НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»

Інститут прикладного системного аналізу

Кафедра системного проектування

ЗВІТ

з виконання лабораторної роботи №2

з дисципліни «*Еколого-економічна організація виробництва*»

на тему: «*Кластерний аналіз соціально-економічних індикаторів, що впливають на рівень доходів населення у регіонах України*»

Виконав:

студент 4 курсу

групи ДА-82

Муравльов Андрій

Київ — 2021

Метою **кластерного аналізу** є розділення об’єктів на відносно однорідні групи, судячи з індикаторів, що їх характеризують. Такі об’єкти в групах є відносно схожими та відносно різняться від об’єктів у інших групах.

Наведемо таблицю з вхідними параметрами (рис. 1). Згідно даних Держкомстату України наведені значення регіональних індикаторів за 2016 р. Серед них: рівень доходів населення у регіонах України (РД), заробітна плата (ЗП), прибуток та змішаний дохід (ПЗД), доходи від власності (ДВ), трансферти (соціальні допомоги (СД) та соціальні трансферти в натурі (СТН)), млн. грн.

	A	B	C	D	E	F	G
1	Регіон	РД	ЗП	ПЗД	ДВ	СД	СТН
2	Вінницька	69 654	23 458	19 043	2 447	12 522	10 746
3	Волинська	39 359	13 537	9 255	1 071	8 064	6 733
4	Дніпропетровська	184 138	86 057	33 836	7 462	29 886	21 892
5	Донецька	111 547	55 007	11 260	3 370	27 512	11 588
6	Житомирська	51 920	18 436	11 822	1 424	10 509	8 634
7	Закарпатська	42 235	14 501	10 474	804	8 034	7 243
8	Запорізька	94 160	37 880	22 191	3 282	16 472	11 402
9	Івано-Франківська	54 492	16 483	15 608	1 219	10 155	8 950
10	Київська	87 937	39 426	17 543	2 252	16 030	11 529
11	Кіровоградська	40 427	14 247	9 073	1 999	7 899	6 418
12	Луганська	38 022	17 685	3 094	1 193	10 434	5 014
13	Львівська	112 697	44 323	24 725	3 707	19 520	17 117
14	Миколаївська	50 728	20 881	9 976	1 718	8 977	7 191
15	Одеська	115 025	44 524	21 667	3 675	17 654	13 606
16	Полтавська	69 789	28 707	12 483	4 335	12 319	10 502
17	Рівненська	45 716	16 201	11 027	1 133	8 966	7 609
18	Сумська	50 951	18 803	11 858	1 943	9 030	8 256
19	Тернопільська	38 727	12 275	8 922	1 072	7 330	7 506
20	Харківська	131 681	52 212	28 455	4 281	23 011	18 727
21	Херсонська	42 707	13 768	10 773	1 391	7 756	5 694
22	Хмельницька	55 542	18 123	14 690	2 180	10 433	8 786
23	Черкаська	51 710	18 901	8 840	3 004	10 792	8 929
24	Чернівецька	32 397	9 664	8 471	825	6 215	5 544
25	Чернігівська	44 283	16 288	8 934	1 797	8 985	7 566
26	м. Київ	333 927	185 379	34 193	17 868	29 268	33 378

Рис. 1. Значення вхідних показників за 2016 рік

У роботі необхідним є дослідження методичних особливостей кластерного аналізу, зокрема, чи впливає використання того чи іншого методу та процедура нормування на структуру кластерного розподілу регіонів за вибраними індикаторами. Виберемо п’ять методів кластерного аналізу:

1. Метод центроїдної класифікації.
2. Метод центроїдної класифікації із застосуванням процедури нормування.
3. Метод ближнього сусіда.
4. Метод ближнього сусіда із застосуванням процедури нормування.
5. Метод Уорда із застосуванням процедури нормування.

Короткий опис методів

Ієрархічні методи дозволяють поєднувати елементи на базі понять відстані чи подібності між точками в багатомірному просторі ознак. Результатом такої розбивки є дендрограма, що показує етапи об’єднання об’єктів в групи за характеристиками.

Опишемо метод центроїдної класифікації. *Кластерний центроїд* – середнє значення змінних для всіх об’єктів у конкретному кластері. Відстань між двома кластерами визначається як евклідова відстань між центрами цих кластерів. Кластеризація йде поетапно, на кожному з n-1 кроків об'єднують два кластери. Якщо n_1 більше n_2, то центри об'єднання двох кластерів близькі один до одного і характеристики другого кластера при об'єднанні кластерів практично ігноруються. Іноді цей метод називають методом зважених груп.

Метод ближнього сусіда. Тут два об'єкти, які належать одній і тій самій групі, мають коефіцієнт подібності, який менше деякого порогового значення S. В термінах евклідової відстані d це означає, що відстань між двома точками (об'єктами) кластеру не повинна перевищувати деякого порогового значення h. Таким чином, h визначає максимально допустимий діаметр підмножини, що утворює кластер.

Метод Уорда. В якості цільової функції застосовують внутрішньогрупову суму квадратів відхилень, що є сумою квадратів відстаней між кожною точкою (об'єктом) і середньою по кластеру, який містить цей об'єкт. На кожному кроці об'єднуються такі два кластери, які призводять до мінімального збільшення цільової функції, тобто внутрішньогрупової суми квадратів. Цей метод направлений на об'єднання близько розташованих кластерів.

Порівняння методів

Перевагою ієрархічних методів кластеризації порівняно з неієрархічними методами є їх наочність і можливість одержати детальне подання структури даних. Використовуючи ієрархічні методи, можливо доволі легко ідентифікувати викиди в наборі даних й, у результаті, підвищити якість даних.

Однак, використання ієрархічних методів супроводжується наступними недоліками, зокрема:

- обмеженням обсягу набору даних;
- обмеженням вибору міри близькості;
- негнучкості отриманих класифікацій об’єктів.

Розглянемо детальніше порівняльну характеристику обраних методів (табл. 1).

Таблиця 1. Таблиця порівняння методів

Метод	Переваги	Недоліки
Метод центроїдної класифікації	Простота використання результатів. Рішення не є унікальними для конкретної ситуації. Пошук не гарантовано дасть вірне рішення, але найкраще з можливих	Висока залежність результатів класифікації від обраної метрики. Обчислювальна трудомісткість. Тільки для невеликої розмірності.
Ближнього сусіда	Простота використання одержаних результатів. Рішення не є унікальними для конкретної ситуації, можливе їх використання для інших випадків. Метою пошуку є не гарантовано вірне рішення, а найкраще з можливих	У виборі рішення правила ґрунтуються на всьому масиві доступних даних, тому неможливо сказати, на якій основі будуються відповіді. Існує складність вибору міри "близькості" (метрики). Є висока залежність результатів класифікації від обраної метрики. Завдання даного методу – невеликої розмірності.
Уорда	Дозволяє отримати компактні добре виражені кластери, добре працює з групами подібних розмірів, ефективно виділяє структуру, «приховану» випадковою мінливістю ознак	Обмеження обсягу набору даних; негнучкість отриманих класифікацій об’єктів.

Опис кроків кластеризації

Крок 1. Завантаження даних. На рис. 1 наведено дані для кластеризації.

	Регіон	РД	ЗП	ПЗД	ДВ	СД	СТН
1	Вінницька	69654	23458	19043	2447	12522	10746
2	Волинська	39359	13537	9255	1071	8064	6733
3	Дніпропетровська	184138	86057	33836	7462	29886	21892
4	Донецька	111547	55007	11260	3370	27512	11588
5	Житомирська	51920	18436	11822	1424	10509	8634
6	Закарпатська	42235	14501	10474	804	8034	7243
7	Запорізька	94160	37880	22191	3282	16472	11402
8	Івано-Франківська	54492	16483	15608	1219	10155	8950
9	Київська	87937	39426	17543	2252	16030	11529
10	Кіровоградська	40427	14247	9073	1999	7899	6418
11	Луганська	38022	17685	3094	1193	10434	5014
12	Львівська	112697	44323	24725	3707	19520	17117
13	Миколаївська	50728	20881	9976	1718	8977	7191
14	Одеська	115025	44524	21667	3675	17654	13606
15	Полтавська	69789	28707	12483	4335	12319	10502
16	Рівненська	45716	16201	11027	1133	8966	7609
17	Сумська	50951	18803	11858	1943	9030	8256
18	Тернопільська	38727	12275	8922	1072	7330	7506
19	Харківська	131681	52212	28455	4281	23011	18727
20	Херсонська	42707	13768	10773	1391	7756	5694
21	Хмельницька	55542	18123	14690	2180	10433	8786

Рис. 1. Завантажені дані для кластеризації програмою SPSS

Крок 2. Формування набору даних для кластеризації. З метою підготовки даних до кластерізації необхідно в поле "Метить наблюдения значениями:" перенести з лівого списку пункт "Регіон", а в список "Переменные" – всі інші фактори (рис. 2).

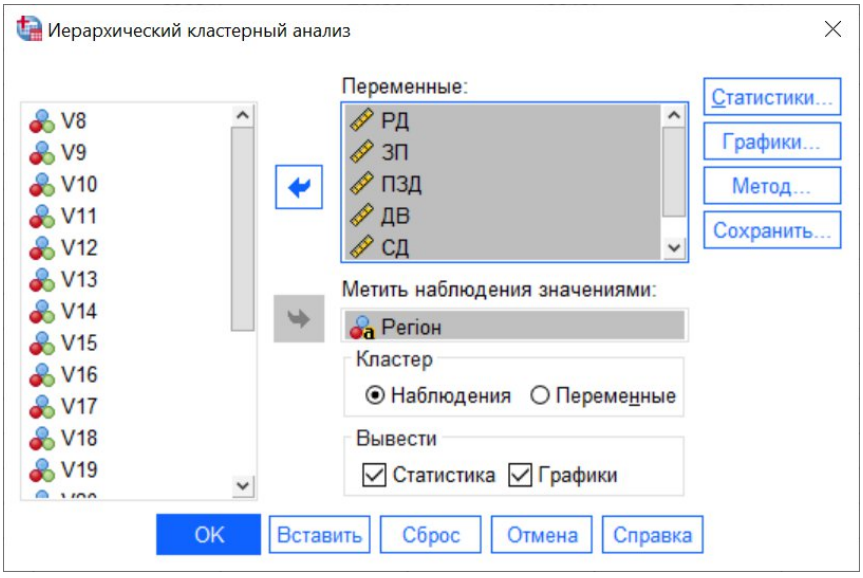

Рис. 2. Підготовка даних до кластеризації програмою SPSS

Після вибору методу та кількості кластерів отримаємо вікно виводу результатів кластеризації (рис. 3).

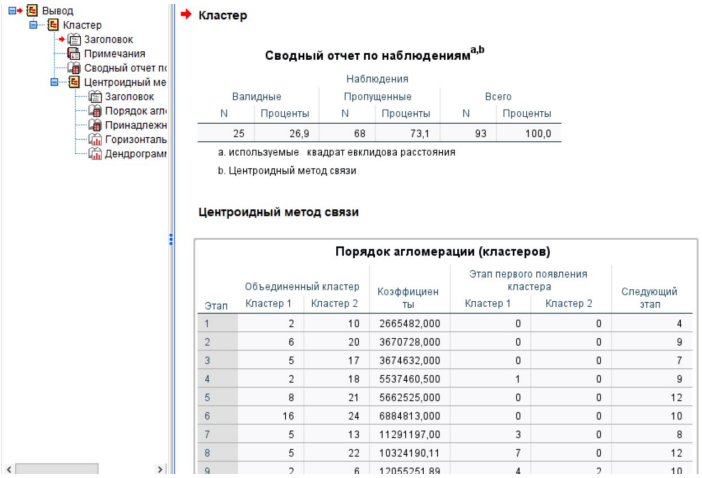

Рис. 3. Вікно виводу результату кластеризації програмою SPSS

На рис. 4 представлено результат кластеризації методом центроїдної класифікації. Тут перший стовпчик відповідає назві об’єкту, в другому вказується номер кластеру, до якого належить об’єкт.

Принадлежность к кластерам

Наблюдение	Кластеры 5
1:Вінницька	1
2:Волинська	2
3:Дніпропетровська	3
4:Донецька	4
5:Житомирська	2
6:Закарпатська	2
7:Запорізька	1
8:Івано-Франківська	2
9:Київська	1
10:Кіровоградська	2
11:Луганська	2
12:Львівська	4
13:Миколаївська	2
14:Одеська	4
15:Полтавська	1
16:Рівненська	2
17:Сумська	2
18:Тернопільська	2
19:Харківська	4
20:Херсонська	2
21:Хмельницька	2
22:Черкаська	2
23:Чернівецька	2
24:Чернігівська	2
25:м. Київ	5

Рис. 4. Результат кластеризації методом центроїдної класифікації

На рис. 5 наведена дендрограма кластеризації методом центроїдної класифікації.

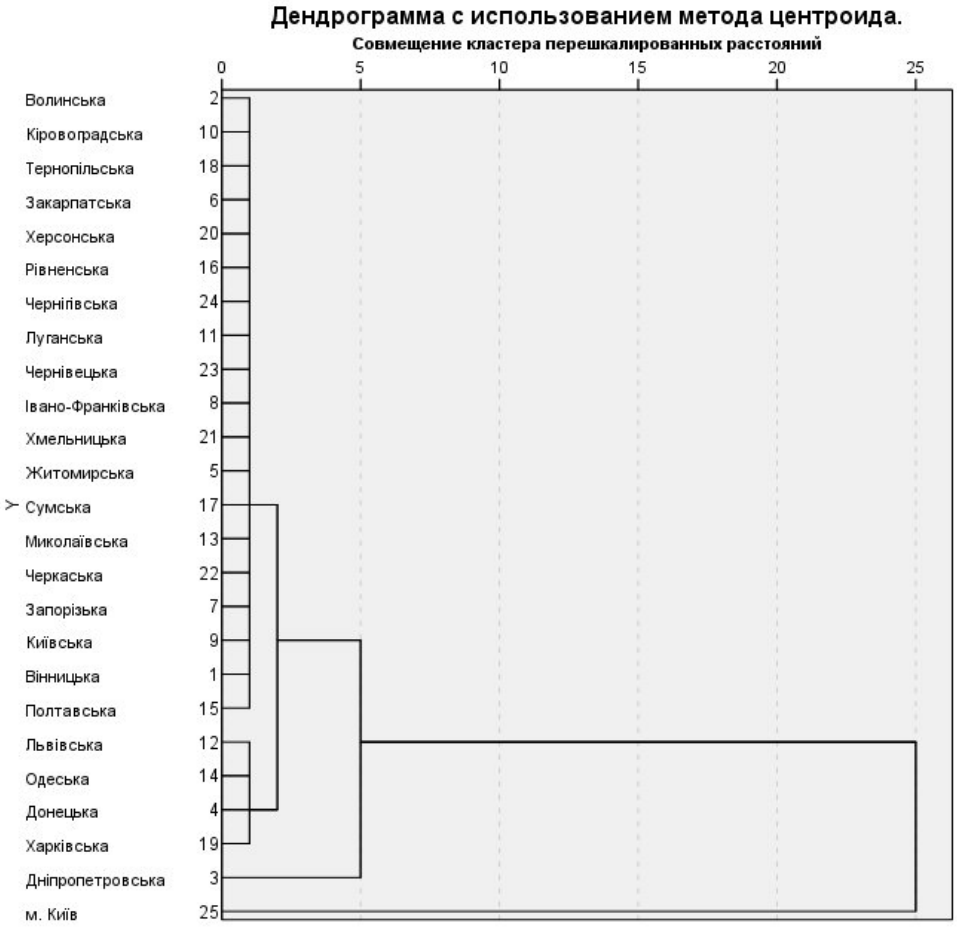

Рис. 5. Дендрограма кластеризації методом центроїдної класифікації без нормалізації

Результати кластерного розподілу регіонів з використання різних методів кластерного аналізу наведено у таблиці 2.

Таблиця 2. Результати кластерного розподілу регіонів різними методами

№ кластера	Методи із застосуванням нормалізації			Методи без застосування нормалізації	
	Метод центроїдної класифікації	Метод ближнього сусіда	Метод Уорда	Метод центроїдної класифікації	Метод ближнього сусіда
1	Вінницька, Запорізька, Київська, Львівська, Одеська, Харківська	Інші (21)	Вінницька, Запорізька, Київська, Львівська, Одеська, Харківська	Вінницька, Запорізька, Київська, Полтавська	Інші (17)
2	Інші (16)	Дніпропетровська	Інші (16)	Інші (15)	Дніпропетровська
3	Дніпропетровська	Донецька	Дніпропетровська	Дніпропетровська	Донецька, Запорізька, Київська, Львівська, Одеська
4	Донецька	Луганська	Донецька	Харківська, Одеська, Львівська, Донецька	Харківська
5	м. Київ	м. Київ	м. Київ	м. Київ	м. Київ

Дані таблиці 2 показують, що методи дають різні результати. Усіх об’єднує лише 5 кластер, до якого завжди входить виключно м. Київ. Найбільш схожий (ідентичний) розподіл за кластерами спостерігається між методами центроїдної класифікації (з нормалізацією) та Уорда. Процедура нормалізації доволі впливає на кластерний розподіл: деякі регіони помітно мігрують між кластерами (див. таблицю 2). Далі у роботі для прикладу буде використаний метод центроїдної класифікації.

На рис. 6 показано кластерний розподіл регіонів України за 2016 р. за методом центроїдної класифікації та середні значення показників.

	A	B	C	D	E	F	G	H
1	Кластер	Регіон	РД	ЗП	ПЗД	ДВ	СД	СТН
2	1	Вінницька	69 654	23 458	19 043	2 447	12 522	10 746
3		Запорізька	94 160	37 880	22 191	3 282	16 472	11 402
4		Київська	87 937	39 426	17 543	2 252	16 030	11 529
5		Львівська	112 697	44 323	24 725	3 707	19 520	17 117
6		Одеська	115 025	44 524	21 667	3 675	17 654	13 606
7		Харківська	131 681	52 212	28 455	4 281	23 011	18 727
8	**Середнє значення:**		**101 859**	**40 304**	**22 271**	**3 274**	**17 535**	**13 855**
9	2	Волинська	39 359	13 537	9 255	1 071	8 064	6 733
10		Житомирська	51 920	18 436	11 822	1 424	10 509	8 634
11		Закарпатська	42 235	14 501	10 474	804	8 034	7 243
12		Івано-Франківська	54 492	16 483	15 608	1 219	10 155	8 950
13		Кіровоградська	40 427	14 247	9 073	1 999	7 899	6 418
14		Луганська	38 022	17 685	3 094	1 193	10 434	5 014
15		Миколаївська	50 728	20 881	9 976	1 718	8 977	7 191
16		Полтавська	69 789	28 707	12 483	4 335	12 319	10 502
17		Рівненська	45 716	16 201	11 027	1 133	8 966	7 609
18		Сумська	50 951	18 803	11 858	1 943	9 030	8 256
19		Тернопільська	38 727	12 275	8 922	1 072	7 330	7 506
20		Херсонська	42 707	13 768	10 773	1 391	7 756	5 694
21		Хмельницька	55 542	18 123	14 690	2 180	10 433	8 786
22		Черкаська	51 710	18 901	8 840	3 004	10 792	8 929
23		Чернівецька	32 397	9 664	8 471	825	6 215	5 544
24		Чернігівська	44 283	16 288	8 934	1 797	8 985	7 566
25	**Середнє значення:**		**46 813**	**16 781**	**10 331**	**1 694**	**9 119**	**7 536**
26	3	Дніпропетровська	184 138	86 057	33 836	7 462	29 886	21 892
27	4	Донецька	111 547	55 007	11 260	3 370	27 512	11 588
28	5	м. Київ	333 927	185 379	34 193	17 868	29 268	33 378

Рис. 6. Кластерний розподіл регіонів за методом центроїдної класифікації

Розглянемо детальніше рис. 6 та спробуємо знайти закономірності при формуванні кластерів. Перше, що помічається, — особливість 2 кластеру, що полягає у найменших показниках серед усіх (див. Середнє значення). Далі, у 1 кластер потрапили регіони, що мають показники вище (до того ж, не аномально різні). Регіони з аномальними показниками розташувалися у 3-5 кластерах: це Дніпропетровська обл. з порівняно вищими усіма показниками; Донецька обл. з високими доходами, але низькими СТН та ПЗД та м. Київ з дуже високими показниками, що не можуть бути порівняні з іншими регіонами.