https://doi.org/10.31891/2219-9365-2025-82-60
УДК 681.128+681.518.2

ІВАЩЕВ Даниїл
Дніпровський національний університет імені Олеся Гончара
https://orcid.org/0009-0002-0556-9418
ivashchev_d@365.dnu.edu.ua
ГЕРАСИМОВ Володимир
Дніпровський національний університет імені Олеся Гончара
https://orcid.org/0000-0002-1366-715X
herasymov_v@365.dnu.edu.ua

АЛГОРИТМ СЕГМЕНТАЦІЇ ДАНИХ ВИМІРЮВАНЬ РІВНЯ ПАЛИВА НА ТРАНСПОРТІ

У статті описано алгоритм сегментації даних про рівень палива транспортних засобів, який використовує згладжування, Z-оцінку, суму абсолютних змін та лінійну апроксимацію для виділення однорідних ділянок цих даних для подальшого аналізу з метою визначення характеристик шуму, обумовленого похибками вимірювання рівня палива, коливання поверхні палива в баку транспортного засобу та іншими впливами на ці дані. Проведено зіставлення результатів сегментації з емпіричними оцінками, що підтвердило надійність підходу для виділення сегментів даних з різними режимами роботи транспортного засобу для подальшого дослідження шуму з метою розробки алгоритмів ефективного позбавлення даних від цього шуму.

Ключові слова: сегментація даних, паливна система, аналіз даних, обробка даних, шум.

IVASHCHEV Daniil, GERASIMOV Vladimir
Oles Honchar Dnipro National University

DATA SEGMENTATION ALGORITHM FOR VEHICLE FUEL LEVEL MEASUREMENTS

This paper presents a data segmentation algorithm designed for analyzing fuel level measurements in vehicles. The proposed approach aims to identify and classify noise caused by sensor inaccuracies, fuel surface oscillations, vibrations, and other external factors that complicate signal interpretation. To achieve this, the algorithm employs a combination of smoothing, Z-score analysis, the sum of absolute changes, and linear approximation techniques, which enable the detection of homogeneous segments within the data. These segments correspond to different operating modes of a vehicle, such as refueling, motion, or idling.

The preprocessing stage includes the removal of non-relevant points and the application of a Gaussian smoothing filter, optimized with a 200-point window. This step reduces short-term fluctuations while preserving significant variations. Subsequently, the segmentation process relies on calculating Z-scores and the sum of absolute changes within overlapping windows to identify potential transition points between modes. Threshold-based selection ensures robustness against insignificant fluctuations, preventing over-fragmentation of the dataset. Detected segments are further classified through linear approximation, which determines the slope and trend type (increasing, decreasing, or stable). Similar neighboring segments are merged to reduce redundancy while maintaining accuracy.

Experimental validation on datasets containing up to one million points demonstrates the reliability of the algorithm. The segmentation results were compared with visual (manual) assessments, showing a high degree of consistency. Notably, the method accurately detects refueling events and effectively distinguishes between noisy motion data and stable idle periods. Statistical analysis confirms that the algorithm provides minimal deviation from empirical boundaries, particularly for stable and decreasing trends, while intentionally extending growing segments to avoid misclassification.

The proposed algorithm proves adaptive, efficient, and resilient to common types of noise. It enables precise division of large datasets into meaningful segments, facilitating subsequent noise analysis and the development of advanced denoising techniques. This, in turn, contributes to improving the accuracy of fuel consumption monitoring and optimization of vehicle fuel systems.

Keywords: data segmentation, fuel system, data analysis, signal processing, noise reduction

Стаття надійшла до редакції / Received 16.04.2025
Прийнята до друку / Accepted 11.05.2025

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

У статті представлено алгоритм сегментації даних вимірювань рівня палива транспортних засобів для їх подальшого аналізу з метою виявлення та класифікації шумів. Сучасні системи моніторингу збирають значні обсяги даних, які часто ускладнені шумами, спричиненими вібраціями, похибками вимірювань та іншими зовнішніми факторами [1]. Запропонований алгоритм сегментації використовує методи згладжування, аналіз Z-оцінки, суми абсолютних змін та лінійної апроксимації для виділення однорідних ділянок даних, що відповідають різним режимам роботи транспортного засобу. Це забезпечує ефективне відокремлення ділянок із мінімальним шумом, таких як заправка, від ділянок із підвищеним шумовим

забрудненням, зокрема під час руху, що є важливим для оптимізації обробки даних та підвищення точності визначення рівня палива на транспортному засобі. Проведено порівняння результатів сегментації з емпіричними оцінками, що підтверджує надійність та адаптивність алгоритму для задач поділу масивів даних на сегменти з різними характеристиками даних вимірювань рівня палива на транспорті.

ФОРМУЛЮВАННЯ ЦІЛЕЙ СТАТТІ

Метою дослідження є створення алгоритму автоматичної сегментації даних вимірювань рівня палива транспортних засобів для виділення ділянок, що відповідають різним режимам роботи (заправка, рух, стоянка), з метою аналізу шумів та створення ефективного методу обробки даних для точного визначення витрат палива транспортом.

Завдання для алгоритму

1. **Виділення сегментів заправки.** Автоматично ідентифікувати ділянки даних, що відповідають процесу заправки, які характеризуються мінімальним рівнем шуму та стабільними умовами, використовуючи методи згладжування (наприклад, гаусів фільтр) та Z-оцінку та сума абсолютних змін для виявлення різких змін рівня палива.
2. **Ідентифікація сегментів руху та стоянки.** Виокремити ділянки даних, пов’язані з рухом або стоянкою транспортного засобу, де присутні шуми, спричинені вібраціями, динамічними впливами чи зовнішніми факторами, шляхом обчислення Z-оцінки та суми абсолютних змін для визначення меж сегментів.
3. **Класифікація сегментів.** Виконати лінійну апроксимацію для кожного сегмента, щоб класифікувати їх за трендом (зростаючий, спадаючий, стабільний), з подальшим об’єднанням схожих сегментів для уникнення надмірної фрагментації.

МЕТОДИ І МАТЕРІАЛИ ДОСЛІДЖЕНЬ

Дослідження базується на даних, отриманих від датчиків рівня палива, які фіксують рівень палива з частотою один вимір за хвилину. Така роздільна здатність забезпечує детальне відображення динаміки роботи паливної системи, але супроводжується значним обсягом шумів і можливих викидів, спричинених похибками вимірювань, коливання поверхні палива чи збоями в передачі даних.

ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

Для налаштування алгоритму було відібрано декілька наборів по 50 000 точок в кожному. Попередня підготовка даних включала:

- Виключення точок із рівнем палива нижче порогового значення, визначеного технічними характеристиками транспортного засобу.
- Згладжування даних для усунення короткочасних коливань, спричинених вібраціями чи похибками вимірювань.

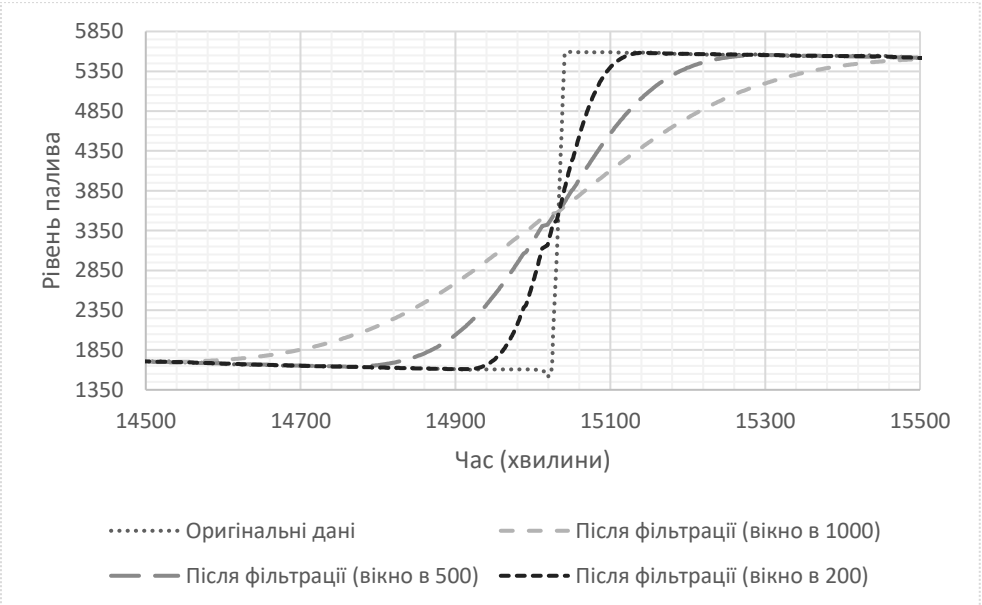

Рис. 1. Порівняння оригінальних і згладжених даних рівня палива для вікон розміром 1000, 500 та 200 точок

Для згладжування даних застосовано гаусів фільтр із вікном у 200 точок, що ефективно пригнічує шум, зберігаючи плавні зміни в даних. Метод використовує зважене середнє значень у вікні, де ваги базуються на гаусовій функції, забезпечуючи оптимальне збереження структури даних [2].

Експериментально встановлено розмір вікна у 200 точок. Вікна розміром 1000 та 500 надмірно згладжують дані, розмиваючи межі заправок, тоді як вікно у 100 точок не забезпечує достатнього приглушення шуму (рис. 1, 2). Розмір 200 точок оптимально поєднує достатньо чітке виділення заправок із ефективною фільтрацією шуму (рис. 2).

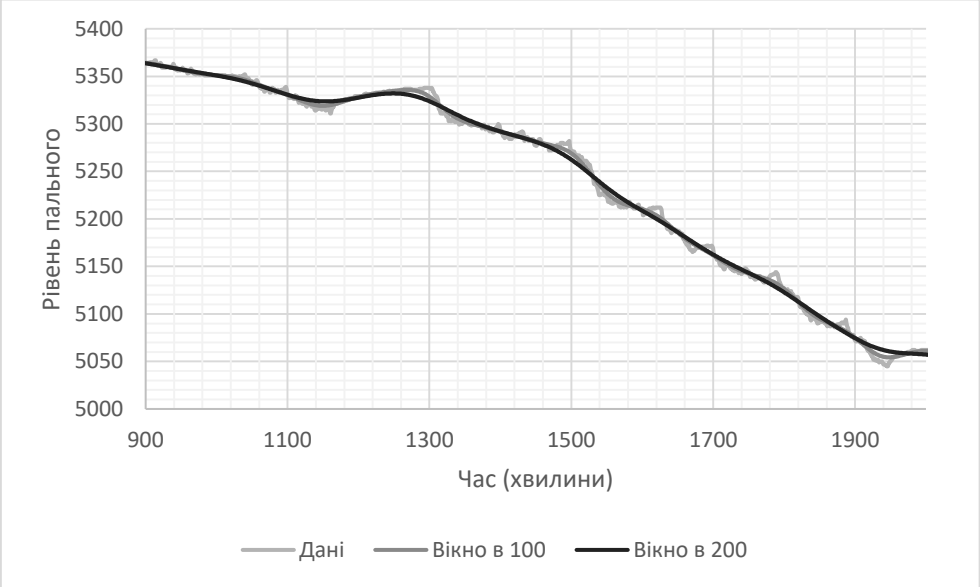

Рис. 2. Порівняння оригінальних даних рівня палива із згладженими для вікон розміром 200 та 100 точок

Для знаходження границь сегментів було розраховано Z-оцінку та суму абсолютних змін даних. Для цього дані згладженого набору вимірювань рівня палива поділяються на перекриваючі вікна розміром 200 точок, із кроком в одну точку. Використання перекриття з кроком в одну точку забезпечує безперервність аналізу, дозволяючи точно виявляти переходи між режимами без втрати важливої інформації.

Z-оцінка потрібна для кількісного визначення значущості змін між сусідніми вікнами даних, що допомагає ідентифікувати потенційні межі сегментів, де відбувається перехід між режимами роботи транспортного засобу, такими як заправка чи рух, шляхом порівняння середніх приростів рівня палива відносно локальної варіативності. Це значення було розраховано за наступною формулою [3]:

$$Z_i = \frac{|\mu_{i+1} - \mu_i|}{\sigma_i}$$

В формулі для розрахунку Z-оцінки застосовані наступні величини: μ_i — середній приріст у вікні i; μ_{i+1} — середній приріст у вікні i+1; σ_i — локальне стандартне відхилення. Локальне стандартне відхилення було розраховано за наступною формулою [4]:

$$\sigma_{\text{local}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{\text{local}})^2}$$
$$\mu_{\text{local}} = \frac{1}{N}\sum_{i=1}^{N} x_i \quad (1)$$

де x_i — значення даних у точці i; N — розмір вікна (кількість точок); μ_{local} — локальне середнє значення у вікні (1).

Сума абсолютних змін даних необхідна для оцінки загальної величини змін рівня палива між двома точками, що дозволяє відфільтрувати незначні коливання, спричинені шумами, і уникнути надмірної фрагментації сегментів, забезпечуючи лише суттєві переходи для класифікації режимів. Локальне стандартне відхилення було розраховано за наступною формулою [4]:

$$\text{C} = \sum_{k=i}^{j-1}|x_{k+1} - x_k|$$

де C — сумарна зміна даних за модулем між точками i та j; x_k — значення даних у точці k.

Якщо Z-оцінка більше 1,25, точка між вікнами позначалася як кандидат на межу сегменту. Це значення було обрано виходячи з даних графіку наведеному на рис. 3, де межа 1,25 відсікає більшість значень.

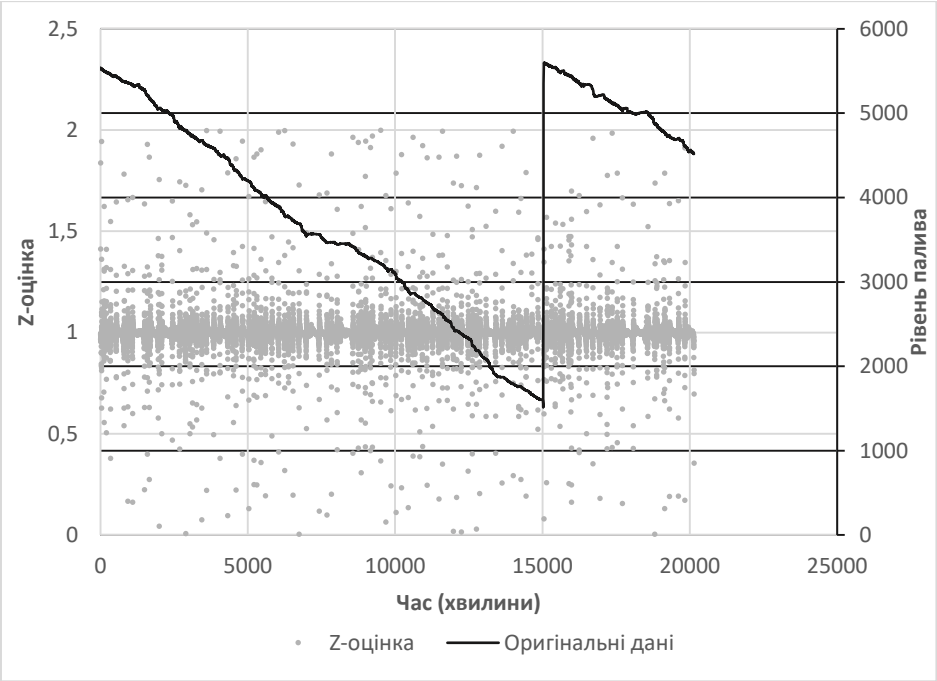

Рис. 3. Z-оцінка та дані рівня пального з потенційними межами

Обчислена сума абсолютних змін даних між поточною точкою перегину та попередньою (4). Якщо отримане значення було меншим за 50 одиниць, така точка виключалась із подальшого аналізу як несуттєва. Порогове значення 50 одиниць визначено як компромісне: при більшому порозі (100 одиниць) частина переходів залишалася не зафіксованою, тоді як менші значення (25 і 10 одиниць) призводили до надлишкового виявлення меж. Візуалізація впливу кількісного значення порогу наведена на рис. 4. Для зручності подання впливу порогу на одному графіку точки було зсунуто по осі у. Для порогу в 25 – на 200 літрів, для 50 – на 400 літрів і для 100 – на 600 літрів.

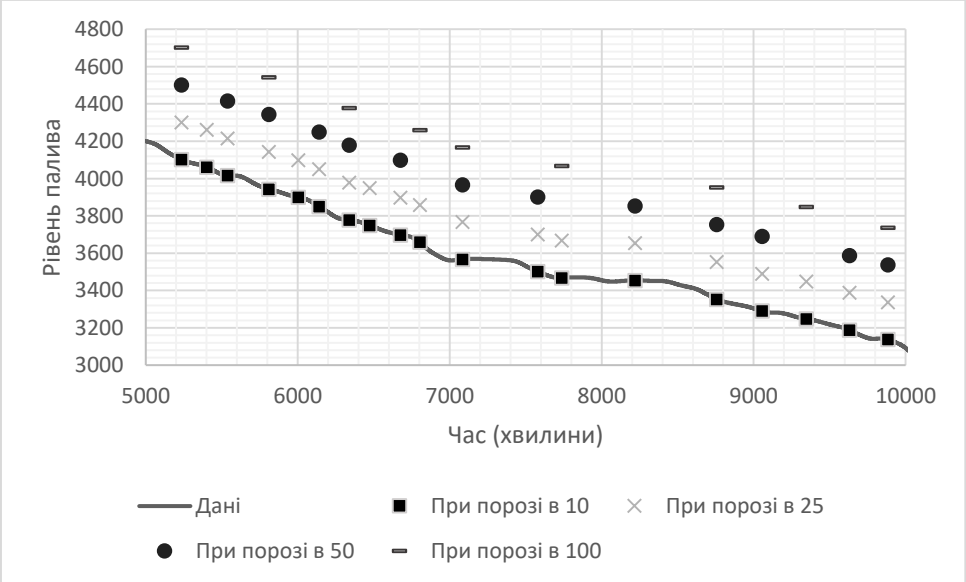

Рис. 4.. Графік оригінального набору даних та меж сегментів при порогах в 10, 25, 50 та 100 одиниць

Експериментальна перевірка розміру перекриваючого вікна в 200 точок необхідна для підтвердження оптимальності цього параметра, оскільки він забезпечує баланс між приглушенням шумів, спричинених вібраціями чи похибками датчиків, і збереженням чіткості меж між вікнами, що відповідають різним режимам

роботи транспортного засобу. Для цього була проведена серія випробувань, результати яких наведено в таблиці 1.

Таблиця 1.

Результати експерименту зі зміною розміру вікна

Розмір вікна	Кількість кандидатів	Характеристика
1000	51	Деякі границі було виявлено з розбіжність більше ніж 1000 точок
500	92	Були виявлені деякі неточність між переходами від стоянки до руху, тобто дані поділилися з помилкою
200	144	Усі переходи до заправок було виявлені, дані добре було поділено
100	175	Всі сегменти з заправкою були поділені порівну

На рис. 5 графічно зображено розділення даних на кандидати меж при вікнах в 100, 200, 500 та 1000 точок. Для подання на одному полотні кандидати на межі було зрушено по осі y: для вікна в 200 – на 200 літрів, для вікна в 500 – на 400 літрів, для вікна в 1000 – на 600 літрів.

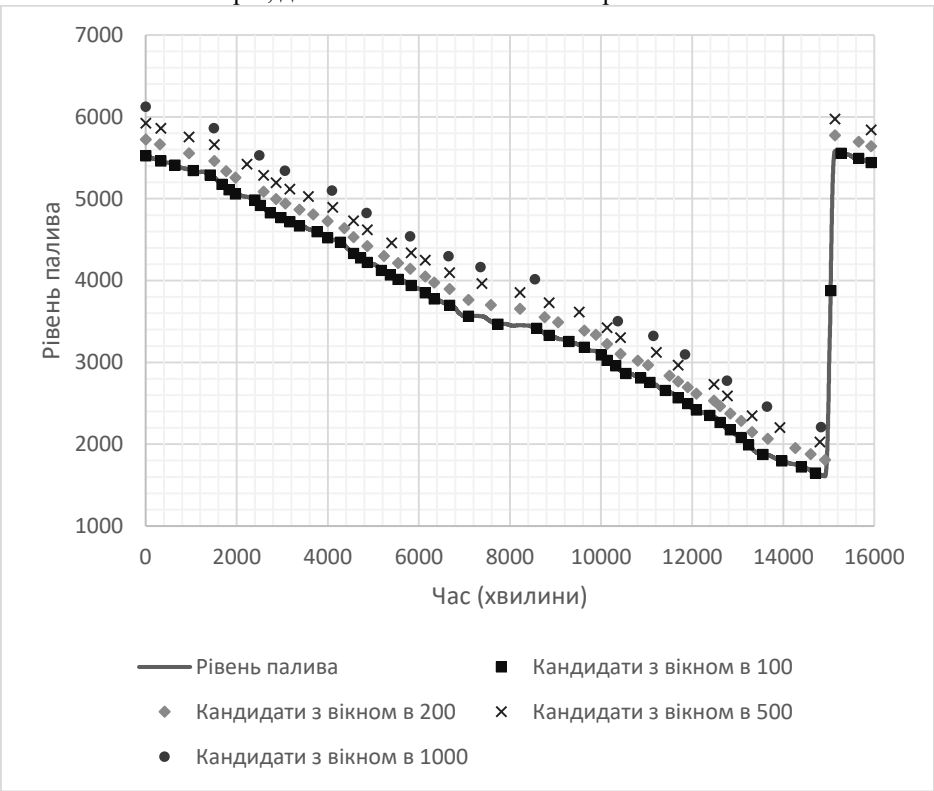

Рис. 5. Графік оригінальних даних рівня палива та кандидатів на межі сегментів для вікон розміром 100, 200, 500 та 1000 точок

Таким чином вдалося з'ясувати, що вікно в 200 точок досягає оптимальний баланс між збереженням деталізації сигналу, що відображає зміни рівня палива, та високою точністю визначення меж між сегментами, які відповідають різним режимам роботи транспортного засобу, таким як заправка, рух чи стоянка.

Наступний крок полягав в тому, щоб усі точки перегину, що пройшли порогові перевірки, були включені до переліку меж сегментів разом із початковою та кінцевою точками даних. На основі сформованого списку виконано поділ сигналу на сегменти, кожен із яких містить згладжені значення рівня палива та відповідні часові мітки.

Для кожного сегменту було здійснено лінійну апроксимацію залежності рівня палива від часу для знаходження коефіцієнта нахилу, за допомогою якого далі була проведена класифікація. Лінійна апроксимація була розрахована за формулою:

$$y(t) = k \cdot t + m$$

де k — це коефіцієнт нахилу, а m — зсув по осі y.

Отримані значення коефіцієнту нахилу використано для класифікації сегментів [5]:

- зростаючий: якщо нахил більше ніж 0,1;
- спадаючий: якщо нахил менше ніж -0,02;
- стабільний: якщо в межах [-0,02, 0,1].

З метою зменшення надмірної фрагментації даних було здійснено об'єднання сусідніх сегментів із подібними характеристиками. Об'єднання виконувалось за умови, що різниця між абсолютними значеннями нахилів не перевищувала 60% від нахилу першого сегменту і при цьому типи тренду збігалися. Після об'єднання дані повторно апроксимувалися, вираховувався коефіцієнт нахилу та тип тренду, а процес класифікації повторювався для всіх оновлених сегментів.

Внаслідок цього отримано оптимізовану структуру сегментів, яка точно відображає характер змін у даних, зокрема тривалі заправки або сталі періоди руху. Для перевірки параметрів об'єднання було проведено серію експериментів із різними порогами різниці нахилу, а саме при значеннях 1; 0,8; 0,6 та 0,4. Результати продемонстрували, що об'єднання сегментів 1 та 0,8 було з надмірною деталізацією, тоді як при порозі 0,6 досягнуто балансу між точністю і узагальненням, а при 0,4 деякі подібні сегменти залишались не об'єднаними (рис. 6). Для відображення впливу порогу на одному графіку точки було зсунуто по осі у. Для нахилу в 0,6 – на 200 літрів, для 0,8 – на 400 літрів і для 1 – на 600 літрів.

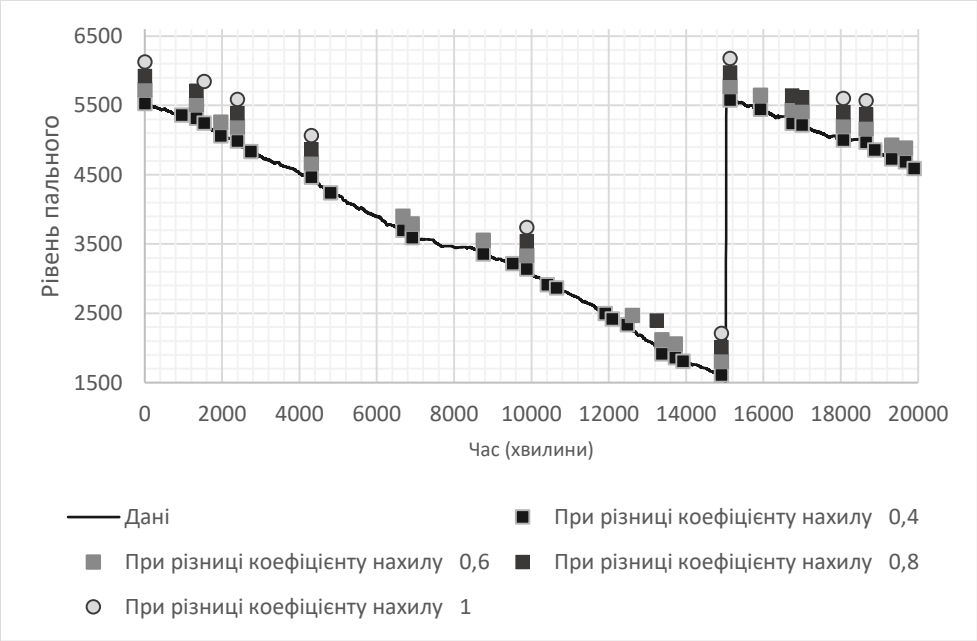

Рис. 6. Графік оригінального набору даних та меж сегментів при різницях коефіцієнту нахилу 0,4, 0,6, 0,8 та 1

Результати досліджень

Для перевірки роботи запропонованого алгоритму було проаналізовано вибірку порядку мільйона точок. На першому етапі виконано розбиття даних вручну на сегменти шляхом візуального аналізу. Надалі результати програмної сегментації було порівняно з візуальними оцінками. У таблиці 2 наведено значення меж сегментів, виділених візуально та програмно.

Таблиця 2.

Значення границь сегментів та їх тип

№ п/п	Візуально				Програмно			
	Початок	Кінець	Довжина	Тип	Початок	Кінець	Довжина	Тип
1	70205	105639	35434	Стоянка	70205	105616	35411	Стоянка
2	105639	108798	3159	Заправка	105616	108826	3210	Заправка
3	108798	164993	56195	Стоянка	108826	164785	55959	Стоянка
4	164993	165039	46	Заправка	164785	165053	268	Заправка
5	165039	179820	14781	Рух	165053	179635	14582	Рух
6	179820	192366	12546	Стоянка	179635	196649	17014	Стоянка
7	192366	197753	5387	Рух	196649	197708	1059	Рух
8	197753	215050	17297	Стоянка	197708	215042	17334	Стоянка
9	215050	217378	2328	Рух	215042	217378	2336	Рух

В цій таблиці, в стовбцях:

- "Початок" та "Кінець" — це порядкові номери точок, що визначають межі сегменту;

- «Довжина» – кількість точок у сегменті;
- «Тип» – класифікація сегменту за трендом (зростаючий, стабільний, спадаючий).

Далі проведено статистичне порівняння між візуально та програмно отриманими межами. Підсумкові дані наведено у таблиці 3.

Таблиця 3.

Статистика порівняння результатів візуального та програмного виділення сегментів

Тип сегмента	$\Delta_{сер}$, початок сегмента	$\Delta_{сер}$, кінець сегмента	$\left(\frac{l_в}{l_п}\right)_{сер}$
Стабільний	−93,33	−143,33	0,94
Зростаючий	135,5	−38,56	7,95
Спадаючий	−64,68	86,93	0,98

Аналіз показав, що для сегментів із зростаючим трендом програмні межі зміщені відносно емпіричних у середньому на 136 точок на початку та на 39 точок на завершенні сегмента. Така поведінка алгоритму забезпечує охоплення всієї фази зростання сигналу й дозволяє уникнути помилкової інтерпретації цих сегментів як ділянок із шумами. У випадках стабільного та спадного тренду спостерігається близька відповідність меж, що підтверджує надійність алгоритму.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

У результаті розробки та тестування алгоритму сегментації даних паливної системи було досягнуто таких результатів:

1. Розроблений комбінований алгоритм демонструє високу ефективність у сегментації даних, адаптивно враховуючи зміни сигналу та локальні особливості, зокрема завдяки використанню Z-оцінки та суми абсолютних змін для точного виявлення змін тренду.
2. Порівняння візуально та програмно визначених меж показало високу відповідність, особливо для спадаючих та статичних ділянок. Алгоритм навмисно розширює межі зростаючих сегментів, забезпечуючи уникнення їх включення до аналізу шумів.
3. Алгоритм є стійким до переважної більшості типових шумів, хоча близько 5% сегментів можуть бути помилково класифіковані внаслідок різких аномалій, що не були повністю згладжені на попередньому етапі обробки.

У підсумку, запропонований метод дозволяє ефективно поділяти сигнал на сегменти з різним режимом роботи транспортного засобу (рух, стоянка, заправка), що є ключовим для якісного аналізу шуму в даних.

References

[1]. Croll, A., Yoskovitz, B. Lean Analytics: Use Data to Build a Better Startup Faster. O'Reilly Media, 2013.
[2]. MathWorks Documentation. Filtering and Smoothing Data [Електронний ресурс]. URL: https://www.mathworks.com/help/curvefit/smoothing-data.html
[3]. Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York, NY: Springer, 2009.
[4]. Proakis, J. G., Manolakis, D. G. Digital Signal Processing: Principles, Algorithms, and Applications. 4th ed. Upper Saddle River, NJ: Pearson Prentice Hall, 2006.
[5]. Paskhaver, B. Data Analysis with Pandas and Python [Video]. Birmingham: Packt Publishing, 2018. URL: https://www.packtpub.com/product/data-analysis-with-pandas-and-python-video/9781788622394
[6]. Müller, A. C., Guido, S. Introduction to Machine Learning with Python: A Guide for Data Scientists. Sebastopol, CA: O'Reilly Media, 2016.
[7]. Suryawanshi, A. R., Bathe, G. N., Patil, S. B. Vehicle Fuel Activities Monitoring System Using IoT. Journal of Transportation Technologies, 2019, Vol. 9, No. 4, pp. 479-489. URL: https://www.scirp.org/journal/paperinformation?paperid=95896
[8]. Brownlee, J. Machine Learning Mastery with Python: Understand Your Data, Create Accurate Models and Work Projects End-To-End. Melbourne: Machine Learning Mastery, 2017. URL: https://machinelearningmastery.com/machine-learning-with-python/
[9]. Zhao, T., Yang, Y., Wang, E. Separated Noise Suppression and Speech Restoration: LSTM-Based Speech Enhancement in Two Stages. IEEE Access, 2019, Vol. 7, pp. 184988-184997. URL: https://ieeexplore.ieee.org/document/8937222