

Харківський національний університет імені В.Н. Каразіна

Факультет математики і інформатики

Кафедра прикладної математики

Кваліфікаційна робота  
**бакалавра**

на тему «Математичні моделі керування рівнем забруднень повітря на  
урбанізованій території»

Виконала:  
студентка групи МП 41  
4 курсу  
спеціальність 113 – прикладна математика  
освітньо-професійна програма  
«Прикладна математика»

Леонова К.О.

Науковий керівник: професор ЗВО, доктор ф.-м.  
наук, Кізілова Н.М.

Рецензент: доктор технічних наук, професор  
кафедри екології, Внукова Н.В.

Харків - 2025 рік

## Анотація

Леонова Ксенія Олександрівна

«Математичні моделі керування рівнем забруднень повітря на урбанізованій території»

Метою цього дослідження є проведення всебічного аналізу концентрацій дев'яти основних забруднювачів повітря в промислових районах України під час військових дій. У дослідженні зібрано описові статистичні дані, які показують чіткі піки нерегулярних викидів та оцінюють природність розподілу. Аналіз часу та спектру показує сезонні цикли та випадкові події (вибухи, пожежі), які впливають на концентрації забруднюючих речовин. Кореляційний аналіз та кластерний аналіз показують, що існує значна кореляція між деякими речовинами, тоді як односторонній дисперсійний аналіз показує, що існують статистично значущі відмінності між сезонними та місячними категоріями. Отримані результати можуть бути використані для прогнозування пікових значень забруднення та розробки ефективних заходів для моніторингу та управління якістю повітря в умовах збройного конфлікту. Ключові слова: атмосферні забруднювачі, описова статистика, часові ряди, кореляційний аналіз.

Leonova Kseniia Oleksandrivna

“Mathematical models for monitoring air pollution levels in urban areas”

The objective of the study was to conduct a comprehensive analysis of the concentration of nine major atmospheric pollutants in industrial areas of Ukraine during hostilities. Descriptive statistics were created that revealed irregular emissions with clear maximum values and assessed the naturalness of the distributions, which, for most materials, deviated from the Gaussian

distribution. Time series and spectral analysis showed seasonal cycles and the effect of occasional events (explosions, fires) on pollutant concentrations. Correlation analysis and cluster analysis showed significant correlations between some pairs of substances and did not show a clear clustering, while one-way analysis of variance showed statistically significant differences by seasonal and monthly categories. The results obtained allow for the prediction of critical pollution periods and the development of effective measures for air quality monitoring and management in the event of armed conflicts.

Keywords: atmospheric pollutants, descriptive statistics, time series, correlation analysis.

## **Зміст**

|                                                                                                                     |           |
|---------------------------------------------------------------------------------------------------------------------|-----------|
| <b>1. Вступ .....</b>                                                                                               | <b>5</b>  |
| <b>1.1 Постановка проблеми.....</b>                                                                                 | <b>6</b>  |
| <b>1.2 Обсяг і методологія .....</b>                                                                                | <b>7</b>  |
| <b>2. Основна частина .....</b>                                                                                     | <b>8</b>  |
| <b>2.1 Підхід до аналізу даних.....</b>                                                                             | <b>8</b>  |
| <b>2.2 Програмне рішення для автоматизованого<br/>завантаження та розбору даних неузгоджених<br/>форматів .....</b> | <b>8</b>  |
| <b>2.3 Обробка відсутніх даних.....</b>                                                                             | <b>9</b>  |
| <b>2.4 Описова статистика .....</b>                                                                                 | <b>10</b> |
| <b>2.5 Кореляційний аналіз .....</b>                                                                                | <b>13</b> |
| <b>2.6 Перевірка на нормальність розподілу .....</b>                                                                | <b>15</b> |
| <b>2.7 Оцінка щільності ймовірності.....</b>                                                                        | <b>17</b> |
| <b>2.8 Згладжування часових рядів .....</b>                                                                         | <b>18</b> |
| <b>2.9 Аналіз коливань.....</b>                                                                                     | <b>20</b> |
| <b>2.10 Спектральний аналіз .....</b>                                                                               | <b>21</b> |
| <b>2.11 Аутлаєри .....</b>                                                                                          | <b>23</b> |
| <b>2.12 Кластерний аналіз .....</b>                                                                                 | <b>26</b> |
| <b>2.13 Аналіз дисперсій .....</b>                                                                                  | <b>28</b> |
| <b>3. Висновки .....</b>                                                                                            | <b>30</b> |
| <b>4. Список використаних джерел .....</b>                                                                          | <b>34</b> |
| <b>Додаток .....</b>                                                                                                | <b>38</b> |

## 1. Вступ

Забруднення повітря в міських районах становить значну загрозу для здоров'я населення, екологічної стабільності та економічного розвитку. Складність динаміки забруднення, на яку впливає поєднання природних та антропогенних факторів, вимагає розробки ефективних стратегій контролю. Математичне моделювання є потужним інструментом для розуміння, прогнозування та управління якістю повітря в містах, що дозволяє приймати обґрунтовані рішення та реалізовувати створення офіційних стратегій, правил чи нормативів, які уряд або відповідні органи впроваджують для покращення якості повітря.

Загальна проблема полягає у побудові математичних моделей, що описують розповсюдження та накопичення забруднюючих речовин у міських середовищах, з урахуванням таких факторів, як джерела викидів, метеорологічні умови, топографія місцевості та хімічні реакції в атмосфері. Ці моделі є необхідними для оптимізації стратегій управління якістю повітря, включаючи плани скорочення викидів, правила зонування та системи моніторингу в режимі реального часу.

Цій темі присвячено значну кількість досліджень, які вивчають підходи теорії керування для динамічного регулювання забруднення та методи асиміляції даних для підвищення точності моделей. Останні роботи акцентують увагу на гібридних моделях, що інтегрують машинне навчання з фізичним моделюванням, забезпечуючи покращені прогностні можливості та адаптивність до складних міських умов.

Незважаючи на значний прогрес, залишаються невирішені проблеми у досягненні високої роздільної здатності, прогнозування якості повітря в

режимі реального часу та ефективної інтеграції моделей в рамки міського планування.

Основною метою цього дослідження є проведення комплексного статистичного аналізу даних про забруднення навколишнього середовища, зібраних у трьох великих промислових містах України: Запоріжжі, Дніпрі та Харкові. Ці міста історично відомі розвитком своєї важкої промисловості, щільною мережею фабрик і заводів та великою кількістю автотранспорту - все це є ключовими чинниками деградації довкілля та забруднення атмосфери.

Мета полягає в змістовному аналізі даних, виявленні тенденцій, визначенні статистично значущих закономірностей та інтерпретації цих закономірностей таким чином, щоб вони могли бути використані в подальшій науковій або політичній роботі.

## **1.1 Постановка проблеми**

Фундаментальною проблемою, яка розглядається в цій роботі, є брак доступних, проаналізованих та інтерпретованих екологічних даних, які можуть бути надійно використані для прийняття рішень або наукових досліджень. Хоча необроблені дані про рівні забруднення можуть бути доступними, вони часто є неструктурованими, непослідовними і їх важко інтерпретувати без належної статистичної обробки.

Дослідження зосереджене на перетворенні необроблених, часто зашумлених наборів даних про забруднення на статистично значущі результати: які забруднювачі присутні, як вони розподілені в часі, чи є кластери високих концентрацій і які тенденції можна визначити за конкретні періоди.

В роботі досліджується, чи демонструють ці міста подібні моделі забруднення, чи є помітні відмінності, які можуть бути пов'язані з промисловою діяльністю, географічним розташуванням або міським плануванням. Розуміння цих факторів за допомогою сучасних статистичних методів є головним внеском цього дослідження.

## **1.2 Обсяг і методологія**

Протягом дослідження були зібрані дані з відкритих джерел про забруднення з загальнодоступних джерел та місцевих систем моніторингу для міст Запоріжжя, Дніпро та Харків. На початковому етапі було проведено очищення та перетворення даних (через невідповідність форматів та відсутність записів).

Основна робота полягала в детальному статистичному аналізі очищених наборів даних. Це включало:

- Описову статистику для розуміння основних характеристик даних.
- Аналіз типів розподілу різних забруднюючих речовин.
- Кластерний аналіз для визначення, чи певні періоди часу демонструють групові моделі забруднення.
- Аналіз часових рядів для виявлення тенденцій, сезонних коливань або аномалій у рівнях забруднення.

## **2. Основна частина**

### **2.1 Підхід до аналізу даних**

У цих розділах представлений детальний аналіз, проведений з використанням різних статистичних методів і методів машинного навчання. Кожен метод наведений у вигляді чіткого визначення, за яким слідує обговорення його переваг і недоліків, огляд часто використовуваних алгоритмів і підсумок отриманих результатів.

Перелік забруднювачів і їх коротке позначення:

- Завислі речовини - SS(Suspended Solids)
- Діоксид сірки – SO<sub>2</sub>
- Оксид вуглецю - CO
- Діоксид азоту – NO<sub>2</sub>
- Оксид азоту - NO
- Фенол - C<sub>6</sub>H<sub>6</sub>O
- Фтористий водень - HF
- Хлористий водень - HCl
- Формальдегід - CH<sub>2</sub>O
- Сажа - CH
- Аміак - NH<sub>3</sub>

### **2.2 Програмне рішення для автоматизованого завантаження та розбору даних неузгоджених форматів**

Щоб проаналізувати дані та глибше їх зрозуміти, спочатку потрібно володіти самими даними. Хоча існує веб-сторінка Міністерства захисту довкілля та природних ресурсів України [2] з архівами, централізованої бази даних екологічної інформації не існує. На сьогоднішній день науковці, які бажають проаналізувати дані про забруднення повітря в Україні, повинні вручну завантажувати та опрацьовувати сотні файлів. Це явно непрофесійно і неефективно.

Щоб вирішити цю проблему, було реалізоване просте, але ефективне рішення з використанням портативних технологій: Python та POSIX-сумісної оболонки.

Це програмне забезпечення значно скорочує час, витрачений на завантаження даних, і дозволяє вченим ефективніше зосередитися на аналізі фактичних даних.

Процес завантаження даних поділяється на кілька етапів:

1. Вилучення даних з веб-сторінки: Сайт[3], [4], [5] містить сотні посилань на .xls файли. Щоб автоматизувати процес і уникнути ручного завантаження, повний HTML-код сторінки можна зберегти і проаналізувати локально за допомогою Bash-скрипту (script.sh, що міститься у додатку до роботи).
2. Завантаження та конвертація файлів: Для кожного витягнутого посилання завантажується відповідний файл .xls і конвертується в .csv, який легше обробляти. Неінформативні заголовки видаляються, а файли об'єднуються.
3. Скрипт Python витягує діапазон дат з кожного імені файлу і коригує відповідні стовпці.

Це рішення спрощує весь процес збору даних і готує набір даних для подальшого аналізу.

## 2.3 Обробка відсутніх даних

Визначення:

Відсутні дані — це значення, яких немає в наборі даних, що може спотворити результати статистичного аналізу та машинного навчання.

Використаний метод: Хронологічні k-найближчі сусіди ( $k = 2$  або  $3$ ) на основі близькості за днями або тижнями.

Інші методи:

- Усереднення/медіана: Простий і швидкий, але може недооцінювати мінливість.

- Лінійна інтерполяція: зберігає тенденції, але не підходить для нелінійних пропусків.
- Сплайн-інтерполяція: фіксує нелінійні закономірності, але може надмірно підганяти шум.
- Багаторазове введення : статистично надійне, але обчислювально інтенсивне.
- Введення на основі моделі (наприклад, з використанням регресії або машинного навчання): адаптивне і точне, але може вимагати маркованих даних і налаштування.

Переваги імпутації KNN:

1. Враховує локальні часові закономірності.
2. Проста у впровадженні.

Недоліки:

Припускає, що подібні часові точки мають подібні значення, що може не відповідати дійсності в нестабільні періоди.

## 2.4 Описова статистика

Визначення:

Описова статистика надає підсумок основних характеристик набору даних, включаючи показники центральної тенденції (середнє, медіана), дисперсії (стандартне відхилення) та форми (асиметрія, ексцес).

У підручнику [6] було введено позначення(розділ 2, стор. 84; розділ 3, стор. 195, стор. 207):

- **Середнє значення** - це величина, що представляє «центр» набору чисел і є проміжною по відношенню до крайніх значень набору чисел.
- **Медіана** набору чисел - це значення, що відокремлює верхню половину від нижньої половини вибірки даних, генеральної сукупності або розподілу ймовірностей.
- **Мода** - це значення, яке найчастіше з'являється в наборі значень даних
- **Ексцес** - це описова статистика, яка використовується для вимірювання того, як дані розсіюються між центром і хвостами розподілу

- **Асиметрія** - це міра асиметрії або спотворення симетричного розподілу. Він вимірює відхилення даного розподілу випадкової величини від симетричного розподілу
- **Стандартне відхилення** - це міра відхилення значень змінної від її середнього значення

Використаний метод:

Для кожного вибраного параметра були обчислені класичні статистичні показники, такі як середнє значення, медіана, мода, мінімальне значення, максимальне значення, ексцес, асиметрія та стандартне відхилення.

Візуалізація отриманих результатів представлена на таблицях 1, 2, 3 та рисунках 1, 2, 3.

*Таблиця 1. Описова статистика для концентрації забруднень у місті Дніпро за період часу 2023-2025 роки*

|                          | SS     | SO2    | CO     | NO2    | NO     | C6H6O  | NH3    | CH2O   |
|--------------------------|--------|--------|--------|--------|--------|--------|--------|--------|
| середнє значення         | 1,746  | 0,231  | 0,703  | 1,726  | 0,594  | 0,734  | 0,975  | 4,542  |
| медіана                  | 1,736  | 0,213  | 0,694  | 1,686  | 0,583  | 0,722  | 0,938  | 4,037  |
| мода                     | 1,583  | 0,169  | 0,667  | 1,500  | 0,500  | 0,806  | 0,875  | 3,444  |
| мінімальне значення      | 0,833  | 0,104  | 0,361  | 0,990  | 0,250  | 0,083  | 0,250  | 2,278  |
| максимальне значення     | 4,500  | 1,049  | 1,139  | 3,500  | 1,111  | 2,074  | 1,917  | 9,259  |
| ексцес                   | 3,264  | 21,956 | 0,736  | 2,074  | 1,089  | 1,347  | 0,161  | 0,095  |
| асиметрія                | 0,853  | 2,980  | 0,525  | 0,887  | 0,290  | 0,539  | 0,476  | 0,937  |
| стандартне відхилення    | 0,398  | 0,074  | 0,104  | 0,321  | 0,121  | 0,237  | 0,245  | 1,466  |
| стандартне відхилення(%) | 22,806 | 31,980 | 14,796 | 18,587 | 20,401 | 32,312 | 25,108 | 32,264 |

*Таблиця 2. Описова статистика для концентрації забруднень у місті Харків за період часу 2023-2025 роки*

|                          | SS      | SO2     | CO      | NO2     | NO      | C6H6O   | CH      | NH3     | CH2O    |
|--------------------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| середнє значення         | 0,3929  | 0,3059  | 0,4294  | 0,7077  | 0,3926  | 0,7283  | 0,4498  | 0,1216  | 0,7912  |
| медіана                  | 0,3500  | 0,2765  | 0,4250  | 0,7042  | 0,3721  | 0,7111  | 0,4000  | 0,1149  | 0,7702  |
| мода                     | 0,2500  | 0,2345  | 0,4250  | 0,7447  | 0,3800  | 0,9139  | 0,4000  | 0,0437  | 0,5226  |
| мінімальне значення      | 0,0056  | 0,1102  | 0,3083  | 0,4239  | 0,0671  | 0,1958  | 0,0000  | 0,0000  | 0,1190  |
| максимальне значення     | 1,3440  | 1,0207  | 0,6500  | 1,1174  | 1,7489  | 1,5407  | 1,2667  | 0,3188  | 1,9926  |
| ексцес                   | 2,9314  | 5,7039  | 0,9521  | -0,1005 | 21,4071 | 0,3442  | -0,6554 | -0,2590 | 0,8253  |
| асиметрія                | 1,2277  | 1,9758  | 0,6422  | 0,2746  | 3,1809  | 0,4388  | 0,1814  | 0,4639  | 0,5742  |
| стандартне відхилення    | 0,2058  | 0,1292  | 0,0477  | 0,1210  | 0,1587  | 0,2272  | 0,2710  | 0,0621  | 0,3047  |
| стандартне відхилення(%) | 52,3759 | 42,2188 | 11,1134 | 17,0923 | 40,4135 | 31,1875 | 60,2552 | 51,0413 | 38,5098 |

*Таблиця 3. Описова статистика для концентрації забруднень у місті Запоріжжя за період часу 2020-2021 роки*

|                          | SS     | SO2    | CO     | NO2    | NO     | C6H6O  | HF     | HCl    | CH2O   |
|--------------------------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| середнє значення         | 0,757  | 0,151  | 0,402  | 1,809  | 0,850  | 2,078  | 0,116  | 0,195  | 1,671  |
| медіана                  | 0,700  | 0,146  | 0,367  | 1,717  | 0,828  | 2,002  | 0,100  | 0,200  | 1,600  |
| мода                     | 0,700  | 0,100  | 0,333  | 1,400  | 0,700  | 1,800  | 0,000  | 0,200  | 1,400  |
| мінімальне значення      | 0,067  | 0,100  | 0,300  | 1,100  | 0,382  | 1,300  | 0,000  | 0,092  | 0,200  |
| максимальне значення     | 1,800  | 0,247  | 0,733  | 3,371  | 1,649  | 3,277  | 0,460  | 0,400  | 3,392  |
| ексцес                   | 0,830  | -1,020 | 0,587  | 1,382  | 1,461  | -0,368 | 0,149  | 2,411  | 0,217  |
| асиметрія                | 0,594  | 0,312  | 1,157  | 1,088  | 0,851  | 0,558  | 0,756  | 0,706  | 0,494  |
| стандартне відхилення    | 0,281  | 0,041  | 0,104  | 0,400  | 0,200  | 0,436  | 0,105  | 0,049  | 0,576  |
| стандартне відхилення(%) | 37,041 | 26,797 | 25,832 | 22,122 | 23,512 | 20,978 | 90,890 | 25,231 | 34,452 |

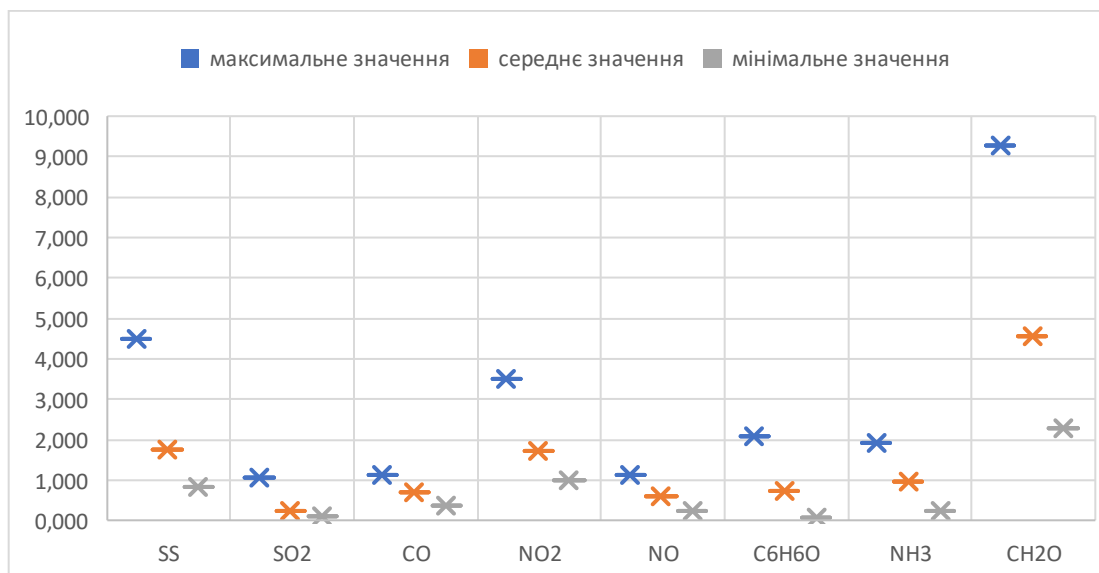


Рис.1. Середні, максимальні і мінімальні значення концентрацій забруднювачів повітря у місті Дніпро за період часу 2023-2025

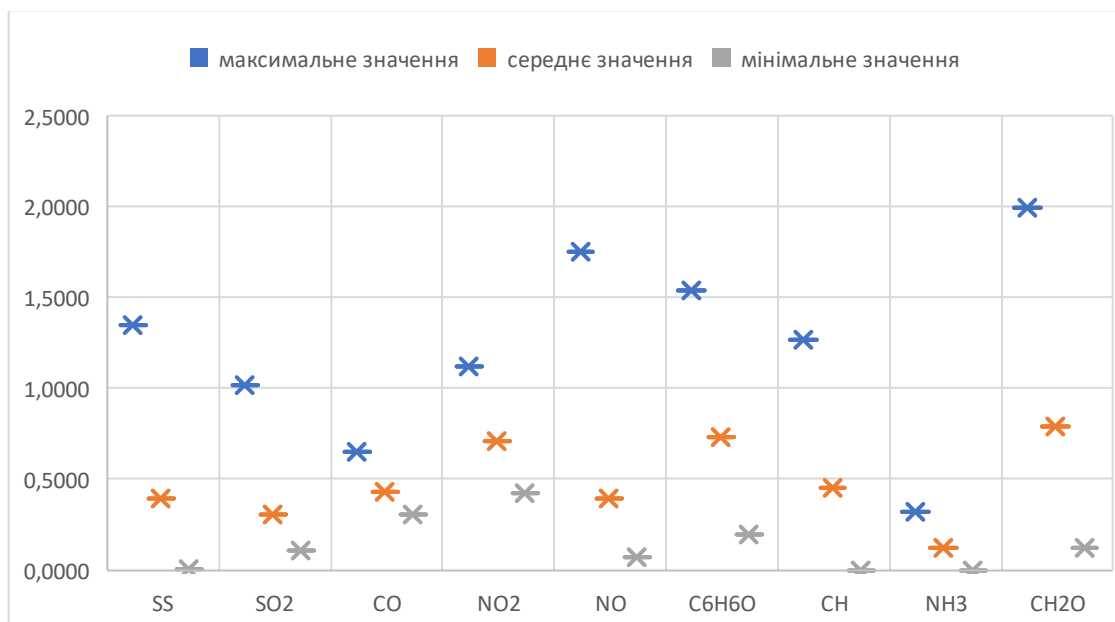
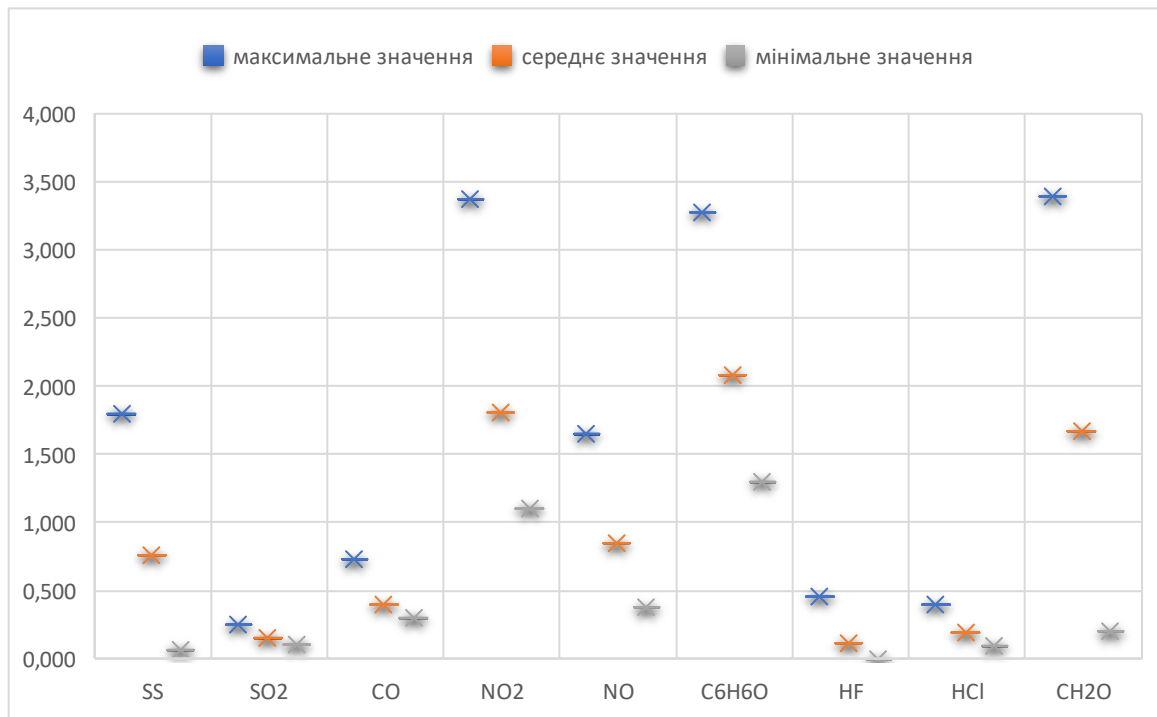


Рис.2. Середні, максимальні і мінімальні значення концентрацій забруднювачів повітря у місті Харків за період часу 2023-2025



*Рис.3. Середні, максимальні і мінімальні значення концентрацій забруднювачів повітря у місті Запоріжжя за період часу 2020-2021*

## 2.5 Кореляційний аналіз

Визначення:

Кореляційний аналіз оцінює силу та напрямок взаємозв'язку між змінними, найчастіше за допомогою коефіцієнта кореляції Пірсона.

Використаний метод:

Були розраховані[6](розділ 3, стор. 232) парні коефіцієнти кореляції Пірсона між показниками забруднення повітря. Для побудови регресійного рівняння була проведена апроксимація даних з використанням методу найменших квадратів.

Метод найменших квадратів[6](розділ 3, стор. 217) - це метод оцінювання параметрів у регресійному аналізі, заснований на мінімізації суми квадратів відхилень (різниці між спостережуваним значенням і значенням, що підбирається за допомогою моделі), отриманих в результаті кожного окремого рівняння. (Простіше кажучи, метод найменших квадратів - це математична процедура для знаходження кривої, що найкраще

відповідає заданому набору точок, шляхом мінімізації суми квадратів зміщень точок від кривої).

Альтернативні методи:

- Рангова кореляція Спірмена: фіксує монотонні (нелінійні) взаємозв'язки.

Переваги коефіцієнта кореляції Пірсона[7](стор. 9):

- Виявляє лінійні залежності
- Корисний для початкового відбору ознак та виявлення мультиколінеарності

Недоліки коефіцієнта кореляції Пірсона:

- Не виявляє нелінійні або часові затримки у взаємозв'язках
- Припускає лінійність та гомоскедастичність

На рисунку 4 показано результати, отримані в ході аналізу.

*Таблиця 4. Матриця кореляцій забруднювачів у місті Дніпро за період часу 2023-2025 роки*

|                                 | SO <sub>2</sub> | CO    | NO <sub>2</sub> | NO    | C <sub>6</sub> H <sub>6</sub> O | NH <sub>3</sub> | CH <sub>2</sub> O |
|---------------------------------|-----------------|-------|-----------------|-------|---------------------------------|-----------------|-------------------|
| SS                              | 0,007           | 0,372 | 0,503           | 0,456 | 0,263                           | 0,560           | 0,669             |
| SO <sub>2</sub>                 |                 | 0,204 | -0,123          | 0,072 | 0,039                           | 0,151           | 0,019             |
| CO                              |                 |       | 0,684           | 0,732 | 0,607                           | 0,500           | 0,533             |
| NO <sub>2</sub>                 |                 |       |                 | 0,839 | 0,564                           | 0,604           | 0,701             |
| NO                              |                 |       |                 |       | 0,640                           | 0,665           | 0,701             |
| C <sub>6</sub> H <sub>6</sub> O |                 |       |                 |       |                                 | 0,473           | 0,469             |
| NH <sub>3</sub>                 |                 |       |                 |       |                                 |                 | 0,874             |

Кореляції двох речовин- на перетині. Виділимо 3 типи:

- <0,1-залежності немає
- 0,1-0,5-слабка кореляція
- >0,5-значуща кореляція

## 2.6 Перевірка на нормальність розподілу

Щоб забезпечити достовірність статистичного аналізу, який передбачає нормальність (наприклад, ANOVA), важливо перевірити, чи підпорядковуються дані нормальному розподілу. З цією метою був застосований критерій узгодженості  $\chi^2$ [8](розділ 11, стор. 199) до кожного забруднювача.

Введемо наступні гіпотези:

H<sub>0</sub>: Дані відповідають нормальному розподілу

H<sub>1</sub>: Дані не відповідають нормальному розподілу

Ми відкидаємо H<sub>0</sub>, якщо р-значення <  $\alpha$  (0,05) Для 8 різних забруднювачів були отримані наступні результати у таблиці 5 та на рисунку 4:

*Таблиця 5. Значення  $\chi^2$  для різних забруднювачів у місті Дніпро за період часу 2023-2025 роки*

| Variable                        | $\chi^2$ | p-value  |
|---------------------------------|----------|----------|
| SS                              | 99.595   | 0.000003 |
| SO <sub>2</sub>                 | 189.251  | 0.000000 |
| CO                              | 5.825    | 0.000045 |
| NO <sub>2</sub>                 | 2.632    | 0.000000 |
| NO                              | 357.284  | 0.154712 |
| C <sub>6</sub> H <sub>6</sub> O | 5.414    | 0.061880 |
| NH <sub>3</sub>                 | 45.703   | 0.000000 |
| CH <sub>2</sub> O               | 6.207    | 0.000027 |

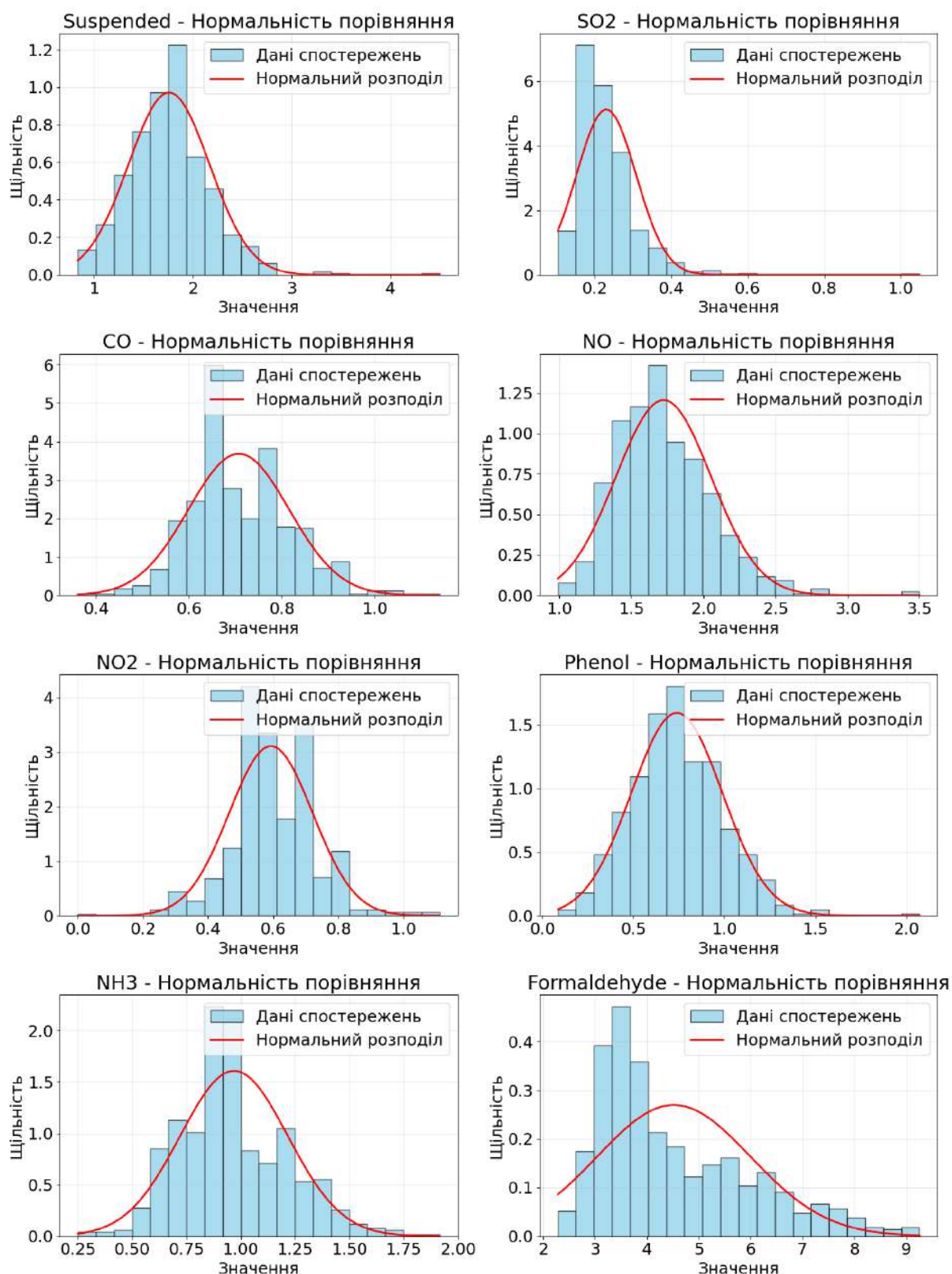


Рисунок 4. Гістограма щільності для різних забруднювачів у місті Дніпро за період часу 2023-2025 роки

Отже, з 8 забруднювачів, 2 підпорядковуються нормальному розподілу.

## 2.7 Оцінка щільності ймовірності

Визначення:

Функція щільності ймовірності – це математична функція, яка описує ймовірність, з якою випадкова змінна, що змінюється безперервно, прийме конкретне значення в заданому діапазоні. Вона використовується для опису ймовірнісного розподілу безперервних випадкових змінних.

Використаний метод:

У цій роботі щільність ймовірності значень забруднення повітря була розрахована за формулою[6](розділ 2, стор. 118) нормальної (гаусової) функції щільності ймовірності на основі вибіркового середнього та стандартного відхилення. Таким чином, перевіряється, наскільки емпіричний розподіл відповідає нормальному.

Альтернативні методи оцінки:

- Оцінка на основі гістограми[9](розділ 3, стор. 140): метод передбачає розбиття діапазону значень на інтервали (біни) та підрахунок кількості спостережень у кожному з них. Він є простим у реалізації, але має обмеження — результати залежать від вибраної ширини бінів і їх вирівнювання: різні параметри можуть суттєво змінити вигляд оцінки щільності, що ускладнює інтерпретацію та порівняння.

Переваги оцінки нормального розподілу:

- Легко обчислюється за допомогою базових інструментів
- Надає стандартний еталон для порівняння.
- Корисний, якщо дані є приблизно симетричними та мають дзвоноподібну форму.

Недоліки:

- Припускає нормальність, яка може не відповідати реальним екологічним даним (наприклад, забруднення часто має асиметричний розподіл)

- Може спотворювати структуру ймовірностей, якщо дані є мультимодальними або мають важкі хвости.

Результати розрахунків функції щільності ймовірності для зважених твердих частинок, оксиду вуглецю, діоксиду азоту, оксиду азоту, фенолу та аміаку у місті Дніпро у 2023-2025 роки наведені на Рис.5.

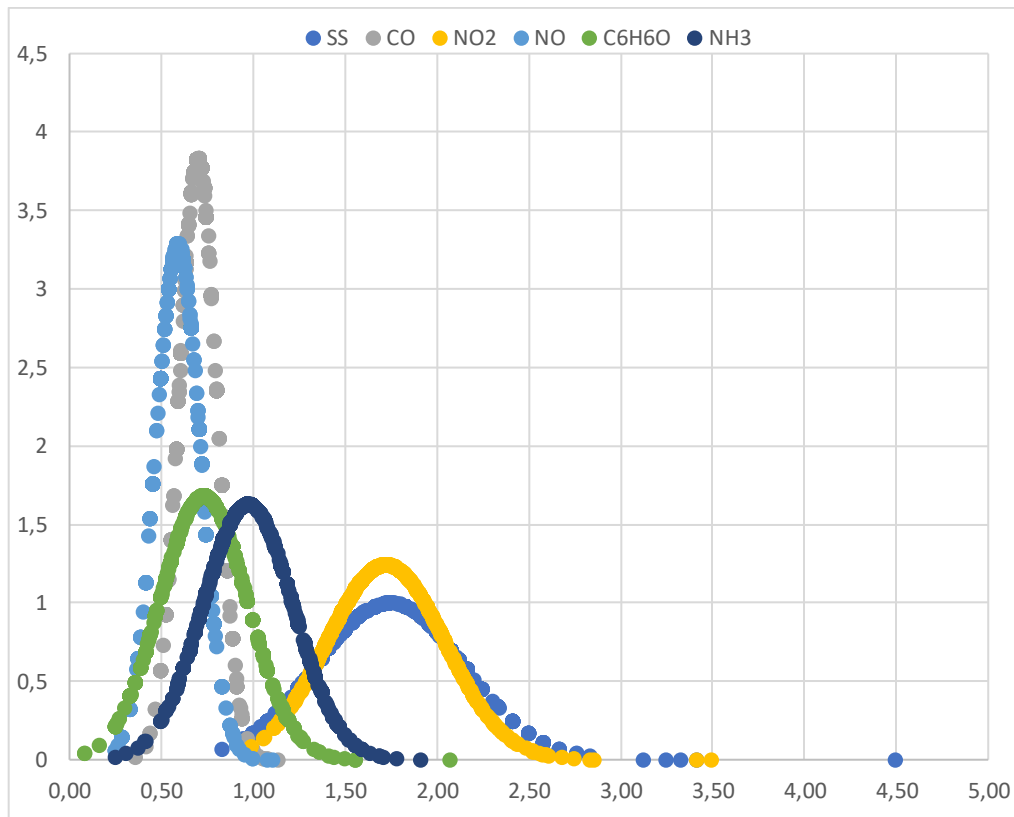


Рисунок 5. Функція щільності ймовірності для даних міста Дніпро у 2023-2025 роки

## 2.8 Згладжування часових рядів

Визначення:

Методи згладжування використовуються для зменшення шуму та виділення основних закономірностей або тенденцій у часових рядах даних.

Використаний метод:

Ми застосували згладжування за допомогою ковзної середньої[10](розділ 3), щоб відфільтрувати короткострокові коливання та краще спостерігати за довгостроковою динамікою рівнів забруднення.

Альтернативні методи:

- Експоненціальне згладжування: надає більшу вагу останнім даним; добре підходить для прогнозування.

Переваги:

- Простий і ефективний для візуалізації тенденцій.
- Покращує інтерпретованість часових закономірностей.

Недоліки:

- Може розмити реальні короткострокові коливання.
- Вибір розміру вікна впливає на результат.

Графічне відображення результатів після згладжування представлено на рисунку 6:

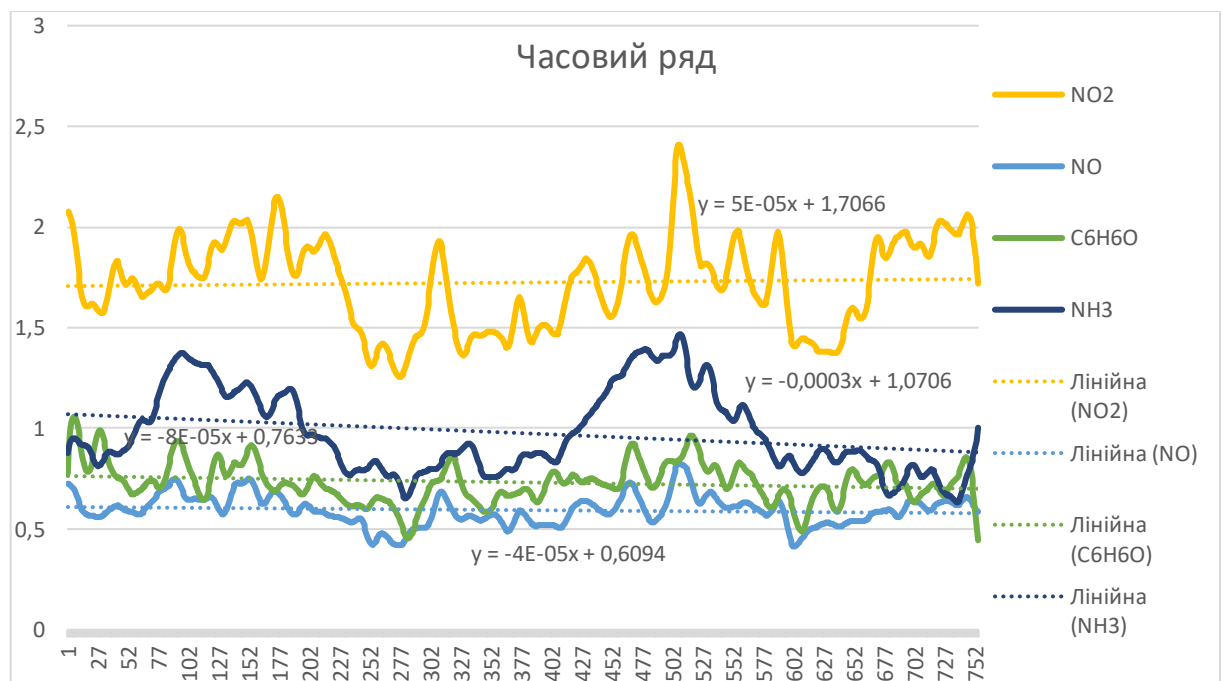


Рисунок 6. Згладжений часовий ряд для даних міста Дніпро у 2023-2025 роки

Згладжування даних дозволило виокремити загальні тенденції в рівнях забруднення повітря, приховані за короткостроковими коливаннями. Отримані криві згладженого ряду забезпечують чіткіше уявлення про довгострокову динаміку, що є особливо корисним для виявлення сезонних змін або стійких зростань/спадів. Це покращує інтерпретацію та може слугувати основою для подальшого аналізу або прогнозування.

## 2.9 Аналіз коливань

Визначення:

Коливання в часовому ряді відображають відхилення спостережуваних значень від базового рівня, що зазвичай використовується для оцінки коливань навколо тренду або середнього значення.

Використаний метод:

У цьому дослідженні коливання були розраховані[13](розділ 5, стор. 31) як різниця між кожним спостережуваним значенням і загальним середнім значенням часового ряду за весь період спостереження. Ці відхилення були зображені у вигляді окремого ряду, щоб візуалізувати, як рівні забруднення коливаються навколо довгострокового середнього значення.

Альтернативні методи[12](розділ 1, стор. 7):

- Розкладення на сезонні та трендові компоненти (наприклад, STL): розділяє часовий ряд на трендові, сезонні та залишкові компоненти.
- Виявлення циклів (аналіз піків/мінімумів): визначає конкретні періодичні структури або режими в даних.

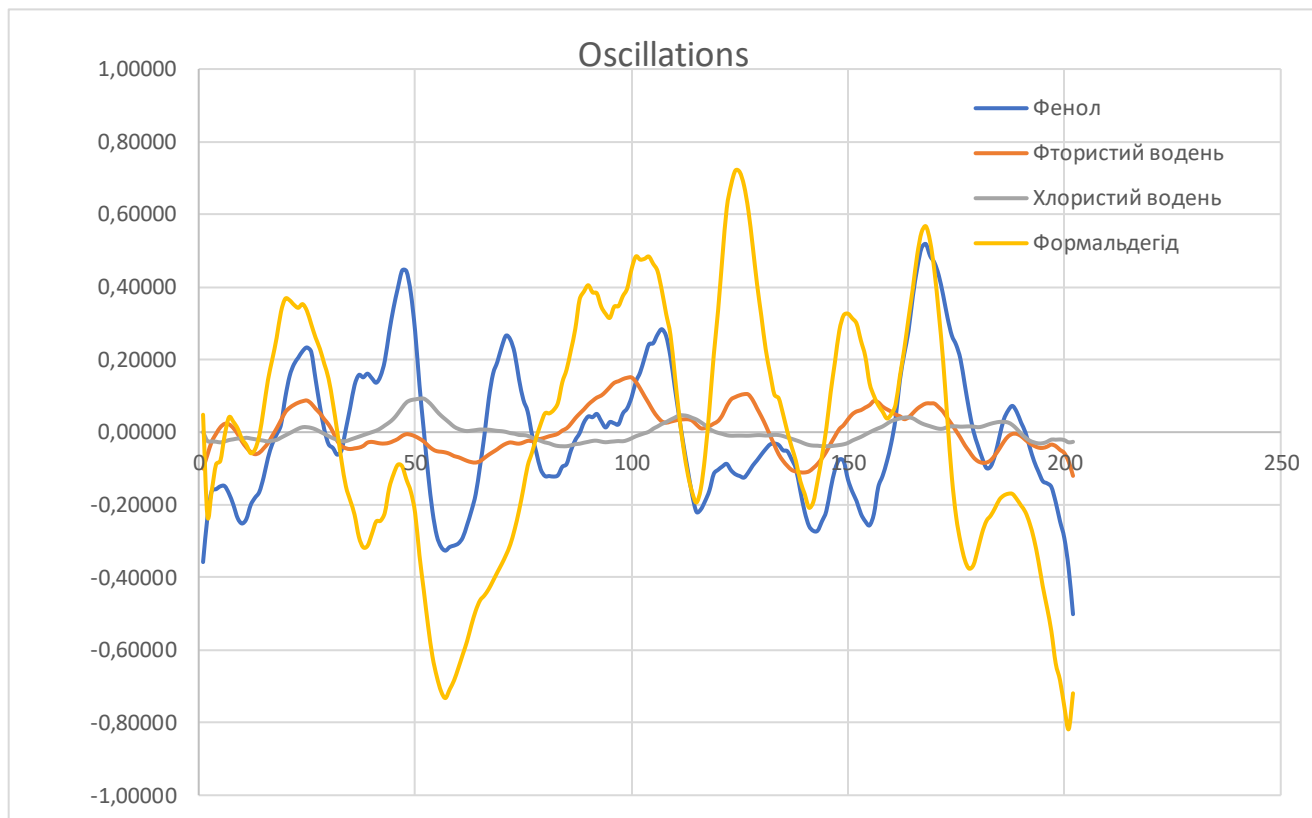
Переваги застосованого методу:

- Простий та інтуїтивно зрозумілий
- Виділяє відхилення від середнього рівня
- Корисний для швидкої візуальної оцінки мінливості

Недоліки:

- Не фіксує сезонні або циклічні структури явно
- Може пропустити часові залежності або інформацію про частоту

Візуалізація отриманих результатів представлена на рисунку 7:



*Рисунок 7. Графіки коливань за 200 діб для даних міста Запоріжжя у 2020-2021 роки*

## 2.10 Спектральний аналіз

Визначення:

Спектральний аналіз використовується для ідентифікації домінуючих частот або періодичних компонентів у часовому ряді шляхом перетворення даних з часової області в частотну область.

Використаний метод:

У цьому дослідженні для аналізу частотного спектру часового ряду забруднення повітря було використано перетворення Фур'є[14](розділ 7, стор. 321), а саме швидке перетворення Фур'є (FFT). FFT дозволило нам розкласти сигнал на синусоїдальні компоненти та виявити домінуючі періодичності протягом спостережуваного періоду.

Переваги методу Фур'є:

- Ефективний і добре відпрацьований

- Забезпечує чітке уявлення про глобальні періодичності
- Підходить для стаціонарних або регулярно відібраних сигналів

Обмеження:

- Припускає стаціонарність сигналу протягом усього періоду
- Не вказує, коли виникають конкретні частоти (відсутність часової локалізації)
- Чутливий до шуму та крайових ефектів

Результати дослідження наведено на відповідному рисунках 8 та 9:

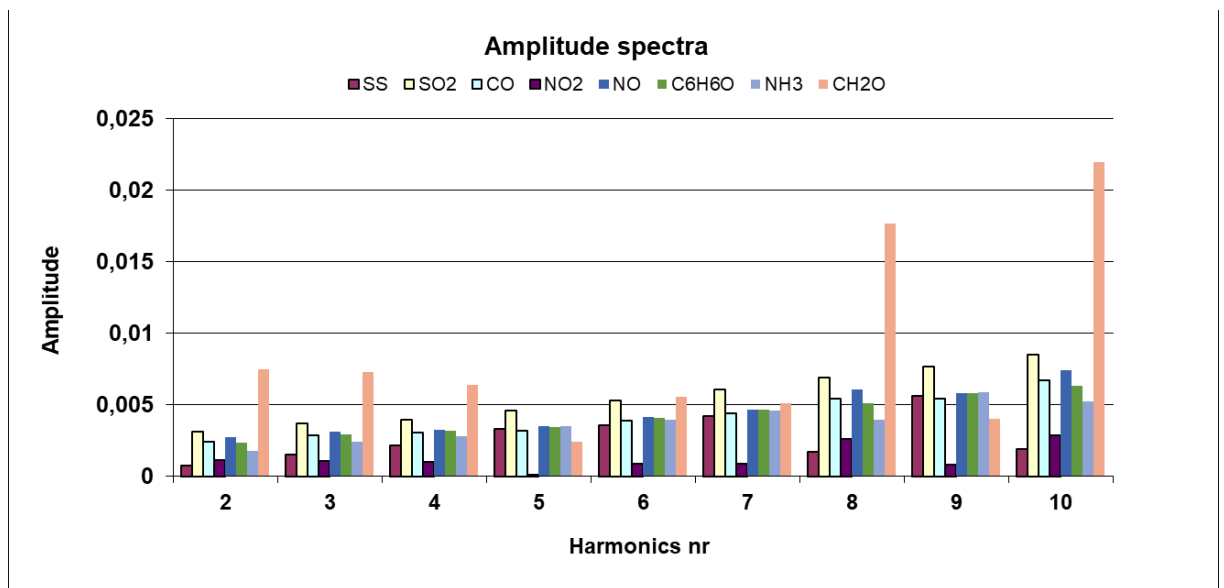


Рисунок 8. Спектр амплітуд для даних міста Дніпро у 2023-2025 роки

CH<sub>2</sub>O (помаранчевий) виділяється - у нього найвища амплітуда на кількох гармоніках (особливо 2, 3, 8 і 10), що вказує на сильну присутність відповідних частот у сигналі

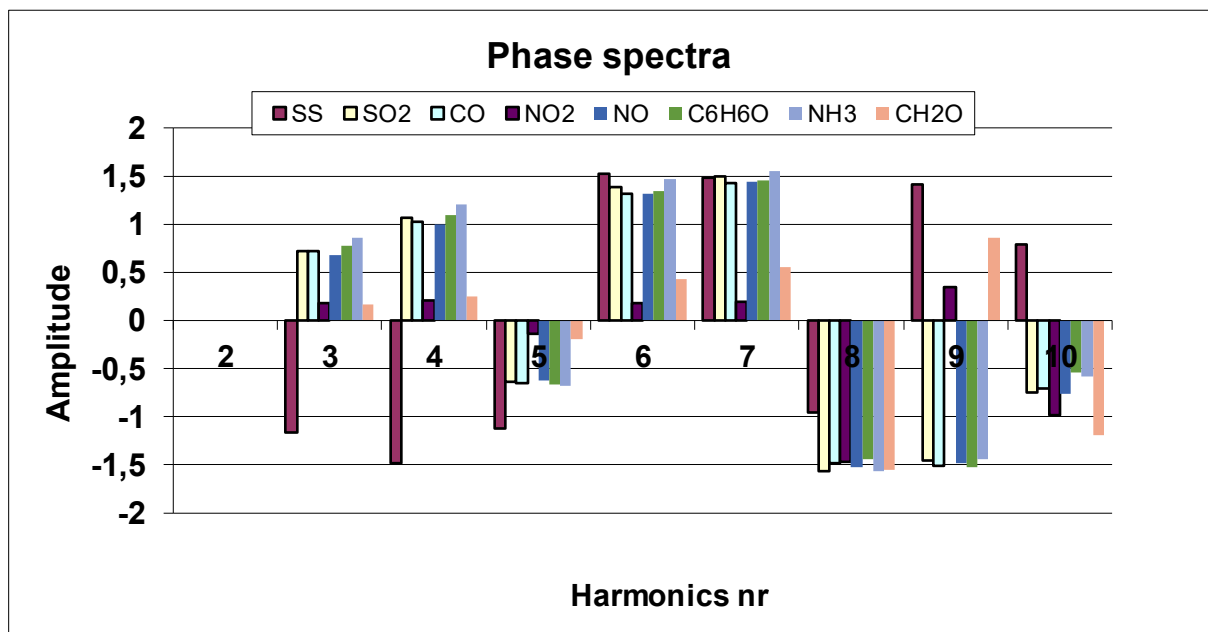


Рисунок 9. Фазовий спектр для даних міста Дніпро у 2023-2025 роки

SS (темно-бордовий) демонструє значні фазові зрушення (особливо негативні на 3-й та 4-й гармоніках і позитивний на 9-й).

## 2.11 Аутлаєри

Аутлаєри- це точки даних, які суттєво відрізняються від інших спостережень у наборі даних. Вони знаходяться далеко від центрального кластера даних і можуть вказувати на мінливість, помилки або нові явища.

Викиди можуть становити значну проблему для аналізу даних, особливо при аналізі шкідливих хімічних речовин у повітрі. У таких завданнях, де у лабораторіях вимірюють хімічні речовини вручну, можуть траплятися помилки. Видалення таких точок даних може значно покращити загальні результати аналізу та продуктивність алгоритмів машинного навчання [16] (стор. 16).

Водночас, у воєнний час ці викиди можуть нести важливу інформацію про лісові пожежі, катастрофи чи обстріли заводських зон. Щоб оцінити ці цінні аспекти, вкрай важливо провести глибокий аналіз цього питання.

Існує кілька методів[15] виявлення таких точок даних :

- IQR: Виявляє викиди як точки за межами діапазону  $Q1 - 1.5 * IQR$ ,  $Q3 + 1.5 * IQR$ , де  $IQR = Q3 - Q1$  (міжквартильний розмах). Це метод, заснований на правилах, що спирається на розкид даних.
- Z-бали: Позначає точки як викиди, якщо їхній стандартизований бал (відстань від середнього в стандартних відхиленнях) перевищує певний поріг (зазвичай  $|z| > 3$ ). Припускає нормальний розподіл.
- Ізоляційний ліс (Isolation Forest): Ансамблевий алгоритм, який ізолює аномалії шляхом випадкового розбиття даних.  
Для ізоляції аномалій потрібно менше розділів, що призводить до скорочення довжини шляху в деревах.

Після виявлення таких точок існує кілька способів[15] їх обробки. Серед інших, були протестовані такі:

- Видалити: Повністю виключити точки, що відхиляються від норми, з набору даних.
- Обмеження (також відоме як «відсікання»): Замінити екстремальні значення фіксованим процентильним значенням (наприклад, обмежити все, що перевищує 95-й процентиль, до 95-го процентилля).
- Інтерполювати: Замінити викиди на 0 і заповнити за допомогою інтерполяції з навколишніх значень, що не є викидами (наприклад, лінійна інтерполяція).

Було отримано наступні результати:

Виявлення викидів за допомогою методу Z-score:

- завислі речовини: 6 відхилень
- SO<sub>2</sub>: 33 відхилень
- CO: 10 відхилень
- NO<sub>2</sub>: 10 відхилень
- NO: 8 відхилень

- фенол: 6 відхилень
- $\text{NH}_3$ : 11 відхилень
- формальдегід: 5 відхилень

та їх подальше обмеження.

Отримані результати наведені на Рис. 9 та 10:

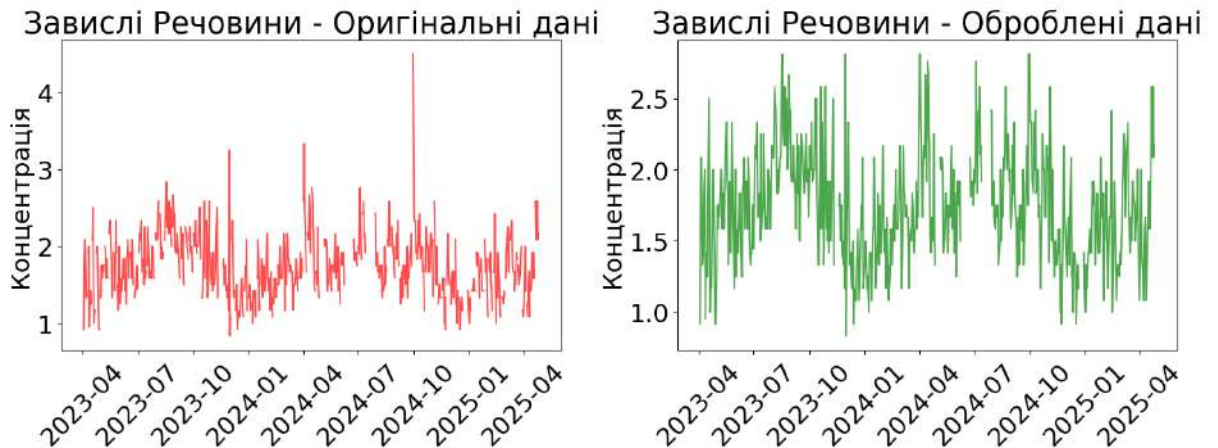


Рисунок 9. Обмеження викидів(часовий ряд), виявлених за допомогою методу Z-score для даних завислих речовин міста Дніпро у 2023-2025 роки

а) оригінальний графік концентрації завислих речовин в повітрі

б) оброблений графік концентрації завислих речовин в повітрі праворуч (були обмежені викиди)

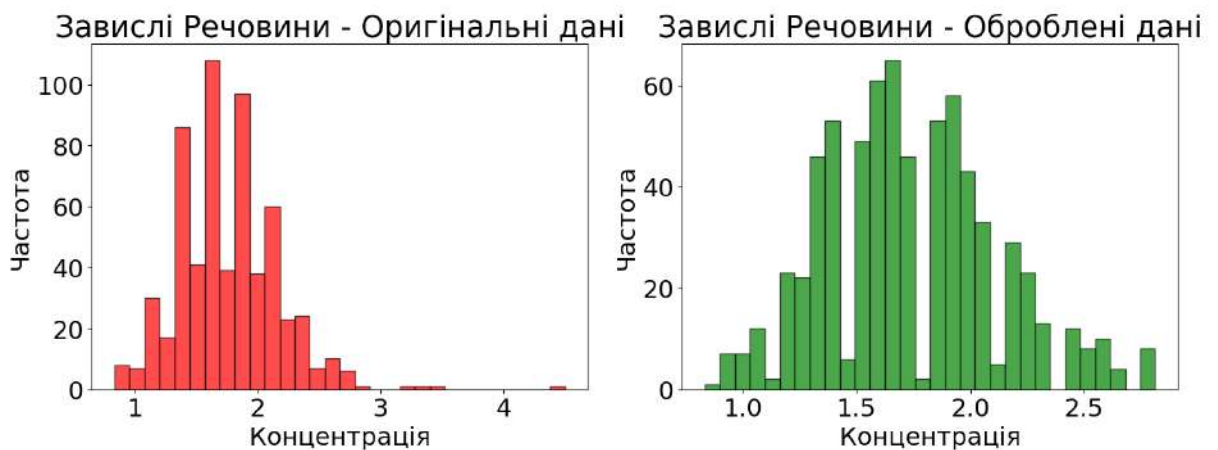


Рисунок 10. Обмеження викидів(гістограма), виявлених за допомогою методу Z-score для даних завислих речовин міста Дніпро у 2023-2025 роки

а) гістограма розподілу концентрацій завислих речовин для оригінальних даних

*б) гістограма розподілу концентрацій завислих речовин для оброблених даних*

## **2.12 Кластерний аналіз**

Кластер - це набір точок даних, згрупованих разом на основі певної схожості. Кластери можуть відрізнятися за розміром, формою і щільністю, а визначення подібності залежить від контексту і конкретного алгоритму, що використовується.

Кластерний аналіз[17](стор. 501-522) - це метод аналізу даних, який визначає та групує схожі точки даних у кластери. Його основна мета - виявити внутрішні структури в даних, допомагаючи зрозуміти закономірності, взаємозв'язки та розподіли в наборі даних. Існують різні методи пошуку оптимальних кластерів, враховуючи їхню кількість, а також численні алгоритми для оцінки якості кластеризації.

Кластерний аналіз особливо важливий у нашому випадку. Якщо дані вдається успішно розділити на групи, це свідчить про наявність глибоких залежностей між забруднювачами, що може сигналізувати про потенційні викиди на заводах.

Були використані наступні методи кластеризації з фреймворку Python SciPy:

1. K-Means розбиває дані на задану кількість кластерів, мінімізуючи відстань між точками та центрами кластерів шляхом ітеративного оновлення цих центрів.

2. Ієрархічна кластеризація - будує ієрархію кластерів, об'єднуючи або розділяючи їх на основі відстаней, не потребуючи заздалегідь визначеної кількості кластерів.

Кількість кластерів налаштовано за допомогою методу ліктя і алгоритмом пошуку центроїдів Kmeans.

Метод ліктя допомагає вибрати оптимальну кількість кластерів, аналізуючи зміну внутрішньої дисперсії — точка "ліктя" на графіку вказує, де подальше збільшення кількості кластерів дає незначне покращення.

Конкретно в нашому випадку, "лікоть" на графіку відсутній через структуру даних — це свідчить про те, що дані не мають чіткої кластерної структури.

Для кластеризації було використано дані усіх забруднювачів:

На рисунку 11, 12 та 13 показано результати, отримані в ході аналізу.

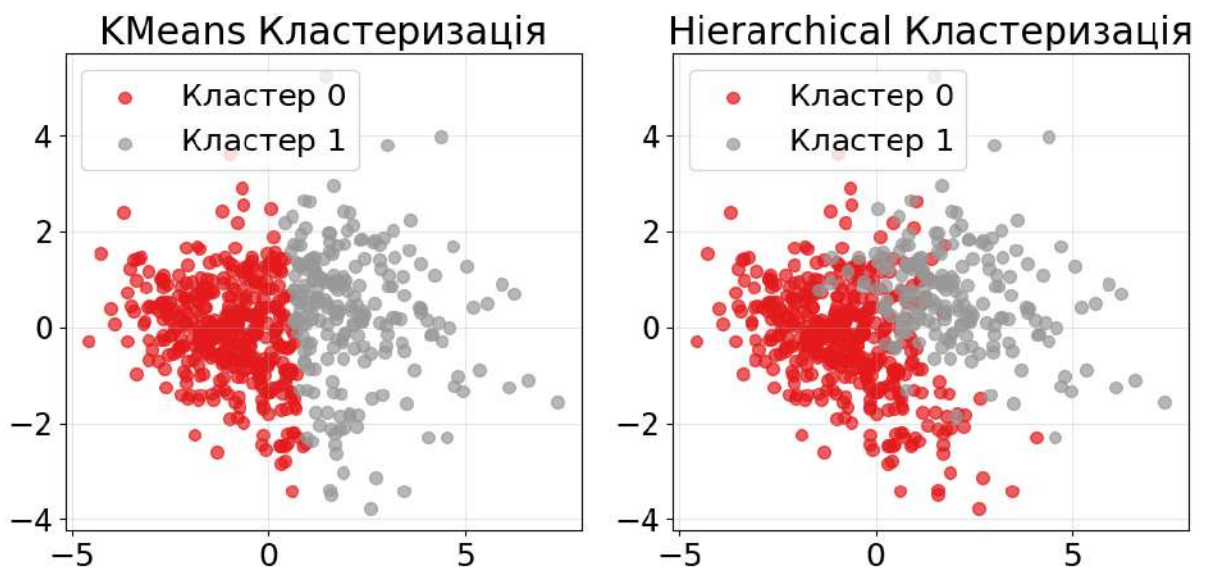


Рисунок 11. Кластеризація для  $k=2$

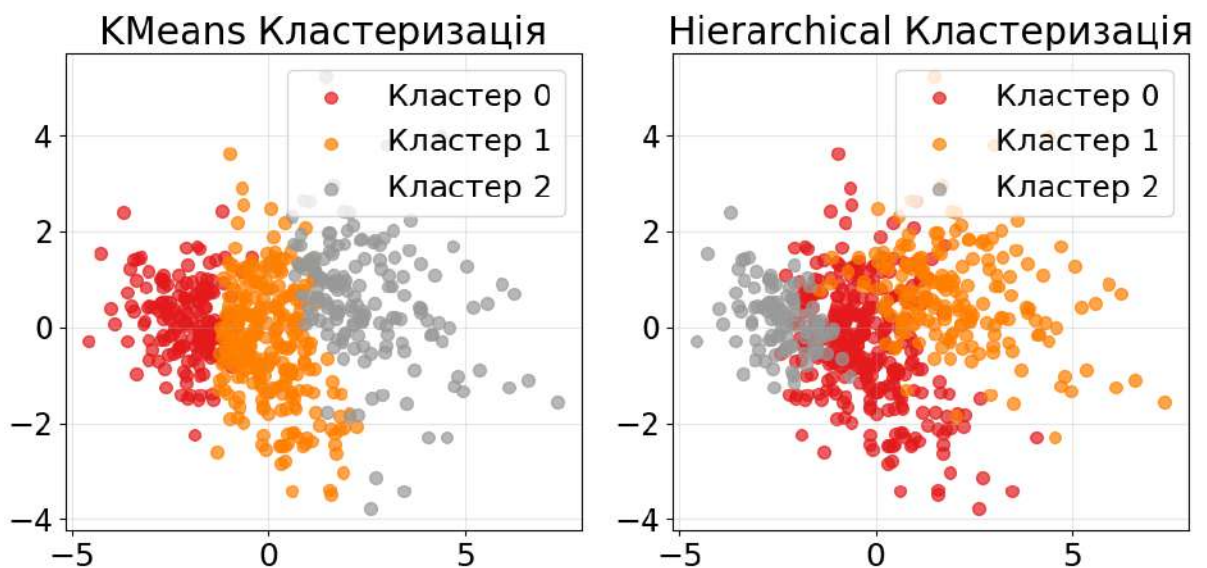


Рисунок 12. Кластеризація для  $k=3$

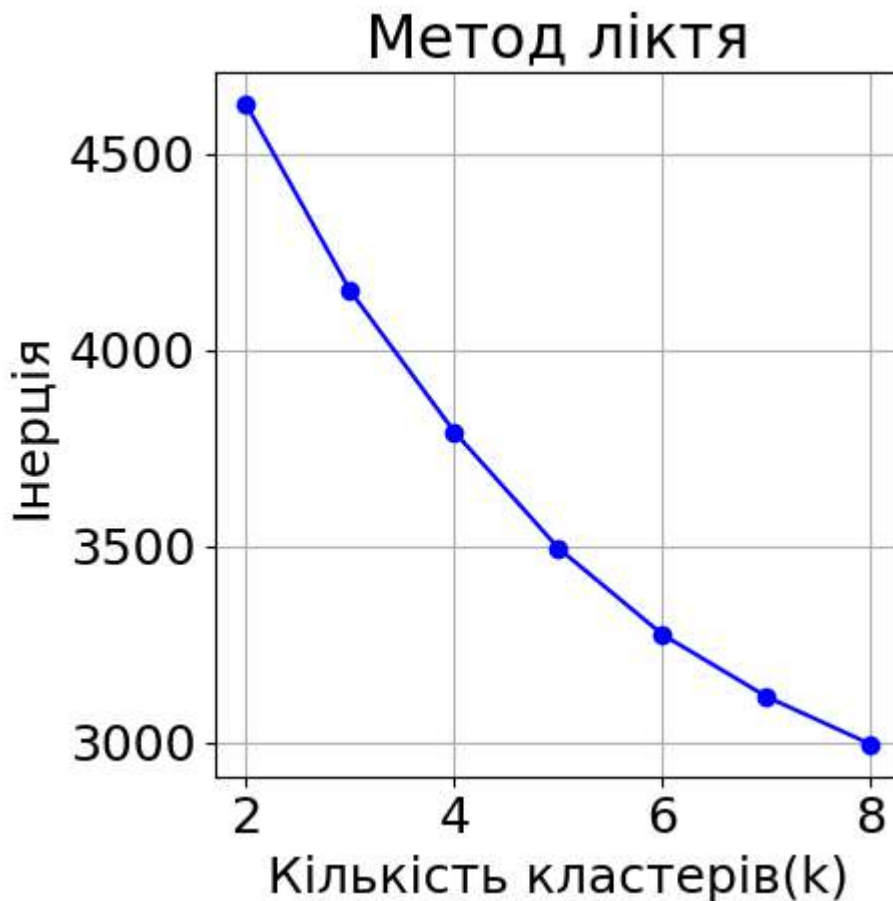


Рисунок 13. Розрахунок кількості кластерів за допомогою методу ліктя

Як видно з рисунка 11, при  $k = 2$  більшість точок зосереджена навколо одного центру, що свідчить про домінування однієї групи. На рисунку 12 при  $k = 3$  межі між кластерами залишаються розмитими, що вказує на неякісний розподіл. Рисунок 13 показує відсутність вираженого «ліктя» на графіку, що свідчить про нечітку кластерну структуру даних.

## 2.13 Аналіз дисперсій

Оскільки опрацьовуються данні з часових рядів, було б корисно знати, чи демонструють ці дані тенденції в певних групах - наприклад, чи залежать рівні забруднювачів від дня тижня, пори року або місяця.

Для таких завдань існує добре відпрацьований математичний апарат, заснований на дисперсійному аналізі (ANOVA). Вимірюючи дисперсії між різними групами, загальну дисперсію та дисперсію в межах усього набору

даних, можна ефективно перевірити, чи суттєво впливають часові фактори на окремі забруднювачі.

Однак цей тип аналізу дуже чутливий[16](стор. 8) до викидів і, особливо, до пропущених даних, яких у нашому наборі даних дуже багато.

Було досліджено три часові групи: дні тижня, місяці та сезони року.

Рисунок 14 нижче демонструє отримані результати

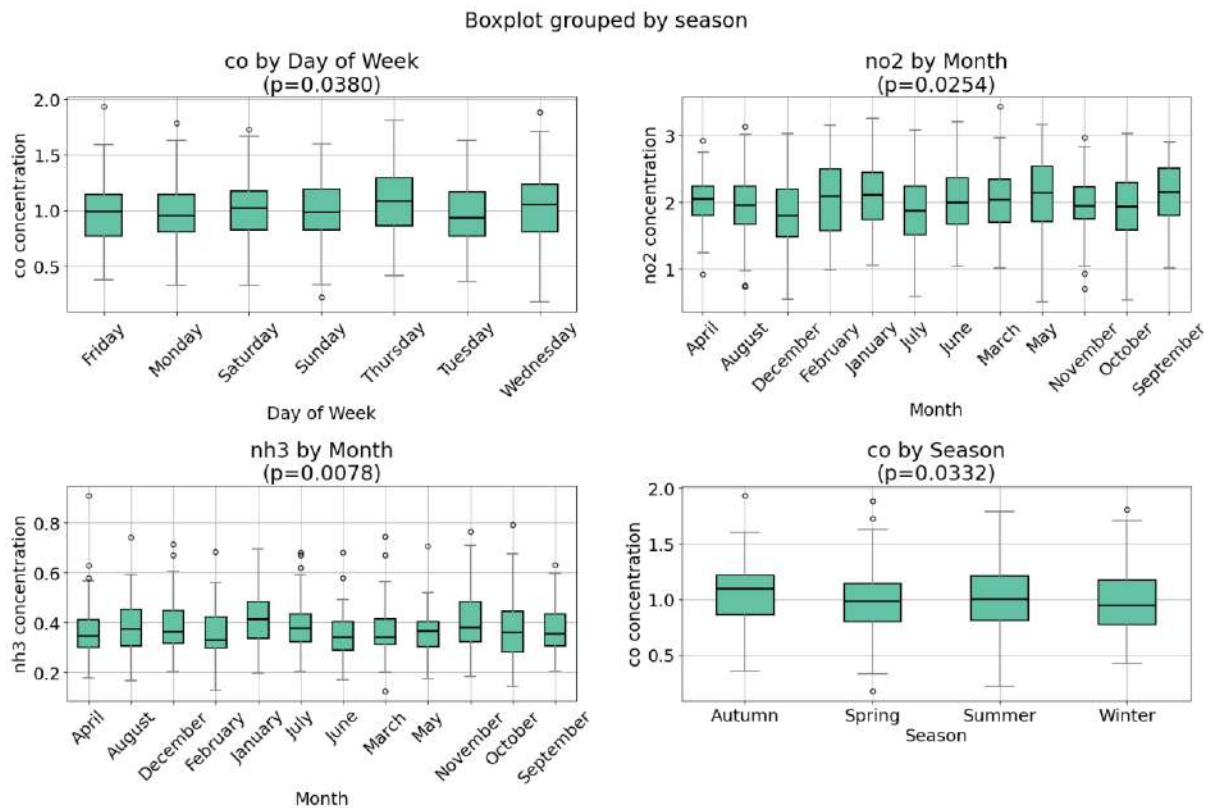


Рисунок 14 а) Залежність CO від днів тижня

б) Залежність NO<sub>2</sub> від місяців

в) Залежність NH<sub>3</sub> від місяців

г) Залежність CO від сезонів

### 3. Висновки

У цьому дослідженні було проаналізовані забруднювачі повітря за допомогою описової статистики, включаючи максимальне, середнє, медіанне, мінімальне значення, асиметрію та ексцес. Для більшості речовин, зокрема для зважених твердих частинок, діоксиду сірки ( $\text{SO}_2$ ), оксиду вуглецю ( $\text{CO}$ ), діоксиду азоту ( $\text{NO}_2$ ), аміаку ( $\text{NH}_3$ ) та формальдегіду ( $\text{HCHO}$ ), максимальні значення значно перевищують відповідні середні значення. Ця закономірність свідчить про те, що викиди не є рівномірними в часі, а включають періодичні піки високої інтенсивності, ймовірно пов'язані з епізодичними подіями, такими як промислові вибухи, великомасштабні пожежі або інтенсивні бойові дії.

Перед застосуванням більш досконалих статистичних методів розподіл кожного забруднювача було оцінено на нормальність за допомогою тесту на відповідність  $\chi^2$ . З перевірених забруднюючих речовин лише оксид азоту та фенол не відхилили нульову гіпотезу нормальності при  $\alpha = 0,05$ . Натомість, зважені частинки,  $\text{SO}_2$ ,  $\text{CO}$ ,  $\text{NO}_2$ , аміак та формальдегід значно відхиляються від гаусового розподілу. Проте для  $\text{CO}$ ,  $\text{NO}_2$  та  $\text{NH}_3$  близькість середніх значень до медіани та лише помірна асиметрія свідчать про те, що методи нормального наближення все ще можуть давати досить точні висновки щодо цих конкретних забруднювачів.

Оцінки щільності ймовірності, змодельовані за припущенням нормального розподілу, підкреслюють, що розподіли декількох забруднювачів демонструють асиметрію та важкі хвости. Графічне накладання теоретичних гаусових кривих на гістограми виявляє лівосторонню асиметрію для  $\text{NO}_2$  і  $\text{CO}$ , тоді як  $\text{NH}_3$  і зважені частинки демонструють правосторонню асиметрію з подовженими хвостами в бік високих значень концентрації. Ці висновки підкреслюють, що припущення

про сувору нормальність ризикує недооцінити ймовірність екстремальних епізодів забруднення в регіоні.

Згладжування часових рядів за допомогою ковзних середніх ефективно відфільтрувало короткостроковий розкид і проявило довгострокові тенденції. Для  $\text{NO}_2$  і  $\text{CO}$  згладжені криві демонструють поступове сезонне зниження, що, ймовірно, відображає зменшення промислового виробництва та обмеження транспорту, тоді як  $\text{NH}_3$  і формальдегід демонструють висхідну тенденцію протягом того ж періоду, що може корелювати зі збільшенням використання азотних боєприпасів або посиленням сільськогосподарської діяльності. Деякі забруднювачі (наприклад, зважені частинки та  $\text{HCHO}$ ) демонструють різкі сплески, розкидані між нижчими базовими рівнями, що відповідає ізольованим бойовим епізодам або епізодичним пожежам.

Аналіз залишкових коливань, обчислених як відхилення кожного спостереження від загального середнього значення, показує виражену мінливість навколо тенденції для зважених частинок та  $\text{NO}_2$ , що вказує на те, що ці забруднювачі зазнають більш нестабільних сплесків. На відміну від цього, коливання  $\text{CO}$  та  $\text{SO}_2$  залишаються порівняно помірними, що вказує на більш стабільні джерела викидів. Візуалізація цих відхилень підкреслює, що екстремальні події, що виходять за межі норми, сильно впливають на часову поведінку певних забруднювачів.

Спектральний аналіз з використанням методу швидкого перетворення Фур'є дозволяє ідентифікувати частотні складові зміни концентрацій забруднюючих речовин. Амплітудні спектри показали, що формальдегід ( $\text{CH}_2\text{O}$ ) має найбільш значущу гармоніку, зі значними максимумами на другій, четвертій, восьмій та десятій гармоніках, що може свідчити про циклічні викиди, які можуть бути пов'язані з регульованими технологічними

процесами. Для більшості інших речовин (включаючи NO, SO<sub>2</sub> та CO<sub>2</sub>) амплітудні розподіли між другою та десятою гармоніками є відносно рівними з невеликими відхиленнями. Це свідчить про те, що це стабільне джерело забруднення з регулярним циклом, який неможливо виявити.

Фазові спектри показують відмінності в часовому розподілі гармонік: у випадку SS спостерігаються значні фазові зсуви, особливо на 4 і 6 гармоніках. Синхронізація фаз NO та C<sub>6</sub>H<sub>6</sub>O відбувається на середніх гармоніках, що відповідає їх спільним джерелам (наприклад, вибухи або вихлопні гази автотранспорту).

Загалом, спектральний портрет показує наявність періодичних (щотижневих або сезонних) та епізодичних (високочастотних) викидів, особливо для NH<sub>3</sub> та CH<sub>2</sub>O, що може бути пов'язано з переривчастою діяльністю в промисловому секторі або військовими подіями.

Виявлення відхилень за допомогою Z-балів ( $|z| > 3$ ) дозволило виявити численні екстремальні спостереження: 6 для зважених частинок, 33 для SO<sub>2</sub>, 10 для CO, 10 для NO<sub>2</sub>, 8 для NO, 6 для фенолу, 11 для NH<sub>3</sub> і 5 для формальдегіду. Хоча деякі з них можуть бути наслідком помилок вимірювання або аномалій введення даних, багато з них відповідають реальним подіям, таким як великомасштабні пожежі, вибухи або концентровані військові удари. Впровадження надійних заходів (наприклад, обмеження на рівні 95-го процентиля) є надзвичайно важливим для запобігання спотворенню подальших аналізів, проте збереження справжніх винятків є не менш важливим для фіксації критичних сигналів щодо стану навколишнього середовища та здоров'я населення в умовах війни.

Кластерний аналіз усіх забруднюючих речовин за допомогою K-Means та ієрархічного кластеризації послідовно вказує на слабку внутрішню структуру групування, причому оптимальний розподіл складається лише з двох кластерів. Це свідчить про те, що рівні забруднення в цілому

утворюють безперервний градієнт, а не дискретні режими, і що екстремальні події (наприклад, промислові аварії або викиди, пов'язані з бойовими діями) не утворюють повністю відокремлених кластерів, а знаходяться на крайньому кінці цього континууму.

Було проведено аналіз дисперсії (ANOVA) для оцінки того, чи відрізняються концентрації забруднюючих речовин залежно від тимчасових факторів — дня тижня, місяця та сезону. Хоча дані чутливі до відсутніх значень та винятків, попередні результати вказують на статистично значущі відмінності ( $p < 0,05$ ) для  $\text{NO}_2$  та зважених частинок як у місячних, так і в сезонних групах, що відображає підвищені концентрації протягом холодніших місяців. Відмінності між буднями та вихідними були менш вираженими, але все ж помітними для  $\text{CO}$  та  $\text{SO}_2$ , що відповідає зменшенню транспортного руху у вихідні дні.

Загалом, комплексний статистичний робочий процес — від описової аналітики та тестування нормальності до просунутих методів часових рядів, спектральних та багатовимірних методів — підтверджує, що забруднення повітря в досліджуваних промислових містах України характеризується періодичними піками високої інтенсивності, негаусовим розподілом та чіткими часовими закономірностями, зумовленими як антропогенною діяльністю, так і подіями, пов'язаними з війною. Ці висновки забезпечують міцну основу для розробки цілеспрямованих заходів щодо поліпшення якості повітря, інформування політиків про критичні періоди підвищеного ризику та вдосконалення моделей прогнозування в умовах швидко мінливої ситуації під час війни.

#### 4. Список використаних джерел

1. Ричак Н.Л., Кізілова Н.М., Майструк В.А., Макаренко О.С., Прогнімак О.С. Математичний аналіз забруднення повітря на території України з використанням даних з відкритих джерел. Вісник Вінницького політехнічного інституту. 2021. №4(157). С. 20-31. doi: 10.31649/1997-9266-2021-157-4-20-32
2. Міністерство захисту довкілля та природних ресурсів України: Стан атмосферного повітря. Режим доступу:  
<https://mepr.gov.ua/diyalnist/napryamky/ekologichnyj-monitoring/stan-atmosfernogo-povitrya/>
3. Стан атмосферного повітря у Запоріжжі. Режим доступу:  
<https://mepr.gov.ua/diyalnist/napryamky/ekologichnyj-monitoring/stan-atmosfernogo-povitrya/zaporizhzhya/>
4. Стан атмосферного повітря у Харкові. Режим доступу:  
<https://mepr.gov.ua/diyalnist/napryamky/ekologichnyj-monitoring/stan-atmosfernogo-povitrya/harkiv/>
5. Стан атмосферного повітря у Дніпрі. Режим доступу:  
<https://mepr.gov.ua/diyalnist/napryamky/ekologichnyj-monitoring/stan-atmosfernogo-povitrya/dnipro/>
6. Горбачук, В. М. Теорія ймовірностей та математична статистика [Електронний ресурс] : підручник для здобувачів ступеня бакалавра за технічними та економічними спеціальностями / В. М. Горбачук, О. І. Кушлик-Дивульська ; КПІ ім. Ігоря Сікорського. – Київ : КПІ ім. Ігоря Сікорського, 2023. – 351 с. – Назва з екрана.
7. Твердохвалова Ю.Л. Математичні методи в психології. Тема №6. Методи кореляційного аналізу; ХНУВС. Режим доступу:  
<https://nm2.univd.edu.ua/discipline/145793>
8. Огірко О. І. Теорія ймовірностей та математична статистика: навчальний посібник / О.І. Огірко, Н.В. Галайко. – Львів: ЛьвДУВС, 2017. – 292 с.

9. Найко Д.А. Шевчук О. Ф. Теорія ймовірностей та математична статистика: навч. посіб. / Д.А. Найко, О.Ф. Шевчук – Вінниця: ВНАУ, 2020. – 382 с.
10. Коцюбинський В. Ю., Кислиця Л. М. Основи моделювання ринкових ситуацій : навч. посіб. / В. Ю. Коцюбинський, Л. М. Кислиця. - Вінниц. нац. техн. ун-т. - Вінниця : ВНТУ. [Електронний ресурс]. Режим доступу: [https://web.posibnyky.vntu.edu.ua/fksa/12kocubynsky,kyslycia\\_on\\_model\\_rynk\\_sytuac/](https://web.posibnyky.vntu.edu.ua/fksa/12kocubynsky,kyslycia_on_model_rynk_sytuac/)
11. Проценко Н. М. Економічна інформатика: навч. посіб. / Н. М. Проценко. - Харк. нац. агр. ун-т ім. В.В. Докучаєва. - Харків, 2020. - 212 с. [Електронний ресурс].
12. А. Т. Яровий, Є. М. Страхов. Аналіз часових рядів. — Одеса : Освіта України, 2019. — 260 с.
13. Юрченко, М. Є. Прогнозування та аналіз часових рядів : метод. вказівки до практ. занять та самост. роботи студентів спец. 051 "Економіка" освітня програма "Економічна кібернетика", "Економічна аналітика" / уклад. М. Є. Юрченко. – Чернігів : ЧНТУ, 2018. – 88 с.
14. Цифрова обробка сигналів. УДУНТ [Електронний ресурс] Режим доступу: [https://nmetau.edu.ua/file/07\\_7.4\\_lbr\\_gr\\_rbr\\_.pdf](https://nmetau.edu.ua/file/07_7.4_lbr_gr_rbr_.pdf)
15. UNDERSTANDING THE INFLUENCE OF OUTLIERS ON MACHINE LEARNING MODEL INTERPRETABILITY (FATIMA ODODO & NICHOLAS ADDOTEY , Trans.). (2025). International Journal of African Sustainable Development Research, 7(2). <https://doi.org/10.70382/tijasdr.v07i2.019>
16. Wilcox R. R. Introduction to Robust Estimation and Hypothesis Testing. – Academic Press, 2021.
17. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. – 2nd ed. – Springer, 2009.

18. Всесвітня організація охорони здоров'я (ВООЗ). [Електронний ресурс].  
Режим доступу: <https://geneva.mfa.gov.ua/posolstvo/2612-who> .
19. Директива 2008/50/ЄС Європейського Парламенту та Ради від 21 травня 2008 р. про якість атмосферного повітря та чистіше повітря для Європи.  
[Електронний ресурс]. Режим доступу: [https://zakon.rada.gov.ua/laws/show/994\\_950#](https://zakon.rada.gov.ua/laws/show/994_950#) .
20. European Environmental Agency, Clean Air For Europe program (CAFE).  
[Electronic resource]. Available: <https://www.eea.europa.eu/themes/air/links/research-projects/clean-air-for-europe-programme-cafe> .
21. World air quality. Air quality and pollution city ranking. [Electronic resource].  
Available: <https://www.iqair.com> .  
Turkey, Ukraine and Western Balkan countries compete for top spot in coal power air pollution in Europe. EMBER report. [Electronic resource].  
Available: <https://ember-climate.org/commentary/2021/05/25/coal-power-air-pollution/> .
22. European Environmental Agency, Air pollution in Europe. [Electronic resource]. Available: <https://www.eea.europa.eu/> .
23. Світлана Пухач. Джерела забруднення повітря у Харкові та перспективи вдосконалення екологічного стану. Режим доступу: [Джерела забруднення повітря у Харкові та перспективи вдосконалення екологічного стану - kharkiv.name](#)
24. Health Effects Institute. 2020. State of Global Air 2020. Data source: Global Burden of Disease Study. IHME. - 2020. [Electronic resource].  
Available: <https://www.stateofglobalair.org/data/#/air/plot>
25. Державні санітарні правила охорони атмосферного повітря населених місць (від забруднення хімічними та біологічними речовинами)  
[Електронний ресурс]. Режим доступу: <https://zakon.rada.gov.ua/rada/show/v0201282-97#Text> .

## 26. Моніторинг повітря у Харкові під час війни

Режим доступу: [Харків: аміак, хлор, радіація – під контролем - Все про повітря](#)

## Додаток

```
#!/bin/bash

# Увімкнути режим, коли скрипт припиняється при будь-
якій помилці

set -e

# Зчитування URL з першого аргументу командного рядка
URL="$1"

[ -z "$URL" ] && echo "Використання: $0 <URL>" && exit
1

# Перевірка, чи існує директорія ./data
if [ -d "./data" ]; then
    echo "Директорія ./data вже існує, будь ласка,
    видаліть її"
    exit 1
fi

# Створення потрібних підкаталогів
mkdir -p data/raw data/csv data/parsed

# Завантаження HTML-сторінки за вказаною URL-адресою
echo "[INFO] Завантаження HTML сторінки з $URL..."
wget -q -O data/page.html "$URL"

# Витягування посилань на .xls файли з HTML
echo "[INFO] Витягування посилань на .xls файли..."
awk '
```

```

BEGIN { IGNORECASE = 1 }

/<a [^>]*href *= *"[^"]*"*.xls"/ {
    match($0, /href *= *"([^"])*\.xls"/, m)
    if (m[1] != "") print m[1]
}

' data/page.html > data/links.txt

# Підрахунок кількості посилань
link_count=$(wc -l < data/links.txt)
echo "[INFO] Знайдено $link_count посилань"
echo "[INFO] Перші 5 посилань:"
head -n 5 data/links.txt

# Завантаження .xls файлів
echo "[INFO] Завантаження .xls файлів..."
count=0
while read -r link; do
    prefix=$(printf "%04d" $count)
    fname="${prefix}-${basename "$link"}"
    wget -q "$link" -O "data/raw/$fname"
    echo "[ЗАВАНТАЖЕНО] $link -> data/raw/$fname"
    count=$((count + 1))
done < data/links.txt

# Конвертація .xls у .csv за допомогою ssconvert
echo "[INFO] Конвертація .xls у .csv..."

```

```

for file in $(ls data/raw | sort); do
    out="data/csv/${file%.xls}.csv"
    ssconvert "data/raw/$file" "$out" >/dev/null 2>&1
    echo "[КОНВЕРТОВАНО] $file -> ${file%.xls}.csv"
done

# Створення порожнього об'єднаного CSV-файлу
touch "data/merged.csv"

# Обробка CSV: видалення перших 7 рядків, нормалізація
# дат, додавання ID
echo "[INFO] Видалення перших 7 рядків, нормалізація
# дат, додавання ID..."

for file in $(ls data/csv | sort --reverse); do
    prefix=$(echo "$file" | cut -d'-' -f1)

    # Витягування дати з назви файлу та перетворення у
    # формат YYYY-MM-DD
    raw_date=$(echo "$file" | grep -oE '[0-9]{2}\.[0-9]{2}\.[0-9]{4}' | head -n1)
    day=$(echo "$raw_date" | cut -d. -f1)
    month=$(echo "$raw_date" | cut -d. -f2)
    year=$(echo "$raw_date" | cut -d. -f3)
    base_iso="$year-$month-$day"

    # Обробка кожного CSV-файлу
    awk -F',' -v id="$prefix" -v fname="$file" -v
base="$base_iso" '

```

```

function to_epoch(iso) {
    split(iso, d, "-")
    return mktime(d[1] " " d[2] " " d[3] " 00
00 00")
}
BEGIN {
    base_epoch = to_epoch(base)
}
NR > 7 {
    raw = $1
    gsub(/"/, "", raw)

    if (raw ~ /^[0-9]+$/) {
        new_epoch = base_epoch + (raw - 1) *
86400
        new_date = strftime("%Y-%m-%d",
new_epoch)
    } else if (raw ~ /^[0-9]{1,2}\/[0-
9]{1,2}\/[0-9]{4}$/) {
        split(raw, p, "/")
        new_date = sprintf("%04d-%02d-%02d",
p[3], p[2], p[1])
    } else {
        new_date = raw
    }

    $1 = new_date
    OFS = FS
}

```

```
        print $0, id, fname
    }
    ' "data/csv/$file" > "data/parsed/$file"
    cat "data/parsed/$file" >> "data/merged.csv"

    echo "[ОБРОБЛЕНО] $file -> data/parsed/$file"
done

echo "[ГОТОВО] Усі файли збережено у ./data"
echo "[ГОТОВО] Результат збережено у ./data/merged.csv"
```