УДК 004.62

В.В. Хайдуров, Б.Я. Яйлимов, А.Ю. Шелестов

МОДЕЛЬ ОЦІНКИ ЯКОСТІ ПОВІТРЯ ЗА СУПУТНИКОВИМИ ДАНИМИ НА ОСНОВІ МЕТОДУ ГРУПОВОГО УРАХУВАННЯ АРГУМЕНТІВ*

Хайдуров Владислав Володимирович

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»,
orcid: 0000-0002-4805-8880

4labs0@gmail.com

Яйлимов Богдан Ялкапович

Інститут космічних досліджень НАН України та ДКА України, м. Київ,
orcid: 0000-0002-2635-9842

yailymov@gmail.com

Шелестов Андрій Юрійович

Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»,
orcid: 0000-0001-9256-4097

andrii.shelestov@gmail.com

У роботі представлено математичну модель на основі методу групового урахування аргументів (МГУА) для оцінки даних про якість повітря на рівні землі за допомогою супутникових спостережень. Забруднення повітря є серйозною екологічною проблемою, яка має значний вплив на екосистеми, здоров'я людини та зміну клімату. Наземні мережі моніторингу якості повітря забезпечують прямі вимірювання рівня забруднення, але у багатьох регіонах світу обмежені кількістю станцій. Супутникове дистанційне зондування пропонує нові можливості для послідовного та детального моніторингу якості повітря як доповнення до наземних спостережень. Однак існують певні обмеження, включно з низьким просторовим розрізненням супутникових даних, невизначеностями вимірювань і низькою частотою зйомки. У цьому дослідженні розроблено модифіковану модель МГУА для співставлення даних супутникових спостережень з наземними даними про якість повітря для дрібних твердих частинок (PM2,5) і твердих частинок розміром менше 10 мк (PM10) у місті Києві, Україна. Модель оптимально реконструює нелінійні функціональні залежності між часовими рядами супутникових і наземних змінних, одночасно оптимізуючи загальну складність моделі. Проведено кілька обчислювальних експериментів на реальних наборах даних. Результати показали сильну кореляцію між прогнозованими та емпірично спостережуваними значеннями на незалежному

* Роботу виконано в рамках проєкту 2020.02/0284 «Геопросторові моделі та інформаційні технології супутникового моніторингу проблем розумного міста» за грантової підтримки Національного фонду досліджень України в межах конкурсу «Підтримка досліджень провідних та молодих учених».

25 %-му тестовому зразку (досягнуто 0,8889 для PM2,5). Для оптимізованої моделі МГУА вимагалось у 2–3 рази менше параметрів, ніж для порівнюваної архітектури нейронної мережі, щоб досягти того самого рівня точності. Це демонструє здатність запропонованого підходу точно оцінювати концентрації забруднення на рівні землі з високою роздільною здатністю на основі супутникових даних, використовуючи МГУА-моделювання. Розроблена модель надає більш повну просторово-часову картину розподілу забруднення для значного покращення можливостей моніторингу навколишнього середовища, інформування громадськості та підтримки науково обґрунтованих політичних рішень щодо стратегій пом'якшення впливу забруднення на довкілля. У дослідженні підкреслюється, що злиття супутникових і наземних даних за допомогою моделювання МГУА дозволяє значно удосконалити можливості оцінки якості повітря, щоб краще зрозуміти дрібномасштабну динаміку забруднення, захистити населення та розробити ефективні рішення для захисту навколишнього середовища.

Ключові слова: математичні моделі, регресійна модель, поліном Колмогорова–Габора, метод групового урахування даних, якість повітря, супутникові дані, кореляційний аналіз.

Вступ

Забруднення повітря є однією з найважливіших екологічних проблем сучасності, які мають серйозний вплив на здоров'я людей та стан природних екосистем. Останнім часом все більше уваги приділяється вивченню та оцінці якості повітря, адже від цього залежить не лише здоров'я населення, а й стан навколишнього середовища [1].

Для моніторингу рівня забруднення повітря та визначення якості атмосферного середовища традиційно використовуються дані наземних станцій. В Україні функціонує кілька систем контролю стану атмосфери, проте їхня щільність значно менша, ніж у європейських державах [2, 3]. Наприклад, державна мережа спостережень налічує лише близько 100 пунктів на всю країну, чого недостатньо для адекватної оцінки рівня забруднення повітря. Відстань між станціями може становити сотні кілометрів, тому вони не забезпечують детального вимірювання показників забруднення [4–6]. Для порівняння у Німеччині нараховується понад 1000 станцій моніторингу якості повітря [7]. Крім цього, в Україні лише частина даних оприлюднюється онлайн. Зокрема, державна система екологічного моніторингу під управлінням Держекоінспекції включає близько 100 станцій по всій країні [8, 9]. Однак на сайті Центральної геофізичної обсерваторії (ЦГО) представлено лише частину даних. Натомість у країнах ЄС більшість даних публікується у зручних для аналізу форматах. Наземні станції спостережень можуть виходити з ладу, тому дані наземних спостережень можуть мати пропуски або містити некоректні дані.

Для розв'язання зазначених проблем можна використовувати дані супутникового моніторингу, які надають нові можливості для отримання об'єктивних та деталізованих даних щодо забруднення повітря. В останні роки супутникові дані стають все більш популярним джерелом інформації про стан атмосфери та якість повітря, проте використання таких даних має певні обмеження. По-перше, просторове розрізнення супутникових знімків обмежене технічними можливостями сенсорів. Зображення з супутника охоплює велику площу, тому важко отримати точні дані щодо невеликих ділянок чи окремих об'єктів. Наприклад, розрізнення продуктів CAMS складає 10–40 км [10, 11], а Sentinel-5P — 7,5 км [12]. Це ускладнює моніторинг забруднень невеликого масштабу. По-друге, супутникові виміри мають певні похибки через вплив атмосферних факторів, що знижує точність оцінок якості повітря. Крім того, дані потребують обробки та інтерпретації, що також вносить додаткові похибки. По-третє, супутники здійснюють спостереження лише

1–2 рази на добу для кожної точки. Цього недостатньо для своєчасного виявлення швидких змін якості повітря, наприклад, під час аварійних викидів. Отже, незважаючи на переваги, супутникові дані не можуть повністю замінити дані, отримані в процесі наземних спостережень. Найефективнішим є спільне використання даних з різних джерел для максимально точного моніторингу якості атмосферного повітря.

Саме з цієї причини актуальною є задача оцінювання даних наземних станцій моніторингу якості повітря на основі супутникових даних. Це дозволяє проводити більш детальний аналіз динаміки змін якості повітря та джерел забруднення, інформувати населення та забезпечувати підтримку прийняття рішень щодо поліпшення якості повітря. Ця задача ставиться в даній статті. Для її розв'язання пропонується використовувати МГУА, який вважається одним із перших методів глибокого навчання [13, 14].

Математична постановка задачі

Задача полягає у побудові математичної моделі регресійного типу для встановлення залежності між супутниковими даними x і наземними станціями y з урахуванням просторового розрізнення супутникових даних, яка має таке узагальнене представлення:

$$y(t) = F\begin{pmatrix} x_1(t),\, x_2(t),\, x_3(t),\ldots,\, x_m(t), \\ x_1(t-1),\, x_2(t-1),\, x_3(t-1),\ldots,\, x_m(t-1), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_1(t-k),\, x_2(t-k),\, x_3(t-k),\ldots,\, x_m(t-k) \end{pmatrix},\ m \in N,\ k \in N,$$

де $x_i(t-k),\ \ i=\overline{1;\, m},$ — значення показника x_i за супутниковими даними в момент часу $(t-k);\ \ m$ — ступінь просторового розрізнення для супутникових даних.

Математичний апарат побудови регресійної моделі

Для приведення супутникових даних до наземних застосуємо модифікацію класичного МГУА [15]. Вона полягає у виборі оптимальних моделей на кожному селективному шарі, що призводить одночасно до спрощення загальної математичної моделі відповідності між супутниковими та наземними даними і зменшення загальної похибки моделі у процесі її валідації на тестовій вибірці реальних геопросторових даних.

Цей метод забезпечує прогнозування показників довільного походження. З його допомогою виконується процедура відновлення нелінійної функціональної залежності між факторними змінними та прогнозованим показником, що від них залежить. Залежність $F(x)$ прогнозованого показника від набору факторних даних, що залежать від часу, представлено на рис. 1.

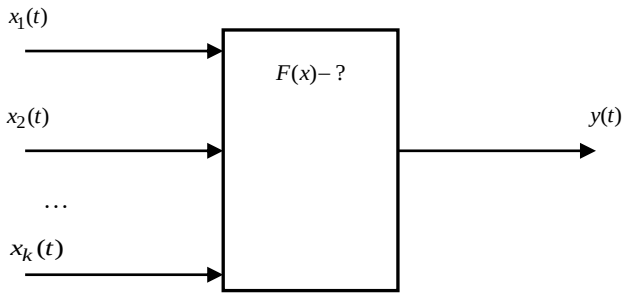

Рис. 1

Дані, отримані з супутників та наземних постів, можна подати у класичному вигляді протягом деякого дискретного часу t:

$$\left[\begin{array}{cccc|c} x_{11} & x_{12} & \cdots & x_{1m} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2m} & y_2 \\ \cdots & \cdots & \ddots & \vdots & \vdots \\ x_{t1} & x_{t2} & \cdots & x_{tm} & y_t \end{array}\right].$$

На основі статистичних даних потрібно визначити деяку функціональну залежність $y = F(x),$ загальний вигляд якої не є наперед відомим. Для найбільш чіткого опису залежності між факторними змінними *X* та прогнозованою величиною *Y* як базову математичну модель використовують узагальнений поліном Колмогорова–Габора. Для вибірки вигляду

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \ddots & \vdots \\ x_{t1} & x_{t2} & \cdots & x_{tm} \end{bmatrix}; \; X_k = \begin{bmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{tk} \end{bmatrix}; \; Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_t \end{bmatrix} \tag{1}$$

класичний поліном Колмогорова–Габора [15] може бути представлений у вигляді

$$Y = a + \sum_{k=1}^{m} b_k X_k + \sum_{k=1}^{m} \sum_{l=k}^{m} c_{kl} X_k X_l + \sum_{k=1}^{m} \sum_{l=k}^{m} \sum_{s=l}^{m} d_{kls} X_k X_l X_s + \ldots, \tag{2}$$

де значення коефіцієнтів $a, b_k, c_{kl}, d_{kls}, k = \overline{1, m}, l = \overline{1, m}, s = \overline{1, m}, \; k \le l \le s,$ невідомі. Для знаходження невідомих коефіцієнтів моделі (2) може бути застосований класичний метод найменших квадратів (МНК). Оскільки функції типу (2) є лінійними відносно невідомих коефіцієнтів, для МНК не потрібно застосовувати ніяких додаткових умов, тобто процедура знаходження цих коефіцієнтів зводиться до розв'язання звичайної системи лінійних алгебраїчних рівнянь із симетричною додатно визначеною матрицею.

Для побудови будь-якої математичної моделі типу (2) на статистичних даних, отриманих з реальних джерел, як критерій якості (точності) використовується помилка вигляду

$$err = \frac{1}{t} \sum_{k=1}^{t} (Y_k - F(x_{k1}, x_{k2}, \ldots, x_{km}))^2 \to \min, \tag{3}$$

де t — кількість вимірювань у часі (кількість рядків даних в (1)).

Основним завданням прогнозування є пошук найкращої математичної моделі з використанням МГУА.

Метод групового урахування аргументів передбачає побудову складної моделі на основі базових (опорних) поліномів Колмогорова–Габора (2). Також слід зазначити, що кожен поліном типу (2) має два параметри — кількість змінних, від яких залежить поліном, і ступінь даного поліному. У такому разі поліном *m*-го ступеня від *n* змінних можна позначити як $P(m; n).$ Тоді, наприклад, базова модель на основі поліному Колмогорова–Габора $P(2; 3)$ матиме вигляд

$$P(2; 3) = a + b_1 X_1 + b_2 X_2 + c_{11} X_1^2 + c_{12} X_1 X_2 + c_{22} X_2^2 +$$

$$+ d_{111} X_1^3 + d_{112} X_1^2 X_2 + d_{122} X_1 X_2^2 + d_{222} X_2^3, \tag{4}$$

ISSN 2786-6491

а базова модель на основі поліному Колмогорова–Габора $P(3;\,2)$ матиме вигляд

$$P(3;\,2) = a + b_1X_1 + b_2X_2 + b_3X_3 + c_{11}X_1^2 + c_{22}X_2^2 + c_{33}X_3^3 + \\ + c_{12}X_1X_2 + c_{13}X_1X_3 + c_{23}X_2X_3. \qquad (5)$$

Математичні моделі базових (опорних) моделей типів (4) і (5) також можуть бути використані в комбінації при побудові загальної нелінійної моделі $y = F(x)$. Звичайно, МГУА — це універсальний метод, тому замість поліномів Колмогорова–Габора можна використовувати інші функції, не лише поліноміального типу.

Для формування складної моделі $y = F(x)$ на основі поліномів типу (2) з метою мінімізації (4) застосовується принцип багаторядності та селекції.

Поділимо всю вибірку на дві частини: $t = \{t_{study}; t_{verification}\}$. Для побудови базової моделі типу (2) будемо використовувати частину вибірки t_{study} для визначення коефіцієнтів моделі. Верифікацію отриманої моделі на основі (3) будемо проводити з використанням залишкової частини вибірки $t_{verification}$:

$$err_{verivication} = \frac{1}{t_{verification}} \sum_{k=1}^{t_{verification}} (Y_k - F(x_{k1}, x_{k2}, \ldots, x_{km}))^2 \to \min. \qquad (6)$$

Побудова часткових модельних описів
на основі поліномів Колмогорова–Габора

1. Якщо математична модель, яка описує процес, має m факторів, а базові поліноми типу (2) мають вигляд $P(3, n)$, то це означає, що загальна кількість таких поліномів буде дорівнювати

$$\binom{m}{3} = \frac{m!}{3!(m-3)!} = \frac{m(m-1)(m-2)}{6},$$

а якщо поліноми типу (2) мають вигляд $P(k, n)$, то

$$\binom{m}{k} = \frac{m!}{k!(m-k)!} = \frac{m(m-1)(m-2)\cdots(m-k+1)}{k!}.$$

Як зазначалось вище, коефіцієнти цих поліномів знаходять з використанням МНК на основі вибірки t_{study}.

2. На основі вибірки $t_{verification}$ проводиться аналіз кожної моделі з п. 1 на основі (6) для визначення пристосованості моделі за критерієм точності.

3. Кращі N_1 моделей з п. 2 на основі (3) переходять на наступний етап. Виходи цих моделей є входами для нових моделей, які потрібно будувати, як вказано в п. 1. Це означає, що нових базових моделей виду $P(k, n)$ на новому кроці (на основі поліномів типу (2)) буде

$$\binom{N_1}{k} = \frac{N_1!}{k!(N_1-k)!} = \frac{N_1(N_1-1)(N_1-2)\cdots(N_1-k+1)}{k!}.$$

4. Відбір кращих моделей для побудови продовжується на основі пп. 1–3 до тих пір, поки значення (6) кращої моделі на попередньому кроці буде меншим за значення (3) кращої моделі на останньому кроці.

5. Після проходження даними базовими моделями з кінця до початку будується загальна нелінійна залежність $y = F(x)$, яка є композицією кращих базових моделей усіх кроків відбору. Деревовидну структуру залежностей базових моделей на основі поліномів Колмогорова–Габора виду (2) показано на рис. 2.

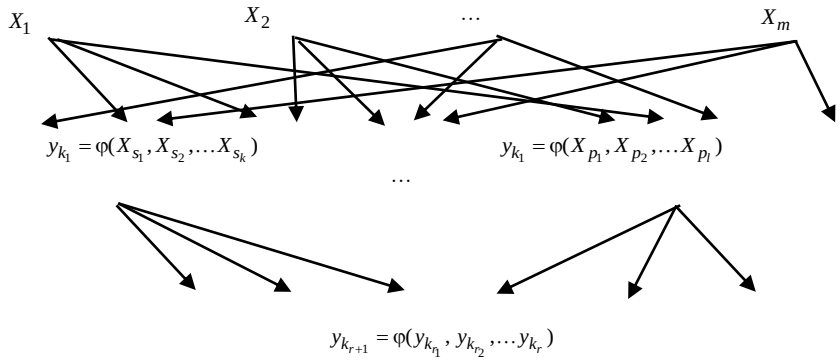

Рис. 2

На рис. 2 показано, що кожна наступна модель приймає на вхід результат попередньої. Тобто, наприклад, на деякому кроці селекції є дві моделі:

$$\begin{aligned} P_1(2;3) &= a^{(1)} + b_1^{(1)} X_1 + b_2^{(1)} X_2 + c_{11}^{(1)} X_1^2 + c_{12}^{(1)} X_1 X_2 + c_{22}^{(1)} X_2^2 + \\ &+ d_{111}^{(1)} X_1^3 + d_{112}^{(1)} X_1^2 X_2 + d_{122}^{(1)} X_1 X_2^2 + d_{222}^{(1)} X_2^3, \\ P_2(2;3) &= a^{(2)} + b_1^{(2)} X_1 + b_2^{(2)} X_2 + c_{11}^{(2)} X_1^2 + c_{12}^{(2)} X_1 X_2 + c_{22}^{(2)} X_2^2 + \\ &+ d_{111}^{(2)} X_1^3 + d_{112}^{(2)} X_1^2 X_2 + d_{122}^{(2)} X_1 X_2^2 + d_{222}^{(2)} X_2^3. \end{aligned} \tag{7}$$

Тоді модель $P_3(2;3)$, яка залежить від двох змінних, $P_1(2;3)$ та $P_2(2;3)$, на основі (7) матиме такий вигляд:

$$\begin{aligned} P_3(2;3) &= \varphi(P_1(2;3); P_2(2;3)) = \\ &= a^{(3)} + b_1^{(3)} P_1(2;3) + b_2^{(3)} P_2(2;3) + c_{11}^{(3)} P_1^2(2;3) + c_{12}^{(3)} P_1(2;3) P_2(2;3) + \\ &+ c_{22}^{(3)} P_2^2(2;3) + d_{111}^{(3)} P_1^3(2;3) + d_{112}^{(3)} P_1^2(2;3) P_2(2;3) + \\ &+ d_{122}^{(3)} P_1(2;3) P_2^2(2;3) + d_{222}^{(3)} P_2^3(2;3). \end{aligned} \tag{8}$$

Використовуючи (8) і (7), отримуємо остаточний вираз для $P_3(2;3)$:

$$\begin{aligned} P_3(2;3) &= \varphi(P_1(2;3); P_2(2;3)) = \\ &= a^{(3)} + b_1^{(3)} \left(\begin{array}{c} a^{(1)} + b_1^{(1)} X_1 + b_2^{(1)} X_2 + c_{11}^{(1)} X_1^2 + c_{12}^{(1)} X_1 X_2 + c_{22}^{(1)} X_2^2 + \\ + d_{111}^{(1)} X_1^3 + d_{112}^{(1)} X_1^2 X_2 + d_{122}^{(1)} X_1 X_2^2 + d_{222}^{(1)} X_2^3 \end{array} \right) + \\ &+ b_2^{(3)} \left(\begin{array}{c} a^{(2)} + b_1^{(2)} X_1 + b_2^{(2)} X_2 + c_{11}^{(2)} X_1^2 + c_{12}^{(2)} X_1 X_2 + c_{22}^{(2)} X_2^2 + \\ + d_{111}^{(2)} X_1^3 + d_{112}^{(2)} X_1^2 X_2 + d_{122}^{(2)} X_1 X_2^2 + d_{222}^{(2)} X_2^3 \end{array} \right) + \end{aligned}$$

ISSN 2786-6491

$$+c_{11}^{(3)}\left(\begin{array}{c}a^{(1)}+b_1^{(1)}X_1+b_2^{(1)}X_2+c_{11}^{(1)}X_1^2+c_{12}^{(1)}X_1X_2+c_{22}^{(1)}X_2^2+\\+d_{111}^{(1)}X_1^3+d_{112}^{(1)}X_1^2X_2+d_{122}^{(1)}X_1X_2^2+d_{222}^{(1)}X_2^3\end{array}\right)^2+$$

$$+c_{12}^{(3)}\left(\begin{array}{c}a^{(1)}+b_1^{(1)}X_1+b_2^{(1)}X_2+c_{11}^{(1)}X_1^2+c_{12}^{(1)}X_1X_2+c_{22}^{(1)}X_2^2+\\+d_{111}^{(1)}X_1^3+d_{112}^{(1)}X_1^2X_2+d_{122}^{(1)}X_1X_2^2+d_{222}^{(1)}X_2^3\end{array}\right)\times$$

$$\times\left(\begin{array}{c}a^{(2)}+b_1^{(2)}X_1+b_2^{(2)}X_2+c_{11}^{(2)}X_1^2+c_{12}^{(2)}X_1X_2+c_{22}^{(2)}X_2^2+\\+d_{111}^{(2)}X_1^3+d_{112}^{(2)}X_1^2X_2+d_{122}^{(2)}X_1X_2^2+d_{222}^{(2)}X_2^3\end{array}\right)+$$

$$+c_{22}^{(3)}\left(\begin{array}{c}a^{(2)}+b_1^{(2)}X_1+b_2^{(2)}X_2+c_{11}^{(2)}X_1^2+c_{12}^{(2)}X_1X_2+c_{22}^{(2)}X_2^2+\\+d_{111}^{(2)}X_1^3+d_{112}^{(2)}X_1^2X_2+d_{122}^{(2)}X_1X_2^2+d_{222}^{(2)}X_2^3\end{array}\right)^2+$$

$$+d_{111}^{(3)}\left(\begin{array}{c}a^{(1)}+b_1^{(1)}X_1+b_2^{(1)}X_2+c_{11}^{(1)}X_1^2+c_{12}^{(1)}X_1X_2+c_{22}^{(1)}X_2^2+\\+d_{111}^{(1)}X_1^3+d_{112}^{(1)}X_1^2X_2+d_{122}^{(1)}X_1X_2^2+d_{222}^{(1)}X_2^3\end{array}\right)^3+$$

$$+d_{112}^{(3)}\left(\begin{array}{c}a^{(1)}+b_1^{(1)}X_1+b_2^{(1)}X_2+c_{11}^{(1)}X_1^2+c_{12}^{(1)}X_1X_2+c_{22}^{(1)}X_2^2+\\+d_{111}^{(1)}X_1^3+d_{112}^{(1)}X_1^2X_2+d_{122}^{(1)}X_1X_2^2+d_{222}^{(1)}X_2^3\end{array}\right)^2\times$$

$$\times\left(\begin{array}{c}a^{(2)}+b_1^{(2)}X_1+b_2^{(2)}X_2+c_{11}^{(2)}X_1^2+c_{12}^{(2)}X_1X_2+c_{22}^{(2)}X_2^2+\\+d_{111}^{(2)}X_1^3+d_{112}^{(2)}X_1^2X_2+d_{122}^{(2)}X_1X_2^2+d_{222}^{(2)}X_2^3\end{array}\right)+$$

$$+d_{122}^{(3)}\left(\begin{array}{c}a^{(1)}+b_1^{(1)}X_1+b_2^{(1)}X_2+c_{11}^{(1)}X_1^2+c_{12}^{(1)}X_1X_2+c_{22}^{(1)}X_2^2+\\+d_{111}^{(1)}X_1^3+d_{112}^{(1)}X_1^2X_2+d_{122}^{(1)}X_1X_2^2+d_{222}^{(1)}X_2^3\end{array}\right)\times$$

$$\times\left(\begin{array}{c}a^{(2)}+b_1^{(2)}X_1+b_2^{(2)}X_2+c_{11}^{(2)}X_1^2+c_{12}^{(2)}X_1X_2+c_{22}^{(2)}X_2^2+\\+d_{111}^{(2)}X_1^3+d_{112}^{(2)}X_1^2X_2+d_{122}^{(2)}X_1X_2^2+d_{222}^{(2)}X_2^3\end{array}\right)^2+$$

$$+d_{222}^{(3)}\left(\begin{array}{c}a^{(2)}+b_1^{(2)}X_1+b_2^{(2)}X_2+c_{11}^{(2)}X_1^2+c_{12}^{(2)}X_1X_2+c_{22}^{(2)}X_2^2+\\+d_{111}^{(2)}X_1^3+d_{112}^{(2)}X_1^2X_2+d_{122}^{(2)}X_1X_2^2+d_{222}^{(2)}X_2^3\end{array}\right)^3. \qquad (9)$$

З наведеного вище видно, що у процесі композиції моделей 3-го ступеня отримується загальна модель 9-го ступеня (9). Обчислення коефіцієнтів здійснюється поетапно, що дозволяє уникнути машинного округлення, яке призводить до втрати даних та хибних результатів побудованої загальної математичної моделі.

Тобто, якщо дві моделі 3-го ступеня об'єднуються в одну, отримується модель 9-го ступеня, а якщо дві моделі 2-го ступеня об'єднуються в одну — модель 4-го ступеня. Звідси можна зробити висновок про загальну складність моделі:

$$\varphi(P_1(k;n);\ldots;P_s(k;n))=P(k;n^2). \qquad (10)$$

Виходячи з отриманих описів математичних моделей на прикладі (1), можна одержати складну залежність на невеликій вибірці з даними. У зв'язку з тим, що

на першому рівні можуть відсіятися деякі факторні змінні X_i, що мають суттєвий вплив на вихідні дані, на другому рівні селекції на вхід подають y_i та x_j, наприклад, таким чином:

$$\varphi(X, Y) = a_0^{(2)} + a_1^{(2)} Y_k + a_2^{(2)} X_l + a_3^{(2)} Y_k^2 + a_4^{(2)} Y_k X_l + a_5^{(2)} X_l^2. \quad (11)$$

Використовуючи МГУА, слід зазначити, що, крім відновлення складної залежності між вхідними та вихідними даними, легко здійснюється адаптація параметрів математичної моделі при одержанні нових даних експериментів. В обчислювальних експериментах замість (10) буде використовуватись (11) для зменшення похибки моделі $y = F(x)$.

Після застосування МГУА виконується процес визначення коефіцієнта кореляції вхідних спостережень з виходом побудованої математичної моделі за такою формулою:

$$\rho_{yx_i} = \sum_{i=1}^{m} (y_i - \bar{y})(x_{ij} - \bar{x}_i) \Big/ \sqrt{\sum_{i=1}^{m} (y_i - \bar{y})^2 \sum_{i=1}^{m} (x_{ij} - \bar{x}_i)^2}. \quad (12)$$

Крім (12), можна використовувати й інші методи кореляційного аналізу для встановлення міри зв'язку між реальними і модельними даними.

Набори даних

З відкритих джерел [16, 17] отримано масиви даних для міста Києва, що містять показники $PM_{2,5}$ та PM_{10} супутникових даних (IDCams) і наземних даних (IDPost) за період з 01.01.2019 по 19.11.2020 року. Набір наземних даних містить дані вказаних вище показників з дискретизацією в часі 3 хв для кожної доби і займає 30 млн рядків (обсяг даних — 1,4 Гбайт). Набір супутникових даних містить дані вказаних вище показників з дискретизацією в часі 1 год для кожної доби і займає 8269 рядків (загальний обсяг даних — 9,15 Мбайт). Усього для збору даних було задіяно 213 наземних постів (за IDPost), які покриваються 12 об'єктами супутникових знімків з просторовим розрізненням 11 км. У табл. 1 представлено відповідності між IDCams та IDPost.

Таблиця 1

№ п/п	ID посту	*x* ID посту	*y* ID посту	ID CAMS	№ п/п	ID посту	*x* ID посту	*y* ID посту	ID CAMS
1	28	50,44400	30,54000	6563	101	3583	50,43741	30,47888	6563
2	30	50,43400	30,43200	6562	102	3600	50,52125	30,49671	6193
3	43	50,41172	30,61895	6564	103	3601	50,51800	30,48600	6378
4	47	50,47294	30,50825	6378	104	3603	50,37390	30,59883	6749
5	108	50,36259	30,44274	6747	105	3619	50,52100	30,58700	6194
…	…	…	…	…	…	…	…	…	…
95	3514	50,38263	30,45813	6563	209	13802	50,49351	30,50624	6378
96	3535	50,39713	30,45571	6563	210	13803	50,43740	30,59567	6564
97	3541	50,36400	30,49660	6748	211	13811	50,41288	30,60786	6564
98	3547	50,38300	30,47605	6563	212	13820	50,49908	30,57751	6379
99	3572	50,40935	30,63961	6564	213	13853	50,41821	30,46421	6563

Супутникові дані мають вигляд часових рядів, як показано на рис. 3 (значення показника $PM_{2,5}$ отримані з використанням супутникових даних IDCams = 6563) і на рис. 4 (значення показника $PM_{2,5}$ отримані з використанням супутникових даних IDCams = 6563) для $PM_{2,5}$ та PM_{10} відповідно протягом доби.

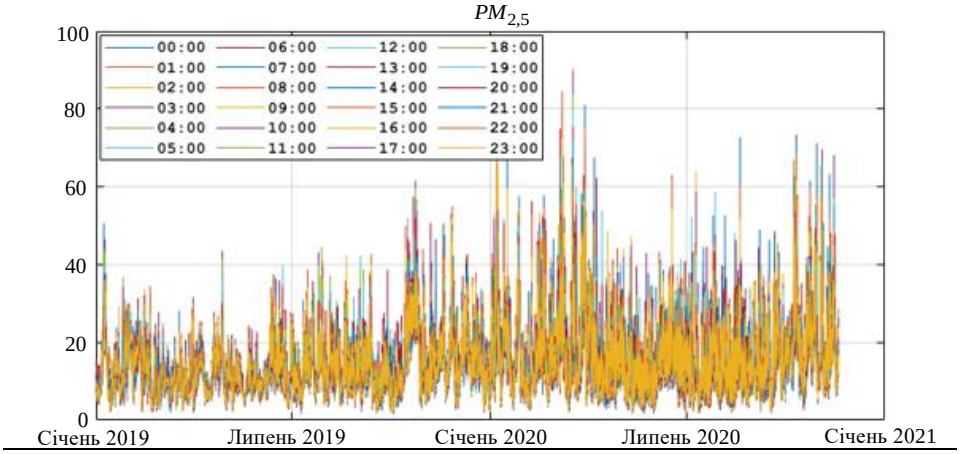

Рис. 3

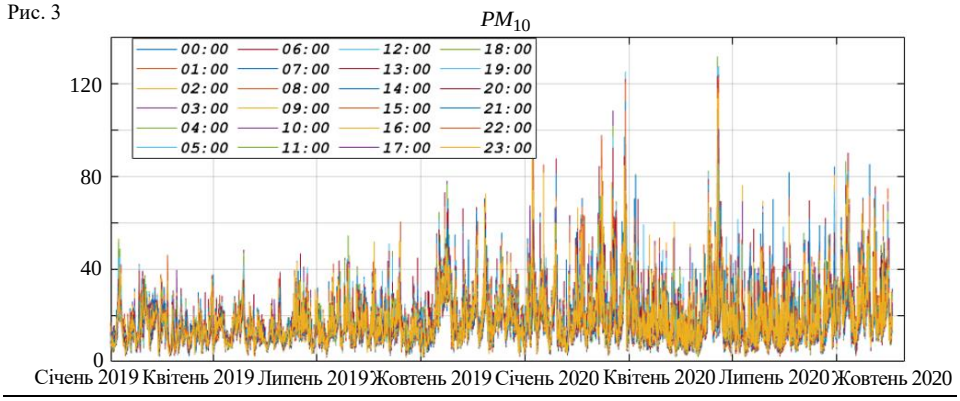

Рис. 4

Після фільтрування наземних даних з частотою дискретизації 3 хв одержуємо наземні дані з частотою дискретизації 1 год. Графічне представлення таких даних для $PM_{2,5}$ наведено на рис. 5, а для PM_{10} — на рис. 6.

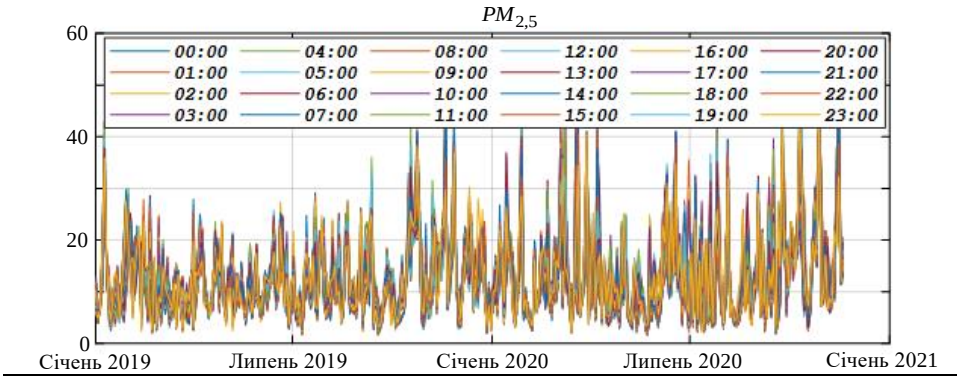

Рис. 5

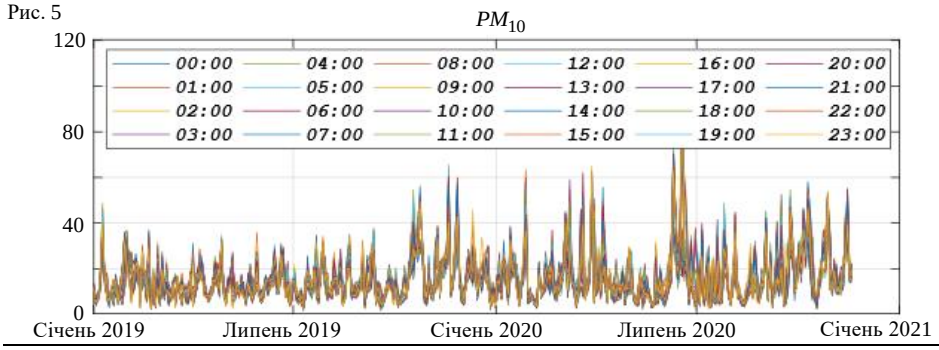

Рис. 6

Часові ряди, які зображено на рис. 3–6, передбачають подання статистичних даних для $PM_{2,5}$ та PM_{10}, отриманих на основі супутникових і наземних даних, у вигляді 24 точок на день.

Приведення супутникових даних до наземних проводиться на основі звичайного подання — у вигляді одного часового ряду для кожного з показників, які було згадано вище.

Попередня обробка даних

Як вже зазначалось, за допомогою МГУА відновлюється деяка невідома нелінійна залежність виду $y = F(x)$, де x — факторні змінні, y — прогнозована (результуюча, залежна від факторів) змінна. Для побудови математичної моделі будемо вважати, що y — це значення показника на основі наземних вимірювань IDPost, яке залежить від 12 значень, що отримуються з супутникових джерел IDCams. Тому базова модель має вигляд

$$y(t) = F(x_1(t), x_2(t), x_3(t), \ldots, x_{12}(t)), \tag{13}$$

де $x_i(t)$, $i = \overline{1;12}$, — значення показника, який отримується на основі i-го супутникового джерела у деякий момент часу t. Очевидно, що наземний показник залежить від значень супутникових показників у деякий попередній момент часу. Це означає, що на основі моделі (13) будується більш узагальнена математична модель, що має такий вигляд:

$$y(t) = F\begin{pmatrix} x_1(t), x_2(t), x_3(t), \ldots, x_{12}(t), \\ x_1(t-1), x_2(t-1), x_3(t-1), \ldots, x_{12}(t-1), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ x_1(t-k), x_2(t-k), x_3(t-k), \ldots, x_{12}(t-k) \end{pmatrix}, k \in N, \tag{14}$$

де $x_i(t-k)$, $i = \overline{1;12}$, — значення показника x_i у момент часу $(t-k)$.

Обчислювальні експерименти

Для побудови остаточної математичної моделі було взято значення затримки в часі (в годинах) $k_1 = 2, k_2 = 2, k_3 = 3$. Початкову вибірку було розбито на навчальну, 75 % (6202 рядки даних), та перевірочну, 25 % (2067 рядків даних), вибірки. Кількість селективних рівнів для побудови деревовидної структури залежності поліномів типу (2) програмно обиралась автоматично на основі принципу спадання похибки на перевірочній вибірці для кожного з показників.

У табл. 2 наведено значення кореляції між модельними та реальними даними, що були отримані під час проведення різних експериментів для даних показників $PM_{2,5}$ та PM_{10} відповідно.

Таблиця 2

Показник $PM_{2,5}$				Показник PM_{10}			
	$k=1$	$k=2$	$k=3$		$k=1$	$k=2$	$k=3$
$P(4; 2)$	0,87041	0,91028	0,92525	$P(4; 2)$	0,90577	0,92352	0,88204
$P(5; 2)$	0,91298	0,92961	0,9234	$P(5; 2)$	0,86751	0,96254	0,94583
$P(4; 3)$	0,90399	0,92363	0,94015	$P(4; 3)$	0,91909	0,91124	0,93776
$P(5; 3)$	0,90617	0,95654	0,94284	$P(5; 3)$	0,94456	0,96008	0,96295

ISSN 2786-6491

У табл. 3 представлено кількість селективних рівнів, на яких МГУА завершує свою роботу для побудови загальної залежності виду (14).

Таблиця 3

Показник $PM_{2,5}$				Показник PM_{10}			
	$k=1$	$k=2$	$k=3$		$k=1$	$k=2$	$k=3$
$P(4;2)$	**7**	**5**	6	$P(4;2)$	8	6	6
$P(5;2)$	6	5	5	$P(5;2)$	6	5	6
$P(4;3)$	5	6	4	$P(4;3)$	6	7	5
$P(5;3)$	4	3	3	$P(5;3)$	4	5	4

На рис. 7 та 8 показано модельні дані кращих регресійних моделей, які побудовані з використанням МГУА за критерієм рівня кореляційного зв'язку (з поліномами $P(5;3), k=3$) для показників $PM_{2,5}$ та PM_{10} відповідно.

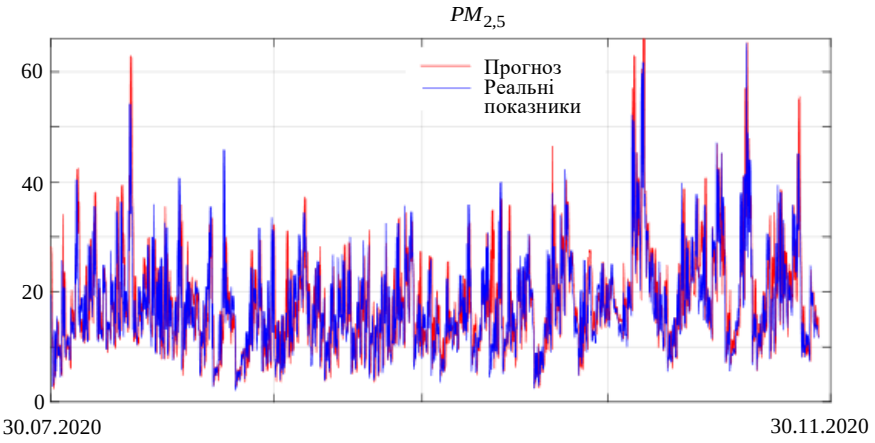

Рис. 7

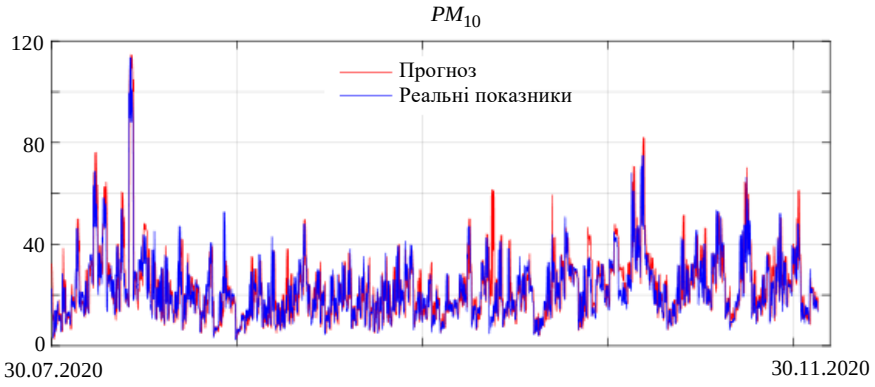

Рис. 8

Кращі результати проведеного моделювання було показано на рівні кореляційного зв'язку 0,8889 на перевірочній вибірці 25 % (2067 точок) для показника $PM_{2,5}$ та 0,8895 — для показника PM_{10}. У табл. 4 представлено кореляційні залежності між прогнозованими станціями та реальними значеннями для деяких станцій одного джерела супутникових даних $(PM_{2,5})$.

Таблиця 4

ID Post	Коефіцієнт кореляції	ID Post	Коефіцієнт кореляції	ID Post	Коефіцієнт кореляції
1161	0,8750	3601	0,8152	13802	0,8231
1274	0,8579	3621	0,8383	13789	0,8305
1306	0,8889	3702	0,7863	13413	0,8407
1387	0,8336	4191	0,8291	13388	0,8069
1425	0,7770	4232	0,8252	13208	0,8001
1651	0,8128	12935	0,8371	13207	0,7456
2741	0,8028	12940	0,8038	13141	0,8021
3483	0,8248	3498	0,8834	12978	0,8059

Виходячи з отриманих результатів, слід зазначити, що побудована математична модель на основі МГУА дає досить непогані результати прогнозування значень на наземних постах на основі супутникових даних.

Переваги застосованого підходу

Застосований підхід до моделювання має суттєві переваги порівняно з використанням штучних нейронних мереж. Першою перевагою є те, що МГУА будує структуру функціональної мережі (залежності) оптимальним способом. Це означає, що виконується не лише оптимальний пошук параметрів моделі; оптимальним також є вигляд моделі (залежність між змінними). Другою суттєвою перевагою застосованого підходу є зниження складності побудованої моделі порівняно з тими ж штучними нейронними мережами. МГУА для досягнення однакової точності моделей на перевірочній вибірці потребує такої кількості параметрів моделі, яка в 2–3 рази менша за кількість параметрів моделі, що будується на основі класичних штучних нейронних мереж. Третьою перевагою підходу є те, що МГУА менш схильний до перенавчання при використанні такого роду даних. Побудована математична модель може бути модифікована методами регресійного аналізу і використана для пошуку уточнених значень забруднення повітря у тих точках, де немає наземних постів [18].

Висновок

Модель групового урахування аргументів дозволяє систематизувати інформацію та оцінити різні фактори роботи об’єктів та систем для прийняття обґрунтованих рішень. У даній роботі з використанням запропонованого методу отримано регресійні математичні моделі приведення даних з супутникових показників до даних наземних показників для $PM_{2,5}$ та PM_{10} відповідно. Проведено низку обчислювальних експериментів, які підтверджують ефективність і коректність підходу до оцінки даних наземних станцій з використанням супутникових даних та моделі МГУА. Такий підхід дозволяє отримати більш повне та точне уявлення про розподіл забруднюючих речовин у повітрі та покращити якість моніторингової системи, а також сприяє вдосконаленню стратегій зменшення забруднення та розробці науково обґрунтованих рішень для збереження природних ресурсів та здоров’я населення.

MODEL OF AIR QUALITY ASSESSMENT ACCORDING TO SATELLITE DATA BASED ON THE GROUP METHOD OF DATA HANDLING

Vladyslav Khaidurov

National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»,

4labs0@gmail.com

Bohdan Yailymov

Institute of Space Research of the NAS of Ukraine and State Space Agency of Ukraine, Kyiv,

yailymov@gmail.com

Andrii Shelestov

National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»,

andrii.shelestov@gmail.com

This paper presents a mathematical model based on the group method of data handling (GMDH) for estimating ground-level air quality data using satellite observations. Air pollution is a serious environmental problem with significant impacts on human health, ecosystems and climate change. Ground-based air quality monitoring networks provide direct pollution measurements, but are limited by the number of stations in many regions of the world. Satellite remote sensing offers new opportunities for consistent and detailed monitoring of air quality as a supplement to ground-based observations. However, there are limitations, including low spatial resolution of satellite data, measurement uncertainties, and low acquisition frequency. This study developed a modified GMDH model to compare satellite observations with ground air quality data for fine particulate matter (PM2,5) and particulate matter less than 10 microns (PM10) in the Kyiv city, Ukraine. The model optimally reconstructs nonlinear functional dependencies between time series of satellite and ground variables, while optimizing the overall complexity of the model. Several computational experiments were performed on real data sets. The results showed a strong correlation between predicted and empirically observed values on an independent 25 % test sample, reaching 0,8889 for PM2,5. The optimized GMDH model required 2–3 times fewer parameters than a comparable neural network architecture to achieve the same level of accuracy. This demonstrates the ability of the proposed approach to accurately estimate ground-level pollution concentrations at high resolution using satellite data through GMDH simulations. The developed model provides a more complete spatiotemporal picture of pollution distribution to significantly improve environmental monitoring capabilities, inform the public, and support science-based policy decisions regarding mitigation strategies. In summary, the study highlights that the fusion of satellite and ground-based data using GMDH modeling significantly improves air quality assessment capabilities to better understand small-scale pollution dynamics, protect public health, and develop effective solutions to protect the environment.

Keywords: mathematical models, regression model, Kolmogorov–Gabor polynomial, group method of data handling, air quality, satellite data, correlation analysis.

ПОСИЛАННЯ

1. Lehmann A., Mazzetti P., Santoro M., Nativi S., Maso J., Serral I., Spengler D., Niamir A., Lacroix P., Ambrosone M., McCallum I., Kussul N., Patias P., Rodila D., Ray N., Giuliani G. Essen-

tial earth observation variables for high-level multi-scale indicators and policies. *Environmental Science & Policy*. 2022. Vol. 131. P. 105–117. DOI: https://doi.org/10.1016/j.envsci.2021.12.024
2. Shelestov A., Yailymova H., Yailymov B., Kussul N. Air quality estimation in Ukraine using SDG 11.6.2 indicator assessment. *Remote Sensing*. 2021. Vol. 13, N 23. P. 4769. DOI: https://doi.org/10.3390/rs13234769
3. Shelestov A., Yailymova H., Yailymov B., Samoilenko O., Shumilo L. Ground based validation of Copernicus atmosphere monitoring service data for Kyiv. *IEEE EUROCON 2021–19th International Conference on Smart Technologies*. Ukraine: Lviv, 2021. P. 88–91. DOI: https://ela.kpi.ua/handle/123456789/48529
4. Jethva H., Chand D., Torres O., Gupta P., Lyapustin A., Patadia F. Agricultural burning and air quality over northern India: a synergistic analysis using NASA's A-train satellite data and ground measurements. *Aerosol and Air Quality Research*. 2018. Vol. 18, N 7. P. 1756–1773. DOI: https://doi.org/10.4209/aaqr.2017.12.0583
5. Cheng J., Su J., Cui T., Li X., Dong X., Sun F., Yang Y., .Tong D., Zheng Y., Li Y., Li J., Zhang Q., He K. Dominant role of emission reduction in $PM_{2.5}$ air quality improvement in Beijing during 2013–2017: a model-based decomposition analysis. *Atmospheric Chemistry and Physics*. 2019. Vol. 19, N 9. P. 6125–6146. DOI: https://doi.org/10.5194/acp-19-6125-2019
6. Kaur G., Gao J., Chiao S., Lu S., Xie G. Air quality prediction: big data and machine learning approaches. *International Journal of Environmental Science and Development*. 2018. Vol. 9, N 1. P. 8–16. DOI: https://doi.org/10.18178/ijesd.2018.9.1.1066
7. Mellios G., Van Aalst R., Samaras Z. Validation of road traffic urban emission inventories by means of concentration data measured at air quality monitoring stations in Europe. *Atmospheric Environment.* 2006. Vol. 40, N. 38. P. 7362–7377. DOI: https://doi.org/10.1016/j.atmosenv.2006.06.044
8. Shelestov A., Kolotii A., Borisova T., Turos O., Milinevsky G., Gomilko I., Bulanay T., Fedorov O., Shumilo L., Pidgorodetska L., Kolos L., Borysov A., Pozdnyakova N., Chunikhin A., Dudarenko M., Petrosian A., Danylevsky V., Miatselskaya N., Choliy V. Essential variables for air quality estimation. *International Journal of Digital Earth*. 2020. Vol. 13, N 2. P. 278–298. DOI: https://doi.org/10.1080/17538947.2019.1620881
9. Shelestov A., Kolotii A., Lavreniuk M., Medyanovskyi K., Vasiliev V., Bulanaya T., Gomilko I. Air quality monitoring in urban areas using in-situ and satellite data within era-planet project. *In IGARSS 2018-2018 IEEE International geoscience and remote sensing symposium*. Spain : Valencia, 2018. P. 1668–1671. DOI: https://doi.org/10.1109/IGARSS.2018.8518368
10. Atmospheric composition forecasts move to higher resolution. 2016. https://atmosphere.copernicus.eu/atmospheric-composition-forecasts-move-higher-resolution
11. Inness A., Ades M., Agustí-Panareda A., Barre J., Benedictow A., Blechschmidt A., Dominguez J.-J., Engelen R., Eskes H., Flemming J., Huijnen V., Jones L., Kipling Z., Massart S., Parrington M., Peuch V., Razinger M., Rémy S., Schulz M., Suttie M. The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics*. 2019. Vol. 19, N 6. P. 3515–3556. DOI: https://doi.org/10.5194/ACP-19-3515-2019
12. Ehret T., Truchis A., Mazzolini M., Morel J., d'Aspremont A., Lauvaux T., Duren R., Cusworth D., Facciolo G. Global tracking and quantification of oil and gas methane emissions from recurrent Sentinel-2 imagery. *Environmental Science & Technology*. 2022. Vol. 56, N 14. P. 10517–10529. DOI: https://doi.org/10.1021/acs.est.1c08575
13. Schmidhuber J. Deep learning in neural networks: an overview. *Neural Networks*. 2015. Vol. 61. P. 85–117. DOI: https://doi.org/10.1016/j.neunet.2014.09.003
14. Farlow S. The GMDH algorithm of Ivakhnenko. *The American Statistician*. 1981. Vol. 35, N 4. P. 210–215. DOI: https://doi.org/10.2307/2683292
15. Anastasakis L., Mort N. The development of self-organization techniques in modelling: a review of the group method of data handling (GMDH). United Kingdom : Department of Automatic Control & Systems Engineering The University of Sheffield Mappin St, Sheffield, S1 3JD. 2001. N 813. 38 p.
16. Український екологічний чат-бот SaveEcoBot. https://www.saveecobot.com
17. Copernicus atmosphere monitoring service. https://atmosphere.copernicus.eu/charts/packages/cams.
18. Yailymova H., Kolotii A., Kussul N., Shelestov A. Air quality as proxy for assesment of economic activity. *IEEE EUROCON 2023-20th International Conference on Smart Technologies*. Italy : Torino, 2023. P. 89–92. DOI: https://doi.org/10.1109/EUROCON56442.2023.10198882

Отримано 25.08.2023