Міністерство освіти і науки України
Київський національний університет імені Тараса Шевченка
факультет соціології
Кафедра методології та методів соціологічних досліджень

КВАЛІФІКАЦІЙНА РОБОТА

НА ТЕМУ:

«Пізнавальний потенціал методу k-means для класифікації типів політичних культур»

Спеціальність: 054 «Соціологія»

Освітня програма: «Соціологія»

Освітній рівень: магістр

Кваліфікація: магістр соціології

Виконавець:

Медвідь Олександра Олександрівна,

студентка магістратури

Науковий керівник:

Сидоров Микола Володимир-Станіславович,

кандидат фізико-математичних наук**,** доцент кафедри методології та методів соціологічних досліджень

Магістерська робота допущена до захисту

рішенням кафедри *методології та методів соціологічних досліджень*

Протокол № ____ від «__» _____ 2020 р.

Зав.кафедри _____ доцент Сидоров М.В.-С.

Київ 2020

Реєстрація

номер	дата	підпис лаборанта кафедри

Рекомендовано до захисту

підпис наукового керівника	ініціали, прізвище наукового керівника

Результат захисту

оцінка	дата захисту

Голова ЕК

підпис	ініціали, прізвище

Члени ЕК

підпис	ініціали, прізвище
підпис	ініціали, прізвище
підпис	ініціали, прізвище
підпис	ініціали, прізвище

Секретар ЕК

підпис	ініціали, прізвище

АНОТАЦІЯ

На сьогодні важко заперечувати значення кластерного аналізу, методи якого дозволяють побудувати класифікації багатовимірних даних, виявити внутрішні зв'язки між одиницями спостережуваної сукупності, а також можуть використовуватися з метою стиснення інформації. Оскільки основною задачею кластеризації є поділ об’єктів на групи таким чином, щоб рівень подібності між об’єктами однієї групи був високий, а рівень подібності між об’єктами різних груп – низький, поняття якості кластеризації складається з таких ознак як компактність, відокремлюваність та концентрація. Однією з найбільших проблем кластеризації є те, що кластери формуватимуться, навіть якщо аналізований набір даних має повністю рандомну структуру. І щоб оцінити кластерне рішення потрібно насамперед провести оцінку загальної схильності наявних даних до об'єднання в кластери, а також провести візуальну оцінку тенденції.

Ключові слова: кластерний аналіз, метод k-середніх, класифікація, політична культура

ANNOTATION

Today it is difficult to deny the importance of cluster analysis, the methods of which allow to build classifications of multidimensional data, to reveal internal relations between the units of the observed population, and can also be used for information compression purposes. Since the main task of clustering is to divide objects into groups so that the level of similarity between objects of one group is high and the level of similarity between objects of different groups is low, the concept of clustering quality consists of such features as compactness, separateness and concentration. One of the biggest problems of clustering is that clusters will form even if the analyzed data set has a completely random structure. And in order to evaluate a clustering solution, you must first assess the overall propensity of the available data to cluster together, as well as perform a visual trend assessment.

Keywords: cluster analysis, k-means method, classification, political culture

ЗМІСТ

ВСТУП

Останнім часом інформація, що росте в колосальних масштабах породжує потребу в обробці великих обсягів даних. В цьому напрямку велике місце відведено інтелектуального аналізу даних. Цей напрямок включає в себе методи, відмінні від класичного аналізу, засновані на моделюванні, ймовірності, вони вирішують завдання узагальнення, асоціювання і відшукування закономірностей. У великій мірі розвитку цієї дисципліни сприяло проникнення в сферу аналізу даних ідей, що виникли в теорії штучного інтелекту. Також однією з причин стрімкого розвитку кластерного аналізу, крім розвитку засобів обчислювальної техніки і зростання обсягів оброблюваної інформації, є поглиблення спеціального знання. Це неминуче призводить до збільшення кількості змінних, що враховуються при аналізі тих чи інших об'єктів і явищ. Внаслідок цього суб'єктивна класифікація, яка раніше спиралася на досить малу кількість врахованих ознак, часто виявляється вже ненадійною. А об'єктивна класифікація, зі все зростаючим набором характеристик об'єкта, вимагає використання складних алгоритмів кластеризації.

На сьогодні важко заперечувати значення кластерного аналізу, методи якого дозволяють побудувати класифікації багатовимірних даних, виявити внутрішні зв'язки між одиницями спостережуваної сукупності, а також можуть використовуватися з метою стиснення інформації.

Хроніка кластерного аналізу налічує менше 100 років, проте він вже встиг стати невід'ємною частиною процедури переробки інформації в численних науках і інших сферах людської життєдіяльності. Настільки інтенсивне формування систематизації знаходиться в зв'язку зі збільшенням обчислювальної техніки та її повсякденному використанні. Сьогодні існує велика кількість методів кластерного аналізу. На даний момент однієї загальноприйнятої класифікації не існує, найпоширенішим є поділ на ієрархічні та неієрархічні групи методів.

Ієрархічні методи зручно застосовувати для невеликих масивів даних через їх наочність, адже поділ на кластери передається дендрограмою. При великій

кількості спостережень ієрархічні методи кластерного аналізу не придатні. У таких випадках використовують неієрархічні методи, засновані на поділі, які представляють собою ітеративні методи дроблення вихідної сукупності. Така неієрархічна кластеризація полягає в поділі набору даних на певну кількість окремих кластерів. Одним з найпопулярнішим методом неієрархічного кластерного аналізу є кластеризація методом k – середніх.

Актуальність дослідження: сьогодні достатньо багато досліджень ставлять за мету організацію отриманих даних у наглядні структури. Кластерний аналіз знаходить застосування в найрізноманітніших наукових напрямках: біологія, соціологія, медицина, археологія, історія, географія, економіка, філологія тощо. Тому не є винятком використання кластеризації у дослідженні такої важливої категорії як політична культура.

Кластеризацію використовують як самостійний інструмент аналізу даних або як попередній етап інших методів аналізу. Термін "кластерний аналіз" вперше був запропонований Тріоном. Помітний поштовх у розвитку робіт по кластерному аналізу дали роботи Р.Розенблатта. Сучасні алгоритми теорії розпізнавання образів, класифікації та кластерного аналізу базуються на роботах С.А.Айвазяна , А.Я.Червоненкіса, В.Н.Вапніка , Р.А. Фішера , В.Н.Фоміна, І.Форджі , К.Фукунагі, Дж.Хартігана, Дж.Хопфілда , Я.З.Ципкіна та ін. Працездатність різних алгоритмів розбиття множини даних на класи істотно залежить від кількості класів (кластерів) і вибору початкового розбиття. При апріорі невідомому кількості кластерів В.Кржановскім і І.Лаем , Дж.Дуном, Л.Хьюбертом і Дж.Шульцом , Р.Калінскім і Дж.Харабазом , Е.Левіне і Е .Домані , А.Бен-Гуром і І.Гійоном, А.Елізівом, В.Волковічем і Г.Вальтером та ін. активно розробляються методи стійкою кластеризації, що точно оцінюють кількість кластерів в різноманітних прикладних завданнях. Але незважаючи на велику кількість досліджень в цій області існує ряд актуальних питань. В роботах багатьох науковців упускається ряд важливих моментів при застосуванні методу k-середніх, таких як мотивація вибору кількості кластерів, попередня обробка даних, особливості вибору початкових центрів кластерів, перевірка валідності

кластерного рішення тощо. Особливо зазнає необхідність в детальному дослідженні та описі оцінка кластерних рішень, оскільки кожен раз, коли ми отримуємо певну кластерну структуру дослідник має бути впевненим, що дана структура є коректною та відповідає дослідницьким вимогам та є придатною для подальшого застосування. Незважаючи на відносну простоту методу кластеризації k-середніх, практика його застосування в дослідженнях і опису в публікаціях має безліч недоліків, що надалі може привести до малої стійкості отриманих класифікацій і невідтворюваних результатів. Спектр застосувань неієрархічної кластеризації дуже широкий, проте універсальність застосування привела до появи великої кількості несумісних термінів, методів і підходів, що ускладнюють однозначне використання і несуперечливу інтерпретацію даних, а також унеможливлюють процес оцінювання результатів, які отримані за допомогою кластеризації. Оцінка якості безсумнівно важлива для всього процесу кластеризації, тому що без неї отримана структура кластерів не може вважатись досить достовірної для того, щоб робити з неї певні висновки або проводити подальший аналіз на її основі.

Розповсюдженим способом оцінки якості кластеризації є перевірка, вироблена аналітиком вручну, наприклад, за допомогою візуалізації отриманої структури кластера. Однак, це не завжди зручно, а в умовах наявності великих обсягів даних або ж даних з високою розмірністю, така перевірка майже неможлива. Альтернативою методам візуальної оцінки якості є автоматичні методи оцінки якості кластеризації. Вони можуть бути розглянуті як функції від отриманої кластерної структури і вихідного безлічі. У літературі методи даної групи носять назви індексів або метрик.

Тенденція відсутності знання про оцінку застосування неієрархічного кластерного аналізу методом k-середніх може призвести до зменшення рівня якості отриманих результатів та розповсюдження некоректних результатів досліджень, які в свою чергу можуть слугувати підґрунтям для наступних досліджень та породжувати ланцюг неякісних даних. Кластерний аналіз традиційно привертає до себе увагу дослідників протягом декількох десятиліть і

до теперішнього моменту не втратив своєї актуальності. Оскільки оцінка якості результатів невіддільна від загального процесу кластеризації, розробка, порівняння або підвищення ефективності автоматичних методів оцінки якості кластеризації є актуальною сферою дослідження так само як й дослідження пізнавального потенціалу методу .

Проблема дослідження полягає у недостатній кількості знань про пізнавальний потенціал методу k-means для класифікації типів політичних культур.

Об'єкт дослідження – неієрархічна кластеризація методом k-середніх.

Предмет дослідження – пізнавальний потенціал методу алгоритмів k-means для класифікації типів політичних культур

Метою даної роботи є опис та узагальнення пізнавального потенціалу методу k-means для класифікації типів політичних культур.

Завдання дослідження:

- зробити теоретико-методологічний аналіз сутності кластерного неієрархічного аналізу методом k-середніх та визначити основні особливості методу;
- розглянути особливості класифікації типів політичних культур методом k-середніх;
- продемонструвати пізнавальний потенціал методу k-середніх для класифікації типів політичних культур.

РОЗДІЛ І. ТЕОРЕТИКО – МЕТОДОЛОГІЧНИЙ АНАЛІЗ СУТНОСТІ НЕІЄРАРХІЧНОГО КЛАСТЕРНОГО АНАЛІЗУ МЕТОДОМ K-MEANS

1.1. Основні положення неієрархічного кластерного аналізу методом k-середніх як методу класифікації інформації

Сьогодні достатньо багато досліджень ставлять за мету організацію отриманих даних у наглядні структури. Одним із способів систематизації та класифікації є кластерний аналіз. Він являє собою набір методів, що використовуються для угрупування об'єктів або подій у відносно однорідні групи, що називають кластерами . Об'єкти в кожному кластері повинні бути схожі між собою і відрізнятися від об'єктів в інших кластерах. Кластерний аналіз дозволяє відкрити в даних раніше невідомі закономірності, які практично неможливо досліджувати іншими способами і представити їх в зручній для користувача формі. [Алексеёнок А.А, 2001: с.42]. Сьогодні література про кластерний аналіз вражає своїм різноманіттям, сучасні алгоритми теорії розпізнавання образів, класифікації та кластерного аналізу базуються на роботах С.А.Айвазяна , А.Я.Червоненкіса, В.Н.Вапніка , Ф.Розенблатта , Р.А. Фішера , В.Н.Фоміна, І.Форджі , К.Фукунагі, Дж.Хартігана , Дж.Хопфілда , Я.З.Ципкіна та ін.

Сьогодні існує велика кількість методів кластерного аналізу. Один з найпоширеніших поділів алгоритмів є поділ на ієрархічні та неієрархічні. Ієрархічний кластерний аналіз - це метод упорядкування даних, шляхом послідовного об'єднання менших кластерів в великі або поділ великих кластерів на менші. Головною відмінною рисою такого методу є те, що процес об'єднання об'єктів при їх використанні має ієрархічний характер і може бути представлений у вигляді дендрограми (деревовидної діаграми), де кожен рівень відповідає одному кроку алгоритму. [Дж.О. Ким, Ч.У. Мьюллер, 1987:с.45] Ієрархічні методи зручно застосовувати для невеликих масивів даних через їх наочність, адже поділ на кластери передається дендрограмою та можливість отримати детальне уявлення про структуру даних. Слід зауважити, що при великій кількості спостережень ієрархічні методи кластерного аналізу не придатні. У

таких випадках використовують неієрархічні методи, засновані на поділі, які представляють собою ітеративні методи дроблення вихідної сукупності. У процесі поділу нові кластери формуються до тих пір, поки не буде виконано правило зупинки. [Енюков И.С.,1997:с.43] На відміну від ієрархічних методів, які не вимагають попередніх припущень щодо числа кластерів, для можливості використання неієрархічних методів необхідно мати гіпотезу про найбільш ймовірне кількості кластерів. Така неієрархічна кластеризация полягає в поділі набору даних на певну кількість окремих кластерів.

Неієрархічних методів більше, хоча працюють вони на одних і тих же принципах. По суті, вони являють собою ітеративні методи дроблення вихідної сукупності. У процесі поділу формуються нові кластери, і так до тих пір, поки не буде виконано правило зупинки. [Глаголєва І.І.,2010:с.65] Процес неієрархічній кластеризації завжди є ітеративним. Ітеративні методи кластеризації розрізняються вибором наступних параметрів:

- початкової точки;
- правилом формування нових кластерів;
- правилом зупинки.

Неієрархічні методи проявляють більші стійкість відносно шумів та викидів, а також некоректного вибору метрик. Відповідно до цього дослідник має заздалегідь визначитись з кількістю кластерів, їх центрами та певним параметром кластеризації. Також у сучасному світі існує поділ алгоритмів кластерного аналізу на:

• Монотетічні та політетичні. Залежно від того, використовуються властивості об'єкта при кластеризації послідовно або одночасно, відповідно. Більшість алгоритмів політетичні. У деяких завданнях монотетичні підходи виявляються більш ефективними, проте з ними виникають великі труднощі при роботі в просторах з великими розмірностями.

• Ті,що не перетинаються(непересічні) та нечіткі. Непересічні алгоритми відносять кожен елемент строго до одного певного кластеру, в той час як нечіткі

алгоритми кожному елементу повертають вектор ступенів належності до того чи іншого кластеру. [An Introduction to Cluster Analysis for Data Mining].

• Детерміновані і стохастичні. Ці методи відносяться до неієрархічних алгоритмів кластеризації, оптимальне рішення для яких шукається за рахунок мінімізації певного функціоналу. Залежно від того, яким способом шукаються оптимальні значення - традиційними методами або методами випадкового пошуку - відбувається такий розподіл. Також можна робити поділ алгоритмів на ті, що працюють в режимі реального часу, залежать або не залежить від початкового розбиття і порядку розгляду об'єктів тощо. [Steinley D. - Br. J. Math. Stat. Psychol,2006:c.32]

Якщо ж повертатись до загальноприйнятого розподілу алгоритмів на ієрархічні та неієрархічні, то слід зауважити, що найбільш поширений серед неієрархічних методів алгоритм є алгоритм k-середніх, також званий швидким кластерним аналізом.

Метод k- середніх - один з найбільш популярних і простих в реалізації методів кластерного аналізу, що був вперше застосований в психології в 30-х роках XX століття К. Тріоном. Більш широке застосування кластерний аналіз отримав з розвитком обчислювальних можливостей комп'ютерної обробки великих баз даних. Зокрема термін «k-Means» був спочатку використаний Джеймсом МакКвіном в 1967, хоча ідея повертається до Хьюго Штейнгаусу в 1957 р., коли польський вчений був перший, хто запропонувати метод об'єднання в кластери k-means. Стандартний алгоритм був спочатку запропонований Стюартом Ллойдом в 1957 як техніка для модуляції кодексу пульсу, хоча це не було видано за межами наукового товариства Bell Labs до 1982 р . У 1965 р. Форджи видав по суті той же самий метод, який наразі, згадується в науковій літературі як алгоритм Ллойда-Форджи. [Tryfos Peter Cluster analysis] Більш ефективна версія була запропонована і видана Хартігеном і Вонгом лише у 1975-1979 рр.

Мета кластерного аналізу методом k-means полягає в знаходженні існуючих в даних структур - кластерів. При цьому об'єкти в кожному кластері

повинні бути максимально схожі між собою і відрізнятися від об'єктів в інших кластерах. [Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data] Кінцеве розбиття даних має задовольняти певний критерій оптимальності, що виражені рівнем бажаності розбивок. [Буреева Н.Н., 2002:с.24] Основний тип задач, що вирішується за допомогою алгоритму k-means це наявне певне припущення стосовно кількості кластерів , які мають бути дуже різними, на скільки це можливо. Вибір числа k може базуватися на результатах попередніх досліджень, теоретичних міркуваннях або інтуїції.

Кластерний аналіз методом k-середніх в якості цільової функції використовує мінімізацію внутрішньокластерної дисперсії тобто оптимальне розбиття досягається за допомогою мінімізації суми квадратичної відстані точок кластерів від центрів цих кластерів за ітеративною процедурою. Кожен центроїд кластера - це сукупність значень функцій, які визначають отримані групи. Вивчення функцій центроїдних функцій може бути використано для якісного інтерпретації того, яка група представляє кожен кластер.

Даний метод дозволяє розділити довільний набір даних на задане число кластерів таким чином, що об'єкти всередині кластера були досить близькі один до одного, а об'єкти різних кластерів не перетиналися. Іншими словами, мета алгоритму - об'єднати в групи подібні дані за певними заздалегідь заданими критеріями. [Захаров В.М., 2007:с.14].

Л. Моріссетт і С. Картьє вказують, що метод К-середніх найчастіше використовується:

1. Для експлораторного аналізу і побудови класифікацій в дослідницькій практиці і інтелектуальному аналізі даних.

2. Для редукції (зменшення складності) даних.

3. Як початковий крок для більш складних в обчислювальному плані алгоритмів, який дає приблизне поділ даних як нові початкові точки (зменшення зашумлення в наборі даних). [Буреева Н.Н.,2004:с.16]

Однією з перших завдань, які вирішує дослідник при плануванні дослідження є вибір методів (методик) дослідження і визначення розміру вибірки. Відносно кластерного аналізу методом К-середніх слід пам'ятати, що рішення буде отримано незалежно від розміру вибірки і кількості змінних І. Муи і М. Сарстедт також вказують, що не існує загальноприйнятих правил щодо мінімально необхідного розміру вибірки, відносин між об'єктами і кількістю змінних для кластеризації [Hans-Hermann Bock,2012:c.23]. Однак це не означає, що кластерний аналіз є універсальним методом для будь-яких даних. А. К. Форман рекомендує розмір вибірки не менше 2k, де k - кількість змінних кластеризації і існує більш сувора рекомендація про те, що бажаний розмір вибірки повинен складати $5 \bullet 2k$.

Алгоритми методу К-середніх також можуть припускати, що дослідник самостійно визначає і задає число кластерів. При цьому реальна кількість природних груп в наборі об'єктів найчастіше невідомо .У дослідника є кілька різних шляхів визначення числа кластерів, яке слід задати алгоритму:

• на основі попередньої інформації ;

• емпіричне визначення числа кластерів ;

• візуальне визначення числа кластерів.

Перший шлях передбачає наявність у дослідника будь-якого теоретичного знання про групи, які повинні вийде в результаті застосування кластерного аналізу або ж є припущення про подальше використання отриманого розбиття. Наприклад, подальше порівняння груп статистичними критеріями може припускати додаткові вимоги до кількості об'єктів в кластерах.

Існує також кілька базових підходів до визначення кількості кластерів в безлічі даних. Вони засновані на:

• індексах, які порівнюють ступеня "розкиду" даних усередині кластерів і між кластерами. [Ким Дж.О.,1989:с.34]

• розрахунку значень евристичних характеристик (функцій стійкості), що показують відповідність призначених кластерів для вибіркових елементів множини. [Coppedge M.,2011:c.72] Г.Муфті, П.Бертранд і Л.Мубаркі визначають

функцію стійкості кластеризації на основі вимірювання ізоляцій Ловінгера. А.Джейн і Дж.Морено використовували дисперсії емпіричних розподілів в якості вимірювання стійкості;

• статистиках, що визначають найбільш ймовірне рішення. [Нейский И.М.,2000:с.38] Т.Ланге, В.Рот, Л.Браун і Дж.Бухман пропонують метод, в якому порівнюються пари кластеризованих даних.

• оцінюванні щільності розподілів. Тут можна згадати роботу Д.Вішарта. Вводиться поняття кластерів високої щільності, кількість кластерів визначається як загальне число непересічних областей, чиї щільності перевищують задане значення.

Отже, у більшості літератури методи для визначення кількості кластерів вважаються методами стійкої кластеризації. Стійкість кластеризації показує, наскільки різними виходять підсумкові розбиття на групи після багаторазового застосування алгоритмів кластеризації для одних і тих же даних. Невелика розбіжність результатів інтерпретується як висока стійкість. Кількість кластерів, яке максимізує кластерну стійкість, може служити гарною оцінкою для реальної кількості кластерів. [Соколова Л.В. ,1999,с.56]

При емпіричному визначенні числа кластерів популярними є стратегії визначення числа кластерів на основі розмірів отриманих кластерів, на основі графіка залежності відносини дисперсій до числа кластерів або зменшення суми внутрікластерних дисперсій (критерій «кам'янистої осипи») на основі результатів ієрархічного кластерного аналізу, при допомоги порівняння різних кластерних рішень від Kmin до Kmax за допомогою статистичних критеріїв [Захаров В.М., 2007:с.14].

При візуальному визначенні числа кластерів рекомендується будувати безліч гістограм розподілів змінних, а також двовимірних графіків розсіювання. У разі занадто великої кількості даних за допомогою методу головних компонент, факторного аналізу або багатовимірного шкалювання можна побудувати проекції даних в простір меншої розмірності [Coppedge M.,1997:с.41]

Є також інші формальні техніки для зменшення ступеня суб'єктивності при вирішенні питання про кількість кластерів. Г. Мілліган і К. Купер описують 30 таких технік. У дослідженнях українських вчених найчастіше виділяють від 2 до 4 кластерів. Як і все методи k-середнє має свої переваги та обмеження. Алгоритм k-середній ефективний перш за все тому, що він не потребує обчисленні всіх попарних відстаней між спостереженнями, на відміну від більшості інших алгоритмів кластеризації, включаючи той, що використовується в процедурі ієрархічного кластерного аналізу.

Якщо говорити про переваги, то вони полягають у тому простоті та швидкості виконання. Метод k-середніх більш зручний для кластеризації великої кількості спостережень оскільки у ієрархічному кластерного аналізі дендограми перевантажують розуміння результатів, тому що втрачають наочність. На відміну від ієрархічних процедур метод k-середніх не вимагає обчислення і зберігання матриці відстаней або подібностей між об'єктами. Алгоритм цього методу передбачає використання тільки вихідних значень змінних. [Захаров В.М. ,2011:с.45]

Для початку процедури класифікації повинні бути задані k обраних об'єктів, які будуть служити еталонами, тобто центрами кластерів. Вважається, що алгоритми еталонного типу зручні та потребують менших часових затрат. У цьому випадку важливу роль відіграє вибір початкових умов, які впливають на тривалість процесу класифікації і на його результати. Метод k-середніх зручний для обробки великих статистичних сукупностей. [Ким Дж.О.,1989:с.34] Слід зауважити, що метод k-середніх має ряд певних обмежень та недоліків. Одним з основних недоліків є той факт, що результат класифікації сильно залежить від випадкових початкових позицій кластерних центрів та залежить від кваліфікації дослідника, адже кількість кластерів повинна бути заздалегідь визначена.

Також алгоритм чутливий до викидів, які можуть викривлювати середнє. Класичний варіант має на увазі випадковий вибір кластерів, що дуже часто було джерелом похибки. [Климчук В.О.,2005:с.56] Як варіант вирішення, необхідно проводити дослідження об'єкта для більш точного визначення центрів

початкових кластерів. Також можливим вирішенням цієї проблеми є використання модифікації алгоритму -алгоритм k-медіани.

Не слід забувати про те, що даному методу характерне порушення умови зв'язності елементів одного кластера, тому розвиваються різні модифікації методу, а також його нечіткі аналоги , у яких на першій стадії алгоритму допускається приналежність одного елемента множини до декількох кластерів (із різним ступенем приналежності).[Ломидзе О.Н.,2004:с.33]

К. Захаров у свої публікації наводить такі недоліки методу [Захаров В.М. ,2011:с.45]:

- метод завжди сходиться, але може привести до знаходження локального мінімуму і неоптимальному поділу даних;

- метод прагне до створення кластерів рівного розміру, навіть якщо це не є найкращим відображенням реально існуючих груп;

- не справляється із завданням, коли об'єкт належить до різних кластерів в рівній мірі або не належить жодному.

- в алгоритмі Форджа-Ллойда можливе створення порожніх кластерів, а в алгоритмі Маккуїна і Харігана і Вонга - рішення чутливе до порядку, в якому висуваються точки;

- вибір різних початкових центрів призводить до різних рішень;

Отже, не зважаючи на недоліки, алгоритм k-means є найбільш поширеним інструментом кластеризації на практиці. Це простий і масштабований метод, який володіє достатньою ефективністю і здатний працювати як доповнення до більш складних методів. Даний метод дозволяє розділити довільний набір даних на задане число кластерів таким чином, що об'єкти всередині кластера були досить близькі один до одного, а об'єкти різних кластерів не перетиналися. Для коректного застосування даного методу необхідно дотримуватися вимог однорідності та повноти змінних та постійно перевіряти стійкість як результатів, так і кластеризації.

1.2.Опис алгоритму методу k-середніх та його варіацій

Перш ніж описувати алгоритми важливо розрізняти такі властивості, властиві алгоритмам. По-перше, необхідно чітко зрозуміти різницю між класифікацією і кластеризацією. Класифікація - це віднесення кожного документа в певний клас із заздалегідь відомими параметрами, отриманими на етапі навчання. Число класів строго обмежена. Кластеризація - розбиття безлічі документів на кластери - підмножини, параметри яких заздалегідь невідомі. [Ворона В.М.,1998:с.10] Кількість кластерів може бути довільним або фіксованим.

Загальна ідея алгоритму: заданий фіксоване число k кластерів спостереження зіставляються кластерам так, що середні в кластері (для всіх змінних) максимально можливо відрізняються один від одного. У загальному вигляді зміст алгоритму методу К-середніх полягає в реалізації наступних кроків:

Етап № 1 - Початковий розподіл об'єктів на кластери. На цьому етапі маємо масив спостережень (об'єктів), кожен з яких має певні значення по ряду ознак. Відповідно до цих значень об'єкт розташовується у багатовимірному просторі. В рамках цього етапу дослідник визначає кількість кластерів, що необхідно утворити. Вибір кількості кластерів відбувається на основі дослідницької гіпотези. Якщо її немає, то рекомендують створити 2 кластери, далі 3,4,5, порівнюючи отримані результати. [An Introduction to Cluster Analysis for Data Mining] А рамках цього етапу вибираються k кластерних центрів, які будуть відповідати k випадково обраним кластерам, після чого кожна точка призначається до найближчого центру кластерів.

Для того, щоб задати початкові наближення центрів кластерів найчастіше використовують такі способи: [Coppedge M.,1997:с.41]

- безпосередньо задають центри кластерів;
- задають кількість кластерів k та беруть як центри, координати k перших точок;

- задають кількість кластерів k та беруть як центри, координати k випадково обраних точок (доцільно здійснювати розрахунки для декількох випадкових запусків алгоритму).

Етап № 2 - Перерозподіл елементів. На даному етапі перераховується центри кластерів, використовуючи поточний розподіл кластерів. Потім - якщо критерій збіжності не задовільнений, то йде повернення до кроку, де кожний елемент призначався до найближчого центру кластера. Типовий критерій збіжності - відсутність або мінімальна порогова зміна нових кластерних центрів або мінімальне зменшення квадратичної помилки [Steinley D.,2006:c.5]. Для розрахунків відстаней між об'єктами найчастіше використовується форма евклідовавідстані.Процес обчислення центрів і перерозподілу об'єктів триває до тих пір, поки не виконується одна з умов:

- кластерні центри стабілізувалися, тобто відбувається така кількість ітерацій, поки кластерні центри стануть стійкими (тобто при кожній ітерації в кожному кластері опинятимуться одні й ті самі об'єкти)
- дисперсія всередині кластера буде мінімізована, а між кластерами — максимізована
- число ітерацій дорівнює максимальному числу ітерацій

Отже, як зазначалось раніше алгоритм k-середніх розділяє спостереження ітеративним способом, включаючи 2 кроки:

Крок № 1: присвоювання даних. Кожній точці даних присвоюється її найближчий представник при тому, що зв'язки порушуються довільно. Це призводить до поділу даних.

Крок № 2: переміщення «середніх». Кожен представник кластера переміщається до центру усіх точок даних, наданих йому. Пояснення цього кроку засноване на спостереженні, що дане безліч точок, єдиний кращий представник для цього безлічі (в сенсі мінімізації суми квадрата Евклидова відстані між кожною точкою і представником) не що інше, як серединна точка даних. Саме тому представник кластера часто взаємозамінні називають

серединним елементом кластера або центроїдом кластера, звідси і назва цього алгоритму [Дюран Н.,1997:с.62].

Наступне важливе питання, який не слід обходити стороною, - це попередня обробка даних перед застосуванням кластерного аналізу. Оскільки в якості міри відстані між об'єктами в методі К-середніх застосовується евклідова відстань, важливо врахувати випливають звідси обмеження. Використання даної метрики має сенс, коли:

• спостереження беруться з генеральної сукупності, що мають багатовимірний нормальний розподіл, змінні взаємно незалежні і мають рівні дисперсії;

• змінні однорідні за своїм фізичним змістом і однаково важливі для класифікації;

• всі змінні мають однакові одиниці виміру.

Крім цього, алгоритми не дають успішних результатів у разі наявності «мостів» між кластерами або в разі, коли кластери не мають сферичної форми. [Steinley D. - Br. J. Math. Stat. Psychol,2006:с.32].Мова йде про випадки, коли крім об'єктів, які представляють групи схожі між собою і відмінні від інших, є безліч об'єктів, які в багатовимірному просторі розташовуються між цими групами. Із зазначених проблем з’являється необхідність попереднього експлораторного аналізу первинних даних, за результатами якого досліднику необхідно обґрунтувати і прийняти рішення:

• про виключення з об'єктів явних викидів (в тому числі багатовимірних викидів) в разі їх наявності;

• стандартизація або зважування даних для забезпечення «однакових» одиниць виміру;

• робота з картельованими змінними - відбір змінних для аналізу або застосування редукції даних.

Безпосереднє застосування змінних при використанні методу К-середніх може привести до того, що змінні, які мають більший розмах значень і великі стандартні відхилення, можуть стати вирішальними для класифікації. Проте,

рішення про видалення викидів потрібно приймати, виходячи з конкретної ситуації. Цілком імовірна ситуація, коли викиди можуть бути представниками підгруп, про які просто недостатньо інформації у вибірці. [Захаров В.М., 2007:с.14]

Оскільки в багатьох дослідженнях методом К-середніх змінні, що описують об'єкт, вимірюють в різних одиницях, часто застосовується стандартизація змінних. Найчастіше застосовують так зване Z-перетворення, рідше зустрічаються варіанти перетворення за відхиленням абсолютних значень від медіани або за розмахом даних [Ворона В.М.,1998:с.10].

У той же час стандартизація змінних може привести до втрати ваги змінної з ростом дисперсії і величина впливу кожної змінної на рішення може змінюватися від одного набору даних до іншого.

Однозначної позиції з приводу стандартизації не існує. У 1996 році К. Шефер і П. Грін провели дослідження на десяти реальних наборах даних і показали, що дані не слід стандартизувати жодним чином. Пізніше в 2004 році Д. Стенлі піддав дане дослідження критиці, показавши, що результати є ненадійними, так як реальна структура в наборах даних була невідома дослідникам, і вказав, що найбільш оптимальною є стандартизація за розмахом значень. [Hans-Hermann Bock,2012:с.23]

Після підготовки даних перед дослідником стоїть завдання вибору алгоритму, яким буде реалізований метод К-середніх. На практиці дослідник вибирає конкретне програмне забезпечення, в якому алгоритм вже реалізований. Загальною рекомендацією тут є точна вказівка назви застосованого ПО, номер його версії, а також налаштування алгоритму (або вказівка про те, що використовувалися налаштування за замовчуванням). [Енюков И.С.,1998:с.51]

Це є важливим з огляду на те, що різне програмне забезпечення за замовчуванням застосовує різні алгоритми К-середніх:

• пакет IBM SPSS використовує за замовчуванням алгоритм Ллойда і в якості початкових центрів використовує перші k елемента набору даних, при

цьому є можливість перейти до алгоритму Маккуїна («Running means») і самостійно задати центри початкових кластерів ;

• програмне забезпечення Statistica використовує за замовчуванням алгоритм Ллойда і вибирає k перших елемента набору даних в якості початкових центрів;

• SAS має функції FASTCLUS (в якій реалізований алгоритм Маккуїна) і PROC FASTCLUS (в якій реалізований алгоритм Хартігана), початкові центри вибираються зі зменшенням щільності даних ;

• в Mathematica існує функція FindClusters, яка імплементує кластеризацию з альтернативним алгоритмом, званим К-медоіди. Цей алгоритм еквівалентний алгоритму Форджа / Ллойда, але він використовує об'єкти з набору даних в якості центрів кластерів, замість арифметичного середнього ;

• У Matlab функція "kmeans" використовує серійний алгоритм в першій фазі і потім застосовує ітеративний алгоритм в подальшій ;

• статистичний мову програмування R має функцію kmeans, в якій за замовчуванням застосовується алгоритм Хартігана-Вонга, початкові точки вибираються випадковим чином. [Кузнецов Д.Ю.,2011:с.56]

Як уже згадувалося, алгоритм К-середніх має кілька різних варіацій, і хоча мета кластеризації полягає в знаходженні структури, на ділі кожен метод привносить якусь структуру в дані. Тому важливо знати про різних алгоритмах методу К-середніх і їх особливості. Найбільш популярними є алгоритми Форджа і Ллойда, Маккуїна і Хартігана-Вонга. Відзначимо, що існують також більш нові специфічні алгоритми - генетичний метод К-середніх, сферичний метод К-середніх і ядерний метод К-середніх [Методы классификации и анализа климатических полей]

Алгоритм Ллойда також відомий як ітерація Вороного чи релаксація. Цей алгоритм названий на честь Стюарта П. Ллойда, який знайшов спосіб знаходження рівномірного розподілу множин точок у підмножини Евклідових просторів і розділення цих підмножин на структуровані опуклі комірки рівномірного розміру.

Алгоритм Ллойда послідовно знаходить центри кожного набору розподілу, а тоді перерозподіляє вхідні дані відповідно до того, які з цих центрів знаходяться найближче. [Наслєдов А.Д. ,2005:с.34] Відмінність між іншими алгоритмами полягає в тому, що вхідними даними для алгоритму Ллойда є неперервна геометрична площина, в той час як для багатьох алгоритмів — дискретна множина точок. Тому під час перерозподілу вхідних даних алгоритм Ллойда використовує діаграми Вороного(Діаграма Вороного — це особливий вид розбиття метричного простору, що визначається відстанями до заданої дискретної множини ізольованих точок цього простору. Вона названа на честь українського математика Георгія Вороного. Інші назви — теселяція Вороного, декомпозиція Вороного, чи теселяція Діріхле) , а не просто визначає найближчий центр до кожної скінченної множини точок, як це відбувається під час кластеризації іншими алгоритмами. [Нейский И.М.,2000:с.52]

Хоча даний алгоритм безпосередньо застосовується в Евклідовій площині, схожі алгоритми можна застосовувати і в багатовимірних просторах чи в просторах з неевклідовою метрикою. Алгоритм збігається повільно або відповідно до обмеження числової точності (може не збігатись взагалі). Саме тому у реальних програмах алгоритм Ллойда зазвичай зупиняється як тільки досягнуто "достатньо хорошого" розподілу. [Подвальный Е.С.,2005:с.49] Один із загальноприйнятих критеріїв завершення алгоритму - це зупинка, якщо максимальна відстань, на яку під час ітерації переміщається будь-яка точка, стає меншою за попередньо встановлену межу.

Зазвичай алгоритм Ллойда використовується в Евклідовому просторі. Евклідова відстань відіграє дві ролі в алгоритмі: вона використовується для задання комірок Вороного, але вона також відповідає за вибір центроїда як представницької точки для кожної комірки, так як центроїд - це точка, що мінімізує піднесену до квадрату середню Евклідову відстань до точок в її комірці. [Райзин Дж. Вэн. ,1980:с.5] Альтернативні відстані і альтернативні центральні точки (не центроїди) також можуть використовуватись.

Алгоритм Форджи має багато спільного з алгоритмом Ллойда, це обумовлюється тим, що алгоритми Ллойда і Фордж мають серійну модель центроїд, де центроїд розуміється, як геометричний центр набору об'єктів. Серійність алгоритму передбачає, що крок трансформації кластерів використовує всі об'єкти одночасно. [Соколова Л.В.,2003:с.11]При цьому іноді можливі різні варіанти ініціалізації кластерів: метод Фордж передбачає випадковий вибір n початкових центроїдів для кластерів з існуючого набору об'єктів; метод випадкового поділу, навпаки, випадковим чином формує набори кластерів, потім з них визначає початкові центроїди.

Алгоритм Хартігана і Вонга намагається знайти поділ об'єктів таким чином, щоб забезпечити мінімальне значення внутрішньогрупової дисперсії кластерів. Відзначимо, що окремий об'єкт при цьому може бути віднесений до іншого підпростору, навіть якщо належить до підпростору найближчого центроїда. [Сокэл Р.Р. ,1980:с.51]

Алгоритм МаКквіна є найбільш популярним, і головна відмінність від попередніх полягає в тому, що розташування центроїда перераховується після призначення кожного нового об'єкта в кластер. Однак, існує вдосконалені алгоритми МакКвіна. Ці алгоритми засновані на можливості обчислити центроїд кожного кластера. В основі K-means лежить ітеративний процес стабілізування центроїдів кластерів. Основною характеристикою кластера є його центр ваги і вся робота алгоритму спрямована на стабілізований або, в кращому випадку, повне припинення зміни центроїда кластера. У зв'язку з тим, що багато типів даних не належать просторам, в яких визначені їх значення, був розроблений ще один подібний алгоритм 'k - medioids'. На відміну від алгоритму k-mean, k-medoids вибирає точки даних як центри (medoids або exemplars) і працює з узагальненням мангеттенської норми, щоб визначити відстань між даними точками. Даний алгоритм більш стійкий до шуму та викидів в порівнянні з k-means, оскільки він мінімізує суму попарних розбіжностей замість суми квадратів евклідової відстані. [Суслов С.А. ,2001:с.51]

Найпоширенішою реалізацією k-медоідних кластеризацій є алгоритм Partitioning Around Medoids (PAM). PAM був створений Леонардом Кауфманом і Пітером Руссівом і він дуже схожий на алгоритм K-means, в основному тому, що обидва є алгоритмами кластеризації, іншими словами, обидва поділяють безліч об'єктів на групи (кластери) і робота обох заснована на спробах мінімізувати помилку, але PAM працює з об'єктами, які є частиною вихідної безлічі і представляють групу, в яку вони включені, а K-means працює з центроїдами - штучно створеними об'єктами, що представляють кластер. PAM є модифікацією алгоритму k-середніх, алгоритмом k-медіани (k-medoids). Цей алгоритм менш чутливий до шумів і викидів даних, ніж алгоритм k-means, оскільки медіана менше піддається впливам викидів. PAM ефективний для невеликих баз даних, але його не слід використовувати для великих наборів даних. Попереднє скорочення розмірності. Алгоритм працює з матрицею відстаней, його мета - мінімізувати відстань між представниками кожного кластера і його членами. [Coppedge M.,2011:c.72]

Існує також безперервний алгоритм К-середніх. Цей алгоритм швидше, ніж стандартна версія і, отже, розширює обсяг даних, які можуть бути кластеризовані. Від стандартного алгоритму він відрізняється методом вибору початкових опорних точок, і способом вибору даних для процесу оновлення. [Steinley D.,2006:c.63] При стандартному підході початкові опорні точки вибираються більш-менш довільно. У безперервному алгоритмі опорні точки являють собою випадкову вибірку з початкового безлічі даних. Якщо вибірка досить велика, розподіл цих початкових опорних точок повинно відображати розподіл точок у всьому наборі.

Ще одна відмінність між стандартним і безперервним алгоритмом k-середніх в способі обробки даних. Під час кожної повної ітерації стандартний алгоритм перевіряє всі дані точки поспіль. [Almond G., Verba S.,2000:c.29] З іншого боку, безперервний алгоритм розглядає тільки випадкову вибірку точок. Якщо набір даних великий і вибірка є репрезентативною, алгоритм повинен сходитися набагато швидше, ніж алгоритм, який перевіряє кожну точку по

порядку. Насправді, безперервний алгоритм модифікує метод оновлення центроїдів алгоритму Маккуїна на етапі ініціалізації, коли точки вихідних даних асоціюються з кластерами. [Coppedge M.,2011:c.72]

Отже, в будь-якому методі необхідно враховувати лише кілька точок при розрахунку і порівнянні відстаней. Вибір того чи іншого методу буде залежати від числа вимірювань даних, тому вибір та обґрунтування алгоритму кластеризації залежить в більшій мірі від компетенції дослідника.

1.3. Теоретико-методологічні засади поняття якості та оцінки кластерних рішень

Оскільки основною задачею кластеризації є поділ об’єктів на групи таким чином, щоб рівень подібності між об’єктами однієї групи був високий, а рівень подібності між об’єктами різних груп – низький, поняття якості кластеризації складається з таких ознак як компактність, відокремлюваність та концентрація. Однією з найбільших проблем кластеризації є те, що кластери формуватимуться, навіть якщо аналізований набір даних має повністю рандомну структуру. І щоб оцінити кластерне рішення потрібно насамперед провести оцінку загальної схильності наявних даних до об'єднання в кластери , а також провести візуальну оцінку тенденції (VAT, Visual Assessment of cluster Tendency) [Шитіков В. К., Мастіцкій С. Е.,2017]. Питання оцінки якості кластеризації не має наразі точного трактуванн, оскільки наприклад, тільки І.Д.Манделем дано огляд 45 функціоналів якості. Це свідчить про те, що не існує універсального критерію оптимізації кластерного рішення. У такій ситуації найкращим способом впевнитись в тому, що знайдене кластерне рішення є на даному етапі дослідження оптимальним, є тільки узгодженість цього рішення з висновками, отриманими за допомогою інших методів. Але повернемося до таких ознак як компактність, відокремлюваність та концентрація.

Компактність відображається у тому, що елементи кластера повинні знаходитись як можна ближче один до одного. Відповідно виміряти це можна як відстань всередині кластера , щільністю всередині кластера або об’ємом, що

займає кластер у багатовимірному просторі. Що стосується відокремлюваності, то вона характеризується найбільшої відстанню між кластерами, яка в свою чергу може бути виміряна як відстань між найближчими елементами кластера, як відстань між найбільш віддаленими елементами двох кластерів та як відстань між кластерними центрами або за допомогою матриці приналежності. Правило концентрації відображається у тому , що елементи кластера повинні бути сконцентровані біля центра кластера. Цей пункт використовують та перевіряють менше за усіх оскільки не в усіх алгоритмах кластеризації присутнє поняття центра.

Після створення кластерного рішення зазвичай виникає питання, наскільки воно стійке і статистично значиме. Тут існує емпіричне правило - стійка група повинна зберігатися при зміні методів кластеризації: наприклад, якщо результати ієрархічного кластерного аналізу мають частку збігів більше 70% з угрупованням за методом k середніх, то припущення про стійкість приймається.

У теоретичному плані проблема перевірки адекватності кластеризації не вирішена, принаймні, без використання іншого виду аналізу або апріорного знання приналежності об'єктів до відповідних груп. У літературі запропоновано безліч методів і критеріїв оцінки якості результатів кластеризації (clustering validation). Існують різні підходи оцінки кластерних рішень, які дають уявлення про ефективність вибору методу кластеризації. Серед найпопулярніших підходів оцінки результатів кластеризації виділяють три різні категорії оцінки кластерних рішень: «Внутрішня», «Зовнішня» та «Відносна».

- Внутрішня кластерна оцінка використовує внутрішню інформацію про кластеризації для оцінки ефективності структури кластеризації без посилання на будь-яку зовнішню інформацію (даний метод також може бути застосований для оцінки кількості кластерів та відповідного алгоритму кластеризації без зовнішніх даних). Тобто сюди відносяться метрики, що для оцінки якості використовують вже відому інформацію про структуру кластера.

- Зовнішня кластерна оцінка характеризується тим, що сюди відносяться метрики, які апріорі не мають знання про структуру кластера та під час оцінювання спираються лише на ту інформацію, яку можна отримати спираючись на множину даних.
- Відносна (непряма) оцінка кластеризації, яка оцінює структуру кластеризації, порівнюючи декілька кластерних структур між собою, не маючи апріорної інформації й беручі до уваги тільки інформацію про кластерну структуру. Тобто відносна оцінка якості кластеризації може відбуватись змінюючи різні значення параметрів для одного алгоритму (наприклад, змінюючи кількість кластерів).

Цей поділ є певним чином не чітким, в тому сенсі що якщо зовнішня метрика застосовується до різних структур, то вона може вважатись як відносна метрика, і навпаки – відносна метрика може вважатись зовнішньою, якщо вона основана на порівнянні показників, що отримуються для кожної структури окремо.

Мета алгоритмів внутрішньої оцінки вибору кластерів розділити набір даних на кластери, таким чином щоб, по-перше, об'єкти в одному кластері були максимально подібні, та, по-друге, об'єкти в різних кластерах були максимально відмінні. Таким чином, середня відстань у кластері повинна бути якомога меншою, а середня відстань між кластерами повинна бути якомога більшою.

Внутрішні заходи оцінки часто відображають компактність, зв’язність та відокремленість частин кластера. Компактність кластера вимірює наскільки близько розташовані об'єкти в одному кластері. Менша дистанція в кластері - це показник ефективної кластеризації. Зв’язність кластера вимірює як саме елементи та їх найближчі сусіди розміщуються в одному кластері і в просторі даних. Значення може набувати показника від 0 до нескінченності і це значення потрібно мінімізувати. Відокремленість кластера вимірює як сильно даний кластер відокремлений від інших кластерів. Даний показник включає в себе відстані між центрами кластерів, а також парні мінімальні відстані між об'єктами

в різних кластерах. Два найбільш часто використовувані індекси для оцінки ефективності кластеризації це «Аналіз силуету» та «Індекс Данна». Ці внутрішні міри можуть бути використані також для визначення оптимальної кількості кластерів у даних. [Климчук В.О.,2000:с.17]

Аналіз силуету вимірює, наскільки добре спостереження кластеризовано, і він оцінює середню відстань між кластерами. Графік силуету відображає міру того, наскільки близька кожна точка в одному кластері до точок в сусідніх кластерах. [Глаголєва І.І., 2003:с.198] Аналіз силуету може бути використаний для визначення ступеню відмінностей між кластерами.

Для цього потрібно:

1.Обчислити середню відстань від усіх точок даних в одному кластері (ai).

2.Обчислити середню відстань від усіх точок даних у найближчому кластері (bi).

3.Обчислити коефіцієнт.

Коефіцієнт може приймати значення в інтервалі [-1, 1]. [11]

- Якщо він дорівнює 0, то вибірка дуже близька до сусідніх кластерів.
- Якщо він дорівнює 1, то вибірка знаходиться далеко від сусідніх кластерів.
- Якщо він дорівнює -1, то зразок присвоюється неправильним кластерам.

Потрібно щоб коефіцієнти були максимально близькими до 1, тому що таким чином вибірка є максимально правильною. [Дод.3] Метод “Аналіз силуету” використовує середню відстань між кластером (a) та середню відстань кластера (b) для кожного зразка. Коефіцієнт силуету для зразка дорівнює (b - a) / max (a, b). Для уточнення, b - відстань між вибіркою та найближчим кластером, до якої вибірка не входить. Ми можемо обчислити середній коефіцієнт для всіх зразків і використовувати його як метрику для вибору кількості кластерів.

Оптимальну кількість кластерів використовуючи Метод “Аналіз силуету” можна визначити наступним чином:

1. Обчислити алгоритм кластеризації для різних значень k. Наприклад, змінюючи k від 1 до 10 кластерів.

2. Для кожного k обчислити середній силует спостережень.

3. Накреслити криву відповідно до кількості кластерів k.

4. Розташування максимуму вважається відповідною кількістю кластерів.

Щоб оцінити кластерне рішення можна також використати Індекс Данна. Для цього насамперед потрібно обчислити відстань між кожним із об’єктів кластера та об'єктами в інших кластерах. Наступним кроком буде обчислення мінімальна відстань між кластерами, а також відстані між об'єктами в одному кластері. Потім максимальна внутрішня кластерна відстань (тобто максимальний діаметр) використовується як компактність кластеру. [Сокэл Р.Р.,1987:с.199] Останнім кроком буде обчислення індексу Данна (D) використовуючи визначені дані за наступною формулою [Дод.1]

Якщо набір даних містить компактні та добре розділені кластери, то діаметр кластерів повинен бути невеликим, а відстань між кластерами навпаки. Таким чином, чим величина індексу Данна більша тим краще.

Як вже було описано в розділі вище, алгоритм k-means розділяє набір даних на окремі підгрупи, де кожен окремий елемент даних належить лише одній групі. Таким чином, кожен окремий елемент даних між кластерами групується із найбільш схожими елементами при цьому зберігаючи інші кластери якомога далеко один від одного. K-means призначає елемент даних кластеру таким чином, що сума відстані між точками даних в квадраті та центроїдом кластера (середнє арифметичне всіх точок даних, що належать до цього кластеру) є мінімальним. Чим менше варіацій у в кластерах, тим більш однорідними (подібними) є точки даних в межах одного кластера. Саме тому визначення

оптимальної кількості кластерів є неабияк важливим елементом в питаннях методології оцінки кластерних рішень та кластеризації, адже вибір оптимальної кількості кластерів в деякій мірі суб'єктивний і точність кінцевого результату і залежить від методу що використовується для вимірювання подібності та параметрів.

«Не існує правильних або неправильних результатів кластеризації, оскільки, за визначенням, це метод навчання без вчителя». Інформація визначається відповідністю отриманого рішення до поставленої задачі, яка на практиці в більшості випадків зводиться до того, щоб приблизно оцінити, на скільки груп доцільно розділити дані. При цьому ступінь цієї відповідності завжди буде суб'єктивним.

Заходи якості кластеризації можна умовно розділити на дві групи. Перша група - це якість по відношенню до деякого еталонного розбиття. У такому випадку для деяких наборів даних вже відомо справжнє розбиття на кластери і ми хочемо просто порівняти отриману кластеризацию з цим істинним розбиттям. Такі методи оцінок називаються External measures або мірою якості розбиття. Друга група - це заходи якості по відношенню до деяких наших уявлень про те, що таке хороша кластеризація, прийнята назва цих заходів - міри валідності розбиття або Internal measures.

Тепер вивчимо заходи валідності кластерів. Міра валідності кластерів вимірюється по відношенню до подання дослідника про те, що таке хороша кластеризація, тому нижче опишемо заходи валідності кластеризації.

По-перше, об'єкти всередині одного кластера повинні бути компактними, тобто дуже схожі один на одного, і, при цьому, різні кластери повинні розташовуватися далеко один від одного, тобто бути максимально несхожими. І різні заходи якості вважають ось ці критерії по-різному і тому їх дуже багато. Ці заходи можна використовувати для, наприклад, підбору параметрів алгоритмів, тобто ви проганяєте один і той же алгоритм кластеризації з різними параметрами, дивіться на заходи і вибираєте найкращу настройку алгоритму по відношенню до цього заходу. Друга міра - це індекс даних. Для того щоб її отримати, нам

потрібно навчитися рахувати відстань між кластерами і ми його визначимо як мінімальна відстань між об'єктами з двох різних кластерів. І друга величина - це діаметр кластерів, тобто ми будемо вважати уже максимальна відстань між усіма парами об'єктів, але вже всередині одного кластера. І тоді індекс буде дорівнювати просто відношенню мінімальної відстані між кластерами, поділене на максимальний діаметр кластера.

Для перевірки якості обраної кількості кластерів використовуєтсья Ліктьовий метод. Даний метод передбачає виконання алгоритму кілька разів по циклу, зі збільшенням кількості кластерів. Для кожного значення k визначаємо суму квадратичних помилок (в різних джерелах Dsitortion або SSE - sum of squared errors). Після цього, на основі цих даних будується графік, шкала кластеризації, як функції від кількості кластерів. Цей графік дає уявлення про те, яка кількість кластерів була б найбільш вдалою виходячи із суми відстані у квадраті (SSE) між елементами даних та центроїдами призначених кластерів. Вибираємо k там, де SSE починає зменшуватися і вирівнюватися. [Суслов С.А.,2005:с.17]

Щоб скористатися даним методом потрібно вибирати кількість кластерів, щоб додавання іншого кластеру не дало значно кращого моделювання даних. Себто, потрібно вибрати таку кількість кластерів, що утворять кут. Але "лікоть" не завжди можна однозначно ідентифікувати. Незначна варіація цього методу побудує викривлення дисперсії всередині групи. Ліктьовий метод дає змогу визначити кількість кластерів в алгоритмі K-means. Назву даний алгоритм отримав через свою візуальну складову, адже при правильному виборі кількості кластерів кінцевий графік набуває форми ліктя або ж кута. [Bock Hans-Hermann,2008:с.89] Таким чином, наша мета при використанні даного методу - вибрати мінімальне значення кластерів k, яке все ще має невеликий показник SSE.

Оптимальну кількість кластерів використовуючи Ліктьовий метод можна визначити наступним чином: 1) Обчислити алгоритм кластеризації для різних значень k. Наприклад, змінюючи k від 1 до 10 кластерів; 2) для кожного k

обчислити загальну суму в межах кластера квадрата (SSE); 3)Провести криву відповідно до кількості кластерів k; 4)Розташування згину (ліктя) на ділянці, як правило, потрібне нам число. [Бондаренко О.С.,2011:с.9]

Отже, Ліктьовий метод - це метод інтерпретації та перевірки узгодженості послідовності в рамках кластерного аналізу, призначений допомогти знайти відповідну кількість кластерів у наборі даних. Часто цей метод неоднозначний і не надійний через недолік неоднозначності вибору кількості кластерів у відповідності до SSE.

Індекс Девіса-Боулдіна заснований на приблизній оцінці відстаней між кластерами та їх дисперсності для отримання кінцевого значення. Обчислюється за формулою [Дод.2] Для того, щоб його порахувати, потрібно навчитися рахувати розкид даних усередині кластера. Але це буде просто середня відстань між об'єктами всередині кластера і центроїдом цього кластера. Тоді для того, щоб порахувати індекс, для кожного кластера нам потрібно знайти інший кластер, для якого максимальна величина сума розкиду цих кластерів, поділена на відстань між центроїдами цих двох кластерів. Ну, і потім це просто усереднювати. Отримуване співвідношення між відстанями в кластері, міжкластерними відстанями та обчисленням середніх загальних кластерів і є індексом Девіса-Боулдіна. Чим нижчий результат тим краще. [Глаголєва І.І.,2010:с.73] Недоліком індексу є не здатність зафіксувати багатовимірність даних, спричиняючи, наприклад, те, що два кластери з різною дисперсією в одному з вимірів простору характеристик мають однакове значення. Так як даний індекс він вимірює відстань між центроїдами кластерів, він обмежений використанням функції евклідової відстані.

Метою зовнішньої оцінки вибору кластера є порівняння ідентифікованих кластерів на основі попередніх даних. Найрозповсюдженим серед індексів є індекс Rand. Цей індекс в статистиці, і зокрема в кластеризації даних, є мірою подібності між двома кластеризації даних. Індекс Rand оцінює, наскільки багато з тих пар елементів, які перебували в одному кластері, і тих пар елементів, які

перебували в різних класах, зберегли цей стан після кластеризації алгоритмом. Насамперед потрібно побудувати матрицю із сполученням. Однак, індекс Ренда не позбавлений недоліків і одним з таких недоліків є те, що для випадкових розбиття ми можемо отримати досить високе значення індексу Ренда. Викорінити цей недолік покликане коригування.

Даний показник має область визначення від 0 до 1, де 1 - повний збіг кластерів із заданими класами, а 0 - відсутність збігів. Існує також модифікований варіант індексу. На відміну від звичайного індексу Rand, індекс Adjusted Rand може приймати негативні значення.

У застосуванні процедур кластерного аналізу також важливим аспектом є стійкість структури кластерів, що відображає реальну об'єктивність класифікації. В якості одного з можливих способів перевірки стійкості результатів кластерного аналізу може бути використаний метод порівняння результатів отриманих для різних алгоритмів кластеризації. Інші шляхи - бутстреп-метод, методи «складного ножа» і «ковзного контролю». [Захаров В.М. ,2011:с.45] Традиційними заходами якості є χ-квадрат і статистики Колмогорова-Смирнова, але розмірність даних зазвичай вимагає використання більш простих заходів,так як аналітичне рішення для оптимального розподілу відсутнє, немає можливості вивести точне рішення для розподілу вибірок.

Найбільш простий засіб перевірки стійкості кластерного рішення може полягати в тому, щоб вихідну вибірку випадковим чином розділити на дві приблизно рівні частини, провести кластеризацію обох частин і потім порівняти отримані результати. Більш трудомісткий шлях передбачає послідовне виключення спочатку першого об'єкта з кластеризації , таким чином щоб залишилися (m - 1) об’єктів. Далі,послідовно проводячи цю процедуру з виключенням другого, третього і т.д. об'єктів, аналізується структура всіх m отриманих кластерів. Інший алгоритм перевірки стійкості передбачає багаторазове розмноження, дублювання вихідної вибірки з m об'єктів, потім об'єднання всіх дубльованих вибірок в одну велику вибірку (псевдогенеральную сукупність) і випадкове витяг з неї нової вибірки з m об'єктів. Після цього

проводиться кластеризація цієї вибірки, далі витягується нова випадкова вибірка і знову проводиться кластеризація і т. Д. Очевидно, що це також досить трудомісткий шлях. [Ким Дж.О.,1989:с.34]

Рішення алгоритму К-середніх багато в чому також визначається вибором початкових точок в якості центрів первинних кластерів. При їх невдалому виборі може вийде неоптимальна кластеризація .

У разі невеликої кількості об'єктів можливий комбінаторний пошук набору початкових точок для оптимального розбиття, але при збільшенні набору даних такий пошук стає скрутним в обчислювальному плані [Дюран Н.,1997:с.62].В якості можливих рішень існують такі варіанти:

- призначати початкові точки на основі існуючих знань і теорій;
- використовувати в якості початкових точок центри кластерів, отримані в результаті ієрархічної кластеризації. Слід пам'ятати про те, що при великій кількості даних цей метод також не є оптимальним;
- обмежувати вибір початкових центрів ділянками з високою щільністю даних. У SAS реалізований алгоритм, який розраховує відстань між усіма точками, на основі цих відстаней обчислює щільність даних і привласнює центри кластерів, виходячи з найменшої щільності даних;
- «перевизначати» початкові центри за допомогою процедури бутстрепа;
- призначити випадковим чином до кластерних центрів або провести велику кількість випадкових розбиття (більше 5000) на групи, виходячи з розбиття визначити початкові точки кластерів і вибрати підсумкове рішення, яке дозволить отримати мінімальне значення функціоналу розбиття.

1.4. Висновки до розділу

Отже в рамках даного розділу було розглянуті основні положення поняття кластеризації методом k-means відповідно до яких кластеризацією методом k-середніх вважаємо неієрархічний метод кластеризації, що дозволяє розділити n спостережень на k кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього середнім

значенням. Цей метод базується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера, тобто функції.

Основана особливість метода полягає у тому, що кластерний аналіз методом k-середніх в якості цільової функції використовує мінімізацію внутішньокластерної дисперсії, тобто оптимальне розбиття досягається за допомогою мінімізації суми квадратичного відстані точок кластерів від центрів цих кластерів за допомогою ітеративної процедури, тобто алгоритм повторюється до тих пір, поки не буде виконано умови виходу. Також у даному розділі були зазначений покроковий алгоритм реалізації даного методу, який включає в себе 2 етапи : присвоєння даних та переміщення середніх.

Важливим етапом кластеризації є визначення кількості кластерів та перевірка результатів. Якщо підсумувати шляхів визначення числа кластерів, то їх можна поділити на умовні 3 групи : на основі попередньої інформації ; емпіричне визначення числа кластерів ; візуальне визначення числа кластерів. Завершальним етапом кластеризації а також даного розділу є перевірка отриманих результатів за допомогою внутрішніх, зовнішніх та відносних метрик валідності.

Отже, даний розділ описує та підсумовує теоретико-методологічне надбання сутності кластерного аналізу методом k-means

РОЗДІЛ ІІ. ОСОБЛИВОСТІ ЗАСТОСУВАННЯ МЕТОДУ К-СЕРЕДНІХ ДЛЯ КЛАСИФІКАЦІЇ ТИПІВ ПОЛІТИЧНИХ КУЛЬТУР

2.1. Теоретико-методологічні засади поняття політична культура

Вивчення особливостей типів політичної культури сучасного суспільства є найважливішою умовою формування ефективної політики, спрямованої на розв'язання основних завдань реформування держави. У сучасній політичній науці категорія політичної культури є однією з ключових у вивченні особливостей функціонування політичних систем, у поясненні й прогнозуванні політичних процесів. У сучасній науці існує широкий діапазон визначень терміна «політична культура», їх більше 50. Така різноманітність пояснюється тим, що, як зазначив Е. Баталов, «незважаючи на те, що проблематика політичної культури в її дифузійної формі вивчається ще з часів Платона і Конфуція, як цілісний феномен, що володіє автономним статусом, вона стала сприйматися наукою зовсім недавно [Стегній О.,2005:с.5].

Розробка категорії «політична культура» знайшла відображення в численних працях зарубіжних вчених (Г. Алмонд, С. Верба, Г. Пауелл, Л. Пай, У. Розенбаум, Е. Баталов, К. Гаджієв і ін.), а також у вітчизняних (Є. Головаха, В. Матусевич, Н. Паніна, О. Рудакевич, О. Кокорская, В. Кокорскій та ін.). Теоретичне обґрунтування і розроблення емпіричних показників політичної культури були здійснені в американській політології та політичній соціології у 1950–1970-х роках. [Головаха Є., Владико О., Щербак М.,2005:с.31]У цей період і в подальші роки були розроблені різні підходи до інтерпретації й дослідження феномену політичної культури. Основні відмінності між цими підходами полягають у розумінні того, яке коло явищ охоплює політична культура у сфері політичного життя суспільства, соціальних груп та особистості.

Існує широкий спектр підходів західних і вітчизняних вчених до трактування політичної культури. Півнева Л.Н. в величезній кількості її визначень виділяє чотири основні підходи:

Перший підхід пов'язаний з концепцією Г. Алмонда, яка визнана класичною переважною більшістю дослідників. Перше обґрунтування концепту політичної культури можна знайти у працях Й. Гердера й Ш.А. де Токвіля, проте джерела її сучасної ґенези пов’язані із соціологічною наукою і, насамперед, з статею Г. Алмонд, що була опублікована в 1956 р «Порівняльні політичні системи» [Головаха Е.И., Пухляк В.А,2001:с.17]. Під політичною культурою Г. Алмонд мав на увазі зразок орієнтацій на політичні дії, який відображає особливості кожної політичної системи. Базуючись на ідеях К. Клакхон, Н. Литтона, Т. Парсонса та ін., Г. Алмонд прийшов до висновку, що будь-яка політична система грунтується на «особливій формі орієнтацій на політичні дії». Цю орієнтацію на політику Г. Алмонд і назвав політичною культурою [Глаголєва І.І. ,2000:с.30]. Класичне визначення, дане Г. Алмонд спільно з Г. Пауеллом, таке: «Політична культура - це зразки індивідуальних позицій і орієнтацій щодо політики учасників даної політичної системи. Це суб'єктивна сфера, що утворює основу політичних дій і надає їм значення ».

Зазначені індивідуальні орієнтації, на думку американських вчених, включають в себе декілька компонентів:

- пізнавальна орієнтація - істинне або хибне знання про політичні об'єкти та ідеї;
- афективна орієнтацію - відчуття зв'язку, ангажованості, протидії і т.д. щодо політичних об'єктів;
- оціночна орієнтація - судження і думки про політичні об'єктах, які зазвичай передбачають використання оціночних критеріїв по відношенню до політичних об'єктів і подій »[Стегній О.,2005:с.17].

У 1963 р. Г. Алмонд і С. Верба опублікували працю «Громадянська культура» – кроснаціональне дослідження, в якому запропонували теорію політичної стабільності та демократії. Саме в цій праці політико-культурний підхід дістав найповніше втілення. Згідно з визначенням Г. Алмонда і С. Верби, яке вже стало класичним у соціології, політична культура – це політична система, яка інтеріорізована в знаннях, відчуттях, оцінках населення. Складовою

політичної культури є політичні орієнтації, що містять установки стосовно політичної системи та власної ролі в цій системі.

Таким чином, Г. Алмонд, С. Верба, Г. Пауелл обмежують політичну культуру сферою свідомості, вони зводять її до сукупності психічних станів індивіда, що проявляються на трьох рівнях - когнітивному (пізнавальному), афективному (емоційному) і оціночному. Інакше кажучи, політична культура зводиться до сукупності стійких політичних уявлень, переконань, почуттів і оцінок.

Другий підхід включає в політичну культуру зразки політичного поведінки. Ряд дослідників, наприклад, Е. Вятр і Д. Пол, вважають, що поряд з «зразками» політичної свідомості до політичної культури слід віднести і зразки політичної поведінки. Адже не все в нашій діяльності контролюється і фіксується свідомістю, а значить, не всі моделі поведінки індивідів і груп можна вивести з моделей їх свідомості. Найвдаліше за інших цю позицію сформулював відомий політолог Р. Такер. Він писав: «культура є звичний спосіб життя суспільства, що включає як прийняті способи мислення, а також переконання, так і прийняті зразки поведінки. Політична культура - це ті елементи культури, які мають ставлення до правління та політиці ».

Цікаві аргументи канадського вченого Р. Престуса, який включає в політичну культуру наступні компоненти [Стегній О.,2005:с.17].

- відношення мас до політики;
- ступінь їх участі в політичному житті;
- розуміння громадянами ефективності політики і відчуження від неї;
- законність, приписувану політичним елітам;
- характер політики всередині соціальної системи.

Третій підхід пов'язаний з широким трактуванням політичної культури. Наприклад, Л. Пай включає в неї «політичну ідеологію, національний характер і дух, національну політичну психологію і фундаментальні цінності народу». Ряд авторів трактують політичну культуру настільки широко, що це викликає заперечення. Л. Дітмер застосовує її щодо «національного характеру, впливу

колективного історичного досвіду на національну специфіку, а також емоційних або нормативних рамок взаємозв'язку між державою і громадянами». До вельми далекосяжних висновків дійшов Р. Патнем. [Стегній О.,2005:с.17] За його словами, політична культура покликана дати відповіді на питання: в чому полягає сутність людини, що таке суспільство і що лежить в його основі - гармонія чи конфлікт, що таке політична система і ін. Комплекс відповідей, за

Отже, політична культура - це частина духовної культури народу, що включає елементи, пов'язані з суспільно-політичними процесами, вона активно взаємодіє з іншими видами суспільної культури: економічною, правовою, релігійною і т.д. Політична культура — не статична система, вона здатна містити значний перетворювальний компонент. Але ця динаміка, як уже зазначалося, виявляється радше не на рівні системи загалом, а на рівні зміни внутрішніх ієрархічних зв’язків. У цьому разі зміна політичної культури полягає не у виникненні якісно іншої системи з абсолютно новими елементами, а в переході елементів з одного «рівня впливу» на інший. Саме в цьому разі й забезпечується наслідування політичної історії, оскільки політична культура є механізмом підтримки цілісності всієї політичної системи. І чим складнішою та багатобічною є політична культура, тим більшими адаптаційними ресурсами володіє політична система. [Ворона В.М.,1998:с.62]

Політична культура нерозривно пов’язана із загальнонаціональною культурою, соціокультурними, національно-історичними, релігійними, національно-психологічними традиціями, звичаями, стереотипами, міфами, установками. Невід’ємною складовою політичної культури є елементи політичної свідомості, які домінують у певному суспільстві, або є найпоширенішими серед членів певної соціальної групи. До них, насамперед, необхідно віднести сталі уявлення про різні аспекти політичного життя суспільства: про політичну систему, її окремі інститути, політичний режим тощо. Сталі політичні уявлення є частиною політичної культури, вони можуть відіграти істотну роль у соціальній практиці, багато в чому визначаючи стан політичної свідомості.

Як компонент політичної системи, вона несе в собі соціально-психологічні чинники політичного життя. Важливу роль у ній відіграє політична свідомість (сукупність уявлень, цінностей, переконань, установок і т.п.). Воно відображає владні відносини в суспільстві і політичні інтереси громадян. [Дембицкий С.,2016:с.140]

Нині політична культура є певним теоретичним поняттям, що не має однозначного й єдиного змісту. Здебільшого ця категорія використовується для пояснення того, чому відбулися ті чи інші події, склалися ті чи інші ситуації, і практично не виправдує себе в передбаченні змін політичної системи і перебігу соціально-політичних процесів. Концепція політичної культури зараз більше походить на енциклопедичний довідник — у ній зібрано уривчасті, несистематизовані дані з різних дисциплін і теорій: політичної психології, теорії національного характеру, концепції політичної соціалізації, антропології, етнографії, культурології, історії тощо. Але акумуляція знань ще не означає нового знання. [Ворона В.М.,1998:с.62] Тому вивчення трансформації політичних культур різних країн за умов змін, що відбуваються в посттоталітарних суспільствах, має стати необхідною ланкою, яка не тільки узагальнить нагромаджений досвід цієї проблематики, а й допоможе на підставі цього узагальненого матеріалу скласти об’єктивні оцінки і прогнози щодо розвитку сучасних світових тенденцій у сфері політичного життя. Адже за сорок років від часу виникнення першої концепції політичної культури проведено чимало як національних, так і порівняльних досліджень. Утім, головні висновки цих наукових праць, як правило, не виходили за межі зауважень на кшталт того, що за одними показниками є подібність, а за іншими — немає. Однак головна мета вивчення політичної культури не зводиться до виявлення установок, цінностей, орієнтації, очікувань і навіть моделей поведінки. Найважливішим є те, що вона може вможливити виявлення закономірностей розвитку і функціонування політичної системи, а зрештою — і соціальної.

Міркування з приводу специфіки політичних орієнтацій і політичної участі в тих чи інших суспільствах послугували відправним пунктом для формування

різних підходів і точок зору на структуру, типологію та характерні риси політичної культури в різних політичних системах, країнах і цивілізаціях.

Тому окремим дискусійним питанням є типологія політичної культури. Багато дослідників намагаються взагалі обходити цю проблему, віддаючи перевагу іншим аспектам теорії політичної культури. Найпоширенішою виявилася класифікація Г. Алмонда і С. Верби, які головним критерієм обрали, насамперед, міру участі громадян у політичному житті. Тип політичної культури можна визначити як переважно «підданський» з елементами «провінціалістської» та «патиціпаторної» складових масової політичної свідомості та поведінки. [Стегній О.,2005:с.17].

Якщо критерієм розвитку політичної культури вважати її відповідність вимогам сучасного демократичного суспільства, то тип політичної культури суспільства можна визначити як «амбівалентний», в якому співіснують суперечливі, а іноді й взаємовиключні елементи тоталітарної та сучасної демократичної політичної свідомості та поведінки [Головаха Є., Владико О., Щербак М.,2005:с.31]. Однак попри всю оригінальність і обґрунтованість їхня схема є звуженою, бо критерій, на якому вона побудована, є важливим, але показує лише один бік такого надзвичайно багатогранного й багатовимірного явища, як політична культура.

Інша класифікація належить У. Розенбауму, він запропонував, залежно від взаємовідносин між політичними суб'єктами, «інтегровані» і «фрагментарні» типи політичної культури. Інтегрований тип передбачає погодження більшістю суспільства базових цінностей, тоді як фрагментарна політична культура відрізняється гострою конфліктністю різних суб- культур по відношенню одна до одної, що є характерним для нестабільної політичної систе- ми з неусталеними політичними інститутами [Стегній О.,2005:с.17].

Особливий підхід до типології політичних культур розробив французький соціолог Р. Шварценберг, запропонувавши як критерій використовувати ступінь сприйнятливості політичної культури до інновацій, можливості адаптувати політичний досвід інших країн до наявних соціально-економічних і політичних

умов. На підставі даного підходу він виокремив «відкриту» і «закриту» політичну культуру [Дембицкий С.,2016:с.140]

Ще одним критерієм виокремлення типів політичної культури є базові цінності, на які орієнтується та чи інша спільнота в політичній діяльності або в політичному процесі. Відповідно до цього виокремлюють три типи політичної культури: громадянська, елітарна та архаїчна.

Альтернативну типологію запропонував польський соціолог Е. Вятр, який взяв за основу зв'язок політичних культур з політичними системами і суспільно-політичними формаціями, на яких вони ґрунтуються. У результаті він виокремив в якості основних три типи: традиційну, буржуазно-демократичну та політичну культуру соціалістичної демократії, які доповнюються другорядними: політичною культурою станової демократії, автократичної та реліктової автократичної [Головаха Є., Владико О., Щербак М.,2005:с.31].

У середині 90-х років С. Хантінгтон вказав на зв'язок демократичної політичної культури з ширшою системою культурних цінностей, насамперед, традиційних, пов'язаних з доктринальними та структурними аспектами тієї чи іншої релігії. Він відзначав існування відмінностей у характері функціональних взаємозв'язків традиційної й демократичної систем цінностей, зумовлених особливостями їхніх релігійних норм і цінностей. На підставі даних таксономічного аналізу С. Хантінгтон виокремив два основні типи політичної культури.

Перший тип – завершені політичні культури, що вирізняються релігійно-світоглядною позицією, відповідно до яких проміжна і кінцева мета в житті людини є взаємопов'язаними та взаємозалежними. Завершені культури є менш сприйнятливими до демократичних цінностей. Це системи, що ґрунтуються на релігійних цінностях католицизму, мусульманства, конфуціанства тощо.

Другий тип – інструментальні політичні культури, які характеризуються наявністю в системі цінностей значного сектору нормативних релігійних традицій, згідно з якими, проміжні цілі в життєдіяльності людини є відокремленими та незалежними від кінцевих цілей і не впливають на будь-яку

конкретну діяльність суб'єкта. Інструментальні політичні культури розглядаються як більш відкриті демократичним цінностям [Ворона В.М.,1998:с.62].

Серед новітніх доробків зарубіжних дослідників політичної культури варто виокремити роботу Ж.-К. Барб'є, присвячену комплексному підходу до політичних культур і різноманітності в Європі. На підставі великого масиву емпіричних даних, отриманого в багатьох країнах ЄС, демонструються можливості соціологічного й етнографічного аналізу в розумінні поточних і майбутніх завдань європейської інтеграції, яка сьогодні зводиться, насамперед, до функціонування економіки . Словенський дослідник Я. Коленц звертається до антропологічного підходу у вивченні політичної культури в транзитивний період, коли істотно зростає вагомість суб'єктивних чинників політичної консолідації. У свою чергу, Б. Вегенер у дослідженні політичних культур посткомуністичного транзиту пропонує використовувати евристичний потенціал підходу соціальної справедливості [Стегній О.,2005:с.17].

Якщо говорити про сучасну вітчизняну соціологічну думку, то тут сформувалися два теоретичних напрями аналізу політичної культури. Для першого (Л. Нагорна, В. Матусевич) характерними є уявлення про політичну культуру як сукупність знань, цінностей, політичного досвіду та традицій, принципів і способів політичної діяльності, функціонування політичних інститутів. Представники другого напряму (В. Бебик, О. Рудкевич) досліджують політичну культуру з позиції узагальнюючих політико-психологічних характеристик індивіда, ступеня його політичної розвиненості, вміння реалізувати власні політичні знання в межах функціонуючої системи. Тут першочерговим є звернення до понять «національна культура» і «менталітет», які сприяють формуванню певних політичних орієнтацій та в кінцевому результаті реалізуються в політичній поведінці [Дембицкий С.,2016:с.140].

2.2. Методика соціологічного вимірювання типів політичної культури за допомогою соціологічного тесту «Типи політичних культур»

У сучасній політологічній науці категорія політичної культури є однією із основних у процесі вивчення особливостей функціонування політичних систем, під час пояснення й прогнозування політичних процесів. Теоретичне обгрунтування і розроблення емпіричних показників політичної культури були здійснені в американській політології та політичній соціології у 1950-х –1970-х роках. Існують різні підходи до інтерпретації й дослідження феномена політичної культури. Головні розбіжності в цих підходах полягають у розумінні того, яке коло явищ охоплює політична культура в сфері політичного життя суспільства, соціальних груп та особистості. [Стегній О.,2005:с.17].

Традиційний підхід, пов'язаний із першими дослідженнями політичної культури, обмежує об'єкт вивчення винятково суб'єктивними компонентами — сукупністю установок соціальних суб'єктів щодо певної політичної системи. Ширший підхід включає до політичної культури також особливості політичної поведінки людей, чим виводить її до сфери об'єктивних феноменів політичного життя. Якщо говорити про перші кроки , пов’язані с методологією вимірювання типів політичних культур, то слід зазаначити, що Ф. Хьюнкс і Ф. Хікспурс на основі концепції Алмонда та Верби запропонували певну сукупність індикаторів даного явища. Як індикатор орієнтації щодо політичної системи в цілому вони розглядали інтерес індивідів до політики. Як індикатор орієнтації щодо "виходу" системи використовувався показник довіри до державних інституцій і управлінському апарату. Індикатором орієнтації щодо власної політичної компетентності виступала оцінка можливості особистої участі в політичному житті і впливу на політику. Кожен їх виділених типів політичних культур описувався особливим поєднанням значень даних індикаторів.

Сучасна українська картина вимірювання типів політичних культур тісно пов’язана з іменем Є. Головахи, оскільки для визначення основного типу політичної культури українського суспільства дослідницький колектив відділу методології та методів соціології Інституту соціології НАН України впровадив спеціальну тестову методику, що робить можливим вимірювання за двома шкалами: 1) тоталітарна – демократична; 2) активна – пасивна. Перша шкала

визначає напрям політичного розвитку суспільства, його декларовану мету, друга – ступінь готовності брати участь у реалізації цієї мети, активно протидіяти тенденціям розвитку суспільства в іншому напрямі. [Ворона В.М.,1998:с.62].

У результаті виділено чотири основних типи політичної культури, що характеризують минуле, дійсне і декларованe майбутнє політичної системи суспільства:

1) активна тоталітарна – є характерною для епохи «сталінізму», коли масова політична свідомість орієнтована на тоталітарну систему і готова активно підтримувати її існування;

2) пасивна тоталітарна – характерна для періоду так званого «застою», коли тоталітарна система залишається основним орієнтиром політичної свідомості, але брати активну участь у її відтворенні та захищати від ворожих посягань «мовчазна більшість» не має наміру;

3) пасивна демократична – сучасна політична культура, у рамках якої в основному при- ймаються декларовані демократичні принципи, але для їхнього практичного втілення немає критичної маси політично активних суб’єктів;

4) активна демократична – тип політичної культури в розвинутих демократичних державах, де більшість громадян готові активно захищати відповідну йому політичну систему та протидіяти різним формам деградації демократичних інститутів.

Що стосується методології даного інструментарію то на попередньому етапі створення інструментарію емпіричного дослідження сформульовано 50 суджень стосовно ставлення до демократії, та 50 суджень стосовно ставлення до політичної активності. [Головаха Є., Владико О., Щербак М.,2005:с.31].Інструмент фіксував ступінь згоди з кожним із цих суджень за п’ятибальною шкалою. Для реєстрації відповідей була використана шкала Лайкерта: "повністю згоден", "скоріше згоден", "важко сказати, згоден чи ні", "швидше за не згоден "," абсолютно не згоден ". Щоб уникнути монотонності у відповідях на послідовність питань до першого блоку включено 25 суджень, згода з якими є індикатором підтримки демократії, та 25 таких суджень, незгода

з якими є індикатором підтримки демократії. Відповідно, до другого блоку увійшли 25 суджень, згода з якими є індикатором позитивного ставлення до політичної активності, та 25 суджень, незгода із якими свідчить про позитивне ставлення до політичної активності респондента. Після відповідного перекодування обчислювали адитивні індекси підтримки демократії та політичної активності. Обидва індекси змінюються в інтервалі від 1 до 5. У цьому разі 1 означає повну підтримку демократії або ж високий ступінь готовності до активних політичних дій, а 5 – неприйняття демократії та, відповідно, неготовність до активних політичних дій. [Дембицкий С.,2016:с.140]

Для визначення конструктної валідності інструменту вимірювання колектив відділом методології та методів соціології Інституту соціології НАН України було проведено пілотажне опитування десяти респондентів з різними соціально-демографічними характеристиками. Опитування засвідчило, що формулювання запитань не викликали в респондентів складнощів у розумінні або несприйняття.

Після цього був проведений аналіз стійкість окремих показників, надійність адитивних індексів (модель надійності, що ґрунтується на узгодженості окремих пунктів шкали і оцінюється коефіцієнтом альфа Кронбаха), кореляцію адитивного індексу з відповідним однофакторним рішенням (фактори виділяли методом головних компонентів). [Стегній О.,2005:с.17]

Враховували також змістовну збалансованість шкали та показники зовнішньої валідності, яку оцінювали кореляцією отриманих у ході аналізу варіантів шкали зі спеціально включеними в опитувальник питаннями стосовно політичних орієнтацій респондентів.

Надалі структура отриманих шкальних оцінок досліджувалася за допомогою факторного аналізу та аналізу внутрішньої узгодженості шкали. Згідно з вихідною гіпотезою, всі судження за шкалою «демократія – тоталітаризм» та всі судження за шкалою «активність – пасивність» мають давати однофакторне рішення. [Головаха Є., Владико О., Щербак М.,2005:с.31].

Такий результат говорив би про те, що ці судження групуються у загальні фактори, які характеризують певну загальну «демократичну» та «активну» орієнтованість особи.. При цьому дотримувалася умова змістовної «збалансованості» шкали.

Кластер 1 – «активно-демократичний тип політичної культури». Респонденти, що потрапили до цієї типологічної групи, сприймають демократію і є активно політично зорієнтованими.

Кластер 2 – «активно-недемократичний тип політичної культури». Респонденти, що потрапили до цієї групи, є помірно політично активними і не сприймають демократію (проте їх недемократизм – теж помірний).

Кластер 3 – «пасивно-недемократичний тип політичної культури». Респонденти, що потрапили до цієї типологічної групи, не сприймають демократію і є політично пасивними.

Кластер 4 – «пасивно-демократичний тип політичної культури». Респонденти, що по- трапили до цієї групи, сприймають демократію, проте є політично неактивними.

За результатами пілотажних досліджень побудовано дві шкали, кожна з яких містила по шість пунктів (тобто шість суджень). Побудовані на підставі цих суджень обидва індекси мали високу надійність (вимірювалися в рамках моделі альфа Кронбаха) та дуже високо корелювали з відповідними однофакторними рішеннями (фактори виділяли методом головних компонентів). Саме цей інструмент застосовано в загальноукраїнському дослідженні і саме в такому двовимірному просторі виконували емпіричну класифікацію респондентів для віднесення їх до одного з чотирьох виділених теоретично типів політичної культури. Опитувальник створювали за безпосередньої участі та під керівництвом Є. І. Головахи. [Головаха Є., Владико О., Щербак М.,2005:с.31] Аналіз результатів пілотажу, оцінювання надійності та оптимізацію інструменту – за безпосередньої участі та під керівництвом А.П. Горбачика.

Побудований інструмент пройшов апробацію на вибірці України у міжнародному порівняльному проекті «Європейське соціальне дослідження».

Враховуючи можливі чотири типи політичної культури, було побудовано чотири кластери. В основу кластерізації покладено факторні рішення.

Хоча всебічне вивчення політичної культури на основі використання різних методів і підходів політичної науки триває вже понад чотири десятки років, все одно залишається відкритим питання про сутність і зміст цього поняття та підстави його концептуалізації, також не менш актуальним залишається використання в даній проблематиці методів сучасного аналізу таких як неїєрархічна кластеризація методом k-середніх.

2.3. Висновки до розділу

У даному розділі було розглянуто теоретико-методологічні засади поняття політичної культури та типів політичної культури відповідно до яких політична культура - це частина духовної культури народу, що включає елементи, пов'язані з суспільно-політичними процесами, вона активно взаємодіє з іншими видами суспільної культури: економічною, правовою, релігійною і т.д Оскільки існує безліч різних підходів до інтерпретації й дослідження феномена політичної культури у розділі наведені дослідження.

Окремим питанням у розділі розглянуто питання типології політичної культури згідно з яким найпоширенішою виявилася класифікація Г. Алмонда і С. Верби. Сучасна українська картина вимірювання типів політичних культур тісно пов’язана з іменем Є. Головахи, оскільки для визначення основного типу політичної культури українського суспільства дослідницький колектив відділу методології та методів соціології Інституту соціології НАН України впровадив спеціальну тестову методику, що робить можливим вимірювання даного феномену за двома шкалами: 1) тоталітарна – демократична; 2) активна – пасивна. Це та багато іншого доводить актуальність дослідження поняття типології політичних культур сучасного українського суспільства.

РОЗДІЛ ІІІ. ПРАКТИЧНЕ ЗАСТОСУВАННЯ ПІЗНАВАЛЬНОГО ПОТЕНЦІАЛУ МЕТОДУ КЛАСТЕРИЗАЦІЇ K-MEANS ДЛЯ КЛАСИФІКАЦІЇ ТИПІВ ПОЛІТИЧНИХ КУЛЬТУР

3.1. Реалізація кластерного аналізу методом k-means для класифікації типів політичних культур.

Мета практичної частини полягає у демонстрації пізнавального потенціалу методу неієрархічної кластеризації k-means на прикладі класифікації типів політичної культури.

Тобто використовуючи типологію, що була запроваджена під керівництвом Є. Головахи в межах розробки соціологічного тесту «Типи політичних культур» провести кластеризацію методом k-середніх за допомогою трьох найбільш розповсюджених алгоритмів: Макквіна, Ллойда та Хартігана-Вонга , оцінити отримані результати для того щоб обрати найбільш приданий алгоритм для класифікації типів політичних культур методом k-means.

Завдання:

- провести кластеризацію методом k-середніх за допомогою алгоритмів: Маккуїна, Ллойда та Хартігана-Вонга та описати отримані результати
- порівняти отримані результати за допомогою зовнішніх та внутршніх метрик валідності. В якості зовнішньої метрики валідності виступає узгодженість з теоретичним розумінням кількості та структури кластерів. Індекс Дана та Аналіз Силуету відповідають за внутрішні метрики валідності відповідно.
- обрати та описати особливості найбільш придатного алгоритму k-means для класифікації типів політичних культур.

Джерело інформації – дані, що були отримані в межах соціологічного моніторингу Інституту соціології НАН України «Українське суспільство» за 2018 рік.

Гіпотези :

1) Результати кластеризації алгоритмами Маккуїна, Ллойда та Хартігана-Вонга будуть відрізнятись різною наповненістю кластерів, оскільки вони за різною формулою рахують міжгрупову та міжкластерну відстань.
2) Алгоритм Хартігана-Вонга є найбільш придатним для класифікації типів політичних культур, оскільки він намагається забезпечити мінімальне значення внутрішньогрупової дисперсії кластерів та є стійким.
3) Кластер респондентів з пасивно недемократичними поглядами буде найменшим.

Хід роботи:

Для того, щоб провести кластеризацію мною було обрано три найбільш розповсюджені алгоритми кластеризації методом k-means - Макквіна, Ллойда та Хартігана-Вонга, вибір саме цих алгоритмів ґрунтується на тому, що ці три алгоритми найбільше різняться с точки зору вибору центру для кластера.

Оскільки алгоритм Ллойда послідовно знаходить центри кожного набору розподілу, а тоді перерозподіляє вхідні дані відповідно до того, які з цих центрів знаходяться найближче. Алгоритм Хартігана і Вонга намагається знайти поділ об'єктів таким чином, щоб забезпечити мінімальне значення внутрішньогрупової дисперсії кластерів. А алгоритм Макквіна перераховує розташування центроїда після призначення кожного нового об'єкта в кластер.

Слід зауважити, що ми не обирали алгоритм Форджи, оскільки алгоритм Форджи має багато спільного з алгоритмом Ллойда, це обумовлюється тим, що алгоритми Ллойда і Фордж мають серійну модель центроїд, де центроїд розуміється, як геометричний центр набору об'єктів, тому використання хоча і розповсюдженого алгоритма не є доречним у даному випадку.

Що стосується програмного забезпечення, то у своїй роботі я користувалась статистичною мовою програмування Rstudio, оскільки пакет IBM SPSS використовує за замовчуванням тільки алгоритм Ллойда і в якості

початкових центрів використовує перші k елемента набору даних, а статистична мова програмування R за замовчуванням використовує алгоритм Хартігана-Вонга, де початкові точки вибираються випадковим чином, але є можливість виконувати кластеризацію і за допомогою алгоритму Макквіна та Ллойда. Також у Rstudio є можливість використовувати індекси перевірки отриманих результатів.

Попередня робота з даними

Оскільки ми отримали масив даних формата sav, для роботи в Rstudio для початку нам необхідно було імпортувати дані з SPSS, що досягається за допомогою використання скрипта для імпортування даних [Додаток №4].

Початковий масив мав 1800 спостережень, однак, після того, як ми відфільтрували невідповіді ми отримали 1691 спостережень. Цікавим фактом є те, що у 2018 році автори дослідження внесли зміни у набір запитань, які відповідають шкалі демократії/тоталітаризму. Запитання були перероблені таким чином, що відображали полярне значення попередніх версій. Ці зміни було пов'язані з змістовним розумінням респондентами поняття «демократії», оскільки більш рані версії описували «ідеальну модель демократії» не беручи до уваги українські реалії. Але не зважаючи на зміну формулювань автори тесту зберегли внутрішню узгодженість та факторну валідність для нової шкали. Отже, змінні, що використовувались у кластеризації в рамках практичної роботи :

V3	Демократія надає надто багато свободи багатим людям
V21	Мені все одно, яка буде влада, якби не стало гірше
V4	Демократичні вибори — це фарс, що не захищає інтереси простих людей
V22	Я не сподіваюсь на вибори, тому що не вірю, що від їх результатів зміниться моє життя
V5	Я свій вибір давно уже зробив, тому не хочу брати участь у сьогоднішньому політичному житті

V23	Немає сенсу боротися за свої права, якщо влада своїми діями явно їх ігнорує
V6	Демократія — це тільки слова, якими прикриваються ті, хто має доступ до влади для забезпечення власних інтересів
V24	Шляхом голосування ми обираємо владу, ну а далі від нас уже нічого не залежить
V7	Демократичні вибори, як правило, приводять до влади найбільш корисливих людей
V25	Будь-яка спроба щось змінити у політичному житті країни потребує від людини занадто великих жертв, які найчастіше виявляються марним
V8	Демократія надає людині оманливу свободу вибору, якою вона не може скористатися
V26	Демократія несе простій людині непевність у завтрашньому дні

Оскільки у кластерному аналізі методом k-means необхідно мати гіпотезу стосовно кількості кластерів, ми користаємся методологією СТ ТПК та виокремлюємо 4 кластери. Для того, щоб говорити чи є отримані результати певним алгоритмом оптимальними для інтерпретації вводимо «ідеальну модель» кластерів з зазначеними центрами кластерів. Під «ідеальною моделлю клатерів» розуміємо кластери, що мають центри з крайніми координатами за шкалою демократичності та активності СТ ТПК.

Оскільки вищезазначені змінні вимірюються за допомогою п’ятибальної шкали Лайкерта (де 1 – повна згода з твердженням, а 5 – повна незгода з ним), де сума оцінок кожного окремого судження дозволяє виявити установку респондента з якого-небудь питання, то будемо вважати за ідеальні центри кластерів мають такі координати :

С1 = (1;1) – центр кластера, що характеризує пасивні демократичні установки;

С2 = (1;5) – центр кластера, що характеризує активні демократичні установки;

С3 = (5;1) – центр кластера, що характеризує пасивні антидемократичні установки;

С4 = (5;5) – центр кластера, що характеризує активні антидемократичні установки.

Після того,як дані готові для використання і ми визначились з кількістю кластерів, ми можемо перейти до самої кластеризації.

Для виконання практичної роботи окрім завантаженого R та R-studio (версія 3.4.1) нам знадобились деякі бібліотеки : library (ltm), library(tidyverse), library(magrittr), library(cluster), library(cluster.datasets), library(cowplot), library(ggfortify), library(clustree).

3.2 Реалізація кластеризації алгоритмом Макквіна

Виконуючи кластеризацію методом Макквіна ми отримуємо таблицю центрів чотирьох кластерів за кожною ознакою. Аналізуючи таблицу бачимо,що:

До Кластера № 1 увійшли ті, хто підримує ідеї демократії (оскільки координати центрів за змінними V 3,4,5,6,7,8(шкала демократії) коливаються від 1.2 до 1.4) але не готовий активно відстоювати свою погляди (оскільки координати центрів за змінними V 21,22,23,25,27,28(шкала активності) коливаються від 1,6 до 2) Отже, Кластер № 1 можна характеризувати як пасивних демократів, процент яких дорівнює 27,2 %. [Додаток № 5]

Якщо говорити про Кластер № 2 , то до нього потрапили 22,7 % респондентів, котрі не визначились зі своїм ставленням до демократії та не визначились стосовно своєї установки на активні та пасивні дії, оскільки координати центрів за шкалою демократії коливаються від 3 до 3,4, а за шкалою активності: від 2,6 до 3,5. На мою думку, цей кластер неможливо співвіднести з наявними кластерами СТ ТПК. Але заперечувати його також не слід, оскільки

ми не проводили попередню перевірку на оптимальну кількість кластера а використовували вже наявну установку про кількість кластерів у розмірі чотирьох.

Кластер № 3 утворюють респонденти, що підтримують ідеї демократії (координати центрів за шкалою демократії: від 1,7 до 1,9), але відстоювати свої ідеї не готові, що демонструє нам шкала активності, яка коливається від 2,6 до 3,7. Процентна частка таких респондентів дорівнює 24,3. Якщо порівнювати з результатами СТ ТПК то Кластер № 3 важко співвіднести з одним з чотирьох кластерів СТ ТПК.

З певною невизначеністю, але тяжіє до демократичних погладів поєднуючі з пасивністю дій (координати центрів за шкалою демократії коливаються від 1,7 до 2,8, а за шкалою активності: від 1,7 до 2) установки респондентів, що потрапили до Кластеру № 4 , частка якого становить 25,8 %. Цей кластер також важко віднести до існуючої типології СТ ТПК, оскільки він не належним чином відтворює наповненість кластерів.

Отже, за допомогою алгоритму Макквіна ми отримали один кластер,який з легкістю можна співвіднести з класифікаціює СТ ТПК та 3 певним чином схожих між собою кластерів ,які важко інтерпретувати тим паче в рамках методології СТ ТПК. Також суттєвим недоліком даного алгоритму є те, що алгоритм прагне створювати кластери однакого розміру, що досить сильно спотворює інтерпретацію результатів.

3.3 Реалізація кластеризації алгоритмом Ллойда

Реалізовуючи кластеризацію даним методом ми отримали чотирьохкластерне рішення [Додаток № 6], де:

Кластер № 1 характеризується пасивними демократичними поглядами оскільки за шкалою демократії значення центрів варіюється від 1,3 до 1,9 , кластер № 1 становить 28 % опитаних респондентів.

Кластер № 2 становить абвівалентне наповнення (41%) , оскільки ми бачимо дуже велику відстань між крайніми координатами центрів за шкалою

демократії (від 1,2 до 4,1) , тобто алгоритм зарахував у цей кластер і тих, хто підримує демократичні погляди, і тих, хто тяжіє до протилежних. Що стосується шкали активності – погляди є невизначеними. Через те, що Кластер № 2 становить дві різні групи респондендентів стосовно їх ставлення до демократії, Кластер № 3 опинився порожнім. Тобто Кластер № 2 вирогідніше за усе містить частково і тих респондентів, що мали опинитись у Кластеру № 3.

Кластер № 4 утворюють 31% респондентів, що з певною невизначеністю, але тяжіють до демократичних поглядів , але у проявах залишаються пасивними, оскільки за шкалою активності координати центрів варіюється від 1,6 до 2,6.

Отже, якщо підсумувати результати кластеризації алгоритмом Ллойда, то можна зазначити, що на відміну від алгоритму Макквіна даний алгоритм не прогне утворювати кластери однакового розміру, але він утворює порожній кластер та кластер, в якому прослідковуються центри кластерів, що за нашими умовами мали становити окремі групи. Тим самим цей алгоритм ускладнює інтерпретацію кластеризації.

3.4 Реалізація кластеризації алгоритмом Хартігана-Вонга

Використовуючи алгоритм Хартігана –Вонга ми отримуємо [Додаток № 7,8] , що до Кластера № 1 увійшли ті, хто підримує ідеї демократії (оскільки координати центрів за шкалою демократії коливаються від 1.3 до 1.7) але не готові активно відстоювати свою погляди (оскільки координати центрів за шкалою активності коливаються від 1.5 до 2.1) . Отже Кластер № 1 можна характеризувати як пасивних демократів, процент яких дорівнює 26,1.

Якщо говорити про Кластер № 2 , то до нього потрапили 18,5 % респондентів, котрі не визначились зі своїм ставленням до демократії та не визначились стосовно своєї установки на активні та пасивні дії. На мою думку, цей кластер неможливо співвіднести з наявними кластерами СТ ТПК.

Кластер № 3 характеризується демократичними поглядами (центри кластерів змінних,що характеризують шкалу демократії сягають від 1.5 до 1.9) та невизначенністю стосовно установки на напрям дій, але скоріше тяжіють до

активних намерів (оскільки координати центрів за шкалою активності сягають відмітки від 2.7 до 3.8) , тобто цю групу можна охарактеризувати як активних демократів і їх частка становить 27, 3%

З певною невизначеністю, але тяжіє до демократичних поглядів поєднуючі з пасивністю дій характеризується Кластер № 4 , частка якого становить 28, 1 % . Даний кластер дуже схожий за наповненням на Кластер № 2, тому я висуваю гіпотезу про те, що якщо виокремити ще один , Кластер № 5 , до нього увійдуть ввсі ті, хто не визначився зі своїми поглядами , та також це вирішить проблеми рівномірного розподілу по кластерам, оскільки ми бачимо , що алгоритм частково прогне наповнити кластери однаковим чином.

Для того, щоб зрозуміти чи є сенс у виділенні п’ятого кластера скористаємось критерієм кам'янистої осипи. Цей критерій має витоки з факторного аналізу. Для графічного визначення оптимальної кількості факторів/ кластерів Каттель у 1966р. запропонував використовувати графік кам'янистої осипи. Даний критерій полягає в пошуку точки, де спадання власних значень сповільнюється найбільш сильно. [Додаток № 10]

Знову проводимо кластеризацію , але додаємо ще один кластер, який характеризує невизначеність з приводу питань демократії та активності , центри такого цнтри становлять C5 = (3;3). Отримуємо такі результати:

До Кластеру № 1 потрапили пасивні демократи, процентна частка яких становить 21,8. Кластер № 2 становить 15,6 % респондентів, що є невизначеними стосовно своїх поглядів. До Кластеру № 3 потрапили респонденти, чиї погляди є демократичними, і хто готовий активно відстоювати свою позицію. Кластер №4 характеризується пасивністю та невизначеністю стосовно позиції, однак, V4= 3,8 може свідчити про тяжіння до антидемократичних поглядів. Для того , щоб більш точно дослідити , чи існує кластер, який характеризується антидемократичними поглядами необхідно звузити п’ятибальну шкалу Лайкерта до трьохбальної, центри кластерів в такому випадку будуть виглядати наступним чином [Додаток № 9]:

С1 = (1;1) – центр кластера, що характеризує пасивні демократичні установки;

С2 = (1;3) – центр кластера, що характеризує активні демократичні установки;

С3 = (3;1) – центр кластера, що характеризує пасивні антидемократичні установки;

С4 = (1;3) – центр кластера, що характеризує активні антидемократичні установки.

С5 = (2;2) – центр кластера, що характеризує невизначеність стосовно і активної, і демократичної позиції.

Кластер № 1 становить 38,6 % респондентів, що дотримуються демократичних поглядів, при цьому активно відстоювати свою позицію респонденти даного кластеру не готові. Цей кластер умовно можно співвіднести з СТ ТПК як пасивні демократи.

До Кластеру № 2 потрапили 7,1 % респондентів, чиї погляди є протилежними до демократичних та респонденти даної групи поділяють активну позицію, цю групу можна охарактеризувати як активні антидемократи.

Кластер № 3 становить 15 % респондентів, що підримують ідеї демократії та готові активно їх відстоювати, тобто Кластер № 3 відповідає активним демократам відповідно до типології СТ ТПК.

Пасивними антидемократами можна назвати 3,8 % респондентів, що утворюють Кластер № 4, оскільки координати центрів за шкалою активності варіюються від 1,1 до 1,5, а за шкалою демократії – від 2,8 до 2,9. Враховуючи такі результати можна вважати підтвердженою гіпотезу № 3, відповідно до якої даний кластер є найменшим серед усіх.

Найбільшим за розміром – 35,5 % респондентів, є Кластер № 5 , до якого увійшли ті респонденти, чиї погляди є невизначеними і за шкалою демократії,і за шкалою активності.

3.5 Оцінка результатів кластеризації за допомогою внутрішніх метрик валідності.

Наразі ми перевірили тільки зовнішню валідність, яка відображається у узгодженості кластерів з теоретичним описом. Відповідно до цих результатів отримали що алгоритм Хартигана-Вонга є найбільш придатним. У деяких випадках, звісно, зовнішньої валідності може бути достатньо, наприклад якщо наявна велика кількість схожих досліджень та обгрунтове теоретичне розуміння кількості та наповненості кластерів. Однак, оскільки наразі не так багато літератури в рамках класифікації типів політичних культур методом неієрархічної кластеризації та немає усімавизнаної кількості кластері для того щоб говорити про кінцеві висновки нам необхідно застосувати внутрішні метрики валідності. В якості таких метрик я обрала Індекс Дана та Аналіз Силуету Індекс Дана працює таким чином, що він порівнює міжкластерну відстань відповідно до діаметру кластера, в той час аналіз Силуету оцінює середню відстань між кластерами. Для цього я використовували пакет cValid.

Даний пакет дає можливість використовувати одночасно декілька алгоритмів кластеризації, для того щоб виявити найкращий або найбільш відповідний цілям алгоритм. Також він може використовуватись як інструмент для визначення кількості кластерів за необхідністю. В результатів отримуємо, що індекс Дана найвищий при алгоритмі Хартігана-Вонга, оптимальної кількістю вважається 4 кластера, Аналіз силуета також демонструє найкращій результат (він є найбільш наближеним до 1) під час кластеризації алгоритмом Хартігана-Вонга, проте оптимальним відображає п’ятикластерне рішення.

```
Cluster sizes:
 2 3 4 5 6 7 8 9 10

Validation Measures:
                                  2       3       4       5       6       7       8       9      10

MacQueen            Connectivity  4.1829 10.5746 13.2579 20.1579 22.8508 25.8258 32.6270 35.3032 38.2905
                    Dunn          0.3595  0.3086  0.3282  0.2978  0.3430  0.3430  0.4390  0.4390  0.5804
                    Silhouette    0.5098  0.5091  0.4592  0.4077  0.4077  0.3664  0.3484  0.4060  0.3801
HartiganandWong     Connectivity  7.2385 10.5746 15.8159 20.1579 22.8508 25.8258 33.5198 35.3032 38.2905
                    Dunn          0.3670  0.3086  0.6884  0.5978  0.3430  0.3430  0.3861  0.4390  0.2804
                    Silhouette    0.5122  0.5091  0.4260  0.4077  0.4077  0.3664  0.3676  0.4060  0.3801
Lloyd               Connectivity  7.2385 14.1385 17.4746 24.0024 26.6857 32.0413 33.8913 36.0579 38.6607
                    Dunn          0.2070  0.1462  0.2180  0.2180  0.2978  0.2980  0.4390  0.4390  0.4390
                    Silhouette    0.5122  0.3716  0.4250  0.3581  0.3587  0.3318  0.3606  0.3592  0.3664

Optimal Scores:

             Score  Algorithm       Clusters
Connectivity 1.2829 MacQueen         2
Dunn         2.127  HartiganandWong  4
Silhouette   0.7533 HartiganandWong  5
```

Для того щоб перевірити гіпотезу про наявність 5 кластера, який наявний у кластеризації методом Хартигана-Вонга необхідно перевірити чи не є виокремлений п’ятий кластер шумом. Це можна зробити за допомогою k-medoids або алгоритму PAM, оскільки k-medoids обирає точки даних як центри (medoids або exemplars) і працює з узагальненням мангеттенської норми, щоб визначити відстань між даними точками, тобто він мінімізує суму попарних розбіжностей замість суми квадратів евклідової відстані. PAM є менш чутливий до шумів і викидів даних, оскільки його мета - мінімізувати відстань між представниками кожного кластера і його членами, тому медіана менше піддається впливам викидів. Але достатнім у даному випадку також можна вважати аргументація п’ятого кластера С. Дембіцьким , який у роботі «Конструювання, валідизація та інтерпретація соціологічного тесту Типи політичної культури – II» наводить ряд доказів наявності ще одного кластера. У своїй роботі С. Дембіцький [С. Дембіцький, 2018 с.4] наводить приклади апріорной кластеризації, що представлена ідеальної точкої поділу та концептуальною типологією, та апостеріорной кластеризацією, що відображена у латентному аналізі відповідно до яким виокремлюється п’ятий невизначений кластер.

3.6 Висновки до розділу

В рамках практичної частини було продемонстровано пізнавальний потенціал кластеризації k-means трьома найбільш розповсюдженими алгоритмами: Хартігана- Вонга, Макквіна та Ллойда. Використовуючи ці три алгоритми ми отримали різну наповненість кластерів, що дозволяє підтвердити нашу Гіпотезу № 1. Підтвердженням Гіпотези № 2 є те, що найбільш придатним для застосування для класифікації типів політичних культур виявився алгоритм Хартігана-Вонга, в результаті кластеризації яким ми отримали п’ятикластерне рішення, де :

Кластер №1 – «пасивно-демократичний тип політичної культури». Респонденти, що потрапили у цю групу, сприймають демократію, проте є політично неактивними. Вони становлять 38,6% опитаних.

Кластер №2 можна інтерпретувати як «активно-недемократичний тип політичної культури». Респонденти, що потрапили у цю групу, помірно політично-активні і не сприймають демократію (проте їх недемократизм теж помірний). Вони становили 7,1% опитаних.

Кластер №3 змістовно інтерпретується як «активно-демократичний тип політичної культури». Респонденти, що потрапили у цю типологічну групу, сприймають демократію і активно політично зорієнтовані. Вони становили 15,2 % опитаних.

Кластер №4 – «пасивно-недемократичний тип політичної культури». Респонденти, що потрапили у цю типологічну групу, не сприймають демократію і є політично пасивними. Вони становили 3,6% опитаних.

Кластер №5 – тип сучасної політичної культури, що характеризується невизначеністю поглядів. Даний тип є найбільшим та становить 35,5 % респондентів. Звісно, використовуючи лише кластеризацію ми не можемо з певністю стверджувати, що в даний кластер потрапили люди, яким байдуже чи їх погляди є амбівалентними, адже для того, щоб дати відповідь на це запитання необхідно проводите більш комплексне дослідження, що буде ставити за мету пояснення каузальних зв’язків. [Додаток 11]

Тобто, використовуючи алгоритм Хартігана – Вонга ми отримали п’ятикластерне рішення, яке в повній мірі відтворює типологізацію СТ ТПК та є найбільш оптимальним з точки зору застосування внутрішніх метрик валідності. Нижче представлена порівняльна таблиця застосування трьох алгоритмів в якій продемонстровані основні висновки застосування різних алгоритмів під час класифікації типів політичних культур.

	Опис отриманої структури та узгодженість з теоретичної концепцією ТПК	Виявлені особливості застосування при класифікації ТПК	Результати внутрішніх метрик валідності

Макквіна	4х кластерне рішення, де лише 1 кластер узгоджений з класифікаціює СТ ТПК та 3 певним чином схожих між собою кластери ,які важко інтерпретувати	алгоритм прагне створювати кластери однакого розміру, що досить сильно спотворює інтерпретацію результатів	Індекс Дана 1,752 Аналіз Силуету 0,4621
Хартіга на-Вонга	5ти кластерне рішення, в якому 4 кластери повністю відповідають класифікацію СТ ТПК + кластер, що характеризується невизначеністю поглядів респондентів.	Потребує у перевірці гіпотези про кількість кластерів.	Індекс Дана 2, 127 Аналіз Силуету 0,7533
Ллойда	4х кластерне рішення, в якому 1 кластер повністю відповідає СТ ТПК , 1 кластер відтворює невизначену групу, 1 кластер, який важко інтерпретувати + 1 порожній кластер	утворює порожній кластер та кластер, в якому прослідковуються центри кластерів, що за нашими умовами мали становити окремі групи. Тим самим цей алгоритм ускладнює інтерпретацію кластеризації.	Індекс Дана 1,583 Аналіз Силуету 0,4392

Необхідно зазначити, що використовуючи будь-який з цих алгоритмів кластеризації ми отримуємо різні результати, але кожен з цих результатів може бути придатним для інтерпретації в більшій або меншій мірі, але для цього необхідно підбирати індивідуальні кроки кластеризації , бути готовим експериментувати та можливо виходити за рамки використання 1 методу. (сюди можна віднести і готовність піддати сумніву гіпотезу про кількість кластерів, і часткове змінення шкали ознаки, і використання більш сучасних методик задля боротьби з «пустими» кластерами тощо). Дуже важливою особливістю є планування необхідного розміру вибірки, виходячи з кількості змінних, які будуть застосовуватися в аналізі та чітке розуміння кількості та наповненості кластерів. Основним моментом залишається необхідність проводити не тільки змістовну інтерпретацію відмінностей в різних кластерах, а й проводити оцінку надійності та валідності кластерних рішень за допомогою інших методів аналізу та більш сучасних методик.

Отже, кластерний аналіз методом k-середніх залишається найбільш поширеним методом кластеризації, що володіє достатньою ефективністю і здатний працювати як самостійний метод та як доповнення до більш складних методів, застосування якого було наведено нами на прикладі класифікації типів політичних культур.

ЗАГАЛЬНІ ВИСНОВКИ

Кластеризація методом k-середніх - неієрархічний метод кластеризації, що дозволяє розділити *n* спостережень на *k* кластерів, так щоб кожне спостереження належало до кластера з найближчим до нього середнім значенням. Цей метод базується на мінімізації суми квадратів відстаней між кожним спостереженням та центром його кластера, тобто функції. Помітний поштовх у розвитку робіт по кластерному аналізу дали роботи Р.Розенблатта. Сучасні алгоритми теорії розпізнавання образів, класифікації та кластерного аналізу базуються на роботах С.А.Айвазяна , А.Я.Червоненкіса, В.Н.Вапніка , Р.А. Фішера , В.Н.Фоміна, І.Форджі , К.Фукунагі, Дж.Хартігана, Дж.Хопфілда , Я.З.Ципкіна та ін.

Даний метод дозволяє розділити довільний набір даних на задане число кластерів таким чином, що об'єкти всередині кластера були досить близькі один до одного, а об'єкти різних кластерів не перетиналися. Іншими словами, мета алгоритму - об'єднати в групи подібні дані за певними заздалегідь заданими критеріями.

Кластерний аналіз методом К-середніх в якості цільової функції використовує мінімізацію внутішньокластерної дисперсії, тобто оптимальне розбиття досягається за допомогою мінімізації суми квадратичного відстані точок кластерів від центрів цих кластерів за допомогою ітеративної процедури. Алгоритм k-середній ефективний перш за все тому, що він не потребує

обчисленні всіх попарних відстаней між спостереженнями, на відміну від більшості інших алгоритмів кластеризації, включаючи той, що використовується в процедурі ієрархічного кластерного аналізу. Основними перевагами методу є його простота, швидкість виконання, наявна можливість паралелізації, зручність для кластеризації великої кількості спостережень. Але не зважаючи на переваги метод має й певні недоліки до яких можна віднести те, що результат класифікації сильно залежить від випадкових початкових позицій кластерних центрів, залежність від компетенції дослідника, адже кількість кластерів повинна бути заздалегідь визначена, від вибору різних початкових центрів та прагнення методу створити кластери рівного розміру, навіть якщо це не оптимально.

Слід зазначити, що алгоритм k-середніх має кілька різних варіацій, і хоча мета кластеризації полягає в знаходженні структури, кожен метод привносить якусь певну структуру в дані, вона може відрізнятися, але це не означає перевагу якогось певного алгоритму. Адже вибір алгоритму кластеризації це індивідуальний крок, який має зробити дослідник, розуміючи наслідки кожного алгоритму. Вибір того чи іншого алгоритму має ґрунтуватися на меті дослідження, специфіці масиву даних та враховувати функціональні можливості програмного забезпечення.

Найбільш популярними є алгоритми Форджи і Ллойда, Макквіна і Хартігана-Вонга. Відзначимо, що існують також більш нові специфічні алгоритми - генетичний метод К-середніх, сферичний метод К-середніх і ядерний метод К-середніх.

Якщо говорити про особливості методу, то важливим фактом є вибір початкових центрів кластерів, адже це перший кров в алгоритмі кластеризації і саме від цього в більшій мірі залежать результати кластеризації. Важливим етапом кластеризації є визначення кількості кластерів та перевірка результатів. Якщо підсумувати шляхів визначення числа кластерів, то їх можна поділити на умовні 3 групи : на основі попередньої інформації ; емпіричне визначення числа кластерів ; візуальне визначення числа кластерів.

Важливим моментом є необхідність приводити результати експлораторного аналізу даних (або будь-якого іншого методу перевірки оптимальної кількості кластерів).Також необхідно враховувати, що різне програмне забезпечення, що використовувалося при обчисленні результатів кластерного аналізу дає нам різні функціональні можливості. Основним моментом залишається необхідність проводити не тільки змістовну інтерпретацію відмінностей в різних кластерах, а й проводити оцінку надійності та валідності кластерних рішень за допомогою інших методів аналізу та більш сучасних методик.

Оскільки основною задачею кластеризації є поділ об’єктів на групи таким чином, щоб рівень подібності між об’єктами однієї групи був високий, а рівень подібності між об’єктами різних груп – низький, поняття якості кластеризації складається з таких ознак як компактність, відокремлюваність та концентрація. Однією з найбільших проблем кластеризації є те, що кластери формуватимуться, навіть якщо аналізований набір даних має повністю рандомну структуру. І щоб оцінити кластерне рішення потрібно насамперед провести оцінку загальної схильності наявних даних до об'єднання в кластери , а також провести візуальну оцінку тенденції. У теоретичному плані проблема перевірки адекватності кластеризації не вирішена, принаймні, без використання іншого виду аналізу або апріорного знання приналежності об'єктів до відповідних груп. У літературі запропоновано безліч методів і критеріїв оцінки якості результатів кластеризації (clustering validation). Існують різні підходи оцінки кластерних рішень, які дають уявлення про ефективність вибору методу кластеризації. Серед найпопулярніших підходів оцінки результатів кластеризації виділяють три різні категорії оцінки кластерних рішень: «Внутрішня», «Зовнішня» та «Відносна».

Кластерний аналіз методом k-means має широкий спектр застосування , до якого відноситься й дослідження такого феномену як типи політичних культур. Сучасна українська картина вимірювання типів політичних культур тісно пов’язана з іменем Є. Головахи, оскільки для визначення основного типу політичної культури українського суспільства дослідницький колектив відділу

методології та методів соціології Інституту соціології НАН України впровадив спеціальну тестову методику, що робить можливим вимірювання даного феномену за двома шкалами: 1) тоталітарна – демократична; 2) активна – пасивна. Відповідно до теоретичних надбань СТ ТПК та методологічних засад методу k-середніх в рамках практичної роботи було розглянуто пізнавальний потенціал методу в рамках класифікації типів політичних культур.

Основними висновками практичної частини є те, що використовуючи будь-який з алгоритмів кластеризації для класифікації типів політичної культури ми отримуємо різні результати, але кожен з цих результатів може бути придатним для інтерпретації в більшій або меншій мірі, але для цього необхідно підбирати індивідуальні кроки кластеризації враховуючи недоліки алгоритмів. Для кластеризації будь-яким алгоритмом k-means важливим є чітке розуміння кількості кластерів та їх наповненості. Навить маючи аргументовану теоретичну базу краще перевіряти гіпотезу про кількість кластерів. Це можна зробити за домагаю критерію кам'янистою осипи (на основі графіка залежності відносини дисперсій до числа кластерів або зменшення суми внутрікластерних дисперсій). Важливим моментом залишається необхідність проводити не тільки змістовну інтерпретацію кластерів, а й проводити оцінку валідності кластерних рішень за допомогою внутрішніх метрик валідності. Для класифікації типів політичних культур можна застосовувати індекс Дана та Аналіз Силуету. Індекс Дана працює таким чином, що він порівнює міжкластерну відстань відповідно до діаметру кластера, в той час аналіз силуету оцінює середню відстань між кластерами.

Результати практичної роботи продемонстрували, що застосуваючи кластеризацію методом k-середніх для класифікації типів політичних культур краще використовувати алгоритм Хартігана-Вонга, оскільки він не прагне створювати кластери однакого розміру як алгоритм Макквіна, не утворює порожні кластери як алгоритм Ллойда, показує кращі результати с точки зору перевірки внутрішньої валідності за допомогою індекса Дана та аналізу Силуету. У той час кластеризація алгоритмом Хартігана-Вонга потребує перевірки

гіпотези про кількість кластерів та звуження п’ятибальної шкали Лайкерта до трьохбальної щоб визначити крайні позиції респондентів. Використовуючи алгоритм Хартігана – Вонга при кластеризації типів політичних культур ми отримали п’ятикластерне рішення, яке в повній мірі відтворює типологізацію СТ ТПК та є найбільш оптимальним з точки зору інтерпретації класифікації типів політичних культур. Найбільшими за розміром виявились кластери, що характеризують пасивно-демократичний ТПК та виділений окремий п’ятий кластер невизначених респондентів. Звісно даних результатів недостатньо для релевантних висновків про типи політичних культур у сучасній Україні, але це і не було метою роботи. В рамках роботи ми з’ясували , що метод k-means може використовуватись для класифікації типів політичних культур, оскільки він дозволяє побачити певну структуру кластерів, їх наповненість та дає змогу перевірити гіпотезу про кількість кластерів. Враховуючи можливості та обмеження методу, найбільш придатним для застосування є алгоритм Хартігана-Вонга за вищезазначених причин.

Отже, метод неієрархічної кластеризації методом k-середніх має ряд певних переваг та обмежень, особливостей застосування та специфіки індивідуального вибору алгоритму відповідно до дослідницьких цілей. Але не зважаючи на все це, кластерний аналіз методом k-середніх залишається найбільш поширеним методом кластеризації, що володіє достатньою ефективністю і здатний працювати як самостійний метод та як доповнення до більш складних методів, застосування якого було наведено нами на прикладі класифікації типів політичних культур. Пізнавальний потенціал даного методу не є вичерпним тому потребує більшої уваги зі сторони дослідників. Дані результати практичної роботи можна використовувати в майбутньому для одного з етапів дослідження типів політичних культур України.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ:

1. Алексеёнок А.А. Поиск взаимозависимостей и класификация обьектов эмпиричаского социологического исследования / А.А. Алексеёнок / [Електронний ресурс]. – Режим доступу: http://cyberleninka.ru/article/n/poisk-vzaimozavisimostey-i-klassifikatsiya-obektov-empiricheskogo-sotsiologicheskogo-issledovaniya.pdf
2. Бондаренко О.С. Методи кластерного аналізу/ О.С. Бондаренко, В.В. Слєсарєв / [Електронний ресурс]. – Режим доступу: http://www.rusnauka.com/11_EISN_2011/Informatica/1_84590.doc.htm
3. Буреева Н.Н. Многомерный статистический анализ / Н.Н. Буреева / [Електронний ресурс]. – Режим доступу: http://www.unn.ru/pages/issues/aids/2007/57.pdf
4. Волкова Н.А Кластерный аналіз результатов социологического опроса работников предприятия / Н.А. Волкова, О.В. Стукач - Вестник Ульяновского государственного технического университета, 2005.- С. 68-72.
5. Ворона В.М., Головаха Є.І. Соціологія і політика у суспільстві, що трансформується // Соціологія: теорія, методи, маркетинг. — 1998. — № 1–2. — С.9–17
6. Глаголєва І.І. Застосування кластерного аналізу для опрацювання даних земельного кадастру / І.І. Глаголєва, А. Ю. Берко [Електронний ресурс]. – Режим доступу: http://science.lp.edu.ua/sites/default/files/Papers/plugin-47_58.pdf

7. Головаха Е. Отношение к власти и политический выбор молодежи Украины // Социология: теория, методы, маркетинг. — 2002. — № 1.— С.117–127.
8. Головаха Е.И., Пухляк В.А. Проблемы политической социализации и формирования политической культуры в Украине //Політична думка. - 1994. - N 2. - С. 26-29.
9. Головаха Є., Владико О., Щербак М. Типи політичної культури. Підхід до вимірювання — 2005. — № 1.— С.117–127
10. Головаха Є.І. Народ як суб'єкт політичного процесу та специфічні риси його політичної культури в сучасній Україні // Політична культура і політичні партії України. — К.: НДІ "Проблеми людини", 1997. — С.30–34.
11. Головаха Є.І. Стратегія соціально-політичного розвитку України. — К.: Абріс, 1994, [Електронний ресурс]. – Режим доступу: http://science.lp.edu.ua/sites/default/files/Papers/plugin-47_58.pdf
12. Дюран Н. Кластерный анализ/ Н. Дюран – М.: Статистика, 1977, С. 127.
13. СЕРГІЙ ДЕМБІЦЬКИЙ,. Апріорна та апостеріорнакластеризація результатів соціологічного тесту “Типи політичної культури”: порівняльний аналіз [Електронний ресурс] / СЕРГІЙ ДЕМБІЦЬКИЙ, // Соціологія: теорія, методи, маркетинг. – 2017. – Режим доступу до ресурсу: http://stmm.in.ua/archive/ukr/2017-1/12.pdf?fbclid=IwAR3pZZ53u-79itu4gFRIrXXro2nKSKG7CEghKpVhr_C4O7f0zqK7xymSDmI.
14. СЕРГІЙ ДЕМБІЦЬКИЙ. Ко нструювання, валідизація та інтерпретація соціологічного тесту “Типи політичної культури – II” [Електронний ресурс] / СЕРГІЙ ДЕМБІЦЬКИЙ. – 2018. Режим доступу до ресурсу: http://stmm.in.ua/archive/ukr/20181/7.pdf?fbclid=IwAR3YgzPkx8jpXhYARB58hn-CTkzFL40Kt2QPUBfCB3Ja7udXH9T1V-qGI-I
15. Енюков И.С. Факторный, дискриминантный и кластерный анализ/ И.С. Енюков - М:Финансы и статистика, 1989, С. 234.
16. Захаров В.М. Многомерный анализ данных методами прикладной статистики / В.М. Захаров / [Електронний ресурс]. – Режим доступу:

http://cs.kai.ru/files/Shalagin/Posobie_mnogoparametrich_analis_dannyh_methodami_prikl_stat-ki.pdf

17. Ким Дж.О. Факторный, дискриминантный и кластерный анализ. / Дж.О. Ким, Ч.У. Мьюллер, У.Р. Клекка, М.С. Олдендерфер, Р.К. Блэшфилд.- М.: Финансы и статистика, 1989, С. 215.
18. Климчук В.О. Кластерний аналіз: використання у психологічних дослідженнях / В.О. Климчук / [Електронний ресурс]. – Режим доступу: http://eprints.zu.edu.ua/3297/1/9.pdf
19. Кузнецов Д.Ю. Кластерный анализ и его применение/ Д.Ю. Кузнецов, Т.Л. Трошина / [Електронний ресурс]. – Режим доступу: http://vestnik.yspu.org/releases/uchenue_praktikam/33_4/
20. Ломидзе О.Н. Кластерный анализ в социологических исследованиях / О.Н. Ломидзе / [Електронний ресурс]. – Режим доступу: http://cyberleninka.ru/article/n/klasternyy-analiz-v-sotsiologicheskih-issledovaniyah.pdf
21. Методы классификации и анализа климатических полей / [Електронний ресурс]. – Режим доступу: http://naukovedenie.ru/PDF/74TVN615.pdf
22. Наслєдов А.Д. SPSS: Компьютерный анализ данных в психологии и социальных науках/ А.Д. Наслєдов - СПб: Питер, 2005, С. 416.
23. Нейский И.М. Классификация и сравнение методов кластеризации / И.М. Нейский / [Електронний ресурс]. – Режим доступу: http://it-claim.ru/Persons/Neyskiy/Article2_Neiskiy.pdf
24. Подвальный Е.С. Сравнение алгоритмов кластерного анализа на случайном наборе данных / Е.С. Подвальный, А.В. Плотников, А.М. Белянин / [Електронний ресурс]. – Режим доступу: http://cyberleninka.ru/article/n/sravnenie-algoritmov-klasternogo-analiza-na-sluchaynom-nabore-dannyh.pdf
25. Райзин Дж. Вэн. Классификация и кластер / Дж. Вэн Райзи. – М.: Мир, 1980, С. 390.

26. Соколова Л.В. Використання методів кластерного аналізу у практичній діяльності підприємств/ Л.В. Соколова, Г.М. Верясова, О.Є. Соколов / [Електронний ресурс]. – Режим доступу: http://ena.lp.edu.ua:8080/bitstream/ntb/13865/1/37_240-246_Vis_720_Menegment.pdf
27. Сокэл Р.Р. Кластер-анализ и классификация: предпосылки и основные направления. В кн: Классификация и кластер / Р.Р. Сокэл, Дж.Вэн Райзина - М: Мир, 1980 - С. 7-19.
28. Стегній О. Динаміка політичної культури регіональних спільнот України. — / [Електронний ресурс]. – Режим доступу: *publications.lnu.edu.ua/bulletins/index.php/article1870*
29. Суслов С.А. Кластерный анализ: сущность, преимущества и недостатки / С. А. Суслов / [Електронний ресурс]. – Режим доступу: http://cyberleninka.ru/article/n/klasternyy-analiz-suschnost-preimuschestva-i-nedostatki.pdf
30. Дембицкий С. Социологические тесты: сущность и валидизация / С. Дембицкий // Социология: теория, методы, маркетинг. — 2016. — № 3. — С. 140–155.
31. Forgy E. W. Cluster analysis of multivariate data: efficiency versus interpretability of classifications / Forgy E. W. - Biometrics. - vol 21, 1965. - pp 768–769.
32. Lloyd S. P. Least square quantization in PCM / Lloyd S. P. - IEEE Trans. Inf. Theory. - vol 28, no 2, 1982. - pp 129–137.
33. Steinley D. K-means clustering: A half-century synthesis / Steinley D. - Br. J. Math. Stat. Psychol. - vol 59, no 1, 2006. - pp 1–34.
34. Almond G., Verba S. The Civic Culture. Political Attitudes and Democracy in Five Nations. — Princeton, 1956.
35. Coppedge M. District Magnitude, Economic Performance, and Party-System Fragmentation in Five Latin American Countries // Comparative Political Studies. — 1997. Vol.30. — № 2. — P.301.

36. An Introduction to Cluster Analysis for Data Mining / [Електронний ресурс]. – Режим доступу: http://www-users.cs.umn.edu/~han/dmclass/cluster_survey_10_02_00.pdf
37. Bock Hans-Hermann Origins and extensions of thek-means algorithm in cluster analysis/ Hans-Hermann Bock [Електронний ресурс]. – Режим доступу: http://www.jehps.net/Decembre2008/Bock.pdf
38. Cluster Analysis :Basic Concepts andAlgorithms/ [Електронний ресурс]. – Режим доступу: https://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf
39. Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data / [Електронний ресурс]. – Режим доступу: http://www.tqmp.org/RegularArticles/vol11-1/p008/p008.pdf
40. Tryfos Peter Cluster analysis / [Електронний ресурс]. – Режим доступу: http://www.yorku.ca/ptryfos/f1500.pdf

ДОДАТКИ

Додаток № 1

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right)$$

Додаток № 2

$$DI_m = \frac{\min\limits_{1\leqslant i<j\leqslant m}\delta(C_i, C_j)}{\max\limits_{1\leqslant k\leqslant m}\Delta_k}$$

Додаток № 3

$$DI_m = \frac{\min\limits_{1\leqslant i<j\leqslant m}\delta(C_i, C_j)}{\max\limits_{1\leqslant k\leqslant m}\Delta_k}$$

Додаток № 4

```
library(foreign)

dat<-read.spss("US2016.sav", use.value.labels = F, to.data.frame = T,trim_values = TRUE, reencode = T, use.missings =T)

kdat<-as.data.frame(dat[,46:57])

kdat<-data.frame(kdat[complete.cases(kdat[,1:12]),])

str(kdat)

m1<-kmeans(kdat, 4, iter.max = 10, nstart = 1,algorithm = "MacQueen", trace=FALSE)

nx<-21

ny<-25

x<-kdat[paste("V",nx,sep="")][[1]]

y<-kdat[paste("V",ny,sep="")][[1]]

tt<-table(x,y)
```

```
tt1<-tt/sum(tt)*20
plot(x,y,col=m1$cluster,pch=16,type="p",cex=tt1)
abline(v=m1$centers[,nx-17],col=c(1:4))
abline(h=m1$centers[,ny-17],col=c(1:4))
#points((m1$centers),type="p")
names(dat)
data100<-sapply(dat[c(46:57)],as.numeric)
c1<-c(rep(1,12))
c2<-c(rep(5,12))
c3<-c(1,5,1,5,5,5,1,5,1,5,1,1)
c4<-c(5,1,5,1,1,1,5,1,5,1,5,5)
m<-rbind(c1,c2,c3,c4)
klu<-kmeans(algorithm =MacQueen,na.omit(data100),m)
klu$centers
round(prop.table(table(klu$cluster))*100,digits=1)
mydata <- na.omit(data100)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:10) wss[i] <- kmeans(mydata,
                                 centers=i)$tot.withinss
plot(1:10, wss, type="b", xlab="Кількість кластерів",ylab="Внутрішньогрупова
сума квадратів", pch=20, cex=2)
abline(v=4,lty=2)
names(dat)
data100<-sapply(dat[c(46:57)],as.numeric)
c1<-c(rep(1,12))
c2<-c(rep(5,12))
c3<-c(1,5,1,5,5,5,1,5,1,5,1,1)
c4<-c(5,1,5,1,1,1,5,1,5,1,5,5)
c5<-c(3,3,3,3,3,3,3,3,3,3,3,3)
```

```
m<-rbind(c1,c2,c3,c4,c5)
klu<-kmeans(na.omit(data100),m)
klu$centers
round(prop.table(table(klu$cluster))*100,digits=1)
mydata <- na.omit(data100)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
for (i in 2:10) wss[i] <- kmeans(mydata,
                             centers=i)$tot.withinss
V18<-as.numeric(dat$V18)
V18[V18==3]<-6
V18[V18==1|V18==2]<-3
V18[V18==4|V18==5]<-1
V18[V18==6]<-2
V19<-as.numeric(dat$V19)
V19[V19==3]<-6
V19[V19==1|V19==2]<-1
V19[V19==4|V19==5]<-3
V19[V19==6]<-2
V20<-as.numeric(dat$V20)
V20[V20==3]<-6
V20[V20==1|V20==2]<-3
V20[V20==4|V20==5]<-1
V20[V20==6]<-2
V21<-as.numeric(dat$V21)
V21[V21==3]<-6
V21[V21==1|V21==2]<-1
V21[V21==4|V21==5]<-3
V21[V21==6]<-2
```

```
V22<-as.numeric(dat$V22)
V22[V22==3]<-6
V22[V22==1|V22==2]<-1
V22[V22==4|V22==5]<-3
V22[V22==6]<-2
V23<-as.numeric(dat$V23)
V23[V23==3]<-6
V23[V23==1|V23==2]<-1
V23[V23==4|V23==5]<-3
V23[V23==6]<-2
V24<-as.numeric(dat$V24)
V24[V24==3]<-6
V24[V24==1|V24==2]<-3
V24[V24==4|V24==5]<-1
V24[V24==6]<-2
V25<-as.numeric(dat$V25)
V25[V25==3]<-6
V25[V25==1|V25==2]<-1
V25[V25==4|V25==5]<-3
V25[V25==6]<-2
V26<-as.numeric(dat$V26)
V26[V26==3]<-6
V26[V26==1|V26==2]<-3
V26[V26==4|V26==5]<-1
V26[V26==6]<-2
V27<-as.numeric(dat$V27)
V27[V27==3]<-6
V27[V27==1|V27==2]<-1
```

```
V27[V27==4|V27==5]<-3
V27[V27==6]<-2
V28<-as.numeric(dat$V28)
V28[V28==3]<-6
V28[V28==1|V28==2]<-3
V28[V28==4|V28==5]<-1
V28[V28==6]<-2
V29<-as.numeric(dat$V29)
V29[V29==3]<-6
V29[V29==1|V29==2]<-3
V29[V29==4|V29==5]<-1
V29[V29==6]<-2
data111<-data.frame(V18,V19,V20,V21,V22,V23,V24,V25,V26,V27,V28,V29)
c1_<-c(rep(1,12))
c2_<-c(rep(3,12))
c3_<-c(1,3,1,3,3,3,1,3,1,3,1,1)
c4_<-c(3,1,3,1,1,1,3,1,3,1,3,3)
c5_<-c(rep(2,12))
n<-rbind(c1_,c2_,c3_,c4_,c5_)
klu2<-kmeans(na.omit(data111),n)
klu2$centers
round(prop.table(table(klu2$cluster))*100,digits=1)
m<-rbind(c1,c2,c3,c4,c5)
klu<-kmeans(na.omit(data100),m)
klu$centers
round(prop.table(table(klu$cluster))*100,digits=1)
mydata <- na.omit(data100)
wss <- (nrow(mydata)-1)*sum(apply(mydata,2,var))
```

```
for (i in 2:10) wss[i] <- kmeans(mydata,
                                 centers=i)$tot.withinss

intern <- clValid(mammals_scaled, nClust = 2:24,
             clMethods = c("MacQueen"," HartiganandWong"," Lloyd"), validation =
"internal")# Summary
summary(intern) %>% kable() %>% kable_styling()Clustering Methods:
 hierarchical kmeans pam

plot(1:10, wss, type="b", xlab="Кількість кластерів",ylab="Внутрішньогрупова
сума квадратів", pch=20, cex=2)

abline(v=4,lty=2)

klu<-kmeans(na.omit(data100),m)

fviz_nbclust(mammals_scaled, kmeans, method = "silhouette", k.max = 24) +
theme_minimal() + ggtitle("The Silhouette Plot")
```

Додаток № 5

Алгоритм Макквіна

	V3	V21	V4	V22	V5	V23	V6	V24	V7	V25	V8	V26
1	1.432	2.174	1.567	1.555	1.997	1.777	1.482	1.571	1.549	1.627	1.345	1.378
2	3.075	3.263	3.250	3.069	3.365	3.358	3.128	3.098	3.266	2.766	3.373	3.207
3	1.862	3.858	1.822	2.951	3.561	3.773	1.692	2.957	1.795	2.703	1.913	1.866
4	1.813	2.088	2.971	1.781	2.345	1.911	2.656	1.989	2.738	2.058	2.909	2.769

Додаток №6

Алгоритм Ллойда

1	V3	V21	V4	V22	V5	V23	V6	V24	V7	V25	V8	V26
2	1.578	2.174	1.867	1.555	1.997	1.777	1.482	1.571	1.549	1.627	1.345	1.278
3	3.075	3.263	4.250	3.069	3.365	3.358	1.328	3.098	3.266	2.766	3.373	4.207
4	1.813	2.088	2.771	1.781	2.345	1.911	2.656	1.989	2.738	2.058	2.809	2.769

Додаток №7

Алгоритм Хартігана –Вонга

	V3	V21	V4	V22	V5	V23	V6	V24	V7	V25	V8	V26
1	1.43243	2.17464	1.60915	1.55509	1.99584	1.77755	1.48233	1.57173	1.5447	1.62786	1.57381	1.57381
2	3.07566	3.26316	3.25000	3.06908	3.36513	3.35855	3.12829	3.09868	3.26645	2.76645	3.27303	3.20724
3	1.66263	3.85859	1.82222	2.95152	3.56162	3.77374	1.69293	2.95758	1.79596	2.70303	1.91313	1.86667
4	2.71313	2.08889	2.97172	1.78182	2.34546	1.91111	2.65657	1.9899	2.73737	2.05859	2.90909	2.7697

Додаток №8

Алгоритм Хартігана-Вонга з п’ятикластерним рішенням

	V3	V21	V4	V22	V5	V23	V6	V24	V7	V25	V8	V26
1	1.912	1.424	1.519	1.416	1.912	1.658	1.330	1.441	1.441	1.521	1.475	1.527
2	3.115	3.343	3.277	3.138	3.447	3.494	3.193	3.212	3.254	2.806	3.297	3.212

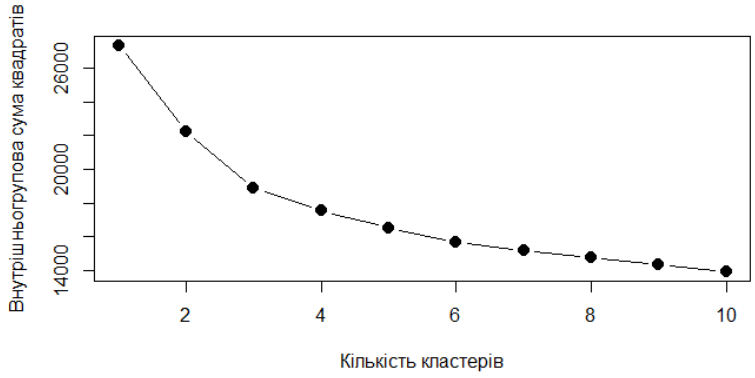
3	1.536	4.081	1.720	4.273	3.762	4.005	1.594	3.106	1.703	2.762	1.782	1.770
4	3.211	1.903	3.361	1.458	1.920	1.528	3.101	1.814	3.118	1.779	3.449	3.207
5	2.183	2.612	2.424	2.132	2.740	2.503	2.205	2.266	2.318	2.338	2.376	2.283

Додаток №9

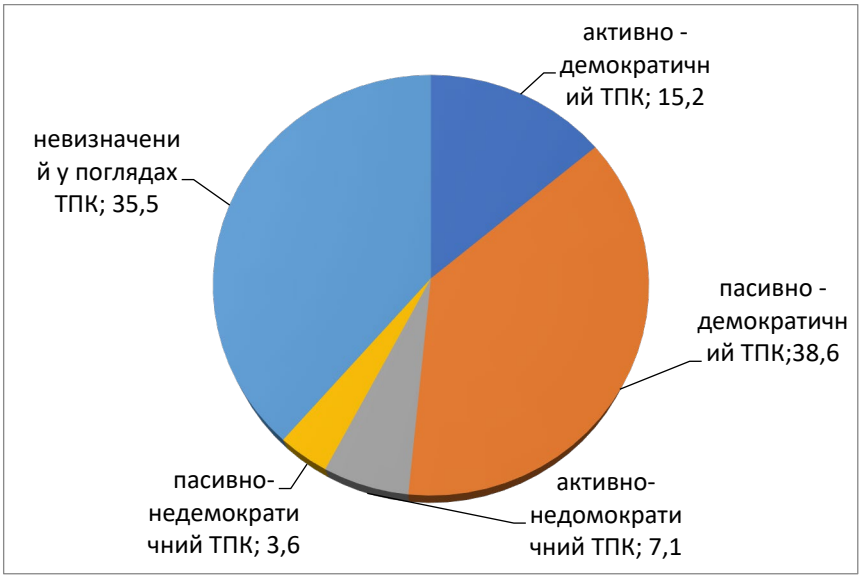
Алгоритм Хартігана-Вонга з п’ятикластерним рішенням для трьохбальної шкали

	V3	V21	V4	V22	V5	V23	V6	V24	V7	V25	V8	V26
1	1.919	1.424	1.824	1.185	1.207	1.249	2.003	1.263	2.007	1.336	1.807	1.996
2	2.907	2.746	2.839	2.16	2.510	2.751	2.879	2.067	2.861	1.824	2.831	2.849
3	1.758	2.409	1.762	2.027	2.288	2.293	1.786	2.246	1.720	1.851	1.665	1.758
4	2.923	1.564	2.807	1.144	1.303	1.267	2.882	1.219	2.867	1.176	2.870	2.870
5	2.640	1.367	2.312	1.581	2.560	1.727	2.607	1.447	2.447	1.563	2.360	2.396

Додаток № 10


Внутрішньогрупова сума квадратів
14000
20000
26000
2
4
6
8
10
Кількість кластерів

Додаток № 11


активно - демократичний ТПК; 15,2
пасивно - демократичний ТПК;38,6
активно-недомократичний ТПК; 7,1
пасивно-недемократичний ТПК; 3,6
невизначений у поглядах ТПК; 35,5