https://doi.org/10.31649/1997-9266-2023-171-6-46-56

УДК 004.9:504.3.054

Д. О. Шмундяк[1]
В. Б. Мокін[1]

МЕТОД ІДЕНТИФІКАЦІЇ ПАРАМЕТРІВ ГАРМОНІК ТА АНОМАЛІЙ ПЕРІОДИЧНОГО ЧАСОВОГО РЯДУ НА ОСНОВІ АДАПТИВНОЇ ДЕКОМПОЗИЦІЇ

[1]Вінницький національний технічний університет

Періодичні часові ряди зустрічаються в багатьох задачах — це і фінансові показники, і показники якості атмосферного повітря, і показники стану води тощо. Відповідно їхнє моделювання та аналіз закономірностей є актуальним і досить поширеним завданням для розуміння можливих тенденцій і змін для коректного та своєчасного реагування. Важливими параметрами періодичних часових рядів є параметри їхнього тренду, сезонних складових та аномалій. І якщо задача визначення тренду часового ряду має багато універсальних методів розв'язання, то ідентифікація одночасно параметрів різних видів сезонності та аномалій різної природи у різні часові проміжки є складною задачею, яка не має універсального розв'язання. Більшість таких розв'язків є специфічними для конкретної предметної області або демонструють не чітку адекватність та точність апроксимації.

Розроблено новий метод ідентифікації параметрів гармонік та аномалій періодичного часового ряду, який базується на адаптивній декомпозиції ряду. Зокрема, запропоновано здійснювати декомпозицію заданого часового ряду з періодом до половини від загальної кількості точок і будувати графік відношень амплітуд сезонної складової до амплітуд самого ряду — так званої «декомпозиційної кривої». А тоді, згладжувати цю криву і знаходити локальні максимуми, які пропонується вважати такими, що відповідають періоду можливих видів сезонності ряду. З урахуванням багаторічного досвіду використання моделі Facebook Prophet запропоновано низку співвідношень між періодом сезонності, порядком ряду Фур'є для її апроксимації та ступенем регуляризації, який варто враховувати. Для кожного виду сезонності у кожному періоді одним з відомих методів слід знаходити аномальні дані та перевіряти їхню статистичну значущість. Статистично значущі аномалії збирати в єдину множину з типовими параметрами. Запропоновано низку можливих варіантів структур таких моделей часового ряду. Наведено алгоритм методу та описано його основні складові.

Здійснено випробування запропонованого методу на Python на базі платформи Kaggle з використанням моделі Facebook Prophet на реальних даних спостережень за якістю атмосферного повітря, отриманих з однієї зі станцій мережі громадського моніторингу EcoCity у межах міжнародної програми «Чисте повітря для України». Випробування показали, що порівняно з моделлю з параметрами і видами сезонності за замовчуванням, запропонований метод дозволив покращити точність апроксимації оптимальної моделі за метрикою R2 у 1,7 рази, а за метрикою MSE — у 2 рази. Це підтвердило ефективність запропонованого методу.

Ключові слова: аналіз часових рядів, моделювання, машинне навчання, аномалії часових рядів, сезонність, гармоніки ряду Фур'є, якість атмосферного повітря, EcoCity.

Вступ

Одним з найбільших класів часових рядів є періодичні ряди з аномальними значеннями — це майже усі види метеопараметрів, дані екологічного моніторингу, дані продаж у магазинах протягом тривалого періоду тощо [1]—[3]. І важливо вміти не тільки прогнозувати такі дані, а й просто апроксимувати адекватну модель. З використанням такої моделі можна виявити багато цінних закономірностей, які важко отримати із самого ряду.

Існує багато методів моделювання періодичних часових рядів та R чи Python-бібліотек для їхньої автоматизації. Переважна більшість дослідників використовує різні моделі на основі авторегресії та проінтегрованого ковзного середнього (АРПКС — англ. «ARIMA») [1]—[3]. Останнім часом більшої популярності набуває модель Facebook Prophet (FB Prophet), яка вперше запропоно-

© Д. О. Шмундяк, В. Б. Мокін, 2023

вана у роботі [4]. Також, особливо у разі доволі довгих рядів, використовуються нейромережеві моделі RNN чи LSTM та ін. Всі ці моделі дуже залежать від правильної ідентифікації аномальних значень та їхніх параметрів, особливо багато параметрів аномалій враховує модель FB Prophet, основними з яких є такі: «вікно впливу», тобто з якого кроку починається вплив аномалії і коли закінчується, ступінь регуляризації prior_scale, тобто ступінь врахування цього впливу [4]—[8]. Ця ж модель має можливість моделювання різних видів сезонності гармоніками ряду Фур'є. Параметри сезонності слід задавати — вони не визначаються автоматично. А для періодичних рядів важливо знати, як мінімум, період ряду і порядок ряду Фур'є.

Складність пошуку одночасно параметрів аномалій та сезонності ряду ще полягає і в тому, що це взаємозалежні задачі. Наприклад, якщо аналізувати середньодобову температуру протягом багатьох років, то температура 0 °С влітку — це аномалія, а взимку — ні. І навпаки: температура взимку +20 °С — аномалія, а влітку — ні. Отже, аномалії у різних періодах ряду — різні. А для цього треба знати ці періоди. Однак, щоб знайти період ряду, слід спочатку відфільтрувати аномалії. Особливо задача ускладнюється, якщо має місце декілька різних видів сезонності одночасно або сезонності поєднуються мультиплікативно. Єдиного підходу розв'язання такої задачі не існує.

Мета дослідження — підвищення точності апроксимації періодичного часового ряду з використанням універсального методу ідентифікації параметрів періодичних складових та аномалій цього ряду. Універсальність означає «без знання предметної області». Метод повинен працювати автоматично без експертного втручання.

Формалізація задачі і традиційні підходи до її розв'язання

Періодичний часовий ряд $y(t)$, в залежності від часу t з аномаліями, зазвичай, записують у такому вигляді [4], [6]:

$$y(t) = g(t) + s(t) + h(t) + \varepsilon_t \tag{1}$$

або

$$y(t) = g(t)s(t)h(t)\varepsilon_t, \tag{2}$$

де $g(t)$ — тренд, $s(t)$ — сезонна складова, апроксимована рядом Фур'є, $h(t)$ — складова, яка враховує вплив аномалій з певним «вікном», тобто в діапазоні певних дат (до і після кроку з аномальним значенням), ε_t — похибка.

У статті [6] запропоновано деталізованіший варіант, який враховує основні параметри часових рядів у моделі FB Prophet. Запишемо модель на прикладі виразу (1)

$$y(t) = g\left(R_g, t\right) + \prod_{i=1}^{\phi} s\left(R_{si}, P_i, n_i, t\right) + h\left(\bigcup_{j=1}^{\psi} H_j\left(R_{hj}, t_{oj}, t_{1j}, t\right)\right) + \varepsilon_t, \tag{3}$$

де R_g — ступінь регуляризації тренду, тобто наскільки гнучко тренд має підлаштовуватися до різких змін значень ряду; ϕ — кількість сезонних складових, які одночасно враховуються; для кожної *i*-ї сезонної складової окремо задаються такі параметри: R_{si} — ступінь регуляризації, P_i — період у добах, n_i — порядок ряду Фур'є, яким описується ця складова; ψ — кількість різних видів аномалій (враховують різні знання предметної області); для кожної *j*-ї аномалії окремо задаються такі параметри: R_{hj} — ступінь регуляризації, $\left[t_{0j}, t_{1j}\right]$ — «вікно» впливу аномалії, де t_{0j} — на яку кількість діб до дати аномалії вже починає проявлятись її вплив, t_{1j} — через яку кількість діб після дати аномалії її вплив вже практично не проявляється; $H_j(\bullet)$ — матриця з датами аномалій та їхніми параметрами.

В оригінальній формулі (3) статті [6] ще був параметр, який враховував можливий зсув прояву аномалії в часі, оскільки її розробляли для опису хворих на коронавірус, в яких симптоми проявляються з певним запізненням, але це вважаємо вже надмірною деталізацією, як для нашого загального випадку.

Гіперпараметрами *K* моделі (3), тобто параметрами, які слід задавати вручну і які, зазвичай, не можуть автоматично визначати відомі Python-бібліотеки, є такі:

$$K = \left[R_g, R_{s1} \ldots R_{s\Phi},\ n_1 \ldots n_\Phi,\ P_1 \ldots P_\Phi,\ R_{h1} \ldots R_{h\Psi},\ t_0 \ldots t_{0\Psi},\ t_1 \ldots t_{1\Psi}\right]. \tag{4}$$

Традиційний підхід до визначення параметрів K полягає в такому.

1. Декомпозиція ряду, її, зазвичай, проводять з використанням методу seasonal_decompose(P) Python-пакету statsmodels.tsa.seasonal, який розбиває ряд $y(t)$ на складові $g(t)$, $s(t)$ (сезонну складову з періодом P), $h(t)$, ε_t (рис. 1).

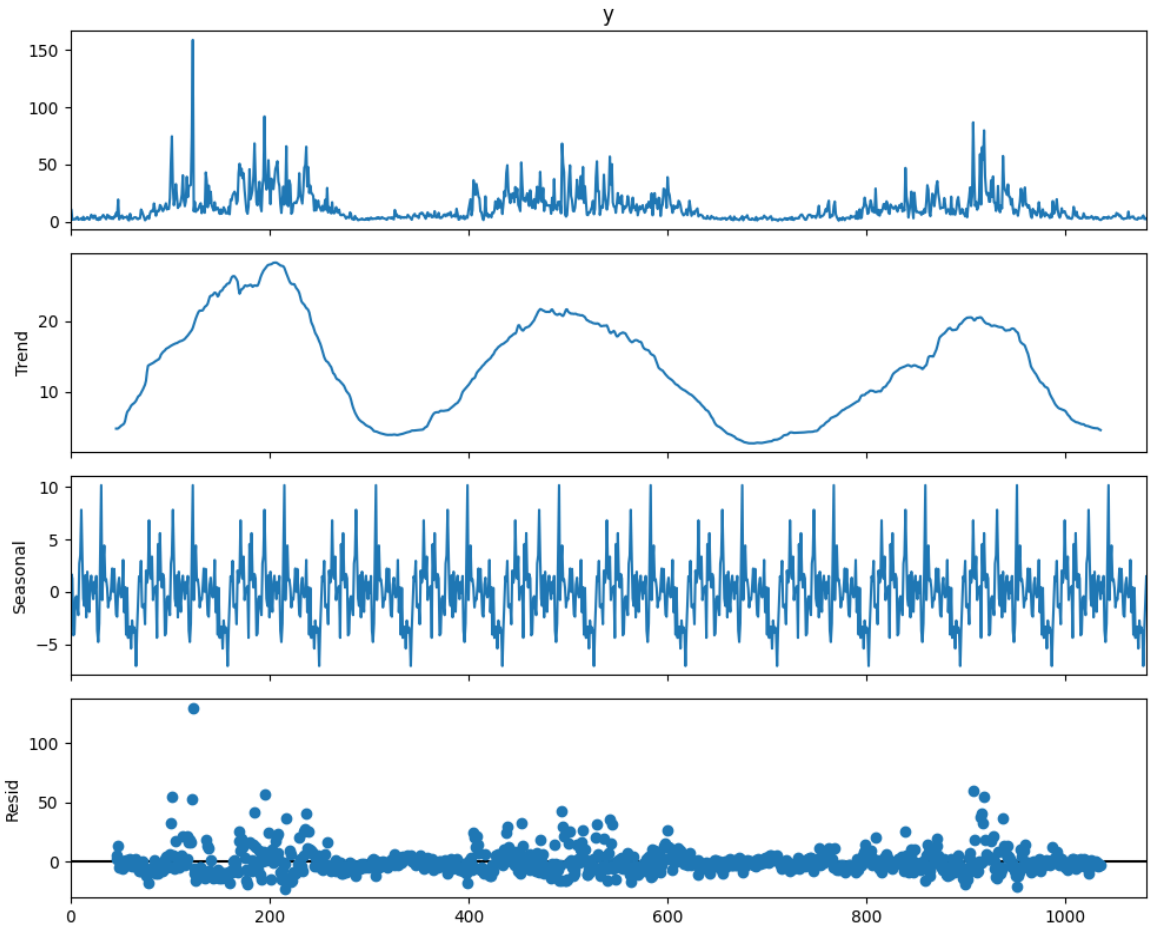

Рис. 1. Декомпозиція періодичного часового ряду *y* показника PM10 на тренд, сезонну складову із заданим періодом 92 доби, складову з аномаліями та похибку [9]

2. Експерт, зазвичай візуально, вивчає за якого значення періоду P ця декомпозиція дасть відчутний ефект. Один зі способів — це порівняння амплітуди сезонної складової $S(P)$ (різниці між максимальним і мінімальним значенням) з амплітудою самого ряду значень y [10]. Якщо це відношення є відчутним, тоді сезонна складова є достатньо значущою. Такий аналіз дає значення періоду. Є й інші способи, до прикладу, запропоновані у статтях [7], [8].

3. Потім аналізують ряд. Виявляють аномалії, які дають розуміння предметної області. Аномалії, які суттєво перевищують певний поріг, наприклад, коли максимальне значення в 10 разів є більшим за перцентиль P90 чи ін. Існує низка універсальних методів виявлення аномальних даних в часовому ряді: метод стандартної оцінки (англ «Standard Score» або «z-score») [11], метод міжквартильного розмаху (англ. «Interquartile range», скорочено — IQR) [12], метод ізольованого лісу («Isolation Forest») [13], метод *k*-найближчих сусідів (англ. «*k*-Nearest Neighbors», скорочено — «*k*-NN») [14] тощо. Ці та інші відомі методи базуються на основі аналізу статистичних показників, методів кластеризації, побудови дерев. Кожен метод має свої переваги, недоліки, обмеження та сферу застосування.

4. Для кожної аномалії визначають параметри «вікна», з урахуванням знань предметної області, та формують матрицю $H_j(\bullet)$.

5. Перебирають варіанти значень K з (4). Іноді виконують низку спрощень, наприклад, одні параметри виражають через інші, як у роботі [6]. Для визначення параметрів або використовують метод GridSearchCV з повним перебором, або — багатоітеративний підхід, або байєсівську оптимізацію з використанням методу HyperOpt, чи ін. [6], [15]. Врахування більшої кількості парамет-

рів збільшує точність та адекватність моделі, але суттєво впливає на тривалість обчислень. А ще зростає ризик перенавчання (англ. «overfitting»).

Новий метод розв'язання задачі

Для розв'язання поставленої задачі пропонується максимально автоматизувати кожний етап її традиційного розв'язання з використанням автоматичних процедур та низки критеріїв.

Етап 1. *Визначення кількості та періоду можливих видів сезонності.*

У програмі [10] одним зі співавторів цієї статті запропоновано будувати графік усіх варіантів відношень амплітуд $S_Y(P)$ з періодом P від 1 до 50 % від кількості усіх значень ряду

$$S_Y(P) = \frac{S(P)}{Y}, \quad P = 1,...0,5N, \tag{5}$$

де N — кількість значень ряду.

Приклад такого графіка (назвемо його «декомпозиційною кривою») $S_Y(P)$ показано на рис. 2.

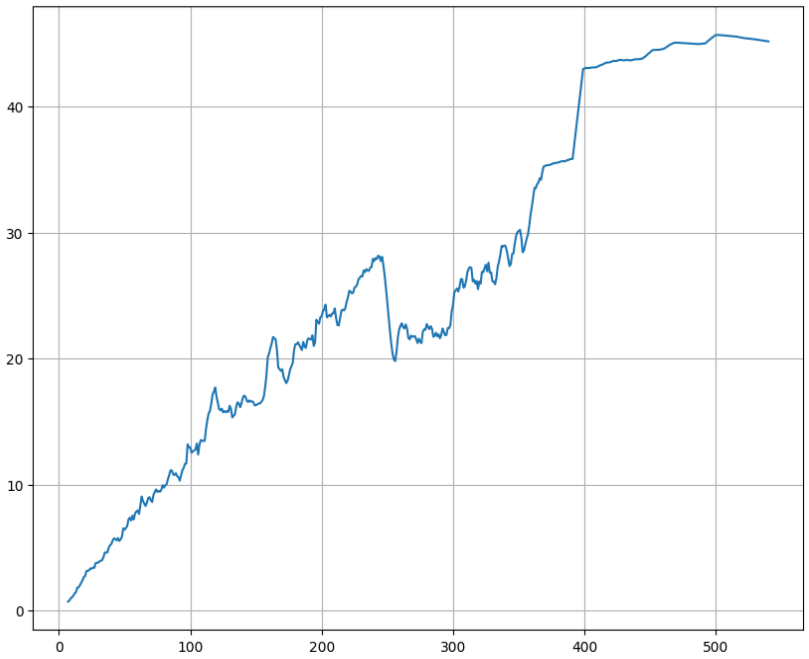

Рис. 2. Декомпозиційна крива часового ряду значення PM10 за даними станції № 650 мережі громадського моніторингу атмосферного повітря EcoCity за 2020—2023 рр. [9]

Подібний графік вже будував А. В. Лосенко (теж — співавтор програми [10]) у своїй статті [6], на основі суттєво нелінійного вигляду якого він зробив висновок про можливу наявність багатьох видів сезонності, а потім пропонував шукати їхні параметри експертним шляхом. Однак, перспективнішим є пошук цих параметрів в автоматичний спосіб. Для цього пропонується такий алгоритм:

1) згладити декомпозиційну криву;

2) знайти усі номери кроків, яким відповідають локальні максимуми, тобто місця, де зростання значення $S_Y(P)$ змінюється на зменшення, це і будуть періоди P_i, $i = 1,\ldots,\phi$

$$P_i : S'_Y(P_i) = 0, \;\; S''_Y(P_i) < 0, P_i, \;\; i = 1,\ldots,\phi; \tag{6}$$

3) дослідити різні варіанти поєднання цих видів сезонності — по одному, усі разом, за моделлю (1) або (2) і з різними параметрами.

Етап 2. *Визначення параметрів видів сезонності.*

Значний досвід авторів протягом 2020—2022 років роботи з моделлю Prophet показав, що у випадку, коли у моделі поєднується декілька видів сезонності з різним періодом, для мінімізації ризику перенавчання ефективними є такі прийоми:

1) сезонність з найбільшим періодом варто описувати рядом Фур'є найменшого порядку $n_{\min}$ і

з найбільшою регуляризацією $R_{s\min}$ (найменше значення саме забезпечує найбільші обмеження на коефіцієнти, тобто — найбільшу регуляризацію), тобто вона має бути якнайгладкішою;

2) сезонність з найменшим періодом варто описувати рядом Фур'є найбільшого порядку $n_{\max}$ і з найменшою регуляризацією $R_{s\max}$, тобто вона має бути найнегладкішою, тобто якнайкраще наблизитися до ряду;

3) порядок n_i ряду Фур'є варто брати між $n_{\min}=2$ і $n_{\max}=12$, інші межі (наприклад, 0, 1 чи 20) менш ефективні;

4) регуляризацію сезонностей варто брати між $R_{s\min}=0{,}1$ і $R_{s\max}=0{,}5$, інші межі (до прикладу, 0,01, 0,05 чи 0,9) менш ефективні;

5) у разі вибору між різними варіантами сезонних складових першими слід враховувати складові з більшим періодом, послідовно додаючи ті, в яких він менший.

Ці прийоми дають можливість вивести універсальні співвідношення для розрахунку параметрів усіх видів сезонності. Зокрема, пропонуються такі нові співвідношення для i-ї складової, враховуючи, що період P_0 є найбільшим значенням

$$n_i=\frac{n_{\min}-n_{\max}}{\phi-1}i+n_{\max};\tag{7}$$

$$R_{si}=\frac{R_{s\max}-R_{s\min}}{\phi-1}i+R_{s\min},\quad i=0,\ldots,\phi-1.\tag{8}$$

Пропонується досліджувати різні варіанти врахування сезонності:

– по одній, за чергою;

– 0 і 1; 0, 1, 2; … усі разом;

– адитивно (див. (1)) чи мультиплікативно (див. (2)).

Порівняння похибок чи метрик дозволить точніше ідентифікувати структуру моделі.

Етап 3. *Визначення параметрів аномалій.*

Для забезпечення більшої універсальності методу пропонується спрощення формули (3) шляхом відмови від різних видів аномалій. До того ж, пропонується під видами аномалій розрізняти не різні за формалізацією чи природою аномалії, а аномалії у різні періоди ряду

$$y(t)=g\left(R_g,t\right)+\prod_{i=1}^{\phi}s\left(R_{si},P_i,n_i,t\right)+h\left(\bigcup_{i=1}^{\phi}\bigcup_{j=1}^{\phi}H_{ij}\left(R_h,t\right)\right)+\varepsilon_t,\tag{9}$$

тобто для кожного періоду, визначеного на попередньому етапі, пропонується робити послідовні вибірки по одному періоду, скільки їх є у ряді (1-й, 2-й, 3-й…, останній може бути неповним) і в кожному окремо шукати аномалії, а потім усі об'єднувати в єдину множину. Всі аномалії брати з нульовим вікном, тобто вважати, що вони не впливали до чи після, а якщо впливали, то це — інші аномалії, які треба враховувати окремо. Регуляризацію для усіх аномалій пропонується брати однакову R_h, наприклад, 1 або 10, або враховувати як гіперпараметр.

Більше того, потрібний критерій для перевірки чи дійсно вибрані у такий спосіб точки є аномаліями. Тому пропонується для кожної вибірки для другої складової виразу (9) перевіряти нульову гіпотезу:

H_0 — нульова гіпотеза про те, що відібрана множина значень не є аномаліями відносно інших значень вибірки;

H_1 — альтернативна гіпотеза про те, що відібрана множина значень є аномаліями відносно інших значень вибірки.

Для перевірки можна використати, наприклад, такі статистичні тести:

– тест Вілкоксона — непараметричний статистичний критерій для порівняння медіан двох вибірок [16];

– тест Манна–Уїтні — непараметричний статистичний критерій для оцінювання наявності різниці між двома вибірками [17].

На рис. 3 подано приклад застосування цих критеріїв для аналізу даних.

У разі, якщо нульова гіпотеза підтверджується, то варто пробувати інший метод пошуку аномалій, або відкидати результат, тобто не враховувати знайдені точки як аномальні. Якщо ж гіпоте-

```
Segment [1014:1081]
--------------------------
mannwhitneyu
Статистика тесту: 192.0
P-значення: 0.0992
Гіпотеза підтвердилась.
--------------------------

ranksums
Статистика тесту: 1.6681
P-значення: 0.0953
Гіпотеза підтвердилась.
```

а)

```
Segment [0:506]
--------------------------
mannwhitneyu
Статистика тесту: 12025.0
P-значення: 0.0
Гіпотеза не підтвердилась.
Вибірка може вважатися аномальною.
--------------------------

ranksums
Статистика тесту: 7.9328
P-значення: 0.0
Гіпотеза не підтвердилась.
Вибірка може вважатися аномальною.
```

б)

Рис. 3. Приклад застосування критеріїв Манна–Уїтні та Вілкоксона для перевірки нульової гіпотези про те, що відібрані точки не є аномальними: *а* — нульова гіпотеза підтвердилась; *б* — нульова гіпотеза не підтвердилась

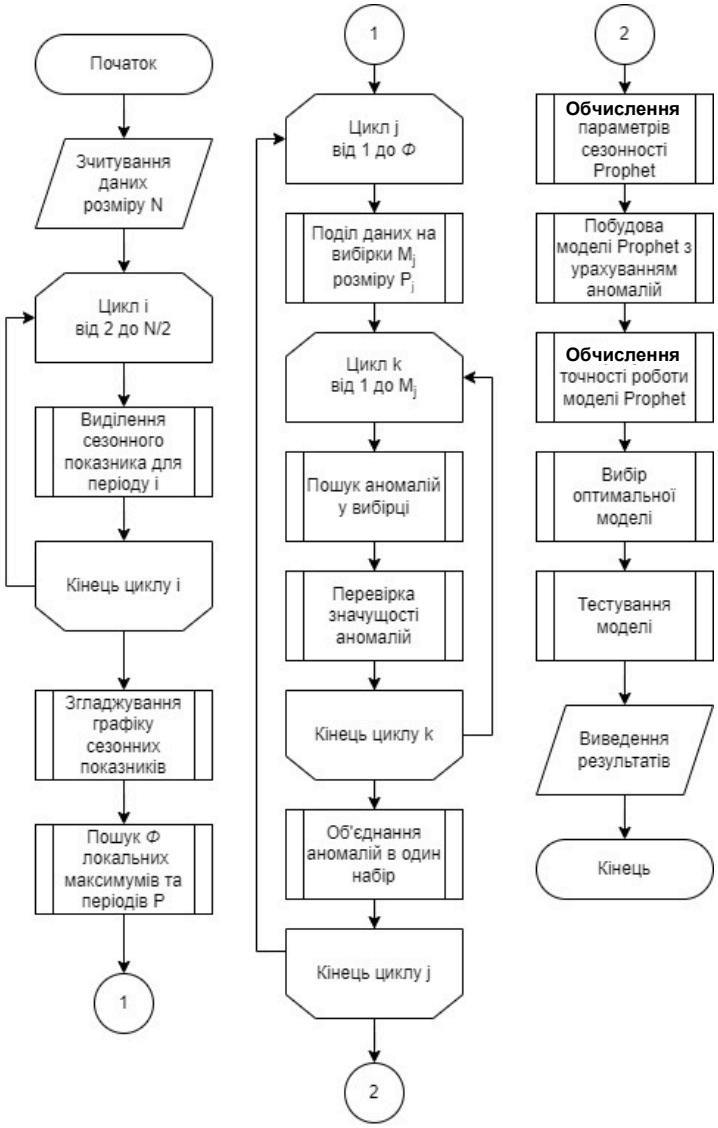

Рис. 4. Блок-схема алгоритму запропонованого методу ідентифікації параметрів гармонік та аномалій періодичного часового ряду на основі адаптивної декомпозиції

за не підтверджується, тоді аномалії слід враховувати у третій складовій $h(\bullet)$ формули (9).

Етап 4. *Ідентифікація гіперпараметрів.*

На етапах 1—3 визначаються більшість параметрів, але частина має залишатись гіперпараметрами для формування множини можливих моделей і вибору оптимальної за точністю. Критерієм точності може бути відносна похибка або відомі метрики: середня абсолютна похибка (англ. «Mean absolute error» або скорочено — MAE), середньоквадратична похибка (англ. «Mean squared error» або скорочено — MSE) та коефіцієнт детермінації (зазвичай, позначається як «r2_score», або «R2» чи «r2») [18].

Гіперпараметрами моделі (9) можуть бути такі:

– регуляризація тренду R_g (має бути сильніша — на рівні $R_{s\min}$);

– кількість ϕ видів сезонності, які варто брати до уваги з числа відібраних на етапі 1;

– граничні значення порядку Фур'є $n_{\min}$ і $n_{\max}$ та регуляризації $R_{s\min}$ і $R_{s\max}$;

– регуляризація аномальної складової R_h.

Якщо ж усі їх задати числами, тоді варто вибирати між моделями, які враховують різну кількість комбінацій видів сезонності.

Пропонується назвати цей метод — методом ідентифікації параметрів гармонік та аномалій періодичного часового ряду на основі адаптивної декомпозиції. «Гармонік», оскільки він передбачає формалізацію періодичних складових саме рядами Фур'є, тобто — рядами пар гармонік. А «адаптивної декомпозиції» — тому, що ключовим моментом методу є адаптація до кількості видів сезонності та визначення їхнього періоду по декомпозиційній кривій.

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

Розглянемо приклад.

Приклад розв'язання задачі

Для визначення ефективності запропонованого методу використано дані якості атмосферного повітря, надані мережею громадського моніторингу EcoCity (https://eco-city.org.ua/) у межах міжнародної програми «Чисте повітря для України». Дані для дослідження безпосередньо отримані за допомогою сервісу «Кабінет дослідника», до якого автори мають доступ, завдяки угоді між EcoCity і ВНТУ. «Кабінет дослідника» — це веб-система, яка дозволяє отримати доступ та використовувати у своїх дослідженнях інформацію, отриману від станцій моніторингу атмосферного повітря. На рис. 5 показано приклад інтерфейсу веб-сайту EcoCity, де зображена одна зі станцій моніторингу, встановлена в смт Турбів Вінницького району.

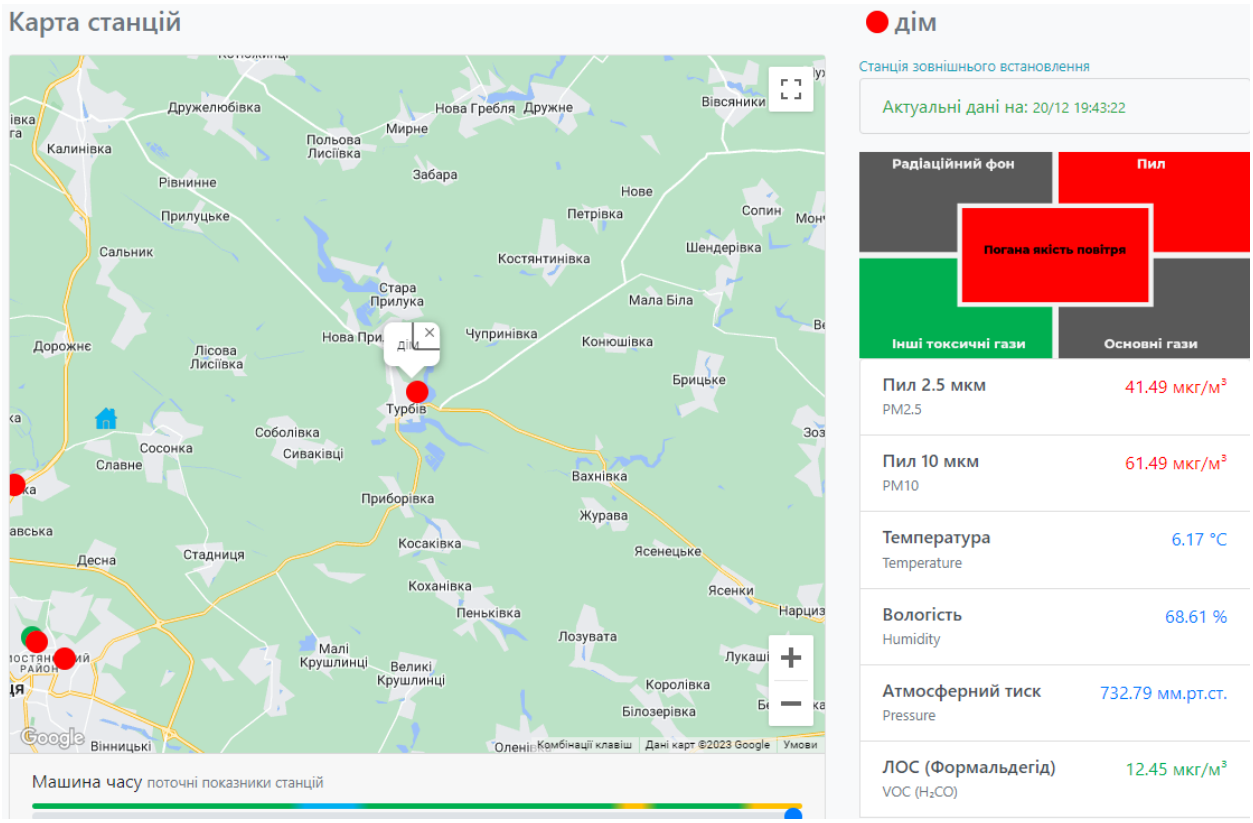

Рис. 5. Станція моніторингу якості атмосферного повітря

Кожна станція, в залежності від комплектації, дозволяє визначати та передавати різні параметри якості атмосферного повітря. Для тестування запропонованого методу вибрано станцію, розташовану у смт Турбів Вінницького району, та її показник пилу PM10 (мікроскопічні тверді частинки) за 3 роки (рис. 6).

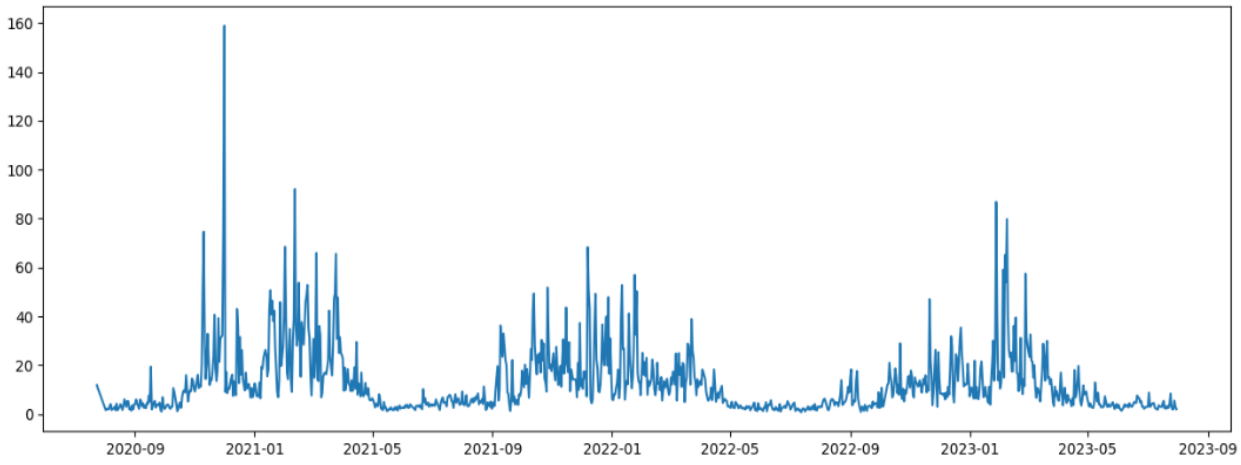

Рис. 6. Дані показника PM10 станції у смт Турбів Вінницького району, надані мережею громадського моніторингу якості атмосферного повітря EcoCity (https://eco-city.org.ua/)

Застосуємо запропонований метод.

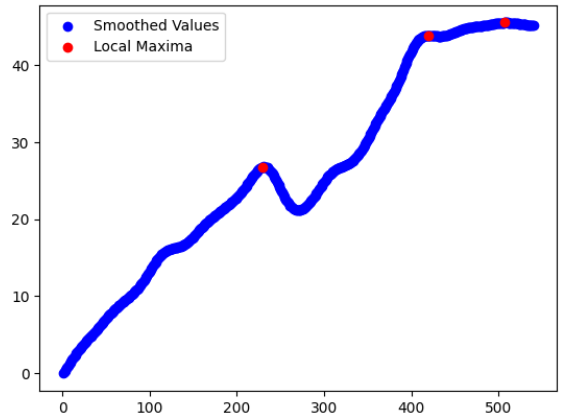

Рис. 7. Локальні максимуми для графіка з рис. 2, згладженого методом Баттерворта

```
period: 507, prior_scale: 0.1, fourier_order: 2
period: 419, prior_scale: 0.3, fourier_order: 7
period: 230, prior_scale: 0.5, fourier_order: 12
```

Рис. 8. Налаштування сезонності для моделі Prophet

Eman 1. Декомпозиційний графік для ряду з рис. 6 показаний на рис. 2. На рис. 7 показано результат його згладжування фільтром Баттерворта за використання Python-бібліотеки scipy.signal. Пошук локальних максимумів методом find_ peaks бібліотеки scipy.signal дозволив з'ясувати, що $\phi = 3$: $P_1 = 230$, $P_2 = 419$, $P_3 = 507$ діб.

Eman 2. Визначимо параметри сезонності з трьома різними періодами. Використовуючи формули (7) та (8), обчислимо інші параметри цих сезонних складових (рис. 8).

Eman 3. Визначаємо параметри аномалій. Вхідний набір даних розділяється на рівні частини, де розмір частини відповідає визначеному на етапі 1 періоду. І ця операція повторюється для кожного виду сезонності окремо. Розглянемо, до прикладу, сезонну складову з періодом P_2. Оскільки $P_2 = 419$, а загальна кількість даних становить 1080, то отримуємо дві вибірки розміром 419 та одну неповну вибірку з 186 даними. Для кожної вибірки, застосовуючи метод Isolation Forest на основі дерев рішень, визначаємо аномалії, а далі перевіряємо нульову гіпотезу про те, що ці значення не є аномальними. В нашому випадку для перших двох вибірок нульова гіпотеза не підтвердилась, тому вважаємо аномалії на цих двох вибірках статистично значущими. Як зазначено вище, ці аномалії додаються в загальну вибірку усіх аномалій для відповідної періодичної складової. Результат пошуку аномалій показано на рис. 9, де червоними точками позначені відповідні елементи графіка показника PM10, ідентифіковані як аномалії. Для частини знайдених потенційних аномалій нульова гіпотеза підтвердилась, а тому вони не додані до множини аномальних.

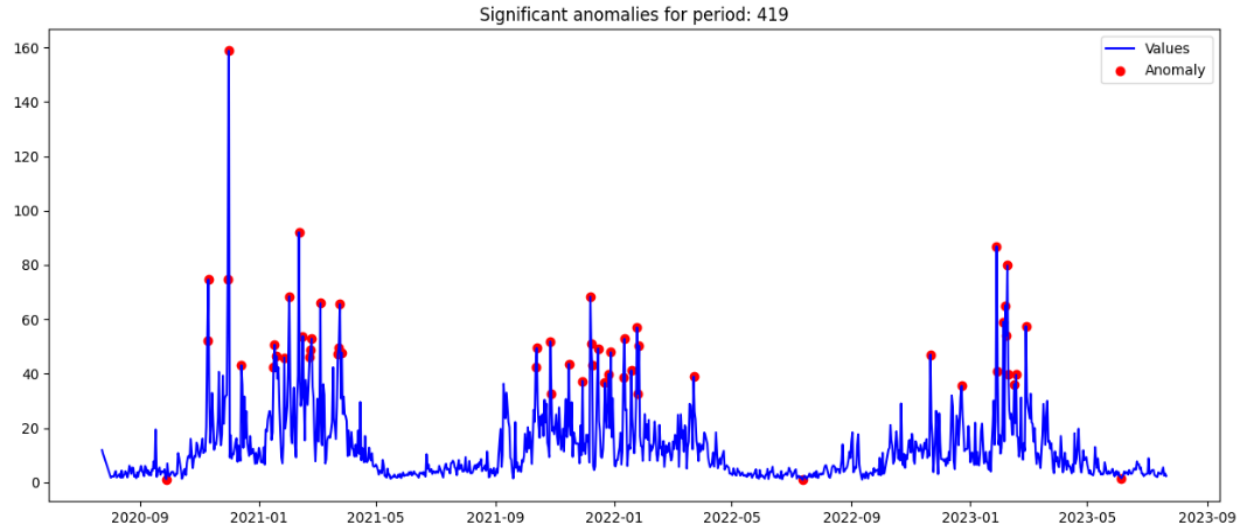

Рис. 9. Результат пошуку аномалій показника PM10 для періоду 419

Eman 4. Сформувавши усі параметри на попередніх етапах, будуємо модель FB Prophet та, використовуючи метрики, зазначені раніше, обчислюємо точність роботи моделі FB Prophet для різних варіантів (рис. 10):

Варіант 0 «Without default seasonality». Модель FB Prophet з параметрами за замовчуванням, але без сезонних складових.

Варіант 1. Варіант 0 з 3-ма видами сезонності (англ. «default seasonality»), які ідентифікуються за замовчуванням (добова, тижнева та річна).

Варіант 2. Варіант 0 з 3-ма визначеними на етапі 1 сезонними складовими одночасно ($\phi = 3$ з

періодами $P_1 = 230$; $P_2 = 419$; $P_3 = 507$ діб) (англ. «custom seasonality»), але без аномалій («no holidays», оскільки в моделі FB Prophet аномалії визнано умовно називати «свята»).

Варіант 3. Варіант 2 з усіма аномаліями («all anomalies»), визначеними у вибірках, розміром 230 діб.

Варіант 4. Варіант 3, але тільки зі статистично значущими аномаліями (англ. «significant anomalies»).

Варіант 5. Варіант 2 з усіма аномаліями, визначеними у вибірках, розміром 419 діб.

Варіант 6. Варіант 5, але тільки зі статистично значущими аномаліями.

Варіант 7. Варіант 2 з усіма аномаліями, визначеними у вибірках, розміром 507 діб.

Варіант 8. Варіант 7, але тільки зі статистично значущими аномаліями.

	name	r2	mean_squared_error	mean_absolute_error
0	Without default seasonality	0.264846	126.149253	7.053936
1	Default seasonality	0.401954	102.622188	6.125823
2	Custom seasonality + no holidays	0.431548	97.543824	5.901587
3	Period: 230 + all anomalies + custom seasonality	0.673193	56.078709	4.751458
4	Period: 230 + significant anomalies + custom s...	0.673193	56.078709	4.751458
5	Period: 419 + all anomalies + custom seasonality	0.699386	51.584003	4.647670
6	Period: 419 + significant anomalies + custom s...	0.699386	51.584003	4.647670
7	Period: 507 + all anomalies + custom seasonality	0.651996	59.715948	4.904736
8	Period: 507 + significant anomalies + custom s...	0.669741	56.671101	4.824506

Рис. 10. Таблиця порівняння точності апроксимації моделі FB Prophet для різних варіантів її архітектури (кількості сезонних складових) і параметрів

З рис. 10 видно, що для двох із трьох видів сезонності точність між статистично значущими аномаліями та усіма наявними аномаліями однакова. Причиною цього стало те, що перевірка нульової гіпотези визначила на усіх вибірках цих періодів аномалії суттєвими, відповідно для них ці два набори аномалій — абсолютно однакові. В той же час, для періоду 507 діб перевірка нульової гіпотези відкинула деякі з аномалій і в результаті це дозволило досягти кращої точності роботи моделі Prophet.

Отже, оптимальною моделлю є модель зазначена як Варіант 5 — модель для періоду 419 діб зі статистично значущими аномаліями. З порівняння точності оптимальної моделі з точністю моделі FB Prophet з параметрами і видами сезонності за замовчуванням (добова, тижнева та річна) видно, що оптимальна модель є значно точнішою, зокрема, за метрикою R2 (чим ближче до 1, тим — краще) — в 1,7 рази більше, а за метрикою MSE (чим ближче до нуля, тим — краще) — у 2 рази менше. Результат роботи оптимальної моделі показано на рис. 11.

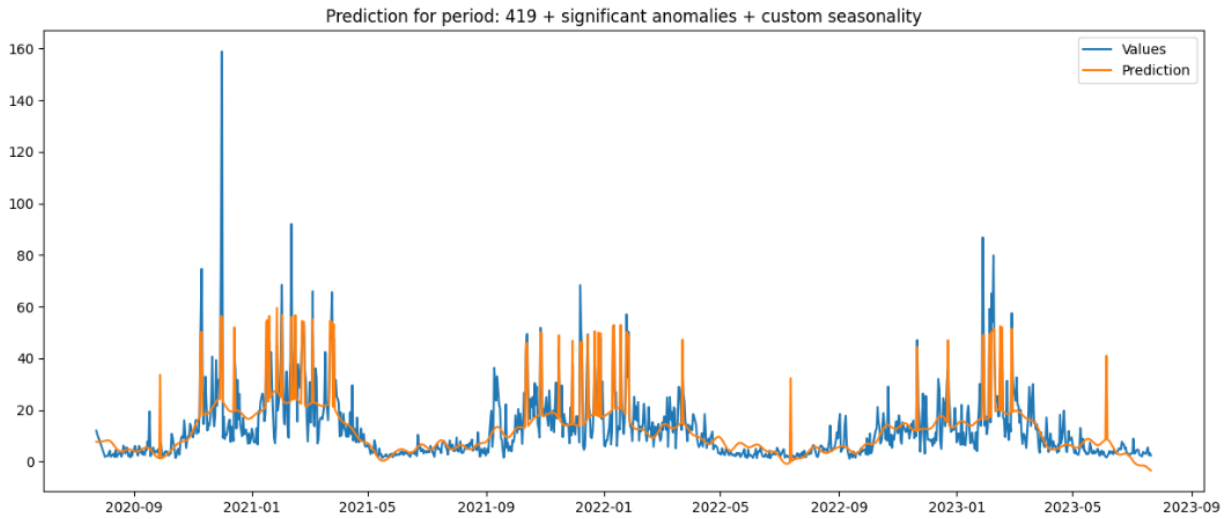

Рис. 11. Результат роботи оптимальної моделі FB Prophet, ідентифікованої за запропонованим методом

Усі результати розрахунку на Python доступні у публічній програмі на платформі Kaggle [9].

Висновки

Запропоновано новий метод ідентифікації параметрів гармонік та аномалій періодичного часового ряду на основі адаптивної декомпозиції. Описано алгоритм роботи цього методу та на прикладі показано результат його роботи.

Для тестування використовувалися реальні дані показника пилу PM10 зі станції моніторингу якості атмосферного повітря, розташованої у смт Турбів Вінницькій області, від мережі громадського моніторингу EcoCity спільно з міжнародною програмою «Чисте повітря для України». Тестування показало, що використання запропонованого методу дозволили покращити точність апроксимації оптимальної моделі на основі Facebook Prophet за метрикою R2 — у 1,7 рази, а за метрикою MSE — у 2 рази у порівнянні з моделлю Prophet з параметрами і видами сезонності за замовчуванням.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Robert Shumway, and David Stoffer, *Time Series Analysis and Its Applications With R Examples*, 2011 https://doi.org/10.1007/978-1-4419-7865-3 .

[2] Б. І. Мокін, О. Б. Мокін, і В. Б. Мокін, *Методологія та організація наукових досліджень*, підруч., 3-е вид., змін. та доп. Вінниця: ВНТУ, 2023, 230 с. ISBN 978-617-8163-01-3 (електр. ресурс).

[3] Terence C. Mills, *ARMA Models for Stationary Time Series, Chapter 3*, 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 .

[4] Sean J Taylor, and Benjamin Letham, "Forecasting at scale," *Peer J. Preprints*, 5, 2017. https://doi.org/10.7287/peerj.preprints.3190v2 .

[5] 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, 2020. https://doi.org/10.31649/1997-9266-2020-151-4-15-24 .

[6] А. В. Лосенко, «Інформаційна технологія прогнозування часового ряду кількості хворих на коронавірус на основі моделі Facebook Prophet,» *Вісник Вінницького політехнічного інституту*, № 5, с. 50-59, 2023. https://doi.org/10.31649/1997-9266-2023-170-5-50-59 .

[7] 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 .

[8] 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 .

[9] 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.

10. 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.

[11] V. Aggarwal, V. Gupta, P. Singh, K. Sharma, and N. Sharma, "Detection of Spatial Outlier by Using Improved Z-Score Test," *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, India, 2019, pp. 788-790. https://doi.org/10.1109/ICOEI.2019.8862582 .

[12] H. P. Vinutha, B. Poornima, and B. M. Sagar, "Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset," S. Satapathy, J. Tavares, V. Bhateja, J. Mohanty, Ed. *Information and Decision Sciences. Advances in Intelligent Systems and Computing*, vol. 701, 2018, Springer, Singapore. https://doi.org/10.1007/978-981-10-7563-6_53 .

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

[14] Salima Omar, Md Ngadi, Hamid Jebur, and Salima Benqdara, "Machine Learning Techniques for Anomaly Detection: An Overview," *International Journal of Computer Applications*, vol. 79, no. 2, 2013. https://doi.org/10.5120/13715-1478 .

[15] В. Б. Мокін, і А. В. Лосенко, «Інформаційна технологія короткострокового прогнозування кількості нових хворих на коронавірус на основі моделі Facebook Prophet. Інформаційно-комунікаційні технології для перемоги та відновлення,» у *Колективна монографія за матеріалами XXII Міжнародної науково-практичної конференції «Інформаційно-комунікаційні технології та сталий розвиток»* (Київ, 14-15 листопада 2023 р.), С. О. Довгий. Ред. Київ, Україна: ТОВ «Видавництво «Юстон», 2023, с. 27-30.

[16] R. F. Woolson, *Wilcoxon Signed-Rank Test. In Wiley Encyclopedia of Clinical Trials*, R. B. D'Agostino, L. Sullivan and J. Massaro, Ed., 2008. https://doi.org/10.1002/9780471462422.eoct979 .

[17] P. E. McKnight, and J. Najab, "Mann-Whitney U Test," In *The Corsini Encyclopedia of Psychology* , I. B. Weiner and W. E. Craighead, Ed., 2010. https://doi.org/10.1002/9780470479216.corpsy0524 .

[18] *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 13.12.2023

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

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій, e-mail: vbmokin@vntu.edu.ua

D. O. Shmundiak[1]
V. B. Mokin[1]

Method of Harmonics Parameters Identification and Anomalies of a Periodic Time Series Based on Adaptive Decomposition

[1]Vinnytsia National Technical University

Periodic time series have many applications — financial indicators, indicators of air quality, indicators of the state of water, etc. Accordingly, simulation of time series and pattern analysis are relevant and quite common tasks for understanding possible trends and changes for correct and timely actions. Important parameters of periodic time series are their trends, seasonal components, and anomalies. There exist numerous methods to determine the trend of a time series, but when it comes to the simultaneous identification of parameters of various types of seasonality and anomalies of different nature in different periods, this task is not trivial and there is no universal solution for this problem. Most of the solutions are specific to a specific subject area or demonstrate insufficient adequacy and accuracy of approximation.

New method of identifying parameters of harmonics and anomalies of a periodic time series, based on the adaptive decomposition of the series, has been developed. It is proposed to decompose a given time series with a period up to half of the total number of time series records and to plot the ratio of the amplitudes of the seasonal component to the amplitudes of the series itself — the so-called "decomposition curve". Then, smooth this curve and find local maxima, which are proposed to be considered as corresponding to the period of possible types of seasonality of the series. Considering many years of experience using the Facebook Prophet model, a set of relations between values of the seasonality period, the order of the Fourier series for its approximation, and the degree of regularization that should be taken into account are proposed. For each type of seasonality in each period, one of the known methods should be used to find anomalous data and check their statistical significance. Statistically significant anomalies are collected in a combined set with typical parameters. A few possible variants of the structures of such time series models are proposed. The algorithm of the method is developed, and its main components are described.

The offered method was tested in Python in the notebook of the Kaggle platform. This notebook uses the Facebook Prophet model on real data of air quality observations obtained from one of the EcoCity public monitoring network stations within the international program "Clean Air for Ukraine". Tests showed that compared to the model with default parameters and default parameters of seasonality, the optimal model of the proposed method improved the accuracy of the approximation for the R2 metric — by 1,7 times, and for the MSE metric — by 2 times. This confirms the effectiveness of the offered method.

Keywords: time series analysis, simulation, machine learning, time series anomalies, seasonality, Fourier series harmonics, air quality, EcoCity.

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

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technologies, e-mail: vbmokin@vntu.edu.ua