

# ДОСЛІДЖЕННЯ ТРАДИЦІЙНИХ СТАТИСТИЧНИХ МЕТОДІВ ВІДНОВЛЕННЯ ДАНИХ В УМОВАХ МАЛИХ ВИБІРОК

Вінницький національний технічний університет

## Анотація

У роботі здійснено критичний аналіз методологічних бар'єрів застосування класичного статистичного інструментарію для відновлення пропусків у малих вибірках. Систематизовано фундаментальні обмеження традиційних підходів: параметричну нестійкість, структурну упередженість та штучне придушення варіабельності. Доведено, що механічне перенесення асимптотичних властивостей великих даних на малі вибірки призводить до виникнення «ілюзії точності» та кумулятивного ефекту похибки в системах підтримки прийняття рішень. Обґрунтовано необхідність зміни парадигми відновлення: від точкової імпутації до реконструкції ймовірного простору станів на основі ентропійно-екстремальних моделей.

**Ключові слова:** малі дані, відновлення пропусків, параметрична нестійкість, ілюзія точності, ентропійний підхід, системи підтримки прийняття рішень.

## Abstract

The paper critically analyzes the methodological barriers to the use of classical statistical tools for restoring missing data in small samples. The fundamental limitations of traditional approaches are systematized: parametric instability, structural bias, and artificial suppression of variability. It is proven that the mechanical transfer of asymptotic properties of big data to small samples leads to the emergence of an "illusion of accuracy" and the cumulative effect of error in decision support systems. The need to change the restoration paradigm is substantiated: from point imputation to the reconstruction of the probability space of states based on entropy-extreme models.

**Keywords:** small data, gap recovery, parametric instability, illusion of accuracy, entropy approach, decision support systems.

## Вступ

Сучасна парадигма аналізу даних та прийняття управлінських рішень переживає трансформацію, зумовлену зростанням ролі досліджень, які проводяться в умовах жорстких інформаційних обмежень. Попри глобальний тренд на Big Data, у багатьох критичних сферах — від клінічної медицини та біоінженерії до аналізу кризових економічних процесів — формування репрезентативних масивів даних часто є неможливим через об'єктивні фінансові, часові або фізичні бар'єри. У таких ситуаціях дослідники стикаються з проблемою «малих даних», де наявність пропусків у спостереженнях має не локальний, а системний вплив на кінцевий результат моделювання [1].

Проблема полягає в тому, що методологічний апарат, який масово використовується для попередньої обробки таких даних, здебільшого запозичений з класичної статистики. Ці методи базуються на законі великих чисел та асимптотичній збіжності оцінок, які гарантують точність лише при значних обсягах вибірки. В умовах малих вибірок ці гарантії втрачають силу, що призводить до методологічного конфлікту між природою даних та інструментами їх аналізу [2].

## Основна частина

Аналіз застосування статистичних алгоритмів (детермінованих, регресійних, імовірнісних) в умовах інформаційної невизначеності дозволяє виділити три групи фундаментальних обмежень, які не можуть бути усунуті шляхом простої алгоритмічної оптимізації:

1. Параметрична надчутливість. Більшість імовірнісних методів відновлення, таких як ЕМ-алгоритм або регресійні моделі, вимагають оцінювання низки параметрів розподілу (середнього, дисперсії, коваріаційної матриці). У малих вибірках виникає критична диспропорція між кількістю невідомих параметрів та обсягом доступних спостережень. Це призводить до ефекту «параметричного резонансу»: модель стає надзвичайно нестійкою, коли додавання або вилучення навіть одного

спостереження може кардинально змінити параметри відновленого розподілу [1]. У прикладних задачах це означає низьку відтворюваність результатів, перетворюючи аналіз на стохастичний процес, залежний від випадкових флуктуацій вибірки [5].

2. Проблема структурної ідентифікації та упередженість. Класичні методи часто апіорі припускають певну форму розподілу даних. У великих масивах даних це припущення можна верифікувати за допомогою статистичних критеріїв згоди. Однак у малих вибірках потужність таких критеріїв є недостатньою для коректного відхилення нульової гіпотези. Як наслідок, дослідник змушений «наосліп» нав'язувати даним теоретичну структуру, яка може не відповідати фізичній природі процесу [2]. Відновлені в такий спосіб дані стають артефактами обраної математичної моделі, а не об'єктивною інформацією про об'єкт, що створює хибне уявлення про закономірності в даних.

3. Зменшення природної мінливості даних. Використання поширених детермінованих методів імпутації (заміщення середнім, медіаною) або простих лінійних регресій призводить до штучного зменшення дисперсії вибірки. Алгоритми ефективно відновлюють «центр» розподілу, але систематично ігнорують його «хвости» — зони екстремальних значень. Для систем підтримки прийняття рішень (СППР), орієнтованих на управління ризиками або виявлення аномалій, таке згладжування є критичним [4]. Воно створює «стерильну» картину процесу, приховуючи реальні загрози та знижуючи чутливість прогностичних моделей до граничних станів системи.

Ігнорування специфіки малих даних призводить до виникнення ланцюгової реакції похибок. Процес відновлення даних не є кінцевою метою, а лише етапом підготовки до подальшого моделювання. Похибки, внесені на етапі відновлення традиційними методами, не зникають, а збільшуються на наступних етапах [5]. Це формує феномен, який можна охарактеризувати як «ілюзія точності»: користувач СППР отримує кількісні оцінки з вузькими довірчими інтервалами, які, однак, не враховують початкову невизначеність відновлених даних [3]. Ухвалення стратегічних рішень на основі такої «надто впевненої» аналітики несе високі ризики, особливо в задачах, де ціна помилки є критичною.

Подолання зазначених обмежень вимагає концептуального переходу від детерміністського «вгадування» пропущених значень до відновлення допустимого ймовірнісного простору станів. Найбільш релевантним інструментом для цього є ентропійно-екстремальні моделі. Застосування принципу максимуму ентропії дозволяє синтезувати розподіл, який задовольняє всім відомим обмеженням, але залишається максимально невизначеним у всіх інших аспектах [1, 3]. Такий підхід забезпечує «методологічну чесність» дослідження: модель не генерує хибного знання там, де інформація відсутня, що дозволяє отримувати стійкі інтервальні оцінки замість оманливо точних точкових значень.

## Висновки

Проведений аналіз засвідчує, що традиційні статистичні методи виявляють суттєву методологічну неспроможність при роботі з малими вибірками через параметричну нестійкість, неможливість структурної верифікації та схильність до викривлення варіабельності. Розвиток методів відновлення даних у сучасних СППР має відбуватися у напрямку формалізації невизначеності через використання ентропійних критеріїв, що забезпечують надійність оцінок без необхідності жорстких апіорних припущень.

## СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. 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. URL: <https://www.cell.com/action/showPdf?pii=S2405-8440%2824%2900739-4>
2. Gongnet E. E., Agbangba C. E., Affossogbe S. A. T., Vihotogbé R., Glèlè Kakai R. Assessing the impact of hard data patterns on Bayesian Maximum Entropy: a simulation study. *Scientific Reports*. 2024. Vol. 14, № 1. Article 28214. DOI: <https://doi.org/10.1038/s41598-024-70518-z>
3. Zhang X., Low Y. M., Koh C. G. Maximum entropy distribution with fractional moments for reliability analysis. *Structural Safety*. 2020. Vol. 83. Article 101904. DOI: <https://doi.org/10.1016/j.strusafe.2019.101904>
4. Sarmadi H., Entezami A., De Michele C. Probabilistic data self-clustering based on semi-parametric extreme value theory for structural health monitoring. *Mechanical Systems and Signal Processing*. 2023. Vol. 187. Article 109976. DOI: <https://doi.org/10.1016/j.ymssp.2022.109976>
5. Tang T., Yuan H. The capacity prediction of Li-ion batteries based on a new feature extraction technique and an improved extreme learning machine algorithm. *Journal of Power Sources*. 2021. Vol. 514. Article 230572. DOI: <https://doi.org/10.1016/j.jpowsour.2021.230572>

**Мельник Сергій Миколайович** — аспірант кафедри системного аналізу та інформаційних технологій; e-mail: [serhiimelnyk8@gmail.com](mailto:serhiimelnyk8@gmail.com);

**Козачко Олексій Миколайович** — доцент, кандидат технічних наук, доцент кафедри САІТ, Вінницький національний технічний університет, Вінниця, e-mail: [pkom@vntu.edu.ua](mailto:pkom@vntu.edu.ua).

**Melnyk Serhii M.** — Post-graduate student of the Chair of System Analysis and Information Technology, e-mail: [serhiimelnyk8@gmail.com](mailto: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](mailto:pkom@vntu.edu.ua).