Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження методів кластеризації геоданих
з використанням дискретизації
(тема)

Виконав:
студент 2 курсу, групи СШІм-22-1
Котелевець К.А.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва спеціалізації)
Керівник доц. Чала Л.Е.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Котелевцю Кирилу Андрійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження методів кластеризації геоданих з використанням дискретизації

затверджена наказом університету від 1 квітня 20 24 р. № 260Ст
2. Термін подання студентом роботи до екзаменаційної комісії 12 червня 20 24 р.
3. Вихідні дані до роботи Література про методи дискретизації, методи дискретизації на основі систем географічної індексації BingTiles, S2, H3. Література про методи кластеризації, методи K-Means, спектральної кластеризації, DBSCAN, OPTICS.

4. Перелік питань, що потрібно опрацювати в роботі
1) Аналіз предметної області та постановка задачі, вибір і обгрунтування методів дискретизації та кластеризації для геоданих
2) Проведення експериментальних практичних досліджень в контексті локацій IoT пристроїв міста Києва.

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	01.04.2024	виконано
2	Аналіз предметної галузі та постановка задачі	01.04.2024 – 12.04.2024	виконано
3	Аналіз методів вирішення поставленої задачі	12.04.2024 – 27.04.2024	виконано
4	Вибір методу та вирішення задачі	27.04.2024 – 12.05.2024	виконано
5	Розробка базового прототипу дослідження	12.05.2024 – 30.05.2024	виконано
6	Оформлення пояснювальної записки	30.05.2024 – 12.06.2024	виконано
7	Захист кваліфікаційної роботи	12.06.2024	виконано

Дата видачі завдання 1 квітня 20 24 р.

Студент _____
(підпис)

Керівник роботи _____ доц. Чала Л.Е.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 90 с., 58 рис., 2 дод., 20 джерел.

АНАЛІЗ ДАНИХ, ГЕОДАНІ, ДИСКРЕТИЗАЦІЯ, КАРТОГРАФІЯ, КЛАСТЕРИЗАЦІЯ, МАШИННЕ НАВЧАННЯ, ПРОЕКЦІЯ, DBSCAN, EPSG, IOT, K-MEANS, OPTICS, OSM, WGS.

Об'єктом дослідження є порівняння ефективності методів кластеризації K-Means, спектральної кластеризації, DBSCAN та OPTICS для роботи з географічними даними та аналіз покращення результатів за допомогою використання дискретизації.

Предмет дослідження – результати аналізу методів кластеризації геоданих з використанням дискретизації.

Мета роботи полягає в аналізі та порівнянні ефективності різних методів кластеризації з використанням дискретизації для аналізу і отримання прихованої інформації з географічних даних. Результати можуть бути використані для вдосконалення методів обробки географічних даних різного типу і природи та для підтримки прийняття рішень у різних галузях, таких як міське планування, транспортна логістика, аналіз місцевого розвитку тощо.

Методи дослідження – аналіз літератури, інтернет джерел, технічної документації та інструментарію для роботи з географічними даними.

У ході виконання даної роботи був проведений аналіз та порівняння методів K-Means, спектральної кластеризації, DBSCAN та OPTICS, що використовуються для кластеризації геоданих.

ABSTRACT

Master's thesis contains: 87 p., 58 fig., 2 ann., 20 references.

DATA ANALYSIS, CLUSTERING, DBSCAN, DISCRETIZATION, EPSG, GEODATA, IOT, K-MEANS, MACHINE LEARNING, MAPPING, OPTICS, OSM, PROJECTION, WGS.

The object of research is to icompare the effectiveness of K-Means, spectral clustering, DBSCAN and OPTICS clustering methods for working with geographic data and to analyze the improvement of results by using discretization techniques.

The subject of the study is the results of the analysis of geodata clustering methods using discretization.

The purpose of the study is to analyze and compare the effectiveness of different clustering methods using discretization to analyze and extract hidden information from geographic data. The results can be used to improve methods for processing geographic data of various types and nature and to support decision-making in various fields, such as urban planning, transport logistics, local development analysis, etc.

Research methods: analysis of literature, Internet sources, technical documentation and tools for working with geographic data.

In the course of this work, we analyzed and compared K-Means, spectral clustering, DBSCAN and OPTICS methods used for clustering geodata.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

DBSCAN – Density-Based Spatial Clustering of Applications with Noise – алгоритм кластеризації в просторі даних, який базується на їх густині;

EPSG – European Petroleum Survey Group – організація, що спеціалізується на стандартизації геодезичних параметрів і систем географічних координат;

GPS – Global Positioning System – система глобального позиціонування на поверхні Землі;

IoT – Internet of Things – Інтернет речей;

OPTICS – Ordering Points To Identify the Clustering Structure – алгоритм пошуку кластерів на основі щільності в просторових даних;

OSM – Open Street Map – вільна та відкрита карта світу, яка створюється, підтримується та оновлюється волонтерами по всьому світу;

UTC – Coordinated Universal Time – всесвітній координований час;

UUID – Universally Unique Identifier – універсально унікальний ідентифікатор;

WSG – World System of Geocentric Coordinates – система геоцентричних координат, яка використовується для визначення точок на поверхні Землі або в космічних дослідженнях.

ВСТУП

У сучасному цифровому світі об’єм географічних даних стрімко збільшується, що ставить перед науковими дослідниками та практиками складне завдання забезпечення ефективної обробки та аналізу великих масивів інформації. Один із ключових етапів у цьому процесі – кластеризація геоданих, яка полягає у групуванні схожих об’єктів з метою подальшого аналізу. Завдяки застосуванню методів кластеризації, вдається виявляти різноманітні зв’язки та закономірності у географічних даних, що забезпечує підставу для ухвалення обґрунтованих рішень у різних галузях, починаючи від геології й закінчуючи маркетингом.

Проте, обробка великих масивів геоданих може виявитися важкою задачею через їхню величезну розмірність та складність структури. У таких умовах на передній план виходить необхідність розробки та вдосконалення методів кластеризації, які були б не лише ефективними, а й масштабованими для роботи з великими обсягами геоданих.

Один із потенційних шляхів вдосконалення методів кластеризації геоданих використання дискретизації – перетворення неперервних даних у дискретні форми, що може допомогти зменшити обсяг інформації та спростити її подальшу обробку, при цьому зберігаючи важливість висновків та результатів аналізу.

Об’єктом дослідження даної кваліфікаційної роботи є аналіз та порівняння методів кластеризації геоданих на предметній галузі даних GPS сигналів з використанням дискретизації для покращення аналітичної цінності отриманих результатів. Мета даної роботи полягає у порівнянні ефективності алгоритмів кластеризації з використанням дискретизації та без, що відкриває нові можливості для підвищення якості та точності аналізу географічних даних.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Геодані у сучасному світі

У сучасному світі геодані виступають як ключовий компонент як у різноманітних аспектах нашого повсякденного життя так і для суспільства в цілому. Постійний прогрес технологій та надзвичайно швидке зростання обсягу географічної інформації підкреслюють їх важливість для прийняття обґрунтованих рішень у різних сферах, починаючи від наукових досліджень і закінчуючи бізнесом та забезпеченням національної безпеки.

Протягом останніх десятиліть геодані набули, без перебільшення, шаленого попиту, що особливо помітно у контексті стрімкого розвитку технологій та значного зростання загального обсягу географічної інформації. Цей вибух популярності зумовлений рядом ключових факторів, серед яких варто відзначити доступність новітніх інструментів і технологій для їх збору, обробки та аналізу. Широке поширення портативних і мобільних пристроїв, вдосконалення супутникових систем позиціонування та розвиток сучасних геоінформаційних систем роблять цей тип даних більш доступним і зрозумілим для звичайних користувачів.

З кожним роком збільшується кількість джерел, звідки можна отримати геодані, та розмаїття їх форматів і типів. Це говорить про те, що доступ до географічної інформації стає все більш широким і різноманітним. Нові технології збирання даних, такі як дрони, супутникові системи та датчики на мобільних пристроях, розширюють можливості отримання геоданих у реальному часі та в різних масштабах. Цей постійний розвиток сприяє зростанню популяризації геоданих і розширює їхнє застосування в сучасному світі.

Використання геоданих на сьогоднішній день настільки широке, що їх можна розглядати як важливий каталізатор новаторських змін у найрізноманітніших галузях. Від найбільших корпорацій до маленьких

стартапів, від державних установ до громадських організацій – усі вони використовують геодані для отримання нових знань, виявлення закономірностей і прийняття стратегічних рішень.

Наукові дослідження, що базуються на аналізі геоданих, допомагають розкривати та розуміти складні географічні взаємозв'язки між об'єктами, а застосування геоданих у галузі бізнесу стає все більш невід'ємною складовою стратегічного планування, маркетингових досліджень та управління логістикою і ресурсами. У сфері громадської безпеки геодані використовуються для моніторингу природних катастроф, запобігання аварійних ситуацій та ефективного реагування на них, що дозволяє забезпечувати безпеку громадян та мінімізувати ризики виникнення негативних наслідків. Для галузей штучного інтелекту та машинного навчання, геодані використовуються для тренування найрізноманітніших алгоритмів, від розробки автономних систем і до досягнення медичних проривів у розумінні соціальних та екологічних явищ.

1.2 Концепція та представлення геоданих

Геодані представляють собою специфічний тип інформації, який визначається прив'язкою до географічних координат або розташуванням об'єктів у просторі. Крім географічних координат, геодані також включають атрибутивну інформацію про об'єкти, таку як їх характеристики, класифікації та описи. Ця додаткова інформація дозволяє більш детально вивчати та аналізувати різні аспекти географічних об'єктів.

Однією з ключових особливостей геоданих є їх прив'язка до конкретної географічної локації на поверхні Землі або у просторі. Ця особливість гарантує точність визначення локації об'єктів та їх просторових взаємозв'язків. Для забезпечення цієї прив'язки використовуються спеціальні системи проекцій, які перетворюють географічний простір на плоску поверхню карти. Ці системи проекцій є методами перетворення

тривимірного географічного простору на плоску поверхню карти. Серед найпоширеніших систем проекцій варто виділити географічні (геодезичні) та проекційні системи.

Географічні системи проекцій базуються на сферичній моделі Землі, де координати об'єктів виражаються у вигляді широти і довготи. З іншого боку, проекційні системи використовуються для перетворення географічного простору на плоску площину шляхом застосування різноманітних математичних формул і алгоритмів. Візуалізація перетворення координат у проекційній системі наведена на рисунку 1.1.

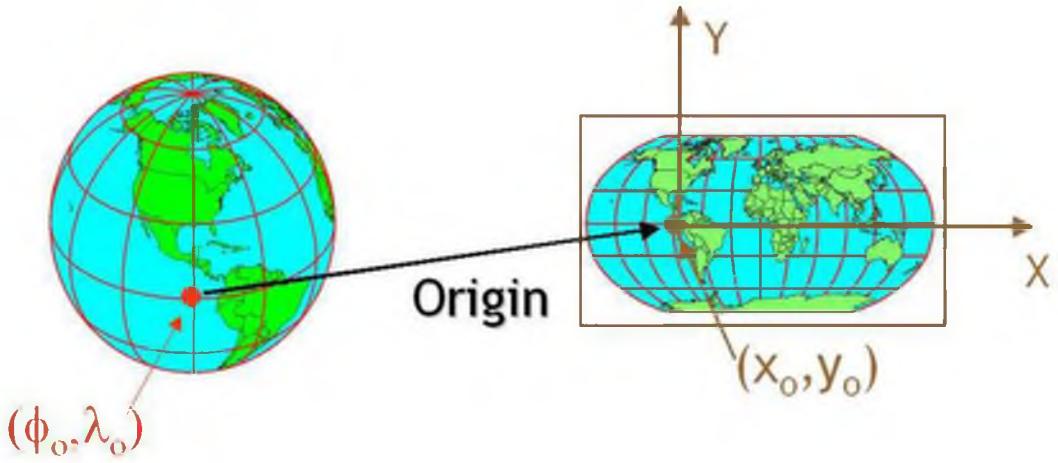

Рисунок 1.1 – Візуалізація проекції координат

EPSG (European Petroleum Survey Group) є організацією, яка спеціалізується на створенні та підтримці бази даних проекцій та систем координат. У їхньому кодовому реєстрі містяться ідентифікатори для різноманітних систем координат, які охоплюють як географічні, так і проекційні системи. Ці коди знаходять широке застосування у галузі геодезії, географічних інформаційних системах та картографії.

Коди EPSG використовуються для однозначної ідентифікації конкретних систем координат. Кожен код відповідає певній системі координат, включаючи проекційні параметри, такі як проекція, одиниці

вимірювання та інші параметри, необхідні для правильного відображення просторових даних на карті.

Застосування кодів EPSG забезпечує однорідність та стандартизацію в галузі геопросторових даних, дозволяючи різним системам та програмам обмінюватися і використовувати геодезичну інформацію без втрати точності або спотворень. Це робить їх важливим інструментом для розвитку та обміну географічних даних в усьому світі.

EPSG:4326, також відома як WGS 84 (World Geodetic System), є однією з найбільш поширених географічних систем координат. Ця система базується на використанні географічних координат у форматі широти і довготи. Діапазон значень широти знаходиться в межах від -90 до +90 градусів, а довготи – від -180 до +180 градусів. Ця система вважається загальним стандартом для представлення географічних даних у багатьох геоінформаційних системах та сервісах. Ця система широко застосовується через свою універсальність та можливість однозначної ідентифікації місцезнаходження об'єктів на поверхні Землі. На рисунку 1.2 наведено приклад відображення карти Землі з використанням EPSG:4326, де можна зауважити, що карта є відчутно приплюснутою на полюсах.

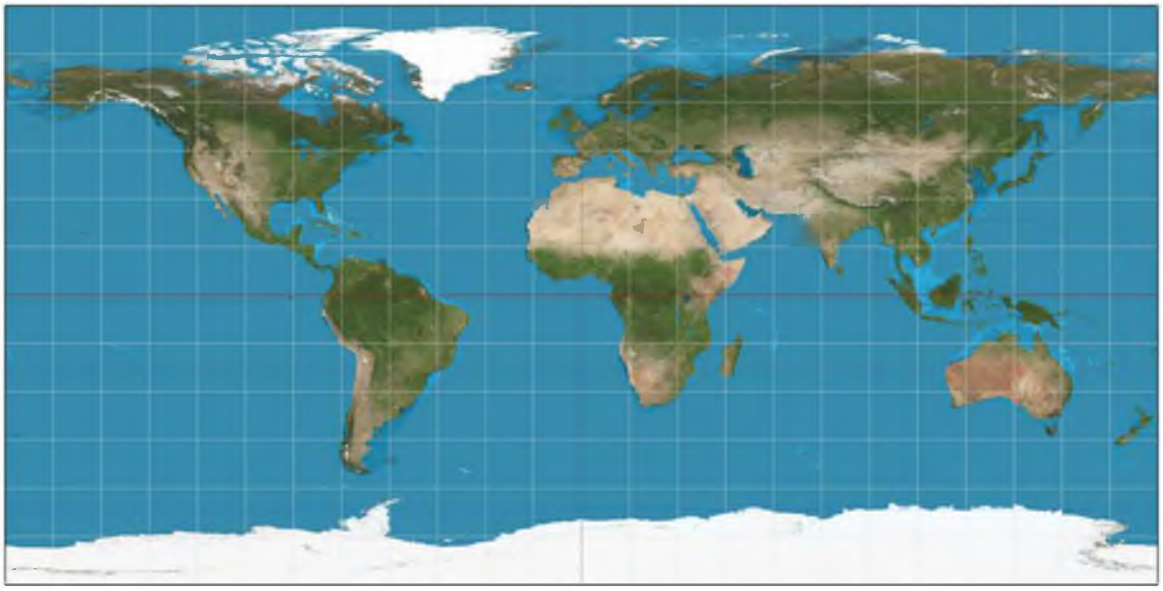

Рисунок 1.2 – Карта Землі з використанням EPSG:4326

Інша широко використовувана система проекції – EPSG:3857, відома як Web Mercator. Ця система стала стандартом для веб-карт та онлайн картографічних сервісів, таких як Google Maps та OpenStreetMap (OSM). EPSG:3857 використовує меркаторську проекцію для передачі географічних даних на плоску поверхню, що дозволяє швидко та ефективно відображати глобальні карти на веб-картах.

На рисунку 1.3 наведено відображення карти Землі з використанням проекції EPSG:3857. За рахунок проекції координат без урахування округлості земної кулі з'являється викривлення масштабів ближче до полюсів.

Рисунок 1.3 – Карта Землі з використанням EPSG:3857

1.3 Кластеризація геоданих

1.3.1 Визначення кластеризації геоданих

Кластеризація – це технологія машинного навчання та важливий інструмент аналізу даних, який використовується для групування схожих об’єктів разом у певну кількість класів або «кластерів». Основна мета кластеризації полягає у розподілі даних на окремі групи, де об’єкти в межах однієї групи максимально подібні між собою, але максимально відрізняються від об’єктів, що належать до інших груп.

Кластеризація спрямована на виявлення внутрішніх структур і закономірностей в наборі даних. Вона допомагає виявити приховані зв’язки і патерни, які можуть бути корисними для подальшого дослідження. Методи кластеризації можуть бути застосовані до різних типів даних, таких як числові, категоріальні і текстові дані, що робить їх універсальним інструментом для аналізу в різних сферах науки та бізнесу.

Основна мета кластеризації полягає у виявленні груп схожих об’єктів у великих наборах даних, що спрощує їхню інтерпретацію та допомагає приймати обґрунтовані рішення на основі отриманих результатів. Крім того, за допомогою кластерного аналізу можна виявляти аномалії та виділяти важливі підгрупи в даних, що дозволяє отримувати додаткові інсайти та використовувати їх у подальшому дослідженні або прийнятті рішень.

Аналіз геоданих з використанням кластеризації є важливим методом для виявлення структур та закономірностей у географічних даних за допомогою групування географічних об’єктів у кластери на основі їхньої просторової схожості та інших властивостей. Використання кластеризації геоданих може мати різноманітні цілі, включаючи аналіз розподілу об’єктів, виявлення взаємозв’язків та ідентифікацію аномалій. Цей дозволяє виявити складні структури у великих наборах геоданих, сприяючи розумінню просторових взаємозв’язків та розв’язанню різноманітних завдань у галузях геоінформатики, містобудування, екології тощо.

1.3.2 Сфери застосування кластеризації геоданих

Сьогодні кластеризація геоданих широко застосовується у різноманітних галузях. Вона використовується не лише в класичних географічних інформаційних системах і геології, а й у містобудуванні, екології, хімії, транспорті, бізнес-аналітиці, маркетингу, логістиці, медицині, сільському господарстві та у багатьох інших сферах діяльності [1].

У географічних інформаційних системах кластеризація геоданих допомагає зрозуміти нетривіальні просторові взаємозв'язки та виявити складні географічні патерни. Вона також допомагає з аналізом і подальшою інтерпретацією великих обсягів географічних даних, що сприяє більш ефективному використанню цієї інформації, оскільки людина краще сприймає великий обсяг даний саме візуально.

В інших галузях застосування кластеризації геоданих розширюється на досить широкий спектр прикладних завдань. У містобудуванні, наприклад, вона відчутно допомагає при плануванні розвитку територій через оптимізації розташування об'єктів інфраструктури та вивчення густоти населення і переміщення людей. В екологічних дослідженнях кластеризація геоданих використовується для оцінки впливу людської діяльності на навколишнє середовище та визначення зон екологічного ризику. У сфері бізнес-аналітики, цей метод дозволяє проводити сегментацію ринку, прогнозування попиту та виявлення нових неочевидних можливостей на конкретному ринку. Для сфери медицини кластеризація геоданих допомагає аналізувати розподіл захворювань та вибирати ефективні стратегії контролю їх поширення. І, нарешті, в аграрній науці та сільському господарстві цей метод може бути надзвичайно корисним для оптимізації обробки сільськогосподарських культур, аналізу ґрунтів та кліматичних умов для вибору оптимальних методів обробки землі і висадження культур [2].

1.3.3 Проблеми кластеризації геоданих

У галузі геоінформатики та суміжних галузях кластеризація геоданих відіграє критичну роль у виявленні структур та закономірностей, що існують у географічних даних. Незважаючи на її широкий спектр застосувань і численні переваги, здійснення такого аналізу вимагає вирішення ряду складних проблем. Кластеризація геоданих, хоча й має значний потенціал у геоінформатиці, стикається зі значними викликами, які потребують комплексного та системного підходу для їх вирішення.

Одна з основних проблем – проблема управління та аналізу геоданих, на сьогоднішній день виникає через їх значний обсяг. Зі збільшенням кількості та різноманітності доступних даних стає важко їх зберігати, обробляти та проводити аналіз. Через те що об'єм вибірки геоданих може бути величезним, це робить обробку та подальший аналіз надзвичайно дорогими як за ресурсами так і за часом.

Обсяг геоданих, що потребує обробки та аналізу, обумовлений їхньою великою різноманітністю та складністю. Різноманітність даних проявляється у гетерогенності їхніх форматів та структур, що є наслідком їх походження з різних джерел. Ці аспекти можуть призвести до неточностей під час використання методів кластеризації та ускладнити виявлення структур і закономірностей в даних.

Також використання методів кластеризації у контексті геоданих досить часто стикається з проблемою зашумленості та наявністю аномалій у зібраних даних. Це створює складнощі для досягнення надійних та точних результатів при аналізі географічної інформації. Зашумлені або аномальні дані можуть відображати несподівані або не репрезентативні патерни, що можуть призвести до неточностей та викривлення інтерпретації результатів кластеризації. Причини походження таких даних можуть бути досить різноманітними, включаючи помилки вимірювань, випадкові відхилення, а

також систематичні спотворення, що зустрічаються у реальних наборах географічних даних.

Для ефективного управління цією проблемою необхідно використовувати певні стратегії та методи. Серед них варто відзначити виявлення та вилучення аномалій перед проведенням процесу кластеризації. Це може бути досягнуто за допомогою встановлення порогових значень для визначення аномалій або використання спеціальних алгоритмів, спрямованих на виявлення віддалених точок даних або груп аномальних об'єктів.

Проте, важливо враховувати, що деякі аномалії можуть нести важливу інформацію або вказувати на унікальні особливості даних, тому їх вилучення має бути обґрунтованим і базуватися на уважному аналізі контексту. Також, для підвищення стійкості до зашумленості та аномалій важливо використовувати алгоритми, які є менш чутливими до впливу шуму, наприклад, DBSCAN або OPTICS [3].

Оскільки аналіз геоданих за допомогою використання методів кластеризації є складною задачею через різноманітність географічних факторів та їхню велику кількість, то це може призводити до проблем з інтерпретацією отриманих результатів.

В першу чергу виникають труднощі у встановленні відповідності між кластерами та конкретними географічними об'єктами або територіями. Значна кількість факторів, що впливають на розподіл об'єктів у просторі, може ускладнювати точне визначення їх географічних властивостей.

По-друге, для правильної інтерпретації результатів кластеризації необхідно враховувати контекстуальні чинники. Географічні дані включають в себе різноманітні характеристики, які можуть бути взаємопов'язані, і важливо враховувати цей контекст при аналізі та інтерпретації кластерів.

Крім того, складність візуалізації та обробки великої кількості геоданих може ускладнити їхнє сприйняття та аналіз. Часто результати

кластеризації представляються у вигляді складних географічних карт або великої кількості даних, що потребує додаткового часу та зусиль для їх інтерпретації.

Також не менш визначальну роль для кластерного аналізу даних має масштабованість та швидкодія методів, що відіграє вирішальне значення в аналізі великих та складних географічних наборів даних.

На сьогоднішній день є потреба в методах кластеризації, які можуть ефективно працювати з великими обсягами геоданих, враховуючи наведені вище особливості та проблеми, такі як обсяг, різноманітність та складність. Чим більше дані, тим більше часу потрібно на їх обробку та подальший аналіз, і саме тому методи кластеризації повинні бути здатними масштабуватись та працювати ефективно навіть при надзвичайно великих обсягах вхідних даних [4].

Отже, швидкодія грає ключову роль у реальних сценаріях застосування, де оперативний аналіз та реагування на дані є надзвичайно важливими. Наприклад, у галузі транспорту або надання екстреної медичної допомоги швидке прийняття рішень може врятувати життя. Через це методи кластеризації повинні бути достатньо швидкими для обробки геоданих у реальному часі або близько до цього [5].

1.4 Дискретизація геоданих

1.4.1 Проблеми кластеризації геоданих

Дискретизація даних – це важливий процес у сучасному цифровому світі, який використовується для перетворення неперервних або аналогових даних у дискретні форми. У геоінформатиці, де дані можуть бути у формі географічних карт, зображень або векторних об’єктів, дискретизація є ключовим елементом для зручного зберігання, обробки та подальшого аналізу геоданих.

Основна мета дискретизації полягає у розбитті неперервних даних на окремі об’єкти, які рівномірно розподілені в часі або просторі. Це дозволяє зберігати дані у формі дискретних значень, які вже можна ефективно обробляти за допомогою комп’ютерних систем. Приклад візуального порівняння функцій розподілу дискретних та неперервних даних наведено на рисунку 1.4.

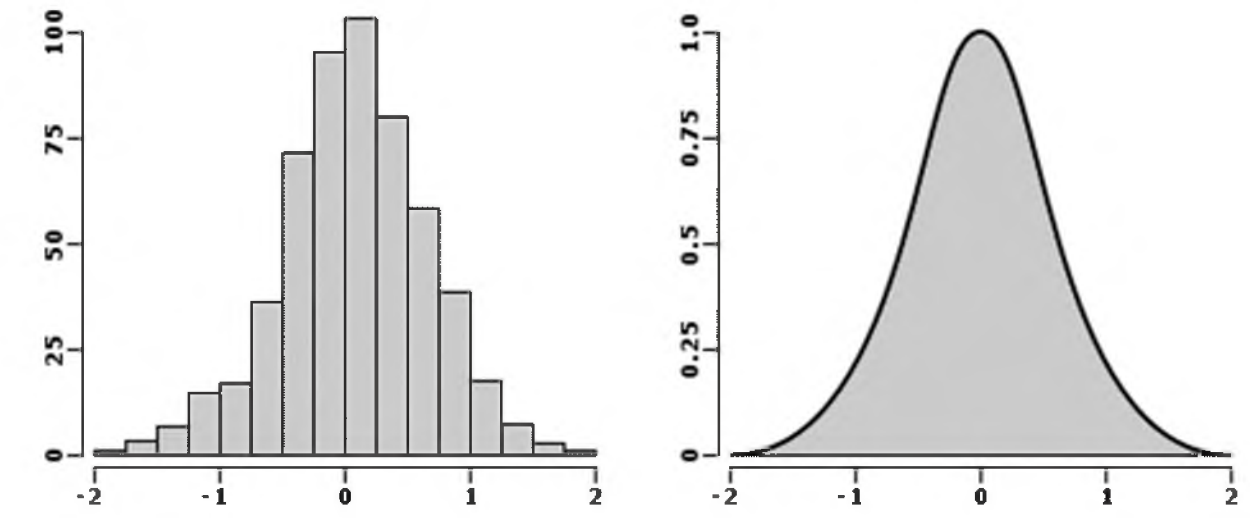

Рисунок 1.4 – Функції розподілу для дискретних та неперервних даних

Одним з найбільш важливих аспектів дискретизації є вибір оптимального співвідношення між деталізацією даних та обсягом втраченої інформації, особливо в контексті обробки географічних даних, зображень, сигналів тощо. Неправильно підібрані параметри можуть призвести або до втрати важливих деталей, або до перевищення обсягу обробки, що ускладнює аналіз та використання даних.

У геоінформатиці, де точність та ефективність аналізу геоданих є критичними, дискретизація є необхідним етапом, оскільки саме вона дозволяє ефективно зберігати та аналізувати геодані з високою точністю та дозволяє зменшити або повність анулювати вплив згаданих проблем на кластерний аналіз.

1.4.2 Вирішення проблем кластеризації дискретизацією

Застосування дискретизації географічних даних в контексті кластеризації є дієвим інструментом для подальшого вдосконалення процесу аналізу та отримання корисної інформації з різноманітних наборів даних. Цей підхід дозволяє розв'язати ряд ключових проблем, що виникають при застосуванні кластерного аналізу до геоданих.

Однією з основних задач у галузі геоінформатики для кластеризації геоданих є зменшення їх обсягу без суттєвої втрати інформативності, оскільки обробка великих обсягів даних може бути викликом з точки зору швидкодії та ефективності. Для вирішення цієї проблеми широко використовується метод дискретизації даних, що значно зменшує обсяг інформації для подальшого аналізу. Для геоданих, таких як картографічні зображення або супутникові знімки, дискретизація може включати перетворення неперервних значень висоти чи інтенсивності на дискретні категорії або рівні. Це сприяє значному зменшенню обсягу даних, необхідних для подальшого аналізу та дозволяє ефективно зберігати та обробляти геодані, відчутно зменшуючи загальні витрати часу та обчилювальних ресурсів.

Для кластерного аналізу геоданих, зменшення кількості точок шляхом дискретизації допомагає спростити процес групування об'єктів у групи, оскільки після застосування дискретизації кожен об'єкт представлений меншою кількістю точок або значень, що полегшує розрахунок відстаней між ними. Зменшення кількості точок дозволяє знизити обсяг обчислень та складність аналізу, що в свою чергу сприяє більш швидкому та ефективному проведенню кластеризації.

Додатково, важливою перевагою дискретизації є збереження суттєвої інформації при зменшенні обсягу даних. Завдяки правильно підібраним методам та параметрам дискретизації можна зберегти ключові аспекти геоданих, такі як рельєф місцевості, зміни в рослинності та інші ландшафтні

особливості, які є важливими для подальшого аналізу та прийняття необхідних рішень.

Не менш ключову роль для кластеризації геоданих відіграє боротьба з шумами та викидами, оскільки останні можуть значно спотворювати результати аналізу даних, відчутно погіршуючи їхню точність. Для географічних даних шуми є розповсюдженою проблемою, оскільки вони можуть виникати з різних джерел, таких як помилки вимірювань, перешкоди при передачі даних або випадкові аномалії.

Одним із найбільш ефективних методів виявлення та усунення шумів для геоданих є використання дискретизації. Оскільки під час застосування дискретизації неперервні або аналогові дані перетворюються на дискретні форми, це дозволяє з легкістю виділити та видалити шуми з даних. Наприклад, при конвертації географічних даних у векторні форми або растрові карти, аномальні значення можуть бути ідентифіковані як викиди або нетипові структури.

Після виявлення шумів за допомогою дискретизації, можна використати різноманітні методи фільтрації або обробки сигналів для їх усунення. Наприклад, застосування фільтрів середнього значення або медіанних фільтрів дозволяє ефективно очистити дані та позбутися шумів. Також можна використовувати методи видалення викидів або відновлення даних на основі навколишніх значень для підвищення якості та достовірності отриманих результатів.

Однією з переваг дискретизованих даних є те, що вони виявляються більш зручними для подальшого використання та аналітики, через те що вони набувають структури, яка сприяє більш легкій інтерпретації та обробці. Перетворення неперервних даних на дискретні форми дозволяє подати їх у вигляді конкретних значень або категорій, що значно спрощує їх розуміння людиною. Наприклад, векторні об’єкти або растрові дані можуть бути представлені у вигляді точок, ліній або полігонів, що робить їх доступними для аналізу та порівняння, що сприяє стандартизації та

однорідності даних. Це полегшує порівняння між різними джерелами даних та сприяє обміну інформацією між різними платформами та користувачами.

Також, дискретизовані дані можна легко обробляти за допомогою різних алгоритмів та програмних засобів. Це відкриває широкі можливості для застосування різноманітних методів аналізу та моделювання в геоінформатиці. Такі дані можуть бути використані у різних програмах та середовищах для геоаналізу, що дозволяє ефективно виконувати різноманітні завдання обробки та виведення результатів.

Отже, дискретизація дозволяє значно зменшити об'єм вхідних даних без значної втрати інформативності, попередньо відфільтрувати шуми та аномалії і покращити загальне розуміння та інтуїтивне сприйняття природи складних географічних даних. Яскравим прикладом може слугувати використання дискретизації на прикладі вибірки даних поїздок таксі у Нью-Йорку за період між 2015 та 2019 роками. На рисунку 1.5, наведено візуалізації вибірки до застосування дискретизації та після, з якого чітко видно контури вулиць міста та скупчення викликів у центральних районах міста на візуалізації дискретизованих даних у порівнянні з візуалізацією оригінальних значень, які утворюють малоінформативну пляму та важко сприймаються візуально.

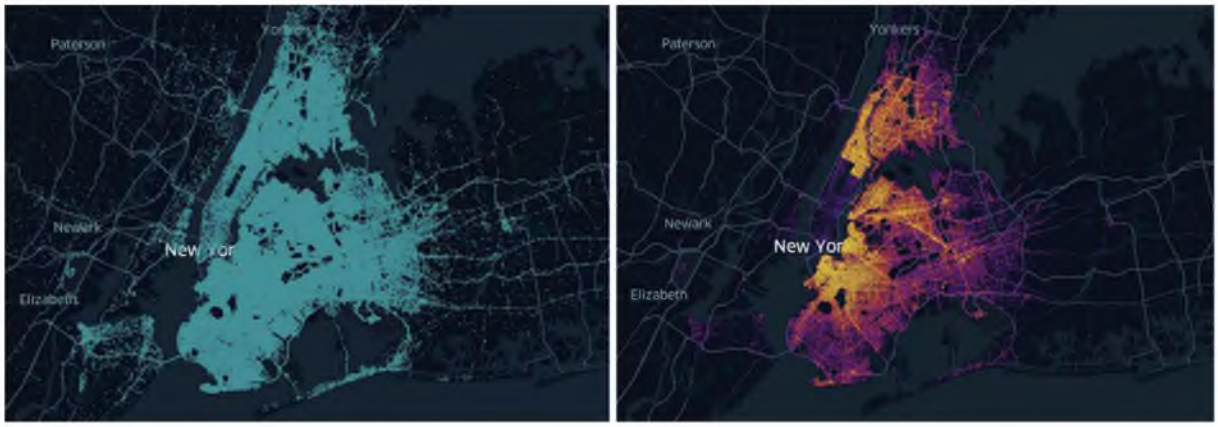

Рисунок 1.5 – Візуалізація даних про поїздки на таксі у Нью-Йорку

1.5 Актуальність, новизна та можливе застосування

У сучасному світі об’єм географічних даних зростає експоненційно, вимагаючи розробки ефективних стратегій їх обробки та аналізу. З підвищенням доступності сучасних технологій та методів збору даних, таких як різноманітні датчики, супутники та сенсори, надходить все більше інформації з географічними атрибутами. Це породжує потік різноманітних даних, включаючи картографічні дані, зображення з супутників, GPS-дані, соціальні медіа з географічними мітками та багато іншого. Великий обсяг цих даних потребує ефективних методів їх обробки та аналізу для виняткового розуміння географічних явищ і забезпечення наукового прогресу та практичного використання в різних галузях, включаючи геоінформатику, геологію, транспорт, екологію, туризм та інші.

Одним із ключових методів для вивчення цих даних виступає кластерний аналіз, який дозволяє групувати схожі об’єкти в кластери та є ефективним інструментом для аналізу географічних даних, так як дозволяє виявити приховані структури в даних та встановити закономірності у розподілі географічних об’єктів. Проте, великий обсяг геоданих може стати викликом через їхню складність та об’єм, що вимагає швидких та ефективних методів обробки. У такому контексті дискретизація геоданих стає актуальним напрямком досліджень, оскільки може сприяти зменшенню обсягу даних та полегшити проведення кластерного аналізу.

Новизна дослідження полягає у застосуванні дискретизації для покращення ефективності та точності кластеризації геоданих. Використання цього методу дозволяє перетворити неперервні або аналогові значення на дискретні форми, що спрощує обробку та аналіз. Цей підхід особливо ефективний при обробці великих обсягів геоданих, де швидкість та ефективність є ключовими вимогами.

Можливе застосування дискретизації для кластеризації геоданих є різноманітним та відкриває широкі можливості. По-перше, її можна

використовувати для попередньої обробки великих обсягів геоданих перед застосуванням алгоритмів кластеризації, що спрощує їхню обробку та зменшує обсяг. По-друге, дискретизація може покращити якість результатів кластеризації, зменшуючи вплив шуму та випадкових аномалій на процес аналізу даних.

Таким чином, дослідження застосування дискретизації для кластеризації геоданих відкриває перспективи для подальших досліджень у галузі обробки та аналізу географічних даних та надає нові можливості для вдосконалення і оптимізації процесів обробки та подальшого аналізу геоданих, а також для виявлення нових залежностей та закономірностей, які раніше могли бути недосяжними. Застосування дискретизації для кластеризації геоданих має широкий спектр можливих застосувань, від планування територій та розташування об'єктів інфраструктури до вивчення екологічних змін та аналізу соціально-економічних процесів.

1.6 Постановка задачі

Мета даної роботи – проведення детального аналізу та порівняння методів кластеризації геоданих з використанням дискретизації для покращення результатів і вирішення основних проблем кластеризації географічних даних.

Для реалізації поставленої мети необхідно виконати наступні задачі:

– детально вивчити предметну область обраної теми;

– обрати методи кластеризації та дискретизації для дослідження;

– проаналізувати переваги і обмеження використання обраних методів кластеризації і дискретизації для геоданих;

– провести експериментальний порівняльний аналіз якості кластеризації обраних методів з використанням дискретизації на реальних географічних даних.

2 ТЕОРЕТИЧНИЙ ОГЛЯД МЕТОДІВ КЛАСТЕРИЗАЦІЇ

2.1 Вимоги до методів кластеризації геоданих

Методи кластеризації геоданих мають відповідати ряду вимог для забезпечення ефективного, урахування усіх технічних особливостей та точного аналізу географічних даних. По-перше, вони повинні бути ефективними для роботи з великими обсягами географічних даних, які можуть містити значну кількість об'єктів або атрибутів. Такі методи повинні мати можливість оптимально обробляти велику кількість даних, забезпечуючи швидкий і ефективний аналіз.

Друга важлива вимога до методів кластеризації геоданих – це здатність враховувати географічні особливості, оскільки геодані можуть містити просторові зв'язки між об'єктами та інші географічні атрибути, методи кластеризації повинні мати змогу працювати таким чином, щоб враховувати ці особливості під час процесу кластеризації.

Нарешті, методи кластеризації геоданих повинні бути достатньо гнучкими, щоб вони могли враховувати різноманітність географічних даних та вимог користувачів. Вони повинні мати можливість працювати з різними типами даних та адаптуватися до різноманітних завдань аналізу та вимог конкретних застосунків у галузі геоінформатики [6].

2.2 Метод K-середніх (K-Means)

Метод K-середніх (K-Means) – це найбільш поширений базовий метод кластеризації даних. Він застосовується для групування елементів, виходячи з їхньої подібності. Цей алгоритм розподіляє дані на k кластерів, або груп, таким чином, що середня відстань між об'єктами в межах кожного кластера менша, ніж між об'єктами з різних кластерів.

Метод k-середніх працює шляхом послідовних ітерацій. Процес починається з обрання k початкових кластерних центрів, також відомих як центри груп. Їх можна обрати випадковим чином або за допомогою спеціальних алгоритмів ініціалізації [7].

Після вибору центрів кластера, алгоритм призначає кожен об'єкт даних до найближчого кластера, використовуючи певну міру відстані, зазвичай евклідову (2.1). Таким чином, об'єкти групуються в кластер, центр якого знаходиться найближче до кожного з них.

$$d(x_i + \mu_j) = \sqrt{\sum_{k=1}^{m} (x_{ik} - \mu_{jk})^2} \tag{2.1}$$

де $d(x_i + \mu_j)$ – відстань між об'єктом x_i та центром кластера μ_j;

x_i – і-й об'єкт з вибірки;

μ_j – центр j-того кластеру;

m – розмірність простору.

Після призначення об'єктів до кластерів, алгоритм k-середніх розраховує нові центри кожного кластера. Визначення нового центру кластера здійснюється шляхом обчислення середнього значення об'єктів, що входять до цього кластера (2.2). Таким чином, центри поточних кластерів постійно оновлюються, щоб краще представляти середнє значення даних у кожному кластері.

$$\mu_j = \frac{1}{N_j} \sum_{i=1}^{N_j} x_i \tag{2.2}$$

де μ_j – новий центр j-того кластеру;

N_j – кількість об'єктів j-тому кластері;

x_i – і-й об'єкт j-того кластеру;

Після оновлення центрів кластерів алгоритм повторно розподіляє об’єкти до кластерів, і процес оновлення центрів кластерів та перерозподіл об’єктів продовжується. Цей цикл триває до тих пір, поки центри кластерів не стабілізуються для мінімізації функції вартості (2.3), або поки не буде досягнуто максимальної кількості ітерацій. Візуалізація роботи алгоритму наведено на рисунку 2.1.

$$J = \sum_{j=1}^{k} \sum_{i=1}^{N_j} d(x_i, \mu_j)^2 \qquad (2.3)$$

де J – функції вартості;

k – кількість кластерів;

N_j – кількість об’єктів j-тому кластері;

$d(x_i + \mu_j)$ – відстань між об’єктом x_i та центром кластера μ_j;

x_i – і-й об’єкт з вибірки;

μ_j – центр j-того кластеру.

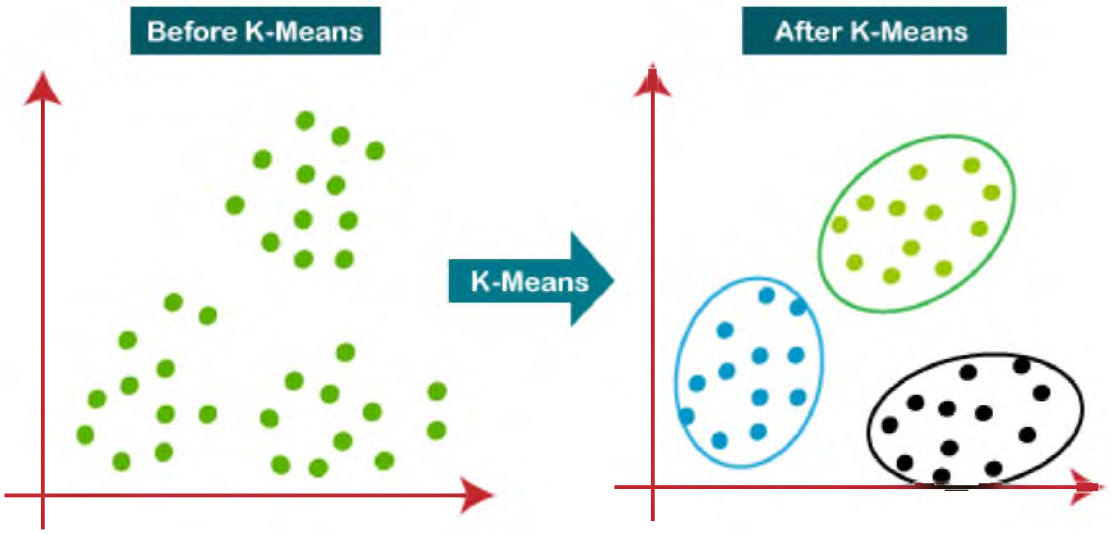

Рисунок 2.1 – Візуалізація роботи алгоритму K-Means

Як правило, алгоритм закінчує роботу за відносно невелику кількість ітерацій, проте, результати кластеризації можуть варіюватися в залежності від вибору початкових центрів кластерів. Саме тому важливо ретельно обирати початкові центри або використовувати методи ініціалізації, які мінімізують вплив цього фактору на результати кластеризації.

Метод k-середніх відомий своєю простотою в розумінні та реалізації, а також швидкістю обробки даних, що робить його ідеальним для достатньо великих обсягів інформації. Також він широко застосовується для кластеризації різних типів даних, включно з геоданими. Однак його використання в цьому контексті може породжувати низку проблем через специфіку самого методу і природу геоданих.

По-перше, геодані часто мають нерівномірний розподіл, включаючи зони з високою концентрацією даних і галузі, де даних занадто мало. Метод k-середніх, орієнтований на створення кластерів з рівною щільністю, може мати труднощі з обробкою таких нерівномірних даних, що може призвести до неточних результатів.

По-друге, метод k-середніх краще підходить для кластерів круглої форми і однакового розміру, тоді як геодані можуть містити кластери різноманітної форми, наприклад, витягнуті або розгалужені області. Цей метод не здатен ефективно обробляти такі складні форми кластерів, що може призвести до поганої інтерпретації структури даних.

Третя проблема полягає в чутливості методу k-середніх до вибору початкових центрів кластерів. Погане визначення початкових центрів може призвести до поганої кластеризації та зниження точності результатів. Це особливо важливо для геоданих, де вибір початкових центрів може істотно вплинути на результат кластеризації.

Додатково, вибір кількості кластерів k є важливим для методу k-середніх. Неправильний вибір k може призвести до неправильного групування даних і знизити якість результатів.

Загалом, метод k-середніх може бути недостатньо ефективним для великих наборів геоданих або багатовимірних даних через вимоги до обчислювальних ресурсів і часу для досягнення конвергенції, проте є непоганим базовим алгоритмом для подальшого порівняння якості.

2.3 Спектральна кластеризація

Спектральна кластеризація є методом кластеризації, який використовує властивості графу для групування даних відповідно до їхньої схожості. Завдяки цьому підходу можна ефективно виявляти складні приховані структури в даних, такі як кластери неправильної форми або розгалужені області.

Метод спектральної кластеризації розпочинається з представлення даних у вигляді графу, де кожен об’єкт даних розглядається як вершина графу, а зв’язки між вершинами відображають рівень подібності або відстані між об’єктами. Ці зв’язки зазвичай представлені у формі вагової матриці подібності, яка визначає ступінь зв’язку між об’єктами [8].

Спектральна кластеризація використовує спектральні властивості матриці для аналізу та групування даних. Зокрема, метод аналізує власні значення та власні вектори вагової матриці подібності, що дозволяє виділити ключові характеристики графового представлення та, відповідно, виявити кластери у наданих даних. Приклад графового представлення даних наведено на рисунку 2.2.

Завдяки своїй здатності працювати з складними структурами даних, спектральна кластеризація є потужним інструментом для аналізу даних у різних сферах, зокрема, в обробці зображень, обробці тексту, дослідженнях соціальних мереж, геоінформатиці, а також в інших різноманітних областях, де потрібно групувати складні дані для забезпечення подальшого аналізу та отримання аналітично цінних висновків.

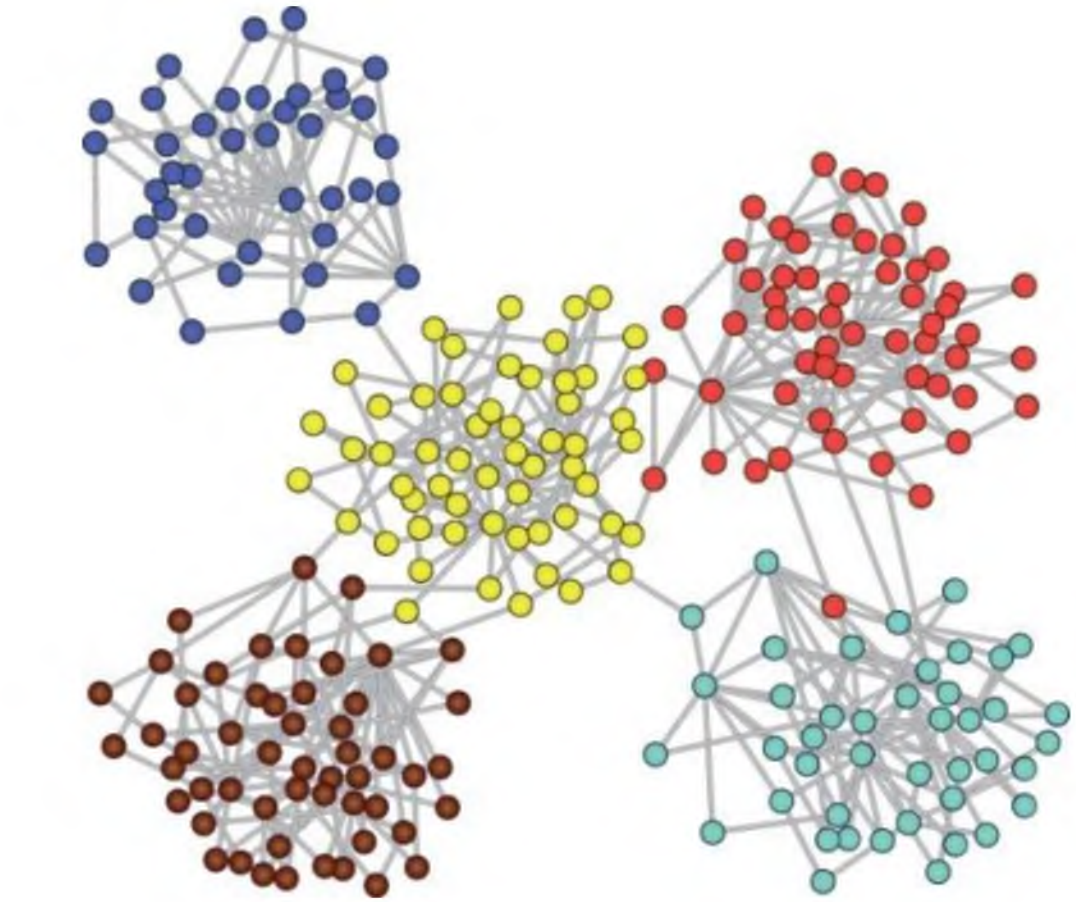

Рисунок 2.2 – Приклад графового представлення даних

Спочатку дані перетворюються на матрицю подібності або вагову матрицю W, де кожен елемент W_{ij} представляє подібність між двома об'єктами даних i та j. Вагова матриця зазвичай є симетричною, де $W_{ij} = W_{ji}$. Далі, обчислюється діагональна матриця ступенів D, де кожен діагональний елемент D_{ii} є сумою всіх ваг у рядку і вагової матриці W (2.4).

$$D_{ii} = \sum_{j=1}^{n} W_{ij} \tag{2.4}$$

де D_{ii} – і-й діагональний елемент матриці ступенів;

W_i – і-й рядок вагової матриці;

n – кількість колонок вагової матриці;

Лапласіан графу L обчислюється як різниця між діагональною матрицею ступенів і ваговою матрицею (2.5).

$$L = D - W \tag{2.5}$$

де L – лапласіан графу;

D – діагональна матриця ступенів;

W – вагова матриця;

Наступний крок представляє собою розкладання лапласіану для знаходження власних значень і власних векторів. Лапласіан може бути нормалізованим для підвищення точності (2.6).

$$L_{\text{норм}} = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} \tag{2.6}$$

де $L_{норм}$ – нормалізований лапласіан;

D – діагональна матриця ступенів;

L – лапласіан.

Спектральний розклад лапласіану передбачає знаходження власних векторів і власних значень лапласіану (2.7).

$$Lv_i = \lambda_i v_i \tag{2.7}$$

де $L_{норм}$ – нормалізований лапласіан;

v_i – власний вектор лапласіану;

λ_i – власне значення лапласіану.

Отже, спектральна кластеризація є передовим методом кластеризації геоданих, завдяки своїм унікальним властивостям, що забезпечують переваги перед традиційними методами. Вона виявляє складні та нерівномірні форми кластерів, властиві геоданим, включно з розгалуженими, витягнутими або неправильною геометрією кластерами.

Цей метод використовує графові підходи для аналізу структури даних і оптимізує їх розподіл по кластерам. Спектральна кластеризація враховує глобальні характеристики даних, досліджуючи власні вектори і власні значення лапласіану графу. Це дозволяє врахувати ширший контекст і знайти адекватні кластери, оцінюючи взаємозв'язки між об'єктами даних на великій площі.

Завдяки методам лінійної алгебри, таким як спектральний розклад лапласіану, спектральна кластеризація здатна ефективно працювати з великими обсягами геоданих. Вона також чутлива до різних мір схожості, включаючи відстань, подібність або кореляцію. Це дозволяє більш точно моделювати взаємозв'язки між геоданими, підвищуючи загальну якість результатів кластеризації.

Окрім того, аналіз власних векторів дозволяє виявляти дані, що виділяються, тобто об'єкти, які не відповідають основним кластерам. Цей процес може бути корисним для ідентифікації аномалій або особливих випадків у геоданих.

Загалом, спектральна кластеризація є ефективним методом для роботи з геоданими, особливо коли вони мають складну структуру або нерівномірний розподіл. Тим не менш, варто зазначити, що цей підхід може бути обчислювально інтенсивним, особливо при обробці великих наборів даних, і може вимагати спеціальних параметрів для оптимального налаштування алгоритму.

2.4 Метод DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) – це метод кластеризації, який групує об'єкти даних на основі їхньої щільності. Цей метод є особливо ефективним для обробки різноманітних просторових, оскільки дозволяє виявляти кластери з різною формою та щільністю, що характерно для географічних даних.

Метод DBSCAN працює, визначаючи області з високою щільністю та об'єднуючи об'єкти даних, які знаходяться поблизу один одного і перевищують заданий поріг щільності. Цей підхід дозволяє виявляти кластери, які можуть мати нерівномірну форму або щільність, що відрізняє DBSCAN від інших методів кластеризації.

Кластеризація методом DBSCAN ґрунтується на двох ключових параметрах: радіусі (ε), який визначає околиці точки та якщо інші точки знаходяться в межах цього радіусу, вони вважаються сусідами початкової точки, і мінімальній кількості точок у кластері (minPts), що визначає скільки точок має бути присутні у радіусі (ε) від початкової точки, щоб вона вважалася ядром кластера та, якщо в межах радіусу знаходиться менше ніж minPts точок, початкова точка розглядається як шумова [9].

DBSCAN починає кластеризацію з обрання довільної точки з набору даних. Після цього алгоритм визначає точки, що знаходяться в радіусі ε від початкової точки, щоб встановити її околиці. Якщо кількість цих точок дорівнює або перевищує мінімальний поріг (minPts), то початкова точка стає ядром кластера, і алгоритм починає розширювати кластер, включаючи всі сусідні точки, які також задовольняють вимоги щільності.

Коли кластер розширюється, алгоритм продовжує додавати всі нові точки, що знаходяться в межах радіуса ε від уже доданих точок кластера. Цей процес розширення кластеру триває, поки не будуть перевірені всі сусідні точки, які відповідають критерію щільності.

Якщо початкова точка не є ядром кластера, тобто кількість точок в її околиці менша за мінімальний поріг, то вона розглядається як шум або аномалія. DBSCAN продовжує кластеризацію, переходячи до наступної довільної точки та повторюючи процес, поки не буде розглянуто весь набір наданих даних.

В цілому, у алгоритмі DBSCAN виділяють три різновиди точок. Точки-ядра мають достатню кількість сусідніх точок у межах радіуса ε та, отже, є центрами кластерів. Точки-ядра виконують ключову роль у

розширенні кластеру, оскільки вони є відправною точкою для включення нових точок у кластер. Точки-грані, які знаходяться в межах радіуса ε від точок-ядер, але самі не є ядрами. Вони належать до кластеру, проте не мають достатньої кількості сусідів для самостійного розширення кластеру. Точки-грані зазвичай знаходяться на межі кластерів і є складовою частиною кластеру, хоч і не виконують центральної ролі. Шумові точки не належать до жодного кластеру, оскільки вони не задовольняють критерію щільності. Ці точки розглядаються як шум або аномалії і залишаються поза межами кластерів. Візуалізація роботи алгоритму наведена на рисунку 2.3.

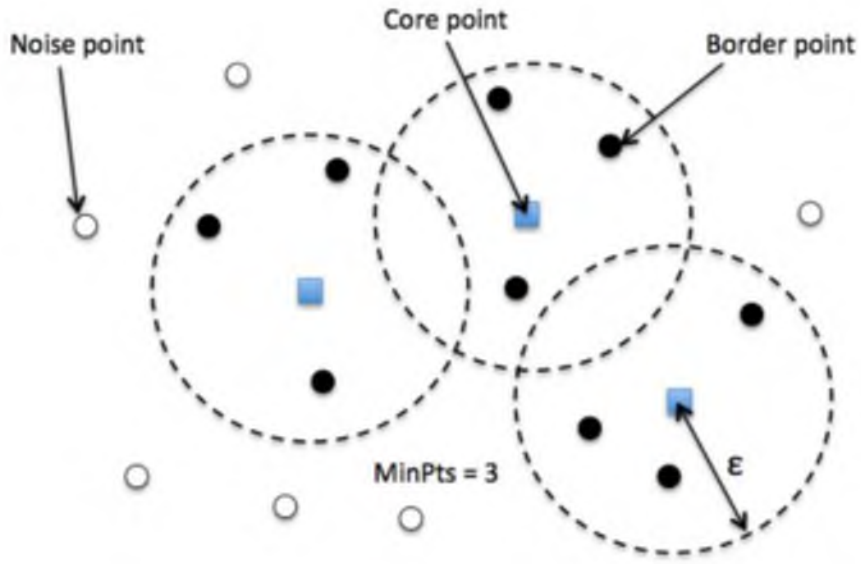

Рисунок 2.3 – Візуалізація роботи алгоритму DCSCAN

Тобто, алгоритм DBSCAN може розпізнавати та виділяти шумові точки, тобто ті, які не входять до жодного кластеру. Це дозволяє виявляти аномальні дані та підвищувати якість кластеризації. Крім того, алгоритм не вимагає попереднього визначення кількості кластерів, оскільки він самостійно встановлює їх на основі щільності даних.

Також, однією з ключових переваг алгоритму DBSCAN є його здатність виявляти кластери з різною щільністю в даних, що особливо важливо для геоданих, які зазвичай характеризуються нерівномірним

розподілом. Цей метод може визначати різноманітні форми кластерів, включаючи складні нерегулярні, розгалужені або витягнуті області, часто присутні в геоданих.

Проте, вибір оптимальних параметрів ε (радіуса) та MinPts (мінімальної кількості точок) може бути складним завданням, оскільки він впливає на якість кластеризації. Крім того, алгоритм чутливий до масштабу відстаней між даними, що вимагає обережності під час попереднього оброблення геоданих та якісного налаштуванні метрики відстані. DBSCAN може вимагати значних обчислювальних ресурсів і часу, особливо при роботі з великими наборами або складними геоданими.

Для досягнення точності кластеризації важлива попередня обробка даних, включаючи нормалізацію та масштабування. Вибір правильних значень ε та MinPts впливає на результат кластеризації, тому важливо експериментувати з різними значеннями, щоб знайти оптимальні. DBSCAN також можна поєднувати з іншими методами кластеризації або алгоритмами машинного навчання для того, щоб досягти кращих результатів у складних завданнях чи даних.

Загалом, DBSCAN є ефективним інструментом для кластеризації геоданих, особливо якщо дані мають нерівномірний розподіл або складну структуру, проте правильний вибір параметрів та попередня обробка даних все ж відіграють критичну роль у забезпеченні якісної кластеризації.

2.5 Метод OPTICS

Ідентифікація OPTICS (Ordering Points to Identify the Clustering Structure) представляє собою метод кластеризації, який розширює підхід DBSCAN, пропонуючи більш гнучке та детальне виявлення кластерів. Як і DBSCAN, OPTICS базується на принципі щільності точок та подібності між ними. Однак головна перевага OPTICS полягає в здатності виявляти різні рівні щільності кластерів.

Алгоритм кластеризації методом OPTICS починається з вибору випадкової точки в наборі даних та оцінки її щільності на основі відстані до навколишніх точок. Використовується параметр радіусу епсілон (ε) для визначення мінімальної відстані, в якій об'єкти можуть бути приєднані до кластеру. Також визначається мінімальна кількість точок (MinPts), які мають перебувати в межах радіусу ε, щоб точка вважалася ядром кластера.

Після вибору початкової точки, OPTICS продовжує аналіз інших точок у порядку їх віддаленості від початкової точки. Алгоритм створює послідовність точок, відому як графік досяжності (reachability plot), який відображає відстань між точками відповідно до порядку їх обробки. Цей графік дає змогу візуалізувати різні рівні щільності кластерів [10].

На відміну від фіксованого значення ε, яке використовується в DBSCAN, OPTICS створює послідовність віддаленості до точок, що дозволяє виявляти багаторівневі кластери. Після виконання кластеризації, послідовність віддаленості можна візуалізувати, щоб розглянути природні розриви між кластерами різної щільності. Це дає можливість краще розуміти структуру даних та виявляти більш точні кластеризації. Візуалізація відстаней наведена на рисунку 2.4.

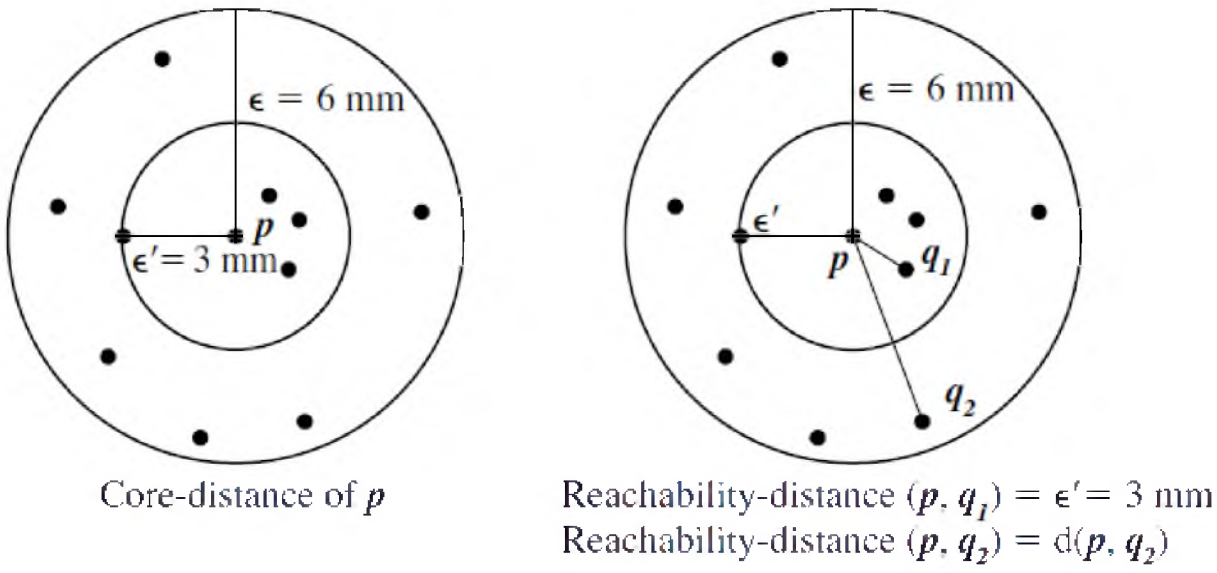

Рисунок 2.4 – Візуалізація відстаней для алгоритму OPTICS

Ідентифікація Метод OPTICS є дуже ефективним інструментом для кластеризації геоданих завдяки своїй здатності виявляти кластери різної щільності та складної структури. Цей метод розширює можливості DBSCAN і дозволяє краще дослідити структуру даних. Одна з його головних переваг полягає в здатності забезпечувати багаторівневе виявлення кластерів і розподіляти дані на основі їх щільності. Це особливо важливо для геоданих, які можуть мати нерівномірний розподіл та складні форми кластерів.

Оскільки OPTICS використовує послідовність віддаленості (reachability plot) для візуалізації різних рівнів щільності кластерів, це допомагає зрозуміти структуру даних і дозволяє дослідити різні рівні кластеризації в одному наборі даних. Метод також здатний виявляти шумові точки, які не належать до жодного кластеру, що сприяє покращенню якості кластеризації та очищенню даних.

Однак використання OPTICS може мати свої недоліки. Це обчислювально дорогий метод, особливо для відносно великих наборів даних. Крім того, вибір оптимальних параметрів, таких як мінімальна кількість точок і значення радіусу, є критичним завданням для забезпечення високої якості кластеризації. Метод також може бути чутливим до масштабу відстаней між даними, тому попередня обробка, включаючи нормалізацію, є відчутно важливою.

Загалом, використання OPTICS для кластеризації геоданих дозволяє дослідити різні рівні кластеризації в одному наборі даних і краще розуміти структуру даних. Хоча цей метод може вимагати великих обчислювальних ресурсів, його здатність виявляти складні кластери та шум робить його цінним інструментом для аналізу геоданих.

3 ТЕОРЕТИЧНИЙ ОГЛЯД МЕТОДІВ ДИСКРЕТИЗАЦІЇ

3.1 Класи методів дискретизації даних

Дискретизація у розрізі геоданих – це процес розбиття безперервних даних на дискретні категорії або інтервали, що може спростити аналіз та інтерпретацію даних. Існує кілька основних класів методів дискретизації, кожен з яких підходить для певних типів геоданих та цілей, що поставлені для дослідження.

Перший клас методів дискретизації – це розподіл на рівні інтервали. У цьому підході весь діапазон даних розбивається на інтервали однакової довжини, що зручно для порівняння різних категорій. У контексті геоданих до цього класу відносяться географічна індексація та грід-системи.

Другий клас – розподіл вибірки даних на рівні частоти. Цей метод передбачає розбиття даних на інтервали так, щоб кожен інтервал містив приблизно однакову кількість спостережень, що забезпечує збалансоване розподілення, яке може бути необхідним для аналізу.

Третій клас – класифікація на основі природних розривів. У цьому методі інтервали визначаються відповідно до природних змін у даних, що дозволяє виділяти окремі категорії. Для геоданих такими категоріями можуть виступати як різні типи ландшафту так і райони одного міста.

Вибір методу дискретизації залежить від специфіки геоданих і цілей дослідження. Кожен метод має свої переваги та обмеження, тому важливо обрати той, який найкраще відповідає поставленим завданням. Деякі методи можуть краще підходити для виявлення природних розривів у даних, тоді як інші більше підходять для створення рівномірних інтервалів або збалансованих категорій. Оскільки кожен підхід має свої унікальні характеристики, оптимальний вибір залежить від конкретних вимог аналізу та мети використання дискретизованих даних.

Для використання кластеризації геоданих основним класом методів дискретизації буде виступати саме географічна індексація, оскільки вона має кілька суттєвих переваг для цього типу даних та його особливостей.

Таким чином, географічні сітки дозволяють покрити всю поверхню Землі, забезпечуючи універсальний та цілісний підхід до дискретизації геоданих та надаючи можливість роботи з різними масштабами, дозволяючи аналізувати дані на різних рівнях деталізації (різні рівні ієрархії). Це корисно для різних завдань, наприклад, від глобального аналізу до локального дослідження [11].

Кожна клітинка у географічних сітках має унікальний ідентифікатор, що спрощує зберігання, запит та управління геоданими. Ці методи дозволяють ефективно обробляти різні типи геоданих, включаючи просторові та темпоральні дані, що робить їх універсальними для різноманітних завдань.

Географічна індексація, пропонує прості та інтуїтивно зрозумілі підходи до дискретизації даних. Вона також добре підходять для обробки великих обсягів геоданих. Методики географічної індексації добре інтегруються з популярними географічними інформаційними системами (ГІС) та інструментами, що дозволяє легко аналізувати та візуалізувати дискретизовані геодані [12].

Отже, дискретизація геоданих, а саме географічна індексація, пропонує ефективний підхід до дискретизації геоданих завдяки своїй універсальності, простоті, гнучкості та можливості обробки великих обсягів даних на різних рівнях деталізації.

3.2 Система географічної індексації BingTiles

BingTiles – це система географічної індексації, розроблена компанією Bing Maps, яка розділяє поверхню Землі на квадратні тайли різних масштабів для оптимізації відображення на різноманітних мапах. Така

система є корисною для організації геоданих, оскільки забезпечує ефективний доступ до них і зручність обробки [13].

Кожен квадратний тайл у системі BingTiles має унікальний ідентифікатор (наприклад, Quadkey), що дозволяє ідентифікувати кожен тайл та працювати з ним окремо. Тайли організовані ієрархічно, що дає змогу виконувати навігацію на мапах різного рівня деталізації, від глобального масштабу до локальних областей, візуалізація верхніх рівнів ієрархії тайлів наведено на рисунку 3.1.

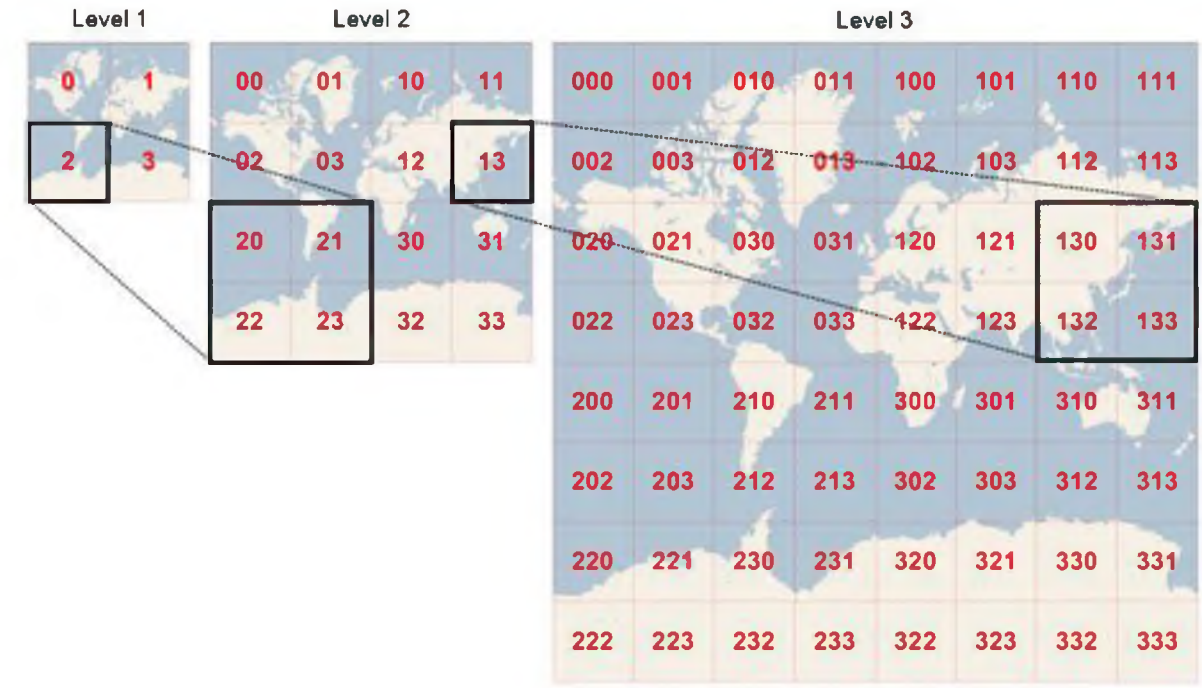

Рисунок 3.1 – Візуалізація ієрархії BingTiles

Головною перевагою BingTiles є ефективність і простота організації геоданих. Тайли різних масштабів дозволяють користувачам легко змінювати масштаб мапи, при цьому забезпечуючи швидке завантаження відповідних даних. Це покращує користувацький досвід у картографічних додатках, мобільних навігаційних системах та інших додатках, пов’язаних з геоданими.

Система BingTiles також забезпечує сумісність з популярними географічними інформаційними системами (ГІС) та інструментами, що спрощує інтеграцію з іншими даними або системами. Використання квадратних тайлів спрощує швидкий пошук та аналіз геоданих на кожному рівні масштабування, забезпечуючи зручну візуалізацію та дослідження.

Хоча BingTiles ефективний для більшості завдань, пов'язаних з картографією та аналізом геоданих, існують певні проблеми, такі як потенційне викривлення під час відображення на плоских картах і труднощі при роботі з великими обсягами даних. Незважаючи на ці проблеми, BingTiles залишається зручною та потужною системою для роботи з геоданими в більшості випадків.

3.3 Система географічної індексації S2

S2 (Spherical GeoHash) – це система географічної індексації, розроблена компанією Google, що дозволяє поділити поверхню Землі на дискретні клітинки геометричних форм. Її основна мета – забезпечити ефективну організацію та адресацію геоданих для швидкого пошуку та різноманітної обробки [14].

Кожна клітинка у системі S2 має унікальний ідентифікатор, який може бути представлений у вигляді рядка з числовим значенням. Це надає можливість точно адресувати будь-яку точку на Землі. Такий підхід робить S2 ідеальним для використання в широкому спектрі застосувань, включаючи картографію, геоаналітику, мобільну навігацію та інші геопрограми.

Однією з ключових переваг S2 є його спроможність працювати з географічними формами на сфері, що дозволяє адресувати точки у будь-якій точці планети. Крім того, система підтримує різні рівні деталізації, що дозволяє адаптувати рівень дискретизації до конкретних потреб додатку або

аналітичного завдання, візуалізація верхніх рівнів ієрархії S2 наведено на рисунку 3.2.

Іншою перевагою S2 є його широка підтримка та інтеграція з іншими інструментами та сервісами Google, такими як Google Maps та Google Earth. Це дозволяє розробникам легко інтегрувати S2 з іншими додатками та сервісами Google.

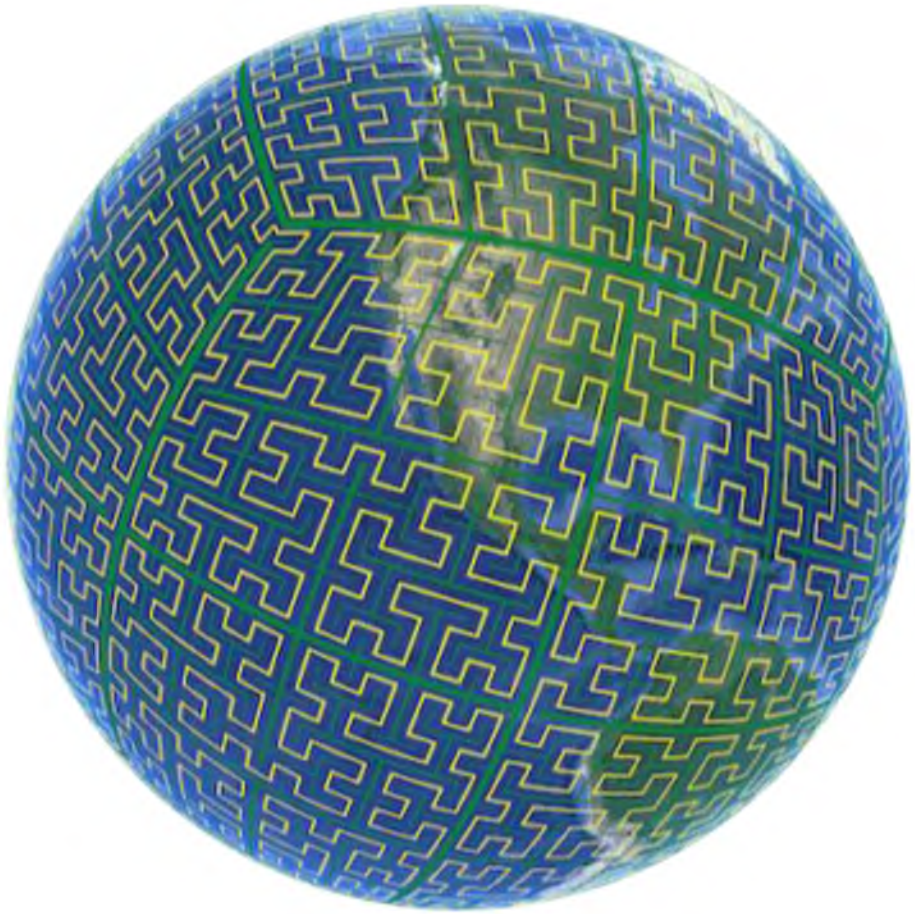

Рисунок 3.2 – Візуалізація ієрархії S2

Незважаючи на свої переваги, варто враховувати, що використання S2 може потребувати додаткових обчислювальних ресурсів через складність обробки географічних форм. Також слід враховувати можливі виклики, пов’язані з конвертацією даних між різними форматами та системами координат. Однак у більшості випадків S2 залишається потужним та ефективним інструментом для роботи з геоданими.

3.4 Система географічної індексації H3

H3 (Hierarchical Hexagonal Indexing) представляє собою інноваційну систему географічної індексації, розроблену компанією Uber Technologies. Вона базується на використанні гексагональної сітки для поділу поверхні Землі на дискретні клітинки, що дозволяє ефективно організовувати та адресувати геодані для різноманітних застосувань, включаючи геолокацію, аналіз даних та розрахунки маршрутів [15].

Кожен гексагон у системі H3 має свій унікальний ідентифікатор, який складається з шістнадцятирічного рядка шестнадцяткових цифр. Ця унікальність дозволяє точно адресувати будь-яку область на Землі з високою точністю, використовуючи гексагональну форму, яка особливо зручна для опису географічних областей, оскільки гексагони легко з'єднуються та обробляються. Приклад покриття географічної області гексагонами наведено на рисунку 3.3.

Рисунок 3.3 – Покриття H3 гексагонами Флориди

Однією із головних переваг H3 є його ієрархічна структура, яка дозволяє різним рівням деталізації. Це означає, що можна використовувати більш грубу дискретизацію для швидкої адресації та обробки геоданих на великих територіях, а також більшу деталізацію для досягнення високої точності на менших масштабах, візуалізація верхніх рівнів ієрархії H3 наведено на рисунку 3.4.

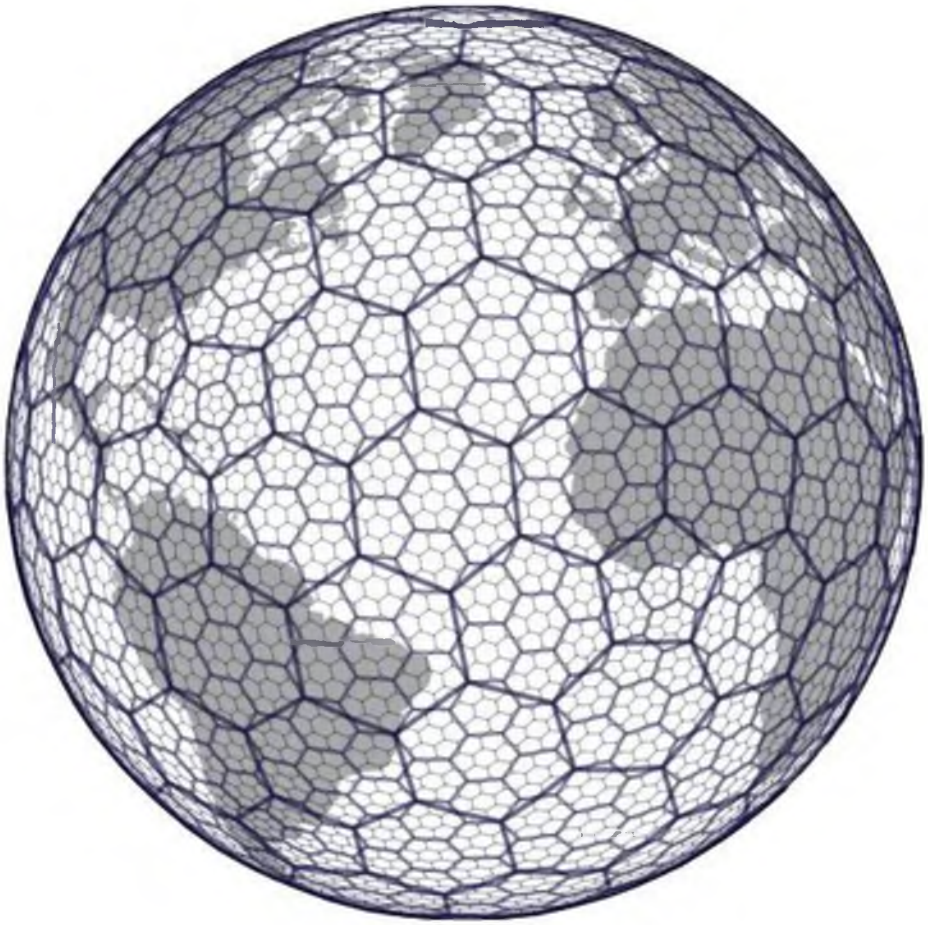

Рисунок 3.4 – Візуалізація ієрархії H3

Додатковою перевагою H3 є його здатність працювати на будь-якій поверхні, включаючи полюси та океани. Це робить його універсальним інструментом для адресації та обробки геоданих у різних географічних областях, надаючи зручність та ефективність для роботи з геоданими на всіх рівнях деталізації.

Незважаючи на всі переваги, використання H3 може потребувати додаткових обчислювальних ресурсів через складність обробки географічних форм, особливо на великих масштабах. Також варто враховувати, що велика кількість гексагонів може призвести до збільшення обсягу даних та потребувати додаткових зусиль для аналізу та візуалізації. Однак у більшості випадків H3 є потужним інструментом для роботи з геоданими на різних масштабах, надаючи зручність та ефективність для широкого спектру застосувань.

3.5 Порівняння ієрархічних систем

Розглянуті системи географічної індексації, такі як BingTiles, S2 та H3, пропонують різні підходи до поділу поверхні Землі для подальшої обробки геоданих [16].

BingTiles відомий своєю простотою та легкістю використання, особливо в контексті інтеграції з іншими картографічними сервісами Bing Maps. Ця система використовує стандартизований підхід, розділяючи поверхню на квадратні тайли, що дозволяє зручно працювати з картографічними даними різних масштабів.

У свою чергу, S2, розроблений Google, пропонує більш гнучкий підхід до географічної індексації. Використовуючи геометричні форми на сфері, S2 забезпечує можливість адресації будь-якої точки на Землі та підтримує різні рівні деталізації, що робить його корисним для широкого спектру застосувань, від глобального картографування до локального аналізу географічних даних.

З іншого боку, H3 вирізняється своєю гексагональною структурою, що дозволяє рівномірно розподіляти геодані на поверхні Землі. Це особливо корисно для кластеризації геоданих, оскільки гексагональна форма дозволяє ефективно описувати різні географічні області. Головною перевагою H3 є його особливість – однакова відстань між сусідніми клітинками. Така

однорідність структури дозволяє однаково враховувати географічні особливості незалежно від місця на Землі, що робить H3 ідеальним для кластеризації, де важлива однорідність та точність визначення географічних областей. Візуалізація відстаней між сусідніми клітинками для S2, BingTiles та H3 наведена на рисунку 3.5 [17].

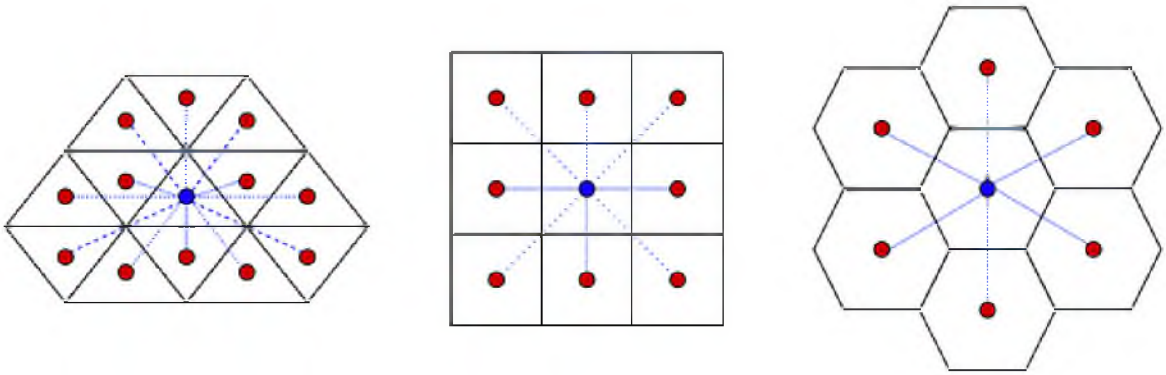

Рисунок 3.5 – Візуалізація відстаней між клітинками S2, BingTiles та H3

Отже, H3 є потужним інструментом дискретизації геоданих для використання у парі з кластеризацією, особливо завдяки його гексагональній структурі з однаковою відстанню між будь-якими сусідніми клітинками, що дозволяє точно визначати та рівномірно представляти географічні області на поверхні Землі.

4 ПРАКТИЧНЕ ДОСЛІДЖЕННЯ КЛАСТЕРИЗАЦІЇ

4.1 Огляд вибірки

4.1.1 Предметна область

За останні роки в галузі IoT (Інтернету речей) значно зросла активність використання інформації про місцезнаходження, що забезпечує великий обсяг даних для аналізу геоданих. Поширення IoT-пристроїв з вбудованими модулями геолокації сприяло стабільному зростанню обсягу інформації про місцезнаходження, відкриваючи нові можливості для подальших досліджень та різноманітного аналізу [18].

Дослідження кластеризації на основі даних про місцезнаходження дозволяє аналізувати розміщення об’єктів у просторі та виявляти патерни їх скупчень. Це сприяє кращому розумінню закономірностей поведінки та потреб користувачів IoT-пристроїв.

Крім того, кластеризація даних за подібністю географічного розташування та часу дозволяє виявляти географічні області зі схожою активністю. Це може бути корисно для виявлення популярних місць, зон інтересу або географічних областей із подібним рівнем активності.

Аналіз кластерів допомагає підприємствам удосконалювати спрямованість своїх маркетингових стратегій, зокрема, орієнтуючись на конкретні географічні області та групи користувачів IoT. Крім того, ідентифікація кластерів корисна для управління транспортними потоками, розробки оптимальних маршрутів та удосконалення систем керування громадським транспортом.

Загалом, використання даних про місцезнаходження для кластеризації геоданих в області IoT відкриває широкі можливості для аналізу руху, розуміння поведінки користувачів та оптимізації найрізноманітніших бізнес-процесів [20].

4.2.1 Огляд вибірки даних

Вибірка для проведення практичного дослідження методів кластеризації геоданих з використанням дискретизації представляє собою набір сигналів різноманітних IoT пристроїв в Києві та Київській області.

Інформацію щодо локацій було отримано та обфусковано від анонімного провайдера, який збирає локації з різних джерел, таких як побутова техніка, портативні девайси, техніка для клімат-контролю, автомобільні навігаційні системи та інші пристрої, які мають доступ до мережі та передають місцеположення.

Такий підхід до розповсюдження даних дозволяє збирати та обробляти дані про місцезнаходження без порушення конфіденційності користувачів, що є важливим аспектом у сучасному цифровому середовищі. Саме тому ці можна використовувати для різних цілей, включаючи аналітику місцезнаходження, без порушення конфіденційності.

Ця вибірка даних охоплює дані місцеположення портативних пристроїв та транспортних засобів у різних частинах міста Києва та Київської області і містить понад 7.550.000 унікальних записів, приклад невеликої підмножини яких наведено на рисунку 4.1.

uuid	latitude	longitude	accuracy	signal_strength	created_at	updated_at
0dc34eb1-bce3-4a35-8a30-e66d42179c3f	49.925806	30.260805	30.0	32.176128	2024-04-23 07:40:05.501000	2024-04-22 16:24:42.148000
47c11be7-3cbc-450f-a4bf-b8b33070a3d8	50.078311	30.357967	109.0	31.826744	2024-03-13 09:45:23.227000	2024-04-17 18:38:16.126000
d7d86d02-a6c7-4299-9c8a-7f201b7df209	50.817316	30.675971	31.0	32.873218	2024-03-26 13:46:08.455000	2024-04-23 09:17:47.572000
42e2997d-b4eb-4cbc-85a1-bae49c6369a6	50.227432	30.798450	75.0	21.756226	2024-03-13 09:11:42.805000	2024-04-23 09:25:58.267000
929ee41e-277f-4730-8d14-4152aeeb4801	50.214936	30.427374	26.0	56.725271	2024-03-13 09:01:37.123000	2024-04-22 10:56:56.130000
...	...	...	...	...	...	...
84894872-e194-4926-b2f8-deef1c3d37f4	50.382514	30.225324	30.0	23.610031	2024-03-13 07:49:47.653000	2024-04-22 16:29:12.145000
601618f2-b6a9-4329-b748-c4433d55e458	49.834609	30.779188	30.0	26.839607	2024-04-23 07:40:05.501000	2024-04-22 17:14:54.918000
5fe19d80-01b3-4662-bb14-15bbc458755a	50.904172	31.126593	30.0	25.682363	2024-03-13 07:55:50.405000	2024-04-22 16:47:52.620000
e3b9f18d-a7fd-4ca5-820e-e608a9be2c62	50.409314	29.969821	30.0	21.635085	2024-03-13 08:18:22.726000	2024-04-22 16:47:40.161000
88ea73bf-a973-478f-81b6-0ad1cdffceae	50.624015	30.877409	87.0	15.138005	2024-03-13 09:00:15.696000	2024-04-18 13:04:01.090000

7559708 rows × 6 columns

Рисунок 4.1 – Приклад частини вибірки сигналів локацій IoT пристроїв

Первинним ключем виступає UUID (Universally Unique Identifier) четвертої версії, що був детермінованим чином згенерований з підмножини прихованих у цілях анонімізації частини атрибутів.

Основні географічні колонки latitude (широта) та longitude (довгота) знаходяться у проекції EPSG:4326 та мають діапазони значень від 49.182615 до 51.530606 та від 29.287285 до 32.109776 для довготи та широти відповідно. Візуалізація координат всіх пар координат з вибірки у вигляді теплової карти на карті знаходиться на рисунку 4.2, на якій можна досить чітко побачити скупчення точок у великих населених пунктах та найбільш густонаселених житлових районах Києва, де переважає використання різноманітних IoT пристроїв.

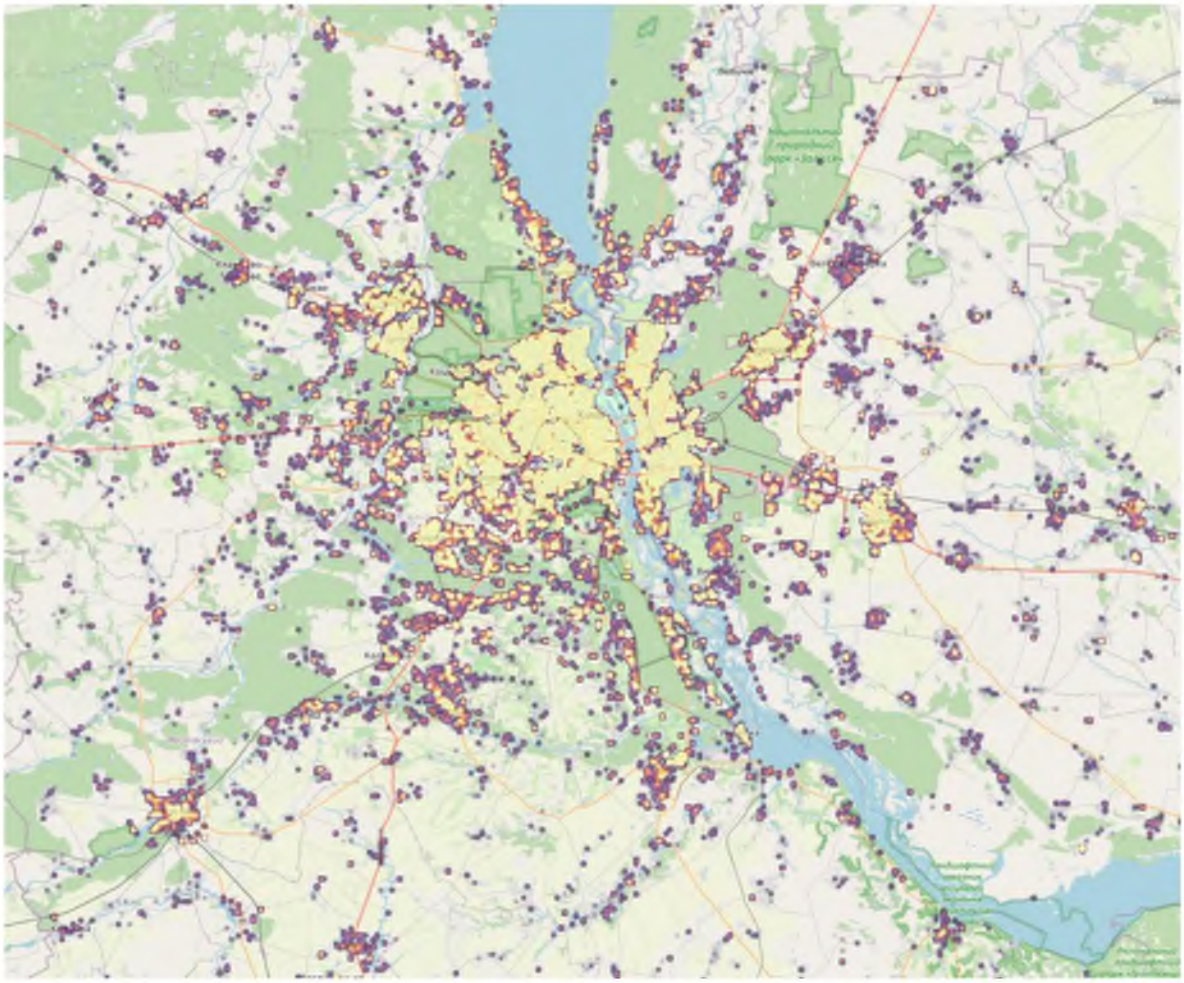

Рисунок 4.2 – Теплова карта вибірки локацій IoT пристроїв для Києва та Київської області

Також для кожної пари координат у вибірці зазначена точність локації accuracy, що представляє собою радіус у метрах та нормована до інтервалу від 0 до 100 сила сигналу signal_strength.

Атрибути created_at та updated_at означають дату і час по UTC першої та останньої фіксації сигналу для окремого унікального ідентифікатора відповідно. Гістограма розподілу сили сигналу місцеположення наведена на рисунку 4.3. На зображеній на рисунку гістограмі можна зауважити розподіл, що апроксимується нормальним з похилим правим краєм.

Такий розподіл може бути пов'язаний з тим, що в деяких областях або у деякий час сигнали можуть бути сильнішими через більшу концентрацію пристроїв або кращу якість сигналу, в той час як у інших областях або в інші періоди сигнали можуть бути слабшими через різноманітні фактори, такі як погодні перешкоди або непередбачені втрати зв'язку.

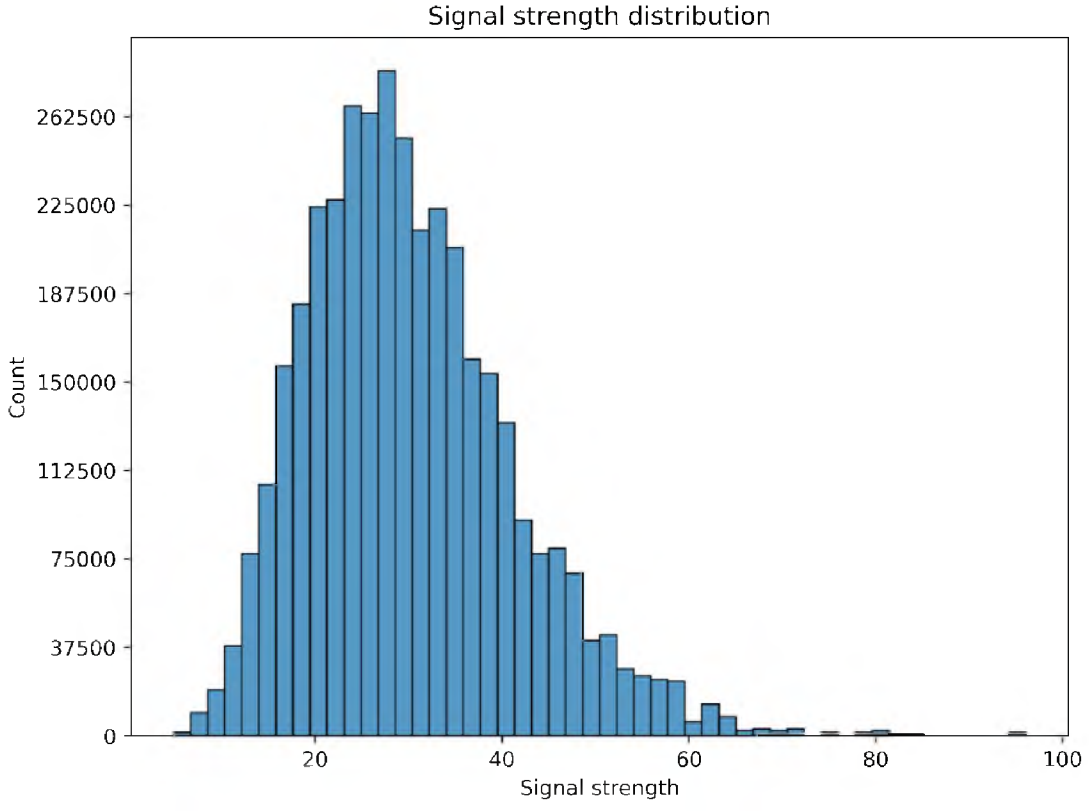

Рисунок 4.3 – Гістограма розподілу частоти сигналу

4.2.3 Огляд вибірки даних

Оскільки об’єм вибірки сигналів достатньо великий і її повний аналіз є малоінформативним через неоднорідне розміщення даних, тому набагато практичніше вибрати декілька районів міста для проведення подальшого дослідження, а саме – Оболонь та Троєщину.

Обрані для проведення дослідження райони є одними із найбільш населених та розвинених районів міста, що включають в себе житлові масиви, комерційні центри, парки та рекреаційні зони. Теплова карта вибірки для міста Києва з позначеними густонаселеними районами наведена на рисунку 4.4.

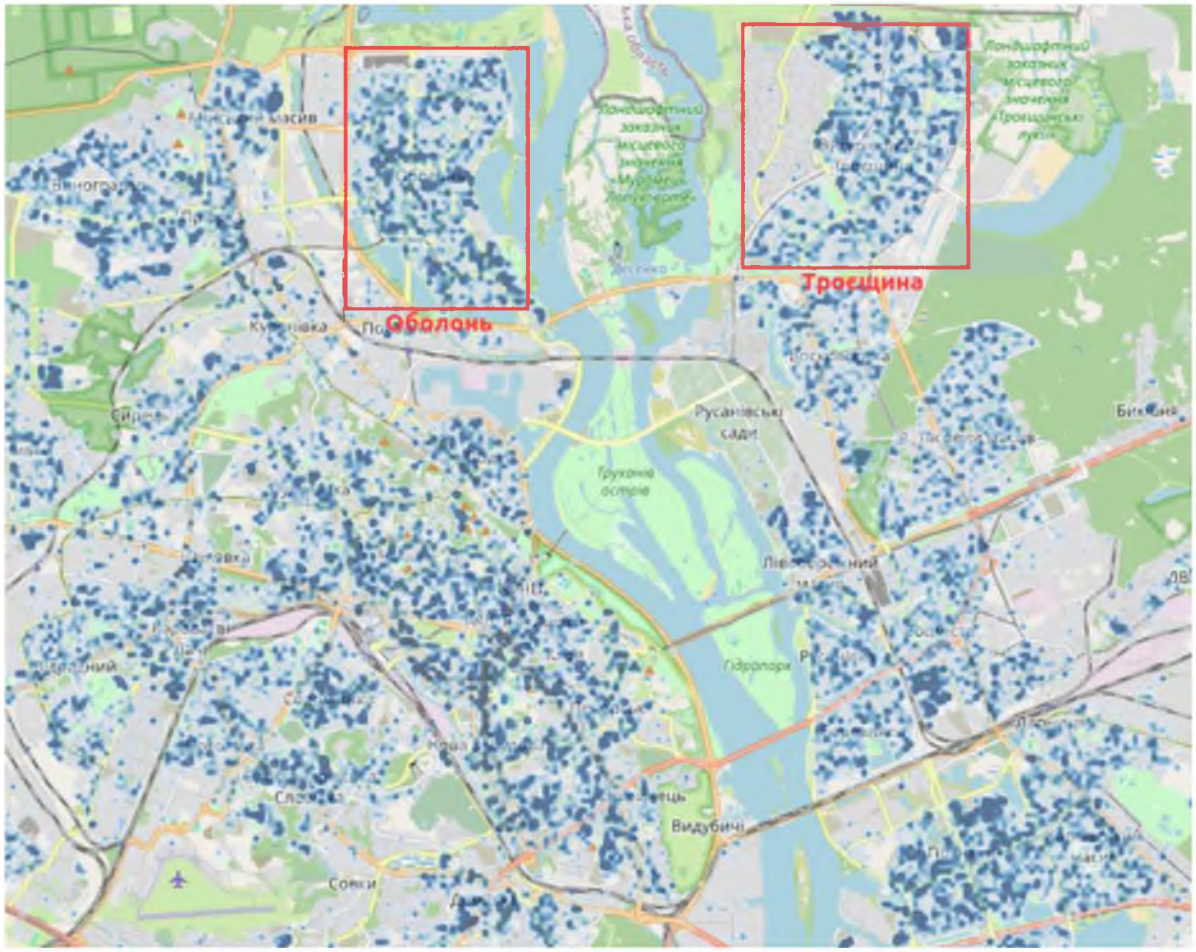

Рисунок 4.4 – Теплова карта вибірки локацій IoT пристроїв для районів міста Києва

Оболонь – це один з найбільш мальовничих та найбільш розвинених районів Києва, розташований на лівому березі річки Дніпро. Цей район приваблює своїми парками, набережними, ресторанами та розважальними закладами, що забезпечує активне життя та багатий потік людей.

Також Оболонь є одним з найбільш густонаселених (понад 300.000 осіб) районів міста Києва, що надає значну кількість даних для аналізу та проведення досліджень.

Проте, навіть окремий район Оболонь налічує забагато точок, саме тому для проведення дослідження був взятий окремий мікрорайон Оболоні, який містить близько 20.000 точок, що зображені на рисунку 4.5.

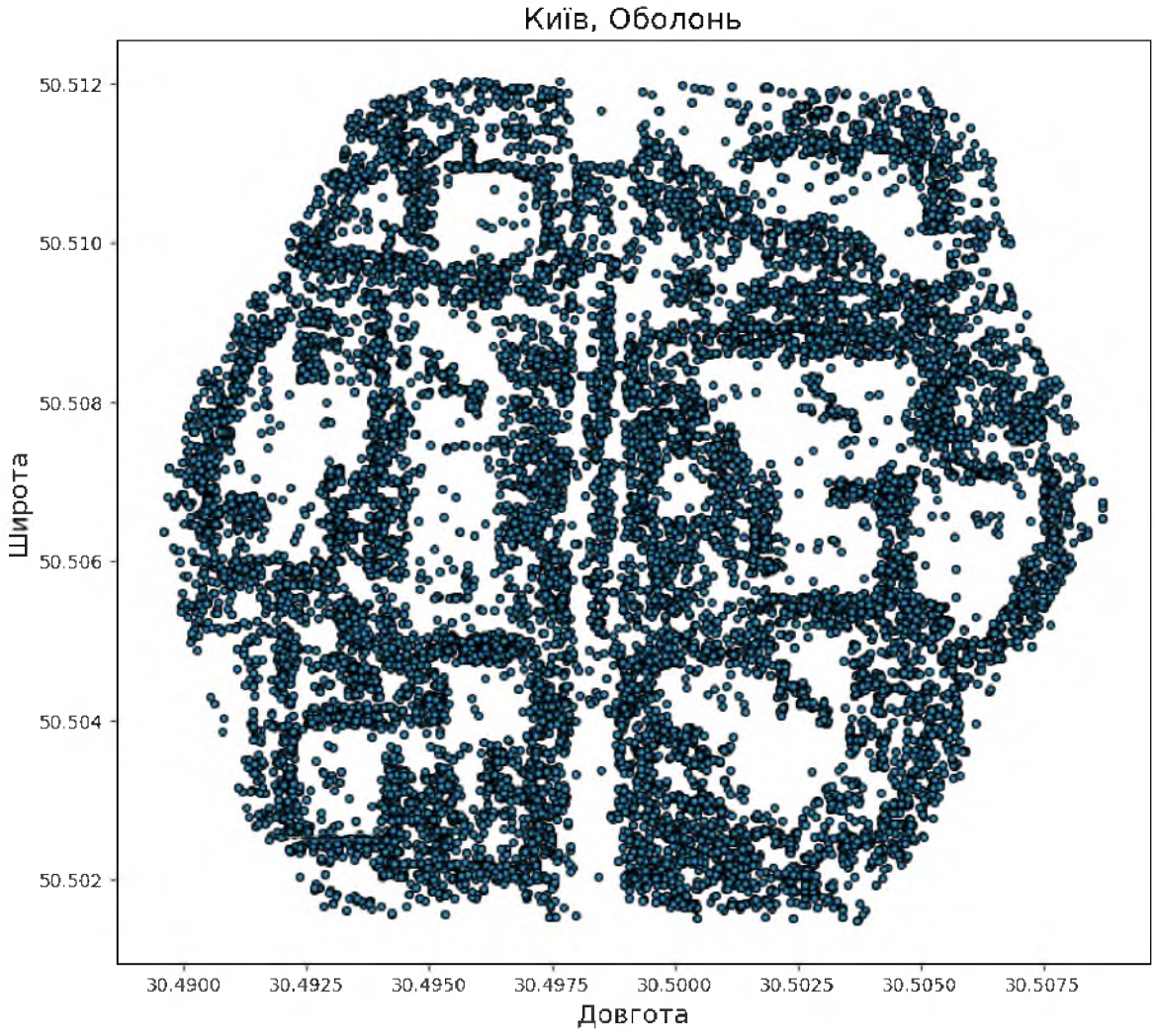

Рисунок 4.5 – Візуалізація локацій IoT пристроїв для мікрорайону Оболоні

Троєщина – це історичний та культурний район Києва, розташований на північному заході лівого берега міста, відомий своєю розвиненою інфраструктурою, багатством зелених зон та великими житловими масивами. Цей район є одним з найбільших та найбільш густонаселених у місті (близько 280.000 осіб), що робить його ідеальним об’єктом для дослідження в контексті кластеризації геоданих на прикладі вибірки з локаціями IoT пристроїв.

Обраний район включає в себе усю багатоповерхову забудову Троєщини та містить близько 110.000 точок для аналізу, візуалізація яких наведена на рисунку 4.6.

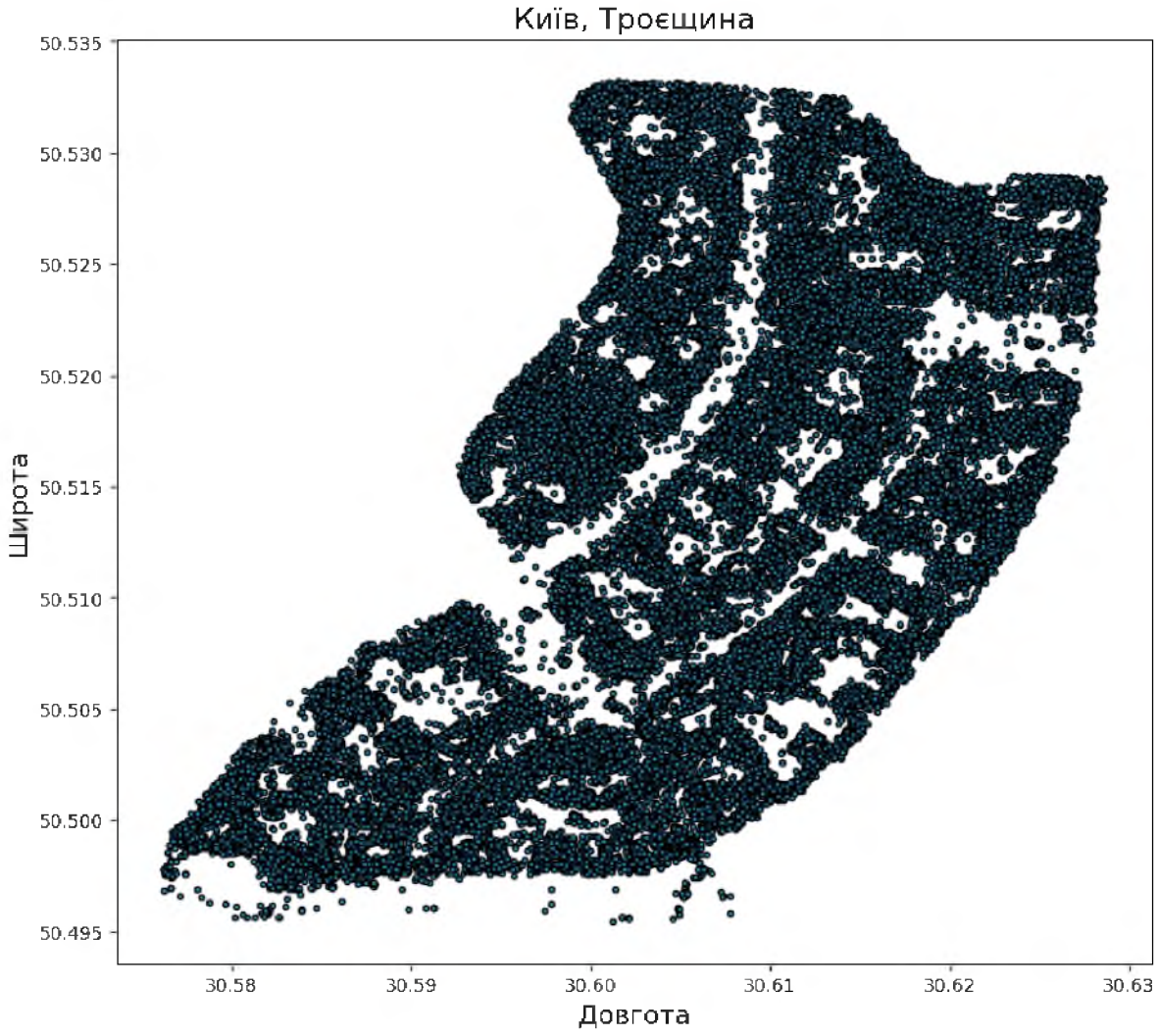

Рисунок 4.6 – Візуалізація локацій IoT пристроїв для району Троєщини

4.2 Кластеризація без використання дискретизації

4.2.1 K-Means

На рисунках 4.7 та 4.8 наведені найкращі результати кластеризації, що були отримані за допомогою метода K-Means для Оболоні та Троєщини відповідно, інші результати наведені на рисунках А.1, А.2, А.3 та А.4.

Для метода був встановлений максимальний поріг у 500 ітерацій та значення точності у 0.000001. Незважаючи на те, що метод потребує чітко зазначеної кількості кластерів, для кожного з обраних районів можливо визначити оптимальне значення, оскільки скупчення сигналів тісно корелюють з житловими приміщеннями та іншою нерухомістю. Таким чином, для розглянутого мікрорайону Оболоні у якому знаходиться понад 100 житлових та інших будівель оптимальним значенням кількості кластерів стало значення близьке до 150, а для усієї Троєщини, де знаходиться понад 850 об’єктів нерухомості, таким значенням було виявлено саме 950. На візуалізації результатів кластеризації можна побачити досить чітко окреслені контури багатоповерхових житлових будинків та торговельних центрів, де наявні скупчення різноманітного роду IoT пристроїв.

В цілому, отримані результати застосування алгоритму K-Means вказують складну та нелінійну природу кластерів і наявність достатньо високого відсотку (10-20%) шуму у даних. Складна форма кластерів свідчить про різноманіття та складність просторового розподілу даних, проте наявність шуму у даних створює виклики для точності кластеризації. Наявність високого відсотку шуму може бути пов’язана з нерегулярними аномаліями або похибками визначення геолокації, що впливає на якість та інтерпретацію результатів.

Отже, з огляду на виявлену нелінійну природу отриманих кластерів та високий відсоток шуму у даних, результати кластеризації методом K-Means

можуть бути не найкращими. Це пов'язано з обмеженнями самого алгоритму, який передбачає лінійно роздільні кластери. Тим не менш, отримані результати можуть стати основою для подальших порівнянь з іншими методами кластеризації, які враховують нелінійність та наявність шуму у даних.

Таким чином, хоча K-Means може не забезпечувати оптимальні результати у даному випадку, його використання дозволяє встановити базові точки відліку та розуміння для подальших досліджень і аналізу інших методів кластеризації.

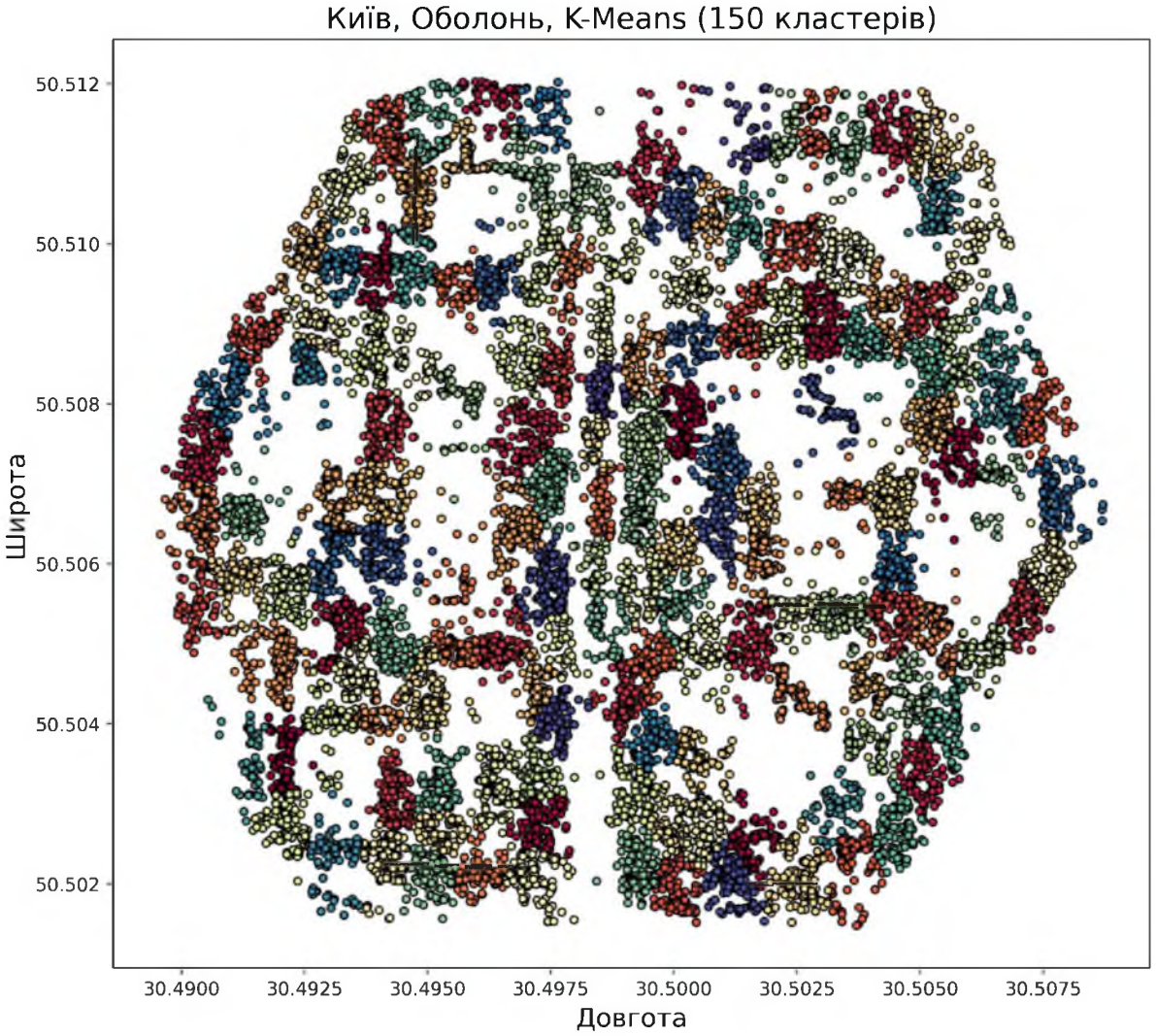

Рисунок 4.7 – Візуалізація кластерів мікрорайону Оболоні отриманих за допомогою методу K-Means

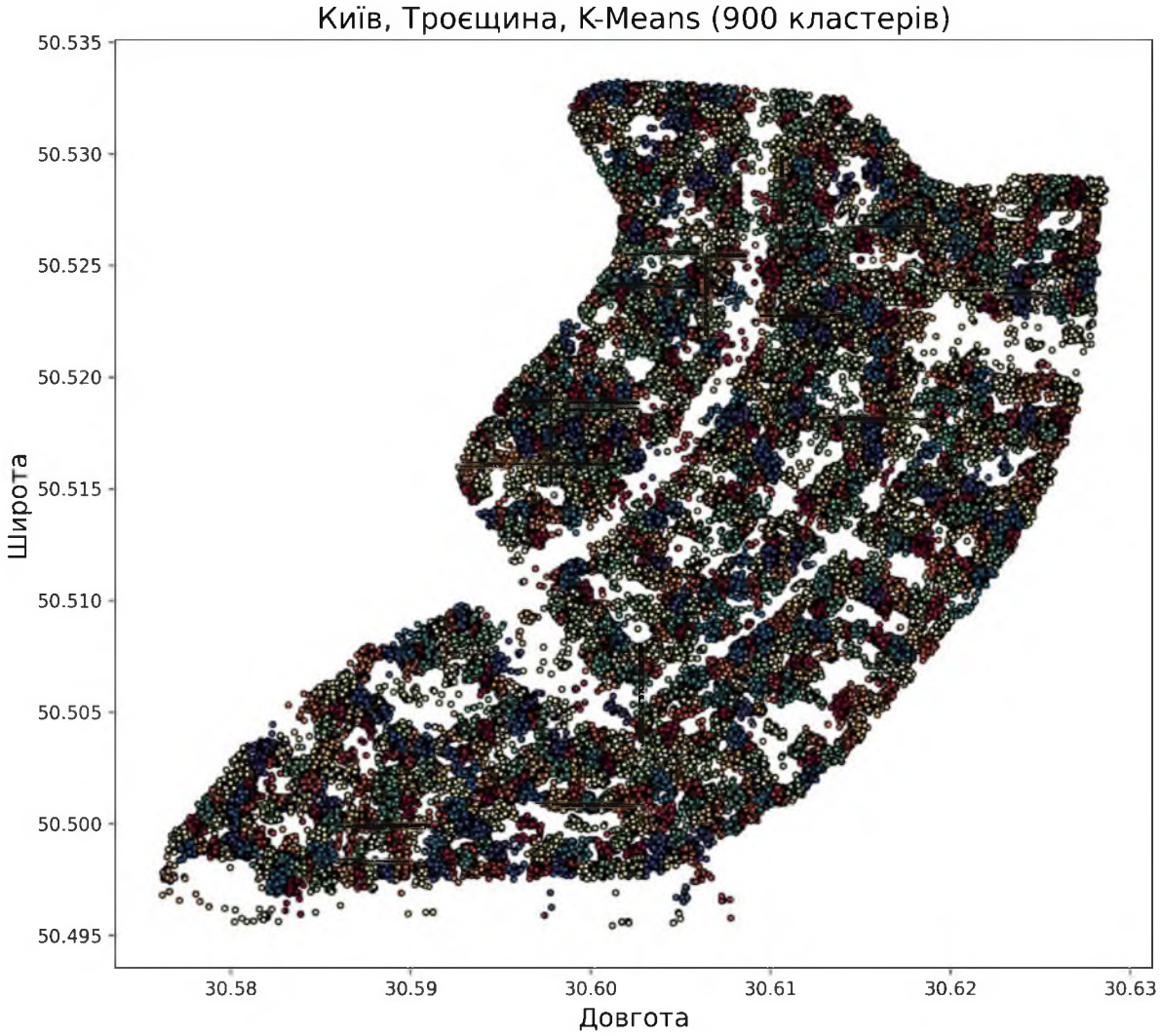

Рисунок 4.8 – Візуалізація кластерів району Троєщини отриманих за допомогою методу K-Means

4.2.2 Спектральна кластеризація

На рисунках 4.9 та 4.10 наведені найкращі наявні результати кластеризації, що були отримані з використанням методу спектральної кластеризації для Оболоні та Троєщини відповідно.

Спектральна кластеризація виявила низьку ефективність для обох розглянутих вибірок, оскільки метод не зміг чітко виділити географічні зони та створив лише кілька великих лінійно розділених зон.

Такі результати свідчать про обмежену здатність метода адаптуватися до складних просторових структур у даних та можуть бути пов'язані з неспроможністю методу працювати зі складними або нелінійно розділеними кластерами. Додатковою причиною незадовільної якості отриманих кластерів може виступати наявність високого відсотку аномалій та шумів в даних, що викривлюють межі кластерів.

Саме тому спектральна кластеризація не є оптимальним методом для аналізу даних вибірок і не може бути використана для подальших аналітичних або бізнес-застосувань.

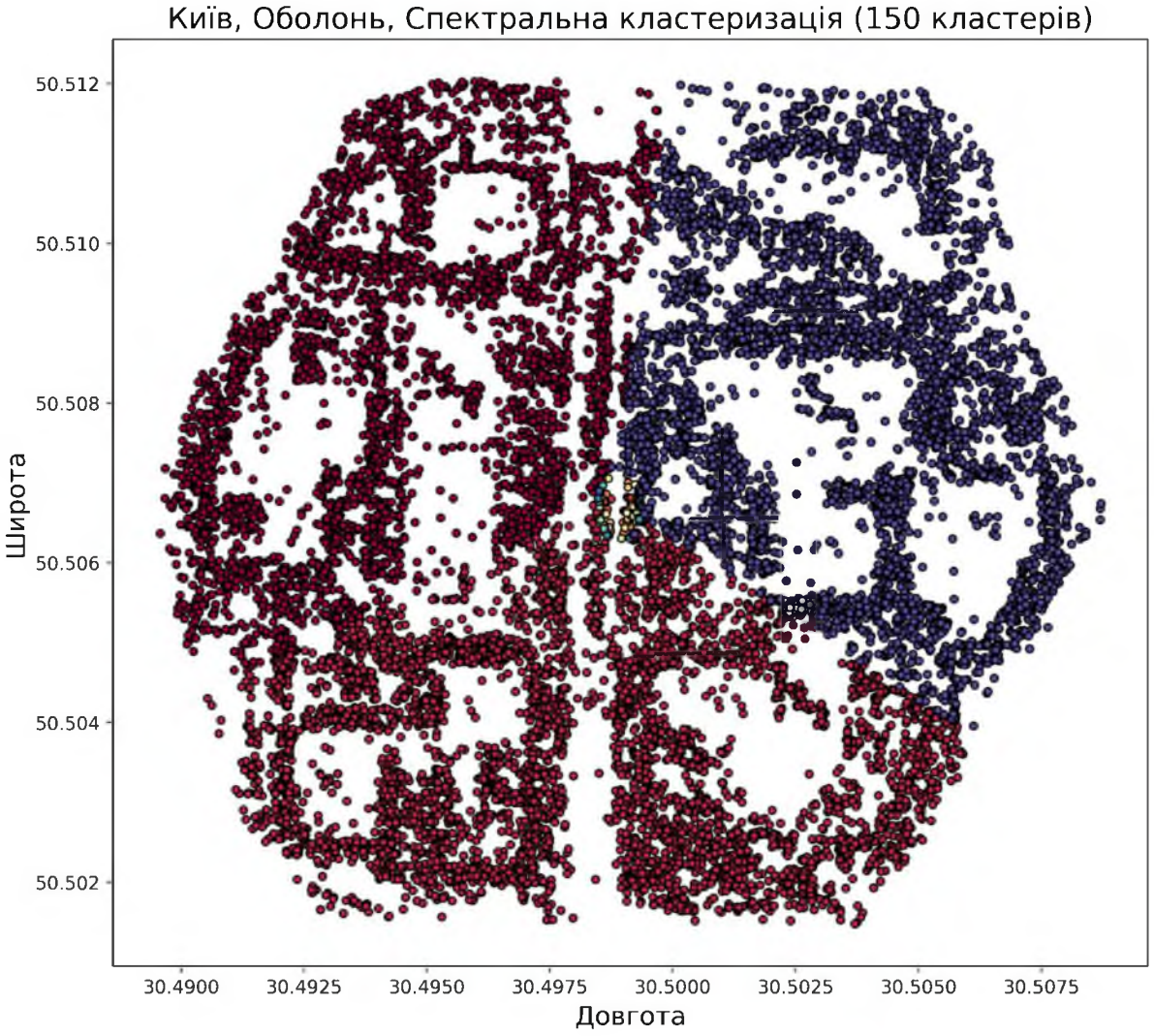

Рисунок 4.9 – Візуалізація кластерів мікрорайону Оболоні отриманих за допомогою методу спектральної кластеризації

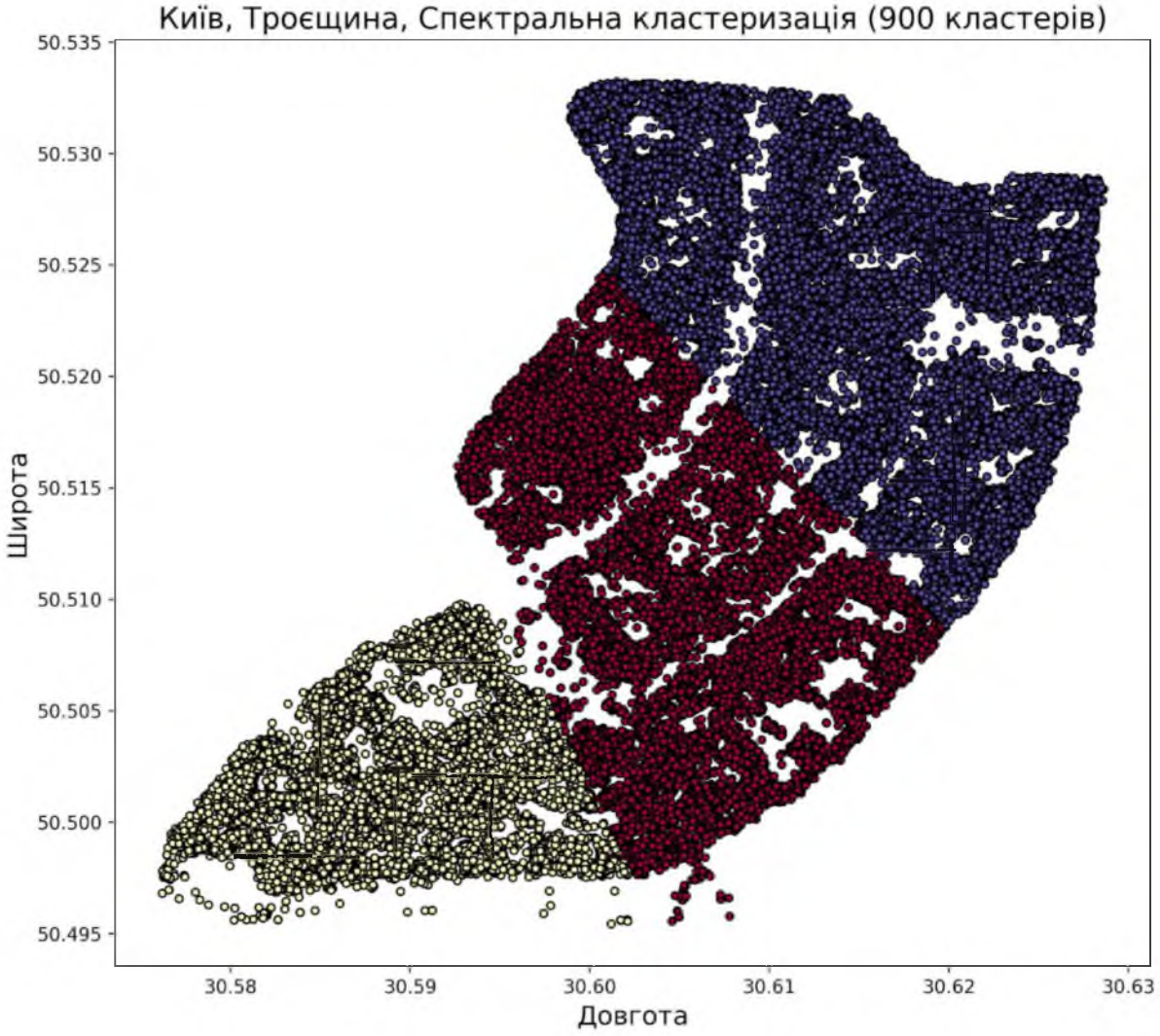

Рисунок 4.10 – Візуалізація кластерів району Троєщини отриманих за допомогою методу спектральної кластеризації

4.2.3 DBSCAN

На рисунках 4.11 та 4.12 наведені найкращі результати кластеризації, що були за допомогою метода DBSCAN для Оболоні та Троєщини відповідно, інші результати наведені на рисунках А.5, А.6, А.7 та А.8.

Використання DBSCAN привело до помітного поліпшення результатів порівняно з попередніми методами, такими як K-Means та спектральна кластеризація, оскільки цей метод знайшов близьку до прогнозованої кількості кластерів (188 проти 150 для Оболоні та 947 проти

900 для Троєщини) та відкинув частину даних, позначивши їх як шуми, що становили близько 20% відповідних вибірок, що значно покращило загальні результати.

Також отримані кластери відзначаються більш рівномірною та збалансованою формою у порівнянні з попередніми методами, що свідчить про краще врахування географічних особливостей даних. Це означає, що кластери, сформовані за допомогою даного методу, мають менші варіації у розмірі та геометрії, що полегшує їх інтерпретацію та аналіз, а також спрощує подальше використання.

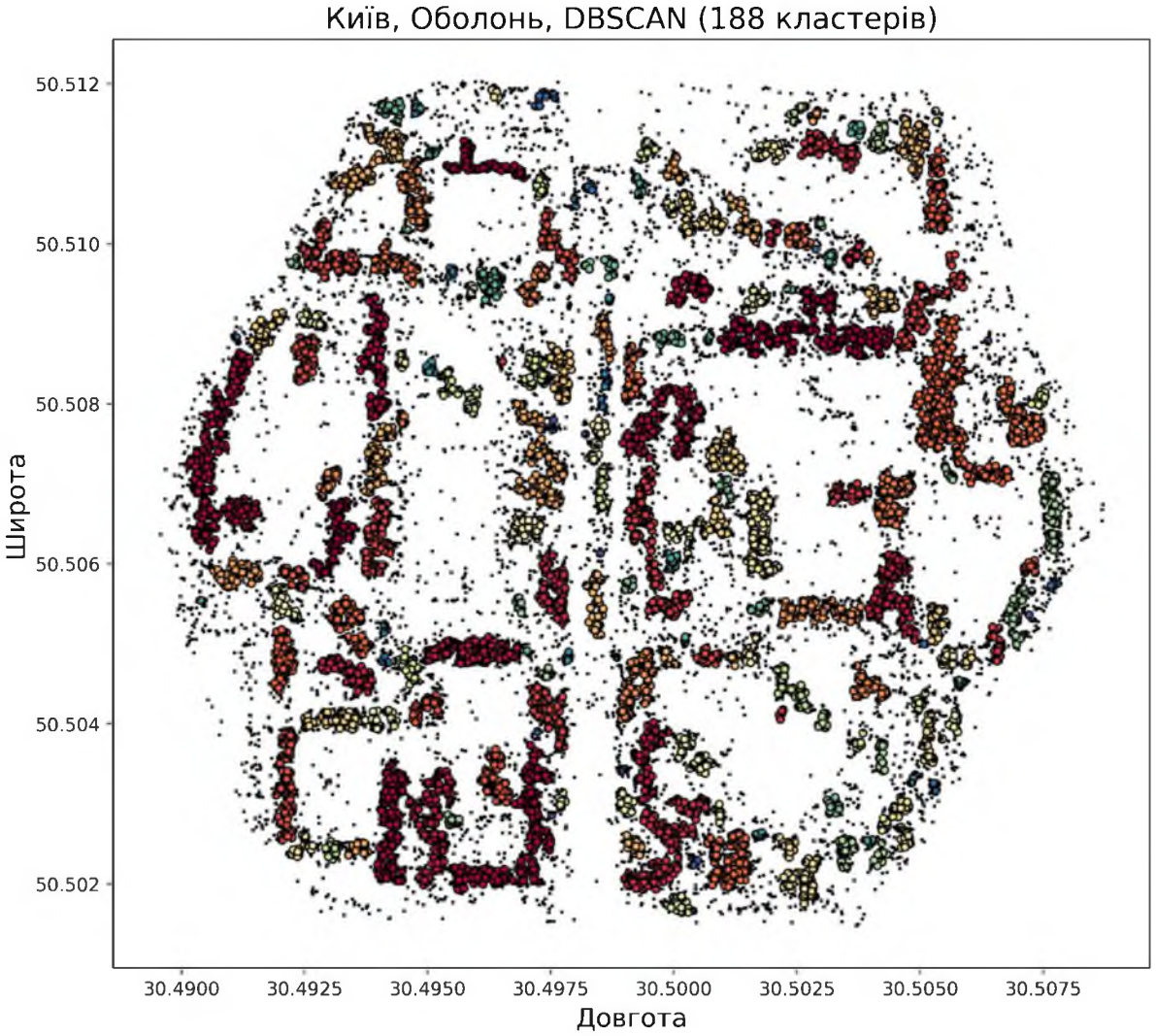

Рисунок 4.11 – Візуалізація кластерів мікрорайону Оболоні отриманих за допомогою методу DBSCAN

Значення параметру максимальної відстані від кластеру max_eps для DBSCAN було встановлено у 0.000145 через те, що координати знаходяться у проекції EPSG:4326, тобто у радіанах, і таке значення параметру встановлює відстань у близько 12 метрів, на якій сигнали будуть згруповані до одного кластеру. Така відстань з урахуванням похибки визначення геолокації в цілому є репрезентативною для аналізу.

Параметр мінімальної кількості точок, які можуть сформувати новий кластер min_pts був встановлений у 15 зразків, тобто кластер може складатись як мінімум з 15 різних пристроїв.

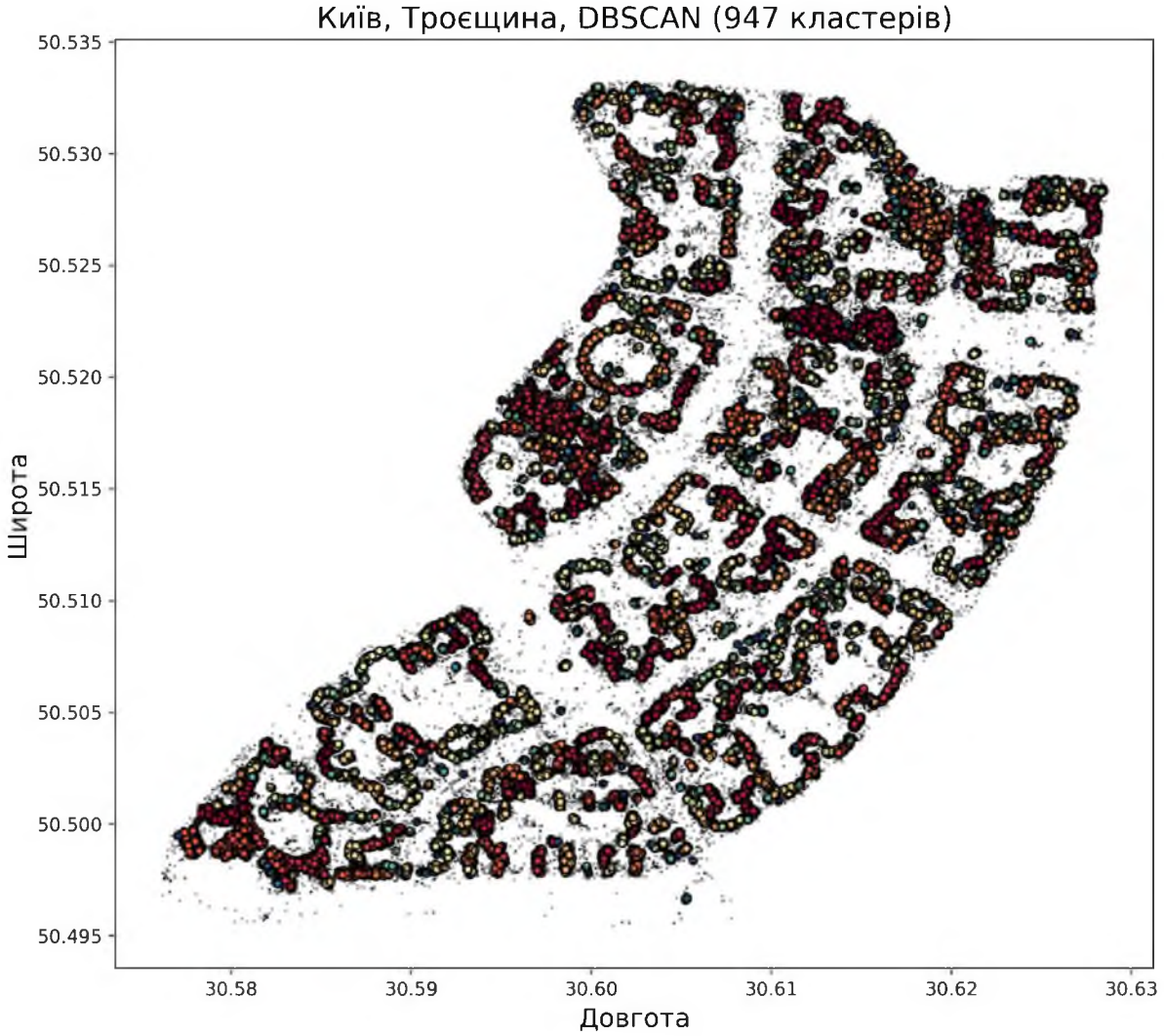

Рисунок 4.12 – Візуалізація кластерів району Троєщини отриманих за допомогою методу DBSCAN

4.2.4 OPTICS

На рисунках 4.13 та 4.14 наведені найкращі результати кластеризації для метода OPTICS для Оболоні та Троєщини відповідно, додаткові і проміжні результати наведені на рисунках А.9, А.10, А.11 та А.12.

Візуально результати використання OPTICS дуже подібні до результатів використання DBSCAN через схожість алгоритмів, проте OPTICS відсікає близько половини вибірки, розглядаючи її як свого роду аномалії чи викиди.

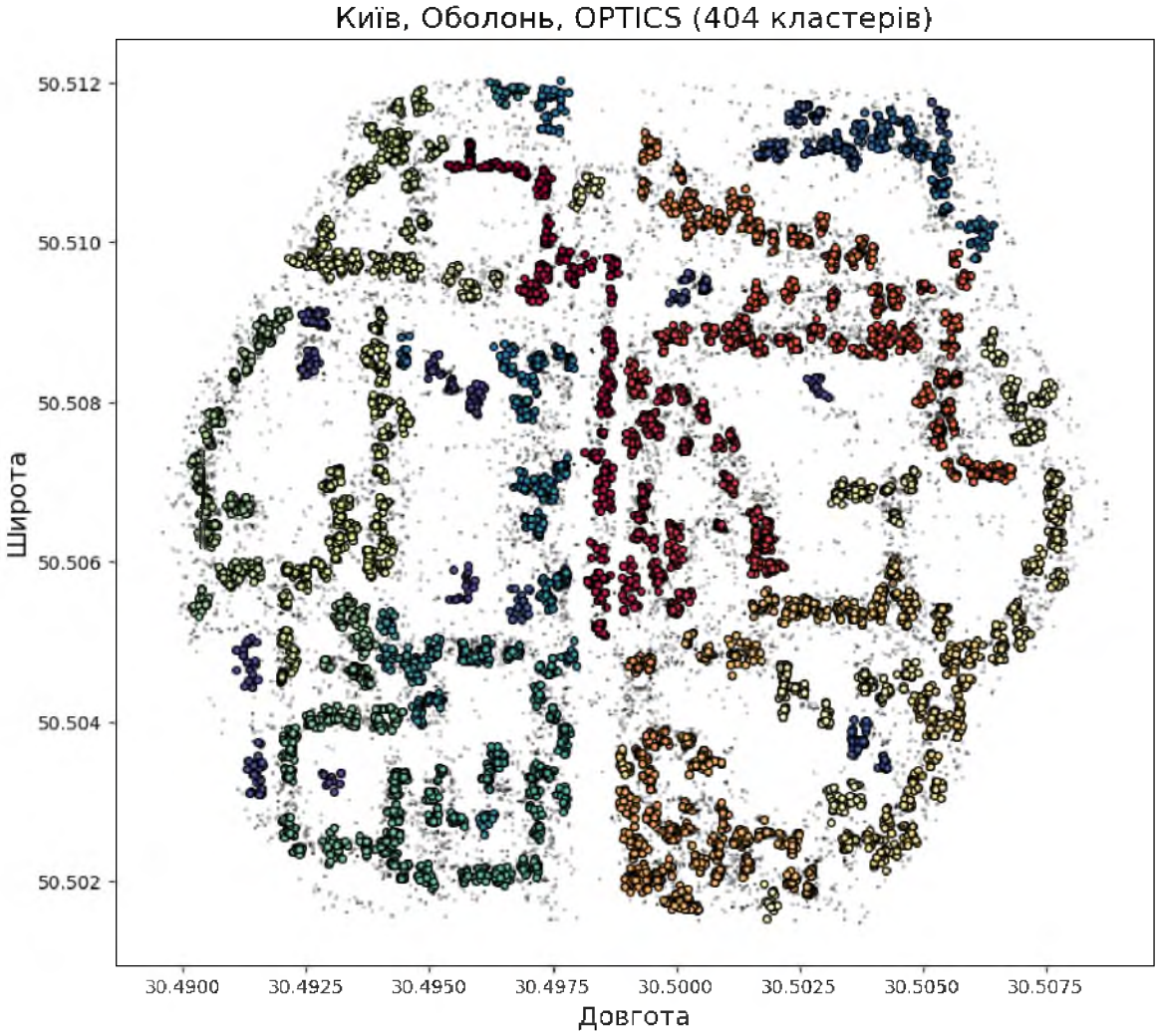

Рисунок 4.13 – Візуалізація кластерів мікрорайону Оболоні отриманих за допомогою методу OPTICS

Також, OPTICS виділяє більш як удвічі більше кластерів, ніж очікувалося (404 проти 150 для Оболоні та 2042 проти 900 для Троєщини), що може бути наслідком здатності алгоритму виявляти більш дрібні географічні зони та розділяти суцільні великі кластери на більш дрібні окремі підгрупи.

Тим не менш, кластери, сформовані у результаті використання метода OPTICS, мають більш чітку та виражену геометричну форму та краще описують особливості наведеної географічної зони, що полегшує їх подальшу інтерпретацію та аналіз.

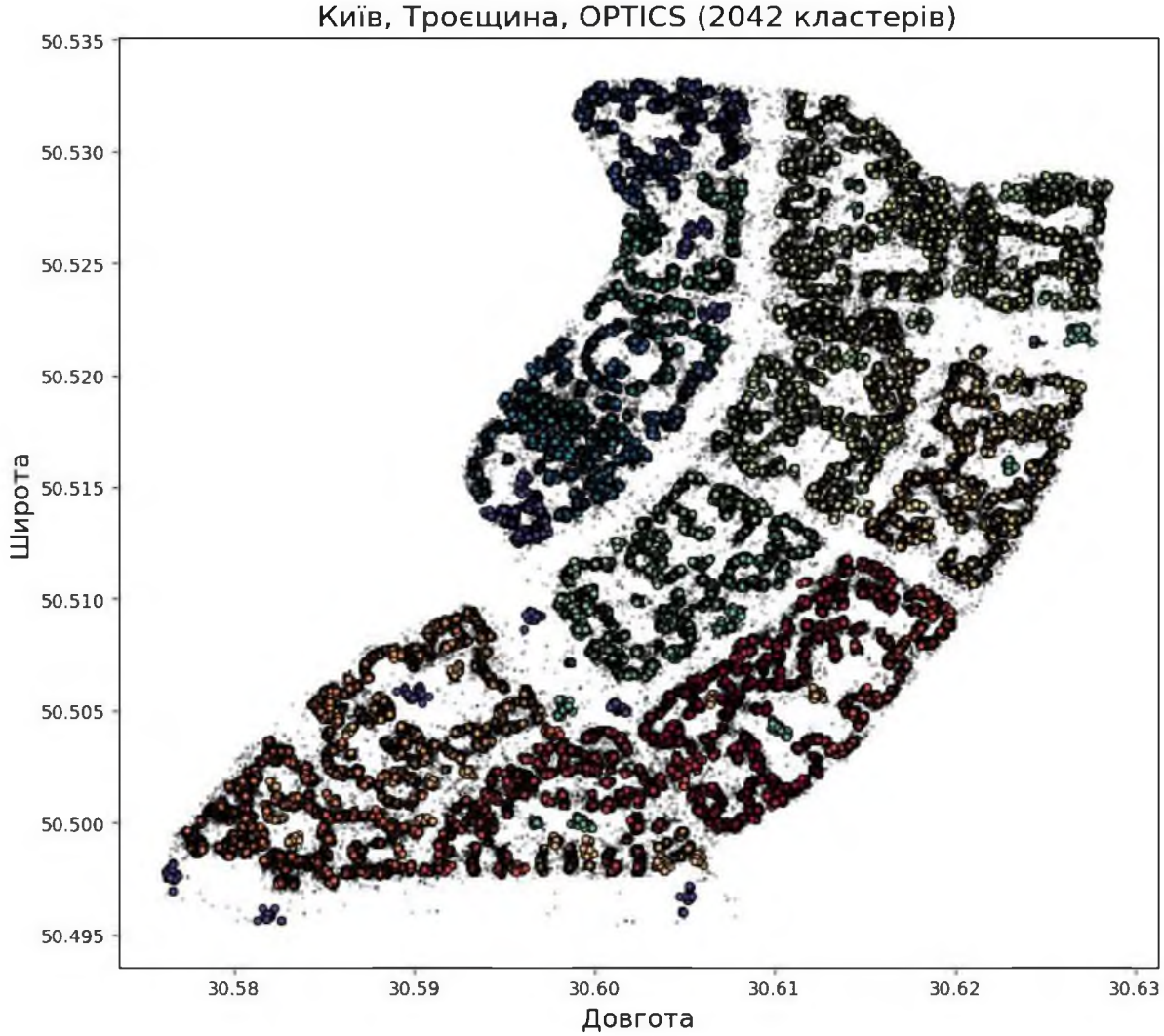

Рисунок 4.14 – Візуалізація кластерів району Троєщини отриманих за допомогою методу OPTICS

4.3 Кластеризація з використанням дискретизації

4.3.1 Фільтрація шумів

Отже, для обраних вибірок Оболоні та Троєщини основною проблемою при застосуванні таких методів кластеризації як K-Means та спектральна кластеризація стала наявність досить високого (близько 20%) відсотку шуму у даних. Для методів, які здатні фільтрувати шуми, до яких відносяться розглянуті DBSCAN та OPTICS наявність високого рівня шуму теж погіршує кінцеву якість кластеризації і ускладнює як обробку так і інтерпретацію результатів.

Для фільтрації шумів із вибірки досить ефективним виявляється застосування H3 дискретизації, оскільки кожна географічна точка відображається на конкретному дискретному сегменті і шуми, які потрапляють у віддалені або малочисельні гексагони, можуть бути легко виявлені та відкинуті з аналізу.

Основною підготовкою до використання дискретизації у якості інструмента для очистки вибірки є правильне визначення рівня деталізації, оскільки занадто високий рівень видалить забагато даних, у тому числі інформативних, а замалий рівень не може охопити достатню площу для групування і подальшої фільтрації точок.

Таким чином, основною підготовкою до використання дискретизації у якості інструмента для очистки вибірки є правильне визначення рівня деталізації, оскільки занадто високий рівень видалить забагато даних, у тому числі інформативних, а замалий рівень не може охопити достатню площу для групування і подальшої фільтрації точок.

Як для Оболоні так і для Троєщини потрібно обрати такий рівень деталізації, щоби сторона H3 гексагона була приблизно рівною до ширини панельного житлового будинку, оскільки саме вони виступають осередками скупчень локацій різноманітних IoT девайсів, а гексагони з надмірно

великою чи малою сторонами будуть викривлювати як візуальне відображення даних так і заважати якісній фільтрації шумів без втрати інформативності.

Таким чином, кращим рівнем деталізації H3 для візуалізації обраних вибірок стане 12-й рівень з середньою довжиною сторони гексагона у 10.5 метрів та середньою площею близько 200 метрів квадратних на розглянутих широтах. На рисунках 4.15 та 4.16 наведені візуалізації Оболоні та Троєщини відповідно з використанням H3 гексагонів 12-го рівня деталізації, візуалізація з використанням інших рівнів деталізації наведена на рисунках А.13, А.14, А.15, А.16, А.17 та А18.

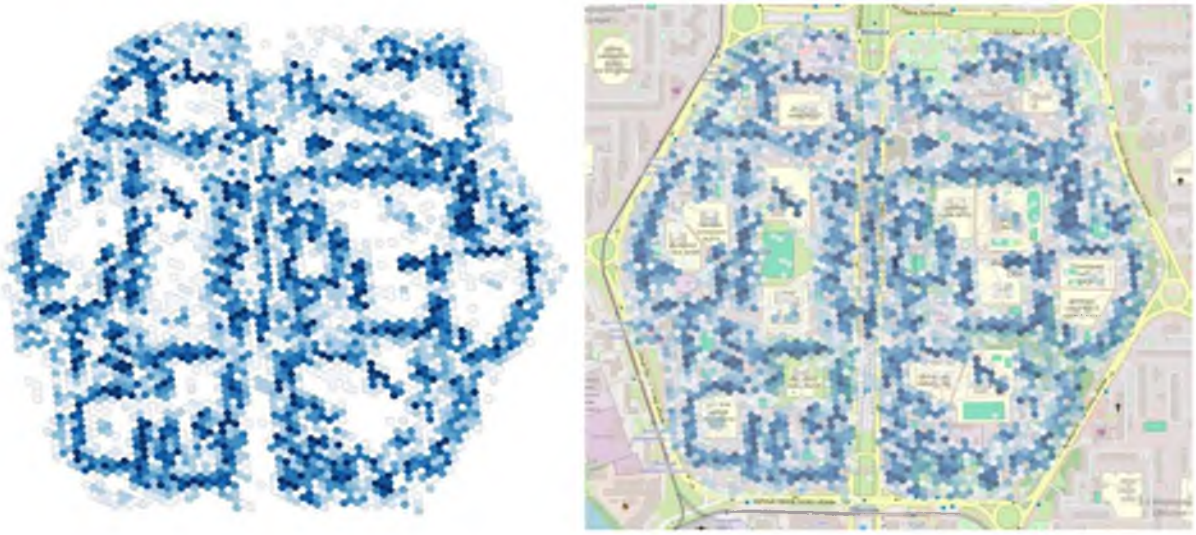

Рисунок 4.15 – Візуалізація місцеположень IoT пристроїв мікрорайону Оболоні з використанням H3 12-го рівня

На наведених візуалізаціях як для Оболоні так і для Троєщини чітко видно скупчення пристроїв у різних частинах житлових кварталів (більш темні гексагони), в основному в околицях будинків. Окрім скупчень можна зауважити і зони з помітно меншою кількістю пристроїв всередині них (світлі гексагони), в основному такі зони знаходяться в нетипових місцях, таких як середини проїжджих частин чи паркових зон, що може свідчити про переважаючу кількість шумів у таких гексагонах.

Рисунок 4.16 – Візуалізація місцеположень IoT пристроїв району Троєщини з використанням H3 12-го рівня

Отже, за допомогою візуалізації з використанням гексагонів H3 можливо виявити зони, що містять переважаючу кількість шумів та не несуть суттєвої інформативності для кластерного аналізу, оскільки представляють собою віддалені або малочисельні групи сутностей. Таким чином, відфільтрувавши ті пристрої, що потрапили у малочисельні гексагони (гексагони що містять менше 10 пристроїв) можна значно скоротити частку шумів у виборці без втрати інформативності. На рисунках 4.17 та 4.18 наведено візуалізація очищених таким методом вибірок з використанням H3 12-го рівня для Оболоні та Троєщини відповідно.

З наведених візуалізацій очищених вибірок для Оболоні і Троєщини легко зауважити значно меншу кількість віддалених гексагонів та гексагонів з невеликою кількістю точок. Також стали більш чіткими контури майбутніх кластерів та скупчень девайсів у житлових будинках, деякі скупчення гексагонів майже повністю повторюють периметри висотних житлових будівель.

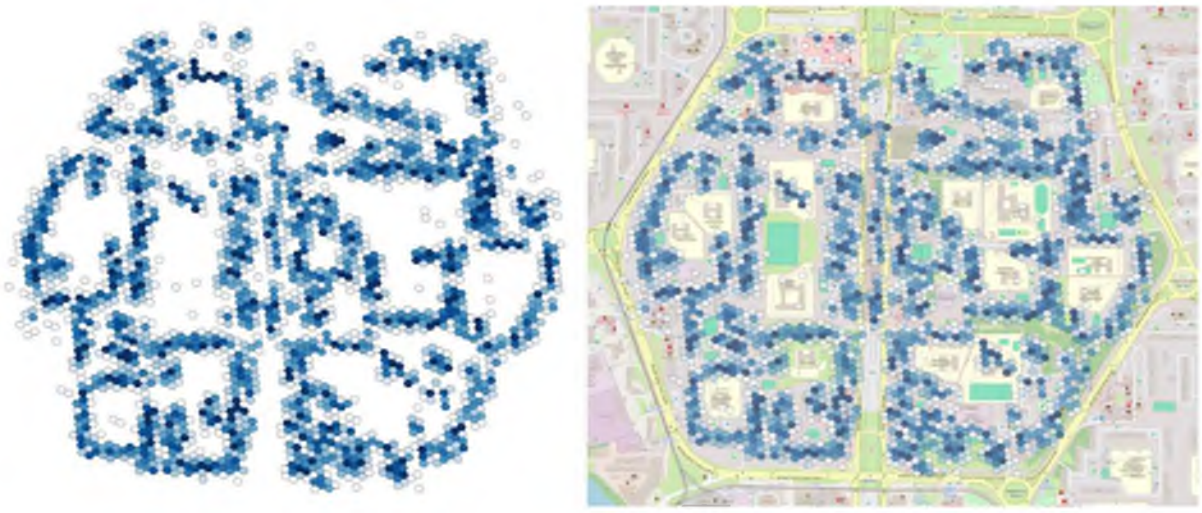

Рисунок 4.17 – Візуалізація знешулених місцеположень IoT пристроїв мікрорайону Оболоні з використанням H3 12-го рівня

Загалом вдалось відфільтрувати близько 18% та 19% шумів для Оболоні та Троєщини відповідно, що значно поліпшує результати кластеризації для чутливих до шумів методів. Крім цього H3 дискретизація дозволила зменшити розмір вибірки на 60% і 90% за рахунок агрегації окремих точок у гексагони для 12-го та 13-го рівнів деталізації відповідно.

Рисунок 4.18 – Візуалізація знешулених місцеположень IoT пристроїв району Троєщини з використанням H3 12-го рівня

4.3.2 K-Means

Завдяки попередній фільтрації шумів для вибірок Оболоні та Троєщини вдалось покращити якість кластеризації за допомогою використання методу K-Means з тими ж налаштуваннями що і в попередньому розділі, 150 кластерів для Оболоні та 900 для Троєщини.

Помітне поліпшення якості кластерів спостерігається для вибірки Оболоні, оскільки видалення шумів методу K-Means працювати з більш збалансованим набором даних, що призвело до формування більш структурованих кластерів та покращило якість кластеризації.

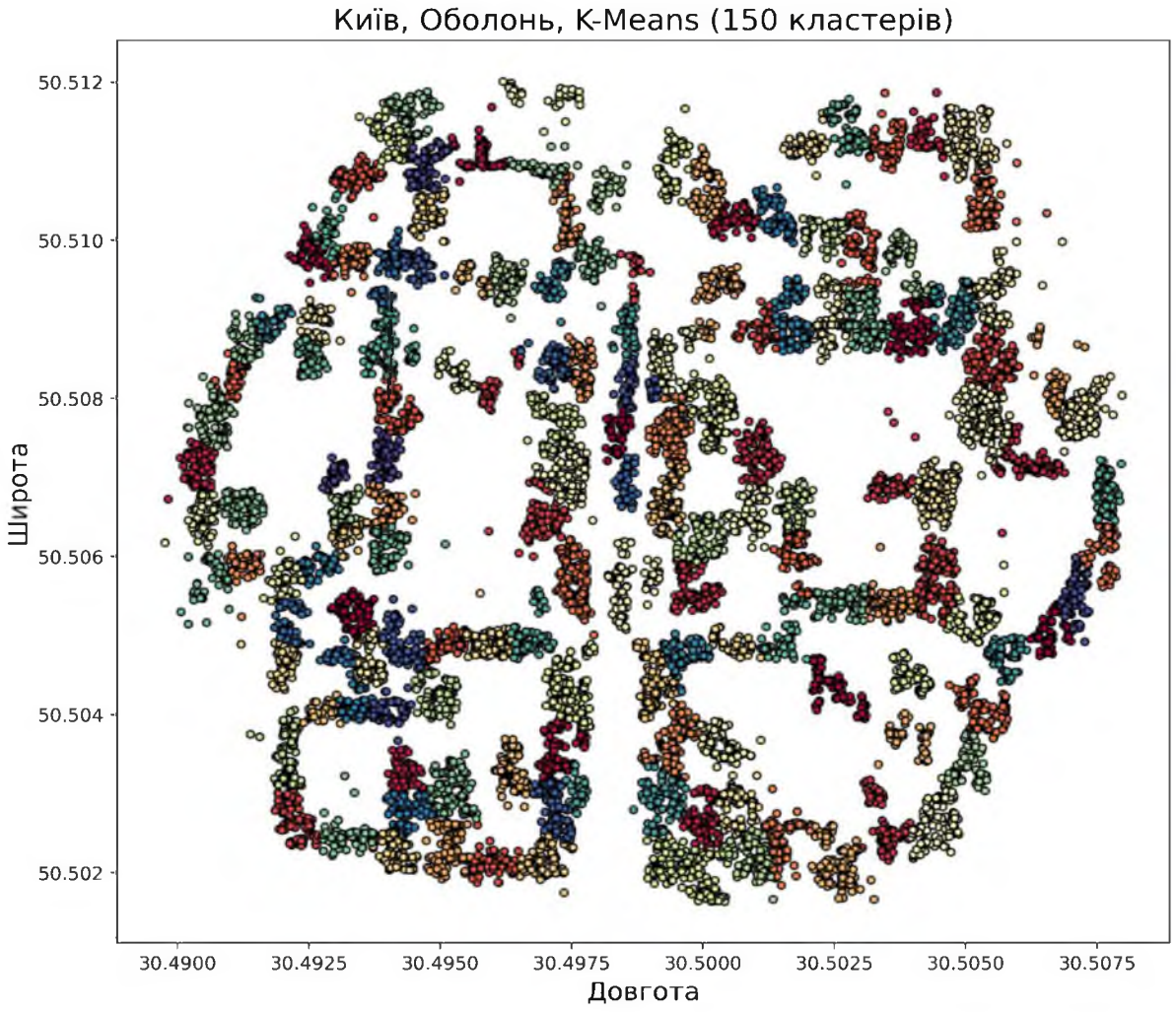

Рисунок 4.19 – Візуалізація кластерів знешумленого мікрорайону Оболоні, отриманих за допомогою методу K-Means

Троєщина є районом зі складною географічною структурою, саме тому попередня фільтрація шумів значно покращила отримані результати, оскільки вона допомогла виявити та видалити непотрібні аномалії, що сприяло кращому розумінню географічних зон і закономірностей розповсюдження скупчень IoT пристроїв.

В цілому, попередня фільтрація шумів як для Троєщини так і для Оболоні при використанні методу K-Means виявилася ефективним інструментом для покращення якості результатів, дозволяючи отримати більш точні та репрезентативні кластери.

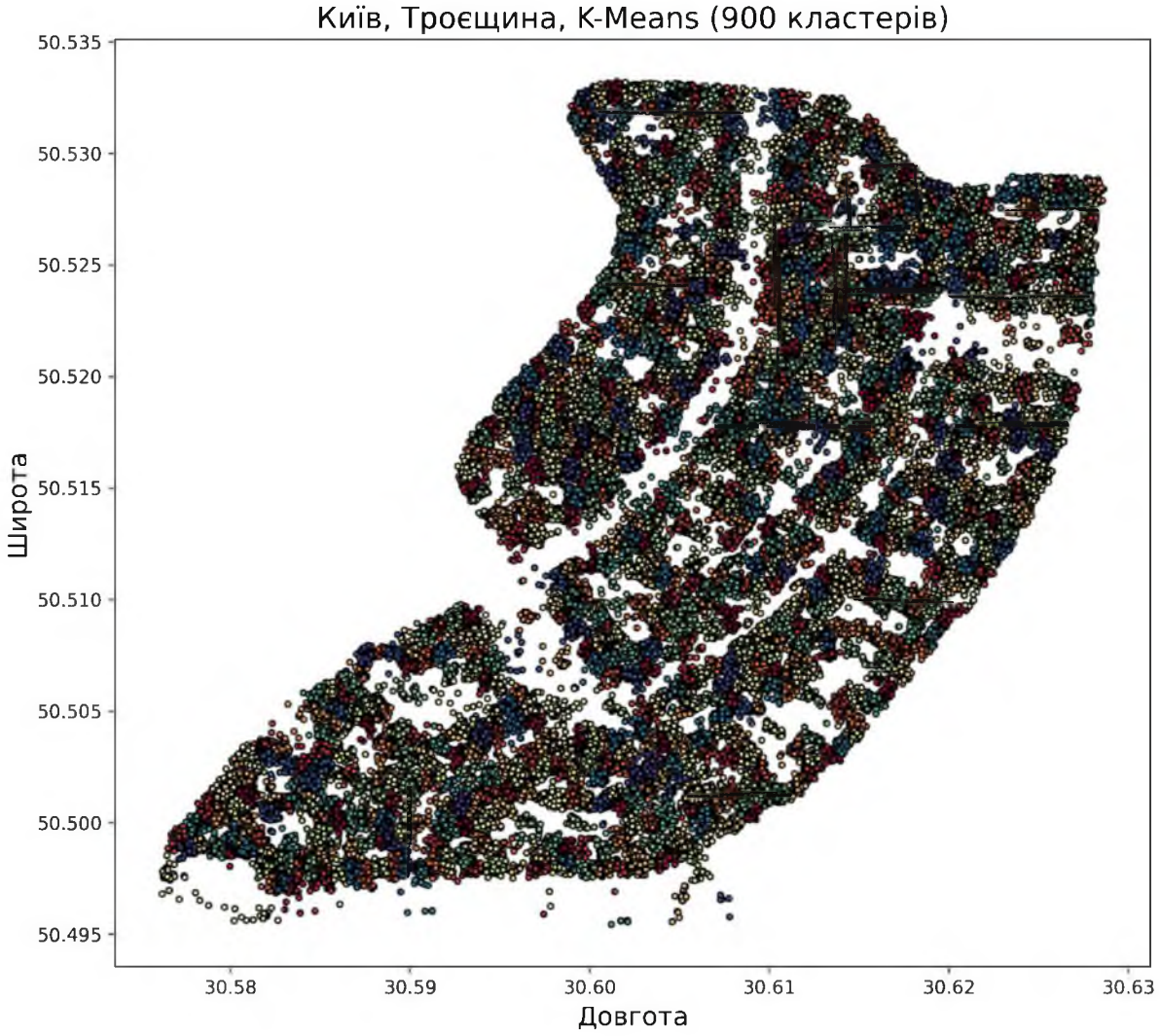

Рисунок 4.20 – Візуалізація кластерів знешумленого району Троєщини, отриманих за допомогою методу K-Means

4.3.3 Спектральна кластеризація

Незважаючи на поліпшення якості кластеризації з використанням метода K-Means, для метода спектральної кластеризації результати залишилися майже незмінними, візуалізація отриманих кластерів для Оболоні та Троєщини наведена на відповідних рисунках 4.21 та 4.22.

Отримана якість кластерів пов'язана з особливостями роботи методу, оскільки він ґрунтується на графових алгоритмах, де кожна точка даних виступає як вершина графа. Фільтрація шумів іноді може не вплинути на структуру графа, особливо якщо шум не утворює чіткі групи.

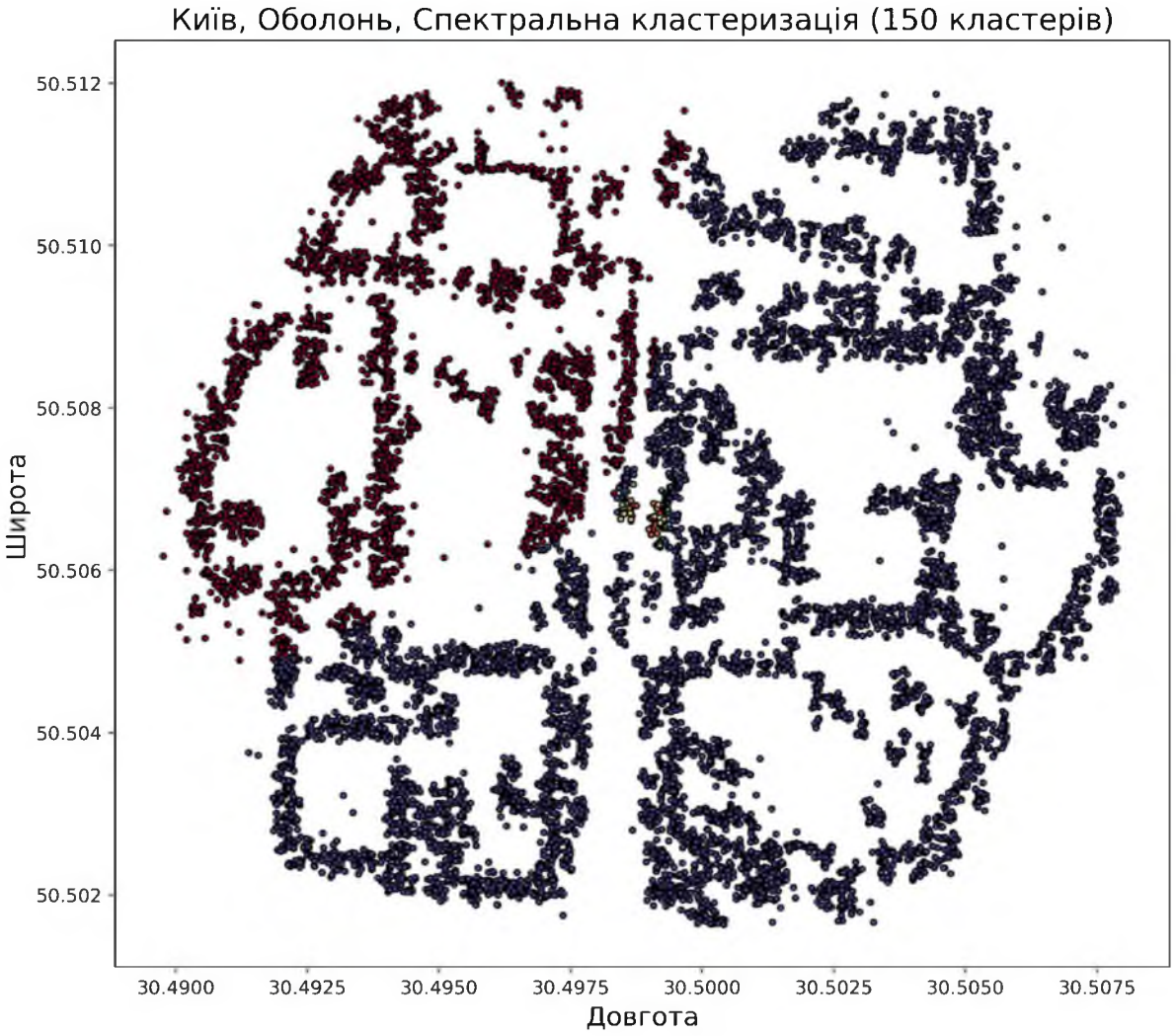

Рисунок 4.21 – Візуалізація кластерів знешумленого мікрорайону Оболоні, отриманих за допомогою методу спектральної кластеризації

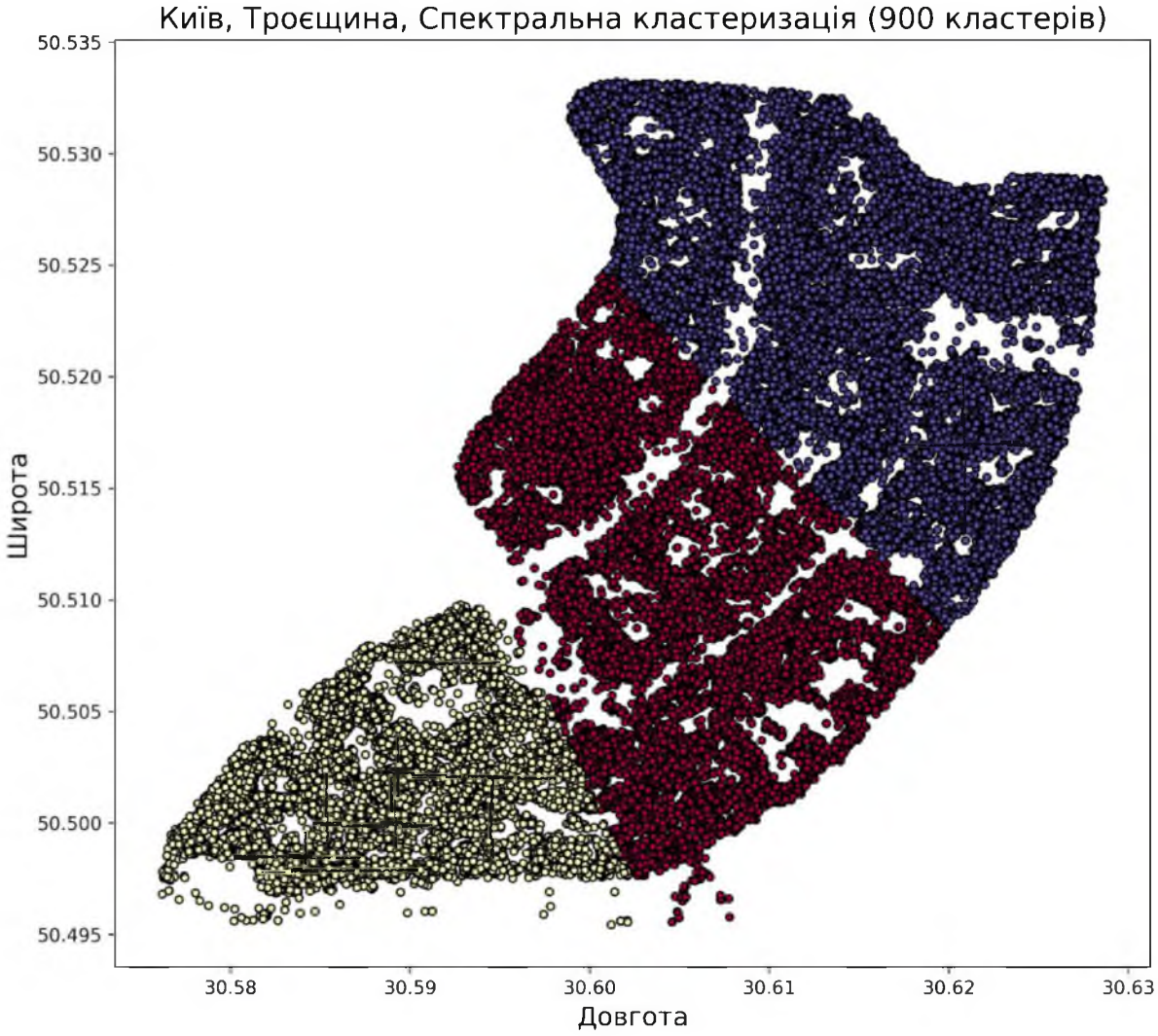

Рисунок 4.22 – Візуалізація кластерів знешумленого району Троєщини, отриманих за допомогою методу спектральної кластеризації

4.3.4 DBSCAN

На рисунках 4.23 та 4.24 наведені результати кластеризації на очищених вибірках з використанням методу DBSCAN для Оболоні та Троєщини відповідно.

Параметри були налаштовані так само як і у попередньому дослідженні з вибірками до фільтрації шумів, отримана кількість кластерів виявилась такою самою, з несуттєвими збільшеннями при зміні початкових параметрів.

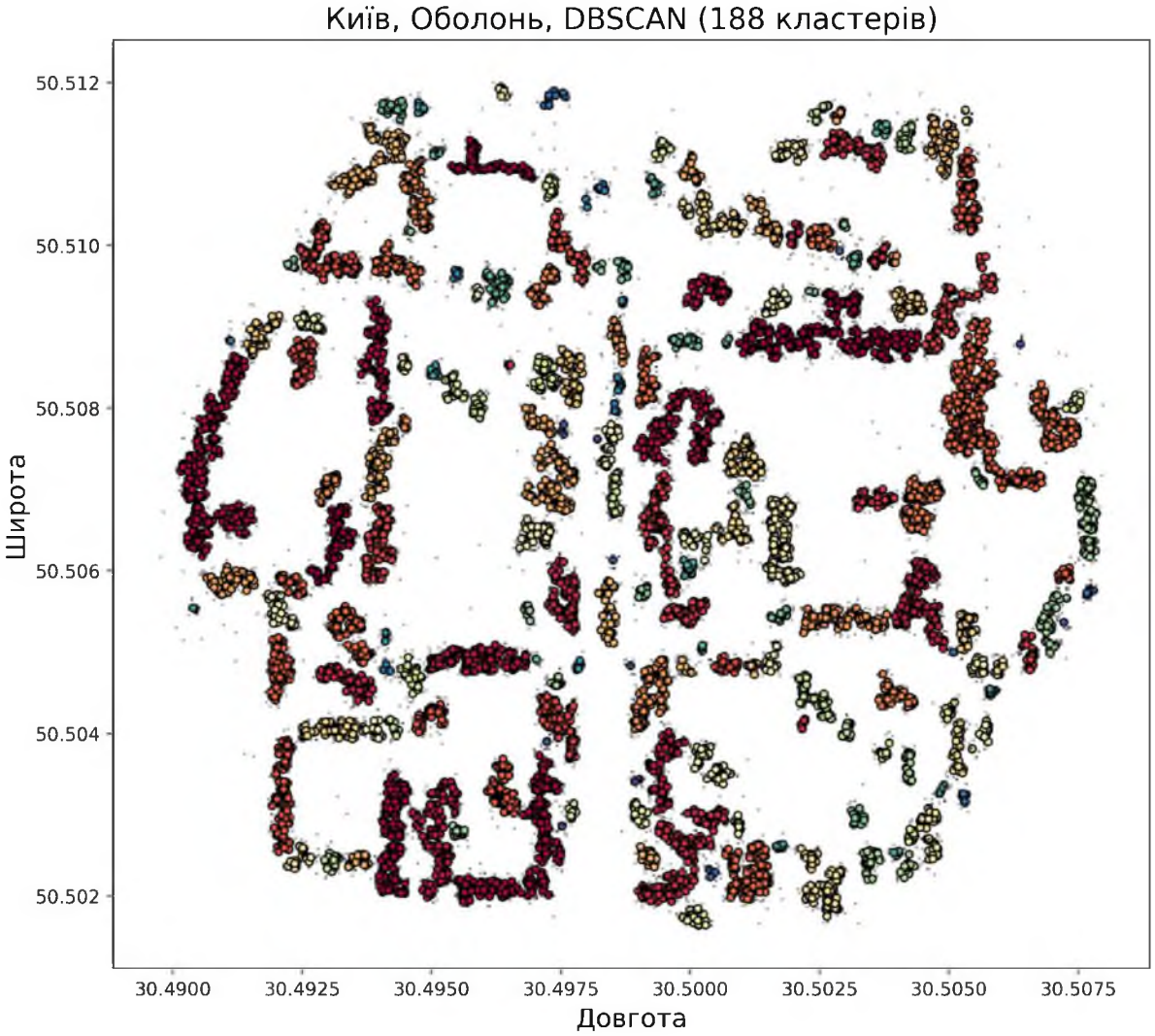

Рисунок 4.23 – Візуалізація кластерів знешумленого мікрорайону Оболоні, отриманих за допомогою методу DBSCAN

Загалом, після попередньої фільтрації шумів й так непогані результати кластеризації методом DBSCAN стали ще більш виразними. Таким чином, видалення шуму призвело до значного зменшення кількості аномальних та малоінформативних точок у наборі даних, що вплинуло на збільшення чіткості та різноманітності утворених кластерів при цьому не вплинувши суттєво на їх кількість.

Це стало можливим у більшості завдяки тому, що алгоритм при роботі зосередився на більш інформативних та репрезентативних структурах даних, позбавившись шумових впливів.

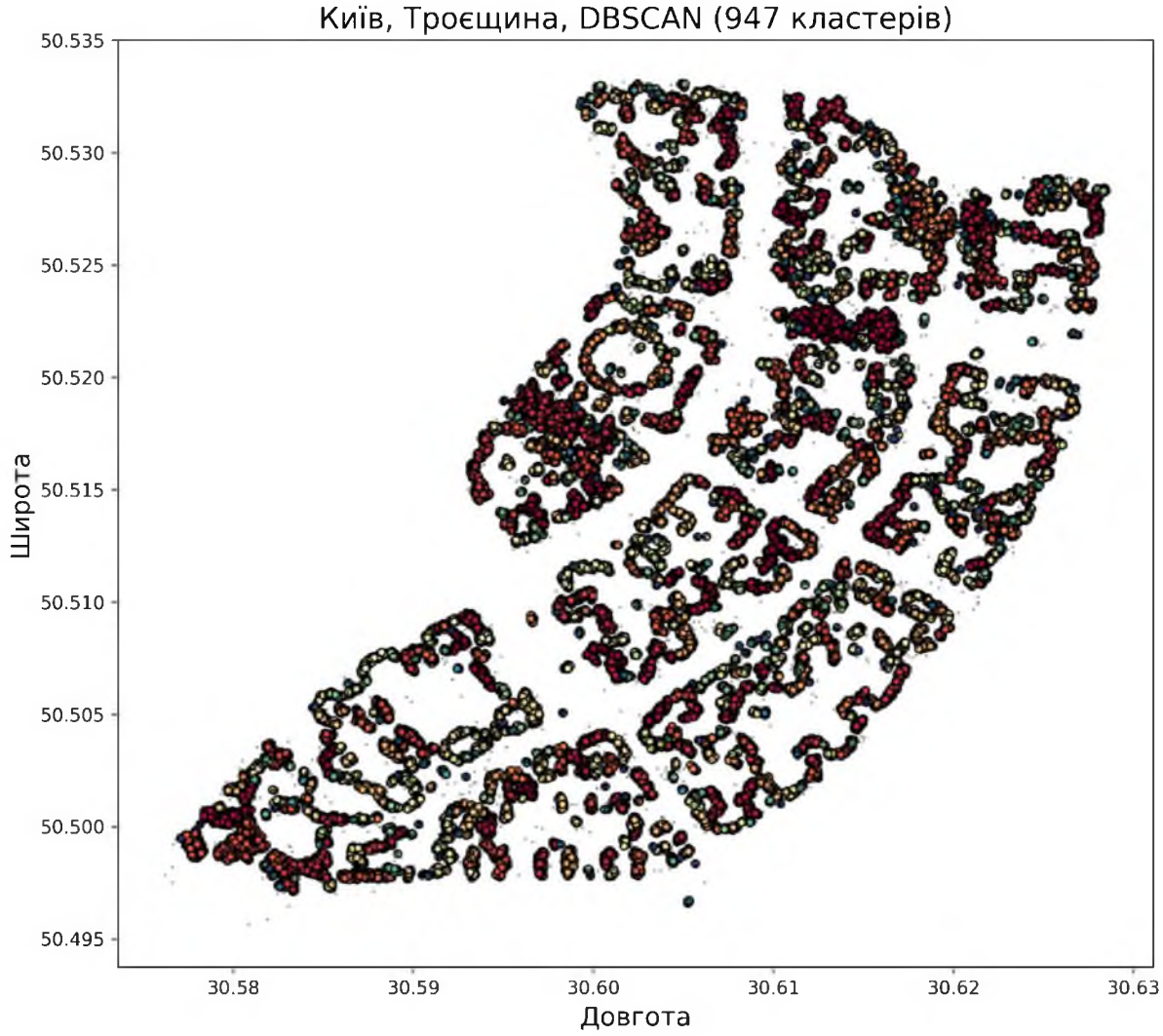

Рисунок 4.24 – Візуалізація кластерів знешумленого району Троєщини, отриманих за допомогою методу DBSCAN

4.3.5 OPTICS

Результати кластеризації з використанням методу OPTICS на очищених вибірках для Оболоні та Троєщини наведені на відповідних рисунках 4.25 та 4.26. Параметри для методу були налаштовані аналогічним попередньому дослідженню чином, проте кількість отриманих кластерів відчутно зросла.

Попередня фільтрація шумових точок дозволила методу OPTICS краще виділяти структури даних і формувати більш однорідні кластери.

Проте, це призвело до небажаного збільшення кількості кластерів (близько 20%), оскільки відфільтровані шумові точки могли об’єднувати менші кластери у єдині групи, що дозволяли методу OPTICS формувати порівняно більші кластери.

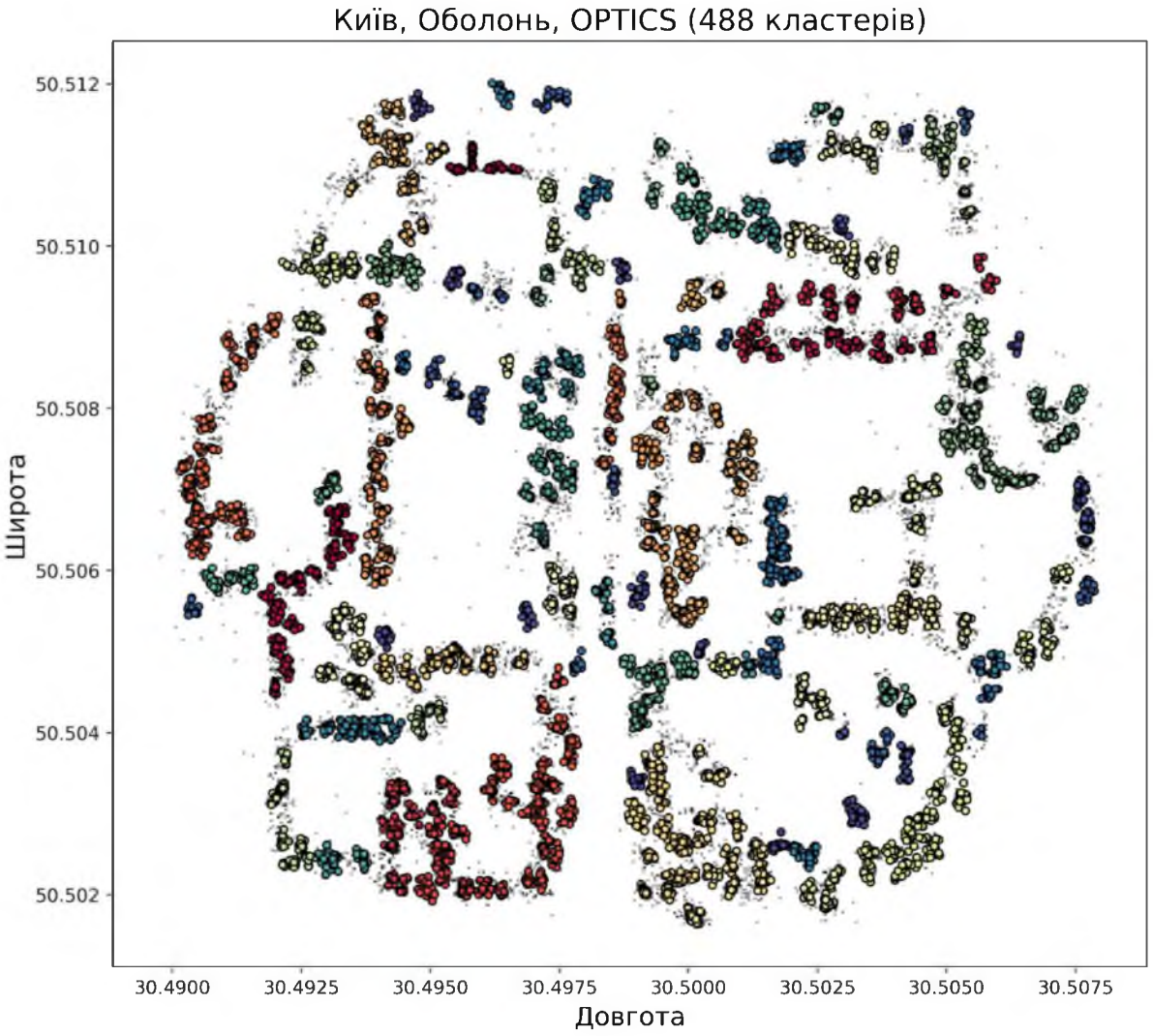

Рисунок 4.25 – Візуалізація кластерів знешумленого мікрорайону Оболоні, отриманих за допомогою методу OPTICS

Проте, незважаючи на збільшення кількості кластерів, попередня фільтрація шумів все ж поліпшила якість кластеризації, забезпечивши більш точне виявлення нетривіальних географічних зон та закономірностей розміщення різноманітних скупчень IoT пристроїв. Результати стали більш

структурованими і зрозумілими для подальшого аналізу та використання аналітиками.

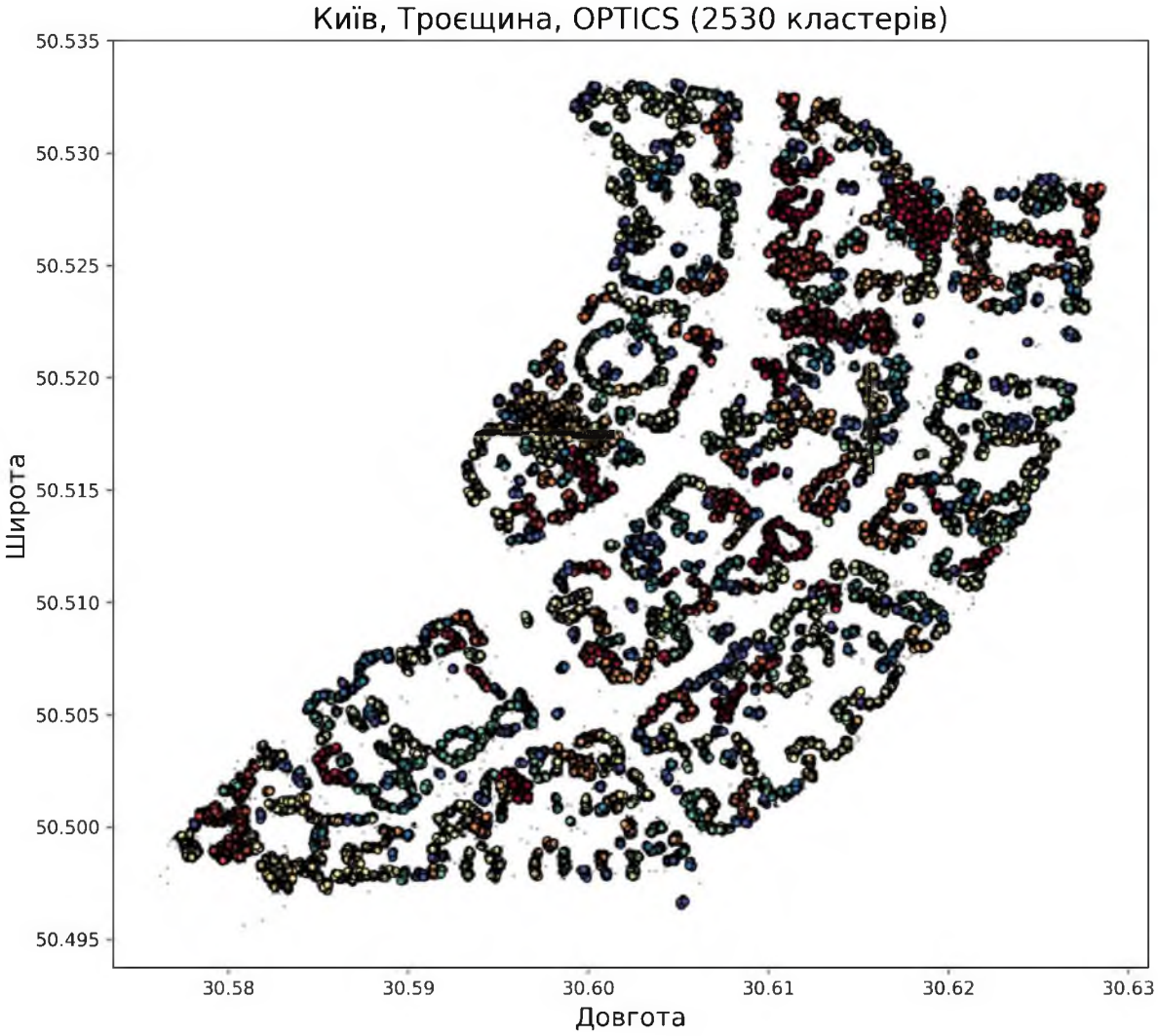

Рисунок 4.26 – Візуалізація кластерів знешумленого району Троєщини, отриманих за допомогою методу OPTICS

4.4 Висновки до практичного дослідження

Отже, практичне дослідження базувалось на вивченні кластеризації геоданих з використанням дискретизації на основі H3 гексагонів для фільтрації шумів та візуалізації методів отриманих кластерів для методів K-Means, спектральної кластеризації, DBSCAN та OPTICS.

Підходи, засновані на методі K-Means та DBSCAN, показали найкращі результати після проведення фільтрації шумів та застосування дискретизації. Кластери, сформовані за допомогою K-Means, стали більш чіткими та однорідними, що призвело до значного покращення результатів цього методу. DBSCAN також показав позитивні зміни, суттєво покращивши результати кластеризації та відкидаючи на 25% менше даних, вважаючи їх шумами.

Щодо спектральної кластеризації, вона продемонструвала незадовільні результати як до, так і після фільтрації шумів. Це може бути пов'язане з чутливістю цього методу до параметрів та складності визначення структури графа, що ускладнює виявлення кластерів.

Нарешті, метод OPTICS, хоч і відкинув на 20% менше даних як шуми, сформував на 15% більше дрібних кластерів у порівнянні з оригінальною вибіркою та більш ніж на 180% у порівнянні з прогнозованою кількістю. Незважаючи на те, що кластери стали чіткішими, цей метод не показав такого значного покращення результатів, як DBSCAN.

Загалом, після фільтрації шумів методи K-Means та DBSCAN продемонстрували найкращі результати, тоді як спектральна кластеризація та OPTICS потребують подальшого вдосконалення для ефективного використання в аналізі геоданих.

ВИСНОВКИ

У ході проведення наукового-практичного дослідження було виявлено, що використання кластеризації геоданих з застосуванням дискретизації має великий потенціал у багатьох сферах, що вимагають аналізу географічної інформації. Застосування дискретизації дозволяє очистити вибірку від шумів а також зменшити обсяг даних, зберігаючи при цьому суттєву частину інформації, що полегшує подальший аналіз та обробку великого даних. Такий підхід може покращити точність кластеризації геоданих, зменшуючи вплив шуму та випадкових аномалій на результати аналізу.

Кластеризація геоданих з використанням дискретизації має широкий спектр застосувань, включаючи геологічні дослідження, маркетингові аналізи, управління ресурсами та інші галузі. Незважаючи на досягнені результати, існує потреба у подальших дослідженнях для вдосконалення існуючих методів кластеризації для геоданих з використанням дискретизації, а також для розробки нових підходів і алгоритмів.

Використання дискретизації спрощує обробку та аналіз геоданих, забезпечуючи зручність та ефективність використання результатів в різних сферах діяльності. Таким чином, науково-практичне дослідження кластеризації геоданих з використанням дискретизації підкреслює можливості у покращенні аналізу та використанні географічної інформації для найрізноманітніших цілей.

Практичне дослідження кластеризації для методів K-Means, DBSCAN та OPTICS на прикладі вибірок локацій IoT пристроїв для районів Києва Оболоні та Троєщини показало ефективність використання дискретизації на основі H3 гексагонів як для фільтрації шумів так і для візуалізації агрегованих результатів.

Таким чином, результати практичного дослідження підтвердили, що застосування методу H3 для дискретизації географічних даних є

ефективним підходом для підготовки даних до кластеризації. Він дозволяє не лише виявити та видалити шуми з даних, а й забезпечує зручну форму представлення для подальшого аналізу. Крім того, дискретизація на основі H3 дозволяє візуалізувати результати кластеризації в зрозумілому для подальшої інтерпретації форматі. Гексагональні клітини створюють графічний зображення географічних зон, що допомагає зрозуміти розподіл пристроїв та їх кластери.

Отже, результати дослідження свідчать про перспективність та потенціал методів кластеризації геоданих з використанням дискретизації в різних областях науки та практики. Ці методи можуть стати ефективним інструментом для виявлення взаємозв'язків та шаблонів у географічних даних, що відкриває широкі перспективи для застосування в різноманітних сферах, від екології та геології до бізнесу та міського планування.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Haining, R.P., Spatial Data Analysis: Theory and Practice, 2003.

2. Anderberg, M. Cluster Analysis for Applications. Academic Press, New York, USA, 1973.

3. Anil, K.J., Murthy, M.N., Flynn, P.J., Data Clustering: A Review. ACM Computing Surveys (CSUR), vol. 31, no. 3, pp. 264-323, 1999.

4. Aggarwal, C.C., Reddy, C.K., Data clustering: Algorithms and applications. CRC Press, 2013.

5. Halkidi, M., Batistakis, Y., Vazirgiannis, M., On clustering validation techniques. Journal of Intelligent Information Systems, 2001.

6. Clustering by measuring local direction centrality for data with heterogeneous density and weak connectivity. URL: https://www.nature.com/articles/s41467-022-33136-9 (дата звернення: 22.04.2024).

7. Steps to calculate centroids in cluster using K-means clustering algorithm. URL: https://www.datasciencecentral.com/steps-to-calculate-centroids-in-cluster-using-k-means-clustering (дата звернення: 23.04.2024).

8. A Tutorial on Spectral Clustering. URL: https://people.csail.mit.edu/dsontag/courses/ml14/notes/Luxburg07_tutorial_spectral_clustering.pdf (дата звернення: 24.04.2024).

9. DBSCAN Clustering Algorithm in Machine Learning. URL: https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html (дата звернення: 25.04.2024).

10. OPTICS: Ordering Points to Identify the Clustering Structure. URL: https://www.researchgate.net/publication/221214752_OPTICS_Ordering_Points_to_Identify_the_Clustering_Structure (дата звернення: 26.04.2024).

11. Tiling Tools. URL: https://spectus.ai/web-apps/tiling-tools/ (дата звернення: 28.04.2024).

12. Using discretization for extending the set of predictive features. URL: https://asp-eurasipjournals.springeropen.com/articles/10.1186/s13634-018-0528-

x (дата звернення: 28.04.2024).

13. Bing Maps Tile System. URL: https://learn.microsoft.com/en-us/bing-maps/articles/bing-maps-tile-system (дата звернення: 29.04.2024).

14. S2 Geometry. URL: http://s2geometry.io (дата звернення: 29.04.2024).

15. Hexagonal hierarchical geospatial indexing system. URL: https://h3geo.org (дата звернення: 29.04.2024).

16. Selecting a geo-representation. URL: https://medium.com/@claude.ducharme/selecting-a-geo-representation-81afeaf3bf01 (дата звернення: 01.05.2024).

17. Geospatial Indexing Explained: A Comparison of Geohash, S2, and H3. URL: https://benfeifke.com/posts/geospatial-indexing-explained/ (дата звернення: 01.05.2024).

18. Internet of Things (IoT): Opportunities, issues and challenges towards a smart and sustainable future. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7368922/ (дата звернення: 02.05.2024).

19. Internet of Things is a revolutionary approach for future technology enhancement: a review. URL: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0268-2 (дата звернення: 03.05.2024).

20. IoT Maps: Charting the Internet of Things. URL: https://www.researchgate.net/publication/331292415_IoT_Maps_Charting_the_Internet_of_Things (дата звернення: 04.05.2024).

ДОДАТОК А

Додаткові результати кластеризації

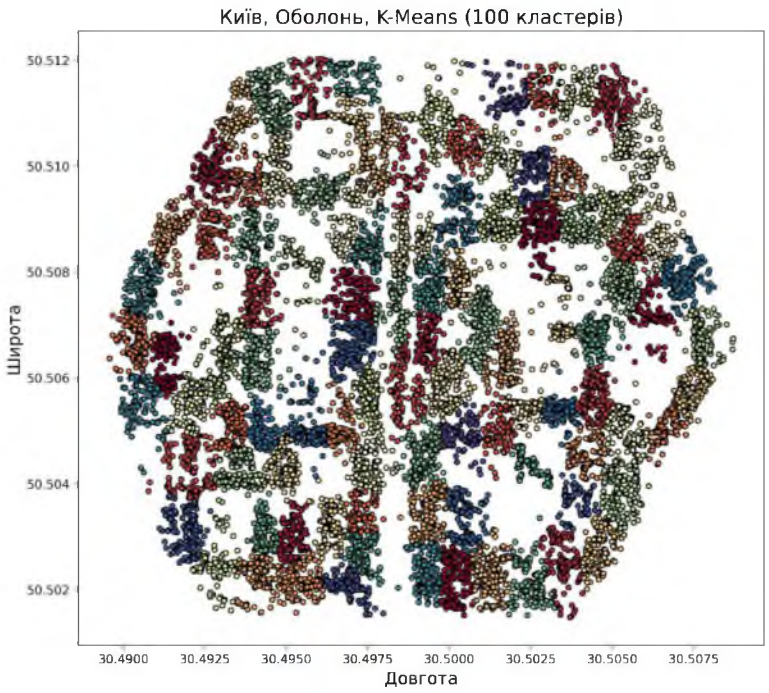

Рисунок А.1 – Кластери Оболоні для методу K-Means з k=100

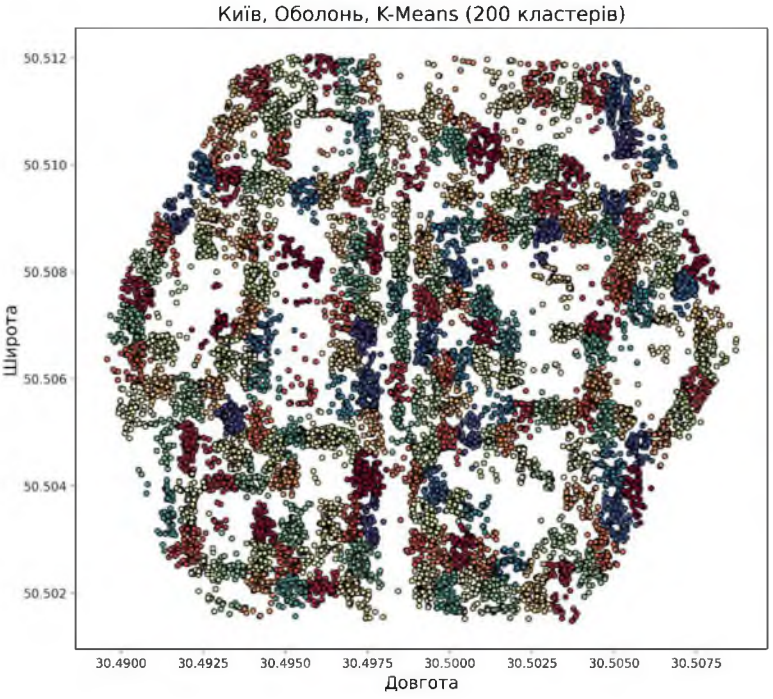

Рисунок А.2 – Кластери Оболоні для методу K-Means з k=200

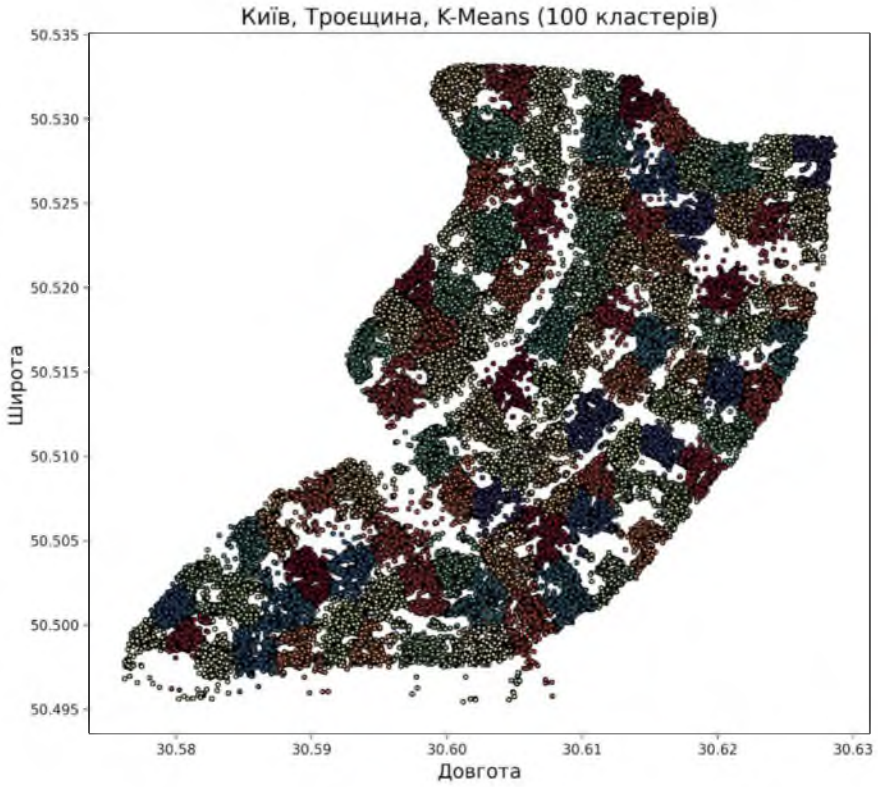

Рисунок А.3 – Кластери Троєщини для методу K-Means з k=100

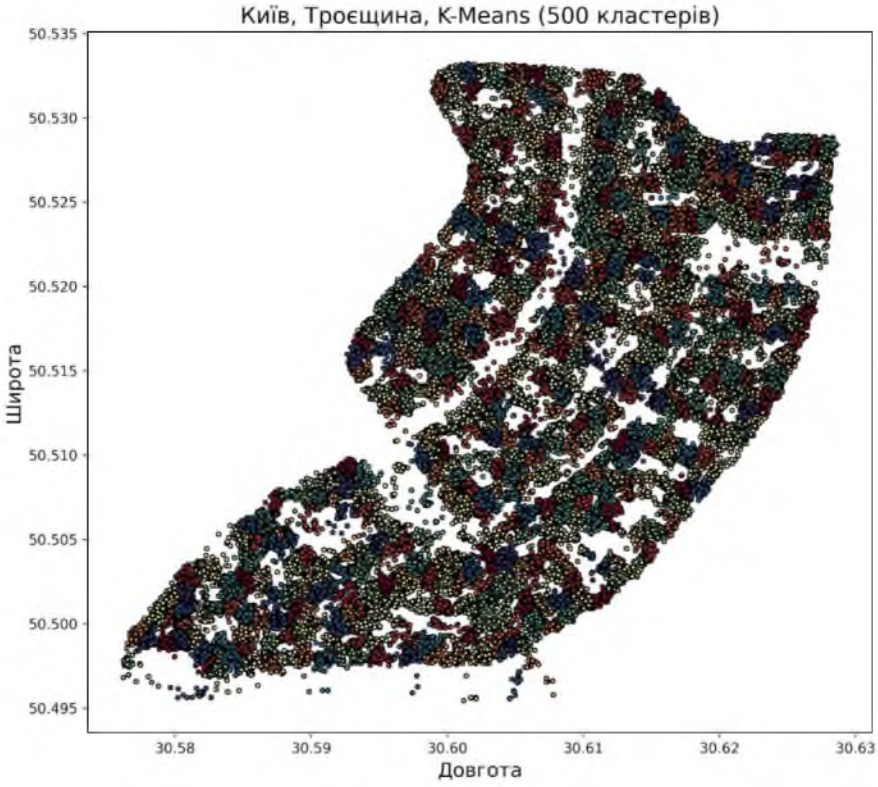

Рисунок А.4 – Кластери Троєщини для методу K-Means з k=500

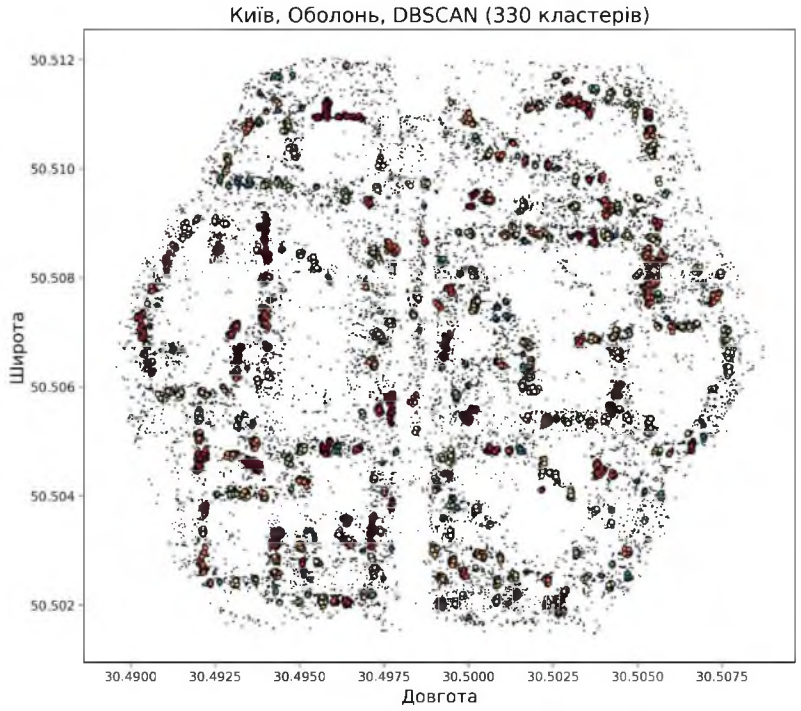

Рисунок А.5 – Кластери Оболоні для методу DBSCAN з max_eps=0.0001

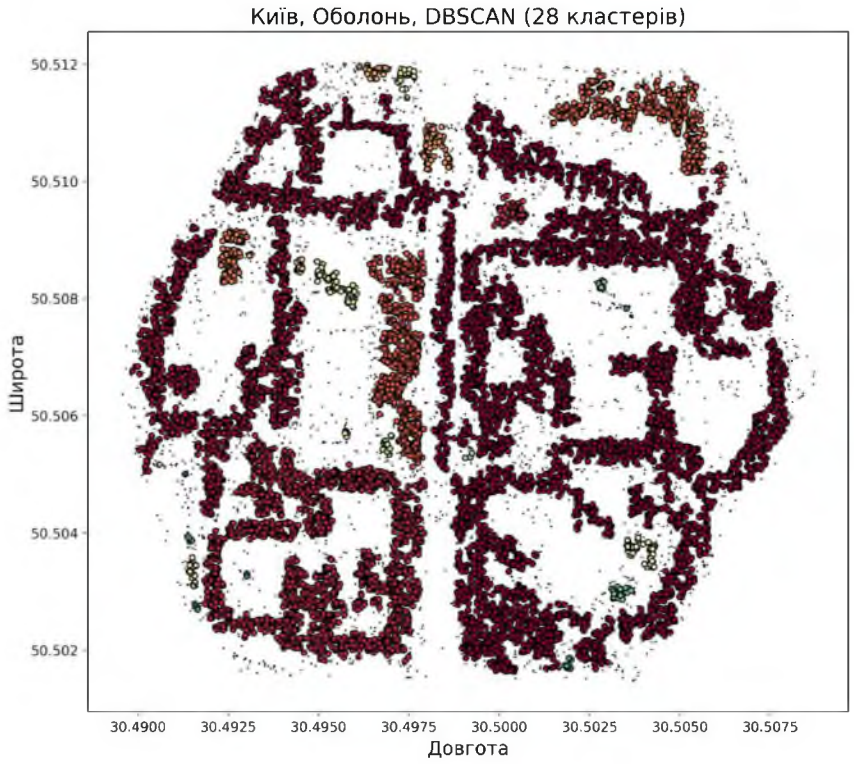

Рисунок А.6 – Кластери Оболоні для методу DBSCAN з max_eps=0.0002

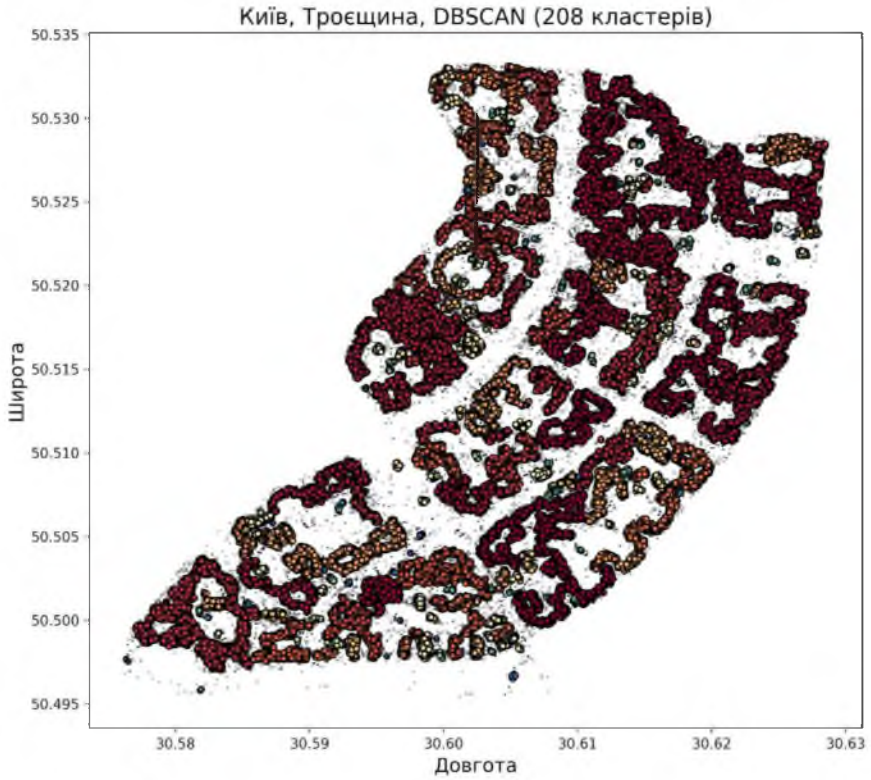

Рисунок А.7 – Кластери Троєщини для методу DBSCAN з max_eps=0.0001

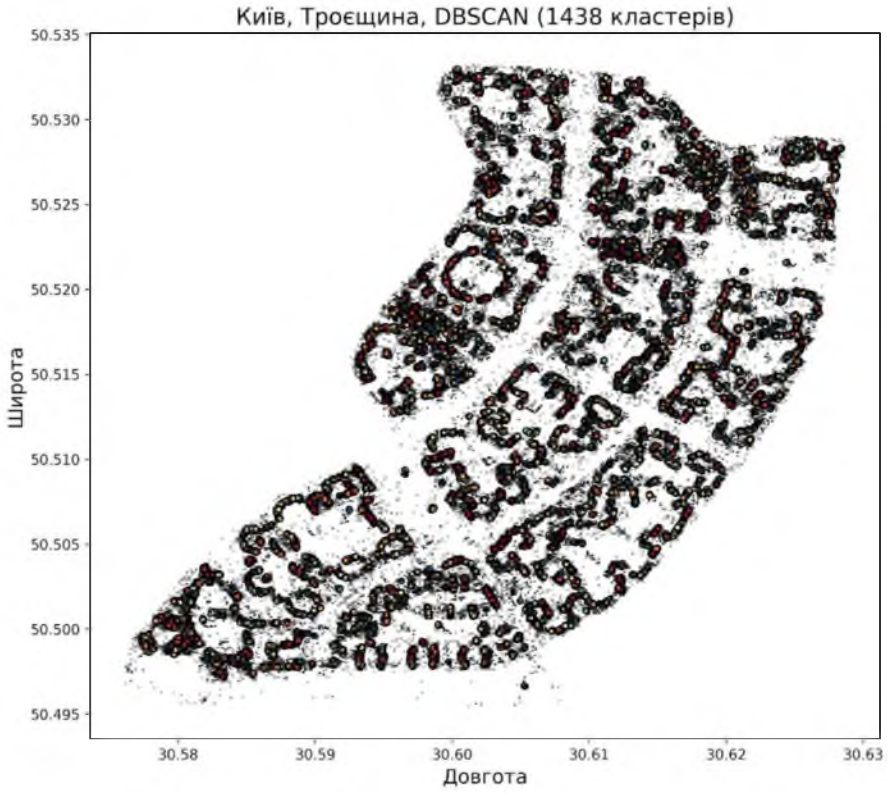

Рисунок А.8 – Кластери Троєщини для методу DBSCAN з max_eps=0.0002

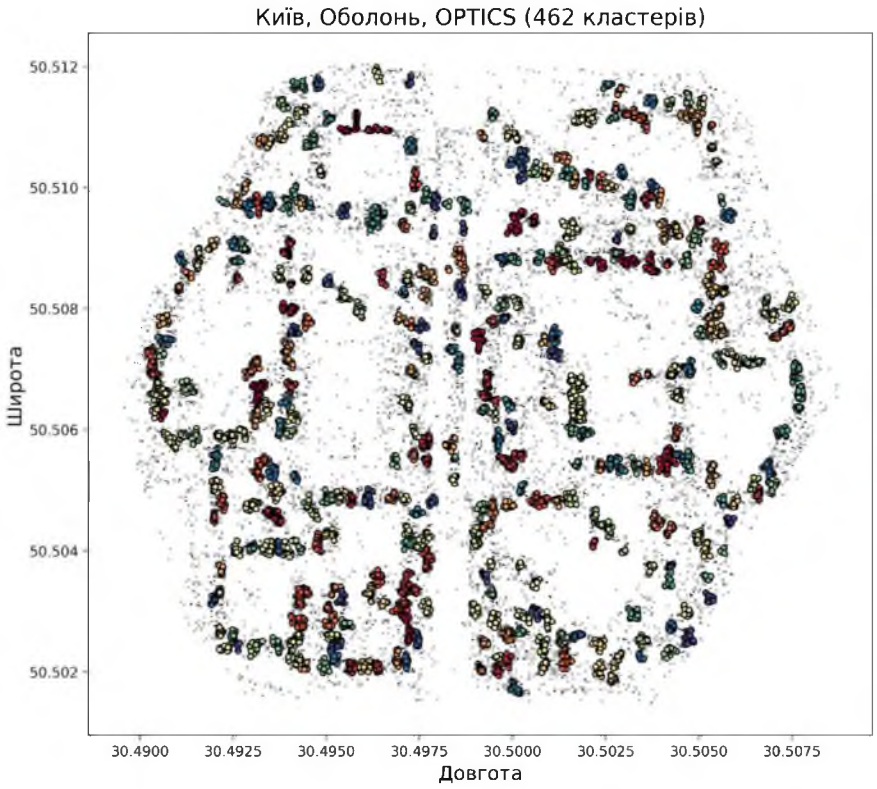

Рисунок А.9 – Кластери Оболоні для методу OPTICS з max_eps=0.0001

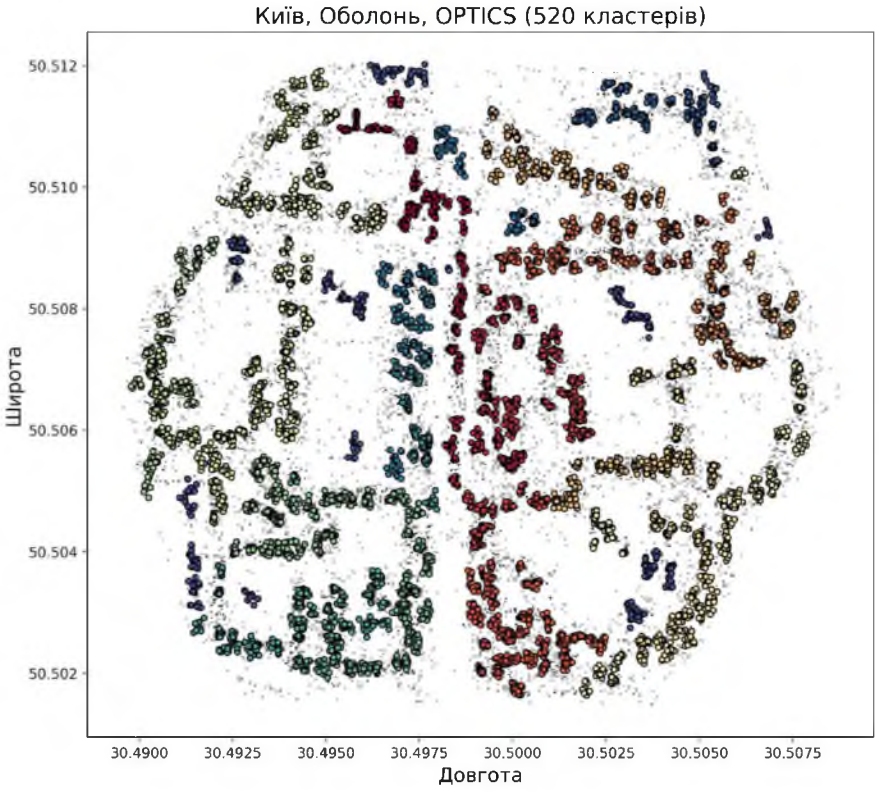

Рисунок А.10 – Кластери Оболоні для методу OPTICS з max_eps=0.0002

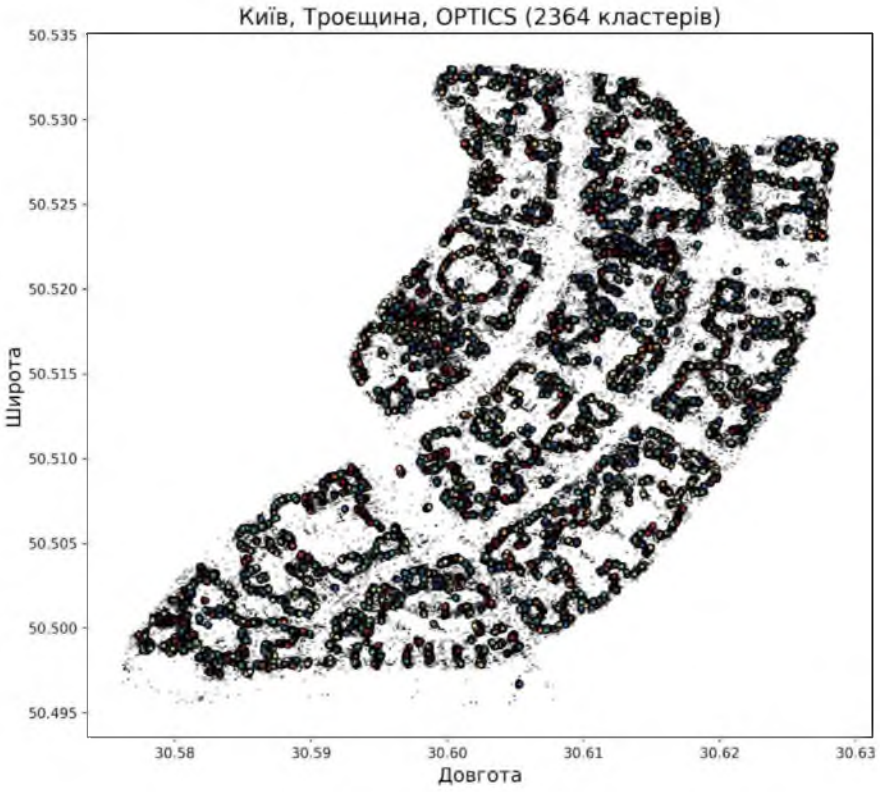

Рисунок А.11 – Кластери Троєщини для методу OPTICS з max_eps=0.0001

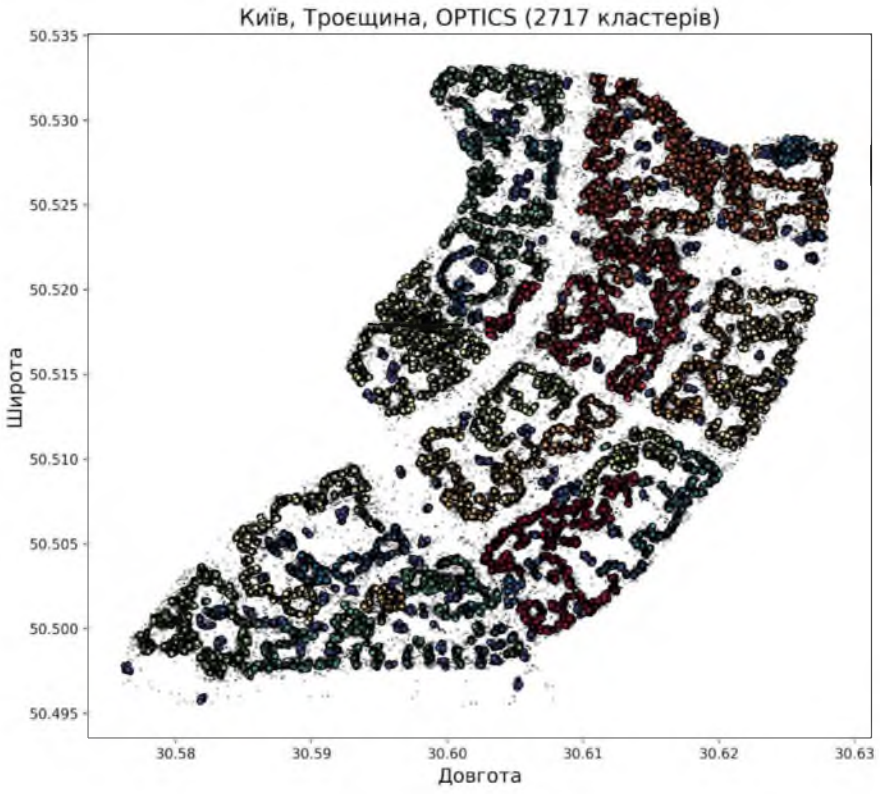

Рисунок А.12 – Кластери Троєщини для методу OPTICS з max_eps=0.0002

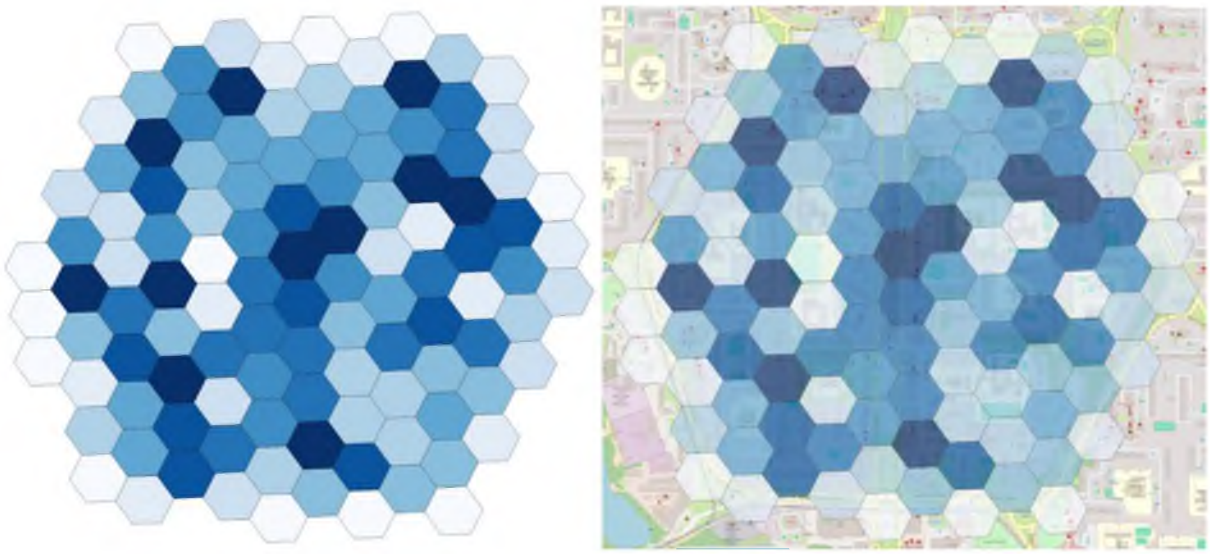

Рисунок А.13 – Візуалізація місцеположень IoT пристроїв мікрорайону Оболоні з використанням H3 10-го рівня

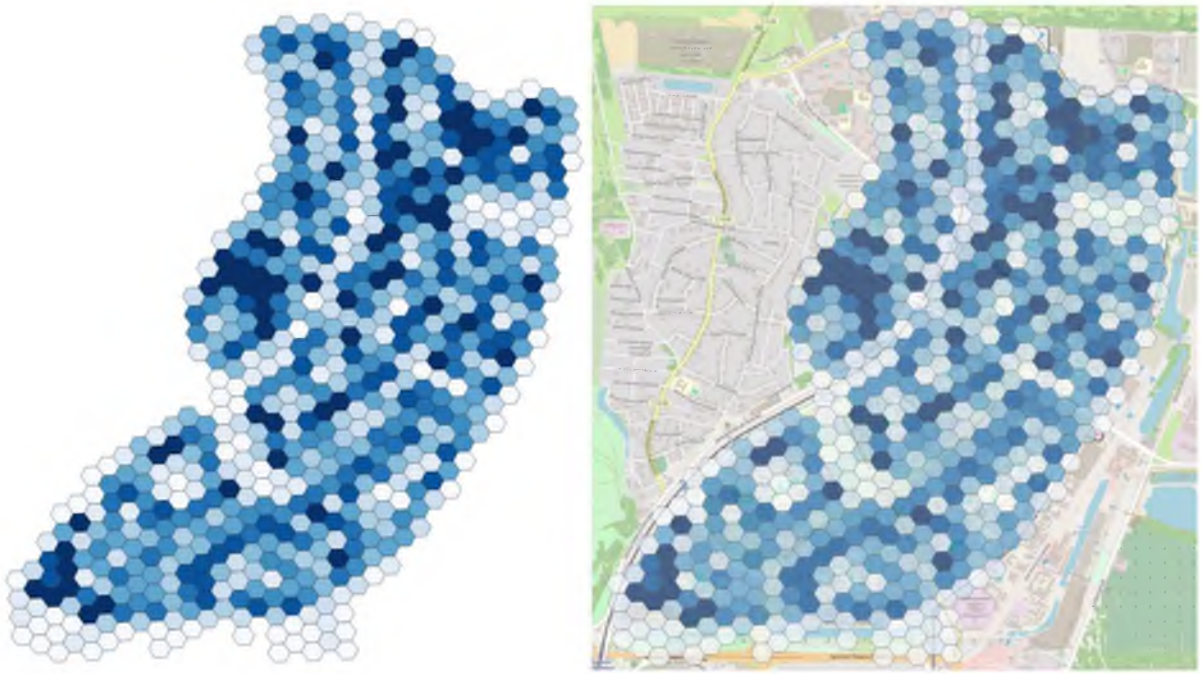

Рисунок А.14 – Візуалізація місцеположень IoT пристроїв району Троєщини з використанням H3 10-го рівня

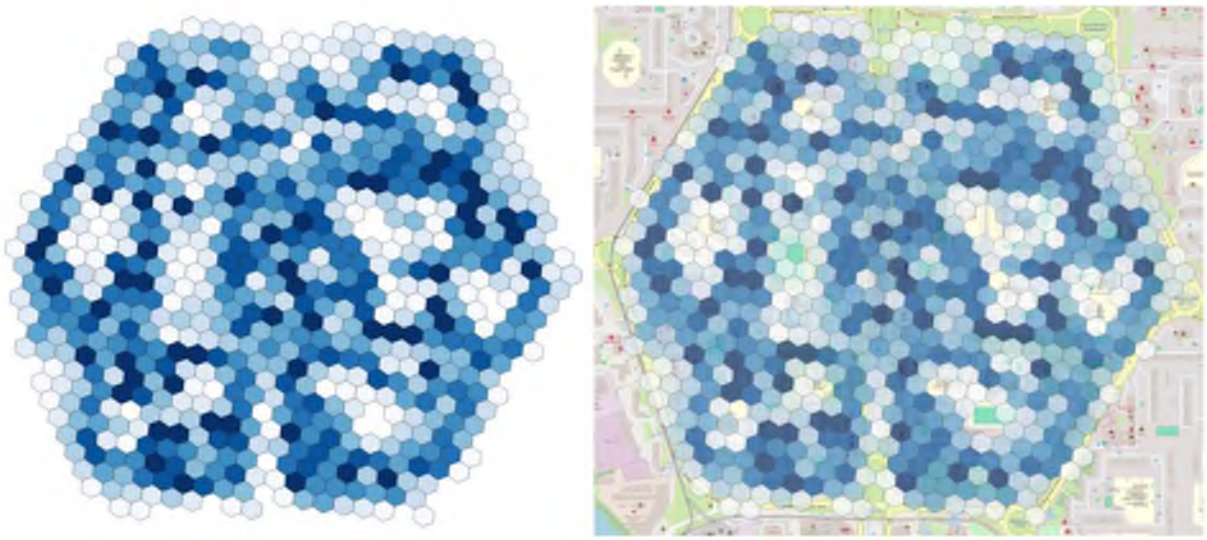

Рисунок А.15 – Візуалізація місцеположень IoT пристроїв мікрорайону Оболоні з використанням H3 11-го рівня

Рисунок А.16 – Візуалізація місцеположень IoT пристроїв району Троєщини з використанням H3 11-го рівня

Рисунок А.17 – Візуалізація місцеположень IoT пристроїв мікрорайону Оболоні з використанням H3 13-го рівня

Рисунок А.18 – Візуалізація місцеположень IoT пристроїв району Троєщини з використанням H3 13-го рівня