи. У багатовимірному просторі класифікація більш обгрунтована. Але це означає, що при безмежному зростанні числа ознак в тій же пропорції зростає точність класифікації.

Існує багато математичних методів і прийомів, які на основі інформації закладеної в матриці даних, дозволяють об'єктивно класифікувати економіко-географічні об'єкти (в тому числі регіоналізувати їх). Такі методи і прийоми об'єднані в багатомірний аналіз, розглядається у вузькому і широкому сенсах. У вузькому сенсі – це аналіз матриці даних, яка має два і більше стовпців, в широкому сенсі – це аналіз матриці даних, коли кількість стовпців не обмежена знизу. В цьому випадку багатовимірний аналіз включає в себе і одновимірний. Отже, одновимірний аналіз можна розглядати як окремий випадок багатовимірного, а багатовимірний – як узагальнення одновимірного. Часто в процесі обробки матриці даних доводиться їх згортати, тобто зводити багатовимірний простір до n-k-мірного, і навіть до одновимірного. У цьому випадку багато ознак замінюються декількома, але такими, кожна з яких синтезує інформацію відповідної групи ознак або є найбільш репрезентативною.

Кластерний аналіз передбачає групування об'єктів за подібністю певних характеристик. Термін вперше введений Трайеном / Tryon /, 1939 р. Це один з методів класифікації, що передбачає поділ вихідної сукупності об'єктів на кластери (класи, групи). Кластер – це група територіальних одиниць (регіонів), що мають схожі тенденції або особливості розвитку. З математико-статистичної точки зору, кожен кластер повинен мати такі властивості: густота об'єктів в межах кластера повинна бути більша за густоту поза ним, можливість відокремлення від інших кластерів і т.д. Складність завдань кластерного аналізу полягає в тому, що реальні суспільно-географічні об'єкти є багатовимірними, тобто описуються не одним, а певною сукупністю параметрів, тому об'єднання в групи здійснюється в просторі багатьох вимірів. Згідно з критерієм об'єднання регіонів в кластери є мінімум відстані в просторі показників, які їх описують. Звідси – поняття відстані між регіонами в просторі даних.

В ході практичної роботи виконується кластерний аналіз в програмі Statistica 8.0 «Statistics» → «Multivariate Exploratory Techniques» → «Cluster Analysis» (рис. 1).

Існують наступні групи методів кластерного аналізу:

1) ієрархічні методи;

2) ітеративні методи;

3) факторні методи;

4) методи згущень;

5) методи, які використовують теорію графів.

До поширених в економіці відносять ієрархічні і ітеративні.

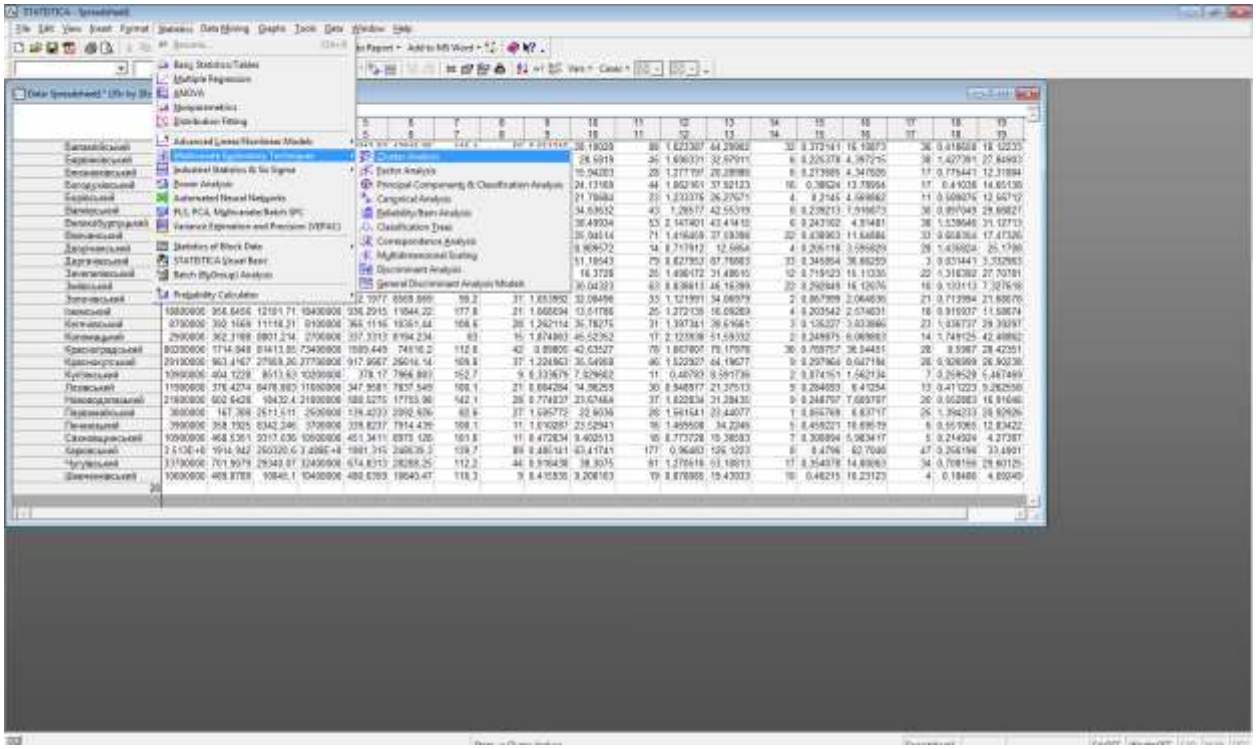

Рис. 1. Алгоритм вибору кластерного аналізу в програмі Statistica 8.0

Для проведення класифікації необхідно ввести поняття подібності об'єктів по спостережуваних змінних. У кожен кластер (клас, таксон) повинні потрапити об'єкти, що мають подібні характеристики. Вибір відстані між об'єктами є вузловим моментом дослідження, від нього багато в чому залежить остаточний варіант розбиття об'єктів на класи при даному алгоритмі розбиття.

Алгоритм ієрархічного агломеративного кластерного аналізу можна представити у вигляді послідовності процедур:

1) нормуються вихідні дані;

2) розраховується матриця відстаней або матриця заходів подібності;

3) знаходиться пара найближчих кластерів, за обраним алгоритмом об'єднуються ці два кластери. Новому кластеру присвоюється менший з номерів поєднуваних кластерів;

4) процедури 2, 3 і 4 повторюються до тих пір, поки всі об'єкти не

будуть об'єднані в один кластер або до досягнення заданого "порога" подібності.

На розрахунок відстаней між об'єктами істотний вплив має вибір одиниць вимірювання показників. Так, наприклад, якщо брати показник обсягів виробництва продукції в мільйонах гривень, різниця між двома територіальними одиницями за цією координатою буде однією, якщо ж в тисячах гривень – в тисячу разів більше, що впливає на кінцеве угруповання. Тому кластерний аналіз передбачає здійснення процедури нормалізації даних.

У кластерному аналізі для кількісної оцінки подібності вводиться поняття метрики. Подібність або відмінність між класифікованими об'єктами встановлюється в залежності від метричної відстані між ними. Якщо кожен об'єкт описується k ознаками, то він може бути представлений як точка в k-вимірному просторі, і схожість з іншими об'єктами буде визначатися як відповідна відстань. У кластерному аналізі використовуються різні міри відстані між об'єктами.

Найбільш поширеними є такі види відстаней:

- Евклідовова відстань (Euclidean distances), розраховується по теоремі Піфагора: відстані по кожній з координат вводяться в квадрат, а потім з їх суми визначається корінь квадратний;
- Манхеттенська відстань (відстань міських кварталів, city- block / Manhattan / distances), розраховується як сума різниць по кожній з координат;
- відстань Чебишева (Chebychev distance metric), регіони визначають як «різні», якщо вони розрізняються за якоюсь однією координатою;
- відсоток незгоди (percent disagreement), використовується, коли вихідні дані не мають кількісного вираження.

Методи кластеризації діляться на дві великі групи – агломеративні (від англ. Agglomerate – скупчення) і дівізивні (від англ. Division – поділ).

Агломеративні методи передбачають послідовне об'єднання найвизначніших регіонів на основі розрахованих відстаней між ними. Процедура кластеризації така. На першому кроці кожен регіон утворює окремий кластер, далі в новий кластер об'єднуються два регіони, ступінь подібності яких є найбільшою. На останньому кроці усі регіони об'єднуються в один кластер.

Міра подібності кластерів і регіонів визначається наступними способами:

- одиничного зв'язку (single linkage, «метод найближчого сусіда»): мінімум найменших відстаней до будь-якого одного регіону в кластері;
- повного зв'язку (complete linkage, «метод найвіддаленіших сусідів»): мінімум найбільших відстаней до будь-якого одного регіону в кластері;
- «середнього» зв'язку: мінімум середньоарифметичного значення відстаней до всіх регіонів в кластері;
- центроїдний – мінімум відстаней до центрів тяжкості кластерів (рис. 2).

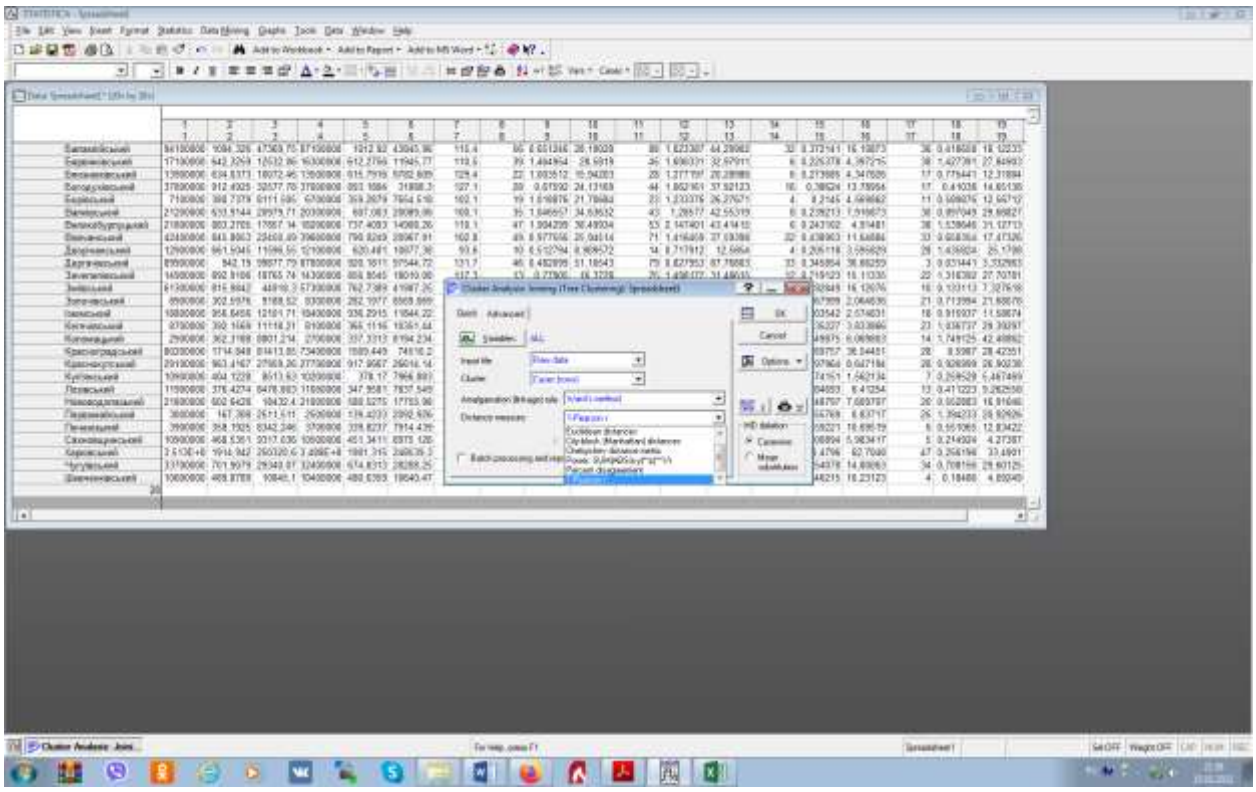

Рис. 2. Вибір відстані при виконанні кластерного аналізу

Центр тяжіння кластера визначається як середнє по кожному параметру.

Результат кластеризації візуалізується у вигляді дендрограмми кластеризації (tree diagram, дерево об'єднання), на одній осі якої відкладаються регіони, на другий – відстані об'єднання (linkage distance). Дивізійні методи передбачають поетапне (ітераційне) поділ регіонів на задану кількість кластерів. Поширеним серед дивізійних є метод k-середніх (kmeans clustering), який вирішує завдання виділення заздалегідь заданої кількості кластерів, які максимально відрізняються, тобто знаходяться на найбільших відстанях один від одного (рис. 3).

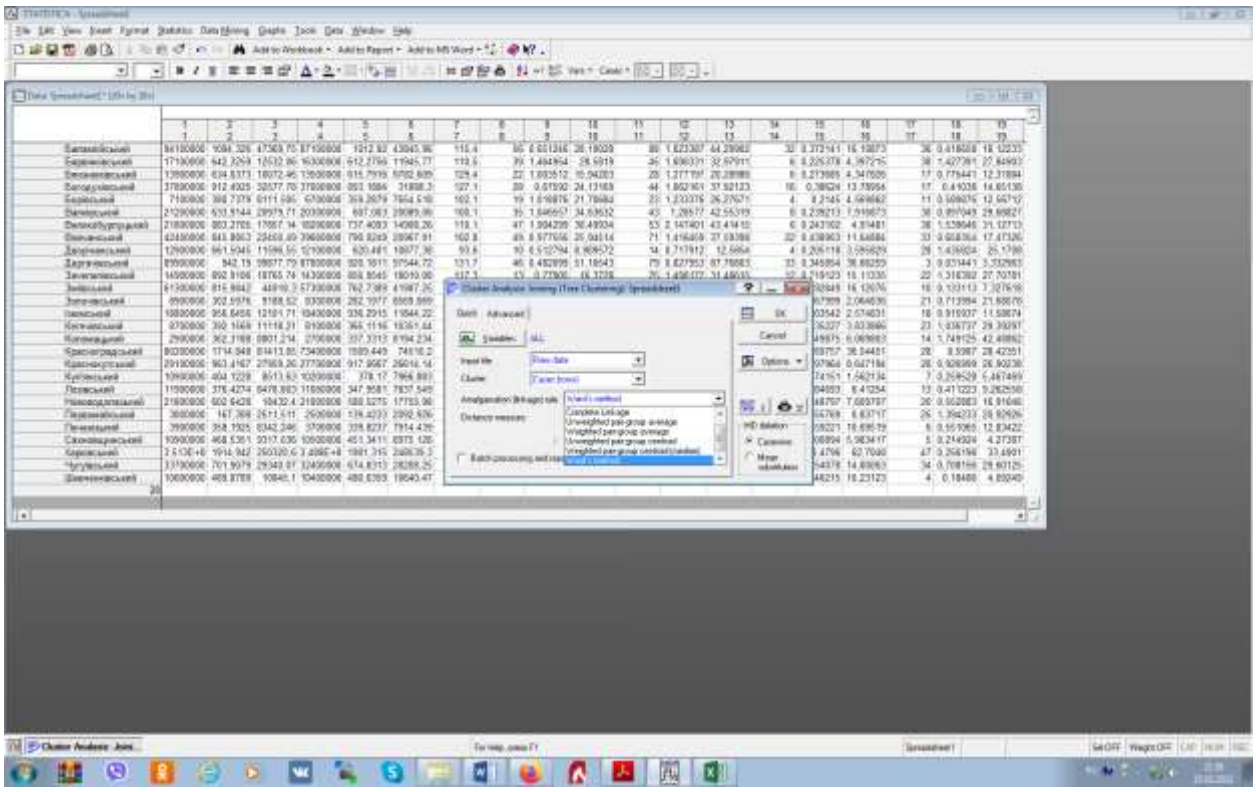

Рис. 3. Вибір методу при виконанні кластерного аналізу

Процедура кластеризації така. На першому кроці задається деякий випадкове поділ даних на задану кількість кластерів (k), розраховуються центри тяжіння кластерів. Далі здійснюється переміщення регіонів: кожен регіон відноситься до того кластеру, відстань до центра ваги якого мінімальне, розраховуються центри тяжіння нових кластерів. Ця процедура повторюється, поки не буде знайдена стабільна конфігурація, тобто склад кластерів перестане змінюватися.

Для візуалізації результатів будуються графіки, що представляють собою проекції будь-якої пари показників на площину, на яких точки, що відносяться до одного кластеру, окантовуються. Одним з найбільш важливих і складних питань при кластеризації є вибір оптимальної кількості кластерів. Зазвичай згідно вихідної гіпотези визначається початкова кількість кластерів, а потім змінюючи його, емпіричним шляхом вибирають остаточний варіант кластеризації. Для того, щоб оцінити, наскільки виділені кластери відрізняються, розраховують середні значення базисних показників для кожного кластера.

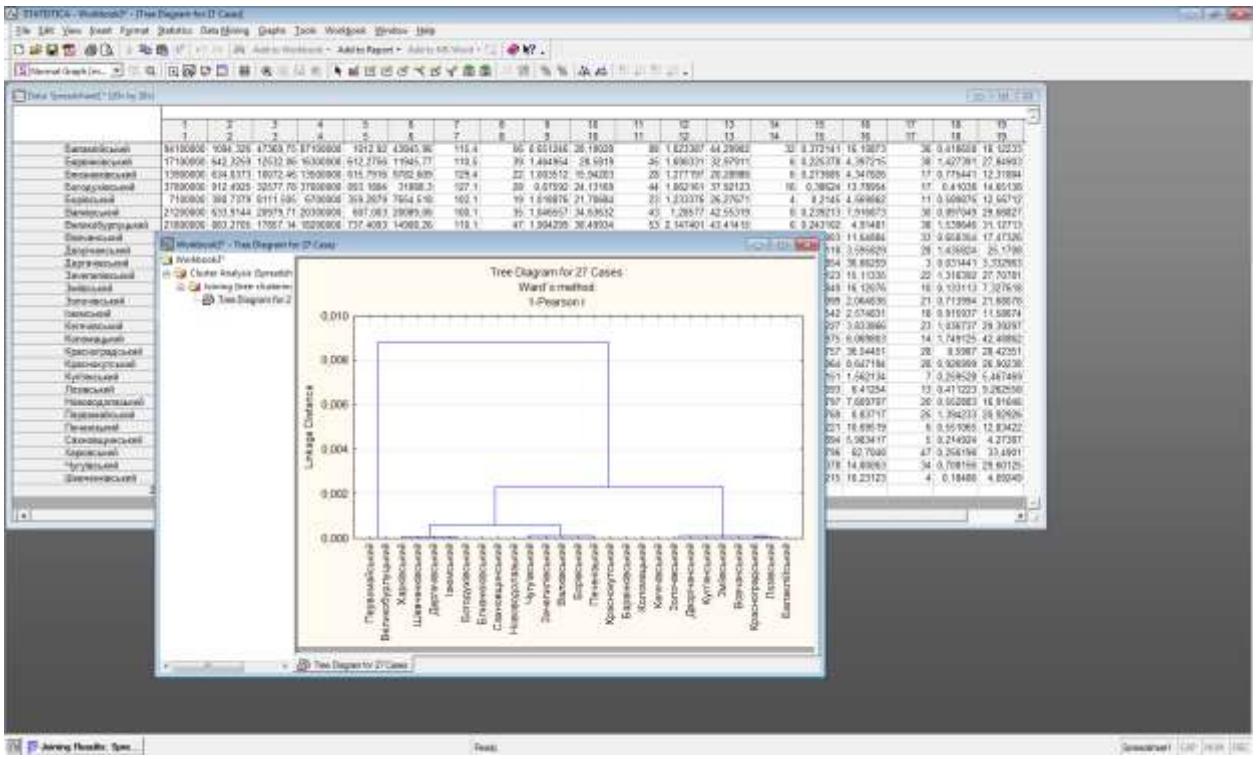

Рис. 4. Приклад результату кластерного аналізу

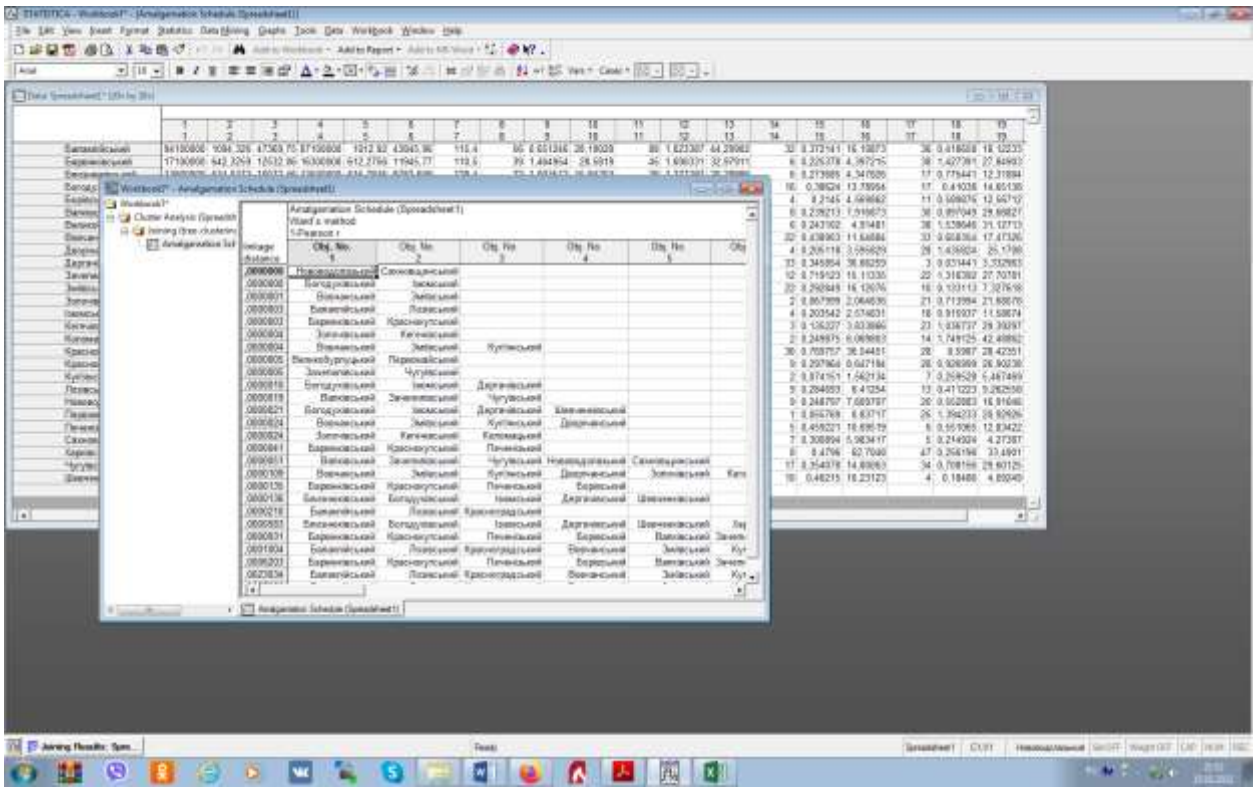

Рис. 5. Приклад протоколу кластерного аналізу

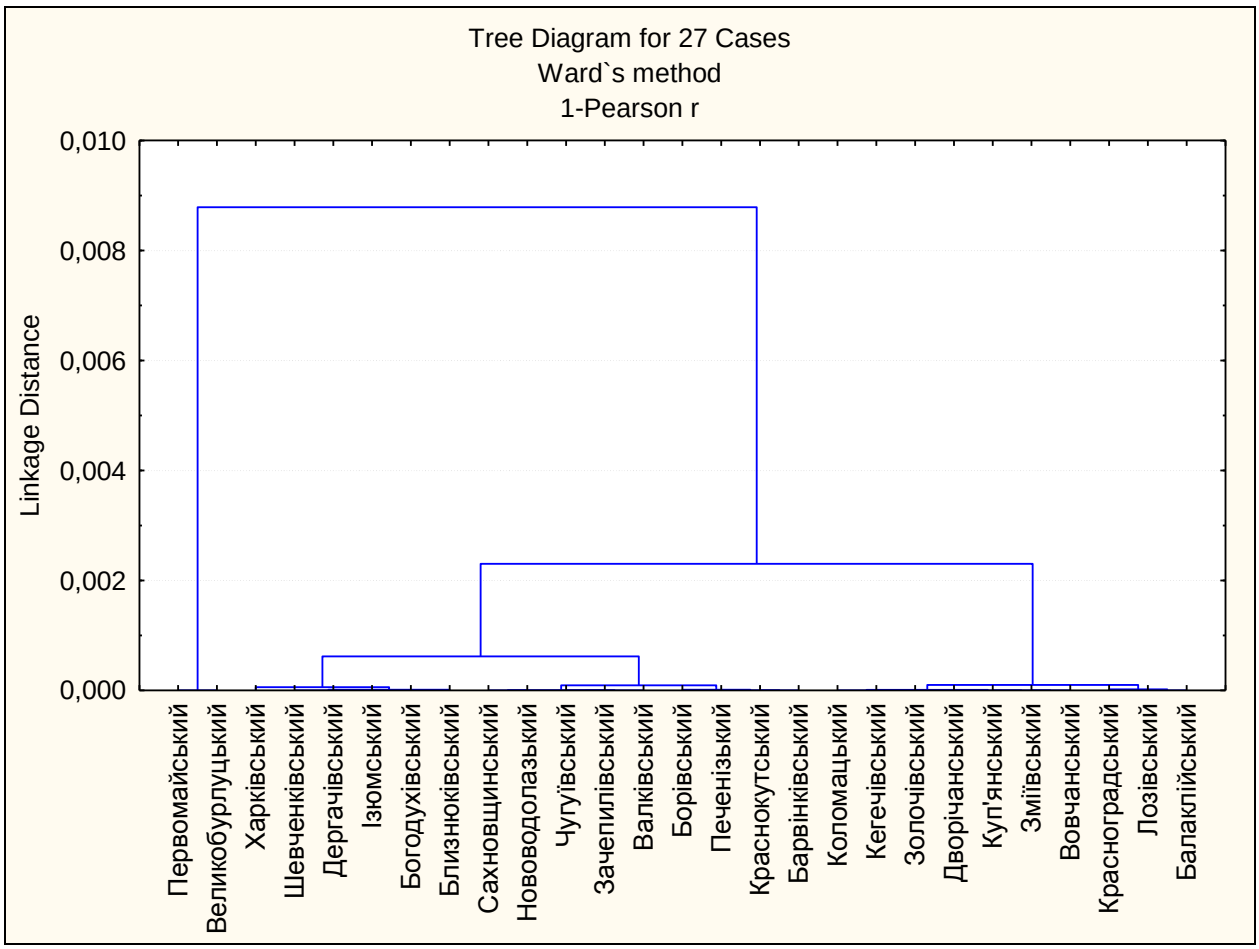

Рис. 6. Приклад дендрограмми кластерного аналізу

Кластер-аналіз застосовується при наявності багатовимірних сукупностей статистичних показників, сутність його полягає в об'єднанні їх в групи (кластери) за принципом мінімального розміру в багатовимірному просторі. Залежно від кількості об'єктів кластеризація і угруповання виконуються послідовно в кілька кроків таким чином, щоб на останньому кроці в одну загальну групу потрапили всі об'єкти. На перших кроках класифікації формуються найбільш однорідні групи по об'єктах, які мають найбільшу схожість. Поступово, «послаблюючи» критерій щодо подібності об'єктів, об'єднується все більшу кількість об'єктів. З кожним кроком до кластерів вищого порядку включаються цілі групи адміністративних районів, які все сильніше розрізняються між собою. На останньому кроці всі об'єкти об'єднуються в один кластер. В кінці процедури отримують групи неоднорідні. В результаті кластерного аналізу отримують багаторівневу ієрархічну класифікацію, яка відображає найбільш суттєві особливості взаємини між об'єктами. Таким чином, отримані кластери – це група територіальних одиниць, які мають подібні особливості розвитку. Такий аналіз проводиться для територіальних об'єктів по ряду показників, для подальшого угруповання районів і виявлення стійких груп.

План виконання практичної роботи:

1. Відкриваємо програму Statistica 8.0
2. Натискаємо кнопку New ( ) або Ctrl+N. Натискаємо ОК.
3. З'являється таблицю, в яку необхідно додати дані з бази даних.
4. Для цього копіємо в таблиці Excel базу даних і вставляємо її у таблицю програми Statistica.
5. Дана операція роться наступним чином. У першій ячійці правою кнопкою миші натискуємо у вікні, що відкривається, натискуємо «Paste With Headers → Paste With Both». Таким чином, ми копіюємо у таблицю програми Statistica наші показники, а також назви показників та назви районів/регіонів (рис. 7).

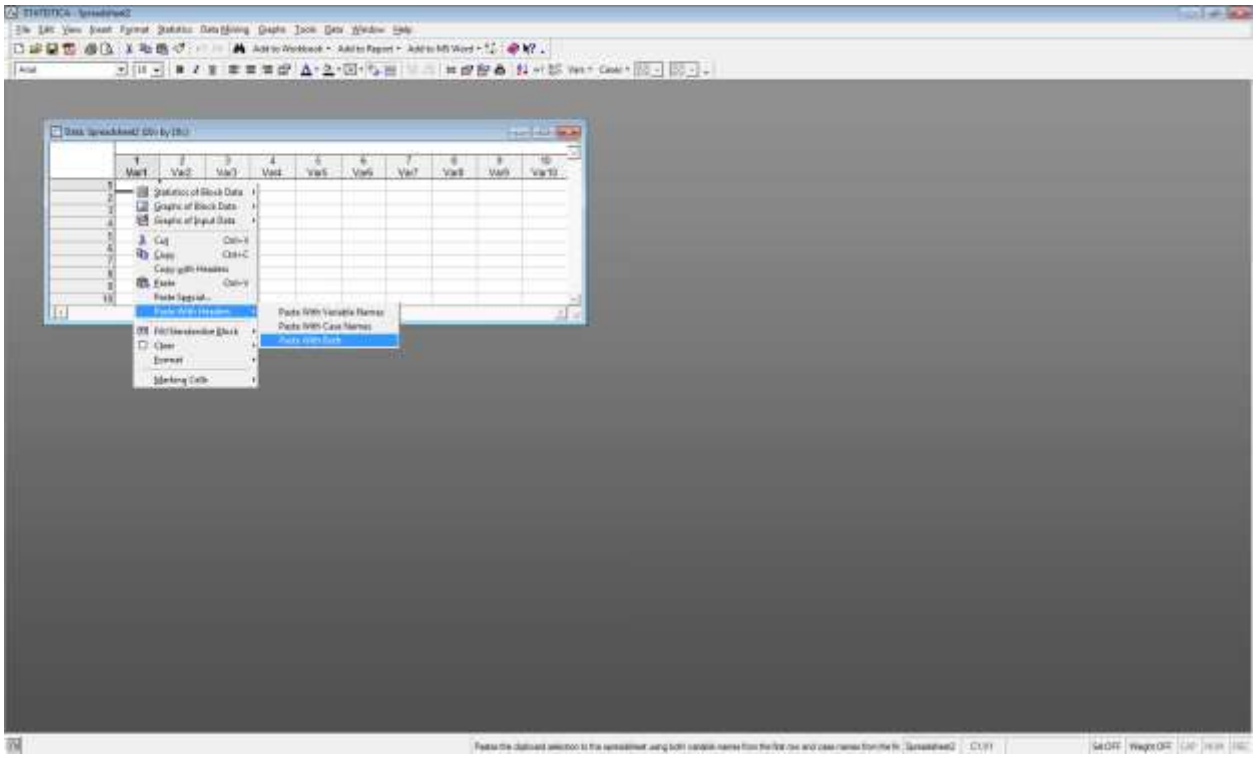

Рис. 7. Додавання бази даних у таблицю програми Statistica

6. Після цього на панелі задач натискаємо «Statistics» → «Multivariate Exploratory Techniques» → «Cluster Analysis» (рис. 8).
7. Після цього зявляється діалогове вікно «Clustering method» / «Метод кластерізації)». Ми обираємо метод «Joining (tree clustering)» та натискаємо «ОК» (рис. 9).
8. Відкривається діалогове вікно «Cluster Analysis: Joining (tree clustering)». Натискаємо вкладку «Advanced», натискаємо кнопку «Variables». Обираємо показники для аналізу (рис. 10). Натискаємо «ОК».

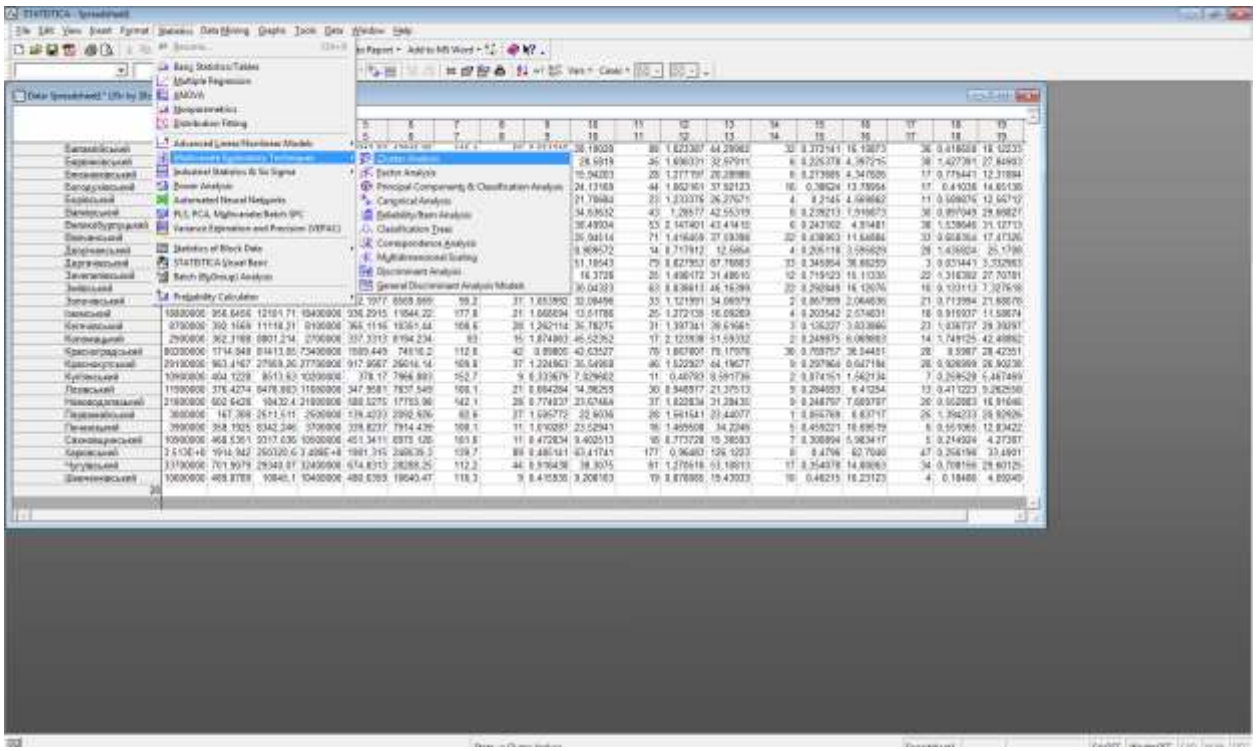

Рис. 8. Алгоритм вибору кластерного аналізу в програмі Statistica 8.0

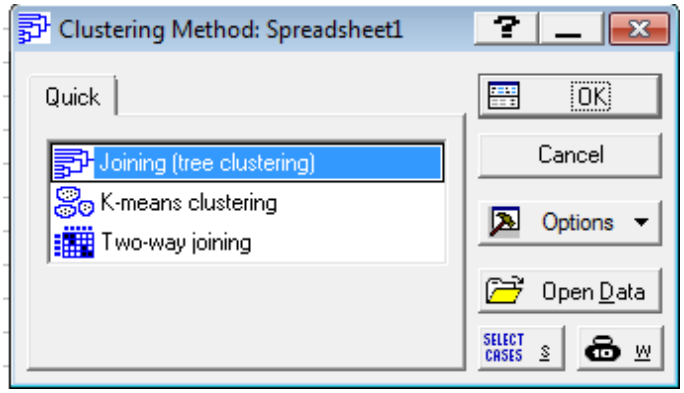

Рис. 9. Вибір методу кластерізації

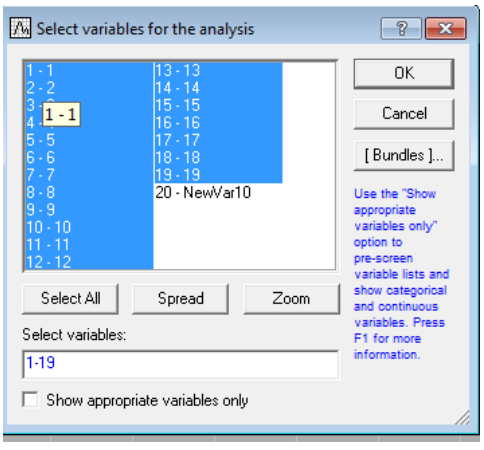

Рис. 10. Вибір показників для аналізу

9. Наступним кроком є вибір методів виділення кластерів та міри відстані.

У полі «Input file» має бути «Raw data», «Cluster» - «Cases (rows)» (кластерний аналіз буде виконуватися за рядками, тобто групуватися будуть адміністративно-територіальні одиниці), «Agglomeration (linkage) rule» - «Ward's method» (метод Варда), «Distance measure» - «1 Pearson r» (відстань Пірсона). Після цього натискаємо «ОК» (рис. 11).

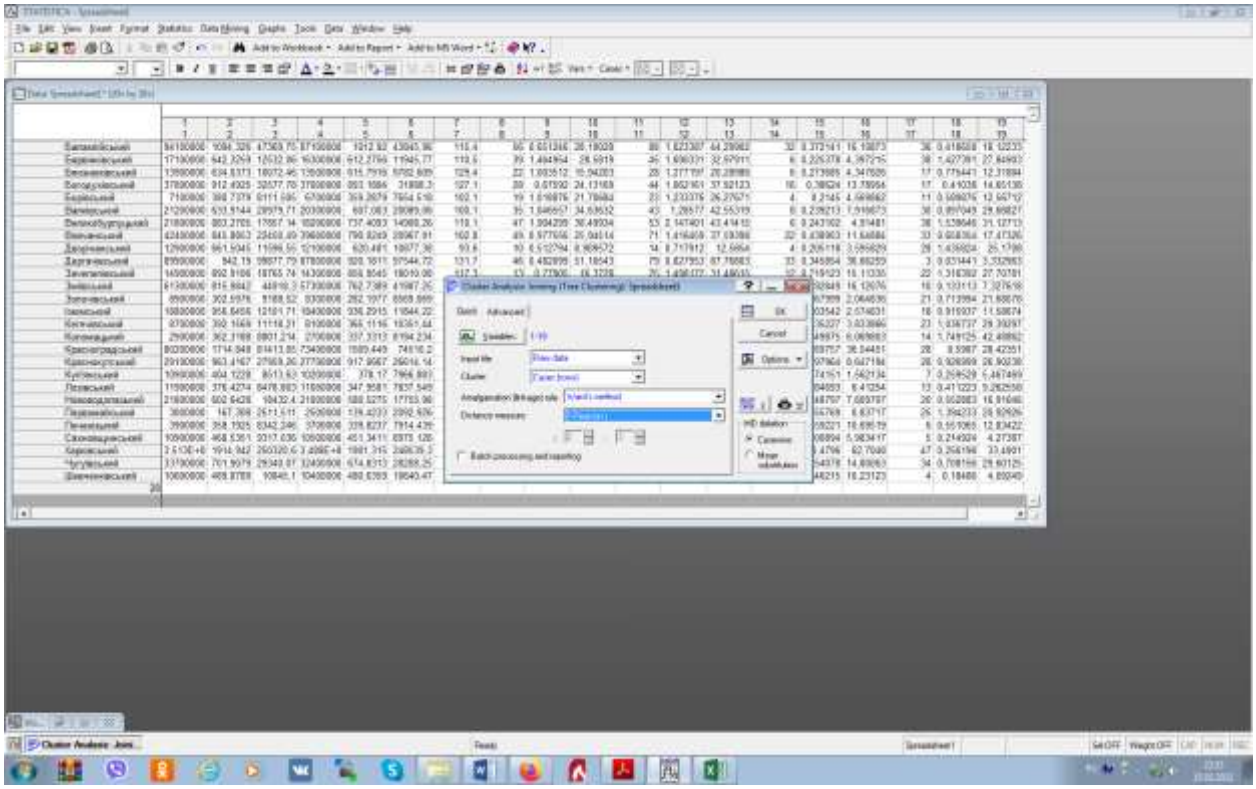

Рис. 11. Вибір методу виділення кластерів та міри відстані

10. Далі відкривається діалогове вікно «Joining results». Галочка має стояти навпроти поля «Rectangular branches». Натискаємо на кнопку «Vertical icicle plot» (рис. 12).

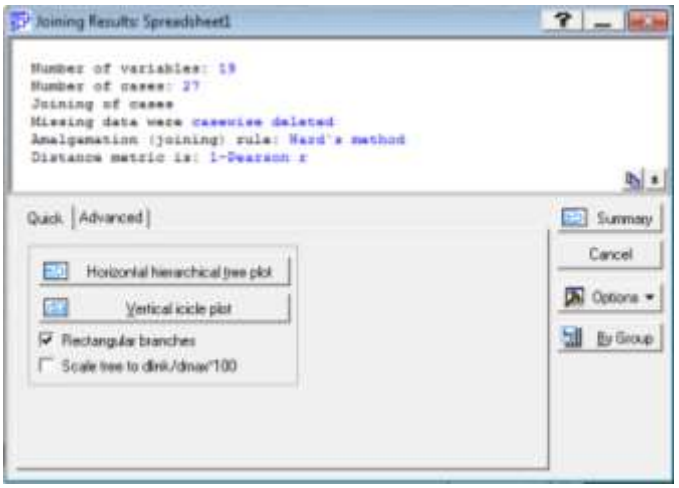

Рис. 12. Вибір напрямку дендрограми кластерізації

11. Таким чином, ми отримуємо вертикальну діаграму (рис. 13).
12. Також потрібно повернутися (у нижному лівому куті шукаємо вкладку «Joining results») та нажати кнопку «Alaglamation». В результаті отримуємо таблицю з результатами об’єднань адміністративно-територіальних одиниць за ступенем міри відстані (рис. 14). Обираємо певний рівень. Наприклад, ми обрали рівень об’єднання 0,002. Візуально на дендрограмі ми бачимо виділення 3 груп. В таблиці Amalgamation Schedule ця відстань лежить між відстанями об’єднання 0,0006167 та 0,0022899. Також ми маємо брати до уваги відстані зв’язку, що є більшими, оскільки класифікація обєктів відбувається з найбільної відстані. В даному випадку, такою відстанню є 0,0087346. Таким чином, для виділення груп районів ми обираємо три відстані зв’язку (рис. 14).

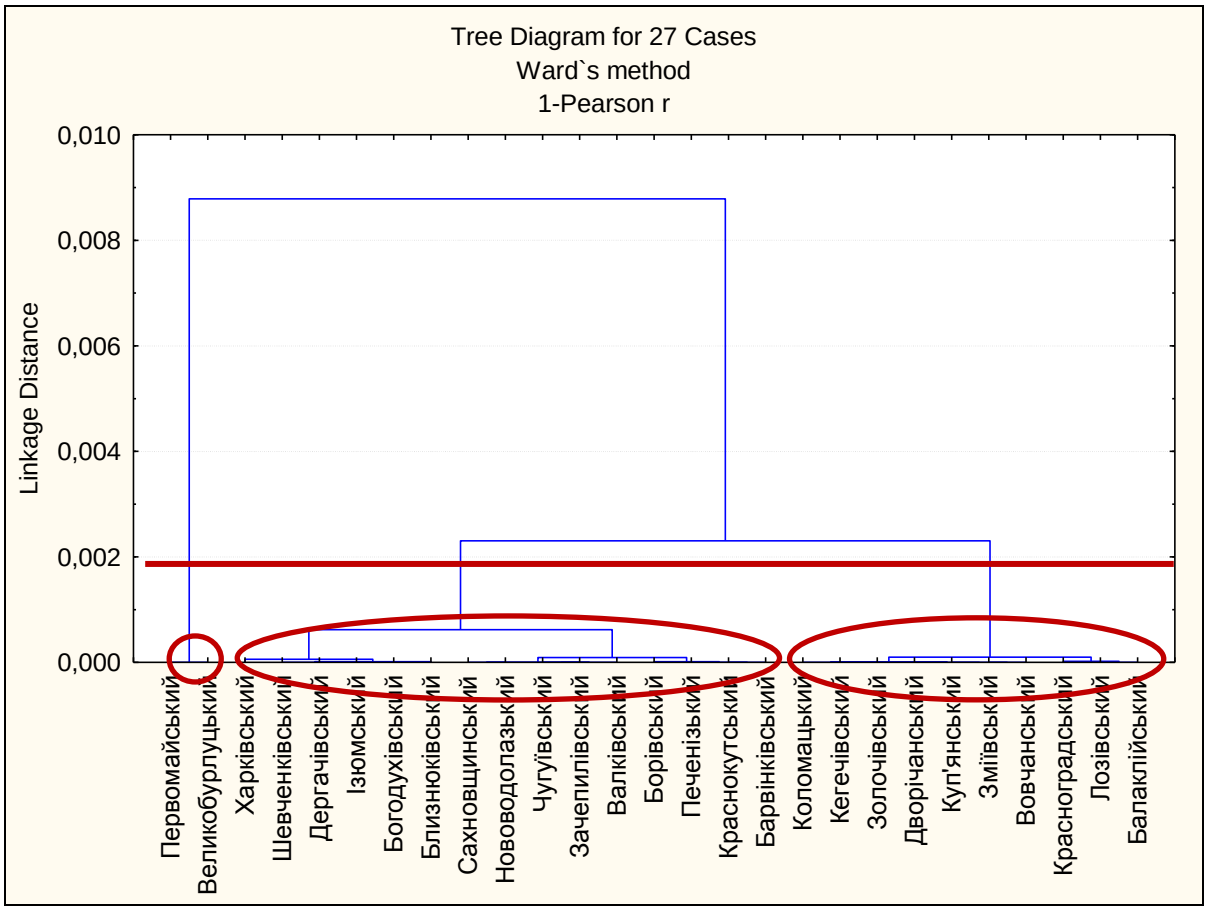

Рис. 13. Приклад дендрограмми кластерного аналізу

13. На трьох обраних відстанях зв’язку виділяємо групи районів, які обєдналися на кожному з 3-х рівнів. Враховуючи ієрахічність класифікації, на кожному рівні перелік районів може бути подібний. Нашою задачею є виокремлення районів та вилучення районів, що вже повторювалися на інших рівнях об’єднання. Найліпше таблицю Amalgamation Schedule скопіювати у Word і там провести виокремлення. В результаті ви отримаєте таку таблицю (рис. 15, нижня

таблиця). Після цього, для наочності, виділені групи на дендрограмі можна позначити, як це зроблено на рис. 13.

14. В залежності від отриманої дендрограми, можливе виділення груп та підгруп. При групуванні районів необхідно дивитися на таблицю Amalgamation Schedule та дендрограму.

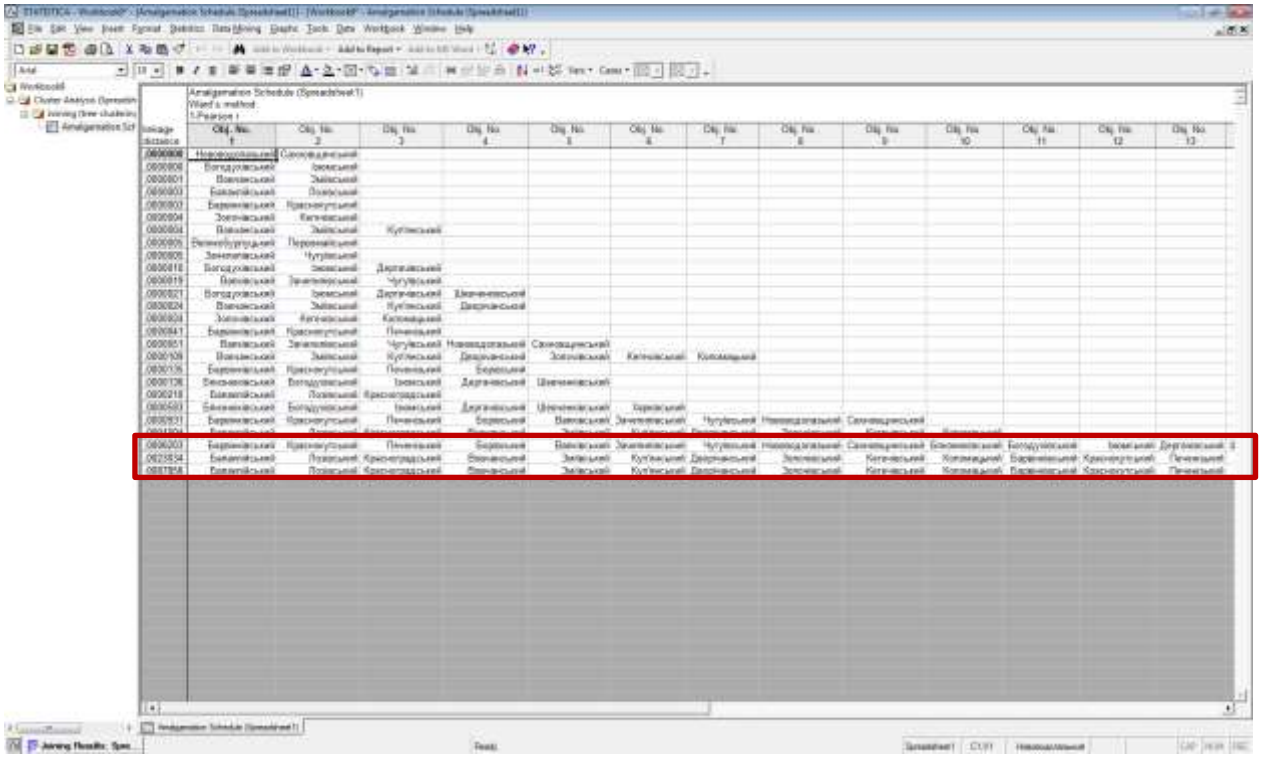

Рис. 14. Ступінь обєднання районів у кластери за мірою відстані

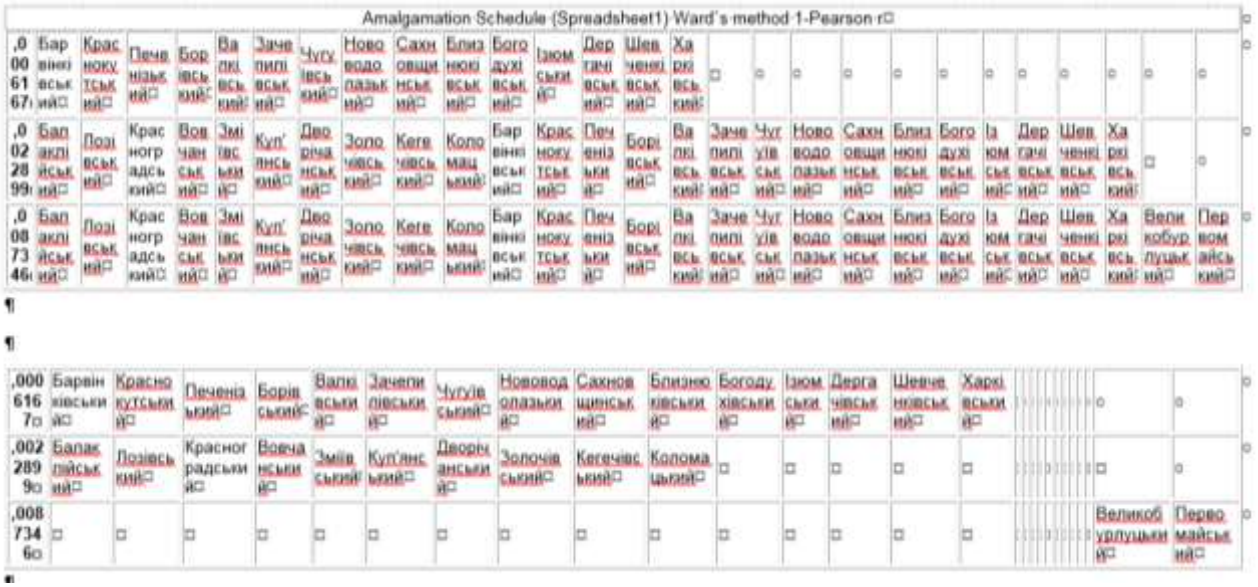

Рис. 15. Ступінь обєднання районів у кластери за мірою відстані та виділення груп районів (повторювані райони були видалені)

15. Завдання для студентів: побудувати кластрну дендрограму та обгрунтувати сформоване групування районів / регіонів, тобто пояснити, чому той чи інший район/регіон потрапив до тієї чи іншої групи. факторам. Використовуати усі показники, які ви обрали для бази даних.

Практичне заняття 20-22
ФАКТОРНИЙ АНАЛІЗ ВИХІДНИХ ДАНИХ. ІНТЕРПРЕТАЦІЯ ОТРИМАНИХ РЕЗУЛЬТАТІВ

Рекомендації до підготовки та проведення заняття

Факторний аналіз – це багатовимірний статистичний метод, який використовується для вивчення взаємозв'язків між значеннями змінних. Він дозволяє визначити групи факторів впливу і досліджувати значення кожної змінної в дії фактора в різний час.

В ході практичної роботи виконується факторний аналіз методом головних компонент, в Statistica 8.0 шляхом Statistics» → «Multivariate Exploratory Techniques» → «Factor Analysis» (рис. 1). Факторний аналіз дозволяє оцінити чинники розвитку.

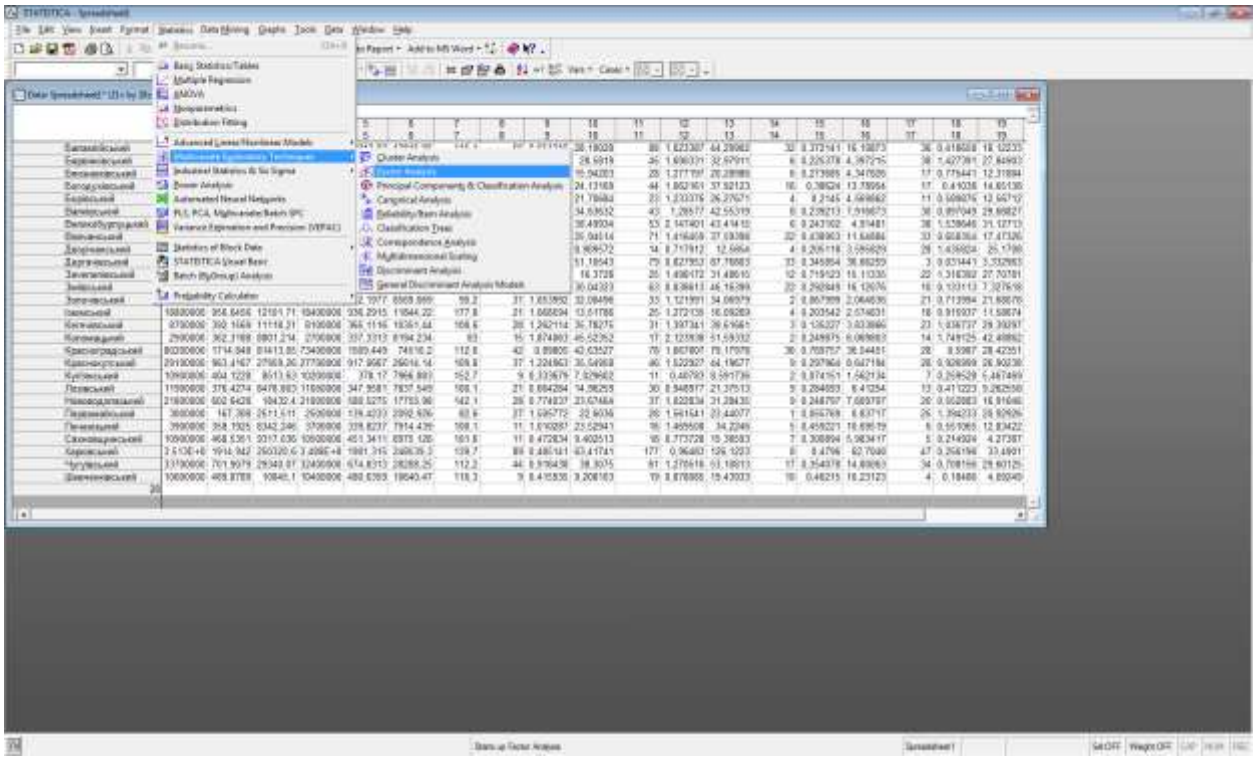

Рис. 1. Алгоритм вибору факторного аналізу в програмі Statistica 8.0

В основі побудови моделей факторного аналізу лежить твердження про те, що безліч взаємопов'язаних показників, що характеризують певний процес, можна представити меншою кількістю гіпотетичних змінних факторів і безліччю незалежних залишків.

Інформаційною базою розрахунків служать просторові ряди показників в розрізі адміністративно-територіальних одиниць регіону / регіонів країни за базові роки. Набір показників є суб'єктивним; основним принципом вибору показників були достовірність і порівнянність (рис. 2). Після обрання оптимальної кількості факторів, виконується інтерпретація результатів, враховуючи показники, які сформували виявлені чинники

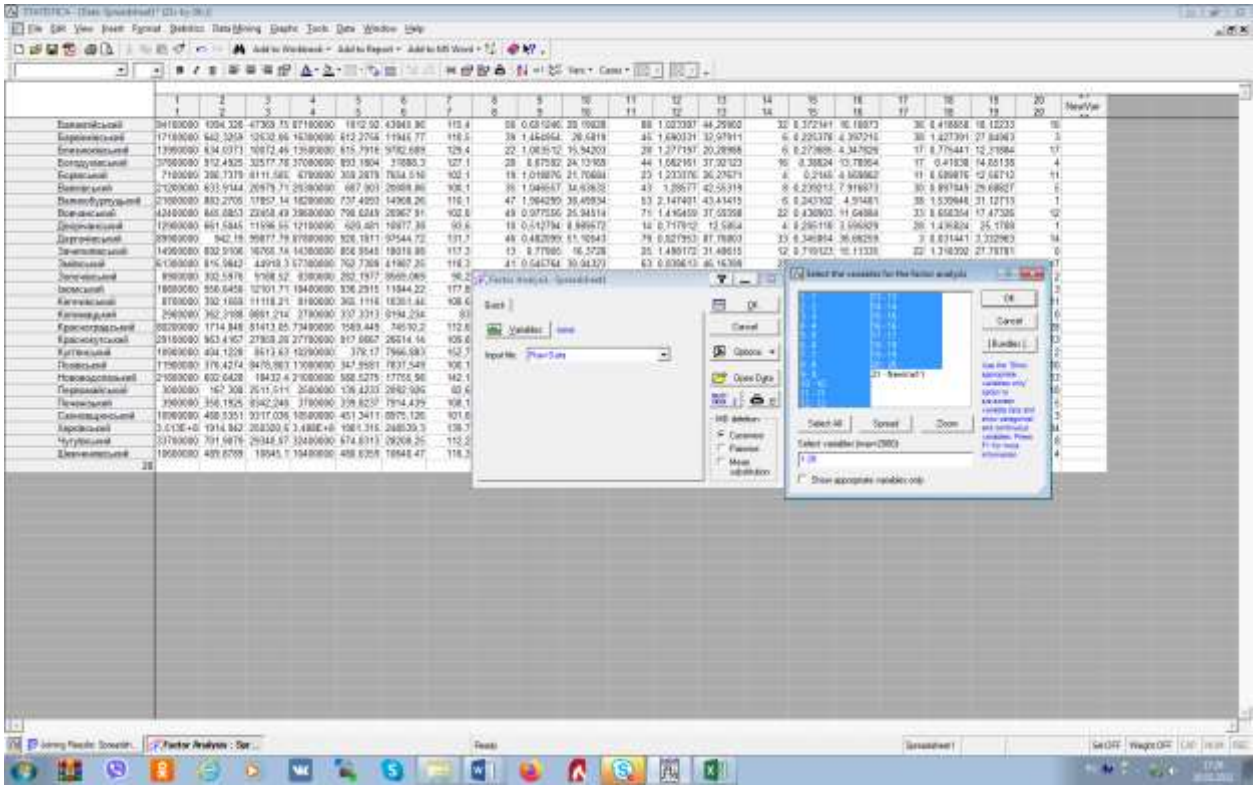

Рис. 2. Вибір вихідних даних і елементів виконання факторного аналізу

Зазвичай фактори, що впливають на розвиток суспільно-просторових процесів, характеризуються не одним, а безліччю показників, що мають між собою певний зв'язок (корелюють). Тому виникає потреба в їх об'єднанні в певну кількість груп за подібністю впливу, в результаті чого визначаються на перший погляд "приховані" (латентні) фактори, кількість яких, зрозуміло, менше кількості вихідних показників. Тобто чинники ідентифікуються.

Застосування факторного аналізу передбачає:

- по-перше, скорочення кількості показників (змінних), що на мові математичної статистики називається редукцією даних;

- по-друге, визначення структури взаємозв'язків між показниками (змінними), тобто класифікацію показників, інтерпретацію факторів.

Таким чином, факторний аналіз використовується і як метод зменшення кількості даних, і як метод класифікації одночасно. В основі побудови моделей факторного аналізу лежить твердження про те, що множину взаємозв’язаних показників, які характеризують певний процес, можна представити меншою кількістю гіпотетичних змінних – факторів та множиною незалежних залишків. Зміст факторного аналізу полягає у лінійному перетворенні n-вимірного простору у k-вимірний. Іншими словами, за допомогою факторного аналізу систему n показників можна замінити значно меншою кількістю (k) факторів.

Факторні навантаження знаходяться в межах від -1 до +1. Знак «+» чи «-» вказує на наявність прямої або оберненої залежності між показником і фактором. Зміст факторів визначають показники (змінні), що мають

найбільші факто-рні навантаження (найближчі по модулю до одиниці).

Процедура факторного аналізує включає такі складові:

1) визначення кількості факторів. На першому кроці розрахунків кількість факторів обрано рівною кількості показників та обраховано абсолютні, відносні та кумулятивні значення дисперсії кожного з факторів. Визначають, яка кіль-кість факторів є оптимальною. Зазвичай користуються одним з трьох критеріїв:

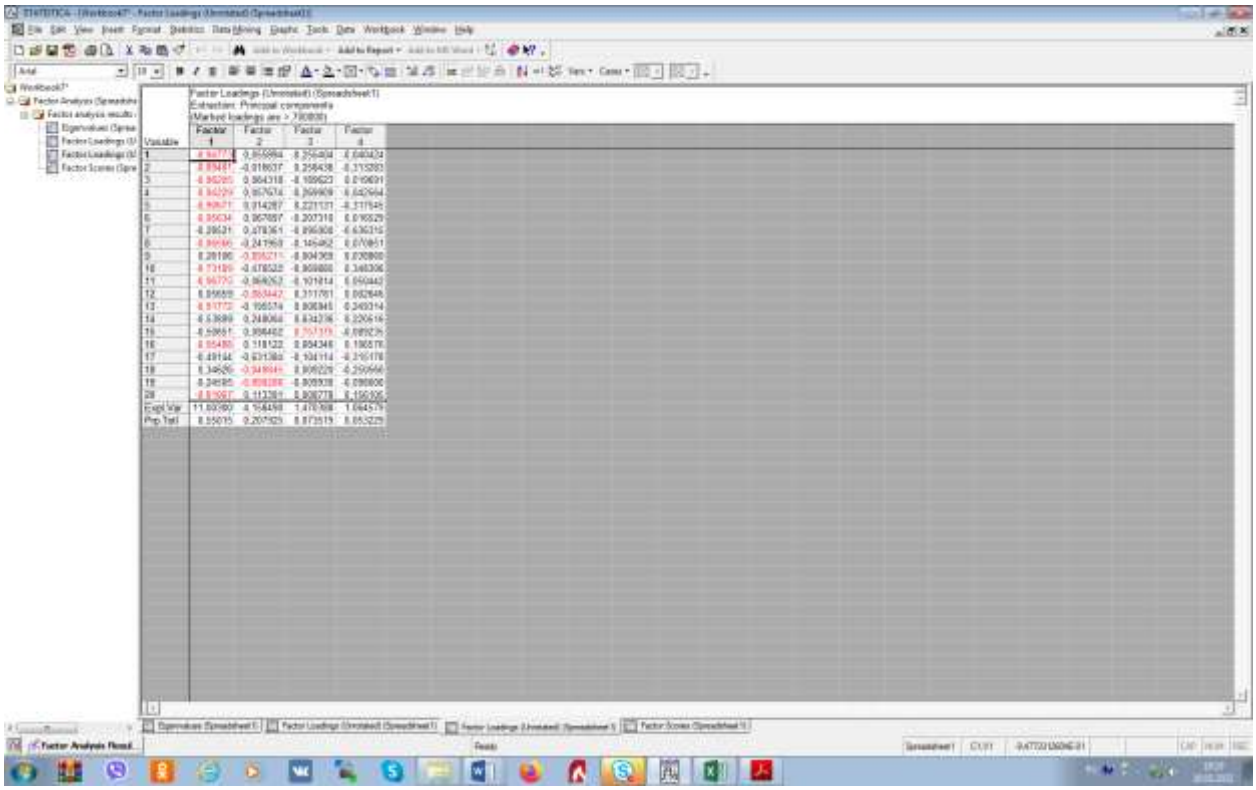

Рис. 3. Приклад протоколу факторного аналізу

– за критерієм Кайзера (Kaiser, 1960): обираються тільки фактори із дисперсією більше 1. Враховується те, що, якщо фактор не виділяє дисперсію, еквівалентну, хоча б дисперсії одного показника, то він відкидається.

– за кумулятивним відсотком: в якості визначальних обираються ті фактори, що сумарно охоплюють приблизно три чверті вихідної інформації (кумулятивний відсоток має перевищувати 75%). У наведеному прикладі два перших фактори пояснюють 90,5% загальної дисперсії – виділяємо два фактори;

– за критерієм «кам’янистого осипу» Кеттеля (Cattell, 1966): на графіку дисперсій (plot of eigenvalue, scree plot) знаходиться таке місце, де зменшення дисперсії зліва направо максимально уповільнюється. Передбачається, що справа від цієї точки знаходиться лише «факторіальний осип» (термін «осип» запозичений з геології, де означає уламки гірських порід, що накопичуються в нижній частині скелястого схилу). Далі здійснюється обернення осей координат (factor rotation), ідентифікація та інтерпретація факторів. У

результаті обернення отримується остаточна матриця факторних навантажень.

2) обрахунок факторних ваг. Факторні ваги (factor scores) – це показники, що відіграють роль оцінок вкладів територіальних одиниць у кожний з факторів. Матриця факторних ваг обчислюється шляхом множення матриці вихідних даних на матрицю факторних навантажень. Вони трактуються як відносні оцінки вияву певного фактору і служать основою для їх групування;

3) аналіз тенденцій.

Останнім завданням практичної роботи є інтерпретація отриманих (в ході виконання всіх видів аналізу) результатів.

План виконання практичної роботи:

1. Відкриваємо програму Statistica 8.0
2. Натискаємо кнопку New ( ) або Ctrl+N. Натискаємо ОК.
3. З'являється таблицю, в яку необхідно додати дані з бази даних.
4. Для цього копіємо в таблиці Excel базу даних і вставляємо її у таблицю програми Statistica.
5. Дана операція роться наступним чином. У першій ячійці правою кнопкою миші натискуємо у вікні, що відкривається, натискуємо «Paste With Headers → Paste With Both». Таким чином, ми копіюємо у таблицю програми Statistica наші показники, а також назви показників та назви районів/регіонів (рис. 4).

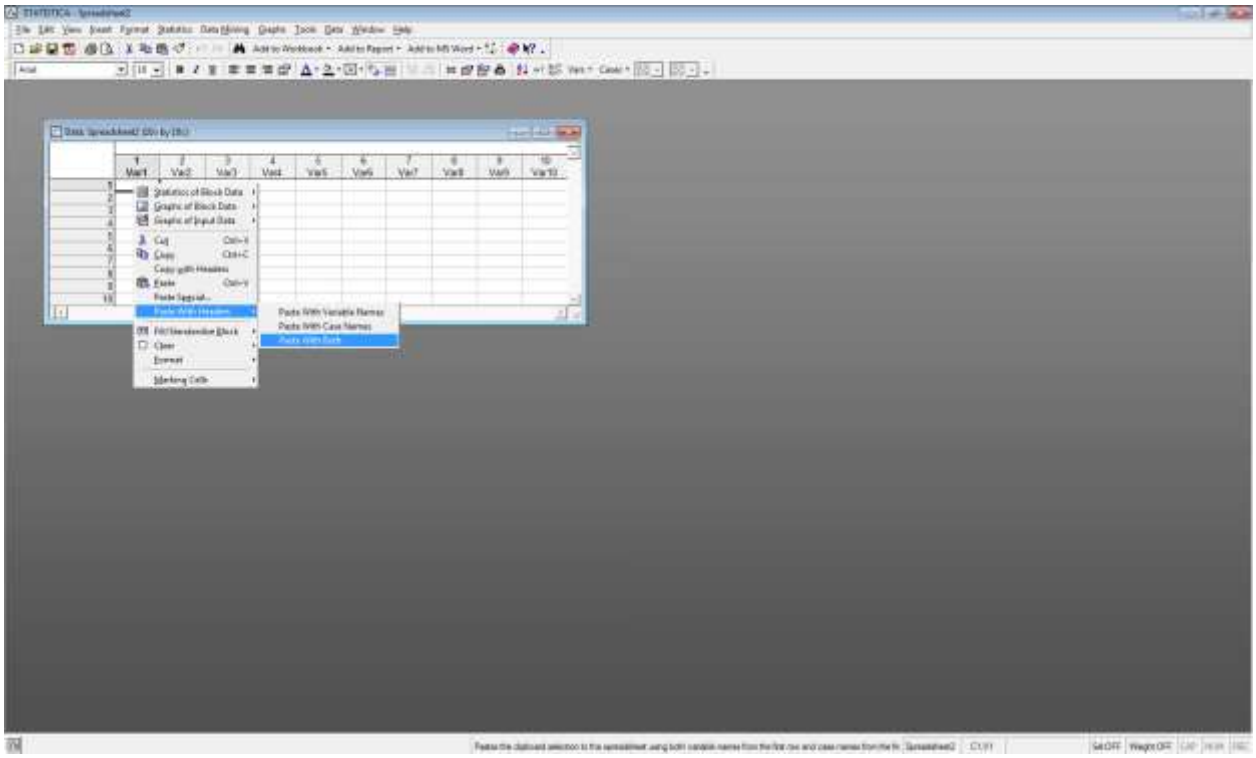

Рис. 4. Додавання бази даних у таблицю програми Statistica

6. Після цього на панелі задач натискаємо «Statistics» → «Multivariate Exploratory Techniques» → «Factor Analysis» (рис. 5).
7. Відкривається вікно факторного аналізу. Натискаємо на кнопку «Variables» та обираємо показники для аналізу (рис. 6). Натискаємо «ОК» і потім ще «ОК».

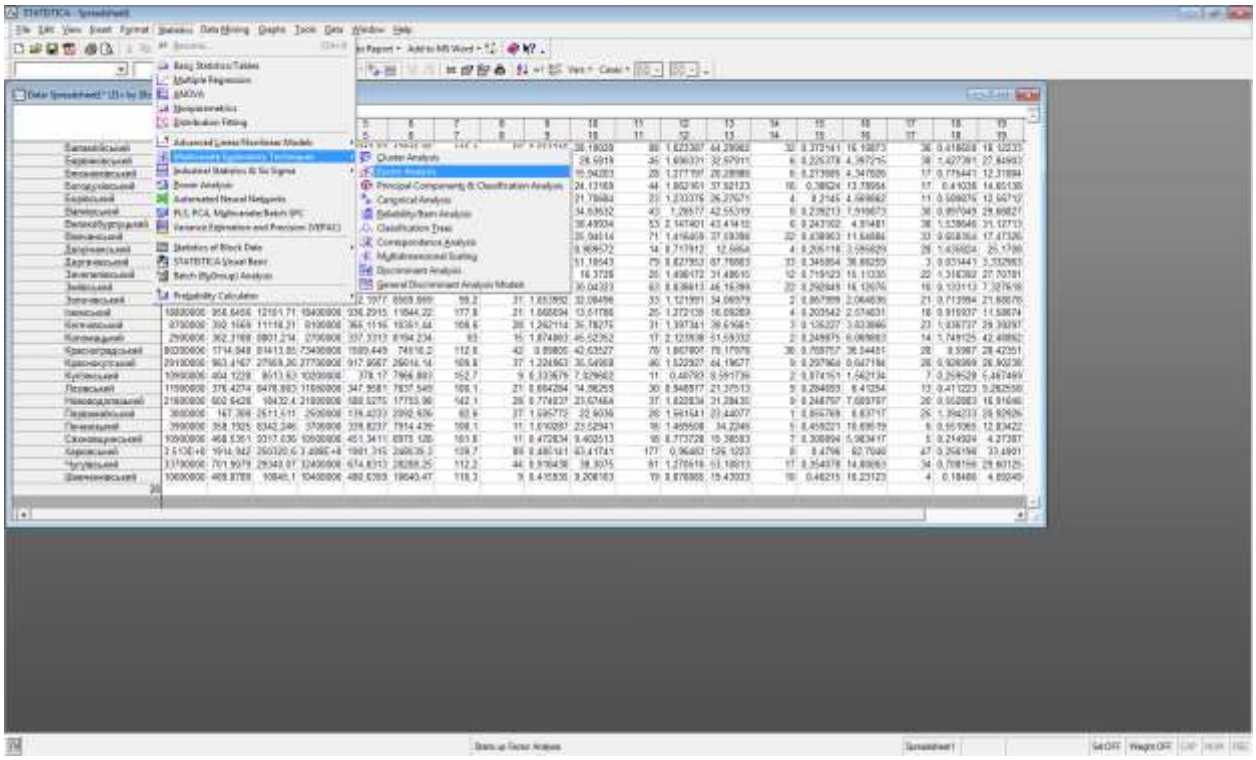

Рис. 5. Алгоритм вибору факторного аналізу в програмі Statistica 8.0

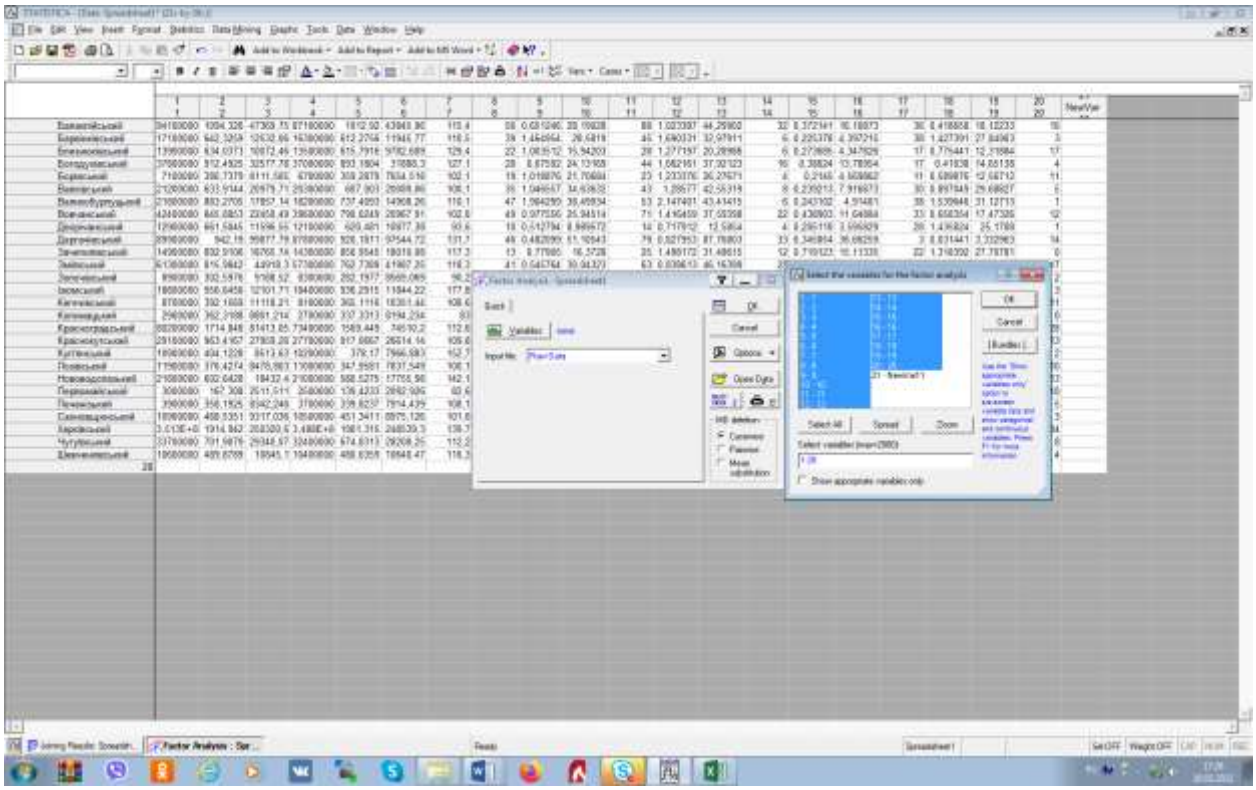

Рис. 6. Вибір вихідних даних і елементів виконання факторного аналізу

8. В результаті відкривається діалогове вікно «Define method of factor extraction» (рис. 7). Обираємо максимальну кількість факторів (наприклад, 10), в полі «Minimum eigenvalue» має бути 1,000. Це і власне значення фактору, яке має бути більше 1. Натискаємо «ОК».

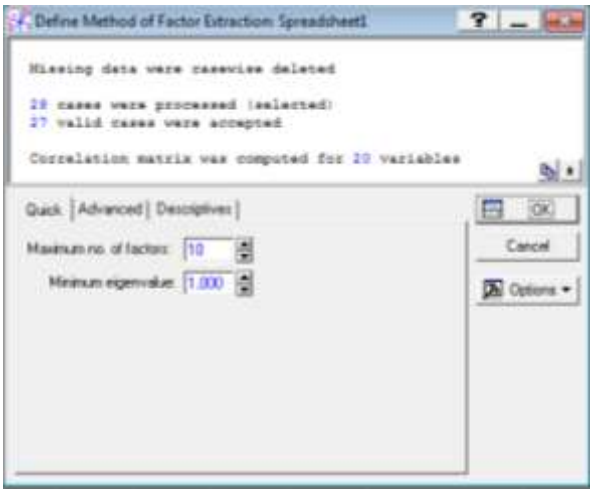

Рис. 7. Вибір методу виділення факторів

9. У вкладці «Advanced» у полі «Extraction Method» обираємо «Principal components». Натискаємо «ОК».
10. Відкривається діалогове вікно «Factor Analysis Results». У полі «Factor rotation» обираємо «Varimax normalized». Тобто ми обрали режим обертання факторів, про що згадувалося вище (рис. 8).

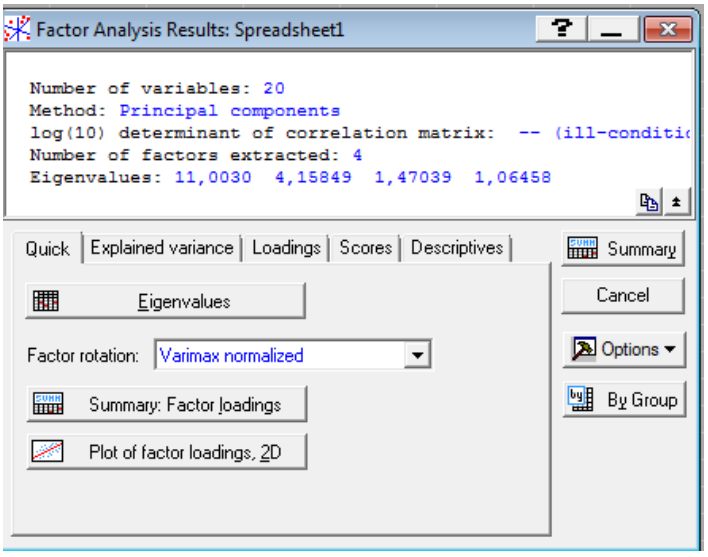

Рис. 8. Вибір обертання фактору

11. У вкладці «Loading» в полі «Highlight factor loadings greater than:» задати погове значення факторних навантажень, які будуть підсвічуватися, якщо значення навантажень буде більше цього значення. Ці значення потрібно враховувати при аналізі.

12. Повертаємося до вкладки «Quick» та натискаємо кнопку «Eigenvalue».
13. Відкривається таблиця із власними значеннями факторів. Eigenvalue має бути більше за 1. Таким чином, у даному випадку за цим критерієм можна виділити 4 фактори. Cumulative % має бути вище 75%. У даному випадку за цим критерієм можна виділити 2-4 фактори (рис. 9). Копіюємо цю таблицю в практичну роботу.
14. У лівому нижньому куту шукаємо «Factor Analysis Result» та повертаємося діалогового вікна і натискаємо кнопку «Summary». Перед нами відкривається реультуюча таблиця (рис. 10). Програма виділила 4 фактори. Кожний з досліджуваних показників має факторне навантаження фактору і можна зрозуміти, які показники входять до того чи інфшого фактору. Беремо до уваги факторні навантаження, що виділені червоним кольором.

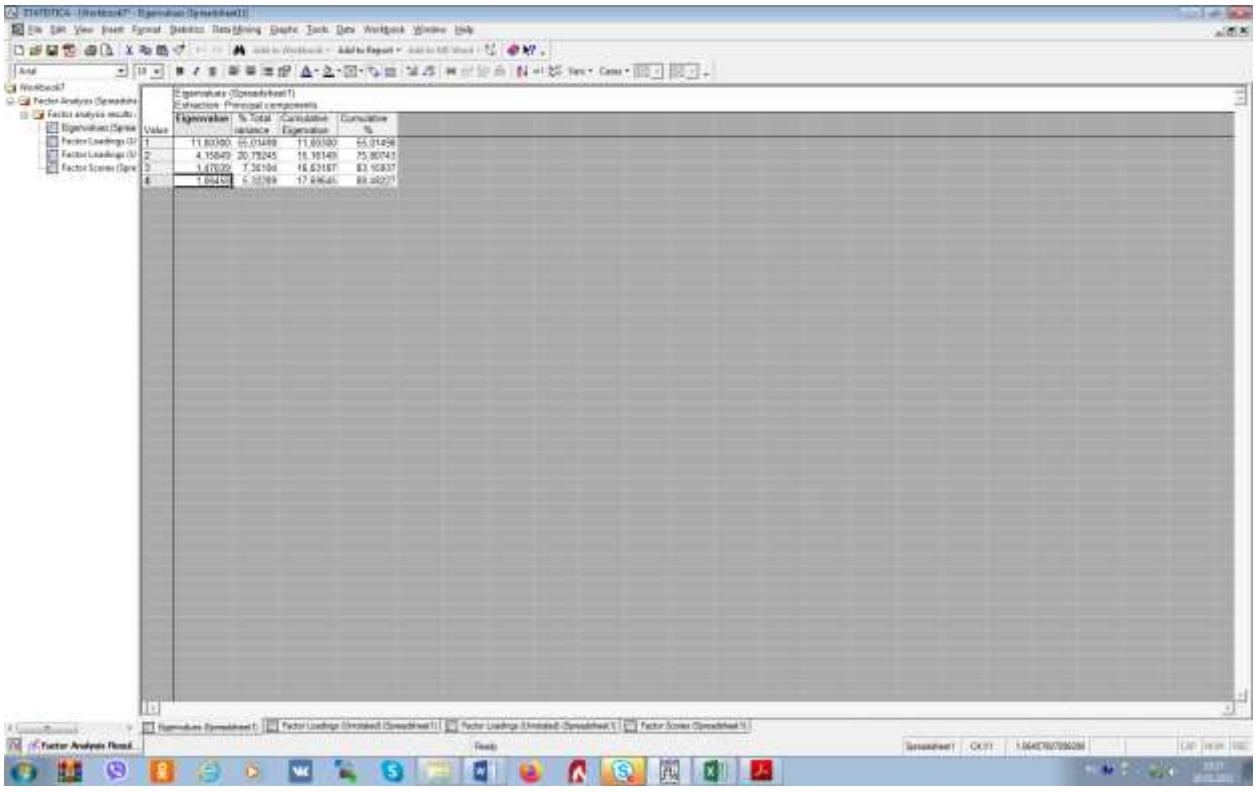

Рис. 9. Таблиця власних значень факторів

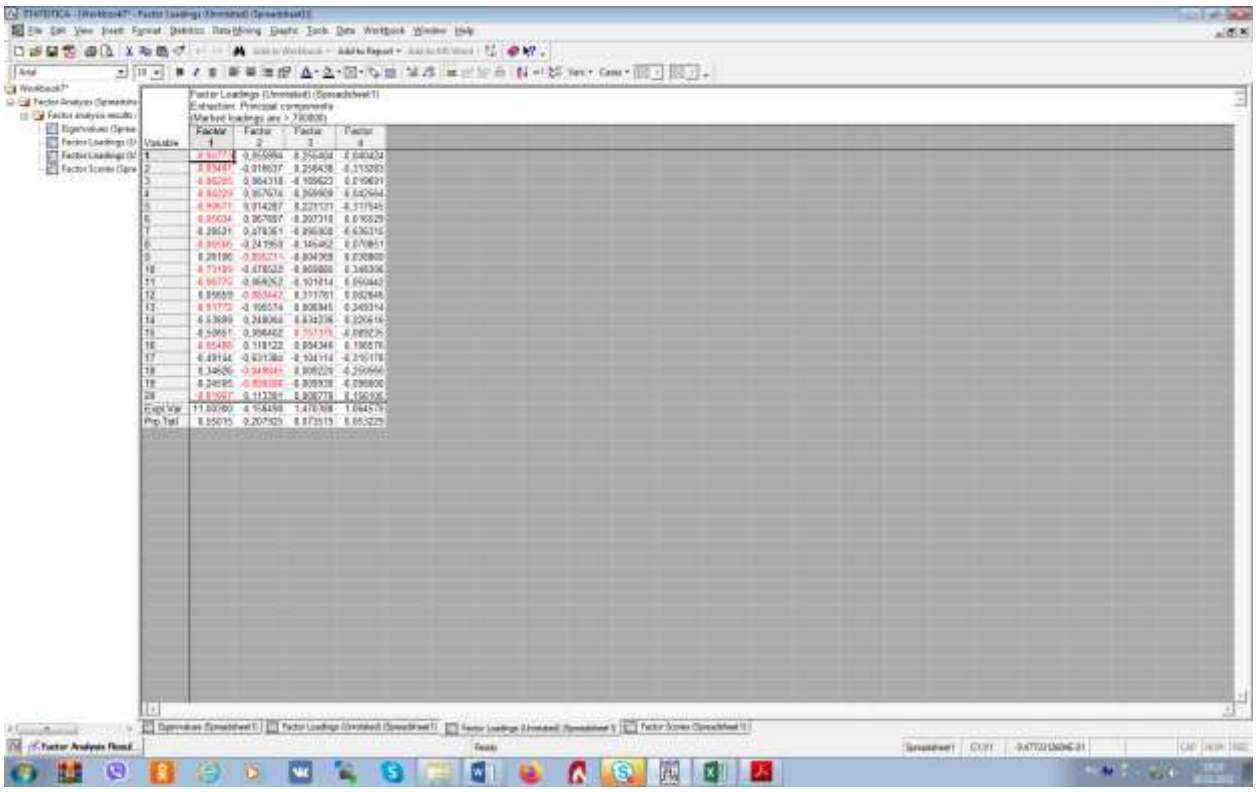

Рис. 10. Приклад протоколу факторного аналізу

15. У лівому нижньому куту шукаємо «Factor Analysis Result» та повертаємося діалогового вікна. Переходимо до вкладки «Scores» та натискаємо кнопку «Factor scores». Відкривається таблиця з факторними вагами (рис. 11). З цієї таблиці можна проаналізувати вклад територіальних одиниць у кожний з факторів.

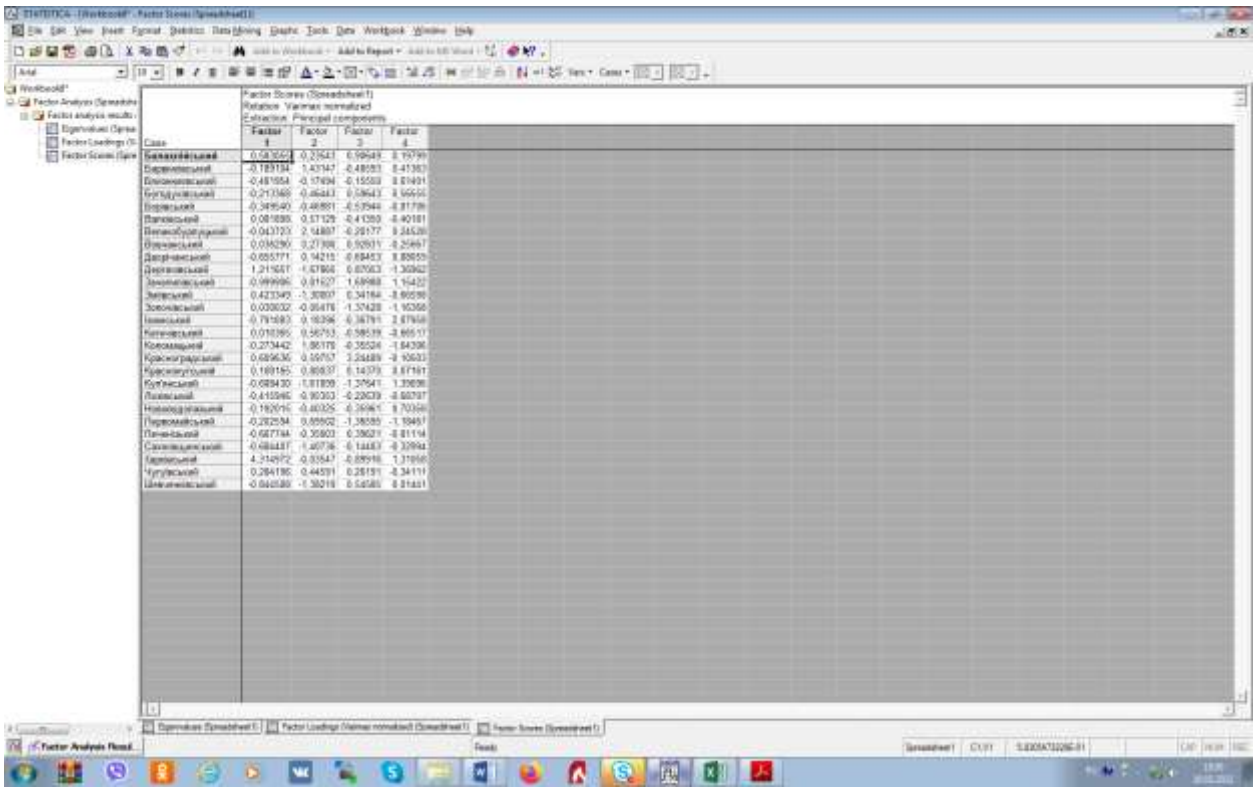

Рис. 11. Визначення факторних вагів

16. Завдання студентам: проаналізувати отримані результати, дати назву отриманим факторам. Пояснити, чому той чи інший показник відноситься до певного фактора та вклад регіонів / районів у той чи інший фактор. Використовуати усі показники, які ви обрали для бази даних.