Лекція 1

Обробка даних у моніторингу навколишнього середовища майже в реальному часі: система та огляд вибраних методів

1 . вступ

Дані моніторингу якості води в режимі реального часу є цінними для побудови інноваційних досліджень, які реагують на динамічні часові зміни, таких як прогнозування якості води, оцінка якості води та управління навколишнім середовищем [1] . Високочастотні вимірювання дозволяють вченим отримати повне розуміння динамічних часових коливань, які зазвичай не можуть бути виявлені традиційними методами вибірки. Однак проблема відсутності даних є повсюдною при моніторингу якості води в режимі реального часу, що часто спричинено несправністю обладнання, покриттям мережі або пошкодженням даних [2] .

Відсутні дані є поширеною проблемою в моделюванні на основі даних. Відсутність даних може спричинити упередження в статистичному аналізі, що призведе до недійсних висновків [3] . Крім того, втрачені дані роблять багато методів моделювання даних неефективними, оскільки вони передбачають повну інформацію для всіх включених змінних [4] . Отже, терміново потрібні ефективні способи обробки відсутніх даних.

Загальновідомі методи роботи з відсутніми значеннями варіюються від пропуску даних до складних алгоритмів імпутації [5] . Метод пропуску даних відкидає зразки з відсутніми значеннями з подальшого аналізу. Хоча його легко застосувати, він різко зменшує ефективний розмір вибірки . Крім того, видалення зразків спричинить переривчасті дані часових рядів, що принесе більше труднощів при аналізі тимчасової інформації.

Статистичний аналіз є ще одним широко використовуваним підходом для оцінки відсутніх даних датчиків якості води. Кабір та ін. [6] застосували методи середнього, медіанного та лінійного імпутації для оцінки відсутніх даних із водорозподільної мережі міста Калгарі. Їх експерименти демонструють, що середнє та медіанне імпутування мають тенденцію недооцінювати дисперсію даних. Среботняк та ін. [7] запровадив метод імпутації гарячої колоди для покращення бази даних якості води Європейського агентства з навколишнього середовища. Згідно з їхнім підходом, знання області якості води має вирішальне значення для визначення повних зразків, які точно відповідають відсутнім зразкам. Загалом, для статистичних методів імпутації інтеграція знань про певну область у процес імпутації є важливою для досягнення багатообіцяючої продуктивності.

Величезна кількість даних датчиків якості води відкриває нову можливість для відкриття на основі даних. На відміну від моделей, заснованих на процесах, які базуються на добре встановлених математичних або фізичних законах, моделі, керовані даними, будують зв’язки між змінними стану системи без явного знання фізичної поведінки системи [8] . Ратолоджанахарі та ін. [9] оцінили різні методи імпутації, такі як K-Nearest Neighbors (KNN), Random Forest (RF) і Multivariate Imputations by Chained Equations (MICE) для обробки високочастотних пропусків у наборі даних якості води, зібраних у Франції. Експериментальні результати

продемонстрували, що гібридизація кількох алгоритмів машинного навчання може досягти кращої продуктивності, ніж вихідний MICE окремо. Кім та ін. [10] порівнювали такі методи імпутації, як нейронна мережа прямого зв'язку та карта самоорганізації в оцінці спостережень за течією річки Таехва, Корея . Методи, засновані на машинному навчанні, демонструють багатообіцяючу продуктивність в обробці подій з великим потоком. Однак більшість методів, керованих даними, розглядають дані датчика якості води як послідовність числових значень. Враховуючи, що багато змінних якості води мають передбачувану часову мінливість, ігнорування тимчасової інформації може значно знизити точність імпутації.

Нейронні мережі з рекурентними одиницями широко використовуються в обробці даних часових рядів завдяки їх здатності демонструвати часову динамічну поведінку. Багато досліджень застосовували рекурентну нейронну мережу , таку як довготривала короткочасна пам'ять (LSTM) або Gated Recurrent Unit(GRU), для оцінки даних часових рядів. Чжан та ін. [11] розробили мережу імпутації на основі GRU та залишкового швидкого підключення. Експериментальні результати показують, що модель забезпечує вищу точність імпутації відсутніх даних, ніж базові методи. Верма та Кумар [12] запропонували точний метод прогнозування відсутніх даних на основі моделі LSTM. Модель, заснована на LSTM, працює добре порівняно з лінійною регресією таГаусана наборі даних охорони здоров'я.

Більшість сучасних досліджень розробляють методи імпутації для конкретних змінних якості води. Експериментальні результати вказують на кращу ефективність запропонованих методів за певних обставин. Однак важко визначити, який метод імпутації стабільно працює найкраще у широкому спектрі прикладних сценаріїв [13] . Зокрема, більшість тестів систематично не описують робочий процес для обробки реалістичних даних моніторингу , що ускладнює проведення повторюваних експериментів для оцінки різних методів імпутації.

На відміну від попередніх оглядових досліджень, ми спочатку розробляємо нову систему імпутації даних, яка забезпечує високу сумісність у реалізації різних алгоритмів імпутації даних. Потім, використовуючи цю систему, ми оцінили вибрані алгоритми імпутації на основі даних двох систем моніторингу якості води в режимі реального часу . Основні внески цієї статті підсумовані таким чином.

- 1.

Ми розглядаємо найсучасніші методи імпутації, аналізуємо їхні переваги та обмеження та обговорюємо вибір методу для імпутації даних про якість води. Крім того, ми оцінюємо ці вибрані методи на двох наборах даних про якість води, зібраних з Америки та Австралії.

- 2.

Ми розробляємо нову хмарну систему імпутації відсутніх даних, яка може працювати з різними алгоритмами імпутації. Розроблена система здатна відновлювати відсутні дані майже в реальному часі. Він підтримує обробку кількох потоків моніторингу якості води одночасно.

Решта цієї статті організована таким чином. Розділ 2 розглядає відповідну роботу та мотивує дослідження. Розділ 3 представляє статистичні концепції проблеми імпутації. Розділ 4 охоплює детальний опис та компоненти запропонованої

системи імпутації. Розділ 5 представив вибрану кількість алгоритмів імпутації. Розділ 6 показує валідність, а також штрафні санкції вибраних алгоритмів імпутації. Розділи 7 Експериментальні випадки , 8 Результати оцінюють ефективність імпутації вибраних алгоритмів. Нарешті, Розділ 9 завершує статтю.

2 . Сродна робота

Послідовні пропуски вимірювань знижують якість і ефективність моніторингу навколишнього середовища в реальному часі та ефективність аналізу даних. Для обробки відсутніх даних у моніторингу навколишнього середовища було застосовано низку методів імпутації даних щодо різних змінних якості води.

Чен та ін. [14] запропонував TrAdaBoost-LSTM, який може фіксувати довгострокові залежності між часовими рядами та використовувати відповідні знання з повних наборів даних для заповнення послідовних відсутніх даних. Результати показують, що запропонований метод покращує точність імпутації приблизно на 20% порівняно з альтернативними тестами. Ламріні та ін. [15] застосували методи на основі самоорганізуючої карти (SOM) для реконструкції відсутніх даних під час обробки питної води . Експериментальні результати показали ефективність і надійність алгоритму SOM. Сребotняк та ін. [7] пояснює мотивацію та методологію індексу екологічної ефективності (EPI) індексу якості води (WATQI) і застосовує методи гарячої деки для приписування відсутніх WATQI у ширших географічних регіонах . Результати імпутації розширюють початковий WATQI на 39 країн до 131 країни, таким чином збільшуючи географічне охоплення на 42%. Хоча розроблено різні методи для заповнення відсутніх даних про якість води, вони оцінюються лише за конкретними змінними якості води. Отже, дуже потрібний систематичний спосіб впровадження цих методів імпутації у великомасштабні дані моніторингу .

Крім того, багато досліджень зосереджені на порівнянні різних методів імпутації даних про якість води. Наприклад, Ratolojanahary et al. [9] об’єднав MICE (багатомірні імпутації за допомогою ланцюжкових рівнянь) з випадковим лісом (RF), посиленими регресійними деревами (BRT), K-найближчими сусідами (KNN) і опорною векторною регресією (SVR), щоб вирішити проблему імпутації даних у контекст оцінки якості води . Результати показали, що MICE-SVR є найкращим у тому, що він сходиться швидше, ніж три інші, і забезпечує найкращу продуктивність. Betrie та ін. [16] порівнювали три методи імпутації, такі як ітераційна надійна імпутація на основі моделі (IRMI), багаторазова імпутація неповних багатовимірних даних (AMELIA) і послідовна імпутація для відсутніх значень (IMPSEQ) щодо даних про якість води, що наповнюється, зібраних із мідь-молібден-золота – срібно-ренієвий рудник. Результати показали, що IMPSEQ та IRMI придатні для визначення відсутніх значень у базах даних якості води на шахтах, тоді як AMELIA ні. Tabari та Talaee [17] перевірили ефективність мереж багатошарового персептрона (MLP) і радіальної базисної функції (RBF) для відновлення відсутніх значень 13 параметрів якості води на основі даних з п’яти станцій, розташованих уздовж річки Марун, Іран. Було також виявлено, що моделі MLP перевершують моделі RBF для реконструкції відсутніх даних про якість води. Наскільки нам відомо, більшість порівняльних досліджень не охоплюють розширені моделі імпутації на основі нейронних мереж. Враховуючи, що

моделі глибокої нейронної мережі перевершують традиційні методи імпутації в багатьох дослідженнях [2] , [18] , [19] , [20] , [21] , ігнорування цього типу методу не може забезпечити повну оцінку продуктивності для імпутації даних про якість води за реалістичних умов. випадки використання.

На відміну від попередніх досліджень, у цій статті ми пропонуємо хмарну систему імпутації відсутніх даних, яка може підтримувати різні алгоритми імпутації. Це пропонує користувачеві системний підхід до виконання завдань імпутації на великомасштабних даних моніторингу якості води . Крім того, ми розглядаємо вибрану кількість методів імпутації, які охоплюють рішення на основі статистичних даних, моделей на основі даних і нейронних мереж.

3 . Проблема з відсутністю даних про якість води

На якість статистичної аналітики може сильно вплинути частка відсутніх даних [22] . Відповідно до дослідження, запропонованого Рубіном [23] , відсутні дані часто класифікують на такі три типи:

- •

Зниклий випадково (MCAR). Якщо кожне вимірювання в наборі даних має однакову ймовірність бути відсутнім, набори даних визначено як відсутні випадковим чином. Це означає, що причини відсутності даних не пов’язані з даними. MCAR є ідеальним припущенням, але воно рідко зустрічається на практиці.

- •

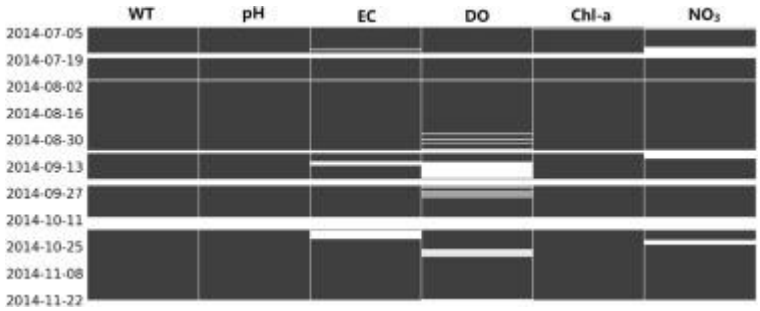
Випадково зниклі (MAR). Припустімо, що лише групи вимірювань у наборах даних мають однакову ймовірність бути відсутніми, а спостережувані дані визначають цю ймовірність. У цьому випадку ми випадково визначаємо набір даних відсутнім. MAR є більш загальним і реалістичним припущенням, ніж MCAR. Згідно з цим припущенням, відсутність може бути змодельовано за допомогою спостережених даних.

- •

Пропущені не випадково (MNAR). Це відноситься до випадку, коли ні MCAR, ні MAR не виконуються. Коли набір даних є MNAR, той факт, що дані відсутні, систематично пов’язується з даними, які не спостерігаються. Важко впоратися з цим відсутнім типом даних , оскільки для цього знадобляться сильні припущення щодо моделей відсутності.

Варто зазначити, що відсутність даних про якість води, як правило, відповідає механізму MAR [24] . Отже, доцільно оцінити відсутню інформацію про якість води, застосовуючи різні аналітичні підходи та підходи до моделювання.

На рис. 1 показано, як бракує даних на станції East Russell River у Північному Квінсленді, Австралія [25] . Ця станція є частиною програми моніторингу водозбірного басейну Великого Бар’єрного рифу , яка буде описана в Розділі 7.1 . Як видно, кожна змінна мала дані моніторингу. Серед цих змінних деякі змінні, такі як DO,HI3і EC мають більшу кількість відсутніх даних, ніж інші змінні. Відсутність послідовної кількості даних між декількома змінними створює значні труднощі для точної оцінки даних моніторингу якості води.



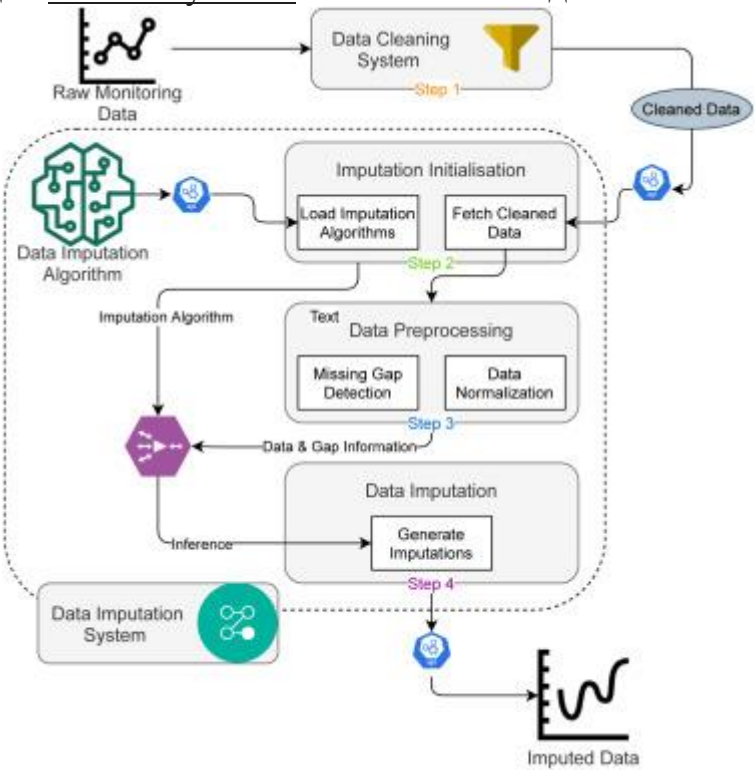
1. Завантажити: завантажити зображення високої роздільної здатності (87 КБ)
2. Завантажити: завантажити повнорозмірне зображення

Рис. 1 . Приклад відсутності даних у системі моніторингу якості води. Ця система вимірює сім змінних якості води, таких як температура води (WT), pH, електропровідність (EC), розчинений кисень (DO), хлорофіл-a (Chl-a) і нітрати (HI3). Проміжок while і сірий блок представляють відсутні та доступні дані.

4 . Система імпутації даних

У цій статті ми розробили та впровадили нову систему імпутації відсутніх даних. На відміну від попередніх досліджень, розроблених для конкретних алгоритмів імпутації [26] , [27] , ми абстрагувалися від критичних етапів обробки в імпутації даних і побудували систему на основі принципів модульного проектування (рис. 2).

Система використовує PyTorch як серверний механізм для моделей глибокого навчання та пакет імпутації Python, такий як impyute [28] , як серверний механізм для алгоритмів без глибокого навчання . Система розроблена на Amazon Cloud для масштабування сотень потоків даних моніторингу навколишнього середовища.



1. Завантажити: завантажити зображення високої роздільної здатності (275 КБ)
2. Завантажити: завантажити повнорозмірне зображення

Малюнок 2 . Хмарна система імпутації даних. Система приймає вимірювання якості сирої води як вхідні дані. Він працює з системою очищення

даних для видалення очевидних викидів. Після вибору алгоритму імпутації даних і отримання очищених даних необхідний етап обробки даних для створення вхідних даних для алгоритму. Після цього модель відповідним чином заповнює відсутні прогалини.

Детальне пояснення основних кроків використання системи , зображених на рис. 2, є таким:

- 1.

Очищення даних моніторингу якості води: вхідними даними системи є необроблені дані моніторингу. Під час збору даних виникає багато шумів або інформації про помилки. Враховуючи, що алгоритми імпутації даних потребують якісних вхідних даних, ми спочатку пропускаємо вхідні набори даних через систему очищення даних. Він забезпечує основні функції очищення за допомогою фільтрації порогових значень і перевірки контрольного значення датчика . Експерти з якості води встановлюють максимальні та мінімальні порогові значення для різних змінних, щоб виключити викиди. Багато виробників датчиків, наприклад NICO [29], надають еталонне значення датчика, щоб вказати якість вимірювання. Таким чином, він також використовується для видалення недійсних вимірювань. У цій системі до кожного потоку даних додається конфігурація метаданих, і процес очищення даних запускається автоматично, коли буде зібрано достатньо вхідних даних.

- 2.

Ініціалізація системи імпутації: на цьому кроці система завантажує вибраний алгоритм імпутації даних. Кожен потік даних може застосовувати різні методи імпутації щодо його конфігурації. Крім того, очищені дані витягуються через API даних під час ініціалізації.

- 3.

Попередня обробка даних : нормалізація даних є важливою для моделювання на основі даних. Він перемасштабував різні вхідні змінні в той самий діапазон, який необхідний для вимірювань з різними одиницями. Крім того, система також позначила всі прогалини у вхідних даних. Після цього застосовується стратегія ковзного вікна для створення вхідних даних алгоритмів.

- 4.

Генерація імпутаційних даних: за допомогою вибраного алгоритму імпутаційних даних і вхідних даних можна згенерувати імпутаційне значення для кожного пропуску. За допомогою інформації про прогалини, зібраної на кроці 3, система може заповнити значення відповідних прогалин і вивести остаточні повні дані.

У наведеному вище дизайні система очищення даних гарантує відсутність викидів у необроблених даних. Це важливо для імпутації даних, оскільки викиди з екстремальними значеннями можуть сильно ввести в оману результати імпутації [30] . У системі імпутації кожен метод імпутації працює як плагін, що забезпечує гнучкість розширення системи для підтримки більш просунутих алгоритмів імпутації.

На основі цієї запропонованої системи ми можемо реалізувати різні алгоритми імпутації даних і оцінити їх ефективність відповідно до імпутованих результатів.

5 . Огляд методів імпутації

У цьому розділі ми перерахували та описали широко використовувані методи імпутації для оцінки даних якості води. Їх можна розділити на три групи: методи на основі статистики, методи на основі моделі та методи на основі нейронної мережі.

5.1 . На основі статистики

Детермінована імпутація замінює відсутні дані правдоподібними значеннями, які можна отримати шляхом заміни значень із доступних спостережуваних змінних [31] . Тут ми перерахували три популярні використовувані методи імпутації: середня імпутація, перенесення останнього спостереження (LOCF) і лінійна імпутація.

5.1.1 . Середнє приписування

Врахування середнього означає заміну відсутнього значення середнім арифметичним усіх інших доступних значень.(1)xˆ=1n∑i=1nxi

5.1.2 . Останнє спостереження перенесено

Перенесене останнє спостереження (LOCF) зазвичай використовується для роботи з відсутніми значеннями. У цьому методі відсутнє значення враховується з останнього спостереження в наборі даних. Цей метод робить нереалістичне припущення про те, що з моменту останнього виміряного спостереження взагалі немає змін [32] . Метод LOCF часто використовується для роботи з безперервною вартістю в умовах MCAR.

5.1.3 . Лінійна імпутація

Лінійна інтерполяція оцінює відсутні значення на основі суміжних доступних значень. Це бажано для оцінки постійно відсутніх даних протягом короткого інтервалу часу. Для відсутнього значеннярі, лінійна інтерполяція генерує оцінку на основі найближчих попередніх і наступних доступних значеньрчірj,(2)piˆ=pч+pj−pчxj−xч(x−xч),xч<xi<xj.dexч,xiixjпредставляють попереднє, поточне та наступне значеннях.

Лінійне імпутування є простим, швидким і потребує лише двох доступних вибірок для імпутації кожного періоду відсутніх даних. З іншого боку, точність лінійного імпутації зазвичай знижується зі збільшенням тривалості періоду відсутніх даних.

5.2 . На основі моделі

Імпутація на основі моделі спрямована на створення прогнозних моделей для кожної цільової змінної, яка містить відсутні значення. У цьому підрозділі пояснюється кілька поширених методів імпутації. Це включає в себе максимізацію очікування, багаторазове врахування за допомогою ланцюжкових рівнянь і k-найближчого сусіда.

5.2.1 . Очікування-максимізація

Максимізація очікування (EM) — це параметричний метод для визначення відсутніх значень на основі оцінки максимальної правдоподібності . EM генерує оцінені значення для відсутніх даних за допомогою етапів очікування та максимізації.

На етапі очікування відсутні дані оцінюються на основі всіх спостережених даних і поточних параметрів моделі оцінки. Математично розрахунок можна виразити так:(3)Q(θ|θi)=∫л(θ|Ю)f(Юміс|Юobc,θi)dЮмісдел(θ|Ю)це логарифмічна функція

правдоподібності повних даних,л(θ|Юobc)є логарифмічної функцією правдоподібності спостережуваних даних, if(Юміс|Юobc,θ)є прогнозним розподілом відсутніх данихθ.

На кроці максимізації очікування ймовірності повного журналу даних від попереднього кроку оцінки максимізується, щоб допомогти отримати наступне припущенняθ:(4)θi+1=argmaxθQ(θ|θi)

Цей двоетапний процес повторюється до тих пір, поки не буде досягнуто конвергенції, і можна буде остаточно оцінити відсутні дані.

5.2.2 . Кілька імпутацій за допомогою ланцюжкових рівнянь

Множинні імпутації за допомогою ланцюжкових рівнянь (MICE) є одним із принципових методів адресації відсутніх даних. Три етапи MICE описані нижче:

На початку всі відсутні значення заповнюються даними випадкової вибірки із заміною з існуючих значень. Перша змінна з відсутніми значеннямих1генерується на основі всіх інших змінниххх2,…,xk, обмежено для осіб із спостережуванимх1. Відсутні значення вх1замінюються змодельованими розіграшами з відповідного апостеріорного прогнозного розподілух1. Потім наступна змінна з відсутніми значеннямих2регресується на основі всіх інших змінниххх1,х3,…,xk, обмежено для осіб із спостережуванимх2, і використовуючи вписані значеннях1. Знову відсутні значеннях2замінюються розіграшами з апостеріорного прогнозного розподілух2. Процес застосовується до всіх інших змінних із відсутніми значеннями. Цю процедуру повторювали б кілька ходів, щоб отримати єдиний врахований набір даних.

5.2.3 . K-найближчі сусіди

K-найближчий сусід (KNN) є популярним підходом у програмах обробки даних. Він призначений для заміни відсутніх значень за допомогою k-найбільш подібних непропущених даних. Категоричне відсутнє значення приписується більшості серед k найближчих сусідів, а числове відсутнє значення заповнюється обчисленням середнього значення k найближчих сусідів.

Щоб вибрати k кількість найближчих сусідів, подібність між даними та його найближчими сусідами має бути максимальною. Для вимірювання відстані між даними A і B використовуються різні функції відстані. У більшості досліджень вибирається функція евклідової відстані . Наприклад A =(x1,x2,…,хм)і B =(p1,p2,…,рм), де m — розмірність простору ознак . Щоб обчислити відстань між точками A і B, нормалізована евклідова метрика обчислюється як:(5)відст(хі,хj)=∑стор=1п(хістор−хjстор)2п

Крім того, зазвичай використовуваним методом є відстань Мінковського (або його варіанти) наступним чином:(6)відст(хі,хj)=(∑стор=1п|хістор−хjстор|r)1rдеrє цілим невід’ємним числом, яке називається коефіцієнтом Мінковського. Відстань Мінковського розглядається як Манхеттенська відстаньr=1і як евклідова відстань, ar=2.

5.3 . На основі нейронної мережі

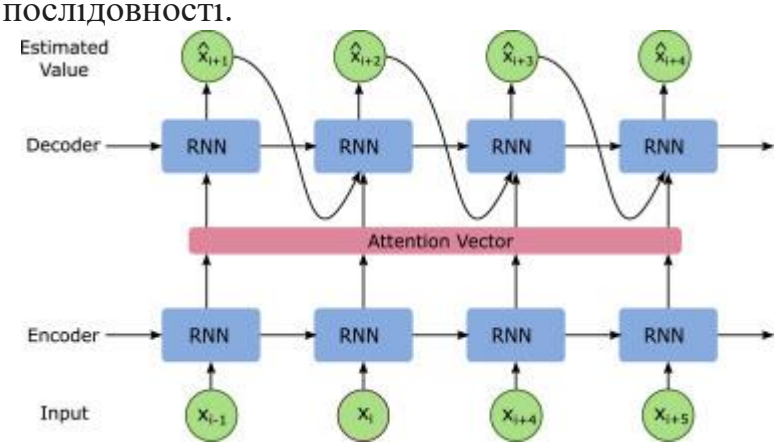
Глибинні нейронні мережі (DNN) нещодавно досягли найсучаснішої продуктивності в різних задачах розпізнавання мовлення та комп’ютерного зору . Це спонукає досліджувати застосування DNN для оцінки відсутніх даних. Тут ми

перерахували моделі імпутації даних із двома основними архітектурами: модель «послідовність до послідовності» та рекурентна нейронна мережа.

5.3.1 . Модель «послідовність до послідовності».

Навчання від послідовності до послідовності стає ефективною парадигмою роботи з входами та виходами змінної довжини. Він спрямований на пряме моделювання умовної ймовірності стор(р|х)відображення вхідної послідовності,х1,…,хп, у вихідну послідовність,р1,…,рм [33] . Цей процес виконується структурою кодера-декодера, запропонованою Чо та ін. [34] . Оскільки під час збору даних випадковим чином можуть виникати відсутні значення, важливо створити довільну кількість оціночних значень.

На рис. 3 ілюструється стандартна модель «послідовність до послідовності» для задач імпутації. На рис. 3 кодер обчислює представленнясдля кожної вхідної послідовності. На основі цього вхідного представлення декодер генерує вихідну послідовність, одну одиницю за один раз. У цьому підході умовна ймовірність розкладається як:(7)журналстор(р|х)=∑j=1млогстор(pj|p<j,x,c)дехірє вхідною та вихідною послідовностями,спозначає представлення для кожної вхідної послідовності.



1. Завантажити: завантажити зображення високої роздільної здатності (160 КБ)
2. Завантажити: завантажити повнорозмірне зображення

Рис. 3 . Модель послідовності для імпутації даних.

Коли справа доходить до проблеми імпутації, вхідними даними є послідовність доступних точок даних навколо відсутнього проміжку. Виходом моделі є оцінка значень для кожного індексу часу відсутнього розриву. У моделі послідовності до послідовності блоки рекурентної нейронної мережі (RNN) застосовуються в компонентах кодера та декодера для обробки даних часових рядів. Крім того, механізм уваги дозволяє моделі навчитися зосереджуватися на певному діапазоні вхідної послідовності для різних виходів.

- 1.

SSIM

Модель імпутації послідовності до послідовності (SSIM) [2] — це перша модель імпутації даних, заснована на архітектурі послідовності до послідовності та механізмі уваги. SSIM використовує мережу довготривалої короткочасної пам’яті (LSTM) для захоплення доступної тимчасової інформації між проміжками, а механізм глобальної уваги застосовано, щоб дозволити SSIM зосередитися на певних частинах вхідних даних для оцінки різних відсутніх значень.

- 2.

Дві SSIM

Dual-SSIM [18] розширює звичайну модель SSIM, маючи два окремих кодери для обробки вхідної послідовності навколо відсутнього проміжку. Кодери на основі Gated Recurrent Unit (GRU) можуть природним чином розділяти інформацію до та після відсутнього проміжку. Отже, не потрібно створювати додаткові вектори проміжків, щоб знайти розташування відсутніх значень. Крім того, механізм глобальної уваги розширено для підтримки обробки часових представлень, отриманих від двох різних кодерів.

5.3.2 . Рекурентна нейронна мережа

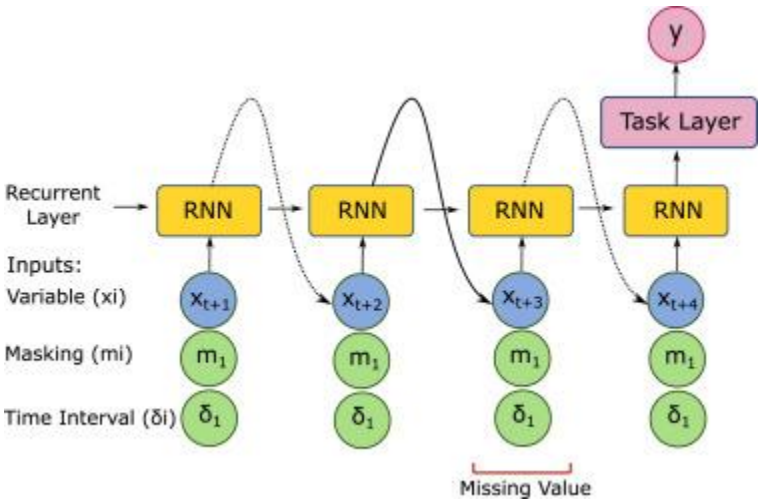
Замість прямого врахування відсутніх значень інша ідея полягає в оцінці відсутніх даних під час обчислення інших корельованих завдань прогнозування. Ця ідея широко застосовується в практичних додатках, таких як охорона здоров'я та біологія [19] . Поєднуючи процес імпутації із завданнями прогнозування, точність оцінки відсутніх значень можна покращити за допомогою посібника з корельованих завдань прогнозування.

У цьому підході рекурентний компонент і регресійний компонент працюють разом для генерації вихідних даних імпутації. Зазвичай рекурентний компонент досягається рекурентною нейронною мережею, а регресійний компонент досягається повністю пов’язаною мережею. Стандартну рекурентну мережу можна представити у вигляді:(8)чt=σ(Bччt−1+Uчxt+bч)деσє сигмоїдною функцією ,Bч,Uчibчпараметри моделі, ічtє прихованим станом попередніх часових кроків .

Для проблеми імпутації,xtможуть мати відсутні значення, тому їх не можна використовуватихtяк вхідні дані безпосередньо, як у наведеному вище рівнянні. Натомість у багатьох дослідженнях [19] , [20] , [21] використовується «комплементний» вхідxtвколихtвідсутній.xtвможна обчислити в різних стратегіях, таких як середнє значення набору даних, те саме, що його останнє вимірювання або інші або проміжні результати всередині моделі. Тут ми беремо третю стратегію як приклад. У цьому випадку вхідxtвможна розрахувати наступним чином:(9)xˆt=Bxчt−1+bx,(10)xtв=мt⊙xt+(1−мt)⊙xˆtдехˆt— це оцінений вектор на основі прихованого станучt−1.⊙представляє поелементне множення.mtвказує, чи відсутні вхідні дані на кроці часуt.Bx,bxє параметрами моделі.

На рис. 4 зображено налаштовану рекурентну модель нейронної мережі для обробки часових рядів із відсутніми значеннями. У цьому підході вхідні дані з відсутніми значеннями використовуються для навчання моделі для завдання прогнозування. Щоб отримати результати прогнозу, модель повинна оцінити відсутні значення як проміжні результати.

Для того, щоб допомогти моделі визначити прогалини у вхідних даних, додатковий часовий інтервал і вектори маскування повинні бути надані як додаткова інформація. Вектор інтервалу часу призначений для вимірювання відстані для кожної змінної з моменту її останнього спостереження. Маскування вектормзастосовується, щоб вказати, які змінні відсутні на кроці часуt.



1. Завантажити: завантажити зображення високої роздільної здатності (195 КБ)
2. Завантажити: завантажити повнорозмірне зображення

Рис. 4 . Індивідуальна рекурентна модель нейронної мережі для імпутації даних.

- 1.

БРИТАНЦІ

BRITS [20] — це рекурентний метод на основі нейронної мережі для імпутації відсутнього значення в даних часових рядів. Шляхом адаптації рекурентної нейронної мережі BRITS розглядає відсутні значення як змінні обчислювального графіка та оновлює оцінки під час процесу зворотного поширення . Таким чином, інформація про градієнти як у прямому, так і в зворотному напрямках використовується разом для оновлення відсутнього значення, що призводить до більш точної оцінки.

BRITS використовує процедуру імпутації на основі даних для оцінки відсутніх даних. Для використання BRITS завдання імпутації даних і завдання прогнозування зазвичай виконуються спільно. Експерименти демонструють, що ця стратегія може значно підвищити точність імпутації та точність завдання класифікації.

- 2.

M-PHH

M-RNN [21] — це багатонаправлена рекурентна нейронна мережа . Він використовує інформацію в одному потоці даних для інтерполяції даних, а також вносить відсутні значення в потоки даних. Він містить блок інтерполяції та блок імпутації. Блок інтерполяції створює функцію інтерполяції , яка працює в потоці даних. Блок імпутації створює функцію імпутації, яка працює між потоками. M-RNN перевершує кілька контрольних показників у п’яти реальних наборах медичних даних.

6 . Переваги та обмеження

У цьому розділі ми підсумували основні відносні переваги та обмеження цих методів імпутації, а також їх придатність для цілей моделювання (див. таблицю 1).

- 1.

Статистичні методи

При використанні статистичних методів пропущені значення замінюються значенням, визначеним певним правилом. Цей підхід є обчислювально простим,

але ігнорує зв’язок між змінними в наборах даних. Таким чином, він часто недооцінює мінливість, оскільки кожне неспостережене значення має таку ж вагу в аналізі, як і відомі спостережувані значення [37] . Крім того, деякі статистичні методи припускають, що всі відсутні дані дотримуються постійної моделі. Наприклад, усі вони близькі до середнього значення (середня імпутація) або попереднього доступного значення (LOCF). Таким чином, методи, засновані на статистиці, часто є потенційно упередженими і повинні використовуватися з великою обережністю [38] .

- 2.

Методи на основі моделі

Імпутація на основі моделі враховує зв’язок між різними змінними шляхом побудови регресійних моделей для відсутніх ознак, які беруть невідсутні ознаки як вхідні дані [39] . Однак він має значні обчислювальні витрати. По-перше, витрати на виконання моделі на великомасштабних наборах даних можуть бути непомірно високими. По-друге, це значною мірою залежить від типу та характеру даних і не може використовуватися як готовий етап попередньої обробки.

- 3.

Методи на основі нейронних мереж

Нейронні мережі з рекурентною архітектурою мають здатність фіксувати довгострокові тимчасові залежності та спостереження змінної довжини, які не можуть підтримуватися іншими технологіями імпутаційного моделювання. Однак моделі на основі нейронних мереж мають такі обмеження: (1) через багатошарову нелінійну структуру моделі нейронних мереж часто критикують за те, що вони непрозорі, а результати не відстежуються людьми [40] . (2) зі швидким збільшенням обсягів даних відповідно збільшується час, необхідний для навчання нейронної мережі [41] . (3) порівняно з традиційними алгоритмами машинного навчання глибоке навчання сильно залежить від налаштування гіперпараметрів [42] . Отже, щоб отримати надійні рішення імпутації, необхідно отримати розуміння механізму нейронної мережі.

Таблиця 1 . Сильні сторони та обмеження методів, перелічених у розділі 5 .

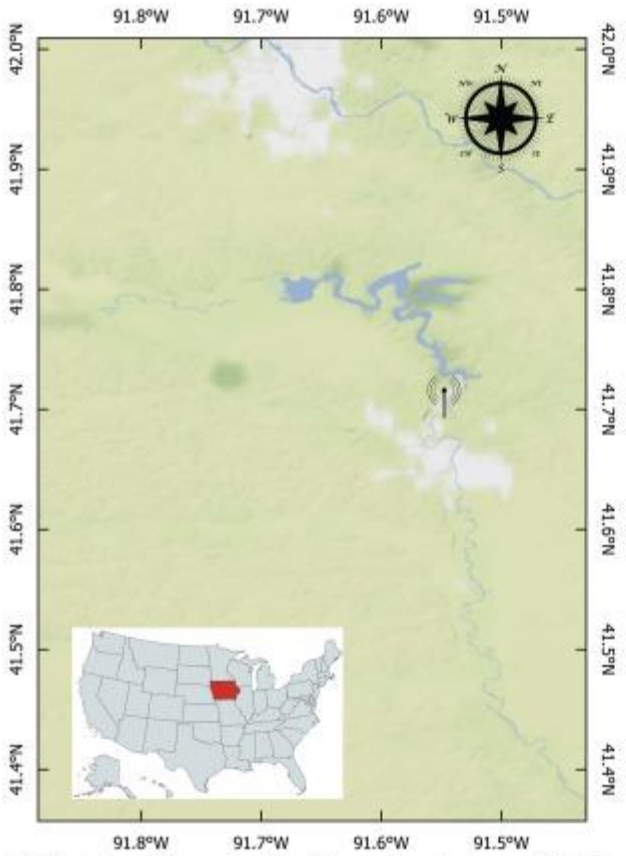
метод	**Сила**	**Обмеження**
Середня імпутація		Варіабельність даних зменшується; стандарт
Порожня клітинка		відхилення та оцінки дисперсії можуть отримати
LOCF	Легко зрозуміти,	недооцінений [35] ,
Порожня клітинка	Ефективний у застосуванні	Припущення здебільшого нереалістичне [36] ,
Лінійна імпутація		Між предиктором має бути лінійна залежність
Порожня		і змінні відповіді

метод	Сила	Обмеження
клітинка		
ЕМ	Хороша інтерпретація,	Високий ризик перевиконання тренувальних даних
МИШІ	Ледачий Навчання, не будуйте	Обчислювально дорого для великих наборів даних
КНН	моделі з навчальних даних	
SSIM	Може захоплювати та використовувати	Метод чорного ящика,
Дві SSIM	тимчасова інформація,	Продуктивність значною мірою залежить від
БРИТАНЦІ	Глибока архітектура приносить	налаштування гіперпараметрів,
М-РНН	сильна репрезентативність навчання	Високі обчислювальні витрати на створення моделі

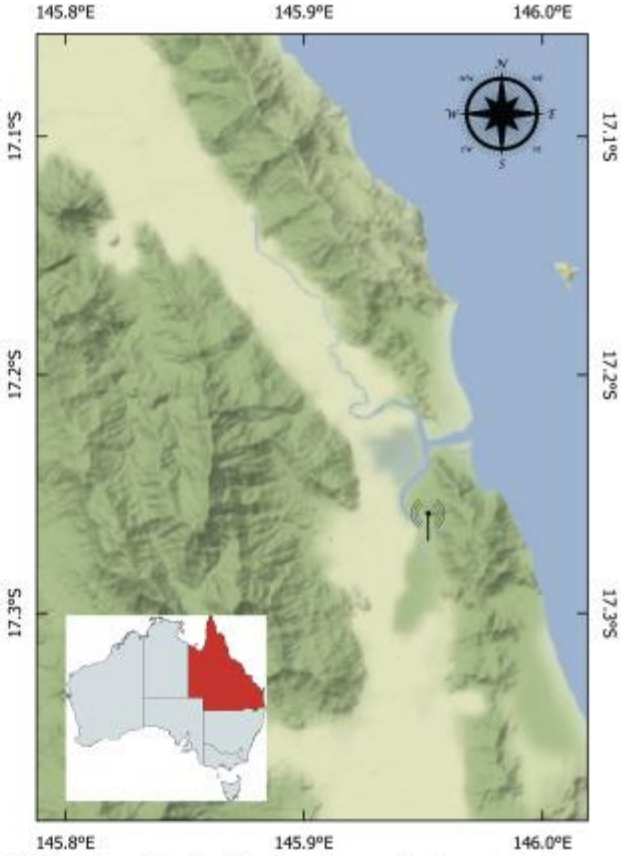
7 . Експериментальні випадки

7.1 . Мережі моніторингу якості води

В експериментальному розділі ми вирішили оцінити методи імпутації даних за допомогою вимірювань якості води, зібраних з двох систем моніторингу якості води , розташованих як у США , так і в Австралії (див. рис. 5).



(a) The Iowa Water Quality Information System (IWQIS). The black icon presents the selected monitoring station located in the upstream of Iowa River.



(b) The Great Barrier Reef catchment loads monitoring program. The black icon represents the in-situ monitoring station located in the upstream of Russell River.

1. Завантажити: завантажити зображення високої роздільної здатності (847 КБ)
2. Завантажити: завантажити повнорозмірне зображення

Рис. 5 . Дві системи моніторингу якості води в США та Австралії.

7.1.1 . Інформаційна система якості води Айови

Інформаційна система якості води Айови (IWQIS) [43] — це мережа моніторингу якості води в штаті Айова, США. Він пропонує вимірювання в реальному часі змінних якості води, таких якНІхконцентрація, pH, каламутність , електропровідність , розчинений кисень і температура. Дані, використані в цьому розділі, були зібрані з однієї станції моніторингу, розташованої у вододілі Кліар-Крік.

7.1.2 . Програма моніторингу навантажень на водозбір GBR

Програма моніторингу навантаження водозбірного басейну Великого Бар'єрного рифу спрямована на те, щоб допомогти відстежити довгострокові та короткострокові тенденції якості води в Північному Квінсленді, Австралія [44] . Програма здійснює моніторинг усіх водозборів інтенсивного землекористування. Він включає 43 контрольовані ділянки в 20 ключових водозбірних районах для моніторингу опадів і поживних речовин, а також 20 місць для пестицидів.

Таблиця 2 . Дані про якість води зібрані з двох систем моніторингу.

Змінні	**одиниця**	**Хв**	**Макс**	**Середній**	**Стандартна розробка**
IWQIS					

Змінні	одиниця	Хв	Макс	Середній	Стандартна розробка
Температура води	°C	2.2	29.6	18.9	7.4
провідність	µS/см	265,8	683,7	554.7	68.9
селітра	мг/л	1.9	14.9	7.8	2.5
GBR					
Температура води	°C	15.7	33.7	24.4	3.4
провідність	µS/см	61.9	1021.4	403.3	205.9
селітра	мг/л	0,002	3.3	0,3	0,3

7.2 . Дані моніторингу якості води

Три змінні якості води, такі як температура води, електропровідність і вміст нітратів, вимірюються в обох системах (Таблиця 2). У наступному розділі ми вирішили оцінити відсутні дані щодо температури води та концентрації нітратів. Дані датчика були нормалізовані та очищені, щоб видалити очевидні викиди. Крім того, у цьому експерименті ми повторили вибірку вимірювань якості води до однієї години. Крім того, усі три змінні використовуються як вхідні дані при обчисленні температури води та концентрації нітратів. Одна річ, яку варто згадати, полягає в тому, що ми плануємо врахувати відсутні вимірювання за кілька годин, оскільки вхідні дані збираються щогодини. Коли подаються щотижневі, місячні або річні дані, моделі в цьому документі можуть генерувати імпутації у відповідному часовому масштабі відповідно.

Щоб оцінити точність імпутації, ми спочатку готуємо дані навчання та тестування без пропущених значень, вибираючи різні періоди з даних, зазначених у таблиці 2 .

Потім для створення всіх зразків використовується алгоритм ковзного вікна. Для кожного зразка ми маскуємо послідовну кількість даних як базову істину, щоб ми могли оцінити, наскільки хороші результати імпутації.

7.3 . Метрики оцінювання

Ми оцінюємо ефективність відновлення відсутніх даних на основі середньоквадратичної помилки (RMSE) і середньої абсолютної помилки (MAE).(11)PMCE=1n∑i=1n(|fi−f^i|)2,(12)MAE=1n∑i=1n|fi−f^i|,defiif^iє істинними та розрахунковими значеннями змінної якості води, що підлягає моніторингу, відповідно.

8 . Результати

8.1 . Дані про температуру води

У таблиці 3 порівнюється продуктивність оцінки відсутньої температури води для десяти різних методів імпутації. І RMSE , і MAE використовуються для кількісного

визначення точності імпутації. Зрозуміло, що Dual-SSIM забезпечує найкращу продуктивність як для RMSE , так і для MAE у двох різних наборах даних. Наприклад, Dual-SSIM отримує 0,004 і 0,015 балів RMSE під час обробки даних, зібраних із систем моніторингу США та Австралії відповідно. Наступними найкращими моделями імпутації є методи на основі нейронних мереж, такі як SSIM, BRITS і M-RNN, які значно перевершують як статистичні, так і модельні рішення. Результати в таблиці 3 демонструють, що методи імпутації на основі нейронної мережі здатні використовувати чіткі часові закономірності, які з'являються в багатьох змінних якості води.

Таблиця 3 . Точність імпутації для температури води з розміром проміжку 6.

Порожня клітинка	**Температура води**			
Порожня клітинка	**США**		**AU**	
Порожня клітинка	**RMSE**	**MAE**	**RMSE**	**MAE**
Середня імпутація	0,48 (±0,041)	0,4 (±0,063)	0,409 (±0,062)	0,348 (±0,05)
LOCF	0,479 (±0,042)	0,399 (±0,064)	0,413 (±0,06)	0,351 (±0,048)
Лінійна імпутація	0,48 (±0,04)	0,4 (±0,062)	0,413 (±0,062)	0,353 (±0,051)
EM	0,48 (±0,042)	0,4 (±0,064)	0,417 (±0,052)	0,356 (±0,04)
МИШІ	0,48 (±0,041)	0,4 (±0,063)	0,409 (±0,062)	0,348 (±0,05)
КНН	0,649 (±0,015)	0,583 (±0,038)	0,552 (±0,166)	0,483 (±0,227)
Дві SSIM	**0,004 (± 0,001)**	**0,004 (± 0,001)**	**0,015 (± 0,004)**	**0,013 (± 0,004)**
SSIM	0,007 (±0,004)	0,007 (±0,004)	0,031 (±0,017)	0,028 (±0,016)
БРИТАНЦІ	0,007 (±0,004)	0,006 (±0,003)	0,03 (±0,009)	0,022 (±0,008)
M-РНН	0,026 (±0,002)	0,021 (±0,001)	0,066 (±0,007)	0,056 (±0,004)

Навпаки, методи імпутації, такі як імпутація середнього значення, LOCF і лінійна імпутація, не показали належних результатів при імпутації відсутніх даних про температуру води. Ці методи ігнорують часові закономірності, і більшість із них

передбачає наявність лінійного зв'язку між змінними якості води. Отже, вони мають низьку точність імпутації в експериментах.

8.2 . Дані про нітрати

Порівняно з температурою води, концентрація нітратів не має чіткої зміни в денній або тижневій шкалі часу. Часові варіації концентрації нітратів можна визначити лише при перевірці тенденції протягом кількох місяців. Отже, дуже складно оцінити відсутні вимірювання нітратів.

Таблиця 4 . Точність імпутації для концентрації нітратів із розміром пропуску 6.

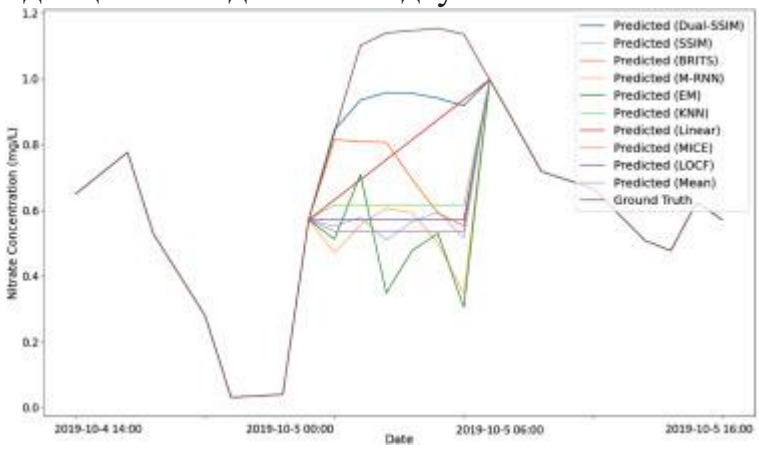
Порожня клітинка	**селітра**			
Порожня клітинка	**США**		**AU**	
Порожня клітинка	**RMSE**	**MAE**	**RMSE**	**MAE**
Середня імпутація	0,007 (±0,002)	0,005 (±0,001)	0,149 (±0,07)	0,096 (±0,05)
LOCF	0,011 (±0,003)	0,007 (±0,002)	0,155 (±0,07)	0,084 (±0,04)
Лінійна імпутація	0,012 (±0,001)	0,007 (±0,001)	0,144 (±0,044)	0,086 (±0,027)
EM	0,023 (±0,006)	0,015 (±0,003)	0,194 (±0,089)	0,122 (±0,058)
МИШІ	0,007 (±0,002)	0,005 (±0,001)	0,149 (±0,07)	0,096 (±0,05)
КНН	0,712 (±0,123)	0,661 (±0,166)	0,552 (±0,166)	0,483 (±0,227)
Дві SSIM	**0,002 (± 0,001)**	**0,002 (± 0,001)**	**0,078 (± 0,067)**	**0,069 (± 0,06)**
SSIM	0,003 (±0,001)	0,003 (±0,001)	0,098 (±0,083)	0,089 (±0,076)
БРИТАНЦІ	0,005 (±0,002)	0,004 (±0,002)	0,117 (±0,107)	0,077 (±0,072)
M-РНН	0,035 (±0,021)	0,029 (±0,017)	0,269 (±0,009)	0,233 (±0,044)

Таблиця 4 підсумовує точність імпутації для концентрації нітратів, виміряної в системах моніторингу США та АС. У цьому експерименті Dual-SSIM все ще показує найкращі показники як для RMSE, так і для MAE. Він має 0,002 і 0,041 RMSE для даних США та Австралії відповідно. Точність імпутації залежить від цих двох наборів даних. Дві системи моніторингу працюють у різних кліматичних

зонах і залежать від різних видів землекористування та сільськогосподарської діяльності. Отже, концентрація нітратів не завжди буде відповідати подібній моделі, і модель імпутації має бути специфічною для різних наборів даних якості води.

При імпутації даних про нітрати деякі прості методи, такі як середнє імпутування, можуть отримати кращу ефективність, ніж методи моделювання, такі як лінійне імпутування, EM і KNN. Наприклад, середня імпутація має оцінку RMSE 0,007 у застосуванні до даних США, тоді як лінійна імпутація та EM отримують 0,012 та 0,023 оцінки RMSE відповідно. Цей результат вказує на те, що методи моделювання на основі даних можуть не досягти очікуваної продуктивності, якщо часова інформація не використовується повністю.

На рис. 6 ми порівняли результати імпутації всіх методів, проаналізованих у цій статті. Концентрація нітратів коливалася протягом цього періоду, що зазвичай обумовлено внесенням добрив. У цьому прикладі відсутні 6 послідовних точок даних навколо піку. Прогалина заповнюється використанням різних методів імпутації. Очевидно, що LOCF, Mean і KNN генерують пряму лінію , що вказує на низьку точність імпутації. Лінійне імпутування також не підходить, якщо значення змінюються протягом періоду. Серед методів на основі нейронної мережі Dual-SSIM генерує імпутації з правильною тенденцією. Більшість методів значно недооцінюють діапазон відсутніх значень.

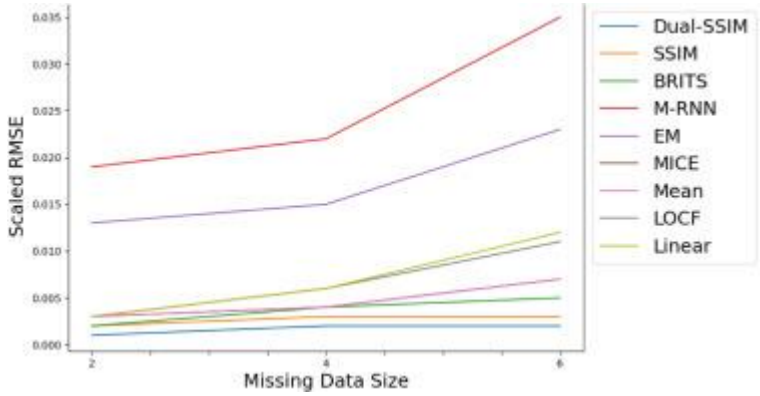


1. Завантажити: завантажити зображення високої роздільної здатності (161 КБ)
2. Завантажити: завантажити повнорозмірне зображення

Мал. 6 . Результати моделі в імпутуванні 6 послідовних відсутніх значень для концентрації нітратів з мережі моніторингу GBR. Суцільна червоно-коричнева лінія представляє наземні дані. Інші рядки представляють результати імпутації, створені різними моделями. 20 доступних даних до і після розриву використовуються як вхідні дані для моделі.

8.3 . Вплив розміру пропуску даних

Обсяг відсутніх даних про якість води може суттєво вплинути на точність імпутації. Чим довші прогалини в даних, тим меншу допомогу можуть надати доступні дані. Отже, ми оцінили, як змінюється точність імпутації при оцінці відсутніх даних різного розміру.



1. Завантажити: завантажити зображення високої роздільної здатності (116 КБ)
2. Завантажити: завантажити повнорозмірне зображення

Мал. 7 . Точність імпутації для оцінки даних різного розміру різними методами.

На рис. 7 показано, як працює модель імпутації при оцінці відсутніх даних різного розміру. В експериментальних умовах було оцінено дев'ять методів імпутації для відсутніх прогалин розміром 2, 4 і 6 відповідно. Загалом похибка імпутації зростає, коли обсяг відсутніх даних зростає.

Методи імпутації мають близьку ефективність при роботі з відсутніми даними розміру 2. Завдяки великій кількості доступної інформації більшість методів імпутації можуть ефективно обробляти невелику кількість відсутніх даних.

При подвоєнні розміру розриву до 4 методи на основі нейронної мережі все ще пропонують високу точність для імпутації відсутніх даних. У цьому випадку моделі на основі нейронної мережі все ще можуть вивчати часові закономірності з доступних даних, які забезпечують надійний орієнтир для заповнення відсутніх даних. Навпаки, лінійне імпутування, LOCF і EM показують різке зниження продуктивності.

Завдяки постійному збільшенню розміру відсутніх даних точність імпутації методів на основі нейронної мережі все ще залишається на низькому рівні. Однак статистичні та модельні методи не змогли досягти обнадійливих результатів. Для даних часових рядів відсутнє значення має меншу релевантність для доступної інформації на етапах часу, далеких від розриву. Якщо модель імпутації не має потужних можливостей для захоплення часових шаблонів з доступних даних, імпутування відсутніх даних без базової правди на найближчих кроках часу може бути складним.

9 . Висновок

У цьому документі наведено огляд широко використовуваних методів імпутації даних і якісно порівняно їх відповідні переваги та недоліки використання для вимірювання якості води.

Методи імпутації, перелічені в цьому документі, можна згрупувати в три різні типи. Статистичні методи імпутації заповнюють відсутні дані на основі статистичного аналізу даних часових рядів. Методи на основі моделі заповнюють відсутнє значення за допомогою регресійних моделей. Методи на основі нейронних мереж створюють конкретні моделі нейронних мереж для прогнозування відсутніх даних. Методи імпутації, побудовані на різних механізмах, мають свої переваги та обмеження. Отже, для конкретних обставин необхідно вибрати відповідні методи імпутації.

Підсумовуючи, розмір відсутніх даних суттєво впливає на точність імпутації. Більшість методів імпутації добре заповнюють відсутні дані за короткий проміжок часу. Наприклад, лише 1 або 2 дані відсутні в одному періоді. Коли в наборах даних є велика кількість відсутніх значень, методи імпутації, які не можуть використовувати тимчасову інформацію, мають значно нижчу продуктивність. Оскільки вони виграють від повторюваної архітектури, методи на основі нейронної мережі показують багатообіцяючі результати в обробці наборів даних із великими проміжками.