УДК 004.9+504.064.36:574

УКПП
№ держреєстрації 0120U100421

Східноукраїнський національний університет імені Володимира Даля
Кафедра комп'ютерних наук та інженерії
01042, м. Київ, вул. Іоанна Павла II, 17

ЗАТВЕРДЖУЮ
Проректор з наукової роботи
_____ О.Б. Целіщев

ЗВІТ
ПРО НАУКОВО-ДОСЛІДНУ РОБОТУ

ДОСЛІДЖЕННЯ СТРАТЕГІЙ ТА МЕХАНІЗМІВ ПРИЙНЯТТЯ РІШЕНЬ ДЛЯ ІНТЕГРОВАНОГО УПРАВЛІННЯ ВОДНИМИ РЕСУРСАМИ

(остаточний)

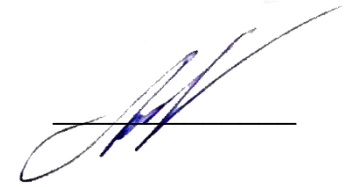
Науковий керівник НДР:
доцент кафедри
комп'ютерних наук та
інженерії СНУ ім. В. Даля,
к.т.н.

_____ Л.В. Барбарук

2023

СПИСОК АВТОРІВ

Керівник НДР:
доцент кафедри
комп’ютерних наук та
інженерії СНУ ім. В. Даля,
к.т.н.



Л.В. Барбарук

РЕФЕРАТ

Звіт про НДР: 103 с., 3 розділи, 130 джерел.

Науково-дослідну роботу 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.

Метою роботи є підвищення ефективності роботи інформаційно-аналітичних систем моніторингу водних об’єктів завдяки розробці та практичному застосуванню моделей та методів інформаційної технології 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

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 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.

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

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)
WQI	Water Quality Index (індекс якості води) Integral Water Quality Index (інтегрований індекс якості води)
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)

ВСТУП

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.

Мета й завдання дослідження. *Метою* роботи 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 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 с-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.

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 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.

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.

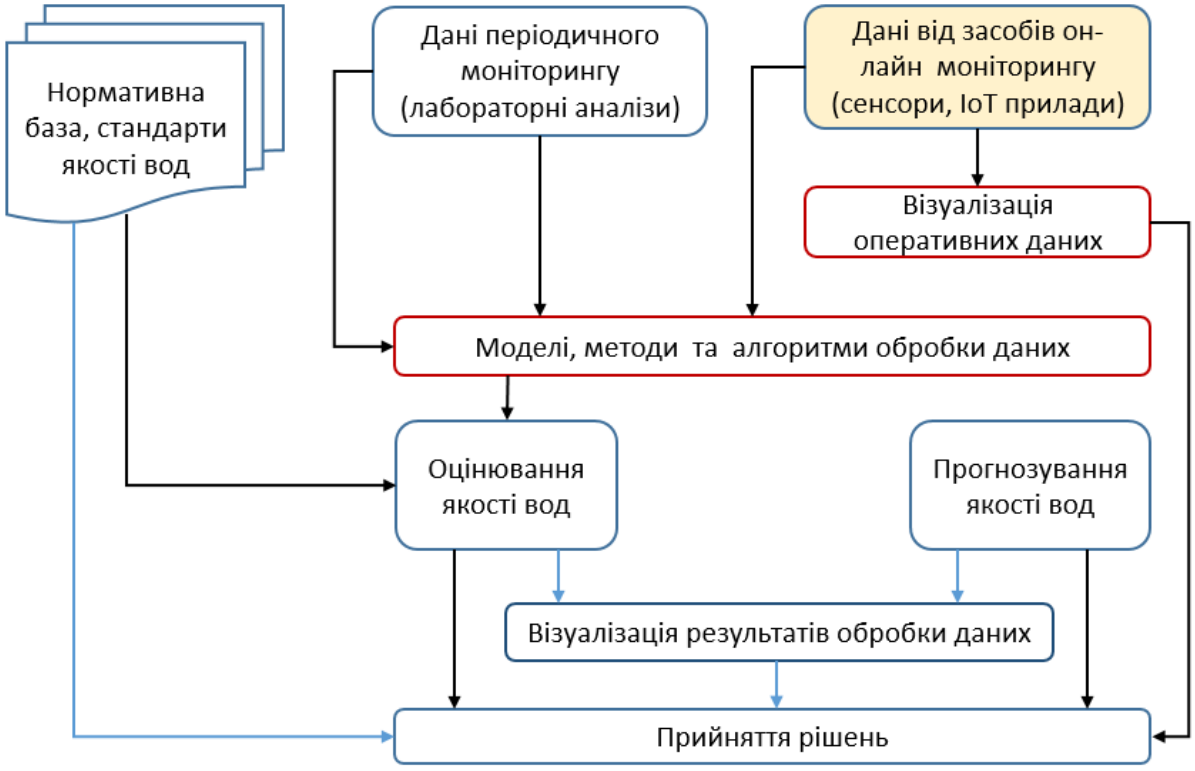


Рисунок 1.2 – Основні компоненти прийняття рішень при обробці моніторингових даних якості води

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).

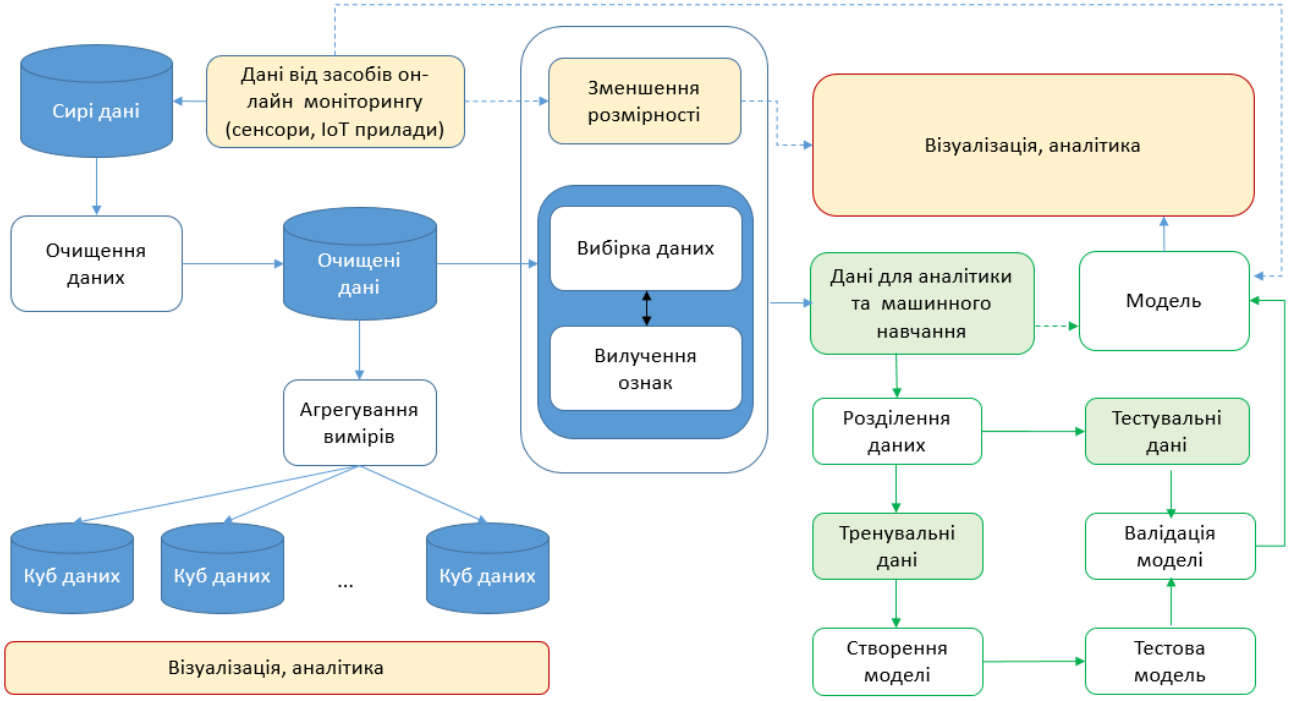


Рисунок 1.3 – Сценарії використання великих даних

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, глибоке навчання, нечіткі множини
Немарковані дані	Активне навчання
Масштабованість	Розподілене навчання, глибоке навчання
Низька достовірність, складні та зашумлені дані	Нечітка логіка, еволюційні алгоритми
Великий обсяг, різноманітність даних	Ройовий інтелект, еволюційні алгоритми, кластерний аналіз, алгоритми узгодження на основі нечіткої логіки
Великий обсяг та розмірність даних	аналіз основних компонентів, факторний аналіз, кластерний аналіз, незалежний компонентний аналіз, незалежний факторний аналіз

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; моделювання стихійних лих [15, 18], розумних водних мереж. Однак моделі запропоновані в [15], не підходять для систем моніторингу водного середовища з одним типом даних; технології глибокого навчання 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. 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] 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., і є математичним інструментом для боротьби з невизначеністю.

В 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.

1.2.2 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.

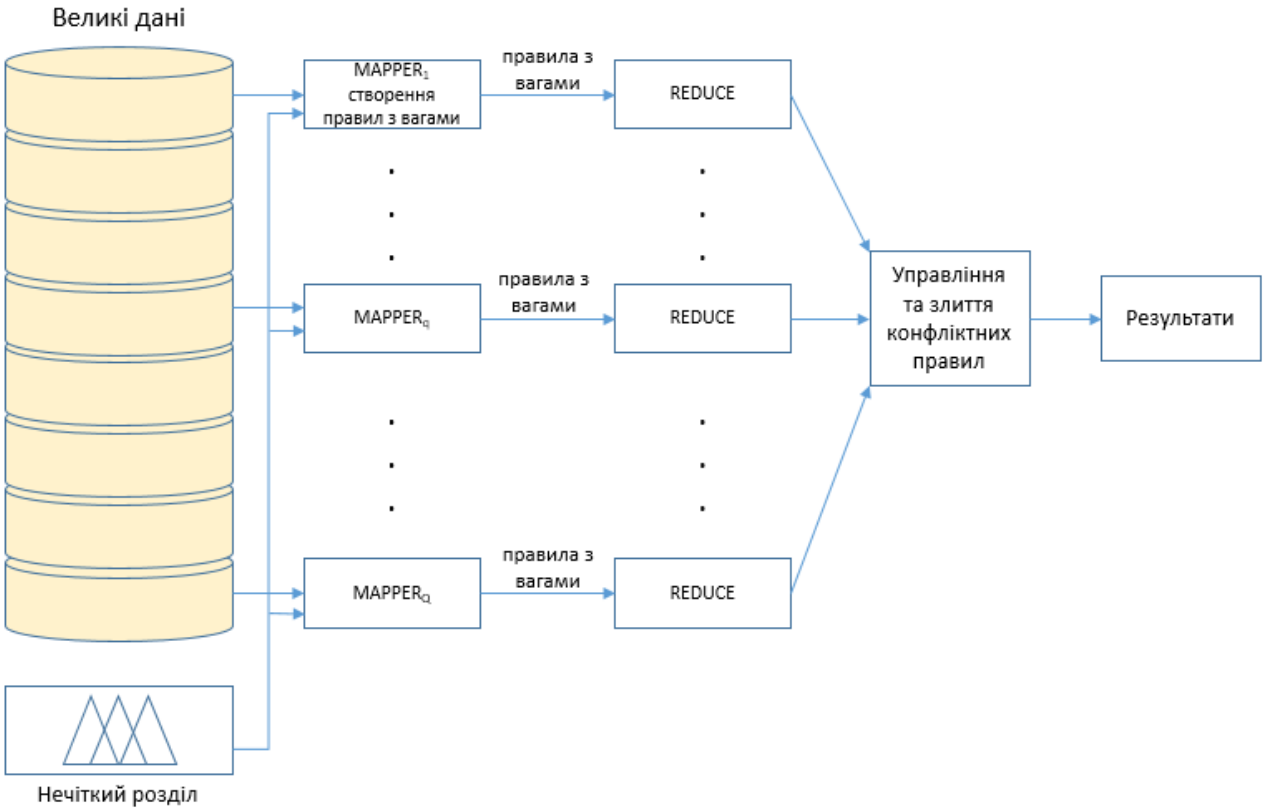


Рисунок 1.5 – Схема MapReduce з урахуванням нечіткого розділу (Адаптовано з [29, 54])

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:

– 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. 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:

– 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.

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.

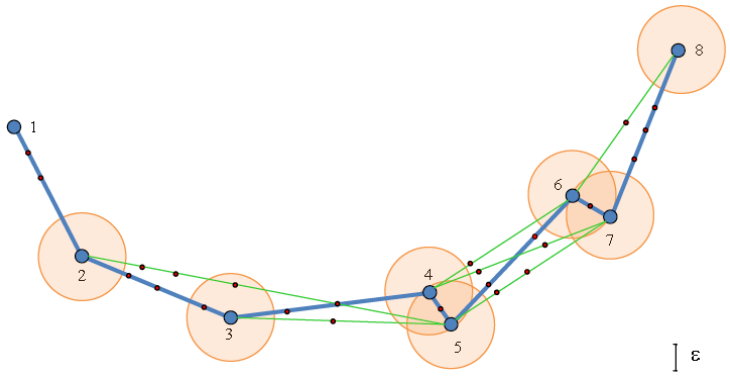


Рисунок 1.7 - Графік для спрощення рядків. Зелені краї представляють дозволені ярлики, точки перетину позначені червоним. Початкова траєкторія коштує 14, найкоротший шлях - 1-2-5-6-8 з вартістю 9. (Джерело [19])

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:

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

1.4 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 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 роботі методика досліджень, означені основні етапи, показані взаємозв’язки з отриманими результатами.

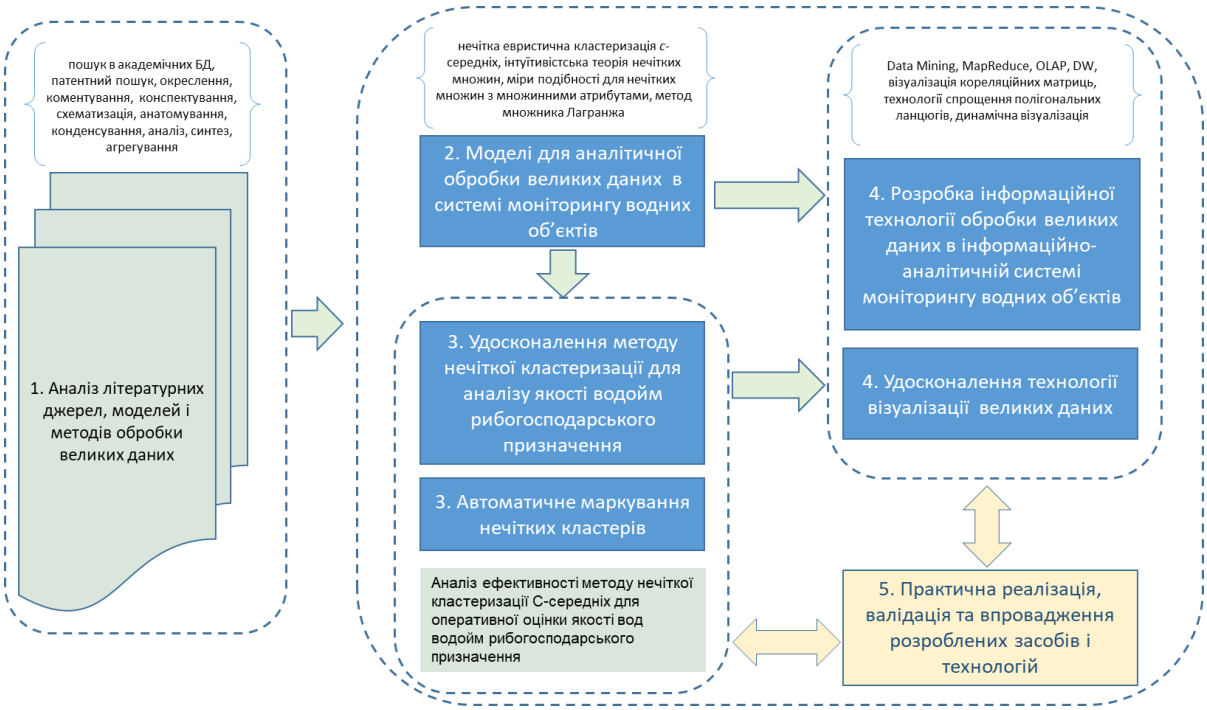


Рисунок 1.8 – Методика дослідження

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 с-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.

Висновок 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 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, 9, 10].

Список літератури до розділу 1

1. 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.

2. 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.

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

4. 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.

5. 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.

6. 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.

7. 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.

8. 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.

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

10. 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.

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

12. 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.

13. 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.

14. 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/

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

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

17. 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.

18. 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

19. 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

20. 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

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

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

23. 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.

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

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

26. 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

27. 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

28. 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.

29. 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.

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

31. 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.

32. 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.

33. 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.

34. 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.

35. 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.

36. 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.

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

38. 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.

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

40. 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.

41. 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.

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

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

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

45. 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.

46. 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.

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

48. 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.

49. 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.

50. 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.

51. 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.

52. 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.

53. 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.

54. 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.

55. 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.

56. 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.

57. 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.

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

59. 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.

60. 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.

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

62. 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.

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

64. 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

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

66. 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 досліджені (1) увагу сконцентровано на якість вод рибогосподарського призначення, (2) змість 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 рибогосподарського 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 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;

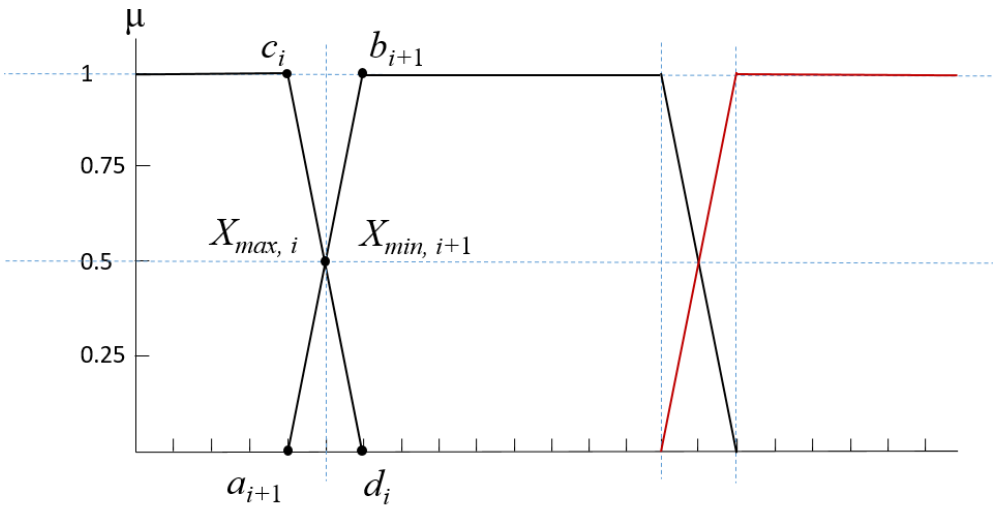


Рисунок 2.1 – Схема розбиття на межі інтервалів $X_i = [X_{min,i}, X_{max,i}]$ і $X_{i+1} = [X_{min,i+1}, X_{max,i+1}]$

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.

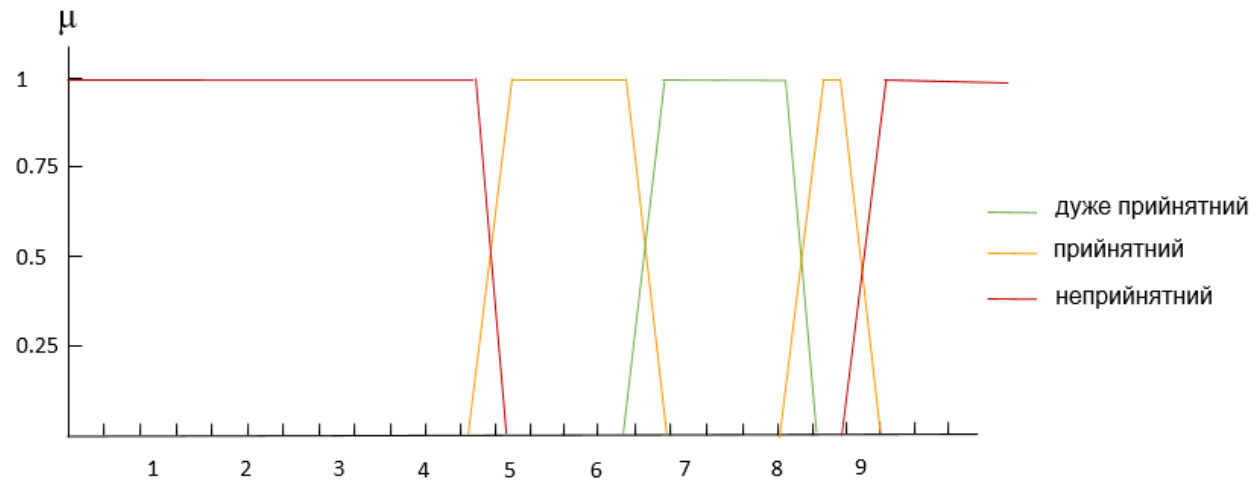


Рисунок 2.2 – Графік функцій термів лінгвістичної змінної P_1 (pH)

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.

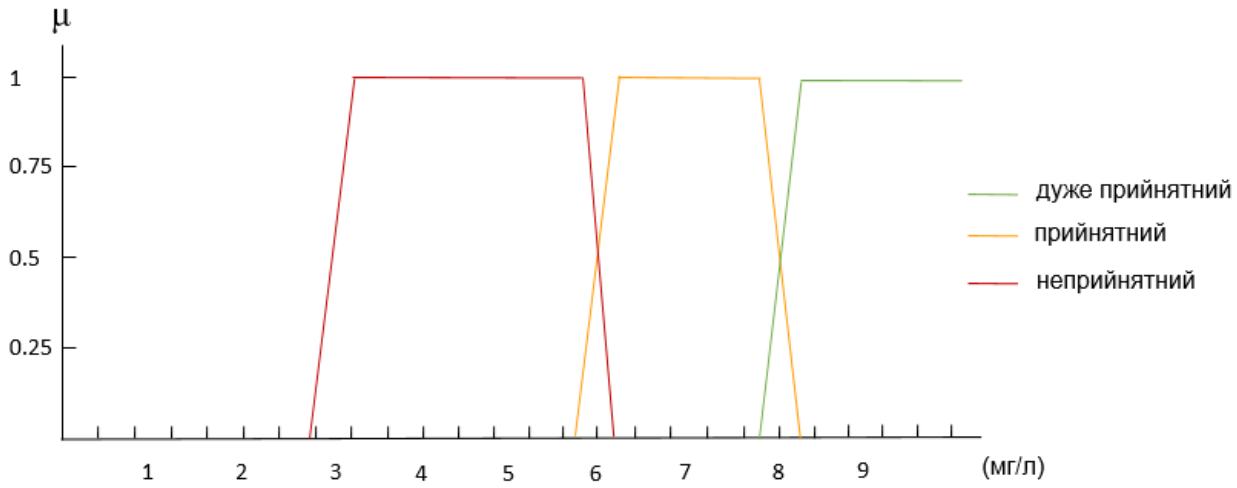


Рисунок 2.3 – Графік функцій термів лінгвістичної змінної P_2 (РК)

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.

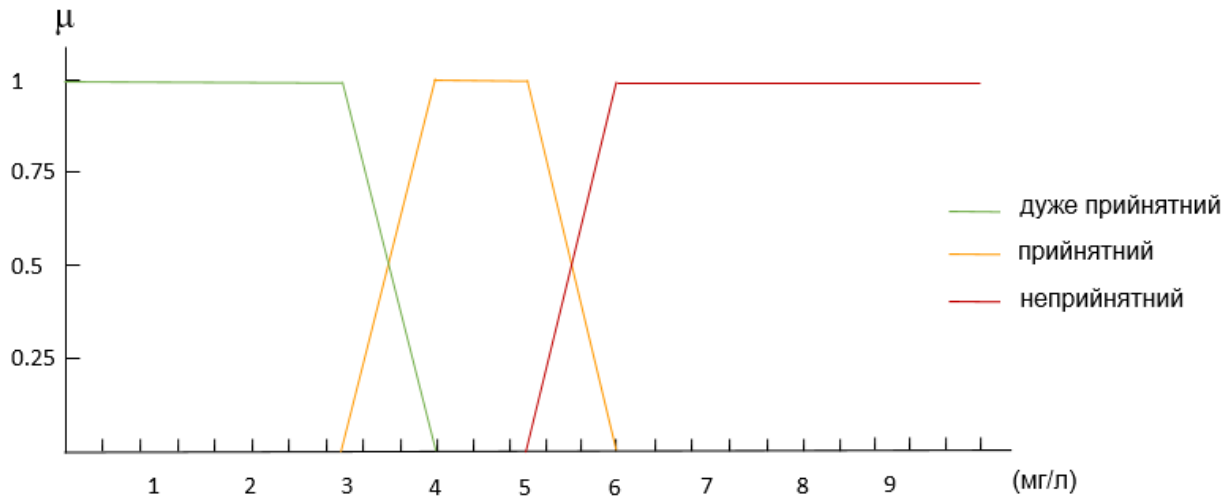


Рисунок 2.4 – Графік функцій термів лінгвістичної змінної P_3 (БСК)

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.

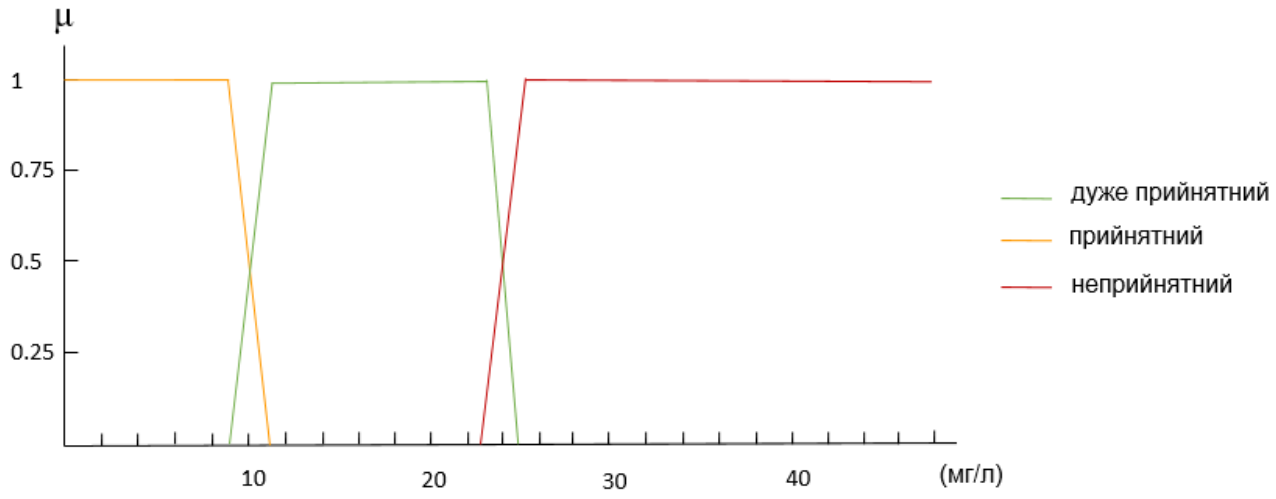


Рисунок 2.5 – Графік функцій термів лінгвістичної змінної P_4 (ХСК)

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.

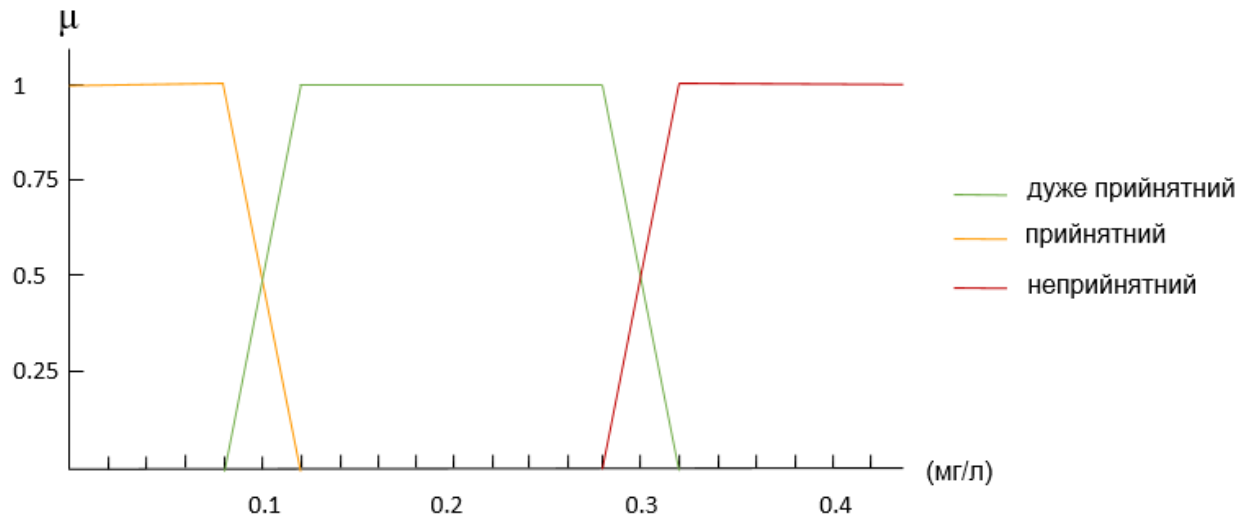


Рисунок 2.6 – Графік функцій термів лінгвістичної змінної P_5 (аміак)

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.

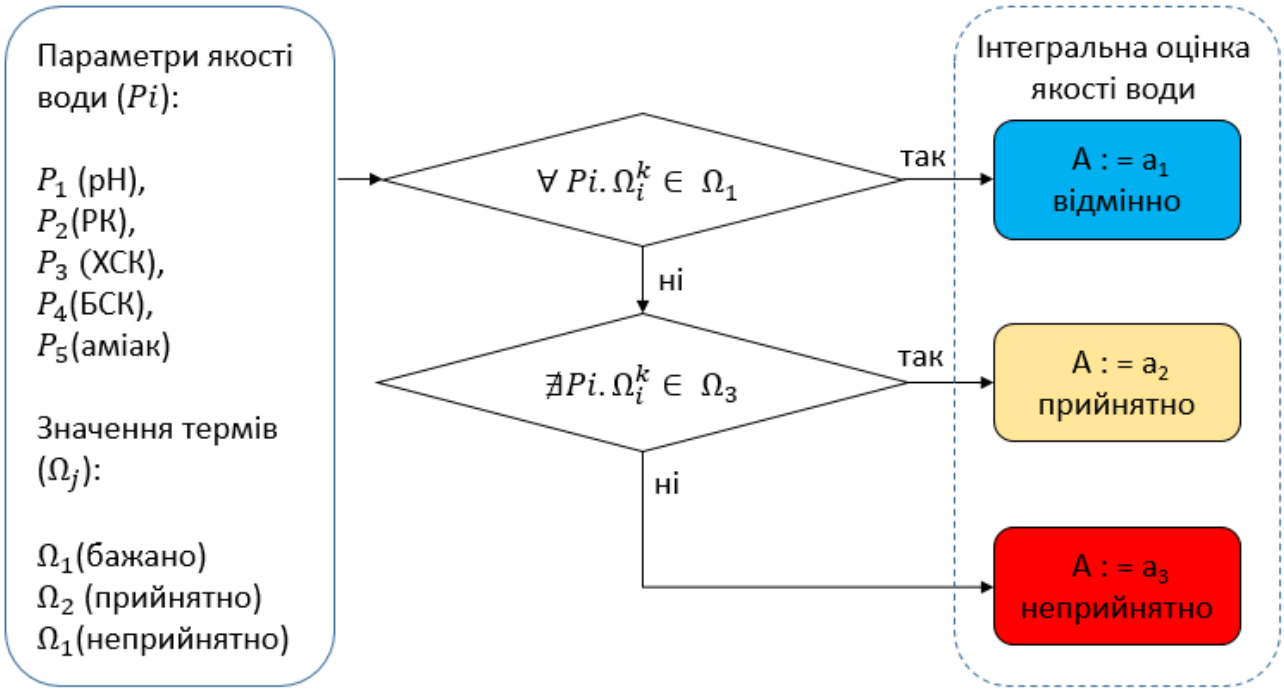


Рисунок 2.7 – Узагальнена структура технології використання нечітких правил для оцінки якості вод рибогосподарського призначення

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).

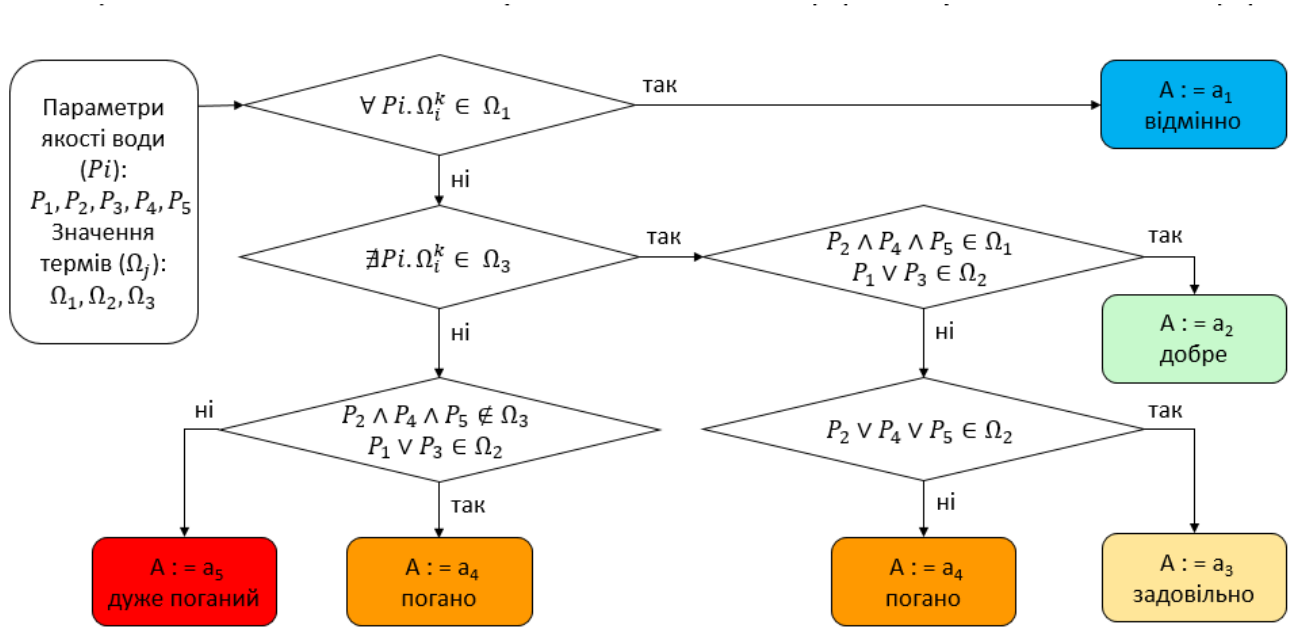


Рисунок 2.8 – Розширена структура технології використання нечітких правил для оцінки якості вод рибогосподарського призначення

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 наступну 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.

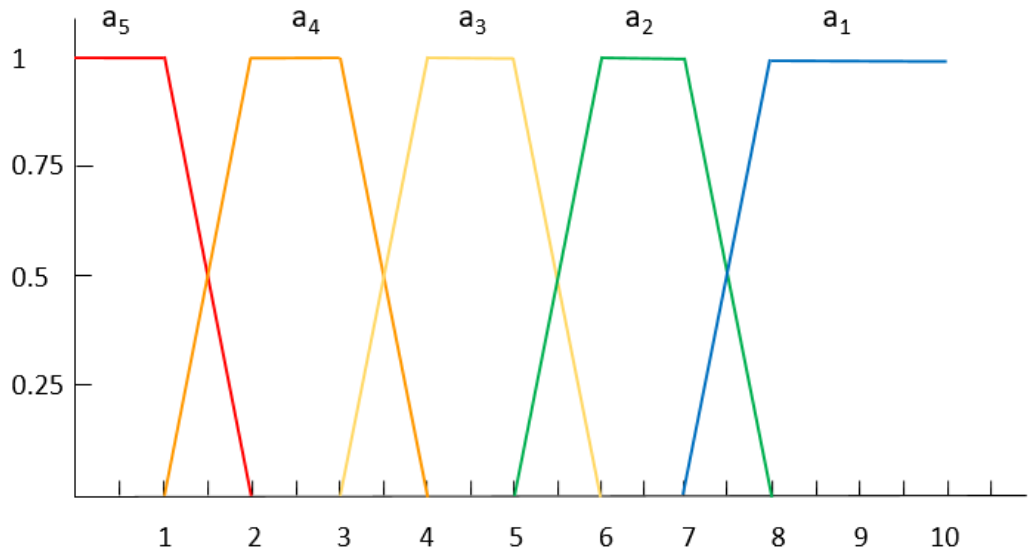


Рисунок 2.9 – Графік функцій термів вихідної лінгвістичної змінної А

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.

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

Третій 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.

3.1 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. 3.1.

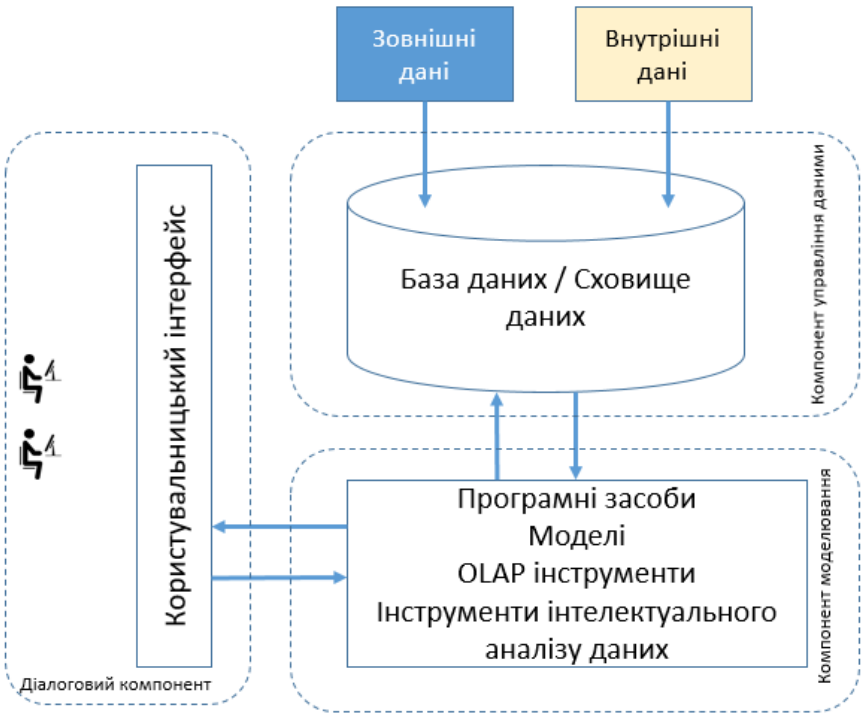


Рисунок 3.1 – Узагальнена структура СППР

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. 3.2 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.

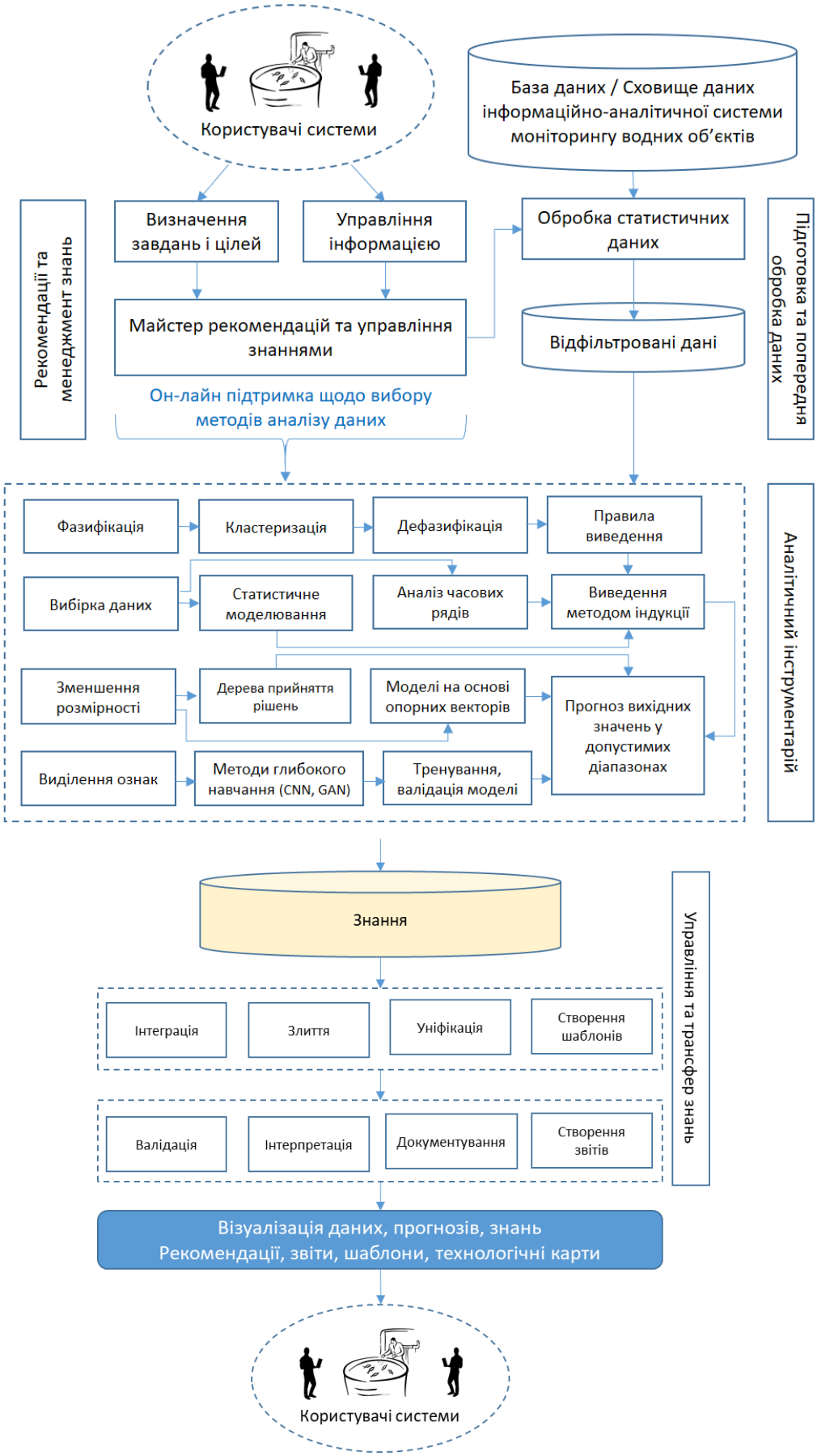


Рисунок 3.2 – Архітектура СППР

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.

3.1.1 Архітектура сховища даних

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

У роботі використано типову архітектуру зберігання даних, запропоновану в [23], яка проілюстрована на рис. 3.3, що має чотири окремі модулі: (1) сирі дані (джерела даних), (2) трансляція – трансформація – видобуток (ETL), (3) інтегрована інформація; та (4) обмін даними.

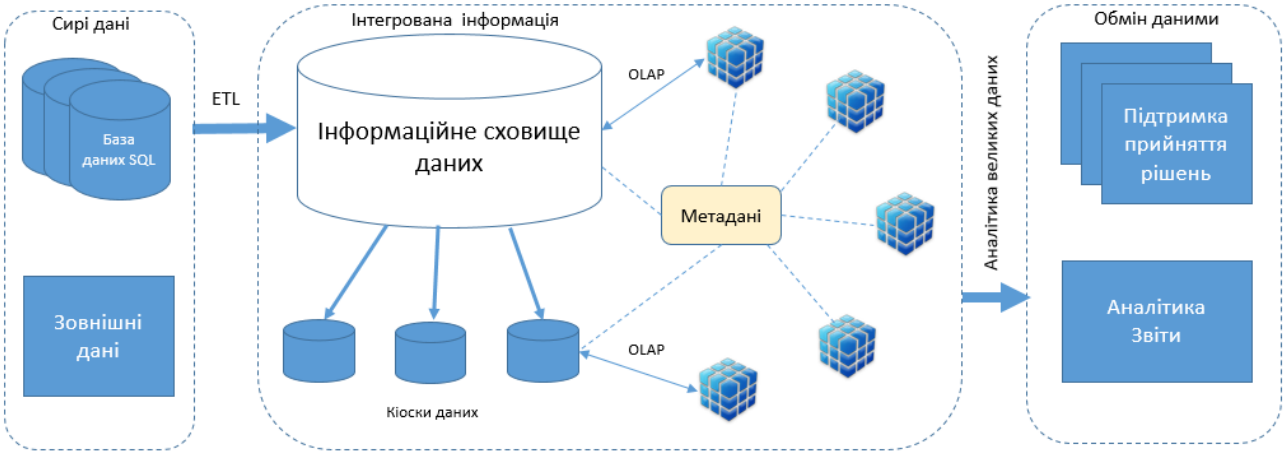


Рисунок 3.3 – Архітектура сховища даних

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. 3.2, 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. 3.4). 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.

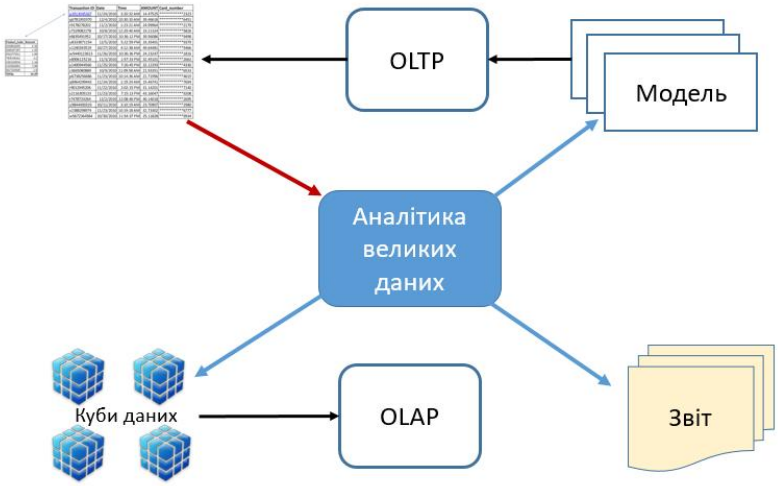


Рисунок 3.4 – Схема обробки даних

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 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. 3.1, 3.4). 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. 3.5).

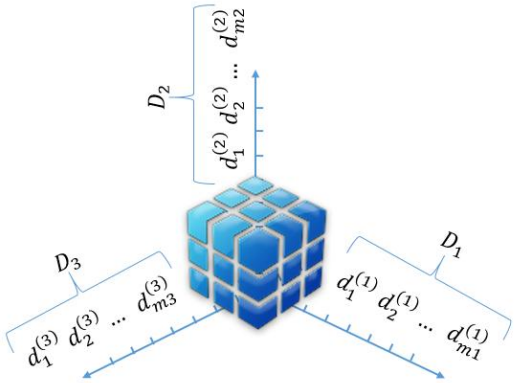


Рисунок 3.5 – Інтуїтивне подання куба даних

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.

3.1.3 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. 3.6 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. 3.6 описана частина схеми сузір'я для систем моніторингу водних об’єктів та підприємств аквакультури, яка включає чотири таблиці фактів та 17-мірні таблиці.

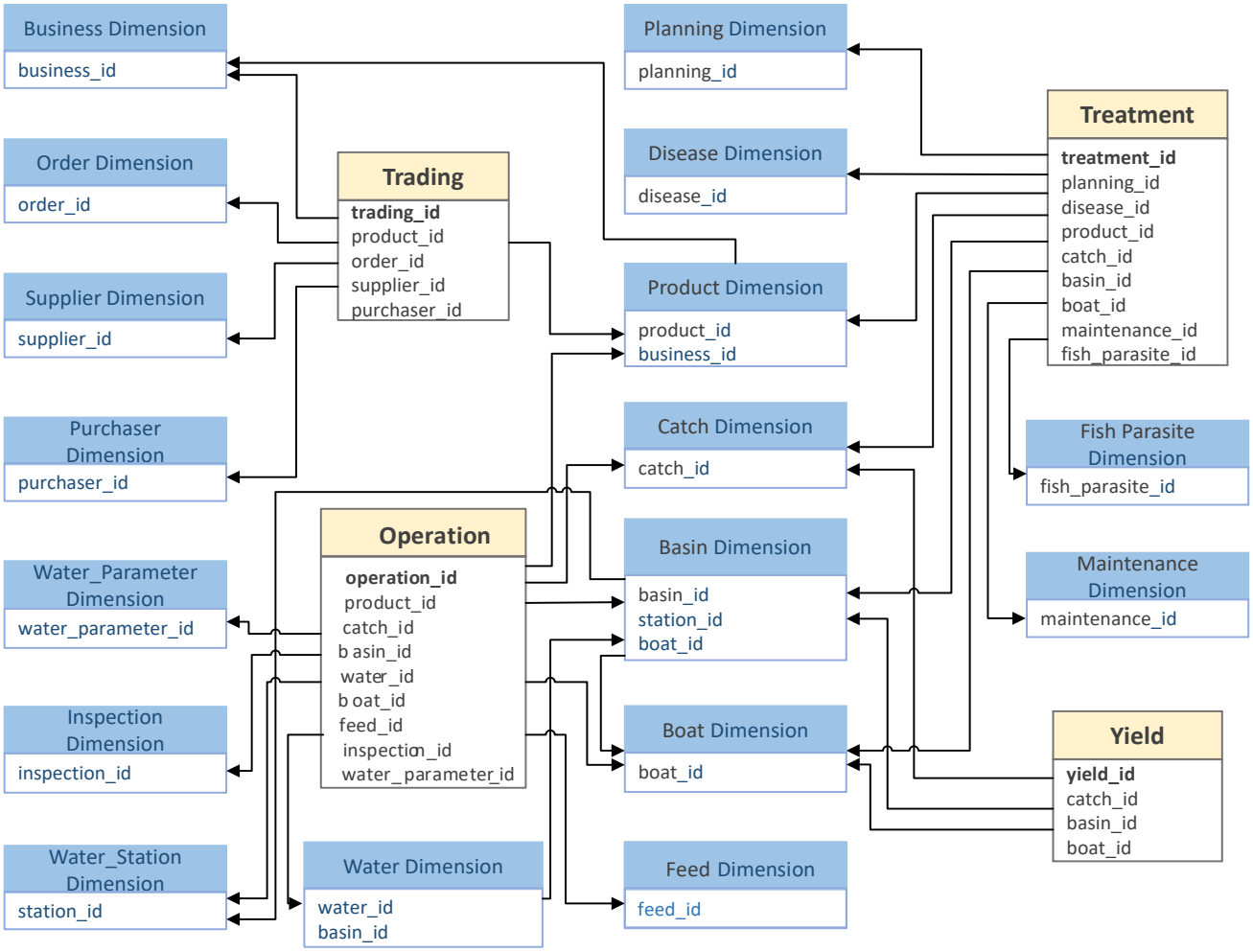


Рисунок 3.6 – Фрагмент схеми сховища даних інформаційно-аналітичної системи моніторингу водних об’єктів та підприємств аквакультури

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 3.1 описує особливості атрибутів 17-ти розмірних таблиць, які представлені на рис. 3.6 та виміри СД системи моніторингу водних об’єктів підприємства аквакультури.

Таблиця 3.1 – 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".

3.2 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 дослідження.

Таким чином, особливості процесу обробки великих даних отриманих від систем моніторингу водних об’єктів визначили наступну структуру сховища даних (рис. 3.7). З огляду на вимоги до подання та обробки даних, представлені в п. 3.1.2, 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;

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.

Вхідні дані для сховища даних, використовувані в 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.

3.4 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 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 3

Третій 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.

Список літератури до розділу 3

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 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 роботи 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, 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 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 роботи апробовано та впроваджено в Управлінні Державного агентства рибного господарства у Луганській області.

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

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 Б

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).