КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023
DOI 10.28925/2663-4023.2026.32.1198
УДК 004.6

Шевченко Дмитро Віталійович
доктор філософії, асистент кафедри комп'ютерних наук,
Національний університет біоресурсів і природокористування України, Київ, Україна
ORCID: 0009-0001-7736-8263
dimashevchenko10021999@gmail.com

Голуб Белла Львівна
кандидат технічних наук, доцент, завідувач кафедри комп'ютерних наук,
Національний університет біоресурсів і природокористування України, Київ, Україна
ORCID: 0000-0002-1256-6138
bellalg@nubip.edu.ua

Бородкіна Ірина Лаврентіївна
кандидат технічних наук, доцент, доцент кафедри комп'ютерних наук,
Національний університет біоресурсів і природокористування України, Київ, Україна
ORCID: 0000-0003-3667-3728
i.borodkina@nubip.edu.ua

КЛАСТЕРИЗАЦІЯ СТАНЦІЙ ДЛЯ ВИЯВЛЕННЯ НЕСТАБІЛЬНОСТІ ДАНИХ У МЕРЕЖІ МОНІТОРИНГУ ЯКОСТІ АТМОСФЕРНОГО ПОВІТРЯ

Анотація. У роботі запропоновано та апробовано підхід до автоматизованого виявлення нестабільності даних у мережі моніторингу якості атмосферного повітря на основі методів інтелектуального аналізу даних. На відміну від традиційних перевірок за пороговими значеннями, підхід орієнтований на поведінкові ознаки «якості потоку вимірювань» (повнота, частка пропусків, варіативність сигналу та ознаки «залипання» сенсора), розраховані на погодинних агрегатах. Об'єктом кластеризації є пари «станція та сенсор», що дає змогу локалізувати проблеми як на рівні станції, так і на рівні окремих вимірювальних каналів. Для групування застосовано алгоритм K-Means із попереднім масштабуванням ознак; оптимальну кількість кластерів визначено за методами «ліктя» та коефіцієнтом силуету. Для інтерпретації кластерів використано проєкцію на двох головних компонентах, що відображають індекс доступності/неповноти даних та індекс динаміки сигналу (варіативність проти «залипання»). Експеримент на реальних даних продемонстрував наявність стійких профілів деградації вимірювань і дозволив сформувати перелік надійних станцій та проблемних каналів (зокрема, сенсорів із високою часткою пропусків або з близькою до нуля погодинною варіацією). Практична цінність полягає у можливості інтеграції методу в інформаційно-аналітичні системи екологічного моніторингу як модуль контролю якості даних із подальшим використанням результатів для відбору референтних сенсорів, калібрування та побудови прогнозних моделей.

Ключові слова: Data Mining; K-Means; екологічний моніторинг; моніторинг атмосферного повітря; інформаційно-аналітична система; інтелектуальна технологія; інформаційні технології; надійність та достовірність даних.

ВСТУП

Атмосферне забруднення залишається одним із найсуттєвіших екологічних ризиків для здоров'я населення та якості життя у містах. Рекомендації щодо граничних рівнів основних забруднювачів ($PM_{2.5}$, PM_{10}, O_3, NO_2, SO_2, CO) фіксують необхідність системного контролю та регулярного оновлення підходів до оцінювання якості повітря. У європейських країнах проблема підсилюється тим, що значна частка міського

населення зазнає впливу концентрацій дрібнодисперсних частинок вище за рівні, рекомендовані для охорони здоров'я [1].

Сучасні мережі моніторингу (включно з IoT/громадськими сенсорними мережами) підвищують просторову деталізацію спостережень, однак одночасно загострюють проблему надійності даних. Практика експлуатації низьковартісних сенсорів показує, що вимірювання можуть бути неповними через збій сенсорів, деградацію, мережеві відмови, енергетичні проблеми або інші операційні причини [2]. У таких мережах критично важливо відокремлювати реальні екологічні події (піки забруднення) від артефактів вимірювання (масові пропуски, «залипання», некоректна варіативність сигналу), оскільки недостовірні дані можуть спотворювати аналітику, інформування населення та управлінські рішення [3].

Отже, постає задача створення автоматизованого механізму, який на основі журналів вимірювань класифікує профілі якості потоку даних, виявляє нестабільні станції/сенсори та формує інтерпретовані групи (кластери) з близькою поведінкою. Такий механізм має бути масштабованим, працювати на агрегованих даних і бути придатним для інтеграції в інформаційно-аналітичні системи моніторингу.

Аналіз останніх досліджень і публікацій. У роботах і рекомендаціях щодо використання сенсорів якості повітря наголошено, що низьковартісні пристрої слід застосовувати із чітким планом дослідження, розумінням обмежень та обов'язковою процедурою інтерпретації/контролю якості, у т.ч. з урахуванням неповноти даних та особливостей роботи сенсора [4]. Окремо підкреслюється, що правдоподібність даних з великих мереж сенсорів є дискусійною без регулярної верифікації та калібрування, тому потрібні формальні підходи до оцінювання надійності потоків вимірювань [3].

У контексті IoT-моніторингу якості повітря дослідники систематизують чинники ненадійності даних і підкреслюють потребу в оцінюванні/покращенні даних для забезпечення їхньої довіри та корисності для осіб, що приймають рішення [5]. Одним із практичних напрямів є методи, які працюють із пропущеними значеннями та послідовними провалами даних у сенсорних мережах, оскільки неповнота суттєво впливає на калібрування і коректність подальших моделей [6].

Методологічно проблема нестабільності сенсорних потоків тісно пов'язана з поняттям багатовимірної якості даних. Класичні підходи виділяють такі базові виміри: точність, повнота, узгодженість та своєчасність. Стандартизація моделей якості даних, зокрема ISO/IEC 25012, надає концептуальну рамку для формулювання вимог до даних та планування оцінювання якості (у тому числі повноти та точності), що є релевантною для екологічних інформаційних систем [7].

Водночас для оперативного контролю якості у великих потоках даних практичним є застосування методів Data Mining (неконтрольоване навчання), які здатні автоматично виділяти типові режими роботи сенсорів без попереднього маркування. Кластеризація є одним із базових інструментів такого аналізу, оскільки дозволяє групувати об'єкти за структурною подібністю ознак і виявляти аномальні групи або «крайні» профілі поведінки.

Метою статті є розробити та перевірити підхід до кластеризації станцій моніторингу (через профілі пар «станція та сенсор») за ознаками повноти та варіативності даних для автоматичного виявлення нестабільних вимірювальних каналів і формування переліку надійних станцій/сенсорів для подальшої аналітики та підтримки прийняття рішень.

МЕТОДИКА ДОСЛІДЖЕННЯ

У межах дослідження реалізовано інтегрований конвеєр формування ознак, вибору параметрів кластеризації та інтерпретації результатів, орієнтований на подальшу інтеграцію в інформаційно-аналітичну систему екологічного моніторингу.

Вхідні дані та погодинна агрегація. Первинні сенсорні вимірювання надходять з фіксованою очікуваною частотою; у проведеному експерименті прийнято інтервал 10 хв, що відповідає 6 очікуваним вимірам на годину. Для кожної пари «станція та сенсор» дані агрегуються у погодинні вікна, для яких обчислюються: кількість вимірів у межах години та погодинне стандартне відхилення значень, що відображає активність сигналу та наявність внутрішньо-годинної варіативності.

Для коректного врахування пропусків будується повна часова сітка годин для кожної пари «станція та сенсор» на заданому інтервалі аналізу; відсутні години доповнюються нульовими значеннями $count_h$ та std_h. Погодинна заповненість визначається як відношення фактичної кількості вимірів до очікуваної з обмеженням зверху:

$$fill_h = min(1, \frac{count_h}{expected\_count}), \quad (1)$$

де:
$count_h$ – кількість вимірів у межах години,

$expected\_count$ – очікувана кількість вимірювань у межах однієї години.

$$expected\_count = \frac{60}{freq\_minutes}, \quad (2)$$

де:
$freq\_minutes$ – номінальний інтервал надходження вимірювань у хвилинах.

На основі погодинних характеристик у межах аналізованого періоду обчислюються агреговані ознаки, що відображають повноту потоку вимірювань та стабільність/варіативність сигналу.

Середня наповненість (*fill_mean*) розраховується як середнє значення *fill_h* за період і характеризує типовий рівень надходження даних. Десятий перцентиль заповненості (*fill_p10*) використовується як індикатор ризику короткочасних провалів: показник відображає заповненість у найгірших 10% годин і дозволяє ідентифікувати сенсори з періодичними збоями навіть за прийнятного середнього рівня.

Медіана погодинної варіативності (*std_median*) визначається як медіанне значення *std_h* за період та відображає типову внутрішньо-годинну змінність сигналу. Низькі значення *std_median* можуть бути ознакою слабкої динаміки показника або потенційної технічної аномалії. Частка годин без даних (*offline_ratio*) обчислюється як відношення кількості годин із *fill_h* дорівнює 0 до загальної кількості годин у періоді й характеризує частоту повних відмов передачі (збоїв зв'язку, живлення або роботи сенсора). Частка годин із даними, але без варіативності (*stuck_ratio*) визначається як частка годин, у яких дані наявні (*fill_h більше 0*), але внутрішньо-годинна варіативність відсутня (*std_h*

дорівнює *0*), відносно годин із наявними даними. Для коректності оцінювання stuck_ratio обчислювався лише для годин із достатньою кількістю вимірів (наприклад, *count_h* ≥ *3*), щоб уникнути трактування одиничних вимірів як відсутності варіативності.

Наведені метрики є інтерпретованими та узгоджуються з підходами до оцінювання якості даних у термінах повноти (completeness) і стабільності потоку спостережень [8]. Оскільки сформовані ознаки мають різні числові масштаби (зокрема, *std_median* та відносні частки), перед застосуванням алгоритму *K-Means* виконується стандартизація ознак. Під стандартизацією розуміємо приведення кожної ознаки до нульового середнього значення та одиничного стандартного відхилення. Така нормалізація усуває домінування ознак із більшим діапазоном значень у методах і забезпечує коректне порівняння об'єктів у просторі ознак. Реалізаційно використано підхід StandardScaler, який центрує дані та масштабує дисперсію до одиниці [9].

Масштабування ознак. Оскільки сформовані ознаки мають різні числові масштаби (зокрема *std_median* та відносні частки), перед застосуванням алгоритму *K-Means* виконано стандартизацію ознак. Стандартизація полягає у приведенні кожної ознаки до нульового середнього значення та одиничного стандартного відхилення (нормування), що є критично важливим для таких методів, оскільки запобігає домінуванню ознак із більшим діапазоном значень у метриці відстані. Реалізаційно використано *StandardScaler (scikit-learn)*, який центрує дані та масштабує дисперсію до одиниці.

Кластеризація методом K-Means. Для кластеризації даних застосовано метод K-Means, який мінімізує суму квадратів відстаней об'єктів до відповідних центроїдів (WCSS / inertia). Він є одним із найпоширеніших алгоритмів неконтрольованого навчання (*unsupervised learning*) і широко використовується для розв'язання задач кластеризації. Реалізація базується на бібліотеці scikit-learn [10].

Завдання K-Means полягає в пошуку такої конфігурації кластерів, що мінімізує сумарну відстань усіх точок до центрів своїх груп. Алгоритм послідовно оновлює координати центрів та склад кластерів до досягнення стабільності.

У математичному формулюванні мета алгоритму полягає у мінімізації функції:

$$J = \sum_{k=1}^{k} \sum_{x_i \epsilon C_k} ||x_i - \mu_k||, \quad (3)$$

де:
K - кількість кластерів,
C_k - множина точок, що належать кластеру k,
μ_k - центр (центроїд) кластера k,
x_i - вектор ознак (набір виміряних екологічних параметрів).

Вибір кількості кластерів є критичним етапом застосування алгоритму K-Means, оскільки значення параметра k безпосередньо впливає на якість сегментації та інтерпретованість результатів. У разі надто малого k різнорідні об'єкти штучно об'єднуються в укрупнені групи, що приховує аномальні профілі сенсорів. Натомість надто велике k призводить до "перерозбиття" вибірки, утворення дрібних кластерів і зниження стійкості кластеризації. Тому до виконання фінальної кластеризації доцільно обґрунтувати k за допомогою внутрішніх критеріїв якості, які не потребують наявності еталонних міток класів.

Для первинної оцінки застосовується показник інерції (within-cluster sum of squares), який характеризує компактність кластерів і визначається як сумарна внутрішньо кластерна дисперсія. Це спостереження лежить в основі методу «ліктя»: на графіку залежності інерції від k обирають значення, після якого додаткові кластери майже не зменшують внутрішньо кластерний розкид. Використання методу «ліктя» дозволяє отримати збалансоване рішення між точністю опису даних і складністю моделі та забезпечує наочне, інтерпретоване обґрунтування вибору k (рис. 1).

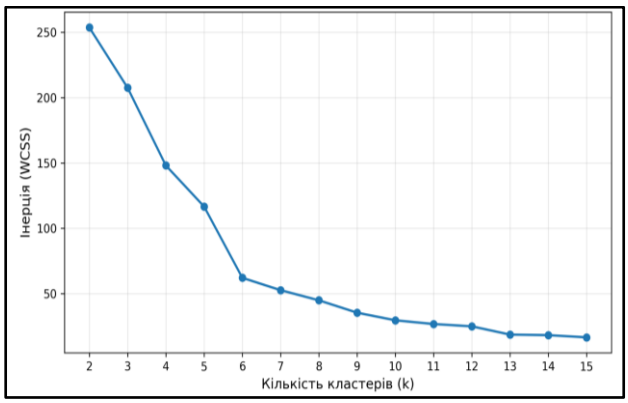

Рис. 1. Метод ліктя (Elbow method)

Для порівняльної оцінки якості розбиття додатково використовується silhouette score (силуетний коефіцієнт), який одночасно враховує дві властивості кластеризації: згуртованість об’єктів у межах свого кластера та відокремленість від сусідніх кластерів. На відміну від інерції, силуетний коефіцієнт не є монотонним за k і дозволяє визначити значення k, при якому кластери найбільш добре розділені. Практично це зручно для вибору k у ситуаціях, коли “лікоть” на графіку інерції виражений слабко або допускає декілька інтерпретацій. Комбіноване використання інерції (для контролю компактності) та silhouette score (для контролю роздільності) підвищує обґрунтованість вибору параметра k перед запуском фінальної моделі K-Means (рис. 2).

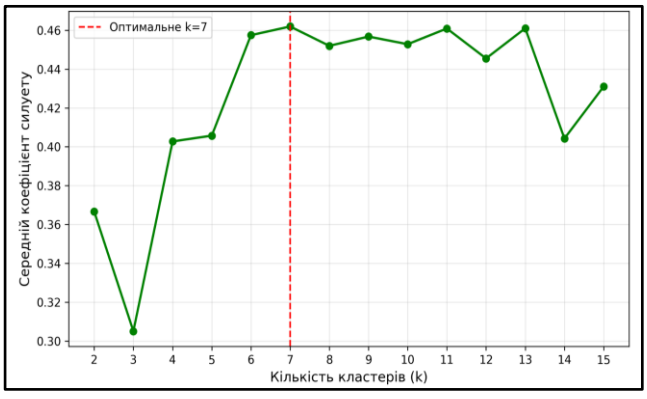

Рис. 2. Silhouette score (силуетний коефіцієнт)

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

За результатами оцінювання кількості кластерів встановлено, що на графіку інерції (рис. 1) спостерігається інтенсивне зменшення внутрішньо кластерної дисперсії при збільшенні k до приблизно 6, після чого темп зниження суттєво сповільнюється. Додатково, за критерієм роздільності кластерів максимальне значення середнього силуетного коефіцієнта отримано при k = 7 і становить 0,462. З огляду на узгодженість результатів обох критеріїв та з метою забезпечення достатньої деталізації профілів якості даних, для подальшого аналізу прийнято k = 7 кластерів.

Результати кластерного аналізу узагальнено подано на рис. 3 у вигляді двовимірної проєкції об'єктів кластеризації (пари «станція та сенсор») на площину головних компонентів. Візуалізація демонструє формування кількох відносно відокремлених груп, що підтверджує наявність типових профілів якості потоку даних у мережі моніторингу.

Рис. 3 відображає структуру даних у зменшеному просторі ознак і слугує засобом пояснюваності кластеризації. Перша вісь X (52,9% дисперсії) переважно характеризує відмінності у доступності даних, зокрема градієнт «повнота проти пропусків». Друга вісь Y (22,2% дисперсії) додатково відображає відмінності, пов'язані з внутрішньо годинною варіативністю сигналу та ознаками «залипання». Таким чином, візуальна картина узгоджується з інтерпретацією кластерів як профілів надійності, нестабільності доступності та потенційної деградації сенсорного сигналу.

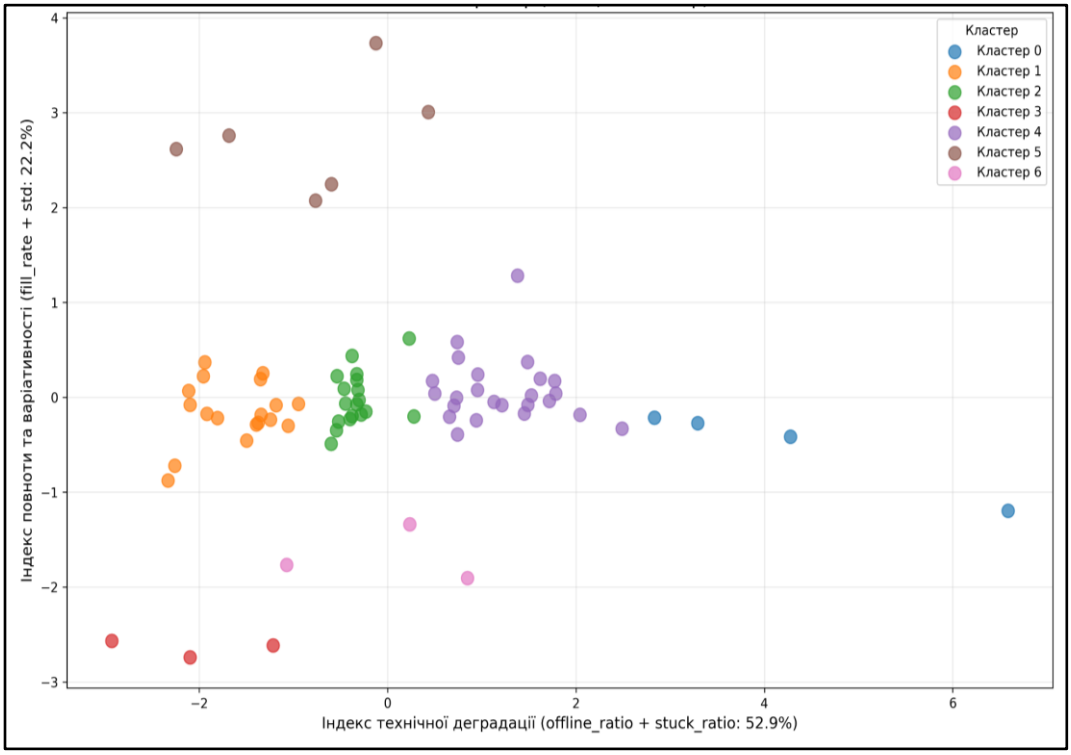

Рис. 3 – Розподіл кластерів K-means

Для змістовної інтерпретації кластеризації сформовано профілі кластерів на основі медіанних значень ключових метрик (табл. 1). Використання медіанних оцінок знижує вплив поодиноких викидів і краще відображає типову поведінку потоку на заданому часовому інтервалі. Додатково наведено кількість пар у кожному кластері.

Таблиця 1

Профілі кластерів за медіанними значеннями ознак

Кластер	К-ть пар	fill_mean	fill_p10	std_median	offline_ratio	stuck_ratio
0	4	0,739	0,000	0,002	0,252	0,013
1	18	0,914	0,667	0,110	0,071	0,016
2	18	0,895	0,333	0,314	0,095	0,016
3	3	0,911	0,667	0,586	0,078	0,087
4	24	0,852	0,000	0,109	0,138	0,011
5	6	0,895	0,333	9,401	0,092	0,013
6	3	0,871	0,000	0,577	0,119	0,064

Отримані профілі дозволяють виділити групи з різними характеристиками якості даних. Кластер 1 відповідає надійним каналам із високими значеннями заповненості та низькою часткою годин без даних, що робить його придатним як референтний профіль. Кластер 2 характеризується загалом доброю повнотою, проте нижчий *fill_p10* і вищий *offline_ratio* вказують на періодичні короткочасні провали.

Кластер 4 відображає нестабільну доступність, для якої характерне *fill_p10* = 0, тобто наявність годин з повною відсутністю вимірів при відносно прийнятному середньому рівні. Кластер 0 демонструє критичний профіль з високою часткою пропусків і практично нульовою варіативністю, що узгоджується зі сценаріями деградації каналу або некоректного сигналу. Кластери 3 і 6 мають підвищені значення *stuck_ratio* за прийнятної повноти, що може відповідати «залипанню» сенсора або низькій чутливості окремих параметрів. Кластер 5 характеризується дуже високою варіативністю (*std_median*) і потребує предметної перевірки, оскільки може відповідати як реальній високій динаміці показника, так і шумності сигналу.

Оскільки кластеризація виконувалася на рівні пар «станція та сенсор», для практичного використання результати агрегуються на рівень станції. Узагальнення здійснювалося через частку каналів, що належать до кластерів з вираженою нестабільністю доступності або критичними ознаками деградації (у даному експерименті – кластери 0 і 4 відповідно до табл. 1). Такий підхід дозволяє сформувати ранжування станцій за придатністю даних до подальшої аналітики.

Найнадійнішими за обраними показниками є станції 20 та 22, які демонструють високу середню наповненість, низьку частку годин без даних і мінімальну частку проблемних каналів. Найбільш проблемними є станції 17 та 18, для яких переважають профілі з нестабільною доступністю, а також наявні канали з критичними ознаками деградації. Водночас результати показують, що проблеми можуть бути локалізовані на рівні окремих сенсорних каналів, а не всієї станції, що є важливим для практичної експлуатації: система може формувати рекомендації щодо обслуговування конкретних сенсорів без повного виключення станції з моніторингу.

Таблиця 2

Приклад агрегованого ранжування станцій за показниками повноти та частки проблемних профілів

Станція №	К-ть каналів	avg(fill_mean)	avg(offline_ratio)	bad_share
22	11	0,917	0,071	0,091
20	11	0,904	0,081	0,091
21	11	0,874	0,115	0,182
16	10	0,863	0,122	0,200
19	11	0,882	0,107	0,273
17	11	0,859	0,132	0,818
18	11	0,829	0,163	0,909

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У статті запропоновано підхід до виявлення нестабільності даних у мережі моніторингу якості атмосферного повітря на основі кластеризації даних «станція та сенсор» за ознаками повноти та варіативності потоку вимірювань. Сформовано набір із п’яти ознак якості потоку (*fill_mean, fill_p10, std_median, offline_ratio, stuck_ratio*), що відображають повноту даних, частоту пропусків та ознаки можливої деградації сигналу. Оптимальну кількість кластерів обґрунтовано за метриками інерції (метод «ліктя») та силуетного коефіцієнта. Для експериментальних даних прийнято k = 7, середній silhouette score становить близько 0,462.

У результаті отримано типові профілі якості даних, які дозволяють відокремлювати надійні та нестабільні канали, а також ідентифікувати критичні випадки з високою часткою пропусків і низькою варіативністю (зокрема ознаки можливого «залипання» сенсора). Агрегація кластерних міток на рівень станцій дала змогу сформувати рейтинг надійності та показала, що значна частина проблем має локальний характер і стосується окремих сенсорних каналів. Запропонований підхід придатний для інтеграції в інформаційно-аналітичні системи як модуль контролю якості даних і може використовуватися як підготовчий етап перед калібруванням сенсорів, розрахунком індексів якості повітря та побудовою прогнозних моделей.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. European Environment Agency. (2022). *Air quality in Europe 2022*. https://doi.org/10.2800/488115
2. Agbo, B., Al-Aqrabi, H., Hill, R., & Alsboui, T. (2022). Missing data imputation in the Internet of Things sensor networks. *Future Internet*, 14(5), Article 143. https://doi.org/10.3390/fi14050143
3. Jiao, W., Hagler, G., Williams, R., Sharpe, R., Brown, R., Garver, D., Judge, R., Caudill, M., Rickard, J., Davis, M., Weinstock, L., Zimmer-Dauphinee, S., & Buckley, K. (2016). Community

Air Sensor Network (CAIRSENSE) project: Evaluation of low-cost sensor performance in a suburban environment in the southeastern United States. *Atmospheric Measurement Techniques*, 9(11), 5281–5292. https://doi.org/10.5194/amt-9-5281-2016

4. U.S. Environmental Protection Agency. (2025, May 1). *How to use air sensors: Air sensor guidebook*. https://www.epa.gov/air-sensor-toolbox/how-use-air-sensors-air-sensor-guidebook
5. Buelvas, J., Múnera, D., Tobón V., D. P., Aguirre, J., & Gaviria, N. (2023). Data quality in IoT-based air quality monitoring systems: A systematic mapping study. *Water, Air, & Soil Pollution*, 234(4), Article 248. https://doi.org/10.1007/s11270-023-06127-9
6. Chen, M., Zhu, H., Chen, Y., & Wang, Y. (2022). A novel missing data imputation approach for time series air quality data based on logistic regression. *Atmosphere*, 13(7), 1044. https://doi.org/10.3390/atmos13071044
7. International Organization for Standardization. (2008). *ISO/IEC 25012:2008. Software engineering—Software product quality requirements and evaluation (SQuaRE)—Data quality model*. https://www.iso.org/standard/35736.html
8. Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86–95. https://doi.org/10.1145/240455.240479
9. Scikit-learn developers. (n.d.). *StandardScaler*. In *Scikit-learn*. Retrieved March 2026, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler
10. Scikit-learn developers. (n.d.). *KMeans*. In *Scikit-learn*. Retrieved March 2026, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans
11. Shevchenko, D. V., & Holub, B. L. (2025). Air quality monitoring in real time. *Mathematical Machines and Systems*, (1), 103–112.
12. Shevchenko, D. V., & Holub, B. L. (2025). Application of data mining methods for multidimensional analysis of atmospheric air quality based on environmental data. *Science and Technology Today. Series: Engineering*, 8(49), 1801–1810.
13. Shevchenko, D. V., & Holub, B. L. (2025). Multidimensional analytics of environmental data: Application of OLAP in monitoring systems. *Mathematical Machines and Systems*, (3–4), 54–65.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 4 (32), 2026
ISSN 2663 – 4023

Dmytro Shevchenko
PhD, Assistant, Department of Computer Science,
National University of Life and Environmental Sciences of Ukraine
ORCID: 0009-0001-7736-8263
dimashevchenko10021999@gmail.com

Bella Holub
Candidate of Engineering Sciences, Associate Professor,
Head of the Department of Computer Science,
National University of Life and Environmental Sciences of Ukraine
ORCID: 0000-0002-1256-6138
bellalg@nubip.edu.ua

Iryna Borodkina
Candidate of Engineering Sciences, Associate Professor, Department of Computer Science,
National University of Life and Environmental Sciences of Ukraine
ORCID: 0000-0003-3667-3728
i.borodkina@nubip.edu.ua

STATION CLUSTERING FOR DETECTING DATA INSTABILITY IN AN AIR QUALITY MONITORING NETWORK

Abstract. This paper proposes and validates an approach for automated detection of data instability in an atmospheric air quality monitoring network using data mining methods. Unlike traditional threshold-based checks, the approach focuses on behavioral features describing the "measurement stream quality" (completeness, missingness rate, signal variability, and sensor "sticking" indicators) computed from hourly aggregates. The clustering objects are "station–sensor" pairs, which enables localization of issues both at the station level and at the level of individual measurement channels. K-Means clustering is applied with prior feature scaling; the optimal number of clusters is selected using the elbow method and the silhouette coefficient. For cluster interpretation, a projection onto two principal components is used, reflecting a data availability/incompleteness index and a signal dynamics index (variability versus "sticking"). Experiments on real-world data reveal stable degradation profiles of measurements and allow identification of reliable stations and problematic channels (in particular, sensors with a high missingness rate or near-zero hourly variation). The practical value of the study lies in the ability to integrate the proposed method into environmental information-and-analytical systems as a data quality control module, and to further use the results for selecting reference sensors, calibration, and building predictive models.

Keywords: Data Mining; K-Means; environmental monitoring; atmospheric air monitoring; information and analytical system; intelligent technology; information technologies; data reliability and validity.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. European Environment Agency. (2022). *Air quality in Europe 2022*. https://doi.org/10.2800/488115
2. Agbo, B., Al-Aqrabi, H., Hill, R., & Alsboui, T. (2022). Missing data imputation in the Internet of Things sensor networks. *Future Internet*, 14(5), Article 143. https://doi.org/10.3390/fi14050143
3. Jiao, W., Hagler, G., Williams, R., Sharpe, R., Brown, R., Garver, D., Judge, R., Caudill, M., Rickard, J., Davis, M., Weinstock, L., Zimmer-Dauphinee, S., & Buckley, K. (2016). Community Air Sensor Network (CAIRSENSE) project: Evaluation of low-cost sensor performance in a suburban environment in the southeastern United States. *Atmospheric Measurement Techniques*, 9(11), 5281–5292. https://doi.org/10.5194/amt-9-5281-2016

4. U.S. Environmental Protection Agency. (2025, May 1). *How to use air sensors: Air sensor guidebook*. https://www.epa.gov/air-sensor-toolbox/how-use-air-sensors-air-sensor-guidebook
5. Buelvas, J., Múnera, D., Tobón V., D. P., Aguirre, J., & Gaviria, N. (2023). Data quality in IoT-based air quality monitoring systems: A systematic mapping study. *Water, Air, & Soil Pollution*, 234(4), Article 248. https://doi.org/10.1007/s11270-023-06127-9
6. Chen, M., Zhu, H., Chen, Y., & Wang, Y. (2022). A novel missing data imputation approach for time series air quality data based on logistic regression. *Atmosphere*, 13(7), 1044. https://doi.org/10.3390/atmos13071044
7. International Organization for Standardization. (2008). *ISO/IEC 25012:2008. Software engineering—Software product quality requirements and evaluation (SQuaRE)—Data quality model*. https://www.iso.org/standard/35736.html
8. Wand, Y., & Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), 86–95. https://doi.org/10.1145/240455.240479
9. Scikit-learn developers. (n.d.). *StandardScaler*. In *Scikit-learn*. Retrieved March 2026, from https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler
10. Scikit-learn developers. (n.d.). *KMeans*. In *Scikit-learn*. Retrieved March 2026, from https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans
11. Shevchenko, D. V., & Holub, B. L. (2025). Air quality monitoring in real time. *Mathematical Machines and Systems*, (1), 103–112.
12. Shevchenko, D. V., & Holub, B. L. (2025). Application of data mining methods for multidimensional analysis of atmospheric air quality based on environmental data. *Science and Technology Today. Series: Engineering*, 8(49), 1801–1810.
13. Shevchenko, D. V., & Holub, B. L. (2025). Multidimensional analytics of environmental data: Application of OLAP in monitoring systems. *Mathematical Machines and Systems*, (3–4), 54–65

Отримано редакцією журналу / Received: 28.01.26
Прорецензовано / Revised: 17.02.26
Схвалено до друку / Accepted: 26.03.26