

ЕКСТРАПОЛЯЦІЯ ЯКОСТІ ПОВІТРЯ В МІСЬКОМУ СЕРЕДОВИЩІ З ОБМЕЖЕНОЮ КІЛЬКІСТЮ СТАНЦІЙ

Д. К. Городецька^{1,a}, Г. О. Яйлимова^{1,2}

¹ Навчально-науковий Фізико-технічний інститут

² Інститут космічних досліджень НАНУ та ДКАУ

Анотація

Сьогодні процес урбанізації виступає як каталізатор зростання міст та збільшення питомої ваги міського населення. Проте цей процес супроводжується багатьма проблемами, а саме погіршення якості повітря, забруднення водойм, зменшення зелених насаджень. Це все є однією з головних загроз для здоров'я міських мешканців, ставши причиною мільйонів передчасних смертей по всьому світу. Саме тому моніторинг якості повітря стає надзвичайно важливим інструментом для забезпечення здоров'я населення та планування дій для покращення якості повітря. В Україні, як і в багатьох інших країнах, для моніторингу якості повітря традиційно використовуються наземні станції. Проте, з огляду на значні території міст та обмежену кількість таких станцій, існує важлива потреба в розробці та впровадженні більш комплексних методів оцінки якості повітря. Це стосується не лише великих урбанізованих агломерацій, а й менших міських поселень, де моніторинг якості повітря часто є обмеженим або недоступним.

Ключові слова: якість повітря, недостатнє покриття, екстраполяція даних

Вступ

З початком пандемії COVID-19 та повномасштабного вторгнення російської федерації в Україну 24 лютого 2022 року, Україна зіткнулася з багатьма проблемами, але одна з них досі не виноситься на загал. Це проблема погіршення якості повітря. Пандемія спричинила багато змін в екології, а саме у якості повітря через зміни в промисловому виробництві та транспортних звичках. Воєнні дії призвели до додаткового забруднення навколишнього середовища, що вплинуло і на якість повітря. У результаті цих подій, місто Київ, як і багато інших регіонів України, зіткнулося із суттєвими викликами щодо моніторингу та оцінки стану якості повітря.

1. Проблема обмеженості моніторингових станцій та світовий досвід їх розширення

Проблема обмеженості моніторингових станцій виходить з того, що традиційні моніторингові системи складаються з невеликої кількості станцій, розташованих у ключових локаціях міста або регіону. Моніторингові системи оснащені високоточним обладнанням для вимірювання рівнів забруднюючих речовин у повітрі, таких як діоксид азоту, озон, тверді частки тощо. Однак, через високу вартість та складність обслуговування такого обладнання, їхня кількість залишається обмеженою, що призводить

до недостатнього географічного покриття та втрати інформації про якість повітря у багатьох районах. Таким чином ми можемо переглянути стан повітря в деяких районах Києва, проте вийшовши за межі Києва в область або навіть на лівий берег міста переглянути стан повітря в деяких місцях буде неможливим. Дивлячись показники інших міст ситуація стає гіршою, бо в багатьох селах, містах відсутні датчики взагалі. Для рішення цієї проблеми потрібно розглянути існуючі рішення для моніторингу якості повітря на прикладі світового досвіду, а саме розробку та впровадження мобільних та низьковартісних сенсорних мереж, що дозволяють значно розширити мережу моніторингу за рахунок збільшення кількості точок збору даних. Також використання методів математичного моделювання для прогнозування стану повітря на різних відстанях.

В Гонконзі моніторинг якості повітря відбувається за допомогою інноваційних технологій. Проблема нестачі датчиків була розглянута у дослідженні [1], вона вирішується за допомогою Концепції Системи Моніторингу Забруднення Повітря Наступного Покоління (TNGAPMS), яка використовує передові технології сприйняття, MicroElectroMechanical Systems (MEMS), та бездротові сенсорні мережі (WSN). Існують три основні типи сенсорних мереж на основі WSN: Статична Сенсорна Мережа (SSN), Спільнотна Сенсорна Мережа (CSN) і Транспортна Сенсорна Мережа (VSN).

- SSN. Включають статичні сенсори, які розміщу-

^abasyuk.daria@lil.kpi.ua

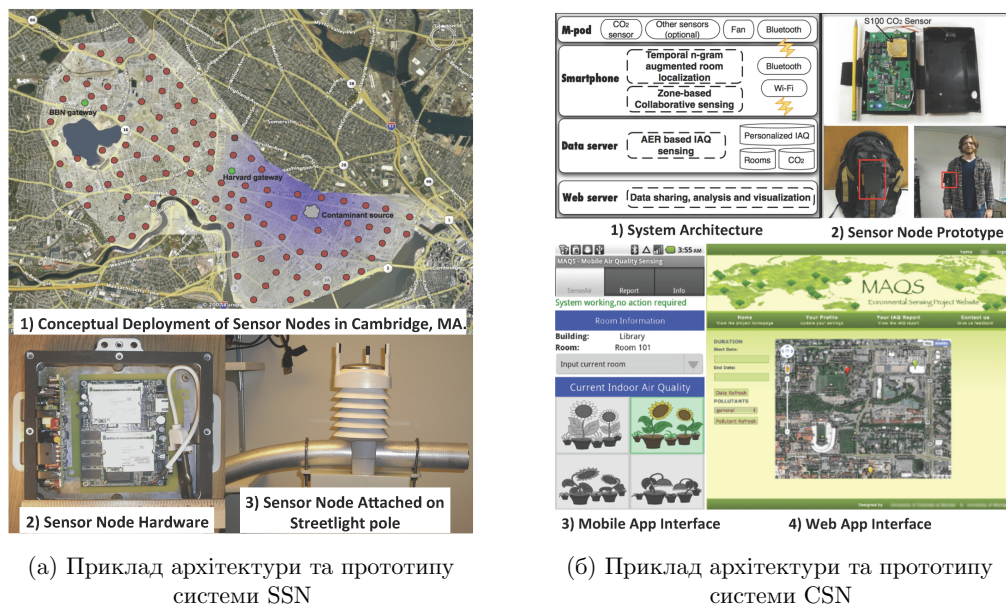


Рис. 1. Приклади архітектури сенсорний мереж

ються на вуличних ліхтарях або стовпах. Вони характеризуються високою надійністю даних і можливістю використання різноманітних сенсорів на одному вузлі, але потребують планування розміщення та затрат на обслуговування. На Рис. 1(а) зображено приклад цієї моделі.

- CSN. Включають мобільні сенсори, що носять з собою люди. Тут вже відсутня проблема частоти вимірів, але стикаються можлива низька точність даних та питань приватності. На Рис. 1(б) зображено приклад цієї моделі.
- VSN. Це сенсори на транспортних засобах (автобуси або таксі), для збору даних по широкій географічній території. Вони дають широке географічне покриття і легкість обслуговування, але мають обмежену часову роздільну здатність і потребують використання GPS для визначення місцезнаходження.

Таким чином, дослідження [1] наголошує на потребу в розробці більш гнучких та ефективних систем моніторингу якості повітря, які можуть забезпечити більш детальну і актуальну інформацію про стан атмосферного повітря у міських умовах.

В деяких країнах проблема вирішується за допомогою розробки інтерактивної веб-картографічної візуалізації, яка охоплює усі основні міста на національному рівні. Дослідження [2] розробляє два нових візуалізаційні інструменти для візуалізації часових рядів з різною зернистістю (timezoom.js) та динамічного контролю символів на карті для тематичного картографування (symadaptive.js).

2. Аналіз недоліків існуючих підходів та обґрунтування необхідності нової моделі

Варто розглянути також новий підхід до проблеми просторово-часової екстраполяції, яка полягає у

передбаченні невідомих даних за допомогою використання інформації з інших точок у просторово-часовій структурі. У статті [3] впроваджено модель Spatio-Temporal Graph Neural Processes (STGNP). Вона є нейронною мережею з прихованими змінними, спеціально розробленою для аналізу графових даних. Модель складається з двох основних етапів:

- Детермінована стадія, на якій відбувається вивчення просторово-часових представлень вузлів за допомогою згорткових мереж і перехресних графових нейронних мереж.
- Стохастичний етап, на якому застосовуються переходи станів для агрегування латентних змінних з урахуванням часової незалежності та довгострокової часової еволюції. Важливою особливістю є використання Graph Bayesian Aggregation (GBA), який агрегує інформацію з контекстних вузлів, враховуючи невизначеності.

Ця модель здатна вловлювати складні просторово-часові зв'язки між даними, що робить її особливо корисною для роботи з даними датчиків, розташованих на великих відстанях один від одного. Однак існують певні виклики, пов'язані зі складністю та масштабованістю моделі, а також залежністю від структури графа, які потребують подальших досліджень та удосконалень.

Недоліки існуючих систем:

1. Обмеження просторового покриття. Традиційні стаціонарні моніторингові станції зосереджені переважно у великих містах та промислових зонах, що залишає значні території без адекватного моніторингу. Це призводить до втрати важливих даних із віддалених або менш населених регіонів.
2. Висока вартість та складність утримання. Використання високоточного обладнання, яке часто є дорогим та складним у обслуговуванні, обме-

жує можливість розширення мережі моніторингу через фінансові та технічні бар'єри.

3. Залежність від структури графа. Моделі, які використовують графові структури, часто залежать від точності та актуальності графа, що може вплинути на якість моделювання та прогнозування при недостатній інтеграції даних або при зміні умов моніторингу.
4. Складність та масштабованість моделі. Розробка та впровадження складних моделей, таких як STGNP, вимагає значних ресурсів, спеціалізованих знань і часу, що може бути важкодоступним для багатьох організацій і країн.
5. Проблеми з точністю та надійністю даних. Хоча мобільні та ком'юніті-орієнтовані сенсорні мережі забезпечують ширше географічне покриття, вони часто страждають від проблем з точністю вимірювань і можуть бути чутливими до зовнішніх перешкод.

Враховуючи зазначені вище недоліки, виникає очевидна потреба у розробці нової моделі моніторингу якості повітря, яка б могла:

- Забезпечувати більш широке і рівномірне географічне покриття.
- Бути економічно ефективною, доступною для широкого впровадження без великих капіталовкладень.
- Бути гнучкою і масштабованою, легко адаптованою до різних умов і потреб.
- Забезпечувати високу точність і надійність даних.
- Бути менш залежною від структури графа, зменшуючи вплив помилок у вихідних даних на якість результатів моделювання.

Така модель дозволить не тільки поліпшити якість моніторингу атмосферного повітря, але й стане важливим інструментом у підвищенні громадської обізнаності, плануванні урбаністичних заходів та розробці ефективних заходів з охорони здоров'я.

3. Джерела даних для екстраполяції

Ефективний моніторинг якості повітря залежить від точності, надійності та різноманіття джерел даних, які використовуються для збору інформації. Використання даних із різних джерел може значно покращити якість та охоплення моніторингових систем, забезпечуючи більш точну і комплексну картину стану атмосферного повітря. Для розробки моделі потрібно мати правильно підготовлені дані, яких в загальному вигляді не було. Саме тому під час роботи потрібно було створити власний датасет з інформацією про стан повітря міста Київ. За основу були взяті дані з сайту SaveEcoBot[4], платформа збирає дані з різних джерел, включаючи державні системи моніторингу, станції місцевої влади, громадські станції, та комерційні пристрої. Усі ці дані публікуються на карті без перевірки, і для прийняття рішень вони потребують додаткової перевірки. Також важливо, що лише державні органи влади мають право оголошувати про аварійні або надзвичай-

ні ситуації. Серед джерел даних сайту є SaveDnipro, DNEPR.com, luftdaten.info, Eco City, AirVisual, Департаменти екології та природних ресурсів різних областей України, а також інші місцеві та міжнародні організації. Робота складалась з етапів, порядок яких зображено на Рис. 2.



Рис. 2. Процес створення датасету

Варто зазначити що проміжок був обраний за весь 2022 рік. Це час відомий всім нам, але в якості аналізу він був найцікавіший. По-перше, інформація була в найбільшому об'ємі, оскільки завдяки вибору саме цього роки ми досягли більшу кількість станцій, а відповідно більше покриття. По-друге, це час завершення пандемії та початок повномасштабного вторгнення російської федерації на територію України. Саме тому обираючи дані за потрібний проміжок часу можна проаналізувати всі події та отримати цікаві кореляції.

Отже, наш датасет містить таку інформацію:

1. device.id. Ідентифікатор пристрою, що збирає дані. Цей стовпець містить унікальний код або номер, який дозволяє ідентифікувати кожен датчик у мережі моніторингу. Це може бути корисно для відстеження вихідних даних кожного пристрою, виявлення потенційних проблем з обладнанням або розміщенням пристроїв. А також розділення даних на навчальну та тестову частину відбувалось саме по індивідуальному номеру щоб виключити однакові дані в різних частинах.
2. phenomenon. Цей стовпець описує тип феномену, який вимірюється. Це можуть бути такі показники як рівень діоксиду азоту, озону, твердих часток PM2.5 або PM10, вологість, температура та різні значення. Знання феномену дозволяє розуміти, які забруднювачі або параметри повітря аналізуються.
3. value. Конкретне значення, зареєстроване датчиком. Цей стовпець містить кількісні дані, що відображають інтенсивність або концентрацію забруднювача, зазначену у стовпці phenomenon. Ці дані використовуються для оцінки якості повітря та розрахунку індексу якості повітря в певному місці.
4. logged.at. Дата та час реєстрації кожного виміру. Цей стовпець дозволяє аналізувати дані у

часовій послідовності та спостерігати за динамікою змін якості повітря. Також можна виявити циклічні зміни або тренди пов'язані з певними годинами доби, днями тижня, сезонами року тощо.

5. latitude та longitude. Географічні координати, де знаходиться датчик. Ці стовпці містять важливі просторові дані, що дозволяють аналізувати географічний розподіл якості повітря, ідентифікувати особливо забруднені зони або регіони, де якість повітря є кращою. Вони також допомагають інтегрувати дані моніторингу в геопросторові системи аналізу та картографічні сервіси.

3.1. Обробка даних

Передпроцесинг даних є фундаментальним етапом у будь-якому процесі аналізу даних. У цьому дослідженні ми стикаємося з екологічним моніторингом, а саме з моніторингом якості повітря. Цей етап підготовки даних забезпечує, що аналіз буде виконаний на чистих, точних та консистентних даних, що, у свою чергу, підвищує якість та надійність результатів аналізу.

У роботі було використано серію функцій для обробки датасету. Розглянемо детальний опис кожної з функцій:

1. `clean.data(data)`. Ця функція конвертує дані у колонках `latitude`, `longitude`, та `value` до числового формату, що дозволяє уникнути проблем з нечисловими даними в цих полях. Опція `errors='coerce'` замінює всі нечислові значення на `NaN`, що спрощує їх ідентифікацію та видалення. Після цього, функція видаляє всі рядки, які містять `NaN` у вказаних колонках. В результаті, ця функція забезпечує, що всі дані в колонках `latitude`, `longitude`, та `value` є числовими та повними (без відсутніх значень).
2. `filter.data(data)`. Ця функція фільтрує дані за географічними та категорійними критеріями. Вона відбирає рядки, де значення широти (`latitude`) знаходяться в діапазоні від 50.30 до 50.60, а значення довготи (`longitude`) – від 30.20 до 30.70. Також функція відбирає тільки ті дані, де феномен (`phenomenon`) дорівнює 'pm10'. Це дозволяє зосередитись на специфічній забруднювальній речовині та певній географічній зоні.
3. `remove.outliers(data)`. Функція видаляє викиди з датасету, використовуючи метод міжквартильного розмаху (`IQR`). Вона обчислює перший (`Q1`) і третій квартилі (`Q3`) для значень `value`, визначає `IQR` як різницю між `Q3` та `Q1`, і видаляє всі рядки, де значення `value` знаходяться за межами від $[Q1 - 1.5IQR]$ до $[Q3 + 1.5IQR]$. Це стандартна процедура для видалення викидів, яка допомагає уникнути впливу аномально високих або низьких значень на аналіз.
4. `scale.data(data)`. Ця функція застосовує масштабування `Min-Max` до колонки `value`, перетворюю-

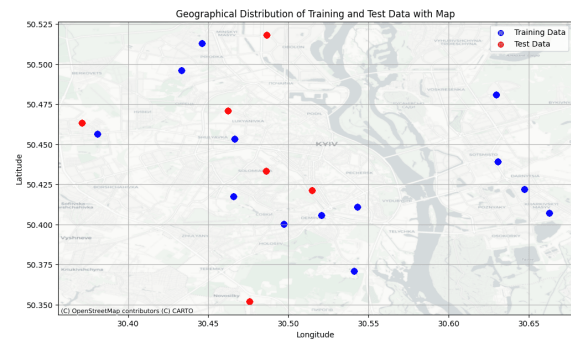


Рис. 3. Географічний розподіл навчальних і тестових даних зі значеннями показників якості повітря

ючи всі значення так, щоб вони були у діапазоні між 0 та 1. Масштабування використовується для нормалізації даних перед використанням у багатьох статистичних та машинному навчанні моделях, оскільки це забезпечує, що відмінності в масштабах різних змінних не впливають непропорційно на результати аналізу.

Після обробки даних йде розділення їх на тренувальну та тестову частину (70/30), проте це розділення має бути за окремим ідентифікатором станції, щоб в різних групах не зіткатись з проблемою однакових даних, яка призводить до помилок у роботі моделі. На Рис. 3 зображено розподіл станцій на мапі Києва.

4. Розробка моделі екстраполяції якості повітря

Процес розробки моделі екстраполяції якості повітря на основі інверсного відстаневого зважування (`IDW`) і стратифікованої крос-валідації включає кілька ключових етапів, які дозволяють забезпечити точність і надійність результатів прогнозування. На Рис. 4 зображено п'ять основних кроків у розробці моделі.

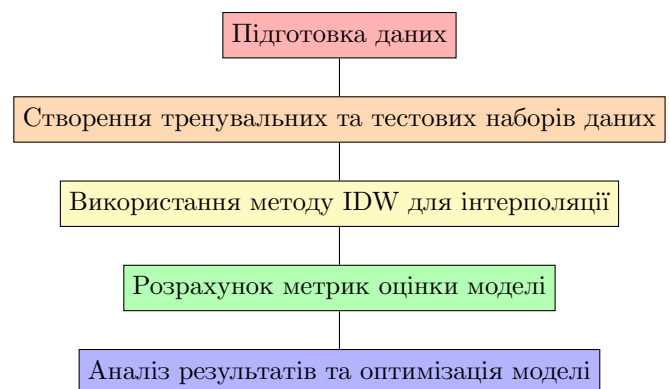


Рис. 4. Основні етапи роботи

Модель екстраполяції якості повітря реалізована через метод інверсного відстаневого зважування (`Inverse Distance Weighting, IDW`). `IDW` використовується для передбачення значень в невідомих точках, базуючись на відомих значеннях з навколи-

шніх місць. Цей метод заснований на припущенні, що вплив забруднюючого фактора зменшується зі збільшенням відстані від його джерела.

Основні компоненти моделі:

1. **Визначення відстаней.** Відстань між кожною цільовою точкою прогнозування та доступними тренувальними точками обчислюється за допомогою евклідової формули:

$$\text{distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2},$$

де x і y відповідають широті та довготі відповідно.

2. **Обчислення ваг.** Ваги для кожної тренувальної точки обчислюються за формулою:

$$\text{weight} = \frac{1}{\text{distance}^{\text{power}}},$$

де power – це параметр, що контролює швидкість зменшення впливу з відстанню. Для точок, що знаходяться за межами визначеного радіуса згладжування, ваги призначаються як 0.

3. **Нормалізація ваг і обчислення прогнозованих значень.** Ваги нормалізуються, а прогнозоване значення в кожній цільовій точці обчислюється за допомогою зваженої суми:

$$\text{interpolated value} = \frac{\sum(\text{weights} \times \text{train values})}{\sum(\text{weights})}$$

Реалізована модель включає в себе наступні дії:

1. Операції з даними. Перед використанням дані очищаються і масштабуються, викиди видаляються для забезпечення якості тренувального набору.
2. Стратифікована крос-валідація. Використання стратифікованої крос-валідації допомагає оцінити надійність моделі за різних умов, розділяючи дані на кілька тренувальних та тестових наборів таким чином, щоб вони представляли розподіл цільової змінної (value.bin).
3. Оцінка моделі. Після проведення крос-валідації обчислюються основні метрики якості моделі, такі як середньоквадратична помилка (MSE) та середня абсолютна помилка (MAE). Ці метрики дозволяють оцінити точність та надійність прогнозів моделі.

Висновки

Враховуючи отримані результати оцінки точності, наша модель інверсного відстаневого зважування (IDW) для екстраполяції рівнів забруднення PM10 продемонструвала достатньо точні результати. Висновок про ефективність моделі базується на вивченні показників середньоквадратичної помилки (MSE) та середньої абсолютної помилки (MAE), які отримані після крос-валідації. Також було отримано суцільну карту забруднень, яка демонструє більш наглядно результати.

MSE (0.08503895425404585) вказує на досить низьку середню помилку прогнозів моделі відносно діапазону значень датасету, що свідчить про добру здатність моделі передбачати значення в навколишніх точках на основі відомих даних.

MAE (0.24033852786960433) підтверджує, що модель має досить високу точність передбачень, оскільки середні абсолютні відхилення між фактичними та прогнозованими значеннями є невеликими.

На Рис. 5 зображено карту забруднень, вона показує графічне представлення результатів моделі. Візуалізація ілюструє просторове розподілення концентрацій PM10 за січень-лютий 2022 року, де червоні точки вказують на розташування моніторингових станцій. Контурні лінії і кольорова шкала представляють екстрапольовані рівні забруднення повітря, дозволяючи візуально оцінити області з вищою та нижчою концентрацією забруднювачів.

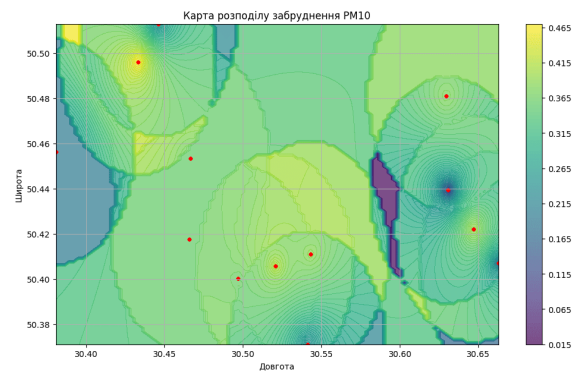


Рис. 5. Географічний розподіл тренувальних даних зі значеннями показників якості повітря

Перелік використаних джерел

1. A Survey of Wireless Sensor Network Based Air Pollution Monitoring Systems / W. Y. L. Yi, K. Ming, T. Mak, K. S. Leung, Y. Leung, M. L. Meng // Sensors. — 2015. — Т. 15, № 12. — С. 31392–31427. — ISSN 1424-8220. — URL: <https://www.mdpi.com/1424-8220/15/12/29859>.
2. An Interactive Web Mapping Visualization of Urban Air Quality Monitoring Data of China / W. Lu, T. Ai, X. Zhang, Y. He // Atmosphere. — 2017. — Т. 8, № 8. — ISSN 2073-4433. — URL: <https://www.mdpi.com/2073-4433/8/8/148>.
3. Graph Neural Processes for Spatio-Temporal Extrapolation / J. Hu, Y. Liang, Z. Fan, H. Chen, Y. Zheng, R. Zimmermann. — 2023. — DOI: 10.1145/3580305.3599372. — URL: <https://doi.org/10.1145/3580305.3599372>.
4. SaveEcoBot. — URL: <https://www.saveecobot.com/>.