УДК 519.254

О.Г. Байбуз, М.Г. Сидорова, Л.П. Сидорова

Дніпропетровський національний університет імені Олеся Гончара

ІНФОРМАЦІЙНО-АНАЛІТИЧНА СИСТЕМА МОНІТОРИНГУ ПОВЕРХНЕВИХ ВОД «ANALIT»

Запропоновано інформаційну технологію кластерного аналізу результатів моніторингу для обґрунтування пунктів спостережень, об'ємів і періодичності гідрохімічних випробувань та подано опис розробленої інформаційно-аналітичної системи моніторингу поверхневих вод «AnalIT».

Ключові слова: *гідрохімічний моніторинг, кластерний аналіз, інформаційна технологія, часові ряди, прогнозування, оцінка якості.*

Предложена информационная технология кластерного анализа результатов мониторинга для обоснования пунктов наблюдений, объемов и периодичности гидрохимических испытаний, а также представлено описание разработанной информационно-аналитической системы мониторинга поверхностных вод «AnalIT».

Ключевые слова: *гидрохимический мониторинг, кластерный анализ, информационная технология, временные ряды, прогнозирование, оценка качества.*

Information technology of cluster analysis of the monitoring results to justify the observation points, volume and frequency of hydro-chemical tests has been offered, a description of the developed information-analytical system of monitoring of surface water «AnalIT» has been submitted.

Keywords: *hydrochemical monitoring, cluster analysis, information technology, time series, forecasting, assessment of quality.*

Вступ. Внаслідок значного техногенного навантаження змінюються гідрохімічні процеси водних об'єктів. Тому актуальним є проведення гідрохімічного моніторингу з метою збереження, поліпшення і стабілізації якості поверхневих вод для забезпечення оптимальних умов функціонування екосистем та підвищення ефективності природно-господарського комплексу. Особливо складним є гідрохімічний моніторинг водних об'єктів у районах з підвищеним техногенним навантаженням [1].

У зв'язку з удосконаленням технологій запису і зберігання даних моніторингу спостерігається тенденція накопичення великої кількості

інформації. Виникає потреба обробки наборів даних значних об’ємів з метою виявлення прихованих у них знань, закономірностей, властивостей, тенденцій, кращого розуміння структури. Це призводить до необхідності розробки інформаційних систем та програмних засобів, що дозволять вирішувати такі задачі.

Аналіз досліджень та постановка задачі. Одне з основних місць у системі гідрохімічного моніторингу займає обгрунтування пунктів спостережень, об’ємів і періодичності гідрохімічних випробувань. Основою розміщення гідрологічних пунктів спостережень, де проводяться спостереження гідрохімічних, гідрометричних та гідрологічних характеристик водостоків і водойм, є принцип основних характеристик водного режиму – рівня води і річкового стоку. Кількість і щільність розміщення пунктів спостережень визначають природно-кліматичними факторами. Використовуючи гідрохімічне районування, можна до певної міри уніфікувати водоохоронні заходи в межах виділених груп та районів. Визначивши пріоритетні водоохоронні заходи для одного об’єкта, планувати і впроваджувати їх для всієї виділеної групи. Для вирішення цієї задачі запропоновано застосовувати методи кластерного аналізу [2–3], що дозволяють поділити сукупність об’єктів на однорідні за певним формальним критерієм подібності групи (кластери). Основною властивістю цих груп є те, що об'єкти, які належать одному кластеру, подібніші між собою, ніж об’єкти з різних кластерів. Таку класифікацію можна виконувати одночасно за досить великою кількістю ознак. Крім того, методи кластерного аналізу дозволять зрозуміти структуру багатовимірних даних, спростити подальшу обробку, скоротити вихідну вибірку, виявити нетипові об’єкти, сформулювати або перевірити гіпотези на підставі отриманих результатів.

Найчастіше об’єкти моніторингу характеризуються набором досліджуваних ознак, що змінюються у часі, тобто дані для аналізу подаються у вигляді багатовимірних часових рядів [4]. Виділення однорідних груп часових рядів для подальшого їх аналізу та прогнозування в останні роки стає все більш актуальною проблемою.

Огляд існуючих інформаційних технологій показав, що майже не існує систем, орієнтованих на кластерний аналіз даних моніторингу; досить мало уваги приділено оцінюванню якості та аналізу отриманих угруповань; більшість програмних засобів не забезпечує підтримку прийняття рішень щодо вибору найкращого результату чи кількості кластерів; задача кластерного аналізу часових рядів не розглядається або розглядається частково; більшість систем потребують від користувача високого рівня кваліфікації. Таким чином, актуальною є

розробка нових інформаційних технологій кластерного аналізу результатів моніторингу.

З метою забезпечення збору, обробки, збереження та аналізу інформації про стан поверхневих вод, прогнозування його змін та розробки рекомендацій для прийняття ефективних управлінських рішень поставлено задачу розробити інформаційно-аналітичну систему моніторингу поверхневих вод.

Основний матеріал. Для проведення інформаційно-аналітичного контролю якості поверхневих вод запропоновано ряд послідовних операцій: відбір проб; підготовка проб до їх транспортування й збереження, доставка у лабораторію; підготовка проб до аналізу безпосередньо у лабораторії; кількісний аналіз у лабораторних умовах; обробка отриманих результатів засобами розробленої інформаційної технології з метою виявлення закономірностей, властивостей, тенденцій, кращого розуміння структури.

Об'єкти моніторингу характеризуються набором досліджуваних ознак, що змінюються у часі, тобто дані для аналізу подаються у вигляді багатовимірних часових рядів $X = \{x_1, x_2, ..., x_N\}$, $x_i = \{u_1, u_2, ..., u_p\}$, $i = \overline{1, N}$, де N – кількість об'єктів, $u_l^{(i)} = \{u_{lt}^{(i)}\}$, $l = \overline{1, p}$, p – кількість ознак, $t = \overline{1, T}$, T – кількість моментів спостереження, $u_{lt}^{(i)}$ – значення l-ї ознаки i-го об'єкта у t-й момент часу.

Структура програмного забезпечення «AnalIT» складається з п'яти основних блоків та представлена на рис. 1.

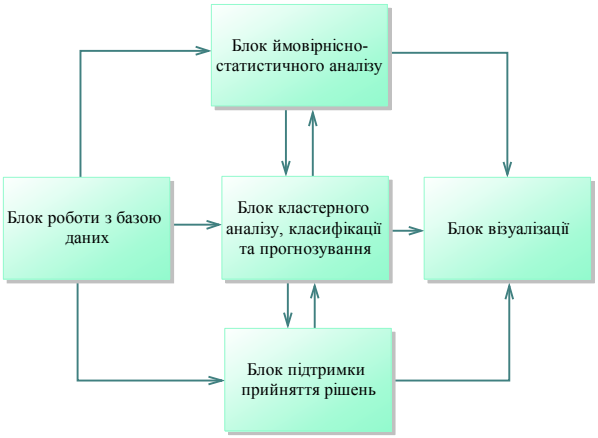

Рисунок 1 – Структура програмного забезпечення «AnalIT»

Блок роботи з базою даних. Система дозволяє завантажувати, зберігати, обробляти, редагувати дані, реалізує наступні можливості:

- формування локальних баз даних, здійснюючи запити на вибір даних за певними ознаками, моментами часу, об'єктами;
- оцінювання інформативності досліджуваних ознак методами апроксимації матриць близькості та «Гойдалки»; автоматичного вибору ознак заданої інформативності;
- проведення стандартизації даних.

Блок кластерного аналізу, класифікації та прогнозування реалізує наступні можливості:

- проведення кластеризації на основі методів: агломеративних ієрархічних ближнього зв'язку, дальнього зв'язку, центрального зв'язку, середнього зв'язку та Уорда, швидких ієрархічних методів, K-середніх у варіантах Болла-Холла та Мак-Кіна, графового методу найкоротшого незамкненого шляху, Forel та генетичного алгоритму, а також методів нечіткої кластеризації Уіндхема та Даве-Сена;
- проведення кластерного аналізу багатовимірних часових рядів;
- визначення метрики близькості між об'єктами: евклідової, манхеттенської, Чебишева, Канберра, Брея-Картіса, а також для часових рядів обчислення запропонованої авторами метрики TSS на основі коефіцієнтів кореляції та міри близькості [5];
- дослідження оптимальної кількості кластерів за критеріями якості та критерієм різниці між рівнями об'єднання на дендрограмі;
- оцінювання якості кластеризації шляхом порівняння отриманих різними методами угрупувань на основі відносних критеріїв якості: Калінського – Гарабача, Данна, Беджека – Данна, суми внутрішньокластерних дисперсій за всіма ознаками, суми квадратів відстаней до центрів класів, суми внутрішньокластерних відстаней, відношення середньої внутрішньокластерної і середньої міжкластерної відстаней та проведення багатокритеріальної оцінки якості результатів кластеризації на основі відносних критеріїв якості та методів теорії прийняття рішень: Борда, Коупленда, плюралітарної процедури та множинного аналізу;
- прогнозування значень нових спостережень на основі моделей адаптивних методів прогнозування: Бокса – Дженкінса, Тейла – Вейджа, лінійного зростання з адитивною сезонністю, лінійного зростання з мультиплікативною сезонністю, експоненціального зростання з адитивною сезонністю, експоненціального зростання з мультиплікативною сезонністю, а також регресійних моделей з

можливістю вибору найкращої моделі та проведення кластерного аналізу на їх основі;

– проведення класифікації нових спостережень на основі наступних методів: байєсовського класифікаційного правила, метода найближчих сусідів, лінійної дискримінантної функції, квадратичної дискримінантної функції, методу еталонів, потенціальної функції. Здійснюється вибір найкращої класифікаційної моделі на основі ковзного контролю.

Блок ймовірнісно-статистичного аналізу дозволяє аналізувати як вихідні дані, так і дані кожного окремого кластера. Реалізовано наступні можливості:

– проведення первинного статистичного аналізу та обчислення комплексних індексів забрудненості води як для усіх даних, так і для кожного виявленого класу;

– відтворення одновимірних розподілів: нормального, сплайн-нормального з одним та двома вузлами склеювання, Вейбулла, рівномірного, експоненціального та вибір найбільш вірогідного типу розподілу за використанням критеріїв згоди Колмогорова та $\chi 2$ Пірсона; визначення вірогідних меж значень показника на основі відтвореного розподілу;

– дослідження даних щодо виявлення суттєвих зв'язків поміж показниками на основі використання методів кореляційного аналізу, а саме: коефіцієнтів кореляції Пірсона, Спірмена, Кендала, кореляційного відношення з перевіркою їх значущості.

Блок візуалізації забезпечує представлення результатів аналізу у вигляді таблиць, графіків, дендрограм, гістограм, діаграм розсіювання, списків та текстових коментарів для полегшення їх інтерпретації; забезпечує зручний інтерфейс, взаємодію з користувачем у діалоговому режимі, надаючи можливість змінювати налаштування.

Покладена в основу системи технологія кластерного аналізу результатів моніторингу [6] дозволяє вирішити чотири типи задач:

1) виявлення кластерної структури об'єктів за всіма досліджуваними ознаками на певний момент спостереження;

2) знаходження угруповань моментів спостереження для кожного об'єкта, що характеризуються схожими значеннями досліджуваних показників;

3) групування об'єктів, які схожі між собою, за деякою ознакою у часовому діапазоні її змін;

4) визначення груп схожих об'єктів за всіма досліджуваними ознаками у часовому періоді спостережень.

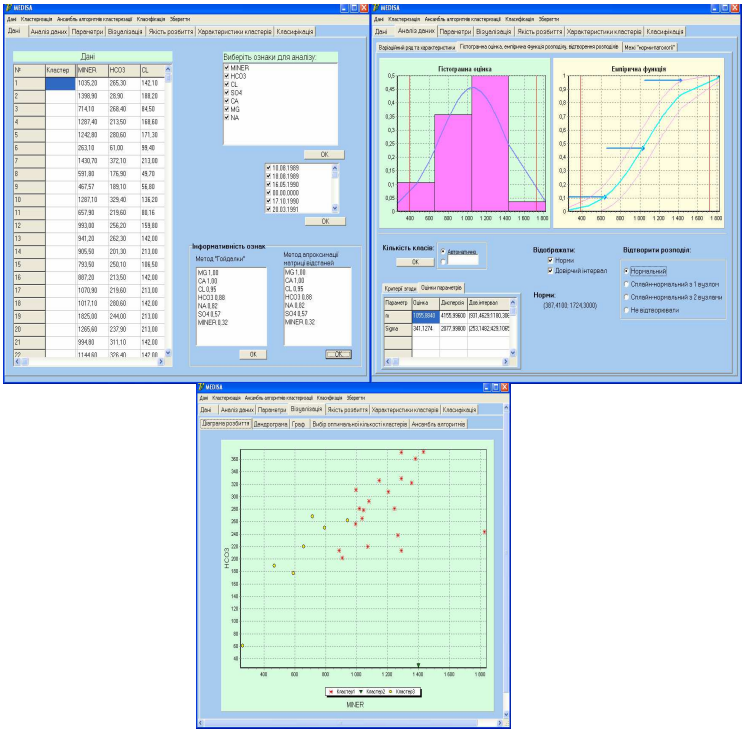

Рисунок 2 – Інтерфейс програмного забезпечення «AnalIT»

На рис. 2 за допомогою діаграми станів представлено перетворення даних при кластерному аналізі баготовимірних часових рядів (БЧР), що ілюструє етапи переходу від вихідної множини до кластерів.

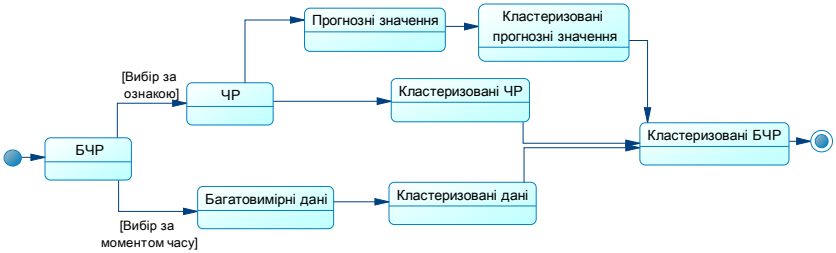

Рисунок 2 – Діаграма станів перетворення даних при кластеризації БЧР

Запропоновану технологію було застосовано до даних гідрохімічного моніторингу р. Самара (Західно-Донбаський регіон, Дніпропетровська область, Україна). Метою роботи було визначення груп контрольних створів, що характеризуються схожим хімічним складом води у р. Самара за досліджуваними компонентами для правильного планування природоохоронних заходів та керування якістю вод річки.

Проби води відбиралися у дев'яти контрольних створах: с. Коханівка (у межах населеного пункту); с. Маломиколаївка (6 км вище населеного пункту, 1 км вище скиду шахтних вод зі ставка-накопичувача б. Косьмінна); с. Петрівка (в межах населеного пункту, 1 км нижче скиду шахтних вод зі ставка-накопичувача б. Косьмінна); с. Богуслав (у межах населеного пункту); с. Тернівка (в/з Федора, 2 км нижче населеного пункту, 1 км вище скиду шахтних вод зі ставка-накопичувача б. Свідовок); с. Вербки (в межах населеного пункту, 1 км нижче скиду шахтних вод з б. Свідовок, 1 км вище скиду з очисних споруд комунального підприємства м. Павлограда); м. Павлоград, с. В'язовок (в межах населеного пункту, 1 км нижче скиду з очисних споруд комунального підприємства м. Павлограда); с. Кочережки (в межах населеного пункту) тричі на рік протягом 6 років.

Для кожної проби фізико-хімічними методами аналізу визначалися наступні показники: водневий показник (РН), розчинений у воді кисень (O_2), біохімічне споживання кисню (БСК), хімічне споживання кисню (ХСК), нітрати (NO_3), нітрити (NO_2), фосфати (PO_4), сухий залишок (СЗ), завислі речовини (ЗР), хлориди (Cl), сульфати (SO_4), аміак (NH_4), нафтопродукти (НП). Аналіз проводився на кафедрі аналітичної хімії хімічного факультету Дніпропетровського національного університету імені Олеся Гончара.

Згідно з розробленою технологією контрольні створи були розподілені на групи за схожістю змін вмісту досліджуваного показника у пробах води; згруповані моменти спостереження, що характеризуються близькими за значеннями ознак пробами води; визначено розбиття контрольних створів на кластери за набором досліджуваних показників для кожного моменту спостереження, а також нечітке розбиття об'єктів аналізу, що враховує часові зміни ознак на всьому проміжку спостережень.

При дослідженні інформативності ознак у кожен з моментів спостереження за методом апроксимації матриці відстаней було виявлено, що усі ознаки, окрім СЗ, є інформативними, найбільшу інформативність мають наступні показники: NO_2, PO_4, NH_4, O_2, НП.

Аналізуючи отримані результати, можна зробити висновок, що контрольний створ, що знаходиться поблизу м. Павлограда, значно відрізняється від усіх інших майже в усі моменти спостереження. Таким чином, і в розбитті, що відповідає всьому проміжку спостережень, виділяється у окремий кластер. Також у окремий кластер виділяються контрольні створи у селах Маломиколаївці, Петрівці та Богуславі, враховуючи, що останній майже з рівною мірою приналежності відноситься до усіх кластерів.

Висновки. Запропоновано інформаційну технологію кластерного аналізу даних моніторингу. Технологія передбачає можливості кластеризації багатовимірних часових рядів, кластерного аналізу даних, що характеризують будь-який момент спостереження або часові зміни деякої ознаки, а також забезпечує оцінку якості отриманих результатів та підтримку прийняття рішень.

Розроблено інформаційно-аналітичну систему моніторингу поверхневих вод «AnalIT», ядро якої містить обчислювальні схеми кластерного аналізу, класифікації, прогнозування, підтримки прийняття рішень, ймовірнісно-статистичного аналізу, а також широкий спектр засобів візуалізації.

Засобами розробленої системи проведено аналіз даних гідрохімічного моніторингу р. Самара (Західно-Донбаський регіон, Дніпропетровська область, Україна). Визначено схожі закономірності зміни вмісту кожного з досліджуваних показників у пробах води контрольних створів у часовому проміжку спостережень; угруповання моментів спостереження для кожного контрольного створу, що характеризуються схожими значеннями ознак; угруповання контрольних створів за схожістю ознак у пробах води для кожного моменту спостереження, а також нечітке розбиття контрольних створів за схожістю набору досліджуваних ознак у пробах води на всьому проміжку спостережень.

Бібліографічні посилання

1. **Шерстюк Н. П.** Особливості гідрохімічних процесів у техногенних та природних водних об‘єктах Кривбасу / Н. П. Шерстюк, В. К. Хільчевський – Д. : ТОВ «Акцент ПП», 2012. – 263 с.
2. **Мандель И. Д.** Кластерный анализ / И. Д. Мандель. – М., 1988. – 176 с.
3. **Миркин Б. Г.** Методы кластер-анализа для поддержки принятия решений: обзор / Б. Г. Миркин. – М. : Изд. дом НИУ «Высшая школа экономики», 2011. – 88 с.

4. **Хеннан Э.** Многомерные временные ряды / Э. Хеннан. – М. : Мир, 1974. – 576 с.

5. **Байбуз О. Г.** Інформаційна технологія нечіткої кластеризації багатовимірних часових рядів на прикладі гідрохімічного моніторингу ріки Самара / О. Г. Байбуз, М. Г. Сидорова // Науковий вісник Національного гірничого університету. – 2014. – № 4. – С. 11–18.

6. **Байбуз О. Г.** Інформаційна технологія кластеризації даних у часовому періоді спостережень / О. Г. Байбуз, М. Г. Сидорова // Системні дослідження та інформаційні технології. – 2013. – № 4. – С. 59–66.

Надійшла до редколегії 15.11.15