ЕКОЛОГІЯ ТА ЕКОЛОГІЧНА БЕЗПЕКА

https://doi.org/10.31649/1997-9266-2020-151-4-15-24

УДК 004.9+556

В. Б. Мокін[1]
О. В. Слободянюк[1]
О. М. Давидюк[1]
Д. О. Шмундяк[1]

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ПОШУКУ МОЖЛИВИХ ДЖЕРЕЛ ПІДВИЩЕНОГО ЗАБРУДНЕННЯ РІЧКИ З ВИКОРИСТАННЯМ МОДЕЛІ PROPHET

[1]Вінницький національний технічний університет

Зміни клімату зумовили низку маловодних років та, відповідно, зменшення кількості води для розбавлення антропогенних забруднень. Отже, актуальнішими стають дослідження, спрямовані на виявлення основних джерел забруднення з метою їх негайного регулювання. Більше того, відповідно до Водної рамкової директиви ЄС, яку, згідно з Угодою про асоціацію з ЄС, зобов'язана виконувати й Україна, необхідно найближчим часом виробити комплекс дій для досягнення чи стабілізації не нижче доброго екологічного стану в усіх масивах вод. В Україні, як і в багатьох інших країнах Європи, система моніторингу якості вод не забезпечує достатньої кількості даних регулярних спостережень для локалізації у просторі та часі усіх, у т.ч. незареєстрованих, джерел підвищеного забруднення, що ускладнює реалізацію політики їх регулювання. Отже, важливо створити інформаційну технологію пошуку можливих джерел підвищеного антропогенного навантаження на річку за даними регулярних спостережень якості води у басейні заданої річки. Проведений аналіз показав, що таким даним властива зміна періодичності спостережень (особливо у довгостроковій перспективі у десятки років), існує практика одномоментних спостережень (раз на квартал чи півроку, щоразу в різний час), багато пропущених даних тощо, що унеможливлює застосування типових для подібних задач множинних регресій та моделей часових рядів на основі авторегресії та проінтегрованого ковзного середнього (АРПКС). Запропоновано використовувати модель та пакет програм Prophet компанії Facebook для R та Python, яка позбавлена усіх зазначених недоліків і є оптимальною для розв'язання поставленої задачі. Розроблено та охарактеризовано методологію її застосування, яка полягає в моделюванні даних моніторингу з фільтруванням різних видів сезонності та виділенні лінійного тренду між точками зміни, перше наближення кожної з яких задається на початку інтервалів в один чи декілька років, залежно від кількості наявних даних. Ідентифіковані тренди між цими точками зіставляються за різними показниками на кожному посту спостережень та за спеціально розробленим алгоритмом виявляються найбільші прирости трендів («імпульси»), які потім спричиняють монотонне наростання забруднення аж до сьогодні. Виявлені дати таких «імпульсів» масштабуються та агрегуються за різними показниками, що дозволяє на кожній ділянці між постами визначити дату появи джерела забруднення і потім, за іншими даними, із залученням відповідних контролюючих служб, точніше ідентифікувати джерело підвищеного забруднення річки у певний час. Розроблено програму на Python, на якій перевірена працездатність створеної технології виявляти такі «імпульси» на прикладі ділянки р. Південний Буг від витоку до м. Вінниця за даними державної системи моніторингу якості вод за 2002—2019 роки та подано успішні результати її роботи.

Ключові слова: інформаційна технологія, якість води, часовий ряд, модель Prophet, джерело забруднення річки, Python.

Вступ

Згідно з Водною рамковою директивою, одне з ключових завдань політики управління водними ресурсами країн Європи, у т.ч. України, є досягнення доброго екологічного стану в усіх маси-

вах вод [1]. План дій передбачає проведення моніторингу, аналіз сучасного стану, розроблення програми заходів для кожного масиву вод для досягнення або стабілізації не нижче як доброго екологічного стану вод. Але ключовим етапом є, передусім, виявлення найпроблемніших ділянок та причин неможливості досягнення такого стану у найближчій перспективі. А в цій задачі, у свою чергу, головним є аналіз наявних даних моніторингу та виявлення джерел антропогенного забруднення. Останнім часом, ця проблема набуває все більшого значення в Україні, через маловоддя. Адже, мала кількість опадів в останні 5 років спричиняє збільшення концентрації забруднень ще й через природні обставини.

Однією з таких проблемних ділянок є масив вод, де розташований водозабір КП «Вінницяводоканал», який є основним джерелом водопостачання більшості населення м. Вінниця. Водночас, останніми роками спостерігається суттєве погіршення якості води, через те, що очисні споруди цього підприємства були розраховані на добрий стан вод у місці водозабору, а він за деякими показниками вже є незадовільним. Подібна проблема є і в інших населених пунктах, де джерелом постачання є саме поверхневі води. Однак, проблема м. Вінниця ускладнюється ще й тим, що досі немає офіційного роз'яснення і єдиної консолідованої думки, що ж є джерелом такого погіршення стану вод.

Державне агентство водних ресурсів України виклало у вільний доступ дані Державного моніторингу якості вод України на Єдиному державному веб-порталі відкритих даних [2] та на вебресурсі «Моніторинг та екологічна оцінка водних ресурсів України» [3], де наразі доступні дані по постах за 1920—2020 роки (по різних постах — різна періодичність та тривалість спостережень).

Отже, є актуальним створення інформаційної технології, яка дозволить за даними регулярних спостережень якості вод максимально локалізувати джерела підвищеного забруднення річки в часі (хоча б з точністю до року) та просторі (хоча б з точністю до ділянки між сусідніми постами системи моніторингу).

Мета дослідження — створення інформаційної технології пошуку можливих джерел підвищеного антропогенного навантаження на річку за відкритими даними спостережень якості води.

Вибір моделі та методології розв'язання задачі

Проведений аналіз наявних відкритих даних Державного моніторингу якості вод України [2], [3] показує, що:

– ці дані мають різну періодичність;

– має місце значна кількість пропущених дат;

– періодичність у рік, квартал чи місяць означає не те, що вони усереднені за цей період, а те, що вони виміряні лише 1 раз у довільну дату протягом кожного року, кварталу чи місяця і часто геть в різну дату (в один квартал це був перший місяць, в інший – третій), час вимірювань не вказаний, тобто теж може бути різним, тобто ці дані слід моделювати з використанням не стільки детермінованих – скільки стохастичних моделей.

Крім того, реалії української системи моніторингу є такими, що пости моніторингу якості води часто розташовані досить віддалено від метеопостів та постів системи моніторингу кількості (витрат) води, що унеможливлює їх коректне зіставлення та використання моделей на основі множинної регресії. А різна періодичність унеможливлює використання класичних моделей часових рядів типу авторегресії та проінтегрованого ковзного середнього (АРПКС) та відповідних пакетів програм, основаних на застосуванні моделей ARIMA (англійський варіант назви АРПКС).

Все це не применшує цінність наявних відкритих даних про якість вод, але вимагає пошуку досконаліших і ефективніших методів їх аналізу. Пропонуємо використовувати модель Prophet, реалізовану у пакеті програм Prophet компанії Facebook для R та Python. Останнім часом вона стає дедалі більш поширеною, завдяки своїм унікальним можливостям [4]:

– працює і з періодичними (сезонними), і з неперіодичними рядами даних;

– допускає чималу кількість пропущених даних;

– дає можливість додавати довільні складові сезонності (період задається у днях, але допускаються дробові числа, наприклад, 36,525 тощо), які моделюються, як правило, рядами Фур'є заданого користувачем порядку і можуть оброблятись як адитивні або як мультиплікативні складові;

– враховує «holidays» (свята) і вікна їх впливу, тобто дати аномальних значень (на прикладі впливу на продажі) і те за скільки днів до вказаної дати вже починається їх вплив та за скільки днів потому зникає;

– використовує і лінійну, і логістичну модель тренду на різних інтервалах, які обмежуються точка-

ми зміни тренду (changepoints), що можуть задаватись як кількістю цих точок, так і у вигляді набору дат точок, але вони використовуються як початкове наближення, яке потім може зазнавати змін;

– здійснює прогнозування даних на заданий інтервал з побудовою зони невизначеності, чим більше тренд усього ряду відхиляється від прямої, тим більшою є ця зона невизначеності та меншим є можливий інтервал прогнозування.

Звичайно, ця модель не позбавлена і недоліків, наприклад, вона не дозволяє застосовувати експоненціальне згладжування даних у вибраному ковзному вікні, як це може робити, наприклад пакет програм forecast в R. Але для поставленої задачі цей недолік не є суттєвим, а важливими є саме переваги моделі Prophet та пакету програм її реалізації.

Пропонується така методологія розв'язання задачі з використанням пакету програм Prophet:

1. Аналізуємо реальні дані про спецводокористування та шукаємо на яких ділянках між постами спостереження та в які роки потужні джерела були введені в експлуатацію чи почали нарощувати забруднюючий вплив за певними показниками якості води, згідно з сумарними за рік офіційними даними — ці дані використаємо і як тренувальні для налаштування параметрів моделі Prophet, і як контрольні, для перевірки ефективності роботи методу.

2. Вибираємо параметри моделі Prophet (варіанти сезонності, порядок коефіцієнтів рядів Фур'є для їх описання) та ін.

3. Встановлюємо точки зміни тренду (CP – англ. «changepoints») на 1 січня кожного року і будуємо лінійну модель Prophet між цими CP для різних показників якості води по різних постах спостережень окремо, таким чином, щоб вона дозволяла виявляти початок забруднюючого впливу джерел із тренувальної вибірки (див. п.1). Цей вплив визначаємо таким чином:

3.1. Аналізуємо нахил тренду для кожного такого випадку (між кожними CP для кожного показника на кожному посту) та шукаємо серед них найбільші прирости, тобто найбільші нахили, але сходинкового типу, тобто, коли мало місце збільшення навантаження по заданому показнику, виявлене на певному посту, а потім воно не зменшувалось, а тільки наростало аж по останній день, тобто підприємство як розпочало роботу і свій забруднюючий вплив, так і продовжує по цей час. Результатом аналізу є приріст забруднення у дату, з якої все почалось (як короткий імпульс), решта приростів, де ця умова не виконується, прирівнюються до нуля.

3.2. Усім місяцям, які потрапляють в діапазон між CP, ставиться у відповідність значення цього імпульсу. Зазвичай, по різних показниках після моделювання CP будуть мати різні значення. Тому такий підхід дозволить їх зіставлення з точністю до місяця.

3.3. Масштабуємо у діапазон [0, 1] і додаємо ці виявлені забруднюючі імпульси за різними показниками по кожному посту — саме за тими показниками, за якими кожне підприємство звітує про забруднюючий вплив, відповідно до його технологічного процесу.

4. Зіставляємо виявлені дати, ділянки річки та сумарний забруднюючий вплив за різними показниками якості води з контрольними даними, відібраними у п.1, за певною метрикою. Змінюємо структуру і параметри моделі та повторюємо пп. 2, 3, поки похибка не стане прийнятною, або виконуємо повний перебір допустимих варіантів моделі (час роботи програми тут не є обмеженням) і вибираємо модель з найменшою похибкою.

Усі виявлені місця і дати забруднюючого впливу за певними показниками, які не збігаються з відомими даними про спецводокористування, можуть належати незареєстрованим джерелам скидання, а тому передаються відповідним контролюючим органам для перевірки.

Для підвищення точності роботи технології варто на етапі 2 правильно задавати параметри «holidays», які враховують вплив аномальних даних, до яких відносяться, наприклад, відомі аварії, коли погіршення якості води відбувалось на певних ділянках епізодично, але суттєво (наприклад, вихід з ладу очисних споруд на водоканалі у м. Хмільницький через надмірне забруднення від одного з підприємств міста, яке спричинило загибель бактерій в аеротенках, та ін.). І такі забруднення можуть потім осідати у водосховищах та бути відчутними ще деякий час, тобто у параметрах holidays слід правильно вказувати «вікно» часу такого впливу, що мав місце постфактум. Запропонована технологія дозволяє фільтрувати такі впливи, але досконаліше налаштування параметрів підвищує точність її роботи.

Подібний метод може бути досить неточним, враховуючи недосконалість вхідних даних, але він має низку переваг:

– має високу швидкодію, адже дозволяє достатньо швидко проаналізувати дані по усіх постах за усіма показниками якості вод заданої річки;

– адаптація до змінної водності річки, коли менший обсяг води для розбавлення забруднень

спричиняє підвищення концентрації, адже не може мати місце постійне зменшення опадів в усі місяці багато років підряд;

– є високоефективним, оскільки дає результат за вже відомими історичними (ретроспективними) даними.

Важливо зазначити, що запропонована технологія не розрізняє причини забруднюючих впливів (і антропогенні, і природні) і тільки зіставлення з даними з п.1 та подальше вивчення результатів технології дозволяє ідентифікувати їх природний чи антропогенний характер.

Метод є дуже чутливим і дозволяє зафіксувати, навіть, незначне підвищення тренду, але таке, що значно відрізняється від тренду у попередні моменти часу.

Ключовим у цьому алгоритмі, окрім застосування моделі Prophet та правильного налаштування її параметрів, є алгоритм виявлення так званих «patterns» (з англ. – шаблонів, ситуацій, закономірностей), тобто імпульсів, які характеризують початок появи сталого забруднення річки, тобто такого, що має місце тривалий час після появи, та актуального, тобто такого, що має місце і по цей час, тому зупинимось на ньому детальніше.

Розроблення математичного та алгоритмічного апарату виявлення моментів початку сталого актуального забруднення річки

Модель Prophet повертає початковий тренд k_0 ряду до першої точки зміни d_0 і m приростів $\Delta k_i, i=\overline{1,m}$ тренду між наступними m точками аж до кінця тренувального інтервалу даних. За необхідності, потім по цій моделі можуть передбачатись дані на подальші періоди. При цьому, дати $d_i, i=\overline{1,m}$ розташування цих точок змін вже не зазнають.

Для нас головним є знаходження моментів приростів, тому початковий тренд значення не має і слід аналізувати тільки прирости тренду відносно нього. Результатом роботи моделі Prophet є

$$\hat{y}_i = \left(k + \Delta k_i\right) d_i, \quad i=\overline{1,m}, \tag{1}$$

де $\hat{y}_i$ — стандартизоване значення тренду в i-й точці зміни d_i, тобто масштабоване між мінімальним і максимальним значеннями та поділене на середньоквадратичне відхилення — це стандартне позначення пакета програм Prophet, яке в результатах так і називається: “yhat” (рис. 1).

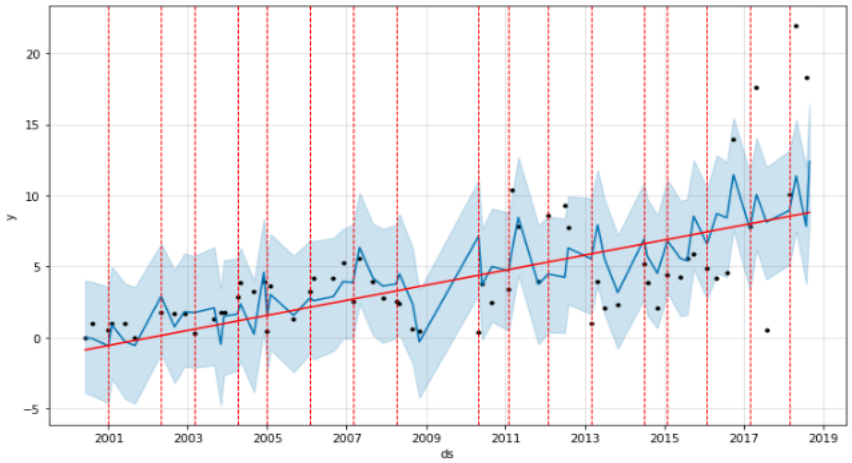

Рис. 1. Значення азоту амонійного $\hat{y}_i$ на посту «Меджибіж» за 2002—2019 рр. (чорні точки), результат роботи моделі (сині лінії), зона невизначеності моделі чи довірчий інтервал між двома обвідними (блакитного кольору), тренд (червона лінія) після фільтрування річної сезонності рядом Фур’є другого порядку між точками зміни тренду (вертикальні пунктирні лінії)

На рис. 2 показано прирости Δk_i тренду між точками зміни d_i.

Необхідно сформувати умову для виявлення ділянок, де прирости трендів є суттєвими (наприклад, більшими, ніж на 5 % від максимального значення. Як відомо, для виявлення імпульсів зазвичай використовується інтегрування. Здійснимо інтегрування, тобто знайдемо S — суми приростів тренду з накопиченням

$$S = \sum_{i=1}^{m} \Delta k_i. \tag{2}$$

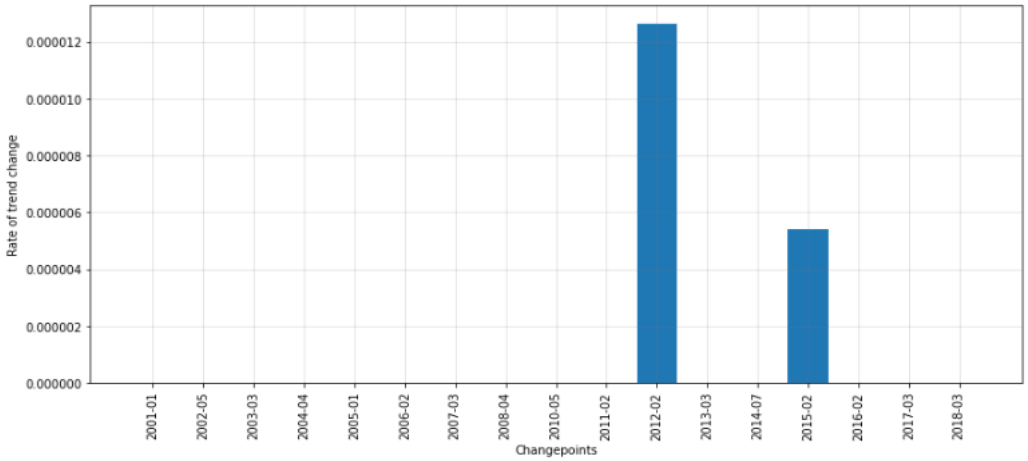

Рис. 2. Значення приростів Δk_i тренду азоту амонійного на посту «Меджибіж» за 2002—2019 рр. між точками зміни d_i

На рис. 3 показано результат оброблення даних рис. 2 за формулою (2).

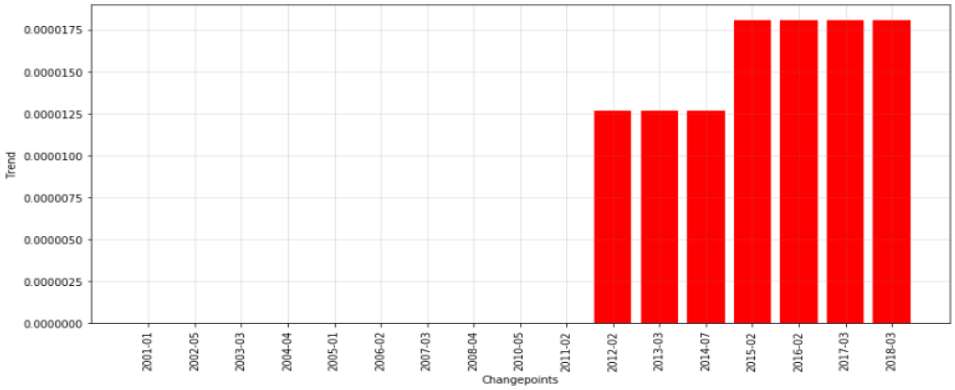

Рис. 3. Значення сум приростів тренду з накопиченням азоту амонійного на посту «Меджибіж» за 2002—2019 рр. між точками зміни d_i

Але аналіз показує, що деякі прирости Δk_i тренду можуть бути знакозмінними. Водночас, важливо знайти саме такі імпульси, вплив яких є і досі актуальним, а не епізодичним. Тобто треба знайти ситуації, коли після такого імпульсу концентрація далі весь час тільки наростала. Значення S можуть змінюватись синусоїдально і за ними важко буде однозначно виявити місця постійного наростання концентрації. Для цього пропонуємо здійснити ще одне інтегрування (рис. 4)

$$J = \sum_{j=i}^{m} S_i = \sum_{q=j}^{m} \sum_{j=i}^{m} \Delta k_{qj}, \qquad (2)$$

а тоді слід лише знайти такі імпульси на графіку на рис. 2, після яких функція J далі лише монотонно наростала.

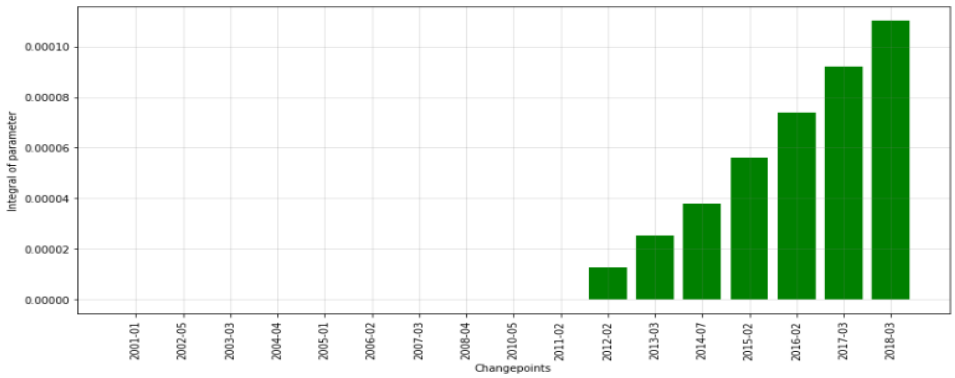

Рис. 4. Значення сум J приростів тренду з накопиченням азоту амонійного на посту «Меджибіж» за 2002—2019 рр. між точками зміни d_i

Розроблення нової інформаційної технології пошуку місць та дат можливого антропогенного навантаження на річку за даними моніторингу якості вод

Пропонуємо такий алгоритм інформаційної технології пошуку місць та дат можливого антропогенного навантаження на річку за даними моніторингу якості вод:

1. Вибрати показники якості води, з яких часто має місце перевищення гранично допустимих значень у контрольній точці та по яких немає впевненості щодо джерел їх походження.

2. Сформувати та очистити датасети по цих показниках від помилкових та пропущених даних.

3. Сформувати датасети типу «holidays» (в термінології пакету програм Prophet) з відомими датами зафіксованих залпових скидань забруднених вод, в яких можливі перевищення саме тих показників, що досліджуються.

4. Виділити з кожного датасету тренувальні дані (75…80 %), на яких будуть ідентифікуватись параметри моделі, і тестові дані, на яких буде вибиратись оптимальна модель з числа ідентифікованих на тренувальних даних.

5. Для кожного показника якості води задати початкове наближення точок зміни тренду changepoints, встановивши їх на першу дату кожного року з числа тренувальних даних.

6. Для кожного показника якості води здійснити налаштування моделі на основі технології Prophet з лінійним трендом, врахування даних типу «holidays» та врахування сезонності, яка мінімізує похибку на тестових даних. Побудувати різні моделі, варіюючи такі параметри:

6.1 Адитивний чи мультиплікативний варіант врахування тренду і сезонних компонент моделі

6.2. Різне врахування сезонності — тільки річна, яка є обов'язковою, чи ще й по порах року (період 365,25/4).

6.3. Різні значення коефіцієнтів рядів Фур'є, якими апроксимуються сезонні компоненти ряду.

6.4. Різна кількість точок зміни тренду.

7. Вибрати оптимальну за структурою і параметрами модель, ідентифіковану на етапі 6, яка забезпечить мінімальну похибку на тестових даних, відібраних на етапі 4.

8. Для кожного *i*-го показника якості води для кожної ділянки між *j*-ми постами спостережень і для усіх дат між *k*-ми точками changepoints показника якості води x_{ijk}, масштабованого по цьому показнику, визначити бальну y_{ijk} оцінку приросту тренду *x* цього показника, яка більшим значенням дає значно більшу вагу, наприклад, шляхом взяття квадрату від їх значень

$$y_{ijk} = \left(\frac{x_{ijk} - \min_i x_{ijk}}{\max_i x_{ijk} - \min_i x_{ijk}} \right)^2$$

або простіший варіант для випадку, коли мінімальне значення береться таким, що дорівнює нулю

$$y_{ijk} = \left(\frac{x_{ijk}}{\max_i x_{ijk}} \right)^2$$

1. Для кожної *k*-ї дати кожної ділянки між *j*-ми постами спостережень слід додати усі ці бальні оцінки

$$J_{jk} = \sum_i y_{ijk}$$

та ранжувати їх за зменшенням для кожної ділянки річки.

2. Проаналізувати ділянки та дати з найбільшими значеннями J_{jk} з урахуванням даних звітності водокористувачів про їх водокористування, дозволами на скидання стічних вод, результатами екоінспекційних перевірок, даними супутникових спостережень, даними громадського екологічного моніторингу та іншою інформацією (див. [5]—[7]), здійснити пошук, коли у визначені дати на визначених ділянках могли з'явитись джерела скидання вод, та рекомендувати органам Держекоінспекції та Держводагентства провести ретельнішу перевірку виявлених фактів.

Блок-схема алгоритму запропонованої інформаційної технології показана на рис. 5.

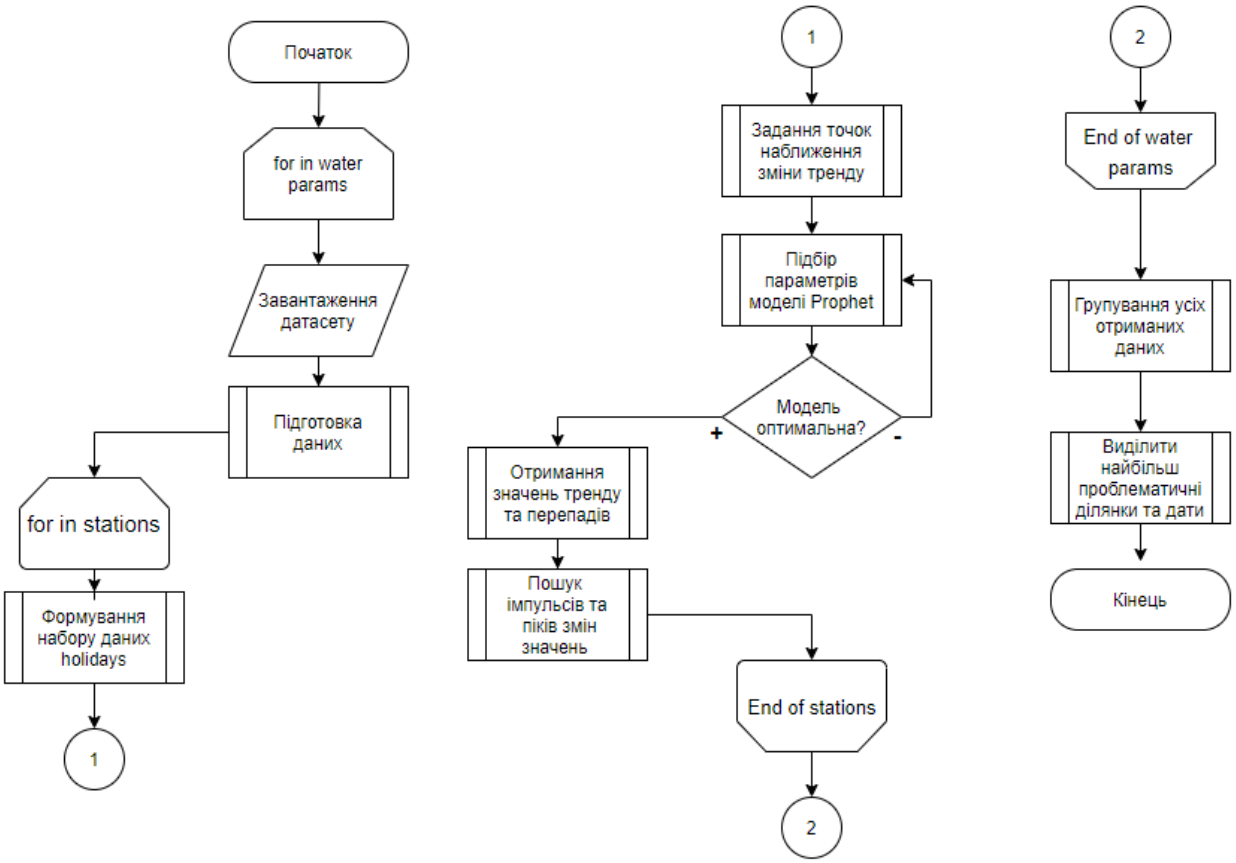

Рис. 5. Блок-схема алгоритму запропонованої інформаційної технології

Розглянемо приклад застосування запропонованої інформаційної технології.

Приклад застосування розробленої інформаційної технології на реальних даних

Працездатність запропонованої інформаційної технології перевірена на реальних даних 8 постів державної системи моніторингу вод від витоку р. Південний Буг до м. Вінниці, використовуючи дані 2002—2019 рр., за показниками $БСК_5$ (біохімічне споживання кисню за 5 днів, гранично допустиме значення 3,0 мг/л), концентрація азоту з перерахунку на нітрати NO_2 (гранично допустима концентрація 45,0 мг/л) та концентрація азоту амонійного NH_4 (2,0 мг/л), за якими останніми роками часто фіксуються наднормативні значення.

Випробовувались лише етапи 2 та 3 методології, тобто наскільки технологія дозволить ранжувати наявні забруднення для випадково вибраних параметрів. Виявлення можливих забруднювачів (тренувальна і контрольна вибірки) та пошук можливих незареєстрованих джерел забруднення за допомогою розробленої технології — це тема окремого дослідження, яке має починатися із запиту на публічну інформацію до Державного агентства водних ресурсів України щодо отримання необхідних даних зі спецводокористування.

Результат застосування розробленої технології до 8 постів ділянки р. Південний Буг від витоку до м. Вінниці за даними системи моніторингу якості вод Держводагентства за 2002—2019 рр. по одному з показників якості води наведено на рис. 6. З рисунку чітко видно на яких ділянках і в які роки почалось стале збільшення забруднення річки. Далі планується передати ці дані в Басейнове управління водних ресурсів річки Південний Буг та Держекоінспекції для порівняння з відомими даними обліку спецводокористування та звітами 2-ТП (водгосп) для прийняття відповідних рішень. Також, результати будуть винесені на розгляд басейнової ради Південного Бугу.

Окремо варто дослідити чому у 2019 р. почалось зменшення приросту забруднення, яке, перед тим, було сталим і постійно зростало. Нагадуємо, що результатом технології є бальна оцінка приросту забруднення, тобто пік меншої висоти означає, що забруднення зростало, але не так сильно, як перед тим.

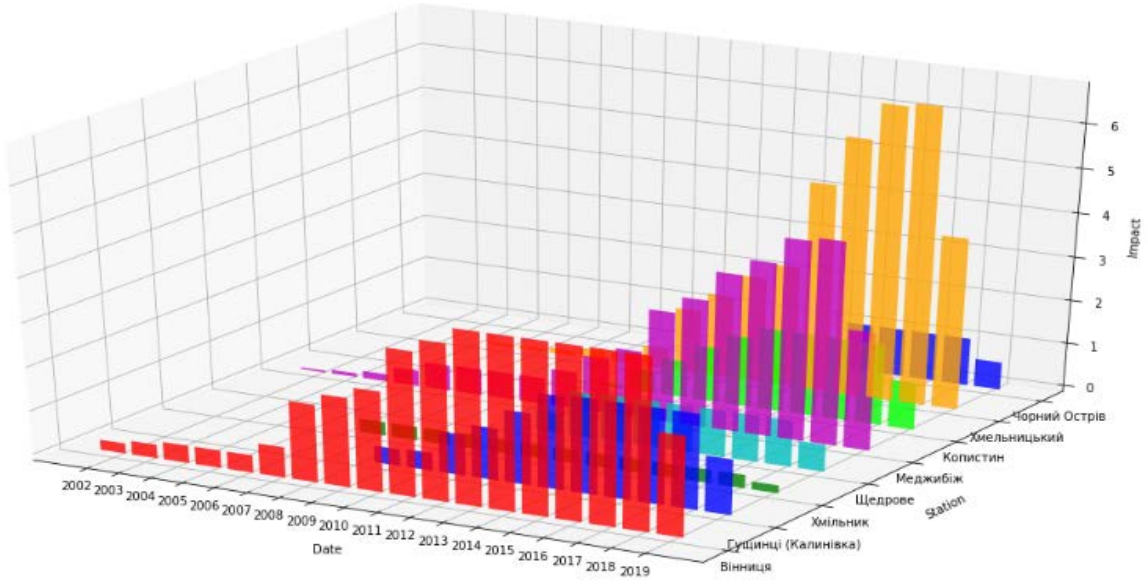

Рис. 6. Результат застосування розробленої технології до 8 постів ділянки р. Південний Буг від витоку до м. Вінниці за даними системи моніторингу якості вод Держводагентства — бальна оцінка приросту забруднення

Висновки

Розглянуто проблему виявлення основних антропогенних джерел забруднення водойм за регулярними багаторічними даними моніторингу вод України. Проаналізовано особливості відкритих даних державної системи моніторингу України за 1920—2020 рр. та обґрунтовано вибір моделі Prophet для розв'язання поставленої задачі, передусім, через її можливості роботи із зашумленими даними, які містять чимало пропусків даних та аномальних даних.

Вперше запропонована інформаційна технологія пошуку можливих джерел підвищеного антропогенного навантаження на річку за даними багаторічних спостережень якості води у басейні заданої річки з використанням моделі Prophet, основана на локалізації у часі та просторі забруднюючого впливу джерел забруднення на основі кусково-лінійної апроксимації тренду. Особливістю запропонованої технології є її стійкість до аномальних завад з різкою зміною якості чи кількості води (аварії, межені, повені тощо), які спеціальним чином враховує блок параметрів «holidays» моделі Prophet, та зміни водності річки, за рахунок аналізу багаторічних даних та виявлення сталих трендів, що мали місце тривалий час і є актуальними і до сьогодні. Технологія дозволяє виявити основні забруднювачі заданих масивів вод басейну заданої річки, що є ключовим завданням під час складання Плану управління цим басейном, у т.ч. розроблення комплексу заходів, спрямованих на досягнення чи стабілізацію доброго екологічного стану річки у довгостроковій перспективі, згідно з вимогами Водної рамкової директиви ЄС.

Розроблено й охарактеризовано математичний апарат та алгоритм запропонованої технології. Авторами статті створена програма для реалізації цієї технології на Python з використанням пакета програм Prophet компанії Facebook. Проведено її успішне випробування на ділянці р. Південний Буг на реальних даних 8 постів державної системи моніторингу вод від витоку р. Південний Буг до м. Вінниці за даними 2002—2019 рр. по показниках $БСК_5$ (біохімічне споживання кисню за 5 днів), концентрація нітратів NO_2 та концентрація азоту амонійного NH_4, за якими останніми роками часто фіксуються наднормативні значення. Результати аналізу будуть передані у відповідні установи та експертні ради для вивчення і прийняття необхідних рішень.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] *Водний Кодекс України.* Введений в дію Постановою ВР № 214/95-ВР від 06.06.95. *Відомості Верховної Ради* (ВВР), № 24, ст. 189, 1995.

[2] *Єдиний державний веб-портал відкритих даних*. Дані державного моніторингу поверхневих вод. Державне агентство водних ресурсів України. [Електронний ресурс]. Режим доступу: https://data.gov.ua/dataset/ee2bc3b0-42d4-4f19-8d96-913cd9d1f02a .

[3] *Моніторинг та екологічна оцінка водних ресурсів України.* Державне агентство водних ресурсів України, Інститут розробки інформаційних систем, 2020. [Електронний ресурс]. Режим доступу: http://monitoring.davr.gov.ua/EcoWaterMon/GDKMap/Index .

[4] *Package «Prophet». Automatic Forecasting Procedure. Version 0.6.1*, 2020-04-28. [Electronic resource]. Available: https://github.com/facebook/prophet .

[5] В. Б. Мокін, Л. М. Скорина, і А. Р. Ящолт, «Удосконалення технології аналізу даних дозвільної документації зі спеціального водокористування в системі Держводагенства,» *Вісник Вінницького політехнічного інституту*, № 4, с. 22-31, 2017.

[6] *Інструкція про порядок розробки та затвердження гранично допустимих (ГДС) речовин у водні об'єкти із зворотними водами*. Харків, Україна: УкрНЦОВ, 1994.

[7] В. Б. Мокін, і А. Р. Ящолт, *Комп'ютеризовані регіональні системи державного моніторингу поверхневих вод: моделі, алгоритми, програми*, монографія. Вінниця, Україна: ВНТУ, 2005, 78-85 с.

Рекомендовано кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 9.09.2020

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

Слободянюк Олена Валеріївна — канд. пед. наук, доцент кафедри інженерних систем у будівництві, e-mail: olenas8@gmail.com ;

Давидюк Оксана Миколаївна — аспірантка кафедри системного аналізу та інформаційних технологій, e-mail: davidyuk-ok@ukr.net ;

Шмундяк Дмитро Олександрович — студент факультету комп'ютерних систем і автоматики, e-mail: dimashmund@gmail.com

V. B. Mokin[1]
O. V. Slobodyanyuk[1]
O. M. Davidyuk[1]
D. O. Shmundiak[1]

Information Technology for Finding Possible Sources of Increased River Pollution Using the PROPHET Model

[1]Vinnytsia National Technical University

Climate change has led to many low-water years and, consequently, a decrease of the volume of water to dilution anthropogenic pollution. Thus, research aimed at identifying the main sources of pollution to regulate them immediately is becoming increasingly important. Moreover, according to the EU Water Framework Directive, which, according to the Association Agreement with the EU, Ukraine is obliged to comply with, it is necessary to develop a set of actions soon to achieve or stabilize at least good environmental status in all water bodies. In Ukraine, as in many other European countries, the water quality monitoring system does not provide a sufficient amount of regular observation data for localization in space and time of all, including unregistered, sources of increased pollution, which complicates the implementation of the policy of their regulation. Therefore, it is important to create information technology to find possible sources of increased anthropogenic pressure on the river according to regular observations of water quality in the basin of a given river. The analysis showed that such data is characterized by a change in the frequency of observations (especially in the long run for decades), there is a practice of one-time observations (once a quarter or six months, each time at different times), many missed data, etc., which makes it impossible to use typical similar problems of multiple regressions and time series models based on autoregression and integrated moving average (ARIMA). It is proposed to use Facebook's Prophet model and package for R and Python, which is devoid of all these short-comings and is optimal for solving this problem. The methodology of its application is developed and characterized, which consists in the modeling of monitoring data with filtering of different types of seasonality and allocation of a linear trend between change points, the first approximation of each of which is set at the beginning of intervals in one or several years, depending from the amount of available data. The identified trends between these points are compared by different indicators at each observation post and a specially developed algorithm reveals the largest increases in trends ("pulses"), which then cause a monotonous increase in pollution up to this time. The detected dates of such "pulses" are scaled and aggregated by different indicators, which allows to determine the date of occurrence of the source of pollution at each section between posts and then, according to other data with the involvement of relevant control services, more accurately identify the source of increased river pollution, at present. A program in Python was developed, which tested the efficiency of the technology to detect such "impulses" on the example of the Southern Bug River from its source to Vinnytsia according to the state water quality monitoring system for 2002-2019 and presents the successful results of its work.

Keywords: information technology, water quality, time series, Prophet model, source of river pollution, Python.

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

Slobodyanyuk Olena V. — Cand. Sc. (Pedag.), Associate Professor of the Chair of Engineering Systems in Construction, e-mail: olenas8@gmail.com ;

Davidyuk Oksana M. — Post-Graduate Student of the Chair of System Analysis and Information Technologies, e-mail: davidyuk-ok@ukr.net ;

Shmundiak Dmytro O. — Student of the Department of Computer Systems and Automation, e-mail: dimashmund@gmail.com

В. Б. Мокин[1]
О. В. Слободянюк[1]
О. М. Давидюк[1]
Д. О. Шмундяк[1]

Информационная технология поиска возможных источников повышенного загрязнения реки с использованием модели Prophet

[1]Винницкий национальный технический университет

Изменения климата обусловили ряд маловодных лет и, соответственно, уменьшение количества воды для разбавления антропогенных загрязнений. Поэтому, все более актуальными становятся исследования, направленные на выявление основных источников загрязнения с целью немедленного регулирования. Более того, согласно с Водной рамочной директивой ЕС, которую, в соответствии с Соглашением об ассоциации с ЕС, обязана выполнять и Украина, необходимо в ближайшее время выработать комплекс действий для достижения или стабилизации не ниже хорошего экологического состояния во всех массивах вод. В Украине, как и во многих других странах Европы, система мониторинга качества вод не обеспечивает достаточного количества данных регулярных наблюдений для локализации в пространстве и времени всех, в т.ч. незарегистрированных, источников повышенного загрязнения, что затрудняет реализацию политики их регулирования. Поэтому важно создать информационную технологию поиска возможных источников повышенной антропогенной нагрузки на реку по данным регулярных наблюдений качества воды в бассейне заданной реки. Проведенный анализ показал, что таким данным присуще изменение периодичности наблюдений (особенно в долгосрочной перспективе в десятки лет), существует практика одномоментных наблюдений (раз в квартал или полгода, каждый раз в разное время), много пропущенных данных и др., что делает невозможным применение типичных для подобных задач множественных регрессий и моделей временных рядов на основе авторегрессии и проинтегрированного скользящего среднего (АРПСС). Предложено использовать модель и пакет программ Prophet компании Facebook для R и Python, лишенную всех указанных недостатков и являющуюся оптимальной для решения поставленной задачи. Разработаны и охарактеризованы методология ее применения, заключающаяся в моделировании данных мониторинга с фильтрацией различных видов сезонности и выделении линейного тренда между точками изменения, первое приближение каждой из которых задается в начале интервалов в один или несколько лет, в зависимости от количества имеющихся данных. Идентифицированные тренды между этими точками сопоставляются по разным показателям на каждом посту наблюдений и по специально разработанному алгоритму выявляются самые большие приросты трендов «импульсы», вызывающие затем монотонное нарастание загрязнения до сего времени. Обнаруженные даты таких «импульсов» масштабируются и агрегируются по разным показателям, что на каждом участке между постами дает возможность определить дату появления источника загрязнения и затем, по другим данным, с привлечением соответствующих контролирующих служб, более точно идентифицировать источник повышенного загрязнения реки в настоящее время. Разработана программа на Python, на которой проверена эта способность созданной технологии выявлять такие «импульсы» на примере участка р. Южный Буг от истока до г. Винница по данным государственной системы мониторинга качества вод за 2002—2019 гг. и приведены успешные результаты ее работы.

Ключевые слова: информационная технология, качество воды, временной ряд, модель Prophet, источник загрязнения реки, Python.

Мокин Виталий Борисович — д-р техн. наук, профессор, заведующий кафедрой системного анализа и информационных технологий, e-mail: vbmokin@gmail.com ;

Слободянюк Елена Валерьевна — канд. пед. наук, доцент кафедры инженерных систем в строительстве, e-mail: olenas8@gmail.com ;

Давидюк Оксана Николаевна — аспирант кафедры системного анализа и информационных технологий, e-mail: davidyuk-ok@ukr.net ;

Шмундяк Дмитрий Александрович — студент факультета компьютерных систем и автоматики, e-mail: dimashmund@gmail.com