УДК 332.1

Кластерний аналіз областей України за випуском продуктів харчування

Олеся Тоцька

Здійснено кластерний аналіз областей України за випуском продуктів харчування. Його реалізація відбувається ієрархічним методом та методом К-середніх. Для автоматизації розрахунків використовується програмний пакет StatSoft Statistica 6.0. Актуальність питання полягає в тому, що кластерний аналіз регіонів України за такими показниками не проводився.

Ключові слова: *кластерний аналіз, ієрархічний метод, метод К-середніх.*

Останнім часом багато розмов точиться навколо адміністративно-територіальної реформи України. Суть її полягає у зменшенні чи збільшенні областей (районів) країни. Вагомий внесок у визначення порядку їх об'єднання та оптимальної кількості може внести кластерний аналіз.

Кластерний аналіз є сукупністю методів класифікації багатовимірних спостережень. Основною його метою є розподіл сукупності вхідних даних на однорідні групи так, щоб об'єкти всередині групи були подібними між собою згідно з деяким критерієм, а об'єкти із різних груп відрізнялися один від одного. Причому класифікація об'єктів проводиться одночасно за декількома ознаками на основі введення певної міри сумарної близькості за всіма ознаками класифікації [1, с. 129].

Вперше термін «кластерний аналіз» був вжитий дослідником Тріоном у 1939 р. Походить він від англійського слова cluster – гроно, згусток. Поряд з ним вживаються також терміни "автоматична класифікація", "таксономія" чи розпізнавання образів без навчальної вибірки.

Кластерний аналіз дозволяє розглядати достатньо великий об'єм інформації і різко скорочувати великі масиви соціально-економічної інформації, робити їх компактними і наочними [2, с. 50].

У даний час він широко використовується у сфері медицини (класифікації захворювань, їх лікуванні), психіатрії (групуванні симптомів захворювань), археології (класифікація кам'яних знарядь), біології (групуванні тварин за видами), а також маркетингових дослідженнях, хоча вперше його застосували у соціології.

Використанню кластерного аналізу для класифікації соціально-економічних об'єктів присвячено ряд наукових публікацій Я.І. Єлейка, Р.Т. Грищук, Н.Ю. Подольчака, А.О. Цапіна, Л. Антонюк, В. Сацик. Так, зокрема, його використовують для класифікації споживачів електроенергії [3], оцінки економічного ризику підприємств [4], побудови моделі розвитку [5], аналізу міжнародної конкурентоспроможності країн [6] тощо.

Метою написання цієї статті є проведення кластерного аналізу областей України за випуском продуктів харчування, а також визначення, до якого кластера належить Волинський регіон. Основні завдання статті:

© *Тоцька Олеся Леонтіївна* – старший викладач Волинського державного університету ім. Л. Українки.

1) побудувати матрицю вхідних даних;
2) провести стандартизацію вхідних даних;
3) побудувати симетричну матрицю евклідових відстаней;
4) провести об'єднання об'єктів у кластери (ієрархічну класифікацію);
5) визначити оптимальну кількість кластерів;
6) проаналізувати отримані кластери.

Актуальність її написання полягає в тому, що кластерний аналіз регіонів України за такими показниками не проводився.

Задачі кластерного аналізу прийнято поділяти на два основні види залежно від обсягу сукупності вхідних даних.

До першого виду належать задачі класифікації порівняно невеликих за об'ємом сукупностей спостережень, які, як правило, складаються з декількох десятків спостережень. Сюди належать задачі класифікації таких макрооб'єктів, як країни, області, міста, підприємства, типи технологічних процесів тощо.

До другого виду належать задачі класифікації великих за об'ємом сукупностей спостережень, які складаються з сотень і тисяч спостережень. Зокрема це задачі класифікації таких мікрооб'єктів, як індивіди, сім'ї, вироби тощо [7, с. 504].

Розрізняють такі три типи методів кластерного аналізу, як ієрархічні, К-середніх, двовходове об'єднання.

Ієрархічні методи поділяються на агломеративні (послідовно об'єднують об'єкти у більші групи) і дивизимні (послідовно розділяють об'єкти на менші групи).

При застосуванні *методу К-середніх* вибираються К-випадкові кластери, а потім змінюється приналежність до них об'єктів таким чином, щоб:

1) мінімізувати змінність всередині кластерів;
2) максимізувати змінність між кластерами.

Тобто при цьому потрібно наперед задавати кількість кластерів, яку бажаємо отримати.

Двовходове об'єднання проводить кластеризацію як спостережень, так і змінних, тобто у двох напрямках. Воно використовується досить рідко, порівняно до інших методів, і тільки тоді, коли передбачається, що і спостереження, і змінні одночасно вносять вклад у визначення осмислених кластерів.

Класифікацію регіонів України будемо проводити у два етапи. На першому за допомогою ієрархічного агломеративного методу розглянемо порядок об'єднання областей у кластери. На другому за допомогою методу К-середніх отримаємо кластери з однорідними регіонами за випуском продовольчих продуктів.

Автоматизація розрахунків відбуватиметься за допомогою використання програмного пакета StatSoft Statistica 6.0, який має високий рейтинг серед інших статистичних систем, таких як SPSS, Systal, Minitab, S-Plus [8, с. 114].

Процедури цієї програми вирізняються високою швидкістю і точністю обчислень. Програма має такі загальновизнані переваги:

1) містить повний набір класичних методів аналізу даних;
2) відповідає всім стандартам Windows;
3) легка в освоєнні;
4) підтримує високоякісну графіку тощо [9, с. 46].

Для проведення дослідження побудуємо таблицю 1, у якій усі області України характеризуватимуться десятьма показниками. Кожен з них – це випуск певного основного продукту харчування, зокрема:

показник 1 – м'ясо, включаючи субпродукти 1-ї категорії (тис. т);

Таблиця 1

Матриця вхідних даних

№ з/п	Область України	Показник 1	Показник 2	Показник 3	Показник 4	Показник 5	Показник 6	Показник 7	Показник 8	Показник 9	Показник 10
1.	АР Крим	36,3	3,1	2,4	40,5	0,5	131,7	3,6	146,0	132,0	548,0
2.	Вінницька	40,9	7,6	28,2	48,2	13,7	178,3	62,7	1317,0	234,0	416,0
3.	Волинська	16,1	13,8	4,0	32,8	6,4	116,7	8,2	1042,0	171,0	212,0
4.	Дніпропетровська	55,9	35,3	3,6	169,2	2,8	22,6	58,6	595,0	492,0	686,0
5.	Донецька	27,7	65,6	2,8	102,9	1,2	17,7	164,0	564,0	401,0	1098,0
6.	Житомирська	20,0	3,2	17,3	20,9	9,8	26,8	30,1	1134,0	205,0	401,0
7.	Закарпатська	2,1	1,2	0,0	4,2	0,0	98,2	8,8	522,0	208,0	293,0
8.	Запорізька	16,6	13,9	4,7	30,0	11,3	29,8	25,4	223,0	252,0	544,0
9.	Івано-Франківська	14,4	2,4	0,8	13,3	3,4	10,4	2,7	834,0	151,0	246,0
10.	Київська	59,3	29,5	9,0	239,5	3,4	74,1	101,6	1320,0	333,0	1187,0
11.	Кіровоградська	26,9	12,3	2,0	2,3	1,1	1,4	4,3	434,0	325,0	256,0
12.	Луганська	19,1	19,6	5,0	33,5	3,7	7,3	34,3	356,0	159,0	576,0
13.	Львівська	22,0	3,4	1,9	55,3	2,4	46,5	41,8	1462,0	327,0	516,0
14.	Миколаївська	5,8	2,7	4,2	110,7	7,5	502,5	5,0	193,0	180,0	225,0
15.	Одеська	6,5	7,5	2,3	39,9	2,2	284,2	41,7	136,0	436,0	688,0
16.	Полтавська	34,7	17,4	7,8	77,9	22,0	42,4	97,5	768,0	344,0	557,0
17.	Рівненська	5,6	2,4	3,9	15,6	4,4	32,5	19,2	962,0	170,0	302,0
18.	Сумська	12,1	6,0	8,2	31,1	19,2	1,8	27,6	927,0	173,0	326,0
19.	Тернопільська	19,0	7,0	5,8	23,4	4,9	40,3	2,9	777,0	156,0	293,0
20.	Харківська	25,7	11,9	4,8	116,7	7,4	67,9	45,8	829,0	455,0	637,0
21.	Херсонська	4,0	0,8	2,2	11,6	9,1	121,5	10,8	239,0	437,0	220,0
22.	Хмельницька	25,7	5,8	8,1	16,3	8,9	165,7	7,8	998,0	169,0	182,0
23.	Черкаська	44,1	7,4	10,6	48,3	13,8	121,3	35,5	772,0	320,0	469,0
24.	Чернівецька	14,6	6,8	0,3	2,8	1,2	52,8	6,0	381,0	135,0	235,0

показник 2 – ковбасні вироби (тис. т);
показник 3 – тваринне масло (тис. т);
показник 4 – продукція з незбираного молока (тис. т);
показник 5 – жирні сири, включаючи бринзу (тис. т);
показник 6 – консерви (млн. умовних банок);
показник 7 – кондитерські вироби (тис. т);
показник 8 – картопля (тис. т);
показник 9 – овочі (тис. т);
показник 10 – яйця (млн. шт.).

Дані для обчислень візьмемо за 2003 р. в [10, 87-117].

Оскільки показники подані в різних одиницях виміру, потрібно провести їх стандартизацію (так зване z-перетворення). Вона приводить значення всіх перетворених показників до єдиного діапазону, а саме від –3 до +3 [11, с. 387]. Стандартизоване значення розраховується за формулою:

$$z_i = \frac{\left(x_i - \bar{x}\right)}{s},$$

де $x_i (i = \overline{1,n})$ – вхідне значення показника;

$\bar{x} = \frac{\left(\sum_{i=1}^{n} x_i\right)}{n}$ – його середнє;

$s = \sqrt{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 / n - 1}$ – його стандартне відхилення.

Отримані дані зобразимо в таблиці 2.

Наступним кроком є вибір міри відстані. Їх є сім:

1) квадрат евклідових відстаней, який обчислюється за формулою

$$відстань\,(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2;$$

2) Евклідові відстані, які розраховуються за формулою

$$відстань\,(x, y) = \left(\sum_{i=1}^{n} (x_i - y_i)^2\right)^{\frac{1}{2}};$$

3) відстань міських кварталів (Манхеттенська), яка визначається за формулою

$$відстань\,(x, y) = \sum_{i=1}^{n} |x_i - y_i|;$$

4) показник відстані Чебишева, який обчислюється за формулою

$$відстань\,(x, y) = max|x_i - y_i|;$$

5) степенева відстань, яка розраховується за формулою

$$відстань\,(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{\frac{1}{r}};$$

Таблиця 2

Матриця стандартизованих вхідних даних

№ з/п	Область України	Показник 1	Показник 2	Показник 3	Показник 4	Показник 5	Показник 6	Показник 7	Показник 8	Показник 9	Показник 10
1.	АР Крим	0,8822	-0,5996	-0,5871	-0,2087	-1,0823	0,3842	-0,7965	-1,4114	-1,1289	0,3429
2.	Вінницька	1,1806	-0,2820	3,6966	-0,0731	1,1955	0,8113	0,7326	1,3799	-0,2399	-0,1661
3.	Волинська	-0,4281	0,1555	-0,3214	-0,3444	-0,0642	0,2467	-0,6774	0,7244	-0,7890	-0,9528
4.	Дніпропетровська	2,1537	1,6729	-0,3879	2,0585	-0,6854	-0,6157	0,6265	-0,3412	2,0090	0,8751
5.	Донецька	0,3244	3,8114	-0,5207	0,8906	-0,9615	-0,6606	3,3535	-0,4150	1,2158	2,4639
6.	Житомирська	-0,1751	-0,5926	1,8868	-0,5540	0,5225	-0,5772	-0,1108	0,9437	-0,4926	-0,2240
7.	Закарпатська	-1,3363	-0,7337	-0,9856	-0,8482	-1,1686	0,0772	-0,6619	-0,5152	-0,4665	-0,6405
8.	Запорізька	-0,3957	0,1626	-0,2052	-0,3937	0,7814	-0,5497	-0,2324	-1,2279	-0,0830	0,3275
9.	Івано-Франківська	-0,5384	-0,6490	-0,8528	-0,6879	-0,5819	-0,7275	-0,8197	0,2285	-0,9633	-0,8217
10.	Київська	2,3742	1,2636	0,5087	3,2970	-0,5819	-0,1437	1,7390	1,3870	0,6230	2,8071
11.	Кіровоградська	0,2725	0,0497	-0,6535	-0,8817	-0,9788	-0,8100	-0,7783	-0,7249	0,5533	-0,7832
12.	Луганська	-0,2335	0,5649	-0,1554	-0,3320	-0,5301	-0,7559	-0,0022	-0,9109	-0,8936	0,4509
13.	Львівська	-0,0454	-0,5784	-0,6701	0,0520	-0,7545	-0,3967	0,1919	1,7255	0,5707	0,2195
14.	Миколаївська	-1,0963	-0,6278	-0,2882	1,0280	0,1256	3,7825	-0,7602	-1,2994	-0,7106	-0,9027
15.	Одеська	-1,0509	-0,2891	-0,6037	-0,2193	-0,7890	1,7818	0,1893	-1,4353	1,5208	0,8828
16.	Полтавська	0,7784	0,4096	0,3095	0,4501	2,6278	-0,4342	1,6330	0,0712	0,7189	0,3776
17.	Рівненська	-1,1093	-0,6490	-0,3380	-0,6474	-0,4093	-0,5250	-0,3928	0,5337	-0,7977	-0,6058
18.	Сумська	-0,6876	-0,3949	0,3759	-0,3743	2,1446	-0,8063	-0,1755	0,4502	-0,7716	-0,5132
19.	Тернопільська	-0,2400	-0,3244	-0,0226	-0,5100	-0,3230	-0,4535	-0,8146	0,0927	-0,9198	-0,6405
20.	Харківська	0,1946	0,0215	-0,1886	1,1337	0,1084	-0,2005	0,2954	0,2166	1,6864	0,6861
21.	Херсонська	-1,2131	-0,7619	-0,6203	-0,7178	0,4017	0,2907	-0,6102	-1,1897	1,5296	-0,9220
22.	Хмельницька	0,1946	-0,4091	0,3593	-0,6350	0,3672	0,6958	-0,6878	0,6195	-0,8064	-1,0685
23.	Черкаська	1,3882	-0,2961	0,7744	-0,0713	1,2128	0,2889	0,0289	0,0808	0,5097	0,0383

6) процент незгоди, який визначається за формулою

$$відстань\,(x, y) = \frac{кількість\_x_i \neq y_i}{i};$$

7) коефіцієнт кореляції Пірсона, який обчислюється за формулою

$$відстань\,(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\,s_n\,s_y},$$

де x_i, y_i – значення двох показників;

$\bar{x}, \bar{y}$ – їх середні значення;

s_x, s_y – їх стандартні відхилення;

n – кількість пар значень.

Будемо використовувати формулу евклідової відстані. Згідно з нею показник з більшим значенням домінує над показником з меншим значенням. Для вирішення цієї проблеми також слугує z-перетворення, яке ми провели на попередньому кроці.

Обравши міру відстані, отримаємо симетричну матрицю відстаней. Оскільки вона є надто громіздкою (розмірність 25×25), а відстані від неї будуть зображені на подальших рисунках, вважаємо за доцільне в межах цієї статті її не подавати.

Після цього потрібно вибрати правило об'єднання. Існує сім алгоритмів об'єднання об'єктів у кластери:

1) *одинарне об'єднання (метод ближнього сусіда)* – відстань між двома кластерами обчислюється як мінімальна відстань між усіма парами об'єктів, що їм належать;
2) *повне об'єднання (метод найбільш віддаленого сусіда)* – відстань між двома кластерами розраховується як максимальна відстань між усіма парами об'єктів, що їм належать;
3) *незважене попарне групове середнє* – відстань між двома кластерами визначається як середня відстань між усіма парами об'єктів, що їм належать;
4) *зважене попарне групове середнє* – аналогічний до попереднього, але кількість об'єктів у кластері використовується як ваговий коефіцієнт (чим більша кількість об'єктів, тим більшу вагу він має);
5) *незважений попарний груповий центроїд* – відстань між двома кластерами обчислюється як відстань між їх центроїдами (середніми значеннями показників);
6) *зважений попарний груповий центроїд (медіана)* – аналогічний до попереднього, але кількість об'єктів у кластері використовується як ваговий коефіцієнт;
7) *метод Варда (Уорда)* – мінімізує дисперсію двох кластерів, що об'єднуються на кожному кроці об'єднання.

Для нашого випадку будемо використовувати метод ближнього сусіда, оскільки він вважається єдиним математично коректним, бо результати не залежать від перестановляння рядків чи стовпчиків у матриці відстаней [12, с. 191]. Отримаємо, що на першому кроці алгоритму об'єднуються дві області – Івано-Франківська та Рівненська (відстань об'єднання – 1,006304), на другому до них приєднується Тернопільська (1,053864), на третьому – Волинська область (1,213349), на четвертому об'єднуються Закарпатська і Чернівецька області (1,267254) і т. д.

Дендрограму результатів кластерного аналізу зобразимо на рисунку 1. У ній горизонтальна вісь відображає відстань об'єднання, вертикальна – регіони України.

Графік списку об'єднання областей у кластери зобразимо на рисунку 2. На ньому горизонтальна вісь відображає крок об'єднання, вертикальна – відстань.

Як дендрограма, так і графік більш наочно відображають порядок об'єднання регіонів у кластери.

Провівши ієрархічну класифікацію областей України, можна визначити оптимальну кількість кластерів, у які їх доцільно об'єднати. Вона знаходиться за допомогою підбору, або аналізу списку об'єднання.

При аналізі списку об'єднання оптимальною вважається така кількість кластерів, яка дорівнює різниці кількості спостережень (у нашому випадку – 25) і кількості кроків, після якої відстань об'єднання збільшується скачкоподібно (у нашому випадку – 21, де відбувається скачок від 2,970425 до 3,558813, який добре видно на рисунку 2). Тобто за цією методикою рекомендоване число кластерів – 4.

Але при такій кількості в одному з кластерів буде знаходитись всього одна область (Донецька). Та й логічно робити поділ на 3 кластери: регіони з великим, середнім та малим випуском продукції. Тому проведемо поділ на 3 кластери, застосувавши при цьому метод К-середніх.

Вони матимуть таку структуру:

кластер 1 – Дніпропетровська, Донецька та Київська області (всього 3 об'єкти);

кластер 2 – Вінницька, Житомирська, Львівська, Полтавська, Сумська, Харківська, Черкаська та Чернігівська області (всього 8 об'єктів);

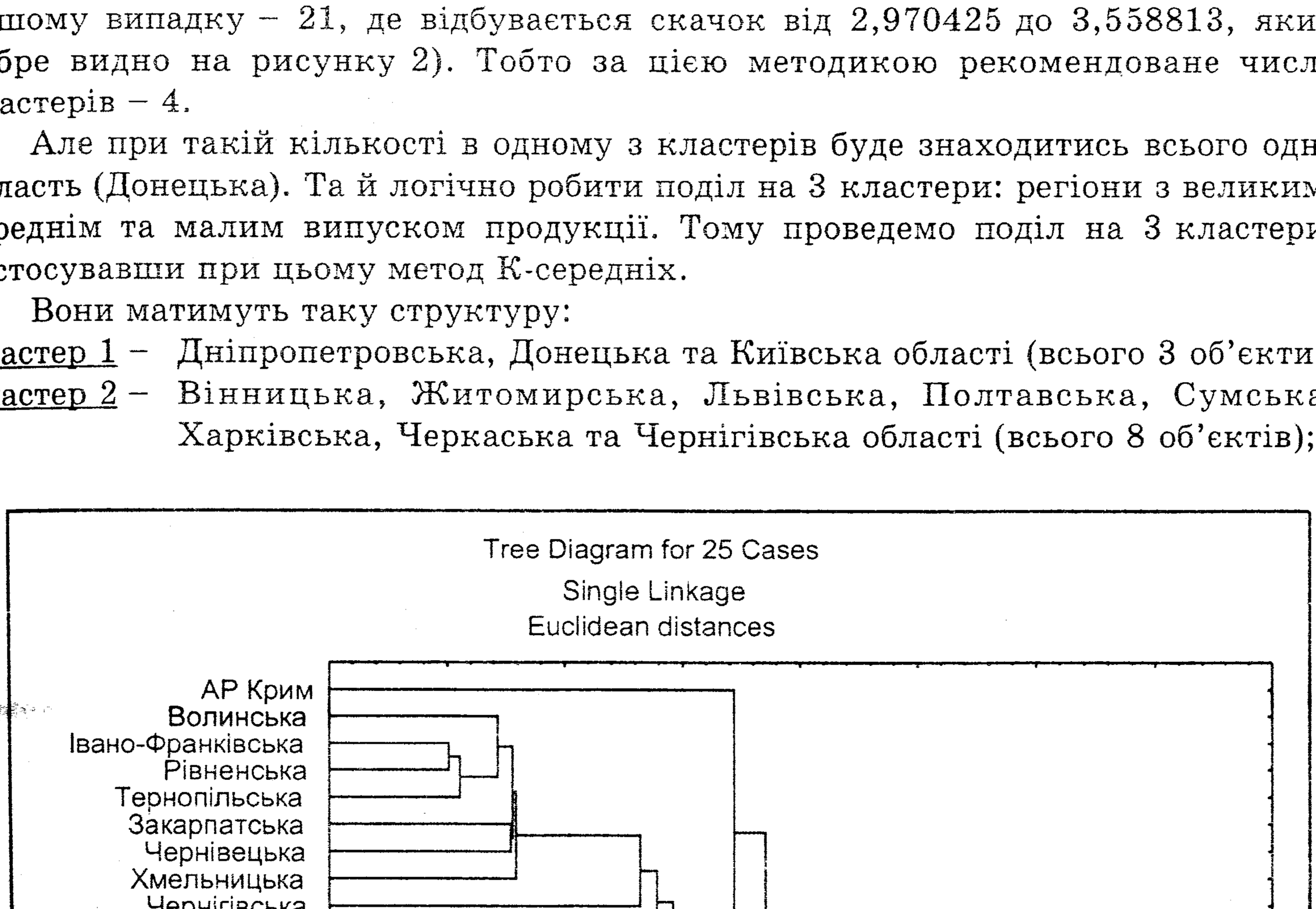

Рис. 1. Дендрограма результатів кластерного аналізу

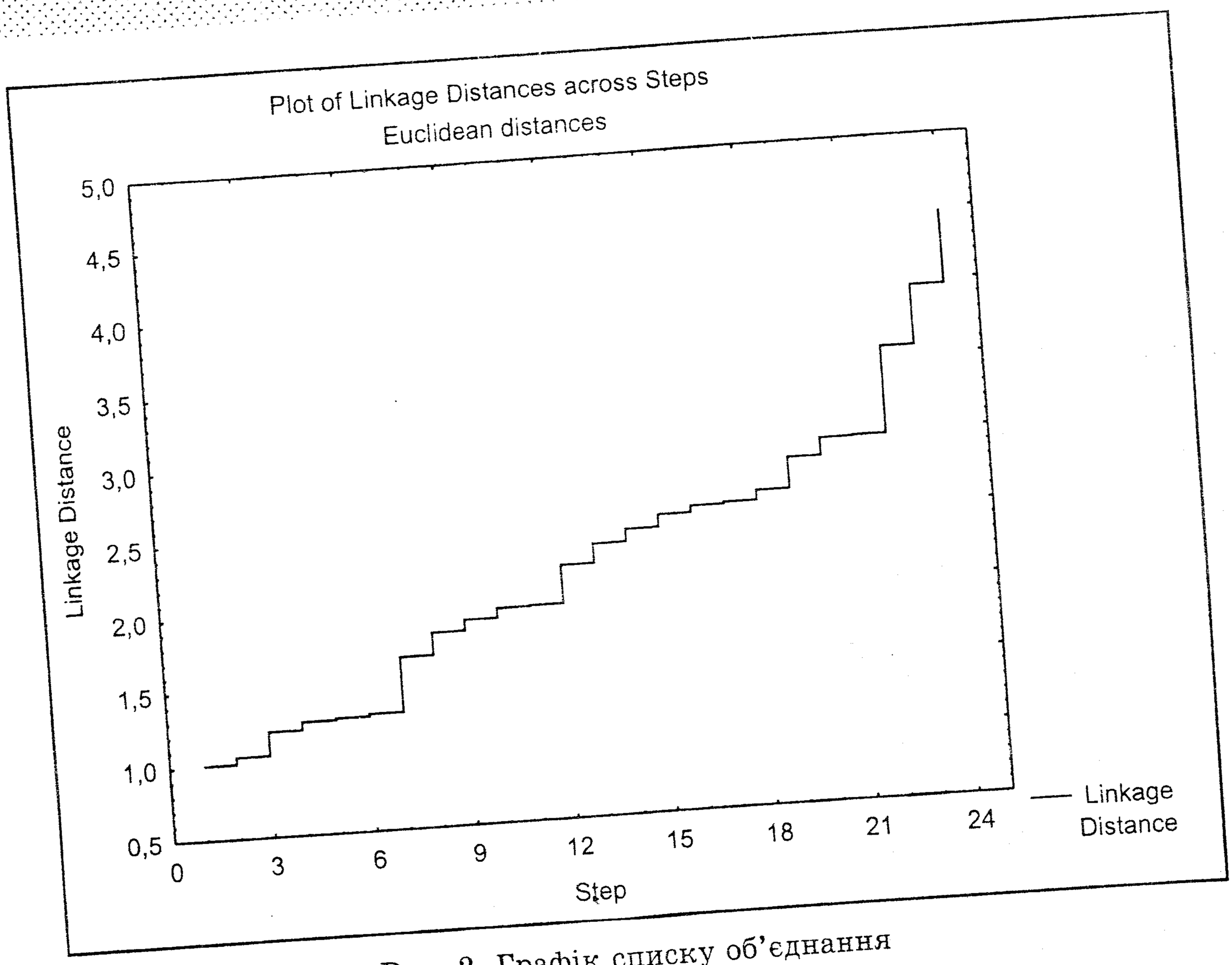

Рис. 2. Графік списку об'єднання

кластер 3 – АР Крим, Волинська, Закарпатська, Запорізька, Івано-Франківська, Кіровоградська, Луганська, Миколаївська, Одеська, Рівненська, Тернопільська, Херсонська, Хмельницька та Чернівецька області (всього 14 об'єктів).

Їхні середні значення зобразимо на рисунку 3.

З нього видно, що у першому кластері містяться об'єкти, в яких середні значення 6-ти показників з 10 є більшими, ніж у інших кластерах. У третьому кластері знаходяться об'єкти, в яких середні значення 9-ти показників є меншими, ніж у інших кластерах. У другому кластері містяться об'єкти, в яких середні значення показників є як більшими, так і меншими, ніж у інших кластерах.

Висновки. В результаті проведеного аналізу можна зробити наступні висновки:

1) області, які знаходяться в першому кластері, більше зосереджені на виробництві м'яса, ковбасних виробів, продукції з незбираного молока, кондитерських виробів, овочів та яєць;
2) області, які знаходяться у другому кластері, більше зосереджені на виробництві тваринного масла, жирних сирів (включаючи бринзу) та картоплі;
3) області, які знаходяться у третьому кластері, більше зосереджені на виробництві консервів.

Тобто, як бачимо, більше продуктів харчування випускають східні та північні області, менше – західні та південні. Волинь, на жаль, належить до третього кластера з невисоким виробництвом товарів харчової промисловості і спеціалізується переважно на випуску консервів.

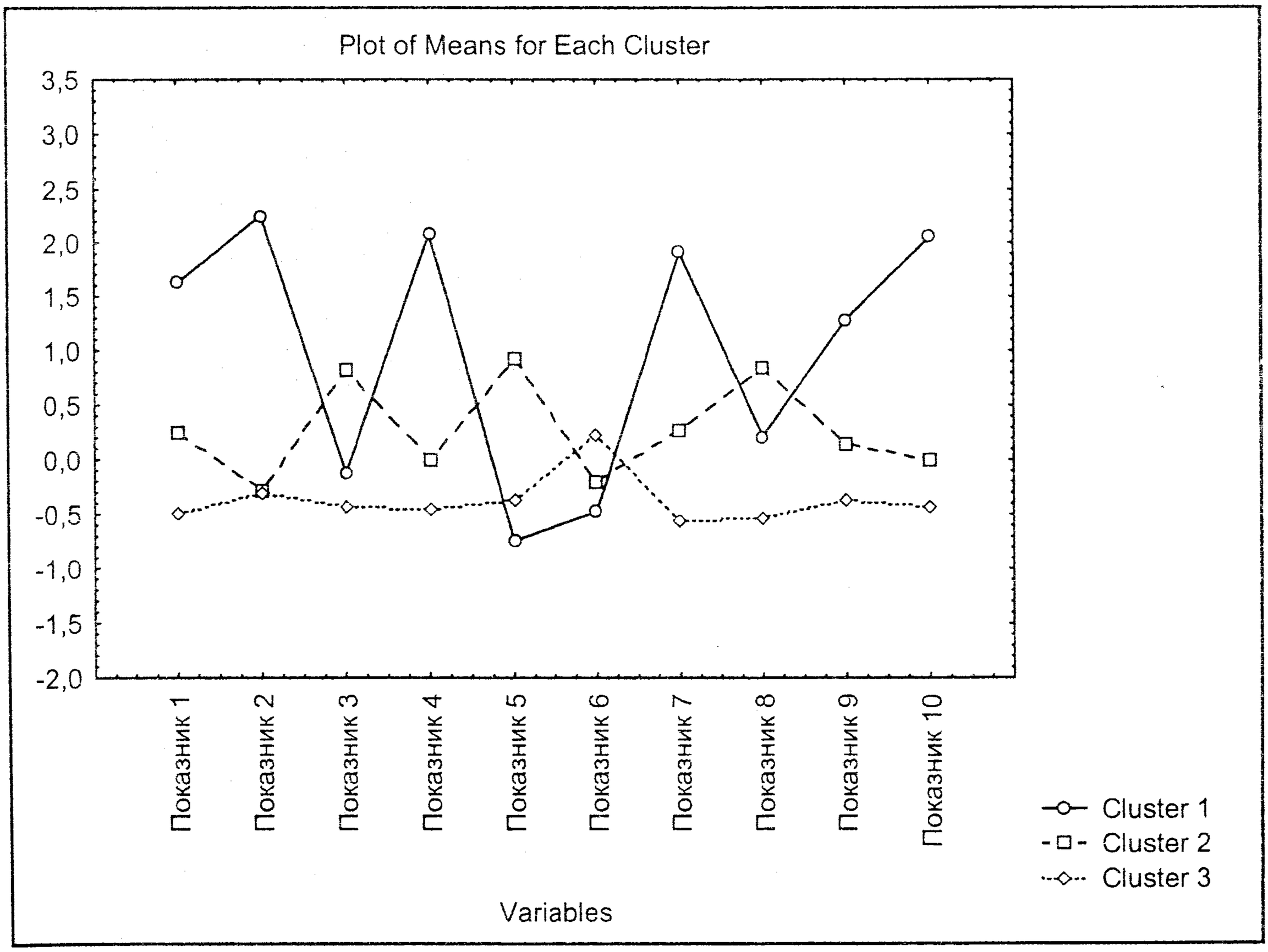

Рис. 3. Графік середніх значень показників для кожного кластера

Отримані дані можуть знадобитися при реформуванні АПК України, зокрема розміщенні та нарощуванні виробничих потужностей. Що ж стосується адміністративно-територіальної реформи, то, звичайно, в процесі її проведення потрібно враховувати значно більшу кількість показників, які характеризуватимуть не тільки харчову промисловість, а й усі галузі виробництва та невиробничу сферу. При цьому буде можливим об'єднання тих областей, які попадуть в один кластер.

Список використаних джерел

1. Єлейко В.І. Основи економетрії. – Львів: Марка Лтд, 1995. – 192 с.
2. Голиков А.П. Экономико-математическое моделирование мирохозяйственных процессов / Учеб. пособие. – Харьков: ХНУ, 2003. – 104 с.
3. Єлейко Я.І., Грищук Р.Т. Класифікація споживачів електроенергії у Львівській області за 2000 рік за допомогою кластерного аналізу // Регіональна економіка. – 2002. – №2. – С. 238-244.
4. Подольчак Н.Ю. Оцінка економічного ризику підприємства на основі кластерного аналізу // Регіональна економіка. – 2002. – №4. – С. 260-266.
5. Цапін А.О. Вивчення можливостей стратегічного управління на основі кластерної моделі розвитку // Наукові записки Національного ун-ту «Острозька академія»: Сер. «Економіка». – 2003. – №5. – С. 250-261.

6. Антонюк Л., Сацик В. Економетричні методи аналізу міжнародної конкурентоспроможності країн // Економіка України. – 2004. – №4. – С. 46-52.
7. Прикладная статистика. Основы эконометрики: Учебник для вузов: В 2 т. 2-е изд., испр. – Т. 1: Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. – М.: ЮНИТИ-ДАНА, 2001. – 656 с.
8. Грабауров В.А. Информационные технологии для менеджеров. – М.: Финансы и статистика, 2002. – 368 с.: ил. – (Прикладные информационные технологии).
9. Боровиков В. STATISTICA: искусство анализа данных на компьютере. Для профессионалов. – СПб.: Питер, 2001. – 656 с.: ил.
10. Україна у цифрах у 2003 році. Короткий статистичний довідник / За ред. О.Г. Осауленка. – К.: ТОВ «Видавництво «Консультант», 2004. – 271 с.
11. Бююль Ахим, Цёфель Петер. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей / Пер. с нем. – СПб.: ООО «ДиаСофтЮП», 2001. – 608 с.
12. Столяров Г.С., Ємшанов Д.Г., Ковтун Н.В. АРМ статистика: Навч. посібн. – К.: КНЕУ, 1999. – 268 с.