https://doi.org/10.31649/1997-9266-2024-172-1-88-100

УДК 004.9:504.3.054

Д. О. Шмундяк[1]
В. Є. Копняк[1]

МЕТОД ІДЕНТИФІКАЦІЇ ЛОКАЛЬНИХ АНОМАЛІЙ ЗНАЧЕНЬ ПОКАЗНИКІВ СТАНУ ДОВКІЛЛЯ З ВИКОРИСТАННЯМ ДЕКОМПОЗИЦІЇ НА ПІВХВИЛІ

[1]Вінницький національний технічний університет

В епоху масової цифровізації всіх існуючих частин діяльності людства, кількість даних невпинно зростає і важливо мати навички з ними працювати для розв'язання різного роду задач. Однією з найпоширеніших структур збереження цих даних є часові ряди — послідовності точок, зазвичай, за певний хронологічний період. До цієї категорії відносяться фінансові показники, дані екологічного моніторингу, медичні показники тощо. Широкий перелік сфер застосування робить задачу аналізу часових рядів актуальною і важливою. Якість зробленого прогнозу часового ряду багато в чому залежить від якості проведеного аналізу, який може включати обробку та стандартизацію самих даних, виділення вагомих показників, пошук взаємозв'язків тощо. Серед цих кроків особливо вагоме місце посідає пошук аномалій. Аномалії — це точки набору даних, які певним чином відрізняються від інших значень або певних шаблонів поведінки. Наявність подібних записів сильно впливає на можливість моделей машинного навчання зробити точний прогноз, тому необхідно мати можливість ідентифікувати ці аномалії.

Розроблено новий метод ідентифікації локальних аномалій значень показників стану довкілля з використанням декомпозиції на півхвилі. Основна ідея методу полягає у декомпозиції часового ряду на півхвилі, використовуючи точки тенденції, де падіння змінюється на зростання, чи навпаки, та у розбитті ряду на фрагменти. Кожен окремий фрагмент аналізується окремо і на ньому виконується пошук аномалій комбінування багатьох методів. Точність роботи цих методів перевіряється за рахунок експертного методу. Описано основні кроки запропонованого методу, наведено приклад його роботи на реальних даних моніторингу якості атмосферного повітря, отриманих з однієї зі станцій мережі громадського моніторингу EcoCity у межах міжнародної програми «Чисте повітря для України».

На базі платформи Kaggle, розроблено та протестовано запропонований метод. Результат пошуку аномалій застосовано для побудови моделі Facebook Prophet, порівняно точність апроксимації з результатами роботи моделі Prophet з параметрами за замовчуванням. Випробування показали зменшення помилки апроксимації часового ряду на 11 % за метрикою RMSE та на 8 % за метрикою MAE. Це дозволило підтвердити ефективність розроблено методу.

Ключові слова: аналіз часових рядів, моделювання, машинне навчання, аномалії часових рядів, якість атмосферного повітря, декомпозиція часового ряду, EcoCity.

Вступ

Наявність аномальних даних у часових рядах часто може ставати проблемою, коли необхідно спрогнозувати значення цього ряду. Це дослідницький напрямок, яким займається багато науковців по всьому світу [1]—[3]. В науковій літературі описується великий перелік методів, які дозволяють ідентифікувати аномалії в даних різного характеру та структури, проте, не завжди можливим є досягнення високої точності. Однією з причин є наявність локальних аномалій, які багато методів не можуть вловити.

Попередній досвід авторів включає в себе дослідження, направлені на ідентифікацію аномалій та прогнозуванні реальних даних якості атмосферного повітря та стану водних ресурсів [4]. Показники стану довкілля зазвичай мають певну циклічність, зумовлену впливом сонячної активності (фотосинтез), забрудненням (у воді — скидання стічних вод, у повітрі — трафіком автотранспорту вдень, особливо у години пік) тощо. Ця циклічність може бути різною у різний час, але спільною є форма сигналу: мінімум, наростання, максимум, спадання, мінімум, що відповідає півхвилі. Зва-

жаючи на це, доцільним вважається пошук цих півхвиль і тим самим розбиття часового ряду на окремі частини. Кожну окрему частину ряду можна аналізувати окремо від інших та застосовувати методи пошуку аномалій на цьому відрізку, щоб ідентифікувати аномалії властиві саме для неї.

Мета дослідження — підвищення точності прогнозування часового ряду показника стану довкілля (якості вод, стану атмосферного повітря тощо) за рахунок виявлення та врахування локальних аномалій.

Формалізація задачі і традиційні підходи до її розв'язання

Як зазначалось раніше, наукова література описує досить великий набір різного роду методів пошуку аномалій. Це можуть бути як і методи, що базуються на відносно простих статистичних метриках, так і складніші моделі та методи, що використовують кластеризацію, дерева рішень, сезонні показники тощо [3], [5].

Розглянемо декілька таких методів, опишемо основні принципи їхньої роботи та перерахуємо переваги та недоліки їхнього застосування. Одним з таких методів є фільтр Хампеля. Цей метод використовує середнє абсолютне відхилення (MAD) та рухливе вікно для пошуку аномалій, через що він менш чутливий у порівнянні з методами, які базуються на середньому значенні та стандартному відхиленні [6]. Спостереження вважатиметься аномалією у тому випадку, коли воно перевищує MAD у n разів (значення n як і розмір вікна можна змінювати за рахунок налаштування методу). Для використання цього методу у Python можна використати бібліотеку Hampel.

Основними перевагами цього методу є:

– можливість застосування для різних типів даних;

– можливість налаштування параметрів.

Серед недоліків можна виділити погану ефективність для великих об'сягів аномалій та за наявності високого рівня шуму.

Ще одним підходом є використання рухомого середнього (Moving Average). Цей метод базується на порівнянні значень ряду з середніми значенням за певний часовий період. Метод дозволяє за рахунок попередніх даних передбачувати наступні. Тобто основна ідея під час пошуку аномалії — це спробувати передбачити подальші дані та перевірити наскільки передбачене значення відхиляється від реального. Якщо відхилення суттєве — більше визначеного порогового значення, то це значення вважається аномальним. У Python цей метод можна використати за рахунок бібліотеки pandas та безпосередньо методу DataFrame.rolling. Перевагами є його ефективність знаходження змін в тренді та сезонності. Водночас серед його недоліків можна виділити:

– можлива наявність затримки у виявленні аномалій, оскільки дані аналізуються в межах вікна значень;

– менша ефективність на великих обсягах даних.

Іншим популярним методом є Isolation Forest. Isolation Forest — це алгоритм виявлення аномалій, який працює шляхом ізоляції аномалій у даних [7]. Покроково цей метод можна описати таким чином:

1. З вхідного набору даних випадковим чином вибирається підвибірка;

2. Вибирається випадкова ознака та поріг — випадкове значення між мінімальними та максимальними значеннями вибраної ознаки;

3. Якщо точка даних менше порогу, то вона назначається в ліву гілку, якщо ж більше — в праву;

4. Крок 3 повторюється допоки усі точки не будуть розділені в окремих гілках дерева або поки не досягнуто максимальної глибини дерева (значення, яке можна налаштувати);

5. Після створення дерев, для кожного елемента визначається середня довжина шляху до гілок дерева;

6. Визначаються аномалії за рахунок перевірки значень середньої довжини шляху з певним порогом.

Серед переваг методу виділяють:

– простота використання алгоритму без складних конфігурацій;

– швидкодія;

– стійкість до викидів та шуму в даних.

Недоліками ж методу є вразливість за великої кількості аномалій та неефективність для багатовимірних даних.

Ще одним методом, який часто застосовується є метод k-найближчих сусідів [8]. Суть методу полягає в знаходженні k-distance — відстані від точки до k-найближчого сусіда. Для визначення

відстані часто застосовують евклідову відстань, хоча в певних випадках застосовують інші метрики, наприклад, відстань Геммінга. Перевірка на аномальність точки виконується за рахунок порівняння середнього значення відстаней з певним пороговим значенням. Порогове значення та показник *k* необхідно вказувати під час налаштування методу. Переваги цього методу:

– можливість працювати з різними типами даних;

– можливість працювати у разі неоднорідного розподілу даних.

Головним же недоліком напевно варто зазначити залежність якості роботи від налаштувань, а саме параметра *k*.

Широкого застосування для пошуку аномалій мають і методи, які базуються на концепції щільності. Серед таких методів можна виділити коефіцієнт локального відхилення (Local outlier factor) — алгоритм, який дозволяє знайти локальну густину точок в наборі [8], [9]. Суть методу полягає в такому:

1. Для кожної точки набору знаходиться k-distance — відстань до *k*-найближчого сусіда (параметр *k* задається у разі створення моделі);

2. Отримане значення використовується для знаходження відстані досяжності за формулою

$$RD_k(A,B) = \max\left\{\text{k-distance}(B), d(A,B)\right\}, \tag{1}$$

де $d(A,B)$ — відстань від об'єкта *A* до об'єкта *B*, а k-distance — відстань об'єкта *B* до його *k*-найближчого сусіда;

3. Обчислюється локальна щільність досяжності об'єкта, яка є оберненою до середньої відстані досяжності відповідного об'єкта до його сусідів.

$$lrd_k(A) = 1\Big/\left(\frac{\sum_{B\in N_k(A)} RD_k(A,B)}{\left|N_k(A)\right|}\right), \tag{2}$$

де $N_k(A)$ — набір *k*-найближчих сусідів об'єкта *A*, а $RD_k(A,B)$ — відстань досяжності від об'єкта A до об'єкта *B*;

4. Визначається коефіцієнт локального відхилення (LOF) шляхом відношення середнього значення lrd від *k*-кількості сусідів точки до lrd цієї точки за допомогою виразу

$$LOF_k(A) = \frac{\sum_{B\in N_k(A)} \frac{lrd_k(B)}{lrd_k(A)}}{\left|N_k(A)\right|} = \frac{\sum_{B\in N_k(A)} lrd_k(B)}{\left|N_k(A)\right| \cdot lrd_k(A)}, \tag{3}$$

де $N_k(A)$ — набір *k*-найближчих сусідів об'єкта A, $lrd_k(B)$ — локальна щільність досяжності об'єкта B, а $lrd_k(A)$ — локальна щільність досяжності об'єкта A

5. Отримане значення, яке дорівнює приблизно 1, вказує на те, що об'єкт можна порівняти зі своїми сусідами, Значення нижче 1 вказує на щільнішу область (що вказує на нормальну точку), тоді як значення, значно більші ніж 1, вказують на аномальні точки.

Цей метод дозволяє знаходити аномалії з урахуванням локальних властивостей набору даних і при цьому може працювати з різними формами розподілу даних. Загалом він дає кращі результати ніж глобальний підхід до пошуку, проте, немає загального порогового значення LOF , тому вибір допустимого відхилення не є найпростішою задачею.

Досить цікавим методом є Seasonal Hybrid ESD, який дозволяє знаходити аномалії в часових рядах з сезонністю [10]. Тест Грабса, також відомий як екстремальний стюдентизований тест відхилень (ESD), є тестом, який використовується для виявлення аномалій в одновимірному наборі даних, підпорядкованому нормальному розподілу. Тест Грабса можна записати так:

$$G = \frac{\max\limits_{i=1,\ldots,N} \left|Y_i - \overline{Y}\right|}{s}, \tag{4}$$

де $\overline{Y}$ — це середнє значення вибірки, а s — стандартне відхилення.

Проблема тесту Грабса в тому, що він передбачає нормальний розподіл, проте, дані далеко не завжди розподілені за таким законом. Щоб вирішити цю проблему, запропоновано метод Seasonal ESD (S-ESD), який використовує сезонну декомпозицію для видалення показників сезонності та тренду та перевірки залишкового компонента тестом Грабса. Проте, це все ще залишає проблему

— ектремальні неправдиві аномалії можуть вплинути на залишковий компонент. Так з'явився Seasonal Hybrid ESD — метод, який пропонує використовувати медіану та середнє абсолютне відхилення замість середнього та стандартного відхилення, оскільки вони дуже чутливі до великих та чисельних аномалій.

Модель дозволяє виявляти як глобальні аномалії, що виходять за межі очікуваних сезонних мінімумів і максимумів, так і локальні аномалії, які інакше були б замасковані сезонністю. Це досягається за рахунок використання декомпозиції часових рядів та використання надійних статистичних показників. Перевагами моделі є:

– врахування сезонних змін в часових рядах;

– можливість гнучкого налаштування.

Серед недоліків можна виділити:

– обмежена робота з незвичайними аномаліями;

– метод є досить повільним.

Також можна згадати методи опорних векторів (SVM) — набір методів машинного навчання, які зазвичай застосовують для класифікації, регресії та пошуку аномалій. Звичайний алгоритм опорних векторів (SVM) в більшості використовується для задач класифікації, де метою є знаходження гіперплощини, яка найкраще розділяє дані на два класи. Проте у разі виявлення аномалії часто маємо лише один клас даних, і цікавим є пошук межі, яка охоплює більшість, виключаючи при цьому аномалії. Саме тому для пошуку аномалій використовують One Class SVM. Його основна ідея полягає у відображенні вхідних даних у просторі функцій більшої розмірності за допомогою функції ядра, подібної до традиційних SVM. Потім він намагається знайти гіперплощину, яка охоплює якомога більше нормальних точок даних, зберігаючи максимальний запас між цією гіперплощиною та точками даних.

Серед його переваг є стійкість до перенавчання та гарна якість роботи для високорозмірних даних. Проте, сам метод досить важко налаштовувати, він вимагає великих обчислювальних ресурсів.

Розв'язання задачі

Для розв'язання пропонується метод, основна ідея якого — це гіпотеза про те, що часовий ряд складається з окремих фрагментів між локальними мінімумами тренду — точками, де графік міняє свій напрям зі спадання на зростання, і на кожному з цих фрагментів ряд поводиться по-різному [11]—[13]. Відповідно замість застосування певних методів пошуку аномалій до усього ряду, пропонується аналізувати кожен з фрагментів окремо, а потім об'єднувати усі отримані аномалії в один набір. Як сказано раніше, показники стану довкілля мають циклічність, зумовлену впливом природних факторів. Як приклад подібного ряду наведемо графік значень показнику пилу PM10 за 2020—2023 рр у Вінницькій області. Графік показано на рис. 1.

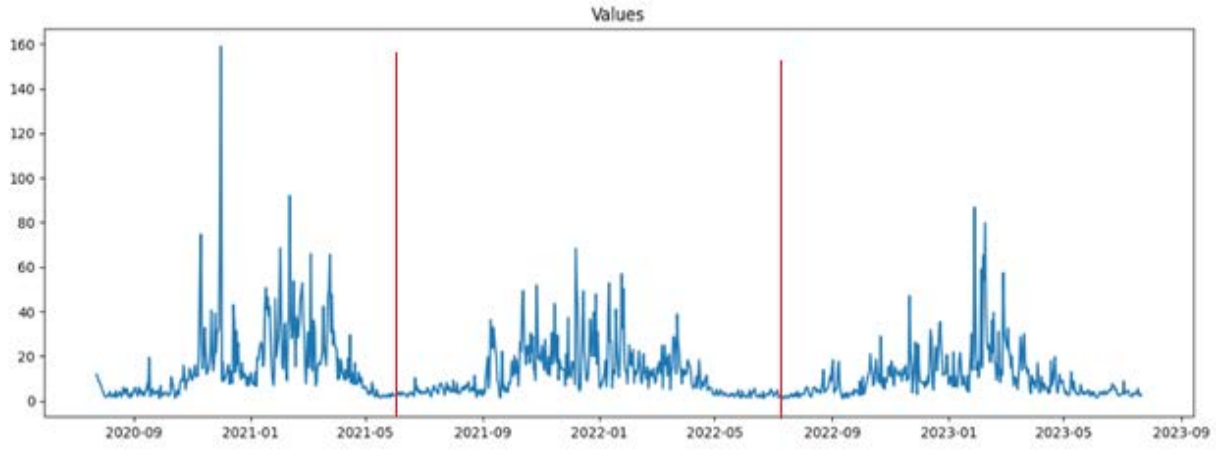


Рис. 1. Часовий ряд значень показника PM10 з приблизним поділом на окремі фрагменти

Як схематично зображено на рис. 1, часовий ряд можна спробувати розділити на декілька фрагментів, які мають схожу поведінку, зумовлену певними зазначеними природними факторами. Попередній досвід авторів показує, що розгляд усього ряду як одного цілого не завжди дає гарний результат, і доцільним є пошук локальних аномалії в кожному окремо декомпозованому фрагменті ряду, а потім об'єднання результатів по кожному фрагменту в одну вибірку [14].

Опишемо етапи роботи запропонованого методу.

Eman 1. Виділення тренду часового ряду.

Для виділення тренду скористаємося відомим прийомом — методом «rolling». Цей метод дозволяє згладити коливання даних та виокремити тренд часового ряду. Щоб обмежитися меншою кількість сегментів та тим самим спробувати захопити більше подібним точок часового ряду, отриману криву тренду додатково згладжуємо певним алгоритмом згладжування даних. Результатом цього етапу для прикладу, поданого на рис 1, буде графік, показаний на рис. 2.

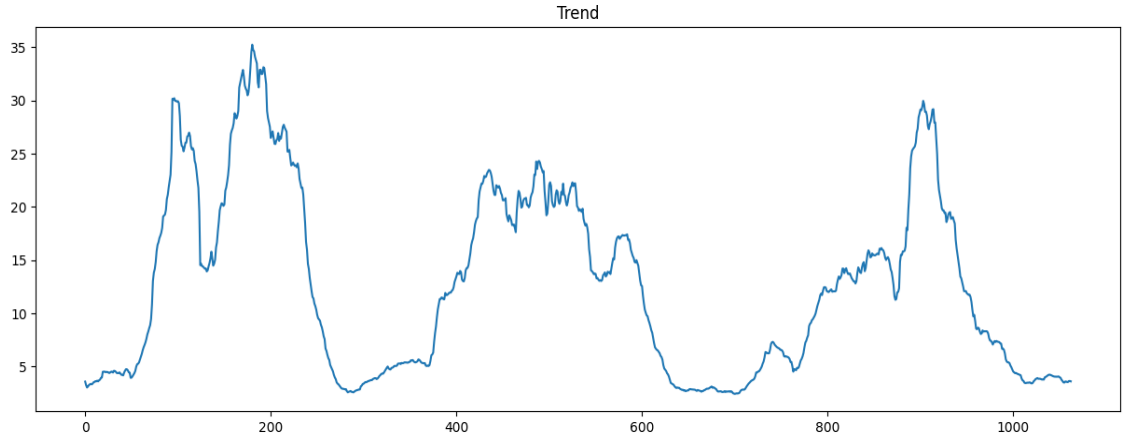


Рис. 2. Згладжена складова тренду для часового ряду показника PM10, наведеного на рис. 1

Eman 2. Пошук локальних мінімумів та максимумів та виділення окремих фрагментів ряду.

На отриманій згладженій кривій тренду часового ряду, необхідно знайти локальні мінімуми та локальні максимуми — точки кривої, де спадання змінюється на зростання та навпаки. До того ж, варто знайти та ігнорувати ті локальні максимуми, де відстань до наступного локального максимуму є малою. Таким чином ми зможемо відкинути незначні коливання та не аналізувати невеликі фрагменти як окремі складові часового ряду, оскільки пошук аномалій на малих часових відрізках не є доцільним. Приклад локальних мінімумів та локальних максимумів для ряду подано на рис. 3.

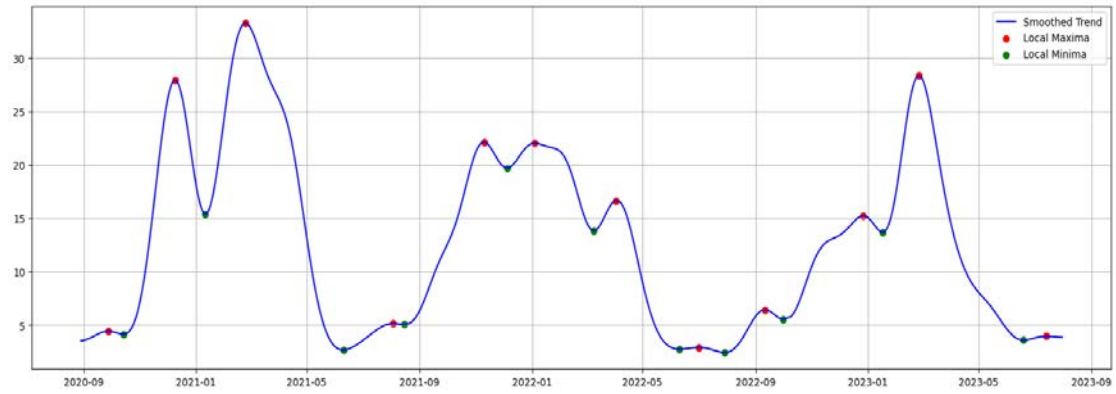


Рис. 3. Локальні мінімуми та локальні максимуми на кривій тренду часового ряду

Виділивши із загального списку вагомі локальні мінімуми, отримуємо точки за якими розділяємо наш часовий ряд на декілька фрагментів. Кількість цих фрагментів залежить від самого часового ряду, параметрів згладжування та визначення вагомості. Останні два значення можуть бути винесені як параметри налаштування методу задля гнучкіших можливостей аналізу відповідного часового ряду.

Eman 3. Пошук аномалій на окремих сегментах часового ряду, використовуючи комбінації методів.

Як описано в попередньому розділі, існує велика кількість різних за своєю природою методів для пошуку аномалій в часових рядах. З порівняльного аналізу видно, що кожен з методів має як свої переваги, так і свої недоліки. Як засіб усунення недоліків одного методу, пропонується одночасне використання декількох за різними стратегіями їхнього комбінування. Подібне поєднання може дозволити досягти більшої точності, оскільки один метод може знайти ті аномалії, які були пропущені іншим методом або ж навпаки, два метода будуть валідувати один одного і відкидати хибні аномалії. Для цього пропонується три стратегії поєднання методів:

Варіант 1. Одночасне застосування двох методів та об'єднання їх результатів. Нехай вхідний набір даних це *X*, тоді аномальні значення, отримані цим комбінованим методом можна позначити як

$$V_1(X) = M_1(X) \cup M2(X), \tag{5}$$

де $M_1(X)$ — аномальні дані, отримані першим методом для даних *X*, а $M_2(X)$ — аномальні дані, отримані другим методом.

Варіант 2. Одночасне застосування двох методів та знаходження перетину їхніх результатів. Нехай вхідний набір даних це *X*, тоді аномальні значення, отримані цим комбінованим методом можна позначити так:

$$V_2(X) = M_1(X) \cap M_2(X), \tag{6}$$

де $M_1(X)$ — аномальні дані, отримані першим методом для даних *X*, а $M_2(X)$ — аномальні дані, отримані другим методом.

Варіант 3. Покрокове застосування обох методів, де після застосування першого методу з набору вхідних даних видаляються аномальні дані і до отриманого набору застосовується другий метод. Результат роботи обох методів об'єднується в один набір. Нехай вхідний набір даних це *X*, а A_1 — аномальні значення, отримані за допомогою першого методу пошуку аномалій, тоді аномальні значення, отримані цим комбінованим методом можна позначити як

$$V_3(X) = A_1 \cup M_2(X \setminus A_1). \tag{7}$$

Усі можливі комбінації методів пошуку аномалій потрібно застосовувати на кожному фрагменті ряду, виділеному на попередньому кроці. Таким чином, дотримуючись нашої гіпотези, що часовий ряд показників стану довкілля поводить себе по-різному на кожному окремому відрізку ряду, можна проаналізувати фрагмент окремо від інших частин часового ряду та ідентифікувати локальні аномалії, властиві безпосередньо цьому відповідному фрагменту ряду.

Блок-схема алгоритму запропонованого методу показана на рис. 4.

Розглянемо приклад застосування методу.

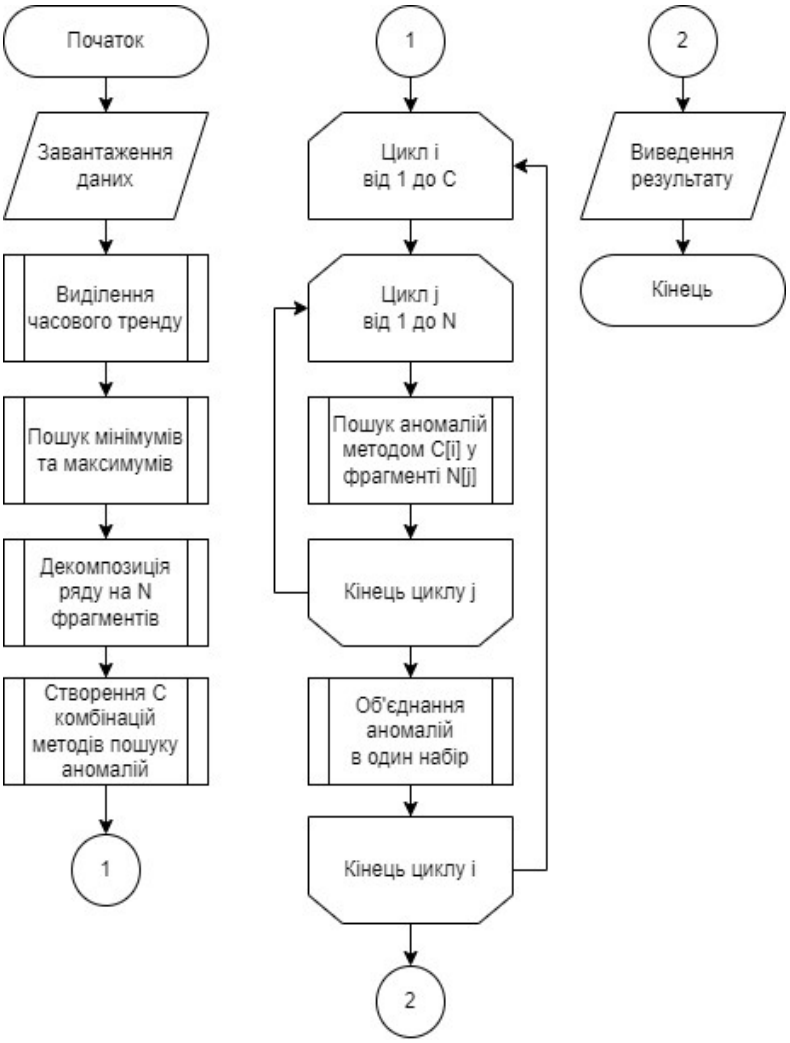


Рис. 4. Блок-схема алгоритму запропонованого методу ідентифікації локальних аномалій

Приклад розв'язання задачі

Для перевірки запропонованого методу, використовувались дані показників якості атмосферного повітря від мережі громадського моніторингу EcoCity (https://eco-city.org.ua/). Дані по Вінницькій області надаються через веб-сервіс «Кабінет дослідника», доступ до якого автори статті мають завдяки договору між ВНТУ та EcoCity. Цей сервіс дає змогу робити запити за певними показниками та за певні часові періоди. На рис. 5 показано мапу з веб-сайту EcoCity, де зображено активні станції моніторингу по місту Вінниця.

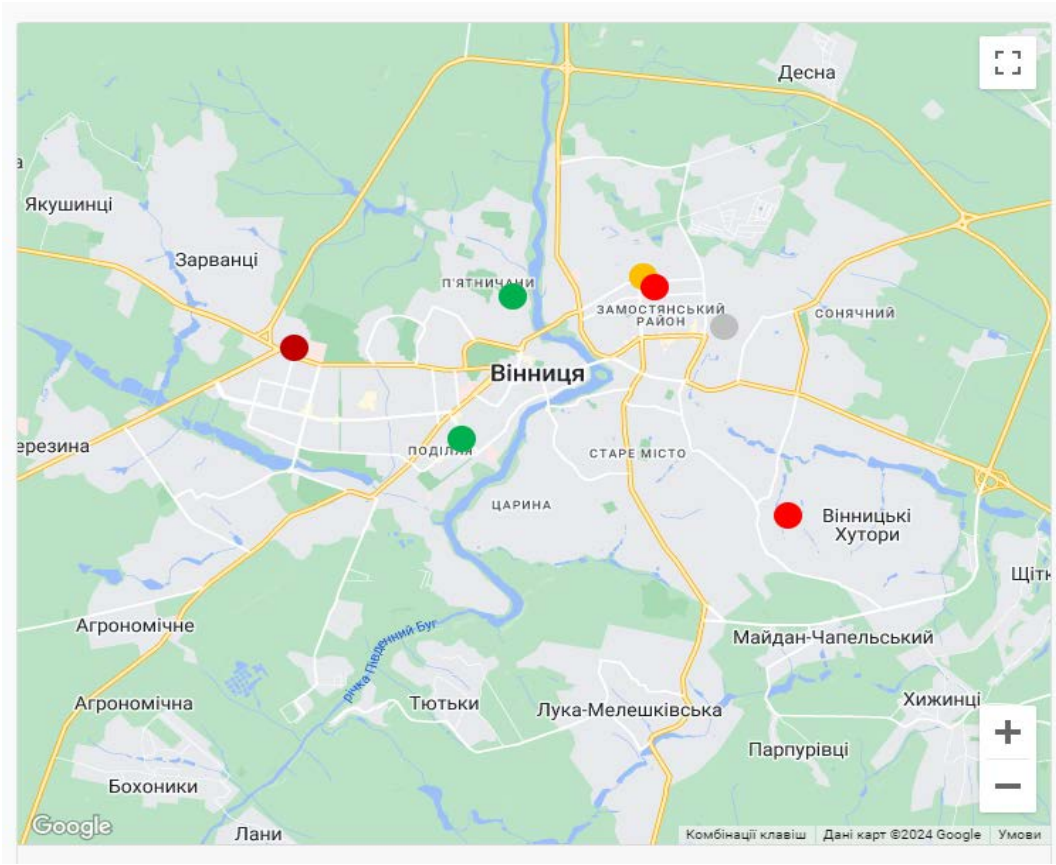


Рис. 5. Мапа станцій моніторингу мережі «EcoCity» , які працюють в межах міста Вінниця станом на лютий 2024 року

Кожна станція є пристроєм з різним набором датчиків, які вимірюють відповідні показники якості атмосферного повітря та передають на сервер EcoCity, який їх обробляє та агрегує. Для даного дослідження було обрано одну зі станцій, розташованих безпосередньо у місті Вінниця та показник пилу PM2.5 (мікроскопічні тверді частинки) за період 2019—2022 рр. Графік цих даних показано на рис. 6.

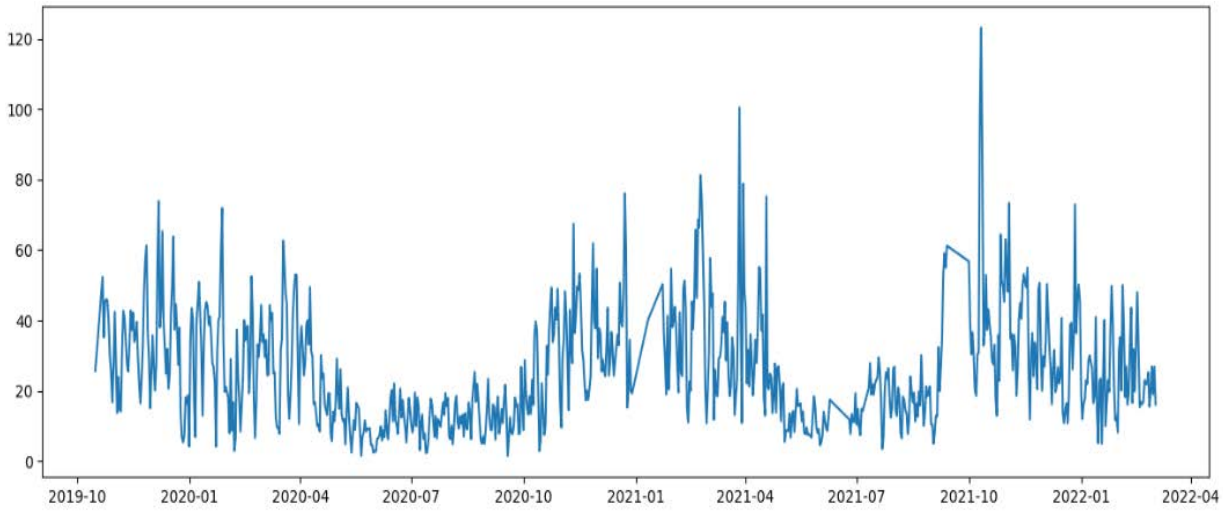


Рис. 6. Дані показника PM2.5 станції у місті Вінниця, надані мережею громадського моніторингу якості атмосферного повітря EcoCity (https://eco-city.org.ua/)

Покроково застосуємо метод ідентифікації.

Eman 1. Знайдемо складову тренду для даних, зображених на рис. 5 та згладимо їх за допомогою фільтру Баттерворта [14], [15] — одного з модулів Python-бібліотеки scipy.signal. Крива тренду та її згладжений варіант показані на рис. 7.

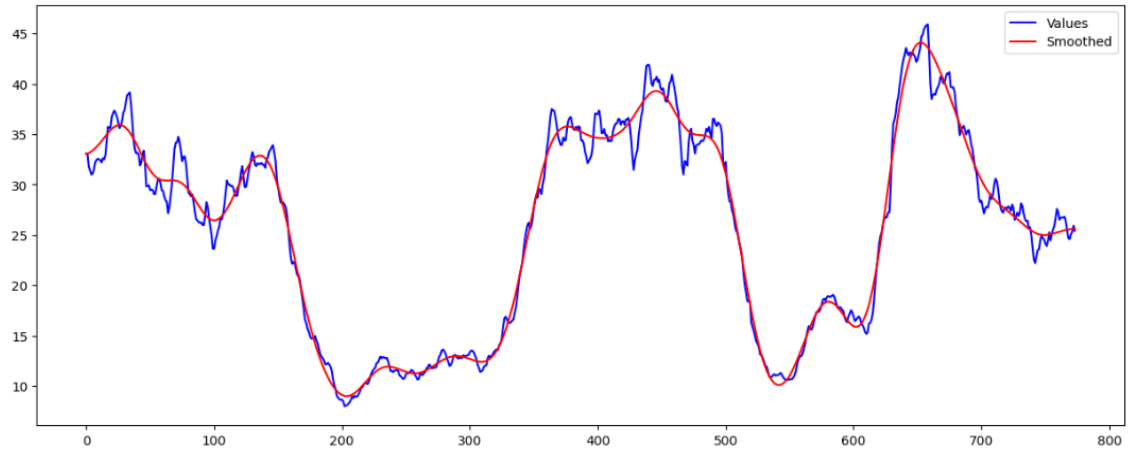


Рис. 7. Складова тренду та її згладжена версія для часового ряду, показаного на рис. 5

Eman 2. Визначимо локальні мінімуми та локальні максимуми на згладженій кривій тренду [10], [11]. Для цього застосуємо методи з раніше згаданої Python-бібліотеки scipy.signal. Отримані значення показано на рис. 8.

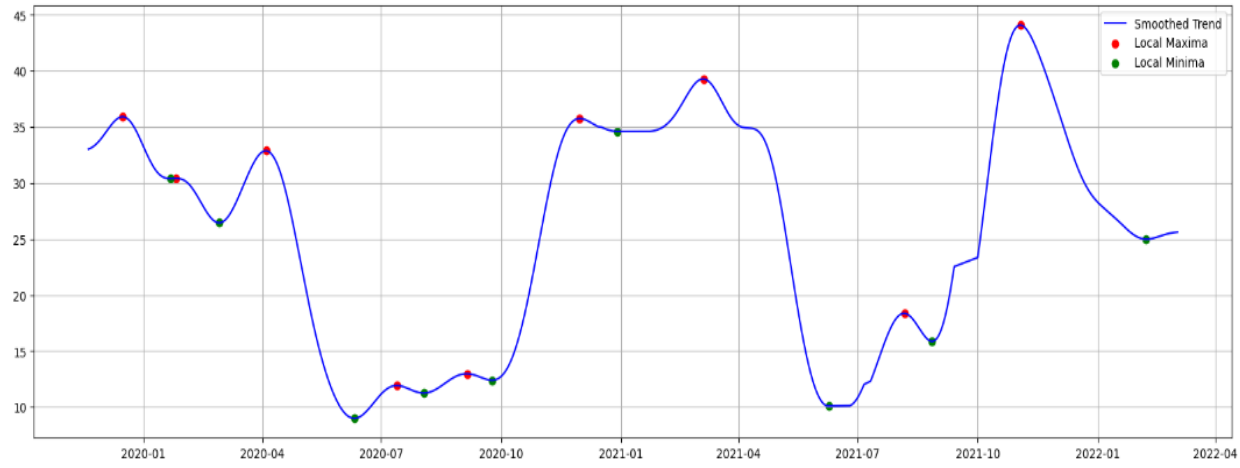


Рис. 8. Локальні максимуми та локальні мінімуми кривої тренду

Порівнюючи сусідні мінімуми та максимуми, необхідно знайти ті з них, які не суттєво відрізняються один від одного. Такі точки локальних мінімумів, як описано в попередньому розділі, будуть ігноруватися та не будуть використовуватися для визначення сегментів часового ряду [15], [16]. Важливі локальні мінімуми показано на рис. 9.

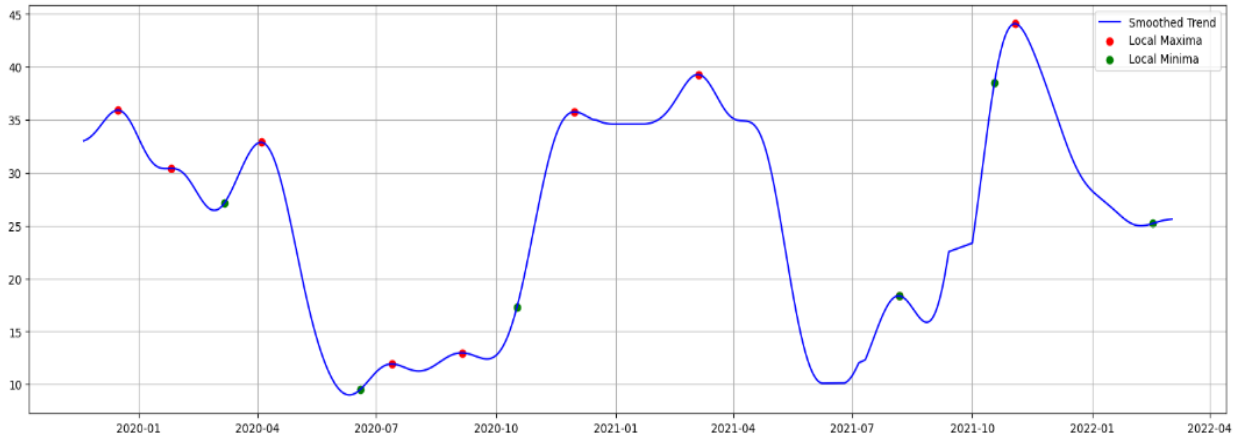


Рис. 9. Значущі локальні мінімуми кривої тренду

Розділимо часовий ряд на окремі сегменти, використовуючи отримані точки значущих локальних мінімумів. Результат показано на рис. 10.

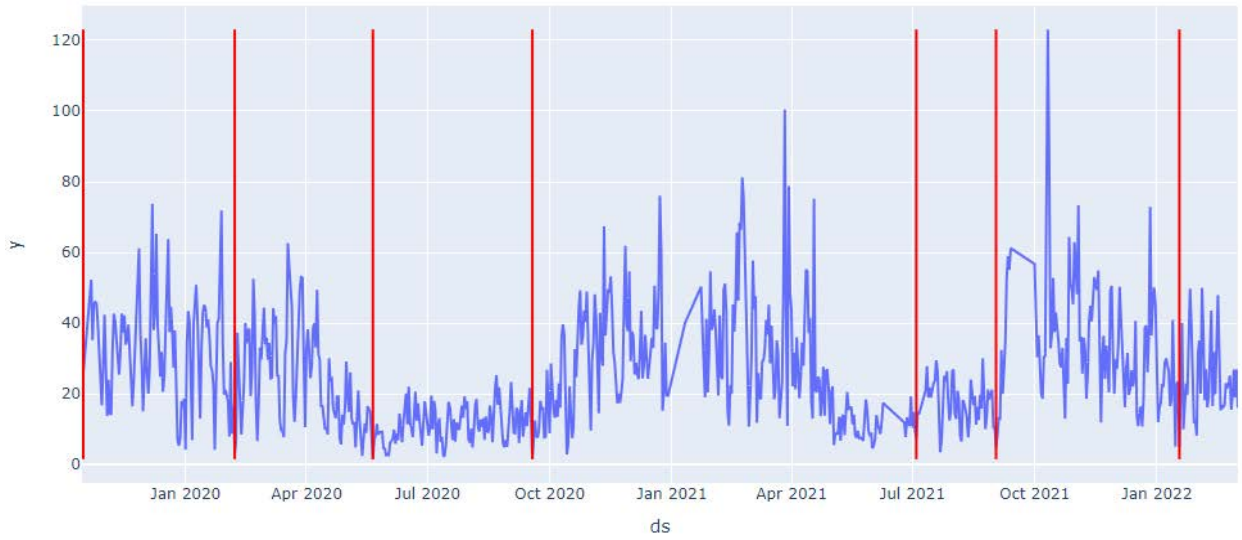


Рис. 10. Окремі сегменти часового ряду розділені червоними лініями — точками локальних мінімумів

Eman 3. Виконаємо пошук аномалій на кожному окремому фрагменті часового ряду, використовуючи наш комбінований підхід. Для цього використаємо 7 методів пошуку аномалій, описаних раніше в порівняльному аналізі. Крім використання різних комбінацій цих методів, варто також спробувати застосувати кожен з них окремо і тим самим порівняти чи дає комбінування методів якусь перевагу у точності порівняно з окремими методами пошуку аномалій. Як описано вище, комбінування методів буде виконуватися за трьома стратегіями. Перші дві стратегії комбінування не залежать від порядку використання методу пошуку, що відповідає вибірці, яка називається комбінація без повторень

$$N_1 = \frac{n!}{r!(n-r)!}, \tag{8}$$

де n — загальна кількість методів, r — кількість методів, які застосовуються разом, а N_1 — загальна кількість можливих унікальних комбінацій.

В нашому випадку усього є 7 методів пошуку аномалій і кожна комбінація включає 2 методи при застосуванні першої стратегії комбінування методів пошуку аномалій. Підставивши ці числа в формулу 8, отримуємо 21 можливу комбінацію. Таке саме число комбінацій є і у разі застосування другої стратегії комбінування методів пошуку аномалій.

Водночас, із застосуванням третьої стратегії комбінування методів пошуку аномалій, результат буде залежати від порядку застосування методу. Таким чином, загальну кількість комбінацій можна знайти за формулою

$$N_2 = n \times (n-1), \tag{9}$$

де n — загальна кількість методів, а N_2 — загальна кількість комбінацій з можливими перестановками.

Оскільки використано 7 різних методів пошуку аномалій, то загальна кількість комбінацій, згідно з формулою 9, дорівнює 42. До того ж, ми також застосовуємо кожен з семи методів окремо, що додає нам ще 7 варіантів пошуку аномалій. Таким чином, загалом наш приклад налічує 91 варіант. Реалізувавши цей перебір за допомогою Python, отримуємо результат, показаний на рис. 11.

	name	anomaly_values
0	Hampel	[13, 49, 75, 76, 93, 130, 170, 215, 239, 260, ...
1	Moving Average	[46, 100, 110, 125, 148, 152, 153, 161, 260, 2...
2	Isolation Forest	[46, 98, 148, 149, 305, 316, 332, 465, 497, 59...
3	LOF	[1, 34, 35, 36, 45, 46, 49, 57, 58, 66, 71, 79...
4	k-neighbors	[1, 34, 35, 36, 45, 46, 49, 58, 79, 97, 98, 10...
...	...	...
86	SVM → Moving Average anomalies	[1, 6, 28, 36, 45, 46, 47, 49, 50, 54, 61, 63,...
87	SVM → Isolation Forest anomalies	[1, 6, 28, 45, 46, 47, 49, 50, 61, 63, 71, 77,...
88	SVM → LOF anomalies	[1, 6, 28, 32, 33, 45, 46, 47, 49, 50, 56, 61,...
89	SVM → k-neighbors anomalies	[1, 6, 27, 28, 31, 32, 33, 45, 46, 47, 49, 50,...
90	SVM → SESD anomalies	[1, 6, 13, 23, 27, 28, 32, 33, 42, 45, 46, 47,...

Рис. 11. Результат пошуку аномалій кожним з 91-го варіанта перебору

Для перевірки точності роботи експертним шляхом знайдемо аномальні записи на кожному з фрагментів часового ряду. В ході експертного аналізу виділено 41 запис, який візуально відрізняється від інших записів відповідного фрагмента часового ряду. Перелік аномалій, знайдених експертним шляхом, показано на рис. 12.

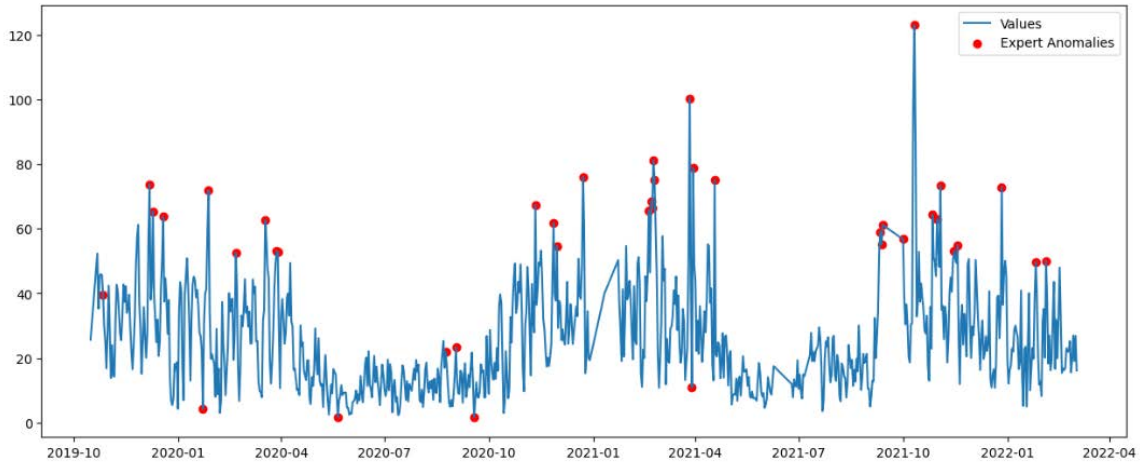


Рис. 12. Аномалії часового ряду, визначені експертним шляхом

Для перевірки точності кожної з 91 комбінації методів пошуку аномалій, перевіримо наскільки список експертних аномалій відрізняється від відповідного списку аномалій, знайдених програмним шляхом. Для цього створимо список значень, який дорівнює розміру нашого часового ряду. Кожен елемент списку це 1, якщо відповідний елемент ряду вважається аномалією, або 0, якщо елемент вважається нормальним. Як критерій точності використаємо декілька статистичних метрик: влучність (англ. «precision» або «precision score»), повнота (англ. «recall» або «recall score») та f-міра — показник, який обчислюється через значення повноти та влучності [17]. Визначивши ці метрики для кожного з списків, ми отримуємо результат, показаний на рис. 13.

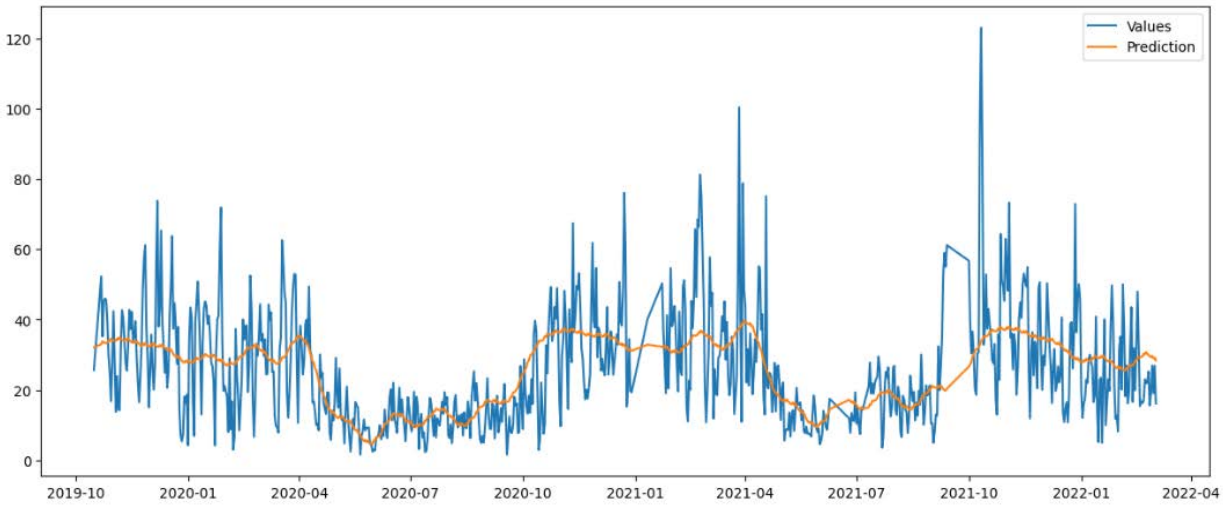
	name	anomaly_values	precision_score	recall_score	f1_score
43	LOF ∩ k-neighbors anomalies	[1, 34, 35, 36, 45, 46, 49, 58, 79, 97, 98, 10...	0.737571	0.830996	0.775075
18	Isolation Forest + LOF anomalies	[1, 34, 35, 36, 45, 46, 49, 57, 58, 66, 71, 79...	0.705752	0.836943	0.751830
3	LOF	[1, 34, 35, 36, 45, 46, 49, 57, 58, 66, 71, 79...	0.705752	0.836943	0.751830
69	LOF → Isolation Forest anomalies	[1, 34, 35, 36, 45, 46, 49, 57, 58, 66, 71, 79...	0.702652	0.836288	0.749021
45	LOF ∩ SVM anomalies	[1, 45, 46, 49, 71, 93, 98, 108, 148, 149, 158...	0.737516	0.748739	0.742974
4	k-neighbors	[1, 34, 35, 36, 45, 46, 49, 58, 79, 97, 98, 10...	0.687918	0.844201	0.737270
19	Isolation Forest + k-neighbors anomalies	[1, 34, 35, 36, 45, 46, 49, 58, 79, 97, 98, 10...	0.687918	0.844201	0.737270
47	k-neighbors ∩ SVM anomalies	[1, 45, 46, 49, 98, 108, 148, 149, 158, 159, 2...	0.736894	0.736894	0.736894
75	k-neighbors → Isolation Forest anomalies	[1, 34, 35, 36, 45, 46, 49, 50, 58, 79, 97, 98...	0.685411	0.843545	0.734748
44	LOF ∩ SESD anomalies	[34, 35, 36, 45, 46, 49, 57, 58, 66, 71, 79, 9...	0.750709	0.702670	0.723725
63	Isolation Forest → LOF anomalies	[1, 34, 35, 36, 45, 46, 48, 56, 57, 65, 70, 78...	0.681653	0.797477	0.722049
22	LOF + k-neighbors anomalies	[1, 34, 35, 36, 45, 46, 49, 57, 58, 66, 71, 79...	0.669843	0.850147	0.719860
64	Isolation Forest → k-neighbors anomalies	[1, 34, 35, 36, 45, 46, 48, 56, 57, 78, 92, 96...	0.668928	0.816579	0.714111
14	Moving Average + LOF anomalies	[1, 34, 35, 36, 45, 46, 49, 57, 58, 66, 71, 79...	0.665431	0.837647	0.713570
76	k-neighbors → LOF anomalies	[1, 34, 35, 36, 45, 46, 49, 50, 58, 63, 79, 84...	0.665431	0.837647	0.713570
46	k-neighbors ∩ SESD anomalies	[34, 35, 36, 45, 46, 49, 58, 79, 148, 149, 237...	0.742520	0.690170	0.712649

Рис. 13. Результати визначення точності кожного з методів пошуку аномалій у порівнянні з аномаліями, отриманими експертним шляхом, посортовані за показником f1_score

Як можна помітити, серед найкращих методів часто зустрічається метод LOF. Також високу точність мають методи з використанням Isolation Forest та k-neighbors. Якщо намагатися виділити які зі стратегій комбінування методів найефективніші, то видно що серед найкращих зустрічаються усі три підходи.

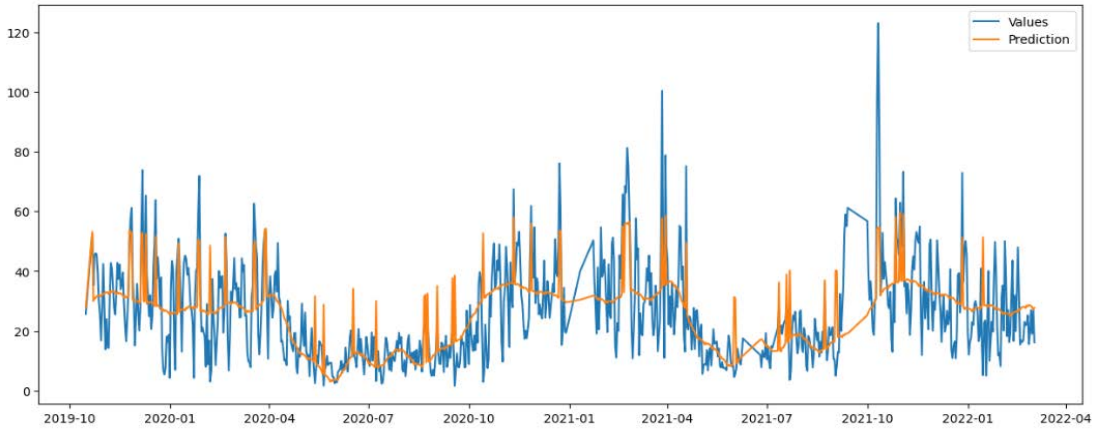
Як додатковий етап перевірки, спробуємо використати найточніший метод для апроксимації часового ряду. Для цього побудуємо модель Facebook Prophet з параметрами за замовчування та з

налаштуванням «holidays», який дозволяє передати в модель список аномальних дат і тим самим покращити роботу моделі. Результати роботи моделі Prophet з параметрами за замовчуванням та з використанням аномалій зображено на рис. 14 та рис. 15 відповідно.



rmse_score=12.953687539166406, mae_score=9.381988320149032

Рис. 14. Результати апроксимації часового ряду показника PM2.5 за допомогою моделі Prophet та параметрами налаштування за замовчуванням



rmse_score=11.60732502268683, mae_score=8.667046412877497

Рис. 15. Результати апроксимації часового ряду показника PM2.5 за допомогою моделі Prophet та аномалій, отриманих найточнішим комбінованим методом

Точність апроксимації визначено за допомогою середньої абсолютної похибки (англ. «Mean absolute error» або скорочено — MAE) та середньоквадратичної похибки (англ. «Root mean squared error» або скорочено — RMSE) [17]. За цим показником похибка апроксимації зменшилась на 11 % за RMSE і на 8 % за MAE з використанням ідентифікованих аномалій у порівнянні з моделлю Prophet з параметрами за замовчуванням. Варто зазначати, що вибраний часовий ряд досить важко аналізувати та передбачувати, але все одно вдалося підвищити точність роботи моделі.

Висновки

Запропоновано новий метод ідентифікації локальних аномалій значень показників стану довкілля з використанням декомпозиції на півхвилі. Описано алгоритм його роботи та розглянуто приклад його застосування на реальних даних показників якості атмосферного повітря мережі громадського моніторингу EcoCity. Для прикладу використано значення показника PM2.5 з однієї зі станцій моніторингу, розташованої у місті Вінниця.

Точність роботи методу перевірялась за допомогою порівняння аномалій, знайдених програмним шляхом з аномаліями, отриманими експертним шляхом. Результат роботи найточнішого з

методів застосувався для апроксимації часового ряду моделлю Prophet. За метриками RMSE та MAE у порівнянні з моделлю Prophet з параметрами налаштування за замовчуванням вдалось зменшити помилку апроксимації на 11 % та 8 % відповідно.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Б. І. Мокін, О. Б. Мокін, і В. Б. Мокін, *Методологія та організація наукових досліджень*, підруч., вид.3-е, змін. та доп. Вінниця, Україна: ВНТУ, 2023, 230 с.

[2] Terence C. Mills, *Chapter 3, ARMA Models for Stationary Time Series*, Terence C. Mills. Ed, Applied Time Series Analysis, Academic Press, 2019, pp. 31-56. ISBN 9780128131176. https://doi.org/10.1016/B978-0-12-813117-6.00003-X .

[3] Omar Salima , Ngadi Md, Jebur Hamid, and Benqdara Salima, "Machine Learning Techniques for Anomaly Detection: An Overview," *International Journal of Computer Applications*, 79, 2013, https://doi.org/10.5120/13715-1478 .

[4] В. Б. Мокін, О. В. Слободянюк, О. М. Давидюк, і Д. О. Шмундяк, «Інформаційна технологія пошуку можливих джерел підвищеного забруднення річки з використанням моделі Prophet,» *Вісник Вінницького політехнічного інституту*, № 4, с. 15-24, Верес. 2020. https://doi.org/10.31649/1997-9266-2020-151-4-15-24 .

[5] О. Б. Мокін, В. Б. Мокін, і Б. І. Мокін, «Алгоритм методу ідентифікації моделі авторегресії — ковзного середнього, який узагальнює методику Юла–Уокера, та його програмна Python-реалізація,» *Вісник Вінницького політехнічного інституту*, № 4, с. 41-55, 2022. https://doi.org/10.31649/1997-9266-2022-163-4-41-55 .

[6] R. K. Pearson, et al., "Generalized Hampel Filters," *EURASIP J. Adv. Signal Process*, 87, 2016. https://doi.org/10.1186/s13634-016-0383-6 .

[7] Julien Lesouple, Cédric Baudoin, Marc Spigai, and Jean-Yves Tourneret, "Generalized isolation forest for anomaly detection," *Pattern Recognition Letters*, vol. 149, 2021, pp, 109-119. ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2021.05.022 .

[8] Yumin Chen, Duoqian Miao, and Hongyun Zhang, "Neighborhood outlier detection," *Expert Systems with Applications*, vol. 37, issue 12, pp. 8745-8749, 2010. ISSN 0957-4174. https://doi.org/10.1016/j.eswa.2010.06.040 .

[9] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. SIGMOD Rec. 29, no. 2, pp. 93-104, June 2000. https://doi.org/10.1145/335191.335388 .

[10] Vieira, Rafael G.; Leone Filho, Marcos A.; Semolini, Robinson, "An Enhanced Seasonal-Hybrid ESD Technique for Robust Anomaly Detection on Time Series," in Simpósio Brasileiro De Redes De Computadores E Sistemas Distribuídos (SBRC), 36, 2018, Campos do Jordão. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2018. pp. 281-294. ISSN 2177-9384. https://doi.org/10.5753/sbrc.2018.2422 .

[11] А. В. Лосенко, «Інформаційна технологія прогнозування часового ряду кількості хворих на коронавірус на основі моделі Facebook Prophet,» *Вісник Вінницького політехнічного інституту*, вип. 5, с. 50-59, 2023. https://doi.org/10.31649/1997-9266-2023-170-5-50-59 .

[12] В. Б. Мокін, А. В. Лосенко, і А. Р. Ящолт, «Інформаційна технологія аналізу та прогнозування кількості нових випадків хвороби на коронавірус SARS-COV-2 в Україні на основі моделі Prophet,» *Вісник Вінницького політехнічного інституту*, № 5, с. 71-83, 2020. https://doi.org/10.31649/1997-9266-2020-152-5-71-83 .

[13] В. Б. Мокін, А. В. Лосенко, і А. Р. Ящолт, «Інформаційна технологія аналізу та прогнозування багатохвильової кількості нових випадків захворювань на коронавірус COVID-19 на основі моделі Prophet», *Вісник Вінницького політехнічного інституту*, № 6, с. 65-75, 2020. https://doi.org/10.31649/1997-9266-2020-153-6-65-75 .

[14] Д. О. Шмундяк, і В. Б. Мокін, «Метод ідентифікації параметрів гармонік та аномалій періодичного часового ряду на основі адаптивної декомпозиції,» *Вісник Вінницького політехнічного інституту*, № 6, с. 46-56, 2023. https://doi.org/10.31649/1997-9266-2023-171-6-46-56 .

[15] Dmytro Shmundiak, and Vitalii Mokin, "Adaptive decomposition for harmonics and anomalies," *Kaggle Notebook.* [Electronic resource]. Available: https://www.kaggle.com/code/dimashmundiak/adaptive-decomposition-for-harmonics-and-anomalies . Accessed:20.12.2023.

[16] Vitalii Mokin, and Arsen Losenko, "COVID-19 Ukraine daily cases – EDA," *Kaggle Notebook.* [Electronic resource]. Available: https://www.kaggle.com/code/vbmokin/covid-19-ukraine-daily-cases-eda . Accessed:12.10.2023.

[17] *Sklearn*. *API Reference*. [Electronic resource]. Available: https://scikit-learn.org/stable/modules/classes.html. Accessed: 07.12.2023.

Рекомендована кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 20.02.2024

Шмундяк Дмитро Олександрович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: dimashmund@gmail.com ;

Копняк Володимир Євгенович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: vkopnyak@gmail.com .

Вінницький національний технічний університет, Вінниця

D. O. Shmundiak[1]
V. Ye. Kopniak[1]

Method of Local Anomalies Identification for Environmental Indicators Values Using Half-Wave Decomposition

[1]Vinnytsia National Technical University

In the era of mass digitalization of all existing spheres of human activity, the amount of data is constantly growing and it is crucial to be able to work with such volume of data for the solution of various problems. One of the most common data structures is a time series — a sequence of data points, collected over some period of time, usually in chronological order. The time series comprise various financial indicators, environmental monitoring data, medical information, etc. Wide range of application areas makes the problem of time series analysis important and relevant. The quality of the time series forecast greatly depends on the quality of the performed analysis, which may include data standardization, detection of significant indicators, correlation analysis, etc. Anomaly detection occupies very important place among these steps. Anomalies are data points that differ in some way from other values in the dataset or violate certain data behavior patterns. The presence of similar records greatly affects the ability of machine learning models make accurate predictions, is why it is necessary to have the possibility for the identification of these anomalies.

New method of local anomalies identification of the environment state indices using half-wave decomposition has been developed. Main idea of the method is to decompose the time series into half-waves, using trend points where the fall changes growth or vice versa and split the series into fragments. Each fragment is analyzed separately and is checked for anomalies by combining numerous methods. The accuracy of the methods is verified, applying the expert method. Main steps of the proposed method are described and the example of the method usage on real air quality monitoring data obtained from one of the stations of the EcoCity public monitoring network within the international program "Clean Air for Ukraine" is given.

The proposed method was implemented and tested on the Kaggle platform's notebook. The result of the anomaly detection was used for the construction of the Facebook Prophet model and the accuracy of the time series approximation was compared with the results of the Prophet model operation with the default parameters. Tests have shown 11 % decrease of approximation error of time series for RMSE metric and 8 % decrease for MAE metric. This result confirms the effectiveness of the method.

Keywords: time series analysis, simulation, machine learning, time series anomalies, air quality, time series decomposition, EcoCity.

Shmundiak Dmytro O. — Post-Graduate Student of the Chair of System Analysis and Information Technologies, e-mail: dimashmund@gmail.com ;

Kopniak Volodymyr Ye. — Post-Graduate Student of the Chair of System Analysis and Information Technologies, e-mail: vkopnyak@gmail.com