**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.

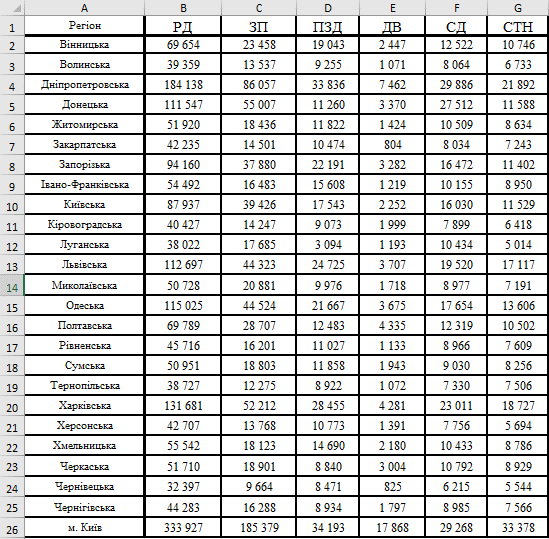


Рис. 1. Значення вхідних показників за 2016 рік

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 n1 більше n2, 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.

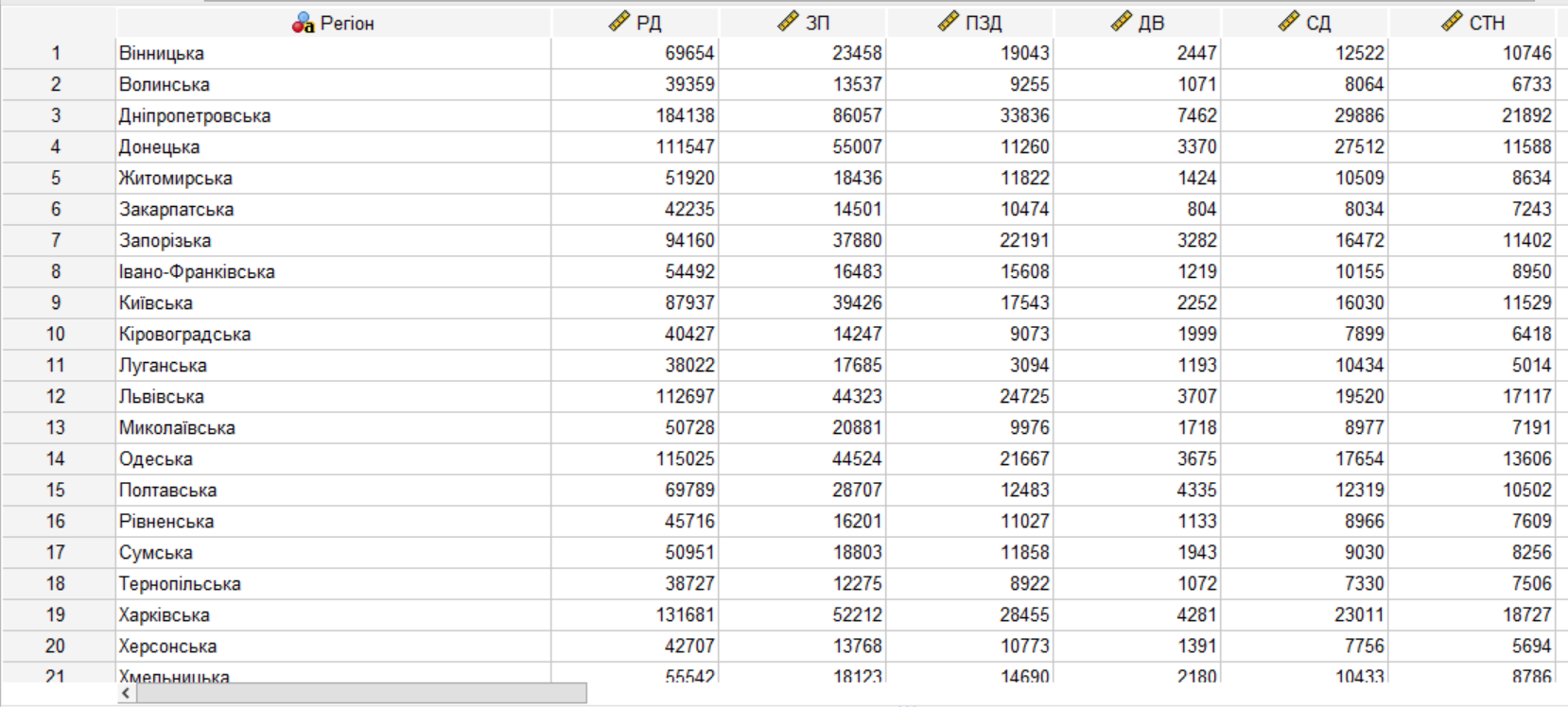


Рис. 1. Завантажені дані для кластеризації програмою SPSS

***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).

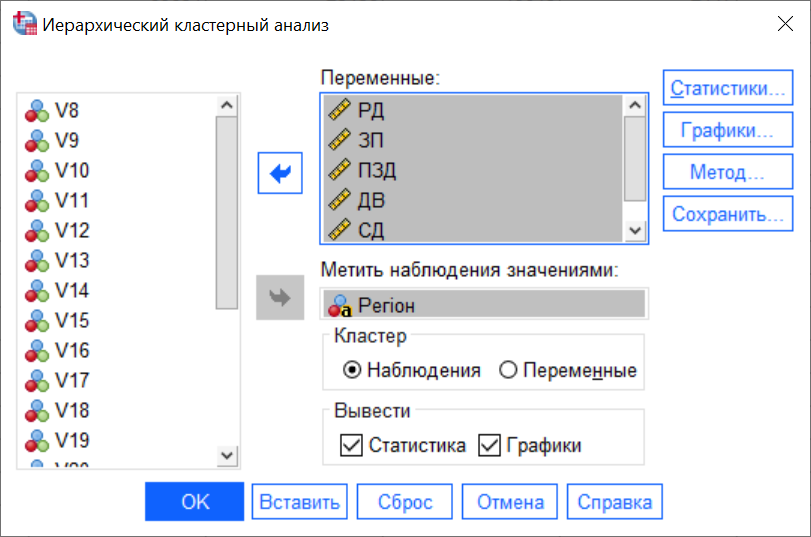


Рис. 2. Підготовка даних до кластеризації програмою SPSS

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.



Рис. 4. Результат кластеризації методом центроїдної класифікації

На рис. 5 наведена дендрограма кластеризації методом центроїдної класифікації.

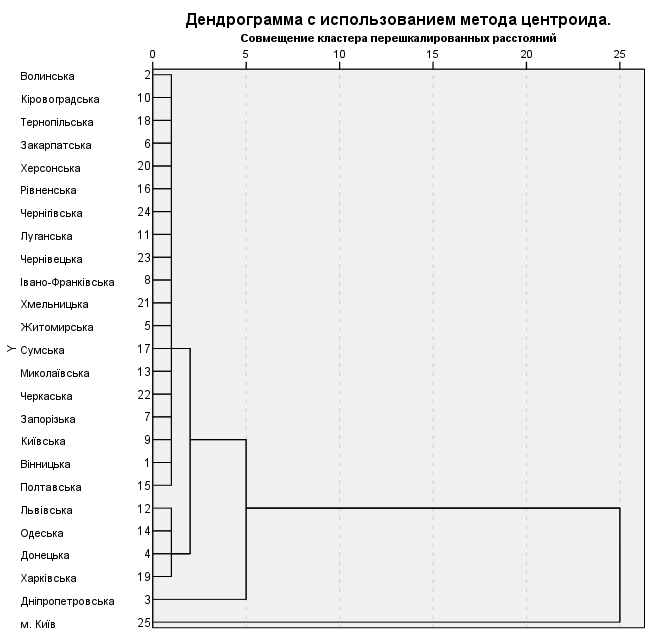


Рис. 5. Дендрограма кластеризації методом центроїдної класифікації без нормалізації

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.

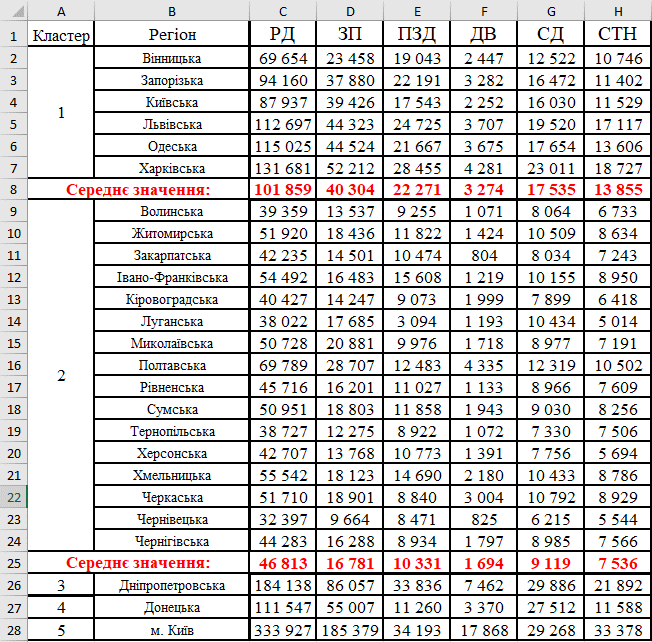


Рис. 6. Кластерний розподіл регіонів за методом центроїдної класифікації

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.