УДК 519.233:519.21:004.8

С. М. Мельник
О. М. Козачко

АНАЛІЗ ТРАДИЦІЙНИХ МЕТОДІВ ВІДНОВЛЕННЯ ПРОПУСКІВ У МАЛИХ ВИБІРКАХ ДАНИХ

Вінницький національний технічний університет

Анотація

У тезах досліджено проблему відновлення пропусків у малих вибірках технічних даних. На основі комплексного бенчмаркінгу продемонстровано критичну вразливість багатометричних та регресійних алгоритмів відновлення даних (MICE, KNN) до дефіциту інформації. Обґрунтовано необхідність застосування альтернативних непараметричних підходів для збереження структурної репрезентативності даних при оцінюванні станів систем.

Ключові слова: обробка даних, відновлення даних, малі вибірки, структурна невизначеність, MICE, KNN, моделювання систем.

Abstract

The abstract investigates the problem of filling gaps in small samples of technical data. Based on comprehensive benchmarking, the critical vulnerability of multimetric and regression imputation algorithms (MICE, KNN) to information deficit is demonstrated. The necessity of applying alternative nonparametric approaches to preserve the structural representativeness of data when evaluating system states is substantiated.

Keywords: data processing, data imputation, small samples, structural uncertainty, MICE, KNN, system modeling.

У задачах технічної діагностики, моніторингу станів систем та міждисциплінарних дослідженнях часто виникає проблема фрагментованості вибірок. Робота з малими масивами даних істотно ускладнює застосування класичних алгоритмів статистичного відновлення [1, 3]. За умов обмеженої кількості спостережень поодинокі та групові пропуски здатні критично спотворити статистичні характеристики вибірки, що призводить до неадекватних оцінок параметрів системи.

Для об’єктивної оцінки ефективності алгоритмів в умовах інформаційної невизначеності було проведено порівняльне тестування на основі вибірки абсолютних значень параметра стану системи (N=25). У процесі експерименту на еталонній структурі (Ground Truth) було штучно змодельовано як поодинокі пропуски (на індексах спостереження 5, 15, 17 та 20), так і груповий блок пропущених даних (індекси 10–11), що відповідають зонам локальних екстремумів (рис. 1).

Для бенчмаркінгу обрано групу галузевих стандартів: статистичне заміщення (середнім), лінійну інтерполяцію, алгоритм K-найближчих сусідів (KNN) та регресійну множинну імпутацію (MICE).

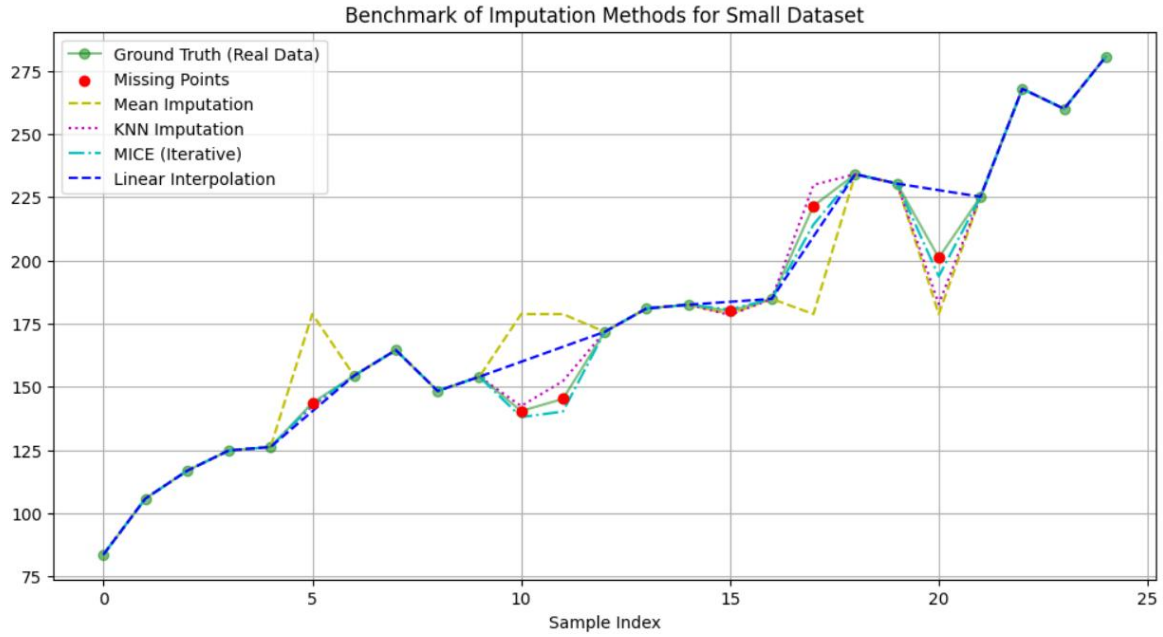

Рис. 1. Порівняльний бенчмаркінг методів відновлення пропусків для малої вибірки

Порівняння відновлених значень з еталонними даними (зелена суцільна лінія) виявило суттєву нестійкість традиційних методів. Лінійна інтерполяція (синя пунктирна лінія) продемонструвала критичну неспроможність відтворювати локальні екстремуми. На ділянках блокових пропусків (наприклад, індекси 15–17) цей метод просто з'єднав крайні відомі точки прямою, повністю «зрізавши» прихований пік. Аналогічно на індексі 20 метод проігнорував глибину локального мінімуму, штучно занизивши варіабельність системи.

Статистичне заміщення середнім значенням (жовта лінія) генерує величезні штучні стрибки (аномалії на індексах 5, 10–11), оскільки алгоритм підставляє глобальне середнє значення замість урахування локального тренду. Своєю чергою, ітераційні методи машинного навчання (MICE та KNN) намагаються відтворити складну динаміку, проте за умов дефіциту інформації працюють нестабільно. Зокрема, на проміжку 15–17 алгоритм KNN генерує завищені значення, а MICE відхиляється від справжньої траєкторії.

Проведений аналіз наочно доводить, що традиційні інтерполяційні та багатокрокові методи не здатні забезпечити адекватну реконструкцію пропусків на малих вибірках. Їх використання призводить або до неприпустимого згладжування критичних екстремумів системи, або до генерації штучних відхилень. Це актуалізує перехід до методології ентропійно-екстремального моделювання. Дана методологія дозволяє реконструювати ймовірнісну структуру вибірки без жорстких параметричних чи лінійних припущень.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Kovtun V., Grochla K., Al-Maitah M., Aldosary S., Kozachko O. Entropy-extreme concept of data gaps filling in a small-sized collection. Egyptian Informatics Journal. 2025. Vol. 29. Article 100621. DOI: https://doi.org/10.1016/j.eij.2025.100621

2. Kovtun V., Altameem T., Al-Maitah M., Kempa W. Entropy-metric estimation of the small data models with stochastic parameters. Heliyon. 2024. Vol. 10, № 2. Article e24708.

3. Luo H., Zhang P., Su J., Zheng D. Evaluation of subsurface soil water content estimate methods: Maximum entropy vs. exponential filter. Journal of Hydrology. 2024. Vol. 643. Article 132007. DOI: https://doi.org/10.1016/j.jhydrol.2024.132007

Мельник Сергій Миколайович — аспірант кафедри системного аналізу та інформаційних технологій; e-mail: serhiimelnyk8@gmail.com;

Козачко Олексій Миколайович — доцент, кандидат технічних наук, доцент кафедри CAIT, Вінницький національний технічний університет, Вінниця, e-mail: pkom@vntu.edu.ua.

Melnyk Serhii M. — Post-graduate student of the Chair of System Analysis and Information Technology, e-mail: serhiimelnyk8@gmail.com;

Kozachko Oleksiy M. — Associate Professor, Candidate of Technical Sciences, Associate Professor of the Department of SAIT, Vinnytsia National Technical University, Vinnytsia, e-mail: pkom@vntu.edu.ua.