

Ю. МАЦЕЛЮХ, В. ЛИТВИН

МЕТОД К-СЕРЕДНІХ ДЛЯ АНАЛІЗУ ДАНИХ З ОРГАНІЗАЦІЇ ПАСАЖИРСЬКИХ ПЕРЕВЕЗЕНЬ У РОЗУМНОМУ МІСТІ

Щороку методи кластеризації великих даних набувають популярності в прийнятті рішень з організації пасажирських перевезень у розумному місті, забезпечуючи ефективність, адаптивність і екологічність транспортної системи. Актуальність зазначених методів зумовлена зростанням обсягів даних, зміною попиту та негативним впливом транспорту на довкілля. **Об'єктом дослідження** є процес кластеризації масивів даних з організації пасажирських перевезень. **Предмет дослідження** – засади вивчення метрик кластеризації в процесі обчислення кількості кластерів виконання графіків перевезень. **Мета роботи** полягає в застосуванні з огляду на метрики якості методу К-середніх для кластеризації даних з організації пасажирських перевезень у розумному місті. Для досягнення мети необхідно виконати такі завдання: дослідити особливості методів кластеризації та їх метрик; проаналізувати масштабний неоднорідний набір даних щодо тривалості перегонів електротранспортом у середньому за розмірами міста; розробити ефективний алгоритм вибору методу обчислення кількості кластерів на основі метрик оцінювання якості кластеризації даних. Упроваджено **методи** аналізу, синтезу, узагальнення, порівняння, групування, кластерного аналізу, системного аналізу, метод К-середніх. **Досягнуті результати.** Установлено, що вибір методу кластеризації залежить від особливостей завдання, характеристик даних і цілей аналізу транспортних потоків. Виявлено складну, неоднорідну й необроблену структуру даних щодо тривалості перегонів електротранспортом. Кластерний аналіз методом К-середніх зумовлено потребою в точному розподілі даних між кластерами. Запропоновано алгоритм вибору методу обчислення кількості кластерів на основі метрик оцінювання якості кластеризації даних, серед яких метод ліктя, метод силуету та індекс Калінського – Харабаша. Рекомендовано застосовувати кластеризацію для створення маршрутів із скороченням часом очікування, меншою кількістю пересадок та відповідністю до потреб пасажирів. **Висновки.** Метод К-середніх використано для аналізу тривалості виконання перегонів електротранспортом. Аналіз даних виявив ділянки маршрутів із різною інтенсивністю транспортних потоків, що залежить від сезонності, їх розташування в міських зонах тощо. Запропоновано алгоритм вибору методу обчислення кількості кластерів на основі внутрішніх метрик.

Ключові слова: пасажирські перевезення; розумне місто; кластерний аналіз; метод К-середніх; системний аналіз.

Вступ

Питання організації пасажирських перевезень у розумному місті тісно пов'язано з пошуком ефективних методів і засобів, що забезпечують оптимальну динамічну взаємодію з транспортом під час організації пасажирських перевезень. В організації пасажирських перевезень важливо брати до уваги різноманітні показники, зокрема маршрути руху, час очікування, тривалість виконання графіків, зміни трафіку, завантаженість транспортних засобів, погодні умови, екологічну ефективність, індивідуальні потреби пасажирів тощо. Зібрані великі масиви даних потребують належного зберігання, відповідного аналізу, ефективного групування та якісної кластеризації.

Актуальність проблеми задекларована на світовому рівні в багатьох документах (угодах, комюніке, деклараціях тощо). Міжнародні організації, зокрема ООН і ЄС, активно підтримують ініціативи щодо зниження викидів парникових газів

у міських агломераціях. Нині проблема організації низьковуглецевих пасажирських перевезень є надзвичайно актуальною через глобальне потепління та зростання потреби в екологічно чистих транспортних засобах. Розумне місто, що використовує сучасні технології та системи управління, може стати ефективним інструментом для досягнення визначених цілей.

Розв'язання окресленої проблеми неможливе без використання методів і засобів теорії системного аналізу та інструментів інформаційних систем. З погляду системного аналізу та інформаційних систем проблема організації низьковуглецевих пасажирських перевезень є складником комплексних систем управління транспортними потоками. Системний підхід дає змогу розглядати всі компоненти системи (транспортні засоби, інфраструктуру, пасажирів) як взаємопов'язані елементи, що впливають один на одного. Інформаційні системи відіграють провідну роль у зборі, обробленні та аналізі даних для оптимізації

маршрутів і часу очікування на транспорт, а також для моніторингу екологічних показників.

Проблема має неабияке **практичне значення**, що збільшується щороку й щоразу із впровадженням інтелектуальних систем динамічної взаємодії пасажирів з транспортними засобами в розумному місті, серед яких виокремимо такі: зниження викидів вуглецю, покращення якості життя, економія ресурсів і формування розумної міської інфраструктури. Зібрані масиви даних потребують належного та якісного аналізу, що неможливий без застосування ефективних алгоритмів кластеризації під час оптимізації маршрутів, упровадження екологічно чистих транспортних засобів на маршрутах, а також ефективної організації перевезень, спрямованої на зменшення часу очікування та завантаженості транспортних засобів, на покращення комфорту пасажирів, зниження викидів парникових газів, зменшення витрат на перевезення, оптимізацію використання транспортної інфраструктури та енергоефективні технології тощо. Загалом розв'язання окресленої проблеми робить вагомий внесок у розвиток розумних міст, які використовують сучасні технології для покращення якості життя мешканців. Тому пошук шляхів упровадження методів кластеризації, серед яких метод *K*-середніх, в організації пасажирських перевезень у розумному місті є важливим складником загальної проблеми розроблення методів і засобів динамічної взаємодії пасажирів із транспортними засобами й має вагоме практичне значення для розвитку низьковуглецевих пасажирських перевезень у громадському транспорті великих, середніх та малих міст. **Об'єктом дослідження** є процес кластеризації масивів даних з організації пасажирських перевезень, а **предметом** – засади вивчення метрик кластеризації під час обчислення кількості кластерів виконання графіків перевезень.

Аналіз останніх досліджень і публікацій

Провівши детальний огляд методів і засобів кластерного аналізу [1, 2], що використовуються для моделювання [3], оптимізації маршрутів [4], аналізу великих даних [5] з метою розроблення адаптивних алгоритмів для організації транспортної мережі в розумному місті [6, 7], з'ясували, що особлива увага приділяється моделям прийняття рішень для оптимізації пасажирських потоків з огляду на сучасні підходи, що ґрунтуються як на висхідних

(агломератних) [8, 9], так і на низхідних (дивізивних) [10] методах кластеризації, а також методам роздільної [11], нечіткої кластеризації [12], DBSCAN [13] та самоорганізованих карт [14]. Сучасні дослідження у галузі системного аналізу [15–17] підтверджують, що інтеграція методів кластеризації до системи управління транспортом сприяє створенню адаптивних, ефективних і низьковуглецевих моделей перевезень [18, 19]. Розвитку фундаментальних основ методів кластеризації в організації пасажирських перевезень, розробленню сучасних інформаційно-комунікаційних систем пасажирських перевезень громадським транспортом присвячені дослідження відомих науковців як в Україні, так і за її межами. Серед учених, чий внесок сприяв розвитку теоретичних засад і практичного досвіду застосування кластерного аналізу даних з організації пасажирських перевезень на основі концепції розумного міста, доцільно згадати таких: Дж. Бездек [20], М. Бублик [21, 22], М. Естер [23], А. Джейн [24], Т. Кохонен [25], Д. Коштура [26], В. Литвин [27], А. Лов [28], Н. Нат [29], Л. Сан [30], Р. Тібішрані [31].

Основними методами кластеризації, що використовують для аналізу великих даних у сфері організації пасажирських перевезень розумного міста [32–36], є ієрархічна кластеризація, роздільна кластеризація, кластеризація на основі щільності, сітки, штучні нейронні мережі. Вони дають змогу не лише зрозуміти індивідуальні потреби пасажирів і патерни споживання послуг, а й сприяють оптимізації ресурсів для задоволення цих потреб. Щоб виокремити, які методи кластеризації використовують для організації пасажиропотоків, було проведено порівняння основних методів кластеризації у сфері організації пасажирських перевезень розумного міста. Результати порівняльного аналізу наведено в табл. 1.

У разі ієрархічної кластеризації агломератні, або, як їх інакше називають, висхідні методи виходять з того, що кожен елемент у наборі даних є окремим кластером. Сам процес об'єднання двох найближчих кластерів до одного відбувається за певними правилами (за вказаною метрикою), доки не утвориться лише один кластер. Застосовуються висхідні методи для організації пасажиропотоків, коли потрібно визначити структуру маршрутів, і об'єднання відбувається за схожими ділянками маршрутів або територіями. У дивізивних, або низхідних, методах, навпаки, починають з одного

кластера, що містить усі дані, і розділяють його на менші кластери. Їх застосовують для організації пасажиропотоків, коли потрібно виокремити конкретні

маршрути або їх ділянки (перегони, зони) з різною інтенсивністю пасажиропотоку, щоб оптимізувати послуги в конкретних зонах.

Таблиця 1. Результати порівняльного аналізу основних методів кластеризації у сфері організації пасажирських перевезень у розумному місті (складено авторами за джерелами [1–36])

Група кластеризації	Метод кластеризації	Сутність методу	Вимоги до даних	Особливості застосування в організації пасажирських перевезень
Ієрархічна	Агломератна ієрархія, або висхідний метод	створює на дендрограмі ієрархічну структуру, відстежує процес об'єднання або розділення кластерів знизу вгору на основі найближчих сусідів	потребує від невеликого до середнього набору даних	підходить для визначення структури транспортних маршрутів і аналогій у розподілі пасажиропотоків з огляду на динамічні зміни
	Дивізивна ієрархія, або низхідний метод	поєднує кластери зверху вниз на основі найближчих сусідів	потребує повної матриці відстаней між об'єктами	підходить для динамічного моделювання пасажиропотоків у різних часових інтервалах
Розділяюча	K -середніх	поділяє дані на k -кластери за найменшою відстанню до центрів, оптимізуючи внутрішню структуру	чисельні показники зі складним просторовим розподілом, попередньо нормалізовані, погано працює з якісними характеристиками	використовується для оптимізації маршрутизації транспорту за зонами попиту, а в статичних системах для розподілу пасажиропотоків за визначеними маршрутами
	K -медоїд	групує об'єкти на k -медоїдних точках, що є класичними представниками	потребує попередньої нормалізації даних	підходить для визначення центрів маси для оптимального розподілу транспортних засобів
	Нечітких C -середніх	дає змогу елементам належати до кількох кластерів одночасно з різними ступенями належності	працює в складних системах із великою неоднозначністю та перекриттям між даними, працює з якісними характеристиками	використовується для створення більш гнучких та адаптивних зон міського транспорту, де пасажир одночасно належить до кількох зон
На основі щільності	DBSCAN	виявляє кластери будь-якої форми на основі щільності даних у просторі, знаходить шум	найкраще працює з даними з високою щільністю у певних ділянках, потребує значення мінімальної кількості точок у кластері	корисна для оброблення інформації про пасажиропотоки з різноманітними густинами на різних маршрутах, ідентифікує вузькі місця маршрутів і зони із значною концентрацією пасажирів
Сітка	STING або CLIQUE	поділяє простір на ґрид, кластеризує дані у вказаних окремих комітках за допомогою структурованих ієрархічних інтервалів	розмір і розділ ґрида впливає на ефективність, розбиття даних на менші групи дає кращі результати	аналізує пасажиропотоки у великому місті на основі зонування, корисна для класифікації маршрутів за часовими інтервалами та густиною пасажиропотоків
Штучні нейронні мережі	Самоорганізованих карт	використовує алгоритми навчання для самоорганізації вхідних даних у кластери, використовує самоорганізаційні карти	потребує значної кількості інформації для навчання нейронної мережі, попередньої нормалізації даних та розбиття на менші групи для аналізу	використовується для оптимізації ресурсів на основі здобутих даних з попередньої інформації у процесі прогнозування пасажиропотоків і створення динамічних моделей пасажиропотоків з огляду на часові інтервали

У разі роздільної кластеризації метод K -середніх розподіляє дані на кластери, знаходячи середню

точку в кожному кластері та повторюючи цей процес до досягнення стабільного розподілу. Зазначений

метод обов'язково точно розподіляє дані, де кожен об'єкт належить лише одному класу. Проте він погано адаптований для даних з непростим розподілом або з якісними характеристиками. Метод *K*-середніх ефективний для зонування транспортної мережі, наприклад, для оптимізації маршрутів за зонами попиту. Метод *K*-медоїд виявляє центри кластерів, що мають найбільш можливий відрив від загального пасажиропотоку. Особливістю цього методу є більша стійкість до шуму. Використовується він для оптимізації робочого графіку транспортних засобів, визначення оптимального розташування зупинок на маршрутах, наприклад станцій метро, здатних покривати значні потреби пасажирів. Метод нечіткої кластеризації *C*-середніх допомагає створити більш гнучкі та адаптивні стратегії управління пасажиропотоками, що важливо в сучасних міських середовищах із динамічною природою попиту на транспортні послуги. Цей метод відрізняється від традиційного методу *K*-середніх тим, що дає змогу елементам належати до кількох кластерів одночасно з різними ступенями належності. Це означає, що кожен елемент має певну ймовірність, за якою він належить до кожного з *k*-кластерів. У зазначеному методі використовується функція належності, що визначає ступінь належності елемента до кожного з можливих кластерів. Метою оптимального розподілу є мінімізація функції, яка бере до уваги як відстань до центрів кластерів, так і ступені належності. Дає найкращі результати в складних системах з великою неоднозначністю та перекриттям між даними. Використовується для створення більш гнучких та адаптивних зон міського транспорту, де пасажир одночасно належить до кількох зон, зважаючи на непередбачуваність попиту, наприклад, зміни пасажиропотоку протягом доби або за різних погодних умов. Виявляє патерни, невидимі для інших методів, що можуть бути важливими в розробленні стратегій оптимізації. Використання *C*-середніх дає точнішу картину сегментування користувачів транспорту, знижує ризик надлишкових або недостатніх обсягів послуг на певних ділянках транспортної мережі. Метод також ефективний для передбачення змішаних потреб та їх впливу на розподіл кластерів, покращуючи цим прийняття стратегічних рішень в організації транспорту.

У класифікації на основі щільності найбільш відомим є метод DBSCAN, що найчастіше використовується для аналізу пасажиропотоків

із різноманітною густиною на маршрутах та для визначення кластерів з огляду на часову динаміку. У класифікації на основі сітки метод STING розбиває для аналізу дані на більш малі групи, щоб створити моделі часових інтервалів та густину. З-поміж штучних нейронних мереж, що використовують для кластеризації, найпоширенішим є метод самоорганізованих карт, який дає змогу створити динамічні карти пасажиропотоків, зважаючи на часові інтервали та густину на різних маршрутах, проте він потребує значної кількості даних для навчання нейронної мережі, попередньої їх нормалізації та розбиття на менші групи для аналізу.

Отже, кожен метод кластеризації має переваги й може бути застосований у різних сценаріях для ефективної організації та оптимізації пасажирських перевезень у розумному місті. Вибір методу залежить від особливостей завдання, характеристик даних і цілей, що висуваються для аналізу транспортних і пасажирських потоків. Висхідні методи кластеризації застосовуються, коли потрібно визначити структуру маршрутів, об'єднати схожі ділянки маршрутів або прилеглих до маршруту територій. Низхідні методи – коли потрібно виокремити деякі маршрути або їх ділянки (перегони, зони) з різною інтенсивністю пасажиропотоків у конкретних зонах. Метод роздільної кластеризації *K*-середніх корисний для зонування транспортної мережі в процесі оптимізації маршрутів за тривалістю виконання перегонів, за зонами попиту тощо. Метод роздільної кластеризації *K*-медоїд завдяки більшій стійкості до шуму застосовують для оптимізації робочого графіка транспортних засобів, визначення оптимального розташування зупинок на маршрутах для задоволення потреб пасажирів. Метод нечіткої кластеризації *C*-середніх – у розробленні більш гнучких і адаптивних стратегій управління пасажиропотоками в сучасних міських середовищах з динамічною природою попиту на транспортні послуги. Метод DBSCAN найчастіше використовують для аналізу пасажиропотоків із різноманітною густиною на різних маршрутах та для кластеризації з огляду на часову динаміку. Метод самоорганізованих карт – для створення динамічних карт пасажиропотоків, зважаючи на часові інтервали та густину на різних маршрутах.

У розумному місті створюється величезна кількість даних із різних джерел, зокрема GPS-систем, мобільних застосунків, сенсорів, соціальних мереж та відеоспостережень. Оброблення таких значних

обсягів інформації є критично важливим для прийняття обґрунтованих рішень у сфері пасажирських перевезень. Потоки пасажирів постійно змінюються залежно від часу доби, дня тижня, погоди, соціальних подій тощо. Методи кластеризації дають змогу виокремити ключові групи або патерни в таких потоках для кращого розуміння їх природи. Методи кластеризації є основою багатьох сучасних алгоритмів штучного інтелекту. Використання інтелектуальних транспортних систем (ITS) [37–41] потребує складних алгоритмів аналізу та моделювання для виявлення оптимальних шляхів і управління транспортною мережею. Ефективна кластеризація даних сприяє оптимізації маршрутів, зменшенню часу простою та обсягів викидів, що є важливим у контексті боротьби зі змінами клімату. Таке групування є надзвичайно важливим для розроблення адаптивних алгоритмів оптимізації маршрутів, адже воно дає змогу ефективно розподіляти транспортні ресурси, знижувати час очікування та мінімізувати викиди вуглецевмісних сполук.

Як зазначає М. Бублик [42], концепція розумної спеціалізації для трансформації української економіки передбачає не лише оптимізацію господарської діяльності транспортних компаній, але й перехід до зеленої економіки, де значну роль відіграє зниження викидів CO₂ завдяки впровадженню інноваційних рішень у транспортну індустрію. В основі інноваційних моделей для зменшення обсягів викидів в атмосферу лежить концепція техносолітону, що розробив М. Бублик [43, 44]. У межах концепції оцінено шкоду та збитки в дуже забруднювальних секторах економіки, яким тривалі роки залишається транспорт. Концепція техносолітону набуває особливого значення для розроблення стратегій організації пасажирських перевезень, оскільки оптимізація маршрутів за допомогою методу *K*-середніх дає змогу не лише покращити якість обслуговування, але й зменшити викиди в атмосферу, що має вирішальне значення для досягнення цілей сталого розвитку [45–50].

Визначення не розв'язаних раніше частин загальної проблеми. Мета роботи, завдання

Проведений вище аналіз останніх досліджень проблеми застосування методів кластеризації, серед яких метод *K*-середніх, в організації пасажирських перевезень у розумному місті дає змогу підсумувати, що недостатньо розв'язаним

залишається питання впливу методів визначення закономірностей пасажиропотоків, оптимізації транспортних маршрутів і підвищення ефективності мережі в режимі реального часу на покращення комфорту пасажирів, зниження викидів парникових газів, зменшення витрат на перевезення, оптимізацію використання транспортної інфраструктури та енергоефективні технології тощо. Недостатньо також приділено увагу пошуку ефективних способів застосування алгоритмів кластеризації під час оптимізації маршрутів, у впровадженні екологічно чистих маршрутів, а також під час ефективної організації перевезень, спрямованої на зменшення часу очікування пасажирів загалом чи завантаженості транспортних засобів зокрема. Це вказує на необхідність наукових досліджень у зазначеному напрямі, а саме застосуванні з огляду на метрики якості методу *K*-середніх для кластеризації даних з організації пасажирських перевезень у розумному місті, що і є **метою цієї роботи**.

Відповідно до мети окреслено такі **завдання**: дослідження особливостей методів кластеризації та їх метрик; аналіз масштабного неоднорідного набору даних щодо тривалості перегонів електротранспортом у середньому за розмірами місті; розроблення ефективного алгоритму вибору методу обчислення кількості кластерів на основі метрик оцінювання якості кластеризації даних.

Матеріали й методи

Для аналізу даних після їх порівняння та групування було обрано метод кластерного аналізу *K*-середніх. Особливістю застосування методів кластеризації даних є вибір метрики відстані, яку серед багатьох інших різних показників необхідно обирати з огляду на її відповідність до конкретного прикладу. У нашому випадку це дослідження набору даних про рух низьковуглецевих транспортних засобів на одному маршруті в середньому за розмірами міста, тому для цього використано евклідову відстань, що описується формулою

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}, \quad (1)$$

де $x = (x_1, x_2, \dots, x_n)$ – вектор характеристики точки x , що містить n компонент; $y = (y_1, y_2, \dots, y_n)$ – вектор характеристики точки y , який містить n компонент; i – індекс для кожного з атрибутів (номер атрибута).

Вибір кількості кластерів безпосередньо впливає на якість кластеризації, тому важливо обрати оптимальну кількість кластерів для набору даних. Ми обрали метод ліктя, що передбачає аналіз залежності суми квадратичних відстаней $SSE(k)$ між точками й центрами їх кластерів від величини k . Сума квадратичних відстаней обчислюється за формулою

$$SSE(k) = \sum_{i=1}^k \sum_{x_j \in C_i} |x_j - \mu_i|^2, \quad (2)$$

де $SSE(k)$ – сума квадратичних відстаней; k – кількість кластерів; x_j – точка в наборі даних, що належить кластеру $C_j (x_j \in C_j)$; μ_i – центр i -го кластера.

Алгоритм методу K -середніх є одним із найбільш поширених, що використовується для розбиття набору даних на k кластерів. Він працює ітеративно, мінімізуючи суму квадратів відстаней точок до центрів кластерів. Алгоритм методу K -середніх має чотири етапи: ініціалізацію, призначення точок до кластерів, оновлення центрів кластерів і перевірку критерію зупинки.

Ініціалізація передбачає вибір k початкових центрів кластерів $\mu_1, \mu_2, \dots, \mu_k$ випадковим чином або за допомогою спеціальних стратегій, а також призначення точок до кластерів. Кожен елемент даних x_i призначається до найближчого центра μ_k кластера за критерієм найменшої відстані, для обчислення якого використовується формула

$$c_i = \arg \min_k \|x_i - \mu_k\|^2, \quad (3)$$

де c_i – кластер, до якого призначається точка x_i ; $\|x_i - \mu_k\|^2$ – квадрат евклідової відстані.

Оновлення центрів кластерів відбувається щоразу, коли додається нова точка до кластера. Розглядається послідовно кожна точка, виділена в кластер, до якого вона розташована ближче за критерієм найменшої відстані. Після призначення всіх точок до кластерів обчислюється новий центр μ_k кожного кластера як середнє значення всіх точок, що до нього належать:

$$\mu_k = \frac{1}{|S_k|} \sum_{x_i \in S_k} x_i, \quad (4)$$

де S_k – множина точок, що належать до k -го кластера.

Центроїд послідовно перераховується щоразу, коли додається нова точка до кластера, тобто коли

змінюється розбиття точок на кластери, тоді змінюються координати центроїдів на нові. Щоб бути впевненим, що кожна точка була оптимально призначена в правильний кластер, порівнюється відстань кожної точки кластера до центра свого власного кластера та до центра найближчого протилежного кластера за формулою

$$C_{opt} = \sum_{k=1}^K \sum_{x_i \in S_k} \operatorname{argmin}_k |x_i - \mu_k|^2. \quad (5)$$

Ітераційне перенесення точок триває в процесі кожного нового поділу на кластери доти, доки останній поділ не буде визнано остаточним результатом кластеризації.

Перевірка критерію зупинки вказує на зупинку алгоритму. Алгоритм кластеризації завершується, якщо центри кластерів перестають змінюватися або зміни є незначними. Інакше повертаємось до кроку 2. Цілком можливо, що алгоритм K -середніх не знайде остаточного рішення. У цьому разі доцільно встановити зупинку алгоритму після того, як алгоритм досягне попередньо обраного максимального значення ітерацій. Отже, алгоритм K -середніх ітеративно покращує розподіл точок між кластерами, зменшуючи значення функції втрат.

Результати досліджень та їх обговорення

У нашому випадку кластеризація здійснювалася з використанням набору даних про тривалість виконання перегонів низьковуглецевим громадським транспортом у межах середнього за розміром міста. Саме тут аналізується час виконання перегону упродовж однієї години всіма транспортними засобами, що перебувають на маршруті, у межах одного маршруту для кожного тижня протягом року. Структура даних має такий вигляд: Номер запису; Геонона; Планований час прибуття; Фактичний час прибуття; Місяць; День; Час; Дата; Тиждень; Година; День тижня; Робочий / вихідний; Перегін 1; Тривалість 1; Перегін 2; Тривалість 2; Перегін 3; Тривалість 3; ... Перегін 10; Тривалість 10.

Унаслідок аналізу зібраних даних про тривалість виконання кожного перегону транспортними засобами в межах кожної доби для кожного тижня протягом досліджуваного періоду було виявлено, що вони мають складну, неоднорідну й великомасштабну структуру, тому потребують відповідного оброблення перед початком аналізу. Після очищення й групування даних було отримано матрицю виконання

графіків перевезень пасажирів транспортними засобами на кожному з десятих перегонів із середніми значеннями тривалості перегону для кожного тижня впродовж десяти місяців. Як приклад,

у табл. 2 наведено тривалість виконання перегону в середньому за добу протягом кожного тижня досліджуваного періоду.

Таблиця 2. Середньодобова тривалість виконання кожного перегону транспортними засобами для кожного тижня (розрахунок авторів відповідно до зібраної інформації)

Перегін Тиж- день	Час виконання перегону 1	Час виконання перегону 2	Час виконання перегону 3	Час виконання перегону 4	Час виконання перегону 5	Час виконання перегону 6	Час виконання перегону 7	Час виконання перегону 8	Час виконання перегону 9	Час виконання перегону 10
9	00:03:04	00:02:14	00:02:35	00:03:16	00:03:42	00:02:25	00:02:22	00:01:56	00:01:44	00:02:03
10	00:03:10	00:02:17	00:02:48	00:03:21	00:03:24	00:02:22	00:02:25	00:01:59	00:01:46	00:01:57
11	00:03:09	00:02:02	00:02:25	00:03:26	00:03:24	00:02:24	00:02:23	00:02:00	00:01:41	00:01:55
12	00:03:08	00:02:04	00:02:22	00:03:21	00:03:15	00:02:26	00:02:24	00:01:59	00:01:46	00:01:58
13	00:03:28	00:02:08	00:02:09	00:03:18	00:03:58	00:02:22	00:02:25	00:02:01	00:01:43	00:01:58
14	00:03:16	00:02:01	00:02:25	00:03:19	00:03:18	00:02:19	00:02:19	00:01:56	00:01:40	00:01:53
15	00:03:21	00:02:02	00:02:07	00:03:14	00:03:06	00:02:20	00:02:22	00:01:57	00:01:45	00:01:53
16	00:04:00	00:02:56	00:02:42	00:03:24	00:03:09	00:02:16	00:02:25	00:01:56	00:01:39	00:01:52
17	00:04:05	00:02:23	00:02:15	00:03:10	00:03:08	00:02:20	00:02:20	00:01:54	00:01:40	00:01:48
18	00:04:02	00:02:04	00:01:50	00:02:47	00:02:50	00:02:22	00:02:14	00:01:55	00:01:39	00:01:53
19	00:04:05	00:02:05	00:02:12	00:03:08	00:03:05	00:02:23	00:02:20	00:01:58	00:01:42	00:01:55
20	00:03:23	00:02:27	00:02:07	00:03:12	00:03:13	00:02:23	00:02:22	00:01:58	00:01:41	00:01:52
21	00:04:27	00:04:19	00:03:31	00:03:54	00:03:24	00:02:36	00:02:22	00:02:06	00:01:45	00:01:54
22	00:03:26	00:03:51	00:03:17	00:03:19	00:03:52	00:02:32	00:02:26	00:01:58	00:01:41	00:01:52
27	00:04:21	00:03:49	00:02:13	00:03:15	00:03:00	00:02:24	00:02:17	00:01:59	00:01:42	00:01:55
28	00:04:51	00:04:23	00:02:16	00:03:06	00:02:55	00:02:22	00:02:14	00:01:58	00:01:42	00:01:57
29	00:04:44	00:03:44	00:02:06	00:02:57	00:02:55	00:02:22	00:02:16	00:01:59	00:01:42	00:01:57
30	00:04:08	00:04:01	00:02:07	00:03:18	00:02:56	00:02:23	00:02:17	00:01:58	00:01:45	00:01:55
31	00:04:30	00:03:24	00:02:18	00:03:31	00:02:52	00:02:23	00:02:20	00:01:53	00:01:47	00:01:56
32	00:05:13	00:04:00	00:02:03	00:03:12	00:02:59	00:02:25	00:02:17	00:01:53	00:01:43	00:02:00
33	00:05:30	00:03:54	00:01:57	00:03:03	00:02:57	00:02:20	00:02:14	00:01:55	00:01:44	00:01:53
34	00:05:17	00:03:02	00:01:53	00:02:50	00:02:41	00:02:26	00:02:32	00:02:00	00:01:44	00:01:54
35	00:05:23	00:03:47	00:02:09	00:02:55	00:02:54	00:02:25	00:02:19	00:01:58	00:01:48	00:01:54
36	00:05:36	00:03:19	00:02:28	00:03:24	00:03:02	00:02:31	00:02:23	00:01:56	00:01:42	00:01:57
37	00:05:12	00:03:15	00:02:24	00:03:12	00:03:05	00:02:29	00:02:22	00:01:57	00:01:43	00:01:55
38	00:05:16	00:03:14	00:02:57	00:03:27	00:03:11	00:02:26	00:02:22	00:01:57	00:01:49	00:01:54
39	00:04:39	00:03:43	00:02:18	00:03:11	00:03:01	00:02:24	00:02:22	00:01:55	00:01:48	00:01:54
40	00:05:03	00:03:29	00:02:25	00:03:16	00:03:15	00:02:31	00:02:26	00:02:00	00:01:46	00:01:57
41	00:04:20	00:03:07	00:02:27	00:03:10	00:03:15	00:02:28	00:02:21	00:02:02	00:01:45	00:01:57
42	00:04:58	00:02:48	00:02:26	00:03:16	00:03:10	00:02:30	00:02:25	00:01:57	00:01:46	00:01:58
43	00:04:47	00:02:31	00:02:53	00:03:23	00:03:23	00:02:22	00:02:22	00:01:58	00:01:44	00:01:57
44	00:04:36	00:03:25	00:04:47	00:04:16	00:03:03	00:02:21	00:02:18	00:01:56	00:01:42	00:01:55
45	00:04:58	00:04:06	00:06:04	00:04:13	00:03:25	00:02:23	00:02:21	00:01:55	00:01:44	00:01:54
46	00:04:47	00:03:00	00:05:46	00:04:11	00:03:09	00:02:23	00:02:26	00:02:01	00:01:46	00:01:56
47	00:04:03	00:03:05	00:04:19	00:04:05	00:03:04	00:02:22	00:02:20	00:01:57	00:01:48	00:01:56
48	00:04:27	00:03:04	00:04:31	00:04:23	00:03:04	00:02:29	00:02:22	00:02:10	00:01:46	00:01:55
49	00:04:19	00:03:24	00:05:08	00:04:08	00:03:16	00:02:21	00:02:20	00:01:57	00:01:42	00:01:55
50	00:04:35	00:04:08	00:04:44	00:04:06	00:03:15	00:02:17	00:02:25	00:01:55	00:01:44	00:01:56
51	00:04:09	00:04:22	00:05:00	00:04:45	00:03:20	00:02:18	00:02:30	00:01:56	00:01:45	00:02:01
52	00:05:03	00:04:12	00:04:54	00:05:20	00:03:44	00:02:18	00:02:26	00:01:52	00:01:43	00:01:57

Бачимо, що середньодобова тривалість виконання перегонів у середньому за рік найвища для перегону 1 (00:04:22), а найнижча – для перегону 9 (00:01:44), що вказує на залежність пасажирських перевезень у середньому за розмірами міста від трафіку та типу перегону, бо перегін 1 – це перегін у середмісті з високою ймовірністю

виникнення заторів, а перегін 9 – це перегін на відокремленій саме для цього громадського транспорту лінії ізольованого типу. Усереднена для всього маршруту тривалість перегону для кожного тижня вказує на наявність сезонних залежностей кількості транспорту на дорогах, а також на вплив змін погоди на тривалість перегонів.

За результатами аналізу пасажирських перевезень у середньому за розміром місті виявлено, що середньодобова тривалість виконання кожного перегону для кожного тижня зростає з початком осінньо-зимового періоду й досягає максимуму на 52 тиждень року (00:03:21), менші від середніх значень середньодобової тривалості виконання кожного перегону спостерігаються у весняний період, у цьому разі мінімальне значення (00:02:22) припадає на 18 тиждень року (кінець квітня – початок травня).

Кластерний аналіз такого масштабного неоднорідного набору даних щодо тривалості виконання перегонів електротранспортом у межах середнього за розміром міста з розвинутою мережею громадського транспорту проводився методом кластеризації K -середніх з огляду на його особливості обов'язково точного розподілу даних між кластерами.

Зазначимо, що існує кілька варіантів підбору оптимального значення кількості кластерів k , серед яких найчастіше застосовують метод ліктя, метод силуету та індексу Калінського – Харабаша. Метод ліктя розглядає суб'єктивно зрозумілі графіки характеру зміни розкиду точок ($W_{total} \rightarrow \max$) від найбільшого значення для всіх точок в одному кластері до найменшого значення ($W_{total} \rightarrow 0$) зі збільшенням кількості груп k ($k \rightarrow n$).

Метод силуету вимірює, наскільки схожі точки в одному кластері порівняно з іншими кластерами.

Значення індексу силуету перебуває в діапазоні $[-1,1]$, де більші значення вказують на кращу якість кластеризації. Цей метод оцінює, як добре точки розташовані всередині своїх кластерів порівняно з іншими кластерами. Більше значення коефіцієнта силуету вказує на кращу якість кластеризації.

Індекс Калінського – Харабаша, відомий ще під назвою "критерій співвідношення дисперсії", передбачає визначення співвідношення міжкластерного поділу до внутрішньо кластерної дисперсії, нормалізованого на їх число ступенів свободи. Найвище значення індексу Калінського – Харабаша вказує на те, що кластери визначено найбільш чітко. Хоча ця метрика найкраще підходить для розрахунку значення кількості кластерів, але має той самий недолік, що й коефіцієнт силуету – завищує оцінку для опуклих форм кластерів і занижує для складних форм кластерів.

З метою пошуку оптимальної кількості кластерів для набору даних із середньодобовими тривалостями виконання транспортними засобами кожного з перегонів протягом тижня на маршруті в середньому за розміром місті (табл. 2) було використано методи ліктя, силуету та Калінського – Харабаша. Результати оцінювання коефіцієнта загальної варіації точок всередині кластера щодо центра кластера SSE за методом ліктя подано на рис. 1. Оптимальне значення кількості кластерів становить $k = 5$ зі значенням (рис. 1).

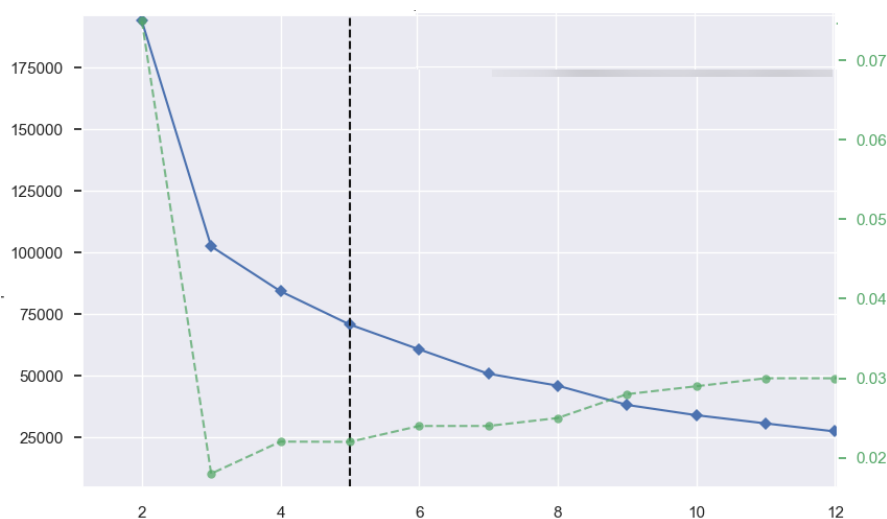


Рис. 1. Залежність величини SSE (*Sum of Squared Errors*) від кількості кластерів k , розрахована за методом ліктя, де синім позначено графік оцінки $SSE(k)$ – ліва вісь, зеленим позначено очікуваний час навчання – права вісь (розрахунок авторів)

Результати оцінювання коефіцієнта силуету S_i за методом силуету наведено на рис. 2. У нашому випадку максимальне значення коефіцієнта силуету $S_i = 0,507$ настає за умови $k = 2$, яке й вважається оптимальним значенням числа кластерів у процесі кластеризації (рис. 2). На рис. 3 наведено результати оцінювання індексу Калінського – Харабаша та відповідних до них значень числа кластерів. У нашому випадку максимальне значення індексу Калінського – Харабаша $S = 56.186$ настає за умови $k = 3$ (рис. 3), що вказує на оптимальне значення кількості кластерів для кластеризації даних.

У процесі кластеризації середньодобової тривалості виконання перегонів для кожного тижня за методом К-середніх було взято до уваги отримані результати розрахунку кількості кластерів k за допомогою методів ліктя, силуету та Калінського – Харабаша, відповідно, $k = 5$, $k = 2$ і $k = 3$ (рис. 1–3). На рис. 4 подано розподіл даних (середньодобових значень тривалості виконання кожного перегону для кожного тижня протягом року) на кластери, отриманий для $k = 2$ (а); $k = 3$ (б) і $k = 5$ (в).



Рис. 2. Залежність коефіцієнта силуету від величини кількості кластерів k , де синім позначено графік оцінки $S_i(k)$ – ліва вісь, зеленим позначено очікуваний час навчання – права вісь (розрахунок авторів)

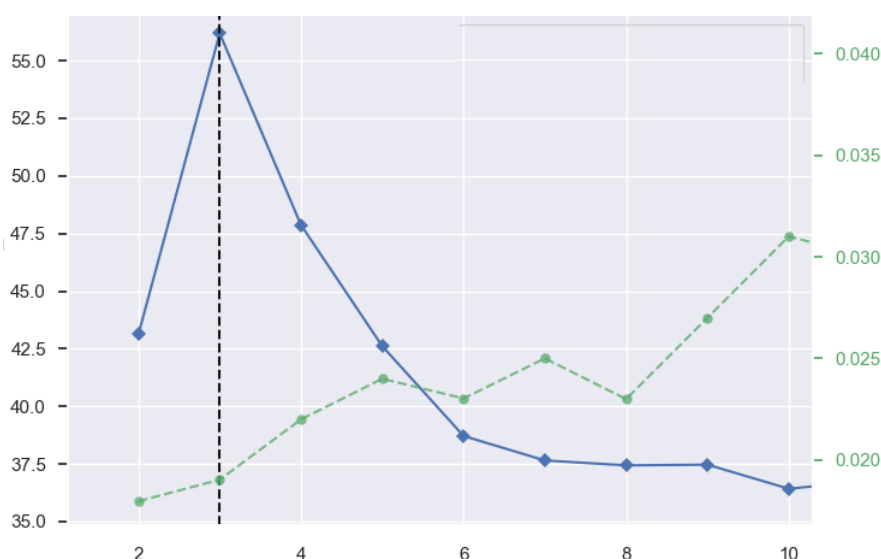


Рис. 3. Залежність індексу Калінського – Харабаша від величини кількості кластерів k , де синім позначено графік оцінки $S(k)$ – ліва вісь, зеленим позначено очікуваний час навчання – права вісь (розрахунок авторів)

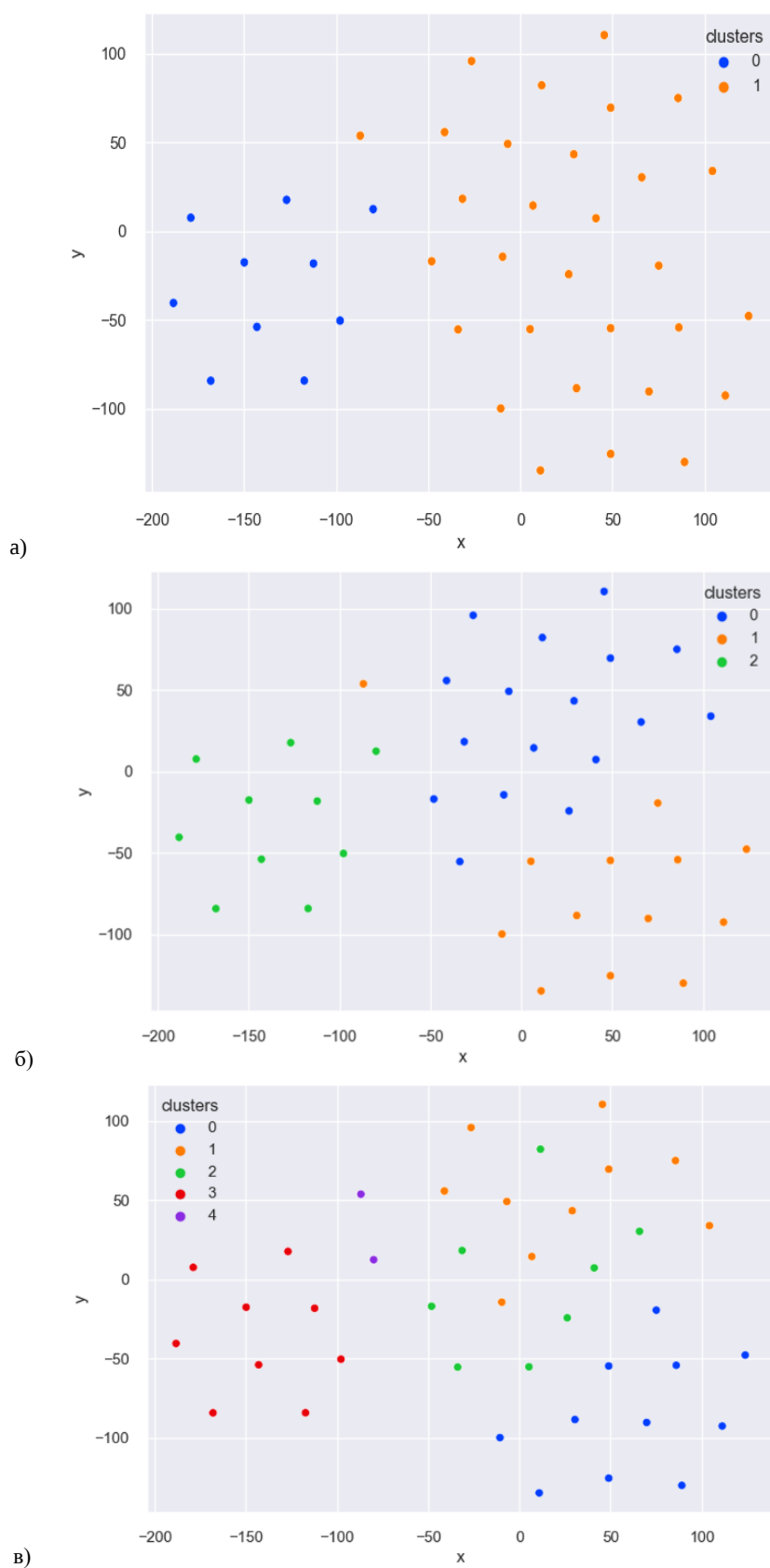


Рис. 4. Результати кластеризації середньодобових значень виконання графіків пасажирських перевезень для кожного тижня впродовж року на кожній ділянці (перегоні), а саме: а – кластеризація набору даних для $k = 2$; б – кластеризація набору даних для $k = 3$; в – кластеризація набору даних для $k = 5$ (розрахунок авторів)

Виконаємо детальний аналіз розподілу даних на кластери. У процесі поділу на два кластери, де значення $k=2$ отримано за методом силуету, маємо кластери з номерами 0 і 1 (відповідно до рис. 4, а). Перший кластер під номером 0 утворюють дані виконання графіків пасажирських перевезень на кожному перегоні для тижнів 9–20 і 22–43 із середньодобовими значеннями, близькими до середнього чи меншими від нього (табл. 2 і 3). Другий кластер під номером 1 утворюють дані виконання графіків пасажирських перевезень на кожному перегоні для тижнів 21 і 44–52 із середньодобовими значеннями, значно вищими від середнього (табл. 2 і 3) як мінімум для двох перегонів. Цей кластер характеризується також наявністю тижнів (21, 45, 46, 50–52) із значним перевищенням (у 1,5–2 рази) середньодобових значень виконання графіків пасажирських перевезень на трьох і більше перегонах. Значна кількість таких перевищень припадає на осінньо-зимовий період, що зумовлено складними погодними умовами.

Унаслідок поділу на три кластери, де значення $k=3$ отримано за методом Калінського – Харабаша, маємо кластери з номерами 0, 1 і 2 (див. рис. 4, б). Перший кластер під номером 0 утворюють дані виконання графіків пасажирських перевезень для тижнів 21 і 44–52 із середньодобовими значеннями, значно вищими від середньорічного на кожному перегоні (табл. 2 і 3) переважно для трьох і більше перегонів. Цей кластер характеризується також наявністю тижнів (21, 44–46, 48, 50–52) із значним перевищенням середньодобових значень виконання графіків пасажирських перевезень на трьох і більше перегонах, що зумовлено складними погодними умовами в осінньо-зимовий період. Це вказує на залежність тривалості виконання перегонів від сезонності. Другий кластер під номером 1 утворюють дані виконання графіків пасажирських перевезень тільки для тижнів 9–20 і 22 із середньодобовими значеннями, меншими від середньорічного на кожному перегоні (табл. 2 і 3). Третій кластер під номером 2 утворюють дані виконання графіків пасажирських перевезень для тижнів 27–43 із середньодобовими значеннями, близькими до середньорічного на кожному перегоні, до того ж незначні перевищення середньорічної величини спостерігаються не більше як на двох перегонах протягом тижня (табл. 2 і 3).

Таблиця 3. Поділ на кластери за методом К-середніх середньодобової тривалості виконання кожного перегону транспортними засобами для кожного тижня (розрахунок авторів)

Тижень \ К-сть кластерів k	2	3	5
9	0	1	0
10	0	1	0
11	0	1	0
12	0	1	0
13	0	1	0
14	0	1	0
15	0	1	0
16	0	1	2
17	0	1	0
18	0	1	0
19	0	1	0
20	0	1	0
21	1	0	4
22	0	1	4
27	0	2	1
28	0	2	1
29	0	2	1
30	0	2	1
31	0	2	1
32	0	2	1
33	0	2	1
34	0	2	1
35	0	2	1
36	0	2	2
37	0	2	2
38	0	2	2
39	0	2	1
40	0	2	2
41	0	2	2
42	0	2	2
43	0	2	2
44	1	0	3
45	1	0	3
46	1	0	3
47	1	0	3
48	1	0	3
49	1	0	3
50	1	0	3
51	1	0	3
52	1	0	3

У процесі поділу на п'ять кластерів ($k=5$), значення якого отримано за методом ліктя (рис. 1), маємо кластери з номерами 0; 1; 2; 3 і 4, наведені на рис. 4, в. Перший кластер під номером 0 утворюють дані виконання графіків пасажирських перевезень на кожному перегоні для тижнів 9–15 і 17–20 із середньодобовими значеннями, меншими від середньорічного (табл. 2 і 3). Другий кластер під номером 1 утворюють дані виконання графіків пасажирських перевезень на кожному перегоні для

тижнів 27–35 і 39 із середньодобовими значеннями, вищими від середньорічного не більше як для двох перегонів (табл. 2 і 3). Третій кластер під номером 2 утворюють дані виконання графіків пасажирських перевезень для тижнів 16, 36–38 і 40–43 із середньодобовими значеннями, близькими до середньорічного майже на кожному перегоні (табл. 2 і 3), до того ж перевищення середньорічного значення спостерігається не більше як для одного перегону. Перевищення спостерігаються тільки для перегону в центрі міста, що вказує на залежність тривалості виконання перегонів від їх розташування в конкретних міських зонах. Четвертий кластер під номером 3 утворюють дані виконання графіків пасажирських перевезень для тижнів 44–52 із середньодобовими значеннями, значно вищими від середньорічного переважно на трьох і більше перегонах (табл. 2 і 3), що зумовлено наявністю сезонності досліджуваної залежності тривалості виконання перегонів. Цей кластер також характеризується наявністю тижнів 45 і 52 із значним перевищенням середньодобових значень виконання графіків пасажирських перевезень на п'ятих перегонах, що може вказувати як на високий вплив трафіку разом із сезонністю. П'ятий кластер під номером 4 утворюють дані виконання графіків пасажирських перевезень тільки для тижнів 21 і 22 у літній період із середньодобовими значеннями, значно вищими від середньорічного на трьох і більше перегонах (табл. 2 і 3). Цей кластер вказує тільки на високий вплив трафіку на середньодобові значення виконання графіків пасажирських перевезень на перегонах у центрі міста. Необхідно звернути увагу, що жодних перевищень середньорічних значень виконання графіків не спостерігалось для перегонів 6–10, розташованих на ізольованій лінії, що виділена тільки для цього виду електротранспорту. Це вказує на оптимальний спосіб розв'язання проблем із пасажирськими перевезеннями громадським транспортом, однак він є складним, бо потребує значних інвестицій в інфраструктуру міста та є тривалим у реалізації.

Отже, кластерний аналіз великомасштабного неоднорідного набору даних щодо тривалості виконання перегонів електротранспортом у межах середнього за розмірами міста виявив окремі ділянки (перегони, зони) маршруту з різною інтенсивністю, на які значно впливають трафік, їх розташування в конкретних міських зонах (середмістя, спальний район тощо), а також сезонність.

Незважаючи на суб'єктивність визначення оптимального значення кількості кластерів за методом ліктя, бачимо, що поділ середньодобової тривалості виконання перегонів для кожного тижня на кластери дав найкращі результати для $k=5$, де значення оцінки внутрішньокластерної загальної варіації точок всередині кластера щодо центра кластера $SSE = 70896.042$ (рис. 1). Важливо звернути також увагу, що за умови $k=5$ значення коефіцієнта силуету $S_i = 0,378$ і значення індексу Калінського – Харабаша $S = 42.5086$ не є набагато меншими від максимальних значень відповідно до коефіцієнта силуету (рис. 2) та індексу Калінського – Харабаша (рис. 3).

Отже, можна стверджувати, що запропонований алгоритм вибору методу обчислення кількості кластерів на основі внутрішніх метрик оцінювання якості кластеризації даних, зібраних з інфраструктури пасажирських перевезень громадським транспортом у середньому за розміром міста, є достатньо простим та ефективним. До метрик кластеризації були додані метод ліктя, метод силуету та індекс Калінського – Харабаша, що дають змогу швидко й просто підібрати оптимальне значення кількості кластерів, а також зважати на особливості даних. Метод ліктя дає змогу брати до уваги внутрішньокластерну загальну варіацію точок усередині кластера щодо його центра, метод силуету вимірює, наскільки схожі точки в одному кластері порівняно з іншими кластерами, а найвище значенням індексу Калінського – Харабаша вказує на те, що кластери визначено найбільш чітко.

Алгоритм вибору оптимальної кількості кластерів можна узагальнити такими етапами:

- 1) вибір методу кластеризації, наприклад, K -середніх;
- 2) визначення діапазону кількості кластерів k , наприклад, від 2 до 10;
- 3) ітераційне застосування методу кластеризації для різних значень k ;
- 4) розрахунок результатів кластеризації та внутрішніх метрик оцінювання якості кластеризації;
- 5) аналіз метрик та вибір оптимальної кількості кластерів, що забезпечує баланс між якістю кластеризації (за метриками) та інтерпретованістю результатів з погляду предметної галузі (пасажирських перевезень);
- 6) візуальне оцінювання кластеризації для різних k з метою кращого розуміння структури даних.

Отже, виявлені методом кластеризації *K*-середніх перегони з високим перевищенням середньодобових значень тривалості виконання, порівняно із середньорічними, також вказують на зростання часу очікування пасажирів на зупинках, що позначається на кількості перевезених пасажирів та на якості наданих послуг. Це вказує на потребу прийняття обґрунтованих рішень у сфері пасажирських перевезень громадським транспортом у місті з метою його оптимізації.

Метод кластеризації *K*-середніх в аналізі середньодобової тривалості виконання кожного перегону транспортними засобами для кожного тижня протягом досліджуваного періоду рекомендуємо застосовувати для прийняття обґрунтованих рішень у сфері пасажирських перевезень громадським транспортом у розумному місті, а саме для оптимізації маршрутів, адаптації самої транспортної мережі, прогнозуванні та плануванні попиту на транспортні послуги, впровадженні персоналізованих послуг, а також в інтегруванні різних видів транспорту для створення єдиної ефективної багатомодальної транспортної системи.

Так, у процесі аналізу пасажиропотоків за допомогою методу кластеризації *K*-середніх виявлені зони високого попиту допоможуть створити оптимальні транспортні маршрути, які відповідають реальним потребам пасажирів у конкретний момент часу, скорочуючи час очікування та кількість пересадок до мінімально можливих. Цей метод кластеризації *K*-середніх також буде ефективний в аналізі змін потреб пасажирів і сприятиме адаптації маршрутів громадського транспорту до змін у попиті, наприклад, додавання нових зупинок, зміна розкладу руху транспортних засобів та їх графіків. Це також дасть змогу керівникам на рівні міської влади краще планувати інфраструктурні проєкти та інвестиції в модернізацію транспортної системи з метою інтегрування різних видів транспорту (електротранспорт, звичайні автобуси, метро, якщо є) для створення єдиної ефективної транспортної системи. У розумному місті також є важливою персоналізація послуг, де основну роль відіграють мобільні застосунки для громадського транспорту, які для надання персоналізованих рекомендацій пасажирам про вибір оптимального маршруту чи часу поїздки використовуватимуть результати кластеризації тривалості виконання графіків перегонів у реальному часі. До основних проблем, які доцільно розв'язувати

із застосування кластеризації великих даних, є виокремлення кластерів пасажирів за типами (робітники, студенти, туристи тощо), визначення гарячих точок (районів із найвищим попитом на транспорт у певний час), виявлення неефективних маршрутів чи низького завантаження окремих ділянок транспортної мережі, аналіз залежності пасажиропотоків від зовнішніх чинників (погоди, подій у місті, соціальних тенденцій), а також побудова динамічних моделей прогнозування змін у потоках.

Отже, досягнуті результати кластерного аналізу середньодобової тривалості виконання кожного перегону транспортними засобами для кожного тижня протягом досліджуваного періоду мають практичну цінність в оптимізації маршрутів, адаптації самої транспортної мережі, прогнозуванні та плануванні попиту на транспортні послуги, впровадженні персоналізованих послуг, а також в інтегруванні різних видів транспорту для створення єдиної ефективної багатомодальної транспортної системи. Рекомендовано застосовувати кластеризацію для оптимізації маршрутів, а саме для створення оптимальних транспортних маршрутів, що мають скорочений час очікування та меншу кількість пересадок, а також відповідають реальним потребам пасажирів у визначений ними момент часу.

Висновки й перспективи подальших досліджень

З метою вивчення можливостей застосування методів кластеризації в організації пасажирських перевезень у розумному місті проаналізовано особливості їх застосування для покращення організації пасажирських перевезень громадським транспортом. Це дало змогу встановити, що вибір методу кластеризації залежить від специфіки поставленого завдання, характеристик даних і цілей, що висуваються для аналізу транспортних і пасажирських потоків. Так, висхідні методи кластеризації застосовують, коли потрібно визначити структуру маршрутів, об'єднати схожі ділянки маршрутів або прилеглих до маршруту територій. Низхідні методи впроваджують для виокремлення маршрутів або їх ділянок (перегонів, зон) із різною інтенсивністю пасажиропотоку з метою подальшої оптимізації послуги в конкретних зонах. Метод роздільної кластеризації *K*-середніх ефективний для зонування транспортної мережі, наприклад, під час оптимізації маршрутів за тривалістю виконання

перегонів, за зонами попиту тощо. Метод роздільної кластеризації *K*-медоїд є більш стійким до шуму, тому його використовують для оптимізації робочого графіку транспортних засобів, визначення оптимального розташування зупинок на маршрутах, щоб найкраще покрити потреби пасажирів. Метод нечіткої кластеризації *C*-середніх застосовують у розробленні більш гнучких та адаптивних стратегій управління пасажиропотоками, що важливо в сучасних міських середовищах з динамічною природою попиту на транспортні послуги. Метод DBSCAN, що класифікує елементи на основі щільності, найчастіше впроваджують для аналізу пасажиропотоків з різноманітною густиною на різних маршрутах та для кластеризації з огляду на часову динаміку. Метод самоорганізованих карт використовують у процесі кластеризації для створення динамічних карт пасажиропотоків, зважаючи на часові інтервали й густину на різних маршрутах.

Унаслідок проведення кластерного аналізу пасажирських перевезень у середньому за розміром місті з розвиненою мережею громадського транспорту виявлено, що зібрані дані про тривалість виконання кожного перегону транспортними засобами в межах кожної доби для кожного тижня протягом досліджуваного періоду мають складну, неоднорідну й масштабну структуру, тому потребують відповідного оброблення перед початком аналізу. Кластерний аналіз такого масштабного неоднорідного набору даних щодо тривалості виконання перегонів електротранспортом у межах середнього за розміром міста проводиться методом кластеризації *K*-середніх, оскільки цей метод, зменшуючи значення функції втрат, реалізує обов'язково точний розподіл даних, де кожен об'єкт належить лише одному класу.

Запропоновано простий та найбільш ефективний алгоритм вибору методу обчислення кількості кластерів на основі внутрішніх метрик оцінювання якості кластеризації даних, зібраних з інфраструктури пасажирських перевезень громадським транспортом у середньому за розміром місті. До метрик кластеризації додано метод ліктя, метод силуету та індекс Калінського – Харабаша, що дають змогу

швидко й просто підібрати оптимальне значення кількості кластерів. Метод ліктя допомагає встановити внутрішньокластерну загальну варіацію точок всередині кластера щодо його центра, метод силуету вимірює, наскільки схожі точки в одному кластері порівняно з іншими кластерами, а найвище значенням індексу Калінського – Харабаша вказує на те, що кластери визначено найбільш чітко.

Досягнуті результати кластерного аналізу середньодобової тривалості виконання кожного перегону транспортними засобами для кожного тижня протягом досліджуваного періоду мають практичну цінність в оптимізації маршрутів, адаптації самої транспортної мережі, прогнозуванні та плануванні попиту на транспортні послуги, впровадженні персоналізованих послуг, а також в інтегруванні різних видів транспорту для створення єдиної ефективної багатомодальної транспортної системи. Було рекомендовано застосовувати кластеризацію для оптимізації маршрутів, а саме для створення оптимальних транспортних маршрутів, що мають скорочений час очікування та меншу кількість пересадок, а також відповідають реальним потребам пасажирів у визначений ними момент часу.

Отже, метод кластеризації *K*-середніх в аналізі середньодобової тривалості виконання кожного перегону транспортними засобами для кожного тижня протягом досліджуваного періоду доцільно застосовувати для оптимізації організації пасажирських перевезень громадським транспортом у розумному місті.

Перспективою подальших досліджень є застосування кластеризації великих даних з метою виокремлення кластерів пасажирів за типами (робітники, студенти, туристи тощо), визначення гарячих точок (районів із найвищим попитом на транспорт у певний час), виявлення неефективних маршрутів чи низького завантаження окремих ділянок транспортної мережі, аналіз залежності пасажиропотоків від зовнішніх чинників (погоди, подій у місті, соціальних тенденцій), а також побудова динамічних моделей прогнозування змін у потоках.

Список літератури

1. Saxena A., Prasad M., Gupta A., Bharill N., Patel O. P., Tiwari A., Er M. J., Ding W., Lin C. A review of clustering techniques and developments. *Neurocomputing*. 2017. No. 267. P. 664–681. DOI: 10.1016/j.neucom.2017.06.053
2. Isoli N., Chaczykowski M. Net energy analysis and net carbon benefits of CO₂ capture and transport infrastructure for energy applications and industrial clusters. *Applied Energy*. 2025. No. 382, 125227 p. DOI: 10.1016/j.apenergy.2024.125227

3. Kowalska-Styczeń A., Bublyk M., Lytvyn V. Green innovative economy remodeling based on economic complexity. *Journal of Open Innovation: Technology, Market, and Complexity*. 2023. V.9 (3), 100091 p. DOI: 10.1016/j.joitmc.2023.100091
4. Podlesna L., Bublyk M., Grybyk I., Matseliukh Y., Burov Y., Kravets P., Lozynska O., Karpov I., Peleshchak I., Peleshchak R. Optimization model of the buses number on the route based on queueing theory in a Smart City. *CEUR Workshop Proceedings*. 2020. Vol. 2631. P. 502–515. URL: <https://ceur-ws.org/Vol-2631/paper37.pdf> (дата звернення: 01.02.2025).
5. Bianchini D., De Antonellis V., Garda M. A big data exploration approach to exploit in-vehicle data for smart road maintenance. *Future Generation Computer Systems*. 2023. No. 149. P. 701–716. DOI: 10.1016/j.future.2023.08.004
6. Katrenko A., Krislata I. Veres O., Oborska O., Basyuk T., Vasyliuk A., Rishnyak I., Demyanovskyi N., Meh O. Development of traffic flows and smart parking system for smart city. *CEUR Workshop Proceedings*. 2020. Vol. 2604. P. 730–745. URL: <https://ceur-ws.org/Vol-2604/paper50.pdf> (дата звернення: 01.02.2025).
7. Matseliukh Y., Bublyk M., Vysotska V. Development of intelligent system for visual passenger flows simulation of public transport in Smart City based on neural network. *CEUR Workshop Proceedings*. 2021. Vol. 2870. P. 1087–1138. URL: <https://ceur-ws.org/Vol-2870/paper82.pdf> (дата звернення: 01.02.2025).
8. Visan M., Negrea S. L., Mone F. Towards intelligent public transport systems in Smart Cities; Collaborative decisions to be made. *Procedia Computer Science*. 2021. No. 199. P. 1221–1228. DOI: 10.1016/j.procs.2022.01.155
9. Ezugwu A. E., Ikotun A. M., Oyelade O. O., Abualigah L., Agushaka J. O., Eke C. I., Akinyelu A. A. A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. *Engineering Applications of Artificial Intelligence*. 2022. No. 110, 104743 p. DOI: 10.1016/j.engappai.2022.104743
10. Chavent M., Lechevallier Y., Briant O. DIVCLUS-T: A monothetic divisive hierarchical clustering method. *Computational Statistics & Data Analysis*. 2007. No. 52(2). P. 687–701. DOI: 10.1016/j.csda.2007.03.013
11. Celebi M. E., Kingravi H. A., Vela P. A. A comparative study of efficient initialization methods for the k-means clustering algorithm. *Expert Systems With Applications*. 2012. No. 40(1). P. 200–210. DOI: 10.1016/j.eswa.2012.07.021
12. Bublyk M., Kowalska-Styczeń A., Lytvyn V., Vysotska V. The Ukrainian economy transformation into the circular based on fuzzy-logic cluster analysis. *Energies*. 2021. No. 14(18). 5951 p. DOI: 10.3390/en14185951
13. Singh J., Singh D. A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects. *Advanced Engineering Informatics*. 2024. No. 62, 102799 p. DOI: 10.1016/j.aei.2024.102799
14. Yan J., Liu J., Tseng F. An evaluation system based on the self-organizing system framework of smart cities: A case study of smart transportation systems in China. *Technological Forecasting and Social Change*. 2020. No. 153, 119371 p. DOI: 10.1016/j.techfore.2018.07.009
15. Matseliukh Y., Vysotska V., Bublyk M. Intelligent system of visual simulation of passenger flows. *CEUR Workshop Proceedings*. 2020. Vol. 2604. P. 906–920. URL: <https://ceur-ws.org/Vol-2604/paper60.pdf> (дата звернення: 01.02.2025).
16. Prasetyo E. A., Novizayanti D., Putri A. N. A. Cluster analysis of potential autonomous vehicle (AV) adopters in Indonesia's new capital. *Transportation Research Interdisciplinary Perspectives*. 2024. No. 29, 101318 p. DOI: 10.1016/j.trip.2024.101318
17. Singh J., Singh D. A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects. *Advanced Engineering Informatics*. 2024. No. 62, 102799 p. DOI: 10.1016/j.aei.2024.102799
18. Liu J., Li J., Chen Y., Lian S., Zeng J., Geng M., Zheng S., Dong Y., He Y., Huang P., Zhao Z., Yan X., Hu Q., Wang L., Yang D., Zhu Z., Sun Y., Shang W., Wang D., Chen X. Multi-scale urban passenger transportation CO2 emission calculation platform for smart mobility management. *Applied Energy*. 2023. No. 331, 120407 p. DOI: 10.1016/j.apenergy.2022.120407
19. Matseliukh Y., Bublyk M., Bosak A., Naychuk-Khrushch M. The role of public transport network optimization in reducing carbon emissions. *CEUR Workshop Proceedings*. 2024. Vol. 3723. P. 340–364. URL: <https://ceur-ws.org/Vol-3723/paper19.pdf>
20. Bezdek J. C. *Pattern Recognition with Fuzzy Objective Function Algorithms*. Bon, Springer, 1981, 245 p.
21. Bublyk M., Vysotska V., Matseliukh Y., Mayik V., Nashkerska M. Assessing Losses of Human Capital Due to Man-Made Pollution Caused by Emergencies. *CEUR Workshop Proceedings*. 2020. Vol. 2805. P. 74–86. URL: <https://ceur-ws.org/Vol-2805/paper6.pdf> (дата звернення: 01.02.2025).
22. Бублик М.І., Бабій Т.І. Розвиток логістики в сучасних умовах функціонування ринку. *Науковий вісник НЛТУ України*. 2009. No. 19(6). С. 138–142.
23. Ester M., Kriegl H.-P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. 1996. P. 226–231.
24. Jain A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*. 2010. No. 31(8). P. 651–666. DOI: 10.1016/j.patrec.2009.09.011
25. Kohonen T. *Self-Organizing Maps* (3rd ed.). Bon, Springer, 2001. 554 p.
26. Koshtura D., Bublyk M., Matseliukh Y., Dosyn D., Chyrun L., Lozynska O., Karpov I., Peleshchak I., Maslak M., Sachenko O. Analysis of the demand for bicycle use in a smart city based on machine learning. *CEUR workshop proceedings*. 2020. Vol. 2631, P. 172–183. URL: <https://ceur-ws.org/Vol-2631/paper13.pdf> (дата звернення: 01.02.2025).

27. Литвин В., Бублик М., Висоцька В., Мацелюх Ю. Технологія візуальної симуляції пасажиропотоків у сфері громадського транспорту Smart City. *Радіоелектроніка, інформатика, управління*. 2022. №4. С. 106–121. DOI: 10.15588/1607-3274-2021-4-10
28. Law A. M. *Simulation Modeling and Analysis* (5th ed.). Bon, McGraw-Hill, 2015. 495 p.
29. Nath N., Nitanai R., Manabe R., Murayama A. A global-scale review of smart city practice and research focusing on residential neighbourhoods. *Habitat International*. 2023. Vol. 142, P. 102963 p. DOI: 10.1016/j.habitatint.2023.102963
30. Sun L., Zhao J., Zhang J., Zhang F., Ye, K., Xu C. Activity-based individual travel regularity exploring with entropy-space K-means clustering using smart card data. *Physica A: Statistical Mechanics and its Applications*. 2024. No. 636, 129522 p. DOI: 10.1016/j.physa.2024.129522
31. Tibshirani R., Walther G., Hastie T. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 2001. No. 63(2). P. 411–423.
32. Subudhi S., Panigrahi S. Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection. *Journal of King Saud University – Computer and Information Sciences*. 2020. No. 32(5). P. 568–575. DOI: 10.1016/j.jksuci.2017.09.010
33. Duan Y., Liu C., Li S., Guo X., Yang C. An automatic affinity propagation clustering based on improved equilibrium optimizer and t-SNE for high-dimensional data. *Information Sciences*, 2023.No. 623. P. 434–454. DOI: 10.1016/j.ins.2022.12.057
34. Bide P., Shedje R. Improved document clustering using K-means algorithm. *2015 IEEE Int. Conf. on Electrical, Computer and Communication Technologies (ICECCT)*. New York: IEEE, 2015. 1048 p.
35. Zhong C., Miao D., Wang R., Zhou X. DIVFRP: An automatic divisive hierarchical clustering method based on the furthest reference points. *Pattern Recognition Letters*. 2008. No. 29(16). P. 2067–2077. DOI: 10.1016/j.patrec.2008.07.002
36. Elassy M., Al-Hattab M., Takruri M., Badawi S. Intelligent transportation systems for sustainable smart cities. *Transportation Engineering*. 2024. No. 16, 100252 p. DOI: 10.1016/j.treng.2024.100252
37. Zhu W. A spatial decision-making model of smart transportation and urban planning based on coupling principle and Internet of Things. *Computers and Electrical Engineering*. 2022. No. 102, 108222 p. DOI: 10.1016/j.compeleceng.2022.108222
38. Bushuev S., Inna L., Alla B., Alexander L., Khusainova M. Creating Urban Transportation Networks Grounded In the Principles of the Smart Port-City Paradigm. *Procedia Computer Science*. 2023. No. 231. P. 323–328. DOI: 10.1016/j.procs.2023.12.211
39. Balbin P. P., Barker J. C., Leung C. K., Tran M., Wall R. P., Cuzzocrea A. Predictive analytics on open big data for supporting smart transportation services. *Procedia Computer Science*. 2019. No. 176. P. 3009–3018. DOI: 10.1016/j.procs.2020.09.202
40. Khemakhem S., Krichen L. A comprehensive survey on an IoT-based smart public street lighting system application for smart cities. *Franklin Open*. 2024. No. 8, 100142 p. DOI: 10.1016/j.fraope.2024.100142
41. Vidović K., Čolić P., Vojvodić S., Blavicki A. Methodology for public transport mode detection using telecom big data sets: Case study in Croatia. *Transportation Research Procedia*. 2021. No. 64. P. 76–83. DOI: 10.1016/j.trpro.2022.09.010
42. Bublyk M., Udovychenko T., Medvid R. Concept of smart specialization in the context of the development of Ukraines economy. *Economics. Ecology. Socium*. 2019. Vol. 3. No 2. P. 55–61. DOI: 10.31520/2616-7107/2019.3.2-6
43. Bublyk M., Matseliukh Y. Small-batteries utilization analysis based on mathematical statistics methods in challenges of circular economy. *CEUR workshop proceedings*. 2021. Vol. 2870, P. 1594–1603. URL: <https://ceur-ws.org/Vol-2870/paper118.pdf> (last accessed: 01.02.2025).
44. Бублик М.І. Модель економічного оцінювання техногенних збитків в національному господарстві. *Black Sea Scientific Journal of Academic Research. Economic Science*. 2014. Vol. 12. No. 05, С. 44–50.
45. Dai Y., Hasanefendic S., Bossink B. A systematic literature review of the smart city transformation process: The role and interaction of stakeholders and technology. *Sustainable Cities and Society*. 2024. Vol. 101, 105112 p. DOI: 10.1016/j.scs.2023.105112
46. Cai J., Luo J., Wang S., Yang S. Feature selection in machine learning: A new perspective. *Neurocomputing*. 2018. No.300. P. 70–79. DOI: 10.1016/j.neucom.2017.11.077
47. Chung S. Applications of smart technologies in logistics and transport: A review. *Transportation Research Part E: Logistics and Transportation Review*. 2021. Vol. 153, 102455 p. DOI: 10.1016/j.tre.2021.102455
48. European Commission. *Sustainable Urban Mobility Package*. URL: <https://ec.europa.eu/transport/themes/urban/> (last accessed: 01.02.2025).
49. Kidmose B. A review of smart vehicles in smart cities: Dangers, impacts, and the threat landscape. *Vehicular Communications*. 2025. No. 51, 100871 p. DOI: 10.1016/j.vehcom.2024.100871
50. Sood S. K. A scientometric analysis of quantum driven innovations in intelligent transportation systems. *Engineering Applications of Artificial Intelligence*. 2024. No. 138, 109258 p. DOI: 10.1016/j.engappai.2024.109258

References

1. Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., Er, M. J., Ding, W., Lin, C. (2017), "A review of clustering techniques and developments", *Neurocomputing*, No. 267, P. 664–681. DOI: <https://doi.org/10.1016/j.neucom.2017.06.053>

2. Isoli, N., Chaczykowski, M. (2025), "Net energy analysis and net carbon benefits of CO₂ capture and transport infrastructure for energy applications and industrial clusters", *Applied Energy*, No. 382, 125227 p. DOI: <https://doi.org/10.1016/j.apenergy.2024.125227>
3. Kowalska-Styczeń, A., Bublyk, M., Lytvyn, V. (2023), "Green innovative economy remodeling based on economic complexity", *Journal of Open Innovation: Technology, Market, and Complexity*. No.9(3), 100091 p. DOI: <https://doi.org/10.1016/j.joitmc.2023.100091>
4. Podlesna, L., Bublyk, M., Grybyk, I., Matseliukh, Y., Burov, Y., Kravets, P., Lozynska, O., Karpov, I., Peleshchak, I., Peleshchak, R. (2020), "Optimization model of the buses number on the route based on queueing theory in a Smart City", *CEUR Workshop Proceedings*, Vol. 2631, P. 502–515, available at: <https://ceur-ws.org/Vol-2631/paper37.pdf> (last accessed: 01.02.2025).
5. Bianchini, D., De Antonellis, V., Garda, M. (2023), "A big data exploration approach to exploit in-vehicle data for smart road maintenance", *Future Generation Computer Systems*, No. 149, P. 701–716. DOI: <https://doi.org/10.1016/j.future.2023.08.004>
6. Katrenko, A., Krislata, I. Veres, O., Oborska, O., Basyuk, T., Vasyliuk, A., Rishnyak, I., Demyanovskyi, N., Meh, O. (2020), "Development of traffic flows and smart parking system for smart city", *CEUR Workshop Proceedings*, Vol. 2604, P. 730–745, available at: <https://ceur-ws.org/Vol-2604/paper50.pdf> (last accessed: 01.02.2025).
7. Matseliukh, Y., Bublyk, M., Vysotska, V. (2021), "Development of intelligent system for visual passenger flows simulation of public transport in Smart City based on neural network", *CEUR Workshop Proceedings*, Vol. 2870, P. 1087–1138, available at: <https://ceur-ws.org/Vol-2870/paper82.pdf> (last accessed: 01.02.2025).
8. Visan, M., Negrea, S. L., Mone, F. (2021), "Towards intelligent public transport systems in Smart Cities; Collaborative decisions to be made", *Procedia Computer Science*, No. 199, P. 1221–1228. DOI: <https://doi.org/10.1016/j.procs.2022.01.155>
9. Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., Akinyelu, A. A. (2022), "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects", *Engineering Applications of Artificial Intelligence*, No. 110, 104743 p. DOI: <https://doi.org/10.1016/j.engappai.2022.104743>
10. Chavent, M., Lechevallier, Y., Briant, O. (2007), "DIVCLUS-T: A monothetic divisive hierarchical clustering method", *Computational Statistics & Data Analysis*, No. 52 (2), P. 687–701. DOI: <https://doi.org/10.1016/j.csda.2007.03.013>
11. Celebi, M. E., Kingravi, H. A., Vela, P. A. (2012), "A comparative study of efficient initialization methods for the k-means clustering algorithm", *Expert Systems With Applications*, No. 40 (1), P. 200–210. DOI: <https://doi.org/10.1016/j.eswa.2012.07.021>
12. Bublyk, M., Kowalska-Styczeń, A., Lytvyn, V., Vysotska, V. (2021), "The Ukrainian economy transformation into the circular based on fuzzy-logic cluster analysis", *Energies*, No. 14 (18), 5951 p. DOI: <https://doi.org/10.3390/en14185951>
13. Singh, J., Singh, D. (2024), "A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects", *Advanced Engineering Informatics*, No. 62, 102799 p. DOI: <https://doi.org/10.1016/j.aei.2024.102799>
14. Yan, J., Liu, J., Tseng, F. (2020), "An evaluation system based on the self-organizing system framework of smart cities: A case study of smart transportation systems in China", *Technological Forecasting and Social Change*, No. 153, 119371 p. DOI: <https://doi.org/10.1016/j.techfore.2018.07.009>
15. Matseliukh, Y., Vysotska, V., Bublyk, M. (2020), "Intelligent system of visual simulation of passenger flows", *CEUR Workshop Proceedings*, Vol. 2604, P. 906–920, available at: <https://ceur-ws.org/Vol-2604/paper60.pdf> (last accessed: 01.02.2025).
16. Prasetyo, E. A., Novizayanti, D., Putri, A. N. A. (2024), "Cluster analysis of potential autonomous vehicle (AV) adopters in Indonesia's new capital", *Transportation Research Interdisciplinary Perspectives*, No. 29, 101318 p. DOI: <https://doi.org/10.1016/j.trip.2024.101318>
17. Singh, J., Singh, D. (2024), "A comprehensive review of clustering techniques in artificial intelligence for knowledge discovery: Taxonomy, challenges, applications and future prospects", *Advanced Engineering Informatics*, No. 62, 102799 p. DOI: <https://doi.org/10.1016/j.aei.2024.102799>
18. Liu, J., Li, J., Chen, Y., Lian, S., Zeng, J., Geng, M., Zheng, S., Dong, Y., He, Y., Huang, P., Zhao, Z., Yan, X., Hu, Q., Wang, L., Yang, D., Zhu, Z., Sun, Y., Shang, W., Wang, D., Chen, X. (2023), "Multi-scale urban passenger transportation CO₂ emission calculation platform for smart mobility management", *Applied Energy*, No. 331, 120407 p. DOI: <https://doi.org/10.1016/j.apenergy.2022.120407>
19. Matseliukh, Y., Bublyk, M., Bosak, A., Naychuk-Khrushch, M. (2024), "The role of public transport network optimization in reducing carbon emissions", *CEUR Workshop Proceedings*, Vol. 3723, P. 340–364, available at: <https://ceur-ws.org/Vol-3723/paper19.pdf> (last accessed: 01.02.2025).
20. Bezdek, J. C. (1981), *Pattern Recognition with Fuzzy Objective Function Algorithms*, Bon, Springer, 245 p.
21. Bublyk, M., Vysotska, V., Matseliukh, Y., Mayik, V., Nashkierska, M. (2020), "Assessing losses of human capital due to man-made pollution caused by emergencies", *CEUR Workshop Proceedings*, Vol. 2805, P. 74–86, available at: <https://ceur-ws.org/Vol-2805/paper6.pdf> (last accessed: 01.02.2025).
22. Bublyk, M.I., Babiy, T.I. (2009), "Development of logistic in modern operating of market conditions", *Scientific Bulletin of UNFU*, No. 19 (6), P. 138–142.
23. Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996), "A density-based algorithm for discovering clusters in large spatial databases with noise", *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, P. 226–231.

24. Jain, A. K. (2010), "Data clustering: 50 years beyond K-means", *Pattern Recognition Letters*, No. 31 (8), P. 651–666. DOI: <https://doi.org/10.1016/j.patrec.2009.09.011>
25. Kohonen, T. (2001), *Self-Organizing Maps*, 3rd ed., Springer, Bon, 554 p.
26. Koshtura, D., Bublyk, M., Matseliukh, Y., Dosyn, D., Chyrun, L., Lozynska, O., Karpov, I., Peleshchak, I., Maslak, M., Sachenko, O. (2020), "Analysis of the demand for bicycle use in a smart city based on machine learning", *CEUR workshop proceedings*, Vol. 2631, P. 172–183, available at: <https://ceur-ws.org/Vol-2631/paper13.pdf> (last accessed: 01.02.2025).
27. Lytvyn, V., Bublyk, M., Vysotska, V., Matseliukh, Y. (2022), "Visual simulation technology for passenger flows in the public transport field at smart city", *Radio Electronics, Computer Science, Control*, No. 4, P. 106–121. DOI: <https://doi.org/10.15588/1607-3274-2021-4-10>
28. Law, A. M. (2015), *Simulation Modeling and Analysis*. 5th ed., McGraw-Hill, Bon, 495 p.
29. Nath, N., Nitani, R., Manabe, R., Murayama, A. (2023), "A global-scale review of smart city practice and research focusing on residential neighbourhoods", *Habitat International*, Vol. 142, 102963 p. DOI: <https://doi.org/10.1016/j.habitatint.2023.102963>
30. Sun, L., Zhao, J., Zhang, J., Zhang, F., Ye, K., Xu, C. (2024), "Activity-based individual travel regularity exploring with entropy-space K-means clustering using smart card data", *Physica A: Statistical Mechanics and its Applications*, No. 636, 129522 p. DOI: <https://doi.org/10.1016/j.physa.2024.129522>
31. Tibshirani, R., Walther, G., Hastie, T. (2001), "Estimating the number of clusters in a data set via the gap statistic", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, No. 63 (2), P. 411–423.
32. Subudhi, S., Panigrahi, S. (2020), "Use of optimized Fuzzy C-Means clustering and supervised classifiers for automobile insurance fraud detection", *Journal of King Saud University – Computer and Information Sciences*, No. 32 (5), P. 568–575. DOI: <https://doi.org/10.1016/j.jksuci.2017.09.010>
33. Duan, Y., Liu, C., Li, S., Guo, X., Yang, C. (2023), "An automatic affinity propagation clustering based on improved equilibrium optimizer and t-SNE for high-dimensional data", *Information Sciences*, No. 623, P. 434–454. DOI: <https://doi.org/10.1016/j.ins.2022.12.057>
34. Bide, P., Shedje, R. (2015), "Improved document clustering using K-means algorithm", *2015 IEEE Int. Conf. on Electrical, Computer and Communication Technologies (ICECCT)*, IEEE, New York, 1048 p.
35. Zhong, C., Miao, D., Wang, R., Zhou, X. (2008), "DIVFRP: An automatic divisive hierarchical clustering method based on the furthest reference points", *Pattern Recognition Letters*, No. 29 (16), P. 2067–2077. DOI: <https://doi.org/10.1016/j.patrec.2008.07.002>
36. Elassy, M., Al-Hattab, M., Takruri, M., Badawi, S. (2024), "Intelligent transportation systems for sustainable smart cities", *Transportation Engineering*, No. 16, 100252 p. DOI: <https://doi.org/10.1016/j.treng.2024.100252>
37. Zhu, W. (2022), "A spatial decision-making model of smart transportation and urban planning based on coupling principle and Internet of Things", *Computers and Electrical Engineering*, No. 102, 108222 p. DOI: <https://doi.org/10.1016/j.compeleceng.2022.108222>
38. Bushuev, S., Inna, L., Alla, B., Alexander, L., Khusainova, M. (2023), "Creating urban transportation networks grounded in the principles of the smart port-city paradigm", *Procedia Computer Science*, No. 231, P. 323–328. DOI: <https://doi.org/10.1016/j.procs.2023.12.211>
39. Balbin, P. P., Barker, J. C., Leung, C. K., Tran, M., Wall, R. P., Cuzzocrea, A. (2019), "Predictive analytics on open big data for supporting smart transportation services", *Procedia Computer Science*, No. 176, P. 3009–3018. DOI: <https://doi.org/10.1016/j.procs.2020.09.202>
40. Khemakhem, S., Krichen, L. (2024), "A comprehensive survey on an IoT-based smart public street lighting system application for smart cities". *Franklin Open*, No. 8, 100142 p. DOI: <https://doi.org/10.1016/j.fraope.2024.100142>
41. Vidović, K., Čolić, P., Vojvodić, S., Blavicki, A. (2021), "Methodology for public transport mode detection using telecom big data sets: Case study in Croatia", *Transportation Research Procedia*, No. 64, P. 76–83. DOI: <https://doi.org/10.1016/j.trpro.2022.09.010>
42. Bublyk, M., Udovychenko, T., Medvid, R. (2019), Concept of smart specialization in the context of the development of Ukraine's economy", *Economics. Ecology. Socium*, Vol. 3, No. 2, P. 55–61. DOI: <https://doi.org/10.31520/2616-7107/2019.3.2-6>
43. Bublyk, M., Matseliukh, Y. (2021), "Small-batteries utilization analysis based on mathematical statistics methods in challenges of circular economy", *CEUR workshop proceedings*, Vol. 2870, P. 1594–1603, available at: <https://ceur-ws.org/Vol-2870/paper118.pdf> (last accessed: 01.02.2025).
44. Bublyk, M.I. (2014), "Model of economic evaluation of man-made damage to the national economy", *Black Sea Scientific Journal of Academic Research. Economic Science*, No. 12 (05), P. 44–50.
45. Dai, Y., Hasanefendic, S., Bossink, B. (2024), "A systematic literature review of the smart city transformation process: The role and interaction of stakeholders and technology", *Sustainable Cities and Society*, Vol. 101, 105112 p. DOI: <https://doi.org/10.1016/j.scs.2023.105112>
46. Cai, J., Luo, J., Wang, S., Yang, S. (2018), "Feature selection in machine learning: A new perspective", *Neurocomputing*, No. 300, P. 70–79. DOI: <https://doi.org/10.1016/j.neucom.2017.11.077>
47. Chung, S. (2021), "Applications of smart technologies in logistics and transport: A review", *Transportation Research Part E: Logistics and Transportation Review*, No. 153, 102455 p. DOI: <https://doi.org/10.1016/j.tre.2021.102455>
48. European Commission, "Sustainable Urban Mobility Package", available at: <https://ec.europa.eu/transport/themes/urban> (last accessed: 01.02.2025).

49. Kidmose, B. (2025), "A review of smart vehicles in smart cities: Dangers, impacts, and the threat landscape. *Vehicular Communications*, 51, 100871 p. DOI: <https://doi.org/10.1016/j.vehcom.2024.100871>
50. Sood, S. K. (2024), "A scientometric analysis of quantum driven innovations in intelligent transportation systems", *Engineering Applications of Artificial Intelligence*, No. 138, 109258 p. DOI: <https://doi.org/10.1016/j.engappai.2024.109258>

Надійшла (Received) 10.02.2025

Відомості про авторів / About the Authors

Мацелюх Юрій Романович – Національний університет "Львівська політехніка", аспірант кафедри інформаційних систем та мереж, Львів, Україна; e-mail: indeed.post@gmail.com; ORCID ID: <https://orcid.org/0000-0002-1721-7703>

Литвин Василь Володимирович – доктор технічних наук, професор, Національний університет "Львівська політехніка", завідувач кафедри інформаційних систем та мереж, Львів, Україна; e-mail: vasyl.v.lytvyn@lpnu.ua; ORCID ID: <https://orcid.org/0000-0002-9676-0180>

Matseliukh Yurii – Lviv Polytechnic National University, PhD Student of Information Systems and Networks Department, Lviv, Ukraine.

Lytvyn Vasyl – Doctor of Sciences (Engineering), Professor, Lviv Polytechnic National University, Head of Information Systems and Networks Department, Lviv, Ukraine.

APPLICATION OF THE K-MEANS CLUSTERING METHOD IN THE ORGANISATION OF PASSENGER TRANSPORT IN A SMART CITY

Every year, big data clustering methods are gaining popularity for decision-making on organizing passenger transportation in a smart city, ensuring the efficiency, adaptability and environmental friendliness of the transport system. Their relevance is due to the growth of data volumes, changing demand and the negative impact of transport on the environment. The object of the study is the process of clustering data sets on organizing passenger transportation. The subject of the study is the principles of studying clustering metrics when calculating the number of clusters for executing transportation schedules. The purpose of the study is to apply the K-means algorithm based on quality metrics for clustering data on organizing passenger transportation in a smart city. The following tasks are solved in the article: study of the features of clustering methods and their metrics; analysis of a large-scale heterogeneous data set on the duration of electric transport trips in an average-sized city; development of an effective algorithm for choosing a method for calculating the number of clusters based on metrics for assessing the quality of data clustering. The methods of analysis, synthesis, generalization, comparison, grouping, cluster analysis, system analysis, K-means method were used. The following results were obtained: It was established that the choice of the clustering method depends on the specifics of the task, data characteristics and the objectives of the analysis of transport flows. A complex, heterogeneous and raw data structure was revealed regarding the duration of electric transport journeys. Cluster analysis using the K-means method was due to the need for accurate data distribution between clusters. An algorithm for choosing a method for calculating the number of clusters based on metrics for assessing the quality of data clustering, including the elbow method, the silhouette method and the Kalinsky-Kharabash index, was proposed. It is recommended to use clustering to create routes with reduced waiting time, fewer transfers and compliance with passenger needs. Conclusions: the K-means method was used to analyze the duration of electric transport journeys. Data analysis revealed sections of routes with different traffic flow intensity, which depends on their location in urban areas, seasonality, etc. An algorithm for choosing a method for calculating the number of clusters based on internal metrics is proposed.

Keywords: passenger transport; smart city; clustering analysis; K-means method; systems analysis.

Бібліографічні описи / Bibliographic descriptions

Мацелюх Ю. Р., Литвин В. В. Метод К-середніх для аналізу даних з організації пасажирських перевезень у розумному місті. *Сучасний стан наукових досліджень та технологій в промисловості*. 2025. № 1 (31). С. 83–101. DOI: <https://doi.org/10.30837/2522-9818.2025.1.083>

Matseliukh, Y., Lytvyn, V. (2025), "Application of the K-means clustering method in the organisation of passenger transport in a smart city", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (31), P. 83–101. DOI: <https://doi.org/10.30837/2522-9818.2025.1.083>