УДК 519.25

Р. М. Камінський, Н. Е. Кунанець, В. В. Пасічник, А. М. Худий
Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж

ВІДНОВЛЕННЯ ПРОПУСКІВ У РЕЗУЛЬТАТАХ ТЕСТУВАННЯ ТА ІДЕНТИФІКАЦІЇ ОПЕРАТОРСЬКОГО ПЕРСОНАЛУ

Розглянуто завдання відновлення пропущених значень у результатах тестування реципієнтів, поданих часовими рядами. Як експериментальні дані наведено штучні часові ряди із пропущеними значеннями. Ефективність відновлення оцінюється відносною похибкою відновленого значення. Наведено приклади відновлення пропусків у таблиці часових рядів та в індивідуальному часовому ряді. Використано прості методи заміни пропуску середнім, зваженим середнім та медіаною. Для цих рядів доволі добрі результати забезпечують медіана вихідного часового ряду з пропущеними значеннями та заповнення значеннями моделі тренду – полінома третього степеня.

Ключові слова: часові ряди, таблиці даних з пропусками, пропущені значення, відновлення пропусків.

The article addresses the problem of the restoration of missed values in the results of testing the recipients given by the time series. As experimental data, the time series with spaces are given. Recovery efficiency is estimated by the relative error of the recovered value. Examples of restoration missing data in the time series table and the individual time series are given. Used simple methods for replacing missed value by average, weighted average and median. For these time series, the good results provide the median of the output time series with the missing values and fill the values of the trend model – the polynomial of the third degree.

Key words: time series, data tables with spaces, missing values, restoration missing data

Вступ. Загальна постановка проблеми

У різних дослідженнях людино-машинних систем для оцінювання кваліфікації операторського персоналу широко використовують методи тестування. Отримані результати часто подають у вигляді таблиць та окремих часових рядів. Проте в отриманих даних, як правило, часто виявляються відсутніми деякі значення. У таких випадках вважають, що маємо дані з пропусками. Пропуск значень істотно ускладнює обробку даних, оскільки отримані в результаті статистичної обробки оцінювані показники мають зміщення, а також неприйнятними виявляються деякі методи ідентифікації, оскільки обсяги вибірок індивідуальних характеристик різні. Вважають, що величина зміщення основних статистичних характеристик прямо пропорційна до кількості пропусків. Тому встановлення пропущених значень є важливою і першочерговою процедурою для подальшої обробки даних, причому відновлення пропущених даних передбачає таке застосування тих чи інших відповідних методів, яке, принаймні істотно, не змінює статистичного характеру вже отриманих даних. Для відновлення пропущених значень існує доволі великий арсенал різних методів – від найпростіших, наприклад, замінити пропущене значення деяким вибраним з тих чи інших міркувань числом або середнім значенням, до доволі складних, які потребують докладного аналізу і спеціальної обробки: нейронні мережі, еволюційні алгоритми, методи прогнозування тощо.

Проблема пропущених значень даних є вельми актуальною у соціології, розпізнаванні образів, у кластерному аналізі тощо. Найчастіше вона виникає під час ідентифікації різних об'єктів, коли апріорна інформація про значення параметрів є неповною. У результаті отримують неповні дані, а у випадку часових рядів замість еквідистантного одержують нееквідистантний часовий ряд, оброблення якого і сьогодні є доволі складним.

Аналіз проблеми за результатами наукових публікацій

Характеристики багатьох методів відновлення пропущених даних доволі докладно описано в [1]. В цій роботі згадано чотири методи заповнення пропусків: виключення некомплектних об'єктів, методи із заповненням, зважування та методи, основані на моделюванні. В [2] описано базовий алгоритм ZET заповнення пропусків, у основу якого покладено процедуру циклічного формування "компетентної матриці", побудову моделі прогнозування та прогнозування пропущеного значення. Особливістю цього алгоритму є його локальність, тобто для обчислення пропущеного значення використовується не вся таблиця даних, а лише її частина. Крім того, розглянуто ЕМ-метод, який для заповнення пропущених значень використовує моделі.

Відновлення пропущених даних в екологічних задачах автори [3] здійснюють з використанням МГУА. Для відновлення пропущених значень у [4] розроблено програмний комплекс на основі методів нечіткого моделювання та описано можливості нечітких систем для таких задач. Оригінальний гібридний адаптивний метод відновлення даних з нейронечітким управлінням запропоновано в [5] та подано оцінку ефективності різних методів відновлення стосовно до масивів даних, що містять інформацію про процеси в складних динамічних системах. В [6] розроблено інформаційну технологію імпутації даних змішаної природи, яка підвищує якість первинних даних у задачах інтелектуального і соціально-мережевого аналізу. Технологія містить комплекс математичних моделей, методів і систему інформаційних процесів, що реалізують процедури отримання, оброблення, зберігання і видавання інформації, яка використовується під час функціонування. В [7] наведено опис алгоритму ковзної віконної апроксимації на основі поліномів другого степеня, яка дає змогу під час відновлення і корекції часових рядів супутникових даних одночасно розв'язувати задачі відновлення пропущених значень. У [8] для відновлення пропущених значень часових рядів розроблено метод ковзного двобічного експоненційного згладжування. В [9] наведений метод неієрархічного розбиття, який узагальнює неієрархічні дивізивні методи кластеризації та групування. Цей метод використано для побудови простих алгоритмів відновлення пропусків у даних.

Програмний комплекс Amelia II, описаний у [10], забезпечує роботу з пропущеними даними та різними методами їх відновлення. В [11] автори використали три методи для заповнення пропусків у даних, а саме: відновлення середнім значенням, множинної імпутації (Multiple imputation (MI)) та повної інформації максимальної правдоподібності (Full information maximum likelihood (FIML). Їх результати показали, що відновлення пропусків середнім значенням є не найкращим методом.

Технічний звіт [12] містить основні поняття та методи, що використовують для відновлення відсутніх даних. Розглядають основні загальноприйняті методології та байєсівські методи. В [13] автори розглядають такі способи обробки відсутніх даних: метод Йєтса, параметричний метод, оснований на максимальній вірогідності, та метод Маркова – Монте-Карло. Результати показують, що найменша невідповідність у разі застосування методу Йєтса. В [14] показано, що для заповнення пропущених значень у даних порівняно найкращий результат дає використання методу MCMC (Markov Chain Monte Carlo). Автори [15] для відновлення пропущених значень даних віддають перевагу програмному забезпеченню, яке використовує оцінки максимальної правдоподібності лінійних моделей з відсутніми даними і яке є в багатьох автономних пакетах. У [16] наведено критичний огляд методів, які на той час використовували у медичних дослідженнях для відновлення пропущених даних. У статті [17] подано результати імітаційного моделювання процесу відновлення пропущених значень часового ряду результатів тестування знань студентів. Використовуваний в роботі метод аналізу часових рядів з пропусками дає точні результати лише за доволі жорстких припущеньх і виявляється застосовним до реальних рядів, проте приводить, в цьому випадку, до наближених результатів. Метод множинного відновлення даних, наведений у [18], забезпечує порівняно добрі результати, проте доволі складний як стосовно математичного апарату, так і щодо змістової інтерпретації результатів. В [19] запропоновано еволюційний метод, оснований на композиції використання нейронної мережі та генетичного алгоритму. Відновлення

цим методом пропущених значень не потребує дотримання обмежень на вхідні дані, тобто вони можуть мати довільну розмірність і структуру пропусків. Підхід, запропонований у [20], полягає у тому, що спочатку простими методами визначають поодинокі пропущені значення, а потім складними методами заповнюють групові пропуски, а це робить таку методику новою і доволі ефективною для відновлення пропущених значень у масивах даних. Експериментальні дослідження методу Бартлета та Resampling методу, наведені в [21], показали, що останній простіший в алгоритмічному плані й дає результати такої самої якості, як і метод Бартлета. Як програмне забезпечення використано макроси табличного процесора MS Excel. В [22] вказано, що добру якість у відновленні даних для вибірки обсягом 24 значення дають методи сплайн-інтерполяції та експоненціального згладжування і, крім того, як зазначають автори, ці методи забезпечують добру якість відновлення поодиноких пропущених даних. У [23] відновлення пропущених значень здійснено для часових рядів із сезонною компонентою і коротко охарактеризовано прості методи заповнення пропусків для цього випадку. Візуально подано результати заповнення пропусків простими методами та запропонованим методом розрахунку відсотка від знакової величини (максимуму чи мінімуму), який виявився найвідповіднішим для такого типу часових рядів. У [24] наведено модифікацію алгоритму ZET та результати порівняння модифікованого алгоритму із оригіналом. В [25] розглянуто спосіб відновлення пропусків у даних за допомогою аналітичних моделей із застосуванням табличного процесора MS Excel, причому використано метод фіктивних змінних із побудовою регресійної.

Причини і проблеми відновлення пропущених даних

Складні системи – технічні, економічні, екологічні, соціальні – часто потребують оперативного прийняття рішення на основі даних відповідних спостережень за їхніми характеристиками. В одних випадках використовують автоматичне спостереження за допомогою спеціальних давачів, в інших це робить спеціально підготовлений персонал, здійснюється опитування експертів та пошук інформації у відповідних документах. Усе це виконують з певним ступенем відповідальності, різними методами збору інформації, по-різному здійснюється консолідація даних. Тому неминуча втрата інформації у вигляді пропущених даних або помилок як значень, що різко виділяються серед інших елементів вибірок та часових рядів.

Наявність пропущених значень у даних істотно обмежує можливості використання різних методів обробки. Важливою проблемою є частота появи пропусків та існування у пропусках будь-яких закономірностей.

На практиці найчастіше розглядають два види даних: таблиці та часові ряди.

Таблиці подають так: рядки – це характеристики (параметри) досліджуваних об’єктів (держави, комп’ютери, методи тощо), а стовпчики – це значення конкретної характеристики для кожного об’єкта і, в принципі, вони є числовими значеннями відповідної шкали.

Часові ряди характеризують динаміку об’єкта у вигляді зміни деякого конкретного визначального показника, значення якого найадекватніше відображають поведінку цього об’єкта.

У першому випадку йдеться про декілька об’єктів, які подані значеннями даних у рядках таблиці. Такі дані в таблиці стають в певному сенсі залежними від клітинок, у яких вони локалізовані. Вони є відповідно впорядкованими за номіналами, а кожне значення характеризує якусь якість об’єкта або властивість.

У другому випадку йдеться про часові ряди, подані значеннями одного показника, прив’язаними до моментів часу його реєстрації.

Відомі методи заповнення пропусків істотно різняться і по-різному забезпечують такі режими роботи.

1. Заповнення усіх пропусків у таблицях чи часових рядах.

2. Заповнення лише тих пропусків, для яких помилка від заповнення цього пропуску не перевищує заданої величини.

3. Заповнення пропусків на основі інформації, що є в таблиці.

4. Заповнення кожного наступного пропуску з використанням початкової інформації та отриманих у результаті прогнозування значень, враховуючи раніше заповнені пропуски.

Найважливішою є якість заповнення пропусків. Варто зазначити, що в задачах тестування або ідентифікації операторського персоналу за допомогою комп'ютерного тренажера є особливість: насамперед тут фігурують індивідуальні часові ряди і лише потім, для цілої групи реципієнтів, дані подаються таблицею "оператор – час розпізнавання" конкретного тестового зображення. Тому, якщо даних багато, а пропусків мало, то характеристики отриманих даних незначно відрізнятимуться від істинних значень для сукупності даних, тобто за відсутності пропусків. Для такого випадку знайти заміну пропущеному значенню не дуже складно, оскільки відома природа даних, принаймні статистична (розподіл, описова статистика). В іншому разі необхідно використати декілька методів і вибрати найкращий за прийнятим критерієм якості, наприклад, за характеристиками описової статистики, середньоквадратичним відхиленням від тренду тощо. Крім того, можна використовувати різні моделі трендів.

Ще одним методом заповнення пропусків є використання методів інтерполяції, які доволі добре розроблені для таблиць з дискретними значеннями складних функцій. Ці методи рідко використовують для заповнення пропусків у реальних даних, оскільки вони дають добрі результати лише для даних, які монотонно зростають чи зменшуються, складних, переважно гладких, функцій, тоді як реальні значення мають значний розкид. Тому для заповнення пропущених значень використовують простіші методи, які розглянуто нижче, для таблиць та часових рядів, хоча істотних відмінностей між ними практично немає.

Формулювання мети

У задачах ідентифікації операторського персоналу в різних системах професійного відбору, навчання та атестації для перевірки професійних навичок широко використовують набори тестових зображень, які послідовно подають реципієнту. В розглянутому випадку для тестування використано послідовність стилізованих зображень деякої гіпотетичної робочої ситуації. Суть робочої ситуації у тому, щоб на випадковому тлі виявити і розпізнати об'єкт уваги заданого класу. Особливість цих зображень у тому, що тлом слугує квадратне фрактальне зображення ділянки земної поверхні, яке подається у восьми положеннях: за допомогою поворотів 0º, 90º, 180º, 270º, в прямому та дзеркальному відображеннях. На такому тлі випадково локалізоване стилізоване зображення літака, мінімально можливих розмірів, орієнтованого вздовж вертикальної осі в напрямку вверх. Під час тестування експозиція кожного зображення становить 30 с, а загальна кількість зображень у послідовності тестів 60 × 3 = 180, тобто сама послідовність містить 60 зображень і повторюється підряд тричі, що дає можливість порівняти результати на початку і в кінці тестування. Запам'ятовування будь-якого зображення практично виключене. Процедура тестування полягає в тому, що в момент появи зображення на моніторі вмикається комп'ютерний секундомір, який вимикає реципієнт, виявивши і розпізнавши об'єкт уваги, навівши мишку на цей об'єкт спеціального візира і кліком мишки зупиняє його. Цей час є часом розпізнавання ситуації, а послідовність його значень подається еквідистантним часовим рядом з обсягом рівнів n = 180. Отже, тривалість тестування становить 90 хв.

Отримані в результаті тестування дані у вигляді часових рядів містять 10–30 % відсотків пропущених рівнів, що істотно позначається на ідентифікаційних показниках та параметрах моделі тренду часового ряду.

Тому метою цієї роботи є використання простих методів відновлення пропущених значень рівнів у задачах ідентифікації та відбору реципієнтів.

Методи заповнення пропусків у таблицях даних

Результати експериментальних досліджень загалом подають таблицею "об'єкт–властивість", в якій рядки відповідають досліджуваним об'єктам, а стовпці – значенням їх ознак та характеристикам. У цьому випадку об'єктами є реципієнти, а властивостями чи ознаками – значення рівнів. Кожен рівень відповідає конкретному зображенню робочої ситуації, а оскільки

зображення не перемішуються, то можна стверджувати, що кожен стовпець відповідає конкретній ознаці, тобто конкретному тесту.

Під час тестування, з різних причин, заданий об'єкт уваги на тестовому зображенні може бути не виявленим або не розпізнаним, в результаті чого буде відсутнє значення для цього тесту. В такий спосіб утворюються пропуски, особливо коли зображення об'єкта маскується зовнішніми (шум, завади) та внутрішніми (контраст, яскравість, розмитість) чинниками. Очевидно, що в таких випадках обробка даних стає доволі складною, навіть проблематичною, оскільки більшість методів математичної статистики не можуть працювати з неповними даними, зокрема кореляційний та факторний аналіз.

Отримана таблиця може містити значну кількість пропусків, наприклад, як зображена на рис. 1. Ознаками об'єктів у цій таблиці є тестові зображення, кожен з яких має, відповідно до реципієнта, своє значення часу розпізнавання і прийняття рішення щодо цього об'єкта уваги, виражений мілісекундами. Тому для заповнення пропусків, зображених у таблиці символом порожньої множини, можна застосувати статистичний підхід і використовувати числові характеристики описової статистики та близькі за простотою до них, такі як середнє арифметичне, медіана, мода, зважене середнє тощо. В реальній таблиці значення в рядку реципієнта відповідають конкретним тестам, як час розпізнавання на ньому виявленого об'єкта. Зауважимо, що у цей час входять: час спостереження зображення до виявлення об'єкта уваги, час його розпізнання і вибору рішення та час психомоторної реакції наведення мишкою візира до завершення кліку мишки.

Об'єкти	**Ознаки**					
	Тест 1	Тест 2	Тест 3	...	Тест $n-1$	Тест n
Реципієнт R_1	t_1^1	t_2^1	t_3^1	...	t_{n-1}^1	t_n^1
Реципієнт R_2	t_2^2	$\varnothing$	t_3^2	...	t_{n-1}^2	t_n^2
...	...	...	...	...	...	...
Реципієнт R_{m-1}	t_1^{m-1}	t_2^{m-1}	$\varnothing$	...	t_{n-1}^{m-1}	$\varnothing$
Реципієнт R_m	t_1^m	$\varnothing$	t_3^m	...	t_{n-1}^m	$\varnothing$

Рис. 1. Таблиця "об'єкт–ознака" даних тестування з пропущеними значеннями

Для заповнення пропусків сьогодні є дуже багато методів та методик, які відновлюють пропущені значення у даних, проте у кожного з них є переваги та недоліки залежно від того чи іншого класу задач. За великої кількості пропусків та малої кількості даних навряд чи існує метод їх ефективного відновлення, проте якщо пропусків кілька, багато методів зможуть, відповідно до заданого критерію, "відновити" пропущені значення.

У цьому дослідженні використано такі найпоширеніші прості методи відновлення пропусків:

1. З*аповнення пропусків середнім значенням* за стовпчиком, тобто пропуск заповнюється середнім значенням наявних у таблиці значень цієї ознаки за умови, що об'єкти є аналогами. Інакше кажучи, якщо в таблиці об'єкти подані рядками, а значення ознак об'єктів локалізовані в стовпчиках, середнє шукають у тому стовпчику, в якому є клітинка з пропущеним значенням. Цей метод нескладно реалізувати, але його застосування можливе лише за умови MAR (missing at random), коли пропуски є випадковими величинами, причому сам механізм утворення пропуску можна проігнорувати. Недоліком методу є спотворення розподілу даних та зменшення дисперсії початкових даних.

2. *Зваженим середнім.* Під зваженим середнім, в цьому випадку, розумітимемо значення, обчислене для відновлення даних за віконного ковзного середнього, тобто розраховане за конкретними формулами. У такому разі декілька значень до пропуску або після пропуску вважають "вікном", в якому і здійснюють обчислення. Можна також використати середнє з цих двох обчислень.

3. ***Заповнення пропуску медіаною.*** Медіана є найстабільнішою (робастною) характеристикою вибірки, оскільки за будь-яких перетворень вона залишається незмінною. Її з успіхом можна використати для таблиці. Щоб знайти значення медіани, будують варіаційний ряд, серединою якого і є медіана. Проте за малої кількості об'єктів і декількох пропусків значення медіани може доволі суттєво змінюватись, тому його потрібно визначати після кожного заповненого пропуску, а самі пропуски заповнювати як послідовно, так і випадковим чином.

У таких задачах заповнення пропущених значень можна скористатися такими варіантами. Оскільки в цьому випадку таблиця містить лише дані індивідуальних часових рядів реципієнтів, серед яких можуть бути також часові ряди повторних проходжень тієї самої послідовності тестових зображень, то для відновлення пропущених значень у часовому ряді можна зробити так.

1. Для встановлення пропущеного значення для тесту можна використати середнє значення часу розпізнавання для цього тесту.

2. Якщо в таблиці є часові ряди повторного тестування цього реципієнта і в таких рядах пропуски не збігаються щодо того чи іншого тесту, то можна відповідно до тестів використати значення одного ряду для заповнення пропусків іншого. Крім того на множині значень таких рядів можна моделювати різні пропущені значення, характерні для цього реципієнта.

3. Для часового ряду з пропущеними значеннями рівнів знайти в таблиці найподібніший до нього за статистичними характеристиками часовий ряд, в якому є дані для відповідних тестів, і ними заповнити пропущені значення рівнів.

Приклад. Для ілюстрації методів розглянемо гіпотетичний приклад заповнення в таблицях пропущених значень у середовищі табличного процесора Ms Excel.

Нехай для декількох об'єктів відсутня певна ознака, тобто у деякому стовпчику таблиці "об'єкт–ознака" є пропущені значення. Як вектор використаємо послідовність випадкових чисел з рівномірним розподілом у межах значень ознак $x_i \in [1, 10]$, як зображено на скриншоті Excel на рис. 2.

Послідовність B3: B32 містить $n = 30$ значень і є, практично, мінімально репрезентативною, а тому висновки на її основі прийнятні (достовірні). Визначимо для цього вектора показники описової статистики, використовуючи: "Сервіс → Аналіз даних → Описова статистика". Основні характеристики, за якими можна порівнювати результати застосування того чи іншого методу, такі: середнє арифметичне з усіх значень вектора, значення медіани, стандартне відхилення або дисперсія, ексцес, асиметрія, сума. Значення моди в цьому прикладі не є інформативним, оскільки рівномірний розподіл не має моди, а максимальне і мінімальне значення у такому разі залишатимуться незмінними, якщо тільки вони не є пропусками, так само незмінним буде інтервал. Для імітації пропусків вилучимо з цієї послідовності такі значення B7, B13, B20 і B24. У результаті отримаємо вектор ознаки з пропусками E3: E32.

Для заповнення пропусків використовуємо такі методи: заповнення середнім значенням для неповних даних, зваженим середнім, розрахованим за формулою для ковзного середнього з вікном $k = 5$ та заповнення значенням медіани, визначеної за неповними даними.

Для з'ясування результатів заповнення пропусків вказаними методами використано показники описової статистики, наведені в скриншоті на рис. 2 у рядках 35–47. Для оригінального вектора і цього вектора з утвореними пропусками описова статистика подає незначні зміни основних показників, причому відносна помилка, визначена як відношення абсолютної різниці між показником для оригінальних даних і цих даних з пропусками до значення показника оригінальних даних дає значення, наведені в табл. 1. В рядку "пропуски" подано порівняння оригіналу без пропусків і оригіналу з пропусками.

Оскільки критерієм для оцінювання якості методу вибрано відносну похибку значення показника, то чим менше це значення, тим кращою є заміна пропуску значенням, встановленим цим методом. Тут потрібно мати на увазі: якщо значення оригінального вектора є випадковими величинами з рівномірним чи з будь-яким іншим законом розподілу, який не має моди, то значення показників моди, ексцесу й асиметрії можна зігнорувати.

Як критерії порівняння можна використати коефіцієнт кореляції між моделлю обвідної варіаційного ряду оригінального часового ряду з пропусками і самого оригінального ряду з

відновленими пропущеними значеннями. Проте цей критерій використано для побудови моделі обвідної варіаційного ряду з пропущеними значеннями застосування методів, які можуть працювати з пропусками, або необхідно погодитись із тим, що обвідна такого варіаційного ряду для оригінального часового ряду з пропусками практично відповідає оригінальному часовому ряду без пропущених значень.

	A	B	C	D	E	F	G	H	I	J	K	L
1		number	value		missing		average		weighted		median	
2		object	original						average			
3		1	9,096		9,096		9,096		9,096		9,096	
4		2	3,407		3,407		3,407		3,407		3,407	
5		3	9,567		9,567		9,567		9,567		9,567	
6		4	3,031		3,031		3,031		3,031		3,031	
7		5	8,652				6,156		2,859		5,770	
8		6	8,712		8,712		8,712		8,712		8,712	
9		7	1,371		1,371		1,371		1,371		1,371	
10		8	8,556		8,556		8,556		8,556		8,556	
11		9	9,977		9,977		9,977		9,977		9,977	
12		10	5,895		5,895		5,895		5,895		5,895	
13		11	3,758				6,156		3,838		5,770	
14		12	9,332		9,332		9,332		9,332		9,332	
15		13	5,573		5,573		5,573		5,573		5,573	
16		14	5,101		5,101		5,101		5,101		5,101	
17		15	9,855		9,855		9,855		9,855		9,855	
18		16	4,809		4,809		4,809		4,809		4,809	
19		17	2,122		2,122		2,122		2,122		2,122	
20		18	2,687				6,156		2,339		5,770	
21		19	3,872		3,872		3,872		3,872		3,872	
22		20	9,990		9,990		9,990		9,990		9,990	
23		21	2,187		2,187		2,187		2,187		2,187	
24		22	5,756				6,156		2,972		5,770	
25		23	5,645		5,645		5,645		5,645		5,645	
26		24	8,462		8,462		8,462		8,462		8,462	
27		25	2,236		2,236		2,236		2,236		2,236	
28		26	7,692		7,692		7,692		7,692		7,692	
29		27	1,127		1,127		1,127		1,127		1,127	
30		28	4,522		4,522		4,522		4,522		4,522	
31		29	8,411		8,411		8,411		8,411		8,411	
32		30	9,499		9,499		9,499		9,499		9,499	
33												
34	Descriptive Statistics											
35	Mean		6,030		6,156		6,156		5,735		6,104	
36	Standard Error		0,544		0,600		0,519		0,557		0,519	
37	Median		5,700		5,770		6,156		5,337		5,770	
38	Mode		#N/A		#N/A		6,156		#N/A		5,770	
39	Standard Deviatior		2,981		3,059		2,840		3,049		2,844	
40	Sample Variance		8,885		9,359		8,068		9,297		8,086	
41	Kurtosis		-1,487		-1,484		-1,191		-1,582		-1,217	
42	Skewness		-0,117		-0,214		-0,228		0,085		-0,169	
43	Range		8,863		8,863		8,863		8,863		8,863	
44	Minimum		1,127		1,127		1,127		1,127		1,127	
45	Maximum		9,990		9,990		9,990		9,990		9,990	
46	Sum		180,897		160,044		184,666		172,052		183,123	
47	Count		30,000		26,000		30,000		30,000		30,000	

Рис. 2. Результати заповнення пропусків

Результати заповнення наведено в табл. 1. Для порівняння результатів заповнення пропущених значень використано показники описової статистики, наведені в першому стовпчику, обчислені значення цих показників – у другому і третьому стовпчиках. Як критерій відповідності відновленого значення істинному використано відносну похибку у відсотках для кожного відновленого відповідним значенням, обчисленим за конкретними і відомими формулами. Значення похибки наведено в останніх трьох стовпчиках табл. 1.

Таблиця 1

Результати заповнення пропусків

Показники описової статистики	Показники оригінального часового ряду	Показники оригінального часового ряду з пропусками	Відносна похибка показників часового ряду, % для відновлених значень		
			середнім	зваж. середнім	медіаною
Середнє	6.030	6.156	2.089	4.892	1.227
Середнє квадр.	2.981	3.059	4.730	2.281	4.595
Медіана	5.700	5.770	8.000	6.368	1.228

У цьому випадку метод заповнення середнім виявився найменш придатним.

Зауваження. Приклад ілюструє лише процедуру, а не розв'язання конкретної задачі.

Методи заповнення пропусків у часових рядах

Подання даних часовими рядами та їх аналіз набувають щораз більшої популярності в різних наукових дослідженнях. Це пов'язано передусім із наявністю великої кількості динамічних процесів, об'ємних масивів спостережень з широкими просторовими і часовими діапазонами і неможливістю розглядати багато явищ без урахування часового контексту. Особливо велике значення аналіз часових рядів має для дослідження потоків даних в інтелектуальних динамічних системах, зокрема в людино-машинних системах, соціальних мережах, у задачах моделювання процесів розвитку різних систем і явищ, у прогнозуванні ситуацій та динаміки систем на основі даних моніторингу їх стану.

Основною причиною, що призводить до появи пропусків у часових рядах, є неможливість отримати інформацію в певні моменти часу. Наприклад, аналіз часового ряду може бути пов'язаний з об'єктивними пропусками даних, причина яких полягає у тому, що інформація на момент спостереження або відсутня, або недоступна, або недостовірна. Крім того, можлива ситуація, коли засоби вимірювання, спостереження, реєстрації не налаштовані, зіпсовані, не відповідають вимірюваним величинам або мають невідповідні межі їх вимірювання (невідповідність шкал значень, низька чутливість, тривалість вимірювання значна), дані реєструє некваліфікований персонал.

Характерним для часових рядів є те, що залежно від предметної області природа пропусків має певні особливості. Поява пропусків у фінансових рядах суттєво відрізняється від виникнення пропусків у соціологічних чи біологічних даних або під час психологічних досліджень. Пропуски в часових рядах, утворених технологічними процесами, відрізняються від пропусків даних в інформаційних системах тощо. Проте, заповнюючи пропуски, доволі часто ігнорують їх природу і застосовують один або декілька найбільш доступних і простих методів. Вигляд часового ряду з пропусками зображено на рис. 3.

Сьогодні існує чимало методів відновлення пропусків, але єдиної методології відновлення пропущених значень або оброблення даних з пропусками немає, незважаючи на її необхідність. Вибір необхідного підходу або найприйнятнішого методу заповнення пропусків у кожній конкретній ситуації – часто доволі складне окреме завдання, яка може потребувати значно більше часу і зусиль, ніж саме оброблення даних з відновленими пропущеними значеннями.

Заповнюючи пропуски у часових рядах, можна використовувати методи для заповнення пропусків у таблицях, у кожному конкретному випадку метод заповнення має бути обґрунтований, а результати інтерпретовані. Річ у тім, що, на відміну від таблиць, якщо розглядати сукупність часових рядів (наприклад, у задачах ідентифікації декількох об'єктів, поданих часовими рядами), маємо ту саму таблицю, проте у ній ознаками є рівні часового ряду. У них однакові природа, фізичний зміст, їх величини виміряні в одній шкалі, вони є залежними випадковими величинами з однаковим розподілом і, головне, вони зв'язані впорядкованою послідовністю моментів часу, в які їх зареєстровано, причому кожному рівню відповідає конкретна ситуація.

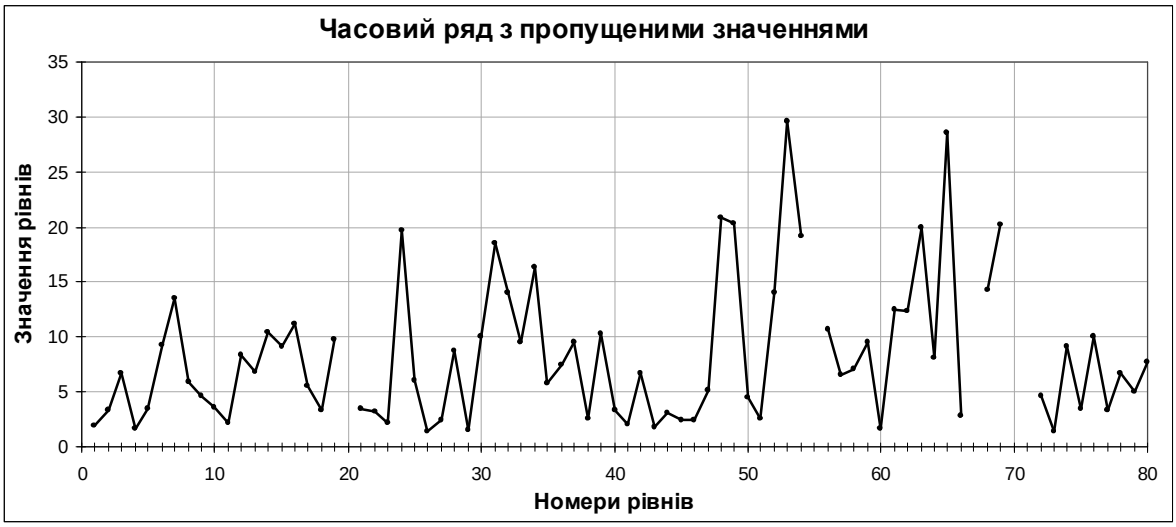

Рис. 3. Часовий ряд з пропусками, що виглядає як послідовність окремих фрагментів

У цьому дослідженні використано метод ***заповненням значеннями моделі даних.*** Суть цього методу полягає в тому, що будують модель тренду у вигляді відповідної функції, переважно нелінійної. Тоді значення пропущених рівнів беруть з цієї моделі (функції) відповідно до номерів цих рівнів. Для заповнення беруть те значення цієї функції, яке відповідає пропуску. Обґрунтування цього методу полягає в тому, що отримані дані *єдині*, інакших не має і саме вони відповідають реципієнту. Тому побудована на основі цих даних модель найбільше відповідає "стану речей" і відновлені дані не повинні істотно змінити її параметри та характеристики часового ряду.

Якщо йдеться про пропущені рівні часового ряду, то існуючі рівні доволі часто розглядають як елементи вибірки. За рекомендаціями [1] можна використати такі способи обробки даних для відновлення пропущених даних.

1. Визначення первинних статистичних характеристик досліджуваної вибірки без урахування пропусків. Очевидно, що за такого способу обсяг вибірки зменшується на кількість пропусків, а разом з тим зменшується і кількість ступенів вільності під час перевірки сформульованих статистичних гіпотез.

2. Визначення первинних статистичних характеристик досліджуваної вибірки у разі заповнення пропущених значень нулями. У такому разі кількість ступенів вільності не змінюється, проте виникає зміщення оцінок статистичних характеристик порівняно з початковою вибіркою рівнів.

3. Визначення первинних статистичних характеристик досліджуваної вибірки у разі заповнення пропущених значень середніми значеннями. Проте і в цьому випадку також можливе зміщення оцінок отриманих характеристик порівняно з оригінальною, тобто початковою вибіркою.

4. Заповнити пропущені значення можна значеннями числових характеристик розподілу рівнів – моди, медіани або його квартилів. Якщо пропусків декілька, то після заповнення пропуску варто знайти нове значення з цих характеристик.

5. Доволі часто пропущені рівні заповнюють квазівипадковими числами, розподіленими за нормальним законом, середнє значення і середньоквадратичне відхилення якого збігаються з відповідними характеристиками початкового ряду з пропусками. Крім того, часто моделюють розподіл рівнів цього ряду, генерують за його допомогою псевдовипадкові числа, якими і заповнюють пропущені рівні.

Приклад. Для часового ряду, зображеного на рис. 12.3, найвідповіднішим методом заповнення пропущених рівнів є обчислення їх значень на підставі моделі його тренду.

Нехай вихідний ряд повинен мати n = 80 рівнів, проте в ньому є пропуски, а саме пропущені рівні x_i для таких моментів часу i = 20, 55, 67, 70, 71.

Пропущені рівні відновлено в такий спосіб. Побудовано модель тренду з використанням як апроксимуючої функції полінома третього степеня. Цей вибір зумовлений такими міркуваннями. Оскільки вигляд тренду невідомий, то вибрана апроксимуюча функція повинна уможливити встановлення, принаймні в першому наближенні, загального характеру часового ряду. Наприклад, вона повинна відобразити зростання, спадання, певні можливі зміни тенденції. Поліном третього степеня є фактично кубічною параболою, яка має точку перегину, а це означає, що нею можна "вловити" найзагальніші зміни в тенденції.

Розглянемо на прикладі заповнення (відновлення) пропусків зображеного на рис. 3 часового ряду. Процедура полягає в тому, що спочатку знаходимо тренд вихідного ряду і апроксимуємо його поліномом третього степеня. Визначаємо з рівняння тренду значення першого пропуску і заповнюємо ним перше пропущене значення. Відтак апроксимуємо ряд із заповненим пропуском таким самим поліномом і визначаємо значення отриманої моделі тренду (апроксимуючої функції) для другого пропуску. В такий спосіб заповнюємо всі пропущені значення. В табл. 2 наведено рівняння поліномів для кожного пропущеного значення у послідовності їх черги.

Таблиця 2

Пропуск	Коефіцієнти полінома третього степеня	Заміна рівнів
x_{20}	$= -0{,}000101\,x^3 + 0{,}010220\,x^2 - 0{,}176069\,x + 6{,}462581$	6,221201
x_{55}	$= -0{,}000101\,x^3 + 0{,}010220\,x^2 - 0{,}176072\,x + 6{,}462590$	10,890255
x_{67}	$= -0{,}000101\,x^3 + 0{,}010217\,x^2 - 0{,}175986\,x + 6{,}462109$	10,168455
x_{70}	$= -0{,}000101\,x^3 + 0{,}010217\,x^2 - 0{,}175986\,x + 6{,}462109$	12,203179
x_{71}	$= -0{,}000102\,x^3 + 0{,}010382\,x^2 - 0{,}181927\,x + 6{,}501850$	9,413773

Після завершення цієї процедури графік часового ряду матиме вигляд, як на рис. 4.

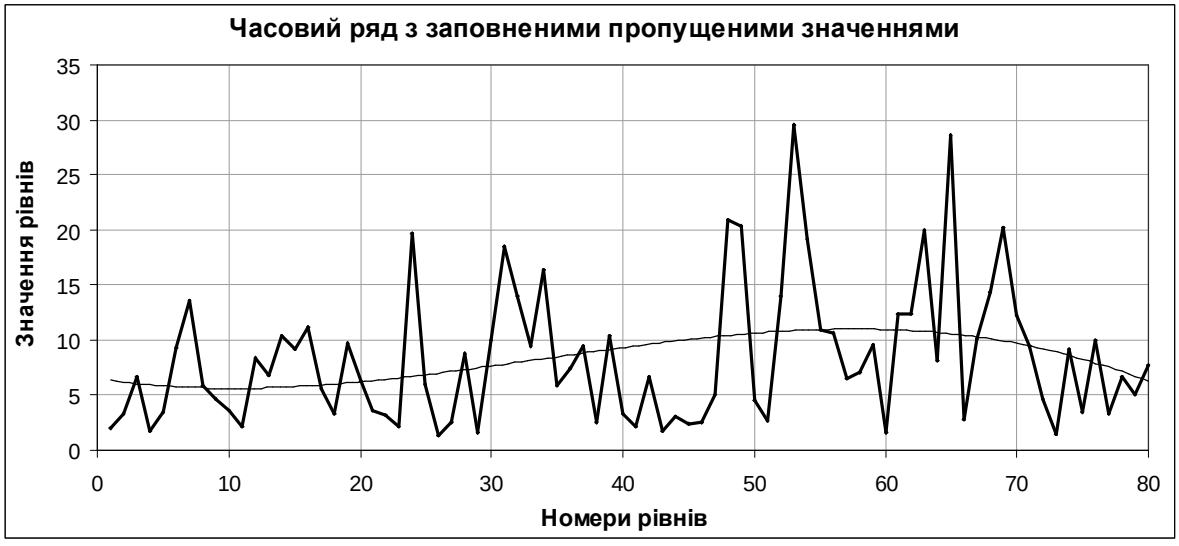

Рис. 4. Часовий ряд із заповненими пропусками

На рис. 4 тонкою лінією зображено тренд, апроксимований поліномом третього степеня. Аналіз коефіцієнтів, що апроксимують тренд, поліномів, показав, що в цьому випадку усі вони дуже мало відрізняються між собою і практично усі лінії трендів зливаються, оскільки відмінність між ними значно менша, ніж можна відобразити за такого масштабу і розмірів пікселів.

Для докладнішого аналізу порівняємо зміни параметрів описової статистики для цього ряду, наведені в табл. 3.

Таблиця 3

Параметри описової статистики	Часовий ряд з пропусками	Після заповнення пропущених значень				
		x_{20}	x_{55}	x_{67}	x_{70}	x_{71}
Середнє	8,235	8,209	8,243	8,268	8,318	8,332
Стандартна помилка	0,739	0,730	0,721	0,712	0,705	0,696
Медіана	6,703	6,688	6,703	6,703	6,703	6,742
Мода	6,703	6,703	6,703	6,703	6,703	6,703
Стандартне відхилення	6,401	6,362	6,327	6,290	6,265	6,227
Дисперсія рівнів	40,967	40,474	40,035	39,563	39,251	38,769
Ексцес	1,660	1,733	1,740	1,766	1,735	1,777
Асиметричність	1,349	1,367	1,355	1,349	1,326	1,326
Інтервал (розмах)	28,187	28,187	28,187	28,187	28,187	28,187
Мінімум	1,360	1,360	1,360	1,360	1,360	1,360
Максимум	29,547	29,547	29,547	29,547	29,547	29,547
Сума	617,625	623,846	634,736	644,905	657,108	666,522
Кількість рівнів	75,000	76,000	77,000	78,000	79,000	80,000

На підставі результатів оцінювання параметрів описової статистики можна зробити висновок, що майже кожен із параметрів “відчуває” зміну часового ряду в разі заміни пропущеного значення значенням, знайденим за допомогою моделі. Не зазнали змін, в межах точності “три знаки після коми”, лише мода, інтервал, мінімальне та максимальне значення. Це можна пояснити тим, що заміна пропусків отриманими з моделей значеннями практично дуже мало впливає на розподіл рівнів часового ряду. Зміну значень ексцесу та асиметричності розподілу можна вважати незначною, оскільки ці параметри визначають особливості форми кривої закону розподілу та функції щільності, які вловити візуально практично неможливо. Інтервал, мінімальне та максимальне значення не змінились тому, що значення моделі лежать всередині інтервалу, тобто не виходять за межі екстремальних значень рівнів.

Висновок

Проблему заповнення пропусків у реальних експериментальних таблицях або індивідуальних часових рядах, що є елементами таких таблиць, можна вирішити різними методами, проте завжди варто спробувати насамперед найпростіші, а також вибрати відповідний критерій відбору як найкращого значення, так і найкращого методу. Найефективнішим способом відновлення пропущених значень є створення штучного часового ряду потрібного обсягу без пропущених значень з аналогічними до оригінального ряду характеристиками, зробити пропуски і на основі його даних вибрати метод відновлення даних, як це і відображено в наведених прикладах.

Науковою новизною можна вважати результат використання того самого підходу до відновлення пропущених значень як для окремих часових рядів, так і для таблиць з часовими рядами однієї і тієї ж структури. Наведені результати отримано під час тестування реципієнтів з-поміж студентів, які виявили інтерес до цієї теми і взяли участь в тестуваннях. Отримані дані за

параметрами описової статистики дали підставу для ідентифікації групи учасників тестування за допомогою агломеративного ієрархічного кластерного аналізу.

1. Литтл Р. Дж. А. Статистический анализ данных с пропусками / Р. Дж. А. Литтл, Д. Б. Рубин. – М.: Финансы и статистика, 1991. – 336 с. 2. Загоруйко Н. Г. Прикладные методы анализа данных и знаний / Н. Г. Загоруйко. – Новосибирск: Изд-во Ин-та математики, 1991. – 278 с. 3. Степашко И. С. Метод відновлення пропущених даних в екологічних задачах на основі МГУА / И. С. Степашко, Ю. В. Коппа, Г. О. Іутинська [Електронний ресурс]. – Режим доступу: http://www. gmdh. net/articles/rus/gaps. pdf. 4. Перемитина Т. О. Программный комплекс восстановления пропущенных значений в многомерных данных на основе методов нечеткого моделирования / Т. О. Перемитина, И. Г. Ященко, С. В. Лучкова // Программные продукты и системы. – 2014. – № 1. – С. – 86–92. 5. Карлов И. А. Восстановление пропущенных данных при численном моделировании сложных динамических систем / И. А. Карлов // Научно-технические ведомости СПбГПУ. Информатика. Телекоммуникации. Управление. – 2013. – 6 (186). – С. 137–144. 6. Слабченко О. О. Інформаційна технологія імпутації даних змішаної природи в задачах інтелектуального аналізу / О. О. Слабченко // Проблеми інформаційних технологій. – 2016. – № 01. – С. 155–161. 7. Плотников Д. Е. Восстановление временных рядов данных дистанционных измерений методом полиномиальной аппроксимации в скользящем окне переменного размера / Д. Е. Плотников, Т. С. Миклашевич, С. А. Барталёв // Современные проблемы дистанционного зондирования Земли из космоса. – 2014. – Т. 11, № 2. – С. 103–110. 8. Братусь О. В. Система підтримки прийняття рішень з адаптивними блоками відновлення та прогнозування сонячних радіофлюксів // Радіоелектроніка, інформатика, управління. – 2017. – № 3. – С. 36–43. 9. Двоенко С. Д. Восстановление пропусков в данных методом неиерархических разбиений // Автомат. и телемех. – 2001. – Вып. 3. – С. 134–140. 10. Honaker J. AMELIA II: a program for missing data / James Honaker, Gary King, Matthew Blackwell // Journal of Statistical Software. – 2011. – Vol. 45, iss. 7. – Mode of access: https://cran. r-project. org/web/packages/Amelia/vignettes/amelia. pdf. 11. Schlomer G. L. Best practices for missing data management in counseling psychology / Gabriel L. Schlomer, Sheri Bauman, Noel A. Card // Journal of Counseling Psychology. – 2010. – Vol. 57, No. 1. – P. 1–10. 12. Soley-Bori M. Dealing with missing data: key assumptions and methods for applied analysis / Marina Soley-Bori // Technical Report. – 2013. – No. 4. –P. 1–20. 13. A comparative study of imputation methods for estimation of missing values of per capita expenditure in central java / Y. Susianto, K. A. Notodiputro, A. Kurnia, H. Wijayanto // IOP Conf. Series: Earth and Environmental Science. – 2017. – 58. – P. 1–10. 14. Sung-Suk Chung. A Study on imputation using adjusted cohen method / Sung-Suk Chung, Young-Min Chun, Sun-Kyung Lee // Journal of the Korean Data & Information Science Society. – 2006. – Vol. 17, No. 3. – P. 871–888. 15. Allison P. D. Missing data / Paul D. Allison. – Mode of access: https://statisticalhorizons. com/wp-content/uploads/2012/01/Milsap-Allison. pdf. 16. Therese D. Pigott. A review of methods for missing data / Therese D. Pigott // Educational Research and Evaluation. – 2001. – Vol. 7, No. 4. – P. 353–383. 17. Бочаров Б. П. Анализ эффективности алгоритма восстановления пропущенных значений временного ряда результатов тестирования знаний / Б. П. Бочаров, М. Ю. Воеводина // Системи обробки інформації. – 2008 – Вип. 3(70). – С. 1–13. 18. Кутлалиев А. Х. Метод множественного восстановления данных / А. Х. Кутлалиев // Социологические методы в современной исследовательской практике: сб. ст., посвященных памяти А. О. Крыштановского [Электронный ресурс]. – М., 2011. – С. 201–207. – Режим доступа: http://www. isras. ru/files/File/Sociologicheskie_methody. pdf. 19. Снитюк В. Е. Эволюционный метод восстановления пропусков в данных: сборник трудов VI Международной конференции "Интеллектуальный анализ информации" [Электронный ресурс]. – К., 2006. – С. 262–271. –

Режим доступу: http://masters. donntu. org/2012/iii/shkarpetkina/library/article2. htm. *20. Абраменкова И. В. Методы восстановления пропусков в массивах данных / И. В. Абраменкова, В. В. Круглов // Программные продукты и системы. – 2005. – № 2. – С. 18 – 22. 21. Злоба Е., Яцкив И. Статистические методы восстановления пропущенных данных / Е. Злоба, И. Яцкив // Computer Modelling & New Technologies, 2002. – Vol. 6, No. 1. – P. 51–61. 22. К вопросу восстановления учетных данных на химических предприятиях / А. В. Волошко, Я. С. Бедерак, Т. Н. Лутчин, М. Ю. Кудрицкий // Известия Томского политехнического университета. – 2014. – Т. 324, № 5. – С. 101–106. 23. Радчикова Е. С. Анализ применения способов заполнения пропусков в данных во временных рядах в экологических исследованиях [Электронный ресурс] / Е. С. Радчикова // Экология и защита окружающей среды: сб. тез. докл. Междунар. науч.-практ. конф., 19–20 марта 2014 г. – Минск, 2014. – С. 112–116. – Режим доступу:* http://elib. bsu. by/handle/123456789/104514. *24. Калинин А. В. Алгоритм восстановления пропусков на поле "плохих" данных / А. В. Калинин, С. В. Ченцов // Сибирский журнал науки и технологий – основное научное издание Сибирского гос. университета науки и технологий им. акад. М. Ф. Решетнева. – 2008. – Т. 18, № 2. – С. 91–95 [Электронный ресурс]. – Режим доступу:* https://cyberleninka. ru/journal/n/sibirskiy-zhurnal-nauki-i-tehnologiy#/939662. *25. Дубинина Е. В. Ежедневная магия* Excel. *Восстановление пропусков в данных / Е. В. Дубинина // Инновационные технологии в науке и образовании [Электронный ресурс]: материалы IX Междунар. науч.-практ. конф. (Чебоксары, 15 янв. 2017 г.): В 2 т. – 2017. – Т. 2, № 1 (9). – Чебоксары: ЦНС "Интерактив плюс", 2017. – С. 29–33. – Режим доступа:* https://interactive-plus. ru/ru/article/116036/discussion_platform.