2. МОДЕЛЮВАННЯ РОЗВИТКУ ІНФРАСТРУКТУРИ СІЛ НА ОСНОВІ ГРАФОВИХ ДАНИХ

Єлизавета Волкова, аспірант
Кафедра математичного моделювання і аналізу даних
Навчально-науковий Фізико-технічний інститут
Національний технічний університет України
«Київський політехнічний інститут імені Ігоря Сікорського»

yelvol-ipt22@lll.kpi.ua

ВСТУП

Розвиток сільської інфраструктури залишається одним із ключових завдань як для України, так і для багатьох інших країн світу. Нерівномірний розподіл ресурсів та відмінності у рівні розвитку інфраструктури між міськими і сільськими районами посилюють соціальні та економічні нерівності, що гальмують загальний розвиток країни. Це питання особливо актуальне для України, де значна частина населення проживає у сільських регіонах із недостатньо розвиненою інфраструктурою. Нерівномірний доступ до базових послуг, таких як охорона здоров'я, освіта, транспорт і зв'язок обмежує можливості цих територій та знижує якість життя населення.

Традиційні підходи до оцінки інфраструктурних потреб сільських громад виявляються недостатніми, оскільки вони часто не враховують специфічні відмінності між селами. Це призводить до узагальнених рішень, які не здатні ефективно адресувати локальні проблеми.

Інтеграція методів геопросторового аналізу з методами машинного навчання пропонує можливість створення більш точних та адаптованих до контексту моделей для оцінки інфраструктурних потреб сільських громад.

2.1. СУЧАСНІ ПІДХОДИ ДО ОЦІНКИ РОЗВИТКУ СІЛЬСКИХ ГРОМАД

Геопросторовий аналіз відіграє важливу роль у вивченні розвитку сільських територій, забезпечуючи аналітичні

інструменти для оцінки інфраструктури та соціально-економічних умов. Сучасні дослідження демонструють, як різні методи геоінформаційних технологій можуть сприяти більш ефективному управлінню сільськими територіями.

В якості прикладу можна навести працю [1], яка фокусується на дослідженні просторових характеристик сільських поселень у провінції Цзянсі (Китай). В цьому дослідженні застосовуються методи просторового аналізу, такі як індекс Морена (*англ.* Moran's I) та ядерна щільність [1], що дозволяє виявити нерівномірний розподіл поселень за схемою «щільно на півночі, рідко на півдні».

Важливим внеском цього дослідження є введення індексу соціально-екологічної оцінки SEI (*англ.* socio-environmental evaluation index), який допомагає ідентифікувати пріоритетні території для інфраструктурного втручання. Наведені індекси дозволяють зконцетрувати увагу на соціально-економічних та природних факторах, що впливають на розташування сільських поселень, і пропонують нові підходи до планування розвитку сільських територій на основі геоінформаційних технологій.

В [2] оцінка якості життя у сільських районах України здійснюється на основі геопросторового аналізу. Авторами запропоновано алгоритм оцінки доступу до об'єктів інфраструктури, таких як лікарні, школи та магазини, в якому також врахована ступінь близькості до природних ресурсів та зони конфліктів. Це дозволило виявити значні регіональні відмінності в якості життя сільських громад, зокрема, найгірші умови спостерігаються на сході та півдні України. Отримані результати [2] також підкреслюють важливість використання методів кластеризації для аналізу інфраструктури та розробки стратегій її покращення.

В праці [3] досліджувалось застосування геопросторових технологій для картографування та аналізу соціальних та інфраструктурних об'єктів у селі Чиннапенд'яла (Індія). Дані дослідження ілюструють, як геоінформаційні системи можуть використовуватись для створення детальних баз даних на місцевому рівні, що допомагає краще розуміти потреби громади і покращувати планування їх розвитку в цілому. Особлива увага приділяється методам картографування та аналізу доступності об'єктів, таких як школи, лікарні та транспортні маршрути. Дослідження демонструє важливість індивідуального підходу до планування на рівні сіл та їх специфічних потреб.

Всі ці дослідження підкреслюють універсальність і важливість геопросторового аналізу для оцінки інфраструктури сільських

районів у різних географічних та соціально-економічних контекстах, демонструючи, як відкриті дані та сучасні технології можуть сприяти рівномірному розвитку сільських територій.

Практичне значення оглянутих досліджень полягає у формуванні ефективних інструментів для прийняття управлінських рішень, що дозволяє визначати пріоритетні напрямки інфраструктурних втручань у сільських районах. Завдяки використанню геоінформаційних технологій та аналізу соціально-економічних даних, ці дослідження надають можливість оптимізувати процеси планування, спрямовані на забезпечення збалансованого розвитку сільських територій.

2.1.1. ГЕОПРОСТОРОВА ХАРАКТЕРИСТИКА СІЛЬСЬКИХ ПОСЕЛЕНЬ У ПРОВІНЦІЇ ЦЗЯНСІ

Розглянемо детальніше застосування геопросторового аналізу для вивчення сільських поселень на прикладі провінції Цзянсі (Китай) [1]. Це дослідження підкреслює важливість інтеграції геоінформаційних технологій та соціально-економічних даних для розробки ефективних стратегій відновлення та покращення умов життя на сільських територіях.

Метою дослідження є виявлення просторових закономірностей розподілу сільських поселень та визначення факторів, що впливають на цей процес. Окрім цього, ставиться завдання розробки індексу соціально-екологічної оцінки SEI, який може використовуватись для оцінювання нерівномірності розвитку сільських територій та виявлення цільових зон для покращення інфраструктури.

Для досягнення зазначених цілей застосовуються наступні методи геопросторового аналізу:

1. *Аналіз ядерної щільності* (*англ.*, Kernel Density Estimation, KDE). Цей метод дозволяє оцінити щільність сільських поселень у визначеному радіусі. Формула ядерної щільності виглядає наступним чином:

$$f(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{d_i}{h}\right), \quad (1)$$

де $f(x, y)$ — оцінка щільності в точці (x, y), n — кількість поселень, h — ширина ядерного вікна, K — ядерна функція, d_i — відстань від i-го поселення до точки (x, y).

2. *Просторова автокореляція* (*англ.*, Spatial Autocorrelation). Для визначення ступеня взаємозв'язку між об'єктами в просторі застосовується індекс Морена, який має наступний вигляд:

$$I = \frac{n}{W}\frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}, \tag{2}$$

де I — індекс Морена, n — кількість поселень, W — сума всіх вагових коефіцієнтів w_{ij}, x_i і x_j — значення змінної для поселень i та j, $\bar{x}$ — середнє значення змінної.

3. *Регресійний аналіз* (*англ.*, Regression Analysis). Для визначення факторів, що впливають на географічний розподіл сільських поселень, використовуються прості та множинні лінійні регресійні моделі. Множинна лінійна регресія описується наступним чином:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon, \tag{3}$$

де Y — залежна змінна (розподіл поселень), X_1, X_2, . . . , X_n — незалежні змінні (соціально-економічні та екологічні фактори), β_0 — вільний член, β_1, β_2, . . . , β_n — коефіцієнти регресії, ϵ — похибка моделі.

4. *Найближча відстань сусіда* (*англ.,* Nearest Neighbor Distance, NND). Метод найближчої відстані сусіда обчислює індекс найближчого сусіда (*англ.*, Nearest Neighbor Ratio, NNR):

$$r_E = \frac{1}{2}\sqrt{\frac{n}{S}} = \frac{1}{2}\sqrt{\omega}, \quad \text{NNR} = \frac{r_i}{r_E}, \tag{4}$$

де r_E — це теоретично очікувана (середня) відстань до найближчого сусіда, n — кількість сільських поселень, S — площа досліджуваної території, ω — густина точок, r_i — це середня фактична відстань до найближчого сусіда, а NNR — індекс найближчого сусіда.

Результати [1] вказують на значну різницю у щільності поселень між північними та південними районами провінції Цзянсі. Одним з ключових внесків даного дослідження є розробка індексу соціально-екологічної оцінки (*англ*,. Socio-Ecological Index, SEI), який може стати основою для стратегій відновлення

сільських територій. Індекс SEI можна представити наступним чином:

$$\text{SEI} = \frac{1}{1 + e^{-(a+bX_1+cX_2+dX_3+eX_4)}}, \quad (5)$$

де SEI — індекс соціально-екологічної оцінки, X_1 — дохід на душу населення, X_2 — ресурси орних земель, X_3 — середня висота, X_4 — середній нахил, a, b, c, d, e — коефіцієнти моделі.

Результати дослідження [1] підтверджують, що сільські поселення провінції Цзянсі мають чітко виражену тенденцію розподілу за принципом «щільні на півночі – рідкі на півдні». Поселення здебільшого розташовані на низьких висотах, рівнинних територіях, в районах з високою щільністю річок та доріг, а також у регіонах, багатих на орні землі. Це пояснюється комплексом фізичних та соціально-економічних факторів.

Запропонований індекс SEI демонструє значний методологічний прогрес у геопросторовому аналізі, надаючи детальніший підхід до вивчення проблем та можливостей розвитку сільських територій.

2.1.2. ГЕОПРОСТОРОВИЙ АНАЛІЗ ЯКОСТІ ЖИТТЯ У СІЛЬСЬКИХ РАЙОНАХ УКРАЇНИ

На кафедрі математичного моделювання та аналізу даних запропоновано алгоритм оцінки якості життя у сільських районах України з використанням агрегації геопросторової інформації з різних джерел [2].

Підхід полягає у комплексній оцінці віддаленості села від життєво важливих інфраструктурних об’єктів (лікарні, навчальні заклади, банки, бібліотеки, магазини, дороги, лінії електропередач тощо), до природних екосистем (водойми, ліси або парки), а також до окупованих територій.

Результати проведених досліджень показали, що найбільша кількість сіл з депресивною якістю життя розташована у східних та південних регіонах країни, тоді як позитивна якість переважає у західній та центральній Україні. Це, безумовно, частково пов’язано з активними бойовими діями.

Для визначення якості життя у селах було використано 16 показників. Для кожного села були розраховані найкоротші

відстані по кожному з показників. Також, для кожного показника було визначено інтервали для трьох класів: 0 – позитивне розташування, 1 – середнє розташування, 2 – депресивне розташування. Для кожного села та кожного показника відстані були розраховані за допомогою геоінформаційних систем.

Градація якості життя для кожного показника i в інтервалі $[0, up\_val]$ було поділено на три рівні частини, а градація якості життя для села j та показника i визначалась наступним чином:

$$Gradation_{j,i} = \begin{cases} 0, & \text{якщо } dist_{j,i} \in \left[0, \frac{up\_val}{3}\right], \\ 1, & \text{якщо } dist_{j,i} \in \left(\frac{up\_val}{3}, 2 \cdot \frac{up\_val}{3}\right], \\ 2, & \text{якщо } dist_{j,i} > 2 \cdot \frac{up\_val}{3}, \end{cases} \quad (6)$$

де $dist_{j,i}$ — відстань для села j та показника i, up_val – верхня межа інтервалу, визначена за правилом 3 сигм.

Наступним кроком було об'єднання отриманих градацій для різних показників в один загальний показник, що відображає загальну картину життя в кожному селі окремо. В якості такого показника було обрано суму отриманих градацій для кожного i:

$$Life\_level_j = \sum_{i=1}^{N} Gradation_{j,i}. \quad (7)$$

Вважалось, що села, які сумарно набрали до 3 балів для всіх показників, мають позитивне розташування, від 3 до 10 – середнє, понад 10 – депресивне:

$$Life\_quality_j = \begin{cases} \text{позитивне, якщо } Life\_level_j \in [0,3], \\ \text{середнє, якщо } Life\_level_j \in (3,10], \\ \text{депресивне, якщо } Life\_level_j > 10. \end{cases} \quad (8)$$

На основі розрахованих градацій показників було визначено рівень якості життя для сіл України.

Проведені експерименти демонструють застосовність запропонованої методології оцінки якості життя у сільських поселеннях України за допомогою геопросторового аналізу. Запропонований підхід дозволяє агрегувати дані про віддаленість сіл від життєво важливих об'єктів і, таким чином, кількісно

оцінити їх рівень доступності. Результати показують значні регіональні відмінності, причому депресивні умови найбільш виражені у східних та південних областях. Градаційний підхід класифікує села за позитивними, середніми та депресивними категоріями на основі накопиченого балу. Це забезпечує відносний рейтинг і дозволяє визначити пріоритетні області для інтервенцій у розвиток. Не зважаючи на те, що в методі використовуються різноманітні показники, алгоритм може бути вдосконалений шляхом включення додаткових шарів даних.

2.1.3. ГЕОПРОСТОРОВІ ТЕХНОЛОГІЇ ДЛЯ КАРТОГРАФУВАННЯ ТА АНАЛІЗУ ПОСЕЛЕНЬ В ІНДІЇ

В [3] представлено детальне дослідження застосовності геопросторових технологій для картографування та аналізу соціальних та інфраструктурних об’єктів на рівні села, зокрема, на прикладі села Чиннапенд'яла в Індії. Дослідження демонструє, як геопросторові технології і, зокрема, геоінформаційні системи, дозволяють ефективно інтегрувати просторові та непросторові дані для побудови детальної геопросторової бази даних. Ця база є важливим інструментом для місцевих органів влади, що дозволяє краще розуміти та задовольняти потреби сільських громад на мікрорівні.

У [3] застосовуються кілька цікавих методів геопросторового аналізу, які спрямовані на покращення картографування та аналізу соціальних та інфраструктурних об’єктів у сільських районах.

Основні методи, що використовуються, включають наступні:

1. *Геоінформаційні системи* (ГІС). Цей метод дозволяє поєднувати просторові та непросторові дані для побудови детальної геопросторової бази даних. ГІС-технології допомагають виявити прогалини в інфраструктурі та точніше відображати наявні соціальні об’єкти, такі як школи, медичні установи, дороги тощо.

2. *Картографування та просторовий аналіз.* За допомогою ГІС виконано картографування інфраструктурних об'єктів для оцінки доступності різних послуг.

3. *Аналіз даних на мікрорівні.* Дослідження використовує інструменти геопросторового аналізу для збору та вивчення даних на рівні окремих сіл, що дає можливість детальніше зрозуміти потреби місцевих громад і точніше планувати розвиток сільських територій.

4. *Інтеграція просторових та непросторових даних.* Цей підхід дозволяє враховувати як географічні, так і соціально-економічні аспекти для прийняття більш обґрунтованих рішень щодо розвитку інфраструктури.

Ці методи допомагають створювати більш ефективну систему оцінки та подальшого розвитку інфраструктурних об'єктів на місцевому рівні.

2.1.4. ПРОСТОРОВО-ЧАСОВИЙ АНАЛІЗ ГЛОБАЛЬНИХ ДАНИХ ПРО МІСЬКІ БУДІВЛІ В OSM

За допомогою даних OpenStreetMap (OSM) аналіз повноти даних про міські будівлі можна проводити на основі наявних в них геопросторових характеристик сільських поселень. Дані OSM дозволяють оцінити масштаб та розподіл міських забудов, вивчаючи відмінності у наявності та якості даних, які можуть вплинути на комплексний аналіз та формування політик.

В [4] описується нерівномірність та неповнота даних про інфраструктуру в глобальній базі OSM. Основна увага зосереджена на аналізі даних для 13 189 урбанізованих територій по всьому світу. Автори використовують машинне навчання для оцінки повноти даних і констатують, що лише для 16% урбанізованих центрів OSM-дані про будівлі мають понад 80% повноти. Натомість для багатьох міст, особливо в країнах із середнім та високим індексом людського розвитку (*англ.* Human Development Index, HDI), дані є недостатньо повними, особливо в Латинській Америці, Південній і Південно-Східній Азії. Водночас в Африці, завдяки гуманітарним проєктам, в яких для розв’язання задач картографування беруть активну участь волонтери, дані є більш повними. β

Формально, запропоновану в [4] модель можна представити наступним чином:

$$\hat{A}_{building} = f(X) + \epsilon, \tag{9}$$

де $\hat{A}_{building}$ — прогнозована площа будівель, X — набір предикторів, що включає дані дистанційного зондування, субнаціональний індекс людського розвитку (*англ.* Sub-national Human Development Index, SHDI), щільність дорожньої мережі та інші змінні, а ϵ — стохастична складова (залишкова похибка).

Для кількісного оцінювання повноти даних про забудову OSM використовується наступна формула:

$$C_{\text{OSM}} = \frac{A_{OSM}}{A_{total}}, \tag{10}$$

де C_{OSM} — повнота даних OSM, A_{OSM} — площа будівельних слідів, зафіксованих в OSM, A_{total} — загальна прогнозована площа будівель.

Оцінка прогалин у даних OSM є критично важливою, оскільки вона безпосередньо впливає на ефективність використання геопросторової інформації в міському плануванні і може сприяти досягненню цілей сталого розвитку (*англ.* Sustainable Development Goals, SDG). Автори запропонували систему підходів для оцінки повноти даних OSM про будівлі, враховуючи такі фактори, як індекс людського розвитку, кількість населення та географічне розташування для побудови складних моделей просторової неоднорідності в покритті даних. Важливою частиною аналізу є використання коефіцієнта Джині (G):

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} \left|x_i - x_j\right|}{2n \sum_{i=1}^{n} x_i}, \tag{11}$$

де x_i і x_j — значення повноти даних у двох різних урбаністичних центрах, n — загальна кількість центрів, та коефіцієнта Морана (I):

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(x_i - \bar{x})\left(x_j - \bar{x}\right)}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2\right) \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}}, \tag{12}$$

де w_{ij} — просторові ваги між об'єктами i та j, $\bar{x}$ — середнє значення повноти даних, для вимірювання рівномірності та просторової автокореляції відповідно.

Таким чином в [4] проілюстрована необхідність інтеграції різних джерел даних та аналітичних методів для всебічного розуміння як сільської, так і міської інфраструктури. Це не лише доповнює сільський фокус попередніх розділів, але й розширює перспективу застосування відкритих геопросторових даних для аналізу інфраструктури на різних масштабах.

Дослідження [4] даних OSM про будівлі надає критичну оцінку поточного стану повноти геопросторових даних, закликаючи до більш справедливого розподілу зусиль по збору

даних. Автори підкреслюють важливість якісних даних для моніторингу розвитку міст і досягнення цілей сталого розвитку, що стосується безпечних та інклюзивних міст. Автори пропонують методи подолання прогалин у даних і покращення якості картографічних даних шляхом залучення місцевих спільнот та гуманітарних ініціатив. Ця праця надає важливу інформацію для дослідників та гуманітарних організацій щодо покращення оцінок та використання даних OSM у своїй роботі.

2.1.5. КЛАСТЕРИЗАЦІЯ ГРАФОВИХ ДАНИХ

Важливим аспектом у вищезазначених публікаціях є використання графових даних. Для прикладу, розглянемо працю [5], в якій були проаналізовані методи побудови графів, такі як граф ϵ-околу та граф k-найближчих сусідів. Автори встановили, що ці методи дозволяють суттєво зменшити чутливість до параметрів і покращити точність кластеризації. Зокрема, використання графу ϵ-околу дозволяє отримати більш надійні кластери, мінімізуючи вплив шуму у великих масивах даних. Цей метод допомагає виявити чіткі межі між кластерами навіть в умовах нерівномірного розподілу даних.

Ще однією важливою перевагою дослідження [5] є застосування методу k-найближчих сусідів, який дозволяє адаптуватися до локальних змін у структурі даних. Це робить його особливо корисним для аналізу складних графових структур у великих мережах, де традиційні методи кластеризації можуть виявитися неефективними. Завдяки цьому підходу вдалося досягти більшої точності в ідентифікації кластерів у багаторівневих мережах.

У [5] автори також підкреслюють важливість багаторівневого аналізу кластерів, яке дозволяє ефективно ідентифікувати структури на різних рівнях складності. Це надає дослідникам інструменти для детальнішого аналізу графів, що охоплюють різні типи взаємодій та зв'язків у мережах - від соціальних мереж до біологічних систем.

У публікації [6] описується використання теорії графів та геопросторового аналізу для планування електрифікації сільських районів. Результати показують, що інструмент GISEle (https://github.com/Energy4Growing/gisele_v01), який використовує кластеризацію на основі густини та теорію графів, оптимізує мережеву топологію для більш ефективного розвитку

інфраструктури. Зокрема, було встановлено, що використання цього інструменту дозволяє знизити вартість електрифікації на 15% завдяки оптимальному розташуванню електричних підстанцій та прокладенню ліній електропередач. Аналіз також виявив, що використання геопросторових даних сприяє більш точному визначенню потреб в електрифікації в різних регіонах, що дозволяє більш ефективно розподіляти ресурси.

У дослідженні [7] аналізуються сільські райони Індії за допомогою кластеризації на основі соціально-економічних характеристик. Вони встановили значні нерівності в розвитку та виявили тенденції, що можуть бути адаптовані для аналізу українських сіл. В [7] для групування сіл на основі 150 різних змінних, зібраних з урядових джерел, використовуються потужні методи кластеризації. В результаті було запропоновано чотирикластерну стратегію, яка дозволяє детальніше зрозуміти різноманітність сільських регіонів Індії та їхні специфічні потреби. Ця стратегія підкреслює важливість індивідуального підходу до планування та розвитку сільських територій, що дозволяє більш точно визначати пріоритети політики та розподілу ресурсів. Ці дослідження демонструють різноманітність підходів до кластеризації графових даних та їх успішне застосування в різних контекстах, що підкреслює їхню користь для аналізу та планування. Всі ці підходи базуються на використанні сучасних методів аналізу даних, що дозволяє досягати більш точних і надійних результатів у вивченні складних соціально-економічних систем.

Огляд сучасних підходів до оцінки розвитку сільських територій показав, що застосування вищезазначених методів є ефективним як на міжнародному рівні, так і в українських реаліях. У наступних підрозділах буде розглянуто методологію, що дозволяє адаптувати та розвинути ці підходи для розв'язання конкретних задач оцінки ступеня розвитку інфраструктури сільских громад в Україні.

2.2. ПОСТАНОВКА ЗАДАЧІ ОЦІНКИ РОЗВИТКУ СІЛЬСЬКОЇ ІНФРАСТРУКТУРИ

В даному розділі запропонована методологія для вирішення задачі оцінки поточного стану та покращення ступеня розвитку сільської інфраструктури в Україні на основі геопросторових даних і методів кластеризації за допомогою побудови системи

класифікації сільських громад за рівнем розвитку їх інфраструктури, що дозволить ідентифікувати розриви в доступності важливих об’єктів інфраструктури та визначити пріоритетні зони для інвестицій і планування подальшого розвитку.

Для цього пропонується використати нові оцінки доступності, отримані з відкритих джерел, таких як OpenStreetMap (OSM) та Humanitarian Data Exchange (HDX). До таких оцінок перш за все слід віднести відстані до ключових об’єктів, таких як дороги, медичні заклади, освітні установи та інші об’єкти інфраструктури.

Ключовим питанням є також застосування методів машинного навчання, таких як KMeans і DBSCAN, для сегментації сіл за рівнем розвитку інфраструктури, а також розробка інструментів візуалізації, що забезпечать легкий доступ до результатів аналізу. Це дослідження спрямоване на формування бази для стратегічного планування розвитку сільських районів, а також на створення інструментів для прийняття рішень стейкхолдерами та спеціалістами по розвитку інфраструктури.

2.3. ВИКОРИСТАНІ ДАНІ

Для досягнення поставленої мети було використано різноманітні геопросторові шари даних, вилучені з OpenStreetMap (OSM) та оброблені у форматі GeoDataFrame (GDF). Дані OSM, отримані на основі роботи волонтерів (citizen science), даних аерофотозйомки, GPS-пристроїв та польових обстежень, є відкритими для використання і містять інформацію про дороги, будівлі, природні об'єкти, що робить цей ресурс цінним для подальшого геопросторового аналізу.

У цьому підрозділі використано шари даних, що охоплюють різні типи доріг (основні, вторинні та сільські), типи земельного покриву, а також розташування об'єктів соціальної інфраструктури, таких як школи, університети, лікарні, аптеки, торгові центри, банки, церкви, бібліотеки та парки різного рівня (місцеві, національні, регіональні).

Додаткові дані про розташування населених пунктів було отримано з платформи Humanitarian Data Exchange (HDX) (https://data.humdata.org/group/ukr) в Україні, що були актуальними на середину 2021 року програми Copernicus (https://www.copernicus.eu). HDX є відкритою платформою для

обміну даними, що стосується гуманітарних криз та стану країн, і керується Офісом ООН з координації гуманітарних питань (*англ.,* Office for the Coordination of Humanitarian Affairs, OCHA). Для цього дослідження використовувалися дані про розташування сіл та міст.

Використаний набір даних містить 28381 записи, кожен з яких відповідає унікальному географічному місцю в сільській Україні. Ці записи ретельно деталізовані по 41 інформативним колонкам, пропонуючи панорамний огляд ландшафту сільської інфраструктури через призму геопросторової аналітики. Однак у цьому дослідженні використано лише 17 параметрів типу «відстань до найближчого об'єкта» та 13 графових описів. Усі ці дані були зібрані в результаті обробки великого обсягу геопросторових даних з баз даних OSM, HDX, інформаційних ресурсів з інформацією про зернові елеватори [9], поштові провайдери [10, 11], а також дані про стільниковий зв'язок [12].

Міри близькості інфраструктури. Значна частина набору даних містить оцінку міри близькості, яка кількісно оцінює близькість кожної сільської громади до різних критичних елементів інфраструктури, і тим самим є фундаментальною складовою для подальшого просторового аналізу. Усі ці об'єкти інфраструктури згруповані за типами та описані в табл. 1. Частково дані міри були сформовані в [2], а також в інших дослідженнях, зокрема [8].

Таблиця 1 Дані про близькість інфраструктури

Тип	Об'єкти	Опис
Дороги	RD_m1_NEAR, RD_m2_NEAR, RD_m3_NEAR	Відстань до основних, регіональних та сільських доріг відповідно.
Міста	CITY2_NEAR, Kyiv_NEAR_	Відстань до найближчого міста та столиці України відповідно.
Парки	LokPark_NE, NatPark_NE, regPark_NE	Відстань до найближчого місцевого, національного та регіонального парку відповідно.

Елеватори	Elevators_	Відстань до найближчого елеватора.
Дитячий садок	Kinder_NEAR	Відстань до найближчого дитячого садка.
Банк	Bank_NEAR_	Відстань до найближчого банку.
Церква	Cerkva_NEA	Відстань до найближчої церкви.
Освіта	Education_	Відстань до найближчого навчального закладу.
Готелі	Hotels_NEA	Відстань до найближчого готелю.
Бібліотека	Library_NE	Відстань до найближчої бібліотеки.
Лікарня	Likarni_NE	Відстань до найближчої лікарні.
Магазин	Magaz_NEAR	Відстань до найближчого магазину.
Пошта	NP_Min_Dist, UP_Min_Dist	Відстань до найближчого відділення Укрпошти та Нова пошти відповідно.
Мобільний зв’язок	kyivstar-4g, kyivstar-3g, vodafone-4g, vodafone-3g, lifecell-4g, lifecell-3g, trimob	Наявність мобільного оператора з певними технологіями.

Графові представлення. У сформованому наборі даних частина параметрів містить елементи опису графових структур, запропонованих в [8, 17]. Наприклад графові колонки (graph_city, graph_local_park тощо) містять дані у форматі JSON у вигляді масиву, що пропонує детальний погляд на найближчі об’єкти різних типів. Це детальне представлення дозволяє проводити розширений аналіз інфраструктурної мережі, включаючи дослідження зв’язності та просторового розподілу основних послуг. Більш детальний опис параметру кожного типу та приклад опису об’єкта наведено в табл. 2.

Таблиця 2 Графові дані

Тип	Приклад Об'єкта та його JSON-представлення	Опис
Graph_city	{ «id_type»: «admin4Pcod», «id»: «UA5323810100», «distance»: 15874.531539053276, «pos_x»: 814898.792138855, «pos_y»: 5565562.434442552 }	Стандартний опис міста, яке знаходиться поблизу села, з усіма основними характеристиками .
Graph_local_park	{ «id_type»: «KodPZF», «id»: «0253UA0708041», «distance»: 3695.6829331930207, «KatObPZF»: «Reserve», «AreaPZF»: 120, «pos_x»: 807527.095508027, «pos_y»: 5577813.50041332 }	Опис місцевого парку, крім стандартного, містить площу та категорію об'єкта KatObPZF.
Graph_national_park, graph_regional_park	{ «id_type»: «KodPZF», «id»: «0153UA0200001», «distance»: 10908.01990446648, «AreaPZF»: 12028.42, «pos_x»: 813282.7981963563, «pos_y»: 5570259.240116742 }	Опис національного та регіонального парків, містить площу об'єкта.
Graph_bank, graph_kindergarten, graph_library	{ «id_type»: «osm_id», «id»: «668736377», «distance»: 24589.561415524415, «pos_x»: 830942.7863028484, «pos_y»: 5593380.706043951 }	Опис банку, дитячих садків та бібліотек.

Graph_church, graph_edu, graph_hotel, graph_medicine, graph_shop	{ «id_type»: «osm_id», «id»: «298759370», «distance»: 6992.808427912616, «fclass»: «class», «pos_x»: 816907.4829874948, «pos_y»: 5580679.407344543 }	Опис церков, навчальних закладів, готелів, медичних установ та магазинів, де у полі «fclass» буде описано клас об'єкта.
Graph_elevator	{ «id_type»: «id», «id»: 746, «distance»: 2657.6559562758794, «pos_x»: 807667.363751653, «pos_y»: 5579215.811847648 }	Опис елеватора.

До кожного об'єкту прив'язаний ідентифікатор id_type. Це пов'язано з тим, що сформований набір даних представляє собою складну структуру, що базується на використанні різних пов'язаних окремих піднаборів даних. Це дозволяє формувати різні представлення без необхідності перебудови загальної структури даних.

2.4. МЕТОДОЛОГІЇ РОЗВ'ЯЗАННЯ ЗАДАЧІ КЛАСТЕРИЗАЦІЇ

В даному підрозділі розглядаються методи, які можуть бути застосовані для виконання дослідження. Вище вже було розглянуто метод KDE як один з можливих підходів. Однак, подальший аналіз показав, що методи кластеризації можуть бути більш придатними для розв'язання поставленої задачі оцінки та планування розвитку сільської інфраструктури.

Кластеризація — це метод машинного навчання, що дозволяє об'єднати схожі об'єкти в окремі групи або кластери. Це дуже важлива частина аналізу даних, оскільки вона допомагає виявити природні структури в даних без їх попереднього маркування. Метою кластеризації є мінімізація внутрішньогрупової варіації та максимізація міжгрупової варіації. Методи кластеризації більш детально розглянуті нижче.

- **KMeans**. Метод KMeans є одним із найпоширеніших алгоритмів кластеризації, метою якого є поділ n об'єктів на k

кластерів таким чином, щоб мінімізувати різницю між об'єктами в межах одного кластера та максимально збільшити відстань між кластерами. Кожен кластер має свій центроїд, і метод намагається мінімізувати відстані між об'єктами та їх центроїдами.

Для оцінки цієї відстані та якості розподілу використовується функція вартості, яка показує сумарну відстань між об'єктами та їхніми центроїдами. Чим менше значення цієї функції, тим краще розподіл об'єктів по кластерах.

Функція вартості, яку необхідно мінімізувати, визначається наступним чином:

$$J = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2, \tag{13}$$

де J — функція вартості, k — кількість кластерів, C_i — набір об’єктів у кластері i, μ_i — центроїд кластера i, x_j — об’єкт, що належить до кластера i.

Алгоритм реалізації методу KMeans можна визначити наступним чином:

1) Вибрати початкові центроїди випадковим чином.

2) Кожен об'єкт призначається до найближчого центроїда μ. Відстань до центроїда обчислюється за формулою:

$$\mu_i = \frac{1}{|C_i|} \sum_{x_j \in C_i} x_j. \tag{14}$$

3) Оновити центроїди μ як середнє значення об’єктів у кожному кластері.

4) Повторювати кроки 2 і 3 до досягнення збіжності, тобто поки центроїди не перестануть змінюватися.

Переваги методу кластеризації KMeans полягають в простоті його реалізації та високій ефективності для обробки великих наборів даних. До недоліків слід віднести чутливість до початкової ініціалізації центроїдів, необхідності попереднього визначення кількості кластерів та можливість роботи лише з числовими даними.

- **Ієрархічна кластеризація.** Ієрархічна кластеризація дозволяє побудувати ієрархію кластерів, яка може бути представлена у вигляді дендрограми. Існують два основні підходи: агломеративна (знизу вгору) та дивізійна (зверху вниз) кластеризація [14].

Алгоритм агломеративної ієрархічної кластеризації має наступний вигляд:

1) Розглянути кожен об'єкт як окремий кластер.

2) Знайти два найближчі кластери C_i та C_j та об'єднати їх у новий кластер C_{ij}, де відстань між кластерами визначається формулою:

$$d(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} \| x - y \|. \quad (15)$$

3) Повторювати крок 2, поки всі об'єкти не будуть об'єднані в один кластер.

Переваги методу ієрархічної кластеризації полягають в відсутності необхідності вказувати кількість кластерів заздалегідь та можливості працювати з будь-яким типом даних.

До недоліків слід віднести високі обчислювальні витрати для великих наборів даних та чутливість до шуму та викидів.

- **DBSCAN** (Density-Based Spatial Clustering of Applications with Noise) — алгоритм кластеризації, що базується на густині даних. Він дозволяє виявити кластери довільної форми та не враховувати шуми [2].

Для кожної точки P в наборі даних D визначаються ϵ-околиці:

$$N_\epsilon(P) = \{ q \in D \mid \text{dist}(P, q) \leqslant \epsilon \}, \quad (16)$$

де $N_\epsilon(P)$ — множина точок в радіусі ϵ від точки P.

Алгоритм DBSCAN має наступний вигляд:

1) Обрати точку P та знайти всі точки в радіусі ϵ від P (ϵ-околиці).

2) Якщо ϵ-околиця містить щонайменше MP точок, створити кластер з P.

3) Повторити процес для всіх точок у кластері шляхом його розширення.

4) Повторити кроки 1–3 для всіх наступних точок.

Переваги методу DBSCAN полягає в відсутності необхідності вказувати кількість кластерів заздалегідь, можливості виявляти кластери довільної форми та в стійкості до шуму та викидів.

З недоліків – можлива складність вибору параметрів ϵ та MP та високі обчислювальні витрати для великих наборів даних.

Для вибору найкращого методу для розв'язання поставленої задачі розглянемо наявні дані більш детально, та підберемо метод,

який дозволить розв'язати задачу розподілу на кластери найкращим чином. Задача розподілу на кластери полягає в групуванні об'єктів, що знаходяться в схожому віддалені від нульової координати.

Для цього введемо вираз (17) для аналізу відмінності відстаней між об'єктами одного кластера та оглянемо кластери, побудовані за допомогою методів графічного представлення, розфарбованих за належністю до кластерів (рис. 1).

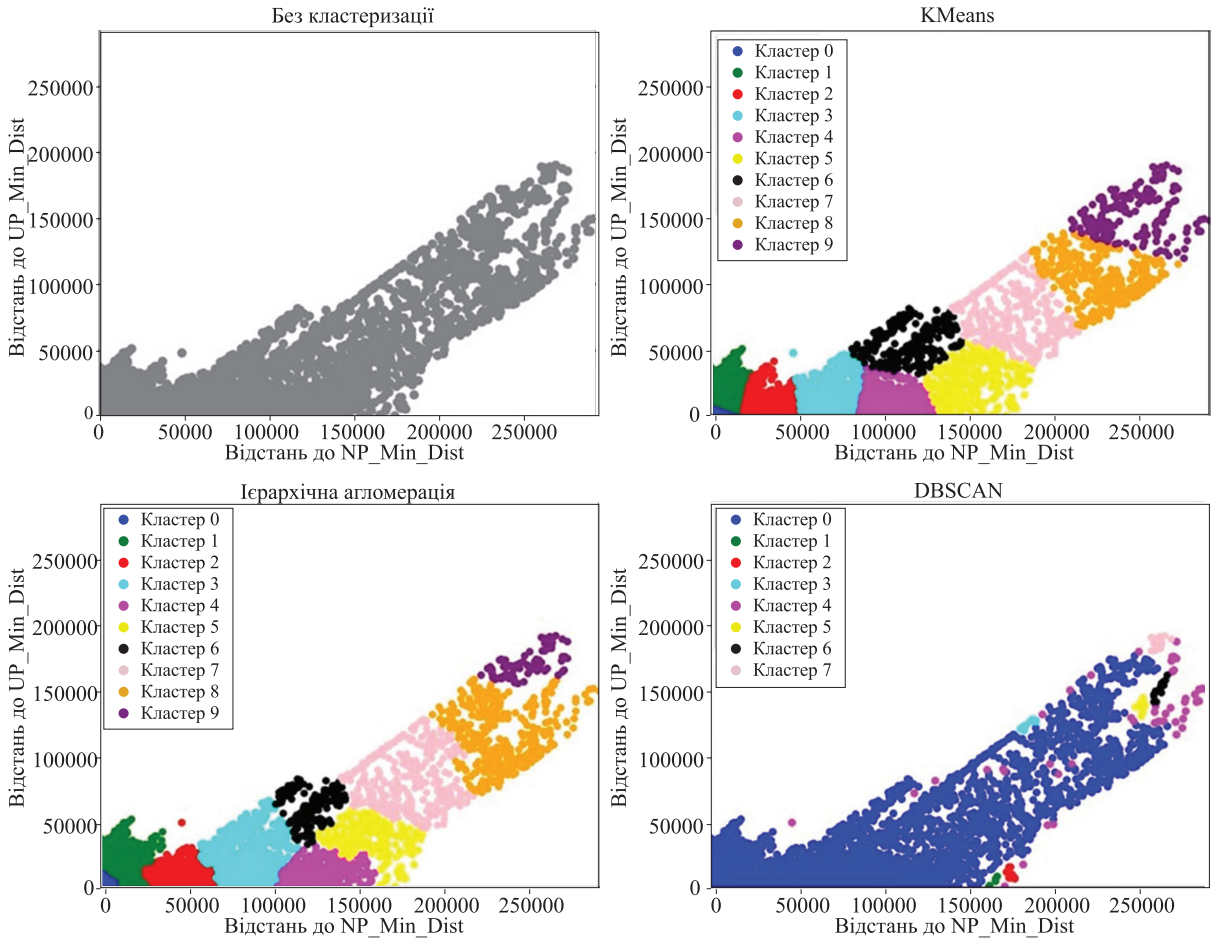


Рис. 1. Порівняння методів кластеризації: (а) без кластеризації, (б) KMeans, (в) ієрархічна агломерація, (г) DBSCAN

На рис. 1 представлені результати застосування методів кластеризації до сформованого набору даних відносно поштових відділень. З наведеного рисунку бачимо, що дані рівномірно розташовані на графіку та фактично є однією групою. Тому можна зробити висновок, що метод DBSCAN проявляє себе значно гірше порівняно з KMeans та ієрархічною агломерацію, які не зважаючи на те, що дані відносяться до одної групи, змогли виокремити окремі кластери.

Введемо наступні позначення:

– $\bar{D}_k$: середнє значення суми абсолютних відхилень для кластера k.

– N_k: кількість об’єктів у кластері k.

– d_i: відстань i-го об’єкта у кластері k від початку координат (0, 0).

– $\bar{d}_k$: середнє значення відстаней всіх об’єктів у кластері k від початку координат (0, 0).

– K: загальна кількість кластерів.

– $\bar{D}$: загальне середнє значення суми абсолютних відхилень для всіх кластерів.

Для оцінки різниці відстаней між об’єктами одного кластера використаємо (17), де $\bar{D}_k$ обчислюється як середнє значення суми абсолютних відхилень відстаней кожного об’єкта в кластері k від середньої відстані всіх об’єктів цього кластера від початку координат.

$$\bar{D}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} \left| d_i - \bar{d}_k \right|. \tag{17}$$

Узагальнюючи це для всіх кластерів, $\bar{D}$ обчислюється як середнє значення всіх $\bar{D}_k$ для кожного кластера, що дозволяє отримати загальне уявлення про відхилення в усіх кластерах.

$$\bar{D} = \frac{1}{K} \sum_{k=1}^{K} \bar{D}_k. \tag{18}$$

Таким чином можна зробити висновок, що для розв’язання поставленої задачі кластеризації краще підходять методи KMeans та ієрархічна агломерація. Проте продовжимо аналіз.

Результати застосування формули (18):

– $\bar{D}$ для KMeans: 10487.38

– $\bar{D}$ для Ієрархічної агломерації: 10579.64

– $\bar{D}$ для DBSCAN: 11127.80

На рис. 2 представлені розподіли відстаней для кожного кластера. Порівнюючи значення на цих графіках, можна побачити,

що в середньому відстані від початку координат для кожного кластера виявилися найбільш щільними для KMeans, середньо щільними для ієрархічної агломерації та найменш щільними для DBSCAN. Також це підтверджується, значеннями, отриманими на основі (18). Таким чином, можна стверджувати, що для подальших досліджень найкраще підходить метод KMeans.

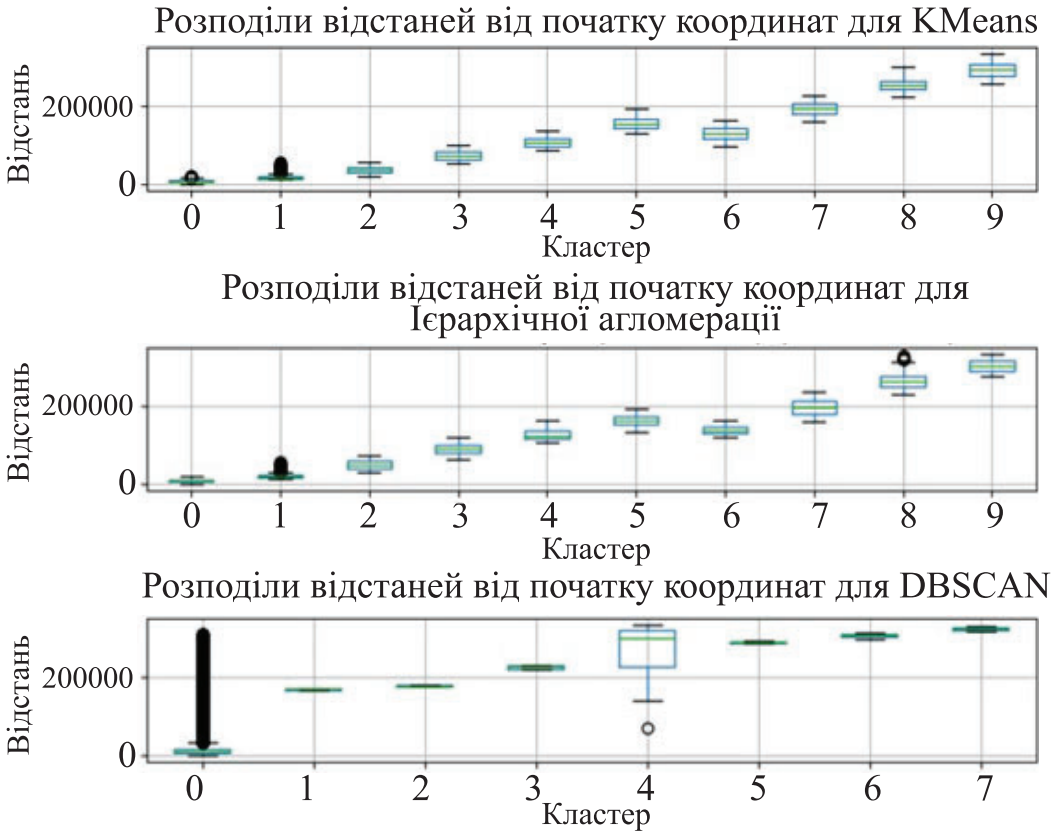


Рис. 2. Розподіли відстаней від початку координат для KMeans, ієрархічної агломерації, DBSCAN

Оптимізація кластеризованих даних. Завдяки алгоритму та природі функцій кластеризації номери кластерів не завжди відображають наближеність даного кластера до нульової координати. Оскільки, як було розглянуто вище, дані містять відстані до різних об'єктів, було б корисним відсортувати дані за віддаленістю від нульової координати, оскільки наближеність до неї означає, що інфраструктура знаходиться ближче до сіл, що в свою чергу означає, що вона є більш доступною, а отже, дане село є більш інфраструктурно розвиненим. Такий підхід також дозволяє виявити найбільш ізольовані села, які потребують пріоритетного розвитку. Врахування цих факторів сприяє більш точному аналізу та плануванню інфраструктурних проєктів.

Допоміжна кластеризація. Для вдосконалення розглянутого методу кластеризації введемо крок, який дозволяє реорганізовати кластери після їх формування. Цей крок ранжує кластери за якістю їх інфраструктури, від найменш до найбільш потребуючої покращення, використовуючи дані точок (центроїдів) з поточних результатів кластеризації. Цей метод не тільки полегшує розуміння стану інфраструктури кожного окремого кластера, а й дозволяє автоматично призначити кластери та отримати їх оцінки.

Введемо наступні позначення:

– $\bar{d}_k$ — середнє значення відстаней всіх об'єктів у кластері k від початку координат (0, 0).

– C_k: номер кластера k до допоміжної кластеризації.

– R_k: номер кластера k після допоміжної кластеризації (ранжування). Процес допоміжної кластеризації включає наступні кроки:

1) Обчислення середньої відстані до початку координат для кожного кластера:

$$\bar{d}_k = \frac{1}{N_k} \sum_{i=1}^{N_k} d_{i,} \tag{19}$$

де N_k - кількість об'єктів у кластері k, d_i - відстань i-го об'єкта до початку координат.

2) Сортування кластерів за зростанням $\bar{d}_k$ та їх перенумерація:

$$R_k = \text{rank}(\bar{d}_k), \tag{20}$$

де $\text{rank}(\bar{d}_k)$ - ранг кластера після сортування.

На рис. 3 зліва показана нумерація та розфарбовка кластерів відразу після кластеризації, а справа після раніше введеного кроку допоміжної кластеризації.

Приклад коду для порівняння допоміжної кластеризації та звичайного підходу:

```
# Function to plot pre-sorted and sorted clusters
def plot_pre_and_post_sorted_clusters(data):
    max_x_y = max(data[:, 0].max(), data[:, 1].max())
    # Determine the limits for the plots
    xlim = (0, max_x_y)
```

```
        ylim = (0, max_x_y)
        fig, axes = plt.subplots(1, 2, figsize=(12, 6))
        # Pre-sorted clustering using simple KMeans
        kmeans = KMeans(n_clusters=10, random_state=0)
        presorted_labels = kmeans.fit_predict(data)
        plot_2d_post_clusters(data, presorted_labels, 'Перед
сортуванням', axes[0], xlim, ylim)
        # Post-sorted clustering using apply_clustering
        sorted_labels, _ = apply_clustering(data, method='kmeans',
params={'n_clusters': 10})
        plot_2d_post_clusters(data, sorted_labels, 'Після сортування',
axes[1], xlim, ylim)
        plt.tight_layout()
        plt.savefig("pre_and_post_sorted_clustering.png",
bbox_inches='tight')
        plt.show()
        # Generate the plots
plot_pre_and_post_sorted_clusters(post_data_scaled)
```

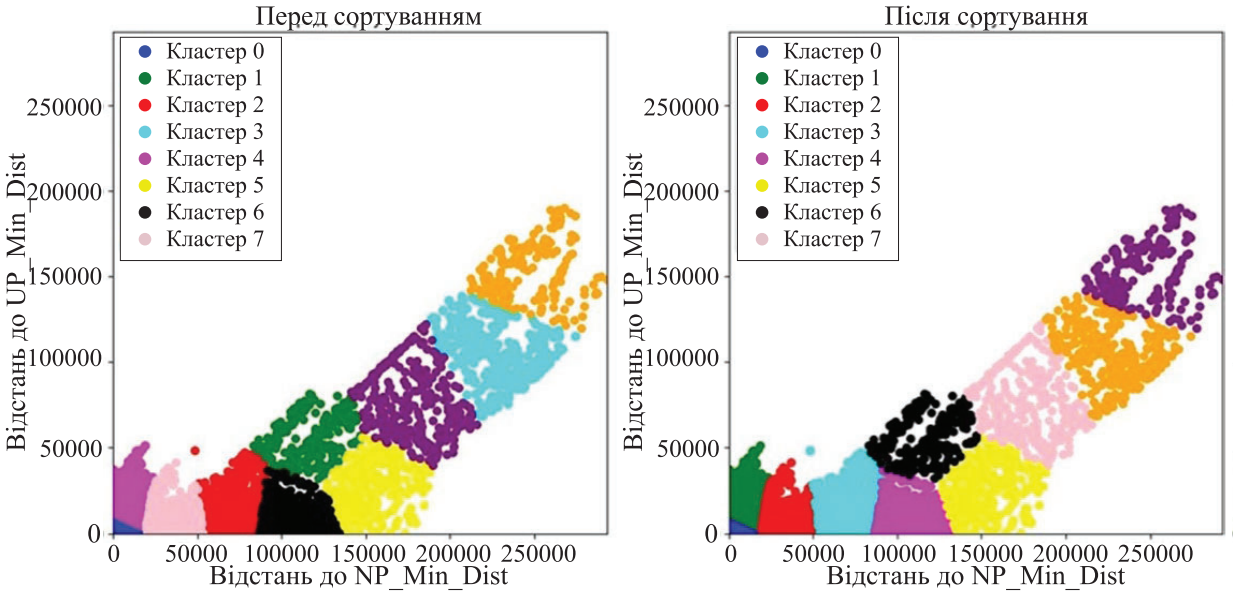


Рис. 3. Порівняння кластерів до та після кроку допоміжної кластеризації

Визначення методу ефективного аналізу інфраструктури для її подальшого покращення. Одним із ключових завдань цього дослідження є вибір методів, які були б ефективними для аналізу інфраструктури в різних сільських населених пунктах. Як вище

зазначено в табл. 1, пропонується додати кластер, що буде створюватися на основі результатів кластеризації відносно кожного типу інфраструктури. В комбінації з раніше розглянутими методами кластеризації та допоміжної кластеризації, можна визначити оптимальні кроки для покращення інфраструктури певного села за допомогою наступного алгоритму:

1) *Ідентифікація сусідніх сіл*. Визначити всі інші села в наборі даних, які були віднесені до різних кластерів на основі їх характеристик інфраструктури.

2) *Аналіз сусідніх кластерів*. Виявити села, що знаходяться в кластерах з вищим рівнем інфраструктурного розвитку. Це можна зробити шляхом порівняння значень кластерів, де вищі значення вказують на кращу інфраструктуру.

3) *Визначення найближчих сіл*. Для кожного кластера з вищим рівнем інфраструктури знайти найближче село до обраного села за характеристиками інфраструктури. Це здійснюється шляхом обчислення відстаней між векторами характеристик інфраструктури обраного села та сіл у вищих кластерах.

4) *Аналіз відмінностей*. Провести порівняльний аналіз між характеристиками інфраструктури обраного села та найближчих сіл з вищих кластерів. Це дозволить виявити конкретні елементи інфраструктури, які потребують покращення.

5) *Визначення дефіциту*. На основі аналізу відмінностей визначити конкретні аспекти інфраструктури, яких не вистачає обраному селу для його віднесення до вищого кластера.

В результаті застосування цього підходу зможемо отримати детальний опис типів інфраструктури, яких не вистачає певному селу для підвищення його кластерної класифікації. В комбінації з використанням додаткових експертних оцінок це дозволить досягнути високої ефективності в покращенні інфраструктури.

Оцінювання загальної якості інфраструктури. Для оцінки загальної якості інфраструктури кожного села пропонується інтегральний підхід, який дозволяє поєднати оцінки різних типів інфраструктури та рейтинг кластера після додаткової кластеризації. Формулу для обчислення інтегральної оцінки якості інфраструктури можна представити наступним чином:

$$\text{ОЯ} = a \times \left(\frac{1}{n} \sum_{i=1}^{n} \text{ОІ}_i \right) + b \times \text{РК}, \qquad (21)$$

де ОЯ – оцінка якості, ОІ – оцінка інфраструктури, РК – рейтинг кластера a, b — параметри, які визначають вагу кожного показника, n — кількість типів інфраструктури, Оцінка інфраструктури i — оцінка якості інфраструктури для кожного типу, рейтинг кластера — рейтинг кластера, який визначається після додаткової кластеризації.

Ці параметри обираються таким чином, щоб вони мали нормальний, логнормальний, чи гамма-розподіл, що дозволило б виокремити села, що є нормальними, та села, що є аномальними. Такий підхід дозволяє не лише оцінити поточний стан інфраструктури, а й виявити потенційні точки для покращення, що є важливим кроком для стратегічного планування розвитку регіону.

2.5. МЕТОДИ ВІЗУАЛІЗАЦІЇ ГЕОПРОСТОРОВИХ ДАНИХ

Для проведення комплексної оцінки геопросторових наборів даних, що стосуються сільської інфраструктури в Україні, застосовуються різні числові та графічні статистичні методи. Комбінація кількох підходів дозволяє отримати комплексне уявлення про розподіл даних та їх особливості.

Значну роль в аналітичному процесі відіграють гістограми, які забезпечують візуалізацію частотних розподілів відстаней до інфраструктурних об’єктів серед сільських поселень. Вони дозволяють швидко виявити характер розподілу — чи є він нормальним, чи має зміщення в напрямку більших або менших значень. Гістограми дозволяють також ефективно виявляти аномалії в даних, такі як викиди. Наприклад, завдяки такій візуалізації сільське поселення, яке знаходиться на аномально великій відстані від найближчої медичної установи, легко ідентифікується.

З технічної точки зору, гістограма — це графічне представлення, яке відображає розподіл частотних значень в наборі даних, надаючи зручний спосіб аналізу ключових характеристик розподілу. Таке представлення будується наступним чином:

$$H(x) = \frac{1}{n}\sum_{i=1}^{n} I\left(x_i \in \left[x_j, x_{j+1}\right)\right), \qquad (22)$$

де $H(x)$ — висота стовпчика гістограми для інтервалу $[x_j , x_{j+1})$, n — загальна кількість даних, I — індикаторна функція, яка дорівнює 1, якщо значення x_i потрапляє у відповідний інтервал, і 0 в іншому випадку.

Коробкові діаграми (*англ.*, box plot) дозволяють лаконічно і наочно відобразити внутрішню структуру розподілу даних, включаючи основні квартилі, викиди та діапазон екстремальних значень. Порівняння коробкових діаграм дозволяє швидко визначити схожості та відмінності медіанних значень, які відповідають різним категоріям інфраструктури. Наприклад, можна легко оцінити, яка частина сільських населених пунктів України знаходиться в межах 10 км від найближчої школи, або яка середня відстань до найближчого великого міста.

Коробкові діаграми показують п'ять наступних основних характеристик розподілу:

1. Нижній квартиль (Q_1),
2. Медіана (Q_2),
3. Верхній квартиль (Q_3),
4. Мінімальне значення (нижній ус),
5. Максимальне значення (верхній ус).

Коробкові діаграми також використовуються для візуалізації міжквартильного розмаху (*англ.*, Interquartile Range, IQR), який розраховується за формулою:

$$IQR = Q_3 - Q_1, \tag{23}$$

де IQR представляє центральну частину розподілу даних, вказуючи на діапазон значень, які охоплюють середні 50% сукупності. Цей інструмент є ефективним для швидкої оцінки варіацій даних та виявлення викидів, що сприяє глибшому розумінню розподілу досліджуваних даних.

У той час як графічні підходи вказують на властивості розподілу, кореляційні матриці дозволяють оцінити безпосередні зв’язки між доступністю лікарень, шкіл, доріг та інших об’єктів. Кореляційні коефіцієнти доволяють ідентифікувати аспекти сільської інфраструктури з найтіснішими зв’язками, спрямовуючи глибші дослідження на найбільш цікаві параметри.

Кореляція між двома змінними X та Y визначається наступним чином:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}, \tag{24}$$

де $\mathrm{cov}(X, Y)$ — коваріація між X та Y, σ_X та σ_Y — стандартні відхилення X та Y.

На додаток до навчання на основі візуалізації даних, використовуються дві фундаментальні статистичні функції.

Основні описові статистики можна формально визначити наступним чином:

– Середнє значення:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{25}$$

– Медіана — значення, яке розділяє набір даних на дві рівні частини

– Стандартне відхилення:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{26}$$

Ці статистичні характеристики надають важливу інформацію про розподіл даних у наборі. Наприклад, середнє значення вказує на типове значення в наборі даних, тоді як медіана дає уявлення про центральну точку розподілу. Стандартне відхилення описує, наскільки значення в наборі даних відхиляються від середнього, що важливо для оцінки варіацій рівня доступності інфраструктури.

Підсумкові статистичні дані, такі як середні значення, медіани та стандартні відхилення, підкреслюють центральні тенденції та варіації для відстаней до інфраструктури та додаткових особливостей. Була використана бібліотека Pandas для Python, яка пропонує зручні вбудовані функції для обчислення описової статистики через синтаксис на кшталт dataframe.describe().

Методи обробки геопросторових даних. Найпростішим є пошук відстані між селом та різноманітними точками інфраструктури. Даний підхід був розглянутий в [2]. Проте, наразі є прагнення розширити даний опис, імплементувавши нові оцінки даних.

Основна суть даного нововведення полягає в отриманні інформації про 5 найближчих об’єктів в певному радіусі від села.

Для цього недостатньо просто використати вбудовані методи мови програмування Python, оскільки методи, наприклад, реалізовані в бібліотеці Pandas, не дозволяють реалізувати ефективний пошук, а роблять повний перебір, що при 28000 тисяч сіл та 50000 об’єктів інфраструктури означає розрахунок 1.4 мільярдів відстаней, що потребує 388 годин обробки при обрахунку 1000 відстаней в секунду.

Підвищити ефективність обчислень можна за рахунок реалізації власного методу обробки даних, що дозволить отримати результати за адекватний час.

Методи створення розширеного набору даних з новими оцінками. Для розробки більш комплексного набору даних з додатковими параметрами оцінки сільських громад України пропонується використати систематичний процес виявлення та інтеграції точок інтересу (point of interest — POI) з вихідних геопросторових шарів даних у групи для кожного села, представлені у графовій формі.

Для кожного типу POI всі об’єкти спочатку сегментуються на буферні зони однакового заздалегідь визначеного розміру, що відповідає максимальній відстані, визначеній для кожного типу у вхідних даних. Після створення цих буферних зон можна перейти до виявлення найближчих об’єктів для кожного села. Це включає визначення, у якому буфері розташоване село, та визначення набору з восьми сусідніх зон.

Процес буферизації POI та визначення буферів пошуку представлено на рис. 4, а процес обчислення відстаней до POI та визначення найближчих — на рис. 5.

Цей підхід дозволяє створити всеосяжний набір POI поблизу кожного села, охоплюючи як різноманітність найближчих об'єктів (включаючи міста, парки тощо), так і визначення їхнього розташування на доступній відстані. Систематично каталогізуючи найближчі POI різних типів до кожного сільського поселення, можна отримати оцінки додаткових параметрів, що охоплюють доступність і наявність ключової інфраструктури та послуг.

Кінцевим результатом є набір даних у графовій структурі, що об’єднує села з їх найближчими POI по визначених категоріях разом з метриками відстані. Цей вихідний набір даних надає детальну інформацію щодо доступу до важливих сервісів на національному рівні. Структура графу, яка з’єднує сільські поселення з найближчими POI різних типів через зв’язки на

основі відстані, формує основу розширених геопросторових оцінок для аналізу сільської інфраструктури.

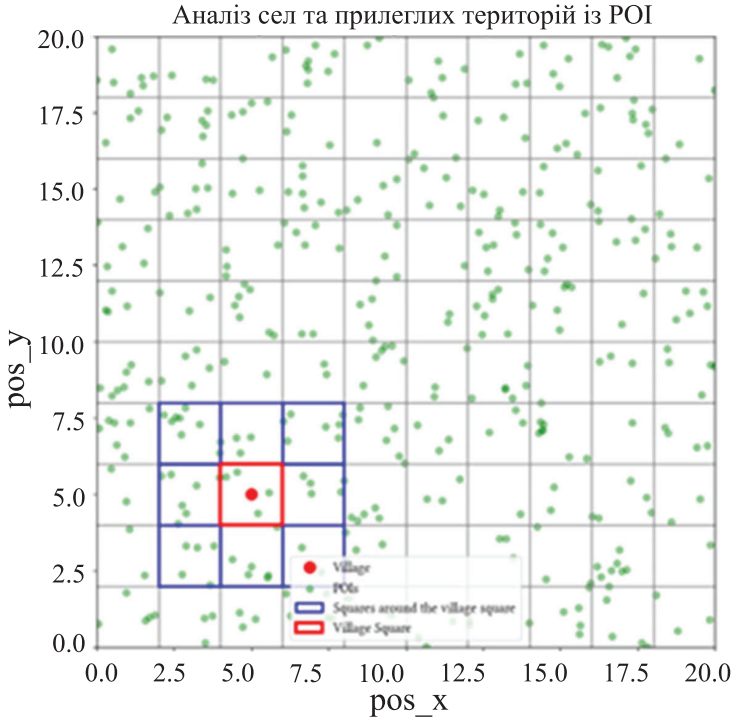


Рис. 4. Запропонований підхід для визначення регіонів для просторового аналізу

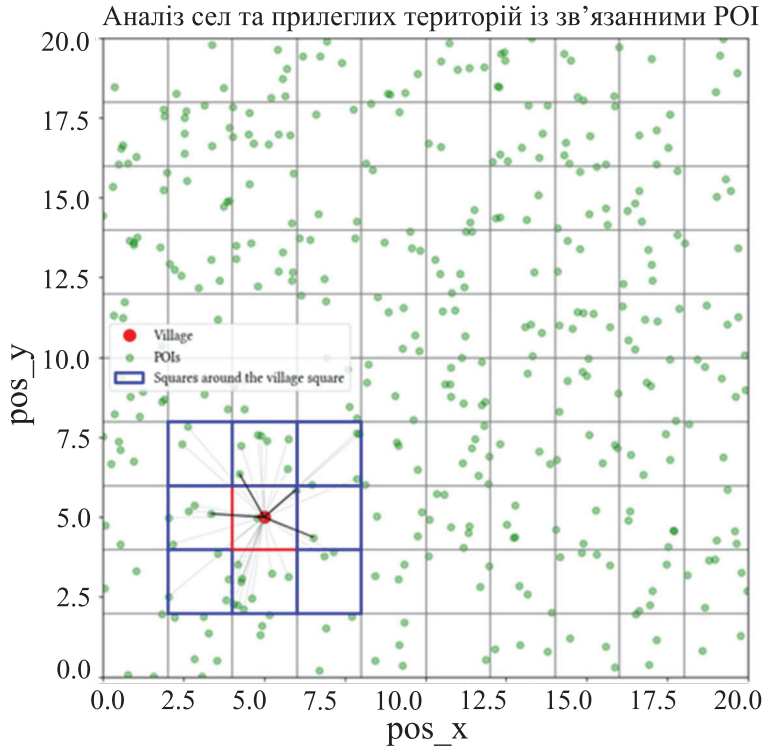


Рис. 5. Обчислення відстаней у визначених регіонах відповідно до запропонованої методології

Основні кроки побудови розширеного набору даних включають наступні:

Розбиття даних POI на окремі блоки. Кожний блок має розмір, що відповідає максимальній відстані, визначеній для кожного типу POI.

Визначення найближчих об'єктів для кожної сільської громади. Це включає визначення, у якому блоці розташоване село, та аналіз восьми сусідніх блоків.

Обмеження кількості об'єктів до заздалегідь визначеної кількості найближчих об'єктів для кожного села. Формула для обчислення евклідової відстані між селом і точкою інтересу (POI) виглядає наступним чином:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}, \tag{27}$$

де d — відстань між селом і POI, (x_1, y_1) — координати сільської громади, а (x_2, y_2) — координати POI.

Цей підхід дозволяє значно скоротити час обчислень і забезпечує точність результатів. Важливо зазначити, що для побудови таких графів використано ефективні алгоритми пошуку найближчих сусідів, такі як KD-дерева (k-d trees) та Ball-дерева, які забезпечують пошук за логарифмічний час:

$$O(n \log n), \tag{28}$$

де n - кількість об'єктів у наборі даних.

При такому підході кількість об'єктів для аналізу зменшилась до 500 об'єктів (в середньому) для кожної сільської громади, що покращило швидкість обчислення в 100 разів. За допомогою паралельних обчислень на 8 ядрах процесора отримано додаткове прискорення в 5 разів. Після оптимізації загального підходу всі обчислення займають близько 0.75 години, що є прийнятним часом для такого обсягу даних.

2.5.1. ВІЗУАЛІЗАЦІЯ РЕЗУЛЬТАТІВ КЛАСТЕРНОГО АНАЛІЗУ

Створення векторів для типів POI дозволяє отримати структуроване представлення просторових характеристик об'єктів, розташованих поблизу кожного села. У процесі проведення експерименту було сформовано вектори, що представляють різні

категорії POI (дороги, міста, парки) та складаються з набору значень, які вказують на відстані до критичних інфраструктурних об'єктів і графові дані.

Наприклад, вектор для доріг включає відстані до кількох найближчих доріг різного типу (RD_m1_NEAR, RD_m2_NEAR тощо), як показано в табл. 3. Такі вектори описують наявну інфраструктуру навколо кожної громади, що дає змогу аналізувати доступність інфраструктури та порівнювати громади між собою.

Таблиця 3 Формат вектора типу POI

Тип	**Вектор**
Дороги	(RD_m1_NEAR, RD_m2_NEAR, RD_m3_NEAR)
Міста	(Kyiv_NEAR_, CITY2_NEAR, obj_1, obj_2, obj_3, obj_4, obj_5)
Місцеві парки	(LokPark_NE, obj_1, obj_2, obj_3, obj_4, obj_5)
Національні парки	(NatPark_NE, obj_1, obj_2, obj_3, obj_4, obj_5)
Регіональні парки	(regPark_NE, obj_1, obj_2, obj_3, obj_4, obj_5)
Об’єднані парки	(значення кластера місцевих парків, значення кластера регіональних парків, значення кластера національних парків)
Банки	(bank_NEAR_, obj_1, obj_2, obj_3, obj_4, obj_5)
Церква	(cerkva_NEA, obj_1, obj_2, obj_3, obj_4, obj_5)
Освіта	(education_, obj_1, obj_2, obj_3, obj_4, obj_5)
Елеватори	(elevators_, obj_1, obj_2, obj_3, obj_4, obj_5)
Готелі	(hotels_NEA, obj_1, obj_2, obj_3, obj_4, obj_5)
Дитячі садки	(kinder_NEA, obj_1, obj_2, obj_3, obj_4, obj_5)
Бібліотеки	(library_NE, obj_1, obj_2, obj_3, obj_4, obj_5)
Медицина	(likarni_NE, obj_1, obj_2, obj_3, obj_4, obj_5)
Магазини	(magaz_NEAR, obj_1, obj_2, obj_3, obj_4, obj_5)
Загальний	(дороги, міста, парки, банки, церкви, освіта, елеватори, готелі, дитячі садки, бібліотеки, медицина, магазини)

Для відсутніх даних використовувалися значення-заповнювачі, що дозволило зберегти цілісність даних та полегшило подальший аналіз.

Одним із завдань, яке виникає при аналізі кластерів, є проблема візуального представлення даних, коли кількість вимірів

перевищує 2 чи 3. Для двовимірних і тривимірних даних можливе створення 2D та 3D візуалізацій, які забезпечують наочне відображення отриманих результатів. Однак для даних, що мають більшу кількість вимірів, доцільним є застосування теплових карт, які дозволяють ефективно візуалізувати результати кластеризації. Для цього використовуються нормалізовані дані, що дозволяє адекватно порівнювати кластери між собою.

Такий підхід спрощує представлення складних багатовимірних даних, підкреслюючи подібності та відмінності між кластерами. Нормалізація даних забезпечує справедливі порівняння, а використання теплових карт виступає ефективним інструментом для швидкого виявлення ключових тенденцій та відхилень у кластеризації.

Процес нормалізації даних включає наступні кроки:

1) Вибір показників, що будуть використовуватись для візуалізації.

2) Нормалізація цих показників з використанням формули:

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}, \tag{29}$$

де x — значення показника, x_{min} та x_{max} — мінімальне та максимальне значення цього показника відповідно.

3) Використання нормалізованих даних для візуалізації.

Генерація теплової карти. Після виконання кластеризації використаємо теплові карти як інструмент візуалізації нормалізованих даних для кожного кластера. Це допомагає наочно відобразити, наскільки подібні чи різні кластери між собою з врахуванням різних інфраструктурних показників. За допомогою додаткових описів, таких як кількість об’єктів у кластері, можна також зробити висновок про загальний стан розвитку певного типу інфраструктури. Приклад теплової карти наведено на рис. 6.

Приклад коду для створення функції відображення теплової карти нормалізованих даних для кластера:

```
    def heatmap_clusters(series, cluster_labels,
custom_row_labels=None, custom_column_labels=None,
file_name="Name"):
    data = np.stack(series.values)
    scaler = MinMaxScaler()
    data_normalized = scaler.fit_transform(data)
```

```
    unique_clusters = np.unique(cluster_labels)
    centroids_normalized = np.array([data_normalized[cluster_labels ==
k].mean(axis=0) for k in unique_clusters])
    if custom_row_labels is None:
        row_labels = [f"Cluster {k}" for k in unique_clusters]
    else:
        row_labels = custom_row_labels
    if custom_column_labels is None:
        column_labels = [f"Feature {j}" for j in range(data.shape[1])]
    else:
        column_labels = custom_column_labels
    centroids_df = pd.DataFrame(centroids_normalized,
index=row_labels, columns=column_labels)
    plt.figure(figsize=(12, 8))
    sns.heatmap(centroids_df, annot=True, fmt=".2f", cmap='viridis')
    plt.title("Heatmap of Normalized Cluster Centroids")
    plt.xlabel("Features")
    plt.ylabel("Clusters")
    plt.xticks(rotation=30)
    timestamp = datetime.now().strftime("%Y-%m-%d_%H-%M-%S")
    filename =
f"Clustering_of_{file_name}_{custom_column_labels[0]}.png"
    save_path = f"{filename}"
    plt.savefig(save_path, bbox_inches='tight')
    plt.show()
```

Застосування запропонованих методів до отриманих векторів та аналіз результатів. Нижче наведено аналіз отриманих результатів, які демонструють створення кластерів та їх відповідність очікуваним результатам. Важливим спостереженням є виявлення менш поширених POI, таких як місцеві парки, церкви, освітні заклади, готелі, дитячі садки та медичні установи, що відображено на кількох рисунках.

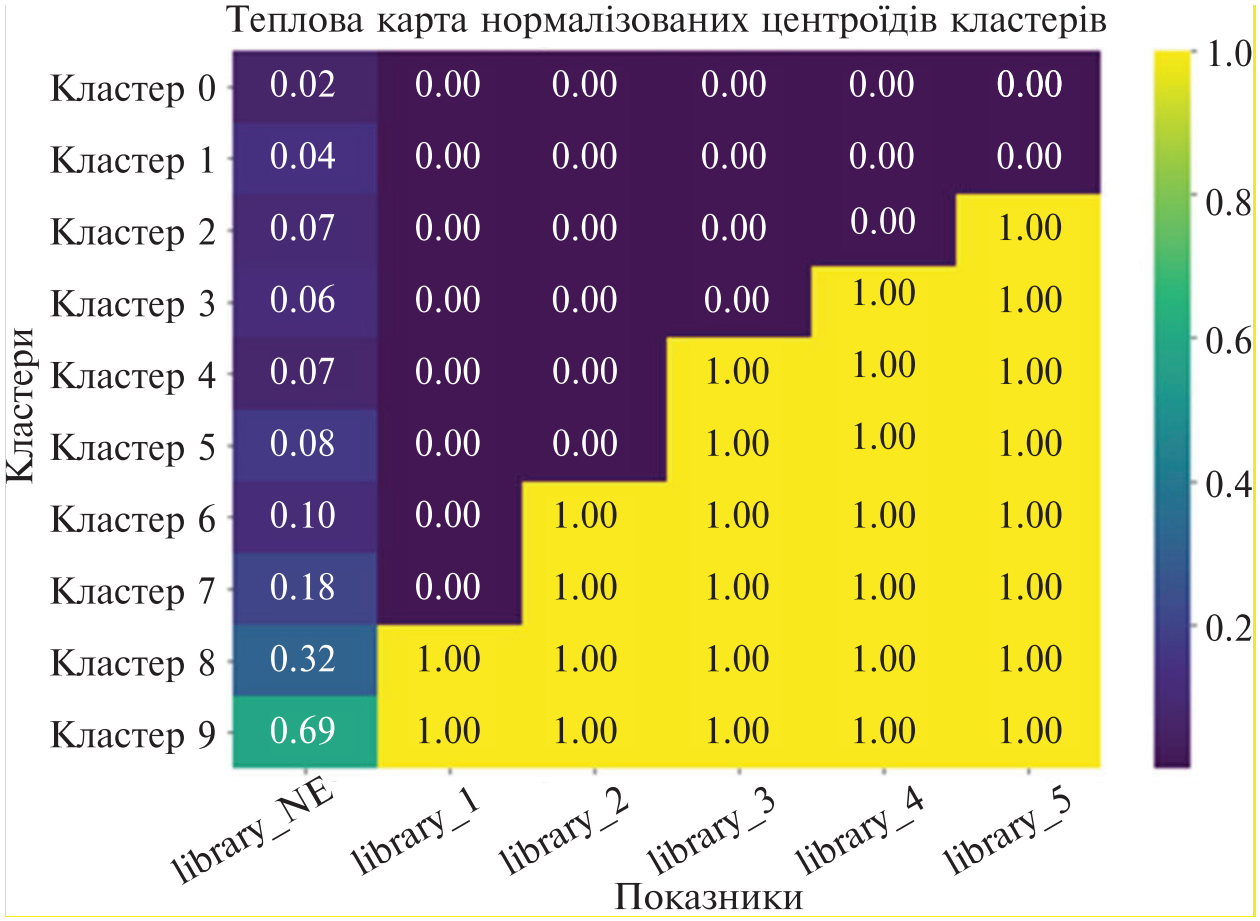


Рис. 6. Приклад теплової карти нормалізованих даних для кожного кластера, інфраструктура типу «бібліотека»

Приклад коду для кластеризації доріг (рис. 7).

```
def road_vector_series(df):
    road_columns = ['RD_m1_NEAR', 'RD_m2_NEAR',
'RD_m3_NEAR']
    if not all(col in df.columns for col in road_columns):
    raise ValueError("Required columns are missing from the
dataframe")
    roads_vec_series = df.apply(lambda row: np.array([row[col] for col
in road_columns]), axis=1)
    return roads_vec_series
    def road_clustering(roads_series, n_clusters=10):
    data = np.stack(roads_series.values)
    kmeans = KMeans(n_clusters=n_clusters,
random_state=0).fit(data)
```

```
    cluster_labels = pd.Series(kmeans.labels_,
index=roads_series.index)
    return cluster_labels
    def plot_3d_road_clusters(roads_series, cluster_labels):
    data = np.stack(roads_series.values)
    fig = plt.figure(figsize=(10, 8))
    ax = fig.add_subplot(111, projection='3d')
    colors = ['blue', 'green', 'red', 'cyan', 'magenta', 'yellow', 'black',
'pink', 'orange', 'purple']
    62
    for i in np.unique(cluster_labels):
    cluster = data[cluster_labels == i]
    ax.scatter(cluster[:, 0], cluster[:, 1], cluster[:, 2], s=50, c=colors[i],
label=f'Cluster {i}')
    ax.set_title('3D Clustering of Roads Data')
    ax.set_xlabel('Distance to Road Type 1')
    ax.set_ylabel('Distance to Road Type 2')
    ax.set_zlabel('Distance to Road Type 3')
    ax.legend()
    plt.savefig("roads_clustered.png", bbox_inches='tight')
    plt.show()
    #%%
    # Clustering
    precategorized_df["roads"] = road_vector_series(df)
    print(precategorized_df["roads"])
    categorized_df["roads"] =
categorize_distances(precategorized_df["roads"], n_clusters=10)
    # Visualization
    plot_3d_road_clusters(precategorized_df["roads"],
categorized_df["roads"])
```

З огляду на рис. 8, більшість сільських громад мають схожий доступ до регіональних та сільських доріг (типи 2 та 3, відповідно). Можна чітко визначити набір сільських громад у кластері 3, які потребують кращого зв'язку з регіональними дорогами.

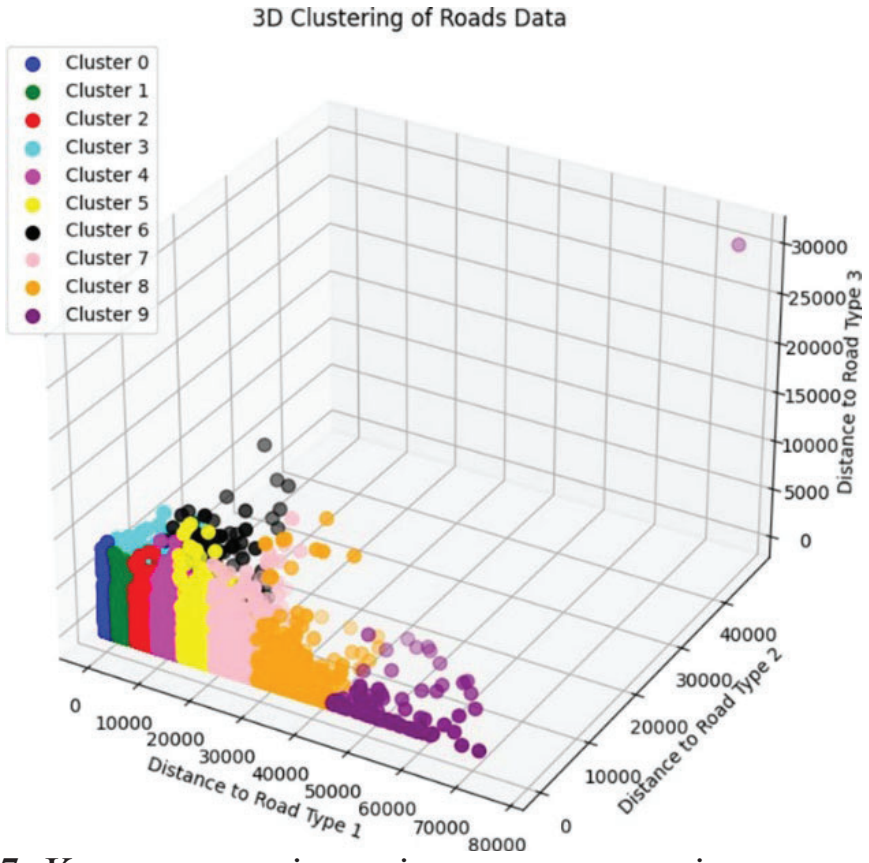


Рис. 7. Кластери типів доріг, де кожен колір представляє різний кластер

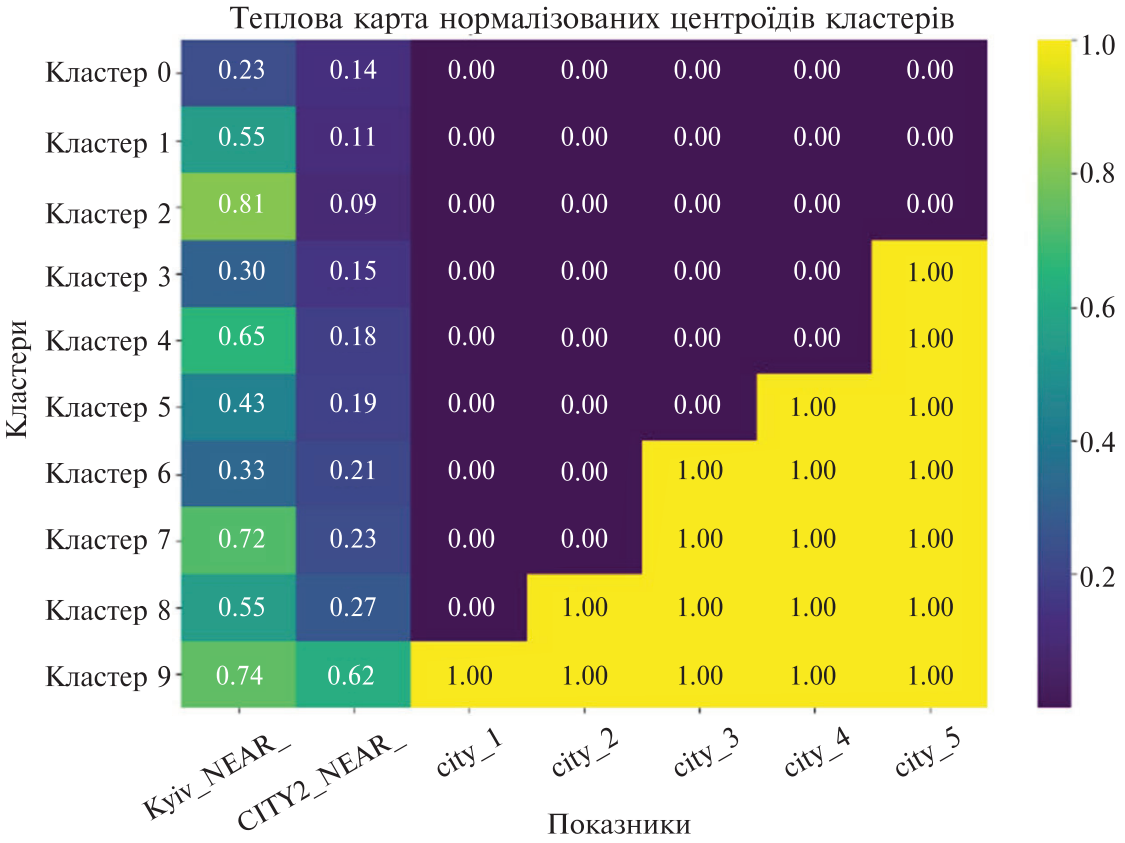


Рис. 8. Теплокарта кластерів міст

Приклад коду для створення карти кластерів міст:

```
    def create_city_distance_vector(df, city_column='graph_city'):
    def process_row(row):
        cities = json.loads(row[city_column]) if
pd.notnull(row[city_column]) else []
        city_distances = [city["distance"] for city in cities]
        city_distances = sorted(city_distances)[:5]
        city_distances.extend([2147483647] * (5 - len(city_distances)))
        return np.array([row['Kiyv_NEAR_'], row['CITY2_NEAR']] +
city_distances)
    df['city_distance_vector'] = df.apply(process_row, axis=1)
    return df['city_distance_vector']
# Clustering
precategorized_df["cities"] = create_city_distance_vector(df, graph_city")
print(precategorized_df["cities"])
categorized_df["cities"] =
categorize_distances(precategorized_df["cities"])
# Visualization
labels = ["Kiyv_NEAR_", "CITY2_NEAR", "city_1", "city_2", "city_3",
"city_4", "city_5"]
heatmap_clusters(precategorized_df["cities"], categorized_df["cities"],
custom_column_labels = labels)
```

На основі аналізу рис. 8 бачимо, що існує три кластери сільських громад, які розташовані поблизу п’яти міст, два кластери з чотирма найближчими містами, один кластер з трьома та два кластери з двома містами поблизу. Такий розподіл свідчить, що значна частина сіл може отримати переваги від близького розташування до кількох міст, що спрощує реалізацію майбутніх інфраструктурних проєктів. Наявність міст у безпосередній близькості свідчить про вищий рівень інфраструктурного розвитку, що, згідно з [6], підвищує загальний добробут мешканців. Окрім того, теплові карти показують, що кластери 0, 3 та 6 мають високу доступність до столиці (K_NEAR), що додатково підтверджує вищий рівень інфраструктурного забезпечення порівняно з іншими сільськими громадами.

Розглядаючи характеристики кластерів національних та регіональних парків на рис. 9, можна побачити, що більшість сіл не мають швидкого доступу до них. Хоча близькість розташування до таких об’єктів може не бути життєво важливою для повсякденного життя мешканців, знаходження поблизу цих зон

безсумнівно сприяє загальному розвитку навколишнього регіону. Це підвищує рівень туризму, що, в свою чергу, стимулює будівництво готелів та магазинів, підтримуючи місцеву економіку.

Приклад коду кластеризації національних парків. Теплові карти для парків наведено на рис. 9.

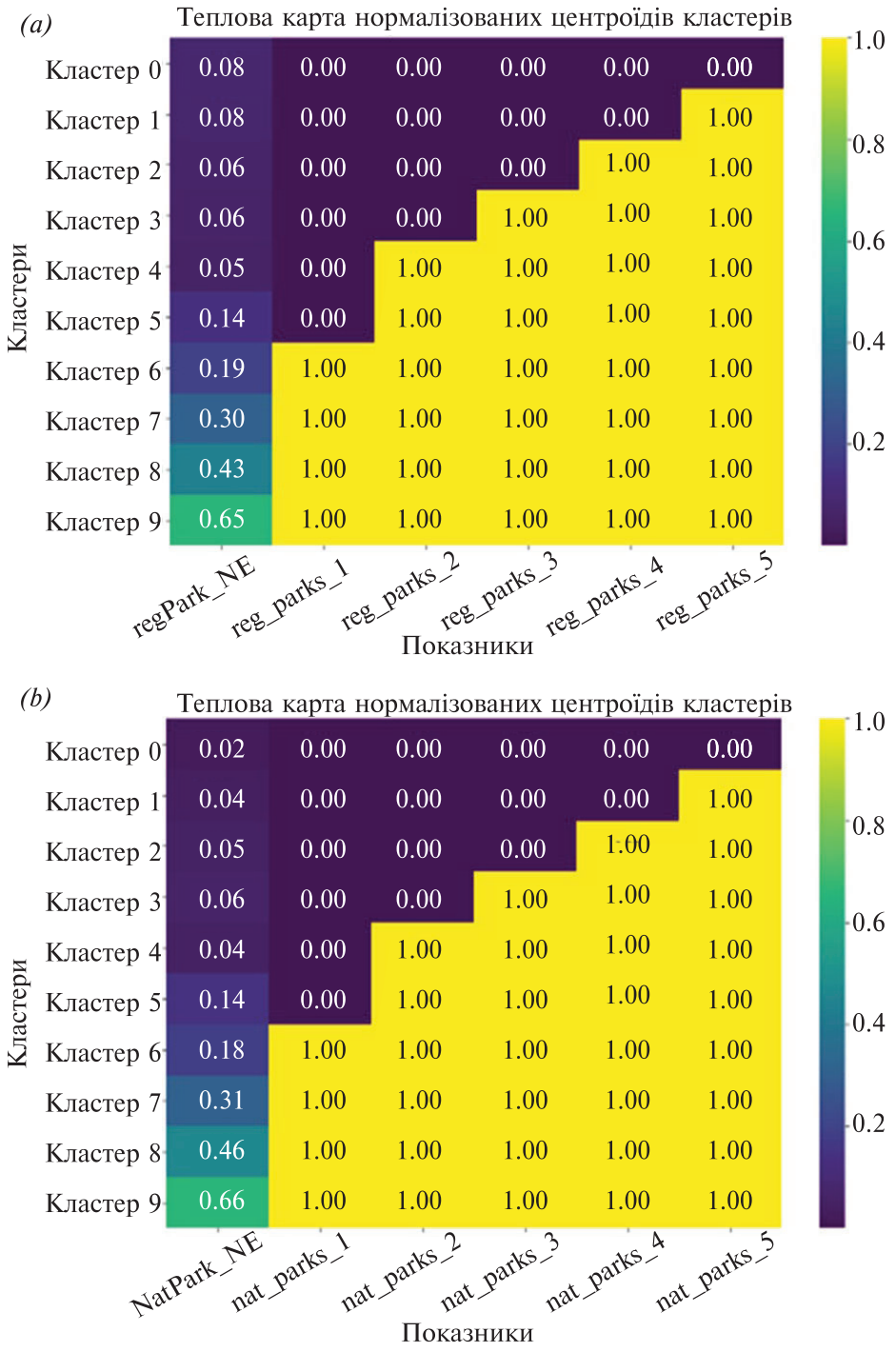


Рис. 9. Теплокарти кластерів: (а) регіональних парків, (б) національних парків

Діаграми на рис. 10-12 ілюструють різні аспекти соціального життя у селах і можуть бути розділені на дві групи: ті, що представляють розвинені та легкодоступні об'єкти (включаючи церкви, як показано на рис. 10e, освітні центри на рис. 10f, елеватори на рис. 10a, готелі на рис. 10b, дитячі садки на рис. 10c, медичні заклади на рис. 10d та магазини на рис. 12), і ті, що є більш складними для доступу, такі як банки та бібліотеки (показані на рис. 11b та рис. 11a відповідно).

```
def create_national_park_distance_vector(df,
park_column='graph_national_park'):
    def process_row(row):
        parks = json.loads(row[park_column]) if
    pd.notnull(row[park_column]) else []
        park_distances = [park["distance"] for park in parks]
        park_distances = sorted(park_distances)[:5]
        park_distances.extend([2147483647] * (5 -
    len(park_distances)))
        return np.array([row['NatPark_NE']] + park_distances)
    df['national_park_distance_vector'] = df.apply(process_row, axis=1)
    return df['national_park_distance_vector']
#%%
precategorized_df["nat_parks"] =
create_national_park_distance_vector(df)
print(precategorized_df["nat_parks"])
labels = ["NatPark_NE", "nat_parks_1", "nat_parks_2", "nat_parks_3",
"nat_parks_4", "nat_parks_5"]
categorized_df["nat_parks"] =
categorize_distances(precategorized_df["nat_parks"])
heatmap_clusters(precategorized_df["nat_parks"],
categorized_df["nat_parks"], custom_column_labels = labels
```

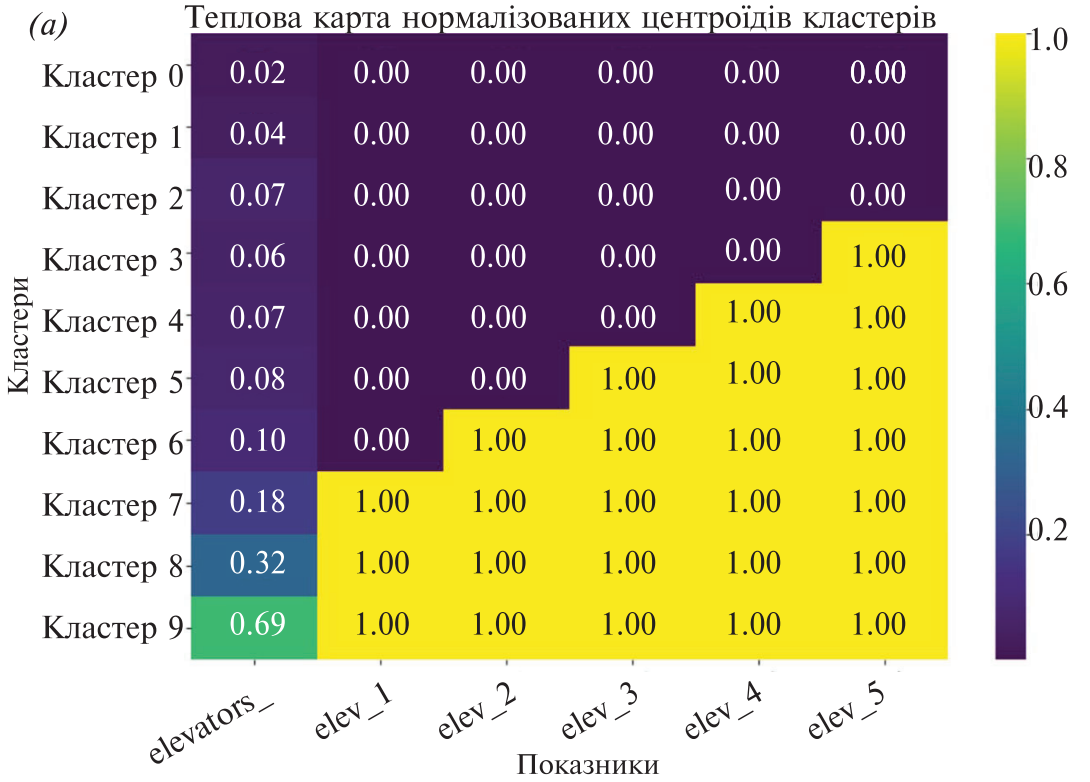
Приклад коду кластеризації регіональних парків:

```
def create_regional_park_distance_vector(df,
park_column='graph_regional_park'):
    def process_row(row):
        parks = json.loads(row[park_column]) if
    pd.notnull(row[park_column]) else []
        park_distances = [park["distance"] for park in parks]
```

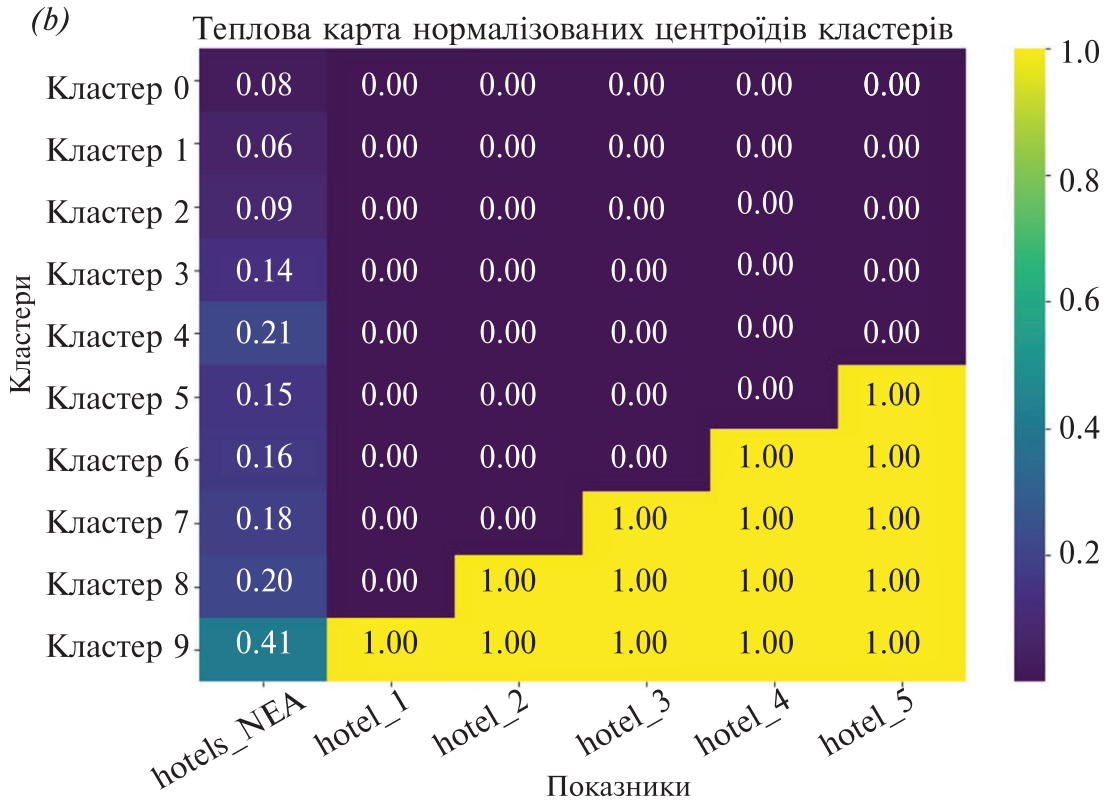
```
            park_distances = sorted(park_distances)[:5]
            park_distances.extend([2147483647] * (5 -
        len(park_distances)))
            return np.array([row['regPark_NE']] + park_distances)
        df['regional_park_distance_vector'] = df.apply(process_row, axis=1)
        return df['regional_park_distance_vector']
        #%%
precategorized_df["reg_parks"] =
create_regional_park_distance_vector(df)
print(precategorized_df["reg_parks"])
labels = ["regPark_NE", "reg_parks_1", "reg_parks_2", "reg_parks_3",
"reg_parks_4", "reg_parks_5"]
categorized_df["reg_parks"] =
categorize_distances(precategorized_df["reg_parks"], 10)
heatmap_clusters(precategorized_df["reg_parks"],
categorized_df["reg_parks"], custom_column_labels = labels)
```

Сільські громади першої групи мають швидкий доступ до важливих об'єктів інфраструктури, таких як освіта, медичні послуги та дитячі садки, що є вирішальним для загального добробуту мешканців. Готелі та магазини, хоч і не є критично важливими, в цілому також сприяють розвитку малого бізнесу та забезпечують доступ до товарів, що підтримує економічну активність регіону. Окрім того, наявність готелів підвищує можливості для прийому туристів, що є важливим чинником для сталого розвитку сільських територій.

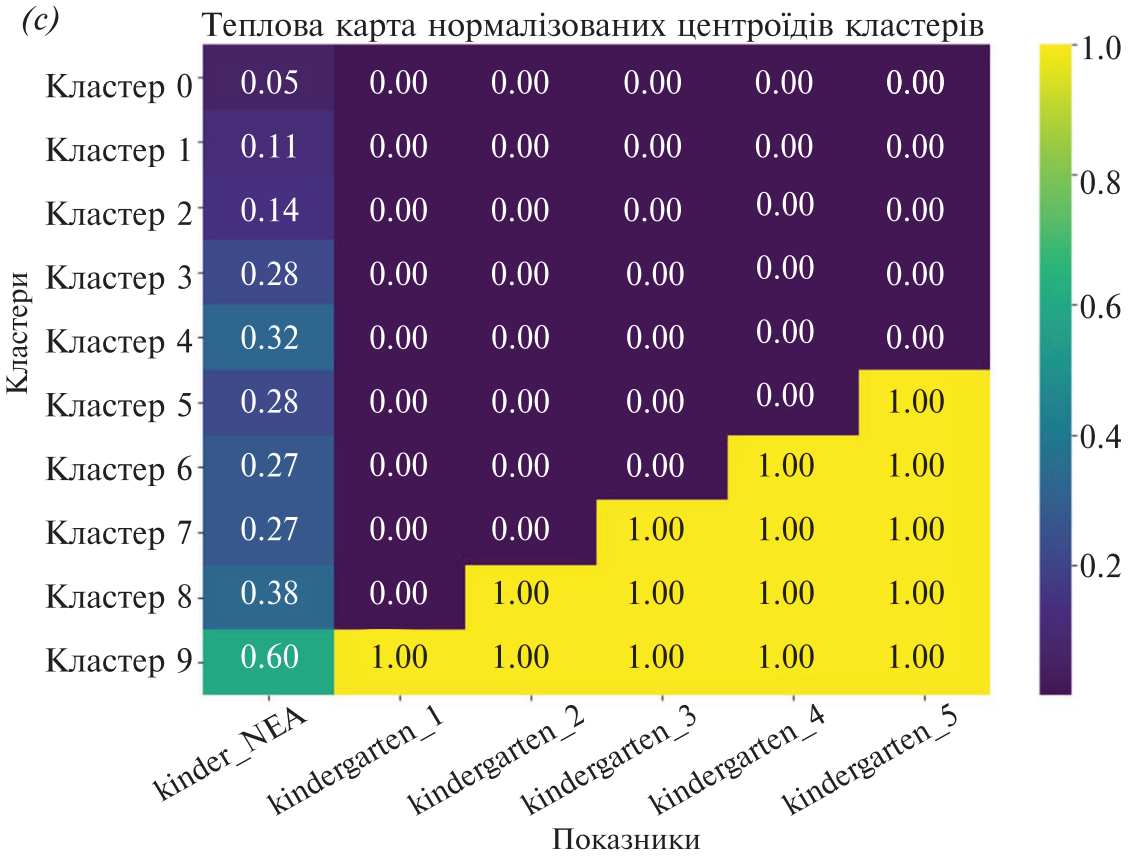
Друга група підкреслює, що хоча роль бібліотек змінюється через поширення Інтернету та мобільних технологій, вони продовжують виконувати важливу соціальну та культурну функцію. Це підкреслює необхідність покращення громадських просторів, таких як місцеві парки, для підтримки соціальної згуртованості. Крім того, розширення банківських послуг, таких як мережа банкоматів та фінансових установ, сприятиме економічній активності, полегшуючи доступ до кредитних та фінансових продуктів для місцевого бізнесу.



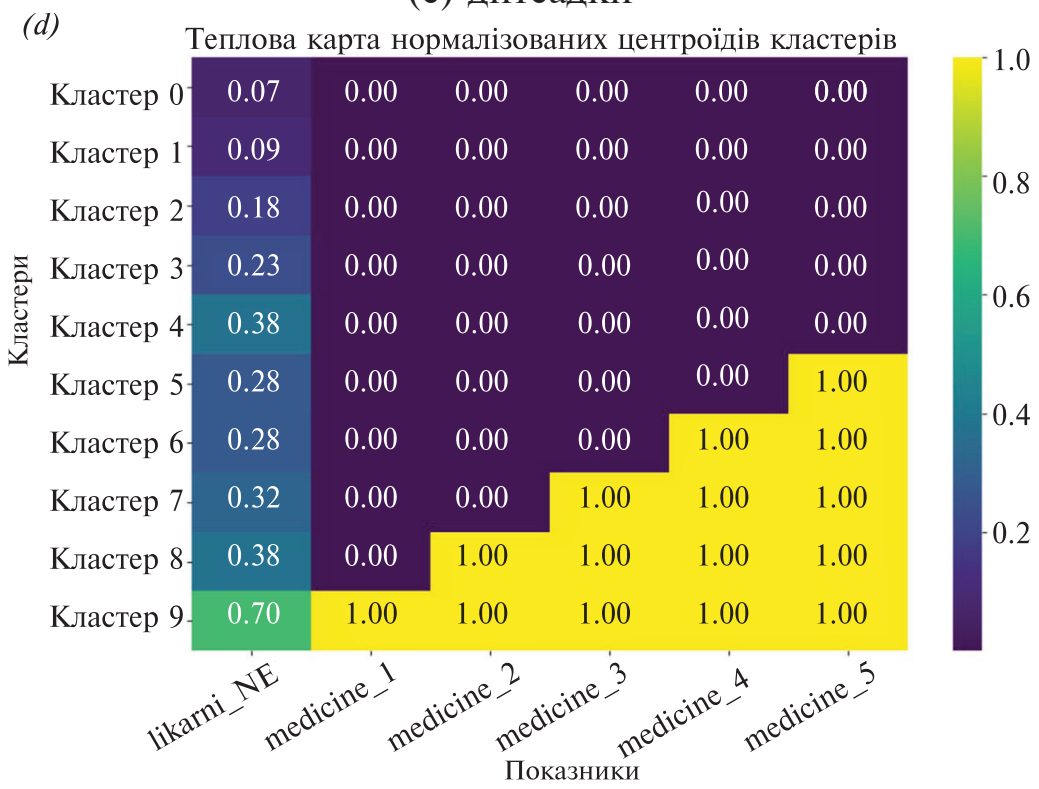
(a) елеватори



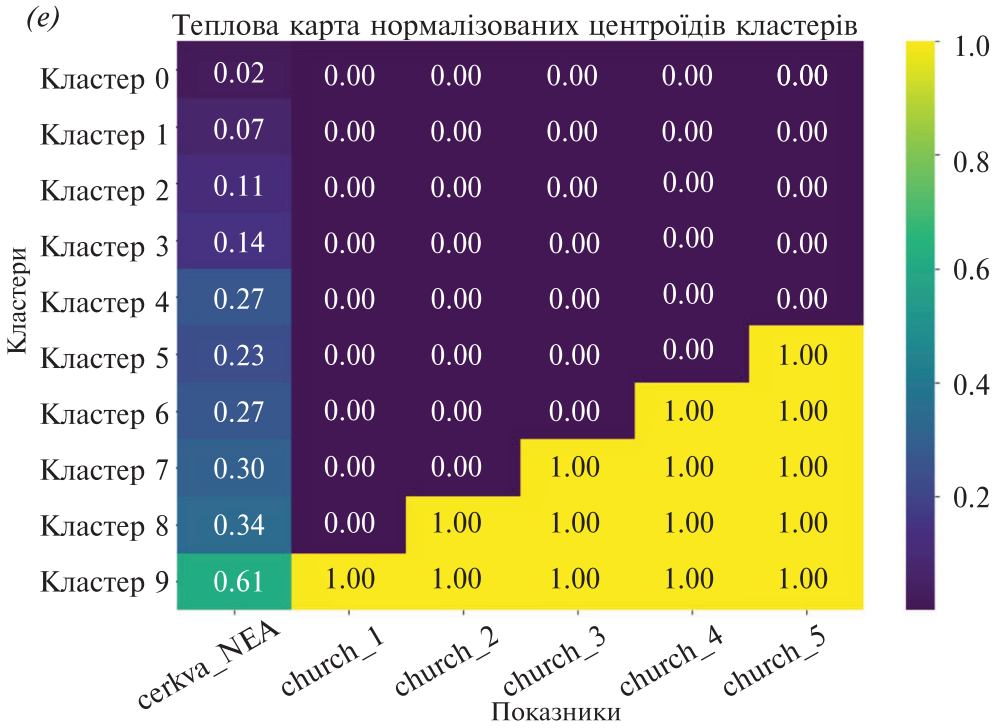
(b) готелі



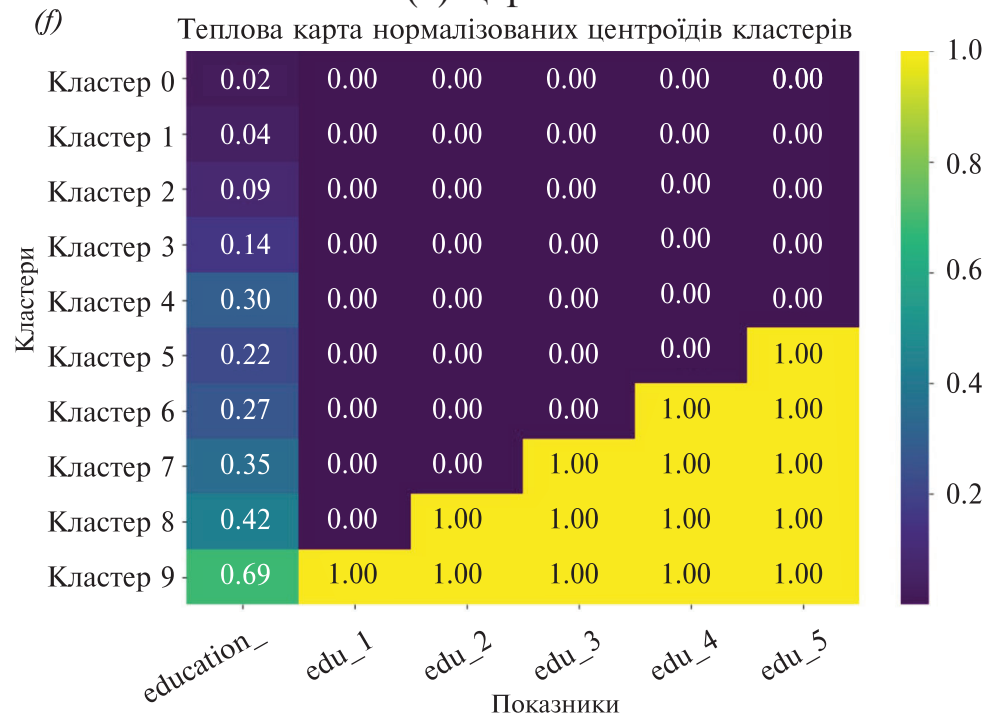
(c) дитсадки



(d) мед. заклади



(e) церкви



(f) об’єкти освіти.

Рис. 10. Теплокарти кластерів різних об’єктів

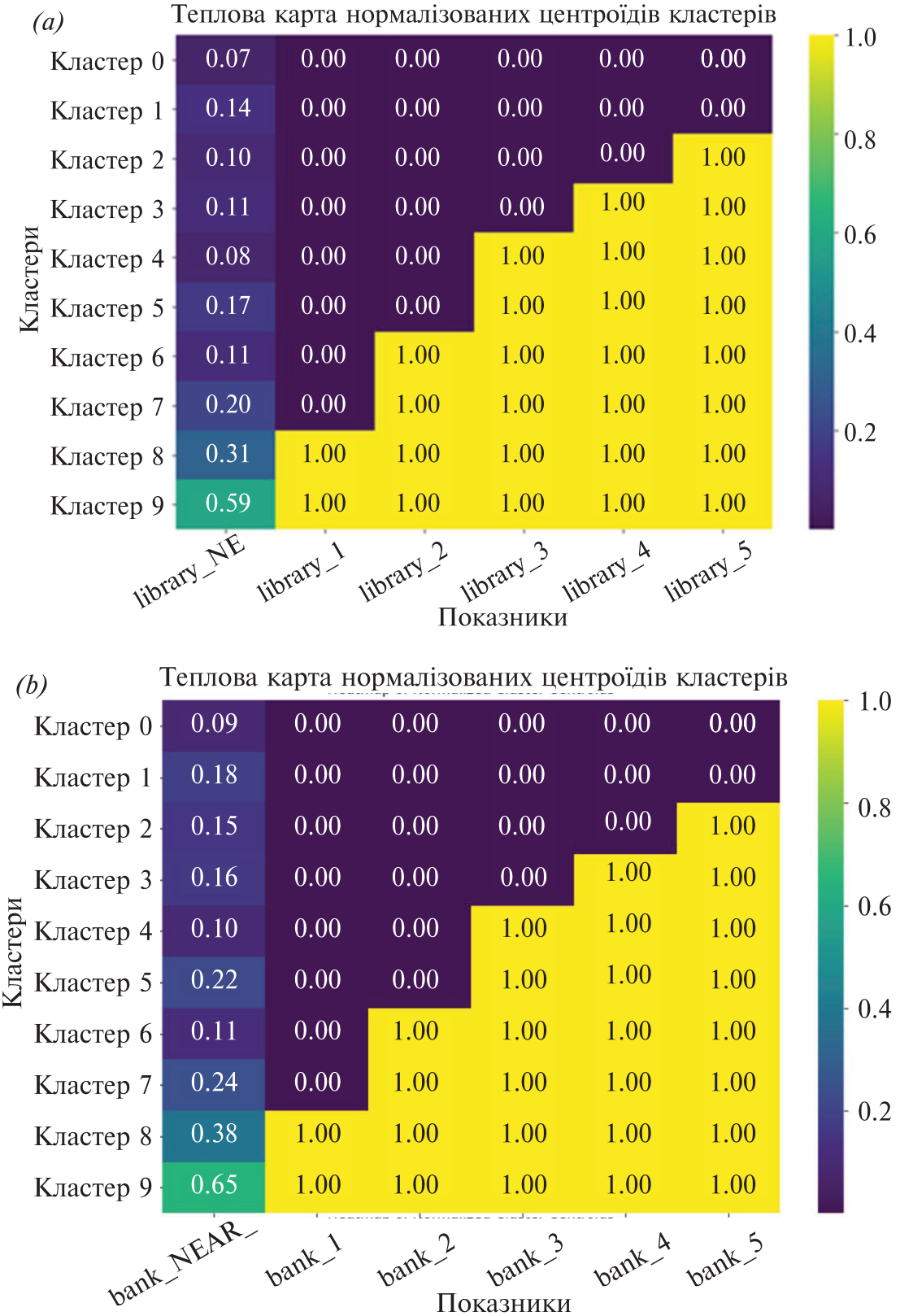


Рис. 11. Теплокарти кластерів: (a) бібліотек, (b) банків

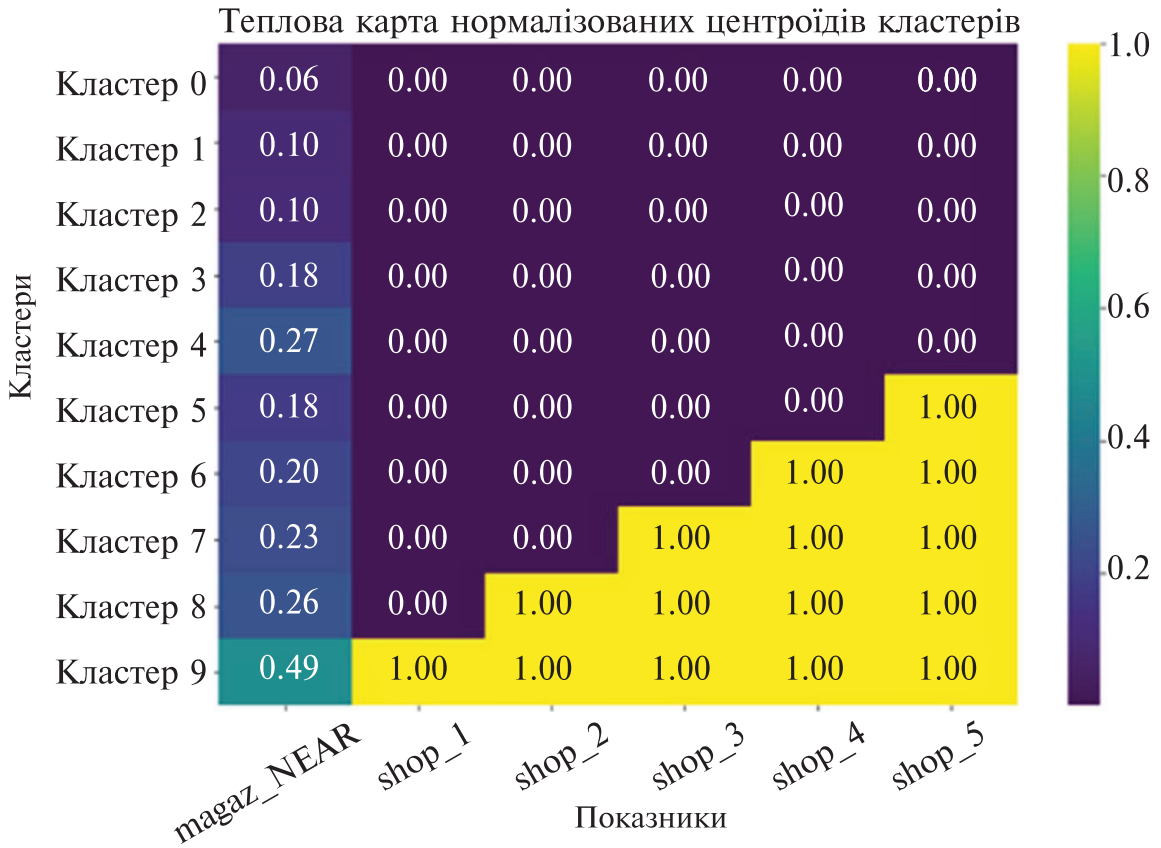


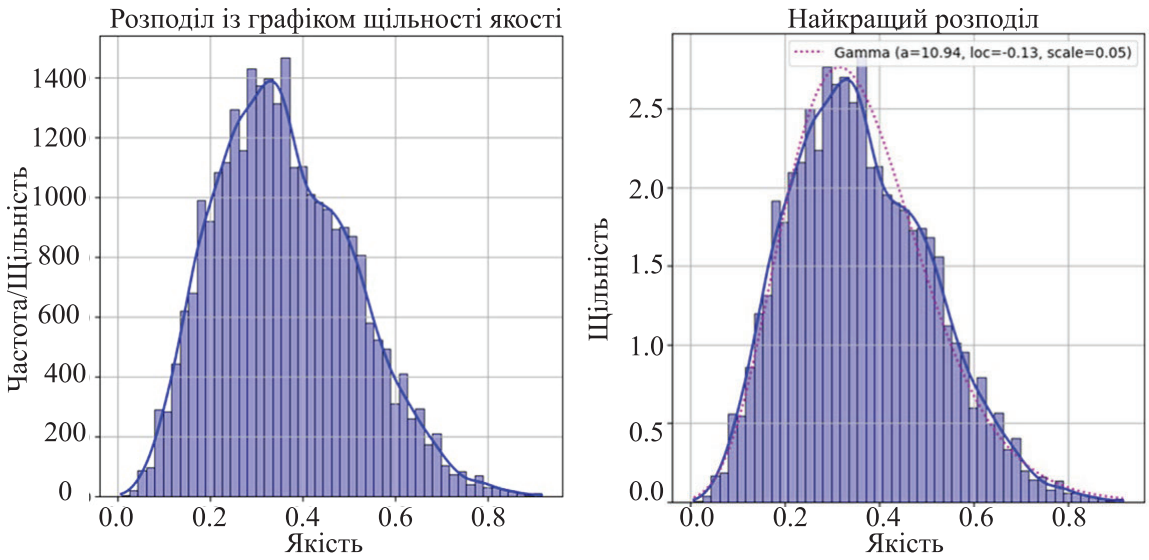
Рис. 12. Теплокарта кластерів магазинів

Важливим аспектом залишається забезпечення зручного доступу до банківських послуг. Незважаючи на розвиток електронного банкінгу в Україні, фізична присутність банків залишається важливою, особливо для літніх людей та туристів, які можуть потребувати особистої взаємодії або допомоги. Тому покращення фізичної банківської інфраструктури може зробити регіони більш привабливими та комфортними для вирішення фінансових питань.

Загальний аналіз кластерів вказує на певні розбіжності, наприклад, в описі бібліотек у кластері 2 та доріг у кластері 1. Це підтверджує, що застосований підхід забезпечив ефективний та детальний аналіз даних.

Аналіз розподілу та обчислення оцінки якості. Перейдемо до більш детального аналізу розподілу якості інфраструктури в селах, представлений на рис. 13а.

Спробуємо визначити, чи задовольняє цей розподіл хоч якусь з статистичних моделей, як оговорювалось раніше. Для цього підберемо найкращу статистику та візуалізуємо її, так на рис. 13б.



(а) Загальна якість всіх сіл (б) Найкраща статистична модель

Рис.13. Гістограми якості інфраструктури в селах та найкраща статистична модель

Не зважаючи на досить низьке значення у тесті Колмогорова-Смірнова, можна побачити, що розподіл досить добре повторює гамма-розподіл з наведеними на рисунку параметрами, а отже, мету досягнуто за допомогою формули (21), де $a = 1.0555$ та $b = 0.0555$.

Таким чином отримано важливий результат експерименту, а саме сформулювано оцінку якості для кожного села, що розрахована наступним чином:

$$\text{ОЯ} = 1.0555 \times \left(\frac{1}{n}\sum_{i=1}^{n} \text{ОІ}_i\right) + 0.0555 \times \text{РК}\,. \qquad (30)$$

Ця формула інтегрує середню оцінку категорії інфраструктури з рейтингом кластера, надаючи детальну міру якості інфраструктури. Ця комплексна оцінка охоплює як різноманітність наявної інфраструктури в сільській громаді, так і її порівняльний стан серед інших сіл.

Для ілюстрації роботи запропонованих раніше методів було розроблено пілотний дешборд, який можна побачити на рис. 14 або ж за відкритим посиланням [15].

На рис. 14 представлено загальний вигляд даного дашборду - мапи України, де відмічені всі сільські громади за допомогою кольорової шкали, що відображає якість інфраструктури, починаючи від зеленого, що означає кращу якість розвитку

інфраструктури, до гіршої якості, що відображається червоним кольором.

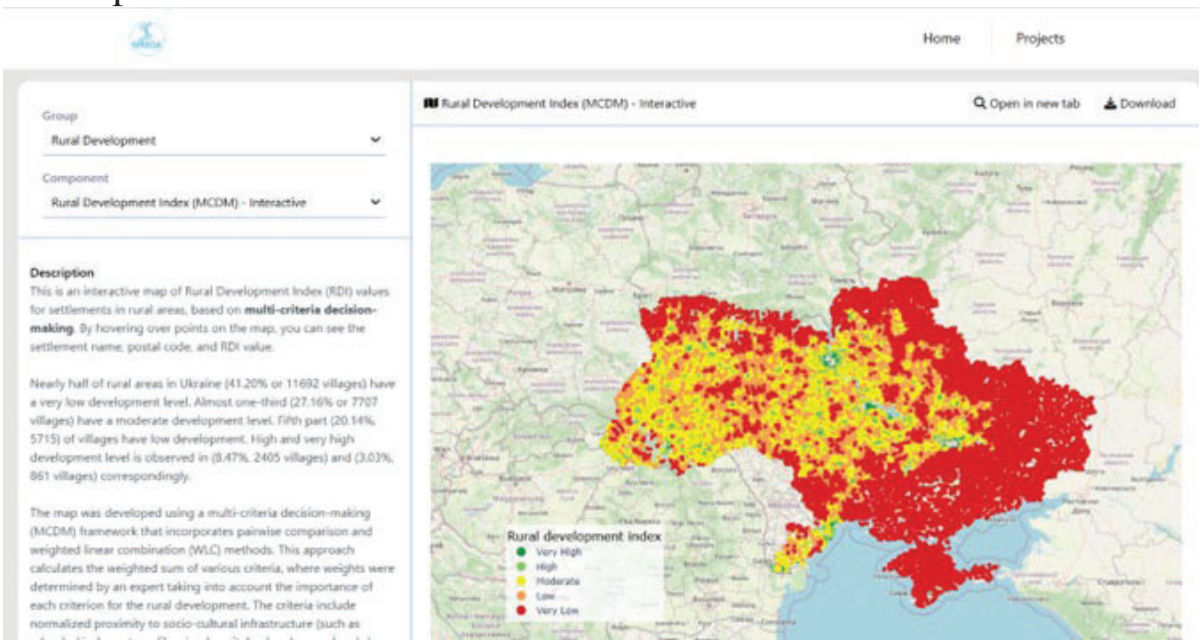


Рис. 14. Дашборд з візуалізацією отриманих результатів дослідження

Однією з важливих функцій цього дашборду є можливість отримання інформації по кожній сільській громаді (рис. 15), наприклад *id*, значення кожного типу інфраструктури, а також загальний рейтинг якості інфраструктури і фінальну оцінку якості.

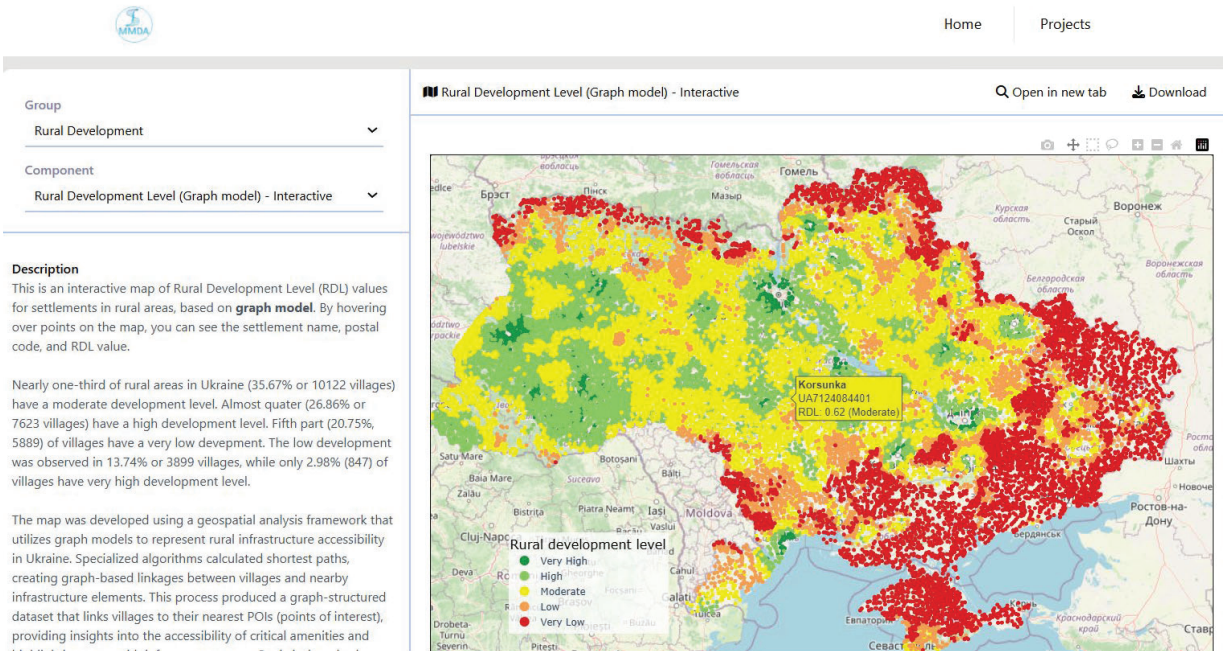


Рис. 15. Детальна інформація про кожне село

Даний дашборд дозволяє як візуалізувати отримані результати, так і детально ознайомитися з кожним селом окремо, що може бути корисним для подальшого аналізу та планування розвитку інфраструктури, при чому важливим є те, що ознайомлення з даним дашбордом має низький рівень входу, що дозволяє використовувати його широкому колу користувачів та стейкхолдерів.

2.5.2. ОПТИМІЗАЦІЯ ОЦІНОК ДОСТУПНОСТІ НА ОСНОВІ ГРАФОВИХ ДАНИХ

Основною метою експерименту є розробка розширеного набору геопросторових оцінок для представлення доступності критичної інфраструктури для сільських громад на основі графової моделі. Для досягнення цієї мети проведемо багатофазний аналіз, що включає наступні кроки.

1) Побудова статистичного профіля базових геопросторових наборів даних про сільські території України для перевірки цілісності та узгодженості з пріоритетами розвитку.

2) Експериментальна реалізація спеціалізованих алгоритмів для побудови графових зв'язків між сільськими поселеннями та навколишніми багатокатегорійними точками інтересу POI, що представляють доступну інфраструктуру.

3) Кількісний та графічний статистичний аналіз для оцінки, чи зберігає графова трансформація якісні властивості попередніх розподілів доступності сільських територій.

Вихідним результатом є вдосконалений графовий набір даних, що кількісно оцінює доступність на рівні сіл до ключових сервісів у межах критичних порогів відстані. Це закладає фундамент для майбутньої аналітики для визначення прогалин у сільській інфраструктурі, що обмежують розвиток.

Перевірка базових геопросторових даних. Спочатку перевіримо повноту та адекватність зібраних геопросторових шарів даних про сільську інфраструктуру в Україні. Для цього було обрано використання GeoDataFrames (GDF) (рис. 16).

GeoDataFrames є основною структурою даних для роботи з геопросторовими даними в бібліотеці geopandas, яка розширює можливості pandas.

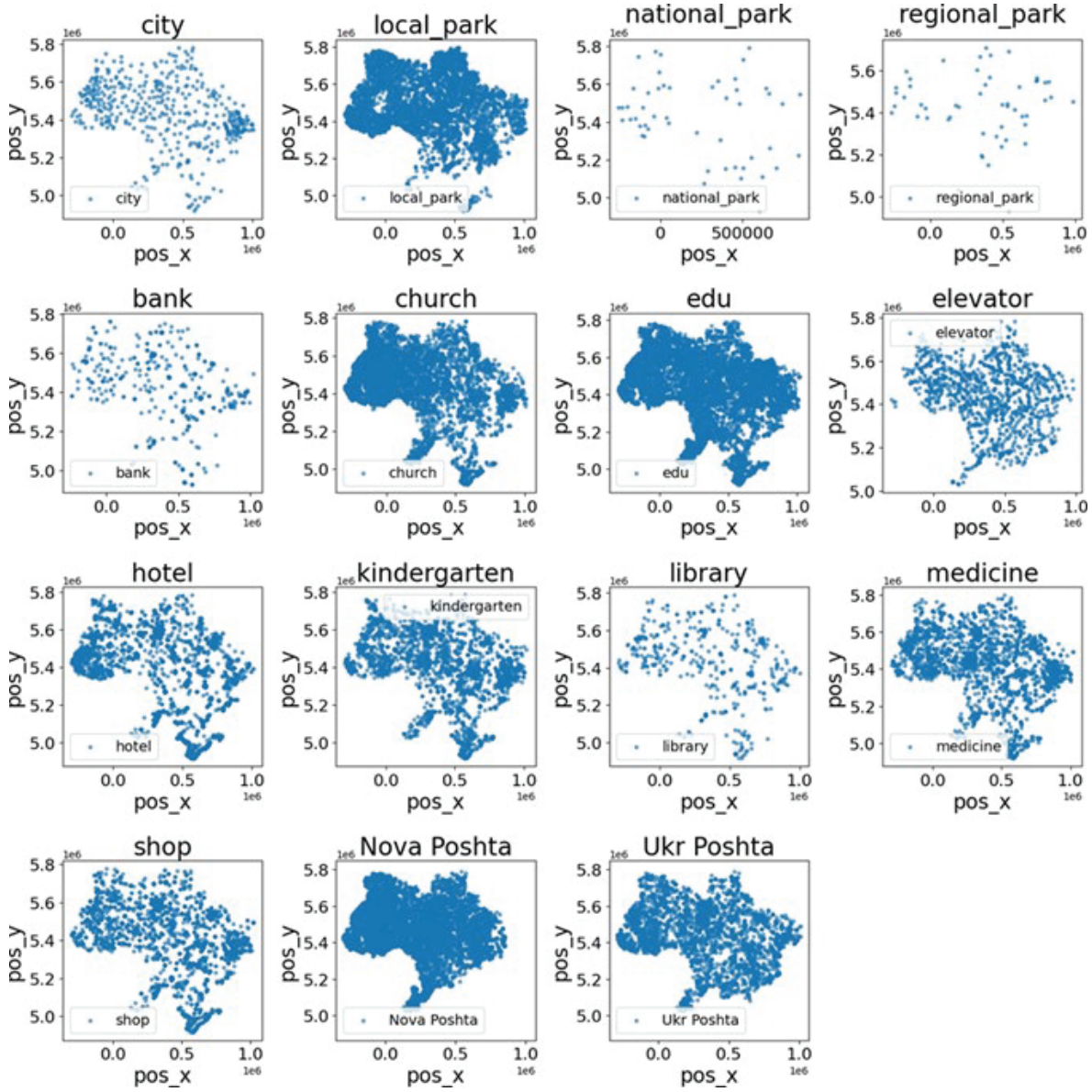


Рис. 16. Графічне представлення груп точок інтересу POI.
Кожний рисунок представляє один тип POI як в табл. 3.

Основні характеристики GeoDataFrame:

Колонка геометрії: У GeoDataFrame є спеціальна колонка geometry, яка містить геометричні об'єкти (точки, лінії або полігони), що представляють просторові дані. Це дозволяє прив’язати кожен рядок до певної географічної форми.

Аналіз і візуалізація: GeoDataFrames підтримують просторові операції, такі як об'єднання, перетини та буферизація, а також дозволяють візуалізувати дані на карті, що спрощує роботу з геопросторовою інформацією.

Сумісність із GIS: GeoDataFrames можуть зберігатися у форматах, сумісних із геоінформаційними системами (наприклад,

Shapefile та GeoJSON), що дозволяє обмінюватися даними з іншими ГІС-додатками.

Після аналізу графічного представлення на рис. 16 стає очевидним, що більшість цих графіків демонструють більш-менш нормальний розподіл по Україні.

«Нормальний» у цьому контексті означає, що немає значних прогалин, смуг або будь-яких інших відсутніх областей у даних, окрім даних поштових операторів, в зв'язку з закриттям своїх відділень через війну на сході України.

Оцінка доступності сільських територій на основі графових даних. Маючи чисті базові дані, реалізуємо запропоновану методологію сегментації зони обслуговування сільських територій, вилучення та відбору POI, розрахунку відстаней та побудови графової бази даних для перетворення попередніх наборів даних у розширені оцінки доступності. Як було описано в підрозділі 2, процес створення графової бази даних включає декілька ключових етапів. Спочатку визначаються всі можливі точки інтересу (POI) в заданому радіусі навколо кожного села. Далі використовуються алгоритми просторового аналізу для розрахунку евклідових відстаней між селами та кожною POI, забезпечуючи тим самим точні географічні зв'язки. Ці відстані конвертуються в атрибути графових вершин і ребер, що дозволяє створити деталізовану мережу доступності. Ця мережа відображає реальні шляхи доступу до різних об'єктів інфраструктури і забезпечує глибоке розуміння просторових структур та можливих прогалин у наявності критичних ресурсів. Експеримент передбачає використання таблиці, яка описує, які дані були включені у новостворені оцінки. Для кожного фрейму визначено конкретні значення, що визначають, які параметри (стовпці) повинні бути включено у вихідні дані.

Таблиця 4 Опис даних, включених в новостворені оцінки POI

Дані	Опис у вихідному графі	Максимальна відстань	Максимальна кількість в кластері
Міста	Id, pos_x, pos_y, distance	50 км	5
Місцеві парки	Id, type,distance, area, pos_x,	30 км	5

	pos_y		
Національні парки	Id, area, pos_x, pos_y, distance	50 км	5
Регіональні парки	Id, area, pos_x, pos_y, distance	50 км	5
Банк	Id, pos_x, pos_y, distance	30 км	5
Церква	Id, type, pos_x, pos_y, distance	30 км	5
Освіта	Id, type, pos_x, pos_y, distance	30 км	5
Елеватори	Id, pos_x, pos_y, distance	30 км	5
Готелі	Id, type, pos_x, pos_y, distance	30 км	5
Дитячі садки	Id, pos_x, pos_y, distance	30 км	5
Бібліотеки	Id, pos_x, pos_y, distance	30 км	5
Медицина	Id, type, pos_x, pos_y, distance	30 км	5
Магазини	Id, type, pos_x, pos_y, distance	30 км	5

Кількісна оцінка відстані в початкових даних. Даний тип оцінки сільських громад [2] включає різні атрибути інфраструктури, детальний опис яких наведено у табл. 3.

Збір та обробка цих даних дозволили створити детальну картину доступності сільських територій до критичної інфраструктури, такої як дороги, міста, елеватори, освітні та медичні заклади, а також інші важливі об'єкти. За допомогою найпростішого типу опису, що дозволяє закласти певний найпростіший опис доступності інфраструктури, розкриваються основні характеристики та значення кореляції, які можна використати в наступних підрозділах при аналізі новостворених оцінок доступності.

Графові оцінки доступності. В результаті виконання експерименту, описаного вище, було отримано графові оцінки доступності для сільських територій. В результаті трансформації

початкових оцінок відстаней у детальне графове з'єднання між сільськими селами та навколишніми багатокатегорійними POI, успішно сконструйовано розширені оцінки доступності. В результаті розроблено наступну структуру, де id_type це назва ключа села, id – це його ідентифікатор, distance - це відстань в метрах до села та pos_x, pos_y - це координати села в EPSG:32636.

{"id_type": "admin4Pcod",
"id": "UA2111000000",
"distance": 3111.931012554032,
"pos_x": 52060.12938924221,
"pos_y": 5420926.907282681 },
{"id_type": "admin4Pcod",
"id": "UA2110100000",
"distance": 23160.7396750182,
"pos_x": 60818.384628206666,
"pos_y": 5441084.188000026 },
{"id_type": "admin4Pcod",
"id": "UA2123210100",
"distance": 41187.70115749954,
"pos_x": 76317.64176459657,
"pos_y": 5451978.68918336},
{"id_type": "admin4Pcod",
"id": "UA2110400000",
"distance": 41376.3746009097,
"pos_x": 90058.87699085698,
"pos_y": 5416411.645983889},
{"id_type": "admin4Pcod",
"id": "UA2110200000",
"distance": 43577.79427222703,
"pos_x": 80594.92931475205,
"pos_y": 5391219.696301765}

Статистичний аналіз. Досліджуючи вторинні та сільські дороги (RD_m2_NEAR та RD_m3_NEAR), видно, що майже всі села мають задовільний доступ до доріг будь-якого типу, як ілюструє рис. 17.

Однак, є значний простір для покращення в основних дорогах RD_m1_NEAR, який демонструє розподіл, зміщений вправо. Kyiv_NEAR показує трикутний розподіл, що вказує на те, що більшість сіл розташовані на відстані 150–500 км, що сприятливо, оскільки це свідчить про рівномірний доступ до столиці та,

відповідно, до можливостей бізнесу з компаніями, які там базуються.

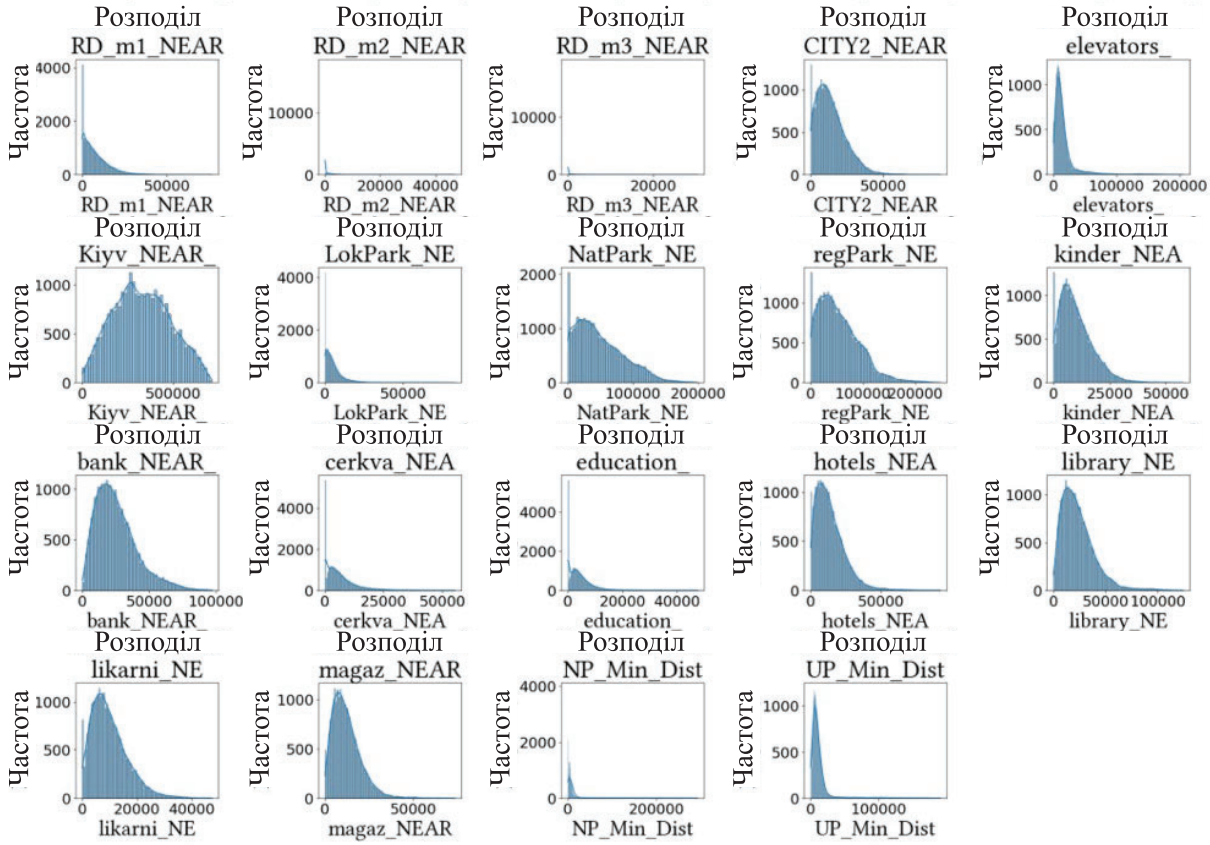


Рис. 17. Гістограма відстані для кожного типу інфраструктури

Розподіли інших змінних ілюструють певну схожість з логнормальним розподілом, що вказує на те, що хоча деякі сільські громади розташовані близько до POI, багато інших мають простір для покращення в плані розташування та доступу до різних елементів інфраструктури. *Слід зазначити, що* «схожість» в даному контексті була отримана, як результат проведення тесту по підбору параметрів для нормального, логнормального гамма- та експоненціального розподілів. Після підбору параметрів і проведення тесту Колмогорова-Смірнова, було визначено, що навіть для найкращого підходу значення p (рівень значущості, який вказує на ймовірність того, що отримані результати могли виникнути випадково, якби нульова гіпотеза була вірною) досить низьке (0.05 – 0.1), що не дозволяє стверджувати, що дані розподіли повторюють обрані розподіли точно, проте можна говорити про їх схожість.

Додатковий аналіз (рис. 18) підтверджує відносну коректність використаних відстаней при побудові графових оцінок, а саме в табл. 4 для параметру «максимальної відстані».

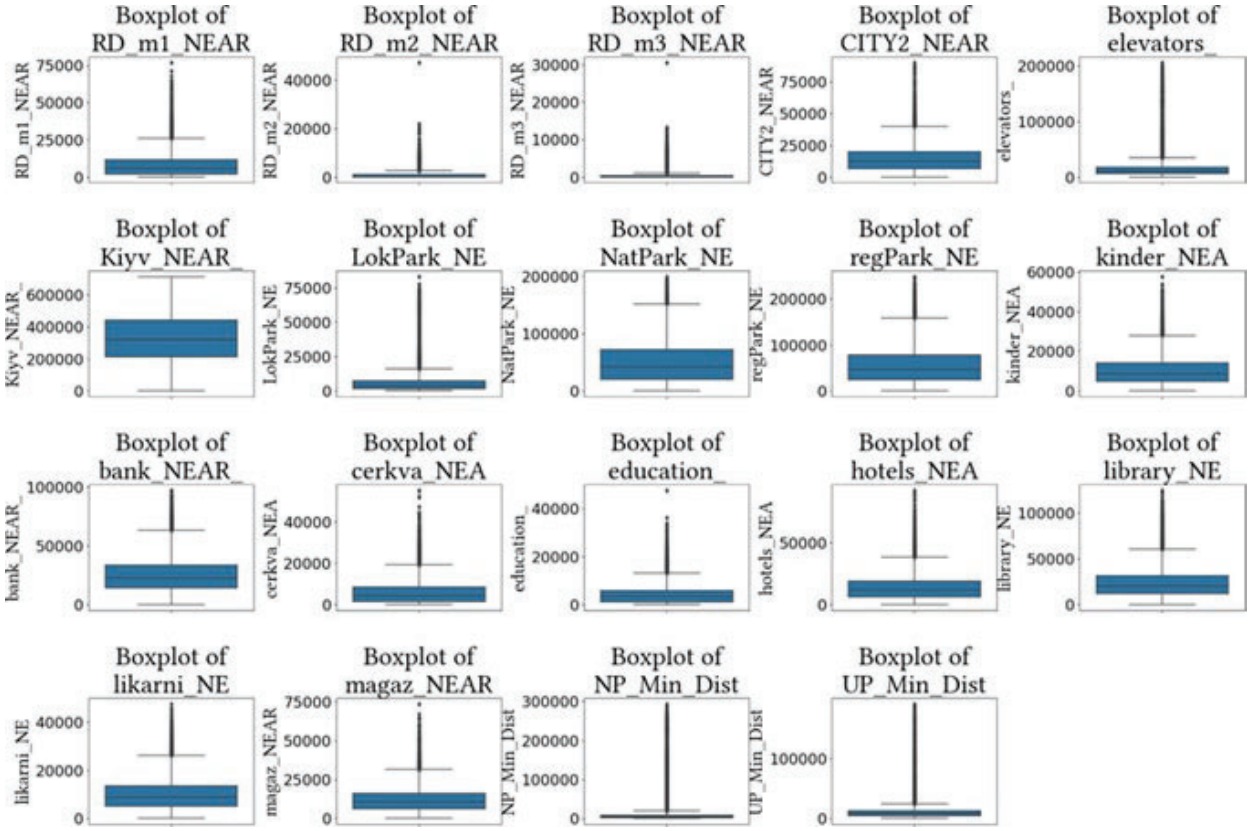


Рис. 18. Коробкові діаграми досліджуваних колонок

Розглянемо кореляції між оцінками відстаней, що зображені на рис. 19, більш детально. Важливим спостереженням є кореляція розташування таких об'єктів, як дитячі садки, банки, церкви та місця надавання освітніх послуг, з близькістю до найближчого міста, оскільки міста зазвичай пропонують більше можливостей та кращі сервіси, ніж села.

Додатковий статистичний аналіз показав, що інші категорії інфраструктури, такі як лікарні та магазини, також демонструють високий рівень кореляції з відстаннями до міст. Це свідчить про те, що сільські райони, які мають кращий доступ до міської інфраструктури, мають значні переваги в доступі до ключових послуг.

Крім того, результати показують, що сервіси, які вважаються менш критичними, такі як місцеві парки та готелі, також демонструють значну кореляцію з ступенем близькості до міста. Це підкреслює важливість комплексного підходу до розвитку сільської інфраструктури, включаючи як критичні, так і

некритичні послуги, для забезпечення високої якості життя в сільських громадах.

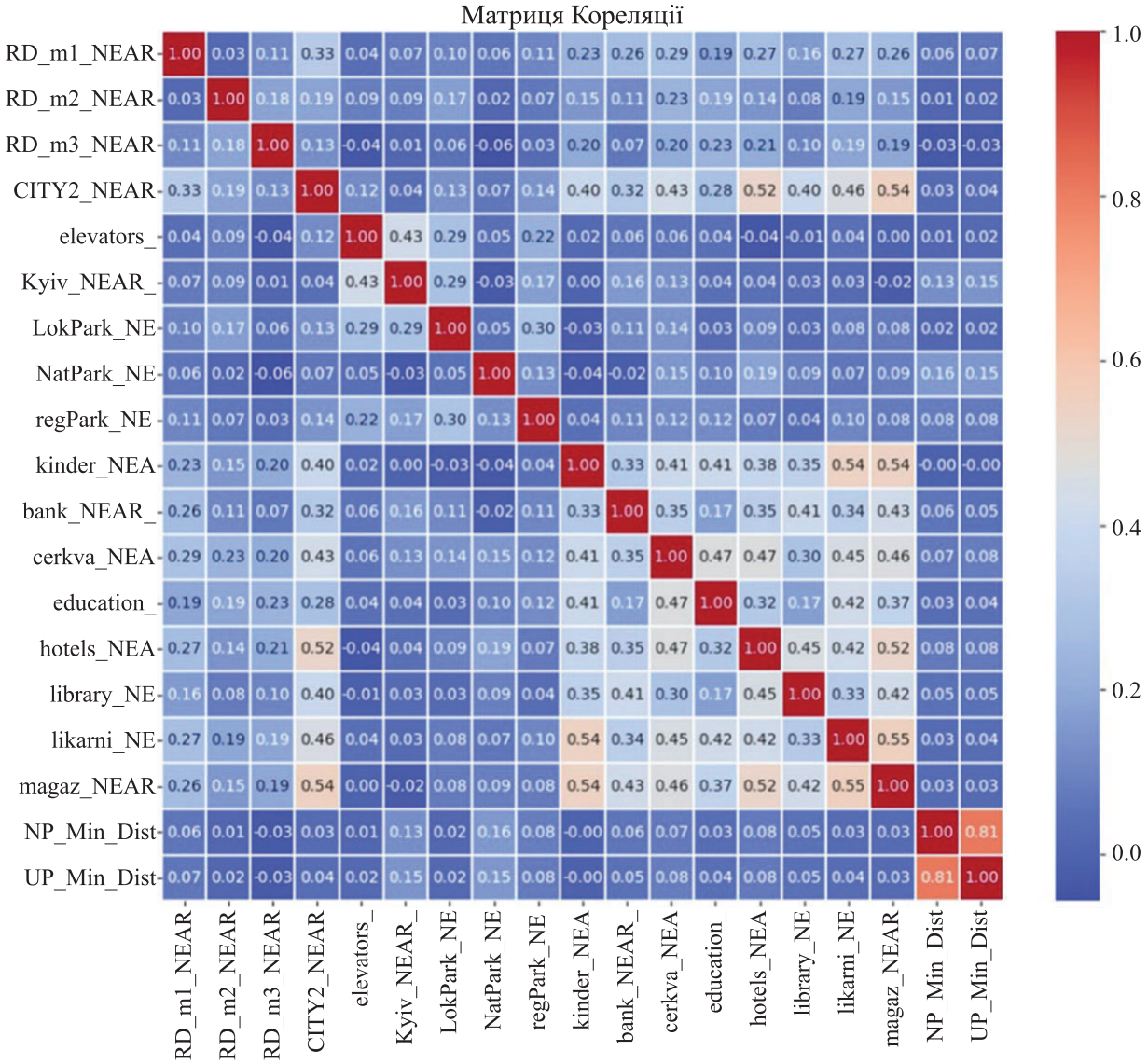


Рис. 19. Кореляційна матриця досліджуваних колонок

Важливо, щоб розширені набори даних не втрачали цілісність або не демонстрували широко варіативні розподіли. Рис. 19 підтверджує актуальність загальної доступності соціальних зручностей. Далі потрібно перевірити, чи залишаються розподіли даних стабільними, або майже такими, створюючи графіки розподілу, як показано на рис. 20. Хоча ці діаграми містять багато даних та містять розбіжності, все ж можна спостерігати, що загальний розподіл не змінився суттєво в розширеному наборі даних.

Це свідчить про те, що запропонований підхід забезпечує стабільність і надійність отриманих результатів, зберігаючи ключові характеристики початкових даних.

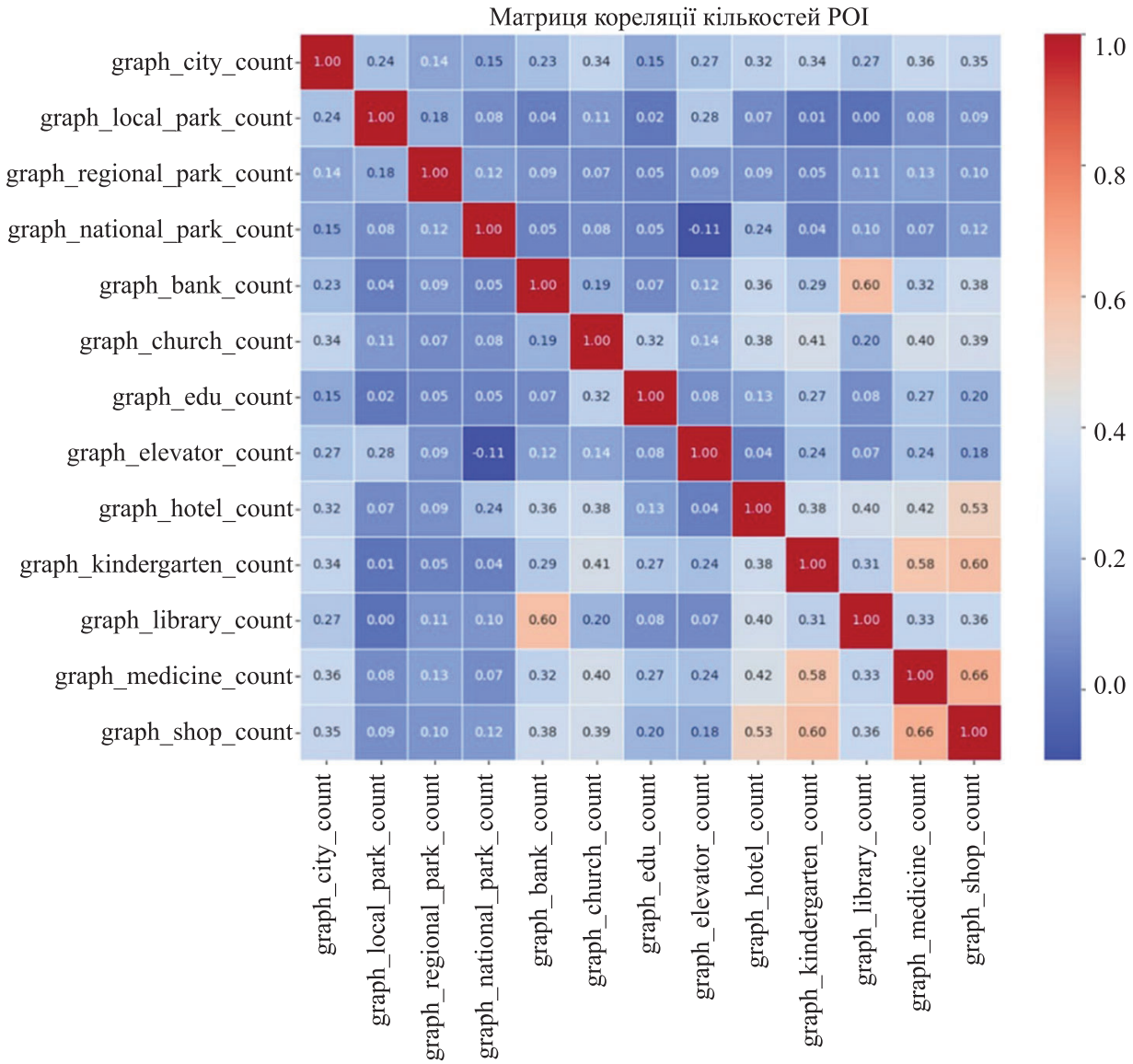


Рис. 20. Кореляційна матриця розширеного набору даних

Також необхідно оцінити, чи зберігається структура в розподілах в графових структурах, а саме чи підтверджується факт того, що сільські громади з більшою кількістю наближених POI серед однієї групи будуть мати перший об’єкт ближче, ніж села з меншою кількістю POI. Та чи підтверджується факт того, що для сіл з декількома наближеними POI, з переходом від першого до останнього об’єкту будуть віддалятися. Для цього було використано коробкові діаграми (рис. 21). Як можна побачити, дані правила виконуються.

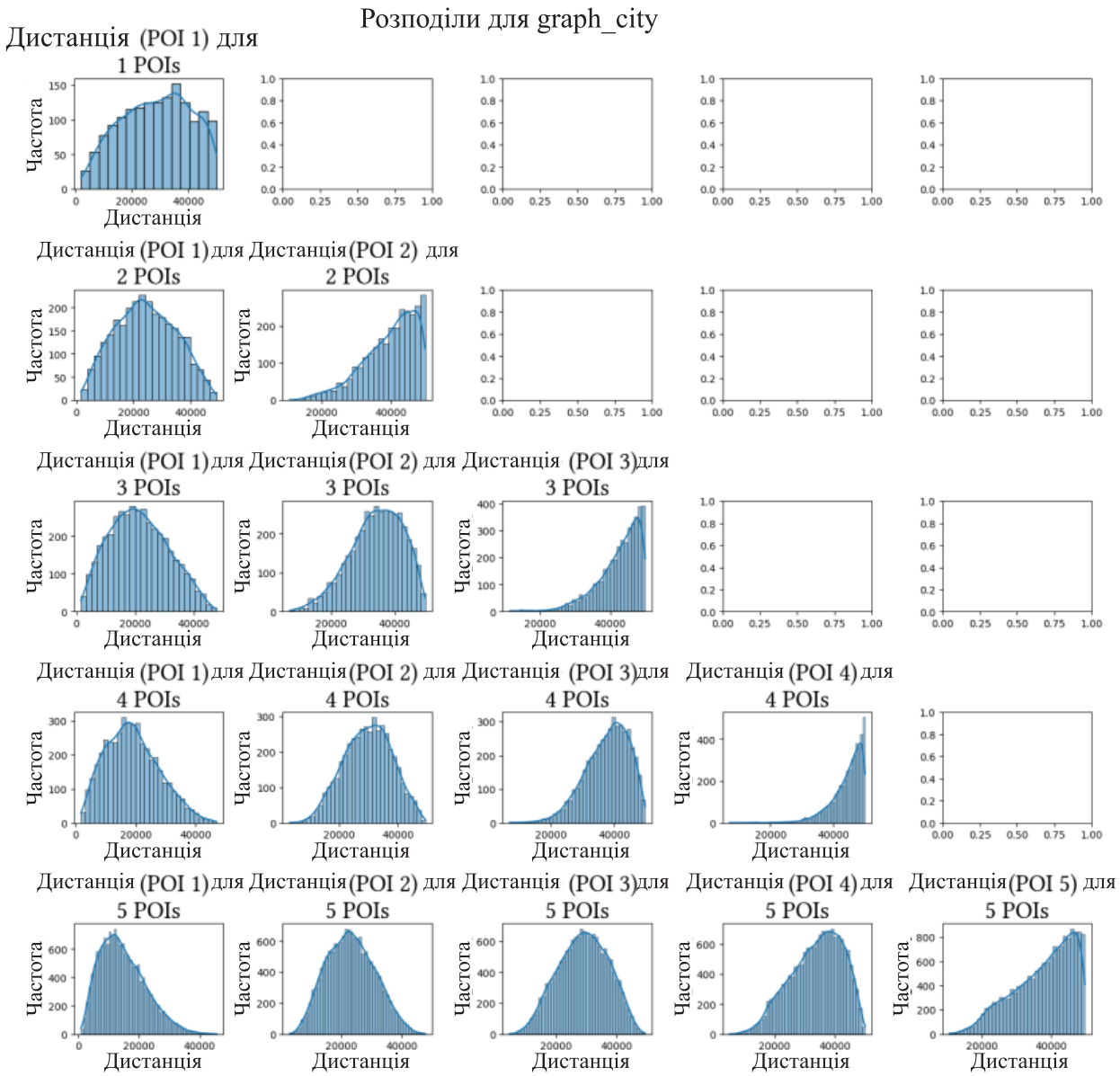


Рис. 21. Матриця розподілу об’єктів для «graph_city» на основі кількості POI на певній відстані для кожного конкретного порядку точки з цього масиву

2.6. ОБГОВОРЕННЯ

У цьому дослідженні проаналізовано доступність інфраструктури для сільських громад, що дозволило краще зрозуміти розподіл об’єктів поблизу сільських територій. Введення нових оцінок інфраструктури на основі даних про POI, зокрема щодо близькості до доріг, міст та соціальних об’єктів, дало змогу створити багатогранний опис інфраструктурних умов. Використання графових дескрипторів із детальними координатами

та відстанями до об’єктів дозволяє оцінити доступність основних ресурсів і виявити, де необхідні покращення.

Статистичний аналіз показав кілька важливих тенденцій, зокрема щодо підключення сільських громад до доріг. Наприклад, відстань до основних доріг має право-зміщений розподіл (RD_m1_NEAR), що вказує на обмеження у сполученні для значної кількості сіл, незважаючи на те, що більшість з них мають доступ до вторинних доріг, як бачимо з гістограми (рис. 17).

Аналіз показує також диспропорцію в доступності важливих послуг, таких як освіта, охорона здоров'я та роздрібна торгівля. Це вказує на необхідність покращень у наданні цих послуг. Рівномірний розподіл доступності до Києва (Kyiv_NEAR) свідчить про можливості для економічного розвитку завдяки кращому доступу до столиці, що може сприяти залученню інвестицій та розвитку місцевого бізнесу.

Коробкова діаграма на рис. 18 показує, що відстань до основних зручностей, як правило, становить менше 50 км, що вказує на відносну доступність послуг, важливих для соціального розвитку, за винятком регіональних та національних парків і відстані до столиці.

Це дозволяє зробити висновок про важливість забезпечення доступності соціальних послуг для громад.

Кореляційний аналіз підкреслює взаємозв’язок між відстанню до міських центрів та доступністю зручностей, підтверджуючи, що міські центри часто забезпечують більше послуг та об’єктів інфраструктури. Оригінальні та нові графові оцінки вказують на те, що введені метрики не змінюють характер розподілу даних, що дозволяє забезпечити достовірність рекомендацій для політики на їх основі.

Крім того, аналіз показав значну кореляцію між кількістю магазинів, медичних закладів, дитячих садків та готелів, що може вказувати на потенційні напрямки для розширення інфраструктури. Оскільки комерційні об’єкти, як магазини та готелі, в Україні здебільшого приватні, це дозволяє побачити, які державні покращення (медицина, дитячі садки, церкви) можуть залучити інвестиції та посилити розвиток.

На основі результатів було створено дешборд з градуюванням POI, що дозволяє виділити найбільш нерозвинені райони для подальших досліджень (рис. 14). Використання цих результатів у майбутньому сприятиме детальному аналізу інфраструктурних потреб у регіонах [15].

Аналіз розвитку сільської інфраструктури, представлений у цьому дослідженні, вимагає обробки великих обсягів даних, що включають геопросторові шари, графові структури та інші параметри, описані вище .

Для ефективного оброблення таких масивів даних необхідно використовувати високопродуктивні обчислювальні системи, такі як GRID або хмарні середовища [16]. Зокрема, системи Google Earth Engine або хмарне середовище CREODIAS дозволяють забезпечити масштабованість та продуктивність при аналізі геоданих великого обсягу.

Використання таких обчислювальних платформ сприяє ефективному плануванню інфраструктурних проєктів, зокрема шляхом ідентифікації пріоритетних регіонів, що потребують розвитку. Особливості застосування таких хмарних платформ будуть детально описані в частині 4 даної монографії.

Використання супутникових даних у поєднанні з розробленими моделями [17] дозволяє проводити соціо-економічний аналіз та оцінювати ризики і вплив надзвичайних ситуацій природного та антропогенного характеру на сільські території.

Зокрема, дані спостереження Землі забезпечують можливість моніторингу змін у землекористуванні, доступності ресурсів та стану інфраструктури в умовах, коли традиційні методи збору інформації є обмеженими або недоступними. Наприклад, у випадках природних катастроф, таких як посухи або повені, можна оцінювати втрати врожаю, зміну водних ресурсів та їх вплив на економіку сільських громад [18, 19].

У контексті антропогенного впливу, наприклад, війни, супутникові дані дозволяють аналізувати руйнування інфраструктури, переміщення населення та зміни у використанні земель [20-22]. Такий підхід сприяє ідентифікації найбільш вразливих регіонів, розробці стратегій пом'якшення наслідків та плануванню заходів для відновлення і сталого розвитку постраждалих територій.

ВИСНОВКИ

Проведені дослідження демонструють важливість використання методів геопросторового аналізу та методів кластеризації для оцінки та планування розвитку сільської інфраструктури, особливо в умовах України, що стикається з

унікальними соціально-економічними та інфраструктурними викликами. Основні отримані результати свідчать про ефективність запропонованих підходів у різних аспектах аналізу даних та класифікації сільських громад.

В даному розділі було запропоновано нові оцінки доступності на основі геопросторових даних, отриманих з джерел OpenStreetMap (OSM) та Humanitarian Data Exchange (HDX). Ці оцінки дозволили більш точно описати відстані до ключових об'єктів інфраструктури, таких як дороги, міста та соціальні об'єкти.

Це створило основу для більш комплексного аналізу доступності та розподілу інфраструктури у сільських районах, що є критично важливим для формування ефективних стратегій подальшого розвитку та відновлення.

Отримані результати кластеризації сільських громад показали, що запропонований підхід, що базується на поєднанні геопросторових даних та технік машинного навчання, є дієвим інструментом для визначення основних типів сільських громад за рівнем інфраструктурного розвитку. Це дозволило виявити ключові прогалини у розвитку інфраструктури та встановити пріоритетні напрями для подальших робіт.

Проведений кількісний та графічний аналіз продемонстрував узгодженість нових оцінок з початковими метриками, підтверджуючи надійність запропонованих методів.

Запропонований підхід до здійснення кластеризації, яка була застосована на реальних даних, показав можливість його використання для стратегічного планування розвитку сільських районів. Побудовані карти інфраструктурних зв'язків допомагають визначити області, які потребують першочергових інфраструктурних інвестицій, і сприяють кращому розумінню просторових нерівностей у доступності ключових послуг.

Подальші дослідження можуть бути спрямовані на подальше розширення набору даних для включення додаткових соціально-економічних показників, таких як рівень доходів та зайнятості, що дозволить отримати більш точну картину розвитку сільських громад. Важливо також інтегрувати дані в реальному часі для оперативного аналізу та моніторингу змін у інфраструктурі сільських територій.

ПЕРЕЛІК ПОСИЛАНЬ

1. Y. Liu та ін. «Geospatial characterization of rural settlements and potential targets for revitalization by geoinformation technology». Англ. В: Scientific Reports 12 (2022), с. 8399. url: https://doi.org/10 .1038/s41598-022-12294-2.

2. Hanna Yailymova та ін. «Geospatial Analysis of Life Quality in Ukrainian Rural Areas». В: 2023 13th International Conference on Dependable Systems, Services and Technologies (DESSERT). 2023, с. 1—5. doi: 10.1109/DESSERT61349.2023.10416517.

3. A. Maryada та V. L. Thatiparthi. «Geospatial technology for mapping and analysis of social and infrastructural facilities at village level: a case study of Chinnapendyala village». Англ. В: Modeling Earth Systems and Environment 6 (2020), с. 1763—1781. url: https://doi.org /10.1007/s40808-020-00788-9.

4. B. Herfort та ін. «A spatio-temporal analysis investigating completeness and inequalities of global urban building data in OpenStreetMap». Англ. В: Nature Communications 14 (2023), с. 3985. url: https://doi.org/10. 1038/s41467-023-39698-6.

5. Zijing Liu та Mauricio Barahona. «Graph-based data clustering via multiscale community detection». В: Applied Network Science 5.1 (2020), с. 3. issn: 2364-8228. doi: 10 . 1007 / s41109 - 019 - 0248 - 7. url: https://doi.org/10.1007/s41109-019-0248-7.

6. Gadelha T. Filho Vinicius та ін. «Rural electrification planning based on graph theory and geospatial data: A realistic topology oriented approach». В: Sustainable Energy, Grids and Networks 28 (2021), с. 100525. issn: 2352-4677. doi: https://doi. org/10.1016/j. segan .2021.100525. url: https://www.sciencedirect.com/science/article/pii/S2352467721000965.

7. Akhbar Sha та ін. «Data-Driven Clustering and Insights for Rural Development in India». В: Procedia Computer Science 233 (2024). 5th International Conference on Innovative Data Communication Technologies and Application (ICIDCA 2024), с. 336—342. issn: 1877-0509. doi: 10.1016/j.procs.2024.03.223. url: https://www.sciencedirect.com/science/article/pii/S1877050924005829.

8. Kussul, N., Svirsh, V., Potuzhnyi, B. Integrated Geospatial Analysis for Rural Development Metrics. CEUR Workshop

ProceedingsVolume 3664, Pages 141 - 160. 2024. Machine Learning Workshop of the 8th International Conference on Computational Linguistics and Intelligent Systems, MLW-CoLInS 2024. url: http://ceur-ws.org/Vol-3664/paper11.pdf.

9. Elevatorist. Elevators in Ukraine. https://elevatorist.com/karta-elevatorov-ukrainy. 2022.

10. Нова Пошта. Національний оператор експрес-доставки України. 2023. Режим доступу: https://novaposhta.ua/.

11. Укрпошта. Національний оператор поштового зв'язку України. 2023. Режим доступу: https://www.ukrposhta.ua/.

12. Mobua.net. Інформаційний портал мобільного зв'язку України. 2023. Режим доступу: https://mobua.net/.

13. James MacQueen. «Some Methods for Classification and Analysis of Multivariate Observations». B: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability 1.14 (1967), с. 281—297.

14. Brian S Everitt. Cluster Analysis. Heinemann Educational Books, 1974.

15. Сайт кафедри ММАД НТУУ "КПІ". Режим доступу: https://mmda.ipt.kpi.ua/portal/MON/Products.html?group=rural_development&infra=rdi-interactive.

16. Куссуль Н. Н. Grid-системы для задач исследования Земли. Архитектура, модели и технологии / Н. Н. Куссуль, А. Ю.Шелестов. — К. : Наукова думка, 2008. — 452 с.

17. Kussul, N. , Potuzhnyi, B., Svirsh, V. Clustering Techniques for Modeling Village Infrastructure Development.CEUR Workshop Proceedings, 2024, 3668, pp. 98–119.

18. Ghazaryan, G., Dubovyk, O., Graw, V., Kussul, N., & Schellberg, J. (2020). Local-scale agricultural drought monitoring with satellite-based multi-sensor time-series. GIScience & Remote Sensing, 57(5), 704-718.

19. Kussul, N., Shelestov, A., Skakun, S., & Kravchenko, O. (2008). Data assimilation technique for flood monitoring and prediction. International Journal Information Theories & Applications, 15, 76-83.

20. Skakun, S., Justice, C. O., Kussul, N., Shelestov, A., & Lavreniuk, M. (2019). Satellite data reveal cropland losses in South-Eastern Ukraine under military conflict. Frontiers in Earth Science, 7, 305.

21. Kussul, N., Drozd, S., Yailymova, H., Shelestov, A., Lemoine, G., & Deininger, K. (2023). Assessing damage to agricultural fields from military actions in Ukraine: An integrated approach using statistical indicators and machine learning. International Journal of Applied Earth Observation and Geoinformation, 125, 103562.

22. S. Skakun et al., "High-Impact Hot Spots of Land Cover Land Use Change in Ukraine" 2022 12th International Conference on Dependable Systems, Services and Technologies (DESSERT), Athens, Greece, 2022, pp. 1-5, doi: 10.1109/DESSERT58054.2022.10018657.