

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ  
“КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ”

**ЕКОЛОГО-ЕКОНОМІЧНА ОПТИМІЗАЦІЯ ВИРОБНИЦТВА:  
методи та засоби кластерного аналізу**

МЕТОДИЧНІ ВКАЗІВКИ  
до виконання лабораторних робіт



Київ – 2016

Національний технічний університет України  
«Київський політехнічний інститут»

Теплоенергетичний факультет  
Кафедра автоматизації проектування енергетичних процесів і систем

**ЕКОЛОГО-ЕКОНОМІЧНА ОПТИМІЗАЦІЯ ВИРОБНИЦТВА:  
методи та засоби кластерного аналізу**

МЕТОДИЧНІ ВКАЗІВКИ  
до виконання лабораторних робіт

*для студентів напряму підготовки 6.050101 “Комп’ютерні науки”  
програм професійного спрямування  
„Комп’ютерний еколого-економічний моніторинг”*

*Рекомендовано Вченою радою теплоенергетичного факультету*

Київ – 2016

Еколого-економічна оптимізація виробництва: методи та засоби кластерного аналізу : методичні вказівки до виконання лабораторних робіт студентів із кредитного модуля «Еколого-економічна оптимізація виробництва» для студентів напряму підготовки 6.050101 «Комп'ютерні науки», програм професійного спрямування “Комп'ютерний еколого-економічний моніторинг” / Укладачі: Н. В. Караєва, І. А. Варава – К. : НТУУ «КПІ», 2016. – 36 с.

*Гриф надано Вченою радою  
теплоенергетичного факультету  
(протокол № 8 від 28 березня 2016 р.)*

Електронне навчальне видання

**ЕКОЛОГО-ЕКОНОМІЧНА ОПТИМІЗАЦІЯ ВИРОБНИЦТВА:  
методи та засоби кластерного аналізу**

**МЕТОДИЧНІ ВКАЗІВКИ  
до виконання лабораторних робіт**

*для студентів напряму підготовки 6.050101 “Комп'ютерні науки”  
програм професійного спрямування  
„Комп'ютерний еколого-економічний моніторинг”*

Укладачі:

*Караєва Наталія Веніамінівна, канд. екон. наук, доцент  
Варава Іван Андрійович, старший викладач*

Відповідальний  
редактор

*С. Г. Карпенко, канд. ф.-м. наук, доцент*

Рецензент

*В. Г. Писаренко, д-р. ф.-м. наук., професор*

*За редакцією укладачів*

© Караєва Н. В., Варава І. А., 2016

## ЗМІСТ

1. Мета і завдання кредитного модуля .....	5
2. Сутність і задачі кластерного аналізу.....	6
3. Класифікація і аналіз методів .....	11
4. Методична основа визначення оптимальної кількості кластерів та обчислення значень міри схожості між об'єктами.....	22
5. Кластерний аналіз інструментами Data Mining (SPSS) .....	27
Список використаної літератури .....	36

Методичні вказівки призначені для якісної організації самостійної роботи студентів при вивченні кредитного модуля, підвищення свідомості студентів у навчанні і поліпшення результатів навчання.



## **1. МЕТА І ЗАВДАННЯ КРЕДИТНОГО МОДУЛЯ**

Кредитний модуль „Еколого-економічна оптимізація виробництва” входить до варіативної частини програми циклу дисциплін за вибором вищого навчального закладу програми підготовки фахівців напряму 6.050101 “Комп’ютерні науки” програми професійного спрямування „Комп’ютерний еколого-економічний моніторинг”. При вивченні кредитного модуля приділяється увага методології економічної оцінки збитків від забруднення довкілля.

Головним завданням еколого-економічної оптимізації виробництва є виважене поєднання виробничих, природо відновних і соціальних функцій геосистем в інтересах досягнення належних просторово-екологічних умов життєдіяльності населення. Загалом, методи оптимізації або, як їх ще називають, методи теорії прийняття рішень є складовими математичних методів, які широко використовуються в екологічних і економічних дослідженнях. Основними методами оптимізації є: аналіз, прогнозування і моделювання, кластерний . Тому важливою складовою методології розробки стратегії планування вирішення еколого-економічних проблем України є прогнозування стану виробництва на всіх рівнях керування. Планування заходів й прогнозування результатів еколого-економічної оптимізації має базуватися на точних розрахунках потреб та їх ресурсному забезпеченні. Еколого-економічна оптимізація характеризується динамікою зміни початкових (вхідних) і похідних (розрахункових) показників ефективності функціонування суб’єктів господарювання (держави, галузей, окремих територій, підприємства) за рівнем використання наявних виробничих потужностей та природних ресурсів, конкурентоспроможності продукції й

обсягу її виробництва, рівня прибутку й платоспроможності, ефективності праці та рівня екологічної безпеки.

Крім того, використання комп'ютерної техніки в процесі обробки еколого-економічної інформації значно розширює можливості аналізу суб'єктів господарської діяльності: підвищує його оперативність, розширює базу для підготовки різних варіантів управлінських рішень, сприяє зростанню якості самого аналізу за рахунок проведення детальнішого аналізу, розширення факторних моделей прогнозування і використання (за необхідності) економічно-математичних прийомів.

У методичних рекомендаціях наведено класифікацію методів кластеризації об'єктів дослідження. Описано методи кластерного аналізу і визначено відмінності між ієрархічними і неієрархічними методами кластеризації. Розглянуто методична основа визначення оптимальної кількості кластерів та обчислення значень міри схожості між об'єктами. Наведено приклади кластерного аналізу інструментами Data Mining.



## **2. СУТНІСТЬ ТА ЗАДАЧІ КЛАСТЕРНОГО АНАЛІЗУ**

Розробка стратегій, планів, програм щодо нейтралізації викликів (або загроз) економічної безпеки України на всіх рівнях управління повинно спиратися на повне й адекватне відображення динаміки розвитку еколого-економічної системи у кожній визначеній одиниці адміністративно-територіального устрою (регіону) на фоні відповідних еколого-економічних індикаторів безпеки. Характерною особливістю аналізу стану регіонів за рівнем економічної безпеки є досить велика кількість еколого-економічних показників (індикаторів), які утворюють багатовимірні вектори. Часто індикатори виміряні в різних шкалах і це є проблемою при виборі алгоритму типізації (класифікації). В такому випадку доцільно використовувати методи багатовимірного, зокрема, кластерного аналізу.

Також, передумовою побудови математично-статистичних моделей еколого-економічного розвитку окремого регіону країни є виявлення однорідних сукупностей районів, представлених системою еколого-економічних показників. Ефективним методом, що дозволяє групувати райони в однорідні сукупності, використовуючи широке коло показників, є кластерний аналіз. Кластерний аналіз не виключає застосування інших методів угруповань у процесі типології районів, але є найбільше могутнім інструментом для проведення багатомірних досліджень.

Рішення цієї задачі безпосередньо пов'язане з необхідністю використання різноманітних методів і алгоритмів класифікації (тобто методів об'єднання кластерів). Принципово всі вони розрізняються алгоритмами обчислень. При цьому результати кластеризації, що одержуються при використанні різних методів кластеризації, можуть істотно відрізнятися один від одного.

Загалом, техніка кластеризації застосовується в дуже різноманітних сферах діяльності. Так наприклад, в медицині – кластеризації піддаються симптоми захворювання чи види лікування, в біології часто метою кластеризації є розбиття сукупності тварин на види і підвиди, у психології – класифікація видів поведінки, у педагогіці – таксономія виховних цілей тощо.

В економіці питанням розвитку та функціонування регіональних та промислових кластерів присвячені роботи вчених європейських країн, таких як: S. Czamanski і L. A. Ablas [1], E.E. Leamer [2], M. Porter [3], J. A. Toleno [4], V. P. Feldman і D. B. Audretsch [5] і багатьох інших. У вітчизняній практиці кластерний аналіз широко застосовуються в задачах групування регіонів та підприємств за показниками (індикаторами) тих чи інших складових економічної безпеки, зокрема: фінансової, інвестиційної, продовольчої, екологічної, соціальної, енергетичній [6-9] та ін.

При аналізі та прогнозуванні еколого-економічних явищ, особливо при їхній класифікації, дослідник досить часто має справу з багатомірністю

їхнього опису. Це відбувається, наприклад, при рішенні задач із сегментації ринку, при побудові типологічної схеми країн за умов необхідності використання досить великої кількості еколого-економічних показників (індикаторів), при прогнозуванні кон'юнктури ринку окремих товарів, при вивченні й прогнозуванні економічної депресії і багатьох інших явищ. У задачах еколого-економічного прогнозування досить перспективне поєднання кластерного аналізу з іншими кількісними методами (наприклад, з регресивним аналізом).

Основною задачею, що виникає в процесі кластеризації спостережень, є шлях неформального, економічного обґрунтування розбивки сукупності певних множин на класи. Як об'єкти можна розглядати райони окремої області України, представлені набором еколого-економічних показників. Для застосування кластерного аналізу не потрібно апріорних значень про розподіл генеральної сукупності. Кожен район представлений сукупністю параметрів, що спостерігаються, які можна інтерпретувати як координати точки в багатомірному просторі.

Термін "кластерний аналіз" походить від англ. *cluster* – "гроно, об'єднання, скупчення". *Кластерний аналіз* (англ. *Data clustering*) – задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини, що називаються кластерами, так, щоб кожен кластер складався з схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися.

*Метою кластерного аналізу* є класифікація об'єктів на відносно гомогенні (однорідні) групи, виходячи із досліджуваної кількості ознак (показників, змінних). Об'єкти в групі є відносно подібними з огляду на досліджуємі показники і відрізняються від об'єктів у інших групах. При використанні кластерного аналізу шляхом групування у меншу кількість кластерів знижується кількість об'єктів, а не кількість змінних.

*Використання кластерного аналізу в задачах еколого-економічної оптимізації* включає проведення аналітики за наступними критеріями:



- план агломерації – дозволяє отримати інформацію про еколого-економічні об'єкти які повинні бути об'єднані у кластери;
- кластерний центроїд – середнє значення змінних для всіх об'єктів у конкретному кластері;
- кластерний центр – початкові точки в кластеризації, навколо яких будують кластер;
- належність до кластерної групи – вказують до якого конкретного кластеру входить досліджуваний еколого-економічний об'єкт;
- деревовидна діаграма (дендрограма) – графічне зображення результатів кластеризації.

Методи кластерного аналізу можна застосовувати у всіляких випадках, навіть у тих, коли мова йде про просте групування, при якому все зводиться до утворення груп за кількісною ознакою (подібністю).

*Кластерний аналіз має одну суттєву особливість* – він не є звичайним статистичним методом, оскільки до нього у більшості випадків незастосовні процеси перевірки статистичної значимості. Тому кластеризація часто використовується, зокрема, при статистичному аналізі даних, векторній квантизації, розпізнаванні образів тощо.

*Перевага кластерного аналізу* в тім, що він дозволяє робити розбивку об'єктів не за одним параметром, а відразу за цілим набором ознак. Крім того, кластерний аналіз на відміну від більшості математико-статистичних методів не накладає ніяких обмежень на вигляд об'єктів і дозволяє розглядати безліч вихідних даних практично довільної природи. Це має велике значення, наприклад, для прогнозування кон'юнктури, коли показники мають різноманітний вигляд, що утруднює застосування традиційних економетричних підходів.

*Задачу кластеризації* можна сформулювати так: заданий набір з  $n$  векторів, кожен з яких має розмірність  $d$ ; необхідно розбити на підмножини відповідно до заданого критерію оптимізації. Як правило, таким критерієм є мінімізація спотворення. Існують різні шляхи оцінювання спотворення, але в

більшості прикладних реалізацій використовують суму середньоквадратичних Евклідових відстаней між вектором і центром кластера (центроїдом), до якого він належить.

Перш ніж перейти до безпосередньо методів та алгоритмів кластеризації, *необхідно зробити декілька зауважень, які слід враховувати, використовуючи кластерний аналіз:*

- використовуючи кластерний аналіз, дослідник має на меті виявлення структури даних. В той же час дія кластерного аналізу полягає у привнесенні структури у аналізовані дані. Тобто, кластеризація може призвести до появи артефактів (виявлення структури в даних, які її не мають);

- більшість методів кластерного аналізу є доволі таки простими евристичними процедурами, які, як правило, не мають статистичного обґрунтування;

- різні методи кластеризації можуть породжувати різні кластерні рішення для одних і тих же даних. Це звичне явище у більшості прикладних дослідженнях, у тому слід по-перше обирати найбільш осмислене рішення, по-друге – завжди вказувати, який саме метод кластеризації було використано;

- осмислене рішення при кластерному аналізі можна обрати лише тоді, коли є базис для його осмислення – теорія. Без теоретичної моделі, без гіпотези стосовно структури даних з'являється небезпека наївного емпіризму, коли результати кластеризації приймаються на істину у кінцевій інстанції. Як і будь-який інший метод, *кластерний аналіз має певні недоліки й обмеження*. Зокрема, склад та кількість кластерів залежить від вибраних критеріїв розбивки. При зведенні вихідного масиву даних до більш компактного вигляду можуть виникати певні перекручування, а також можуть губитися індивідуальні риси окремих об'єктів за рахунок заміни їхніми характеристиками узагальнених значень параметрів кластера. Також при проведенні класифікації об'єктів дуже часто ігнорується можливість відсутності в розглянутій сукупності яких-небудь значень кластерів.

Незалежно від постановки задачі дослідження проблем, *алгоритм застосування кластерного аналізу передбачає виконання основних шістьох етапів:*

- 1) Визначення множини характеристик, по яких будуть оцінюватися об'єкти у вибірці.
- 2) Вибір оптимальної кількості кластерів.
- 3) Обчислення значень тієї чи іншої міри схожості між об'єктами.
- 4) Обґрунтування застосування методів та алгоритмів кластерного аналізу для створення груп схожих об'єктів.
- 5) Перевірка достовірності результатів кластеризації.
- 6) Подання та інтерпретація отриманих результатів.

Якщо кластерному аналізу передує факторний аналіз, то вибірка не потребує коректування – викладені вимоги виконуються автоматично самою процедурою факторного моделювання. В іншому випадку вибірку потрібно коректувати.



### 3. КЛАСИФІКАЦІЯ ТА АНАЛІЗ МЕТОДІВ

Кластерний аналіз є набором різноманітних методів і алгоритмів класифікації (тобто методів об'єднання кластерів).

У сучасній практиці використовується безліч методів, що здійснюють кластеризацію об'єктів дослідження. Деякі методи можуть використовувати кілька альтернативних алгоритмів. Наведемо деякі приклади цих методів:

– метод *Custom Search Folders* (дозволяє звузити результати пошуку шляхом розподілу їх по "папках" (folders). Ця система вже реалізована в пошуковому сервері, що знаходиться за адресою [www.northernlight.com](http://www.northernlight.com), і має назву NorthernLight, – відповідно);

– *LSA/LSI – Latent Semantic Analysis/Indexing* (об'єднання кластерів відбувається шляхом факторного аналізу безлічі об'єктів виявляються латентні (приховані) чинники, які надалі є основою для утворення кластерів документів);

– *STC – Suffix Tree Clustering* (кластери утворюються у вузлах спеціального виду дерева – суффіксного дерева, яке будується із слів і фраз вхідних об'єктів);

– *Single Link – Complete Link, Group Average* ( ці методи розбивають безліч об'єктів на кластери, розташовані в деревовидній структурі, що отримується за допомогою ієрархічною агломеративної кластеризацією);

– *Scatter/Gather* (кластеризація представляється як ітеративний процес, що спочатку розбиває (*scatter*) безліч об'єктів на групи і представлення потім цих груп користувачеві (*gather*) для подальшого аналізу. Далі процес повторюється знову над конкретними групами);

– *K-means* (кластери представлені у вигляді центроїдів , що є "центром маси" усіх об'єктів);

– *SOM – Self-Organizing Maps* (методи класифікації об'єктів з використанням самоналагоджувальної нейронної мережі).

Загалом, наявні методи побудови кластерних моделей за способами обробки даних утворюють два основні типи (*ієрархічний і неієрархічний*) та сім груп методів:

- 1) ієрархічні агломеративні;
- 2) ієрархічні дивізивні;
- 3) ітеративні методи групування;
- 4) факторні;
- 5) методи згущень;
- 6) методи пошуку модальних значень щільності;
- 7) методи, що використовують теорію графів.

Кожний тип включає безліч підходів і алгоритмів.

Вищезазначені положення свідчать, що для розбивки об'єктів еколого-економічної оптимізації на кластери використовуються дві процедури, що використовуються на різних етапах кластерного аналізу.

*Перша – ієрархічна процедура* дозволяє поєднувати елементи кластерів на базі понять відстані чи подібності між точками в багатомірному просторі

ознак. Результатом такої розбивки є *дендрограма (дерево рішень)*, що показує етапи об'єднання об'єктів еколого-економічної оптимізації в групи за еколого-економічними характеристиками.

*Другий підхід для процедури розбивки на кластери – ітеративні (неієрархічні) методи угруповання.* Для застосування цього підходу необхідною є попередня розбивка даних на задане число кластерів і наступна робота з первинними даними. Ітеративна процедура починається з розбивки даних на задане число кластерів. Потім обчислюються центри ваги цих кластерів. Далі кожен об'єкт поміщають у той кластер, центр ваги якого є найближчим. Обчислюються нові центри ваги кластерів. Обчислення повторюються доти, поки кластери не стають стійкими, тобто перестануть змінюватися.

Значне місце в прикладних статистичних дослідженнях займає деревоподібна кластеризація, що полягає в об'єднанні об'єктів у досить великі кластери, використовуючи деяку міру подібності чи відстань між об'єктами. Типовим результатом цього виду кластеризації є ієрархічне дерево. Процедура починається з розгляду кожного об'єкта в класі. Далі поступово знижується поріг, що відноситься до рішення про об'єднання двох чи більше об'єктів в один кластер. Результатом обробки даних є зв'язування разом усе більшого числа об'єктів, що склалися з елементів, що дуже відрізняються. На останньому кроці обчислення всі об'єкти об'єднуються разом.

Авторська спроба побудови класифікації методів кластерного аналізу наведена на рис. 1 [8, 9].

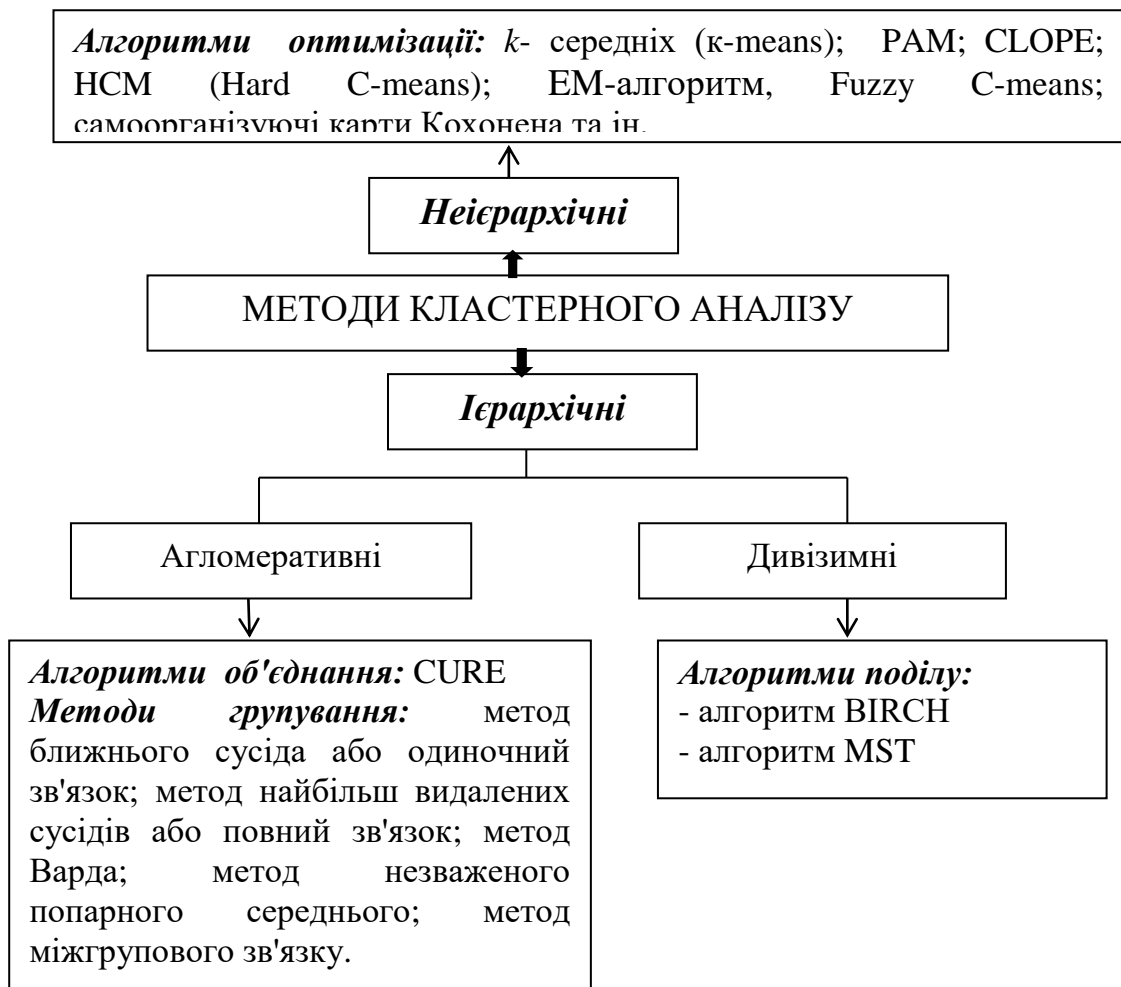


Рис. 1. Класифікація методів кластерного аналізу

У табл. 1 наведені відомості щодо призначення, переваг та недоліків основних алгоритмів кластерного аналізу, які наведено на рис. 1.

Таблиця 1

Призначення, переваги та недоліки основних алгоритмів кластерного аналізу

Алгоритми	Призначення	Переваги	Недоліки
<i>k</i> - середніх ( <i>k-means</i> )	Основна операція цього алгоритму відносно проста: задане фіксоване число (бажане або гіпотетичне) <i>k</i> кластерів, спостереження зіставляються кластерам так, що середні в кластері (для всіх змінних)	простота використання; швидкість використання; зрозумілість і прозорість алгоритму, алгоритм менш чутливий до викидів в порівнянні з <i>k-means</i> .	Необхідно задавати кількість кластерів; повільна робота на великих базах даних.

	максимально відрізняються один від одного.		
PAM ( <i>Partitioning around methods</i> )	Алгоритм аналогічний алгоритму $k$ -середніх. Алгоритм перерозподіляє об'єкти щодо медіани кластера, а не його центру.	Простота використання; швидкість використання; зрозумілість і прозорість алгоритму, алгоритм менш чутливий до викидів в порівнянні з $k$ -means.	Невеликий обсяг даних; необхідно задавати кількість кластерів; повільна робота на великих базах даних.
CURE	Застосовується для дуже великих наборів числових даних (бази даних), але ефективний лише для даних низької розмірності.	Виконує кластеризацію на високому рівні навіть при наявності викидів, виділяє кластери складної форми і різних розмірів, володіє лінійно залежними вимогами до місця зберігання даних і тимчасову складність для даних високої розмірності.	Необхідність у завданні порогових значень і кількості кластерів.
BIRCH	Використовується для кластеризації дуже великих наборів числових даних (бази даних).	Виконує двоступеневу кластеризацію, кластеризацію великих обсягів даних, працює на обмеженому обсязі пам'яті, є локальним алгоритмом, може працювати при одному скануванні вхідного набору даних, використовує той факт, що дані неоднаково розподілені по простору, і обробляє області з великою щільністю як єдиний кластер.	Робота з тільки числовими даними, добре виділяє тільки кластери сферичної форми, є необхідність у завданні порогових значень.
MST	Використовується для кластеризації великих наборів довільних даних	Виділяє кластери довільної форми, в т.ч. кластери опуклої й увігнутої форми, вибирає з кількох оптимальних рішень найоптимальніше.	Наявні недоліки для цього типу задач відсутні
Fuzzy C-means	Є нечітким алгоритмом кластеризації.	Нечіткість при визначенні об'єкта в кластер дозволяє визначати об'єкти, які	Обчислювальна складність, завдання кількості кластерів,

	Використовується для кластеризації великих наборів числових даних	знаходяться на кордоні, в кластери	виникає невизначеність з об'єктами, які віддалені від центрів всіх кластерів.
ЕМ-алгоритм (Expectation-Maximization)	Використовується в математичній статистиці для знаходження оцінок максимальної правдоподібності параметрів імовірнісних моделей, у разі, коли модель залежить від деяких прихованих змінних.	У ньому замість центрів кластерів передбачається наявність функції щільності ймовірності для кожного кластеру з відповідним значенням математичного сподівання і дисперсією.	Великий обсяг обчислень, що, не зважаючи на високу швидкість сучасних обчислювальних машин, розв'язання задачі залишається складним.
Самоорганізуючі карти Кохонена	дозволяє здійснювати проектування багатовимірних даних у простір меншої розмірності (зазвичай двовимірне) і на практиці застосовується при візуалізації даних для визначення наявності або відсутності кластерної структури в даних, кількості кластерів, залежності між змінним.	Візуальне подання інформації з карт Кохонена дозволяє формувати гіпотези, перевіряти їх та приймати обґрунтовані рішення. Візуалізація карт Кохонена піддається вербальному опису, точність якого достатня для використання реальних інструментів у реальні завданнях	Зменшення точності візуалізації багатовимірних даних.

Розглянемо ієрархічні і неієрархічні методи детально.

**Ієрархічні методи** характеризуються побудовою ієрархічної чи деревоподібної структури, коли відбувається послідовне угруповання чи поділ об'єктів щодо інших об'єктів. При ієрархічній кластеризації виконується послідовне об'єднання менших кластерів у великі чи поділ



великих кластерів на менші. Іншими словами, ієрархічні методи кластеризації розрізняються правилами побудови кластерів. У якості правила виступають критерії, що використовуються при вирішенні питання про "схожість" об'єктів при їх об'єднанні в групу (агломеративні методи) або розділення на групи (дивізімні методи) (див. рис. 1).

Група *ієрархічних агломеративних методів (Agglomerative Nesting, AGNES)* характеризується послідовним об'єднанням вихідних елементів і відповідним зменшенням числа кластерів. На початку роботи алгоритму всі об'єкти є окремими кластерами. На першому кроці найбільш схожі об'єкти об'єднуються в кластер. На подальших кроках роботи алгоритму об'єднання продовжується до тих пір, поки всі об'єкти не складатимуть один кластер. Послідовне об'єднання вихідних елементів і зменшенням числа кластерів здійснюється на основі алгоритму *CURE (Clustering Using Representatives)*. Цей алгоритм виконує ієрархічну кластеризацію з використанням набору визначальних точок для визначення об'єкта в кластер.

Ієрархічні агломеративні методи розрізняються переважно за правилами групування кластерів. Існує багато різних правил групування, кожне з яких породжує специфічний ієрархічний метод. Найпоширенішими з них є:

- метод ближнього сусіда або одиночний зв'язок;
- метод найбільш видалених сусідів або повний зв'язок;
- метод Варда;
- метод незваженого попарного середнього;
- метод між групового зв'язку.

*Ієрархічні дивізімні (ділимі) методи (Divisive Analysis, DIANA)* становлять логічну протилежність агломеративним методам. На початку роботи алгоритму всі об'єкти належать одному кластеру, який на подальших кроках ділиться на менші кластери, в результаті утворюється послідовність розщеплюючих груп.

Основними алгоритмами поділу кластерів є:

– алгоритм BIRCH (*Balanced Iterative Reducing and Clustering using Hierarchies*). У цьому алгоритмі передбачений двоетапний процес кластеризації;

– алгоритм MST (*Algorithm based on Minimum Spanning Trees*). Алгоритм мінімального покриває дерева спочатку будує на графі мінімальне покриває дерево, а потім послідовно видаляє ребра з найбільшою вагою. Короткий опис сутності основних методів кластеризації наведено в табл. 2.

Таблиця 2

Сутність методів кластерного аналізу

Назва методу	Сутність методу
Нейронні мережі Кохонена	Клас нейронних мереж, основним елементом яких є прошарок Кохонена. Прошарок Кохонена складається з адаптивних лінійних суматорів («лінійних формальних нейронів»). Як правило, вихідні сигнали прошарку Кохонена обробляються за правилом «переможець забирає все»: найбільший сигнал перетворюється в одиничний, решта звертаються в нуль. За способами налаштування вхідних суматорів і по розв'язуванню завданням розрізняють багато різновидів мереж Кохонена. Найбільш відомі з них: мережі векторного квантування сигналів, тісно пов'язані з найпростішим базовим алгоритмом кластеризації (метод динамічних ядер або К-середніх, тобто K-means) карт Кохонена (Self-Organising Maps, SOM), мережі векторного квантування, що навчаються з учителем (Learning Vector Quantization).
Метод Варда	У цьому методі в якості цільової функції застосовують 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.
ЕМ-алгоритм	Алгоритм, який використовується в математичній статистиці для знаходження оцінок максимальної правдоподібності параметрів імовірнісних моделей, у разі, коли модель залежить від деяких прихованих змінних. Кожна ітерація алгоритму

	<p>складається з двох кроків. На Е-кроку обчислюється очікуване значення функції правдоподібності, при цьому приховані змінні розглядаються як ті, що спостерігаються. На М-кроку обчислюється оцінка максимальної правдоподібності, таким чином збільшується очікувана правдоподібність, яка обчислюється на Е-кроку. Потім це значення використовується для Е-кроку на наступній ітерації. Алгоритм виконується до збіжності. Як правило, ЕМ-алгоритм застосовується для розв'язання задач двох типів. До першого типу можна віднести завдання, пов'язані з аналізом дійсно неповних даних, коли деякі статистичні дані відсутні в силу будь-яких причин. До другого типу завдань можна віднести ті завдання, в яких функція правдоподібності має вигляд, який не припускає зручних аналітичних методів дослідження, але допускає серйозні спрощення, якщо у завдання ввести додаткові «неспостережувані» (приховані, латентні) змінні. Прикладами прикладних завдань другого типу є задачі розпізнавання образів, реконструкції зображень.</p>
Метод дальнього сусіда	<p>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 <math>n - 1</math> кроків і результатом є розбиття, які співпадають з можливими розбиттями в попередньому методі для будь-яких порогових значень.</p>
Центроїдний метод	<p>Відстань між двома кластерами визначається як евклідова відстань між центрами (середніми) цих кластерів. Кластеризація йде поетапно, на кожному з <math>n-1</math> кроків об'єднують два кластери. Якщо <math>n_1</math> більше <math>n_2</math>, 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.</p>
Метод ближнього сусіда	<p>Цей метод полягає в тому, що два об'єкти, які належать одній і тій самій групі (кластері), мають коефіцієнт подібності, який менше деякого порогового значення <math>S</math>. В термінах евклідової відстані <math>d</math> це означає, що відстань між двома точками (об'єктами) кластеру не повинна перевищувати деякого порогового значення <math>h</math>. Таким чином, <math>h</math> визначає максимально допустимий діаметр підмножини, що утворює кластер.</p>

***Порівняльний аналіз ієрархічних і неієрархічних методів кластеризації.*** Перед проведенням кластеризації в аналітика може виникнути питання: якій групі методів кластерного аналізу надати перевагу? Вибираючи між ієрархічними й неієрархічними методами, необхідно враховувати їхні переваги та недоліки.

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. Ця процедура є основою двокрокового алгоритму кластеризації.

Ієрархічні алгоритми пов'язані з побудовою дендрограм (*dendrogram* від гр. *dendron* – "дерево"), які є результатом ієрархічного кластерного аналізу. *Дендрограма* – деревовидна діаграма, що містить  $n$  рівнів, кожен із яких відповідає одному з кроків процесу послідовного укрупнення кластерів. Дендрограму також називають деревовидною схемою, деревом об'єднання кластерів, деревом ієрархічної структури. Дендрограма описує близькість окремих точок і кластерів один до одного, представляє в графічному вигляді послідовність об'єднання (розділення) кластерів. У дендрограмі об'єкти можуть розташовуватися вертикально або горизонтально.

Використовуючи ієрархічні методи, можливо доволі легко ідентифікувати викиди в наборі даних й, у результаті, підвищити якість даних. Ця процедура є основою двокрокового алгоритму кластеризації. Такий набір даних надалі можна використати для проведення неієрархічної кластеризації.

Слід відзначити, що при використанні ієрархічних методів, на відміну від неієрархічних, визначення кількості кластерів (тобто виконання другого етапу алгоритму) не є обов'язковим внаслідок побудови дерева ієрархічної структури вкладених кластерів.

Однак, використання ієрархічних методів супроводжується наступними недоліками, зокрема:

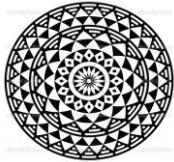
- обмеженням обсягу набору даних;
- обмеженням вибору міри близькості;
- негнучкості отриманих класифікацій об'єктів.

Тобто, ієрархічні методи не можуть працювати з великими наборами даних, а використання деякої вибірки, тобто частини даних, могло б дозволити застосовувати ці методи. Результати кластеризації можуть не мати достатнього статистичного обґрунтування. З іншого боку, під час розв'язання задач кластеризації допустима нестатистична інтерпретація отриманих результатів, а також доволі велика розмаїтість варіантів поняття кластера. Така нестатистична інтерпретація дає можливість аналітикові одержати результати кластеризації, які задовольняють його, що у разі використання інших методів часто доволі складно. В даній ситуації, *неієрархічні методи виявляють вищу стійкість* стосовно некоректного вибору метрики, введення незначущих змінних у набір, що бере участь у кластеризації. Ціною, що доводиться платити за ці переваги методу, є слово “апріорі”. Аналітик повинен заздалегідь визначити кількість кластерів, кількість ітерацій або правило зупинки, а також деякі інші параметри кластеризації. Якщо немає припущень щодо кількості кластерів, рекомендують використати ієрархічні алгоритми. Однак якщо обсяг вибірки не дозволяє це зробити, можливий шлях – провести низку експериментів із різною кількістю кластерів, наприклад, почати розбивку сукупності даних з двох груп і, поступово збільшуючи їх кількість, порівнювати результати. За рахунок такого “варіювання” результатів досягається доволі велика гнучкість кластеризації.

Таким чином, вибираючи між ієрархічними і неієрархічними методами, слід звернути увагу на наступні моменти:

- неієрархічні методи виявляють вищу стійкість по відношенню до викидів, невірному вибору метрики, включенню незначимих змінних у базу для кластеризації та ін. Але платою за це є слово "апріорі";

– дослідник повинен заздалегідь фіксувати результуючу кількість кластерів, правило зупинки і, якщо на те є підстави, початковий центр кластера. Останній момент істотно відбивається на ефективності роботи алгоритму. Якщо немає підстав штучно задати цю умову, взагалі кажучи, рекомендується використовувати ієрархічні методи.



#### 4. МЕТОДИЧНА ОСНОВА ВИЗНАЧЕННЯ ОПТИМАЛЬНОЇ КІЛЬКОСТІ КЛАСТЕРІВ ТА ОБЧИСЛЕННЯ ЗНАЧЕНЬ МІРИ СХОЖОСТІ МІЖ ОБ'ЄКТАМИ

Важливою складовою реалізації алгоритмів неієрархічних методів кластерного аналізу є визначення кількості кластерів, на яку розбиватиметься досліджувана сукупність об'єктів. Найкращою (тобто оптимальною) розбивкою вважається та, що приводить до найбільшої ентропії (невизначеності). Причому відхилення ентропії від максимального значення повинне бути мінімальним.

Розглянемо методичну основу визначення кількості кластерів на прикладі кластерного аналізу стану регіонів України за рівнем загроз енергетичної безпеки.

Вихідні дані значення індикаторів загроз наведено в табл. 3.

*Таблиця 3*

Середні значення індикаторів загроз енергетичної безпеки у  
регіональному розрізі за 2012-2013 рр.

Регіон	Значення індикаторів загроз								
	X <sub>1</sub> *	X <sub>2</sub> **	X <sub>3</sub> **	X <sub>4</sub> **	X <sub>5</sub> **	X <sub>6</sub> **	X <sub>7</sub> **	X <sub>8</sub> **	X <sub>9</sub> ***
АР Крим	49,3	1,34	1859	59.9	1,9	20,5	14,83	4,0	0,22
Вінницька	62,9	2,34	1347	71.0	1,1	26,5	11,60	50,1	2,23
Волинська	56,9	1,24	739	59.6	2,10	18,4	8,95	2,5	0,15
Дніпропетровська	49,5	7,69	6382	62.3	0,01	6,8	5,67	154,6	3,46
Донецька	62,9	10,59	8152	67.0	0,15	12,4	23,81	163,9	4,64
Житомирська	60,2	1,27	1029	60.3	1,19	16,5	13,84	2,6	0,14
Закарпатська	64,3	1,02	764	47.7	2,41	9,9	30,79	1,4	0,21
Запорізька	59,9	2,38	2030	62.7	0,06	21,6	17,43	70,5	2,46
Івано-Франківська	58,1	4,44	1902	76.7	1,01	19,0	10,18	129,0	7,57

Київська	53,0	3,70	2953	55.7	1,09	13,2	74,29	67,1	1,64
Кіровоградська	48,8	1,38	798	73.9	1,40	14,2	12,65	3,5	0,16
Луганська	44,7	7,02	4071	47.5	0,39	7,2	6,00	77,4	3,31
Львівська	57,0	1,87	2653	58.1	1,21	28,7	30,37	23,7	1,22
Миколаївська	45,0	1,70	1377	66.2	0,89	4,4	58,64	5,7	0,24
Одеська	60,8	1,64	2859	41.9	0,40	40,2	83,85	2,6	0,11
Полтавська	45,5	6,46	3509	71.7	0,00	13,1	7,10	14,1	0,39
Рівненська	38,5	1,84	1407	51.1	0,80	14,2	12,68	6,5	0,42
Сумська	52,1	1,79	1425	72.4	2,10	41,5	26,06	7,5	0,42
Тернопільська	58,6	1,23	953	44.8	1,89	32,1	20,39	2,0	0,15
Харківська	46,7	3,13	3837	72.7	0,06	27,2	16,12	66,0	2,01
Херсонська	61,5	1,07	647	77.4	1,34	10,2	40,91	1,3	0,08
Хмельницька	51,2	1,41	1153	49.8	1,73	20,2	24,56	6,6	0,54
Черкаська	43,9	2,85	2653	54.1	0,64	21,2	20,10	26,7	1,01
Чернівецька	62,7	1,01	571	8,80	0,01	12,2	38,16	1,8	0,19
Чернігівська	63,8	1,92	1114	46.2	1,72	27,2	31,98	21,8	1,18
м. Київ	66,1	2,42	5004	58.2	0,03	14,6	38,79	10,0	0,13

Примітка: 1)\* – дані Аналітичного центру «БЕСТ» (<http://www.energy-index.com.ua>); 2)\*\* – дані статистичних бюлетенів Державної служби статистики України, зокрема: "Використання енергетичних матеріалів та продуктів перероблення нафти", "Результати використання палива, теплоенергії та електроенергії", "Про основні показники роботи опалювальних котелень і теплових мереж України", "Виробництво і споживання електроенергії та окремі техніко-економічні показники роботи електростанцій в Україні"; 4)\*\*\* – розраховано авторами; 5)  $X_1$  – енергоефективність, % від ЄС;  $X_2$  – середньодушові витрати енергетичних матеріалів та продуктів перероблення нафти, т у. п. / особу;  $X_3$  – витрати газу природного, тис. т.у.п.;  $X_4$  – ступінь зносу основних засобів виробництва та розподілення електроенергії, газу та води, % (дані за 2011 р.);  $X_5$  – частка втрат при транспортуванні, розподілі та зберіганні енергетичних матеріалів та продуктів перероблення нафти у загальному обсязі витрат, %;  $X_6$  – частка ветхих та аварійних теплових та парових мереж у загальній протяжності, % (міські поселення та сільська місцевість);  $X_7$  – частка капітальних інвестицій в виробництво та розподілення електроенергії, газу та води у загальному обсязі освоєння, %;  $X_8$  – обсяги викидів забруднювальних речовин в атмосферу від енергетики на душу населення, кг/особу;  $X_9$  – енерговикомісткість ВРП (відношення викидів забруднювальних речовин в атмосферу від енергетики до ВРП), кг / 1000 грн.

Ентропія класифікації  $r$  об'єктів розбивається на  $D$  класів і визначається за виразом:

$$H = - \sum_{d=1}^D \frac{r_d}{r} \log_2 \frac{r_d}{r}, \quad (1)$$

де  $H$  – ентропія класифікації, біт;  $r_d$  – кількість регіонів, що потрапили в  $d$ -ий клас, од.

Результати розрахунку ентропії при різних варіантах кількості кластерів методом  $k$ -середніх (за даними табл. 3) наведено в табл. 4.

Таблиця 4

Розрахунок ентропії при різних варіантах кількості кластерів методом  $k$ -середніх

Кількість кластерів	Кількість об'єктів у кластерах, одиниць										ентропія, біт
	1	2	3	4	5	6	7	8	9	10	
3	8	2	16	–	–	–	–	–	–	–	1,24
4	1	7	16	2	–	–	–	–	–	–	1,41
5	7	4	13	1	1	–	–	–	–	–	1,79
6	7	3	13	1	1	1	–	–	–	–	1,91
7	3	7	9	1	1	1	4	–	–	–	2,36
8	7	1	9	1	1	4	1	2	–	–	2,46
9	3	7	6	1	1	1	1	4	2	–	2,78
10	3	4	5	1	1	4	2	1	4	1	3,07

Максимально можливе значення ентропії  $H_{max}$  визначається за формулою (1) при значеннях  $r_{db}$ , рівних між собою, тобто кількість об'єктів рівномірно розподілена в кластерах. Приклад розрахунку максимальної ентропії при різних варіантах кількості кластерів регіонів наведено в табл. 5.

Таблиця 5

Розрахунок максимальної ентропії при різних варіантах кількості кластерів регіонів

Кількість кластерів	Кількість об'єктів у кластерах, одиниць										Максимально можлива ентропія, біт
	1	2	3	4	5	6	7	8	9	10	
3	9	9	8	–	–	–	–	–			1,58
4	7	7	6	6	–	–	–	–			2,00
5	6	5	5	5	5	–	–	–			2,32
6	5	5	4	4	4	4	–	–			2,58
7	4	4	4	4	4	3	3	–			2,80
8	4	4	3	3	3	3	3	3			2,99
9	3	3	3	3	3	3	3	3	2		3,16
10	3	3	3	3	3	3	2	2	2	2	3,30



Відхилення ентропії ( $\Delta H$ ) від максимального значення визначено за виразом:

$$\Delta H = \frac{(H_{\max} - H) \cdot 100}{H_{\max}}. \quad (2)$$

Результати розрахунку відхилення ентропії від максимального значення наведено в табл. 6.

Таблиця 6

Відхилення ентропії від максимального значення

Кількість кластерів	Максимально можлива ентропія, біт	ентропія, біт	Відхилення ентропії від максимального можливого значення, %
3	1,58	1,24	21,73
4	2,00	1,41	29,54
5	2,32	1,79	22,91
6	2,58	1,91	25,81
7	2,80	2,36	15,72
8	2,99	2,46	17,57
9	3,16	2,78	12,02
10	3,30	3,07	6,81

За даними табл. 6. зроблено висновок, що найменше відхилення ентропії від максимального можливого значення спостерігається при групуванні регіонів на 10 кластерів  $H = 6,81\%$ .

### **Методики обчислення значень міри схожості між об'єктами**

Метод кластеризації полягає у формуванні початкової матриці та розрахунках відстаней кластеризації.

У кластерному аналізі використовуються наступні варіанти для вимірювання відстаней між об'єктами в кластерах ( $l$ ).

#### **1. Евклідова відстань (Euclidean distances).**

Найбільш загальний тип відстані. Обчислюється за формулою (за вихідними, а не за стандартизованими даними):

$$l(x, y) = (\sum_i (x_i - y_i)^2)^{1/2}.$$

#### **2. Квадрат евклідова відстані (Squared Euclidean distances).**

Застосовується, щоб надати більші ваги більш віддаленим один від одного об'єктам:

$$l(x, y) = \sum_i (x_i - y_i)^2.$$

3. *Манхеттенська відстань або відстань міських кварталів (City-block (Manhattan) distances).*

У більшості випадків ця міра відстані призводить до таких же результатів, як і для звичайного відстані Евкліда. Однак для цього заходу вплив окремих великих різниць (викидів) зменшується (оскільки вони зводяться в квадрат).

$$l(x, y) = \sum_i |x_i - y_i|.$$

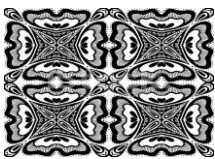
4. *Відстань Чебишева (Chebychev distances metric).*

Це відстань може виявитися корисним, коли бажають визначити два об'єкти як «різні», якщо вони різняться з якої-небудь однієї координати.

$$l(x, y) = \max |x_i - y_i|.$$

5. *Ступенева відстань.* Іноді бажають прогресивно збільшити або зменшити вагу, що відноситься до розмірності, для якої відповідні об'єкти сильно відрізняються. Це може бути досягнуто з використанням ступеневої відстані:

$$l(x, y) = (\sum_i (x_i - y_i)^p)^{1/r}.$$



## 5. КЛАСТЕРНИЙ АНАЛІЗ ІНСТРУМЕНТАМИ DATA MINING (SPSS)

*Data Mining* – міждисциплінарна технологія, що виникла й розвивається на базі досягнень прикладної статистики, розпізнавання образів, методів штучного інтелекту, теорії баз даних тощо. Основна особливість *Data Mining* – це поєднання широкого математичного інструментарію (від класичного статистичного аналізу до нових кібернетичних методів) і останніх досягнень у сфері інформаційних технологій.

У технології *Data Mining* гармонійно об'єдналися строго формалізовані методи і методи неформального аналізу, тобто кількісний та якісний аналіз даних.

До найпоширеніших методів можна віднести такі:

- об'єднання (*association*; іноді вживають термін «*affinity*», що означає подібність, структурну близькість) – виокремлення структур, що повторюються в часовій послідовності. Цей метод визначає правила, за якими можна встановити, що один набір елементів корелює з іншим. Користуючись ним, аналізують ринковий кошик пакетів продуктів, розробляють каталоги, здійснюють перехресний маркетинг тощо;

- аналіз часових рядів (*sequence-based analysis*, або *sequential association*) дає змогу відшукувати часові закономірності між даними (транзакціями). Наприклад, можна відповісти на запитання: купівля яких товарів передуює купівлі даного виду продукції? Метод застосовується, коли йдеться про аналіз цільових ринків, керування гнучкістю цін або циклом роботи із замовником (*Customer Lifecycle Management*);

- кластеризація (*clustering*) — групування записів, що мають однакові характеристики, наприклад за близькістю значень полів у БД. Використовується для сегментування ринку та замовників. Можуть залучатися статистичні методи або нейромережі. Кластеризація часто розглядається як перший необхідний крок для подальшого аналізу даних;

- оцінювання (*estimation*);

- нечітка логіка (*fuzzy logic*);

- статистичні методи, що дають змогу знаходити криву, найближче розміщену до набору точок даних;

- генетичні алгоритми (*genetic algorithms*);

- фрактальні перетворення (*fractal-based transforms*);

– нейронні мережі (*neural networks*) — дані пропускаються через шари вузлів, «навчених» розпізнавати ті чи інші структури.

До *Data Mining* можна віднести ще візуалізацію даних – побудову графічного образу даних, що допомагає у процесі загального аналізу даних вбачати аномалії, структури, тренди. Частково до *Data Mining* примикають дерева рішень і паралельні бази даних. *Data Mining* тісно пов'язана (інтегрована) зі сховищами даних (*Data Warehousing*, DW), які, в якійсь мірі, забезпечують роботу *Data Mining*. Загалом, більшість аналітичних методів, що використовуються в технології *Data Mining* – це відомі математичні алгоритми і методи. Новим у їх застосуванні є можливість використання тих чи інших програмних засобів, що застосовуються при вирішенні тих чи інших конкретних проблем. Зокрема, програмна реалізація алгоритмів кластерного аналізу широко представлена в різних інструментах *Data Mining*, які дозволяють вирішувати завдання досить великої розмірності. Наприклад, агломеративні методи можуть бути реалізовані програмними засобами «SPSS», дивізімні методи – програмними засобами «Statgraphics».

Розглянемо приклад проведення кластерного аналізу програмними засобами «SPSS».

Алгоритм проведення кластерного аналізу стану регіонів України для вихідної статистичної інформації наведено нижче.

### **Крок 1.** Завантаження даних.

Для завантаження даних із книги Excel в програму SPSS необхідно вибрати в меню "Файл" пункт "Открыть", а потім підпункт "Данные...".

У вікні "Открыть данные"(рис. 2) вибрати тип файлу Excel (\*.xls, \*.xlsx, \*.xlsm) та виділити необхідну книгу і натиснути кнопку "Открыть".

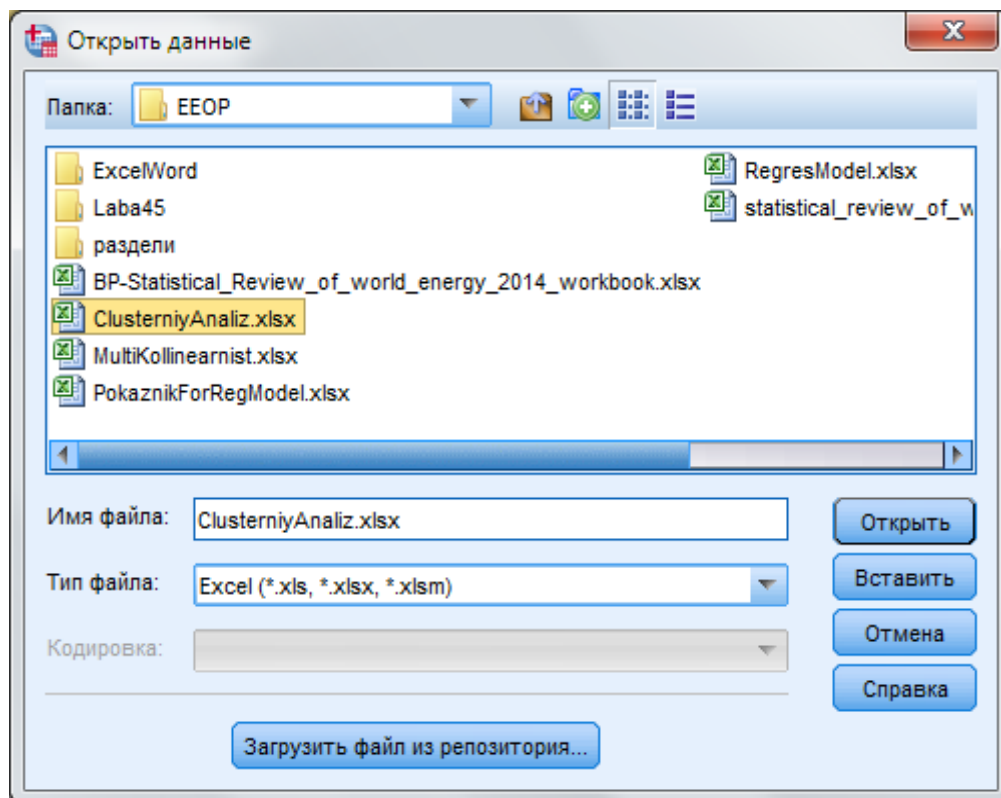


Рис. 2. Вибір книги Excel з даними

У вікні "Открытие файлов Excel" (рис. 3) виставити прапорець "Читает имена переменных из первой строки данных" та натиснути кнопку "ОК".

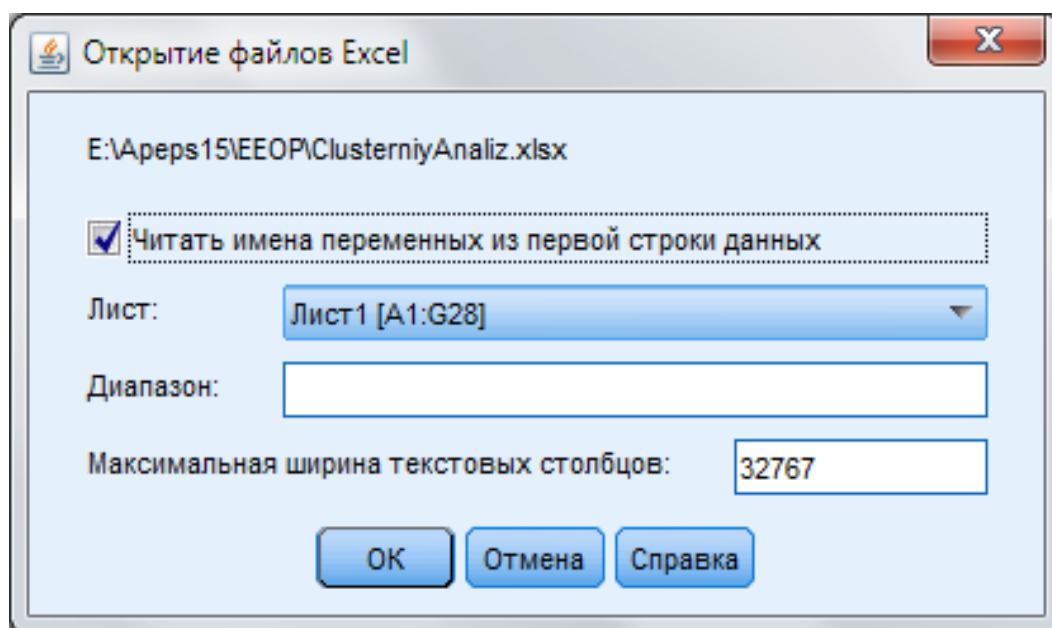


Рис. 3. Вибір назв факторів

Після цього дані завантажаться до закладки "Представление Данные" (рис. 4).

	Регион	Часткагазу природного	Часткаветхита аварійних	Часткавтрат теплоенергії	Часткавикидів від енергетики	Середньодушові обсяги викидів...	Викинуто одним підприємством
1	АР Крим	3,63	20,46	11,02	24,01	4,01	27,9
2	Вінницька	44,63	26,54	13,11	80,08	50,07	324,5
3	Волинська	3,49	18,42	8,64	35,61	2,51	22,6
4	Дніпропетровська	33,93	6,82	16,38	53,10	154,56	1439,0
5	Донецька	46,80	12,42	12,91	47,22	163,92	987,4
6	Житомирська	1,02	16,50	7,26	18,01	2,62	25,2
7	Закарпатська	1,16	9,90	1,62	22,37	1,45	31,2
8	Запорізька	79,80	21,65	12,84	60,59	70,49	590,5
9	Івано-Франківська	56,96	19,02	12,35	90,47	129,05	1521,0
10	Київська	29,32	13,26	9,74	89,04	67,08	509,4
11	Кіровоградська	6,28	14,22	9,90	20,50	3,48	,0

Рис. 4. Завантажені дані для кластеризації

## Крок 2. Формування набору даних для кластеризації.

Вибір даних для кластеризації полягає у визначенні факторів, що описують об'єкти в багатовимірному просторі ознак. Для цього необхідно за допомогою меню "Анализ" вибрати пункт "Классификация", а потім – підпункт "Иерархическая кластеризация" (рис. 5).

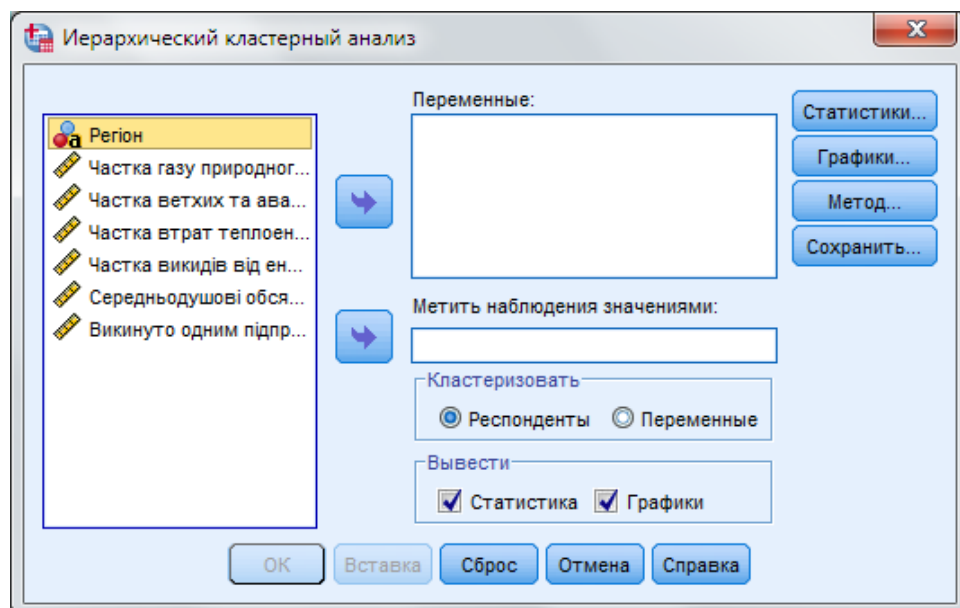


Рис. 5. Вікно "Иерархический кластерный анализ"

З метою підготовки даних до кластеризації необхідно в поле "Метить наблюдения значениями:" перенести з лівого списку пункт "Region", а в список "Переменные" – всі інші фактори (рис. 6).

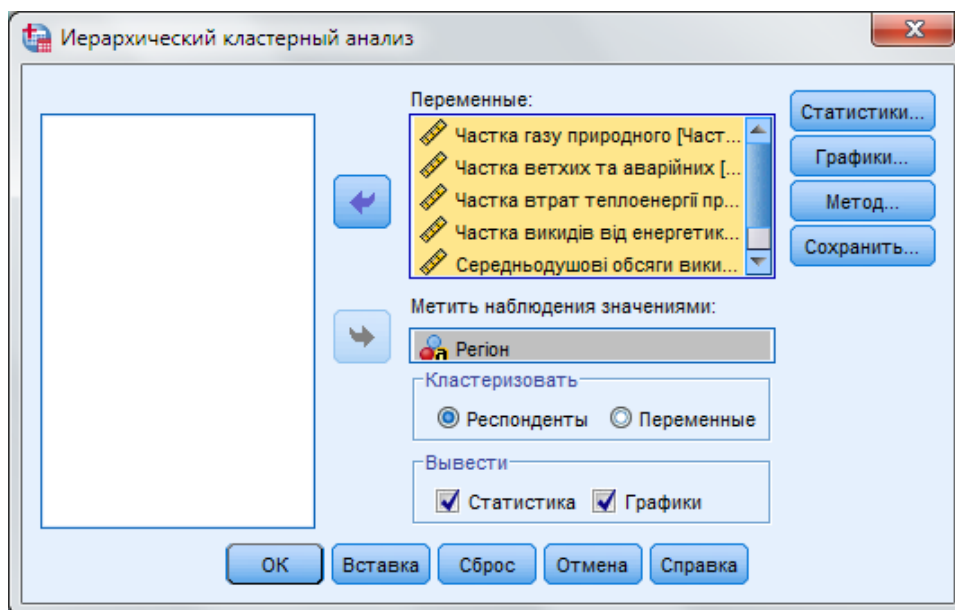


Рис. 6. Підготовка даних до кластеризації

### **Крок 3.** Вибір методу кластеризації.

При натисканні кнопки "Метод..." з'являється вікно "Иерархический кластерный анализ: Метод" (рис. 7). В ньому необхідно вибрати у випадяючому списку "Метод:" один з трьох методів: "Ближайший сосед", "Самый дальний сосед" або "Центроидная кластеризация".

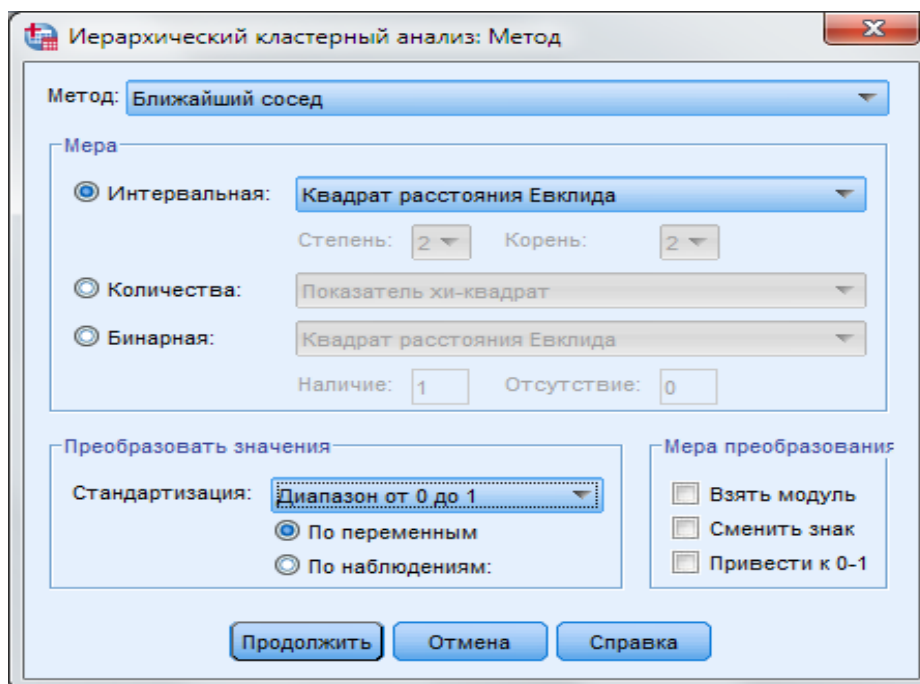


Рис. 7. Вибір методу кластеризації

У випадяючому списку "Стандартизация" залишаємо значення "Нет" (кластеризація по фактичним даним) або обираємо пункт "Диапазон от 0 до 1" (кластеризація по стандартизованим даним). Далі натискаємо кнопку "Продолжить" і повертаємося до попереднього вікна (див. рис. 7).

**Крок 4.** Визначення кількості результуючих кластерів.

При натисканні на кнопку "Статистики" з'являється вікно "Иерархический кластерный анализ: Статистики". В цьому вікні в області "Принадлежность к кластерам" вибираємо пункт "Одно решение" та задаємо кількість кластерів (рис. 8).

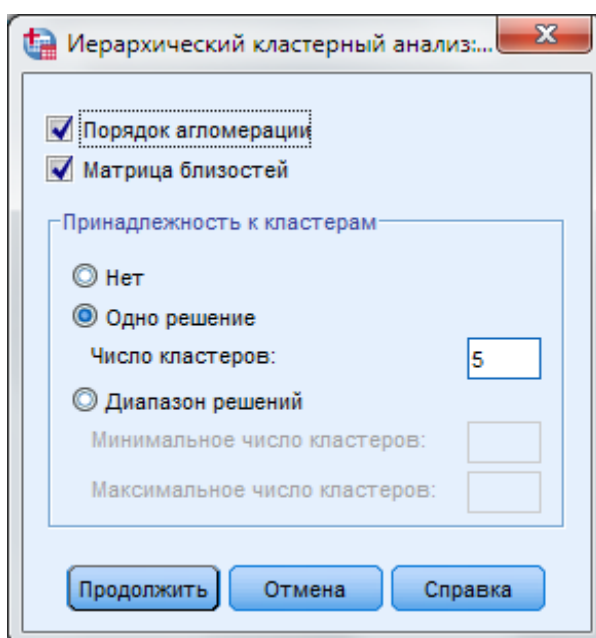


Рис. 8. Визначення кількості кластерів

**Крок 5.** Ініціалізація дендрограми.

При натисканні на кнопку «Графики» з'являється вікно «Иерархический кластерный анализ: Графики». В цьому вікні встановлюємо прапорець «Дендрограмма». При натисканні на кнопку «Продолжить» дане вікно закриється (рис. 9).



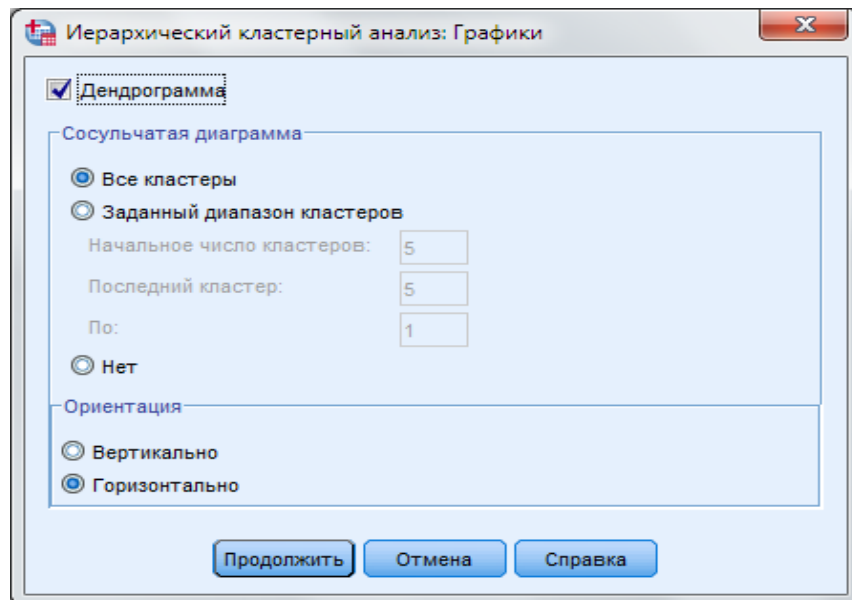


Рис. 9. Вікно ініціалізації дендрограми

### Крок 6. Виконання кластеризації.

Для виконання процедури кластеризації з обраними на попередніх кроках налаштуваннями у вікні рис. 9 натискають кнопку "ОК". Результат кластеризації з'являється у новому вікні виводу (рис. 10).

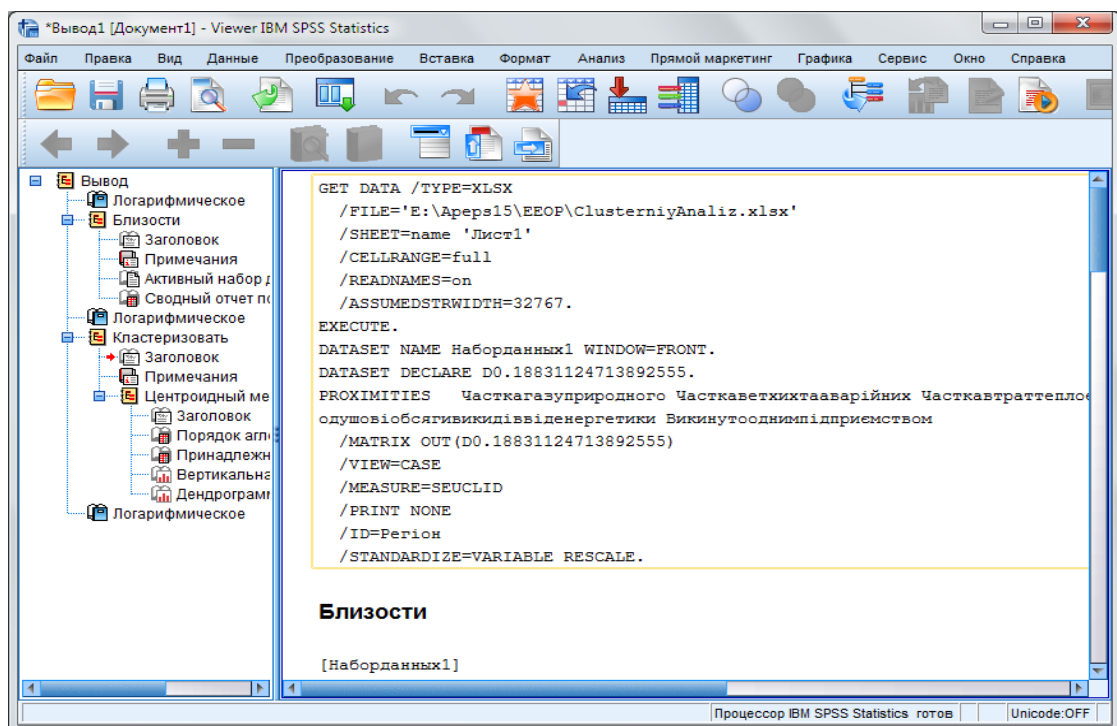
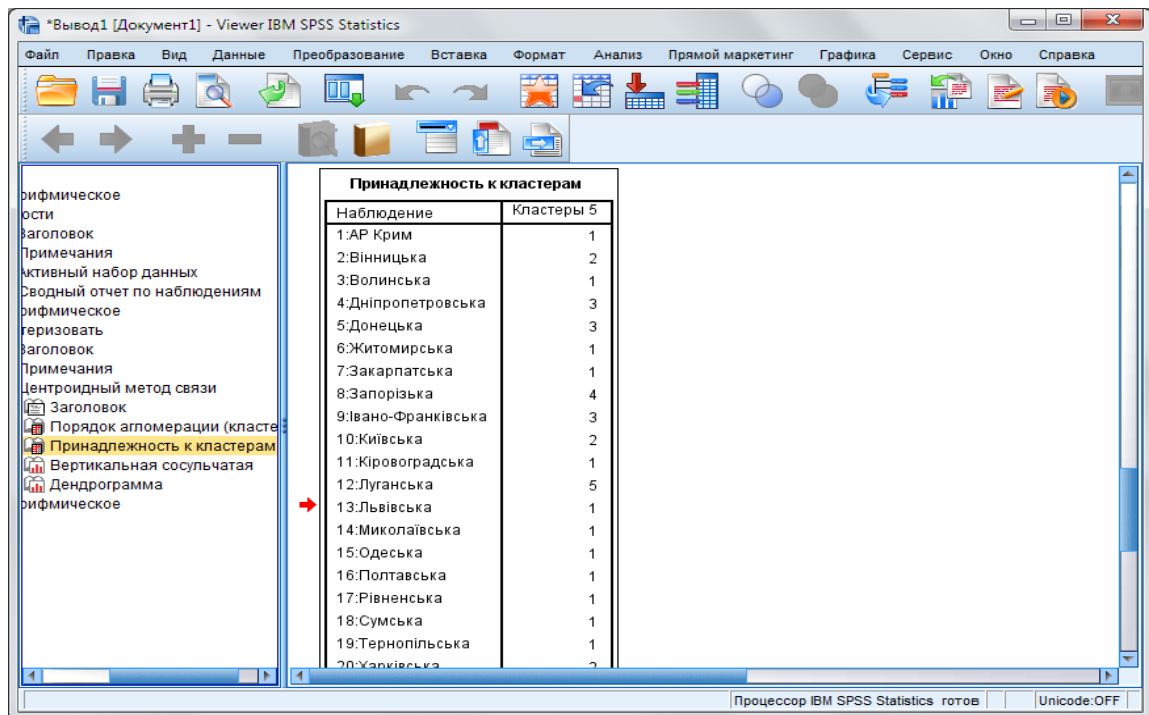


Рис. 10. Вікно виводу результату кластеризації

### Крок 7. Аналіз результатів кластеризації.

У лівій частині вікна в дереві виводу результати кластеризації можна перевірити в табличному вигляді, натиснувши пункт "Принадлежность к кластерам" (рис. 11). У таблиці перший стовпчик відповідає назві об'єкту, в другому вказується номер кластеру до якого належить об'єкт.



Вывод1 [Документ1] - Viewer IBM SPSS Statistics

Файл Правка Вид Данные Преобразование Вставка Формат Анализ Прямой маркетинг Графика Сервис Окно Справка

Принадлежность к кластерам

Наблюдение	Кластеры
1:АР Крим	1
2:Вінницька	2
3:Волинська	1
4:Дніпропетровська	3
5:Донецька	3
6:Житомирська	1
7:Закарпатська	1
8:Запорізька	4
9:Івано-Франківська	3
10:Київська	2
11:Кіровоградська	1
12:Луганська	5
13:Львівська	1
14:Миколаївська	1
15:Одеська	1
16:Полтавська	1
17:Рівненська	1
18:Сумська	1
19:Тернопільська	1
20:Харківська	2

Процессор IBM SPSS Statistics готов Unicode:OFF

Рис. 11. Табличний вигляд результату кластеризації

Також результати кластеризації можна перевірити в графічному вигляді – вузол «Дендрограма» (рис. 12). Дендрограма показує порядок об'єднання об'єктів в кластери.

Для того, щоб порівняти результати кластеризації різними методами (зокрема, з стандартизацією або без стандартизації), з різною кількістю результуючих кластерів необхідно змінити відповідні налаштування для кроку 3 (метод кластеризації, стандартизація) та кроку 4 (кількість кластерів).

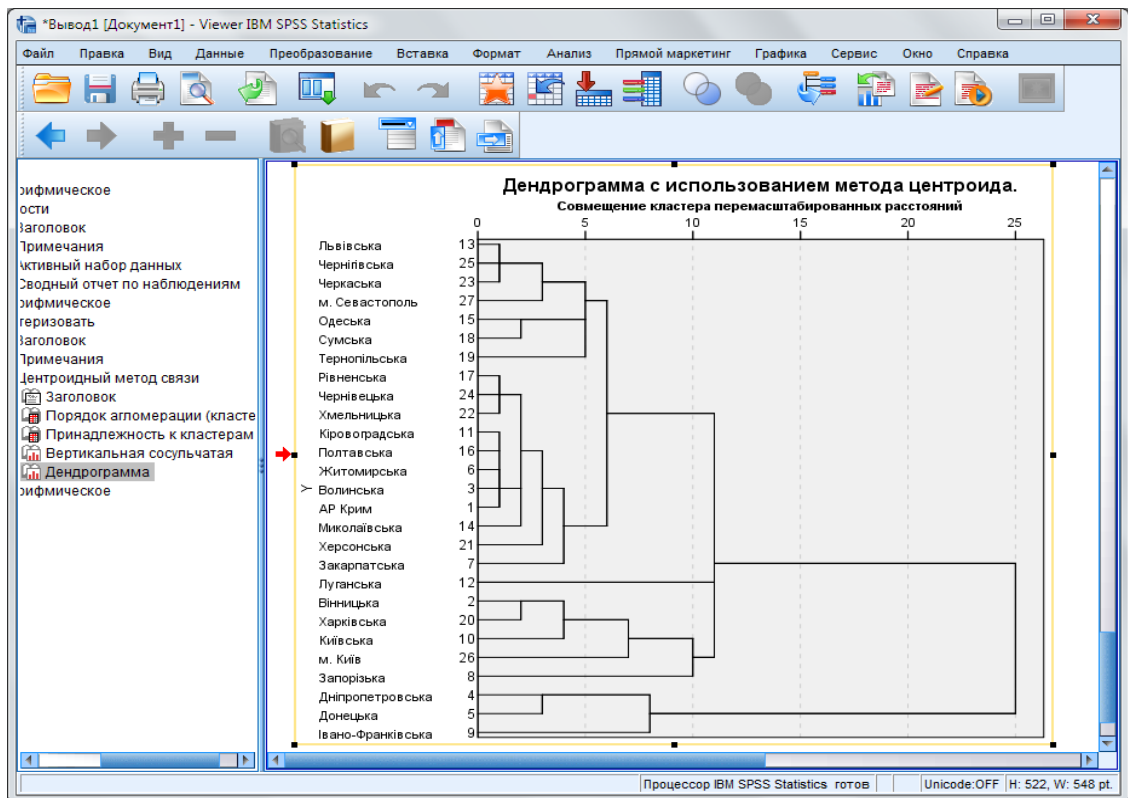


Рисунок 49 – Дендрограма кластеризації

Для того, щоб порівняти результати кластеризації різними методами (зокрема, з стандартизацією або без стандартизації), з різною кількістю результуючих кластерів необхідно змінити відповідні налаштування для кроку 3 (метод кластеризації, стандартизація) та кроку 4 (кількість кластерів).

### Список використаної літератури:

1. Czamanski S. and Ablas L. A. Identification of industrial clusters and complexes: a comparison of methods and findings // Urban Studies. –1979. – V. I6 . – P. 61-80.
2. Leamer E. E. Sources of International Comparative Advantage: Theory and Evidence // Cambridge, MIT Press, 1984. [Electronic resource]. - Access mode : [http://www.anderson.ucla.edu/faculty/edward.leamer/selected\\_research/ucla\\_anderson\\_faculty\\_edward\\_leamer\\_selected\\_research.html](http://www.anderson.ucla.edu/faculty/edward.leamer/selected_research/ucla_anderson_faculty_edward_leamer_selected_research.html).
3. Porter, M. The Competitive Advantage of Nations. New York : Free Press, 1990. (Конкурентные преимущества стран). Портер, М. Конкурентная стратегия. Методика анализа отраслей и конкурентов : пер. с англ. / М. Е.

- Портер. - 2-е изд. – М. : Альпина Бизнес Букс, 2006. – 453 с.
4. Tolentino J. A. *Propos des Filières Industrielles* // *Revue d'Economie Industrielle*. – 1978. – Vol.6, №4. – P. 149-158.
  5. Feldman V. P., Audretsch D. B. *Innovation in Cities: Science based Diversity, Specialization and Localized Competition* // *European Economic Review*. – 1999. – № 43. – P. 31 – 39.
  6. Розен В. П. Використання методу k-середніх кластерного аналізу під час розв'язання задач енергетичної безпеки територій / В. П. Розен, П. П. Іщук, Л. В. Давиденко/ [Електронний ресурс]. – Режим доступу: <http://ena.lp.edu.ua:8080/bitstream/ntb/7693/1/13.PDF>.
  7. Ткаченко О. М. Метод кластеризації на основі послідовного запуску k-середніх з удосконаленням вибором кандидата на нову позицію вставки / О. М. Ткаченко, О. Ф. Грійо Тукало; О. В. Дзісь; С. М. Лаховець. // *Наукові праці ВНТУ*, 2012, № 2. [Електронний ресурс]. – Режим доступу: [http://ot.vntu.edu.ua/content/articles/tkachenko/Наукові%20праці%20ВНТУ\\_2012.pdf](http://ot.vntu.edu.ua/content/articles/tkachenko/Наукові%20праці%20ВНТУ_2012.pdf).
  8. Караєва Н. В. *Методологія кластерного аналізу стану регіонів України за рівнем загроз енергетичної безпеки* / Н.В. Караєва, І.А. Варава, О.В. Красько / *Економічна безпека територіально-виробничих комплексів: енергетика, екологія, інформаційні технології : монографія* / Коцко Т. А. , Чеховська М. М., Лісовські О.Л. [та ін.]; за наук. ред. д.т.н., проф. Лук'яненко С. О., к.е.н., доц. Караєвої Н. В. – К. : «МП Леся», 2015. — С.162-170.
  9. Караєва Н.В. *Методологічні аспекти кластеризації регіонів України за рівнем викликів енергетичної безпеки* / Н.В. Караєва // *Науковий вісник Одеського національного економічного університету*. – Науки: економіка, політологія, історія. – 2016. – No 1 (233). – С. 40-55.