https://doi.org/10.31891/2307-5732-2026-361-16
УДК 519.25:616.98:578.834

КАМІНСЬКИЙ РОМАН
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-8083-4288
e-mail: kaminsky.roman@gmail.com
ШАХОВСЬКА НАТАЛІЯ
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-6875-8534
e-mail: nataliya.b.shakhovska@lpnu.ua
ДМИТРІВ ГАЛИНА
Національний університет «Львівська політехніка»
https://orcid.org/0009-0007-2707-6688
e-mail: dmhaluna83@gmail.com

КЛАСТЕРНЙ АНАЛІЗ МАЛИХ ВИБІРОК БАГАТОВИМІРНИХ ДАНИХ НА ПРИКЛАДІ ТЕРМІНІВ ОДУЖАННЯ ПРИ ЗАХВОРЮВАННІ НА COVID-19

Метою дослідження є виявлення зв'язку тривалості одужання (час перебування в лікувальному закладі) від медико-фізіологічних показників та його представлення лінійними регресійними моделями для окремих підгруп пацієнтів. Для досягнення мети використано такі методи: кореляційний аналіз таблиці «пацієнт – ознака», кластерний аналіз за методом Ланса-Уільямса, побудови лінійних регресійних моделей багатовимірних даних. Проведений аналіз дав такі результати: кореляційний аналіз показав незначущість зв'язку часу одужання з ознаками-показниками, якість кластеризації є низькою, проте моделі множинної регресії є для кожного з трьох кластерів адекватні з коефіцієнто детермінації практично рівним одиниці. Встановлено, що не дивлячись на низьку якість кластеризації моделі адекватно описують зв'язок часу одужання з окремими медико-фізіологічними показниками. Результати можуть бути використані в медико-біологічних дослідженнях, в шкільній педагогічній практиці та інших дослідження.

Ключові слова: *мала вибірка, коронавірус, кластерний аналіз, метод Ланса-Уільямса, дендрограма, множинна регресія.*

KAMINSKY ROMAN, SHAKHOVSKA NATALIA, DMYTRIV GALYNA
National University "Lviv Polytechnic"

CLUSTER ANALYSIS OF SMALL SAMPLE MULTIDIMENSIONAL DATA ON THE EXAMPLE OF RECOVERY TIMES IN COVID-19 DISEASE

The article examines the results of hierarchical agglomerative cluster analysis of a small sample of multidimensional data relating to patients who recovered from Covid-19 coronavirus. The condition of patients at the time of recovery is described by medical and physiological signs. The aim of the study is to identify the relationship between the duration of recovery, i.e. the time spent in a medical institution. To study this relationship more precisely, the group of patients is divided into separate subgroups - clusters. The selection of clusters was carried out using a dendrogram, which resulted in three clusters of patients. The distances between the signs are represented by the Euclidean metric. As medical and physiological signs for cluster analysis, the values of the following indicators were given: physical (age, height, weight); cardiovascular, respiratory, immune and circulatory systems. The final result should be linear regression models of the relationship between the recovery time and the physiological state of patients. In other words, patients should be represented by separate groups – clusters, and linear multiple regression models should be constructed for each cluster. To achieve the goal, the following methods were used: correlation analysis of the "patient – symptom" table, elimination of multicollinearity, hierarchical agglomerate cluster analysis according to the Lance-Williams method, using a flexible strategy and determining the Euclidean distance between normalized symptom values. After combining patients into clusters, linear models of linear multiple regression of multidimensional data were constructed. Based on the analysis, the following results were obtained: correlation analysis showed the insignificance of the relationship between recovery time and symptoms, in the proximity matrix the distances differ little from each other, the quality of the clustering itself is low, however, the multiple regression models are adequate for each of the three clusters with a coefficient of determination close to unity. In addition, as a result of regression analysis, three features were discarded due to their lack of influence on the recovery period. It was found that despite the low quality of clustering, the models adequately describe the relationship between recovery time and individual medical and physiological indicators. The results can be used in medical and biological research, in school pedagogical practice and other studies.

Keywords: *small sample, coronavirus, cluster analysis, Lance-Williams method, dendrogram, multiple regression.*

Стаття надійшла до редакції / Received 12.12.2025
Прийнята до друку / Accepted 11.01.2026
Опубліковано / Published 29.01.2026

Постановка проблеми у загальному вигляді та її зв'язок із важливими науковими чи практичними завданнями

Дане дослідження продовжує вивчення зв'язку біомедичних показників стану пацієнта з терміном одужання при захворюванні на COVID-19 описане в статтях авторів, опублікованих у цьому Віснику том 341, №5 за 2024 р. та том 347, №1 за 2025 р., використовуючи ті ж самі дані. Основною задачею даної публікації є поділ групи пацієнтів на підгрупи – кластери щодо виявлення існуючої відмінності між ними та відмінності за усередненими по кластерах результативними та факторними показниками. Власне ідея полягає в тому чи однакові показники фігурують в цих кластерах.

Вивчення стану здоров'я людей щодо лікування пандемії – захворювання на COVID-19 триває вже понад 7 років, відкриваючи, встановлюючи та моделюючи різноманітні ситуації. Як правило, такі дослідження

проводятьс в різних малих та великих лікувальних закладах де фахівці збирають та агалізують відповідні дані та необхідну інформацію, інтерпретують результати та будують моделі цього неприйнятного і небезпечного явища. Помимо цього і досить часто, провести якісний статистично-математичний аналіз отриманих (зібраних) в лікувальних закладах даних є суттєвою проблемою за відсутності саме таких фахівців, що змушує звертатись в інші, переважно наукові та науково-дослідні заклади. В цьому плані, задача обробки та опрацювання такого набору даних, а точніше розробка простої та ефективної методології попередньої обробки, аналізу та інтерпретації цього типу даних є вельми важливою науково-практичною задачею.

Аналіз досліджень та публікацій

В наукових дослідженнях падемії коронавіруса COVID-19 кластерний аналіз посідає важливе місце і на сьогоднішній день це один з найбільш поширених методів класифікації її локалізації на територіях різних держав та регіонів. Так, в [1], за результатами проведеного кластерного аналізу процесу вакцинації проти COVID-19 найкраща ситуація склалась у більшості країн Європи. Для цих країн характерним є низька швидкість приросту нових випадків захворювань та середній рівень смертності, спричинений COVID-19, разом із достатньо високим рівнем вакцинованого населення. Згідно з проведеним науковим дослідженням, можна прийти до висновку, що проведення масової вакцинації від COVID-19 дійсно має позитивний ефект на зниження захворюваності серед населення, навіть за достатньо короткий термін її впровадження. Тому, тільки при комплексному виконанні всіх необхідних вказівок можна буде уникнути локдаунів та карантинів, які наносять нищівний удар по світовій економіці. Авторами роботи [2] було розроблено спеціальну анкету, яка розповсюджувалась соціальними мережами, відповіді фіксувались у Google forms. В результаті, було виділено 6 кластерів, які загалом описували 6 способів переживання тривоги в умовах цієї пандемії. Відповідно до різних способів та стилів переживання тривоги різними людьми мають бути застосовані і відповідні способи психологічної допомоги особистостям. Їх визначення постає як перспективне завдання подальшої роботи влади. Використання кластерного аналізу на основі нейронної мережі для наукового обгрунтування протиепідемічних заходів з метою зниження захворюваності та для вирішення задачі виділення зон її розповсюдження розглянуто в [3]. Тут, основна задача – це аналіз методів кластеризації територій України за характером епідемічного процесу COVID-19. Були використані дані по областям України, надані Центром громадського здоров'я МОЗ України. Як результат, проведено розбиття областей України на зони зараження вірусом COVID-19 та представлено карту цього розбиття. В [4] автори провели статистичні дослідження пандемії коронавірусної хвороби, використовуючи ієрархічний кластерний аналіз з десятьма кластерами, причому майже всі параметри були значущими. Результати вказують на територіальну, расову та геномну основу пандемії та потенційні зв'язки з міграціями. Отримана інформація може бути передана органам влади, а також відповідним міжнародним органам, для вжиття запобіжних та превентивних заходів у боротьбі з пандемією. На основі даних про кількість зареєстрованих смертей від коронавірусного захворювання в роботі [5] визначено 16 кластерів-муніципалітетів. Серед основних результатів цього дослідження виявлено, що кластери з високими значеннями рівня смертності мають високі значення щільності населення та низький рівень бідності. Навпаки, кластери з низькими показниками щільності та високим рівнем бідності мали низький рівень смертності. Знайдені закономірності, виражені у вигляді муніципальних кластерів можуть бути корисними для прийняття рішень органами охорони здоров'я щодо профілактики захворювань і контролю за ними для посилення заходів охорони громадського здоров'я та оптимізації розподілу ресурсів для зменшення кількості госпіталізацій і смертності. У дослідженні [6] автори провели кластерний аналіз даних пацієнтів із COVID-19 на основі біохімічних вимірювань протягом перших 72 годин після госпіталізації. Клінічні та біохімічні змінні були отримані від 1039 пацієнтів із підтвердженим діагнозом COVID-19. Діагнози пацієнтів та прогнози надзвичайно відрізнялися між трьома отриманими кластерами, що свідчить про те, що технології, керовані даними, розроблені для скринінгу, аналізу, прогнозування та відстеження пацієнтів, відіграють ключову роль у застосуванні індивідуального лікування пандемії COVID-19. Дослідження проведені в [7] спрямоване на аналіз кластерів, пов'язаних зі смертністю від COVID-19 на муніципальному рівні в Мексиці, було розглянуто з точки зору Data Science. Представлено нову програму, яка використовує гібридний алгоритм машинного навчання для генерації кластерів-муніципалітетів. Було визначено два ключові показники, пов'язані зі смертністю від COVID-19 на муніципальному рівні: один – це щільність населення, а інший – відсоток населення, яке живе в бідності. Знайдені закономірності у вигляді муніципальних кластерів можуть бути корисними для прийняття рішень органами охорони здоров'я. У дослідженні [8] проведено кластерний аналіз різних штатів Сполучених Штатів, щоб визначити тяжкість інфекції COVID-19. Набір даних цього дослідження охоплює період з 2020 по 2021 рік. В результаті цього дослідження було визначено три кластери результатів аналізу штатів США як високий, середній та низький рівень інфікування, щоб сприяти об'єктивним рішенням щодо надання пріоритету у розподілі по штатах вакцин та щоб запобігти подальшому поширенню епідемії у разі дефіциту вакцин. У статті [9] розглядаються алгоритми кластеризації, їх переваги та застосування. Показано, як можна поєднувати різні методи для отримання стабільних кластерів, не надто залежних від критеріїв, обраних для аналізу даних. Вказано на те, що кластеризація завжди забезпечує групи, навіть якщо немає групової структури. При застосуванні кластерного аналізу, як правило, висувають гіпотезу про існування груп. Але це припущення може бути хибним або слабким. Результати кластеризації не слід узагальнювати. Випадки в одному кластері схожі лише щодо інформації, на якій базувався кластерний аналіз, тобто вимірів/змінних, що викликають відмінності. У навчальному посібнику [10] розглянуто теоретичні і практичні аспекти кластерного аналізу, такі як: розрахунок відстаней між економічними об'єктами, принципи поєднання

їх у кластери, прийоми віднесення нових об'єктів до існуючих кластерів, порядок зміни параметрів кластерів, розглянуто критерії якості отриманих розбиттів. В збірнику статей [11] містить опис прикладних статистичних методів та їх алгоритмів з прикладми застосування, що відносяться до регресійного та дискримінантного аналізу, методу головних компонент, факторному аналізу, кластерному аналізу, зокрема поданий алгоритм ієрархічного агломеративного кластерного аналізу Ланса-Уільямса щодо його застосування в даному дослідженні, а також представлено методи розпізнаваню образів та аналізу часових рядів. Автори статті [12] розглядають загальний алгоритм Ланса-Уільямса і вказують на його здатність обчислювати ієрархічну дендрограму для багатьох ієрархічних схем кластеризації. Автори статті [13] також вказують на те, що агломеративна ієрархічна кластеризація за допомогою відомої формули Ланса-Уільямса не лише узагальнює одиночні, повні та усереднені зв'язки між об'єктами, але й включає міжкластерні відстані на основі кількох найближчих або найдальших сусідів. Крім того, вони надають деякі умови для генераторів ваг, щоб гарантувати відсутність неестетичних інверсій у результуючих дендрограмах. У дослідженні [14] здіснено порівняння різних підходів до розв'язання регресійних задач. Експериментально доведено, що при побудові регресійних моделей реальних систем і процесів вельми важливим є вибір відповідних припущень, оскільки від них залежить коректність застосування процедур класичного регресійного аналізу.

Формулювання цілей статті

Метою даного дослідження є поділ групи пацієнтів на однорідні підгруп, утворені за допомогою ієрархічного агломеративного кластерного аналізу, розроблення лінійних регресійних моделей для кожної окремої підгрупи-кластера, а також встановрення розбіжностей між ними, в сенсі середніх значень результативної та факторних ознак для кожного з цих кластерів.

Виклад основного матеріалу

1. Підготовка та опис даних. Для проведення даного дослідження надано вибіркові дані про пацієнтів, які одужали від захворювання коронавірусом. Особливістю цих даних є те, що термін одужання кожного з пацієнтів супроводжується ще 29-ма факторними ознаками, які представляють собою показники фізичної, серцево-судинної, дихальної, імунної та кровоносної систем. Саме за цими показниками має бути досягнута мета цього дослідження.

Обсяг вибірки складається з 19 об'єктів – пацієнтів. Термін – «тривалість захворювання» або «час одужання (Duration)» прийнято за результативну ознаку, усі інші ознаки розглядаються як факторні. З статистичної точки зору у відповідності з обсягом дана вибірка відповідає категорії *малих вибірок*. Результативна ознака – час одужання є визначеною кількістю ліжкоднів, що в очевидь можна розгядати як кількість діб перебування в лікувальному закладі.

Після усунення мультиколінеарності та вилучення тих факторних ознак, які практично не зв'язані з терміном одужання (результативною ознакою) їх кількість зменшилась до 13, тобто одна результативна – тривалість одужання та дванадцять факторних, які належать до цих п'яти медико-фіологічних систем. Отже, отримані для дослідження дані про 19 пацієнтів після їх нормалізації за формулою міні-максу, тобто приведення їх до інтервалу [0, 1], представлені в табл. 1. Особливістю цих даних є те, що результативна ознака – термін одужання кожного з пацієнтів супроводжується не 29-ма факторними ознаками, а лише 12-ма, які також представляють собою медико-фізіологічні показники фізичної, серцево-судинної, дихальної, імунної та кровоносної систем. Ознаки тут є такими показниками:

$\boldsymbol{Y}$ – Duration – тривалість одужання;
$\boldsymbol{X_2}$ – Age – вік;
$\boldsymbol{X_3}$ – Hejght – зріст;
$\boldsymbol{X_4}$ – IMT – індекс маси тіла;
$\boldsymbol{X_5}$ – Pulse – частота пульсу;
$\boldsymbol{X_6}$ – Test Walk 6 min – 6-хвилинний тест ходьби;
$\boldsymbol{X_7}$ – SaO2 – тест на астму;
$\boldsymbol{X_8}$ – $МОШ_{75}$ – максимальна об'ємна швидкість повітря на рівні видиху 75%;
$\boldsymbol{X_9}$ – ЖЄЛвдиху – життєва ємність легень;
$\boldsymbol{X_{10}}$ – CD3 – основний маркер Т-лімфоцитів;
$\boldsymbol{X_{11}}$ – CD16 – клітини-ефектори;
$\boldsymbol{X_{12}}$ – IL4 – рівень інтерлейкіну-4,
$\boldsymbol{X_{13}}$ – IL10 – рівень інтерлейкіну-10.

Саме за цими показниками має бути досягнута мета цього дослідження.

Тут ознака $\boldsymbol{Y = X_1}$ є результативною ознакою, а усі решта вважаються факторними.

Таблиця 1

Таблиця «об'єкт – ознаки» нормалізованих даних

Пацієнти	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
П1	0,91	0,72	0,15	0,71	0,88	0	0	0,13	0,38	0,5	0	0,13	0,84
П2	0,55	0,56	0,33	0,44	0,44	0,31	0	0,69	0,86	0,78	0,79	0,55	0,94
П3	0,45	0,56	0	0,17	0,69	0,38	0	0,81	0,32	0,78	1	0,64	0,13
П4	0,73	0,16	0,26	0,63	0,28	0,72	0,75	1	1	1	0,79	0,05	0,47
П5	0,73	0,33	0,07	0,2	1	0,22	0,5	0,29	0,05	0,17	0,43	0,11	0,14

Пацієнти	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
П6	0,82	0,79	0,15	1	0,13	0,1	0,5	0,32	0,43	0,56	0,57	0,05	0,44
П7	0	0,77	0,19	0,24	0,34	0,07	1	0,97	0,84	0	0,43	0,12	0,94
П8	0,18	0,12	0,63	0,39	0,47	1	0,25	0,82	0,86	0,44	0,64	0,13	0
П9	0,18	1	1	0,18	0	0,3	0,5	0,87	0,84	0,17	0,93	0,13	0,04
П10	0,18	0,84	0,44	0,48	0,16	0,61	0,25	0,94	0,76	0,33	0,86	0,04	0,66
П11	0,55	0,7	0,07	0,52	0,38	0,2	0,25	0,92	1	0,83	0,5	0	0,29
П12	0,55	0,58	0,33	0	0,59	0,96	0,5	0,84	0,78	0,89	0,93	0,01	0,41
П13	0,45	0,74	0,44	0,93	0,19	0,45	0,25	0,97	0,62	0,5	0,86	0,12	1
П14	0,45	0,72	0,07	0,99	0,25	0,27	0	0,95	0,03	0,17	1	0,06	0,24
П15	0,36	0,72	0,56	0,09	0,28	0,35	0,75	0,98	0,92	0,44	0,21	0,08	0,09
П16	1	0,28	0,63	0,39	0,03	0,25	1	0,94	0,65	0,44	0,07	1	0,57
П17	0,82	0,67	0,81	0,6	0,25	0,25	0,25	0	0,92	0,33	0,64	0,08	0,74
П18	0,64	0	0,3	0,31	0,22	0,39	0,5	0,27	0	0,5	0,57	0,07	0,51
П19	0,55	0,65	0,7	0,37	0,22	0,38	0,5	0,03	0,46	0,39	0	0,95	0,26

Для розбиття пацієнтів на підгрупи, тобто за найбільш подібними індивідуальними показниками використано ієрархічний агломеративний кластерний аналіз.

2. Проведення кластерного аналізу. Для проведення кластеризації пацієнтів вибрано метод Ланса-Уільямса [ ] – метод ієрархічного агломеративного кластерного аналізу. Цей метод математично обґрунтовує кількість кластерів та дозволяє вибрати стратегію об'єднання набільш близьких об'єктів. Робота алгоритму цього методу базується на аналізі та перерахунку таблиці близькостей між об'єктами, причому в якості відповідної метрики використано метрику Евкліда. В результаті, таблиця близькостей, тобто близькість між пацієнтами, з врахуванням їхніх індивідуальних показників представлена таблицею 2.

Таблиця 2

	Пацієнти																		
	p1	**p2**	**p3**	**p4**	**p5**	**p6**	**p7**	**p8**	**p9**	**p10**	**p11**	**p12**	**p13**	**p14**	**p15**	**p16**	**p17**	**p18**	**p19**
p1	0	1,41	1,74	1,99	1,29	1,21	1,92	2,07	2,26	1,8	1,49	1,95	1,61	1,74	1,87	1,97	1,32	1,49	1,56
p2	1,41	0	1,13	1,28	1,72	1,34	1,57	1,47	1,64	1,04	1,03	1,24	0,92	1,54	1,51	1,62	1,19	1,45	1,57
p3	1,74	1,13	0	1,55	1,35	1,57	1,91	1,42	1,71	1,38	1,23	1,2	1,54	1,32	1,58	1,92	1,81	1,43	1,7
p4	1,99	1,28	1,55	0	1,85	1,43	1,71	1,21	1,74	1,28	1,02	0,97	1,25	1,75	1,32	1,58	1,61	1,46	1,95
p5	1,29	1,72	1,35	1,85	0	1,43	1,78	1,65	1,96	1,76	1,58	1,62	1,87	1,56	1,58	1,89	1,65	1,04	1,55
p6	1,21	1,34	1,57	1,43	1,43	0	1,65	1,77	1,68	1,32	1,11	1,66	1,13	1,17	1,51	1,67	1,1	1,2	1,48
p7	1,92	1,57	1,91	1,71	1,78	1,65	0	1,81	1,51	1,2	1,46	1,7	1,4	1,83	1,19	1,69	1,72	1,74	1,87
p8	2,07	1,47	1,42	1,21	1,65	1,77	1,81	0	1,38	1,16	1,32	1,07	1,52	1,66	1,19	1,87	1,63	1,47	1,7
p9	2,26	1,64	1,71	1,74	1,96	1,68	1,51	1,38	0	1,02	1,47	1,5	1,49	1,64	1,04	1,87	1,49	1,81	1,71
p10	1,8	1,04	1,38	1,28	1,76	1,32	1,2	1,16	1,02	0	1,04	1,1	0,69	1,2	1,15	1,82	1,3	1,49	1,73
p11	1,49	1,03	1,23	1,02	1,58	1,11	1,46	1,32	1,47	1,04	0	1,14	1,15	1,4	1,01	1,71	1,41	1,52	1,73
p12	1,95	1,24	1,2	0,97	1,62	1,66	1,7	1,07	1,5	1,1	1,14	0	1,4	1,75	1,22	1,89	1,61	1,46	1,88
p13	1,61	0,92	1,54	1,25	1,87	1,13	1,4	1,52	1,49	0,69	1,15	1,4	0	1,15	1,52	1,75	1,26	1,5	1,83
p14	1,74	1,54	1,32	1,75	1,56	1,17	1,83	1,66	1,64	1,2	1,4	1,75	1,15	0	1,77	2,13	1,73	1,47	1,98
p15	1,87	1,51	1,58	1,32	1,58	1,51	1,19	1,19	1,04	1,15	1,01	1,22	1,52	1,77	0	1,41	1,54	1,56	1,47
p16	1,97	1,62	1,92	1,58	1,89	1,67	1,69	1,87	1,87	1,82	1,71	1,89	1,75	2,13	1,41	0	1,75	1,62	1,27
p17	1,32	1,19	1,81	1,61	1,65	1,1	1,72	1,63	1,49	1,3	1,41	1,61	1,26	1,73	1,54	1,75	0	1,39	1,36
p18	1,49	1,45	1,43	1,46	1,04	1,2	1,74	1,47	1,81	1,49	1,52	1,46	1,5	1,47	1,56	1,62	1,39	0	1,43
p19	1,56	1,57	1,7	1,95	1,55	1,48	1,87	1,7	1,71	1,73	1,73	1,88	1,83	1,98	1,47	1,27	1,36	1,43	0

Результатом проведення кластерного аналізу є дендрограма, зображена на рис. 1. Як показує графік дендрограми, відтань останнього об'єднання, коли усі об'єкти є єдиним кластером становить 4.84 одиниці, натомість найменша відстань між окремими об'єктами рівна 0,69. Можна також вказати на те, що відстані об'єктами для перших об'єднань: 22, 21, 25, 23, 24, 26, за незначним винятком 28, 20 є майже рівні. Різниці між другими об'єднаннями: 30, 29, 27 також мало різяються між собою. Об'єднання: 33, 31, 32 та 34, 35, а також 36, 37 також близькі, що означає слабку кластерну структуру даних. Тут перше об'єднання об'єкт з об'єктом, друге – об'єкт з групою, далі група з групою.

3. Обговорення характеристик кластерів. Отриману таблицю подібності та побудовану дендрограму можна віднести ко категорії однорідних величин. Дійсно, величини відстаней в таблиці та різниці між вузлами об'єднань знаходяться в межах одиниці, тобто $1 \le \delta_{ij} \le 2$, де $i \in \{p1, \dots, p19\}$ – є номером пацієнта, відстань між вузлами лежить в межах одного інтервалу між сусідніми поділками вертикальної шкали, очевидно за винятком декількох об'єктів.

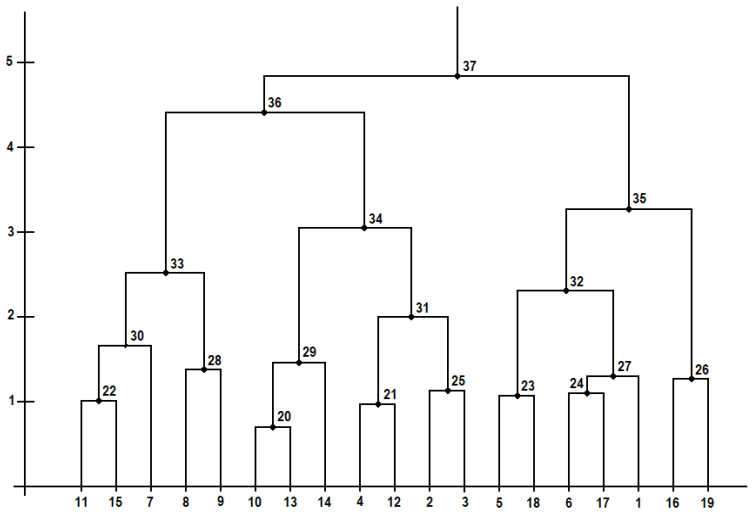

Рис. 1. Дендрограма кластерного аналізу пацієнтів.

Дана ситуація свідчить про невисоку якість кластеризації, проте на рівні вище 3.3 цілком чітко можна виділити три кластери.

Кластер 1. об'єкти: 11,15, 7, 8, 9.

Кластер 2. об'єкти: 10, 13, 14, 4, 12, 2, 3.

Кластер 3. об'єкти: 5, 18, 6, 17, 1, 16, 19.

Нижче представлені три таблиці-кластери (табл. 3, табл. 4, табл. 5) з реальними значеннями їх індивідуальних медико-фізіологічних характеристик пацієнтів, які включені у відповідний кластер. Останній рядок кожної таблиці відповідає середньому значенню по стовпчику, тобто середньому значенню ознаки в цьому кластері.

Таблиця 3

Кластер 1													
	y1	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**	**x10**	**x11**	**x12**	**x13**
p11	20	60	158	30,8	75	409	96	90	94	63	19	4,6	8,1
p15	18	61	171	23,3	72	443	98	94	91	56	15	5,4	5,1
p7	14	63	161	25,9	74	380	99	93	88	48	18	5,8	18
p8	16	35	173	28,4	78	586	96	84	89	56	21	5,9	3,7
p9	16	73	183	24,8	63	432	97	87	88	51	25	5,9	4,3
	16,8	58,4	169	26,6	72,4	450	97,2	89,6	90	54,8	19,6	5,52	7,84

Таблиця 4

Кластер 2													
	y1	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**	**x10**	**x11**	**x12**	**x13**
p10	16	66	168	30,1	68	500	96	91	85	54	24	5	13,7
p13	19	62	168	37,9	69	464	96	93	80	57	24	5,8	18,9
p14	19	61	158	38,9	71	424	95	92	58	51	26	5,2	7,3
p4	22	37	163	32,7	72	525	98	95	94	66	23	5,1	10,8
p12	20	55	165	21,7	82	578	97	85	86	64	25	4,7	10
p2	20	54	165	29,4	77	434	95	76	89	62	23	10,3	18
p3	19	54	156	24,7	85	450	95	83	69	62	26	11,2	5,7
	19,3	55,6	163	30,8	74,9	482	96	87,9	80,1	59,4	24,4	6,76	12,1

Таблиця 5.

Кластер 3													
	y1	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**	**x10**	**x11**	**x12**	**x13**
p1	24	61	160	34	91	365	95	41	71	57	12	5,9	16,4
p5	22	44	158	25,2	95	414	97	51	59	51	18	5,7	5,9
p6	23	64	160	39,1	67	386	97	53	73	58	20	5,1	10,4
p16	25	42	173	28,4	64	420	99	91	81	56	13	14,9	12,3
p17	23	59	178	32,2	71	420	96	33	91	54	21	5,4	14,9
p18	21	30	164	27,1	70	451	97	50	57	57	20	5,3	11,4
p19	20	58	175	28,1	70	450	97	35	74	55	12	14,4	7,6
	22,6	51,1	167	30,6	75,4	415	96,9	50,6	72,3	55,4	16,6	8,1	11,3

4. Визначення якості кластеризації. Візуальний аналіз таблиці «об'єкт-властивість» і таблиці близькостей, а також величини відстаней об'єднань, на отриманій дендрограмі, вказує на те, що використовувані дані цих 19 пацієнтів є досить однорідними, тобто непомітно, особливо на дендрограмі, значних відстаней. Таке зауваження вимагає перевірки якості проведеної кластеризації.

Оцінювання якості проведеного ієрархічного агломеративного кластерного аналізу за методом Ланса–Уїльямса з використанням евклідової метрики (відстані) та гнучкої стратегії об'єднання, здійснене такими, найбільш поширеними методами.

Коефіцієнт силуета **(KS).** В даному випадку коефіцієнт силуета KS = 0.171. Значення цього показник вказує на декілька моментів, а саме: по-перше, якщо його значення є мешими < 0.25, це вказує на слабку кластеризацію; по-друге, об'єкти майже однаково "схожі" і на об'єкти свого кластеру, і на об'єкти сусідніх кластерів, що свідчить про відсутність чітких меж між кластерами, тобто кластери мають розмиті межі.

Індекс Данна **(ID).** Тут індекс Данна ID = 0.078. Значення цього індексу є дуже низьким, що свідчить про слабке розділення кластерів або сильну внутрішню їх розтягнутість. Також вказує і на велику внутрішньокластерну варіативність та мале міжкластерне розділення. Він є типовим для однорідних даних, де немає природних груп.

Індекс Девіса-Боулдена **(IDB).** Значення цього індексу для проведеного кластерного аналізу становить IDB = 1.179, проте, воно повинно бути принаймні 0.5 або менше, хоча це не оптимальне значення, тобто значення має бути < 1. Значення цього індексу тут є більш близький до свого порогового значення 1, а це означає перекриття між кластерами.

Індекс Калінські–Харабаса **(**англ. Calinski–Harabasz index); **(CHI).** В даному дослідженні індекс Калінські–Харабаса = 3.798. Величина цього індексу в даному дослідженні вважається низькою, тобто чим вона вище – тим краще), іншими словами кластерна структура даних є слабо виражена. Низьке значення для цього індексу означає відсутність чіткої кластерної структури даних.

В даному дослідженні, судячи з показників оцінки якості, кластеризація є низькою. Ймовірно, що значення ознак або відстаней не забезпечують чіткого розділення кластерів. Так, за наведеними метриками можна вважати, що група з 19 пацієнтів є досить однорідною, а кластерний аналіз у цьому випадку є малоефективним, оскільки дані не мають виражених підгруп, імовірно, пацієнти достатньо подібні між собою за всіма 13 ознаками, що можливо і є причиною не надійної сегментації. Причиною цього можна вважати і те, що значення самих ознак для 19 пацієнтів не дуже різняться, про це вказує і сама матриця відстаней – практично усі відстані є в межах від одиниці до двох, тобто суттєвих відмінностей немає.

5. Порівняльний аналіз кластерів. Недивлячись на низьку якість кластерів, між ними існує суттєва різниця, а саме: щодо кількості показників, а головне в середніх значеннях кожного показника в кластерах. Нижче приведені таблиці значень показників для кожного кластера. Нижні рядки представляють середні значення показників в кожному кластері

Для зручності порівняльного аналізу, за усередненими значеннями показників, цих трьох кластерів, їх значення зведені у таблицю 6. Саме в цьому плані і проведено порівняльний аналіз кластерів. Такий підхід дає підставу вирізнити кластери за їхніми ознаками.

Таблиця 6.

Зведена таблиця усереднених показників кожного з кластерів

	y1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13
К1	16,8	58,4	169	26,6	72,4	450	97,2	89,6	90	54,8	19,6	5,52	7,84
К2	19,3	55,6	163	30,8	74,9	482	96	87,9	80,1	59,4	24,4	6,76	12,1
К3	22,6	51,1	167	30,6	75,4	415	96,9	50,6	72,3	55,4	16,6	8,1	11,3

За даними таблиці 6 можна зробити такі висновки.

Кластер 1 має найменші значення таких показників X1, X4, X5, X10, X12, X13;
найбільше значення цих показників X2, X3, X7, X8, X9;

Кластер 2 має найменші значення таких показників X3, X7;
найбільше значення цих показників X4, X6, X10, X11, X13;

Кластер 3 має найменші значення таких показників X6, X8, X9, X11;
найбільше значення цих показників X1, X5, X12.

Отже, найменше середнє значення тривалості часу одужання мають пацієнти першого кластера – 16,8 ліжкоднів. Для пацієнтів цього кластера найменшими є індекс маси тіла та частота пульсу. Крім того, стосовно імунної та кровоносної систем, вони мають найменші значення показників основного маркера Т-лімфоцитів та клітин-ефекторів в сенсі імунної системи, а також стосовно кровоносної системи мають найнижчий рівень інтерлейкіну-4 та інтерлейкіну-10. З другого боку, необхідно відмітити і набільші середні значення таких показників: щодо фізичної системи – це вік і ріст (58.4 кг та 169 см), а також характеристику дихальної системи, тобто середні показники: тесту на астму (97.2), максимальну об'ємну швидкість повітря на рівні видиху 75% (89,6), та життєву ємність легень (90).

Середній час одужання пацієнтів другого кластера становить 19,3 ліжкоднів. Їх зріст є найменшим – 163 см, а тест на астму (96) вказує на її відсутність. Серед найбільших середніх значень показників виділяються:

індекс маситіла (30.8), 6-хвилинний тест ходьби (482), основний маркер Т-лімфоцитів (59,4), значення клітин-ефекторів (24,4) та рівень інтерлейкіну-10 (12,1).

Пацієнти третього кластеру мають найбільший час тривалості одужання (22,6 ліжкоднів). Проте, середні значення показників: максимальної об'ємної швидкості повітря на рівні видиху 75% (50,6), життєвої ємності легень (72,3) та клітин-ефекторів (16,6) є найнижчими. Найбільші значення мають показники пульсу (75,4) та рівня інтерлейкіну-4 (8,1).

Загальні середні значення нормалізованих значень ознак для кожного кластера є досить близькими: К1 = 0.46, К2 = 0.52, К3 = 0.43, а це свідчить про те, що кластерна структура є досить слабка, іншими словами, різниця між середніми значеннями у кластерах менша, ніж типовий розкид всередині кластера. Крім того, має місце суттєва різниця між цими значеннями, майже у двічі, для ознак Y, X_8, X_9 та X_{11}, що очевидно впливає на різницю між кластерами.

Оскільки кластери представляють собою малі вибірки багатовимірних даних, для більш конкретного їх опису є сенс представити їх лінійними регресійними моделями.

6. Регресійні моделі кластерів.

Не дивлячись на низькі показники якості кластеризації, відмінність між кластерами можна вважати суттєвою. Як показано вище, кластери відрізняються не лише тривалістю одужання, але і максимальними та мінімальними значеннями одних і тих самих ознак. У межах кожного з отриманих трьох кластерів (обсягами 5, 7 і 7 об'єктів відповідно), за допомогою табличного процесора Microsoft Excel побудовано окремі моделі множинної регресії. В цих моделях кожен об'єкт (пацієнт) є описаний 13 ознаками, з яких одна ознака є результативною (це тривалість одужання), а решта ознак є факторними. Кількість відібраних факторів у моделях варіюється: 4, 6 та 6 відповідно для кластерів 1, 2 та 3. Усі моделі демонструють високий рівень апроксимації, тобто коефіцієнт детермінації має значення $R^2 = 0{,}99$ у кожному випадку. Ці моделі мають такий вигляд:

Кластер 1. $Y = -15.4 + 0.087 \cdot X_3 - 0.01 \cdot x_6 + 0.445 \cdot x_{10} - 0.09 \cdot x_{11}$,

Кластер 2. $Y = -6.48 - 0.13 \cdot X_2 + 0.455 \cdot X_4 + 0.242 \cdot X_5 + 0.027 \cdot X_6 - 0.15 \cdot X_8 + 0.007 \cdot X_9$,

Кластер 3. $Y = -39.6 - 0.09 \cdot X_2 + 0.004 \cdot X_5 - 0.05 \cdot X_6 + 0.025 \cdot X_8 + 0.1 \cdot X_9 - 0.01 \cdot X_{11}$.

Моделі мають різний склад предикторів, що свідчить про варіативність впливу факторів у межах різних груп об'єктів. Така відмінність у структурі моделей підтверджує доцільність попередньої кластеризації, оскільки залежність між факторними та результативною ознаками має специфічний характер: мала вибірка, багато ознак та їх незначущість. Крім того, набір предикторів та їх кількість в кожному кластері може свідчити про відмінності у механізмах, які впливають на результативну ознаку в різних групах пацієнтів. Така варіативність підтверджує доцільність урахування структурної неоднорідності вибірки та обґрунтовує застосування регресійного аналізу окремо для кожного кластеру. Аналіз регресійних рівнянь показав, що в даному випадку такі показники як X_7, X_{12}, X_{13} не впливають на термін одужання.

Проте, високе значення коефіцієнта детермінації, близьке до одиниці, свідчить про тісний зв'язок між результативною ознакою та відібраними факторами.

Загалом результати підтверджують ефективність підходу, який поєднує кластеризацію з подальшим регресійним аналізом у межах однорідних груп об'єктів.

Висновки з даного дослідження
і перспективи подальших розвідок у даному напрямі

В результаті проведеного, попередньо, кореляційного аналізу для усунення мультиколінеарності, кластерного аналізу для розподілу пацієнтів на групи за медико-медикофізіологічним станом та побудовою лінійних регресійних моделей показників на час одужання можна можна вказати на таке.

По-перше, виходячи з таблиці близькостей, середні значення усіх відстаней між об'єктами знаходяться між 1 і 2. Це можна вважати головною причиною низької відокремлюваності кластерів, бо якщо всі пацієнти схожі за ознаками, то й відстані між ними малі, а це означає випадок, коли жодна кластерна модель не зможе бути побудуваною з чітко відокремленими кластерів – бо їх просто немає у даних.

По-друге, обчислені середні значення нормованих значень ознак для кожного кластера є досить близькі, а це свідчить про те, що кластерна структура є досить слабка, іншими словами, різниця між середніми значеннями у кластерах менша, ніж типовий розкид всередині кластера.

По-третє, проблема низької якості кластеризації мабуть і втому, що ознак більше ніж обєктів, тобто 19 об'єктів і кожен має 13 ознак. Це класична проблема, а в такій ситуації відстані «згладжуються» (концентруються навколо певного середнього) – це явище відоме в теорії кластеризації у високих розмірностях (ефект «curse of dimensionality») і тому навіть нормалізовані відстані втрачають інформативність, а кластерні алгоритми починають працювати нестабільно, оскільки в результаті нормалізації дані належать одному і томуж інтервалу значень.

В перспективі є подальше дослідження динаміки одужання та впливу показників.

Література

1. Каща М.О., Чугаєва О.В., Грек К.А. Реакція населення Європи на процес вакцинації проти COVID-19 з використанням кластерного аналізу. Вісник СумДУ. Серія «Економіка». № 1. 2021. С. – 312 – 317.

2. Зімовін О.І. Способи переживання тривоги в умовах пандемії covid-19. Досвід переживання пандемії covid 19: матеріали онлайн семінарів 23 квітня 2020 року «Досвід карантину: дистанційна психологічна

допомога і підтримка» та 15 травня 2020 року «Дистанційні психологічні дослідження в умовах пандемії covid 19 і карантину». Київ. 2020. С. 91 – 96.

3. Базілевич К.О., Меняйлов Є.С., Чумаченко Д.І.. Виділення зон розповсюдження захворюваності на коронавірус covid-19 на основі методів кластерного аналізу. Сучасний стан наукових досліджень та технологій в промисловості. 2021. № 1 (15). С. 5 – 13.

4. Ahmed Al-Imam. Clustering analysis of coronavirus disease 2019 pandemic / Ahmed Al-Imam, Usama Khalid Enjaz, Hend Al-Door. Asian Journal of Medical Sciences, 12(2):108-113. DOI: 10.71152/ajms.v12i2.3572.

5. Pérez-Ortega, J.; Almanza-Ortega, N.N.; Torres-Poveda, K.; Martínez-González, G.; Zavala-Díaz, J.C.; Pazos-Rangel, R. Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemo-graphic Factors at Municipal Level in Mexico. Mathematics 2022, 10, 2167. https://doi.org/10.3390/math10132167.

6. San-Cristobal R., Martín-Hernández R., Ramos-Lopez O. and others. Longwise Cluster Analysis for the Prediction of COVID-19 Severity within 72 h of Admission: COVID-DATA-SAVE-LIFES Cohort / J. Clin. Med. 2022, 11(12), 3327; https://doi.org/10.3390/jcm11123327

7. Pérez-Ortega J., Nely Almanza-Ortega N., Torres-Poveda K., and others. Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico / Mathematics, 10 (13):2167, https://doi.org/10.3390/math10132167.

8. Shih, D.-H.; Shih, P.-L.; Wu, T.-W.; Li, C.-J.; Shih, M.-H. Cluster Analysis of US COVID-19 Infected States for Vaccine Distribution. Healthcare 2022, 10, 1235. https://doi.org/10.3390/healthcare10071235.

9. Soni Madhulatha T. An overview on clustering methods. IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725.

10. Пістунов І.М., Антонюк О.П., Турчанінова І.Ю. Кластерний аналіз в економіці. Навч. Посібник. Дніпропетровськ: Національний гірничий університет. 2008. 84 с.

11. Статистические методы для ЭВМ. Под ред. К. Энслейна, Э. Релстона, Г.С. Уилфа: Пер. с англ. Под ред. М.Б. Малютова. – М.: Наука. Гл. ред. физ.-мат. Лит., 1986. – 464 с.

12. Yarmish G., Distributed Lance-William Clustering Algorithm / Yarmish G., Listowsky P., Kings-borough D.S., 1709.06816v1 PDF (arxiv.org).

13. Gagolewski M., Cena A., James S., Beliakov G., Hierarchical clustering with OWA-based linkages, the Lance–Williams formula, and dendrogram inversions, Fuzzy Sets and Systems 473, 108740, 2023, DOI:10.1016/j.fss.2023.108740.

14. Бойко Н. І., Газдюк К. П. Порівняння регресійних моделей за наявності викидів у наборі різнотипових даних. 2023. Scientific Bulletin of UNFU, 33(2), 84-91. https://doi.org/10.36930/40330212.

References

1. Kashcha M.O., Chuhaieva O.V., Hrek K.A. Reaktsiia naselennia Yevropy na protses vaktsynatsii proty COVID-19 z vykorystanniam klasternoho analizu. Visnyk SumDU. Seriia «Ekonomika». № 1. 2021. S. – 312 – 317.

2. Zimovin O.I. Sposoby perezhyvannia tryvohy v umovakh pandemii covid-19. Dosvid perezhyvannia pandemii covid 19: materialy onlain seminariv 23 kvitnia 2020 roku «Dosvid karantynu: dystantsiina psykholohichna dopomoha i pidtrymka» ta 15 travnia 2020 roku «Dystantsiini psykholohichni doslidzhennia v umovakh pandemii covid 19 i karantynu». Kyiv. 2020. S. 91 – 96.

3. Bazilevych K.O., Meniailov Ye.S., Chumachenko D.I.. Vydilennia zon rozpovsiudzhennia zakhvoriuvanosti na koronavirus covid-19 na osnovi metodiv klasternoho analizu. Suchasnyi stan naukovykh doslidzhen ta tekhnolohii v promyslovosti. 2021. № 1 (15). S. 5 – 13.

4. Ahmed Al-Imam. Clustering analysis of coronavirus disease 2019 pandemic / Ahmed Al-Imam, Usama Khalid Enjaz, Hend Al-Door. Asian Journal of Medical Sciences, 12(2):108-113. DOI: 10.71152/ajms.v12i2.3572.

5. Pérez-Ortega, J.; Almanza-Ortega, N.N.; Torres-Poveda, K.; Martínez-González, G.; Zavala-Díaz, J.C.; Pazos-Rangel, R. Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemo-graphic Factors at Municipal Level in Mexico. Mathematics 2022, 10, 2167. https://doi.org/10.3390/math10132167.

6. San-Cristobal R., Martín-Hernández R., Ramos-Lopez O. and others. Longwise Cluster Analysis for the Prediction of COVID-19 Severity within 72 h of Admission: COVID-DATA-SAVE-LIFES Cohort / J. Clin. Med. 2022, 11(12), 3327; https://doi.org/10.3390/jcm11123327

7. Pérez-Ortega J., Nely Almanza-Ortega N., Torres-Poveda K., and others. Application of Data Science for Cluster Analysis of COVID-19 Mortality According to Sociodemographic Factors at Municipal Level in Mexico / Mathematics, 10 (13):2167, https://doi.org/10.3390/math10132167.

8. Shih, D.-H.; Shih, P.-L.; Wu, T.-W.; Li, C.-J.; Shih, M.-H. Cluster Analysis of US COVID-19 Infected States for Vaccine Distribution. Healthcare 2022, 10, 1235. https://doi.org/10.3390/healthcare10071235.

9. Soni Madhulatha T. An overview on clustering methods. IOSR Journal of Engineering Apr. 2012, Vol. 2(4) pp: 719-725.

10. Pistunov I.M., Antoniuk O.P., Turchaninova I.Iu. Klasternyi analiz v ekonomitsi. Navch. Posibnyk. Dnipropetrovsk: Natsionalnyi hirnychyi universytet. 2008. 84 s.

11. Statystycheskye metodы dlia ЭVM. Pod red. K. Эnsleina, Э. Relstona, H.S. Uylfa: Per. s anhl. Pod red. M.B. Maliutova. – M.: Nauka. Hl. red. fyz.-mat. Lyt., 1986. – 464 s.

12. Yarmish G., Distributed Lance-William Clustering Algorithm / Yarmish G., Listowsky P., Kings-borough D.S., 1709.06816v1 PDF (arxiv.org).

13. Gagolewski M., Cena A., James S., Beliakov G., Hierarchical clustering with OWA-based linkages, the Lance–Williams formula, and dendrogram inversions, Fuzzy Sets and Systems 473, 108740, 2023, DOI:10.1016/j.fss.2023.108740.

14. Boiko N. I., Hazdiuk K. P. Porivniannia rehresiinykh modelei za naiavnosti vykydiv u nabori riznotypovykh danykh. 2023. Scientific Bulletin of UNFU, 33(2), 84-91. https://doi.org/10.36930/40330212