УДК 519.688:556

О.П. Луценко, О.Г. Байбуз, А.О. Чорна

Дніпропетровський національний університет імені Олеся Гончара

ОСОБЛИВОСТІ ЗАСТОСУВАННЯ МЕТОДІВ ВИЯВЛЕННЯ РОЗЛАДНАНЬ У ПРОЦЕСІ ДОСЛІДЖЕННЯ ДАНИХ ЕКОЛОГІЧНОГО МОНІТОРИНГУ В РЕГІОНАХ З ТЕХНОГЕННИМ НАВАНТАЖЕННЯМ

Досліджено статистичні властивості даних моніторингу хімічних показників складу води річки Саксагань та хвостосховища Північного гірничо-збагачуваного комбінату. Запропоновано способи перетворення та обробки даних, а також модифікації та налаштування методів пошуку розладнання, які дозволяють ефективно отримати дані про розладнання процесів для подальших етапів дослідження.

Ключові слова*: нестаціонарні процеси, визначення розладнань, екологічний моніторинг.*

Исследованы статистические свойства данных мониторинга химических показателей состава воды реки Саксагань и хвостохранилища Северного горно-обогатительного комбината. Предложены способы преобразования и обработки данных, а также модификации и настройки методов поиска разладок, которые позволяют эффективно получать данные о разладках процессов для дальнейших этапов исследования.

Ключевые слова: *нестационарные процессы, выявление разладок, экологический мониторинг.*

The statistical properties of the chemical monitoring data of the Saksagan river and the tailings dam of the Northern Mining-enriched plant has been studied. The methods for data conversion and processing, as well as the modification and adjustment of change-point detection methods has been proposed. The methods and adjustments mentioned allow to effectively retrieve change-point data from processes of ecological monitoring for the further stages of the study.

Keywords: *non-stationary processes, change-point detection, ecology monitoring.*

Постановка проблеми. Аналіз часових рядів, утворених нестаціонарними процесами, був і залишається актуальним

напрямом досліджень з огляду на наявність великої кількості задач, що зводяться до роботи з часовими рядами. Задача екологічного моніторингу належить саме до такого класу задач. Існує значна кількість методів та підходів розв'язання задачі аналізу станів та прогнозування нестаціонарних процесів, серед яких можна виділити застосування методів лінічної та нелінійної регресії, нечіткої логіки, нейронних мереж, динамічного та еволюційного програмування.

Окремо слід зазначити напрям, який на сьогоднішній день є перспективним для застосування в галузі аналізу нестаціонарних процесів – ***методи виявлення розладнання***. Даний напрям дослідження представляє інтерес як засіб дослідження часових рядів через здатність методів реагувати на зміни статистичних характеристик ряду з мінімальною затримкою у часі, розмежовуючи ряд на ділянки з подібними статистичними властивостями.

В регіонах з підвищеним технологічним навантаженням такі зміни виникають при різкому впливі на навколишнє середовище, наприклад при екологічній катастрофі, при якій можлива швидка зміна контрольованих фізичних і хімічних показників. При обробці наукових спостережень, використовуючи методи виявлення розладнання, є можливим раннє виявлення відхилень у показниках моніторингу, локалізація джерел збурень і, як наслідок, своєчасне попередження наслідків цих збурень.

Розладнанням випадкового процесу називається стрибкоподібною зміною його властивостей, що відбувається в невідомий момент часу τ, або не відбувається взагалі. Завданням виявлення розладнання є встановлення факту розладнання, і якщо таке сталося, оцінювання моменту часу τ.

Виявлення розладнань у процесі екологічного моніторингу є складовою задачі контролю та мінімізації ризиків форс-мажорних обставин, так як дає можливість вчасно втрутитися у процес керування і зменшити фінансові втрати, пов'язані зі справдженням гіпотези про розладнання.

Аналіз літературних даних і постановка проблеми. Задача послідовного виявлення розладнання з моменту її формальної постановки в 1950-х роках отримала розвиток у працях вітчизняних і зарубіжних вчених. У працях [1; 2] були запропоновані непараметричні модифікації методів, що дозволяють виявляти розладнання при нестачі даних про статистичний розподіл процесу після моменту розладнання. Авторами [3; 4] було здійснено перехід до байєсівських методів аналізу часових рядів у задачі про розладнання. В попередніх працях авторів даної статті, зокрема [5; 6], було запропоновано моделі та

обчислювальні схеми ймовірнісного аналізу часових рядів, які дозволяли обчислювати ймовірність розладнання у часових рядах з множинними розладнаннями.

Якнайшвидше виявлення події розладнання і встановлення моменту часу τ в досліджуваному процесі є значущим чинником у процесі аналізу, а послідовні алгоритми, що використовуються для виявлення розладнань, повинні залишатися працездатними при будь-яких відхиленнях статистичних характеристик спостережуваних сигналів. При цьому слід зазначити, що властивості вхідних часових рядів, а також характер розладнань, які повинні бути виявлені в задачі екологічного моніторингу, обумовлюють виникнення специфічних підзадач, пов'язаних з перетвореннями вхідних послідовностей і налаштуванням методів таким чином, щоб забезпечити достатню якість виявлення. Ідентифікація і вирішення таких підзадач є необхідною складовою вирішення задачі ідентифікації розладнань.

Мета даної статті – на основі комплексного аналізу вхідних даних обрати найбільш придатні для проведення подальшого дослідження алгоритми виявлення розладнань та встановити їхні налаштування для використання у процесі аналізу рядів екологічного моніторингу.

Основний матеріал. Дані, що розглядаються в рамках дослідження – гідрохімічний склад вод районів з підвищеним технологічним навантаженням, зокрема дані хімічних показників складу води річки Саксагань (рис. 1) та хвостосховища Північного гірничо-збагачувального комбінату (рис. 2).

Перед побудовою моделей виявлення розладнань необхідно провести попередній аналіз цих даних для виявлення відмінних статистичних особливостей процесів. Далі наведено результати перевірки статистичних властивостей рядів, з урахуванням яких виконувалися подальші етапи дослідження.

Отримані коефіцієнти кореляції між гідрохімічними показниками вод р. Саксагань та хвостосховища наведено у табл. 1 і 2. За ними можна зробити висновок, що для обох водних джерел між усіма даними проявляється сильна або середня залежність. Тільки показник РН кислотність слабо корельований з іншими показниками. Це можна бачити і за графіком його рівнів.

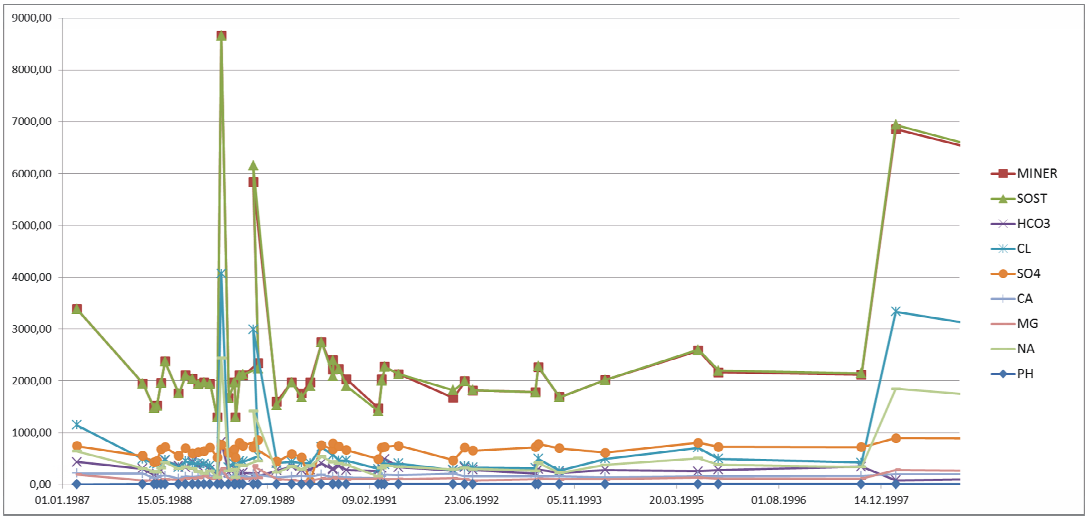

Рисунок 1 – Часові ряди даних річки Саксагань

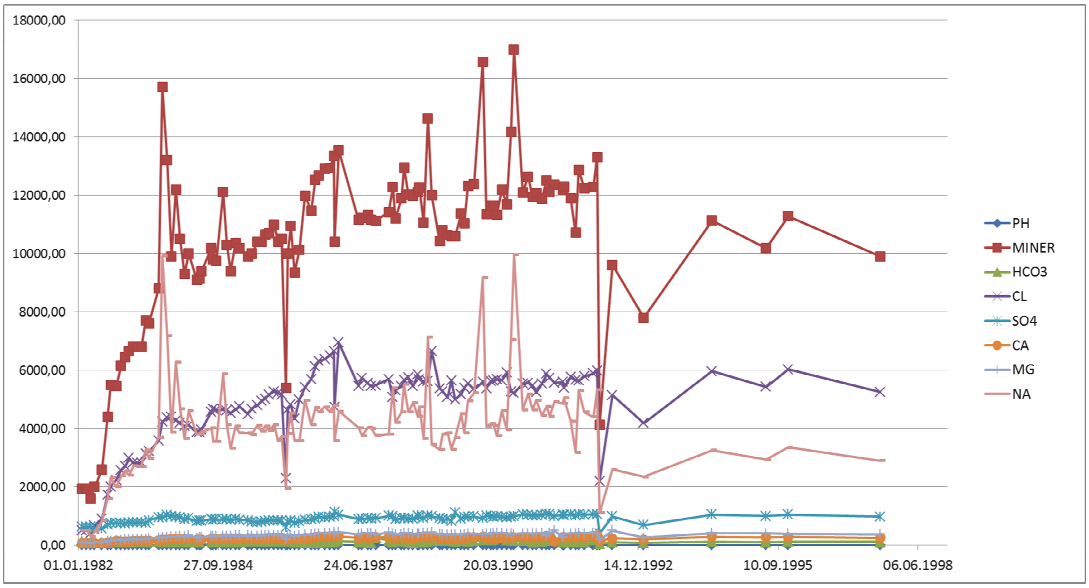

Рисунок 2 – Часові ряди даних хвостосховища Північного гірничо-збагачувального комбінату

Таблиця 1 – Коефіцієнти кореляції між даними р. Саксагань

	Mineral	PH	HCO3	SO4	Cl	Ca	Mg	Na
Mineral	1	-0,04569	0,380594	0,836438	0,829786	0,500989	0,575182	0,940915
PH	-0,04569	1	0,231659	0,019775	-0,26862	-0,17319	0,14119	-0,1378
HCO3	0,380594	0,231659	1	0,172162	0,161664	0,36426	0,13212	0,258212
SO4	0,836438	0,019775	0,172162	1	0,591462	0,40115	0,546197	0,735907
Cl	0,829786	-0,26862	161664	0,591462	1	0,338051	0,426466	0,881301
Ca	0,500989	-0,17319	0,36426	0,40115	0,338051	1	0,258497	0,418449
Mg	0,575182	0,14119	0,13212	0,546197	0,426466	0,258497	1	0,419196
Na	0,940915	-0,1378	0,258212	0,735907	0,881301	0,418449	0,419196	1

Таблиця 2 – Коефіцієнти кореляції між даними хвостосховища

	Mineral	PH	HCO3	SO4	Cl	Ca	Mg	Na
Mineral	1	0,242569	0,262598	0,740274	0,791098	0,776215	0,734273	0,886257
PH	0,242569	1	0,071485	0,104478	0,338202	0,310287	0,356376	0,206522
HCO3	0,262598	0,071485	1	0,292346	0,28468	0,291074	0,311122	0,167929
SO4	0,740274	0,104478	0,292346	1	0,702331	0,786697	0,731661	0,582591
Cl	0,791098	0,338202	0,28468	0,702331	1	0,883866	0,8854	0,516075
Ca	0,776215	0,310287	0,291074	0,786697	0,883866	1	0,822179	0,538863
Mg	0,734273	0,356376	0,311122	0,731661	0,8854	0,822179	1	0,492041
Na	0,886257	0,206522	0,167929	0,582591	0,516075	0,538863	0,492041	1

Перед виділенням тренда було перевірено дані на його наявність за допомогою критеріїв: ранговий критерій Спірмена і критерій, заснований на знаках різниці. Також використовувався критерій випадковості, заснований на серіальній кореляції (рис. 1.3), який будується за коефіцієнтами автокореляції. Аналіз даних на випадковість показав, що покажчики річки Саксагань є випадковими даними, в той час як у рядах хвостосховища виявлено зростаючий тренд.

Виділення тренда здійснювалося:

— методом лінійної регресії: $y(t) = a_0 + a_1 t + e_t$,

— методом нелінійної регресії – модифікована експоненціальна крива зростання: $y(t) = a_0 a_1^t + e_t + \gamma$.

У цих методах значення параметрів a_0 , a_1^t знаходять методом найменших квадратів, e_t – випадкова складова, γ – обчислюється методом трьох точок.

Для перевірки адекватності та значущості побудованих трендів використовувався дисперсійний аналіз. Вибір найкращого тренда проводився за байєсівським інформаційним критерієм Шварца – BIC.

Для даних хвостосховища усі моделі нелінійного тренда виявилися неадекватними. Деякі лінійні моделі тренда були неадекватними, але в них похибка незначна і не перевищує 5 %. Цим можна знехтувати, так як наявність точок розладнань ускладнює виділення тренда. По критерію BIC для усіх часових рядів хвостосховища лінійний тренд визнано кращим.

Також було перевірено два види моделей сезонності:

— в адитивній моделі сезонність виражається у вигляді абсолютної величини, яка додається або віднімається з середнього значення ряду для обліку сезонності;

— в мультиплікативній моделі сезонність виражається як відсоток від середнього рівня, на який множиться середнє значення для обліку сезонності при прогнозуванні.

Сезонність будувалася за допомогою гармонічного аналізу, використовуючи розклад у ряд Фур'є. Виявлення сезонності необхідно проводити по стаціонарному ряду. Тому для тих даних, де є трендова складова, вона попередньо була вилучена. Перевірка адекватності та значущості тренд-сезонної моделі проводилася за допомогою дисперсійного аналізу.

Для даних річки Саксагань після виділення сезонності і побудови тренд-сезонної моделі для усіх покажчиків, крім РН, модель виявилася неадекватною. Для даних хвостосховища ситуація є аналогічною.

Таким чином:

1. Числові ряди гідроекологічних показників, що розглядаються, характеризуються високими значеннями автокореляції та значною кореляцією між окремими показниками.

2. У вхідних даних сезонність відсутня.

3. Дані по р. Саксагань є стаціонарними за середнім значенням, в той час як числові ряди хвостосховища містять у собі зростаючий тренд, який ускладнює коректне виявлення розладнання.

Ключовою особливістю розглядуваних даних є наявність у них моментів часу, коли вони змінюють статистичні властивості. В даних, зібраних з річки Саксагань, спостерігаються єдиноразові викиди показників вгору, в той час як процеси, що протікають у хвостосховищі, утворюють ряди з локально направленими трендами на фоні глобального.

Для вирішення завдання представляється доцільним використання як послідовних методів виявлення розладнання, так і алгоритмів апостеріорного класу. Послідовні методи повинні забезпечити обробку новітніх даних, що надходять, в той час як апостеріорні – слугувати для побудови основної вибірки точок розладнання на вже існуючих даних у минулому, а також при наявності можливості одночасного застосування обох класів методів до однієї і тієї самої вибірки, бути контрольними методами.

Досліджувані процеси належать до класу нестаціонарних випадкових процесів з множинними (повторюваними) розладнаннями. При цьому характер спостережуваної при розладнанні зміни статистичних характеристик відрізняється, і статистичні

характеристики процесу після розладнання невідомі. У зв'язку з таким характером досліджуваного процесу, для подальшого дослідження більшою мірою підходять непараметричні методи, які зовсім не потребують інформації про розподіл процесу після розладнання, або методи з низькими вимогами до параметризації характеру розподілу після розладнання.

У зв'язку з тим, що розладнання в одному з досліджуваних процесів (р. Саксагань) являє собою викиди вгору на фоні незмінного середнього значення на ділянках без розладнання, для використовуваних методів бажана здатність виявлення змін статистичних характеристик (в даному випадку – середнього значення) як в обидві сторони, так і в одну сторону.

З урахуванням названих критеріїв, для вирішення задачі послідовного виявлення була задіяна непараметрична форма алгоритму кумулятивних сум [2].

На кожному кроці алгоритму накопичується сума:

$$S_t = \max(0, S_{t-1} + g_t) .$$

Сигнал про розладнання подається у момент часу:

$$\tau = \inf(t \geq 1 : S_t > b) ,$$

де b – бар'єр чутливості методу.

Для випадку збільшення математичного очікування m:

$$g_t = z(t_i) - m_1 - k \quad при\ m_2 > m_1 ,$$

для випадку зменшення m:

$$g_t = -z(t_i) + m_1 + k \quad при\ m_2 < m_1 ,$$

де m_1, m_2 – математичне очікування величини y до і після розладнання відповідно,

$k \geq 0$ – поріг чутливості методу для відхилення від m_1.

Обрані обчислювальні схеми послідовного виявлення моменту розладнання не можуть бути застосовані у випадку нестаціонарних процесів з направленим характером змін середнього, так як сигнал про розладнання виникатиме на кожному кроці спостережень. При послідовному аналізі процесів з направленим характером змін середнього значення (вираженим трендом), необхідно здійснити таке перетворення часового ряду, при якому зберігалися б постійні середні значення. У якості таких перетворень авторами запропоновано такі:

1) обчислення ряду кінцевих різниць і знаходження розладнань його середньої величини. Оскільки стабільний тренд сам по собі характеризується зміною середнього значення спостережень, у даному випадку доцільно застосовувати АКС не до вихідної кривої

спостережень, а до її першої похідної за часом. Різка зміна значення першої похідної буде означати зміну напряму графіка часового ряду, тобто розладнання початкового процесу;

2) обчислення ряду значень нахилів апроксимуючих відрізків (рис. 2.3). Для кожної точки часового ряду знаходяться рівняння прямої, яка найкраще апроксимує задану кількість точок l часового ряду зліва від заданої точки. Позначимо: y_t – значення точки часового ряду, яку спостерігаємо; t – відповідний момент часу; k , b – коефіцієнти рівняння апроксимаційної прямої.

Досягнемо найкращої апроксимації за допомогою рішення такої системи (мінімізуємо сумарне квадратичне відхилення точок апроксимаційної прямої від кривої спостережень):

$$y_j = kt_j - b,$$

$$\sum_{j=i-l}^{j=i} (y_j - kt_j - b)^2 \to \min,$$

$$\sum_{j=i-l}^{i} y_j t_j = k \sum_{i=t-l}^{i} t_j^2 + b \sum_{j=i-l}^{i} t_j,$$

$$\sum_{j=i-l}^{i} y_j = k \sum_{j=i-l}^{i} t_j + nb.$$

Знайдемо коефіцієнт k, що є тангенціальним коефіцієнтом рівняння прямої. Визначивши кути arctg(k) для усіх точок часового ряду, отримаємо ряд нахилів апроксимуючих прямих до горизонтальної осі. Знаходячи розладнання даного ряду, отримуємо набір точок розладнань для часового ряду з вираженим трендом.

В якості методів послідовного виявлення було використано апостеріорний тест Бродського – Дарховського [1; 7]:

$$Y(\tau) = \frac{\tau}{N}\left(1 - \frac{\tau}{N}\right)\left(\frac{1}{\tau}\sum_{1}^{\tau} y_i - \frac{1}{N-\tau}\sum_{\tau+1}^{N} y_i\right),$$

$$G = \max_{1 \le \tau \le N-1} |Y(\tau)|,$$

$G < h$ => *дефекта немає*,

$G \ge h$ => *дефект*,

$$T(\tau) = Y(\tau + [\varepsilon N]) - Y(\tau),\ 0 < \varepsilon < \frac{\delta}{4}.$$

$$\tau_1 = \begin{cases} \min A_1, A_1 \neq \varnothing \\ N, A_1 = \varnothing \end{cases},$$

$$A_1 = \{\tau \geq 1 : sign(T(\tau)) \neq sign(T(\tau+1))\},$$

$$\tau_i = \begin{cases} \min A_i, A_i \neq \varnothing \\ N, A_i = \varnothing \end{cases},$$

$$A_i = \{\tau \geq \tau_{i+1} + [\delta N / 2] : sign(T(\tau)) \neq sign(T(\tau+1))\},$$

$$i = 2..\tilde{k},$$

$$\tilde{k} = \min\{s : \tau_S = N\} - 1.$$

Даний алгоритм має важливу для поставленої задачі особливість – здатність виявляти повторювані розладнання на часовому ряді. В той же час більш класичні методи, зокрема тест Манна – Уітні [7], виявляють лише одне розладнання на інтервалі:

$$u_{ik} = \begin{cases} 1, y_i \geq y_k \\ 0, y_i < y_k \end{cases},$$

$$G(\tau) = \frac{\sum_{k=\tau+1}^{N} \sum_{i=1}^{n} u_{ik}}{\tau(N-\tau)},$$

$$\tau_0 = \arg \min_{[\alpha N] \leq \tau \leq [N-\alpha N]} G(\tau), \mu_1 < \mu_2,$$

$$\tau_0 = \arg \max_{[\alpha N] \leq \tau \leq [N-\alpha N]} G(\tau), \mu_1 > \mu_2,$$

де τ_0 – момент розладнання.

Слід, однак, зазначити, що алгоритми, засновані на виявленні розладнання як локального мінімума або максимума деякої допоміжної статистики, можуть бути налаштовані на виявлення множинних розладнань. Для цього в рамках вирішуваної задачі був застосований рекурсивний порядок виклику алгоритму на ділянках зліва і справа від розладнання: (0; τ_0), (τ_0+1, N). Умови завершення рекурсії: на ділянці відсутній локальний максимум $G(\tau)$ або локальний максимум $G(\tau)$ не перевищує встановлений бар'єр.

Приклад результату застосування алгоритмів до часового ряду наведено на рис. 3.

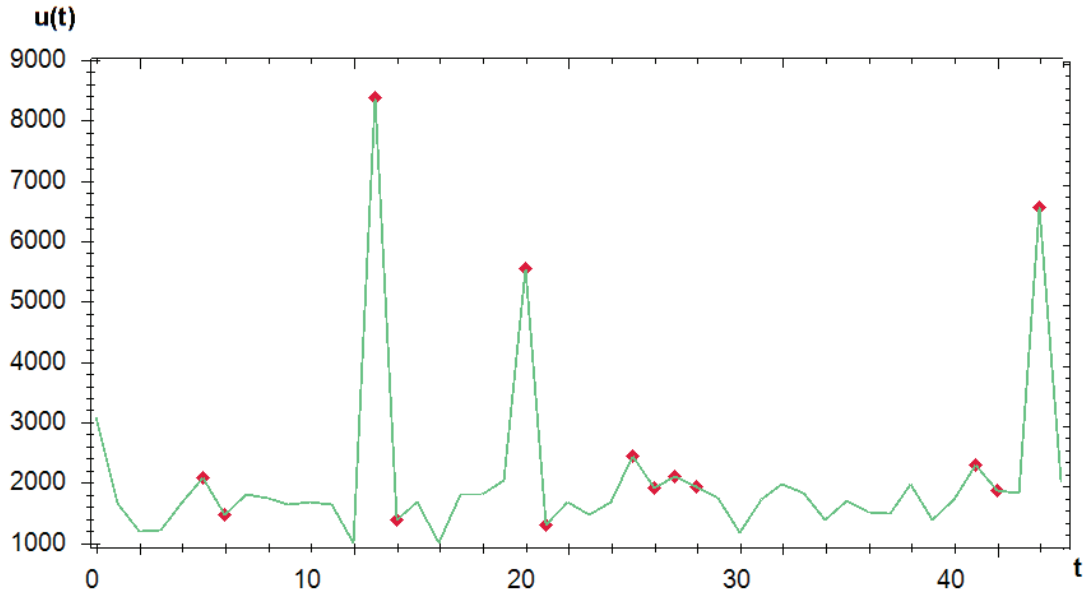

Рисунок 3 – Точки розладнання на часовому ряді

Уточнення результатів виявлення розладнань за допомогою апостеріорних алгоритмів проводилося таким чином: після виявлення розладнання послідовним алгоритмом у точці τ_S апостеріорний алгоритм запускався на ділянці в околі точки розладнання: $(\tau_{Si} - \alpha; \tau_{Si} + \alpha)$, де α – величина допуску, яка налаштовується. За відсутності сигналу про розладнання від апостеріорного алгоритму сигнал про розладнання вважався помилковим. Щоб виявити кількість моментів розладнання, що були пропущені послідовним алгоритмом, послідовний алгоритм запускався на інтервалі часу в минулому, після чого виявлялися точки, знайдені апостеріорним алгоритмом, в околі яких $(\tau_{aj} - \beta; \tau_{aj} + \beta)$ відсутні точки розладнання τ_{Si}, що виявлені послідовно.

Вказані дії дозволили виявити ймовірність помилок першого і другого роду при застосуванні кожного набору налаштувань послідовного алгоритму і обрати налаштування алгоритмів згідно з цими критеріями.

Висновки. Таким чином, отримані налаштування та запропоновані модифікації методів виявлення розладнань, які дозволяють адаптувати методи до застосування в предметній галузі екологічного моніторингу.

Запропоновані методи усунення тренда, придатні для використання при послідовному виявленні розладнань. Дістала подальшого розвитку методика апостеріорного виявлення множинних розладнань. Запропоновано спосіб уточнення результатів послідовного виявлення

множинних розладнань за допомогою апостеріорних алгоритмів.

Отримані результати закладають основу для розширення області застосування підходів, запропонованих у [5; 6], на нові класи процесів, що характеризуються складнощами у процесі виявлення розладнань.

Бібліографічні посилання

1. Brodsky B. E., Darkhovsky B. S. Nonparametric Methods in Change-Point Problems. Dordrecht. Kluver Academic Publishings. 1993. 210 p.

2. Nadler J., Robbins N. B. Some characteristics of Page's twosided procedure for detecting a change in a location parameter. // Ann. Math. Statist. 1971. Vol. 42. N 2. P. 538–551.

3. Adams R. P., McKay D. J. C. Bayesian Online Changepoint Detection // arXiv preprint. 2007. URL: http://arxiv.org/pdf/0710.3742v1.pdf (дата звернення: 28.09.2016).

4. Adams R. P., Murray I., McKay D. J. C. Nonparametric Bayesian density modeling with Gaussian processes // arXiv preprint. 2009. URL: https://arxiv.org/pdf/0912.4896.pdf (дата звернення: 28.09.2016)

5. Lutsenko O., Baybuz O. Model of probabilistic assessment of trend stability at financial market // Eastern-European Journal of Enterprise Technologies. 2013. Vol. 6. N 3 (66). P. 50–54.

6. Луценко О. П., Байбуз О. Г. Оцінка функції ризику розладнання процесу коливань валютних курсів із застосуванням байєсівської оцінки параметра функції умовного розподілу // The European Scientific and Practical Congress «Global scientific unity 2014» (Prague (Czech Republic) 26–27 September 2014). Prague. 2014. С. 126–132.

7. Бродский Б. Е., Дарховский Б. С. Асимптотический анализ некоторых оценок в апостериорной задаче о разладке. // Теория вероятностей и ее применения. 1990. Т. 35. № 3. С. 551–557.

Надійшла до редколегії 01.10.2016.