###### Міністерство освіти і науки України

###### Національний технічний університет України

###### “Київський політехнічний інститут ім. Ігоря Сікорського”

###### Кафедра автоматизації проектування енергетичних процесів та систем

# **Лабораторна робота №2**

з дисципліни : “Еколого-економічна оптимізація виробництва”

на тему :

“ Кластерний аналіз інструментами data mining (spss)”

Виконав студент 4 курсу

Групи ТМ-51

Задачин С.С.

Перевірила

Караєва Н. В.

Київ – 2018

КЛАСТЕРНИЙ АНАЛІЗ ІНСТРУМЕНТАМИ DATA MINING (SPSS)

**Задачі:**

Кластеризувати данні методом ближнього сусіда, циклоїдним методом, методом віддаленого сусіда та методом Варда за допомогою програмних засобів дата-майнінгу (SPSS).

Відстань між об’єктами за певною ознакою вимірюється як квадратний корінь модуля різниці квадратів значень ознаки об’єкта. При нормалізації відстань також зводиться до діапазону 0-1.

ОПИС МЕТОДІВ

**Метод ближнього сусіда**

Цей метод полягає в тому, що два об'єкти, які належать одній і тій самій групі (кластеру), 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.

Переваги методу:

* Простота використання отриманих результатів.
* Рішення не унікальні для конкретної ситуації, можливе їх використання для інших випадків.
* Метою пошуку є не гарантовано вірне рішення, а найкраще з можливих.

Недоліки методу:

* Даний метод не створює будь-яких моделей або правил, узагальнюючих попередній досвід, - у виборі рішення вони ґрунтуються на всьому масиві доступних історичних даних, тому неможливо сказати, на якій підставі будуються відповіді.
* Існує складність вибору заходи "близькості" (метрики). Від цього заходу головним чином залежить обсяг безлічі записів, які потрібно зберігати в пам'яті для досягнення задовільної класифікації або прогнозу. Також існує висока залежність результатів класифікації від обраної метрики.
* При використанні методу виникає необхідність повного перебору навчальної вибірки при розпізнаванні, наслідок цього - обчислювальна трудомісткість.
* Типові завдання даного методу - це завдання невеликої розмірності за кількістю класів і змінних.

**Центроїдний метод**

Відстань між двома кластерами визначається як евклідова відстань між центрами (середніми) 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.

Переваги:

* Менш залежний від викидів

Недоліком центроїдного методу є залежність центроїдних навантажень від шкали, в якій вимірюються змінні.

**Метод дальнього сусіда**

Кожен об'єкт розглядається як одноточковий кластер. Об'єкти групуються за наступним правилом: два кластери об'єднуються, якщо максимальна відстань між точками одного кластеру та точками іншого мінімальна. Процедура складається з n - 1 кроків і результатом є розбиття, які співпадають з можливими розбиттями в попередньому методі для будь-яких порогових значень.

Переваги

* Немає інформації про кількість необхідних кластерів.
* Простота реалізації та найкращий результат у деяких випадках.

Недоліки:

* Чутливість до шуму та викидів
* Розбиття великих кластерів
* Складність обробки різних розмірів кластерів та опуклої форми
* Іноді важко визначити правильну кількість кластерів за допомогою дендрограми.

**Метод Варда**

У цьому методі в якості цільової функції застосовують внутрішньо групову суму квадратів відхилень, що є не що інше, як сума 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 внутрішньо групової суми квадратів. Цей метод направлений на об'єднання близько розташованих кластерів.

Переваги:

* Використовує загальну суму квадратів у кластері

Недоліки:

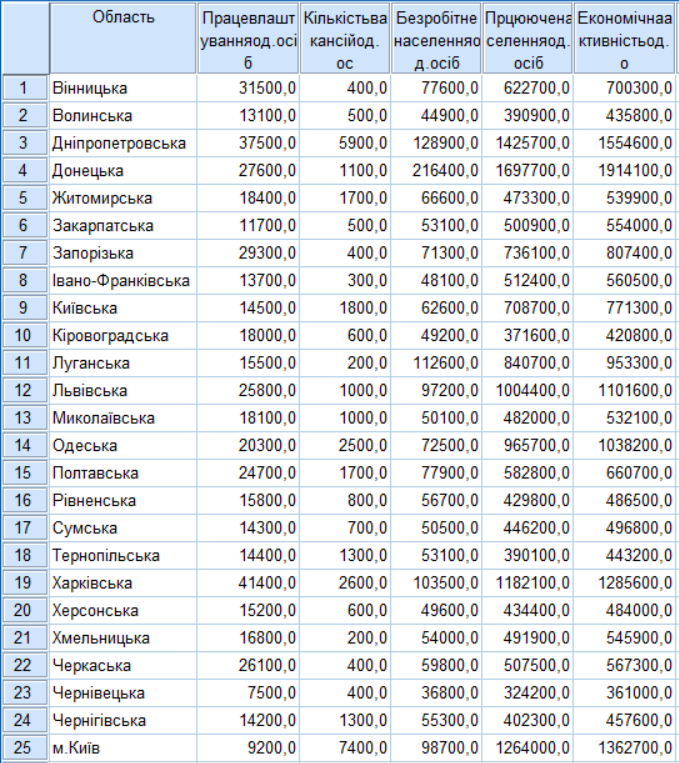
* добре працює, лише якщо з кожної популяції вимальовується однакова кількість об'єктів
* залежність від викидів

Згідно даних Держкомстату України було сформовано таблицю 1 із вхідними параметрами (показниками), які характеризують Працевлаштування населення, кількість вакансій, кількість непрацюючого населення, кількість людей, що працюють, економічну активність за 2014 рік.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **2014** | **Індикатори** | | | | |
| **Область** | **Працевлаштування (од.осіб)** | **Кількість вакансій (од.осіб)** | **Безробітне населення (од.осіб)** | **Прцююче населення (од.осіб)** | **Економічна активність (од.осіб)** |
| **Вінницька** | **31500** | **400** | **77600** | **622700** | **700300** |
| **Волинська** | **13100** | **500** | **44900** | **390900** | **435800** |
| **Дніпропетровська** | **37500** | **5900** | **128900** | **1425700** | **1554600** |
| **Донецька** | **27600** | **1100** | **216400** | **1697700** | **1914100** |
| **Житомирська** | **18400** | **1700** | **66600** | **473300** | **539900** |
| **Закарпатська** | **11700** | **500** | **53100** | **500900** | **554000** |
| **Запорізька** | **29300** | **400** | **71300** | **736100** | **807400** |
| **Івано-Франківська** | **13700** | **300** | **48100** | **512400** | **560500** |
| **Київська** | **14500** | **1800** | **62600** | **708700** | **771300** |
| **Кіровоградська** | **18000** | **600** | **49200** | **371600** | **420800** |
| **Луганська** | **15500** | **200** | **112600** | **840700** | **953300** |
| **Львівська** | **25800** | **1000** | **97200** | **1004400** | **1101600** |
| **Миколаївська** | **18100** | **1000** | **50100** | **482000** | **532100** |
| **Одеська** | **20300** | **2500** | **72500** | **965700** | **1038200** |
| **Полтавська** | **24700** | **1700** | **77900** | **582800** | **660700** |
| **Рівненська** | **15800** | **800** | **56700** | **429800** | **486500** |
| **Сумська** | **14300** | **700** | **50500** | **446200** | **496800** |
| **Тернопільська** | **14400** | **1300** | **53100** | **390100** | **443200** |
| **Харківська** | **41400** | **2600** | **103500** | **1182100** | **1285600** |
| **Херсонська** | **15200** | **600** | **49600** | **434400** | **484000** |
| **Хмельницька** | **16800** | **200** | **54000** | **491900** | **545900** |
| **Черкаська** | **26100** | **400** | **59800** | **507500** | **567300** |
| **Чернівецька** | **7500** | **400** | **36800** | **324200** | **361000** |
| **Чернігівська** | **14200** | **1300** | **55300** | **402300** | **457600** |
| **м.Київ** | **9200** | **7400** | **98700** | **1264000** | **1362700** |

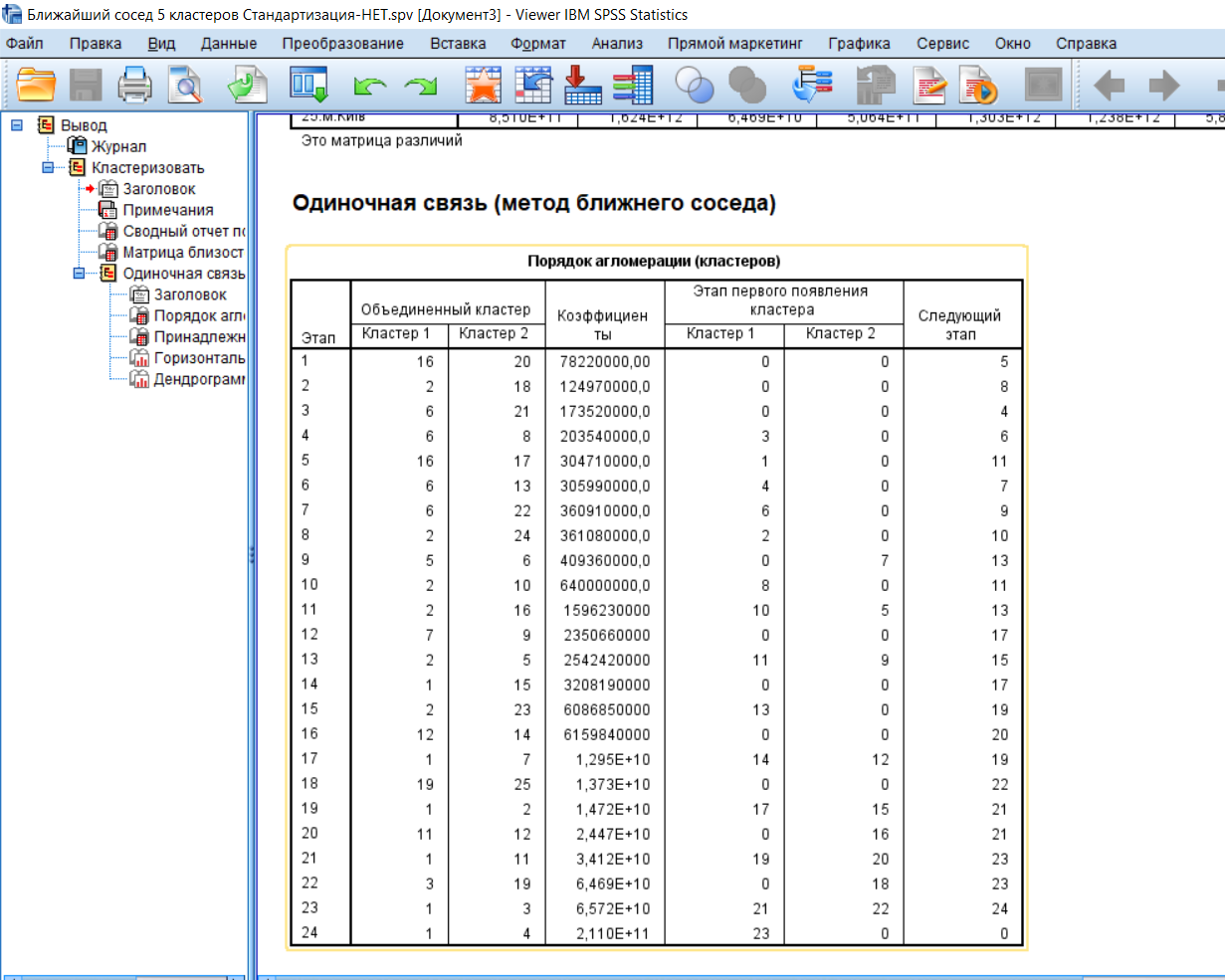
**Таблиця 1**. Значення окремих показників соціально-економічного розвитку регіонів України за 2014 рік

На рис.1 наведено вхідні дані для кластеризації засобами IBM SPSS Statistics 23.



**Рис. 1**. Вхідні дані для кластеризації

Після вибору цікавлячого нас методу та кількості кластерів (в даному випадку використовуємо 5 кластерів) було отримано вікно виводу результатів кластеризації. На рисунку 2 відображено приклад виводу результатів кластеризації методом найближчого сусіда.



**Рис. 2**. Вікно виводу результату кластеризації програмою SPSS методом найближчого сусіда

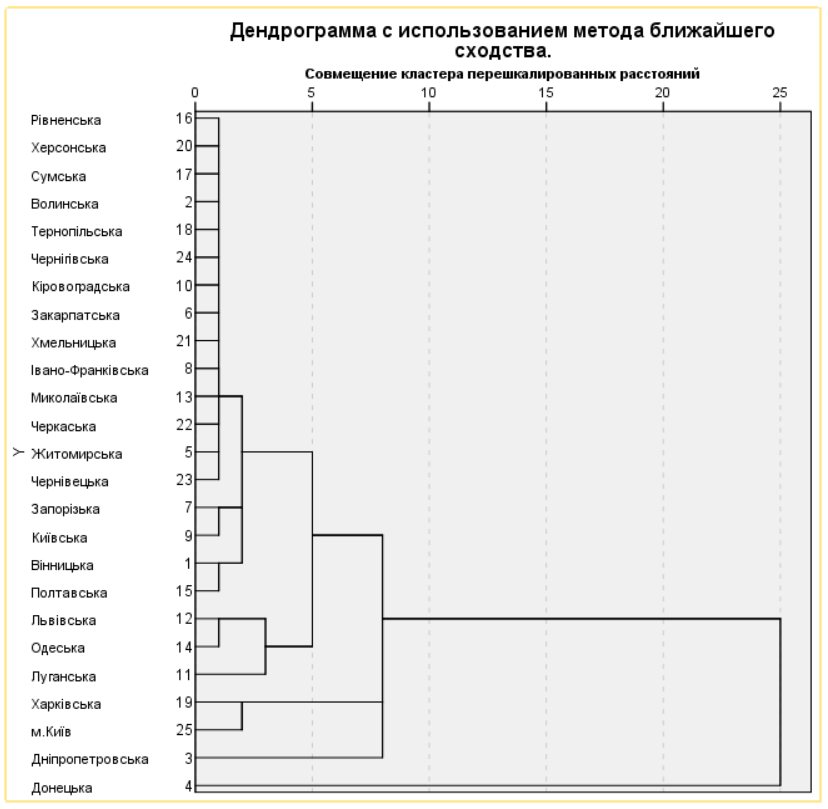
**Метод найближчого сусіда без використання нормалізації**

На рисунку 3 представлено табличний вигляд кластеризації. У таблиці перший стовпчик відповідає назві об’єкту, в другому вказується номер кластеру до якого належить об’єкт.



**Рис. 3.** Табличний вигляд результату кластеризації методом найближчого сусіда.

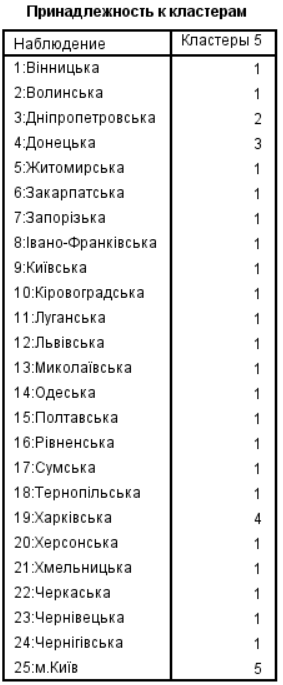
На рисунку 4 було представлено дендрограму кластеризації методом найближчого сусіда без використання нормалізації.



**Рис.4**. Дендрограма кластеризації методом найближчого сусіда без використання нормалізації

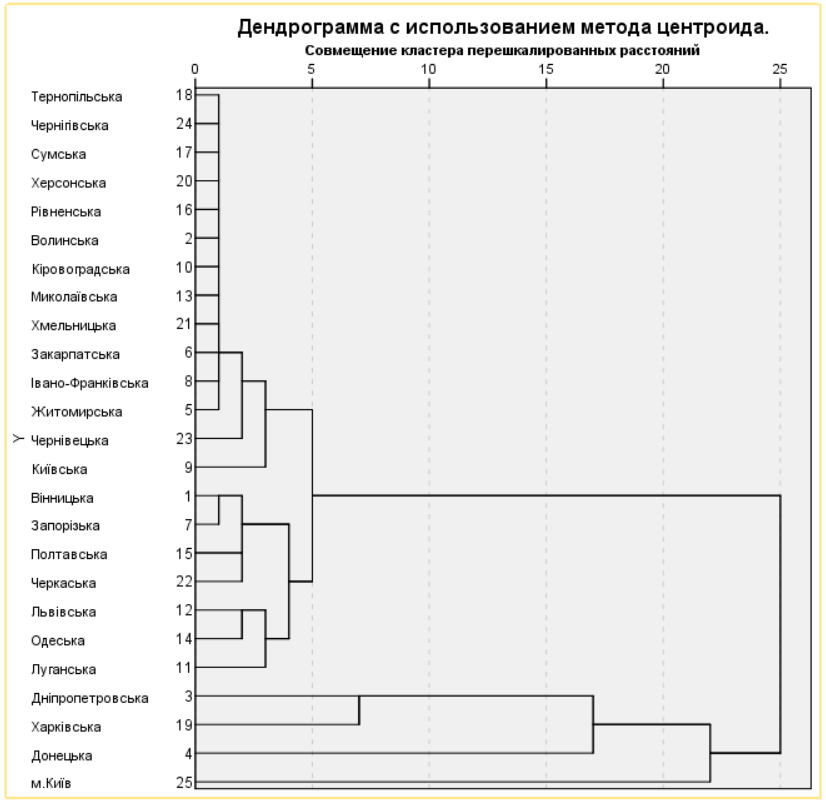
**Центроїдний метод з використанням нормалізації**

На рисунку 5 представлено табличний вигляд кластеризації. У таблиці перший стовпчик відповідає назві об’єкту, в другому вказується номер кластеру до якого належить об’єкт.

****

**Рис. 5**. Табличний вигляд результату кластеризації центроїдним методом

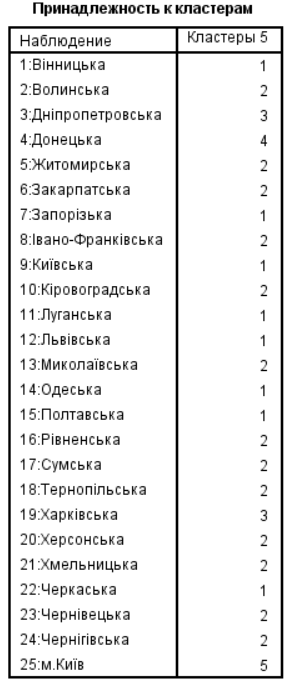
На риcунку 6 було представлено дендрограму кластеризації центроїдним методом з використанням нормалізації.



**Рис.6**. Дендрограма кластеризації центроїдним методом

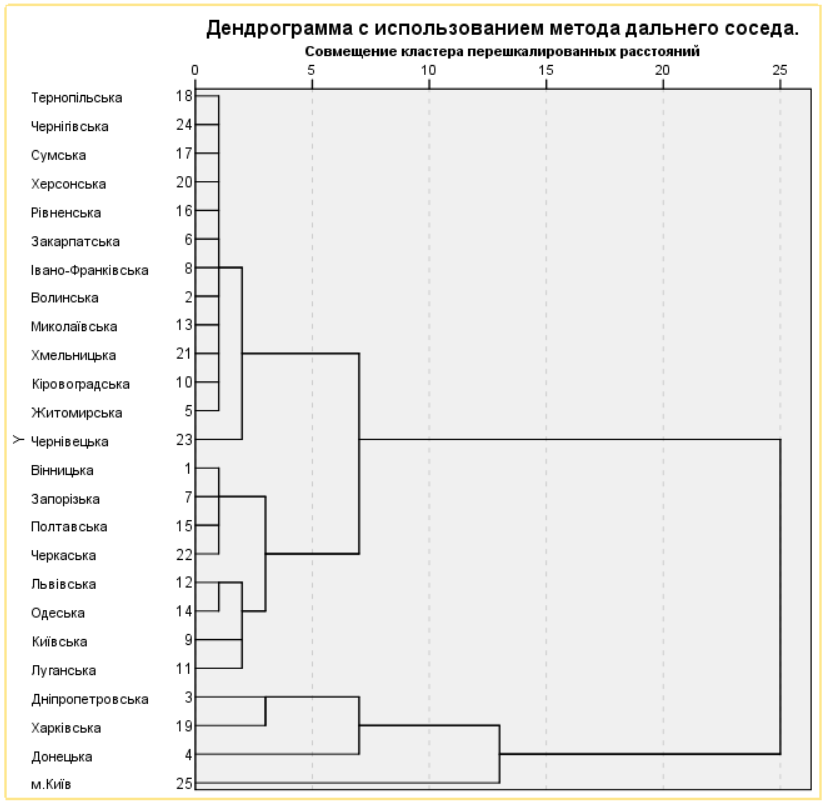
**Метод дальнього** **сусіда зі застосуванням нормалізації**

На риcунку 7 було відображено табличний вигляд кластеризації. У таблиці перший стовпчик відповідає назві об’єкту, в другому вказується номер кластеру до якого належить об’єкт.

****

**Рис. 7**. Табличний вигляд результату кластеризації методом дальнього сусіда

На рисунку 8 відображено дендрограму кластеризації методом дальнього сусіда зі застосуванням нормалізації.



**Рис. 8**. Дендрограма кластеризації методом дальнього сусіда

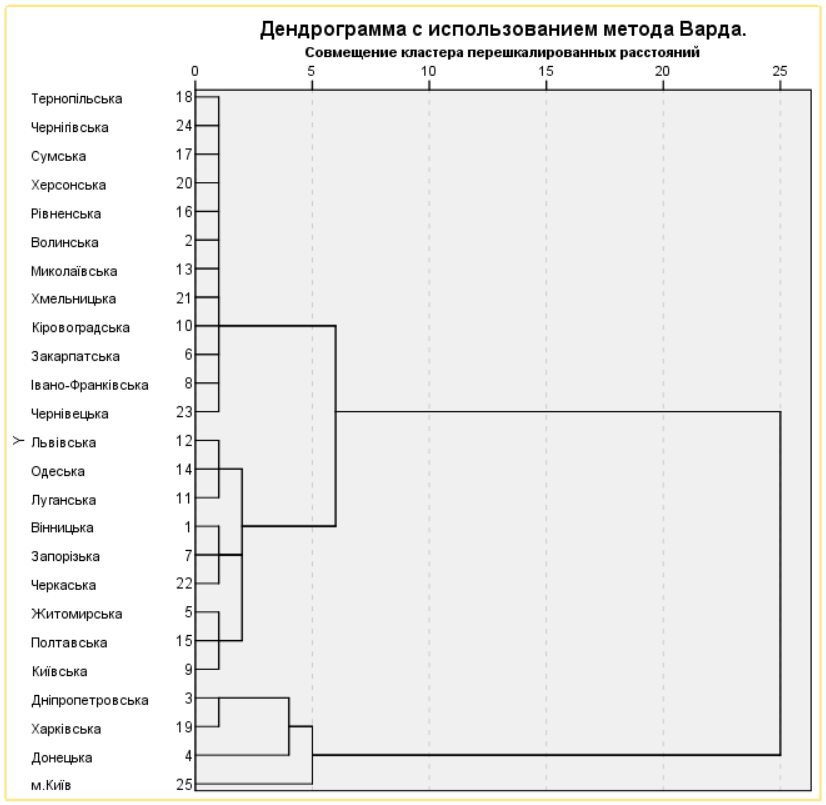
**Метод Варда**

На рисунку відображено табличний вигляд кластеризації. У таблиці перший стовпчик відповідає назві об’єкту, в другому вказується номер кластеру до якого належить об’єкт.

****

**Рис. 9**. Табличний вигляд результату кластеризації методом Варда

На рисунку 10 представлено дендрограму кластеризації методом Варда.



**Рис.10**. Дендрограма кластеризації методом Варда з нормалізацією

На таблиці 2 відображено розподіл областей за кластерами та методами.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **№ Кластеру** | **Методи** | | | |
| **Метод найближчого сусіда** | **Метод дальнього сусіда зі застосуванням нормалізації** | **Центроїдний метод зі застосуванням нормалізації** | **Метод Варда з нормалізацією** |
| **1** | **Вінницька, Волинська, Житомирська, Закарпатська, Запорізька, Івано-Франківська, Київська, Кіровоградська, Миколаївська, Полтавська, Рівненська, Сумська, Тернопільська, Херсонська, Хмельницька, Черкаська, Чернівецька, Чернігівська** | **Вінницька, Запорізька, Київська, Луганська, Львівська, Одеська, Полтавська, Черкаська** | **Вінницька, Волинська, Житомирська, Закарпатська, Запорізька, Івано-Франківська, Київська, Кіровоградська, Луганська, Львівська, Миколаївська, Одеська, Полтавська, Рівненська, Сумська, Тернопільська, Херсонська, Хмельницька, Черкаська, Чернівецька, Чернігівська** | **Вінницька, Житомирська, Запорізька, Київська, Луганська, Львівська, Одеська, Полтавська, Черкаська** |
| **2** | **Дніпропетровська** | **Волинська, Житомирська, Закарпатська, Івано-Франківська, Кіровоградська, Миколаївська, Рівненська, Сумська, Тернопільська, Херсонська, Хмельницька, Чернівецька, Чернігівська** | **Дніпропетровська** | **Волинська, Закарпатська, Івано-Франківська, Кіровоградська, Миколаївська, Рівненська, Сумська, Тернопільська, Херсонська, Хмельницька, Чернівецька, Чернігівська** |
| **3** | **Донецька** | **Дніпропетровська** | **Донецька** | **Дніпропетровська, Харківська** |
| **4** | **Луганська, Львівська, Одеська** | **Донецька** | **Харківська** | **Донецька** |
| **5** | **м. Київ** | **м. Київ** | **м. Київ** | **м. Київ** |

**Таблиця 2**. Результати кластерного розподілу регіонів з використанням різних методів кластерного аналізу

В таблиці 3 відображено кластерний розподіл областей України за кількістю працевлаштування за 2014 рік методом Варда.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№ Кластеру** | **Область** | **Працевлаштування (од.осіб)** | **Кількість вакансій (од.осіб)** | **Безробітне населення (од.осіб)** | **Прцююче населення (од.осіб)** | **Економічна активність (од.осіб)** |
| **1** | **Вінницька** | **31500** | **400** | **77600** | **622700** | **700300** |
| **Запорізька** | **29300** | **400** | **71300** | **736100** | **807400** |
| **Київська** | **14500** | **1800** | **62600** | **708700** | **771300** |
| **Полтавська** | **24700** | **1700** | **77900** | **582800** | **660700** |
| **2** | **Волинська** | **13100** | **500** | **44900** | **390900** | **435800** |
| **Житомирська** | **18400** | **1700** | **66600** | **473300** | **539900** |
| **Закарпатська** | **11700** | **500** | **53100** | **500900** | **554000** |
| **Івано-Франківська** | **13700** | **300** | **48100** | **512400** | **560500** |
| **Кіровоградська** | **18000** | **600** | **49200** | **371600** | **420800** |
| **Миколаївська** | **18100** | **1000** | **50100** | **482000** | **532100** |
| **Рівненська** | **15800** | **800** | **56700** | **429800** | **486500** |
| **Сумська** | **14300** | **700** | **50500** | **446200** | **496800** |
| **Тернопільська** | **14400** | **1300** | **53100** | **390100** | **443200** |
| **Херсонська** | **15200** | **600** | **49600** | **434400** | **484000** |
| **Хмельницька** | **16800** | **200** | **54000** | **491900** | **545900** |
| **Черкаська** | **26100** | **400** | **59800** | **507500** | **567300** |
| **Чернівецька** | **7500** | **400** | **36800** | **324200** | **361000** |
| **Чернігівська** | **14200** | **1300** | **55300** | **402300** | **457600** |
| **3** | **Дніпропетровська** | **37500** | **5900** | **128900** | **1425700** | **1554600** |
| **Харківська** | **41400** | **2600** | **103500** | **1182100** | **1285600** |
| **м.Київ** | **9200** | **7400** | **98700** | **1264000** | **1362700** |
| **4** | **Донецька** | **27600** | **1100** | **216400** | **1697700** | **1914100** |
| **5** | **Луганська** | **15500** | **200** | **112600** | **840700** | **953300** |
| **Львівська** | **25800** | **1000** | **97200** | **1004400** | **1101600** |
| **Одеська** | **20300** | **2500** | **72500** | **965700** | **1038200** |

**Таблиця 3**. Кластерний розподіл областей України за кількістю працевлаштування за 2014 рік методом Варда

**Характеристика кластерів:**

– Особливістю першого кластеру є параметр кількості вакансій, який значно відрізняється від інших;

– Особливістю 2 кластеру є найменша кількість економічно активного населення та найменша кількість вільних вакансій;

– Особливістю 3 кластеру є найбільша кількість населення, що працює;

– Особливістю 4 кластеру є найбільший серед усіх показник економічно активного населення;

– Особливістю 5 кластеру є незначна різниця у працевлаштуванні населення.

Оскільки метод Варда використано без нормалізації, то складно визначити характеристики кластерів. Для більш точного опису кластерів використаємо метод Варда з нормалізацією. Це відображено у таблиці 4.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№ Кластеру** | **Область** | **Працевлаштування (од.осіб)** | **Кількість вакансій (од.осіб)** | **Безробітне населення (од.осіб)** | **Прцююче населення (од.осіб)** | **Економічна активність (од.осіб)** |
| **1** | **Вінницька** | **31500** | **400** | **77600** | **622700** | **700300** |
| **Житомирська** | **18400** | **1700** | **66600** | **473300** | **539900** |
| **Запорізька** | **29300** | **400** | **71300** | **736100** | **807400** |
| **Київська** | **14500** | **1800** | **62600** | **708700** | **771300** |
| **Луганська** | **15500** | **200** | **112600** | **840700** | **953300** |
| **Львівська** | **25800** | **1000** | **97200** | **1004400** | **1101600** |
| **Одеська** | **20300** | **2500** | **72500** | **965700** | **1038200** |
| **Полтавська** | **24700** | **1700** | **77900** | **582800** | **660700** |
| **Черкаська** | **26100** | **400** | **59800** | **507500** | **567300** |
| **2** | **Волинська** | **13100** | **500** | **44900** | **390900** | **435800** |
| **Закарпатська** | **11700** | **500** | **53100** | **500900** | **554000** |
| **Івано-Франківська** | **13700** | **300** | **48100** | **512400** | **560500** |
| **Кіровоградська** | **18000** | **600** | **49200** | **371600** | **420800** |
| **Миколаївська** | **18100** | **1000** | **50100** | **482000** | **532100** |
| **Рівненська** | **15800** | **800** | **56700** | **429800** | **486500** |
| **Сумська** | **14300** | **700** | **50500** | **446200** | **496800** |
| **Тернопільська** | **14400** | **1300** | **53100** | **390100** | **443200** |
| **Херсонська** | **15200** | **600** | **49600** | **434400** | **484000** |
| **Хмельницька** | **16800** | **200** | **54000** | **491900** | **545900** |
| **Чернівецька** | **7500** | **400** | **36800** | **324200** | **361000** |
| **Чернігівська** | **14200** | **1300** | **55300** | **402300** | **457600** |
| **3** | **Дніпропетровська** | **37500** | **5900** | **128900** | **1425700** | **1554600** |
| **Харківська** | **41400** | **2600** | **103500** | **1182100** | **1285600** |
| **4** | **Донецька** | **27600** | **1100** | **216400** | **1697700** | **1914100** |
| **5** | **м.Київ** | **9200** | **7400** | **98700** | **1264000** | **1362700** |

**Таблиця 4**. Кластерний розподіл областей України за кількістю працевлаштування за 2014 рік методом Варда з нормалізацією

У таблиці 5 наведено кластерний розподіл областей України за кількістю працевлаштування за 2014 рік методом Варда з використанням нормалізації. У таблиці відображено середні значення.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№ Кластеру** | **Область** | **Працевлаштування (од.осіб)** | **Кількість вакансій (од.осіб)** | **Безробітне населення (од.осіб)** | **Прцююче населення (од.осіб)** | **Економічна активність (од.осіб)** |
| **1** | **Вінницька** | **22900** | **1122,2** | **77566,6** | **715766,6** | **793333,3** |
| **Житомирська** |
| **Запорізька** |
| **Київська** |
| **Луганська** |
| **Львівська** |
| **Одеська** |
| **Полтавська** |
| **Черкаська** |
| **2** | **Волинська** | **14400** | **683,3** | **50116,6** | **431391,6** | **48151,6** |
| **Закарпатська** |
| **Івано-Франківська** |
| **Кіровоградська** |
| **Миколаївська** |
| **Рівненська** |
| **Сумська** |
| **Тернопільська** |
| **Херсонська** |
| **Хмельницька** |
| **Чернівецька** |
| **Чернігівська** |
| **3** | **Дніпропетровська** | **39450** | **4250** | **116200** | **1303900** | **1420100** |
| **Харківська** |
| **4** | **Донецька** | **27600** | **1100** | **216400** | **1697700** | **1914100** |
| **5** | **м.Київ** | **9200** | **7400** | **98700** | **1264000** | **1362700** |

**Таблиця 5.** Кластерний розподіл областей України за кількістю працевлаштування за 2014 рік методом Варда з використанням нормалізації. Середні значення.

**Характеристика кластерів:**

– Особливістю першого кластеру є найбільша різниця в кількості працюючого і непрацюючого населення;

– Особливістю 2 кластеру є найменша кількість вакансій;

– Особливістю 3 кластеру є найбільша кількість працевлаштування в рік;

– Особливістю 4 кластеру є найбільша кількість працюючого населення та економічно активного населення;

– Особливістю 5 кластеру є найменша кількість працевлаштування в рік.

**Висновок:** на основі отриманих даних ми визначили, що оптимально розділяти данні на п’ять кластерів. В перший кластер повинні входити Вінницька, Житомирська, Запорізька, Київська, Луганська, Львівська, Одеська, Полтавська, Черкаська. В другий – Волинська, Закарпатська, Івано-Франківська, Кіровоградська, Миколаївська, Рівненська, Сумська, Тернопільська, Херсонська, Хмельницька, Чернівецька, Чернігівська. В третій – Дніпропетровська, Харківська. В четвертий – Донецька. В п’ятий- місто Київ.

На основі аналізу чотирьох методів результати, що наведені у таблиці 2 свідчать, що метод дальнього сусіда із застосуванням нормалізації та метод Варда із застосуванням нормалізації дають дуже схожий результат поділу регіонів на кластери. Метод Варда та центроїдний метод здійснюють різний кластерний поділ регіонів у порівнянні з вище згаданими методами.