СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ
МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

Кваліфікаційна наукова
праця на правах рукопису

БАРБАРУК ЛІНА ВІКТОРІВНА

УДК 004.05+004.415.5

Д И С Е Р Т А Ц І Я

МОДЕЛІ ТА МЕТОД ОБРОБКИ ВЕЛИКИХ ДАНИХ В ІНФОРМАЦІЙНО-АНАЛІТИЧНИХ СИСТЕМАХ МОНІТОРИНГУ ВОДНИХ ОБ'ЄКТІВ

05.13.06 – інформаційні технології

12 - інформаційні технології
галузь знань

Подається на здобуття наукового ступеня кандидата технічних наук

Дисертація містить результати власних проваджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело.

_____ Барбарук Ліна Вікторівна
підпис

Науковий керівник Скарга-Бандурова Інна Сергіївна,
доктор технічних наук, професор

Всі примірники дисертації ідентичні за змістом.
Вчений секретар спеціалізованої
Вченої ради К 29.051.16 /Сафонова С.О./

Сєвєродонецьк – 2021

АНОТАЦІЯ

Барбарук Ліна Вікторівна. Моделі та метод обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів. – Кваліфікаційна наукова праця на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 "Інформаційні технології". – Східноукраїнський національний університет імені Володимира Даля, Сєвєродонецьк, 2021.

Дисертацію присвячено вирішенню актуального науково-прикладного завдання підвищення ефективності роботи інформаційно-аналітичних систем моніторингу водних об'єктів завдяки розробці та практичному застосуванню моделей та методів інформаційної технології обробки великих даних.

Об'єктом дослідження є процеси обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів, а предметом досліджень є моделі та метод інформаційної технології обробки та візуалізації великих даних, використовуваних в інформаційно-аналітичних системах моніторингу водних об'єктів.

У роботі виконано аналітичний огляд сучасних методів і технологій обробки великих даних для кращого управління водними ресурсами. Встановлено, що: (1) незважаючи на інтенсивний розвиток технологій інтелектуального зондування та моніторингу, для багатьох водних ресурсів все ще бракує інтегрованих оцінок на основі великих даних, для підтримки динамічної оцінки якості води, моніторингу та управління наглядом; (2) зростаючі вимоги до даних вимагають збільшення потужності та ефективності інструментів та методів їх обробки. З експоненціальним зростанням даних, традиційні алгоритми видобутку та аналізу даних не можуть в повній мірі задовольнити потреби обробки даних, що обумовлює необхідність розробки і використання нових моделей обробки з розумною обчислювальною вартістю; (3) при візуалізації великих даних, сигнали, що надходять до системи моніторингу

можуть бути занадто щільні та розмиті, що призводить до проблем з аналізом тривалих записів не дозволяють швидко орієнтуватися в даних, що вимагає удосконалення технологій візуалізації. За результатами аналізу сучасного стану та тенденцій розвитку систем моніторингу водних об'єктів показано необхідність розробки ефективних моделей і методів інформаційних технологій обробки великих даних. Обґрунтовано проведення дисертаційних досліджень в межах вирішення наступних основних завдань: розробка і використання нечітких моделей для реалізації інтегрованих оцінок води на основі великих даних у реальному часі, удосконалення підходів до нечіткої кластеризації на випадок великих даних, удосконалення технологій візуалізації великих даних, розробка та впровадження інформаційної технології та засобів підтримки прийняття рішень в інформаційно-аналітичній системі моніторингу водних об'єктів. У роботі поставлене та вирішене актуальне науково-прикладне завдання підвищення якості роботи інформаційно-аналітичних систем моніторингу водних об'єктів завдяки розробці та практичному застосуванні моделей та методів інформаційної технології обробки великих даних.

Методологічну основу дисертаційного дослідження становлять методи теоретико-множинного опису, теорії нечітких множин, які використовувались при розробці моделі для аналітичної обробки великих даних; методи нечіткої евристичної кластеризації та інтуїтивістьскої теорії нечітких множин, міри подібності та метод множника Лагранжа, що використовувались для удосконалення методу обробки великих даних та автоматичного маркування нечітких кластерів, в контексті аналізу якості водойм рибогосподарського призначення; методи Data Mining, MapReduce, OLAP, технології сховищ даних, що використовувались для розробки інформаційної технології; методи візуалізації кореляційних матриць, зокрема хордові діаграми, криві Безьє, технології спрощення полігональних ланцюгів, що використовувалися для побудови моделей візуалізації великих даних.

Вперше одержано модель для аналітичної обробки великих даних системи

моніторингу водних об’єктів на основі формалізації її атрибутів та інтерпретації невизначеності оцінки якості води у вигляді лінгвістичних змінних, що дозволило надати інтегровану характеристику стану водних об’єктів, для подальшого прийняття рішення;

Удосконалено метод нечіткої кластеризації с-середніх для великих даних, в якому, на відміну від існуючих, виконано узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів для інтуїтивістських нечітких множин, що дозволило застосовувати автоматичну розмітку при обробці великих даних і покращити швидкість і конвергенцію алгоритму кластеризації;

Удосконалено технологію візуалізації великих даних за рахунок застосування кореляційних матриць та технології спрощення полігональних ланцюгів, що дозволило проводити динамічну візуалізацію та зменшити час на аналіз даних, отриманих з системи моніторингу, зокрема тривалих записів.

Дістала подальшого розвитку інформаційна технологія обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів шляхом її адаптації до завдань контролю та управління рибогосподарських підприємств, що забезпечує семантичну основу для комплексної автоматизації у частині реалізації основних аналітичних функцій.

Усі теоретичні положення дисертації доведено до конкретних інженерних рішень із застосуванням запропонованої інформаційної технології обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів і вибору варіантів її використання в системі підтримки прийняття рішень. Використання моделей та інформаційної технології обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів, зокрема у виробничих процесах рибогосподарських підприємств, дозволило зробити висновок, про їх ефективність в частині підвищення точності інтегральної оцінки якості вод, зменшення часу на прийняття рішень щодо якості вод водойм рибогосподарського призначення та зменшення часу на пошук

невідповідностей і ретроспективний аналіз великих даних.

Розроблені моделі, методи, інформаційне та програмне забезпечення використані при виконанні науково-дослідних проектів “Дослідження стратегій та механізмів прийняття рішень для інтегрованого управління водними ресурсами” 12.2016-07.2020, (№ ДР 0116U005784) (акт впровадження від 12.01.2021 р.), “Проектування системи моніторингу та контролю водних об’єктів на основі технології інтернет речей” 01.2020-12.2023, (№ ДР 0120U100421), м. Сєвєродонецьк; при виконанні міжнародного проекту Європейського Союзу ERASMUS+ ALIOT 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP “Internet of Things: Emerging Curriculum for Industry and Human Applications”.

Результати дисертаційного дослідження впроваджені в управлінні Державного агенства рибного господарства у Луганській області(акт впровадження від 10.02.2021 р.); в навчальний процес кафедри комп’ютерних наук та інженерії Східноукраїнського національного університету ім. В. Даля при вивченні дисциплін “Комп’ютерне моделювання процесів та систем”, “Методи машинного навчання”, “Data-mining та бізнес-аналітика” (акт впровадження від 12.01.2021 р.).

Результати дисертації викладені у 18 друкованих роботах: з них 8 статей у виданнях, які зазначені в переліку фахових видань України з технічних наук, 1 стаття у фаховому виданні іншої держави, що входить до Європейського Союзу, 1 стаття у збірнику, що індексується у базі даних Scopus, 8 тез доповідей всеукраїнських та міжнародних конференцій.

Ключові слова: великі дані, інформаційно-аналітична система, моніторинг, водний об’єкт, підтримка прийняття рішень, аналіз даних.

Список публікацій здобувача за темою дисертації

Праці, які відображають основні наукові результати дисертації

[1] Skarga-Bandurova I., Krytska Y., Velykzhanin A., Barbaruk L., Suvorin O., Shorohov M. “Emerging Tools for Design and Implementation of Water Quality Monitoring Based on IoT”, *Complex Systems Informatics and Modeling Quarterly*. Published online by RTU Press, https://csimq-journals.rtu.lv Article 138, Issue 24, September/October 2020, pp. 1–14, 2020 https://doi.org/10.7250/csimq.2020-24.01.

[2] Skarga-Bandurova I., Krytska Y., Shorohov M., Suvorin O., Barbaruk L., Ozheredova M. “Towards Development IoT-based Water Quality Monitoring System”, *Proceedings of the 7th IEEE International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pp. 140-145, 2019. doi: 10.1109/FiCloudW.2019.00038 (*Scopus*).

[3] Барбарук Л.В., Зубарев Д.В., Медведєв Є.М., Барбарук В.М. “Використання нечітких моделей для планування збиральних робіт у різних кліматичних умовах”, *Вісник Східноукраїнського національного університету ім. В. Даля*, №6 (247), С. 175-179, 2018.

[4] Скарга-Бандурова І.С., Грушка М.О., Барбарук Л.В. “Підходи до ефективного спрощення та візуалізації великих наборів даних”, *Вісник Національного технічного університету “Харківський політехнічний інститут”. Зб. наукових праць. Серія: Інформатика та моделювання.* Харків: НТУ “ХПІ”, №. 50 (1271), С. 55-65, 2017. doi: 10.20998/2411-0558.2017.50.10.

[5] Барбарук Л.В., Суворін О.В. “Система аналізу даних та підтримки прийняття рішень з інвентаризації промислових відходів”, *Вісник Східноукраїнського національного університету ім. В. Даля*, №8 (238). С. 5-12. 2017.

[6] Скарга-Бандурова І.С., Барбарук Л.В., Сіряк Р.В. “Моделі багатовимірних структур даних для аналізу природоохоронної діяльності

промислових підприємств", *Зб. наукових статей Комп'ютерне моделювання в хімії, технологіях і системах сталого розвитку.* Київ: НТУУ "КПІ", С. 185-190, 2014.

[7] Барбарук В.М., Барбарук Л.В. "Методи моделювання складних систем", *Вісник Східноукраїнського національного університету ім. В. Даля*, № 15(186), С. 132-136, 2012.

[8] Барбарук В.М., Барбарук Л.В. "Засоби представлення багатомірних даних в інформаційно-аналітичних системах", *Міжвузівський збірник "Комп'ютерно-інтегровані технології: освіта, наука, виробництво"* Луцьк, №8. С. 5-10, 2012.

[9] Барбарук Л.В. "Применение многомерных моделей данных в системах аналитической обработки данных", *Вісник Східноукраїнського національного університету ім. В. Даля,* №11, Ч. 2, С. 180-184, 2011.

[10] Рязанцев О.І., Барбарук В.М., Барбарук Л.В. "Комплексна оцінка екологічної небезпеки території промислового виробництва", *Вісник Східноукраїнського національного університету ім. В. Даля*, №7(154), Ч.2, С. 136-140, 2010.

Опубліковані праці апробаційного характеру:

[11] Skarga-Bandurova I., Krytska Y.O., Barbaruk L.V. "Application of Internet of Things for long term water quality monitoring", *Проблеми інформатики і моделювання. Тезиси дев'ятнадцятої міжнар. наук.-техн. конф*., Харків: НТУ "ХПІ", С. 77, 2019.

[12] Барбарук Л.В., Михайличенко С.О. Бездротова система віддаленого моніторингу сільськогосподарських параметрів. *Матеріали VII Всеукр. науково-практ. конф. «Електронні апарати та системи. Проблеми створення. Перспективи розвитку»*, Сєвєродонецьк : Східноукр. нац. ун-т ім. В. Даля, 2017. С. 206-208.

[13] Михайличенко С.О., Барбарук Л.В. "Вибір протоколів маршрутизації для організації бездротових сенсорних мереж", *Матеріали всеукр. наук.-практ.*

конф. "Майбутній науковець – 2017" 1 груд. 2017 р., м. Сєвєродонецьк. укл. В. Ю. Тарасов. Сєвєродонецьк : Східноукр. нац. ун-т ім. В. Даля, С. 296-297, 2017.

[14] Барбарук Л.В., Добрецова А.О. "Застосування алгоритму Краскала для синтезу древоподібних глобальних мереж", *Технологія-2016 : матеріали міжнар.наук.-техн. конф.*, 22-23 квіт. 2016 р., м. Сєвєродонецьк. Ч. II [укл. : Тарасов В.Ю.] Сєвєродонецьк : [Східноукр. нац. ун-т ім. В. Даля], С. 152-154, 2016.

[15] Барбарук Л. "Методы формирования многомерных отчетов для задач учета ресурсопотребления", *Theoretical and Applied Computer Science and Information Technology: Proceedings of the first International Conference TACSIT-2015.* Severodonetsk: East Ukrainian National University, P. 35-40, 2015.

[16] Барбарук Л.В., Скарга-Бандурова І.С. Методы оперирования многомерными данными в хранилище данных опасных промышленных отходов. *Системи обробки інформації.* Вип. 2 (118). Т.2 Проблеми і перспективи розвитку ІТ-індустрії. С. 223, 2014.

[17] Барбарук В.М., Барбарук Л.В. "Використання технології багатовимірних баз даних при проектуванні складних інформаційних систем", *Матеріали IX міжнародної конференції "Strategy of Quality in Industry and Education: Part 3."* Varna, Bulgaria: International Scientific Journal Acta Universitatis Pontica Euxinus. Special number, С. 440-443, 2013.

[18] Барбарук Л.В. "Сложные информационные системы", *Матеріали XIV всеукраїнської науково-практичної конференції "Технологія": ч. 2.* Сєвєродонецьк: Технол. ін-т Східноукр. Нац. ун-ту ім. В. Даля (м. Сєвєродонецьк), С. 129-131, 2011.

ANNOTATION

Lina Barbaruk. Models and method for large data processing in integrated monitoring system of water bodies. – Manuscript copyright.

The dissertation on competition of a scientific degree of the candidate of technical sciences on a specialty 05.13.06 "Information technology". - Volodymyr Dahl East Ukrainian National University, Severodonetsk, 2021.

The thesis is devoted to the actual scientific and applied task of increase of efficiency of integrated monitoring system of water bodies by means of the development and practical application of the ad-hock models and methods of information technology for large data processing. The object of research is the processes of large data processing in integrated monitoring system of water bodies, and the subject of research is the models and method of information technology of processing and visualization of large data used in integrated monitoring system.

The thesis presents an analytical review of modern methods and technologies of big and large data processing for better water resources management. The analysis of the current state and trends in the development of water monitoring systems shows the further need to develop effective models and methods of information technology for large data processing. Therefore, the current scientific and applied task of improving the quality of information and analytical systems for monitoring of water bodies through the development and practical application of models and methods of information technology for large data processing is set and solved. The methodological basis of the research comprises the methods of set-theoretic description, fuzzy set theory used in the model development for analytical processing of large data; fuzzy heuristic clustering, intuitive fuzzy set theory, similarity measures, and Lagrange multiplier method used to improve the data processing and automatic labelling of fuzzy clusters, in the context of quality analysis of fishery reservoirs; Data Mining, MapReduce, OLAP, data warehouse technologies used for information technology development; methods of visualization of

correlation matrices, in particular chord diagrams, Bezier curves, technologies of simplification of polygonal chains used to build models of visualization of big data.

The main scientific results are: A new model for analytical large data processing of water monitoring system based on formalization of its attributes and interpretation of uncertainty of water quality assessment in the form of linguistic variables was obtained, which allowed to provide integrated characteristics of fisheries water bodies for further decision making. The method of fuzzy clustering for large data has been improved, which, in contrast to the existing ones, generalizes the procedure of automatic labelling of fuzzy clusters obtained using heuristic algorithms for intuitive fuzzy sets, which allowed the use of automatic mark-up in processing large data and improve convergence of the clustering algorithm. The technology of visualization of large data has been improved by means of correlation matrices and the technology of simplification of polygonal chains, which allowed carrying out dynamic visualization and reducing the time for analysis of data obtained from the monitoring system, including long recordings. The information technology for large data processing in integrated monitoring system of water bodies by its adaptation to tasks of control and management of fisheries enterprises, which provides a semantic basis for integrated automation in terms of implementation of basic analytical functions.

All theoretical provisions of the dissertation are brought to concrete engineering decisions with application of the information and its implementation at the decision support system. This enables to reduce the time to make decisions on the quality of water bodies in fisheries and reduce the time to find discrepancies and retrospective analysis of a large datasets.

Key words: large data, big data, integrated monitoring system, monitoring, water body, decision support, data analysis.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

БСК	Біологічне споживання кисню
РК	Розчинний кисень
СД	Сховище даних
СППР	Система підтримки прийняття рішень
ХСК	Хімічне споживання кисню
DO	Dissolved Oxygen (розчинний кисень)
ETL	Extract, Transform, Load
FCM	Fuzzy C-Means (алгоритм нечітких с-середніх)
FFCM	Fast Fuzzy C-Means (швидкий алгоритм нечітких с-середніх)
HOLAP	Hybrid (Гібридний) OLAP
IFS	Intuitionistic fuzzy sets (інтуїтивістський нечіткий набір)
IoT	Internet of Things (Інтернет речей)
MOLAP	Multidimensional (Багатовимірний) OLAP
OLAP	Online Analytical Processing (аналітична обробка в реальному часі)
OLTP	On Line Transaction Processing (обробка он-лайн транзакцій)
PCA	Principal Component Analysis (аналіз головних компонент)
ROLAP	Relational (Реляційний) OLAP
TDS	Total Dissolved Solids (загальна кількість розчинених твердих речовин)
WQI	Water Quality Index (індекс якості води) Integral Water Quality Index (інтегральний індекс якості води)

ВСТУП

Актуальність теми дослідження. Якість води є одним з найважливіших факторів здорової екосистеми. Чиста вода підтримує різноманітність рослин та дикої природи. Особливі потреби до якості води в аквакультурі, оскільки неякісна вода може впливати на здоров'я та ріст риб. У ставках і штучних водоймах хімічні та фізичні показники води можуть швидко змінюватись, оскільки риби використовують воду для проживання, харчування, розмноження, тощо. Отже, ефективне використання води, її якість, потреби аквакультури та способи управління факторами якості води є основними чинниками екологічного благополуччя. Найкращим рішенням для ефективного використання води є можливість постійного моніторингу в реальному часі її фізико-хімічних параметрів, аналіз обсягів споживання, попиту та результатів господарської діяльності. Таку інформацію можна отримати лише за допомогою моніторингових систем і засобів потужної аналітики.

Дослідженням обробки даних у водній галузі займаються вчені всього світу, зокрема, значний внесок у розвиток систем обробки та аналізу даних зробили видатні вітчизняні та зарубіжні вчені: В.П. Ковальчук, О.М. Клименко, В.Б. Мокін, А.М. Прищепа, А.В. Яцишин, R. Allen, C. Briese, W. Wagner, G. Wong та ін. Стрімкий розвиток інформаційних технологій, систем моніторингу та обробки великих даних показав необхідність організації підтримки рішень в умовах нечіткої інформації. Серед вчених, які займалися розвитком теорії та методів обробки нечіткої інформації варто відзначити L. Zadeh, E. Mamdani, M. Sugeno, В.Д. Шапіро, Є.В. Бодянський, А.О. Каргін, Ю.П. Кондратенко, О.В. Леоненков, Д.О. Поспелов, А.П. Ротштейн та ін. Більшість світових експертів з питань води та водного господарства, наголошують на необхідності інвестувати в аналітичні інструменти та рішення з великими даними як наріжний камінь, для зменшення навантаження на водні ресурси. Разом з тим, будь-яка моніторингова система, що працює в режимі реального часу, зокрема система

моніторингу водного середовища, стикається з великим тиском при обробці даних. З одного боку, традиційні аналітичні інструменти не здатні підтримувати обробку великих обсягів даних, з іншого боку - засоби аналітики Big Data, у більшості випадків, не орієнтовані на обробку даних систем моніторингу водних об'єктів. Хоча багато галузей промисловості вже використовують потужність великих даних та додатків аналітики для безперебійної роботи, є багато областей, куди аналітика все ще не змогла повноцінно проникнути та заощадити природні ресурси, до таких областей відноситься і вода. Таким чином, існує об'єктивне протиріччя між зростаючою кількістю впроваджень високотехнологічних систем он-лайн моніторингу водних об'єктів і наявними технологіями управління та обробки великих наборів даних, які ці системи здатні збирати.

Виходячи з вищевикладеного, *актуальним науковим завданням* є розроблення моделей та методу інформаційної технології, здатних забезпечити ефективну обробку та використання великих даних, отримуваних від систем моніторингу водних об'єктів.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалась у відповідності з науково-технічними планами кафедри комп'ютерних наук та інженерії Східноукраїнського національного університету імені Володимира Даля, у відповідності до державних програм і планів НДР, а також міжнародних проектів і програм:

– НДР "Дослідження стратегій та механізмів прийняття рішень для інтегрованого управління водними ресурсами" (Східноукраїнський національний університет імені Володимира Даля, № ДР 0116U005784, 12.2016-07.2020 рр.);

– НДР "Проектування системи моніторингу та контролю водних об'єктів на основі технології інтернет речей", (№ ДР 0120U100421, 01.2020-12.2023 рр.);

– Проект Європейського Союзу ERASMUS+ ALIOT 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP "Internet of Things: Emerging Curriculum for Industry and Human Applications" (2016-2020 pp.).

Роль автора в зазначених науково-дослідних роботах і проектах, у яких дисертант був безпосереднім виконавцем, полягає у проведені аналізу моделей та методів обробки великих даних в системах моніторингу водних об’єктів; виборі набору базових компонент для прийняття рішень при обробці моніторингових даних якості води; розробленні методів і засобів моніторингу для автоматизації завдань аналітичної обробки великих даних; розробці та імплементації моделей для роботи з великими даними; практичній реалізації розроблених моделей та методів в інструментальний засіб, призначений для підтримки рішень при управлінні водним господарством.

Мета й завдання дослідження. *Метою* дисертації є підвищення ефективності роботи інформаційно-аналітичних систем моніторингу водних об’єктів завдяки розробці та практичному застосуванню моделей та методів інформаційної технології обробки великих даних.

Для досягнення мети дослідження в дисертації сформульовані та вирішені наступні *задачі:*

– аналіз моделей та методів обробки великих даних в системах моніторингу водних об’єктів;

– розроблення математичних моделей для аналітичної обробки великих даних системи моніторингу водних об’єктів на основі формалізації її атрибутів та інтерпретації невизначеності оцінки якості води у вигляді лінгвістичних змінних;

– розроблення методу нечіткої кластеризації c-середніх для великих даних, шляхом узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів для інтуїтивістських нечітких даних;

– розроблення інформаційної технології на основі запропонованих моделей і методу для обробки великих даних та підтримки прийняття рішень в інформаційно-аналітичній системі моніторингу водних об’єктів;

– розроблення засобів візуалізації великих даних;

– реалізація програмних засобів та елементів інформаційної технології

обробки великих даних в інформаційно-аналітичній системі моніторингу водних об’єктів та впровадження отриманих результатів.

Об'єкт дослідження – процеси обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів.

Предмет дослідження – моделі, методи та програмні засоби інформаційної технології забезпечення обробки великих даних.

Методи дослідження. В основу методології дослідження були покладені принципи системного аналізу для декомпозиції процесу обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів; методи пошукового аналізу, технології роботи з літературними джерелами для аналізу перспективних технологій обробки великих даних, оцінювання потенціалу інноваційних технологій моніторингу; методи теоретико-множинного опису, теорії ймовірностей, теорії нечітких множин – для розробки моделі для аналітичної обробки великих даних; нечітка евристична кластеризація с-середніх, інтуїтивістська теорія нечітких множин, міри подібності для нечітких множин з множинними атрибутами, метод множника Лагранжа – для удосконалення методу обробки великих даних та автоматичного маркування нечітких кластерів у контексті аналізу якості водойм рибогосподарського призначення; методи візуалізації кореляційних матриць, зокрема хордові діаграми, криві Безьє, технології спрощення полігональних ланцюгів – для побудови моделей візуалізації великих даних. Для розробки інформаційної технології використовувалися методи Data Mining, MapReduce, OLAP, технології сховищ даних.

Наукова новизна результатів досліджень полягає у наступному.

1) *вперше розроблено* модель для аналітичної обробки великих даних системи моніторингу водних об’єктів на основі формалізації її атрибутів та інтерпретації невизначеності оцінки якості води у вигляді лінгвістичних змінних, що дозволило надати інтегровану характеристику стану водних об’єктів, для подальшого прийняття рішення;

2) *удосконалено метод* нечіткої кластеризації, в якому, на відміну від існуючих, виконано узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів для інтуїтивістських нечітких даних, що дозволило застосовувати автоматичну розмітку при обробці великих даних і покращити конвергенцію алгоритму кластеризації;

3) *удосконалено* технологію та засоби візуалізації великих даних за рахунок використання хордових діаграм та технології спрощення полігональних ланцюгів, що дозволяють зберегти баланс між вимогами до деталізації вихідного зображення та отриманням мінімальної кількості точок на екрані систем он-лайн моніторингу водних об’єктів, знижуючи час на пошук невідповідностей і ретроспективний аналіз великих даних;

4) *набула подальшого розвитку* інформаційна технологія обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів шляхом її адаптації до завдань контролю та управління рибогосподарських підприємств, що забезпечує семантичну основу для комплексної автоматизації водних господарств у частині реалізації основних аналітичних функцій.

Практичне значення одержаних результатів полягає в розробленні програмного забезпечення інформаційної технології для обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів. Усі теоретичні положення дисертації доведено до конкретних інженерних рішень із застосуванням запропонованої інформаційної технології обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів і вибору варіантів її використання в системі підтримки прийняття рішень. Використання моделей та інформаційної технології обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів, зокрема у виробничих процесах рибогосподарських господарств та підприємств аквакультури, дозволило зробити висновок, про їх ефективність в частині підвищення точності інтегральної оцінки якості вод, зменшення часу на прийняття рішень щодо якості

вод водойм рибогосподарського призначення, зниження витрат на виробництво (до 15% під час тестової експлуатації) за рахунок зменшення часу на пошук невідповідностей і ретроспективний аналіз великих даних.

Теоретичні та практичні результати дисертаційної роботи впроваджено:

– в Управлінні Державного агентства рибного господарства у Луганській області (акт впровадження від 12.01.2021 р.);

– у Східноукраїнському національному університеті імені Володимира Даля при розробці навчальних курсів “Комп’ютерне моделювання процесів та систем”, “Методи машинного навчання”, “Data-mining та бізнес-аналітика” на кафедрі комп’ютерних наук та інженерії, виконанні міжнародного проекту ERASMUS+ ALIOT 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP “Internet of Things: Emerging Curriculum for Industry and Human Applications” - при розробленні магістерського курсу “Development and implementation of IoT based systems” (акт впровадження від 10.02.2021 р.).

Особистий внесок здобувача полягає в розробці нової моделі, методу, елементів інформаційної технології та програмних засобів, які забезпечують вирішення поставлених у дисертації завдань. Всі наукові результати, наведені в дисертаційній роботі та сформульовані в положеннях, що виносять на захист, отримані автором особисто. Роботи [9, 15, 18] опубліковано без співавторів.

У друкованих працях, опублікованих у співавторстві, здобувачеві належать: [7, 8] проведено аналіз моделей та методів обробки великих даних, [18] виконано класифікацію моделей обробки великих даних, [3] розроблено і досліджено нечіткі моделі для різних кліматичних умов, [10, 11] розроблено модель комплексної оцінки параметрів, [2] проведено узагальнення процедури автоматичного маркування нечітких кластерів, [2, 5] розроблено елементи інформаційної технології обробки великих даних та підтримки прийняття рішень в інформаційно-аналітичній системі, [6, 9, 15, 16] розроблено моделі багатовимірних структур даних, [4] запропоновано підходи до ефективного спрощення та візуалізації великих наборів даних, [1, 12, 13, 14, 17] досліджено

результати практичної реалізації технології обробки великих даних в системі моніторингу водних об’єктів.

Апробація результатів наукових досліджень. Основні положення та результати дисертаційної роботи обговорювалися на 12 науково-технічних конференціях і тематичних семінарах, у тому числі:

- міжнародній науковій конференції міжнародній науково-технічній конференції “Future Internet of Things and Cloud (FiCloudW)” (м. Стамбул, Турція, 2019 р.);
- міжнародній науково-технічній конференції “Проблеми інформатики і моделювання (ПІМ) ” (м. Харків-Кароліно Бугаз, 2019 р.);
- міжнародних науково-технічних конференціях “Технологія” (м. Сєвєродонецьк, 2016-2018 рр.);
- III міжнародній науково-технічній конференції студентів, магістрів та аспірантів «Інформатика, управління та штучний інтелект» (м. Харків, 2016 р.);
- міжнародних науково-технічних конференціях “Theoretical and Applied Computer Science and Information Technology (TACSIT)” (м. Сєвєродонецьк, 2015, 2017 рр.);
- всеукраїнських науково-практичних конференціях: “Електронні апарати та системи. Проблеми створення. Перспективи розвитку”, “Майбутній науковець” (м. Сєвєродонецьк, 2017 р.);
- міжнародній науково-технічній конференції “Комп’ютерне моделювання в хімії, технологіях і системах сталого розвитку” (м. Київ, 2014 р.);
- міжнародній науково-технічній конференції “Strategy of Quality in Industry and Education” (м. Варна, Болгарія, 2013 р.),
- наукових семінарах кафедри комп'ютерних наук та інженерії СНУ ім. В. Даля (м. Сєвєродонецьк, 2011-2020 рр.).

Публікації. За темою дисертації з викладенням її основних результатів опубліковано 18 наукових праць, з яких 8 статей [3-10] у фахових наукових виданнях України, включених до міжнародних наукометричних баз, 1 стаття [2]

у закордонному виданні, що внесено до міжнародної наукометричної бази Scopus; 1 стаття [1] у фаховому виданні іншої держави, що входить до Європейського Союзу, 8 публікацій [11-18] у збірниках матеріалів і праць міжнародних та всеукраїнських конференцій.

Структура та обсяг дисертації. Дисертація складається зі вступу, 4 розділів, висновків, списку використаних джерел зі 177 найменувань на 22 сторінках та 5 додатків. У додатках наведено перелік публікацій здобувача, акти впровадження та основні нормативні показники якості вод, використані для розрахунків. Загальний обсяг дисертації становить 180 сторінок, у тому числі: анотація на 9 сторінках, зміст на 4 сторінках, перелік умовних позначень на 1 сторінці, основний текст на 148 сторінках. Робота містить 38 рисунків і 20 таблиць.

РОЗДІЛ 1
АНАЛІЗ МОДЕЛЕЙ ТА МЕТОДІВ ОБРОБКИ ВЕЛИКИХ ДАНИХ В СИСТЕМАХ МОНІТОРИНГУ ВОДНИХ ОБ'ЄКТІВ

У розділі подано огляд перспективних технологій обробки великих даних, розглянуто потенціал інноваційних технологій моніторингу для кращого управління водними запасами. На основі аналізу виконано постановку задачі дослідження, обрано основні об'єкти для моделювання та математичний апарат, зокрема вирішено, що основними методами аналітичної обробки великих даних системи моніторингу водних об'єктів доцільно обрати нечіткі моделі, зокрема нечітку евристичну кластеризацію, інтуїтивістську теорію нечітких множин, тощо.

1.1 Особливості великих даних в системах моніторингу водних об'єктів

В епоху великих даних величезний обсяг інформації генерується з дуже високою швидкістю. Технологічні досягнення дозволяють інноваційному обладнанню для моніторингу приєднатися до традиційного обладнання для відбору зразків та збирати більше даних, таких як інформація про екосистему, для кращого управління водними ресурсами. У більшості випадків такі дані збираються з різних джерел, можуть мати різний формат і потребують детальної обробки майже в реальному часі. Однак, широке використання нових технологій все ще обмежено через проблеми взаємодії програмного забезпечення та обладнання для обміну даних, високу вартість обладнання, та обмежений штат спеціалістів, здатних ефективно застосовувати ці інструменти. Крім того, оскільки управління водним господарством стає більш масштабним та цілісним, вимоги до даних та їх аналіз стають все більш складними. Тому обробка великих даних для потреб аналітичних операцій стає однією з ключових проблем сьогодення.

1.1.1 Категорії великих даних

Термін “великі дані” стосується до дуже швидкого зростання неоднорідних потоків даних через збільшення використання інформаційних технологій, зростання Інтернету, використання даних соціальних мереж, мобільних мереж, IoT та інших пов’язаних та об’єктів, що обмінюються інформацією, яка в свою чергу, зростає швидше кожного дня. Ключовими рисами великих даних є обсяг (volume), швидкість (velocity) та різноманітність (variety), що складає оригінальну парадигму з трьома V, запропоновану в 2001 р. [52] для опису управління даними у трьох вимірах. Ця модель все ще діє, але нещодавно вона збагатилася додатковими 5, 7 і навіть 8 Vs (рис.1.1).

Рисунок 1.1 – Модель великих даних з 8 V

Складність використання моделі для аналізу стану водних ресурсів, характеризується:

– Величезними обсягами зібраних, проаналізованих та візуалізованих даних.

– Великою кількістю та різноманіттям джерел даних та їх масштабами.

– Зібрані дані є складними за розмірами, типами, якістю.

– Неоднорідністю даних, отриманих в результаті використання різних складних імітаційних моделей.

– Наявністю просторових даних та даних, зібраних дистанційно в режимі реального часу.

– Необхідністю фільтрації та зменшення розмірів даних та потребами в складному моделюванні у багатьох масштабах.

Останні зауваження є особливо суттєвими, оскільки в складних інформаційно-аналітичних системах, значна частина даних не представляє інтересу, і їх можна відфільтрувати та стиснути за порядком. Однак, однією з головних проблем є визначення цих фільтрів таким чином, щоб вони не відкидали корисну інформацію. Крім того, потрібні методи аналізу в режимі он-лайн, які можуть обробляти такі потокові дані на льоту, оскільки не має сенсу спочатку зберігати, а потім зменшувати дані.

Стосовно обсягів даних, здатних називатися великими даними, існує наступна категоризація [44], що умовно представляє 5 рівнів (табл. 1.1).

Таблиця 1.1 – Категорії, визначені в [44] для великих даних

Параметри			“Великі дані”		
Байт	10^6	10^8	10^{10}	10^{12}	$10^{>12}$
Розмір	середні	великі	величезні	велетенські	дуже великі

Великі дані - явище, яке не має чітких меж, і може бути представлене в необмеженому або навіть нескінченному накопиченні даних. Дані з Інтернету речей (IoT) та веб-трафіку все ще перевершують здатність фіксувати та аналізувати. Навіть більше, накопичені дані можна представити в різних форматах, більшість з яких не є структурованими. Тим не менше, головним питанням є не обсяг даних, а сфера їх застосування [16].

1.1.2 Основні компоненти та сценарії використання великих даних в інформаційно-аналітичних системах екологічного спостереження

Існує декілька підходів до класифікації інформаційно-аналітичних систем екологічного спостереження. Для даного дослідження було використано два основних класи [6, 23]:

(1) Традиційні системи екологічного спостереження, завданням яких є періодичний моніторинг, оцінка екологічного стану територій, аналіз природоохоронної діяльності суб’єктів господарювання в контексті їх впливу на природні водойми, водні ресурси та ін.

(2) Інформаційно-аналітичні системи моніторингу водних об’єктів (зокрема об’єктів аквакультури), завданням яких є моніторинг та візуалізація стану водойм в реальному часі, визначення та оперативне інформування про відхилення параметрів та ін.

Для інформаційно-аналітичних систем моніторингу водних об’єктів (рис. 1.2), які на відміну від традиційних систем екологічного спостереження, містять засоби он-лайн моніторингу, візуалізації, аналітики та підтримки прийняття рішень, нарощування технологій розуміння глибоких даних є життєво необхідним.

Великі дані надають широкі перспективи для прийняття рішень стосовно оптимального контролю, планування водних систем, аналізу впливу кліматичних змін, виявлення змін в екосистемі за допомогою дистанційного зондування, прогнозування природних змін, пом'якшення забруднення навколишнього середовища тощо.

Завдання великих даних у водних ресурсах передбачає не лише боротьбу з високими обсягами даних. Зокрема, проблеми щодо збору, зберігання, управління та аналізу даних також пов'язані з проблемами водних ресурсів, пов'язаних з великими даними. Однак, обробка великих даних не є тривіальною задачею, і вимагає спеціальних методів та підходів. Зокрема, через їх величезний

обсяг неможливо завантажити всі дані в пам'ять однієї машини. Це запобігає виконанню класичних послідовних алгоритмів, включаючи процедури видобування даних та машинного навчання.

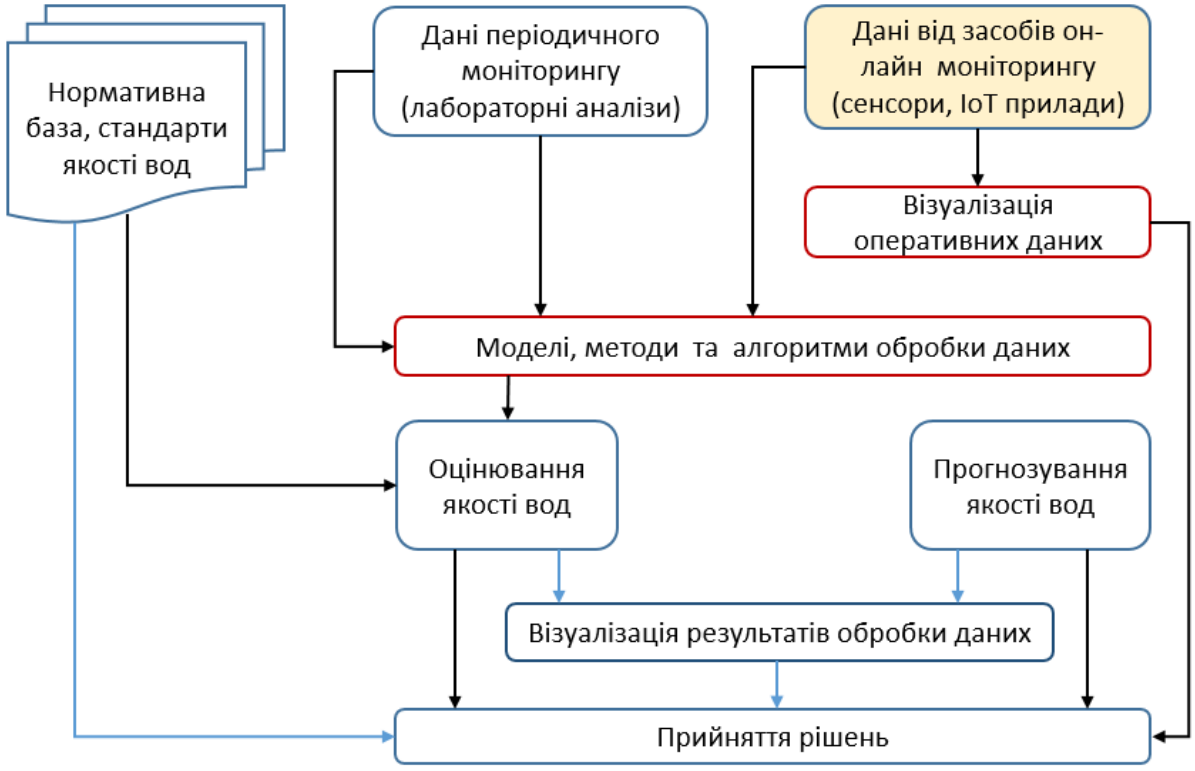


Рисунок 1.2 – Основні компоненти прийняття рішень при обробці моніторингових даних якості води

Згідно [45] аналіз великих даних зазвичай виконується пакетним способом (наприклад, один раз на день), причому, на різних етапах можливі декілька сценаріїв: (1) обробка великих даних через створення та агрегування вимірів; (2) зменшення розмірності й візуалізація; та (3) глибокий аналіз даних (рис. 1.3).

Перший сценарій обробки великих даних зазвичай виконується за допомогою технології Hadoop з класичною реалізацією логіки ETL [46]. Модель MapReduce, яку надає Hadoop, дозволяє лінійно масштабувати обробку, додаючи більше машин до обчислювального кластеру.

Другий сценарій передбачає он-лайн аналітику і вимагає використання

засобів зменшення розмірності даних з їх наступною візуалізацією з метою пошуку аномалій та оперативного реагування на ситуацію.

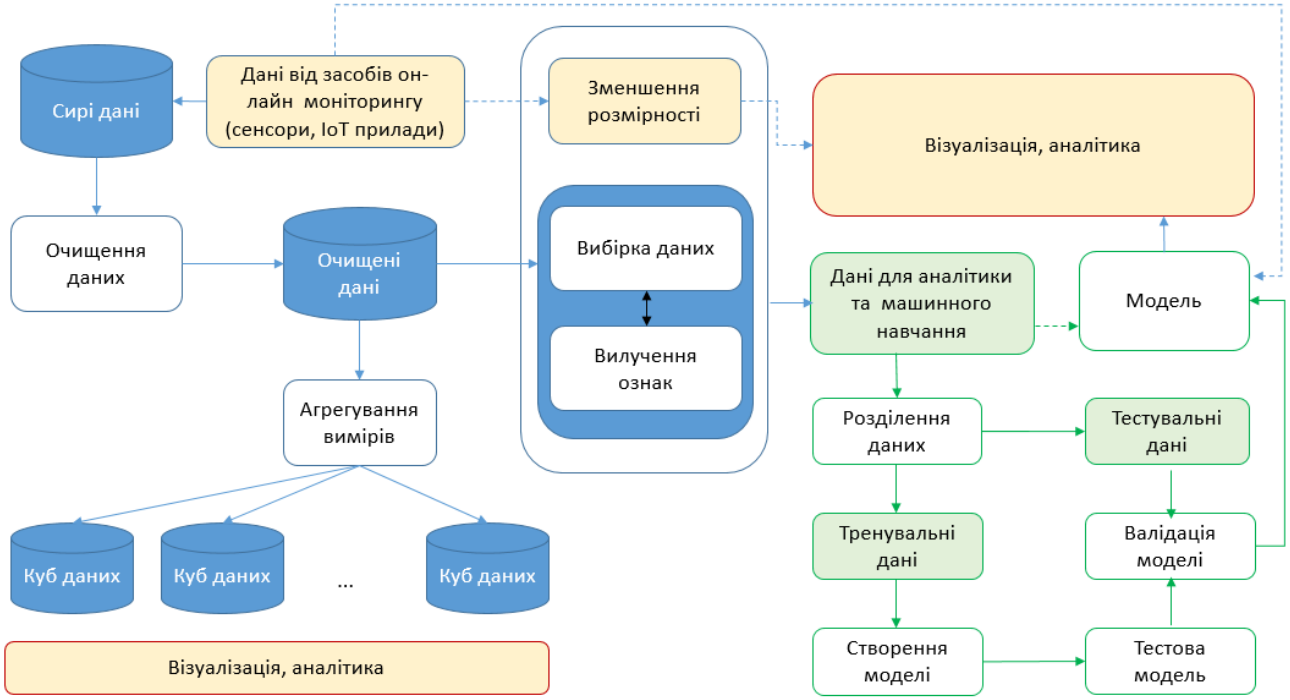


Рисунок 1.3 – Сценарії використання великих даних

Третій сценарій заснований на глибокому аналізі та виконується засобами машинного навчання, використовуючи набагато меншу кількість ретельно відібраних даних, що вміщується в ємності однієї машини (менше сотні тисяч записів даних). Глибокий аналіз передбачає підготовку даних, створення та аналіз моделей (наприклад, лінійна регресія, метод найближчого сусіда, метод опорних векторів, байєсівські мережі, нейронні мережі, дерева рішень та ін.), оцінку моделі, візуалізацію даних.

Варто відзначити, що для моніторингових систем, дані корисні лише в тому випадку, якщо з великих обсягів необроблених даних можливо отримати цінну інформацію і надати її у зручному форматі для прийняття рішень в реальному (або майже реальному) часі.

1.1.3 Великі дані в системах моніторингу якості поверхневих вод та аквакультури

Якість води можна визначити як придатність води для певного застосування на основі її хімічних, біологічних та фізичних характеристик. Моніторинг якості води передбачає виявлення її характерних параметрів та порівняння їх із встановленими стандартами та рекомендаціями.

Моніторинг та управління якістю води передбачають контроль, аналіз, оцінку та звітність про якість води [8, 39].

Правильне визначення стану якості води в річковій чи озерній системі має важливе значення для досягнення цілей екологічного управління [8], але, як зазначалось раніше, екологічні дані загалом і дані моніторингу води зокрема, можуть надходити в різних формах і форматах, таких як структуровані та неструктуровані дані, зображення, метрики, геопросторові, текстові, мультимедійні, моделі, рівняння тощо. Те, наскільки швидко виробляються і збираються дані, є одним з основних викликів для ефективного і своєчасного реагування на зміни стану водних систем. Ще одним складним викликом є достовірність наборів даних про стан водойм, оскільки дуже важливо мати високоякісні дані, перш ніж аналізувати їх. Тенденція даних води може бути невизначеною, оскільки дані збираються і за допомогою датчиків і з використанням ручних процесів. У неоднорідних джерелах часто виявляються непослідовність, неоднозначність, затримка та помилки в даних.

Для українських господарств, що займаються розведенням риби (переважно це гідробіонти (короп, товстолобик, білий амур), а також райдужна форель, європейський сом, щука, карась, білуга) та інших комерційних об'єктів аквакультури, системи, що контролюють параметри води та навколишнього середовища (наприклад, кисень, температура та ін.), фізіологічні показники культурних видів та кінцеві результати технологічного процесу (наприклад, аміак, pH та ін.) дозволять оптимізувати їх ефективність за рахунок зниження

витрат на оплату праці та комунальних послуг. Безумовними перевагами використання інформаційно-аналітичних систем в процесах аквакультури є:

– підвищення ефективності технологічних процесів;
– зменшення втрат енергії та води;
– зменшення витрат на оплату праці;
– знижений стрес і захворювання риби;
– покращений облік;
– поліпшення розуміння процесу.

Разом з тим, враховуючи обсяги та характер великих даних, виникає низка питань щодо організації обробки, зберігання та візуалізації даних, а саме:

– скільки даних можна і потрібно збирати та зберігати;
– скільки даних необхідно для побудови точних моделей прогнозування;
– яким чином можливо реалізувати аналіз та візуалізацію даних, щоб отримані данні найкраще відображали зміни та забезпечували прийняття рішень;
– витрати на зберігання, обчислення та візуалізацію великої кількості даних.

Крім того, варто відзначити, що вибір підходу та використані технології мають великий вплив на ефективність систем моніторингу якості поверхневих вод та аквакультури. В роботі [4] розглянуто приклад, коли обробка одних і тих же даних складу річкової води за різними методиками інтегральної оцінки якості дає діаметрально протилежні результати – від доброї (за методикою розрахунку індексу якості води (water quality index або WQI) Національної санітарної організації (NSF) США [14, 47]) до дуже поганої (за методикою згідно з КНД 211.1.4.010-94 [5]).

1.2 Аналіз моделей і методів аналізу даних для великих даних

Основними математичними підходами до аналізу великих даних є кластеризація, нечітка логіка, еволюційні алгоритми. Кластерний аналіз займає

одне з центральних місць серед методів аналізу великих даних, дозволяючи формувати однорідні класи у довільних проблемних галузях і виконувати пошук закономірностей у великих обсягах багатовимірних даних. В табл. 1.2 надано результати аналізу методів видобування даних і технологій машинного навчання, що дозволяють опрацьовувати великі дані та зменшувати невизначеності у даних в залежності від типу невизначеності (неповнота, відсутність маркування, різні масштаби, низька достовірність, великі обсяги, розмірність та різноманітність даних, тощо).

Таблиця 1.2 – Технології зменшення невизначеності в великих даних

Тип невизначеності	Технології зменшення невизначеності
Неповні дані	Активне навчання [32, 64], глибоке навчання [75, 79], нечіткі множини [26,62]
Немарковані дані	Активне навчання [38, 64]
Масштабованість	Розподілене навчання [17, 50], глибоке навчання [75]
Низька достовірність, складні та зашумлені дані	Нечітка логіка, еволюційні алгоритми [60, 69]
Великий обсяг, різноманітність даних	Ройовий інтелект, еволюційні алгоритми [55, 68, 69], кластерний аналіз [37, 51], алгоритми узгодження на основі нечіткої логіки [36, 80]
Великий обсяг та розмірність даних	аналіз основних компонентів [41, 56, 80], факторний аналіз, кластерний аналіз [22, 39], незалежний компонентний аналіз [58], незалежний факторний аналіз [21]

Значну роль в роботі з великими даними відіграють методи зменшення розмірності даних, серед яких аналіз основних компонентів, факторний аналіз, кластерний аналіз, незалежний компонентний аналіз, незалежний факторний

аналіз. Було проведено багато наукових досліджень щодо використання методів аналізу великих даних для аналізу і прогнозування, зокрема прогнозу кількості сонячної енергії [72]; моделювання стихійних лих [15, 18], розумних водних мереж [79]. Однак моделі запропоновані в [15, 72], не підходять для систем моніторингу водного середовища з одним типом даних; технології глибокого навчання [75, 79] є дуже корисними для аналізу та розпізнавання даних, отриманих з потокового відео але ставлять високі вимоги до систем передачі даних, обробної здатності комп'ютерів, що також, у більшості випадків, не підходить для системи моніторингу водного середовища. Окрім того, враховуючи, що ступінь забруднення води є розмитим поняттям із властивою неточністю, інтервальними критеріями класифікації та розмитими межами між різними класами якості води, існують труднощі класифікації та оцінки якості води у традиційних методологіях оцінки, таких як індекс якості води (WQI), запропонований у 1965 році, як опис інтегрованої якості води [47]. Традиційні методи оцінки якості води [7] не враховують невизначеностей, пов'язаних з вимірюванням параметрів та існуючими обмеженнями, і використовують чіткі значення встановлених значень і концентрацій, близькі або далекі від меж, що входять до одного класу. Така ситуація змусила дослідників шукати методи, засновані на нечіткій комплексній оцінці, призначені для інтерпретації невизначеності оцінки якості води [77]. Цей підхід дозволяє всебічно оцінювати внески різних забруднюючих речовин відповідно до заздалегідь визначених ваг та зменшує нечіткість за допомогою функцій належності. Тому, як визначається в [48, 81] його чутливість вища за інші методи оцінки. Таким чином, методи, що використовують нечіткі комплексні оцінки, здатні покрити невизначеності у процесі відбору проб та аналізу, порівнюючи результати зі стандартами якості для кожного параметра та підсумовуючи значення окремих параметрів. У цьому контексті, ряд методів було успішно застосовано для вирішення екологічних задач шляхом кластеризації та моделювання на основі нечітких множин і нечіткої логіки. Так, у багатьох дослідженнях [37, 44, 53, 59, 62, 66] йдеться про те, що

нечіткі методи широко застосовуються в оцінці якості навколишнього середовища і довели свою ефективність у вирішенні проблем з розпливчастими межами та для контролю ефекту помилок моніторингу, що робить цей підхід перспективним до використання в системах моніторингу якості води.

1.2.1 Нечіткі моделі в обробці великих даних якості вод

Нечітка логіка, як узагальнення класичної логіки та теорії множин була введена Л. Заде в 1965 р. [83], і є математичним інструментом для боротьби з невизначеністю.

В [76] визначено наступні чотири фактори, які обумовлюють використання нечітких моделей в контексті великих даних:

(1) Наявність невизначеності в самих даних та методах обробки великих даних. Наприклад, дані можуть надходити від несправних датчиків; результати інтелектуальних алгоритмів також містять невизначеності.

(2) Надто точне вирішення проблем може бути дуже дорогим або не потрібним. У випадках, коли проблему краще вирішувати на грубому рівні, нечіткі моделі дозволяють реконструювати її на певному рівні деталізації.

(3) Необхідність використання декількох методів прийняття рішень, таких як ймовірнісні, грубі множини, нейронні мережі тощо.

(4) Нечіткі моделі можуть вдосконалити поточні методи обробки великих даних зокрема, забезпечити нову стратегію для абстрагування і представлення знань.

Як зазначено в [34], нечіткі множини особливо підходять для обробки різноманітності та якості походження великих даних (V3, V5 на рис.1.1). Це, головним чином, завдяки їхній здатності справлятися з розпливчастими, неточними та невизначеними поняттями. Більше того, використання нечітких функцій належності, що перекриваються, забезпечує гарне покриття проблемного простору. Ця проблема особливо актуальна при роботі з дуже великими наборами

даних, які можуть бути представлені наборами різнорідних шматків, наприклад, у парадигмі програмування MapReduce. Способи використання технології нечітких множин для обробки великих даних розглядалися в багатьох роботах [31, 35, 43, 76], представлені на рис. 1.4.

Рисунок 1.4 - Способи використання технології нечітких множин для обробки великих даних (адаптовано з [43, 76])

Протягом багатьох років нечіткі моделі успішно застосовуються для проектування екологічних індексів, оскільки вирішують складні ситуації, такі як неоднозначність, суб'єктивні судження та інтерпретація складного набору багатовимірних даних [59, 66]. Результати розрахунку індексу якості води (WQI) найчастіше пов'язані з різними мовними категоріями якості води (наприклад, відмінна, хороша, звичайна або погана якість води). Ці лінгвістичні змінні використовують нечіткі межі, маючи високий ступінь невизначеності. Таким чином, нечітка логіка вважається корисним інструментом для моделювання якості води, оскільки є альтернативним підходом до проблем, де цілі та межі є дифузними або неточними.

В умовах великих даних існує кілька досліджень, присвячених використанню нечіткої логіки але більшість з них використовує дві найважливіші системи нечіткого висновку - Мамдані [57] та Сугено [70]. Система Mamdani є

найбільш широко використовуваним для екологічних застосувань завдяки своїй простоті та практичному застосуванню [65], тому пропонується до використання у цьому дослідженні.

1.2.2 Особливості кластеризації великих даних

Хоча масивні обсяги даних можуть бути дуже корисними для дослідників, підприємств та компаній, через високі обчислювальні витрати вони роблять аналітичні та пошукові операції надзвичайно трудомісткими. Можливе рішення цієї проблеми покладається на кластеризацію великих даних у компактну, але все ще інформативну версію повного набору даних.

Кластерний аналіз є методом поєднання набору даних у групи подібних об’єктів. Це підхід до неконтрольованого навчання, а також одна з основних методик розпізнавання образів. У випадку великих даних, кластеризація повинна створювати зведення, які є значущими і якомога компактнішими. Методи кластерного аналізу добре вивчені в літературі, більшість з них є реляційними [51]. Доступно багато алгоритмів кластеризації, в основному похідних від ієрархічного або розділового підходів. Ієрархічні алгоритми створюють вкладену серію розділів, тоді як алгоритми розділів - лише один [61]. При цьому приймаються до уваги наступні параметри:

– змінні кластеру, тобто компоненти шаблонів, на яких базується класифікація;

– метрика або спосіб за яким алгоритм порівнює шаблони, наприклад евклідова відстань;

– ступінь деталізації опису даних; кластеризація даних це завжди компроміс між дуже детальним (з багатьма кластерами) та спрощеним (кілька кластерів) описом даних.

Незважаючи на велику кількість впроваджень і глибоку теоретичну основу, традиційні методи кластеризації не можуть впоратися з обсягом великих даних

через їх велику складність та обчислювальні витрати. Наприклад, традиційна кластеризація K-means є NP-складною, навіть, коли кількість кластерів k = 2. Отже, масштабованість є основною проблемою для кластеризації великих даних.

Іншою проблемою є те, що звичайні (чіткі) способи кластеризації обмежують кожну точку набору даних точно одним кластером і генерують набори, в яких кожному об’єкту присвоюється рівно один кластер [36]. Однак дуже часто об'єкти не можуть бути належним чином присвоєні строго одному кластеру, оскільки вони розташовані між кластерами. Наприклад, оцінка WQI = 1 вказує на водний об’єкт, якість води якого досягла критичної точки. За такої оцінки класифікацію якості води можна віднести лише до однієї з двох груп – «нормальної», або «ненормальної» групи, що суттєво обмежує простір прийняття рішень.

Головна мета - розширити та пришвидшити алгоритми кластеризації з мінімальними втратами якості кластеризації. Хоча масштабованість та швидкість кластеризації алгоритмів завжди були метою для дослідників у цій галузі, проблеми з великими даними підкреслюють ці недоліки та вимагають більшої уваги та досліджень з цієї теми. Огляд літератури з методів кластеризації показує, усі методи можна класифікувати на дві категорії [13, 84]: методи одиночної машинної кластеризації та методи множинної кластеризації, як показано на рис. 1.5.

Одномашинна кластеризація [13, 84] складається з методів на основі секціонування та методів, заснованих на зменшенні простору вимірів. Перша група методів ділить набір даних в одному розділі, використовуючи відстань для класифікації точок на основі їх подібності; недоліком методів секціонування є те, що вони методи, як зазначалося вище, вимагають від користувача заздалегідь визначеного параметра k (кількості кластерів), який часто не є детермінованим. Існує багато алгоритмів розподілу, таких як алгоритм k-середніх, k-медоїди, PAM, CLARA, CLARANS та FCM. Алгоритм BIRCH, DENCLUE та OptiGrid

більш пристосовані до великих обсягів даних, але вони страждають від низької якості класифікації.

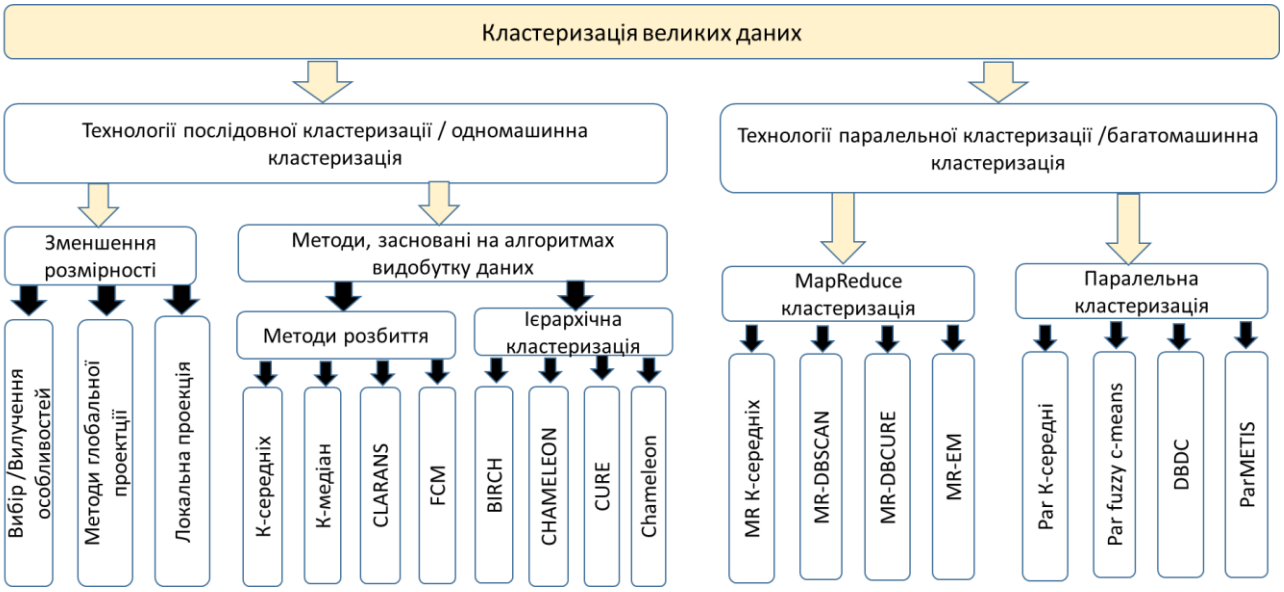


Рисунок 1.5 – Таксономія методів кластеризації великих даних (Адаптовано з [12, 33, 67, 68, 84])

Багатомашинні методи кластеризації в свою чергу складаються з двох груп – методів що використовують неавтоматизований розподіл або паралельної кластеризації та методів автоматизованого розподілу – MapReduce. Цей підхід дозволяє опрацьовувати великі обсяги даних за рахунок їх розбиття на менші частини, які можна завантажити на різні машини, та використання обчислювальних потужностей цих машин.

Різниця між паралельними алгоритмами та структурою MapReduce полягає в тому, що MapReduce забезпечує вирішення багатьох питать балансування навантаження, розподілу даних, відмовостійкість та ін., обробляючи їх автоматично. Ця функція забезпечує величезний паралелізм та простішу та швидшу масштабованість паралельної системи [28, 68].

Дослідження, проведене в [67], оцінює різні методи кластеризації за терміном використання великих даних, використовуючи відповідні критерії для

оцінки сильних та слабких сторін кожного алгоритму щодо тривимірних властивостей великих даних, включаючи обсяг, швидкість та різноманітність. З цього дослідження можна зробити висновок, що не існує ефективного алгоритму класифікації за всіма критеріями оцінки, але EM та FCM показують кращі показники якості у порівнянні з іншими алгоритмами.

Усі ці методи показують, що для великої кількості даних розрахований час вибухає. Тому, щоб вирішити цю проблему, ми повинні скористатися ефективною мовою програмування або вдосконаленим технічним обладнанням.

Ідея часткової належності [83], дозволила використовувати нечітку кластеризацію, результат якої представляє ступінь вибірки, що належить пов'язаному кластеру. Нечітка кластеризація є більш природньою та обґрунтованою. У багатьох методах нечіткої кластеризації найбільш широко застосовується алгоритм кластеризації нечітких c-середніх (fuzzy c-means або FCM) [44]. У науковій літературі з неконтрольованого навчання, FCM є найбільш відомим нечітким методом кластеризації.

У порівнянні з традиційними чіткими методами кластеризація FCM може надати більше вагомої інформації, що набагато точніше відображає фактичний розподіл вибірки. У літературі про нечітку кластеризацію алгоритм кластеризації FCM, запропонований в [30] та розширений у [17], є найбільш відомим і використовуваним методом. Однак, хоча FCM є дуже корисним методом кластеризації, його оцінки не завжди добре відповідають ступеню належності даних, і можуть бути неточними у шумному середовищі. Щоб поліпшити цю ситуацію у роботі [49] було запропоновано потенціалістичний підхід (possibilistic approach) до кластеризації, який використовував потенційований тип функції належності для опису ступеня належності. Автори показали, що алгоритми з можливістю належності є більш надійними для шуму та викидів в даних ніж FCM.

Потенціалістичний підхід до кластеризації застосовується також для поверхневої кластеризації, виявленні меж, наближенні поверхні та функцій. Варто відзначити, що для використання підходу необхідно заздалегідь визначити

кількість кластерів, але в більшості випадків кластерне число невідоме. Проблему пошуку оптимального числа кластерів зазвичай називають валідністю кластера (cluster validity). Як тільки кластеризація завершується і отримано відповідний розділ, функція валідності дозволяє перевірити наскільки точно розділ представляє структуру набору даних. Комбінуючи функцію валідності та нечіткі алгоритми кластеризації можливо значно покращити якість кластеризації.

Отже, кластеризація є одним з найважливіших завдань у видобутку даних, які потребують вдосконалення більше, ніж раніше, в умовах наявності терабайт і петабайт даних. Основним обмеженням усіх алгоритмів кластеризації є їх нестабільність, що означає, що реалізація одного і того ж алгоритму може дати різні результати. З цією метою було запропоновано нові розподілені реалізації алгоритмів аналізу даних та машинного навчання для великих даних, в основному на основі парадигми MapReduce [24]. Системи, засновані на парадигмі MapReduce, особливо Apache Hadoop [11], стали дуже популярними оскільки можуть використовуватися для величезного обсягу розподілених даних і забезпечувати паралельну обробку. Паралельна обробка може застосовуватися у великих водопровідних мережах на рівні області або регіону.

Підводячи підсумок, при традиційній вибірці та зменшенні розмірів алгоритми все ще корисні, але не мають достатньої сили, щоб мати справу з величезними обсягами даних, оскільки навіть після вибірки петабайт даних набір даних все ще дуже великий і не може бути кластеризований звичайними алгоритмами кластеризації, отже, майбутнє кластеризації пов'язане розподіленими обчисленнями. Хоча паралельна кластеризація потенційно дуже корисне, але складність реалізації таких алгоритмів є проблемою. З іншого боку, фреймворк MapReduce забезпечує цілком задовільну базу для реалізації алгоритмів кластеризації. Як показує аналіз літератури [19, 20], алгоритми, засновані на MapReduce, пропонують вражаючу масштабованість і швидкість порівняно з послідовними аналогами.

1.2.3 Парадигма MapReduce

Питання, пов’язані з великими даними, вимагають прийняття нових стратегій для їх опрацювання [29]. З цією метою в 2004 році компанія Google запровадила і прийняла парадигму програмування MapReduce [26], яка стала фактичним стандартом для роботи з великими даними. Це не тільки модель програмування, а й модель планування завдань. На вищому рівні парадигма ділить обчислювальний потік на дві основні фази – Map (мапування) та Reduce (зменшення), визначених користувачами. Функція Map - це обробка пари ключ-значення для отримання проміжної пари ключ-значення. Величезні обсяги даних масштабуються відповідно до стратегії поділу та мапування, після чого виконується розв’язання підзадач та їх поєднання для отримання остаточного рішення поставленого завдання. Функція Reduce поєднує всі проміжні значення з тим самим середнім ключем [25].

Оскільки масивні дані обробляються паралельно, вони спочатку розподіляються по вузлах (комп’ютерах) кластера і зберігаються у розподіленій файловій системі. Дані представляються у вигляді пар (k_i, v_i) = (ключ, значення). Обчислення двох функцій виражається таким чином [74]:

$$map(k_1, v_1) \rightarrow list(k_2, v_2),$$
$$reduce(k_2, list(v_2)) \rightarrow list(k_3, v_3).$$

Приклад структури MapReduce, що використовує нечіткі оцінки розподілу проілюстровано на рис. 1.6, це розширення реалізації класичного алгоритму яке представляє локально розподілену реалізацію запропоновану в [54], кожен мапер генерує базу нечітких правил, використовуючи лише підмножину екземплярів, оброблених цим конкретним мапером. Таким чином, ваги правил залежать від пропорції та розподілу класів у конкретній підгрупі навчальних екземплярів.

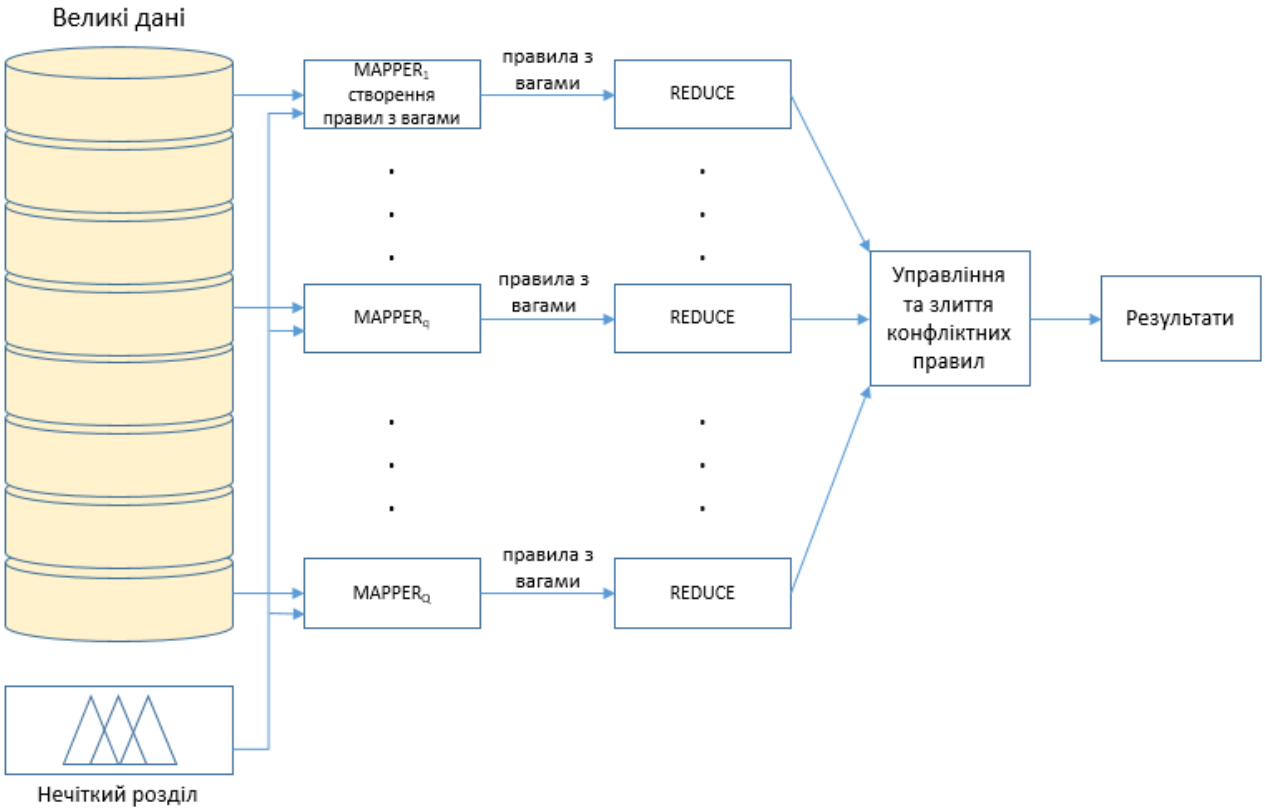


Рисунок 1.6 – Схема MapReduce з урахуванням нечіткого розділу (Адаптовано з [29, 54])

На даний момент не всі алгоритми обробки великих даних можна розпаралелювати. Деякі алгоритми неможливо розпаралелювати в теорії. Інші мають бути пристосовані, до задач розпаралелювання. У цій роботі використано нечіткий алгоритм k-середніх у рамках парадигми MapReduce. Зокрема, адаптовано алгоритм c-середніх у програмному забезпеченні з відкритим кодом для кластеризації великомасштабних даних якості води. Для того, щоб розглянути відображення алгоритму c-середніх на примітиви Map і Reduce, необхідно, щоб він був розділений на два завдання MapReduce. Перше завдання MapReduce обчислює матрицю центроїдів шляхом ітерації над записами даних, а друге завдання MapReduce необхідно, оскільки для наступних розрахунків в якості вхідних даних потрібна повна матриця центроїда. Друге завдання MapReduce також переглядає записи даних і обчислює відстані, які потрібно використовувати для оновлення матриці належності.

1.3 Огляд проблем візуалізації великих даних і підходів до їх вирішення

Стан будь-якого водного об'єкту, що оснащений датчиками з яких постійно знімається інформація може бути описаний послідовностями їх значень, які в свою чергу характеризуються і можуть бути представлені точками, лініями, кривими або поверхнями. Графічне відображення - це простий та дієвий спосіб візуалізації даних, що дозволяє полегшити сприйняття даних та надає достатньо інформації для аналізу і прийняття рішень [42]. Методи візуалізації даних важливі, оскільки вони:

– роблять дані більш природними для сприйняття людиною і, отже, полегшують виявлення тенденцій, закономірностей та відхилень у межах великих наборів даних;

– дають можливість особам, які приймають рішення, швидко зрозуміти, що відбувається;

– отримати інформацію щодо тенденцій - використання відповідних методів може полегшити розпізнавання цієї інформації;

– виявити взаємозв’язки та несподівані зв’язки, які не вдалося знайти з конкретними питаннями;

– забезпечити високоефективний спосіб передачі знання.

Візуалізація даних дає чітке уявлення про те, що означає інформація, надаючи їй візуальний контекст за допомогою карт або графіків, але у випадку з великими даними більшість класичних методів подання даних стають менш ефективними або навіть не застосовними до конкретних завдань.

Сучасні можливості зберігання необроблених даних довгострокових записів за відносно низької вартості дозволили аналізувати ретроспективні дані з доступом до повного запису. Це дозволяє багаторазово вдосконалювати процес аналізу, тестувати альтернативні алгоритми аналізу, виявляти та обробляти несподівані артефакти й різні за часом моделі діяльності. Під час ітеративного

аналізу даних візуальна перевірка даних часових рядів є повторюваним кроком роботи. Спочатку артефакти та / або шум потрібно ідентифікувати та виключити з подальшого аналізу. Згодом результати етапів передобробки повинні бути перевірені.

Завдяки інтерактивному характеру такого аналізу користувач повинен мати можливість швидко переходити між даними та переглядати їх з різним рівнем збільшення. Ця вимога ставить виклик традиційним програмам візуалізації [63]. Технології візуалізації великих даних викликають величезні потреби в ресурсах, що включають високі вимоги до пам'яті та надзвичайно високу вартість розгортання. Ці інструменти можна класифікувати на основі трьох факторів: за типом даних, за типом техніки візуалізації та за сумісністю. Перший стосується різних типів даних, що підлягають візуалізації [71]:

– Одновимірні дані - одновимірні масиви, часові ряди.
– Двовимірні дані - двовимірні графіки, географічні координати.
– Багатовимірні дані - фінансові показники, результати експериментів.
– Тексти та гіпертексти - статті, веб-документи.
– Ієрархічність та зв'язки - структура підпорядкування в організації, електронні листи, документи та гіперпосилання.
– Інформаційні потоки, операції налагодження тощо.

Другий фактор базується на методах візуалізації та зразках для представлення різних типів даних. Методи візуалізації можуть бути як елементарними (лінійні графіки, діаграми, гістограми), так і складними (на основі попередньої обробки). Хоча існують різні способи візуалізації даних часових рядів, найбільш часто використовуваним методом для даних моніторингу є лінійний графік [71]. Канонічним підходом для лінійного сюжету є проектування зразків по черзі на полотно та з'єднання отриманих точок лініями. Тому час побудови графіку залежить від обсягу даних: більша кількість даних призводить до збільшення часу побудови графіку. Через великі обсяги даних, отриманих при тривалих записах, канонічний метод побудови графіків не може забезпечити

необхідну продуктивність, незважаючи на вражаючу обчислювальну потужність поточних процесорів та графічних карт; тому необхідний інший алгоритмічний підхід.

Іншим аспектом є те, що багатьох випадках, при візуалізації великих даних, сигнали з малою амплітудою занадто щільні та розмиті, що призводить до того що невеликі варіації максимальної амплітуди параметрів при тривалих записах не дозволяють орієнтуватися в даних.

В рамках дослідження, приділяючи увагу властивостям наборів даних, можна виділити наступні проблеми візуалізації [73, 78]:

– Візуальний шум.
– Сприйняття великих зображень.
– Надмірне спрощення даних.
– Математичні обмеження алгоритмів спрощення даних.

1.3.1 Візуальний шум

Візуальний шум – це втрата видимості даних внаслідок накладання графічних примітивів. Звичайна візуалізація повного ряду даних може створити повний безлад на екрані, в результаті чого виникає одна велика пляма, що складається з точок, які представляють кожен рядок даних. Ця проблема пов'язана з тим, що більшість об'єктів в наборі даних, занадто пов'язані один з одним, і на екрані спостерігач не може розділити їх у вигляді окремих об'єктів. Так, іноді, аналізуючи складно отримати навіть трохи корисної інформації від всього набору візуалізованих даних без будь-якої додаткової обробки інформації [78].

У загальному випадку проблема візуального шуму може бути вирішена шляхом збільшення розмірів компонентів відображення, але у випадку візуалізації великих наборів даних такий підхід може призвести до появи проблеми сприйняття великого зображення.

1.3.2 Обмеження сприйняття занадто великих зображень

Однією із основних проблем візуалізації великих наборів даних є обмеження сприйняття занадто великих зображень. Існує певний рівень сприйняття людини для різних візуалізацій даних. Незважаючи на те, що цей рівень для візуалізації графічних даних набагато вищий, порівняно з візуалізацією табличних даних, він має свої обмеження. І після досягнення цього рівня сприйняття, людина втрачає здатність отримувати будь-яку корисну інформацію з перевантаженого перегляду даних.

Всі способи візуалізації обмежуються роздільною здатністю пристрою, що відповідає за вивід візуалізації, тому існує межа кількості точок, які повинні відображатися для кожної візуалізації. Звичайно, можна замінити пристрій відображення даних на більш сучасний або групою пристроїв для часткової візуалізації даних, що дозволить представити більш детальне зображення з більшою кількістю точок даних, але навіть якщо можливо повторити цей процес нескінченну кількість разів, обов’язково з’являється проблема обмеження людського сприйняття. Зі збільшенням об'ємів даних, що відображаються відразу, людина буде стикатися з труднощами в розумінні даних та їх аналізу. Тому можна стверджувати, що методи візуалізації даних обмежені не тільки співвідношенням сторін та роздільною здатністю пристрою, але й обмеженнями фізичного сприйняття.

1.3.3 Спрощення даних

Для вирішення проблеми сприйняття великих зображень вихідні дані піддають передобробці задля виділення меншої за обсягом підмножини даних, що якісно відображає поведінку вихідного набору великих даних. Задача відображення великих обсягів даних на точкових та лінійних діаграмах може бути розглянута як окремий випадок задачі спрощення полігональних ланцюгів.

Застосування методів спрощення полігональних ланцюгів є важливою проблемою в контексті візуалізації великих обсягів даних. Суть алгоритмів спрощення полягає у зменшенні кількості точок шляхом видалення тривіальних точок, але без порушення істотної форми вихідної лінії.

Перше застосування алгоритми спрощення полігональних ланцюгів набули у картографії. На сьогоднішній день більшість навігаційних додатків використовують алгоритми спрощення ліній для зменшення об’єму карт та поліпшення швидкості виконання таких операцій як масштабування та прокрутка.

Ключовим моментом у використанні алгоритмів спрощення полігональних ланцюгів є вибір порога спрощення, що визначає силу спрощення. Саме від цього параметру залежить кількість точок та вид отриманої лінії. В різних алгоритмах поріг спрощення визначається по різному. Найпоширенішими варіаціями порогу спрощення є:

1. Відстань від точки до лінії, утвореної сусідніми або крайніми точками.
2. Радіус, в межах якого видаляються усі точки.
3. Номер точки: видаляється кожна N точка.
4. Площа, в межах якої залишається лише зазначена кількість точок.

Чим більшим є поріг спрощення тим менше рівень деталізації отриманої лінії та тим менше точок складають отриману лінію. В межах проблеми спрощення, вибір помилки ε стає дуже важливим, оскільки занадто велике значення спричинить надмірне спрощення рядка, тоді як занадто мале значення робить алгоритм неефективним, зменшуючи недостатню кількість точок перетину.

Рис. 1.7 ілюструє проблему вибору траєкторії з можливими комбінаціями та пов'язаною з ними вартістю. Для знаходження оптимального значення порогу спрощення необхідно дати визначення факторам, що дають змогу вирахувати оптимальну кінцеву кількість точок.

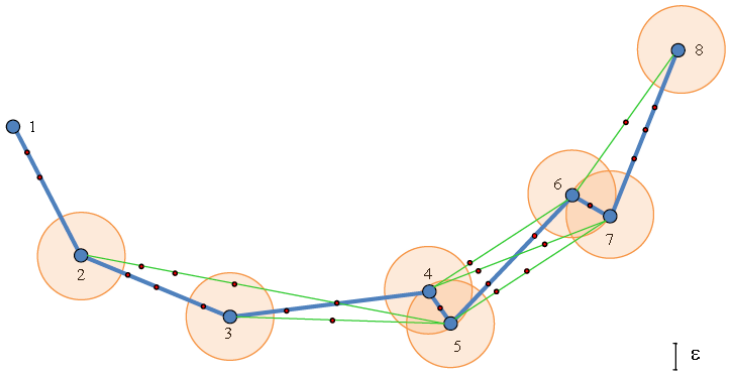


Рисунок 1.7 - Графік для спрощення рядків. Зелені краї представляють дозволені ярлики, точки перетину позначені червоним. Початкова траєкторія коштує 14, найкоротший шлях - 1-2-5-6-8 з вартістю 9. (Джерело [19])

Рівень деталізації спрощеної лінії тісно пов'язаний з порогом спрощення використовуваного методу. Поріг асоціюється із силою спрощення. Для вимірювання сили спрощення зазвичай використовують алгоритм знаходження середньої відстані зсуву. Серед факторів, що впливають на необхідну силу спрощення полігонального ланцюга виділяють наступні:

1) Розподільна здатність дисплея (Hr x Vr), Розмір дисплея (Ds)

Чим більшу розподільну здатність має дисплей тим менше повинна бути сила спрощення і навпаки. При однаковій розподільній здатності: чим більший розмір (діагональ) дисплея – ти більший розмір пікселя.

2) Кількість пікселів на дюйм (PPI)

PPI є показником щільності розташування пікселів на дисплеї. PPI можна розрахувати за допомогою наступної формули:

$$PPI = \frac{\sqrt{{H_r}^2 + {V_r}^2}}{D_s}$$

Фізичний розмір пікселю (D_p) можна розрахувати за допомогою наступної формули:

$$D_p = \frac{1}{PPI} \text{ (дюймів)}$$

3) Просторова розподільна здатність пікселя (SRp)

Просторова розподільна здатність - число незалежних пікселів значень на дюйм. Цей показник головним чином впливає на здатність розрізнити деталі на лінії. Формула для обчислення просторової розподільної здатності:

$$SR_p = \frac{D_p}{PPI * S}$$

Для отримання знаходження сили спрощення у класичному випадку використовується формула:

$$D_{simpl} = SR_p * N_e \ ,$$

де N – мінімальна кількість пікселів, що зможе розлічити людське око на певній відстані від монітору. Оскільки за технічним завданням має бути можливість зміни пропорцій графіків відносно один одного за допомогою компоненти під назвою «сплітер», то N помножимо на співвідношення поточної висоти компоненти відображення графіків до висоти графіка, для якого робимо обчислення. Формула прийме наступний вид:

$$D_{simpl} = SR_p * N_e * \frac{H}{H_{plot}}$$

1.3.4 Математичні обмеження алгоритмів спрощення полігональних ланцюгів

Проблема математичного обмеження алгоритмів спрощення полігональних ланцюгів є найбільш складною. Більшість алгоритмів спрощення передбачають

наявність параметрів, що характеризують поріг спрощення. Різноманіття варіацій порогу спрощення призводить до необхідності детального аналізу характеру вхідних даних. Вибір алгоритму з непідходящим порогом спрощення може призвести до втрати значущих даних та наштовхнути людину, що аналізує візуалізовані дані, до хибних висновків.

Спрощення полігонального ланцюга змінює його форму. Чим вище ступінь спрощення, тим більше спрощений ланцюг відрізняється від оригінального. Метод знаходження позиційних похибок - це спосіб вимірювання похибки спрощення, що полягає в знаходженні відстаней між вихідними точками, що були спрощені, та спрощеним ланцюгом. Оцінка якості спрощення полігональних ланцюгів методом знаходження позиційних похибок.

Підсумовуючи вищевикладене варто відзначити, що однією з головних проблем візуалізації великих даних є розмір набору даних, який може суттєво вплинути на продуктивність. Якщо ми спробуємо відобразити на карті всі траєкторії як лінії, неможливо розрізнити різні сутності або розрізнити транспортні магістралі чи вузли, де зустрічається багато траєкторій. Існує потреба в аналізі та обробці даних, зберігаючи глобальний аспект. Цього можна досягти за допомогою декількох методів, таких як кластеризація, агрегування з картами щільності, зменшення траєкторії, спрощення лінії або за допомогою пост-обробки. Для поліпшення візуалізації, рішення повинно або мінімізувати велику метушню, або знизити її під певним порогом. Цей поріг може бути визначений як обчислювально, так і експериментально, за допомогою людського судження. Візуалізація набору даних траєкторій повинна відповідати вимогам, встановленим на початку, і допомагати людині краще аналізувати великі набори даних. Різноманітність шляхів визначення порогу спрощення у різних алгоритмах та наявність приорітетних точок у архівних даних обумовлює необхідність порівняльного аналізу існуючих алгоритмів спрощення полігональних ланцюгів, та розробку нового алгоритму, на базі існуючих, що має враховувати наявність приорітетних даних.

При цьому, для візуалізації великих даних цілями є розробка засобів, що дозволять візуально дослідити зміни та поведінку окремих параметрів та інтегрального індексу якості вод; виявляти періодичні закономірності; підкреслювати нетипові зміни; і створити візуалізацію, яка дає змогу порівняти різні дні. Засоби візуалізації мають ефективно використовувати дисплейний простір, максимізуючи щільність даних та мінімізуючи утворення скупчень плям на екрані. В контексті даної проблеми, важливою, раніше не розглянутою задачею, є розробка метода вибору алгоритму спрощення полігональних ланцюгів та динамічного визначення сили спрощення в залежності від типу і характеру вхідних даних. У даній роботі запропоновано вирішити задачу вибору алгоритму за допомогою методу знаходження позиційних похибок.

1.4 Постановка наукового завдання та обґрунтування методики досліджень

Результати проведеного аналізу нормативної бази, моделей, методів й інструментальних засобів обробки великих даних показали необхідність подальшої розробки для моделей, методів та інформаційних технологій для систем моніторингу водних об’єктів. Зокрема, виділено наступні три напрями (виклики) дослідження.

Виклик №1: Моніторинг якості води в режимі реального часу.

З розвитком технологій інтелектуального зондування та IoT все більше датчиків навколишнього середовища (включаючи датчики якості води) встановлюють та розгортають для багатьох водних ресурсів. Однак, все ще бракує інтегрованих оцінок води на основі великих даних у реальному часі та моніторингу середовищ для підтримки динамічної оцінки якості води, моніторингу та управління наглядом.

Можливе рішення: розробка і використання нечітких моделей якості води

Виклик №2: Перевантаження аналітичних систем потоковими даними, що надходять з пристроїв IoT та інших джерел

Аналіз досліджень показав, що обробка та розумне використання великих даних має значний потенціал до використання. Разом з тим, зростаючі вимоги до даних вимагають збільшення потужності та ефективності інструментів та методів їх обробки. Обчислювальні ресурси зростають лінійно, тоді як обчислювальні потреби можуть зростати надлінійно або навіть експоненційно.

З експоненціальним зростанням даних, традиційні алгоритми видобутку та аналізу даних не можуть в повній мірі задовольнити потреби обробки даних і тому для використання цього величезного обсягу даних необхідні ефективні моделі обробки з розумною обчислювальною вартістю великих, складних, динамічних та неоднорідних даних.

Можливе рішення: використання підходів до зменшення кількості та розмірності даних, зокрема кластеризації даних.

Виклик № 3: У багатьох випадках, при візуалізації великих даних, сигнали з малою амплітудою занадто щільні та розмиті, що призводить до того що невеликі варіації максимальної амплітуди параметрів при тривалих записах не дозволяють орієнтуватися в даних.

Можливе рішення: розробка засобів, що дозволять візуально дослідити зміни та поведінку окремих параметрів та інтегрального індексу якості вод; що у випадку візуалізації часових рядів може бути досягнуто за допомогою спрощення полігональних ланцюгів та динамічного визначення сили спрощення в залежності від типу і характеру вхідних даних.

1.4.1 Загальна наукова задача

Загальною задачею дисертаційного дослідження є підвищення ефективності роботи інформаційно-аналітичних систем моніторингу водних

об’єктів завдяки розробці та практичному застосуванні моделей та методів інформаційної технології обробки великих даних.

1.4.2 Часткові наукові завдання

1. Провести аналіз моделей та методів обробки великих даних в системах моніторингу водних об’єктів.
2. Розробити моделі для аналітичної обробки великих даних системи моніторингу водних об’єктів на основі формалізації її атрибутів та інтерпретації невизначеності оцінки якості води у вигляді лінгвістичних змінних.
3. Удосконалити метод нечіткої кластеризації с-середніх для великих, шляхом узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів для інтуїтивістських нечітких даних.
4. Удосконалити інформаційну технологію обробки великих даних та підтримки прийняття рішень в інформаційно-аналітичній системі моніторингу водних об’єктів.
5. Удосконалити технологію та засоби візуалізації великих даних.
6. Розробити програмні засоби та елементи інформаційної технології обробки великих даних в інформаційно-аналітичній системі моніторингу водних об’єктів.
7. Виконати практичне впровадження отриманих результатів.

1.4.3 Методика досліджень

Методика досліджень ґрунтується на вимогах до логіки та послідовності рішення поставлених завдань і адекватності обраного математичного апарату та складається з п’яти основних етапів: (1) аналіз розвитку галузі та стану питань з досліджуваної тематики, формування наукових і практичних завдань; (2)

розробка моделей для аналітичної обробки великих даних; (3) удосконалення методу нечіткої кластеризації для обробки великих даних; (4) розробка інформаційної технології для інформаційно-аналітичної системи моніторингу водних об'єктів; і (5) практична реалізація, валідація та впровадження розроблених засобів і технологій. На рис. 1.8 наведена запропонована в дисертаційній роботі методика досліджень, означені основні етапи, показані взаємозв'язки з отриманими результатами.

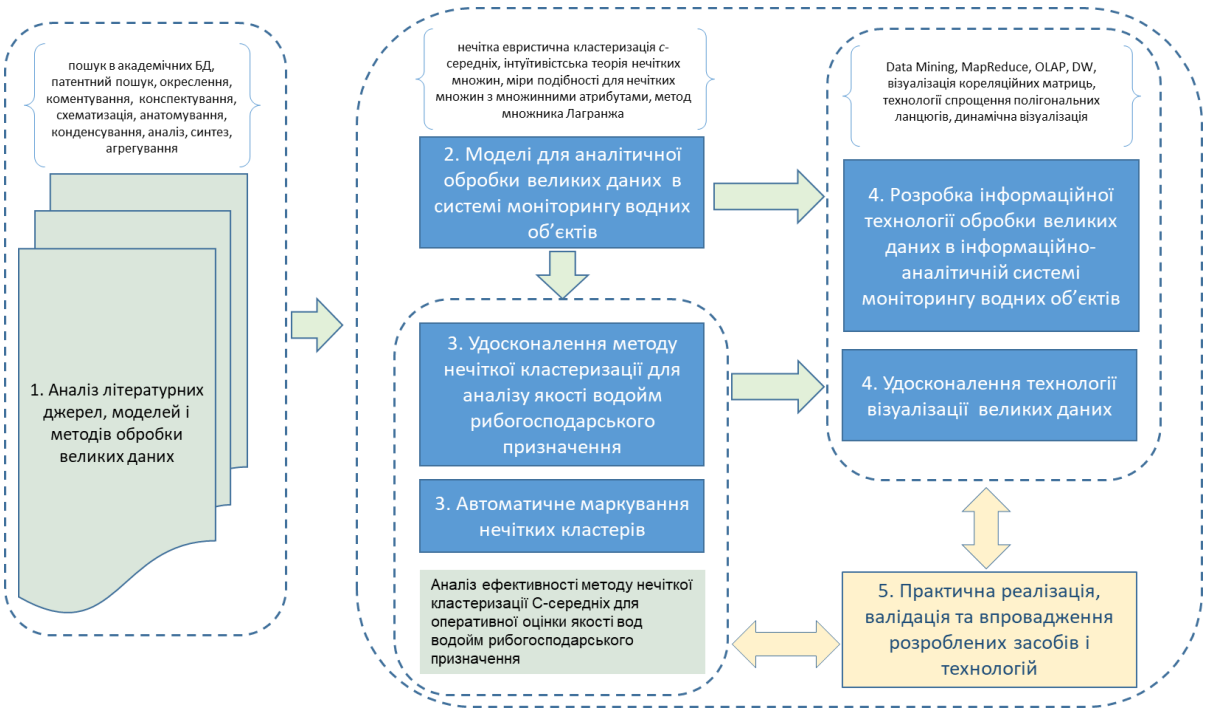


Рисунок 1.8 – Методика дисертаційного дослідження

Методологічну основу досліджень становлять методи системного аналізу, пошукового аналізу, технології роботи з літературними джерелами, зокрема окреслення, коментування, конспектування, схематизація, конденсування, тощо; методи системного аналізу. В процесі проведення досліджень було використано методи теоретико-множинного опису, теорії ймовірностей, теорії нечітких множин, зокрема нечітка евристична кластеризація c-середніх, інтуїтивістська теорія нечітких множин, міри подібності для нечітких множин з множинними атрибутами, метод множника Лагранжа. Для розробки інформаційної технології

використовувалися методи Data Mining, MapReduce, OLAP, DW, методи візуалізації кореляційних матриць, зокрема хордові діаграми, криві Безьє, технології спрощення полігональних ланцюгів, динамічна візуалізація.

Висновки до розділу 1

1. У даному розділі проведено аналіз моделей та методів обробки великих даних. Розглянуто особливості великих даних в системах моніторингу водних об’єктів, проведено аналіз нечітких моделей, особливостей кластеризації великих даних, парадигму MapReduce. Виконано огляд проблем візуалізації великих даних та підходів до їх вирішення.

2. На основі аналізу літератури виділено основні дослідницькі проблеми, завдання та майбутні потреби в оцінці якості та прогнозуванні якості води.

3. Основна мета дослідження полягає у створенні системи для ефективного поводження з даними моніторингу вод та сприяння впровадженню практичних методів видобутку даних на цих наборах даних з метою отримання корисних знань та підтримки прийняття рішень.

4. Результати проведеного аналізу нормативної бази, моделей, методів й інструментальних засобів візуалізації великих наборів даних, показали, що у відомих публікаціях і нормативно-технічних документах процеси формування та візуалізації великих наборів даних розглядаються як різнорідні технічні задачі без загального математичного апарата.

5. На основі проведеного аналізу сформульовано загальне завдання дисертаційного дослідження, яке розбито на ряд часткових завдань, спрямованих на розробку моделей для аналітичної обробки великих даних; удосконалення методу нечіткої кластеризації для обробки великих даних; розробку інформаційної технології для інформаційно-аналітичної системи моніторингу водних об’єктів; і практичну реалізацію.

Основні результати, наведені у першому розділі, опубліковано у [1-3, 9, 10].

Список літератури до розділу 1

1. Барбарук В.М., Барбарук Л.В. Методи моделювання складних систем, *Вісник Східноукраїнського національного університету ім. В. Даля*, № 15(186), С. 132-136, 2012.

2. Барбарук Л.В., Голдін В.А. Організація обробки передачі даних в бездротових сенсорних мережах, *Матеріали VII Всеукр. науково-практ. конф. «Електронні апарати та системи. Проблеми створення. Перспективи розвитку»*, Сєвєродонецьк : Східноукр. нац. ун-т ім. В. Даля, С. 152-154, 2017.

3. Барбарук Л.В. Сложные информационные системы, *Матеріали XIV всеукраїнської науково-практичної конференції "Технологія": ч. 2*. Сєвєродонецьк: Технол. ін-т Східноукр. Нац. ун-ту ім. В. Даля (м. Сєвєродонецьк), С. 129-131, 2011.

4. Вербецька К.Ю. Порівняльний аналіз методик оцінки якості поверхневих вод (на прикладі типової р. Губісцкалі). *Вісник Національного університету водного господарства та природокористування.* Серія «Сільськогосподарські науки». 2011, Вип. 5 (11). С. 91–99.

5. Екологічна оцінка якості поверхневих вод суші та естуаріїв України: Методика: КНД 211.1.4.010-94. – К.: 1994. – 37 с.

6. Еколого-економічні засади раціонального природокористування: теорія та практика реалізації : [кол. моногр.] Л. В. Єлісєєва, Р. С. Стрільчук, О. М. Стрішенець [та ін.] ; за заг. ред. д-ра екон. наук, проф. О. М. Стрішенець. Луцьк : Вежа-Друк, 2015. 236 с.

7. Мальцев В.І., Карпова Г.О., Зуб Л.М. Визначення якості води методами біоіндикації. К.: Науковий центр екомоніторингу та біорізноманіття мегаполісу НАН України, Недержавна наукова установа Інститут екології (ІНЕКО) Національного екологічного центру України, 2011. 112 с.

8. Основні засади управління якістю водних ресурсів та їхня охорона : навч. посібник. В. К. Хільчевський, М. Р. Забокрицька, Р. Л. Кравчинський, О. В.

Чунарьов, за ред. В. К. Хільчевського К. : ВПЦ "Київський університет", 2015. 172 с.

9. Рязанцев О.І., Барбарук В.М., Барбарук Л.В. Комплексна оцінка екологічної небезпеки території промислового виробництва, *Вісник Східноукраїнського національного університету ім. В. Даля*, №7(154), Ч.2, С. 136-140, 2010.

10. Скарга-Бандурова І.С., Грушка М.О., Барбарук Л.В. Підходи до ефективного спрощення та візуалізації великих наборів даних, *Вісник Національного технічного університету "Харківський політехнічний інститут". Зб. наукових праць. Серія: Інформатика та моделювання.* Харків: НТУ "ХПІ", №. 50 (1271), С. 55-65, 2017. doi: 10.20998/2411-0558.2017.50.10.

11. Apache foundation: Hadoop, 2014. URL: http://hadoop.apache.org/

12. Arora, S., & Chana, I. (2014). A survey of clustering techniques for big data analysis. In *5th International Conference - Confluence the Next Generation Information Technology Summit* (Confluence), 59-65.

13. Ashrafi, M.Z., Taniar, D., Smith, K.A. (2007) Redundant association rules reduction techniques. *Int. J. Bus. Intell. Data Min.* 2(1), 29–63.

14. A water quality index – do we dare? (1970) R.M. Brown, N.I. McClelland, R.A. Deininger et al. In *Water and Sewage Works*. Vol. 117, Issue 10. p. 339–343.

15. Belaud J.-P., Negny S., Dupros F., Michéa D., and Vautrin B., "Collaborative simulation and scientific big data analysis: illustration for sustainability in natural hazards management and chemical process engineering," *Computers in Industry*, vol. 65, no. 3, pp. 521–535, 2014.

16. Beracoechea J.A. (2017) Big Data: Volume is not the problem but the generation of value in an environment with such diversity https://www.bbvadata.com/big-data-volume-is-not-the-problem-but-the-generation-of-value-in-an-environment-with-such-diversity/

17. Bezdek J. C. Pattern Recognition with Fuzzy Objective Function Algorithms (1981), Plenum Press, New York.

18. Blöschl, G., C. Reszler, and J. Komma (2008), A spatially distributed flash flood forecasting model, *Environ. Model. Software*, vol. 23(4), pp. 464–478.

19. Böse J.-H., Andrzejak A., and Högqvist M., Beyond online aggregation: parallel and incremental data mining with online map-reduce, in *Proceedings of the Workshop on Massive Data Analytics on the Cloud (MDAC '10)*, pp. 1–6, April 2010.

20. Chao L., Yan Y., and Tonny R., A parallel Cop-Kmeans clustering algorithm based on MapReduce framework, *Advances in Intelligent and Soft Computing*, vol. 123, pp. 93–102, 2011

21. Coletti, C., Testezlaf, R., Ribeiro, T. A. P., Souza, R. T. G. de, & Pereira, D. de A. (2010). Water quality index using multivariate factorial analysis. *Revista Brasileira de Engenharia Agrícola e Ambiental*, 14(5), 517–522. doi:10.1590/s1415-43662010000500009

22. Dabgerwal, D. K., Tripathi, S. K. (2016). Assessment of surface water quality using hierarchical cluster analysis. *International Journal of Environment*, 5(1), 32–44. doi:10.3126/ije.v5i1.14563

23. Davies M.S. (2019) IoT-Based Environmental Monitoring: How Data Becomes Insights https://www.iotacommunications.com/blog/iot-based-environmental-monitoring/

24. Dean J, Ghemawat S. Mapreduce: a fexible data processing tool. *Commun ACM*. 2010;53(1):72–7.

25. Dean J. and Ghemawat S., MapReduce: simplied data processing on large clusters, in *Proceedings of the 6th Symposium on Operating System Design and Implementation (OSDI '04)*, pp. 1– 6, San Francisco, Calif, USA, December 2004.

26. Dean J, Ghemawat S. MapReduce: simplifed data processing on large clusters. *Comm ACM*. 2008;51(1):107–13.

27. Dewanti N. A., Abadi A. M. (2019) *IOP Conf. Ser.: Mater. Sci. Eng.* 546 032005.

28. Dubuc T., Stahl F. and Roesch E. B., (2021) Mapping the Big Data Landscape: Technologies, Platforms and Paradigms for Real-Time Analytics of Data

Streams, in *IEEE Access*, vol. 9, pp. 15351-15374, 2021, doi: 10.1109/ACCESS.2020.3046132

29. Ducange P., Fazzolari M. & Marcelloni F. (2020) An overview of recent distributed algorithms for learning fuzzy models in Big Data classification. *J. Big Data* 7, 19. https://doi.org/10.1186/s40537-020-00298-6

30. Dunn J.C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact WellSeparated Clusters", *Journal of Cybernetics* (1973): 3: 32-57.

31. Elkano, M., Galar, M., Sanz, J., Bustince, H. (2018) CHI-BD: a fuzzy rule-based classification system for big data classification problems. *Fuzzy Sets Syst.* 348, pp. 75–101.

32. El-Naas, M. (2011). Teaching water desalination through active learning. *Education for Chemical Engineers*, 6.

33. Fahad, A., Alshatri, N., Tari, Z., Alamri, A., Khalil, I., Zomaya, A.Y., Foufou, S., & Bouras, A. (2014). A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis. *IEEE Transactions on Emerging Topics in Computing*, *2*, pp. 267-279.

34. Fernandez, A., Carmona, C.J., del Jesus, M.J., Herrera, F. (2016) A view on fuzzy systems for big data: progress and opportunities. *Int. J. Comput. Intell. Syst.* 9(sup1), pp. 69– 80.

35. Ferranti, A., Marcelloni, F., Segatori, A., Antonelli, M., Ducange, P. (2017) A distributed approach to multi-objective evolutionary generation of fuzzy rule-based classifiers from big data. *Inf. Sci.* 415, pp. 319–340.

36. Foody, G. M. (1992) A Fuzzy Sets Approach to the Representation of Vegetation Continua from Remotely Sensed Data: An Example from Lowland Health. *Photogrammetric Engineering and Remote Sensing* 58(2):221-225.

37. Fu-Cheng L. and Xue-Zhao H., Application of Fuzzy C-means Clustering for Assessing Rural Surface Water Quality in Lianyungang City, *2013 Fifth International Conference on Measuring Technology and Mechatronics Automation*, Hong Kong, 2013, pp. 291-295, doi: 10.1109/ICMTMA.2013.75.

38. Gao, F., Yue, Z., Wang, J., Sun, J., Yang, E., & Zhou, H. (2017). A Novel Active Semisupervised Convolutional Neural Network Algorithm for SAR Image Recognition. *Computational Intelligence and Neuroscience*, 2017, 1–8. doi:10.1155/2017/3105053.

39. Gao C., Yan J., Yang S., Tan G. (2011) Applying Factor Analysis to Water Quality Assessment: A Study Case of Wenyu River. In: Li S., Wang X., Okazaki Y., Kawabe J., Murofushi T., Guan L. (eds) *Nonlinear Mathematics for Uncertainty and its Applications. Advances in Intelligent and Soft Computing*, vol 100. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-22833-9_66.

40. Gan G., Ma C., Wu J., (2007) Data Clustering: Theory, Algorithms, and Applications, *SIAM*, Philadelphia, PA.

41. Garcia, C.A., Garcia, H.L., Mendonça, M.C., da, A.F., Silva, Alves, J.P., Costa, S.S., Araujo, R.O., & Silva, I.S. (2017). Assessment of water quality using principal component analysis : a case study of the açude da Macela – Sergipe – Brazil.

42. Gorodov E., V. Gubarev, Analytical Review of Data Visualization Methods in Application to Big Data doi:10.1155/2013/969458.

43. Hariri, R.H., Fredericks, E.M. & Bowers, K.M. (2019) Uncertainty in big data analytics: survey, opportunities, and challenges. *J Big Data* 6, 44 https://doi.org/10.1186/s40537-019-0206-3.

44. Hathaway R. and Bezdek J., (2006) Extending fuzzy and probabilistic clustering to very large data sets, *Comput. Stat. Data Anal.*, vol. 51, no. 1, pp. 215–234.

45. Ho R. (2012) Big data analytics pipeline http://horicky.blogspot.com/2012/08/big-data-analytics.html.

46. Holmes A. (2012) Hadoop in practice. *Manning Publications Co*.

47. Horton R.K. (1965) An index number system for rating water quality *J. Water Pollut. Control Federation*, vol. 37 (3), pp. 300-306.

48. Ji X, Dahlgren RA, Zhang M. (2016) Comparison of seven water quality assessment methods for the characterization and management of highly impaired river systems. *Environ Mon Assess*. Vol. 188, pp. 115–30.

49. Krishnapuram R. and Keller J. M., (1993) A possibilistic approach to clustering, in *IEEE Transactions on Fuzzy Systems*, vol. 1, no. 2, pp. 98-110, doi: 10.1109/91.227387.

50. Kumar, R., Samaniego, L., & Attinger, S. (2013). Implications of distributed hydrologic model parameterization on water fluxes at multiple scales and locations. *Water Resources Research*, vol. 49(1), pp. 360–379. doi:10.1029/2012wr012195.

51. Kung, H., Ying, L., & Liu, Y.-C. (1992). A Complementary Tool to water quality index: Fuzzy clustering analysis. *Journal of the American Water Resources Association*, vol. 28(3), pp. 525–533. doi:10.1111/j.1752-1688.1992.tb03174.x.

52. Laney D. (2001) 3D data management: Controlling data volume, velocity and variety. *Appl. Deliv. Strat*. File 949.

53. Li, R., Zou, Z. and An, Y., (2016). Water quality assessment in Qu River based on fuzzy water pollution index method. *Journal of environmental sciences*, vol. 50, pp. 87-92.

54. López V, del Río S, Benítez JM, Herrera F. (2015) Cost-sensitive linguistic fuzzy rule based classification systems under the MapReduce framework for imbalanced big data. *Fuzzy Sets Syst*, vol. 258, pp. 5–38.

55. Lu, Y., Zhou, J., Qin, H., Wang, Y., Zhang, Y. (2011) Environmental/economic dispatch problem of power system by using an enhanced multi-objective differential evolution algorithm. *Energy Conversion and Management* vol. 52, pp. 1175–1183.

56. Mahapatra, S. S., Sahu, M., Patel, R. K., & Panda, B. N. (2012). Prediction of Water Quality Using Principal Component Analysis. *Water Quality, Exposure and Health*, vol. 4(2), pp. 93–104. doi:10.1007/s12403-012-0068-9.

57. Mamdani, E. H. (1976). Advances in the linguistic synthesis of fuzzy controllers. *International Journal of Man-Machine Studies, 8(6), 669–678.* doi:10.1016/s0020-7373(76)80028-4.

58. Middleton, M. A., Whitfield, P. H., & Allen, D. M. (2015). Independent component analysis of local-scale temporal variability in sediment-water interface

temperature. *Water Resources Research*, vol. 51(12), pp. 9679–9695. doi:10.1002/2015wr017302.

59. Ocampo-Duque W, Osorio C, Piamba C, Shumahmacher M, Domingo J.L. (2013) Water quality analysis in rivers with non-parametric probability distributions and fuzzy inference systems: Application to the Cauca River, Colombia. *Environment International*. Vol. 52, pp. 17–28.

60. Olofintoye O., Adeyemo J., Otieno F. (2013) Evolutionary Algorithms and Water Resources Optimization. In: Schütze O. et al. (eds) EVOLVE - A Bridge between Probability, Set Oriented Numerics, and Evolutionary Computation II. *Advances in Intelligent Systems and Computing*, vol 175. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31519-0_32.

61. Omran, M. G. H., Engelbrecht, A. P., & Salman, A. (2007). *An overview of clustering methods. Intelligent Data Analysis,* vol. 11(6), pp. 583–605. doi:10.3233/ida-2007-11602.

62. Raman, B. V., Bouwmeester, R., & Mohan, S. (2009). Fuzzy Logic Water Quality Index and Importance of Water Quality Parameters. Air, *Soil and Water Research*, vol. 2, ASWR.S2156. doi:10.4137/aswr.s2156.

63. Riss M. FTSPlot: Fast Time Series Visualization for Large Datasets (2014) https://doi.org/10.1371/journal.pone.0094694.

64. Russo, S., Lürig, M., Hao, W., Matthews, B., Villez, K. (2020) Active Learning for Anomaly Detection in Environmental data, *Environmental Modelling and Software*, https:// doi.org/10.1016/j.envsoft.2020.104869.

65. Scannapieco, D., Naddeo, V., Zarra, T., Belgiorno, V. (2012). River water quality assessment: a comparison of binary and fuzzy logic-based approaches. *Ecol. Eng*. 47, 132e140.

66. Sivert W. (1997) Ecological impact classification with fuzzy sets. *Ecologial Modelling*, vol. 96, pp. 1–10.

67. Shirkhorshidi A.S., Aghabozorgi S., Wah T.Y., Herawan T. (2014) Big Data Clustering: A Review. In: *Murgante B. et al. (eds) Computational Science and Its*

Applications – ICCSA 2014. ICCSA 2014. Lecture Notes in Computer Science, vol 8583. Springer, Cham. https://doi.org/10.1007/978-3-319-09156-3_49.

68. Shyam Reddy K. S., Bindu C. S. (2017) A review on density-based clustering algorithms for big data analysis, *in International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, Palladam, 2017, pp. 123-130, doi: 10.1109/I-SMAC.2017.8058322.

69. Sreekanth, J., Datta, B. (2010) Multi-objective management of saltwater intrusion in coastal aquifers using genetic programming and modular neural network based surrogate models. *Journal of Hydrology*, vol. 393, pp. 245–256.

70. Sugeno M. (1985) Industrial Applications of Fuzzy Control. Elsevier Science Pub. Co., Japan.

71. Steed C. A., Ricciuto D. M., Shipman G. et al., Big data visual analytics for exploratory earth system simulation analysis, *Computers and Geosciences*, vol. 61, pp. 71–82, 2013.

72. Tannahill B. K. and Jamshidi M. (2014) System of systems and big data analytics-bridging the gap, *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 2–15.

73. Underwood J. Data Visualization. Best Practices. https://www.datarobot.com/blog/author/jen-underwood/.

74. Vernica R., Carey M. J., Li C. (2010) Efficient parallel setsimilarity joins using MapReduce, in *Proceedings of the International Conference on Management of Data (SIGMOD '10)*, pp. 495–506, Indianapolis, Ind, USA.

75. Venkata V. Prasad D, Lokeswari Y. Venkataramana, P. Senthil Kumar, Prasannamedha G., Soumya K., Poornema A.J. (2020) Water quality analysis in a lake using deep learning methodology: prediction and validation, International *Journal of Environmental Analytical Chemistry*, doi: 10.1080/03067319.2020.1801665.

76. Wang, H. et al. (2017) An overview on the roles of fuzzy set techniques in big data processing: Trends, challenges and opportunities. *Knowl. Based Syst.* Vol. 118, pp. 15-30.

77. Wang X, Pan H. (2014) Fuzzy comprehensive evaluation of water quality management information system. *Adv Mater Res*. Vol. 1044, pp. 486–489.

78. Wang L., Wang G., Cheryl Ann A. (2015) Big Data and Visualization: Methods, *Challenges and Technology Progress,* Vol. 1, No. 1, 2015, pp 33-38.

79. Wu, Z., Elmaghraby, M., & Pathak, S. (2015). Applications of Deep Learning for Smart Water Networks. *Procedia Engineering*, vol. 119, pp. 479-485.

80. Xu S, Cui Y, Yang C et al (2020) The fuzzy comprehensive evaluation (FCE) and the principal component analysis (PCA) model simulation and its applications in water quality assessment of Nansi Lake Basin, China. Environ Eng Res. https://doi.org/10.4491/eer.2020.022

81. Yin, H. L., & XU, Z. X. (2008). Comparative study on typical river comprehensive water quality assessment methods, *J. Resources and Environment in the Yangtze Basin*, vol. 5(011).

82. Ying, L., Shihu, S., Hongyu, W., Xin, Z., & Qi, Y. (2019). Big Data Analysis of Water Quality of Secondary Water Supply. *Procedia Computer Science*, vol. 154, pp. 744–749. doi:10.1016/j.procs.2019.06.116

83. Zadeh, L. A. (1965). Fuzzy sets. Information and Control 8 (3): 338. doi:10.1016/S0019-9958(65)90241-X.

84. Zerhari, B., Lahcen, A. A., & Mouline, S. (2015). Big data clustering: Algorithms and challenges. In *Proc. of Int. Conf. on Big Data, Cloud and Applications (BDCA'15)*.

РОЗДІЛ 2
МОДЕЛІ ДЛЯ АНАЛІТИЧНОЇ ОБРОБКИ ВЕЛИКИХ ДАНИХ В СИСТЕМІ МОНІТОРИНГУ ВОДНИХ ОБ'ЄКТІВ

В розділі представлено нечіткі моделі для оцінювання якості води на прикладі поверхневих вод та вод рибогосподарського призначення. Нечіткий логічний формалізм використаний для опису якості води шляхом розробки індексу якості води на основі нечітких міркувань. Нечіткі умови описані у вигляді відображення від заданого вхідного детермінанта до детермінанта виходу, використовуючи нечіткі логічні міркування. Модель дозволяє приймати рішення на підставі відображення та розпізнаних шаблонів. Процес нечіткого висновку містить вибір функції належності, операції нечіткого набору та правила висновку.

2.1 Моделі оцінювання якості вод рибогосподарського призначення

Об’єкти аквакультури та інтенсивні рибні господарства вимагають постійного контролю якості вод. Як визначається на сайті Державного агентства рибного господарства [6]: «Інтенсивна форма (рибного господарства – авт.) є найбільш технологічною, дозволяє отримувати найкращі результати, але потребує значних капіталовкладень, фахової підготовки суб’єктів аквакультури». Саме високі вимоги до технологій значно стримують розвиток рибних господарств, через що найбільш розповсюдженою зараз в Україні є напівінтенсивна форма аквакультури, що має відносно невисоку рибопродуктивність та значні ризики пов’язані з хворобами риб.

У процесі збору даних, показники якості води мають очевидні характеристики великих даних. Одночасно, дані, зібрані з кожної точки заміру, містять багато різних показників якості, таких як температура, DO, pH, жорсткість, NH_3–N та інші. В ситуаціях, коли дані вимірюються принаймні щохвилини, традиційні методи аналізу даних стикаються з проблемою аналізу

великих даних за короткий час, а також отримання якості поточних водних умов у певному резервуарі. Тому потрібно знайти швидкий та ефективний метод аналізу та обробки великих даних. З цією метою в даному розділі пропонується використання комплексної нечіткої моделі даних за допомогою якої можливо побудувати інтегральну оцінку, ідею якої запозичено в [17-19, 26, 29], що пропонують розрахунок інтегрального індексу якості води для різних умов використання. Але, на відміну від попередніх робіт, в даному дисертаційному досліджені (1) увагу сконцентровано на якість вод рибогосподарського призначення, (2) зміст 3 параметрів запропонованих в роботі [32] використовується 5, (3) вперше розроблено нечіткі моделі для розрахунку інтегрального індексу якості води для вод рибогосподарського призначення.

Однією з найбільших переваг нечітких моделей є те, що вони використовують досвід людини та дозволяють обробляти інформацію, отриману від експертів для визначення стратегії прийняття рішень. Як результат, модель пропонує рішення швидше, ніж традиційні техніки обробки даних моніторингу. В цьому розділі, нечіткі логічні прийоми використані для оперативного оцінювання якості води на основі лінійних даних та бази правил, створених за допомогою галузевих експертів. У досліджуваному випадку, припустимо, що кожен набір даних оцінки якості води має низку характерних параметрів: водневий показник (pH), розчинений кисень (РК), біологічне споживання кисню (БСК), хімічне споживання кисню (ХСК), азот загальний (N), аміак (NH_3) тощо; одиниця вимірювання – параметрів РК, БСК, ХСК, N та NH_3 – мг/л.

2.1.1 Загальна модель для розрахунку індексу якості води

У загальному виді, нечітка модель аналізу якості води основана на функціональній залежності виду:

$$f: \bar{P} = \{P_1, P_2, \dots, P_n\} \to Z, \tag{2.1}$$

де $\bar{P}$ – множина вхідних лінгвістичних змінних, які містять оцінки кожного з вимірюваних параметрів якості води, Z - вихідна змінна, значення якої відповідає інтегральній оцінці якості вод.

Щоб отримати інтегральну оцінку якості вод, розроблена модель (2.1) поєднує в собі набір нечітких моделей окремих параметрів моніторингу:

$$P = \left\{(p_1,\ \Omega_i^j),\ (p_2,\ \Omega_i^j),(p_3,\ \Omega_i^j),...(p_k,\ \Omega_i^j)\right\},\ p_k \in A \tag{2.2}$$

$$\Omega_i^j = \{\omega_i^j \mid \mu_{\Omega_i^j}(\omega_i^j)\},\ \omega_i^j \in \Omega_i^j; \tag{2.3}$$

де $A = \{p_k\}$ – множина ознак якості води, $k \in [1; n]$; k – індекс ознаки якості води; $\Omega_i^j = \{\omega_i^j\}$ – множина значень ознаки p_k, що представляють собою найменування нечітких змінних; і значення індексів $j \in [1; m]$ які позначають множину значень ознаки.

Опис ознак p_k, характеризує якість води. Таким чином, відомості про якість води можна отримати якщо відома P_i – поточна якість, де *i* є відповідний ідентифікатор якості води.

Побудова функцій належності термів лінгвістичних змінних нечіткої моделі аналізу якості вод рибогоподарського призначення виконувалася з використанням наступної методології:

– визначення параметрів що використовуються в моделі якості вод рибогосподарського призначення;

– формування терм-множин лінгвістичних змінних;

– аналіз наборів даних моніторингу води та аналіз діапазонів гранично-допустимих значень кількісних змінних для кожного зразка, що відповідають лінгвістичним змінним;

– визначення форми та меж інтервалів функцій належності;

– побудова функцій належності для параметрів моніторингу.

2.1.2 Визначення параметрів, що використовуються в моделі якості вод рибогосподарського призначення

Для визначення множини «якість води» {Z} використовується мінімальний набір параметрів, а саме: pH; кількість розчинного кисню; БСК, ХСК, аміак, азот амонійний, які є базовими лінгвістичними змінними моделі, нижче розглянуто їх основні складові. Бажані та прийнятні межі для параметрів якості води, визначені в Директивах ЄС [20, 21] (табл. 2.1).

Таблиця 2.1 – Бажані та прийнятні межі для параметрів якості води (Джерела: [5, 20, 21])

Параметр якості води	**Бажана межа**	**Прийнятна межа**
1. Водневий показник, pH	6	9
2. Розчинний кисень, мг/л	8	6
3. БСК, мг/л	0-3	4-5
4. ХСК, мг/л	1-10	10-25
5. Аміак, мг/л	0-0,1	0,1-0,3
6. Азот загальний, мг/л	0,2	1

Більш детальний розподіл значень параметрів для оцінювання якості вод рибогосподарського призначення представлено в [5]. В результаті аналізу нормативів, для задач дисертаційного дослідження використовувалися три рівні якості (термів): I – бажаний; II – прийнятний; III – неприйнятний. Граничні значення для кожного параметра надано у табл.2.2

Таблиця 2.2 – Параметри та їх гранично-допустимі значення за [5], прийняті для побудови функцій належності

Параметр якості води	**Рівні якості**		
	I	**II**	**III**
Водневий показник, pH	6,5-8,4	6,0-6,4 8,6-9,0	0,0- 4,8 9,2-14
Розчинений кисень (РК) (влітку), мг/л	8,0-14,6	6,0-8,0	0,0-6,0
БСК, мг/л	0,0-3,0	4,0-5,0	> 6,0
Аміак NH_3 мг/л	0,0-0,1	0,1-0,3	0,3-2,7
ХСК, мг/л	1,0-10,0	10,0-25,0	> 25,0
Азот загальний, мг/л	≤ 0,04	≤ 1	≤ 0,2
Азот амонійний, NH_4^+ мг/л	≤ 0,005	≤ 0,025	> 0,025

Далі, після побудови для кожного параметра таблично-заданих функцій належності, було проведено їх апроксимацію класичними функціями належностей, а саме: трикутною, трапецієподібною та симетричною функцією Гауса для вибору виду функцій належності кожної терм-множини вхідних лінгвістичних змінних якості води. В результаті, для побудови функцій належності для всіх параметрів якості води було обрано трапецієподібну функцію, що в загальному виді описується рівнянням (2.4).

$$\mu^f(x;a,b,c,d) = \begin{Bmatrix} 0, & x \le a \\ \dfrac{x-a}{b-a}, & a \le x \le b \\ 1, & b \le x \le c \\ \dfrac{d-x}{d-c}, & c \le x \le d \\ 0, & d \le x. \end{Bmatrix}, \qquad (2.4)$$

де x – параметр, що підлягає фазифікації, a, b, c, d – лінгвістичні змінні, які використовуються для розподілу параметрів на різні класи, відповідно, a і d – параметри функції належності, що задають нижню основу трапеції та відповідно песимістичну оцінку значень лінгвістичної змінної, b і c – параметри функції належності, що відповідають верхній основі трапеції та визначають оптимістичні оцінки значень лінгвістичної змінної.

Для опису параметрів в табл. 2.2, що не мають двосторонніх обмежень в якості функції належності використано часткове представлення трапецієподібної функції належності у вигляді R- та L-функції, відповідно ф. (2.5) та ф. (2.6).

R-функція з параметрами $a = b = -\infty$

$$\mu^f(x;c,d) = \begin{cases} 0, & x > d \\ \dfrac{d-x}{d-c}, & c \le x \le d \\ 1, & x < c \end{cases}, \qquad (2.5)$$

L-функція з параметрами $c = d = +\infty$

$$\mu^f(x;a,b) = \begin{cases} 0, & x < a \\ \dfrac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}, \qquad (2.6)$$

Для визначення меж для інтервалів $X = [X_{min}, X_{max}]$ кожного інтервального розбиття параметрів що перекриваються і отримання значень лінгвістичних змінних a, b, c, d трапецієподібної функції належності в місті перетину визначених термів (рис. 2.3) використано формули (2.7):

$$
\begin{aligned}
a_{i+1} &= \frac{\mathrm{X}_{min,i+1} + \mathrm{X}_{max,i}}{2} - k, \\
b_{i+1} &= \frac{\mathrm{X}_{min,i+1} + \mathrm{X}_{max,i}}{2} + k, \\
c_i &= \frac{\mathrm{X}_{min,i+1} + \mathrm{X}_{max,i}}{2} - k, \\
d_i &= \frac{\mathrm{X}_{min,i+1} + \mathrm{X}_{max,i}}{2} + k,
\end{aligned}
\tag{2.7}
$$

де $\mathrm{X}_{\min}, \mathrm{X}_{\max}$ – мінімальні та максимальні значення, для кожного з показників інтервального розбиття;

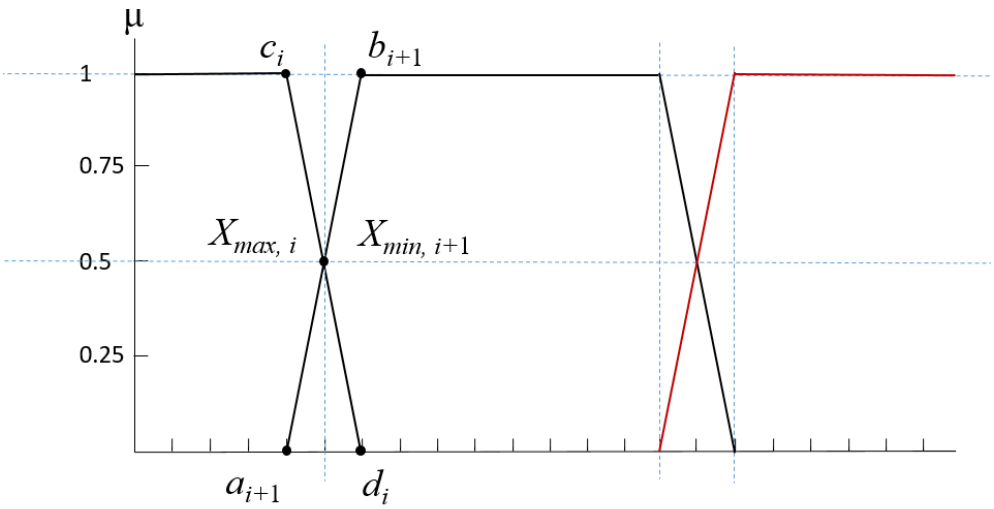


Рисунок 2.1 – Схема розбиття на межі інтервалів $X_i = [X_{min,i}, X_{max,i}]$ і $X_{i+1} = [X_{min,i+1}, X_{max,i+1}]$

2.1.3 Побудова функцій належності для параметрів моніторингу

2.1.3.1 Водневий показник (pH)

Водневий показник або параметр pH визначає кількість іонів водню (H+) у воді й характеризує співвідношення кислоти та лугу в розчині. Він вимірюється за шкалою від 0 до 14. Нейтральне середовище має показник pH=7,0. Значення pH від 0 до 6,9 характеризують кислотне середовище, високі значення показника,

від 7,1 до 14 - лужне. Для вод господарсько-питного водопостачання цей параметр регулюється ДСТУ 4077-2001 [7] з максимально граничним значенням 6,5-8,5 (тобто від слабо кислого до слабо лужного). Допустимий діапазон для вирощування риб також становить від 6,5 до 8,5 (у деяких джерелах [35] - 9,0). Показники pH 6 та 9 є відповідно гранично допустимими межами для рибогосподарських підприємств. За даними компанії Tempcon Instrumentation Ltd [16], значення ≤ 4 та ≥ 11 вважаються абсолютно не припустимими (табл. 2.3).

Таблиця 2.3 – Вплив pH на розвиток риб [16]

pH	Вплив pH на рибу
4	Кислотна точка загибелі
4 - 5	Репродукція не можлива
4 - 6,5	Повільне зростання
6,5 - 9	Бажаний діапазон для розмноження риби
9 - 10	Повільне зростання
≥ 11	Лужна точка загибелі

Враховуючи визначені вище рекомендації а також [31, 35], в якості основної, прийнято наступну шкалу оцінки для параметра pH вод рибогосподарського призначення: від 6,5 до 8,4 оцінюється як клас I «бажаний» (дуже прийнятний), 4,8 - 6,5 та 8,4-9,0 – як II «прийнятний», і 0 - 4,8 та > 9,0 – як клас III «неприйнятний».

В моделі оцінювання якості води (2.1), pH баланс є параметром, за який відповідає лінгвістична змінна P_1, що визначається кортежем $\langle P_1, \Omega(P_1), X\rangle$, де P_1 = «pH баланс», $\Omega(P_1)$={I, II, III}, $X = [X_{min}, X_{max}]$.

Параметри термів для інтервального розбиття X представлені в табл. 2.4. Для розрахунку меж використано ф. (2.7).

Таблиця 2.4 – Параметри термів лінгвістичної змінної P_1 (pH)

Ім'я терма	Ім'я функції	Параметри				Діапазон	
		a	b	c	d	Xmin	Xmax
$\Omega_1^1 = \text{I}$	$\mu_I(x; a, b, c, d)$	6,3	6,7	8,2	8,6	Xmin_1	Xmax_1
$\Omega_2^1 = \text{II}$	$\mu_{II}(x; a, b, c, d)$	4,6	5,0	6,3	6,7	Xmin_2	Xmax_2
$\Omega_2^1 = \text{II}$	$\mu_{II}(x; a, b, c, d)$	8,2	8,6	8,8	9,2	Xmin_3	Xmax_3
$\Omega_3^1 = \text{III}$	$\mu_{III}(x; c, d)$	0	0	4,6	5,0	Xmin_4	Xmax_4
$\Omega_3^1 = \text{III}$	$\mu_{III}(x; a, b)$	8,8	9,2			Xmin_5	Xmax_5

Графік функцій належності для термів лінгвістичної змінної β_1 представлено на рис. 2.2.

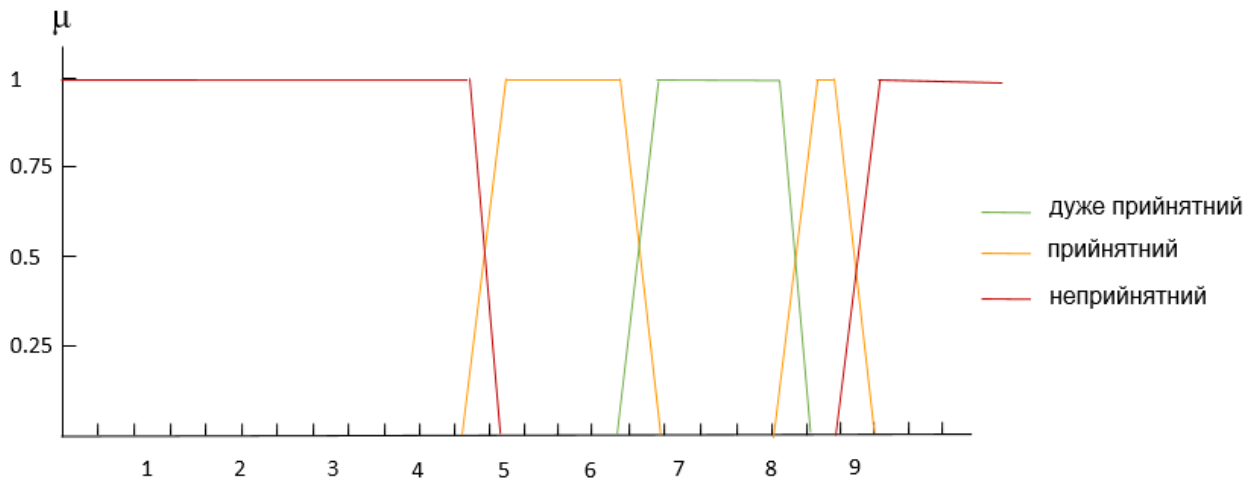


Рисунок 2.2 – Графік функцій термів лінгвістичної змінної P_1 (pH)

2.1.3.2 Розчинений кисень

Розчинений кисень (dissolved oxygen, DO) є одним з найважливіших параметрів, що потребують постійного контролю. Риби «дихають» киснем так само, як це роблять наземні тварини. Однак риби здатні поглинати кисень безпосередньо з води у кров за допомогою зябер.

У загальному випадку, є три основних джерела кисню у водному середовищі [24]: (1) пряма дифузія з атмосфери; (2) дія вітру та хвилі; і (3) фотосинтез.

Концентрація розчиненого кисню у ставках залежить від багатьох параметрів: температури, солоності води, атмосферного тиску, кількості, щільності риби та коливається протягом доби, збільшуючись в світлий час доби, коли відбувається фотосинтез, і зменшуючись вночі, коли дихання риб триває, але фотосинтез не відбувається.

Вміст розчиненого кисню регулюється ДСТУ ISO 5813:2004 [8] і не повинен бути нижчим 4 мг/л незалежно від цілей її використання. Для оптимального здоров'я риб рекомендується концентрація від 5 мг/л DO. Чутливість до низького рівня розчиненого кисню є специфічною для певного виду, однак більшість видів риб страждають, коли DO падає до 2-4 мг/л. Кількість риб, які гинуть під час виснаження кисню, визначається тим, наскільки низьким стає РК і як довго цей показник залишається в такому стані.

В моделі оцінювання якості води (2.1), РК є параметром, за який відповідає лінгвістична змінна P_2, що визначається кортежем $\langle P_2, \Omega(P_2), X\rangle$, де P_2 = «РК», $\Omega(P_2)$={I, II, III}, $X = [X_{min}, X_{max}]$.

Параметри термів для інтервального розбиття X (табл. 2.5):

Таблиця 2.5 – Параметри термів лінгвістичної змінної P_2 (РК)

Ім’я терма	Ім’я функції	Параметри				Діапазон	
		a	b	c	d	Xmin	Xmax
$\Omega_1^2 = \text{I}$	$\mu_I(x; a, b)$	7,8	8,2			$Xmin_1$	$Xmax_1$
$\Omega_2^2 = \text{II}$	$\mu_{II}(x; a, b, c, d)$	5,8	6,2	7,8	8,2	$Xmin_2$	$Xmax_2$
$\Omega_3^2 = \text{III}$	$\mu_{III}(x; a, b, c, d)$	2,8	3,2	5,8	6,2	$Xmin_3$	$Xmax_3$

від > 8 мг/л оцінюється як дуже прийнятний,

6 - 8 мг/л – як прийнятний і

3 - 6 мг/л – як неприйнятний (за умов, коли кількість розчиненого кисню знижується менше 3 мг/л відбувається масовий замор риби).

Графік функцій належності для термів лінгвістичної змінної P_2 представлено на рис. 2.2.

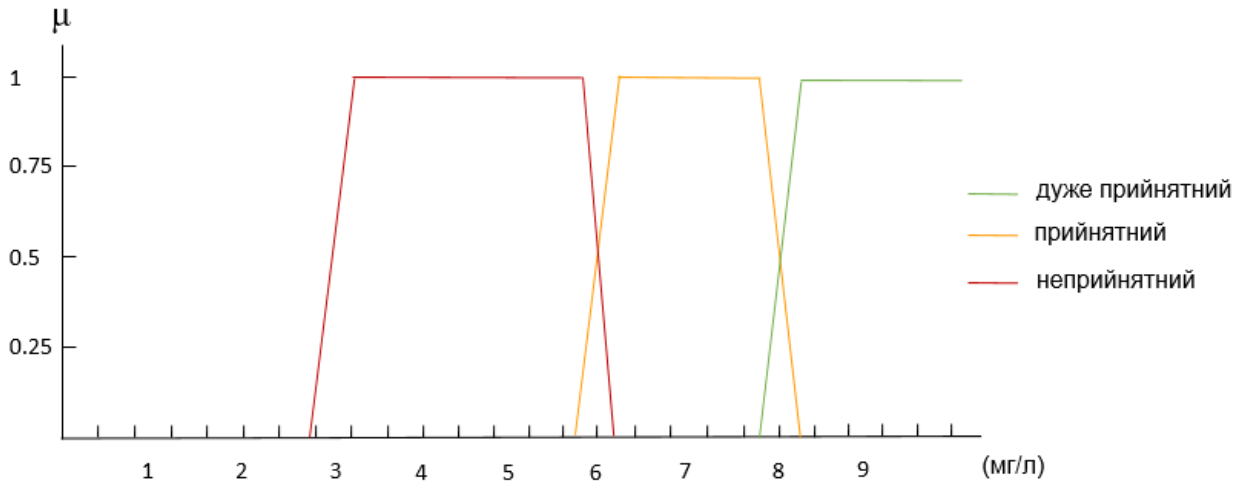


Рисунок 2.3 – Графік функцій термів лінгвістичної змінної P_2 (РК)

2.1.3.3 Біологічне споживання кисню

Біологічне споживання кисню (БСК) - це міра вмісту органічних речовин, як природного (наприклад, рослинного і тваринного матеріалу що руйнується), так і техногенного (нафтопродукти, органічні хімікати, тощо), які для хімічних перетворень використовують кисень, що розчинений у воді.

Природний органічний матеріал може реагувати з хлором на водоочисних спорудах, утворюючи шкідливі побічні продукти що потребують дезінфекції у питній воді.

БСК належить до узагальнених показників якості води, оскільки може служити в якості оцінки її загального забруднення органічними сполуками, що легко окислюються. Для вод господарсько-питного водопостачання цей параметр регулюється ДСТУ ISO 5815-2:2009 [9] з максимально граничним значенням 3 мг/л.

В моделі оцінювання якості води (2.1), БСК є параметром, за який відповідає лінгвістична змінна P_3, що визначається кортежем $\langle P_3, \Omega(P_3), X\rangle$, де P_3 = «БСК», $\Omega(P_3)$={I, II, III}, $X = [X_{min}, X_{max}]$.

Норма шкали позначає рівень БСК від 0 до 3 мг/л як дуже прийнятний, 4-5 мг/л як прийнятний і вище 6 мг/л як неприйнятний (табл. 2.5).

Таблиця 2.5 – Параметри термів лінгвістичної змінної P_3 (БСК)

Ім'я терма	Ім'я функції	Параметри				Діапазон	
		a	b	c	d	Xmin	Xmax
$\Omega_1^3 =$ I	$\mu_I(x; c, d)$			3	4	$Xmin_1$	$Xmax_1$
$\Omega_2^3 =$ II	$\mu_{II}(x; a, b, c, d)$	3	4	5	6	$Xmin_2$	$Xmax_2$
$\Omega_3^3 =$ III	$\mu_{III}(x; a, b, c, d)$	2,8	3,2	5,8	6,2	$Xmin_3$	$Xmax_3$

Графік функцій належності для термів лінгвістичної змінної P_3 представлено на рис. 2.4.

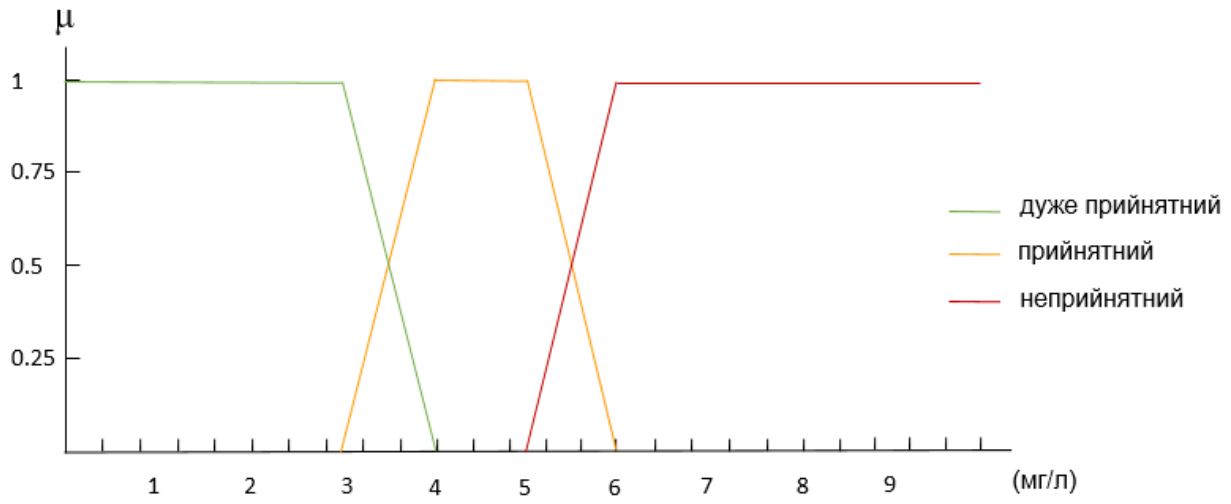


Рисунок 2.4 – Графік функцій термів лінгвістичної змінної P_3 (БСК)

2.1.3.4 Хімічне споживання кисню

Хімічне споживання кисню (ХСК) є мірою вмісту у воді хімічних речовин, як органічного, так й неорганічного походження, які при хімічному перетворенні використовують кисень, розчинений у воді й часто співвідноситься БСК у воді.

Для вод господарсько-питного водопостачання цей параметр регулюється ДСТУ ISO 6060:2003 [10] з максимально граничним значенням 15 мг/л.

В моделі оцінювання якості води (2.1), ХСК є параметром, за який відповідає лінгвістична змінна P_4, що визначається кортежем $\langle P_4, \Omega(P_4), X\rangle$, де P_4 = «ХСК», $\Omega(P_4)$={I, II, III}, $X = [X_{min}, X_{max}]$.

Використана рейтингова шкала позначає рівні ХСК наступним чином:

від 1 до 10 мг/л – дуже прийнятний,

10-25 мг/л – прийнятний

вище 25 мг/л – неприйнятний.

Характеристики застосованих функцій належності до трапеції (a, b, c, d) для параметру ХСК надано в табл. 2.6.

Таблиця 2.6 – Параметри термів лінгвістичної змінної P_4 (ХСК)

Ім’я терма	Ім’я функції	Параметри				Діапазон	
		a	b	c	d	Xmin	Xmax
$\Omega_1^4 = \text{I}$	$\mu_I(x; c, d)$			9,8	10,2	Xmin_1	Xmax_1
$\Omega_2^4 = \text{II}$	$\mu_{II}(x; a, b, c, d)$	9,8	10,2	24,8	25,2	Xmin_2	Xmax_2
$\Omega_3^4 = \text{III}$	$\mu_{III}(x; a, b)$	24,8	25,2			Xmin_3	Xmax_3

Графік функцій належності для термів лінгвістичної змінної P_4 представлено на рис. 2.5.

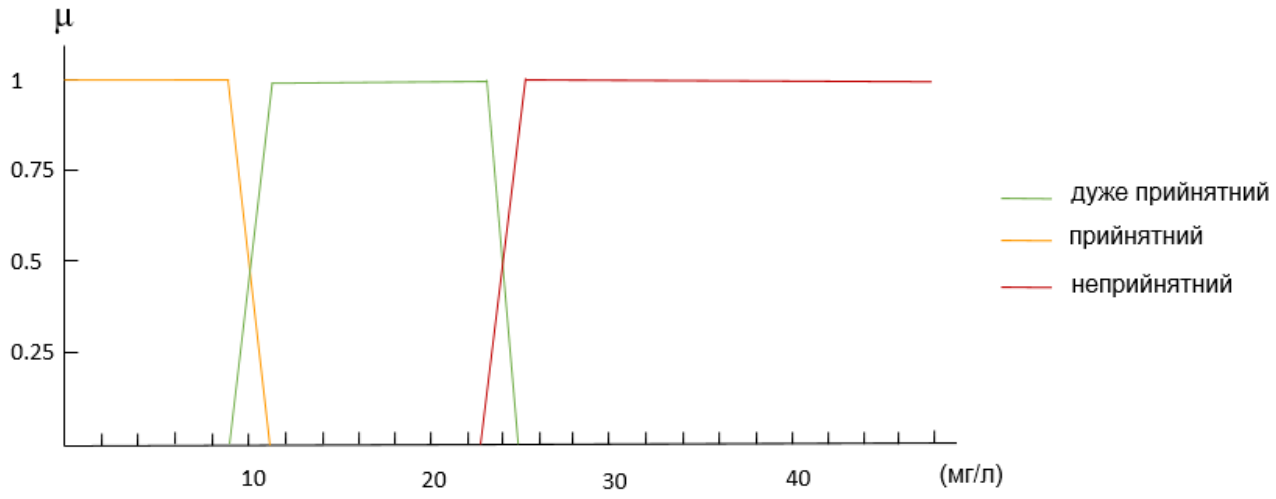


Рисунок 2.5 – Графік функцій термів лінгвістичної змінної P_4 (ХСК)

2.1.3.5 Аміак

Аміак – одне з джерел азоту (NH_3), важлива поживна речовина для рослин та водоростей, а аміак відбувається у нешкідливій формі, але при більш високій температурі або більшому pH аміак змінюється на газ, форму, шкідливу для риб та іншого водного життя. Аміак виводиться з тварин і утворюється під час розкладання рослин і тварин. Аміак є інгредієнтом багатьох мінеральних добрив, а також присутній у стічних водах, стічних водах, деяких промислових стічних водах та стічних водах для тварин.

У більшості річок та озер аміак існує переважно в іонізованому вигляді (NH_4^+). З підвищенням pH та температури іонізований аміак змінюється на неіонізований газ аміаку (NH_3). Аміак може бути токсичним для риб та інших водних організмів при підвищенні концентрації. Якщо присутня достатня кількість розчиненого кисню (DO), аміак легко розщеплюється нітрифікуючими бактеріями, утворюючи нітрити та нітрати.

В моделі оцінювання якості води (2.1), аміак є параметром, за який відповідає лінгвістична змінна P_5, що визначається кортежем $\langle P_5, \Omega(P_5), X\rangle$, де P_5 = «аміак», $\Omega(P_5)$={I, II, III}, $X = [X_{min}, X_{max}]$.

Використовувана шкала оцінки позначає аміак

від 0 до 0,1 мг/л як низький (оцінюється як бажаний або дуже прийнятний),

0,1 - 0,3 мг/л як середній (оцінюється як прийнятний) і

з 0,3 до 2,7 як високий (оцінюється як неприйнятний).

Характеристики застосованих функцій належності до трапеції (a, b, c, d) для параметру «аміак» надано в табл. 2.7.

Таблиця 2.7 – Параметри термів лінгвістичної змінної P_5 (аміак)

Ім'я терма	Ім'я функції	Параметри				Діапазон	
		a	b	c	d	Xmin	Xmax
$\Omega_1^5 = \text{I}$	$\mu_I(x; c, d)$			0,08	0,12	Xmin_1	Xmax_1
$\Omega_2^5 = \text{II}$	$\mu_{II}(x; a, b, c, d)$	0,08	0,12	0,28	0,32	Xmin_2	Xmax_2
$\Omega_3^5 = \text{III}$	$\mu_{III}(x; a, b)$	0,28	0,32			Xmin_3	Xmax_3

Графік функцій належності для термів лінгвістичної змінної P_5 представлено на рис. 2.6.

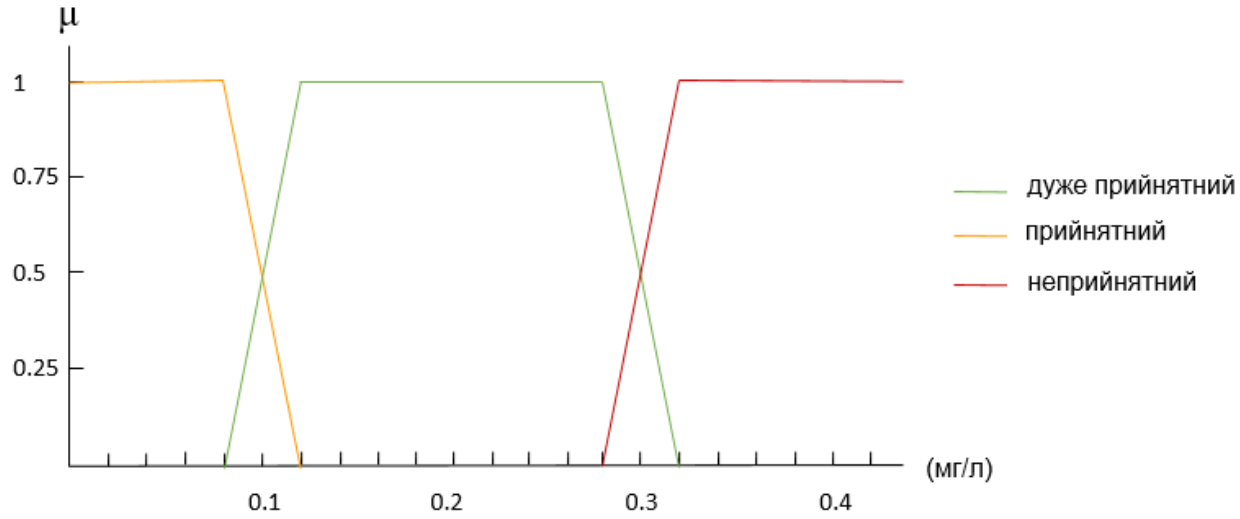


Рисунок 2.6 – Графік функцій термів лінгвістичної змінної P_5 (аміак)

2.2 Правила нечіткого виведення для оцінки якості вод рибогосподарського призначення

2.2.1 Узагальнена структура технології використання нечітких правил

Згідно прийнятого розподілу, вхідні параметри моделі були розділені на три класи: Ω_1 – бажаний (значення концентрації менше або дорівнюють бажаним інтервалам), Ω_2 – прийнятний (значення концентрації між бажаним і допустимим межами) і Ω_3 – неприйнятний (значення концентрації перевищують допустимі межі).

Узагальнену структуру технології використання нечітких правил для оцінки якості вод рибогосподарського призначення на основі інтегрованого нечіткого індексу якості вод надано на рис. 2.7.

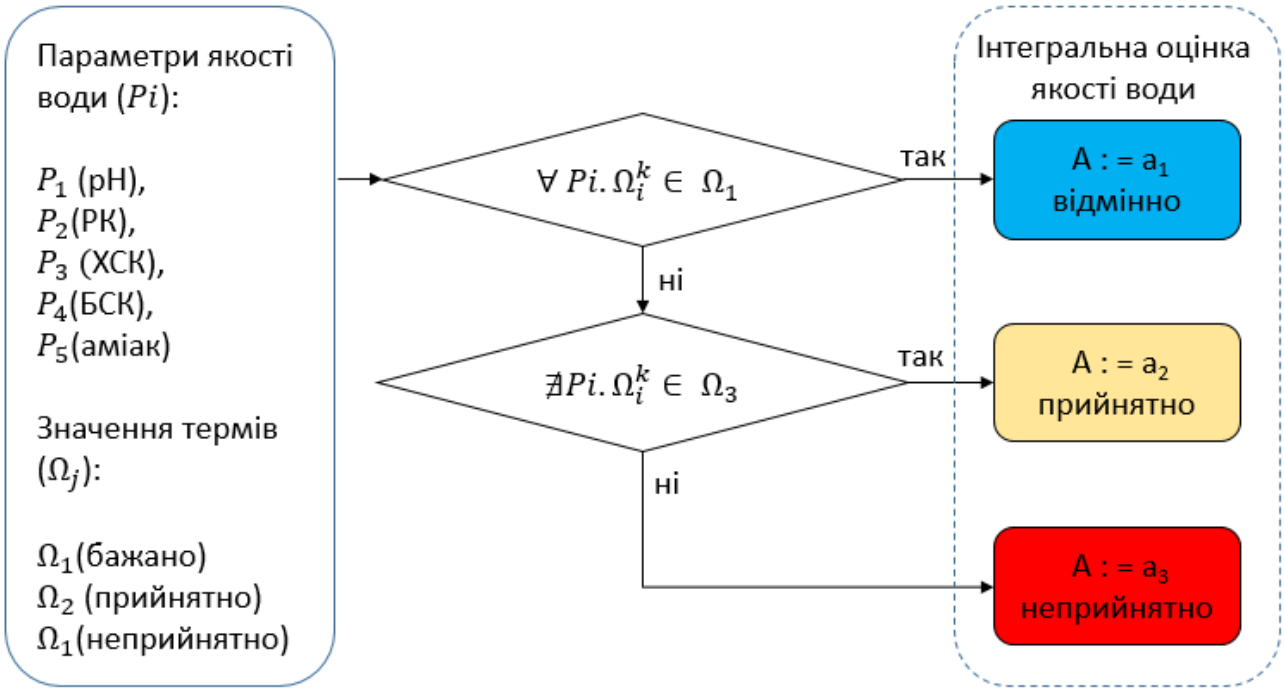


Рисунок 2.7 – Узагальнена структура технології використання нечітких правил для оцінки якості вод рибогосподарського призначення

В якості правил нечіткого виведення для локальних моделей використовуються нечіткого правила *modus ponens* [27]:

$$I : x = A \Rightarrow y = B$$
$$P : x = A' \qquad (2.8)$$
$$C : y = B'$$

де *I* позначає імплікацію, *P* - передумова, *C* - висновок. Це правило дає нам можливість зробити висновок про наступника на основі логічного значення попередника. Імплікація трактується як нечітке відношення, що означає, що A' не повинно бути рівним *A*, а B' не повинно бути рівним *B*. Цілком достатньо, коли A' подібний до *A* і таким чином B' подібний до *B*.

Основною передумовою при розробці правил нечіткого виведення було: «навіть якщо один параметр якості води перевищує допустимі межі, визначені для вод рибогосподарського призначення, тоді вода не підходить для технологічного циклу вирощування й потребує негайного очищення». Попередньо, усі правила нечіткого виведення були розроблені з урахуванням цієї головної передумови. Набір нечітких правил у цьому випадку складається з наступних правил:

Правило 1: «Якщо значення усіх параметрів знаходяться в межах Ω_1 – бажаний, тоді інтегральний показник якості води приймається a_1 – відмінно».

Правило 2: «Якщо значення хоча б одного з параметрів виходять за межі Ω_1 – бажаний та жоден з параметрів не належить до класу Ω_3 – неприйнятний, тоді інтегральний показник якості води приймається a_2 – прийнятно».

Правило 3: «Якщо значення хоча б одного з параметрів належить до класу Ω_3 – неприйнятний, тоді інтегральний показник якості води приймається a_3 – неприйнятно».

Зрозуміло, що такі правила достатньо жорстко нормують інтегральний показник якості вод, і вимагають додаткових оцінок щодо можливого втручання в технологічний цикл.

2.2.2 Розширена структура технології використання нечітких правил

З метою уніфікації, виходи (значення інтегрального індексу) також збільшено до п'яти класів, а саме a_1 – відмінний, a_2 – добрий, a_3 – задовільний, a_4 – поганий, a_5 – дуже поганий (рис. 2.8). Такий розподіл обрано згідно «Методики віднесення масиву поверхневих вод до одного з класів екологічного та хімічного станів масиву поверхневих вод ...» [15], яка визначає п'ять класів: I – відмінний (позначається синім кольором), II – добрий (зелений колір), III – задовільний (жовтий колір), IV – поганий (помаранчевий колір), V – дуже поганий (червоний колір).

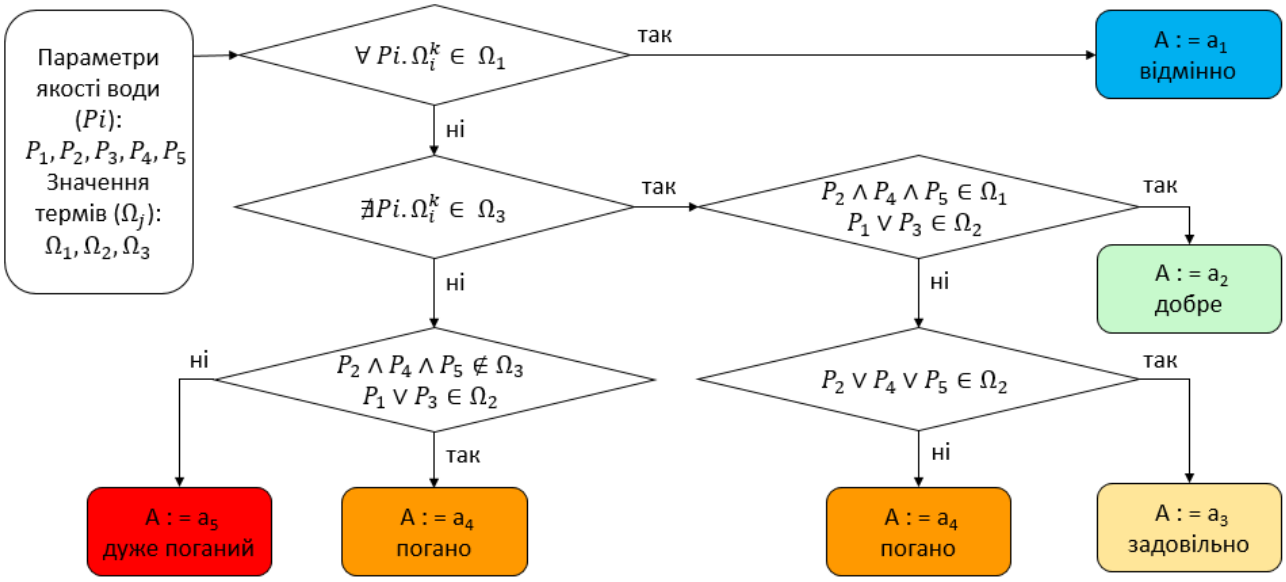


Рисунок 2.8 – Розширена структура технології використання нечітких правил для оцінки якості вод рибогосподарського призначення

Нечіткі правила для досліджуваної водної екосистеми були розроблені за допомогою знання експертів та ретельного розгляду параметрів отриманих від системи моніторингу, що мають потенційний ризик для здоров'я риб (тобто наборів РК, БСК, NO_3 та pH, ХСК). Ця основна передумова забезпечує практичність розробленого індексу якості води при оцінці придатності вод рибогосподарського призначення.

Вихідна лінгвістична змінна «оцінка якості води» визначена у вигляді кортежу $A = \langle a_1, a_2, a_3, a_4, a_5 \rangle$, де терми a_1... a_5 відповідають класам якості: a_1 – відмінно, a_2 – добре, a_3 – задовільно, a_4 – погано, a_5 – дуже погано. Терм-множина A ={"відмінно","добре","задовільно", "погано", "дуже погано"}.

Спільно з експертами галузі інтенсивного рибогосподарства, було введено нуступну шкалу інтервалів для термів лінгвістичної змінної $A \in [0..10]$:

a_1 – відмінно, $a_1 \in [8,9,10]$;

a_2 – добре, $a_2 \in [6..7]$;

a_3 – задовільно, $a_3 \in [4..5]$;

a_4 – погано, $a_4 \in [2..3]$;

a_5 –дуже погано, $a_5 \in [0..1]$.

Критерії віднесення масиву поверхневих вод до одного з класів екологічного стану наведено у Додатку Д.

Графік функцій належності для термів лінгвістичної змінної А представлено на рис. 2.9.

Після дефазифікації, згідно з алгоритмом Мамдані, буде отримано лінгвістична змінна А-«оцінка якості води» у вигляді числового значення в інтервалі, $A \in [0..10]$.

Важливо зазначити, що представлені моделі є універсальними прикладами в яких правила і межі можуть бути легко адаптовані до прийнятих в господарстві стандартів щодо вод рибогосподарського призначення (відносно технологічних вимог до вирощування відповідного виду риб.

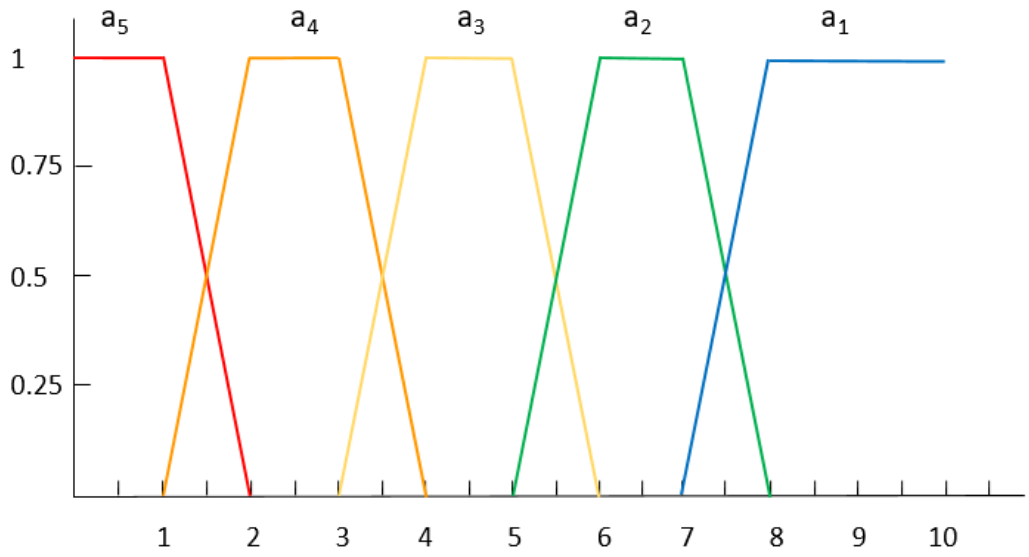


Рисунок 2.9 – Графік функцій термів вихідної лінгвістичної змінної А

2.2.3 Агрегування і дефазифікація нечітких правил

Після розробки правил нечіткого виведення, виконується об’єднання всіх розроблених правил. У якості схеми нечіткого виведення в даній роботі використано алгоритм Мамдані [11, 13, 28].

Алгоритм використовує метод min-активації (2.9)

.

$$\mu'_i(x) = \min \{d_i, \mu_i(x)\}, \qquad (2.9)$$

де $\mu'_i(x)$ – «активізована» функція належності;

$\mu_i(x)$ – функція належності;

d_i – ступінь істинності і-го часткового висновку.

Для обчислення об’єднаного нечіткого виведення необхідна агрегація усіх правил нечіткого виведення. Для представлення та логічної зв’язки усіх правил застосовується нечітка кон'юнкція. Агрегування в свою чергу, виконується за допомогою min-кон'юнкції (2.10):

$$c_j = \min\{b_i\}, \tag{2.10}$$

де $j = 1..n$;

i – число з множини підумов, в яких бере участь j-а вхідна змінна.

Метод max-диз'юнкції [33] був використаний для акумуляції висновків правил, як процедура агрегування що застосовує операцію об'єднання на всіх усічених нечітких наборах вихідних даних (2.11):

$$\mu'_i(x) = \max\ \{\mu_1(x), \mu_2(x)\}, \tag{2.11}$$

де $\mu_1(x)$, $\mu_2(x)$ – функції належності поєднуваних множин.

Дефазифікація сукупного вихідного значення для перетворення нечітких наборів у числове значення досягається центроїдним методом [33], оскільки це найбільш поширений і прийнятний з усіх доступних методів [13].

Математично, дефазифікований вихід y_i буде представлений як

$$X^* = \frac{\int_{Min}^{Max} \mu_i(x) \cdot x dx}{\int_{Min}^{Max} \mu_i(x) \cdot dx} \tag{2.12}$$

де $\mu_i(x)$ – функція належності відповідної нечіткої множини E_i;

Min і *Max* – кордони універсуму нечітких змінних;

X^* – результат дефазифікації.

Остаточний числовий бал є нечіткий індекс якості води.

У загальному випадку, індекси якості води є математичними виразами використовуваними для оцінки якості води шляхом обробки набору даних, зібраного системами моніторингу в різних числових шкалах і приведених до однієї шкали. Індекс є загальновизнаним підходом до оцінки якості води, але для

різних завдань і систем моніторингу використовуються різні набори змінних [18]. Процеси відбору, обробки, перетворення, зважування параметрів та їх інтерпретація також відрізняються. Разом з тим, універсальність та доступність такого підходу дозволяє використовувати його як інструмент комунікації для політиків, дослідників та різних державних органів для надійного повідомлення громадськості про стан водних об’єктів.

2.3 Моделі нечіткої комплексної оцінки поверхневих вод

З метою розширення сфери використання підходу, було розроблено моделі комплексної оцінки поверхневих вод, основою яких є відповідні функції належності. В даний час для обчислення функції належності використовується ступінчастий метод зменшеної половини трапеції.

Відповідно до діючого в Україні стандарту оцінки «Методика екологічної оцінки якості поверхневих вод за відповідними категоріями» [14], поверхневі води поділяються на п‘ять рівнів (класів) і 7 категорій якості води за екологічним станом: I клас (1 категорія) – відмінні; II клас (2 категорія) – дуже добрі II клас (3 категорія) добрі; III клас (4 категорія) – задовільні, III клас (5 категорія) – посередні; IV клас (6 категорія) – погані; V клас (7 категорія) – дуже погані.

За ступенем чистоти (забруднення): I клас (1 категорія) – дуже чисті; II клас (2 категорія) – чисті II клас (3 категорія) досить чисті; III клас (4 категорія) – слабко забруднені, III клас (5 категорія) – помірно забруднені; IV клас (6 категорія) – брудні; V клас (7 категорія) – дуже брудні.

Екологічна класифікація природних вод, що використовується в країнах Європейського союзу (ЄС) також виділяє п'ять класів якості води [12, 22]: I – відмінна якість; II – добра якість; III – задовільна якість; IV – незадовільна якість; V – погана якість. Дані щодо граничних параметрів надано у Додатку В.

Отже, враховуючи прийнятну практику [25], формули для визначення рівня належності якості води можуть бути представлені наступним набором (2.13-2.15):

Клас I

$$\mu_{i1} = \begin{cases} 1 & x_i \le s_{i1} \\ \dfrac{s_{i2} - x_i}{s_{i2} - s_{i1}} & s_{i1} < x_i < s_{i2} \\ 0 & x_i > s_{i1} \end{cases} \quad (2.13)$$

Клас II-IV:

$$\mu_{ij} = \begin{cases} 1 - \dfrac{s_{ij} - x_i}{s_{ij} - s_{ij-1}} & s_{ij-1} \le x_i \le s_{i1} \\ 0 & x_i < s_{ij-1}, \quad x_i > s_{ij+1} \\ \dfrac{s_{ij+1} - x_i}{s_{ij+1} - s_{ij}} & s_{ij} < x_i < s_{ij+1} \end{cases} \quad (2.14)$$

Клас V

$$\mu_{ij} = \begin{cases} 0 & x_i \le s_{i4} \\ 1 - \dfrac{s_{i5} - x_i}{s_{i5} - s_{i4}} & s_{in-1} < x_i < s_{in} \\ 1 & x_i > s_{i5} \end{cases} \quad (2.15)$$

де x_i - вимірювана концентрація i-го індексу оцінки, s_{ij} - стандартне значення рівня j-го і-го індексу оцінювання, а μ_{ij} - ступінь належності індексу оцінювання i-го рівня до j- рівень якості води.

З отриманих функцій належності може бути визначена матриця оцінки нечітких відношень M:

$$M = \begin{bmatrix} \mu_{11} & \mu_{12} & \mu_{13} & \mu_{14} & \mu_{15} \\ \mu_{21} & \mu_{22} & \mu_{23} & \mu_{24} & \mu_{25} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{n1} & \mu_{n2} & \mu_{n3} & \mu_{n4} & \mu_{n5} \end{bmatrix} \quad (2.16)$$

Оскільки різні фактори впливають на якість води по-різному, необхідно обчислювати вагу кожного фактора, щоб зробити модель оцінки більш прийнятною до аналізу.

Етапи визначення коефіцієнтів ваги:

Стандартизація вимірюваних даних. Дані складаються з *n* індексів оцінки та *m* об'єктів оцінювання, які утворюють матрицю *X*:

$$X = \begin{bmatrix} x_{11} & x_{12} \cdots & x_{1m} \\ x_{21} & x_{22} \cdots & x_{2m} \\ \vdots & \vdots \ddots & \vdots \\ x_{n1} & x_{n2} \cdots & x_{nm} \end{bmatrix}, \tag{2.17}$$

Нормалізація X:

$$y_{ij} = \frac{\max\limits_{j}\{x_{ij}\} - x_{ij}}{\max\limits_{j}\{x_{ij}\} - \min\limits_{j}\{x_{ij}\}} \tag{2.18}$$

Стандартизація і розрахунок матриці судження Y:

$$Y = \begin{bmatrix} y_{11} & y_{12} \cdots & y_{1m} \\ y_{21} & y_{22} \cdots & y_{2m} \\ \vdots & \vdots \ddots & \vdots \\ y_{n1} & y_{n2} \cdots & y_{nm} \end{bmatrix} \tag{2.19}$$

Щоб усунути вплив розмірності та порядок величин, вихідна матриця даних X нормалізується за допомогою методу Z-Score.

Нормалізація виконується наступним чином

$$Z = \frac{x_{ij} - \overline{x_j}}{S_j} \tag{2.20}$$

де $\overline{x_j}$ - середня кількість j-х зразків, а S_j - стандартне відхилення j-х зразків. Формули для обчислення $\overline{x_j}$ та S_j:

$$\overline{x_j} = \frac{1}{m}\sum_{i=1}^{m} x_{ij} \tag{2.21}$$

$$S_j = \sqrt{\frac{1}{m-1}\sum_{i=1}^{m}\left(x_{ij} - \overline{x_j}\right)^2} \tag{2.22}$$

Таким чином, визначена нормована матриця оцінки нечітких відношень може бути використана для розпізнавання потенційних забруднювачів. Аналогічні підходи можуть бути використані й для інших важливих показників якості води, наприклад для оцінки вмісту лужних та лужно земельних металів, важких та кольорових металів, фосфатів, хлоридів, сульфатів, сірководню й т.ін.

З метою розпізнавання потенційних забруднювачів в різних дослідженнях використовуються багатофакторні методи, такі як кластерний аналіз і аналіз головних компонент (PCA), [23, 30].

Кластерний аналіз є додатково одним із багатовимірних методів, що застосовуються для оцінки відносної подібності в однорідності оцінюваних параметрів [34]. У Розділі 3 пропонується застосування кластерного аналізу на випадок великих даних.

Висновки до розділу 2

В розділі вперше розроблено нечіткі моделі для оцінювання якості води на прикладі поверхневих вод та вод рибогосподарського призначення. Модель для аналітичної обробки великих даних системи моніторингу водних об’єктів на основі формалізації її атрибутів та інтерпретації невизначеності оцінки якості

води у вигляді лінгвістичних змінних дозволяє надати інтегровану характеристику стану водних об’єктів, для подальшого прийняття рішення.

Нечіткий логічний формалізм використаний для опису якості води шляхом розробки і використання індексу якості води на основі нечітких міркувань. Процес нечіткого виведення містить вибір функції належності, операції нечіткого набору та правила виводу.

Результати розділу опубліковано в роботах автора [1-4].

Список літератури до розділу 2

1. Барбарук Л.В., Зубарев Д.В., Медведєв Є.М., Барбарук В.М. “Використання нечітких моделей для планування збиральних робіт у різних кліматичних умовах”, *Вісник Східноукраїнського національного університету ім. В. Даля*, №6 (247), С. 175-179, 2018.

2. Барбарук Л.В., Суворін О.В. “Система аналізу даних та підтримки прийняття рішень з інвентаризації промислових відходів”, *Вісник Східноукраїнського національного університету ім. В. Даля*, №8 (238). С. 5-12. 2017.

3. Барбарук В.М., Барбарук Л.В. “Методи моделювання складних систем”, *Вісник Східноукраїнського національного університету ім. В. Даля*, № 15(186), С. 132-136, 2012.

4. Барбарук Л.В., Михайличенко С.О. Бездротова система віддаленого моніторингу сільськогосподарських параметрів. Матеріали VII Всеукр. науково-практ. конф. «Електронні апарати та системи. Проблеми створення. Перспективи розвитку», Сєвєродонецьк : Східноукр. нац. ун-т ім. В. Даля, 2017. С. 206-208.

5. Вербецька К.Ю. Порівняльний аналіз методик оцінки якості поверхневих вод (на прикладі типової р. Губісцкалі). *Вісник Національного*

університету водного господарства та природокористування. Серія «Сільськогосподарські науки». 2011, Вип. 5 (11). С. 91–99.

6. Державне агентство рибного господарства. Офіційний веб-сайт https://darg.gov.ua/ (05.01.2021).

7. ДСТУ 4077-2001. Якість води. Визначення pH. Вперше; введ. 2003-07-01. К.: Державний комітет України з питань технічного регулювання та споживчої політики, 2003. 12 с.

8. ДСТУ ISO 5813:2004. Якість води. Визначення розчинного кисню. Йодометричний метод. Вперше; введ. 2006-01-01. К.: Держспоживстандарт, 2005. 8 с.

9. ДСТУ ISO 5815-2:2009. Якість води. Визначення біохімічного споживання кисню після n днів (БСКn). Частина 2. Метод для нерозведених проб. Взамін ДСТУ ISO 5815:2004; введ. 2011-07-01. К.: Держспоживстандарт, 2010. 12 с.

10. ДСТУ ISO 6060:2003. Якість води. Визначення хімічної потреби в кисні. Вперше; введ. 2004-07-01. К.: Держспоживстандарт, 2005. 10 с.

11. Каргин А. А. Введение в интеллектуальные машины. Книга 1. Интеллектуальные регуляторы / А. А. Каргин. – Донецк: Норд-Прессс, ДонНУ, 2010. – 526 с.

12. Клименко М.О., Вознюк Н.М., Вербецька К.Ю. Порівняльний аналіз нормативів якості поверхневих вод. *Наукові доповіді Національного університету біоресурсів та природокористування.* Київ, 2012. Вип. 1(30). Режим доступу: http://nd.nubip.edu.ua/2012_1/12kmo.pdf (01.01.2021)

13. Леоненков А. В. Нечеткое моделирование в среде MATLAB и fuzzy TECH, СПб. : БХВ-Петербург, 2005. 736 с.

14. Методика екологічної оцінки якості поверхневих вод за відповідними категоріями / Романенко В.Д., Жукинський В.М., Оксіюк О.П. та ін. – К.: Символ-Т, 1998. 28с.

15. Методика віднесення масиву поверхневих вод до одного з класів екологічного та хімічного станів масиву поверхневих вод, а також віднесення штучного або істотно зміненого масиву поверхневих вод до одного з класів екологічного потенціалу штучного або істотно зміненого масиву поверхневих вод. Режим доступу: https://zakon.rada.gov.ua/laws/show/z0127-19#Text (12.12.2020).

16. Aquaculture water monitoring Режим доступу: https://www.enviromonitors.co.uk/aquaculture-water-monitoring/ (20.12.2020).

17. Abbasi T, Abbasi SA (2012) Water quality indices. *Elsevier, Amsterdam,* 384 p.

18. Bharti N, Katyal D (2011) Water quality indices used for surface water vulnerability assessment. *Int J Ecol Environ Sci*. vol. 2(1), pp. 154–173.

19. Cao T.S., Thị H. G. N., Trieu P.T., Nguyen H.N., Nguyen T.L., Vo H.C. (2020) Assessment of Cau River water quality assessment using a combination of water quality and pollution indices. *J Water Supply Res Technol Aqua* vol. 69(2), pp. 160–172. https://doi.org/10.2166/aqua.2020.122.

20. Directive 2008/105/EC of the European Parliament and of the Council of 16 December 2008 on environmental quality standards in the field of water policy, amending Directive 2000/60/EC of the European Parliament and of the Council. Режим доступу: http://ec.europa.eu/environment/water/waterdangersub/lib_pri_ substances. htm (23.12.2020).

21. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Режим доступу: https://eur-lex.europa.eu/legal-content/EN/TXT/ ?uri=CELEX%3A02000L0060-20141120 (06.12.2020).

22. Ecological status of surface water bodies. Режим доступу: https://www.eea.europa.eu/themes/water/european-waters/water-quality-and-water-assessment/water-assessments/ecological-status-of-surface-water-bodies (31.12.2020).

23. Fan X, Cui B, Zhao H et al (2010) Assessment of river water quality in Pearl river Delt using multivariate statistical techniques. *Procedia Environ Sci*. vol. 2, pp. 1220–1234.

24. Francis-Floyd R. Dissolved Oxygen for Fish Production http://agrilife.org/fisheries2/files/2013/09/Dissolved-Oxygen-for-Fish-Production1.pdf (20.12.2020).

25. Jiang, Y.; Gui, H.; Yu, H.; Wang, M.; Fang, H.; Wang, C.; Chen, C.; Zhang, Y.; Huang, Y. (2020) Hydrochemical Characteristics and Water Quality Evaluation of Rivers in Different Regions of Cities: A Case Study of Suzhou City in Northern Anhui Province, China. *Water* 2020, vol. 12, 950. https://doi.org/10.3390/w12040950.

26. Jha M.K., Shekhar A., Jenifer M.A. (2020) Assessing groundwater quality for drinking water supply using hybrid fuzzy-GIS-based water quality index. *Water Res.* 2020 Jul 15;179:115867. doi: 10.1016/j.watres.2020.115867. Epub 2020 May 3. PMID: 32408184.

27. Kacprzyk J. (1986) Fuzzy sets in system analysis. Państwowe Wydawnictwo Naukowe, Warsaw.

28. Mamdani E. H. (1974) Application of fuzzy algorithms for control of simple dynamic plant. *Proceedings of the Institution of Electrical Engineers*, vol. 121(12), 1585. doi:10.1049/piee.1974.0328.

29. Matta G., Nayak, A., Kumar, A. *et al.* (2020) Water quality assessment using NSFWQI, OIP and multivariate techniques of Ganga River system, Uttarakhand, India. *Appl Water Sci* vol. 10, 206 https://doi.org/10.1007/s13201-020-01288-y.

30. Misaghi F, Delgosha F, Razzaghmanesh M, Myers B (2017) Introducing a water quality index for assessing water for irrigation purposes: a case study of the Ghezel Ozan River. *Sci Total Environ* vol. 589, pp. 107–116.

31. Raman B. V., Bouwmeester, R., & Mohan, S. (2009). Fuzzy Logic Water Quality Index and Importance of Water Quality Parameters. *Air, Soil and Water Research*, 2, ASWR.S2156. doi:10.4137/aswr.s2156.

32. Rana D., Rani, S. (2015). Fuzzy logic based control system for fresh water aquaculture: A MATLAB based simulation approach. *Serbian Journal of Electrical Engineering*, vol. 12, pp. 171-182.

33. Ross T. J. Fuzzy logic with engineering applications.–3rd ed. 2010 John Wiley & Sons, Ltd. ISBN: 978-0-470-74376-8

34. Shrestha S, Kazama F (2007) Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, *Japan. J Environ Model Softw* vol. 22, pp. 464–475.

35. Swann L. (2000). A fish farmer's guide to understanding water quality. Illinois–Indiana Sea grant program, Purdue University. Режим доступу: http://ag.ansc.purdue.edu/aquanic/publicat/state/il-in/as-503.htm (13.12.2020).

РОЗДІЛ 3
МЕТОДИ ОБРОБКИ ВЕЛИКИХ ДАНИХ ДЛЯ АНАЛІЗУ ЯКОСТІ ВОД ВОДОЙМ РИБОГОСПОДАРСЬКОГО ПРИЗНАЧЕННЯ

У третьому розділі представлено вдосконалений метод нечіткої кластеризації *c*-середніх для великих даних водойм рибогосподарського призначення. Розглянуто проблему інтерпретації результатів кластеризації на випадок застосування евристичних алгоритмів кластеризації для інтуїтивістської нечіткої обробки даних. Показано приклад використання запропонованих методів для аналізу якості вод водойм рибогосподарського призначення. Наведені результати дослідження запропонованих методів.

3.1 Особливості нечіткої кластеризації наборів великих даних

Нехай набір даних X являє собою сукупність N об'єктів, представлених у вигляді векторів у F-вимірному просторі: $X=\{x_1, x_2, \ldots, x_n\} \subseteq \Re^F$. Тоді, у загальному випадку, розділ або кластеризація в X - це набір непересічних кластерів, що розділяє X на K груп: $C=\{c_1, c_2, \ldots c_K\}$, де $\bigcup_{c_k \in C} c_k = X,\ c_k \cap c_l = \emptyset\ \forall k \neq l.$ Центроїд кластера c_k є його середнім вектором

$$c_k = \frac{1}{|c_k| \sum_{x_i \in c_k} x_i},$$

аналогічно, центроїд набору даних є середнім вектором усього набору даних,

$$\bar{X} = \frac{1}{N \sum_{x_i \in X} x_i}.$$

Задачею кластеризації є призначення кожного $x_i \in X$ лише до одного кластеру. Причому, нечітка кластеризація використовується тоді, коли межі між кластерами є невизначеними та заплутаними.

Кластеризований набір даних як у випадку звичайної (жорсткої), так і нечіткої кластеризації може бути представлений матрицею U розмірністю $c \times n$. Дотримуючись позначення [13], матриця виглядає наступним чином:

$$U = \begin{bmatrix} u_{11} & u_{12} \cdots & u_{1n} \\ u_{21} & u_{22} \cdots & u_{2n} \\ \vdots & \vdots \ddots & \vdots \\ u_{c1} & u_{c2} \cdots & u_{cn} \end{bmatrix},$$

де c – кількість кластерів, а n – кількість точок даних у вихідному наборі даних. Кожен запис у U позначається u_{ji}, де $j \in \{1,\ldots, c\}$, $i \in \{1,\ldots, n\}$.

За умов використання звичайної (жорсткої) кластеризації u_{ji} може приймати значення лише 0 або 1. Якщо точка i знаходиться в кластері j, то $u_{ji} = 1$. Інакше $u_{ji} = 0$. Окрім цього, в умовах жорсткої кластеризації є дві умови, яким U має відповідати:

$$\sum_{j=1}^{c} u_{ij} = 1, \tag{3.1}$$

$$\sum_{i=1}^{n} u_{ij} > 0, \tag{3.2}$$

Вираз (3.1) передбачає, що об'єкт даних i може належати лише одному кластеру, тобто лише один запис може приймати значення 1 у будь-якому конкретному стовпці. Для виконання умови (3.2) не має бути порожніх кластерів, тобто кожен рядок повинен містити запис значення 1.

При нечіткій кластеризації умова (3.1) дещо відрізняється, а саме, для нечіткої кластеризації дозволяється $u_{ji} \in [0, 1]$. Тобто, умова (3.1) вимагає, щоб для кожного об'єкта даних сума його ступенів належності у всіх кластерах дорівнювала 1. Умова (3.2), як і раніше, вимагає, щоб не було порожніх кластерів.

Одним із прикладів кластеризації є метод k-середніх (на випадок нечіткої кластеризації – c-середніх). Як і його аналог з жорсткою кластеризацією, метою нечіткого алгоритму є мінімізація заданої функції.

Припустимо, що існує набір даних $D = \{x_1, ..., x_n\}$ і $q \in [0, 1]$, де q характеризує нечіткість, кластерів. Чим більше значення q, тим менші значення належності u_{ji} і, відповідно, нечіткість кластеризації. Цільова функція визначається як

$$E_q = \sum_{i=1}^{n} \sum_{j=1}^{c} u_{ji}^q \, d^2(x_i, V_j), \tag{3.3}$$

де d - метрична функція скалярного добутку, а V_j - центроїди початкової кластеризації даних. В початковій кластеризації D дозволяється перекриття даних допоки всі точки включені щонайменше в один кластер. На будь-якій ітерації ступінь належності точки даних x_i кластеру j дорівнює

$$u_{ji} = \left(\frac{d^2(x_i, V_j)}{\sum_{l=1}^{n} d^2(x_i, V_l)} \right)^{\frac{1}{1-q}}. \tag{3.4}$$

Кожен центроїд V_j перераховується таким чином:

$$V_j = \frac{\sum_{i=1}^{n} u_{ji}^q x_i}{\sum_{i=1}^{n} u_{ji}^q}. \tag{3.5}$$

Після кожної ітерації порівнюються матриці належності послідовних кроків. Якщо різниця між абсолютними значеннями попереднього u_{ji}^{old} і нового u_{ji}^{new} значень не перевищує деякого заданого критерію стійкості ε

$$\max_{ji}\left|u_{ji}^{old}-u_{ji}^{new}\right|<\varepsilon, \tag{3.6}$$

та виконанні умови (3.2), алгоритм нечітких с-середніх вважається завершеним. В іншому випадку обчислюється матриця належності за допомогою нових центроїдів.

Як було визначено в розділі 1, евристичні методи, ієрархічні методи та методи, засновані на об'єктивних функціях, є основними підходами у нечіткій кластеризації. Метод нечітких с-середніх (fuzzy c-means, далі – FCM) [8, 9] є найпопулярнішим підходом у нечіткій кластеризації. Хоча використання підходів нечітких с-середніх, як правило, більш трудомістке та складне, ніж їх аналоги з жорсткою кластеризацією, нечітка кластеризація має важливе застосування завдяки своїй гнучкості у групуванні даних з невизначеними параметрами. У багатьох випадках [14, 16], щоб мати доступ до локальних мінімумів та отримувати оптимальну кластеризацію використовується алгоритм FCM що дозволяє виконувати постійну оптимізацію вибірки за допомогою ітеративної та цільової функції подібності центру с-кластерів. Однак метод кластеризації FCM надзвичайно трудомісткий, оскільки в алгоритмі кластеризації FCM приналежність, що використовується для позначення ступеня кожної точки даних, належить кластеру [14, 18].

Виходячи з цього має сенс скористатися існуючими розширеннями ступеня належності, що надає інтуїтивістський нечіткий набір.

3.2 Основні визначення інтуїтивістського підходу до евристичної кластеризації

У цьому підрозділі розглянуто основні поняття та визначення підходу, заснованого на інтуїтивістській теорії нечітких множин, яку запропонував Атанасов [6] на випадок нечіткої евристичної кластеризації.

3.2.1 Інтуїтивістський нечіткий набір

Нехай $X = \{x_1, x_2, ..., x_n\}$ – універсум, а $\omega = (\omega_1, \omega_2, ..., \omega_n)^T$ – ваговий вектор елементів x_j ($j = 1, 2, ..., n$), де $\omega_j \geq 0$ і $\sum_{j=1}^{n} \omega_j = 1$. Нечіткий набір (3.7)

$$A = \{\langle x_j, \mu_A(x_j)\rangle | x_j \in X\} \tag{3.7}$$

визначений [28] характеризується функцією належності μ_A: $X \rightarrow [0, 1]$, де μ_A (x_j) позначає ступінь належності елемента x_j до множини A.

Тоді, узагальнений інтуїтивістський нечіткий набір IFS A або (IFS)A на X [7], може бути представлено наступним чином:

$$A = \{\langle x_j, \mu_A(x_j), \nu_A(x_j)\rangle | x_j \in X\}, \tag{3.8}$$

де μ_A – функція належності, а ν_A - функція не належності, визначені як

$$\begin{aligned} &\mu_A: X \rightarrow [0,1], \qquad x_j \in X \rightarrow \mu_A(x_j) \in [0,1], \\ &\nu_A: X \rightarrow [0,1], \qquad x_j \in X \rightarrow \nu_A(x_j) \in [0,1], \end{aligned} \tag{3.9}$$

що відповідають умові:

$$\mu_A(x_j) + \nu_A(x_j) \leq 1, \quad \forall\, x_j \in X$$

Для кожного (IFS)A в X, якщо

$$\pi_A(x_j) = 1 - \mu_A(x_j) - \nu_A(x_j), \tag{3.10}$$

де π_A (x_j) – ступінь невизначеності x_j до A.

Інтуїтивістський нечіткий індекс π_A (x_j) можна розглядати як ступінь невпевненості у належності x_i до (IFS)*A*. Видно, що $0 \le \pi_A$ (x_j) ≤ 1 для всіх $x_i \in X$. Очевидно, що за умов (3.11)

$$\pi_A(x_j) = 1 - \mu_A(x_j) - \nu_A(x_j) = 0, \qquad \forall\, x_j \in X, \tag{3.11}$$

(IFS)*A* зводиться до нечіткого набору.

3.2.2 Особливості інтуїтивістського підходу до евристичної кластеризації

Розглянемо інтуїтивістські розширення основних понять нечіткої кластеризації, які були запропоновані в [23].

Нехай {(IFS)A^1, …, (IFS)A^n} – інтуїтивістські нечіткі множини на *X*, які породжені інтуїтивістською нечіткою толерантністю.

Тоді (α, β)-рівень інтуїтивістського нечіткого набору

$$(\text{IFS})\text{A}^l_{\alpha,\beta} = \begin{Bmatrix} \big(x_i, \mu_l(x_i), \nu_l(x_i)\big) \ge \alpha, \\ \nu_l(x_i) \le \beta, \;\; x_i \in X \end{Bmatrix},$$

де $(\text{IFS})\text{A}^l_{\alpha,\beta}$ визначає рівень інтуїтивістського нечіткого (α, β)-кластера або, просто, інтуїтивістський нечіткий кластер, в якому μ_{li} є ступенем належності елемента $x_i \in X$ для якогось інтуїтивістського нечіткого кластера $(IFS)\text{A}^l_{\alpha,\beta}$, $\alpha \in (0,1]$, $\beta \in [0,1)$, $l \in \{1, \ldots, n\}$; а ν_{li} – ступінь не належності елемента $x_i \in X$ до інтуїтивістського нечіткого кластера $(IFS)\text{A}^l_{\alpha,\beta}$.

Ступінь належності елемента $x_i \in X$ інтуїтивістського нечіткого кластера $(IFS)\mathrm{A}^{l}_{\alpha,\beta}$, $\alpha \in (0,1]$, $\beta \in [0,1)$, $0 \le \alpha + \beta \le 1$, $l \in \{1,\ldots,n\}$ можна визначити за (3.12):

$$\mu_{lj} = \begin{cases} \mu_l(x_i), & x_i \in (IFS)A^{l}_{\alpha,\beta}, \\ 0, & \text{у іншому випадку,} \end{cases} \tag{3.12}$$

де α, β-рівнем $(IFS)\mathrm{A}^{l}_{\alpha,\beta}$ інтуїтивістського нечіткого набору $(IFS)\mathrm{A}^{l}_{\alpha,\beta}$ є підтримка інтуїтивістського нечіткого кластера $(IFS)\mathrm{A}^{l}_{\alpha,\beta}$.

З іншого боку, ступінь не належності елемента $x_i \in X$ для інтуїтивістського нечіткого кластера $(IFS)\mathrm{A}^{l}_{\alpha,\beta}$, $\alpha \in (0,1]$, $\beta \in [0,1)$, $0 \le \alpha + \beta \le 1$, $l \in \{1,\ldots,n\}$ може бути визначена як

$$v_{lj} = \begin{cases} v_l(x_i), & x_i \in (IFS)A^{l}_{\alpha,\beta}, \\ 0, & \text{у іншому випадку.} \end{cases} \tag{3.13}$$

Отже, ступінь не належності може трактуватися як міра нетиповості елемента щодо інтуїтивістського нечіткого кластеру.

3.2.3 Міри подібності інтуїтивістських нечітких множин для прийняття рішень з множинними атрибутами

Нехай s: $\Phi(X)^2 \to [0, 1]$ – множина всіх IFS з X. Використаємо поняття міри подібності між двома IFS, запропоноване в [24].

Нехай s – відображення s: $\Phi(X)^2 \to [0, 1]$, тоді ступінь подібності між $A \in \Phi(X)$ і $B \in \Phi(X)$ визначається як $s(A, B)$, що задовольняє наступним властивостям:

(1) $0 \leq s(A, B) \leq 1$,

(2) $s(A, B) = 1$ iff $A = B$,

(3) $s(A, B) = s(B, A)$,

(4) Якщо $s(A, B) = 0$ та $s(A, C) = 0$, $C \in \Phi(X)$, то $s(B, C) = 0$.

В якості мір подібності на основі геометричної моделі відстані можуть бути використані відстань Хемінга, нормалізована відстань Хемінга, Евклідова відстань, нормалізована Евклідова відстань. При розрахунку подібності нечітких множин зазвичай використовується Евклідова відстань:

$$d(A,B) = \sqrt{\sum_{j=1}^{n} \left(\mu_A(x_j) - \mu_B(x_j)\right)^2}. \tag{3.14}$$

Евклідова відстань на випадок *IFS* розраховується наступним чином [22]:

$$d(A,B) = \sqrt{\sum_{j=1}^{n} \left(\mu_A(x_j) - \mu_B(x_j)\right)^2 + \left(\nu_A(x_j) - \nu_B(x_j)\right)^2 + \left(\pi_A(x_j) - \pi_B(x_j)\right)^2}.$$

У ряді випадків, можливо використання узагальненої моделі, запропонованої в [25]:

$$d(A,B) = \left[\frac{1}{2}\sum_{j=1}^{n} \left(\left(\mu_A(x_j) - \mu_B(x_j)\right)^2 + \left(\nu_A(x_j) - \nu_B(x_j)\right)^2 + \left(\pi_A(x_j) - \pi_B(x_j)\right)^2\right)\right]^{\frac{1}{2}}. \tag{3.15}$$

Оскільки для аналізу багатовимірних даних слід враховувати вагу кожного елемента $x_j \in X$, наприклад, при прийнятті рішень з множиною атрибутів, атрибути зазвичай мають різне значення, необхідно враховувати зважену відстань. З цією метою, в дисертаційному дослідженні використано розширений випадок (3.15) де зважену відстань визначено з урахуванням вагових коефіцієнтів:

$$d(A,B) = \left[\frac{1}{2}\sum_{j=1}^{n} \omega_j \left(\left(\mu_A(x_j) - \mu_B(x_j)\right)^{\alpha} + \left(\nu_A(x_j) - \nu_B(x_j)\right)^{\alpha} + \left(\pi_A(x_j) - \pi_B(x_j)\right)^{\alpha}\right)\right]^{\frac{1}{\alpha}}, \quad (3.16)$$

де $\omega = (\omega_1, \omega_2, ..., \omega_n)^T$ – ваговий вектор елементів x_j (j = 1, 2, ..., n), α = 2.

З огляду на вищевикладене, задача нечіткої кластеризації може бути визначення наступним чином.

3.3 Постановка задачі нечіткої кластеризації

Нехай $X = \{x_1, x_2, \ldots, x_n\} \subset \Phi^s$ – набір точок даних без кластеризації, c - кількість кластерів, нечітка матриця належності $U=[u_{ij}]_{c\times n}$ вказує на нечіткий розділ точок даних, u_{ij} – елемент матриці U, що характеризує приналежність точки даних x_i до кластеру. Тоді, задача нечіткої кластеризації зводиться до побудови нечіткої матриці належності U яка має відповідати наступним умовам (3.17-3.19) [26]:

$$\forall i,j \quad u_{ij} \in [0,1], \quad (3.17)$$

$$\sum_{i=1}^{c} u_{ij} = 1, \quad (3.18)$$

$$0 < \sum_{i=1}^{n} u_{ij} > n, \tag{3.19}$$

Згідно [8], узагальнимо задачу нечіткого розподілу C як проблему екстремальних значень, що відповідає цільовій функції в умовах обмеження:

$$G_m(U,V) = \sum_{j=1}^{c}\sum_{i=1}^{n} u_{ij}^m d_{ij}^2, \tag{3.20}$$

де $V = (v_{1i}, v_{2i},\ldots,v_{si})$ – матриця розмірністю $s \times c$, $v_i = (v_{1i}, v_{2i},\ldots ,v_{si})^{\mathrm{T}} \in \Phi^s$ – центр i-го кластера, $d_{ij} = \|x_j - v_i\|$ – Евклідова відстань між центром i-го кластера і j-ю точкою даних, m – нечіткий індекс, який використовується для керування нечітким ступенем належності матриці класифікації.

З (3.20) зрозуміло, що чим більше m, тим вище ступінь нечіткості класифікації.

При m = 1, метод кластеризації FCM перероджується в метод звичайної (жорсткої) кластеризації. Отже, цільова функція $G_m(U, V)$ – це квадратична сума зваженої відстані для кожної точки даних до центру кластера.

Метод FCM - це ітеративна процедура рішення, що мінімізує цільову функцію $G_m(U, V)$. У багатьох випадках, для вирішення задачі оптимізації (3.20) за умов обмеження (3.17) застосовується метод множника Лагранжа. Отже, для розрахунку U і V можна отримати наступні формули:

$$u_{ij}(k) = \frac{1}{\sum_{r=1}^{c}\left(\frac{d_{ij}(k)}{d_{rj}(k)}\right)^{\frac{2}{m-1}}}, \tag{3.21}$$

$$1 \le i \le c;\ 1 \le j \le p.$$

$$V_i(k+1) = \frac{\sum_{j=1}^{p} u_{ij}^{m}(k) Z_j}{\sum_{j=1}^{p} u_{ij}^{m}(k)}. \quad (3.22)$$

В методі кластеризації FCM початковий центр кластера задається випадковим чином. Одночасно, коригуються центри класифікації та кластеризації відповідно до формул (3.21), (3.22) на кожній ітерації. Допоки зміна центру кластера, що виходить з двох суміжних ітерацій, не перевищує заданого ε, вважається, що алгоритм сходиться.

Разом з тим варто відзначити, що за результатами тестових завдань, виконуваних системою SmartWater [19, 20] мало місце велике ітеративне навантаження, що призводило до того, що підхід, описаний вище мав низьку ефективність. Таким чином, зроблено висновок, що традиційні методи кластерного аналізу не підходять для аналізу великих багатовимірних даних. Зокрема, для обробки великих даних, отриманих з системи он-лайн моніторингу якості води такий підхід не відповідає вимогам швидкості та ефективності моніторингу якості води. З метою покращення показників, в роботі пропонується використання удосконаленого методу, наданого в наступному підрозділі.

3.4 Удосконалений метод нечіткої кластеризації с-середніх для великих даних

Для викладення основної ідеї пропонованого підходу, зроблено наступне припущення. У методі кластеризації FCM, при $m \in [1, 5]$, нечіткий центр тяжіння кластера сильно наближений до жорсткого центру тяжіння кластера. Відхилення між ними дещо збільшується зі збільшенням m, а схожість між ними при максимальному відхиленні становить 0,98. Згідно [29] відхилення становить близько 1 для типового $m = 2$. В роботі [8] вказано, що має сенс використовувати діапазон є $m \in [1.5, 2.5]$, тому у подальшому взято значення $m = 2$. У цьому випадку, нечіткий центр тяжіння кластера дуже близький до жорсткого центру

тяжіння. Тому жорсткий центр тяжіння кластера можна розглядати як початкове значення нечіткого центру тяжіння, що дозволить прискорити швидкість конвергенції та зменшити кількість ітерацій

3.4.1 Основні етапи вдосконаленого методу нечіткої кластеризації

Виходячи з вищевикладеного, в роботі був вдосконалений метод нечіткої кластеризації *c*-середніх на випадок роботи з великими даними, що складається з наступних кроків (рис. 3.1).

На першому етапі проводиться аналіз обсягів даних і приймається рішення щодо потреби в розділенні та розпаралелюванні даних. Для кластеризації даних вище 10^{10} розділення є абсолютно необхідним.

За умови вхідного набору даних алгоритм починається з одного кластера S, що відповідає цілому набору даних. S вводиться в чергу пріоритетів, Q. Q містить блоки розділу, який ітеративно будується над набором даних.

Поки Q не порожній, блок B видаляється і розбивається на пару блоків. Якщо розбиття ефективне, пари блоків замінюють B в Q, інакше B стає „остаточним” блоком для цієї фази. Для того, щоб ефективно знайти найкращий розділ для блоку, слід обчислити граничні розподіли блоку. Зокрема, для кожної розмірності i d-вимірного блоку B необхідно обчислити функції $C_B^i: \mathbb{R} \longrightarrow \mathbb{N}$ C $LS_B^i: \mathbb{R} \to \mathbb{R}^d$, визначені наступним чином

$$C_B^i(x) = |p \in B \wedge p[i] = x|$$

$$LS_B^i(x) = \sum_{p \in B \wedge p[i] = x} p$$

де підсумовуються всі значення точок p, що належать до блоку B, які мають значення і-ї координати, рівної x. Ці функції можна представити як карти, або, якщо припустити, що координати точок є цілими числами, як масиви.

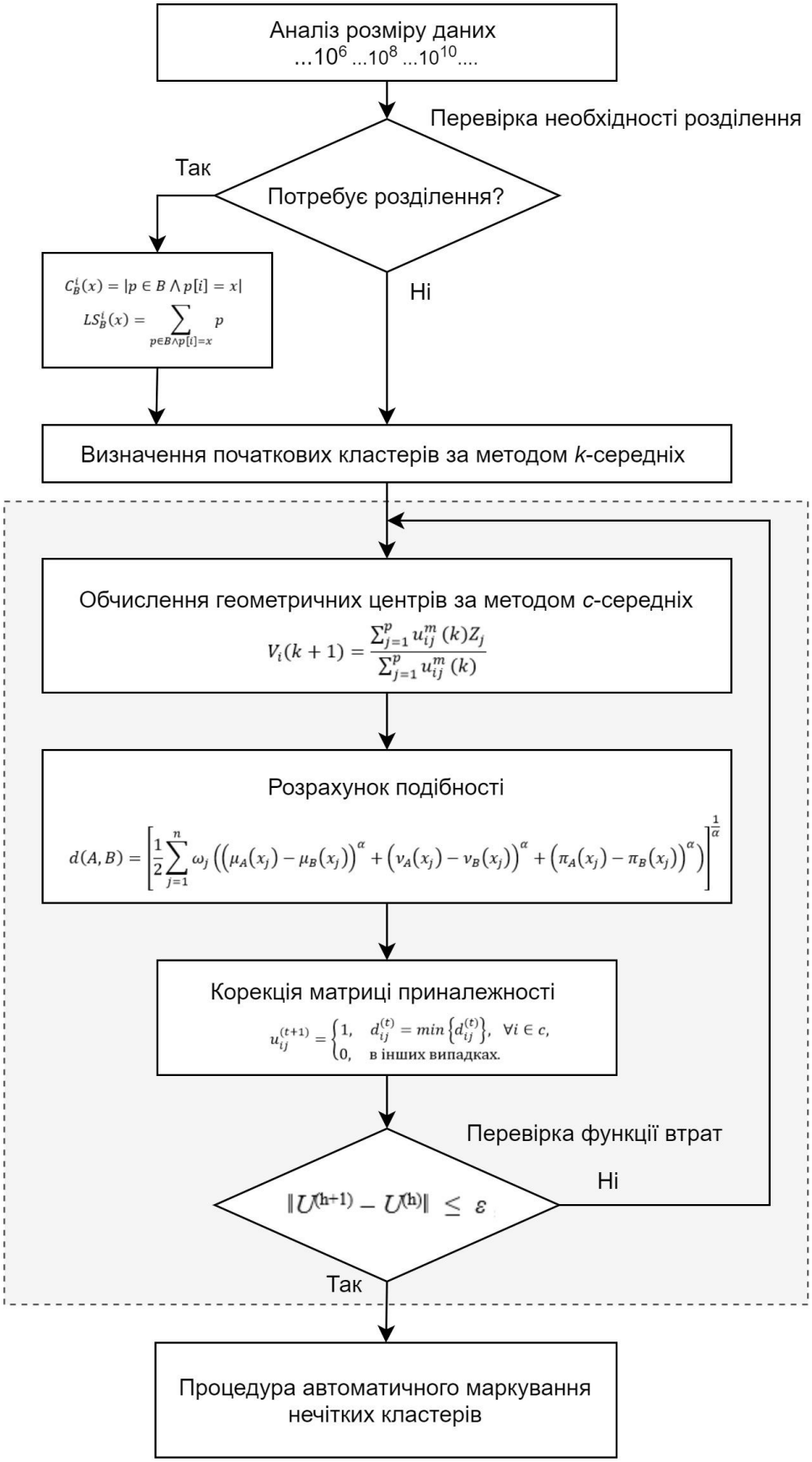


Рисунок 3.1 – Основні етапи вдосконаленого методу нечіткої кластеризації *c*-середніх на випадок великих даних

У роботі [15] показано, що розділення, можна знайти за допомогою лінійного сканування цих карт / масивів.

На етапі 2 застосовується класичний жорсткий метод кластеризації для швидкого визначення центру кластера вхідних даних: на цьому етапі визначається кількість кластерів ($1 < c < n$), та обираються початкові центри кластерів $V(0)=\{V_1, V_2,\dots, V_c\}$.

Отримані центри жорсткого кластера використовується як початковий центр кластера для ітерацій з використанням нечіткої кластеризації. Такий підхід дозволяє прискорити конвергенцію алгоритму.

Основні кроки:

1) Визначити кількість кластерів ($1 < c < n$), вибрати константу ε ($\varepsilon > 0$), встановити кількість ітерацій $t = 0$
2) Обрати початкові центри кластерів $V(0)=\{V_1, V_2,\dots, V_c\}$
3) Розрахувати початкову відстань d між точками даних та кластерами за ф. (3.16)
4) Створити матрицю належності U(0) за ф. (3.23)

$$u_{ij}^{(0)} = \begin{cases} 1, & d_{ij}^{(t)} = min\left\{d_{ij}^{(t)}\right\}, \ \forall i \in c, \\ 0, & \text{в інших випадках.} \end{cases} \qquad (3.23)$$

5) Відкоригувати центр кластера V(1) відповідно до U(0) за ф. (3.22). Перейти до етапу 2.

Етап 3 запускає швидкий алгоритм кластеризації, що містить наступні кроки.

1) Розрахувати початкову відстань d між точками даних та відкоригованими центрами кластерів за ф. (3.16)
2) Для t-ї ітерації обчислити матрицю належності U(t) відповідно до V(t) за ф. (3.21)

3) Розрахувати функцію втрат.

Якщо $\|U^{(t+1)} - U^{(t)}\| \leq \varepsilon$, закінчити обчислення. Вивести центри кластера та матрицю належності *U*.

Якщо $\|U^{(t+1)} - U^{(t)}\| > \varepsilon$, встановити $t = t + 1$ і перейти до кроку (4).

4) Відкоригувати центр кластера V(t+1) відповідно до U(t) за ф. (3.22) і повернутися до кроку (1)

Обчислена матриця належності може визначати належність кожної точки даних до кожного кластеру. Серед них для однієї точки даних ступінь належності відображає кластер, до якого належить точка даних.

Центр жорсткого кластера визначається за допомогою традиційного методу С-кластеризації від кроку (1) до кроку (4). Оскільки центр жорсткого кластера дуже близький до центру нечіткого кластера, центр жорсткого кластера, який отриманий на кроці (5), буде вважатися початковим центром кластера методу кластеризації FCM. Тоді вибірками даних керуватиме кластеризація FCM через наступні кроки, щоб швидкість зближення алгоритму нечіткої кластеризації могла бути прискорена. Кількість ітерацій алгоритму нечіткої кластеризації буде зменшено, а час аналізу кластеризації значно скоротиться. Для моніторингу водного середовища дані про якість води можна швидко згрупувати за допомогою цього підходу, який дозволяє оцінити вимірювані параметри стану води.

3.4.2 Процедура автоматичного маркування нечітких кластерів

Для того, щоб застосувати знайдений прототип кластера у якості класифікаторів, їм потрібно провести маркування. Метод маркування для нечіткого набору полягає в тому, щоб перевірити прототипи кластера та відповідні їм значення належності різних атрибутів та призначити мітку вручну. Отже, виникає проблема швидкого автоматичного маркування.

Процедура маркування для евристичної кластеризації запропонована в [27]. Головною відмінністю даної роботи є узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів кластеризації у випадку інтуїтивістських нечітких даних.

Нехай $\{IA^1_{(\alpha,\beta)},...,IA^n_{(\alpha,\beta)}\}$ - набір інтуїтивістських нечітких множин для яких $\alpha \in (0,1]$ і $\beta \in [0,1)$. Тоді, точка $\tau^l_e \in IA^l_{\alpha,\beta}$, для якої

$$\tau^l_e = \arg\max_{x_i} \mu_{li},\ \forall x_i \in IA^l_{\alpha,\beta} \tag{5}$$

є типовою точкою інтуїтивістського нечіткого кластеру $IA^l_{(\alpha,\beta)}$, де e – індекс типової точки.

Для набору типових точок $\{\tau^1,\ldots,\tau^c\}$ та набору міток $\{label\ 1,\ldots,label\ c\}$ - це вхід для процедури. Припустивши, що результати, отримані за допомогою евристичних алгоритмів кластеризації, є стабільними можливо визначити наступну процедуру маркування.

Для кожної типової точки τ^l, $l=\overline{1,c}$ та кожної мітки $label\ m$, $m=\overline{1,c}$: виконати наступні операції

Алгоритм автоматичного маркування нечітких кластерів

$l \leftarrow 1, m \leftarrow 1$

1. Перевірити наступну умову
   if τ^l відповідає $label\ m$
   then помітити $label\ m$ як мітку типової точки τ^l і перейти до кроку 3
   else $m := m+1$ і перейти до кроку 2
2. Перевірити наступну умову
   if типова точка τ^l помічена
   then $l = l+1$ і перейти до кроку 3
   else перейти на крок 4
3. Перевірити наступну умову
   if всі типові точки τ^l, $l=\overline{1,c}$ помічені

then stop
else перейти до кроку 2
4. Перевірити максимальні входження кластерів
if τ^l відповідає *label* m_i і *label* m_j, і *label* $m_i \neq$ *label* m_j
then призначити τ^l відповідає max *label m*
else перейти до кроку 2

В результаті виконання процедури, мітка може бути призначена всім об'єктам, які відповідають кластеру або є на межі двох кластерів.

3.5 Аналіз ефективності методу нечіткої кластеризації С-середніх для оперативної оцінки якості вод водойм рибогосподарського призначення

Для того, щоб перевірити ефективність пропонованого методу обробки великих даних та його застосування до обробки даних моніторингу водного середовища, в дисертаційному дослідженні було використано дані, зібрані IoT системою SmartWater [19, 21], розробленою на кафедрі комп'ютерних наук та інженерії СНУ ім. В. Даля, з якої було обрано 2026 наборів вибіркових даних для імітаційного тесту. Кожен набір даних має шість характерних параметрів: розчинений кисень (РК), водневий показник (pH), БСК, аміак, NH_3-N, та ХСК; одиниця вимірювання – параметрів РК, БСК, аміак, NH_3-N, та ХСК – мг/л.

3.5.1 Вибір основних параметрів

Відповідно до діючих нормативів якості вод водойм рибогосподарського призначення, якість води може бути представлена у такі способи:

(1) Україна: 1 норматив, що орієнтується на ГДК [4]

(2) ЄС: 2 типи нормативів – G (обов'язкові нормативи) та I (бажані нормативи) [10-12].

Фрагмент нормативів надано у табл. 3.1, повний перелік представлено у Додатку Б.

Таблиця 3.1 – Фрагмент нормативів якості вод рибогосподарського призначення (Директиви 2005/44/EU, 76/464/EU, 78/659/EU)

Параметр	Лососеві		Короп та інші види	
	I	II	I	II
pH	6,5-8,5	6,5-8,6	6,0-9,0	6,0-9,0
Розчинений кисень (РК) (влітку), мгO_2/дм^3	> 9	> 9	> 8	> 6
БСК, мгO_2/дм^3	≤ 3	≤ 3	≤ 6	≤ 6
Азот загальний, мгN/дм^3	≤ 0,04	≤ 1	≤ 0,2	≤ 1
Азот амонійний, мгN/дм^3	≤ 0,005	≤ 0,025	≤ 0,005	≤ 0,025

Ґрунтуючись на 2 типах нормативів та даних [17], якість вод було для поточного завдання поділено на три умовних класи: I – найкращій (бажані нормативи), II - обов‘язкові нормативи, III – невідповідність нормативам.

Таблиця 3.2 – Фрагмент нормативів якості вод рибогосподарського призначення за [17]

Параметр	I	II	III
pH	6,5-8,4	4,8-6,5 8,4-9,0	0-4,8 > 9,0
Розчинений кисень (РК) (влітку), мгO_2/дм^3	> 8	6-8	0-6
БСК, мгO_2/дм^3	0-3	4-6	> 6
Аміак NH_3-N	0-0,1	0,1-0,3	0,3-2,7
ХСК	1-10	10-25	> 25
Азот загальний, мгN/дм^3	≤ 0,04	≤ 1	≤ 0,2
Азот амонійний, мгN/дм^3	≤ 0,005	≤ 0,025	> 0,025

Для подальших розрахунків вводиться припущення, що зібрані зразки води відповідають рівномірному розподілу. Для перевірки ефективності пропонованого методу проводився аналіз зразків на основі відомих методів

кластеризації [k-means, FCM, FFCM] і запропонованого методу кластеризації, та обчислювалася продуктивність алгоритмів.

3.5.2 Вихідні дані для розрахунку

Для розрахунку було використано нормований набір даних, що пройшов етап фазифікації, представлений у Розділі 2. Параметри ступеня належності μ_1 та не належності ν_1 представлені у табл. 3.3, відповідні значення ступеня невизначеності π_l – у табл. 3.4.

Таблиця 3.3 - Вихідні дані для розрахунку

Точка контролю	РК		pH		БСК		Аміак		NH_3-N		ХСК	
	μ_1	ν_1	μ_2	ν_2	μ_3	ν_3	μ_4	ν_4	μ_5	ν_5	μ_6	ν_6
P1	0.8	0.1	0.7	0.2	0.5	0.4	0.5	0.2	0.7	0.2	0.5	0.3
P2	0.5	0.3	0.5	0.2	0.7	0.2	0.1	0.8	0.3	0.6	0.7	0.2
P3	0.4	0.2	0.6	0.1	0.8	0.1	0.2	0.6	0.3	0.6	0.7	0.2
P4	0.6	0.4	0.9	0.1	0.8	0.1	0.7	0.2	0.1	0.7	0.2	0.8
P5	0.7	0.1	0.7	0.2	0.7	0	0.4	0.1	0.8	0.2	0.4	0.6
P6	0.9	0	0.4	0.5	0.5	0.4	0.7	0.1	0.5	0.4	0.4	0.6
P7	0.6	0.4	0.6	0.2	0.7	0.2	0.3	0.6	0.3	0.7	0.6	0.1
P8	0.8	0.1	0.7	0.2	0.4	0.5	0.5	0.4	0.4	0.5	0.8	0.1
P9	0.6	0.4	1	0	0.9	0.1	0.6	0.2	0.2	0.7	0.1	0.8
P10	0.9	0.1	0.8	0	0.6	0.3	0.5	0.2	0.8	0.1	0.6	0.4

Таблиця 3.4 - Відповідні значення ступеня невизначеності x_j до P (π)

Точка контролю	РК	pH	БСК	Аміак	NH_3-N	ХСК
	π_1	π_2	π_3	π_4	π_5	π_6
P1	0.1	0.1	0.1	0.3	0.1	0.2
P2	0.2	0.3	0.1	0.1	0.1	0.1
P3	0.4	0.3	0.1	0.2	0.1	0.1
P4	0	0	0.1	0.1	0.2	0
P5	0.2	0.1	0.3	0.5	0	0
P6	0.1	0.1	0.1	0.2	0.1	0
P7	0	0.2	0.1	0.1	0	0.3
P8	0.1	0.1	0.1	0.1	0.1	0.1
P9	0	0	0	0.2	0.1	0.1
P10	0	0.2	0.1	0.3	0.1	0

Ваговий вектор для обраних параметрів: w = (0.3, 0.1, 0.2, 0.15, 0.15, 0.1).

3.5.3 Реалізація методу кластеризації

Попередньо, кожний потенційний центр кластеру з шести параметрів якості води аналізується за допомогою методу кластеризації к-середніх та методу кластеризації FCM відповідно. Отримані імітаційні центри кластерів двох алгоритмів наведені в табл. 3.5.

Таблиця 3.5 – Порівняння кластерів, отриманих безпосередньо з наборів даних та кластерів, отриманих в результаті першого етапу кластеризації FCM

P9	0.60	0.40	1.00	0.00	0.90	0.10	0.60	0.20	0.20	0.70	0.10	0.80
$P9_{FCM}$	0.60	0.38	0.91	0.07	0.84	0.11	0.61	0.23	0.19	0.68	0.20	0.74
P10	0.90	0.10	0.80	0.00	0.60	0.30	0.50	0.20	0.80	0.10	0.60	0.40
$P10_{FCM}$	0.81	0.11	0.71	0.13	0.60	0.26	0.50	0.22	0.71	0.21	0.54	0.41
P8	0.80	0.10	0.70	0.20	0.40	0.50	0.50	0.40	0.40	0.50	0.80	0.10
$P8_{FCM}$	0.69	0.17	0.63	0.21	0.54	0.35	0.40	0.47	0.39	0.53	0.70	0.18

З табл. 3.5 видно, що центр кластерів HCM знаходиться в значній близькості до центру кластерів FCM. Отже, звичайні центри кластерів можуть використовуватися як початкові центри алгоритму кластеризації FCM.

Далі використано FCM, що містить наступні кроки.

Крок 1: Визначити кількість кластерів ($1 < c < n$), вибрати константу ε та створити матрицю належності U(0) за ф. (3.21).

У загальному випадку, для визначення кількості кластерів можливо використання набору критеріїв: дисперсія, ентропія та кількість помилок. Для досліджуваного набору даних, результати аналізу (рис. 3.2) показали, що найкращий результат дає розподіл набору даних на 5 кластерів, разом з тим, 3 кластери також задовольняють умовам поставленої задачі. Отже, згідно отриманого графіку та попередніх припущень щодо розподілу даних якості вод

водойм рибогосподарського призначення, визначаємо, що в наборі даних є три класи, c = 3.

Рисунок 3.2 – Результати попереднього підрахунку кількості кластерів для набору даних якості вод водойм рибогосподарського призначення

Перша ітерація.

Крок 1. Нехай ε = 0,005. Випадковим чином обираємо три початкові геометричні центри кластерів V(0):

$$V(0) = \begin{bmatrix} P_9 \\ P_{10} \\ P_8 \end{bmatrix}$$

$$V(0) = \begin{vmatrix} <0.6,\ 0.4> & <1.0,\ 0.0> & <0.9,\ 0.1> & <0.6,\ 0.2> & <0.2,\ 0.7> & <0.1,\ 0.8> \\ <0.9,\ 0.1> & <0.8,\ 0.0> & <0.6,\ 0.3> & <0.5,\ 0.2> & <0.8,\ 0.1> & <0.6,\ 0.4> \\ <0.8,\ 0.1> & <0.7,\ 0.2> & <0.4,\ 0.5> & <0.5,\ 0.4> & <0.4,\ 0.5> & <0.8,\ 0.1> \end{vmatrix}$$

Тоді, за розрахунками відстані між відповідними точками і обраними центрами кластерів по ф. (3.16) отримуємо:

$$d = \begin{vmatrix} 0.860 & 0.300 & 0.458 \\ 0.964 & 0.871 & 0.616 \\ 0.889 & 0.831 & 0.640 \\ 0.224 & 0.889 & 0.894 \\ 0.812 & 0.469 & 0.818 \\ 0.860 & 0.624 & 0.632 \\ 0.825 & 0.812 & 0.529 \\ 0.954 & 0.592 & 0 \\ 0 & 0.883 & 0.954 \\ 0.883 & 0 & 0.592 \end{vmatrix}$$

Матриця належності U(0)

$$U(0) = \begin{vmatrix} 0.078 & 0.645 & 0.276 \\ 0.214 & 0.262 & 0.524 \\ 0.245 & 0.281 & 0.473 \\ 0.888 & 0.056 & 0.055 \\ 0.200 & 0.602 & 0.198 \\ 0.211 & 0.400 & 0.390 \\ 0.224 & 0.231 & 0.545 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{vmatrix}$$

Крок 2: За ф. (3.22) оновлюємо геометричні центри кластерів $V=\{V_1, V_2, V_3\}$

$$V(1)= \begin{vmatrix} <0.601,\ 0.376> <0.908,\ 0.067> <0.835,\ 0.110> <0.606,\ 0.231> <0.189,\ 0.678> <0.196,\ 0.742> \\ <0.807,\ 0.110> <0.712,\ 0.125> <0.603,\ 0.263> <0.469,\ 0.219> <0.710,\ 0.209> <0.538,\ 0.409> \\ <0.693,\ 0.174> <0.626,\ 0.211> <0.544,\ 0.352> <0.399,\ 0.468> <0.387,\ 0.529> <0.698,\ 0.179> \end{vmatrix}$$

Крок 3: Розрахунок відстані між точками даних та кластерами

Оскільки $\|U^{(h+1)} - U^{(h)}\| > \varepsilon$, обчислення продовжуються.

Друга ітерація. Матриця відстаней та відповідна відкоригована матриця належності для другої ітерації

$$d = \begin{vmatrix} 1.078 & 0.224 & 0.607 \\ 1.171 & 1.029 & 0.574 \\ 1.018 & 0.928 & 0.547 \\ 0.157 & 1.078 & 1.088 \\ 0.949 & 0.423 & 0.917 \\ 1.065 & 0.701 & 0.857 \\ 0.978 & 0.916 & 0.434 \\ 1.188 & 0.694 & 0.312 \\ 0.179 & 1.097 & 1.164 \\ 1.128 & 0.242 & 0.790 \end{vmatrix} \quad \mathbf{U(1)=} \begin{vmatrix} 0.037 & 0.848 & 0.115 \\ 0.155 & 0.201 & 0.644 \\ 0.176 & 0.212 & 0.611 \\ 0.960 & 0.020 & 0.020 \\ 0.141 & 0.708 & 0.151 \\ 0.206 & 0.475 & 0.319 \\ 0.138 & 0.158 & 0.704 \\ 0.054 & 0.159 & 0.787 \\ 0.952 & 0.025 & 0.022 \\ 0.040 & 0.878 & 0.082 \end{vmatrix}$$

Оновлені геометричні центри кластерів $V=\{V_1, V_2, V_3\}$

$$V(2)= \begin{vmatrix} <0.604,\ 0.383> <0.920,\ 0.066> <0.836,\ 0.109> <0.631,\ 0.215> <0.171,\ 0.684> <0.180,\ 0.768> \\ <0.806,\ 0.099> <0.698,\ 0.161> <0.586,\ 0.272> <0.483,\ 0.193> <0.713,\ 0.210> <0.513,\ 0.418> \\ <0.622,\ 0.226> <0.603,\ 0.196> <0.616,\ 0.282> <0.325,\ 0.546> <0.350,\ 0.575> <0.686,\ 0.171> \end{vmatrix}$$

Розрахована функція втрат для другої ітерації складає $0.05951 > 0.005$

Третя ітерація. Матриця відстаней та відповідна відкоригована матриця належності для третьої ітерації

$$d = \begin{vmatrix} 1.112 & 0.200 & 0.741 \\ 1.217 & 1.055 & 0.404 \\ 1.064 & 0.956 & 0.385 \\ 0.121 & 1.077 & 1.079 \\ 0.980 & 0.411 & 0.965 \\ 1.084 & 0.655 & 0.973 \\ 1.023 & 0.938 & 0.280 \\ 1.224 & 0.705 & 0.479 \\ 0.158 & 1.101 & 1.145 \\ 1.160 & 0.272 & 0.900 \end{vmatrix} \quad \mathbf{U(2)=} \begin{vmatrix} 0.029 & 0.905 & 0.066 \\ 0.088 & 0.117 & 0.795 \\ 0.101 & 0.126 & 0.773 \\ 0.975 & 0.012 & 0.012 \\ 0.129 & 0.737 & 0.134 \\ 0.201 & 0.550 & 0.249 \\ 0.064 & 0.076 & 0.859 \\ 0.095 & 0.286 & 0.619 \\ 0.962 & 0.020 & 0.018 \\ 0.048 & 0.872 & 0.080 \end{vmatrix}$$

Оновлені геометричні центри кластерів $V=\{V_1, V_2, V_3\}$

$$V(3)=\begin{vmatrix} <0.607,\ 0.386> & <0.930,\ 0.063> & <0.837,\ 0.108> & <0.643,\ 0.203> & <0.166,\ 0.687> & <0.167,\ 0.783> \\ <0.816,\ 0.090> & <0.692,\ 0.175> & <0.573,\ 0.284> & <0.498,\ 0.180> & <0.712,\ 0.209> & <0.509,\ 0.422> \\ <0.566,\ 0.264> & <0.586,\ 0.182> & <0.672,\ 0.227> & <0.267,\ 0.602> & <0.326,\ 0.604> & <0.675,\ 0.168> \end{vmatrix}$$

Розрахована функція втрат для третьої ітерації складає 0.038389 > 0.005, отже обчислення продовжуються.

Четверта ітерація. Матриця відстаней та відповідна відкоригована матриця належності для четвертої ітерації

$$d = \begin{vmatrix} 1.130 & 0.188 & 0.850 \\ 1.245 & 1.075 & 0.290 \\ 1.092 & 0.979 & 0.279 \\ 0.114 & 1.082 & 1.092 \\ 0.995 & 0.423 & 1.015 \\ 1.096 & 0.630 & 1.072 \\ 1.050 & 0.955 & 0.206 \\ 1.245 & 0.704 & 0.613 \\ 0.1455 & 1.110 & 1.149 \\ 1.176 & 0.282 & 0.991 \end{vmatrix} \quad \mathbf{U(3)=} \begin{vmatrix} 0.026 & 0.929 & 0.045 \\ 0.048 & 0.065 & 0.887 \\ 0.057 & 0.071 & 0.872 \\ 0.978 & 0.011 & 0.011 \\ 0.133 & 0.738 & 0.128 \\ 0.197 & 0.597 & 0.206 \\ 0.036 & 0.043 & 0.921 \\ 0.121 & 0.379 & 0.500 \\ 0.968 & 0.017 & 0.015 \\ 0.050 & 0.878 & 0.071 \end{vmatrix}$$

Оновлені геометричні центри кластерів $V=\{V_1, V_2, V_3\}$

$$V(4)=\begin{vmatrix} <0.608,\ 0.386> & <0.933,\ 0.062> & <0.837,\ 0.108> & <0.646,\ 0.200> & <0.166,\ 0.687> & <0.165,\ 0.786> \\ <0.820,\ 0.087> & <0.688,\ 0.182> & <0.565,\ 0.294> & <0.505,\ 0.179> & <0.705,\ 0.215> & <0.512,\ 0.418> \end{vmatrix}$$

<0.539, 0.278> <0.578, 0.176> <0.697, 0.202> <0.240, 0.628> <0.316, 0.615> <0.671, 0.169>

Розрахована функція втрат для четвертої ітерації складає **0.01709** > 0.005, отже обчислення продовжуються.

П’ята ітерація. Матриця відстаней та відповідна відкоригована матриця належності для п’ятої ітерації

$$d = \begin{vmatrix} 1.133 & 0.176 & 0.900 \\ 1.251 & 1.075 & 0.245 \\ 1.098 & 0.981 & 0.240 \\ 0.114 & 1.083 & 1.104 \\ 0.997 & 0.438 & 1.041 \\ 1.098 & 0.616 & 1.119 \\ 1.056 & 0.953 & 0.203 \\ 1.250 & 0.691 & 0.673 \\ 0.143 & 1.114 & 1.156 \\ 1.178 & 0.290 & 1.034 \end{vmatrix} \quad \mathbf{U(4)=} \begin{vmatrix} 0.023 & 0.941 & 0.036 \\ 0.035 & 0.048 & 0.917 \\ 0.043 & 0.054 & 0.903 \\ 0.979 & 0.011 & 0.010 \\ 0.141 & 0.729 & 0.129 \\ 0.194 & 0.618 & 0.187 \\ 0.034 & 0.042 & 0.924 \\ 0.129 & 0.424 & 0.446 \\ 0.969 & 0.016 & 0.015 \\ 0.053 & 0.877 & 0.069 \end{vmatrix}$$

Оновлені геометричні центри кластерів $V=\{V_1, V_2, V_3\}$

$$V(5)= \begin{vmatrix} <0.609, 0.386> <0.933, 0.062> <0.837, 0.109> <0.646, 0.200> <0.166, 0.686> <0.165, 0.786> \\ <0.822, 0.087> <0.686, 0.185> <0.561, 0.301> <0.508, 0.181> <0.699, 0.220> <0.515, 0.414> \\ <0.530, 0.281> <0.575, 0.174> <0.705, 0.194> <0.230, 0.636> <0.314, 0.618> <0.670, 0.170> \end{vmatrix}$$

Розрахована функція втрат для п’ятої ітерації складає 0.006923 > 0.005, отже обчислення продовжуються.

Шоста ітерація. Матриця відстаней та відповідна відкоригована матриця належності для шостої ітерації

$$d = \begin{vmatrix} 1.132 & 0.169 & 0.916 \\ 1.252 & 1.070 & 0.232 \\ 1.098 & 0.978 & 0.228 \\ 0.114 & 1.082 & 1.109 \\ 0.996 & 0.449 & 1.049 \\ 1.097 & 0.609 & 1.134 \\ 1.056 & 0.948 & 0.208 \\ 1.249 & 0.680 & 0.693 \\ 0.143 & 1.115 & 1.160 \\ 1.177 & 0.296 & 1.048 \end{vmatrix} \quad \mathbf{U(5)=} \begin{vmatrix} 0.021 & 0.946 & 0.032 \\ 0.032 & 0.043 & 0.925 \\ 0.039 & 0.049 & 0.911 \\ 0.979 & 0.011 & 0.010 \\ 0.146 & 0.721 & 0.132 \\ 0.193 & 0.626 & 0.181 \\ 0.036 & 0.044 & 0.920 \\ 0.131 & 0.443 & 0.426 \\ 0.969 & 0.016 & 0.015 \\ 0.055 & 0.874 & 0.070 \end{vmatrix}$$

Оновлені геометричні центри кластерів $V=\{V_1, V_2, V_3\}$

$$V(6)= \begin{vmatrix} <0.609,\ 0.386> <0.933,\ 0.062> <0.837,\ 0.108> <0.646,\ 0.200> <0.167,\ 0.686> <0.165,\ 0.786> \\ <0.822,\ 0.086> <0.685,\ 0.187> <0.558,\ 0.304> <0.509,\ 0.182> <0.696,\ 0.223> <0.516,\ 0.412> \\ <0.527,\ 0.282> <0.574,\ 0.173> <0.708,\ 0.191> <0.227,\ 0.639> <0.313,\ 0.618> <0.670,\ 0.170> \end{vmatrix}$$

Розрахована функція втрат для шостої ітерації складає D(K5) = 0.003,

Оскільки 0.003 <0.005, обчислення завершуються.

Результуюча матриця належності U.

$$\mathbf{U} = \begin{array}{c} \begin{matrix} C1 & C2 & C3 \end{matrix} \\ \begin{vmatrix} 0.021 & 0.946 & 0.032 \\ 0.032 & 0.043 & 0.925 \\ 0.039 & 0.049 & 0.911 \\ 0.979 & 0.011 & 0.010 \\ 0.146 & 0.721 & 0.132 \\ 0.193 & 0.626 & 0.181 \\ 0.036 & 0.044 & 0.920 \\ 0.131 & 0.443 & 0.426 \\ 0.969 & 0.016 & 0.015 \\ 0.055 & 0.874 & 0.070 \end{vmatrix} \end{array}$$

Для даного випадку, припустивши, що $u_{ij} = 0.6 \Rightarrow A_j \in C_i$ $(1 \leq j \leq 10, 1 \leq i \leq 3)$, де C_i позначає кластер i, отримано наступний розподіл кластерів (табл. 3.6).

Таблиця 3.6 – Результати кластеризації

Точки контролю	Кластер
P4, P9	1
P1, P5, P6, P10	2
P2, P3, P7	3
P8	Не визначено

Як видно з табл.3.5, для точки P8 значення не визначено, що потребує додаткової перевірки.

3.5.4 Процедура автоматичного маркування нечітких кластерів для якості вод водойм рибогосподарського призначення

На цьому етапі, для кожної точки контролю виконується додаткова перевірка належності класу. Для даного випадку, встановлено наступну приналежність результатів замірів у відповідних точках (табл. 3.7).

Таблиця 3.7 – Розподіл кластерів для вхідного набору

Точка контролю	РК		pH		БСК		Аміак		NH_3-N		ХСК		Клас
P1	0.8	0.1	0.7	0.2	0.5	0.4	0.5	0.2	0.7	0.2	0.5	0.3	II
P2	0.5	0.3	0.5	0.2	0.7	0.2	0.1	0.8	0.3	0.6	0.7	0.2	III
P3	0.4	0.2	0.6	0.1	0.8	0.1	0.2	0.6	0.3	0.6	0.7	0.2	III
P4	0.6	0.4	0.9	0.1	0.8	0.1	0.7	0.2	0.1	0.7	0.2	0.8	I
P5	0.7	0.1	0.7	0.2	0.7	0	0.4	0.1	0.8	0.2	0.4	0.6	II
P6	0.9	0	0.4	0.5	0.5	0.4	0.7	0.1	0.5	0.4	0.4	0.6	II
P7	0.6	0.4	0.6	0.2	0.7	0.2	0.3	0.6	0.3	0.7	0.6	0.1	III
P8	0.8	0.1	0.7	0.2	0.4	0.5	0.5	0.4	0.4	0.5	0.8	0.1	II-III
P9	0.6	0.4	1	0	0.9	0.1	0.6	0.2	0.2	0.7	0.1	0.8	I
P10	0.9	0.1	0.8	0	0.6	0.3	0.5	0.2	0.8	0.1	0.6	0.4	II

Згідно встановленої процедури, з трьох кластерів <C_1, C_2, C_3> виключається Сі з найменшими значеннями і встановлюється приналежність до найнижчого класу, тобто у даному випадку Р8 віднесено до класу III.

3.5.5 Аналіз збіжності алгоритмів на наборі якості вод

Для аналізу ефективності, проведено аналіз збіжності на вибраних вибіркових даних. Порівняння проводилося за 3 відомими алгоритмами кластеризації к-середніх, FCM та FFCM. Зближення та ітераційний процес для двох алгоритмів показані на рис. 3.3.

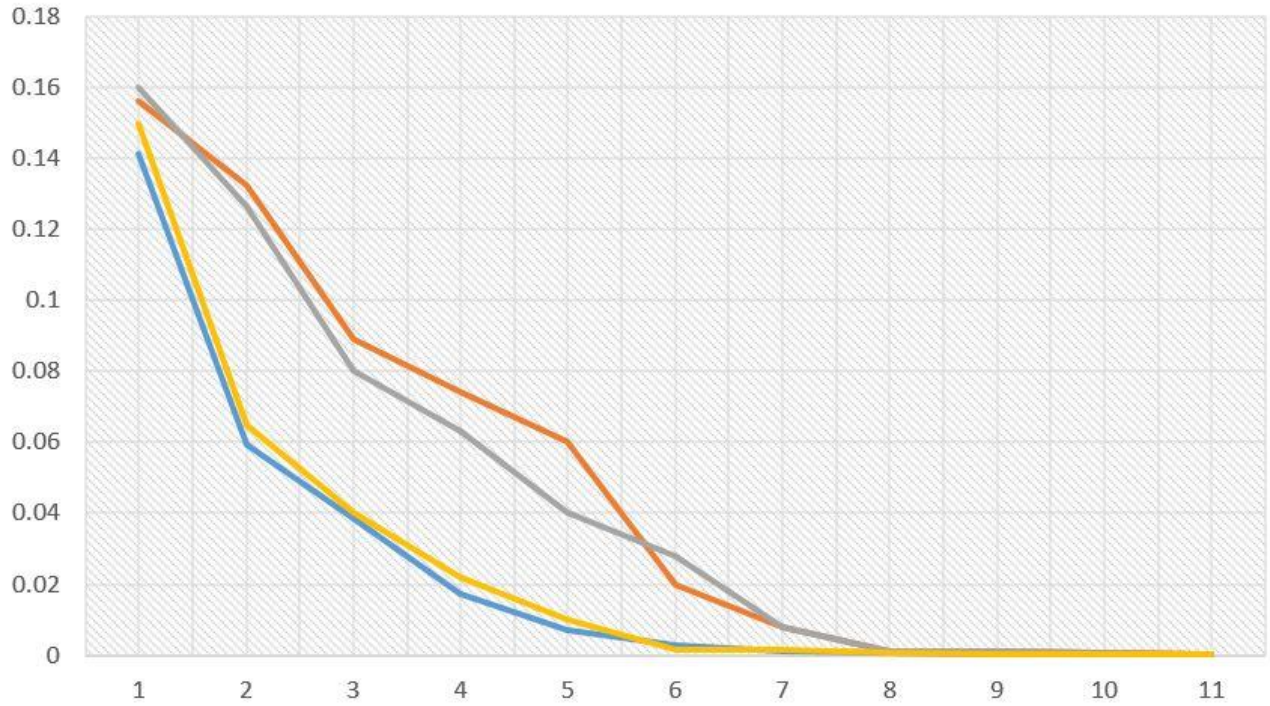


Рисунок 3.3 – Зближення цільової функції в кластеризації в залежності від кількості ітерацій (синій – пропонований метод, сірий – метод нечіткої кластеризації с-середніх (FCM), жовтий – швидкий метод нечіткої кластеризації FFCM, жовтогарячий – метод k-середніх).

Як видно з рис. 3.3, швидкість конвергенції запропонованого методу кластеризації швидша, ніж за традиційними методами FCM, що підтверджує попередні дослідження [25] і порівняний з відомим методом кластеризації FFCM. Це показує, що пропонований метод кластеризації може ефективно скоротити необхідний час кластеризації та підвищити ефективність обробки даних.

Висновки до розділу 3

У розділі запропоновано удосконалений метод кластеризації швидких нечітких С-середніх для оброблення великих даних моніторингу водних об’єктів. Показано, що новий метод дозволяє виконувати аналіз великих даних за рахунок нової процедури оцінювання обсягу та введення розділення об’єктів за технологією MapReduce, проводити автоматичну розмітку за рахунок додаткової процедури розмітки і покращує конвергенцію алгоритму кластеризації.

Використання жорсткої кластеризації на перших етапах роботи дозволяє тримати початкове значення нечіткої кластеризації, і тим самим прискорити швидкість конвергенції та підвищити ефективність обробки великих даних.

Результати моделювання показують, що порівняно з кластеризацією k-means та стандартним алгоритмом кластеризації FCM, цей алгоритм може не тільки ефективно реалізувати нечітку кластеризацію даних, але й мати більш швидку конвергенцію. Швидкість конвергенції порівняна з відомим методом кластеризації FFCM, але останній не розрахований на роботу з великими даними.

Метод може бути застосований як до обробки великих даних, отриманих з IoT систем моніторингу води, але й у інших областях при обробці великих даних. Разом з тим, варто відзначити, що для інших застосувань процедура автоматичного маркування класів має бути переглянута, щоб в найбільшій мірі відповідати розподілу.

Результати розділу опубліковано в роботах автора [1-3, 20, 21].

Список літератури до розділу 3

1. Барбарук Л.В., Зубарев Д.В., Медведєв Є.М., Барбарук В.М. Використання нечітких моделей для планування збиральних робіт у різних кліматичних умовах, *Вісник Східноукраїнського національного університету ім. В. Даля*, №6 (247), С. 175-179, 2018.

2. Барбарук Л.В., Суворін О.В. Система аналізу даних та підтримки прийняття рішень з інвентаризації промислових відходів, *Вісник Східноукраїнського національного університету ім. В. Даля*, №8 (238). С. 5-12. 2017.

3. Барбарук Л.В., Михайличенко С.О. Бездротова система віддаленого моніторингу сільськогосподарських параметрів. *Матеріали VII Всеукр. науково-практ. конф. «Електронні апарати та системи. Проблеми створення. Перспективи розвитку»*, Сєвєродонецьк : Східноукр. нац. ун-т ім. В. Даля, 2017. С. 206-208.

4. Гранично допустимі значення прказників якості води для рибогосподарських водойм. Загальний перелік ГДК і ОБРВ шкідливих речовин для води пибогосподарський водойм : [№ 12-04-11 чинний від 09-08-1990]. 45 с.

5. Рязанцев О.І., Барбарук В.М., Барбарук Л.В. “Комплексна оцінка екологічної небезпеки території промислового виробництва”, *Вісник Східноукраїнського національного університету ім. В. Даля*, №7(154), Ч.2, С. 136-140, 2010.

6. Atanassov K.T. (1986) Intuitionistic fuzzy sets. *Fuzzy Sets and Systems*. vol. 20. pp. 87-96.

7. Atanassov K.. (1999) Intuitionistic fuzzy sets: theory and applications. Heidelberg: Physica-Verlag.

8. Bezdek J.C. Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press, 1981. 272 p.

9. Dunn J. C. (1973). A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters. *Journal of Cybernetics*. Vol. 3 (3), pp. 32–57.

10. European Economic Community (EEC) 1978. Council Directive of 18 July 1978 on the quality of fresh waters needing protection or improvement in order to support fish life (78/659/EEC). Off J, L 222, 14.8.1978, 1-10. Режим доступу: eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=celex:31978l0659: en:html (20.11.2020).

11. European Economic Community (EEC) 1992. Council Directive 92/43/EEC of 21 May 1992 on the conservation of natural habitats and of wild fauna and flora. Off J, L 206, 22.7.1992, 7-50. Режим доступу: eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:31992L0043:en:html (20.11.2020).

12. European Parliament and European Council (EC) 2000. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 establishing a framework for Community action in the field of water policy. Off J, L 327, 22.12.2000, 1-73. Режим доступу: eur-lex.europa.eu/LexUriServ/ LexUriserv.do?uri=celex:32000l0060:en:html (20.11.2020).

13. Gan G., Ma C., Wu J. (2007) Data Clustering: Theory, Algorithms, and Applications, SIAM, Philadelphia, PA, 2007.

14. Li E., Wang L., Song B., Jian, S. (2018). Improved Fuzzy C-Means Clustering for Transformer Fault Diagnosis Using Dissolved Gas Analysis Data. *Energies, 11(9), 2344*. doi:10.3390/en11092344

15. Mazzeo, G. et al. (2017) A fast and accurate algorithm for unsupervised clustering around centroids. *Inf. Sci*. vol. 400, pp. 63-90.

16. Majhi, S.K. Fuzzy clustering algorithm based on modified whale optimization algorithm for automobile insurance fraud detection. Evol. Intel. (2019). https://doi.org/10.1007/s12065-019-00260-3.

17. Raman B.V., Reinier B., Mohan S. (2009). Fuzzy logic Water Quality index and importance of Water Quality Parameters. Air, Soil Water Res., vol. 2, pp. 51–59.

18. Ren Z., Wang C., Fan X., Ye Z. (2018). Fuzzy Clustering Based on Water Wave Optimization. *2018 13th International Conference on Computer Science & Education (ICCSE)*. doi:10.1109/iccse.2018.8468778

19. SmartWater Skarga-Bandurova I., Krytska Y. (2019) IoT based Water Quality Monitoring System. *Internet of Things for Industry and Human Applications. Volume 3. Assessment and Implementation. IoT for Ecology, Safety and Security Monitoring Systems. Section 49.* / Ed. V. S. Kharchenko. – Ministry of Education and Science of Ukraine, National Aerospace University KhAI, 2019, pp. 627-671.

20. Skarga-Bandurova I., Krytska Y., Shorohov M., Suvorin O., Barbaruk L., Ozheredova M. (2019) "Towards Development IoT-based Water Quality Monitoring System", *Proceedings of the 7th IEEE International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pp. 140-145, 2019. doi: 10.1109/FiCloudW.2019.00038.

21. Skarga-Bandurova I., Krytska Y., Velykzhanin A., Barbaruk L., Suvorin O., Shorohov M. (2020) "Emerging Tools for Design and Implementation of Water Quality Monitoring Based on IoT", *Complex Systems Informatics and Modeling Quarterly*. Published online by RTU Press, https://csimq-journals.rtu.lv Article 138, Issue 24, September/October 2020, pp. 1–14, 2020 https://doi.org/10.7250/csimq.2020-24.01.

22. Szmidt, E., & Kacprzyk, J. (2000). Distances between intuitionistic fuzzy sets. *Fuzzy Sets and Systems,* vol. 114, pp. 505–518.

23. Xu Z. (2007). Some similarity measures of intuitionistic fuzzy sets and their applications to multiple attribute decision making. *Fuzzy Optimization and Decision Making,* vol. 6(2), pp. 109–121. doi:10.1007/s10700-007-9004-z

24. Xu Z. S. (2009) Intuitionistic fuzzy hierarchical clustering algorithms. *Journal of Systems Engineering and Electronics,* 2009, vol. 20(1), pp. 90–97.

25. Xu Z., Wu J. (2010) Intuitionistic fuzzy C-means clustering algorithms. *Journal of Systems Engineering and Electronics* Vol. 21, No. 4, pp.580–590.

26. Yang Z.-Y., Huang X.-Y., Du C. H., and Tang M.-X. (2008) Study of urban traffic congestion judgment based on FFCM clustering, *Application Research of Computers*, vol. 25, no. 9, pp. 2768–2770.

27. Viattchenin D.A. Shyrai S., Damaratski A. (2014) Labeling consequents of fuzzy rules constructed by using heuristic algorithms of possibilistic clustering. *Database Systems Journal,* vol. V(3), pp. 3-13.

28. Zadeh, L. A. (1965). Fuzzy Sets. *Information and Control*, vol. 8, pp. 338–353.

29. Zhu L., Chung F.L., Wang S. (2009) Generalized fuzzy C-means clustering algorithm with improved fuzzy partitions. *IEEE Trans Syst Man Cybern B Cybern*. Vol. 39(3), pp. 578-91. doi: 10.1109/TSMCB.2008.2004818.

РОЗДІЛ 4
ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ ОБРОБКИ ВЕЛИКИХ ДАНИХ В ІНФОРМАЦІЙНО-АНАЛІТИЧНИХ СИСТЕМАХ МОНІТОРИНГУ ВОДНИХ ОБ'ЄКТІВ

Четвертий розділ присвячений розробці засобів та інформаційної технології для обробки великих даних отриманих від станцій моніторингу водних об’єктів. Розкрито етапи розробки та реалізації інформаційної технології у вигляді системи підтримки прийняття рішень, надано архітектуру сховища даних, визначено критерії якості видобутку даних, представлено структуру бази даних та основні інструменти для аналітичної обробки даних у вигляді гібридного інструмента інтелектуального аналізу даних та управління непрямими знаннями баз даних спеціалізованої аналітичної системи водних об’єктів. Наведено результати адаптації методів візуалізації великих даних на випадок он-лайн моніторингу водних об’єктів.

4.1 Інформаційна технологія

Система підтримки прийняття рішень (СППР) з управління водними об’єктами може бути визначена як інформаційна аналітична система, що використовується для накопичення екологічних даних діяльності підприємств аквакультури та водного господарства, видобутку знань з даних моніторингу водного середовища, зменшення часу на прийняття рішень та підвищення якості рішень в системах управліннях виробництвом рибної продукції у повністю або частково контрольованих умовах. СППР є класичним рішенням, орієнтованим на вироблення рекомендацій як у сфері охорони навколишнього середовища загалом так і для управління діяльністю об’єктів аквакультури. Базова структура СППР, запропонована у [29, 18] складається з трьох основних компонентів: діалогового

компонента, компонента управління даними і компонента моделювання, як показано на рис. 4.1.

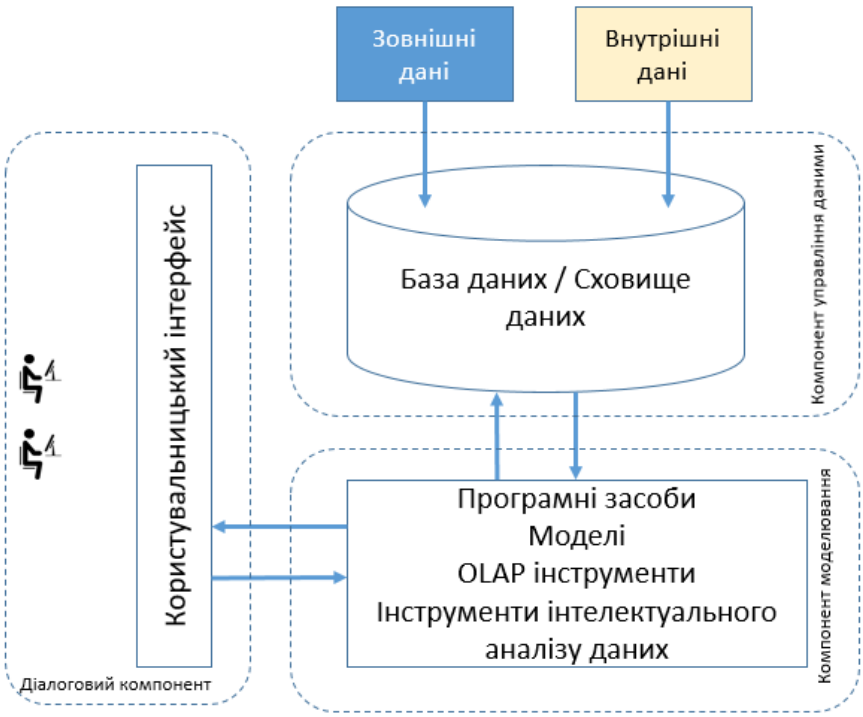


Рисунок 4.1 – Узагальнена структура СППР

Діалоговий компонент забезпечує інтерфейс між системними компонентами та користувачем. Типові можливості діалогового компонента містять обробку різноманітних стилів діалогу для адаптації дій користувача до різноманітних пристроїв введення, представлення даних у різноманітних форматах, надання контекстно-залежної он-лайн допомоги, попереджень про невідповідності параметрів якості води, повідомлень про помилки та дозволяє користувачеві контролювати обробку даних в гнучкій та простій формі.

Компонент управління даними в першу чергу стосується всіх видів діяльності з управління даними. Типові можливості компонента управління даними містять, можливість легко і швидко забезпечувати ведення даних та ведення рибного господарства, збирати / витягувати дані з різних джерел, зображати логічні структури даних в термінах користувача та обробляти особисті

та неофіційні дані, щоб користувачі могли експериментувати з альтернативами, заснованими на власних судженнях.

Компонент моделювання є аналітичною частиною системи. Програмна частина СППР містить програмні засоби, які використовуються для аналізу даних і містить різні інструменти OLAP, засоби аналізу даних, колекцію математичних та аналітичних моделей, доступних для користувачів.

Особливістю СППР є вибірка та застосування накопичених знань, правил і методик, що вочевидь переводить такі системи до рангу інтелектуальних аналітичних систем. Загальна проблема цих систем полягає у ефективності методів отримання знань. Традиційні підходи базуються на отриманні знань за допомогою серії інтерактивних сеансів спільно з експертами. В системах даного напряму більш ефективним є поєднання експертів у галузі інтенсивного рибогосподарства, охорони навколишнього середовища та експертів з організації та проведення технологічних процесів водопідготовки. За наявності сховищ даних, в яких зберігається ретроспективна інформація щодо стану на змін показників навколишнього середовища, існує більш перспективний та наукоємний підхід. Він полягає у застосуванні технологій статистичної обробки даних, видобутку даних та машинного навчання.

Ґрунтуючись на [22] було вирішено, що система буде мати багаторівневу архітектуру з 4 рівнів, яка з'єднує користувача з інформаційно-аналітичною системою:

1. Фільтрація даних
2. Формування рекомендацій та управління знаннями
3. Знаходження знань
4. Управління знаннями

На рис. 4.2 відображено архітектуру системи. Система використовує набір методів, алгоритмів та технологічних рішень для отримання необхідних знань із екологічних та рибопромислових інформаційних систем через наявні бази даних.

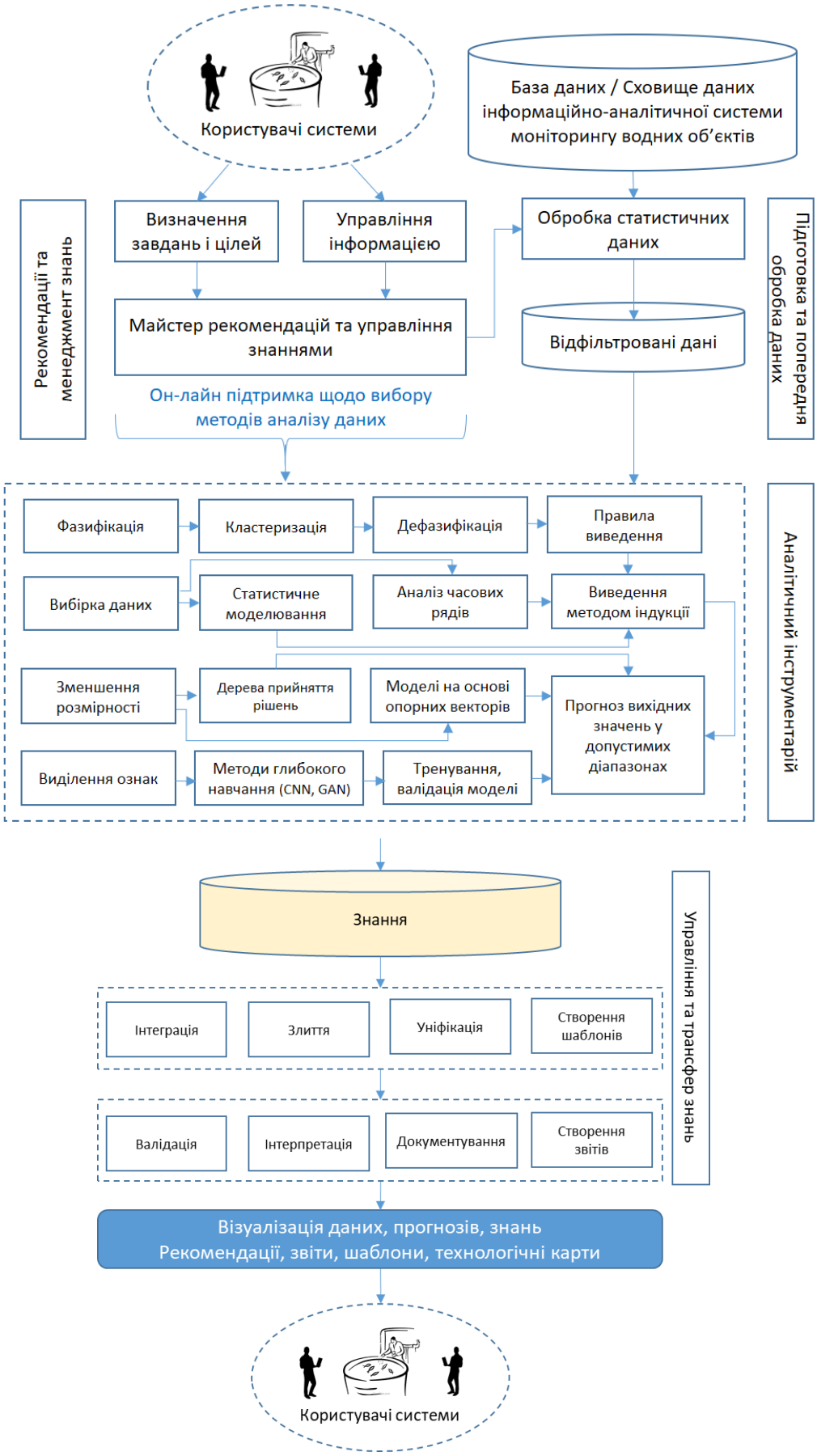


Рисунок 4.2 – Архітектура СППР

СППР є важливою частиною інформаційної технології моніторингу водних ресурсів. Зазвичай дані, зібрані з датчиків, зберігаються в хмарі. На основі аналізу таких даних інструменти ШІ та аналітичної обробки даних здатні надавати рішення або рекомендації щодо управління параметрами якості води, кількості корму для риби, використання ресурсів, оперативного планування тощо. Це допомагає зменшити виробничі витрати для усієї екосистеми, одночасно збільшуючи її продуктивність. Наприклад, фермери, які займаються рибним господарством, можуть вибрати оптимальний час, кількість корму, необхідного для риби, аналізуючи попереднє зібрані дані, що зберігаються в хмарі. Таким чином, можна уникнути забруднення води, що виникає через надмірну кількість корму в рибній екосистемі.

Стратегічним плануванням можна уникнути загибелі риб внаслідок відключення електроенергії, засмічення труб, надмірної кількості небажаного вмісту у воді тощо. Наприклад, кількість агрегатів для фільтрації, які повинні бути присутніми, можна придбати до збирання врожаю, оскільки це необхідне обладнання, яке регулює вміст аміаку та нітратів у резервуарах.

Ці знання можуть бути використані не лише при впровадженні, але й при розширенні розробленої системи. Для проекту використовується база даних MySQL. Bash Script і сценарії Python використовуються для завантаження та очищення даних перед зберіганням їх у базі даних.

Основні завдання, що вирішувалися в роботі представлені в СППР у вигляді трьох компонентів:

- Портал даних: Візуалізує параметри якості води з різних датчиків на кожен день.
- Аналіз параметрів: із заданою періодичністю (щодобово, щомісяця, щорічно, тощо) консолідує параметри якості води та показує порівняння між різними станціями моніторингу по всій ланці, де встановлено відповідні засоби.
- Оперативний аналіз якості води: Якість води розраховується за допомогою різноманітних параметрів, таких як розчинений кисень, БСК, pH,

помутніння, солоність, фосфати, нітрати тощо. Індекс якості, що розраховується за методами запропонованими у розділах 2, 3 використовується для розрахунку.

4.1.1 Архітектура сховища даних

Загалом, сховище даних (СД) - це об'єднане сховище для всіх даних, які можливо збирати через різні джерела даних.

У дисертаційній роботі використано типову архітектуру зберігання даних, запропоновану в [23], яка проілюстрована на рис. 4.3, що має чотири окремі модулі: (1) сирі дані (джерела даних), (2) трансляція – трансформація – видобуток (ETL), (3) інтегрована інформація; та (4) обмін даними.

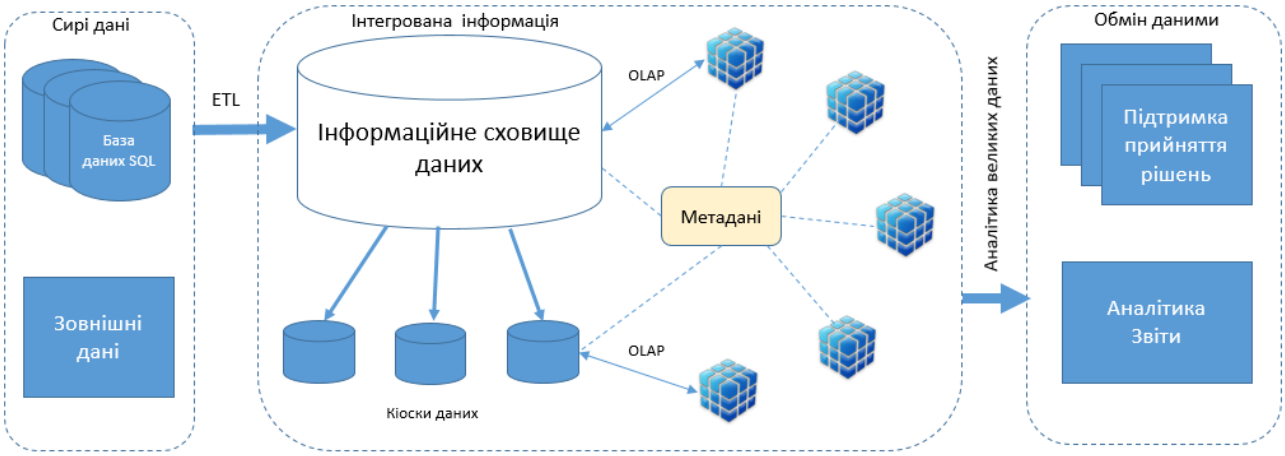


Рисунок 4.3 – Архітектура сховища даних

Сирі дані спочатку зберігаються в операційних системах баз даних або надходять із зовнішніх інформаційних систем. Необхідні дані потребують очищення, щоб забезпечити якість даних перед їх використанням. Для цієї задачі використовується модуль ETL, що містить інструменти для вилучення, перетворення та завантаження даних у сховище даних. Модуль інтегрованої інформації, містить централізоване сховище даних, метадані та механізм OLAP. Зберігання сховища даних впорядковується, зберігається та доступ до них за

допомогою відповідної схеми, визначеної у метаданих. Для великих і середніх рибогосподарств та підприємств аквакультури з розгалуженою структурою дані можуть бути сформовані з використанням набору кіосків даних. Кіоск (вітрина) даних являє собою підмножину сховища даних, зазвичай орієнтовану на певну ділову функцію чи відділ.

Згідно рис. 4.2, дані витягуються у вигляді куба даних, перш ніж вони будуть проаналізовані в модулі обміну даними. Кубічна структура даних дозволяє швидко аналізувати дані відповідно до кількох вимірів, що визначається окремою задачею або бізнес-проблемою.

Нарешті, модуль обміну даними містить набір методів навчання для аналізу великих даних та отримання знань. Знання представлені у формі правил, які можна швидко інтерпретувати користувачами. Точність видобутку даних та методів аналізу даних залежить від якості СД. В [23], згадувалося, що якість СД повинна відповідати наступним критеріям:

– Легка доступність інформації.
– Постійне представлення інформації.
– Інтеграція даних має виконуватись правильно та повністю.
– Адаптація до змін.
– Оперативність подання інформації.
– Захищеність.
– Інформація має бути надійною основою для прийняття рішень.
– Аналітичні інструменти мають забезпечувати правильною інформацією в потрібний час.
– Прийняття користувачами, що використовують СД.

Отже, технології повинні мати деякі ключові характеристики, такі як висока продуктивність, підтримка наукових даних, висока ємність зберігання, масштабованість та безпека. Нижче, представлено результати експериментів зі масштабованості рішень для візуалізації великих наборів даних зі сховища даних на реальних наборах.

Для обробки даних, які потрапляють до системи, організовано конвеєр, що функціонує у такий спосіб (рис. 4.4). У цій моделі дані створюються із системи обробки он-лайн транзакцій або OLTP (On Line Transaction Processing), яка надходить у систему, яка дає різні результати включаючи дані / куби даних для он-лайн аналітичної обробки OLAP, звіти про споживання та витрати, моделі прогнозування, що підтримують прийняття рішень із зворотним зв'язком для OLTP, тощо.

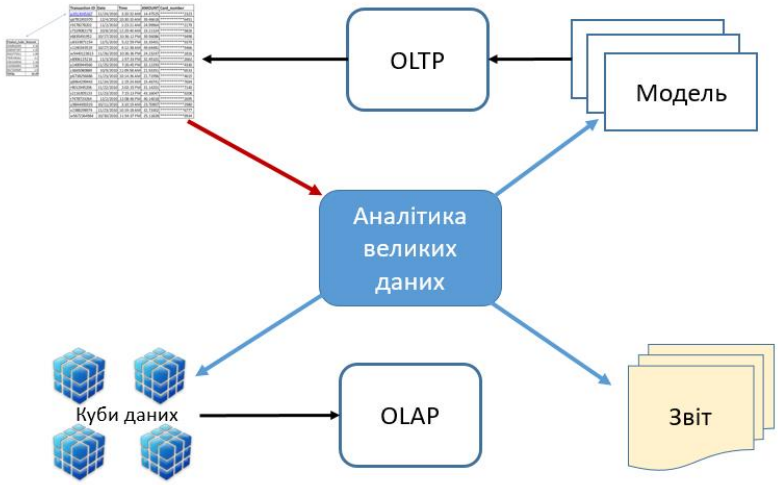


Рисунок 4.4 – Схема обробки даних

Аналіз даних виконується пакетним способом (один раз на день), для завдань системи, виконується два процеси обробки – обробка моніторингових даних та глибокий аналіз даних, кожен з яких відбувається на різних етапах пакетного процесу.

4.1.2 Вибір OLAP для аналітичної обробки даних в реальному часі

Як зазначалося у Розділі 1, відповідно до предметної області дисертаційної роботи, існує необхідність побудови моделі даних для цілей аналітичної обробки інформації про стан водних ресурсів, зокрема стосовно до задачі моніторингу якості вод для рибних господарств і підприємств напівінтенсивної та інтенсивної аквакультури.

За [8] основними вимогами до подання та обробки даних є:

– багатовимірне подання даних з підтримкою ієрархій і множинних ієрархій, здатність врахувати різні шляхи агрегації;

– наявність засобів статистичного, оперативного та інтелектуального аналізу даних, та інших видів аналізу які визначаються бізнес-процесами;

– візуалізація результатів в доступному для кінцевого користувача вигляді;

– однаково висока швидкість виконання всіх запитів до системи.

В запропонованій архітектурі, багатовимірний аналіз даних і можливість складних обчислень, аналіз тенденцій та складне моделювання даних з невеликим часом виконання забезпечує OLAP (рис. 4.1, 4.4). Багатовимірне сховище даних являє собою набір <D, F>, що складається з простору вимірювань D (dimension), і сукупності фактів F, що визначаються мірами M (measure). Технологія багатовимірного моделювання передбачає, що дані зберігаються у вигляді кубів, які представляють собою групу комірок даних, розташованих за вимірюваннями даних (рис. 4.5).

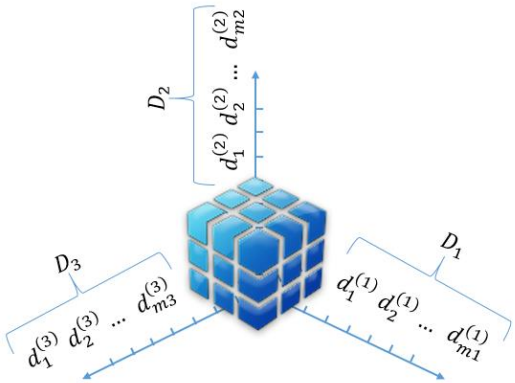


Рисунок 4.5 – Інтуїтивне подання куба даних

Кожен з кубів складається зі списку k-вимірів з ім'ям D_i і значеннями з dom_i $D =< D_1,...,D_k, M >$; виміру M, що являє собою міру куба, $M \in \Omega$ і набору кортежів виду $x =< x_1,...,x_n, m >$ $\forall i \in [1,...,n]$ $x_i \in dom_i(D)$, $m \in dom_i(ML)$, де ML визначає рівень вимірів мір куба. Елементи кортежу визначаються як

відображення R(C) $dom_1 \times dom_2 \times ... \times dom_k$ і можуть приймати значення 0, 1, кортеж з n-елементів [13].

Найпопулярнішими операціями OLAP над багатовимірними даними є згортання (консолідація), розгортання, формування продольних та поперечних зрізів даних (або нарізання кубиків) та поворот (обертання). Згортання (Roll-up) виконує агрегацію даних обчислюється в багатьох вимірах. Розгортання (Drill down) - це зворотний процес зведення, що дозволяє користувачам переходити від менш детальних даних до більш детальних даних. Операції з формування продольних та поперечних зрізів даних (Slice and dice) виконують вибір на одному або декількох вимірах даного куба, в результаті чого утворюється субкуб. Поворот (Pivot) обертає осі даних з огляду, щоб забезпечити альтернативне представлення даних. Для вибору системи OLAP було проаналізовано три типи:

(1) Реляційний OLAP (ROLAP), який використовує реляційну або розширену реляційну базу даних і не вимагає попередніх обчислень;

(2) Багатовимірний Multidimensional OLAP (MOLAP), що використовує багатовимірні системи зберігання даних на основі масиву для багатовимірних представлень даних і часто потребують попередньої обробки для створення кубів даних;

(3) Гібридний OLAP (HOLAP), який є комбінацією ROLAP і MOLAP. Він пропонує більш високу масштабованість ROLAP та швидше обчислення MOLAP.

У контексті великих даних, ROLAP не підходить через свою продуктивність; кожен звіт ROLAP - це SQL-запит у реляційній базі даних, який вимагає значного часу на виконання. Крім того, ROLAP зі своїми операторами SQL не відповідає всім потребам користувачів, особливо при виконанні складних обчислень. Хоча MOLAP є більш традиційною формою двигуна OLAP для сховищ даних, оскільки він долає недоліки ROLAP, а склад даних часто будується на багатовимірній схемі. Однак MOLAP має недолік у тому, що всі обчислення потрібно виконувати під час побудови куба даних. Apache Kylin [16] - це двигун розподіленої аналітики, що забезпечує MOLAP в масштабі з більш ніж 10

мільярдів рядків записів даних. Цей двигун із відкритим кодом реалізує деякі методи покращення зазначеного недоліку, а саме:

(1) зберігання попередньо обчислених результатів для обслуговування запитів аналізу;

(2) створення кубоїдів кожного рівня з усіма можливими комбінаціями вимірів;

(3) обчислення всіх показників на різних рівнях; та

(4) використання розподілених обчислювальних потужностей.

Разом з тим, варто відзначити, що технологія MOLAP Kylin не достатньо ефективна в запитах вимірах високої розмірності. Крім того, поєднання результатів у реальному або майже реальному часі та історичних для прийняття рішень є дійсно необхідно, але MOLAP корисний лише для подання запитів на історичні дані. Отже, для великих даних що поступають від систем моніторингу водних об’єктів та підприємств аквакультури, було обрано HOLAP, оскільки він забезпечує можливість використовувати технології ROLAP в пам'яті для обробки даних у режимі реального часу.

4.1.3 Висока розмірність і особливості атрибутів даних

СД інформаційно-аналітичної системи моніторингу водних об’єктів та підприємств аквакультури використовує схему для логічного опису всіх наборів даних. Схема СД складається з наборів таблиць фактів, їх відповідних розмірних таблиць та залежностей. Вимір являє собою структуру, що категоризує атрибут даних, таких як факти та спостереження. Виміри мають набір основних функцій для забезпечення фільтрації, групування та маркування даних що аналізуються. Рядки в кожній таблиці вимірів однозначно ідентифікуються одним ключовим полем. Таблиця фактів складається з ключів вимірів та вимірювань або метрик певного бізнес-процесу. У наведеній на рис. 4.6 схемі представлено набір даних у вигляді схеми сузір'я, також відомої як галактична схема, шо містить множину

таблиць фактів і таблиць вимірів. В роботі представлено лише деякі конкретні таблиці фактів та розмірності. На рис. 4.6 описана частина схеми сузір'я для систем моніторингу водних об’єктів та підприємств аквакультури, яка включає чотири таблиці фактів та 17-мірні таблиці.

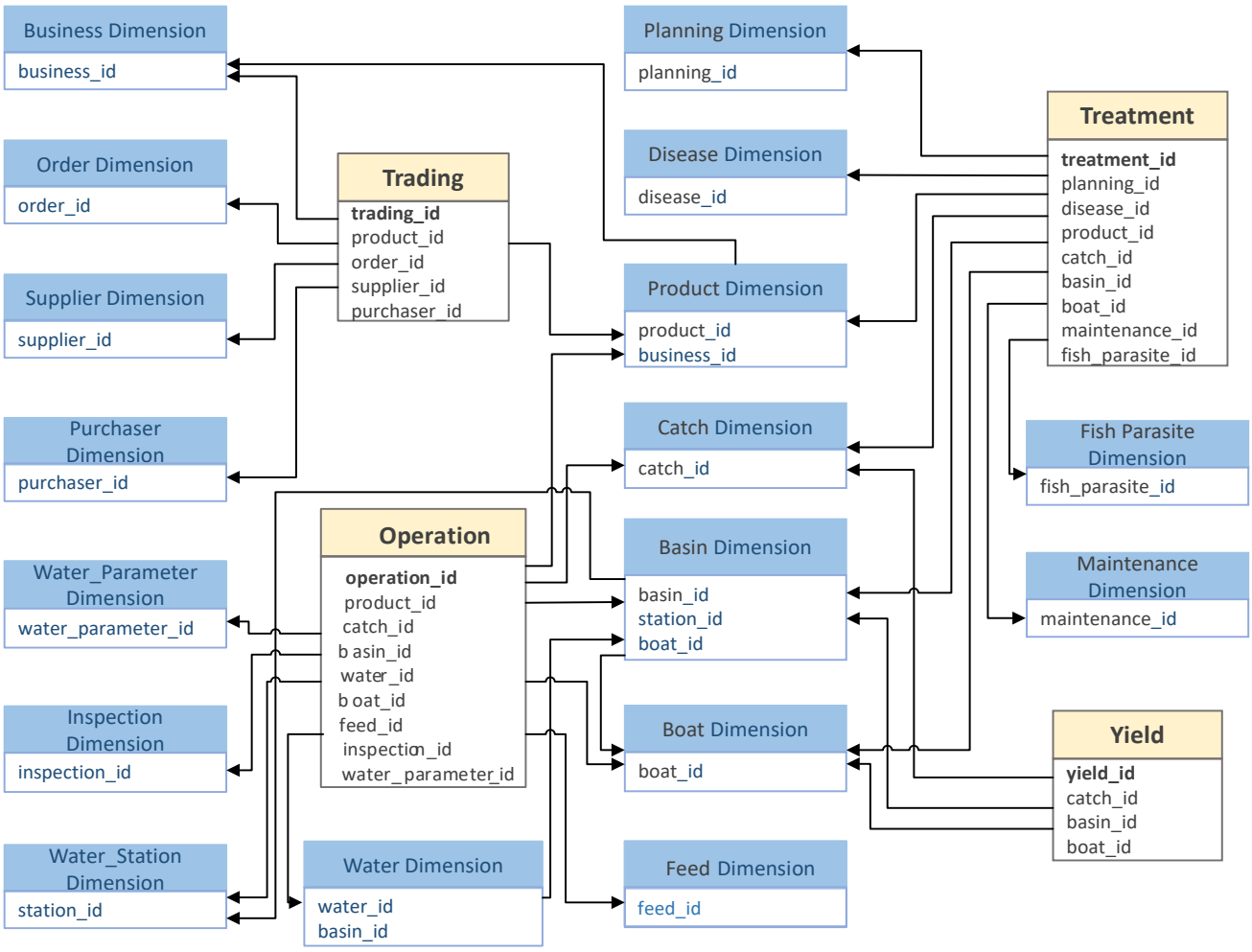


Рисунок 4.6 – Фрагмент схеми сховища даних інформаційно-аналітичної системи моніторингу водних об’єктів та підприємств аквакультури

Таблиці фактів складаються з наступних елементів: Торгівля, Операція, Лікування та Вихід. Таблиця фактів “Торгівля” має чотири виміри, а саме товар, замовлення, постачальник та покупець. Таблиця фактів “Операція” описує експлуатаційні характеристики та має вісім вимірів, а саме: продукт, вилов, водойма, човен, корм, інспекція, параметри води, місце забору.

Таблиця фактів “Лікування” має вісім вимірів, а саме: планування, хвороба, продукт, вилов, водойма, човен, технічне обслуговування та рибні паразити. Нарешті, таблиця фактів “Вихід” має три виміри: вилов, водойма, човен. Кожна таблиця фактів має первинний ключ, що є комбінацією первинного ключа таблиць вимірів.

Таблиця 4.1 описує особливості атрибутів 17-ти розмірних таблиць, які представлені на рис. 4.6 та виміри СД системи моніторингу водних об’єктів підприємства аквакультури.

Таблиця 4.1 – Опис вимірів таблиць

№	Виміри	Атрибути таблиці	Зв’язок з фактами/вимірами
1	Виробництво (business)	ідентифікатор виробництва, назва виробництва, офіційна назва, адреса, телефон, мобільний телефон, електронна адреса	Таблиця фактів "Торгівля"
2	Вилов (catch)	Ідентифікатор вилову, назва, код, назва виду риби, опис виду риби, оцінена кількість, розмірність риби	Таблиці фактів "Операція", "Лікування", "Вихід"
3	Лікування (disease)	disease_id, name, type, features on crop, description, measure	Таблиці фактів "Операція", "Лікування"
4	Човен (boat)	farmer_id, name, sex, birth year, address, field area, phone, email, experiences, skills	Таблиці фактів "Операція", "Лікування", "Вихід", таблиця вимірів "Водойма"
5	Водойма (basin)	ідентифікатор водойми, ідентифікатор станції, ідентифікатор човна, назва, розташування, площа поверхні	Таблиці фактів "Операція", "Лікування", "Вихід", таблиці вимірів "Човен", "Водна станція"
6	Корм (feed)	ідентифікатор корму, назва продукту, покриття водойми, кількість, дата підкормки	Таблиця фактів "Операція"
7	Інспекція (inspection)	ідентифікатор інспекції, опис, тип проблеми, складність, дата, ступінь	Таблиця фактів "Операція"
8	Технічне обслуговування (maintenance)	ідентифікатор технічного обслуговування, рівень, опис, дата	Таблиця фактів "Лікування"
9	Замовлення (order)	ідентифікатор замовлення, дата замовлення, дата транзакції, значення, коментар, посилання	Таблиця фактів "Торгівля"

10	Рибний паразит (fish paratise)	ідентифікатор рибного паразиту, назва, наукова назва, тип, опис, покриття	Таблиця фактів "Лікування"
11	Планування (planning)	ідентифікатор планування, назва, номер плану, назва продукту, дата, нотатки	Таблиця фактів "Лікування"
12	Продукт (product)	ідентифікатор продукту, назва продукту, назва групи, назва типу, дата виробництва, ідентифікатор виробництва	Таблиці фактів "Операція", "Лікування", "Вихід", "Торгівля", таблиця вимірів "Виробництво"
13	Покупець (purchaiser)	ідентифікатор покупця, ім’я, адреса, контактна особа, телефон, мобільний телефон, електронна адреса	Таблиця фактів "Торгівля"
14	Постачальник (supplier)	ідентифікатор постачальника, ім’я, адреса, контактна особа, телефон, мобільний телефон, електронна адреса	Таблиця фактів "Торгівля"
15	Водна станція (water station)	ідентифікатор водної станції, назва станції, дата виміру, температура повітря, температура води	Таблиця вимірів "Водойма"
16	Параметри води (water parameter)	ідентифікатор параметру води, об’єм, джерело, кількість, дата	Таблиця фактів "Операція"
17	Параметри забору води (water)	ідентифікатор забору води, ідентифікатор водойми, глибина забору, дата	Таблиця фактів "Операція", таблиця вимірів "Водойма"

Кожна таблиця вимірів має первинний ключ з тим же найменуванням таблиці. Існують деякі таблиці вимірів, які поділяються між таблицями фактів. Наприклад, таблиці вимірів "Вилов", "Човен" та "Водойма" діляться таблицями фактів "Операція", "Лікування" та "Вихід". Деякі таблиці вимірів пов'язані між собою, наприклад, такі як таблиці вимірів "Водойма" та "Водна станція".

4.2 Аналітичний інструментарій системи моніторингу водних об’єктів та підтримки прийняття рішень

Система моніторингу водних об’єктів для екосистеми рибного господарства забезпечить виживання водного життя, попереджаючи про виникнення незвичайних подій, такі як коливання кольору води, температури,

рівня солоності, концентрації вуглекислого газу, концентрації кисню тощо. Дані, зібрані з датчиків можуть бути використані для аналізу ненормальних подій за допомогою різноманітних аналітичних інструментів. Наприклад, отримують попередження, коли спостерігаються різкі перепади температур, оскільки риби не справляються з коливаннями температури. Іншим важливим аспектом використання системи є прогнозування простою обладнання, яке використовується для управління рибогосподарською екосистемою. Виявлення та відстеження ненормальної поведінки датчиків за допомогою програми інтелектуального технічного обслуговування може допомогти зменшити витрати на технічне обслуговування, понесені через несправність обладнання, що є суттєвим фактором для підприємств інтенсивної аквакультури. Наприклад, може бути надіслано попереднє попередження про необхідність незначного ремонту, коли виявлено ненормальну поведінку. У цьому випадку система мінімізує несподівану і раптову несправність обладнання що контролює параметри води. Останній функціонал закладено в роботу системи але не розглядається докладно оскільки виходить за рамки даного дисертаційного дослідження.

Таким чином, особливості процесу обробки великих даних отриманих від систем моніторингу водних об’єктів визначили наступну структуру сховища даних (рис. 4.7). З огляду на вимоги до подання та обробки даних, представлені в п. 4.1.2, прикладом лінійного упорядкування даних може служити ієрархія вимірювання "Водойма" [станція ← човен ← параметри води].

Кожному рівню вимірювань відповідає набір значень, що належать йому (наприклад, рівень вимірювання "Параметри води" має значення "Контроль pH", "Контроль прозорості", "Дослідження прозорості" та ін.).

Аналітичні інструменти використовують три типи вхідних даних відповідно до впливу, який вони гарантують:

1. Ідентифікаційні дані. Це дані, які дозволяють власникам рибних господарств керувати виробництвом (температура води, температура повітря, параметри води, корма та ін.) та правильно ідентифікувати рибу;

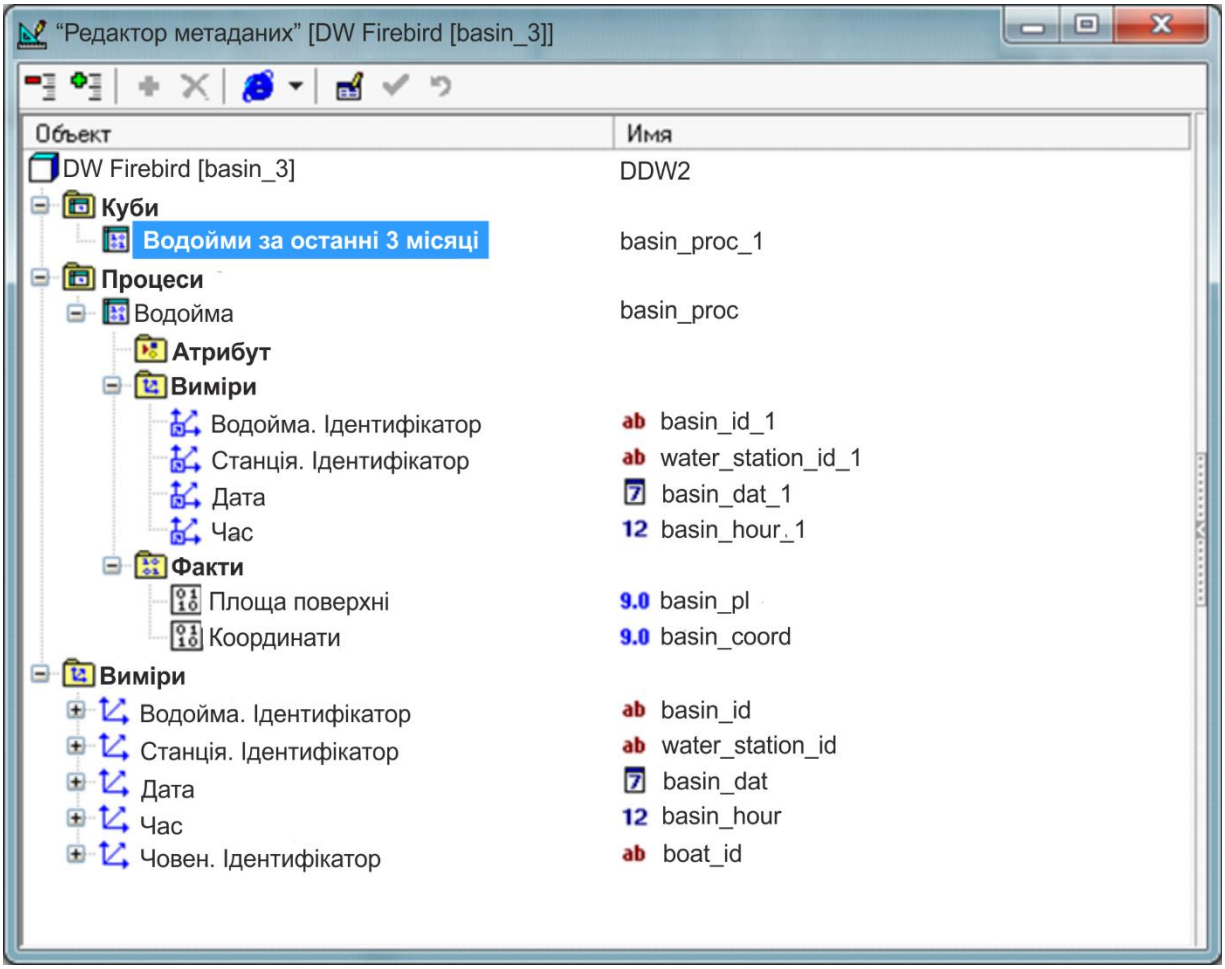


Рисунок 4.7 – Структура сховища даних

2. Щоденні дані. Це дані, які надають рибалки, отримані в результаті щоденного введення даних (наприклад, "дата", "середня вага", "фактичний корм" тощо);

3. Вибірка даних. У заздалегідь визначених точках часової шкали росту робиться зразок риби для підтвердження модельних значень та внесення відповідних коригувань;

4. Життєвий цикл. Це сукупні дані, які обчислюються з часу, коли риба потрапляє в ставок як мальок до дати збору даних, і триватиме до дати збору врожаю.

Введені дані здебільшого відповідають одній партії риби від початку зариблення до кінця виробництва. Одні вхідні дані можуть мати кілька одиниць, але для цілей використовуваних алгоритмів розглядається лише час, витрачений

на виробництво однієї партії. Для деяких використовуваних алгоритмів дані поділяються таким чином (деякі таблиці даних не мають значень у стовпці "збір") з чітким введенням / виведенням в межах одного блоку. Це також перевіряється під час очищення даних.

Дані вибірки служать рибним господарствам та аквафармерам для налаштування та вдосконалення їх початкової моделі господарювання з реальними даними. Ці вихідні дані збираються і дають можливість пристосувати вимірювання до реалій виробництва аквакультури. Окремо реалізовано показники і функції, про які можна дізнатися за певним набором даних, такі як інтегральний індекс якості води, елементи якого розглядалися в Розділі 2. Ці функції відповідають стовпцям, що потенційно впливають на кінцевий результат.

Також вони можуть впливати на виробництво (наприклад, "годівниця"). Програмне забезпечення, вбудоване в SmartWater [2, 3, 5, 28], адаптується до даних та виконує аналіз та прогнозування з наявних даних. Вхідні дані також містять стовпці даних, необов'язкові для аквафермерів (наприклад, для ставків окрім передустановлених показників, вони містять дані про рівень pH, який важливий у закритій системі аквакультури). Це зроблено з метою уніфікації системи, оскільки не можливо наперед передбачити ні релевантність даних у цих стовпцях, ні їх природу, але можливо передбачити їх появу у загальній глобальній аналітиці. Засоби фільтрування статистичних даних відповідають за управління базами даних, статистичний аналіз та візуалізацію даних. Ці засоби забезпечують всю систему потужними методами фільтрації даних для підготовки даних для подальшого знаходження знань.

Початкова версія сховища розроблялася в системі Deductor [19]. Режим управління даними (рис. 4.8) дозволяє додавати новий вимір до сховища даних, видаляти один вимір зі сховища та змінювати характеристики вимірів. Також підтримується керування фактами. Поєднання засобів статистичного аналізу з засобами візуалізації даних надає користувачеві більш детальну інформацію, та поліпшує розуміння трактування остаточних результатів.

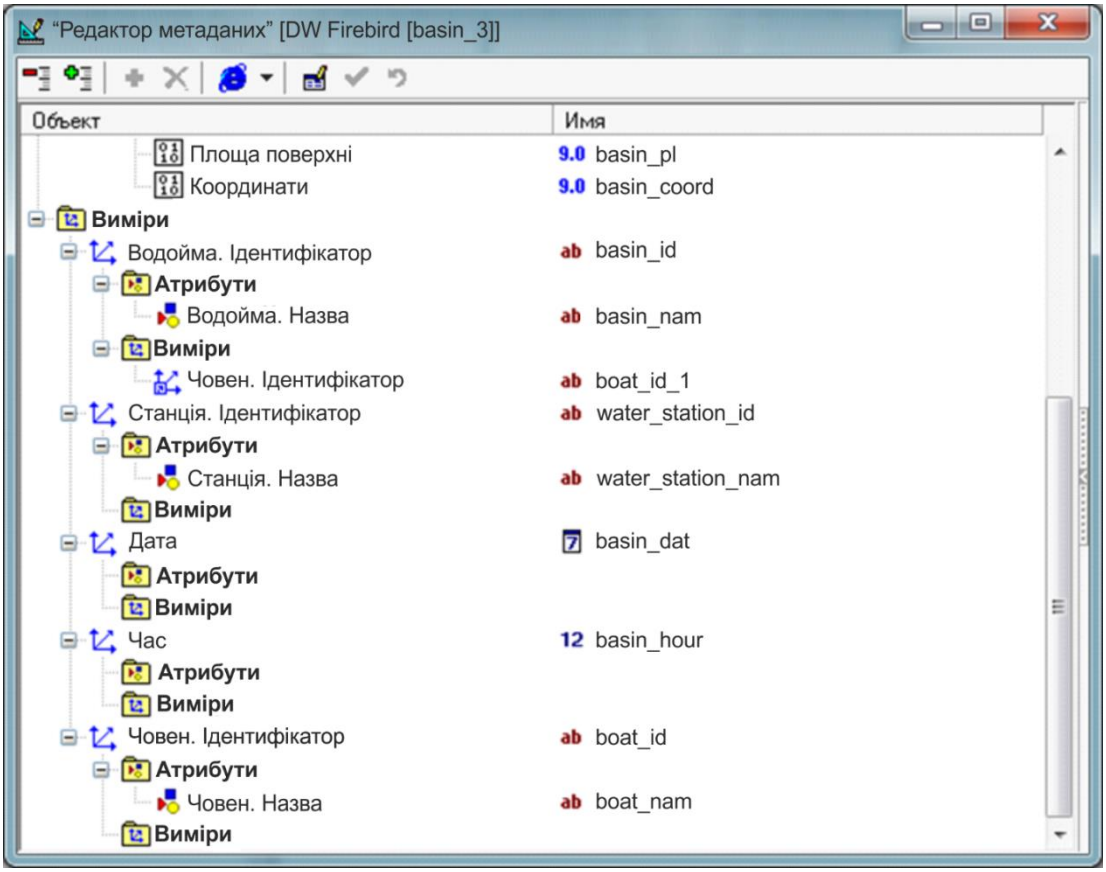


Рисунок 4.8 – Деталізація характеристик вимірів в режимі редактора метаданих

Статистичний аналіз даних реалізується за допомогою оператора навігації над кубом даних. Мета оператора навігації полягає в тому, щоб за рахунок використання статистичної функції (однієї з набору *{сума, усереднення, кількість, мінімум, ступінь (n), порожня операція}*) взяти куб з певного стану, змінити рівень конкретного виміру, упакувати результат і сформувати новий куб з новим станом. Вимірювання нового куба відповідають вимірам старого. Рівні вимірювання також однакові крім того одного виміру, для якого змінений рівень. Графічне представлення результатів аналізу реалізуються як однофакторними ділянками (двомірні графіки), так і двофакторними сюжетами (тривимірні діаграми) (рис. 4.9), а також гістограмами або таблицями для візуалізації розподілу значень вимірюваних параметрів у часі.

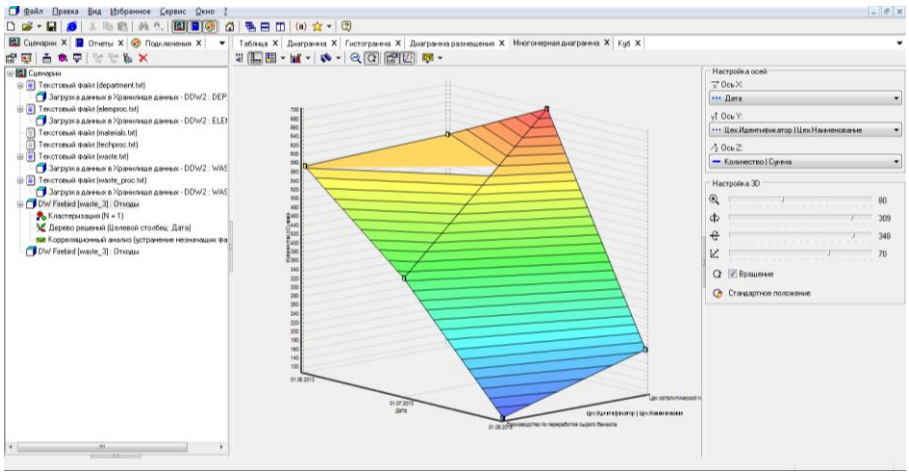


Рисунок 4.9 – Приклад описового статистичного аналізу

Ще однією складовою частиною засобів обробки даних є кластерний аналізатор, математичне забезпечення якого розглядалося у Розділі 3. Для налаштування системи було проведено роботу з алгоритмами k-середніх (рис. 4.10) c-середніх та g-середніх над даними аналітичної системи обліку якості води на прикладі рибного господарства напівінтенсивного типу. Метою застосування методу c-середніх є формування набору даних таким чином, щоб максимально схожі дані певної групи, були належно представлені користувачеві.

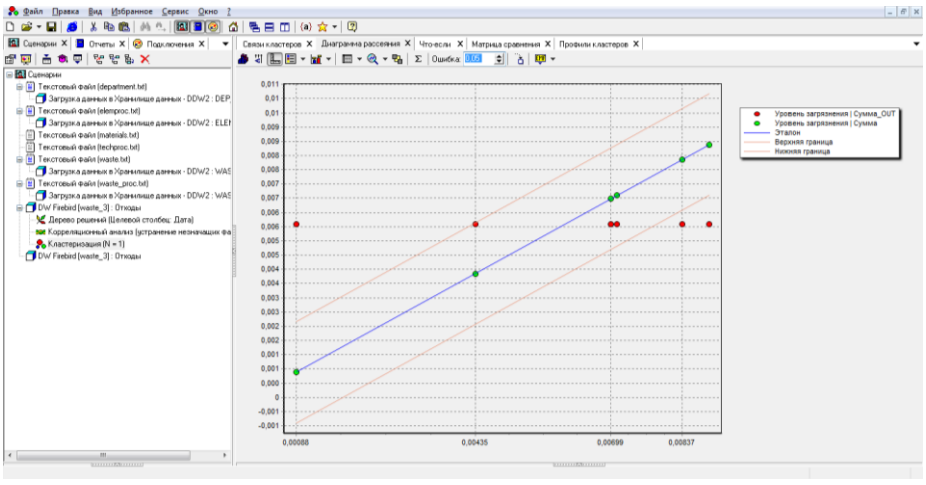


Рисунок 4.10 – Діаграма розсіювання для двох кластерів

Вхідні дані для сховища даних, використовувані в дисертаційному дослідженні, були отримані від рибогосподарських підприємств, включаючи дані

їх моніторингових систем, результати досліджень та випробування на місцях з використанням системи SmartWater [28] та з використанням польових досліджень. Ці набори даних були зібрані з демонстраційних рибних господарств Луганської області. Всього зібрано 29 наборів даних. В середньому кожен набір даних містить 18 таблиць і розміром становить близько 1,4 ГБ. Зрозуміло, що такий обсяг не можна вважати великими даними у загально прийнятному сенсі, але на момент закінчення роботи, процес збору даних продовжується, отже в майбутньому тестування моделей та інформаційної технології буде проведено на наборах, що значно перевершують тестовані.

4.3 Підходи до ефективного спрощення і візуалізації великих наборів даних

За результатами аналізу представленого в Розділі 1, для великих даних виділено також задачу їх візуалізації. Запропоновані у даному підрозділі рішення представлено у вигляді хордових діаграм і методів спрощення полігональних ланцюгів. Перший використовується для візуалізації залежностей, другий для отримання найбільш збалансованого відношення рівня деталізації до кількості точок ліній, що описують динаміку параметрів водного середовища.

4.3.1 Використання хордових діаграм

Хордові діаграми є одним з найбільш прийнятних та зрозумілих засобів візуалізації кореляційних матриць [14], разом з тим, в системах екологічного моніторингу поки що вони не знайшли широкого застосування. Заповнюючи цю прогалину, для системи підтримки прийняття рішень що є частиною запропонованої інформаційної технології, в роботі розглянуто технологію побудови відображення кореляцій параметрів якості води з використанням кривих Безьє. Крива Безьє – параметрична крива, що задається виразом:

$$B(t) = \sum_{k=0}^{n} P_k b_{k,n}(t), 0 \le t \le 1 \ , \qquad (4.1)$$

де P_k – функція-компонент векторів опорних вершин, а $b_{k,n}(t)$ – базисні функції кривої Безьє, що також називаються поліномами Бернштейна. $b_{k,n}(t) = \binom{n}{k} t^k (1-t)^{n-k}$, де $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ – кількість комбінацій із n до k, де n – ступінь полінома, k – порядковий номер опорної вершини.

Квадратична крива Безьє (n=2) задається трьома опорними точками: P_0, P_1 та P_2 (рис.4.11), що описують шлях, розрахований функцією B(t):

$$B(t) = (1-t)[(1-t)P_0 + tP_1] + t[(1-t)P_1 + tP_2], 0 \le t \le 1, \qquad (4.2)$$

яку можливо інтерпретувати як лінійний інтерполянт відповідних точок на лінійних кривих Безьє від P_0 до P_1 та від P_1 до P_2 відповідно.

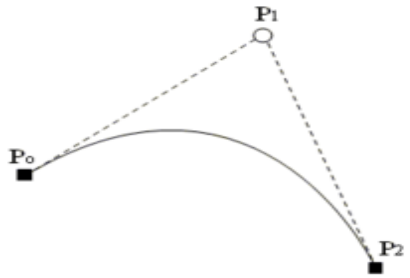


Рисунок 4.11 – Квадратична крива Безьє

Після перестановки рівняння (4.2) отримуємо:

$$B(t) = (1-t)^2 P_0 + 2(1-t)tP_1 + t^2 P_2, 0 \le t \le 1. \qquad (4.3)$$

Вираз можливо переписати таким чином, щоб підкреслити симетрію щодо P_1:

$$B(t) = P_1 + (1-t)^2 (P_0 - P_1) + t^2 (P_2 - P_1), 0 \le t \le 1. \qquad (4.4)$$

Отримуємо похідну кривої Безьє відносно t:

$$B'(t) = 2(1-t)(P_1 - P_0) + 2t(P_2 - P_1). \quad (4.5)$$

Із виразу (4.5) можливо зробити висновок, що дотичні до кривої у точках P_0 та P_2 перетинаються у точці P_1. Коли значення t збільшується від 0 до 1, крива відхиляється від P_0 у напрямку P_1, а потім згинається, щоб досягнути P_2 від напрямку P_1. Друга похідна кривої Безьє за t: $B''(t) = 2(P_2 - 2P_1 + P_0)$.

До основних властивостей кривих Безьє відносяться наступні:

– Інтерполяція кінцевої точки, тобто крива починається у точці P_0 та закінчується у точці P_n;

– Крива є прямою лінією тільки тоді, коли всі контрольні точки колінеарні;

– Початок та кінець кривої є дотичними до першого та останнього перерізів багатокутника Безьє;

– Крива може бути розподілена у будь-якій точці на дві криві або на довільну кількість кривих, кожна з яких є кривою Безьє;

– Кожна квадратична крива Безьє також є кубічною кривою Безьє, та у більш загальному сенсі кожна крива Безьє ступеня n також є кривою ступеня m для будь-якого $m > n$. Тобто, крива ступеня n із контрольними точками $P_0, \dots, P_n$ еквівалентна (включно із параметризацією) кривій ступеня $n + 1$ із контрольними точками $P'_0, \dots, P'_{n+1}$, де $P'_k = \frac{k}{n+1} P_{k-1} + \left(1 - \frac{k}{n+1}\right) P_k$;

– Кривим Безьє присутня властивість зменшення варіації;

– Відсутній локальний контроль. Будь-яка зміна положення контрольної точки вимагає перерахунку та впливає на побудову усієї кривої.

Для побудови квадратичних кривих Безьє необхідне виділення двох проміжних точок Q_0 та Q_1 за умови, що параметр t змінюється від 0 до 1.

– Точка Q_0 змінюється від P_0 до P_1;

– Точка Q_1 змінюється від P_1 до P_2;
– Точка B змінюється від Q_0 до Q_1;

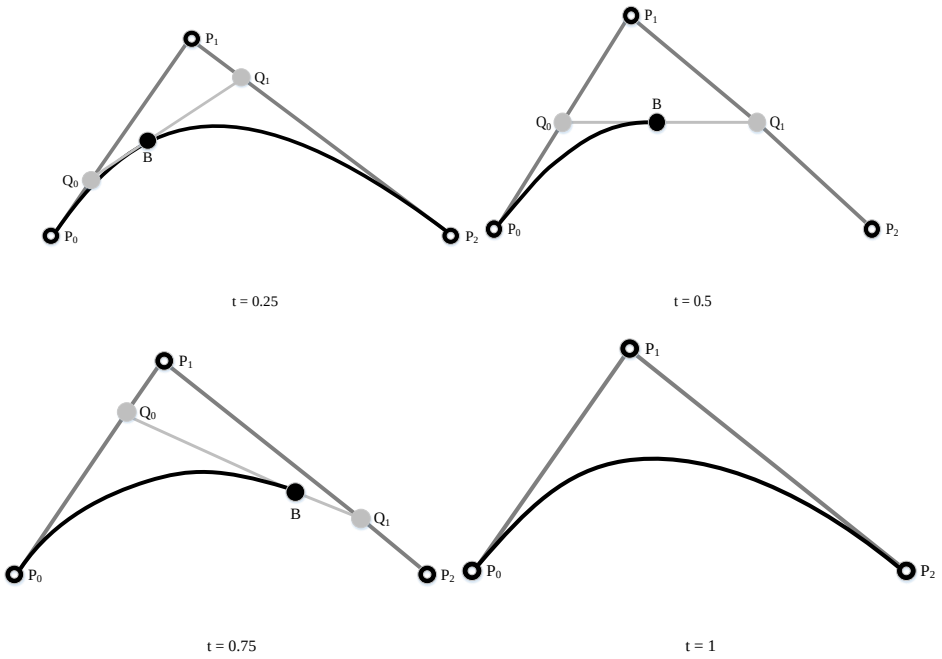


Рисунок 4.12 – Побудова квадратичної кривої Безьє при $t = 0.25$, $t = 0.5$, $t = 0.75$, і $t = 1$

Модифіковані координати положення отримуються шляхом використання алгоритмів інтерполяції прямолінійного руху зі сталою швидкістю. Якщо V_1 – швидкість прямолінійного руху Flying object, а V_2 – швидкість прямолінійного руху Supporting point, то повинна виконуватися наступна умова: $V_1 \leq V_2$.

Зміна положення, під час руху Flying object, потребує виконання наступної умови:

$$\sqrt{(X_{P_0} - X_{P_2})^2 + (Y_{P_0} - Y_{P_2})^2} \geq \sqrt{(X_{P_0} - X_{Q_2})^2 + (Y_{P_0} - Y_{Q_2})^2}$$

При недотриманні умови алгоритм генерує траєкторію, яка є неприпустимою. Необхідно відзначити, що розроблена система обробляє переміщення усіх задіяних об’єктів тільки за двома осями. Подібна особливість

обумовлена ідентичними значеннями координат, відносно третьої осі, для усіх задіяних об’єктів.

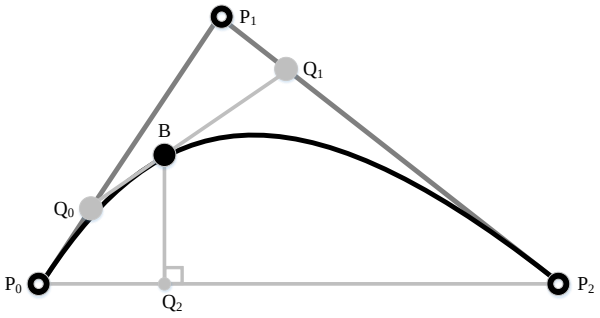


Рисунок 4.13 – Схема побудови квадратичної кривої Безьє

Для обчислення відстані між точками P_0 та Q_2 необхідні додаткові розрахунки.

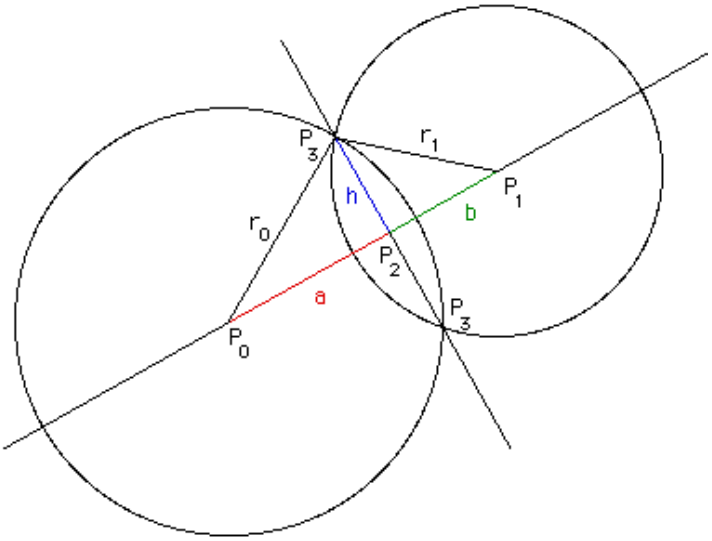


Рисунок 4.14 – Схема із додатковими елементами, де точки P_0, P_2, Q_2, та B (із попередньої схеми) відповідають точкам P_0, P_1, P_2, та P_3

Спочатку розраховується відстань d між центрами кіл. $d = \|P_1 - P_0\|$.

– Якщо $d > r_0 + r_1$ рішень немає, кола відокремлені.
– Якщо $d < |r_0 - r_1|$ рішень немає, бо одне з кіл включає інше.
– Якщо $d = 0\ and\ r_0 = r_1$ кола збігаються, безкінечна кількість рішень.

Розглянувши два трикутника $P_0P_2P_3$ та $P_1P_2P_3$ маємо наступні формули:

$a^2 + h^2 = r_0^2$ та $b^2 + h^2 = r_1^2$. Із рівності $d = a + b$ випливає: $a = \frac{(r_0^2 - r_1^2 + d^2)}{2} * d$. При перетині двох кіл тільки у одній точці, значення r_0 зменшується, тобто $d = r_0 \pm r_1$. Вираз h підстановкою a у першому рівнянні, $h^2 = r_0^2 - a^2$. Отже $P_2 = P_0 + a(P_1 - P_0)/d$. Якщо $P_3 = (x_3, y_3)$ із точки зору $P_0 = (x_0, y_0)$, $P_1 = (x_1, y_1)$ та $P_2 = (x_2, y_2)$, то

$$x_3 = x_2 \pm h(y_1 - y_0)/d$$
$$y_3 = y_2 \pm h(x_1 - x_0)/d \quad (4.6)$$

Для відтворення нахилу на діаграмі кореляційних залежностей використовується метод, що обертає цільовий об’єкт.

У табл. 4.2 надано приклад кореляційної матриці з використанням даних моніторингового експерименту якості води.

Таблиця 4.2 – Матриця кореляції параметрів якості води

	Прозорість	pH	DO	CO_2	NH_3	NO_2-N	Хлориди	Температура	Лужність	Жорсткість	Провідність	Помутніння	TDS
Прозорість	1	0.11	0.33	-0.4	-0.51	-0.2	-0.35	0.2	0.1	-0.34	-0.71	-0.42	-0.8
pH	0.11	1	-0.18	-0.21	0.11	0.02	-0.1	-0.83	-0.51	-0.75	-0.2	-0.01	-0.3
DO	0.33	-0.18	1	0.3	-0.42	-0.41	0.2	-0.03	0.44	0.03	0.1	-0.04	-0.01
CO_2	-0.4	-0.21	0.3	1	0.06	-0.06	0.4	0.1	0.2	0.35	0.5	0.01	0.33
NH_3	-0.51	0.11	-0.42	0.06	1	0.71	0.11	0.2	-0.14	-0.07	0.05	0.97	0.14
NO_2-N	-0.2	0.02	-0.41	-0.06	0.71	1	-0.25	0.07	-0.2	-0.22	-0.16	0.77	-0.15
Хлориди	-0.35	-0.1	0.2	0.4	0.11	-0.25	1	0.13	0.1	0.4	0.22	0.5	0.33
Температура	0.2	-0.83	-0.03	0.1	0.2	0.07	0.13	1	0.62	0.81	0.3	0.27	0.4
Лужність	0.1	-0.51	0.44	0.2	-0.14	-0.2	0.1	0.62	1	0.6	0.15	-0.02	0.11
Жорсткість	-0.34	-0.75	0.03	0.35	-0.07	-0.22	0.4	0.81	0.6	1	0.5	-0.07	0.6
Провідність	-0.71	-0.2	0.1	0.5	0.05	-0.16	0.22	0.3	0.15	0.5	1	-0.01	0.93
Помутніння	-0.42	-0.01	-0.04	0.01	0.97	0.77	0.5	0.27	-0.02	-0.07	-0.01	1	0.04
TDS	-0.8	-0.3	-0.01	0.33	0.14	-0.15	0.33	0.4	0.11	0.6	0.93	0.04	1

- рівень значущості 0.05
- рівень значущості 0.01

Матриця кореляції та хордова діаграма корельованих змінних дозволяють виділити основні рекурентні кореляційні закономірності. Суть хордової діаграми полягає у поданні домінуючих відносин. Такий спосіб представлення інформації дозволяє аналітику переключитись на найважливіші взаємозалежності параметрів якості води. Результуюча хордова діаграма для табл. 4.2 представлена на рис. 4.15. Хоча за допомогою такого підходу можна екстраполювати лише якісні показники, результати показали значну залежність між сімейством показників. Послідовні кольори пов'язані зі збільшенням абсолютного значення обчислених кореляцій.

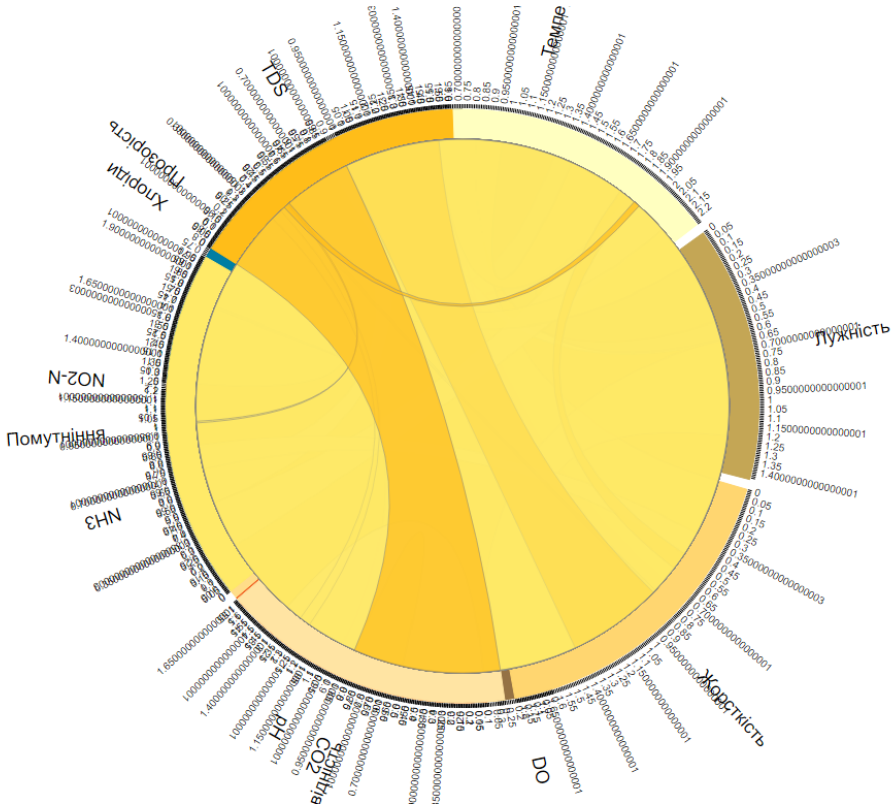


Рисунок 4.15 – Візуалізація кореляційних залежностей у вигляді хордової діаграми

Так, наприклад, провідність, TDS і нітрати негативно корелюють з прозорістю, pH з температурою, жорсткістю і лужністю. Кольори можна відобразити різними способами (+/- спрямованість, значення p тощо).

Важливо відзначити, що найбільш серйозною проблемою відсутності моніторингових даних є значний рівень відсіву в довготривалих дослідженнях, який обмежує потужність інструментів аналітики [15].

Рішення візуальної аналітики у вигляді хордових діаграм може використовуватися для аналізу відсутніх значень, що значно покращує потенціал СППР.

4.3.2 Технології спрощення полігональних ланцюгів

У загальному виді задача спрощення полігональних ланцюгів формулюється наступним чином. Нехай дана полігональна крива $P = \langle p_1, ... p_n \rangle$ і крива $P' = \langle p_{i1}, ... p_{ik} \rangle$ з $1 = i_1 < ... < i_k = n$, що спрощує криву P, де P(i, j) це відрізок шляху з p_i до p_j. Для пари індексів $1 \le i \le j \le n$, $\delta_F(p_i p_j, P)$ позначає помилку спрощення сегмента $p_i p_j$ відносно P (i, j). Тоді,

$$\delta_F(P', P) = \max_{1 \le j \le k} \delta_F(p_i p_{j+1}, P), \qquad (4.7)$$

де P′ є ε-спрощенням P, якщо $\delta_F(P', P) \le \varepsilon$.

Проблема оптимізації відображення даних може бути сформульована у вигляді задачі вибору алгоритму спрощення простих полігональних ланцюгів, а саме: Для наданої полігональної кривої P, необхідно знайти спрощення P′ з мінімальною кількістю вершин.

У більшості випадків ця проблема розглядається головним чином як задача ε- або #-мінімізації.

Проблема ε-мінімізації. Для наданого полігонального ланцюга P і цілого числа $k \le n$, знайти серед усіх апроксимацій P з не більше ніж k вершинами апроксимацію з мінімальною погрішністю.

Проблема #-мінімізації. Для наданого полігонального ланцюга P і дійсного числа $\varepsilon \geq 0$, знайти серед усіх ε -апроксимацій P ε-апроксимацію з мінімальною кількістю вершин.

Надалі в роботі розглянуто проблему ε-мінімізації.

Розрахунки необхідно проводити з урахуванням характеристик дисплею. У рамках поставленої задачі баланс між деталізацією та мінімальністю точок був зсунутий у бік зменшення сили спрощення та збільшення деталізації. Для оцінки застосовності алгоритмів спрощення необхідно розрахувати позиційні помилки.

Методи спрощення умовно поділено на три категорії в залежності від способу визначення порогу спрощення: (1) вибір точок, (2) локальна обробка, (3) обробки секцій.

Методи, що базуються на виборі точок здатні значно зменшити обсяги відображуваних даних, але не враховують математичні відношення обраної точки із сусідніми точками. Прикладами спрощення, що можна віднести до цієї категорії є випадковий вибір точок та вибір кожної N-ної точки. Оскільки дана категорія характеризується відносно невеликою точністю відображення даних моніторингу, методи з цієї категорії не були обрані до порівняння. Методи локальної обробки здійснюють вибір точок ураховуючи математичні відношення поточної точки із попередньою або двома сусідніми точками. До цієї категорії можна віднести методи радіального виключення точок та методи виключення за висотою трикутника.

Методи обробки секцій здійснюють вибір точок серед підмножини точок, ураховуючи математичні відношення точок у поточному сегменті. Виходячи з вищевикладеного, для подальшого аналізу було обрано шість підходів, а саме: (1) Спрощення за методом Рамера-Дугласа-Пекера [20], (2) Реумана-Віткама [27], (3) Опхейма [25], (4) Ланга [24], (5) виключення за висотою трикутника [5], та (6) радіальне виключення точок [1]. Схематично, особливості кожного з підходів зображено на рис 4.16.

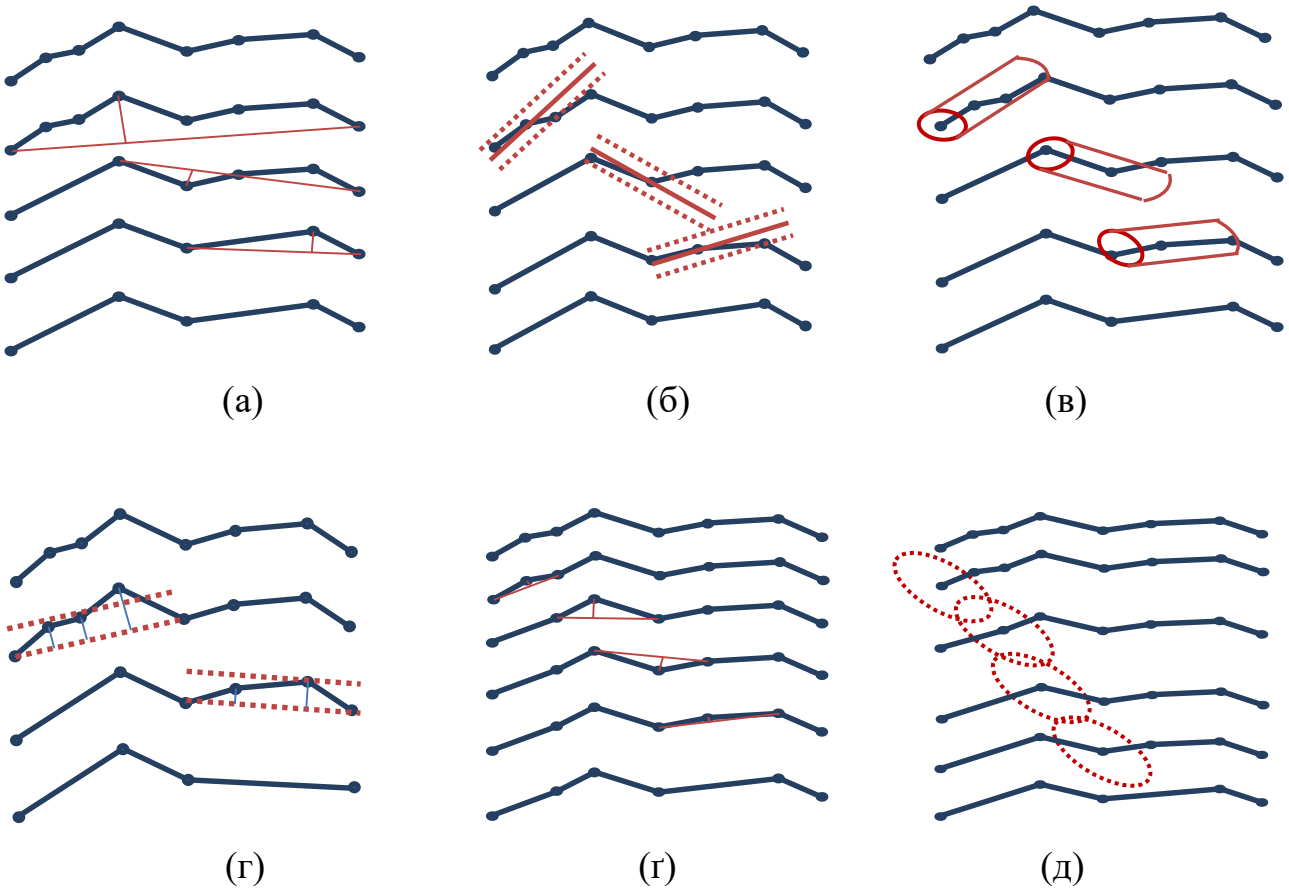


Рисунок 4.16 – Схематична дія алгоритмів спрощення полігональних ланцюгів: (а) Дугласа-Рамера-Пекера, (б) Реумана-Віткама, (в) Опхейма, (г) Ланга, (ґ) виключення за висотою трикутника, (д) радіального виключення точок

4.3.2.1 Вибір порогу спрощення

За умовами завдання, необхідно було надати можливість масштабування компоненти відображення. Для ефективного спрощення даних при масштабуванні необхідно змінювати величину порога спрощення відповідно до масштабу. При виборі порога спрощення необхідно врахувати розміри графічної сцени в пікселях.

Для алгоритмів, що відносяться до категорії локальної обробки, поріг спрощення характеризується однією величиною t (tolerance),що є відношенням між поточною точкою та сусідніми, що обрахується за ф. (4.8).

$$t = \sqrt{(\frac{|y_{max} - y_{min}|}{H})^2 + (\frac{|x_{max} - x_{min}|}{W})^2}\ , \quad (4.8)$$

де H – висота графічної сцени в пікселях, W – ширина графічної сцени в пікселях, y_{max} – верхня межа діапазону значень осі значень, y_{min} – нижня межа діапазону значень осі значень, x_{max} – верхня межа діапазону значень осі часу, x_{min} – нижня межа діапазону значень осі часу.

Для методів, що відносяться до категорії обробки секцій, поріг спрощення характеризується двома величинами: відношенням між поточною точкою та сегментом(tolerance), і довжиною сегмента.

Довжина сегмента і Tolerance обраховуються за ф. 4.9 і 4.10 відповідно.

$$l = 0.05 * |x_{max} - x_{min}|, \quad (4.9)$$

$$t = \frac{|y_{max} - y_{min}|}{H}. \quad (4.10)$$

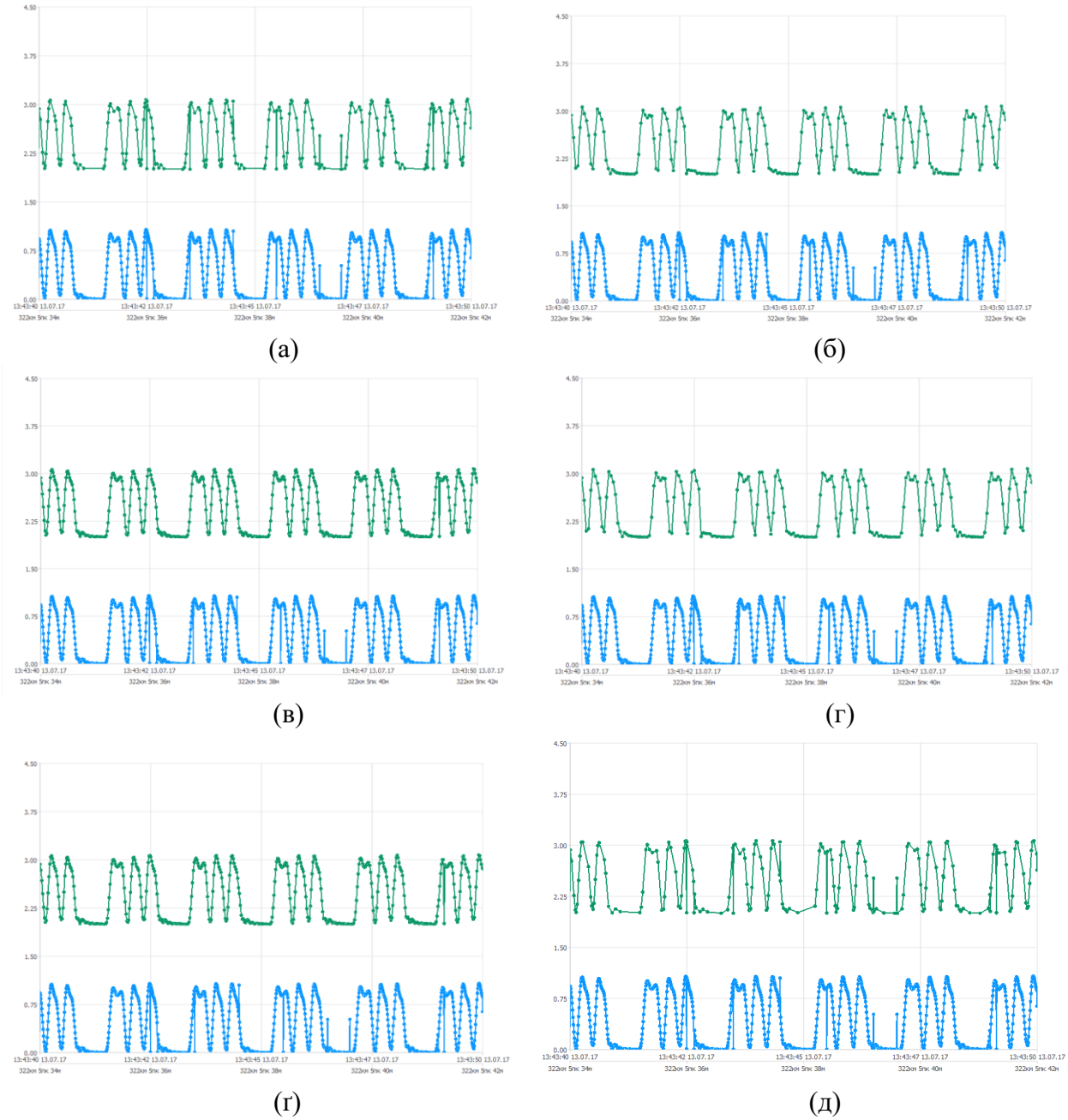
4.3.2.2 Порівняльний аналіз алгоритмів спрощення

Порівняння ефективності алгоритмів спрощення складається із двох етапів: візуального аналізу та розрахунку відхилень. Візуальний аналіз дасть змогу оцінити на скільки спрощений ланцюг передає поведінку(форму) початкового ланцюга. Розрахунок відхилень дозволить обрати з поміж алгоритмів, що однаково добре пройшли візуальний аналіз, алгоритм з найменшим значенням позиційної похибки.

З огляду на об’єктивну відсутність великих наборів даних параметрів якості води, що відзначалося вище, для проведення експерименту було обрано значення середньо-квадратичного відхилення сигналу автоматичної системи блокування, що мають усі ознаки великих даних [10]. Координати точок: по осі X – дата і час в мілісекундах з початку епохи (1970), по осі Y–значення середньо-квадратичного відхилення сигналу.

Візуальний аналіз.

На рис. 4.17 наведено результати застосування алгоритмів спрощення. Синім кольором (нижній ряд) позначено початковий ланцюг. Зеленим кольором (верхній ряд) позначено спрощений ланцюг. Для більш наглядного порівняння розглядається одна й та ж сама ділянка ланцюга даних тривалістю 10 секунд.



(а) (б) (в) (г) (ґ) (д)

Рисунок 4.17 – Візуальний аналіз результатів спрощення за методами (а) Дугласа-Рамера-Пекера, (б) радіального виключення точок, (в) виключення за висотою трикутника, (г) Ланга, (ґ) Опхейма, (д) Реумана-Віткама

За результатами візуального аналізу для спрощення ланцюга досліджуваного сигналу доцільно використовувати алгоритми Реумана-Віткама, алгоритм Радіального виключення та алгоритм Дугласа-Рамера-Пекера. Обрані підходи найбільш точно відображають загальну поведінку ланцюга.

Аналіз позиційних відхилень.

Для порівняння обраних алгоритмів спрощення полігональних ланцюгів було обчислено позиційні відхилення. Значення відхилень обчислювалися окремо для кожної точки, що була спрощена і дорівнювала відстані від відповідної спрощеної точки до спрощеного ланцюга. Зрозуміло, що спрощення полігонального ланцюга змінює його форму. Чим вище ступінь спрощення, тим більше спрощений ланцюг відрізняється від оригінального.

Для оцінки застосовності алгоритмів спрощення до мети роботи було розраховано позиційні помилки. Спосіб вимірювання відхилення спрощення полягає в знаходженні відстаней між вихідними точками, що були спрощені, та спрощеним ланцюгом. Для визначення якості спрощення було використано наступні показники [21]:

– максимальне відхилення (максимальна відстань від оригінальної точки, що була спрощена, до спрощеної лінії)

$$\max_{i=1,N}(p_i - p_i^m) \tag{4.11}$$

– мінімальне відхилення (максимальна відстань від оригінальної точки, що була спрощена, до спрощеної лінії)

$$\min_{i=1,N}(p_i - p_i^m) \tag{4.12}$$

– середньоквадратичне відхилення

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(p_i - p_i^m)^2}{N}} \quad (4.13)$$

– сума відхилень (сума відстаней від оригінальних точок, що були спрощені, до спрощеної лінії)

$$S = \sum_{i=1}^{N}(p_i - p_i^m). \quad (4.14)$$

Для порівняльного аналізу було обрано ланцюг, що має 15425 точок. Результати обчислення позиційних відхилень наведено у табл. 4.3.

Таблиця 4.3 – Результати обчислення похибок

Алгоритм / Показник	Дугласа-Рамера-Пекера	Радіальне виключення	Висота трикутника	Ланг	Опхейм	Реумана-Віткама
Максимальна похибка	0.022	1.062	1.064	1.062	1.064	0.127
Середня похибка	0.005	0.023	0.008	0.023	0.008	0.018
Середньо-квадратична похибка	0.005	0.069	0.064	0.069	0.066	0.019
Сума похибок	82.08	352.09	129.49	356.01	128.89	285.63
Кількість точок після спрощення	3856	3090	7713	3086	7726	2867

За результатами порівняльного аналізу, для використання в СППР та системі моніторингу було обрано алгоритм Дугласа-Рамера-Пекера, бо він найточніше відтворює поведінку початкового ланцюга. У межах поставленої задачі, баланс між вимогами до деталізації вихідного зображення та отриманням мінімальної кількості точок був зсунутий у бік зменшення сили спрощення та збільшення деталізації.

4.4 Індикатори ефективності впровадження інформаційної технології

Як зазначалося в Розділі 1, більша частина рибогосподарських об’єктів та підприємств аквакультури України управляється вручну, водночас існує великий потенціал технологічних систем, зокрема для аналізу факторів витрат та пропозицій щодо вдосконаленого моніторингу, архітектури управління та підтримки прийняття рішень, що дозволяє не тільки автоматизувати всі процеси аквакультури але й забезпечити умови для інноваційного перетворення галузі.

Прикладами успішних проектів є ТОВ «Лаурсен Аквакультура» (Рівненська область, с. Забороль) [12] що має автоматизовану лінію вирощування африканської тилапії та африканського кларієвого сома (мармуровий сом, Clarias gariepinus) і ТОВ «Аква Систем Органік» ТМ Aquafarm (Київська обл., м. Васильків) [11] українська компанія, яка реалізує проект вирощування риби (тилапія, мармуровий сом та сібас (Dicentrarchus labrax)) з використанням аквапоніки, поєднуючи аквакультуру і гідропоніку в системі рециркуляції води [7].

Товарна риба середньою вагою 1,0-1,5 кг вирощується протягом 6-8 місяців. Технологія фільтрації та аерації для ТОВ «Лаурсен Аквакультура» надана компанією Aquaculture ID (Нідерланди) [17].

Як визначається в [26], перевагами використання інформаційно-аналітичних систем моніторингу водних об’єктів, що поєднують у собі системи контролю, візуалізації та підтримки прийняття рішень, зокрема рибних господарств інтенсивного типу або рециркуляційних систем аквакультури є:

- Повністю контрольоване середовище для риб.
- Низьке використання води.
- Ефективне використання енергії.
- Ефективне використання земель.
- Оптимальна стратегія годування.
- Легке сортування та заготівля риби.

– Ефективна боротьба із захворюваннями.

Варто відзначити, що високотехнологічні системи можуть застосовуватися в будь-якому середовищі та кліматі, не обмежуючи географічних меж. Разом з тим, щоб дозволити використання таких систем, існує низка обмежень щодо інфраструктури, кормів та персоналу:

– Постійна потреба в електроенергії 24/7.

– Якісне джерело води, бажано свердловина.

– Хороша якість корму для риб, бажано екструдована дієта з високим вмістом білка та жиру з високою засвоюваністю.

– Технічно кваліфікований персонал, здатний працювати в середньотехнологічних умовах.

Основний фактор витрат на ведення аквакультури можна розділити на три складові, такі як (1) фіксована вартість, що включає витрати на технологічне обладнання, систему моніторингу і управління, датчики, резервуари, воду, корма, насоси, клапани, труби; (2) вартість робочої сили; і (3) змінні витрати, які залежать від багатьох факторів, включаючи рівень загибелі риби.

У звіті [26] було підраховано, що витрати на обладнання таких систем не повинні перевищувати 1,84 доларів США на 1 кг товарної продукції. Близько 20% інвестицій - це постійні витрати. Продуктивність праці, тобто наявність кваліфікованої робочої сили та технічних спеціалістів, таке є суттєвим фактором успішного впровадження. Складові частини витрат також включають капіталомісткість, експлуатаційний ресурс, коефіцієнт конверсії корму та рівень виживання. Значної економії можна досягти за рахунок змінних витрат.

СППР з управління водними об’єктами, що використовує сучасні технології керування і моніторингу, у тому числі SCADA (наглядовий контроль та збір даних), засоби аналізу та візуалізації даних, людино-машинної взаємодії, портативні мобільні пристрої розглядається як найбільш практичне рішення, здатне задовольнити потреби в екологічному регулюванні.

Така високотехнологічна система в Луганській області ще повністю не розроблена, але базові компоненти, розроблені та представлені в дисертаційному дослідженні вже знаходяться на стадії дослідної експлуатації. Наразі виконано аналіз і налаштування найбільш важливих параметрів, які контролюються системою моніторингу: температура, pH, кисень, азот, нітрати, інтенсивність світла, ріст та загальний вихід риби, тощо.

На даний момент перевірено наступні можливості СППР:

– Визначення індексу якості води.

– Аналіз залежностей споживання електроенергії, води, кисню, та рівня pH від щільності риби, типу та кількості кормів.

– Автоматичне визначення дати і номерів резервуарів які необхідно поєднувати, виходячи з щільності риби внаслідок різного зростання.

– Прогноз кількості біологічних відходів в залежності від щільності, типу кормів та сезону.

Наступним етапом роботи є точне визначення інших керованих змінних таких як норми та час подачі корму, параметри переробки біологічних відходів та вторинної води. Оскільки на деякі параметри витрат впливає масштаб ферми, аналіз умов вирощування, попиту та пропозицій за допомогою СППР може істотно усунути непотрібні витрати. Так, наприклад, ціна на рідкий кисень в Україні обумовлює обмежене використання аерації, лише для забезпечення потреб у кисні риб. Це в свою чергу вимагає додаткових розрахунків і точного контролю та дозування кисню, постійного контролю кількості та щільності риб (експериментально встановлено до 55 кг/м3 для риб роду тилапія).

Таким чином, результати роботи, моделі, метод та інформаційна технологія повністю відповідають глобальному руху за розширення технологічної складової у виробництво аквакультури підкреслюючи той баланс, який необхідно досягти для підтримання здоров'я екосистем, одночасно підштовхуючи систему до постійного вдосконалення.

Висновки до розділу 4

Четвертий розділ присвячений розробці засобів та інформаційної технології для обробки великих даних отриманих від систем моніторингу водних об’єктів. У розділі розкрито етапи розробки прототипу інструмента інтелектуального аналізу даних та управління непрямими знаннями баз даних спеціалізованої аналітичної системи обліку вод. Слід зазначити, що велика кількість наборів даних та знань, що зберігатимуться у великих сховищах даних, надходять від систем моніторингу або систем управління динамічними промисловими процесами керування якістю вод. Такими даними виступають хронологічні дані, зібрані про певні метеорологічні явища, про стан роботи технологічних систем, періодичні заміри тощо.

На відміну від багатьох існуючих комерційних систем, аспектом даної роботи є взаємодія існуючих методів Data Mining та розробка змішаних методів, що можуть співпрацювати з існуючими підходами для вилучення знань, що містяться в даних. Другою особливістю є поєднання засобів динамічного аналізу даних та засобів системи підтримки прийняття рішень, здатних запропонувати найкращий метод для досягнення кінцевої мети.

Наведено функціональні моделі інформаційної технології, архітектуру СППР, реалізація інформаційної технології у вигляді програмного комплексу, а також деякі методи візуалізації, які покладено в основу функціонування даної системи.

Архітектура СД містить необхідні модулі для роботи з масштабною та ефективною аналітикою. Представлена схема була оптимізована для наборів даних, які були нам доступні. Схема СД розроблена за схемою сузір'я, щоб полегшити критерії якості, вона гнучка й адаптується до інших наборів даних.
Нарешті, описи та взаємозв'язки конкретних таблиць фактів і вимірів також моделюються та включені в схему. Запропонована система є прототипом СППР

для інформаційно-аналітичних систем моніторингу водних об’єктів. Результати роботи показали перспективність подальшого розвитку описаних методів.

Результати розділу опубліковано в роботах автора [2-6, 9, 10, 28].

Список літератури до розділу 4

1. Алгоритм радіального виключення точок. Режим доступу www. http://psimpl.sourceforge.net/radial-distance.html (12.09.2020).

2. Барбарук В.М., Барбарук Л.В. Використання технології багатовимірних баз даних при проектуванні складних інформаційних систем, *Матеріали IX міжнародної конференції “Strategy of Quality in Industry and Education: Part 3.” Varna, Bulgaria: International Scientific Journal Acta Universitatis Pontica Euxinus. Special number*, С. 440-443, 2013.

3. Барбарук В.М., Барбарук Л.В. Засоби представлення багатомірних даних в інформаційно-аналітичних системах, *Міжвузівський збірник “Комп’ютерно-інтегровані технології: освіта, наука, виробництво” Луцьк*, №8. С. 5-10, 2012.

4. Барбарук Л. Методы формирования многомерных отчетов для задач учета ресурсопотребления. *Теоретичні та прикладні аспекти комп’ютерних наук та інформаційних технологій: Матеріали І міжнародної конференції TACSIT-2015.* Сєвєродонецьк: Східноукраїнський національний університет, 2015. с. 35-40.

5. Барбарук Л.В. Применение многомерных моделей данных в системах аналитической обработки данных, *Вісник Східноукраїнського національного університету ім. В. Даля*, №11, Ч. 2, С. 180-184, 2011.

6. Барбарук Л.В., Суворін О.В. Система аналізу даних та підтримки прийняття рішень з інвентаризації промислових відходів. *Вісник*

Східноукраїнського національного університету ім. В. Даля, №8 (238). С. 5-12. 2017.

7. Новини компаній: AQUAFARM вийшла на повну потужність виробництва риби (2019). Agrevery: Аграрне інформаційне агенство. Режим доступу: https://agravery.com/uk/posts/show/novini-kompanij-aquafarm-vijsla-na-povnu-potuznist-virobnictva-ribi (20.12.2020)

8. Скарга-Бандурова І.С. Моделі, методи та інформаційні технології підтримки прийняття рішень у природоохоронній діяльності. Харків: Вид-во "ТОВ "Щедра садиба плюс", 2014. – 135 с.

9. Скарга-Бандурова І.С., Барбарук Л.В., Сіряк Р.В. Моделі багатовимірних структур даних для аналізу природоохоронної діяльності промислових підприємств, *Зб. наукових статей Комп'ютерне моделювання в хімії, технологіях і системах сталого розвитку*. Київ: НТУУ "КПІ", С. 185-190, 2014.

10. Скарга-Бандурова І.С., Грушка М.О., Барбарук Л.В. Підходи до ефективного спрощення та візуалізації великих наборів даних, *Вісник Національного технічного університету "Харківський політехнічний інститут". Зб. наукових праць*. Серія: Інформатика та моделювання. Харків: НТУ "ХПІ", №. 50 (1271), С. 55-65, 2017. doi: 10.20998/2411-0558.2017.50.10.

11. ТОВ «Аква Систем Органік». Офіційний веб-сайт. Режим доступу: https://aquafarm.com.ua/ (05.11.2020)

12. ТОВ «Лаурсен Аквакультура». Офіційний веб-сайт. Режим доступу: https://www.laursen-aqua.com.ua/ (23.10.2020)

13. Agrawal R., Gupta A., Sarawagi A. (1997) Modeling Multidimensional Databases. *Proc. of ICDE*. pp. 232–243.

14. Alam T., Hussain A., Sultana S., Hasan T. et al., (2015) Water quality parameters and their correlation matrix: a case study in two important wetland beels of Bangladesh. *Ciência e Técnica*. - Vol. 30 (n. 3), pp. 463-489.

15. Alemzadeh S., Niemann U., Ittermann T., Völzke H., Schneider D., Spiliopoulou M., Preim, B. (2017). Visual Analytics of Missing Data in Epidemiological Cohort Studies. *VCBM*.

16. Apache Kylin, An open source, distributed Analytical Data Warehouse for Big Data. Режим доступу: http://kylin.apache.org (07.11.2020)

17. Aquaculture id. Офіційний веб-сайт. Режим доступу: https://www.aquacultureid.com (03.11.2020)

18. Boreisha Yu., Myronovych O. Web-based decision support systems in knowledge management and education, *Proceedings of the 2007 International Conference on Information & Knowledge Engineering, IKE 2007*, June 25-28, 2007, Las Vegas, Nevada, USA.

19. Deductor Studio Academic. Офіціфйний веб-сайт. Режим доступу: https://deductor-studio-academic.software.informer.com/download/ (02.04.2019)

20. Douglas D., Peucker T. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature, *Canadian Cartographer*, vol. 10, pp. 112-122.

21. Ekdemir S. Efficient Implementation of Polyline Simplification for Large Datasets and Usability Evaluation, Режим доступу http://www.diva-portal.org/smash/get/diva2:444686/FULLTEXT01.pdf (08.09.2020)

22. Gibert K., Flores X., Rodríguez-Roda I., Sànchez-Marrè M. (2004). Knowledge Discovery in Environmental Data Bases using GESCONDA.

23. Kimball R., M. Ross The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd Edition 3rd Wiley, 600 p.

24. Lang T. (1969) Rules for robot draughtsmen, *Geographical Magazine*, vol. 42, pp. 50-51.

25. Opheim H. Fast data reduction of a digitized curve, *GeoProcessing*, vol. 2, pp. 33-40.

26. Recirculating Aquaculture System. Режим доступу: https://www.aquacultureid.com/recirculating-aquaculture-system/ (16.09.2020)

27. Reumann K. and Witkam A. P. M. Optimizing curve segmentation in computer graphics, in *Proc. of International Computing Symposium*, pp. 467-472.

28. Skarga-Bandurova I., Krytska Y., Velykzhanin A., Barbaruk L., Suvorin O., Shorohov M. (2020). Emerging Tools for Design and Implementation of Water Quality Monitoring Based on IoT, Complex Systems Informatics and Modeling Quarterly. Published online by RTU Press, https://csimq-journals.rtu.lv Article 138, Issue 24, September/October 2020, pp. 1-14. https://doi.org/10.7250/csimq.2020-24.01.

29. Sprague, R.H. and E.D. Carlson (1982). Building Effective Decision Support Systems. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

ВИСНОВКИ

Актуальність дисертаційного дослідження обумовлена наявністю об’єктивного протиріччя між зростаючою кількістю впроваджень високотехнологічних систем он-лайн моніторингу водних об’єктів і рівнем адаптації технологій управління та обробки великих наборів даних в системах моніторингу водних об’єктів. Метою дисертаційного дослідження визначено підвищення ефективності роботи інформаційно-аналітичних систем моніторингу водних об’єктів базуючись на розробці та практичному застосуванні моделей та методів інформаційної технології обробки великих даних.

Основні результати роботи полягають в наступному:

1. З метою швидкого та ефективного визначення якості води по даним, отриманим через моніторингову систему, у дисертації представлено нову модель обробки великих даних на основі формалізації її атрибутів та інтерпретації невизначеності оцінки якості води у вигляді лінгвістичних змінних, що дозволило надати інтегровану характеристику стану водних об’єктів, для подальшого прийняття рішення.

2. Проведено удосконалення методу нечіткої кластеризації с-середніх на випадок великих даних, шляхом узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів для інтуїтивістських нечітких даних, що дозволило застосовувати автоматичну розмітку при обробці великих даних. Метод швидкої кластеризації заснований на методі нечітких с-середніх та традиційному жорсткому методі кластеризації. Результат жорсткої кластеризації використовується для орієнтування на початкове значення нечіткої кластеризації. Доведено, що запропонований метод дозволяє прискорити швидкість конвергенції. Як теоретичні результати, так і результати імітаційного моделювання показують, що підвищення ефективності кластеризації показників якості води, в свою чергу, дозволяє підвищити ефективність обробки великих даних.

3. Проведено удосконалення технології візуалізації великих даних за рахунок застосування кореляційних матриць та технології спрощення полігональних ланцюгів, що дозволило проводити динамічну візуалізацію та зменшити час на аналіз даних, отриманих з системи моніторингу, зокрема записів довготривалого моніторингу.

4. Дістала подальшого розвитку інформаційна технологія обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів шляхом її адаптації до завдань контролю та управління рибогосподарських підприємств, що забезпечує семантичну основу для комплексної автоматизації водних господарств у частині реалізації основних аналітичних функцій.

5. Усі теоретичні положення дисертації доведено до відповідних інженерних рішень із застосуванням запропонованої інформаційної технології обробки великих даних в інформаційно-аналітичних системах моніторингу водних об’єктів.

6. Достовірність представлених положень підтверджується результатами імітаційного моделювання та практичним впровадженням запропонованих інформаційних технологій. Практичні результати дисертаційної роботи апробовано та впроваджено в Управлінні Державного агентства рибного господарства у Луганській області. Матеріали дисертації також впроваджені в навчальний процес на кафедрі комп'ютерних наук та інженерії Східноукраїнського національного університету ім. В. Даля.

7. Подальші дослідження будуть направлені на розширення технологічної складової у виробництво аквакультури та удосконалення технологій підтримки прийняття рішень для покращення умов вирощування риб, зокрема більш точного контролю та дозування кисню, організації постійного контролю кількості та щільності риб, тощо.

ДОДАТОК А
СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

1. Skarga-Bandurova I., Krytska Y., Velykzhanin A., Barbaruk L., Suvorin O., Shorohov M. “Emerging Tools for Design and Implementation of Water Quality Monitoring Based on IoT”, *Complex Systems Informatics and Modeling Quarterly*. Published online by RTU Press, https://csimq-journals.rtu.lv Article 138, Issue 24, September/October 2020, pp. 1–14, 2020 https://doi.org/10.7250/csimq.2020-24.01.

2. Skarga-Bandurova I., Krytska Y., Shorohov M., Suvorin O., Barbaruk L., Ozheredova M. “Towards Development IoT-based Water Quality Monitoring System”, *Proceedings of the 7th IEEE International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pp. 140-145, 2019. doi: 10.1109/FiCloudW.2019.00038 (*Scopus*).

3. Барбарук Л.В., Зубарев Д.В., Медведєв Є.М., Барбарук В.М. “Використання нечітких моделей для планування збиральних робіт у різних кліматичних умовах”, *Вісник Східноукраїнського національного університету ім. В. Даля*, №6 (247), С. 175-179, 2018.

4. Скарга-Бандурова І.С., Грушка М.О., Барбарук Л.В. “Підходи до ефективного спрощення та візуалізації великих наборів даних”, *Вісник Національного технічного університету “Харківський політехнічний інститут”. Зб. наукових праць. Серія: Інформатика та моделювання.* Харків: НТУ “ХПІ”, №. 50 (1271), С. 55-65, 2017. doi: 10.20998/2411-0558.2017.50.10.

5. Барбарук Л.В., Суворін О.В. “Система аналізу даних та підтримки прийняття рішень з інвентаризації промислових відходів”, *Вісник Східноукраїнського національного університету ім. В. Даля*, №8 (238). С. 5-12. 2017.

6. Скарга-Бандурова І.С., Барбарук Л.В., Сіряк Р.В. “Моделі багатовимірних структур даних для аналізу природоохоронної діяльності

промислових підприємств", *Зб. наукових статей Комп'ютерне моделювання в хімії, технологіях і системах сталого розвитку.* Київ: НТУУ "КПІ", С. 185-190, 2014.

7. Барбарук В.М., Барбарук Л.В. "Методи моделювання складних систем", *Вісник Східноукраїнського національного університету ім. В. Даля*, № 15(186), С. 132-136, 2012.

8. Барбарук В.М., Барбарук Л.В. "Засоби представлення багатомірних даних в інформаційно-аналітичних системах", *Міжвузівський збірник "Комп'ютерно-інтегровані технології: освіта, наука, виробництво"* Луцьк, №8. С. 5-10, 2012.

9. Барбарук Л.В. "Применение многомерных моделей данных в системах аналитической обработки данных", *Вісник Східноукраїнського національного університету ім. В. Даля,* №11, Ч. 2, С. 180-184, 2011.

10. Рязанцев О.І., Барбарук В.М., Барбарук Л.В. "Комплексна оцінка екологічної небезпеки території промислового виробництва", *Вісник Східноукраїнського національного університету ім. В. Даля*, №7(154), Ч.2, С. 136-140, 2010.

11. Skarga-Bandurova I., Krytska Y.O., Barbaruk L.V. "Application of Internet of Things for long term water quality monitoring", *Проблеми інформатики і моделювання. Тезиси дев'ятнадцятої міжнар. наук.-техн. конф*., Харків: НТУ "ХПІ", С. 77, 2019.

12. Барбарук Л.В., Михайличенко С.О. Бездротова система віддаленого моніторингу сільськогосподарських параметрів. *Матеріали VII Всеукр. науково-практ. конф. «Електронні апарати та системи. Проблеми створення. Перспективи розвитку»*, Сєвєродонецьк : Східноукр. нац. ун-т ім. В. Даля, 2017. С. 206-208.

13. Михайличенко С.О., Барбарук Л.В. "Вибір протоколів маршрутизації для організації бездротових сенсорних мереж", *Матеріали всеукр. наук.-практ.*

конф. "Майбутній науковець – 2017" 1 груд. 2017 р., м. Сєвєродонецьк. укл. В. Ю. Тарасов. Сєвєродонецьк : Східноукр. нац. ун-т ім. В. Даля, С. 296-297, 2017.

14. Барбарук Л.В., Добрецова А.О. "Застосування алгоритму Краскала для синтезу древоподібних глобальних мереж", *Технологія-2016 : матеріали міжнар.наук.-техн. конф.*, 22-23 квіт. 2016 р., м. Сєвєродонецьк. Ч. II [укл. : Тарасов В.Ю.] Сєвєродонецьк : [Східноукр. нац. ун-т ім. В. Даля], С. 152-154, 2016.

15. Барбарук Л. "Методы формирования многомерных отчетов для задач учета ресурсопотребления", *Theoretical and Applied Computer Science and Information Technology: Proceedings of the first International Conference TACSIT-2015.* Severodonetsk: East Ukrainian National University, P. 35-40, 2015.

16. Барбарук Л.В., Скарга-Бандурова І.С. Методы оперирования многомерными данными в хранилище данных опасных промышленных отходов. *Системи обробки інформації.* Вип. 2 (118). Т.2 Проблеми і перспективи розвитку ІТ-індустрії. С. 223, 2014.

17. Барбарук В.М., Барбарук Л.В. "Використання технології багатовимірних баз даних при проектуванні складних інформаційних систем", *Матеріали IX міжнародної конференції "Strategy of Quality in Industry and Education: Part 3."* Varna, Bulgaria: International Scientific Journal Acta Universitatis Pontica Euxinus. Special number, С. 440-443, 2013.

18. Барбарук Л.В. "Сложные информационные системы", *Матеріали XIV всеукраїнської науково-практичної конференції "Технологія": ч. 2.* Сєвєродонецьк: Технол. ін-т Східноукр. Нац. ун-ту ім. В. Даля (м. Сєвєродонецьк), С. 129-131, 2011.

19. Шорохов М.М., Суворин О.В., Ожередова М.А., Зубцов Є.І., Барбарук Л.В., Критська Я.О., Мочалов В.В. Патент на корисну модель "Спосіб сумісної утилізації відпрацьованих промивних вод, що містять сполуки шестивалентного хрому, та лужних стічних вод содового виробництва". Державний реєстр патентів України на корисні моделі, реєстраційний номер: № 133168, 25.03.2019.

ДОДАТОК Б

АКТИ ВПРОВАДЖЕННЯ

ЗАТВЕРДЖУЮ
Проректор з наукової роботи
Східноукраїнського національного
університету ім. Володимира Даля
_____ Целіщев О.Б.
«10» 02 2021 р.

АКТ
про впровадження в навчальний процес
Східноукраїнського національного університету
імені Володимира Даля
результатів дисертаційного дослідження старшого викладача кафедри
комп'ютерних наук та інженерії

Барбарук Ліни Вікторівни

"Моделі та метод обробки великих даних в інформаційно-аналітичних системах моніторингу водних об'єктів"

Цим актом підтверджується, що результати досліджень, які проводились у межах наукового напряму "Інформаційні технології в промисловості, екології, медицині" за науково-дослідними роботами "Дослідження стратегій та механізмів прийняття рішень для інтегрованого управління водними ресурсами" 12.2016-07.2020, (№ ДР 0116U005784), "Проектування системи моніторингу та контролю водних об'єктів на основі технології інтернет речей" 01.2020-12.2023, (№ ДР 0120U100421), проектом Європейського Союзу ERASMUS+ ALIOT 573818-EPP-1-2016-1-UK-EPPKA2-CBHE-JP "Internet of Things: Emerging Curriculum for Industry and Human Applications"; тематичними планами науково-дослідних робіт Східноукраїнського національного університету ім. В. Даля протягом 2017-2020 рр. впроваджені на кафедрі комп'ютерних наук та інженерії в процес підготовки здобувачів вищої освіти за спеціальністю "Комп'ютерні науки".

Впровадження результатів дисертаційної роботи полягає в їх використанні при викладанні навчальних дисциплін "Комп'ютерне моделювання процесів та систем", "Методи машинного навчання", "Data-mining та бізнес-аналітика" як окремих розділів лекційних курсів, так і в циклах практичних і лабораторних робіт.

При викладанні цих дисциплін автором використовувались наступні матеріали досліджень, отримані ним особисто:

1) огляд перспективних технологій обробки великих даних;

2) модель для аналітичної обробки великих даних, отриманих від системи моніторингу водних об'єктів на основі формалізації її атрибутів та інтерпретації невизначеності оцінки якості води у вигляді лінгвістичних змінних;

3) метод нечіткої кластеризації с-середніх на випадок великих даних з використанням узагальнення процедури автоматичного маркування нечітких кластерів, отриманих за допомогою евристичних алгоритмів упорядкування інтуїтивістських нечітких множин;

4) основні етапи розробки та реалізації інформаційної технології у вигляді системи підтримки прийняття рішень;

5) сервіс-орієнтована архітектура системи обробки великих даних отриманих від станцій моніторингу водних об'єктів яка є прототипом системи підтримки прийняття рішень для інформаційно-аналітичних систем моніторингу водних об'єктів.

Декан факультету
інформаційних технологій та
електроніки, к.т.н., доц. С.О. Митрохін

В.о. завідувача кафедри
комп'ютерних наук та
інженерії, д.т.н., проф. О.І. Рязанцев

"УЗГОДЖЕНО"
Проректор з наукової роботи
Східноукраїнського національного
університету ім. Володимира Даля

_____ Целіщев О.Б.

"12" ____01____ 2021 р.

"ЗАТВЕРДЖУЮ"
Начальник Управління Державного
агентства рибного господарства у
Луганській області

_____ Поляков О.А.

"12" ____01____ 2021 р.

А К Т
про впровадження результатів дисертаційного дослідження
Барбарук Ліни Вікторівни,

представленого до захисту на здобуття вченого ступеня кандидата технічних наук

Науково-технічна комісія у складі:

заст. начальника Управління - начальник відділу охорони водних біоресурсів «Рибоохоронний патруль» Краснянського А.В.,

зав. кафедри комп'ютерних наук та інженерії СНУ ім.В.Даля Рязанцева О.І.,

зав. кафедри хімічної інженерії та екології СНУ ім.В.Даля Суворіна О.В.

склала цей акт про те, що результати дисертаційної роботи Барбарук Л.В.:

1. Моделі для обробки великих даних системи моніторингу водних об'єктів на основі інтерпретації невизначеності оцінки якості вод у вигляді лінгвістичних змінних.
2. Метод нечіткої кластеризації с-середніх для моніторингових даних якості води.
3. Технологія візуалізації великих даних у вигляді хордових діаграм та спрощених полігональних ланцюгів.
4. Інформаційна технологія обробки великих даних для інформаційно-аналітичної системи моніторингу об'єктів рибогосподарського підприємства.
5. Сервіс-орієнтована архітектура системи аналізу та обробки великих даних використані при розробці і тестуванні системи підтримки прийняття рішень з управління водними об'єктами.

Застосування запропонованих в роботі методів, моделей та інформаційної технології дозволяють зробити наступні висновки:

1. В процесі впровадження інформаційної технології підтверджено достовірність і адекватність математичних моделей, методів та алгоритмів, що застосовувалися для збору, обробки та аналізу даних.
2. Використання моделей та інформаційної технології обробки великих даних в системах моніторингу та підтримки прийняття рішень у виробничих процесах підприємства, довели їх ефективність в частині підвищення точності інтегральної оцінки якості вод,
3. Під час тестової експлуатації системи протягом 05.2019 р. – 12.2020 р., зниження витрат на виробництво склало від 13,5% до 15%, за рахунок зменшення часу на прийняття рішень щодо якості вод водойм рибогосподарського призначення, пошук невідповідностей і ретроспективний аналіз даних моніторингу.

Заст. начальника Управління - начальник відділу охорони водних біоресурсів «Рибоохоронний патруль» _____ Краснянський А.В.

Зав. кафедри комп'ютерних наук та інженерії СНУ ім.В.Даля _____ Рязанцев О.І.

Зав. кафедри хімічної інженерії та екології СНУ ім.В.Даля _____ Суворін О.В.

ДОДАТОК В

ОСНОВНІ НОРМАТИВНІ ПОКАЗНИКИ ЯКОСТІ ВОД РИБОГОСПОДАРСЬКИХ ПІДПРИЄМСТВ

Показник	Нормативні показники якості вод рибогосподарських підприємств				
	ЄС				
	Україна	Директиви 2006/44/ЄС (78/659/ЄС), 76/464/ЄС			
		Лососеві		Короп та інші види	
		G	I	G	I
Температура води, ^{0}C	28	–	21,5 10*	–	28 10*
Прозорість, см	–	–	–	–	–
Мінералізація, мг/дм^3	1000	–	–	–	–
Жорсткість, мг-екв/дм^3	7	–	–	–	–
Хлориди, мг Cl/дм^3	300	–	–	–	–
Сульфати, мг SO_4/дм^3	100	–	–	–	–
Натрій, мг Na/дм^3	120	–	–	–	–
Калій, мг K/дм^3	50	–	–	–	–
Кальцій, Ca/дм^3	180	–	–	–	–
Магній, мг Mg/дм^3	50	–	–	–	–
Завислі речовини, мг/дм^3	20	≤ 25	–	≤ 25	–
Водневий показник, pH	6,5-8,5	6,0-9,0	–	6,0-9,0	–
Розчинний кисень, мг O_2/дм^3	>6,0	50% > 9 100% > 7	50% > 9 >6	50% > 8 100% > 7	50% > 7 >4
БСК$_5$, мг O_2/дм^3	2	≤ 3	–	≤ 6	–
ХСК (Cr), мг O_2/дм^3	2	–	–	–	–
ХСК (Mn), мг O_2/дм^3	20	–	–	–	–
Азот загальний амонійний, мг N/дм^3	0,05	≤ 0,04	≤ 1	≤ 0,2	≤ 1
Азот амонійний, мг N/дм^3	0,39	≤ 0,005	≤ 0,025	≤ 0,005	≤ 0,025
Азот амонійний, мг NH_4/дм^3	0,5	0,8	–	0,8	–
Азот нітратний, мг N/дм^3	9,1	–	–	–	–
Азот нітратний, мг NO_3/дм^3	40	–	–	–	–

Азот нітритний, мг $N/дм^3$	0,02	≤ 0,01	–	≤ 0,03	–
Азот нітритний, мг $NO_2/дм^3$	0,08	–	–	–	–
Азот загальний, мг $N/дм^3$	1,0	≤ 0,04	≤1,0	≤0,2	≤1,0
Фосфати, мг $P/дм^3$	0,2	0,2	–	0,4	
Фосфати, мг $PO_4/дм^3$	3,5	–	–	–	–
Силікати, мг $SiO_3/дм^3$	30	–	–	–	–
Залізо загальне, мкг $Fe/дм^3$	5(100)	–	–	–	–
Кадмій, мкг $Cd/дм^3$	5	–	–	–	–
Кобальт, мкг $Co/дм^3$	10	–	–	–	–
Марганець, мг $Mn/дм^3$	10				
Мідь, мкг $Cu/дм^3$	1	<0,4 11,2**	–	<0,4 11,2**	–
Миш'як, мкг $As/дм^3$	50	–	–	–	–
Нікель, мкг $Ni/дм^3$	10	–	–	–	–
Ртуть, мкг $Hg/дм^3$	0,01 відсутня	–	–	–	–
Свинець, мкг $Pb/дм^3$	100	–	–	–	–
Хром (3+), мкг $Cr/дм^3$	–	–	–	–	–
Хром (6+), мкг $Cr/дм^3$	1	–	–	–	–
Цинк, мкг $Zn/дм^3$	10	≤ 500**	300	≤ 2000**	1000
Ціаніди мкг $CN/дм^3$	50	–	–	–	–
Нафтопродукти, мкг/$дм^3$	50	–	–	–	–
СПАР, мкг/$дм^3$	100	–	–	–	–
Феноли, мкг/$дм^3$	1	–	–	–	–
Пестициди, мкг/$дм^3$	4	–	–	–	–

«–» – норматив не визначено;

*– температура в період розмноження;

**– вміст кальцій гідрогенкарбонату у поверхневих водах перевищує 100 мг $Ca(HCO_3)_2/дм^3$

G – обов'язкові нормативи

I – бажані нормативи

Джерело: Клименко М.О., Вознюк Н.М., Вербецька К.Ю. Порівняльний аналіз нормативів якості поверхневих вод. Наукові доповіді НУБіП України. 2012. Режим доступу www: https://nd.nubip.edu.ua/2012_1/12kmo.pdf

ДОДАТОК Д

АЛГОРИТМ
ВИЗНАЧЕННЯ ЕКОЛОГІЧНОГО СТАНУ МАСИВУ ПОВЕРХНЕВИХ ВОД, ВІДПОВІДНО ДО МЕТОДИКИ

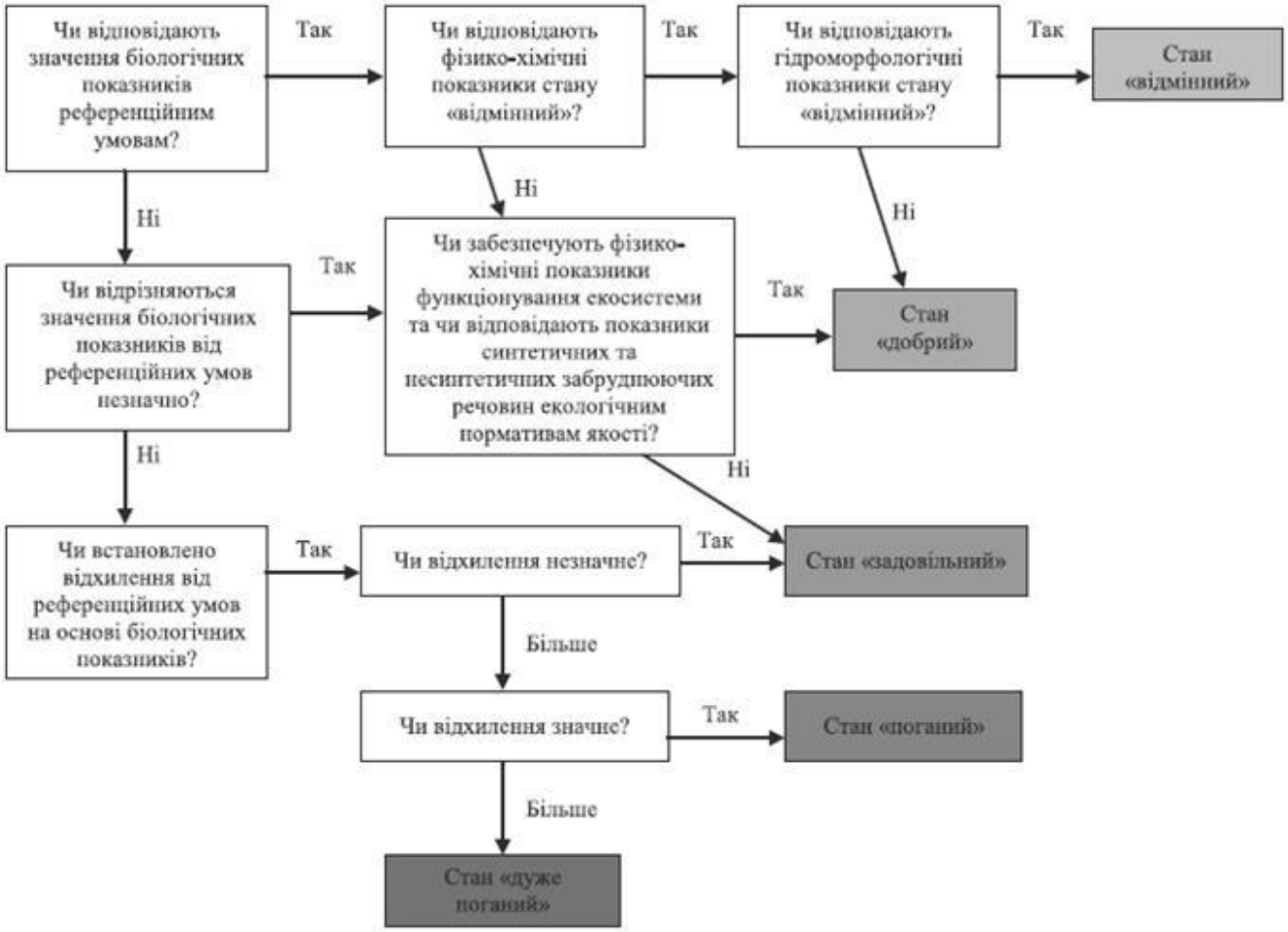


Рисунок Д.1 –Алгоритм визначення екологічного стану

Джерело: Методика віднесення масиву поверхневих вод до одного з класів екологічного та хімічного станів масиву поверхневих вод, а також віднесення штучного або істотно зміненого масиву поверхневих вод до одного з класів екологічного потенціалу штучного або істотно зміненого масиву поверхневих вод. Режим доступу: https://zakon.rada.gov.ua/laws/show/z0127-19#Text (12.12.2020).

ДОДАТОК Г

КРИТЕРІЇ
ВІДНЕСЕННЯ МАСИВУ ПОВЕРХНЕВИХ ВОД ДО ОДНОГО З КЛАСІВ ЕКОЛОГІЧНОГО СТАНУ

Стан «відмінний»	**Стан «добрий»**	**Стан «задовільний»**
Значення біологічних показників відповідають значенням, характерним для масиву поверхневих вод у референційних умовах, мають тенденцію до дуже незначних змін. Відсутні або виявлені дуже незначні антропогенні зміни значень гідроморфологічних, хімічних та фізико-хімічних показників порівняно з величинами, характерними для масиву поверхневих вод в референційних умовах	Значення біологічних показників масиву поверхневих вод вказують на низькі рівні антропогенного впливу і мало відхиляються від значень, характерних для масиву поверхневих вод у референційних умовах. Концентрації хімічних та фізико-хімічних показників не перевищують екологічних нормативів якості, встановлених для екологічного стану «добрий»	Значення біологічних показників масиву поверхневих вод помірно відхиляються від значень, характерних для масиву поверхневих вод у референційних умовах. Ці значення мають помірну тенденцію до відхилення в результаті антропогенного впливу та мають значно більші відхилення порівняно з умовами стану «добрий». Концентрації хімічних та фізико-хімічних показників перевищують екологічні нормативи якості, встановлені для екологічного стану «задовільний»

Стан «поганий»	**Стан «дуже поганий»**
Спостерігаються значні зміни щодо значень біологічних показників та значні відхилення від норм відповідних біологічних популяцій, характерних для масиву поверхневих вод у референційних умовах	Спостерігаються дуже сильні зміни щодо біологічних показників, відсутність великої частини відповідних біологічних ценозів, характерних для масиву поверхневих вод у референційних умовах

Джерело: Методика віднесення масиву поверхневих вод до одного з класів екологічного та хімічного станів масиву поверхневих вод, а також віднесення штучного або істотно зміненого масиву поверхневих вод до одного з класів екологічного потенціалу штучного або істотно зміненого масиву поверхневих вод. Режим доступу: https://zakon.rada.gov.ua/laws/show/z0127-19#Text (12.12.2020).