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.

*Суть методу найближчого сусіда*:

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

*Суть методу найвіддаленішого сусіда зі застосуванням нормалізації*:

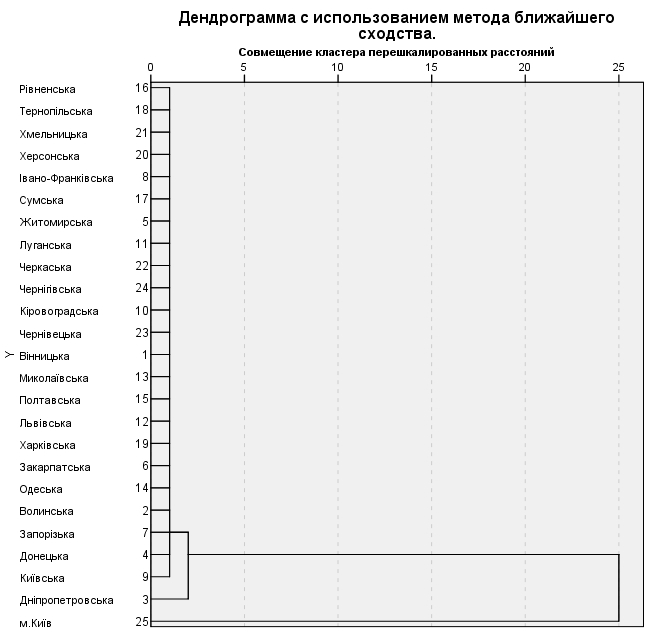
Кожен об'єкт розглядається як одноточковий кластер. Об'єкти групуються за наступним правилом: два кластери об'єднуються, якщо максимальна відстань між точками одного кластеру та точками іншого мінімальна. Процедура складається з n - 1 кроків і результатом є розбиття, які співпадають з можливими розбиттями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

*Суть методу Варда*:

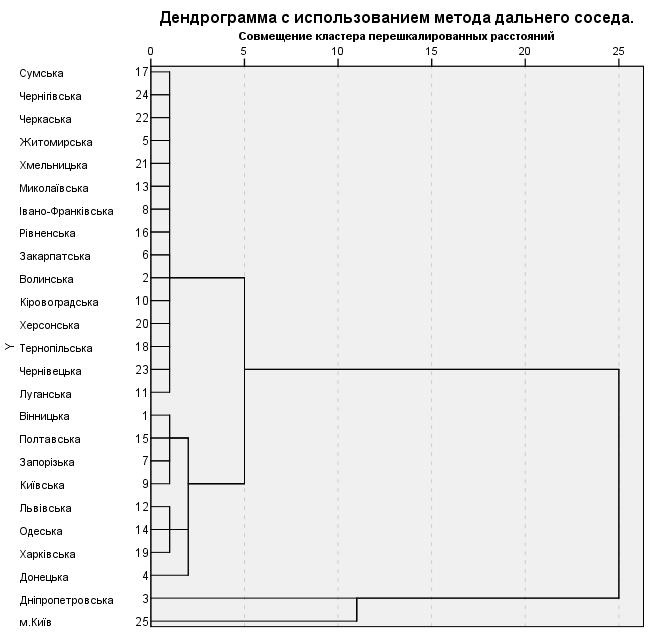
В якості цільової функції застосовуємо внутрішньогрупову суму квадратів відхилень, що є не що інше, як сума квадратів відстаней між кожною точкою

(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 внутрішньогрупової суми квадратів. Цей метод направлений на об'єднання близько розташованих кластерів.

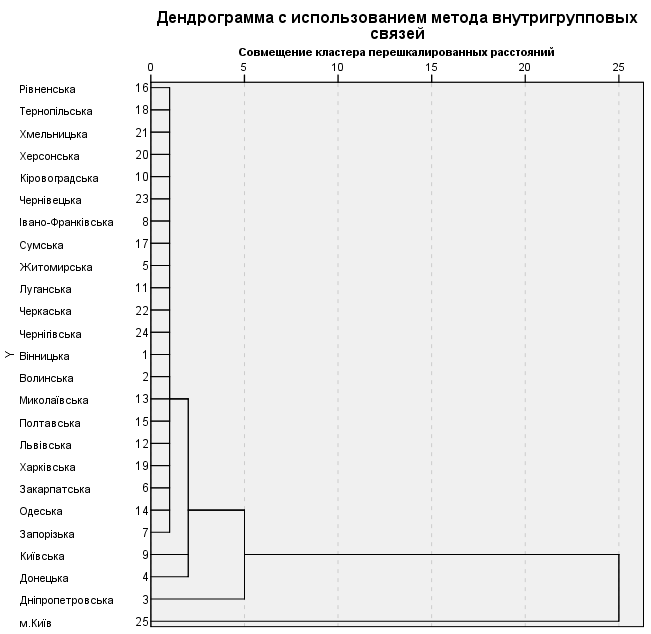
Дендрограмма кластеризації методом найближчого сусіда:



Дендрограмма кластеризації методом найвіддаленішого сусіда зі застосуванням нормалізації:



Дендрограмма кластеризації методом Варла:



Висновок: На основі отриманих даних ми визначили, що оптимально розділяти данні на чотири кластери. В перший кластер повинні входити м. Київ. В другий – Дніпропетровська. В третій – Вінницька, Полтавська, Запорізька, Київська, Львівська, Одеська, Харківська, Донецька. В четвертий – всі інші.

Таблиця 2. Результати кластерного розподілу регіонів з використання різних методів кластерного аналізу

|  |  |  |  |
| --- | --- | --- | --- |
| № | Методи | | |
| Найближчого сусіда | Варда | Найдальшого сусіда, нормалізований |
| 1 | м. Київ | м. Київ | м. Київ |
| 2 | Дніпропетровська | Дніпропетровська | Дніпропетровська |
| 3 | Донецька | Київська | Вінницька, Полтавська, Запорізька, Київська, Львівська, Одеська, Харківська, Донецька |
| 4 | Інші | Інші | Інші |

Таблиця 3. Кластерний розподіл регіонів України за рівнем соціально-демографічного розвитку, 2015 р. За методом найдальшого сусіда з нормалізацією.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № кластеру | Кількість живонароджених (осіб) | Кількість померлих (осіб) | Середня заробітна плата (грн) | Експорт (тис.дол. США) | Імпорт (тис.дол. США) | Обсяг виконаних будівельних робіт (млн.грн) |
| 1 | 36570 | 30666 | 8586 | 8558160 | 16115091 | 91794 |
| 2 | 31000 | 53127 | 5084 | 5859668 | 3431922 | 34241 |
| 3 | 159359 | 250246 | 4888,225313 | 13616883,8 | 10391720 | 174705,3 |
| 4 | 170110 | 249592 | 4163,057611 | 7991507,9 | 6470310 | 82654,2 |

Характеристика клfстерів:

– особливістю першого кластеру є найбільше значення рівню експорту та імпорту і найменша кількість померлих осіб;

– Особливістю 2 кластеру є найменша кількість живонароджених, найменший експорт, імпорт та обсяг виконаних будівельних робіт.

– особливістю 3 кластеру є найбільша кількість померлих і найбільший обсяг виконаних будівельних робіт;

– особливістю 4 кластеру є найбільша кількість живонароджених, найменша середня заробітна плата та обсяг виконаних будівельних робіт.

**Задачі:**

Кластеризувати данні методом ближнього сусіда, віддаленого сусіда та методом

Варла 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.

Суть методу найближчого сусіда:  
Метод полягає в тому, що два об'єкти, які належать одній ітій самій групі (кластері), мають коефіцієнт подібності, якийменше деякого порогового значення S. В термінах евклідовоївідстані d це означає, що відстань між двома точками

(об'єктами) кластеру не повинна перевищувати деякогопорогового значення h. Таким чином, h визначає максимальнодопустимий діаметр підмножини, що утворює кластер.

Суть методу найвіддаленішого сусіда зі застосуванням нормалізації:

Кожен об'єкт розглядається як одноточковий кластер. Об'єктигрупуються за наступним правилом: два кластериоб'єднуються, якщо максимальна відстань між точками одногокластеру та точками іншого мінімальна. Процедураскладається з n - 1 кроків і результатом є розбиття, якіспівпадають з можливими розбиттями в попередньому методідля будь-яких порогових значень.

Суть методу Варда:

В якості цільової функції застосовуємо внутрішньогрупову суму квадратів відхилень, що є не щоінше, як сума квадратів відстаней між кожною точкою

(об'єктом) і середньою по кластеру, який містить цейоб'єкт. На кожному кроці об'єднуються такі два кластери, якіпризводять до мінімального збільшення цільової функції,тобто внутрішньогрупової суми квадратів. Цей метод

направлений на об'єднання близько розташованих кластерів.

Таблиця 1 Дані економічних показників за 2016 рік

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Обл. | Експорт | Індекс споживчих цiн | Чисельнiсть н | Кiлькiсть п | Середня зп |
| Вінницька | 970358,3 | 101,7 | 1583,3 | 9442 | 5097 |
| Волинська | 602263,7 | 102,5 | 1038,2 | 5600 | 5028 |
| Дніпропетровська | 5859667,5 | 101,7 | 3227,1 | 27251 | 5905 |
| Донецька | 3428478,5 | 102,7 | 4231,2 | 11763 | 6503 |
| Житомирська | 461775,3 | 101,5 | 1241,2 | 6607 | 4871 |
| Закарпатська | 1165337,2 | 102,1 | 1255,9 | 6111 | 5191 |
| Запорізька | 2289912,3 | 102,2 | 1738,7 | 14423 | 5831 |
| Івано-Франківська | 545586,4 | 103,3 | 1377,2 | 7942 | 5218 |
| Київська | 1696595,1 | 102,3 | 1728,7 | 18164 | 6160 |
| Кіровоградська | 425137,2 | 102,5 | 959,3 | 7684 | 4919 |
| Луганська | 435547 | 102,5 | 2190,7 | 3354 | 5385 |
| Львівська | 1258257,6 | 102,1 | 2515,7 | 18628 | 5372 |
| Миколаївська | 1664779,6 | 102,1 | 1149,4 | 10569 | 5743 |
| Одеська | 1517770 | 102,4 | 2375,5 | 23883 | 5666 |
| Полтавська | 1434992,9 | 101,6 | 1419 | 10142 | 5466 |
| Рівненська | 310656 | 102,3 | 1161,7 | 5253 | 5214 |
| Сумська | 532107,1 | 102 | 1102,3 | 5842 | 5097 |
| Тернопільська | 287831,2 | 102,7 | 1055,9 | 5096 | 4641 |
| Харківська | 1020519,1 | 101,7 | 2685,6 | 24938 | 5263 |
| Херсонська | 245035,6 | 103,3 | 1054,2 | 8008 | 4867 |
| Хмельницька | 314372,4 | 102,1 | 1282,1 | 6992 | 5027 |
| Черкаська | 468563,2 | 102,3 | 1227,6 | 8296 | 5055 |
| Чернівецька | 105691,9 | 101,9 | 905 | 4054 | 4717 |
| Чернігівська | 426824,1 | 102,4 | 1024,9 | 6018 | 4758 |
| м.Київ | 8558160,2 | 101,8 | 2884,5 | 87380 | 9292 |

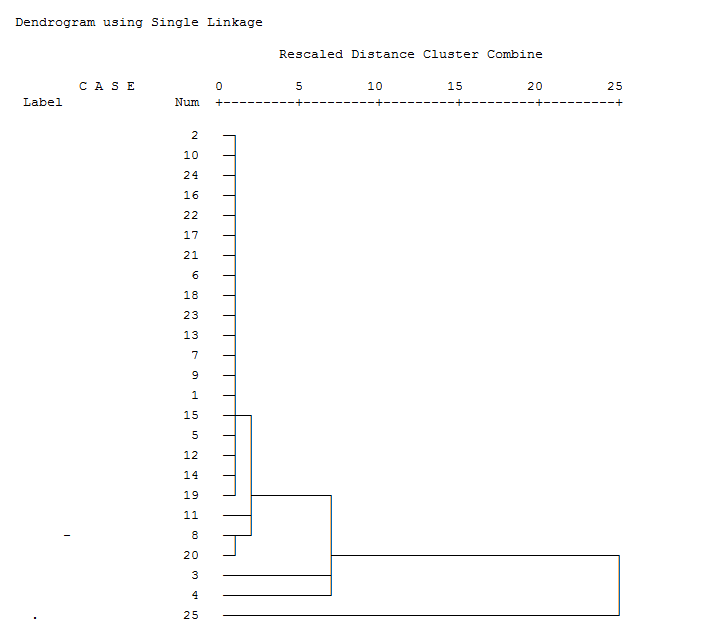


Рис 1. Розподілення областей на кластери методом Варда.

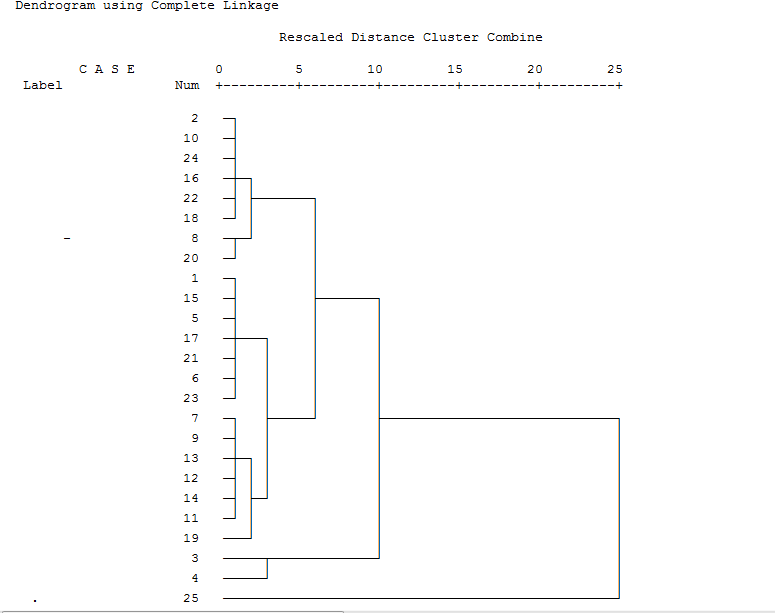
Таблиця 2 Результати розбиття областей на кластери 5 різними способами.



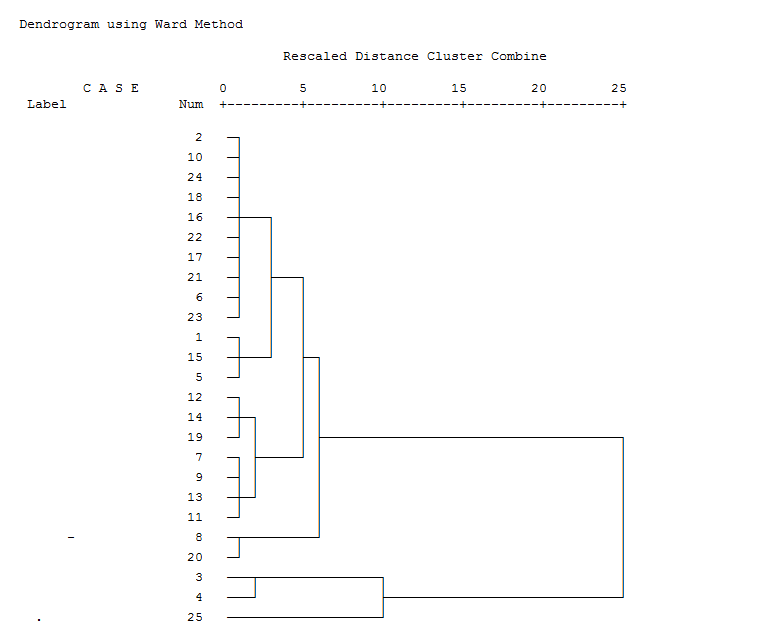
Дендрограмма кластеризації методом найближчого сусіда:



Дендрограмма кластеризації методом найвіддаленішого сусіда зі застосуванням нормалізації:



Дендрограмма кластеризації методом Варда:



Розподіл значень індикаторів по кластерам отриманим методом Варда

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Кластер |  | Експорт | Індекс споживчих цiн | Чисельнiсть н | Кiлькiсть п | Середня зп |
| 1 | м.Київ | 8558160,2 | 101,8 | 2884,5 | 87380 | 9292 |
| % по Україні | 0,237553658 | -0,428 | 0,068006762 | 0,254425809 | 3840,56 |
| 2 | Дніпропетровська | 5859667,5 | 101,7 | 3227,1 | 27251 | 5905 |
| Донецька | 3428478,5 | 102,7 | 4231,2 | 11763 | 6503 |
| % по Україні | 0,257816284 | -0,028 | 0,175841509 | 0,113597717 | 752,56 |
| 3 | Харківська | 1020519,1 | 101,7 | 2685,6 | 24938 | 5263 |
| Одеська | 1517770 | 102,4 | 2375,5 | 23883 | 5666 |
| Львівська | 1258257,6 | 102,1 | 2515,7 | 18628 | 5372 |
| Луганська | 435547 | 102,5 | 2190,7 | 3354 | 5385 |
| Миколаївська | 1664779,6 | 102,1 | 1149,4 | 10569 | 5743 |
| Київська | 1696595,1 | 102,3 | 1728,7 | 18164 | 6160 |
| Запорізька | 2289912,3 | 102,2 | 1738,7 | 14423 | 5831 |
| % по Україні | 0,274338547 | -0,042285714 | 0,339133182 | 0,331816329 | 179,98857 |
| 4 | Івано-Франківська | 545586,4 | 103,3 | 1377,2 | 7942 | 5218 |
| Херсонська | 245035,6 | 103,3 | 1054,2 | 8008 | 4867 |
| % по Україні | 0,021945739 | 1,072 | 0,057324195 | 0,046441882 | -408,94 |
| 5 | Вінницька | 970358,3 | 101,7 | 1583,3 | 9442 | 5097 |
| Волинська | 602263,7 | 102,5 | 1038,2 | 5600 | 5028 |
| Житомирська | 461775,3 | 101,5 | 1241,2 | 6607 | 4871 |
| Закарпатська | 1165337,2 | 102,1 | 1255,9 | 6111 | 5191 |
| Кіровоградська | 425137,2 | 102,5 | 959,3 | 7684 | 4919 |
| Полтавська | 1434992,9 | 101,6 | 1419 | 10142 | 5466 |
| Рівненська | 310656 | 102,3 | 1161,7 | 5253 | 5214 |
| Сумська | 532107,1 | 102 | 1102,3 | 5842 | 5097 |
| Тернопільська | 287831,2 | 102,7 | 1055,9 | 5096 | 4641 |
| Хмельницька | 314372,4 | 102,1 | 1282,1 | 6992 | 5027 |
| Черкаська | 468563,2 | 102,3 | 1227,6 | 8296 | 5055 |
| Чернівецька | 105691,9 | 101,9 | 905 | 4054 | 4717 |
| Чернігівська | 426824,1 | 102,4 | 1024,9 | 6018 | 4758 |
| % по Україні | 0,208345772 | -0,104923077 | 0,359694353 | 0,253718262 | -445,20923 |

Висновок: в ході виконання роботи я ознайомився з проведенням кластерного аналізу за допомогою програми SPSS. Всі 5 використаних методів показують приблизно однаковий результат кластеризації. Ресурси приблизно однаково розподілені по кластерах, але потрібно врахувати, що кількість міст в різних кластерах суттєво різниться. З цих даних можна зробити висновок. що найбільш розвиненими є Київська, Дніпропетровська, Донецька, Харківська області України.