Міністерство освіти і науки України
Національний технічний університет
«Дніпровська політехніка»

Факультет інформаційних технологій

(факультет)

Кафедра системного аналізу та управління

(повна назва)

**ПОЯСНЮВАЛЬНА ЗАПИСКА**

кваліфікаційної роботи ступеня магістра

**Студента** Сабельникова Андрія Вадимовича

**академічної групи** 124 - 21 - 2

**спеціальності** 124 Системний аналіз

**на тему:** «Генерація та аналіз синтетичних маршрутів як метод захисту персональних даних у сервісах суспільного транспорту»

| **Керівники** | **Прізвище, ініціали** | **Оцінка за шкалою** | | **Підпис** |
|---|---|---|---|---|
| | | **рейтинговою** | **Інституційною** | |
| **кваліфікаційної роботи** | *к.т.н., доц. Малієнко А.В.* | | | |
| розділів: | | | | |
| Інформаційно-аналітичний | *к.т.н., доц. Малієнко А.В.* | | | |
| Спеціальний розділ | *к.т.н., доц. Малієнко А.В.* | | | |
| Рецензент | | | | |
| Нормоконтролер | *к.ф.-м.н., доц. Хом'як Т.В.* | | | |

Дніпро
2025

ЗАТВЕРДЖЕНО:
**завідувач кафедри**
*Системного аналізу та управління*
(повна назва)

_____ *к.т.н., доц. Желдак Т.А.*
(підпис) (прізвище, ініціали)

«_____» _____2025 року

**ЗАВДАННЯ**
**на кваліфікаційну роботу**
**ступеня** *бакалавра*

**студенту** *Сабельникову А.В.* **академічної групи** *124 -21-2*
**спеціальності**: *124 Системний аналіз*
**на тему** «*Генерація та аналіз синтетичних маршрутів як метод захисту персональних даних у сервісах суспільного транспорту»*
затверджену наказом ректора НТУ «Дніпровська політехніка»
від *05.05.2025 р. № 336–с*

| Розділ | Зміст | Терміни виконання |
|---|---|---|
| 1. Інформаційно-аналітичний розділ | *Проаналізувати структуру об'єкта дослідження. Визначити предметну область дослідження та проблему, що розв'язується. Обґрунтувати методи виконання поставлених завдань* | 20.03.2025-30.04.2025 |
| 2.Спеціальний розділ | *Розв'язати поставлені задачі: 1) Підготуівати вхідні данні для обробки. 2) Виконати анонімізацію данних за допомогою обраних методів. 3) Проаналізувати отримані результати.* | 30.04.2025-30.05.2025 |

Завдання видано _____ *доц. Малієнко А.В.*
(підпис) (прізвище, ініціали)

Дата видачі: *07.12.2024*

Дата подання до екзаменаційної комісії: _____

Прийнято до виконання _____ Сабельников А. В.
(підпис студента) (прізвище, ініціали)

# РЕФЕРАТ

Пояснювальна записка: 35 с., 7 рис., 3 табл., 5 додатки, 14 джерел.

*Об'єктом дослідження* в роботі є процеси обробки, зберігання та використання персоналізованих маршрутних даних у цифрових сервісах громадського транспорту.

*Предметом дослідження* є методи генерації та аналізу синтетичних маршрутів, що дозволяють знизити ризик ідентифікації користувача при збереженні аналітичної цінності даних.

*Метою* є розробити, реалізувати та дослідити метод генерації синтетичних маршрутів, який забезпечує анонімність персональних даних користувачів у цифрових сервісах громадського транспорту, зберігаючи при цьому корисні статистичні властивості даних.

*Методи дослідження*: Метод локальної диференційної приватності (LDP) — параметричний підхід для забезпечення ε-диференційної приватності у координатних даних шляхом додавання статистичного шуму. Метод обчислення дивергенції Топсе — застосовується для оцінки схожості між оригінальними та синтетичними маршрутами. Сіткове розбиття за допомогою H3-індексації — геопросторове представлення координат у вигляді шестикутників фіксованої ієрархічної структури, що не вимагає попереднього налаштування параметрів. Open Source Routing Machine (OSRM) — побудова маршрутів на основі реальних транспортних мереж без параметричної моделі.

В *інформаційно–аналітичному розділі* наведено аналіз об'єкту дослідження та ключових проблем на ньому. Поставлені задачі дослідження та обрано концепції їх розв'язання.

У *спеціальному розділі* очищено дані, застосовано H3-сітку для агрегації, побудовано синтетичні маршрути через OSRM. Для анонімності використано LDP з ε-диференційною приватністю. Схожість маршрутів оцінено дивергенцією Топсе. Проаналізовано отримані результати.

*Практична цінність* роботи полягає в розробці ефективного методу анонімізації маршрутних даних, який дозволяє захищати приватність користувачів сервісів мобільності без втрати аналітичної цінності даних для досліджень та міського планування.

*Апробація результатів роботи:* основні результати кваліфікаційної роботи були апробовані на міжнародній науково-практичній конференції студентів та молодих учених «Інформаційні технології: теорія та практика», яка відбулася 2 квітня 2025р. За участь у конференції отримано сертифікат (див. ДОДАТОК Д).

*Ключові слова*: АНОНІМІЗАЦІЯ ДАНИХ, H3-СІТКА, OSRM, ДИФЕРЕНЦІЙНА ПРИВАТНІСТЬ, LDP, ДИВЕРГЕНЦІЯ ТОПСЕ, СИНТЕТИЧНІ МАРШРУТИ.

# ЗМІСТ

# ВСТУП

З поширенням цифрових технологій та розширенням інфраструктури міської мобільності зростає кількість сервісів, що збирають дані про пересування користувачів. До таких сервісів належать орендні платформи електросамокатів, велосипедів, таксі, які активно застосовують GPS-моніторинг для зручності використання та аналітики. Ці дані мають цінність для досліджень трафіку, планування інфраструктури та оптимізації маршрутів. Проте вони одночасно містять чутливу інформацію, яка дозволяє ідентифікувати користувача, що створює ризики для конфіденційності.

*Актуальність теми.* У сучасних умовах питання захисту персональних даних користувачів стає дедалі гострішим. Традиційні методи анонімізації, зокрема видалення імен чи ID, є недостатніми, оскільки збереження точних координат переміщення дозволяє здійснити зворотну ідентифікацію особи. Особливо це актуально для міських сервісів, де маршрути мають сталу географічну структуру, що піддається аналітичному розпізнаванню. Необхідним є створення технологій, які дозволяють маскувати точну інформацію без втрати її аналітичної цінності.

*Мета і завдання дослідження.* Метою є розробка методу генерації синтетичних маршрутів для захисту персональних даних із використанням просторового агрегування (H3), побудови маршрутів (OSRM) та локльної диференційної приватності (LDP). Завдання полягає у реалізації та порівнянні ефективності цих інструментів у збереженні структури переміщень при одночасному забезпеченні анонімності.

*Об'єкт дослідження* — процес обробки маршрутних даних користувачів сервісів мікромобільності.
*Предмет дослідження* — методи трансформації маршрутів із метою анонімізації та збереження аналітичної цінності.

*Практичне значення.* Запропонований підхід дозволяє захистити особисті дані користувачів і водночас зберегти цінну інформацію для транспортної аналітики. Така система може бути інтегрована у міські цифрові платформи, надаючи безпечний доступ до даних для дослідників, планувальників і операторів мобільності. Це значно скорочує ризики витоку інформації, підвищує

довіру користувачів і сприяє розвитку безпечного цифрового міського середовища.

Отже, розробка методології анонімізації маршрутних даних є актуальним напрямом, що поєднує інтереси приватності та ефективної обробки просторової інформації.

# РОЗДІЛ 1. ІНФОРМАЦІЙНО-АНАЛІТИЧНИЙ

## 1.1 Аналіз проблеми захисту персональних даних у транспортних сервісах

Сервіси мікромобільності стали невіддільною частиною сучасного міського транспорту. Вони надають користувачам можливість швидко пересуватись містом за допомогою орендованих електросамокатів, велосипедів, електромобілів або таксі. Для функціонування таких сервісів збираються великі обсяги даних: точні координати GPS у режимі реального часу, час початку та завершення поїздки, швидкість, тривалість маршруту, іноді навіть дані про смартфон користувача. Ці дані є основою для логістики, оптимізації трафіку, бізнес-аналітики та планування інфраструктури.

Проте разом з корисністю зібраної інформації виникає критичне питання — захист приватності. Дані, що мають геолокаційний компонент, відносяться до чутливої персональної інформації, адже навіть кілька маршрутів користувача дозволяють ідентифікувати його з високою точністю. Наприклад, якщо маршрут починається щодня з одного й того ж місця — це може бути дім, а завершується в іншому — ймовірно, робота. Навіть у знеособленому вигляді ці дані можуть бути повторно ідентифіковані шляхом перехресного аналізу з іншими відкритими джерелами.

Правові норми, як-от GDPR (Загальний регламент захисту даних) в ЄС або CCPA у Каліфорнії, зобов’язують операторів таких сервісів вживати заходів щодо збереження конфіденційності, включаючи анонімізацію або псевдонімізацію даних. Невиконання цих вимог може призвести до великих штрафів, репутаційних втрат, а також втрати довіри з боку користувачів. [1]

З технічної точки зору, найбільший виклик полягає в тому, щоб знайти баланс між анонімністю та корисністю. Якщо координати просто зашифрувати або сильно спотворити — аналітика стає неможливою. Якщо залишити їх майже без змін — зростає ризик розкриття особи. Тому традиційні підходи — видалення ID користувача, обрізка координат, випадкове зміщення точки — не є ефективними.

Відтак, зростає потреба у використанні сучасних методів математичного захисту даних, таких як диференційна приватність, просторове агрегування,

генерація синтетичних маршрутів та інші підходи, які дозволяють зберегти аналітичну цінність даних без ризику ідентифікації користувача.

### 1.2 Огляд сучасних методів анонімізації маршрутів

В сучасній науковій та прикладній практиці сформувалося декілька підходів до анонімізації маршрутів, які дозволяють зберегти корисність даних для аналізу й водночас зменшити ризики порушення конфіденційності.

У цьому розділі розглянуто три основні групи методів анонімізації маршрутних даних:

- Евристичні методи
- Агрегація та кластеризація координат
- Генерація синтетичних маршрутів

Кожен із зазначених підходів має власні переваги та недоліки, які залежать від контексту використання, типу даних та бажаного рівня захисту.

#### 2.1.1 Евристичні методи анонімізації

Евристичні методи — це найпростіший і найбільш інтуїтивно зрозумілий клас методів анонімізації маршрутних даних. Вони не вимагають складної математичної бази чи машинного навчання, і часто використовуються як базове рішення в обробці геолокаційної інформації. Їхня суть полягає в локальній зміні або спрощенні даних так, щоб знизити ризик ідентифікації користувача, зберігаючи при цьому основну інформацію про структуру руху. [2]

Найбільш поширені евристичні підходи:

*1. Округлення координат (Geographic Rounding)*

Координати кожної точки маршруту округлюються до певної точності — наприклад, до двох знаків після коми. Це дозволяє замінити точку з точністю 1 метр на область радіусом приблизно 1 км. У результаті зникає можливість точного розпізнавання місця (наприклад, будинку), але зберігається напрямок і загальний маршрут. [3]

*2. Вилучення чутливих точок (Endpoint Trimming)*

Початкові та кінцеві точки маршруту зазвичай є найбільш ризикованими з точки зору конфіденційності, оскільки часто збігаються з місцем проживання або роботи. Тому маршрут обрізається на певну кількість точок з обох кінців або замінюється їх розташування на найближчі публічні місця (наприклад, перехрестя чи зупинки). [4]

*3. Штучне шумове зміщення (Location Perturbation)*

Цей метод полягає у випадковому зміщенні кожної точки маршруту на деяку відстань у випадковому напрямку. Наприклад, кожна точка може бути зсунутою в межах 100 метрів. Хоча метод не гарантує захисту від перехресного аналізу, він значно ускладнює точну прив'язку маршруту до реальних адрес. [5]

Евристичні методи мають низку очевидних переваг, завдяки яким їх часто застосовують на практиці. Насамперед, їх вирізняє простота реалізації — вони не потребують складних математичних обчислень чи значних обчислювальних ресурсів. Також такі методи легко інтегрувати у вже існуючі системи, адже вони сумісні з більшістю форматів, у яких зазвичай зберігаються маршрутні або геолокаційні дані.

Однак, попри ці переваги, евристичні підходи мають і суттєві обмеження. Найголовніший їхній недолік — це відсутність формальних гарантій конфіденційності. Це означає, що навіть після застосування таких методів завжди залишається ризик повторної ідентифікації користувача. Крім того, вони виявляються не надто ефективними при роботі з великими масивами даних або з відкритими публічними наборами, де простих технік недостатньо для забезпечення надійного захисту.

### 2.1.2 Методи агрегування та кластеризації

Методи агрегування та кластеризації маршрутних даних є наступним кроком після евристичних підходів і забезпечують вищий рівень анонімізації без значної втрати аналітичної цінності. Вони базуються на ідеї групування просторових точок у більші географічні області, що дозволяє приховати точні координати користувача, зберігаючи загальну структуру та закономірності переміщень.

*Агрегація простору*

Просторова агрегація передбачає поділ території на осередки певного розміру — це можуть бути квадрати, шестикутники, зони покриття базових станцій мобільного зв’язку, адміністративні межі тощо. У межах кожного такого осередку дані користувачів об’єднуються, і аналіз проводиться не на рівні конкретних точок, а на рівні зон.

*Кластеризація*

Кластеризація є статистичним методом, який дозволяє об'єднувати схожі об'єкти — у даному випадку, точки маршруту — на основі певних критеріїв, таких як просторові відстані або щільність розташування. Це один із найефективніших способів узагальнення геоданих, який широко застосовується для підвищення конфіденційності маршрутів [6].

Серед найбільш поширених алгоритмів кластеризації варто виділити K-means, який передбачає попереднє задання кількості кластерів і рівномірне розподілення точок між ними [7]. Інший підхід — DBSCAN — дозволяє виявляти кластери без необхідності визначати їх кількість заздалегідь. Він орієнтований на щільність точок і здатен автоматично ідентифікувати винятки або «шум» [8]. Третій підхід — ієрархічна кластеризація, зокрема агломеративна, яка поступово об’єднує найближчі точки у більші групи на основі певного критерію подібності [9].

Застосування кластеризації до маршрутних даних дозволяє зберегти загальну логіку переміщення, не розкриваючи точних координат користувача. Наприклад, кілька близько розташованих точок, що відображають пересування по вулиці, можна замінити одним усередненим положенням. Такий підхід дає змогу створити узагальнений маршрут, який відповідає реальному за формою та напрямком, але при цьому унеможливлює точну ідентифікацію конкретного місця перебування.

### 2.1.3 Генерація синтетичних маршрутів

Генерація синтетичних маршрутів — це сучасний підхід до анонімізації геолокаційних даних, який дозволяє поєднати захист персональної інформації з високою збереженістю їхньої аналітичної цінності. Суть методу полягає у створенні штучних маршрутів, які не належать жодному реальному користувачу,

але водночас статистично імітують характерні риси реального переміщення у міському середовищі. Такі маршрути зберігають загальні патерни мобільності, але не дозволяють ідентифікувати особу.

На практиці побудова синтетичних маршрутів відбувається шляхом поєднання зашумлених або агрегованих початкових і кінцевих точок з алгоритмами маршрутизації. Ці алгоритми враховують особливості дорожньої мережі, топографію, правила дорожнього руху, дозволені напрямки тощо. Найчастіше для цього використовують інструменти з відкритим кодом, такі як Open Source Routing Machine (OSRM), GraphHopper або Valhalla. Також застосовуються комерційні API, зокрема Mapbox Directions чи Google Directions API [10].

Процес генерації маршруту зазвичай складається з кількох послідовних етапів. Спочатку виконується агрегування координат — реальні точки перетворюються на просторові осередки. Далі вибираються або генеруються анонімізовані початкові та кінцеві зони для кожного маршруту. Заключним етапом є побудова самої траєкторії — синтетичного шляху між точками, що створюється за допомогою інструментів маршрутизації з урахуванням наявної дорожньої інфраструктури.

Основна перевага такого підходу полягає в тому, що жодна координата у маршруті не відповідає реальній позиції користувача, а отже забезпечується високий рівень конфіденційності. Водночас маршрути зберігають реалістичну структуру переміщення містом, що дозволяє використовувати їх у дослідженнях мобільності, аналізі транспортного навантаження та інших завданнях. Такий підхід також добре масштабується — його можна застосовувати як до невеликих наборів даних, так і до мільйонів записів.

### 1.3 Опис методів вирішення задачі

У процесі розробки методів анонімізації маршрутних даних у мікромобільності важливу роль відіграють математичні підходи, що дозволяють моделювати просторові закономірності, зберігаючи при цьому конфіденційність користувачів.

У рамках цього дослідження використовуються такі методи як

дивергенцією Топсое — що дозволяють кількісно оцінити схожість між оригінальними та анонімізованими маршрутами. Також метод локальної диференційної приватності (LDP) для створення зашумлених синтетичних маршрутів .

Застосування комбінованого підходу забезпечує широку гнучкість у виборі рівня деталізації маршрутів та дозволяє підібрати компроміс між конфіденційністю та збереженням корисної структури даних.

### 1.3.1 Дивергенція Топсое

*Дивергенція Топсое*

Дивергенція Топсое є представником класу ƒ-дивергенцій, які використовуються для вимірювання розбіжності між двома ймовірнісними розподілами. Вона базується на ідеї оцінки відхилення кожного розподілу від їх спільного середнього і дозволяє симетрично та формально оцінити ступінь відмінності між дискретними розподілами. [11]

Нехай задано два ймовірнісні розподіли $P = \{p_i\}$ та $P = \{q_i\}$, визначені на скінченній множині X, де $i \in X$, і виконуються умови нормування:

$$\sum_{i \in X} p_i = 1, \quad \sum_{i \in X} q_i, p_i \geq 0, q_i \geq 0. \tag{1.1}$$

Дивергенцію Топсое між розподілами P і Q можна визначити як середнє арифметичне двох дивергенцій Кульбака–Лейблера між кожним розподілом і усередненим розподілом $M = \frac{1}{2}(P + Q)$:

$$D_{Topsøe}(P \parallel Q) = \frac{1}{2}[D_{KL}(P \parallel M) + D_{KL}(Q \parallel M)] \tag{1.2}$$

Де $D_{KL}(P \parallel M)$ позначає дивергенцію Кульбака-Лейблера:

$$D_{KL}(P \parallel M) = \sum_{i \in X} p_i log\left(\frac{p_i}{m_i}\right), \text{ де } m_i = \frac{p_i + q_i}{2} \tag{1.3}$$

На відміну від деяких інших дивергенцій, дивергенція Топсое є завжди скінченною при ненульовому перекритті підтримки розподілів P і Q, що робить її зручною у практичному застосуванні. Зокрема, вона не вимагає заміни нульових значень на регуляризовані оцінки, оскільки за визначенням уникнуто ділення на нуль або логарифмування нуля.

З точки зору інформаційної теорії, дивергенція Топсое вимірює середню відстань кожного з розподілів до їх симетричного усереднення. Її використання особливо доцільне у задачах, де необхідно зберігати баланс між двома апроксимаціями та забезпечити стабільну оцінку незалежно від того, який розподіл обрано за еталон.

Таким чином, дивергенція Топсое є зручною та строго визначеною метрикою для аналізу відмінностей між дискретними розподілами, що поєднує симетричність, додатність та аналітичну стабільність при обчисленнях.

### 1.3.1 Локальна диференційна приватность (LDP)

*Локальна диференційна приватность (LDP)*

Локальна диференційна приватність (Local Differential Privacy, LDP) є концепцією захисту персональних даних, яка гарантує конфіденційність кожного окремого елемента ще до його передавання на сервер. На відміну від глобальної диференційної приватності, яка застосовує шум до агрегованих результатів, локальна приватність додає шум безпосередньо до значення на рівні клієнта. Це забезпечує найвищий рівень захисту даних, навіть у випадку компрометації центрального вузла.

Один із найпростіших способів реалізації LDP — це випадкове перетворення значення з визначеним імовірнісним розподілом. Суть методу полягає в тому, що справжнє значення залишають або замінюють випадково, з певною імовірністю, яка контролюється параметром приватності ε. [12]

Припустимо, що дані користувача є скалярним або категоріальним значенням $x$ з множини можливих значень $X$. Для кожного такого значення застосовується стохастичне перетворення:

$$x \to x' \in X, \text{де вибір здійснюється з ймовірностями:}$$

$$P(x') = \begin{cases} p = \frac{1}{2}, & \text{якщо } x' = x \\ = \frac{1}{e^{\varepsilon} + 1}, & \text{якщо } x' \neq x \end{cases} \tag{1.4}$$

Це означає, що справжнє значення зберігається з фіксованою ймовірністю p, тоді як інші значення вибираються з рівною ймовірністю q, яка експоненційно зменшується в залежності від параметра ε.

Ця умова гарантує, що сторонній спостерігач не може з високою точністю відрізнити, яке саме значення було вхідним, базуючись лише на зашумленому результаті. Параметр ε визначає рівень захисту: при менших значеннях ε приватність вища, але знижується точність; при більших ε – точність підвищується, однак послаблюється конфіденційність.

### 1.4 Постановка задачі

У зв'язку з необхідністю забезпечення конфіденційності користувачів сервісів мікромобільності, а також збереження цінної інформації про характер міського руху, постає задача побудови методу анонімізації маршрутів, який в межах даного дослідження, повинен відповідати кільком ключовим вимогам. По-перше, він має забезпечувати приховування чутливих геолокаційних точок, таких як місце проживання користувача, офіс або інші часто відвідувані локації, які можуть дозволити ідентифікувати особу. По-друге, важливо, щоб метод зберігав загальну структуру переміщень — тобто маршрути після обробки повинні відображати характер руху та основні просторові закономірності без прив'язки до конкретних координат. По-третє, запропонований підхід має передбачати можливість формального аналізу втрати інформації, що дозволить кількісно оцінити, наскільки отримані анонімізовані дані відхиляються від оригінальних і чи зберігають достатній рівень корисності для аналітики.

У якості вихідних даних використовуються GPS-маршрути користувачів з часовими мітками, початковими та кінцевими точками. Дані попередньо агрегуються через H3-індексацію, після чого проводиться побудова синтетичних маршрутів. Для оцінки якості анонімізації застосовуються метрики статистичної дивергенції Топсое.

**Висновки за розділом**

У першому розділі було проаналізовано проблему захисту персональних даних у сфері мікромобільності, що є надзвичайно актуальною в умовах масового використання GPS-трекінгу та сервісів геолокації. Наведено приклади ризиків, пов’язаних з ідентифікацією користувачів на основі маршрутних даних, навіть після їх знеособлення.

Проведено класифікацію сучасних підходів до анонімізації маршрутів. Зокрема, розглянуто округлення координат, обрізку чутливих точок, просторову агрегацію, кластеризацію, генерацію синтетичних маршрутів.

Також описано методи моделювання та оцінки ефективності анонімізації. Особливу увагу приділено дивергенції Топсое як симетричній і стабільній метриці подібності розподілів, що дозволяє формально оцінювати втрату інформації.

Таким чином, у розділі обґрунтовано вибір комбінованого підходу до задачі: застосування просторового агрегування, генерації синтетичних маршрутів та методів диференційної приватності у поєднанні з метриками статистичної оцінки втрат, що забезпечує баланс між захистом приватності та збереженням корисності даних.

# РОЗДІЛ 2. Спеціальний розділ

## 2.1 Опис вхідних даних та їх попередня обробка

Для дослідження було створено програмний код (Додаток Г) у якому використано набір даних про поїздки на електросамокатах у місті Штутгарт. Ці дані зберігаються у форматі JSON у файлі “output_150.json” та містять 145220 записів, де кожен рядок відповідає окремій поїздці. Основні поля цього набору включають унікальний ідентифікатор поїздки (reservationId), пройдену відстань у метрах (drivenDistance), тривалість використання (usageTime), а також часові мітки початку та завершення поїздки (startTime, endTime).

Ключовим елементом для просторового аналізу є географічні координати старту та фінішу маршруту, які представлені полями (startLat), (startLon), (endLat) та (endLon). Крім цього, поле (waypoints) містить список точок, що відображають фактичну траєкторію руху під час поїздки. Ці точки представлені у вигляді словника, де кожна точка містить координати lat і lon (табл.2.1 - зображено 5 рядків з 145220).

Таблиця 2.1

**Початкові дані**

| | reservationId | drivenDistance | usageTime | startTime | endTime | startLat | startLon | endLat | endLon | waypoints |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 53983 | 3000 | 11 | 2017-10-17 13:05:09 | 2017-10-17 16:13:47 | 48.775116 | 9.155653 | 48.770775 | 9.158312 | {'0': {'lat': '48.775116', 'lon': '9.155653', ... |
| 1 | 53982 | 5000 | 15 | 2017-10-17 13:04:53 | 2017-10-17 14:00:10 | 48.764484 | 9.177281 | 48.790157 | 9.204130 | {'0': {'lat': '48.764481', 'lon': '9.177288', ... |
| 2 | 53980 | 3000 | 10 | 2017-10-17 13:04:05 | 2017-10-17 13:19:06 | 48.774185 | 9.166993 | 48.770519 | 9.183026 | {'0': {'lat': '48.77415', 'lon': '9.166971', '... |
| 3 | 53974 | 3000 | 7 | 2017-10-17 12:50:30 | 2017-10-17 13:03:31 | 48.775360 | 9.149068 | 48.774170 | 9.166937 | {'0': {'lat': '48.775365', 'lon': '9.14925', '... |
| 4 | 53971 | 2000 | 4 | 2017-10-17 12:40:20 | 2017-10-17 12:57:18 | 48.771221 | 9.160151 | 48.764496 | 9.162521 | {'0': {'lat': '48.771238', 'lon': '9.160208', ... |

На етапі попередньої обробки дані були очищені від поїздок, які виходять за межі географічного регіону дослідження — центральної частини Штутгарта. Для цього було задано прямокутну область (Bounding Box), яка охоплює координати з півночі на південь та із заходу на схід у межах широти 48.76265–48.80005 та довготи 9.13885–9.21885 відповідно (рис.2.1). Вибірка була відфільтрована таким чином, щоб залишити лише поїздки, які починаються і завершуються всередині цієї області.

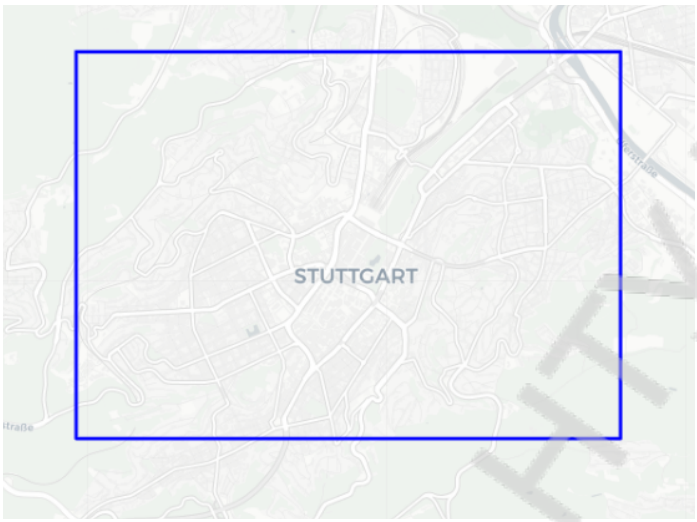


Рис.2.1. Границі міста Штуттгарт (Bounding Box)

## 2.2 Геопросторове кодування з використанням H3

Однією з ключових складових проекту є ефективне подання географічних координат у вигляді дискретних осередків. Для цього використовується геопросторовий індексатор H3, розроблений компанією Uber. H3 забезпечує ієрархічну дискретизацію земної поверхні шляхом розбиття її на шестикутні осередки різної роздільної здатності, що є особливо зручним для аналізу мобільності, просторової анонімізації та агрегування маршрутних даних.[13]

У геопросторових задачах вибір рівня дискретизації (або роздільної здатності) визначає, наскільки точно координати можуть бути представлені у вигляді географічних осередків. У бібліотеці H3 роздільна здатність задається цілим числом від 0 до 15 (табл.2.2)

Таблиця 2.2

**Залежність площі осередків від роздільної здатності**

| H3 Resolution | Площа осередку (приблизно) | Діаметр (км) |
|---|---|---|
| 5 | 2.44 км² | 1.57 км |
| 6 | 0.65 км² | 0.98 км |
| 7 | 0.17 км² | 0.52 км |
| 8 | 0.044 км² | 0.26 км |

| | | |
|---|---|---|
| 9 | 0.011 км² | 0.13 км |
| 10 | 0.003 км² | 0.07 км |

Візуалізувавши H3-осередки на крті, методом підбору було обрано восьмий рівень роздільності, який найкраще підходить для обраної території та подальшої обробки. Такий масштаб є оптимальним для аналізу міського середовища: кожна клітинка покриває декілька кварталів, зберігаючи загальний напрямок маршруту без деталізації до окремих будівель (рис.2.2 - 2.4).

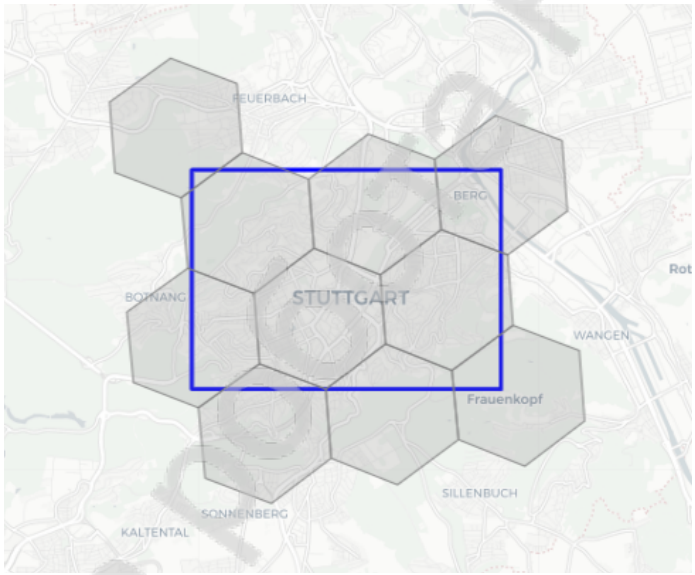


Рис.2.2. H3-осередки з резолюцією “7”

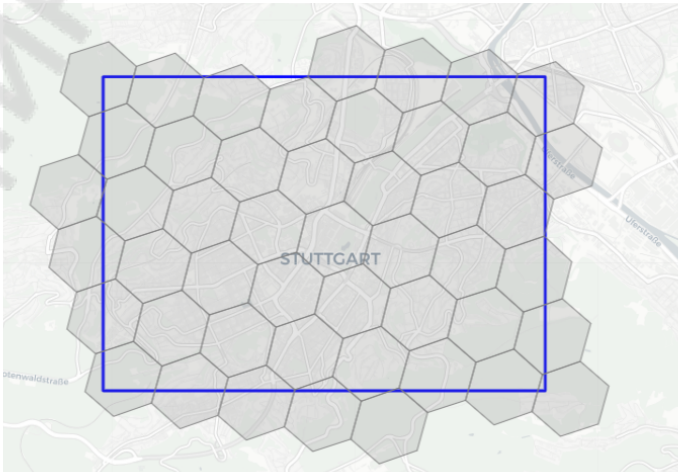


Рис.2.3. H3-осередки з резолюцією “8”

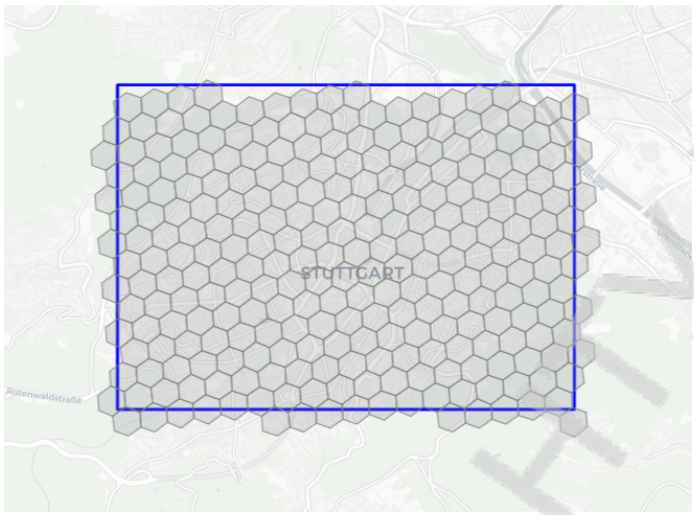


Рис.2.4. H3-осередки з резолюцією “9”

Для кожної стартової та кінцевої точки поїздки було обчислено H3-ідентифікатори на восьмому рівні роздільності. Це дозволяє зручно порівнювати маршрути між собою та виконувати агрегацію даних у просторовому вимірі.

На завершення, з очищених даних було сформовано окрему таблицю з обраними маршрутами для подальшого аналізу та візуалізації. Таким чином, проведена обробка даних дозволила сформувати якісну вибірку, придатну для застосування алгоритмів маршрутизації, додавання шуму та аналізу приватності (табл.2.3- зображено 5 рядків з 133441).

Таблиця 2.3

**Оброблені дані**

| | reservationId | drivenDistance | usageTime | startTime | endTime | startLat | startLon | endLat | endLon | waypoints | start_hex | end_hex |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 53983 | 3000 | 11 | 2017-10-17 13:05:09 | 2017-10-17 16:13:47 | 48.775116 | 9.155653 | 48.770775 | 9.158312 | {'0': {'lat': '48.775116', 'lon': '9.155653', ... | 881faa7a8dfffff | 881faa7a85fffff |
| 1 | 53982 | 5000 | 15 | 2017-10-17 13:04:53 | 2017-10-17 14:00:10 | 48.764484 | 9.177281 | 48.790157 | 9.204130 | {'0': {'lat': '48.764481', 'lon': '9.177288', ... | 881faa7a87fffff | 881faa71a5fffff |
| 2 | 53980 | 3000 | 10 | 2017-10-17 13:04:05 | 2017-10-17 13:19:06 | 48.774185 | 9.166993 | 48.770519 | 9.183026 | {'0': {'lat': '48.77415', 'lon': '9.166971', '... | 881faa7a81fffff | 881faa7a83fffff |
| 3 | 53974 | 3000 | 7 | 2017-10-17 12:50:30 | 2017-10-17 13:03:31 | 48.775360 | 9.149068 | 48.774170 | 9.166937 | {'0': {'lat': '48.775365', 'lon': '9.14925', '... | 881faa7aebfffff | 881faa7a81fffff |
| 4 | 53971 | 2000 | 4 | 2017-10-17 12:40:20 | 2017-10-17 12:57:18 | 48.771221 | 9.160151 | 48.764496 | 9.162521 | {'0': {'lat': '48.771238', 'lon': '9.160208', ... | 881faa7a85fffff | 881faa7a85fffff |

### 2.3 Анонімізація маршрутів шляхом додавання шуму

У рамках цього дослідження реалізовано спрощений та адаптований до геопросторового формату підхід — локальну диференційну приватність (Local Differential Privacy, LDP). На відміну від класичної глобальної DP, яка додає шум до результату запиту або агрегованої інформації, локальна диференційна приватність забезпечує захист на рівні кожного окремого запису. У нашому випадку це означає, що шум вноситься безпосередньо до координат маршруту ще до їх аналізу або збереження, що мінімізує ризик деанонімізації навіть на рівні вихідних даних.

Процедура побудована на використанні H3-індексації, яка розбиває простір на шестикутні області фіксованого розміру. Початкові GPS-координати кожного маршруту перетворюються у відповідні H3-комірки, що дозволяє перейти від точних координат до дискретизованих просторових одиниць. Це вже само по собі забезпечує базовий рівень узагальнення та абстрагування.

Подальша анонімізація відбувається шляхом заміни справжньої H3-комірки, у якій знаходиться маршрутна точка, на іншу — вибрану випадковим чином із множини можливих осередків. Ймовірність вибору кожного з цих осередків задається згідно з певним розподілом, який контролюється параметром ε (епсілон) — ключовим числовим показником, що визначає рівень приватності.

Параметр ε має суттєвий вплив на баланс між точністю даних і приватністю. Якщо значення ε велике (наприклад, ε = 10), то розподіл ймовірностей зосереджується навколо справжньої комірки — тобто з великою ймовірністю вибір не змінює початкову точку. У цьому випадку точність маршруту зберігається, але рівень приватності суттєво знижується. Навпаки, за низьких значень ε (наприклад, ε = 1), розподіл стає більш рівномірним — імовірність вибору іншої комірки зростає, що підвищує захист приватності, але зменшує точність.

Таким чином, застосування локальної диференційної приватності в геолокаційних даних дозволяє гнучко налаштовувати рівень анонімізації в залежності від цілей дослідження, типу даних і вимог до збереження їхньої аналітичної цінності. У запропонованому підході цей механізм виступає ключовим інструментом забезпечення просторової конфіденційності маршрутів.

У проєкті використано ε = 7, що дає баланс між збереженням логіки міських маршрутів і достатньою приватністю користувача. Нехай початковий H3-осередок буде позначений $h_0$ , а зашумлений буде позначатися $h'$.

Застосувавши метод LDP для обраних значень отримаємо наступні розрахунки:

$$P(h') = \begin{cases} p = \frac{1}{2}, \text{ якщо } h' = h_0 \\ q = \frac{1}{e^7 + 1}, \text{ якщо } h' \neq h_0 \end{cases}$$

$$P(h') = \begin{cases} p = 0.5, \text{ якщо } h' = h_0 \\ q = 0.00091188, \text{ якщо } h' \neq h_0 \end{cases}$$

Це означає що з 48 H3-комірок ймовірність того, що після застосування цього методу, буде обрана поточна комірка дорівнює 0.5, а ймовірність що буде обрана будь-яка з решти 47 H3-комірок дорівнює 0.00091188.

Таким чином, справжнє місцезнаходження користувача буде обране лише у половині випадків, а решта — випадково з іншої частини міста, що гарантує приватність маршруту.

### 2.4 Створення синтетичних маршрутів

Після того як початкові координати користувача було замінено на зашумлені з використанням механізму диференційної приватності, виникає потреба побудувати синтетичний маршрут між цими новими точками. Такий маршрут, хоча й не відображає реального шляху користувача, може використовуватись для аналітики транспортних потоків, загального навантаження на інфраструктуру та дослідження мобільності у межах міста.

Першим етапом відновлення маршруту є визначення координат зашумленого старту та фінішу. Оскільки кожна точка була спочатку переведена у шестикутну комірку за допомогою H3-індексації, після застосування механізму шуму визначаються нові комірки. Центри цих комірок і використовуються як нові координати початку та завершення маршруту.

Далі, для побудови маршруту між цими координатами використовується зовнішній геоінформаційний сервіс, зокрема Open Source Routing Machine (OSRM), що базується на відкритих даних OpenStreetMap. Завдяки цьому сервісу формується реалістичний автомобільний маршрут між точками, який враховує дорожню мережу, тип покриття, наявність перехресть та інші навігаційні особливості. Це дозволяє отримати маршрут, який хоча й не є справжнім, проте

зберігає логіку реального руху містом і відповідає реальній транспортній ситуації.[14]

Синтетичні маршрути зберігаються і візуалізуються на карті у вигляді ліній, які за кольором відрізняються від оригінальних, що дозволяє порівняти ступінь спотворення даних. Це дає змогу оцінити, наскільки сильно змінилися координати після зашумлення і як це вплинуло на топологію маршруту (рис.2.5).

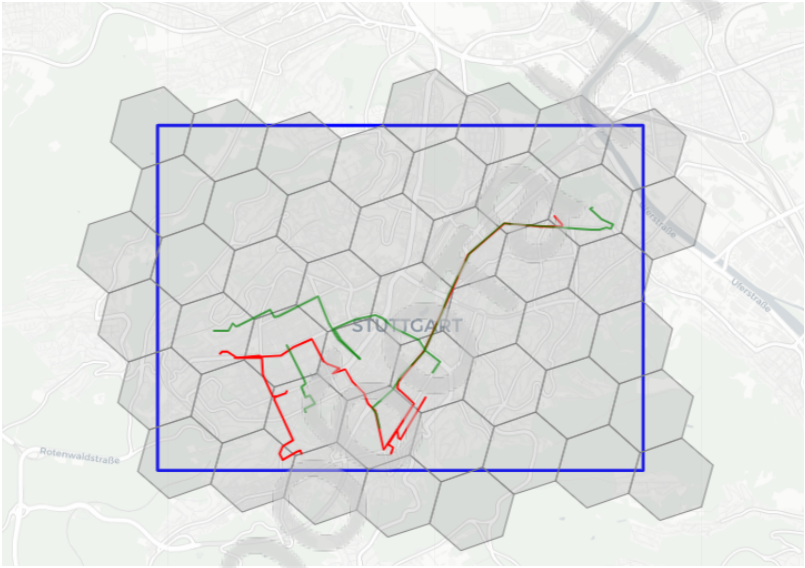


Рис.2.5. Оригінальні маршрути (червоні лінії) та синтетичні маршрути (зелені лінії)

В окремих випадках траєкторії можуть проходити через сусідні райони міста або починатися на декілька кварталів далі від початкової точки. Це залежить від того, яку саме зашумлену комірку було обрано механізмом приватності. Проте завдяки тому, що вибір нових точок підпорядковується імовірнісному розподілу, який гарантує більшу вагу для ближчих осередків, відхилення зазвичай не є надто великим.

### 2.5 Аналіз розподілу маршрутів

Після генерації як реальних, так і синтетичних маршрутів, наступним логічним кроком є порівняння їх просторових характеристик. Для цього необхідно перетворити кожен маршрут на розподіл за географічними зонами — у нашому випадку це H3-осередки певної роздільності. Такий підхід дозволяє не лише стандартизувати просторові дані, але й порівнювати траєкторії між собою з урахуванням частоти перебування в тих чи інших осередках.

Для кожного маршруту — як оригінального, так і синтетичного — визначалося, які H3-осередки він перетинає. Далі підраховувалася кількість появ кожної осередки на маршруті, і на основі цих частот формувався нормалізований розподіл.

Наприклад, якщо певний маршрут проходив через 10 осередків, кожна з яких була відвідана по одному разу, то ймовірність для кожної з них складала 0.1.

Формально, для кожного маршруту будується вектор P, де:

$$P(h_i) = \frac{\text{кількість разів, коли маршрут проходить через } h_i}{\text{загальна кількість осередків}}$$

$h_i -$ множина осередків

Таким чином, ми маємо два ймовірнісні розподіли: один — для реального маршруту, інший — для синтетичного. Щоб оцінити, наскільки ці два розподіли відрізняються один від одного, була застосована метрика Topsoe-дивергенції.

В рамках проєкту для кожної пари маршрутів (реальний — синтетичний) була обчислена Topsoe-дивергенція. Далі для кожного H3-осередку, що з'являвся у цих маршрутах, було усереднено значення дивергенцій — тобто отримано, наскільки в середньому змінюється імовірність присутності маршруту в цьому осередку при переході від реальних координат до зашумлених.

```
📊 Розраховані значення Topsoe-дивергенції по H3-осередках:

H3 cell: 881faa7ae3fffff → Topsoe divergence: 0.3469
H3 cell: 881faa7a8dfffff → Topsoe divergence: 0.4140
H3 cell: 881faa7aebfffff → Topsoe divergence: 0.5048
H3 cell: 881faa7a85fffff → Topsoe divergence: 0.3716
H3 cell: 881faa7ad3fffff → Topsoe divergence: 0.3021
H3 cell: 881faa7a8bfffff → Topsoe divergence: 0.3831
H3 cell: 881faa7a99fffff → Topsoe divergence: 0.3512
H3 cell: 881faa7ad7fffff → Topsoe divergence: 0.4322
H3 cell: 881faa7a83fffff → Topsoe divergence: 0.3706
H3 cell: 881faa7a87fffff → Topsoe divergence: 0.3072
H3 cell: 881faa71a5fffff → Topsoe divergence: 0.1309
H3 cell: 881faa7a89fffff → Topsoe divergence: 0.4334
H3 cell: 881faa7a81fffff → Topsoe divergence: 0.3880
H3 cell: 881faa7ac7fffff → Topsoe divergence: 0.3982
H3 cell: 881faa7aa9fffff → Topsoe divergence: 0.6042
H3 cell: 881faa7a9dfffff → Topsoe divergence: 0.4795
```

Рис.2.6. Розраховані метрики для кожного H3-осередку

Візуалізація цих значень у вигляді теплової карти дозволяє швидко виявити райони, у яких спотворення були найбільшими.

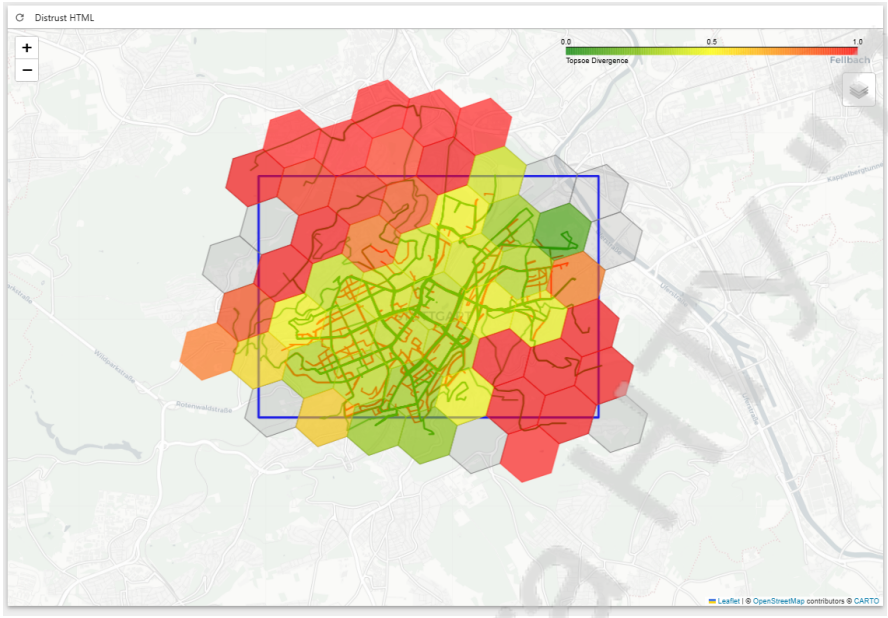


Рис.2.7. Теплова карта дивергенції Топсе

На представленій карті (рис.2.7) відображено результати аналізу анонімізованих маршрутів самокатів у місті Штутгарт із застосуванням гексагональної сітки. Кольорова шкала візуалізує ступінь відхилення дивергенції Топсе між оригінальними та синтетичними даними в кожній зоні: зелені області вказують на високу подібність, тоді як червоні — на значні відмінності.

Найбільша відповідність зберігається в центральній частині міста, де маршрути мають високу щільність та регулярність. Це свідчить про успішне відтворення патернів руху у синтетичних даних без значної втрати аналітичної цінності. У периферійних регіонах, де маршрути менш стабільні або фрагментарні, спостерігається більша розбіжність між реальними та згенерованими маршрутами, що є очікуваним через меншу кількість вхідних даних у цих зонах.

Загалом, отримана візуалізація демонструє ефективність підходу до генерації синтетичних маршрутів. Метод дозволяє зберігати просторові закономірності в зонах високої активності, одночасно забезпечуючи високий рівень приватності користувачів завдяки втраті точності у менш критичних зонах.

**Висновки за розділом**

У спеціальному розділі було проведено комплексне дослідження підходів до забезпечення приватності геолокаційних даних з використанням локальної диференційної приватності (LDP), побудови H3-осередків для просторової агрегації та оцінювання схожості маршрутів за допомогою Topsoe-дивергенції.

В ході аналізу було відновлено маршрути користувачів на основі зашумлених координат, згенерованих за принципом локальної диференційної приватності. Була використана формула ймовірнісного вибору координати, що відповідає механізму Randomized Response, з параметром ε, який задає рівень приватності. Це дозволило зберегти конфіденційність реальних точок маршруту, водночас зберігаючи можливість агрегованого аналізу.

Далі було побудовано розподіли маршрутів у вигляді H3-осередків з певною резолюцією, що дозволило просторово агрегувати дані та спростити їх подальший аналіз. H3-гексагональна сітка забезпечила стабільну та ефективну систему кодування простору, яка добре підходить для задач геоаналітики.

На завершення, для оцінки точності відновлення маршрутів та порівняння розподілів справжніх і зашумлених маршрутів була застосована метрика Topsoe-дивергенції. Результати розрахунків показали прийнятну схожість між реальними та відновленими даними при допустимому рівні приватності, що свідчить про збалансованість між точністю реконструкції та рівнем анонімізації.

# ВИСНОВКИ

У межах цієї кваліфікаційної роботи було проведено комплексне дослідження проблеми захисту персональних геолокаційних даних у сфері мікромобільності, зосереджене на побудові методів анонімізації маршрутів користувачів. В роботі проаналізовано сучасні ризики, пов'язані з використанням геоданих, і доведено, що традиційні методи анонімізації є недостатніми для забезпечення належного рівня конфіденційності.

У відповідь на цю проблему було розроблено методологію, яка поєднує просторову агрегацію координат із використанням H3-індексації, локальну диференційну приватність (LDP) як механізм анонімізації, та генерацію синтетичних маршрутів за допомогою алгоритмів маршрутизації на основі реальних дорожніх мереж. Такий підхід дозволив створювати штучні маршрути, що не містять справжніх координат, але зберігають основні закономірності руху користувачів у міському просторі.

Було впроваджено механізм стохастичного вибору H3-осередків із параметром ε, який формально забезпечує ε-LDP, і дозволяє балансувати між точністю та рівнем приватності. Для оцінки якості збереження структури маршрутів застосовано дивергенцію Топсое, яка дозволила кількісно оцінити ступінь відхилення між розподілами справжніх та синтетичних маршрутів.

Отримані результати свідчать про ефективність запропонованого підходу: синтетичні маршрути демонструють високий рівень просторової подібності в центральних зонах міста, при цьому значно знижуючи ризики деанонімізації. Це дозволяє зберегти аналітичну цінність маршрутних даних для транспортного планування, досліджень мобільності та міської аналітики.

Практична значущість роботи полягає у тому, що розроблена система анонімізації може бути інтегрована у цифрові платформи мікромобільності, забезпечуючи як приватність користувачів, так і інформативність даних. Таким чином, робота робить внесок у розвиток надійних та відповідальних технологій обробки просторових даних у сфері урбаністики та інтелектуального транспорту.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ


1. Regulation (EU) 2016/679 of the European Parliament and of the Council
2. El Emam, K., Jonker, E., Arbuckle, L., & Malin, B. (2011). *A Systematic Review of Re-Identification Attacks on Health Data*. с. 3-5.
3. Monreale, A., et al. (2010). *Privacy-aware Geographical Data Publishing*. с. 4-6.
4. Gedik, B., & Liu, L. (2008). *Protecting Location Privacy with Personalized k-Anonymity: Architecture and Algorithms*. - 8 с.
5. Sweeney, L. (2002). *k-Anonymity: A Model for Protecting Privacy*. Int. J. of Uncertainty, Fuzziness and Knowledge-Based Systems, 8 с.
6. Zang, H., Bolot, J. (2011). Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study. *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking (MobiCom '11)*. с. 145–156.
7. MacQueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, с. 281–297.
8. Ester, M., Kriegel, H.-P., Sander, J., Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96)*, с. 226–231.
9. Rokach, L., Maimon, O. (2005). Clustering Methods. In: *Data Mining and Knowledge Discovery Handbook*. Springer, с. 321–352.
10. Adrian Wöltche “Open source map matching with Markov decision processes: A new method and a detailed benchmark with existing approaches”. с. 5-6.
11. Topsøe, F. (2000). *Some inequalities for information divergence and related measures of discrimination*. IEEE Transactions on Information Theory, 46(4), 1602–1609.
12. Wang, T., Blocki, J., Li, N. & Jha, S. Locally Differentially Private Protocols for Frequency Estimation. 26th USENIX Security Symposium (USENIX Security 17). ст. 9

13. Uber Engineering Blog – Hexagonal Hierarchical Spatial Index (H3)
14. Open Source Routing Machine (OSRM) Documentation

Додаток А.

| № з/п | Позначення | Найменування | Кількість аркушів | Примітки |
|---|---|---|---|---|
| 1 | | | | |
| 2 | | Документація | | |
| 3 | | | | |
| 4 | САУ.КР.25.45.ПЗ | Пояснювальна записка | 35 | Формат А4 |
| 5 | | | | |
| 6 | | Демонстраційний матеріал | 14 | Презентація на CD-R |
| 7 | | | | |
| 8 | | Копія роботи | 1 | Диск CD-R |
| 9 | | | | |
| 10 | | | | |
| 11 | | | | |
| 12 | | | | |
| 13 | | | | |
| 14 | | | | |
| 15 | | | | |
| 16 | | | | |
| 17 | | | | |
| 18 | | | | |

| | | | | | САУ.КР.25.45.ДА.ПЗ. | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | |
| Змін. | Аркуш | № докум. | Підпис | Дата | | | | |
| Розроб. | | Сабельников А.В. | | | **Матеріали кваліфікаційної роботи** | Літ. | Аркуш | Аркушів |
| К. розд. | | Малієнко А.В. | | | | | | |
| Керівн. | | Малієнко А.В. | | | | НТУ «ДП», 12; 124-21-2 | | |
| Н.контр. | | Хом’як Т.В. | | | | | | |
| Зав. каф. | | Желдак Т.А. | | | | | | |

Додаток Б.

ВІДГУК

на кваліфікаційну роботу бакалавра

на тему: «Генерація та аналіз синтетичних маршрутів як метод захисту персональних даних у сервісах суспільного транспорту »

Студента Сабельникова А.В академічної групи 124-21-2

Обсяг кваліфікаційної роботи 30 стор.

Мета кваліфікаційної роботи – розробка, дослідження реалізація методу генерації синтетичних маршрутів, який забезпечує та анонімність персональних даних користувачів у цифрових сервісах громадського транспорту, зберігаючи при цьому корисні статистичні властивості даних.

Актуальність теми обумовлена значенням застосування управління в стратегій розвитку та вдосконалення логістичних послуг та захисту персональних даних в сучасних умовах економіки України. При цьому використання актуальний математичний апарату на основі сучасних методів транспортних задач та проведено комплексне дослідження проблеми захисту персональних геолокаційних даних у сфері мікромобільності, зосереджене на побудові методів анонімізації маршрутів користувачів.

Тема кваліфікаційної роботи безпосередньо пов’язана з об'єктом діяльності бакалавра спеціальності 124 Системний аналіз, оскільки в роботі проведений аналіз та розроблено методологію, яка поєднує просторову агрегацію координат із використанням H3-індексації, локальну диференційну приватність (LDP) як механізм анонімізації, та генерацію синтетичних маршрутів за допомогою алгоритмів маршрутизації на основі реальних дорожніх мереж.

Виконані в кваліфікаційній роботі завдання відповідають вимогам ступеня бакалавра.

Практична цінність отриманих результатів полягає в розробці ефективного методу анонімізації маршрутних даних, який дозволяє захищати приватність користувачів сервісів мобільності без втрати аналітичної цінності даних для досліджень та міського планування.

Висновки підтверджують можливість використання результатів роботи в умовах сучасного підприємства.

Оформлення пояснювальної записки та демонстраційного матеріалу до неї виконано згідно з вимогами але мають зауваження.

Роботу виконано самостійно, відповідно до завдання та у повному обсязі

У роботі відзначено такі недоліки:

1. Не проведений аналіз інших методів які можна використати як альтернативу.

2. Відсуне представлення реальних даних на гео-позиційне налаштування при впровадженні на підприємствах.

3. Відсутні розрахункові дані при впровадженні та не наведений алгоритм захисту даних для підприємства.

Кваліфікаційна робота в цілому заслуговує оцінки: задовільно (64 бали) при відповідному захисті.

З урахуванням висловлених зауважень автор заслуговує присвоєння освітньої кваліфікації «бакалавр з системного аналізу».

Науковий керівник

К.т.н. доц. кафедри СА і У Малієнко А.В.

Додаток В

Рецензія

на кваліфікаційну роботу бакалавра
Студента Сабельникова А.В академічної групи 124-21-2

Тема кваліфікаційної роботи: Генерація та аналіз синтетичних маршрутів як метод захисту персональних даних у сервісах суспільного транспорту

Обсяг кваліфікаційної роботи: 30 *стор.*

Висновок про відповідність кваліфікаційної роботи завданню та освітньо-професійній програмі спеціальності кваліфікаційна робота відповідає перевірці знань та рівня підготовки виконавця за фахом спеціальності 124 Системний аналіз

Загальна характеристика кваліфікаційної роботи, актуальні дані та посилання на іноземні джерела інформації є перспективним до економіки України. А ступінь використання літератури показує відповіднє орієнтування студента в темі роботи. Проаналізований матеріал інтернет ресурсів обраного направлення теми кваліфікаційної роботи є перспективним та актуальним.

Зміст робити відповідає виданому завданню, але не розкритий в повній мірі.

Матеріал у роботі викладені згідно тематики дослідження. Висновки в кінці розділів відповідають аналізу теми.

Позитивні сторони кваліфікаційної роботи: простежується аналіз матеріалу сучасного стану питання. В роботі в певній мірі розкрита тема та вирішені завдання. Також зазначені певні рекомендації на основі результатів, які можуть мати практичну значимість для підприємств України.

Основні недоліки кваліфікаційної роботи:

1. В кваліфікаційній роботі не повністю зрозуміло, які методи існують для вирішеня поставленої задачі та їх практичне застосування. Не наведений реальний стан та приклад розрахунку -хоча є пені числа.
2. Відсутня інформація про можливі впровадження та ефекти, захист інформації та використання альтернатив.

Кваліфікаційна робота в цілому заслуговує оцінки: задовільно -64 балів.

З урахуванням висловлених зауважень автор заслуговує присвоєння освітньої кваліфікації «бакалавр з системного аналізу».

Рецензент,

К.т.н. доц. кафедри СА і У Малієнко А.В.

Додаток Г.

```
import pandas as pd
import numpy as np
import folium
import h3
import requests
import random
from tqdm import tqdm
from collections import defaultdict
import branca.colormap as cm
from ast import literal_eval
from scipy.stats import entropy

# === ПАРАМЕТРИ ===
h3_resolution = 8
epsilon = 7
n_routes_to_visualize = 500
stuttgart_bbox = {"north": 48.80005, "south": 48.76265, "west": 9.13885, "east": 9.21885}

# === ЗАВАНТАЖЕННЯ ДАНИХ ===
df = pd.read_json('output_150.json')
print("Перші 5 рядків початкових даних:")
df.head()[["reservationId", "drivenDistance", "usageTime", "startTime", "endTime", "startLat", "startLon", "endLat", "endLon",
"waypoints"]]

# Фільтрація маршрутів, які повністю знаходяться в межах міста Штутгарт
df_filtered = df[
    (df["startLat"].between(stuttgart_bbox["south"], stuttgart_bbox["north"])) &
    (df["startLon"].between(stuttgart_bbox["west"], stuttgart_bbox["east"])) &
    (df["endLat"].between(stuttgart_bbox["south"], stuttgart_bbox["north"])) &
    (df["endLon"].between(stuttgart_bbox["west"], stuttgart_bbox["east"]))
].copy()

# === ДОДАВАННЯ H3-КОМІРКИ ДЛЯ ПОЧАТКУ ТА КІНЦЯ МАРШРУТУ ===
df_filtered["start_hex"] = df_filtered.apply(lambda r: h3.latlng_to_cell(r["startLat"], r["startLon"], h3_resolution), axis=1)
df_filtered["end_hex"] = df_filtered.apply(lambda r: h3.latlng_to_cell(r["endLat"], r["endLon"], h3_resolution), axis=1)

print("Перші 5 рядків після обробки:")
df_filtered.head()[["reservationId", "drivenDistance", "usageTime", "startTime", "endTime", "startLat", "startLon", "endLat",
"endLon", "waypoints","start_hex","end_hex"]]

# === СТВОРЕННЯ МАПИ ===
stuttgart_center = [48.78135, 9.17885]
stuttgart_map = folium.Map(location=stuttgart_center, zoom_start=12, tiles="cartodbpositron")

# === ДОДАЄМО МЕЖІ МІСТА (BBox) ===
folium.PolyLine([
    [stuttgart_bbox["north"], stuttgart_bbox["west"]],
    [stuttgart_bbox["north"], stuttgart_bbox["east"]],
    [stuttgart_bbox["south"], stuttgart_bbox["east"]],
    [stuttgart_bbox["south"], stuttgart_bbox["west"]],
    [stuttgart_bbox["north"], stuttgart_bbox["west"]]
], color='blue', weight=3).add_to(folium.FeatureGroup(name='Межі Штутгарта').add_to(stuttgart_map))

# Слой для візуалізації H3-комірок
hex_layer = folium.FeatureGroup(name="H3-сітка в межах міста")

# === УНІКАЛЬНІ H3-КОМІРКИ В BBOX ===
hexes = {
    h3.latlng_to_cell(lat, lon, h3_resolution)
```

```
    for lat in np.arange(stuttgart_bbox["south"], stuttgart_bbox["north"], 0.0025)
    for lon in np.arange(stuttgart_bbox["west"], stuttgart_bbox["east"], 0.0025)
}
print(f"Кількість унікальних H3-комірок у місті: {len(hexes)}")

# Візуалізація кожної комірки на карті
for h in hexes:
    boundary = h3.cell_to_boundary(h)
    folium.Polygon(
        locations=boundary,
        color="gray",
        fill=True,
        fill_opacity=0.2,
        weight=1
    ).add_to(hex_layer)

hex_layer.add_to(stuttgart_map)

# === ДОДАВАННЯ ШУМУ ДЛЯ АНОНІМІЗАЦІЇ ===
def apply_privacy_noise(hex_id):
    p = 1 / 2
    q = 1 / (np.exp(epsilon) + 1)
    candidates = list(hexes)
    weights = [p if h == hex_id else q for h in candidates]
    return random.choices(candidates, weights=weights, k=1)[0]

df_vis = df_filtered.head(n_routes_to_visualize).copy()
df_vis.loc[:, "start_hex_noisy"] = df_vis["start_hex"].apply(apply_privacy_noise)
df_vis.loc[:, "end_hex_noisy"] = df_vis["end_hex"].apply(apply_privacy_noise)

# === ДОПОМІЖНІ ФУНКЦІЇ ===
def get_hex_center(h3_id):
    return h3.cell_to_latlng(h3_id)

def get_osrm_route(points, max_points=50):
    path = []
    for i in range(0, len(points) - 1, max_points):
        segment = points[i:i + max_points + 1]
        coords = ";".join([f"{lon},{lat}" for lat, lon in segment])
        url = f"http://router.project-osrm.org/route/v1/driving/{coords}?overview=full&geometries=geojson"
        r = requests.get(url)
        if r.ok:
            data = r.json()
            if data['routes']:
                path.extend([(lat, lon) for lon, lat in data['routes'][0]['geometry']['coordinates']])
    return path or points

def parse_waypoints(w):
    if isinstance(w, str):
        w = literal_eval(w)
    if isinstance(w, dict):
        return [[float(p["lat"]), float(p["lon"])] for p in w.values()]
    elif isinstance(w, list):
        return [[float(p["lat"]), float(p["lon"])] for p in w]
    return []

def route_to_h3_distribution(route):
    counts = defaultdict(int)
    for lat, lon in route:
        h = h3.latlng_to_cell(lat, lon, h3_resolution)
```

```
        counts[h] += 1
    total = sum(counts.values())
    return ({h: c / total for h, c in counts.items()}, counts) if total else ({}, {})

def compute_topsoe(p, q):
    keys = set(p) | set(q)
    p_arr = np.array([p.get(k, 0) for k in keys])
    q_arr = np.array([q.get(k, 0) for k in keys])
    m = (p_arr + q_arr) / 2
    p_arr = np.where(p_arr == 0, 1e-10, p_arr)
    q_arr = np.where(q_arr == 0, 1e-10, q_arr)
    m = np.where(m == 0, 1e-10, m)
    return entropy(p_arr, m) + entropy(q_arr, m)

# === ВІЗУАЛІЗАЦІЯ ОРИГІНАЛЬНИХ МАРШРУТІВ ===
original_routes = []
orig_layer = folium.FeatureGroup(name="Оригінальні маршрути").add_to(stuttgart_map)
for _, row in df_vis.iterrows():
    route = parse_waypoints(row["waypoints"])
    if not route:
        print(f"⚠️ Пропущено: порожній маршрут")
        continue
    original_routes.append(route)
    folium.PolyLine(route, color="red", weight=2).add_to(orig_layer)

# === ВІЗУАЛІЗАЦІЯ ШУМНИХ МАРШРУТІВ ===
noisy_routes = []
noisy_layer = folium.FeatureGroup(name="Шумні маршрути").add_to(stuttgart_map)
for _, row in tqdm(df_vis.iterrows(), total=len(df_vis)):
    start = get_hex_center(row["start_hex_noisy"])
    end = get_hex_center(row["end_hex_noisy"])
    path = get_osrm_route([start, end])
    noisy_routes.append(path)
    folium.PolyLine(path, color="green", weight=2, opacity=0.7).add_to(noisy_layer)

# === ОБЧИСЛЕННЯ ТОПСОЕ-ДИВЕРГЕНЦІЇ ===
div_sum = defaultdict(float)
div_count = defaultdict(int)

for orig, noisy in zip(original_routes, noisy_routes):
    p, _ = route_to_h3_distribution(orig)
    q, _ = route_to_h3_distribution(noisy)
    div = compute_topsoe(p, q)
    for h in set(p) | set(q):
        div_sum[h] += div
        div_count[h] += 1

div_avg = {h: div_sum[h] / div_count[h] for h in div_sum}

print("📊 Розраховані значення Topsoe-дивергенції по H3-осередках:\n")

for h, val in div_avg.items():
    print(f"H3-комірка: {h} → Topsoe-дивергенція: {val:.4f}")

# === ВІЗУАЛІЗАЦІЯ ТЕПЛОВОЇ КАРТИ ДИВЕРГЕНЦІЇ ===
colormap = cm.LinearColormap(['green', 'yellow', 'red'], vmin=0, vmax=1, caption="Topsoe-дивергенція")
heat_layer = folium.FeatureGroup(name="Карта Topsoe-дивергенції").add_to(stuttgart_map)

for h, val in div_avg.items():
    folium.Polygon(
```

```
        locations=h3.cell_to_boundary(h),
        color=colormap(val),
        fill=True,
        fill_opacity=0.6,
        weight=0.5,
        tooltip=f"H3: {h}, Дивергенція: {val:.3f}"
    ).add_to(heat_layer)

colormap.add_to(stuttgart_map)
folium.LayerControl().add_to(stuttgart_map)

# === ЗБЕРЕЖЕННЯ РЕЗУЛЬТАТУ ===
stuttgart_map.save("stuttgart_scooter_routes.optimized500.html")
```

Додаток Д.

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Запорізька політехніка»

# СЕРТИФІКАТ

ФНКТ-0425-34
засвідчує, що
Андрій САБЕЛЬНИКОВ
був учасником II (VIII) міжнародної науково-практичної конференції здобувачів вищої освіти і молодих учених **«Інформаційні технології: теорія та практика»**
**2 - 4 квітня 2025 р.**
**Запоріжжя-Харків-Дніпро**
обсягом 24 години

*В.о. начальника науково-дослідної частини* — *Наталя Висоцька*

# СЕРТИФІКАТ

засвідчує, що

**Сабельников А. В.**

був (була) учасником II (VIII) Міжнародної науково-практичної конференції здобувачів вищої освіти і молодих учених
**«Інформаційні технології: теорія і практика»**

**2-4 квітня 2025 р.**
**Запоріжжя – Харків – Дніпро**

Ректор НУ «Запорізька політехніка» — Віктор ГРЕШТА

ЗАПОРІЗЬКА ПОЛІТЕХНІКА
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ