УДК 519.254

О. Г. Байбуз, М. Г. Сидорова

Дніпропетровський національний університет ім. Олеся Гончара

ГРУПУВАННЯ ПУНКТІВ ГІДРОХІМІЧНОГО СПОСТЕРЕЖЕННЯ ЗА СХОЖІСТЮ ХІМІЧНОГО СКЛАДУ ВОДИ З УРАХУВАННЯМ ЧАСОВИХ ЗМІН

Запропоновано інформаційну технологію визначення груп схожих об'єктів за сукупністю досліджуваних ознак, враховуючи їх зміни у часі. Розроблено обчислювальні схеми та програмне забезпечення для аналізу даних гідрохімічного моніторингу.

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

Предложена информационная технология определения групп похожих объектов по совокупности исследуемых признаков, учитывая их временные изменения. Разработаны вычислительные схемы и программное обеспечение для анализа данных гидрохимического мониторинга.

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

Information technology for determining groups of similar objects on the set of the studied features taking into account their changes over time has been proposed. Computational schemes and software for data analysis of hydrochemical monitoring has been developed.

Key words: *cluster analysis, hydrochemical monitoring, time series, information technology.*

Вступ. За процесом видобутку та збагачення залізних руд, унаслідок 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.

Одне з основних місць у системі гідрохімічного моніторингу

займає обґрунтування пунктів спостережень, об'ємів і періодичності гідрохімічних випробувань. Це може бути вирішено на підставі гідрохімічного районування територій. Використовуючи районування, можна до певної міри уніфікувати водоохоронні заходи в межах виділених груп та районів. Визначивши пріоритетні водоохоронні заходи для одного району, планувати і впроваджувати їх для всієї виділеної групи [1].

Постановка задачі. Метою роботи є визначення групи пунктів спостереження, що характеризуються схожим хімічним складом води р. Інгулець поблизу ВАТ «Центрального гірничо-збагачувального комбінату» за досліджуваними компонентами для правильного планування природоохоронних заходів та керування якістю вод річки. Проби води відбиралися у 5 пунктах спостереження: селище Тернівка, створи балок: Мала Лозоватка, Велика Лозоватка, Завертана, північна частина Карачунівського водосховища. Аналіз проводився за вмістом головних іонів у воді річки Інгулець: HCO_3^-, Cl^-, SO_4^{2-}, Ca^{2+}, Mg^{2+}, Na^+ та мінералізацією протягом наступних років: 1993–1995, 1997, 2001, 2003, 2005–2007. Представимо досліджувані дані у вигляді наступної структури $X = \{x_{ijt}\}, i = \overline{1,N}, j = \overline{1,p}, t = \overline{1,T}$. Тобто маємо N об'єктів, які характеризуються p ознаками, значення яких змінюються протягом T моментів часу, x_{ijt} – значення j-го показника i-го об'єкта в момент часу t. Необхідно виділити групи об'єктів, схожих між собою за усіма досліджуваними ознаками у заданому періоді спостережень.

Таким чином, метою роботи є розробка технології, яка дозволить виділити групи об'єктів, схожих між собою за усіма досліджуваними ознаками, що змінюються у часі та врахувати часові зміни досліджуваних показників.

Основний матеріал. Методи кластерного аналізу, що дозволяють виявити групи схожих між собою об'єктів не можна явним виглядом застосувати до вирішення поставленої задачі, оскільки в якості вхідної інформації вони використовують матрицю «об'єкти-ознаки» $X = \{x_{ij}\}, i = \overline{1,N}, j = \overline{1,p}$, де x_{ij} – конкретне значення, а не часовий ряд, як у нашому випадку. Тому в даній роботі пропонується технологія виділення груп об'єктів, схожих між собою за набором ознак, які змінюються у часі, що ґрунтується на методах колективної кластеризації та складається з наступних етапів: визначення груп об'єктів для кожного моменту часу t, побудова узгодженої матриці подібності, отримання узагальнюючого розв'язку задачі.

Визначення груп об'єктів. Представимо вихідні дані у вигляді групи двовимірних матриць $X^{(t)} = \{x_{ij}^{(t)}\}, i = \overline{1,N}, j = \overline{1,p}, t = \overline{1,T}$. Окремо до кожної з них застосовуємо відомі методи кластерного аналізу [2–6] для визначення структури даних у певний момент часу. Тобто, отримаємо набір угрупувань $G_t = \{g_1^{(t)}, g_2^{(t)}, \ldots, g_k^{(t)}\} : t = \overline{1,T}$, де $g_i^t -$ список об'єктів, що потрапили до *i*-го кластера в *t*-му угрупуванні, *k* – кількість кластерів. Важливо проводити оцінку якості та вибір найбільш вірогідного результату кластеризації, щоб кожне з угрупувань G_t найкраще відповідало структурі досліджуваних даних.

Побудова узгодженої матриці подібності. На основі отриманого набору угрупувань будуємо узгоджену матрицю подібності об'єктів $S = \{s_{ij}; i, j = \overline{1,N}\}$, де N – кількість об'єктів, $s_{ij} -$ частота віднесення *i*-го та *j*-го об'єктів до одного кластера.

Процес побудови можна представити у вигляді наступного алгоритму:

1. Створюємо матрицю $S = \{s_{ij} = 0; i, j = \overline{1,N}\}$.

2. Розглядаємо по черзі угрупування $G_t : t = \overline{1,T}$. Якщо *i*-й та *j*-й об'єкти відносяться до одного кластера в момент *t*, то s_{ij} збільшуємо на одиницю.

3. Зводимо елементи матриці подібності до одиничної шкали $S = \{s_{ij} = \frac{s_{ij}}{T}; i, j = \overline{1,N}\}$. Чим ближче значення s_{ij} до одиниці, тим вища ймовірність віднесення об'єктів і та j до одного кластеру.

Отримання узагальнюючого колективного розв'язку. Для отримання підсумкового узагальнюючого розв'язку необхідно виділити групи об'єктів на основі обчисленої матриці подібності. Для цього застосовуємо графовий алгоритм найкоротшого незамкненого шляху. В якості матриці близькості використовуємо матрицю $S' = \{s'_{ij} = 1 - s_{ij}; i, j = \overline{1,N}\}$. Тобто чим більше подібні об'єкти *i* та *j* за матрицею *S*, тим менша відстань між ними у матриці S'.

Розглянемо результати запропонованої технології застосованої до даних гідрохімічного моніторингу, що проводиться Криворізькою геолого-гідрогеологічною партією по р. Інгулець (Кривбас).

За постановкою завдання маємо дев'ять багатовимірних вибірок «пункти спостереження – досліджувані компоненти», кожна з яких відповідає певній даті відбору проб води з річки. За допомогою

методів кластерного аналізу (ієрархічних, К-середніх, графового, Forel) та технології оцінки якості [7–9] та підвищення стійкості результатів [6; 10] отримуємо угрупування схожих між собою об'єктів для кожного окремого моменту часу. Для зведення даних до єдиного масштабу попередньо проводимо стандартизацію.

На рис. 1, 2 представлені розбиття, що відповідають даним 2001 та 2007 років. За станом води відібраних проб пункти спостереження розподілилися на 2 групи наступним чином: у 2001 році до першого класу увійшли с. Тернівка, Мала Лозоватка, до другого – Велика Лозоватка, Завертана, північна частина Карачунівського водосховища; у 2007 році в перший кластер виділено с.Тернівка, другий містить усі інші об'єкти дослідження.

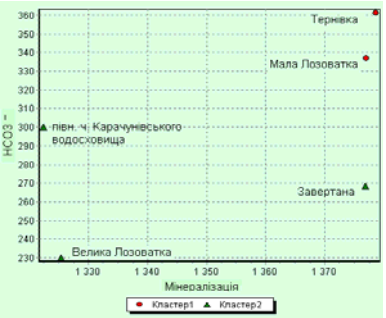

Рис. 1. Результати кластеризації за даними 2001 р.

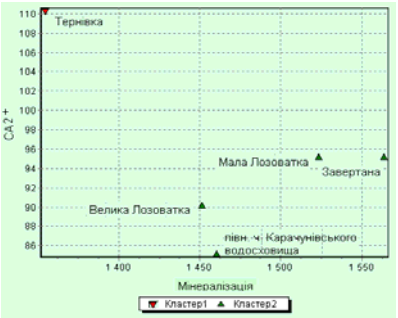

Рис. 2. Результати кластеризації за даними 2007 р.

Такий підхід визначає угрупування пунктів спостереження на певну дату, що дозволяє аналізувати зміни схожості об'єктів за часом.

Для визначення об'єктів схожих між собою на всьому часовому проміжку спостереження за всіма досліджуваними показниками одночасно з метою відображення загальної картини перебігу певних гідрохімічних процесів у воді річки застосовуємо запропоновану технологію часової кластеризації. За результатами аналізу було виділено дві групи об'єктів: перша складається з пункту спостереження у с. Тернівка, друга містить всі інші об'єкти дослідження.

Проведемо кластерний аналіз за даними представленими у вигляді матриці «пункти спостереження – значення фіксованого показника у часі», тобто визначимо пункти спостережень, які є схожими між собою на досліджуваному часовому проміжку за одним з показників. Отримані результати (при виборі будь-якого показника) співпадають з результатами продемонстрованими запропонованою технологією

часової кластеризації, що підтверджує її адекватність. На рис. 3 – 4 представлено діаграми розсіювання об'єктів за значеннями показників HCO_3^- (рис.3) та Ca^{2+} (рис.4) у часовому діапазоні.

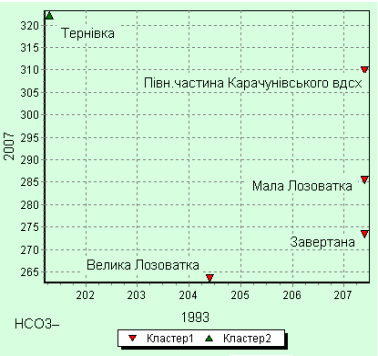

Рис. 3. Результати кластеризації за значеннями показника HCO_3^-

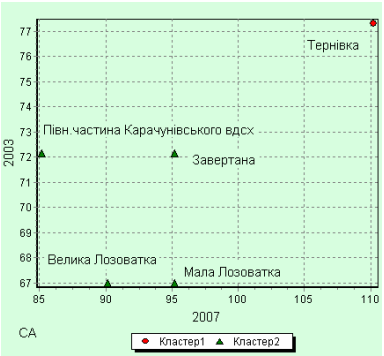

Рис. 4. Результати кластеризації за значеннями показника Ca^{2+}

Отримане поділення на кластери, за думкою фахівців з предметної області, відповідає дійсній гідрологічній та гідрохімічній ситуації на досліджуваній ділянці р. Інгулець. До Карачунівського водосховища подається вода каналом «Дніпро – Інгулець». Найбільший вплив цього каналу відзначається на верхній ділянці, що вивчається, а саме в районі селища Тернівка. Нижче за течією, вплив дніпровської води на формування хімічного складу води у річці Інгулець менший, більший вплив надає гірничо-збагачувальний комбінат (фільтраційні втрати з гідротехнічних споруд комбінату, пиління хвостосховища та інші).

Висновки. У даній роботі запропоновано технологію виділення груп об'єктів, схожих між собою за набором ознак, що змінюються за часом. Розроблено обчислювальні схеми та створено програмне забезпечення, це дозволило вирішити прикладну задачу визначення груп пунктів спостереження, що характеризуються схожим хімічним складом води у р. Інгулець за досліджуваними компонентами для правильного планування природоохоронних заходів та керування якістю вод річки. Запропонована технологія може бути застосована і в інших предметних галузях для виділення груп схожих об'єктів з урахуванням часових змін досліджуваних ознак.

Бібліографічні посилання

1. **Шерстюк Н. П.** Особливості гідрохімічних процесів у техногенних та природних водних об'єктах Кривбасу / Н. П. Шерстюк, В. К. Хільчевський. – Д., 2012. – 263 с.

2. **Мандель И. Д.** Кластерный анализ / И. Д. Мандель. – М., 1988. – 176 с.

3. **Айвазян С. А.** Классификация многомерных наблюдений / С. А. Айвазян, З. И. Бежаева, О. В. Староверов. – М., 1974. – 240 с.

4. **Загоруйко Н. Г.** Прикладные методы анализа данных и знаний / Н. Г. Загоруйко. – Новосибирск, 1999. – 270 с.

5. **Jain A. K.** Data Clustering: A Review / A. K. Jain, M. N. Murty, P. J. Flunn // ACM Computing Surveys, Vol. 31. – № 3. – September 1999. – P.265–323.

6. **Бериков В. С.** Современные тенденции в кластерном анализе / В. С. Бериков, Г. С. Лбов // Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы», 2008. – 26 с.

7. **Приставка О. П.** Підтримка прийняття рішень в задачах кластерного аналізу / О. П. Приставка, М. Г. Сидорова // Актуальні проблеми автоматизації та інформаційних технологій : зб. наук. праць. – 2011. – Т.15. – С.117–125.

8. **Гусарова Л.** Проверка обоснованности кластерного решения / Л. Гусарова, И. Яцкив // Proceedings of International Conference RelStat'03. – Vol. 2.

9. **Halkidi M.** Clustering validity assessment: Finding the optimal partitioning of a data set/ M. Halkidi, M.Vazirgiannis // Proceedings of ICDM, 2001. –
P. 187–194.

10. **Бирюков А. С.** Решение задач кластерного анализа коллективами алгоритмов / А.С. Бирюков, В.В. Рязанов, А.С. Шмаков // Журн. Вычислит. математ. и математ. физики. – 2008. – Т. 48, № 1. – С. 176–192.

Надійшла до редколегії 12.06.2012