УДК 519.22:504.06

ЗГЛАДЖУВАННЯ НЕПОВНИХ РЯДІВ ДАНИХ СТАНЦІЙ ЕКОЛОГІЧНОГО МОНІТОРИНГУ З ВИКОРИСТАННЯМ ПРОГНОЗНИХ МОДЕЛЕЙ

Густера О. М.
к.е.н., старший викладач кафедри комп'ютерних наук
ORCID https://orcid.org/0000-0003-1010-6100
E-mail: o.gustera@nubip.edu.ua
Ніколаєнко Д. В.
к.е.н., старший викладач кафедри комп'ютерних наук
ORCID https://orcid.org/0009-0008-4817-3951
E-mail: d.nikolaenko@nubip.edu.ua
Національний університет біоресурсів і природокористування України, Україна.
Героїв оборони 15, м. Київ, Україна

Анотація. У статті розглядається проблема неповних рядів даних при аналізі даних *станцій екологічного моніторингу, її вплив на достовірність отриманих результатів та прогностичну придатність вхідних даних, а також методи згладжування даних, що дозволяють мінімізувати негативний вплив пропусків даних.*
Відсутність даних в ряді може проявлятись на практиці як хибні нульові значення, які можуть призводити до суттєвих відхилень, а також як відсутні дані, що в деяких випадках приховують тенденції зміни динаміки ряду. При цьому, аналітик може не знати про присутність пустих або нульових значень, що, в результаті, призводить до хибних висновків або прогнозів.
Методи згладжування за допомогою простої ковзної середньої та екстраполяції дозволяють підвищити якість вхідних даних, та, як результат, підвищити прогностичну якість отриманих прогнозних моделей.
Використання локальних прогнозів для заповнення пропущених значень дозволяє отримати найбільш точні результати замість відсутніх даних, і, як результат, підвищити прогностичну якість розроблюваних прогнозних моделей. Точність результатів отриманих замість відсутніх даних перевіряється розрахунком основних статистичних показників ряду з пустими значеннями та повного ряду. Розрахунок параметрів моделей прогнозування для заповнення пустих інтервалів може здійснюватись на основі попередніх даних або тенденції всього ряду.
Отримані в даному дослідженні результати можуть бути використані у подальшому для заповнення неповних рядів при аналізі даних станцій екологічного моніторингу або інших рядів даних, що використовуються для прогнозування або аналітичних розрахунків.
Ключові слова: *екологічний моніторинг, ряд даних, динаміка ряду даних, пропуски даних, прогноз, прогнозні моделі.*

Вступ

У процесі аналізу екологічного стану шляхом отримання даних від станцій моніторингу надійність отриманих результатів напряму залежить від точності початкових даних, які отримані від датчиків та збережені у сховище даних[8]. При цьому, в сучасних умовах може виникати проблема нерівномірних інтервалів дослідження або відсутності частини спостережень, що може бути пов'язано з негативним впливом наступних факторів[6]:

- відсутність електропостачання у екологічній станції;
- відсутність зв'язку з екологічною станцією через проблеми зі зв'язком;

- некоректна робота апаратного або програмного забезпечення після відключень електроенергії;
- інші фактори, що спричиняють збої у роботі екологічної станції (DDOS-атаки, проблеми у провайдера та ін..);
- цілеспрямоване приховування або знищення даних третіми особами.

В результаті, підсумкові дані за досить тривалі періоди можуть бути відсутні, як показано на рисунку 1:

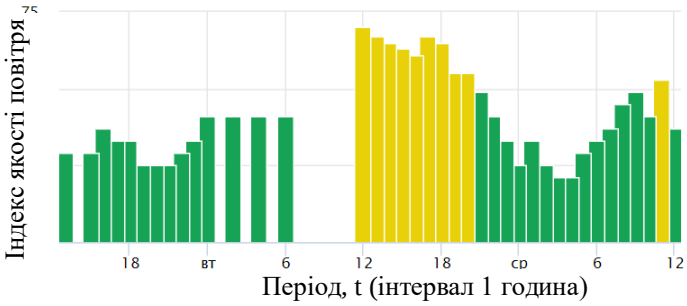


Рисунок **1.** – Індекс якості повітря [18]

Ряд даних показника якості повітря, представлений на Рис. 1, у період з 12.00 10.01.2023 до 12.00 12.01.2023 є неповним. З 6.00 до 12.00 11.01.2023 дані відсутні. В таких умовах отримання достовірних та повних рядів даних від станції екологічного моніторингу ускладнюється. При цьому, неможливо заздалегідь спрогнозувати періоди, коли отримані дані будуть неповними або не коректними. Тобто, дізнатися про те, що отримані дані не можуть бути використані для отримання достовірних результатів аналізу, як правило, можна лише у процесі аналізу[8].

Огляд літератури

На сьогоднішній день питання екологічного моніторингу як необхідної складової частини та інформаційної бази для захисту навколишнього середовища є особливо актуальним, та висвічується в багатьох зарубіжних та вітчизняних джерелах [3,12,14,19].

В умовах постійно зростаючої кількості станцій екологічного моніторингу та інших пристроїв, що дозволяють отримувати дані про окремі показники стану навколишнього середовища зростає актуальність проблеми зберігання, обробки та аналізу накопиченої інформації для її подальшого використання [4,7,11].

Серед сучасних вітчизняних робіт у галузі екологічного моніторингу слід відзначити розробку інформаційно-аналітичної системи оцінювання стану атмосферного повітря [1,2].

Сучасні методи інтелектуального аналізу дозволяють досить ефективно вирішувати задачі інтерпретації необроблених даних за умови їх точності та достовірності [4,5]. Таким чином, першочергове питання, якому слід приділяти достатньо уваги при екологічному моніторингу – саме отримання достовірних та точних вхідних даних, які можуть бути спотворені через ряд об'єктивних чи суб'єктивних чинників.

Мета статті

Методологічне обґрунтування

На практиці найбільш негативний вплив неповних рядів даних проявляється у використанні недостовірних даних для аналізу без усвідомлення їх придатності до аналізу, що може призводити до більш негативних наслідків. Тобто, аналітики або особа що приймає рішення отримує готові результати та не усвідомлює, що вони побудовані на частково хибних даних[10].

Необхідність заповнення пропусків у ряді даних підтверджується тим, що інтервали без значень можуть бути сприйняті як нульові значення, або ж не будуть враховані у загальній тенденції ряду. В деяких випадках відрізнити реальні нульові значення від пропущених спостережень досить складно. Так, наприклад, температура повітря може приймати нульове значення, яке може бути переплутане з пустим значенням.

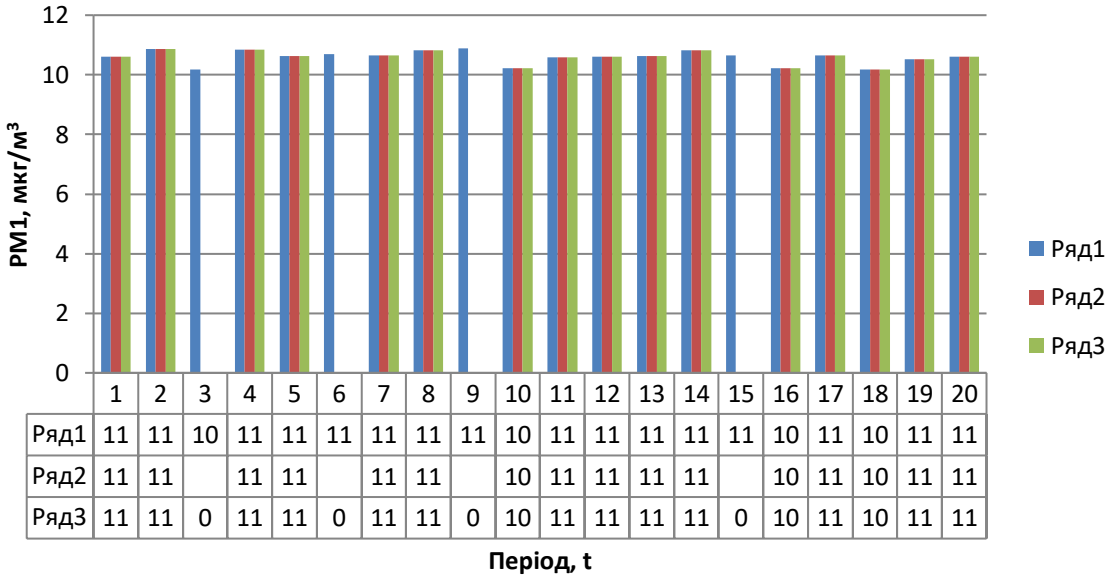
Відновлення пропусків у часових рядах може вирішуватись за допомогою наступних груп методів:

- прості статистичні методи (екстраполяція, ковзна середня),
- ітеративне прогнозування,
- комбіновані схеми прогнозування.

Перевагою простих статистичних методів їх простота в реалізації, що особливо зручно при автоматизованій обробці вхідних даних[12].

Результати обговорення

Для того щоб більш детально проаналізувати можливі наслідки використання неповних або некоректних рядів даних без їх обробки або адаптації розглянемо наступний приклад. Розглянемо один і той самий ряд даних показника PM1 - дрібнодисперсні частки у повітрі, діаметром менше 1 мкм (мікрону) у трьох можливих варіантах спостереження – без пропусків, з пропусками, та з хибними значеннями замість пропусків. Хибні значення в даному прикладі це нульові значення.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Ряд1	11	11	10	11	11	11	11	11	11	10	11	11	11	11	11	10	11	10	11	11
Ряд2	11	11		11	11		11	11		10	11	11	11	11		10	11	10	11	11
Ряд3	11	11	0	11	11	0	11	11	0	10	11	11	11	11	0	10	11	10	11	11

Рисунок 2. – Початковий ряд даних показника PM1(дрібнодисперсні частки у повітрі, діаметром менше 1 мкм (мікрону)), Ряд 1 – без пропусків, Ряд 2 – з пропусками, Ряд 3 - з хибними значеннями замість пропусків (16)

У першому випадку (ряд 1) ми отримуємо повний ряд даних, і на його основі можемо побудувати прогноз на майбутній період, визначити основні статистичні характеристики ряду. В другому випадку (ряд 2) проаналізуємо аналогічний ряд, у якому будуть відсутні дані за певні періоди, тобто зімітуємо неповний ряд даних. У третьому випадку (ряд 3) замість відсутніх даних будемо використовувати некоректні дані - нулі. Для того щоб визначити, наскільки суттєво повнота ряду впливає на отримані результати, аналогічно до повного ряду визначимо його основні статистичні характеристики – середнє значення, дисперсію та середньоквадратичне відхилення (табл. 1).

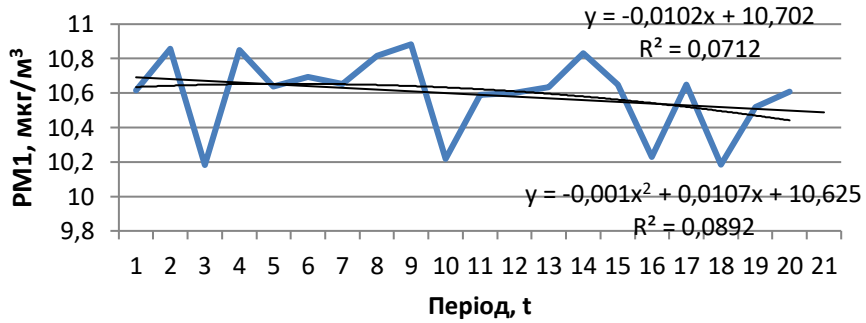
Таблиця 1 – Основні статистичні характеристики досліджуваного часового ряду

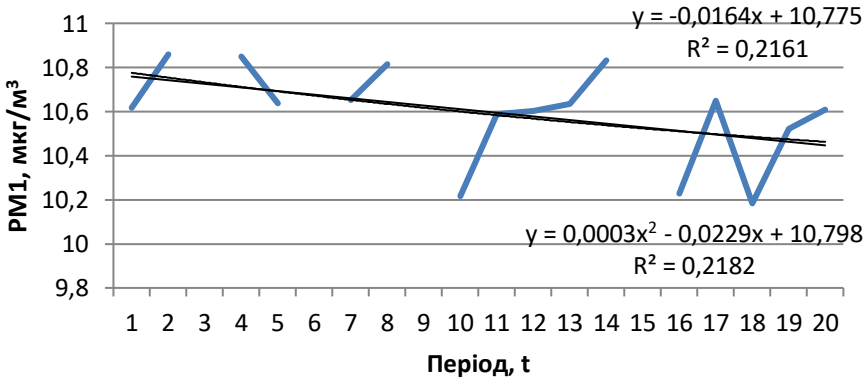
	Ряд 1	Ряд 2	Ряд 3
Середнє	10,59493	10,59324	8,474591
Дисперсія	0,048268	0,043738	17,98966
Стандартне відхилення	0,219701	0,209135	4,241422

Як видно з результатів, представлених у таблиці 1, найбільш суттєво на отримані результати впливає використання для подальшого аналізу чи прогнозування використання некоректних даних. Врахування нульових значень призводить до збільшення дисперсії та стандартного відхилення у декілька разів, що, в результаті, робить побудування прогнозних моделей нераціональним через великі довірчі інтервали або низьку точність прогнозу. При цьому, врахування нульових значень залежить від абсолютних характеристик ряду даних.

Відсутність даних за певні періоди також негативно впливає на точність отриманих результатів. У випадках, коли у пропущених інтервалах відбувались суттєві зміни тенденції ряду даних, це не буде помічено та враховано при побудуванні прогнозних моделей або розрахунку аналітичних показників[13].

Для оцінки впливу врахування нульових та пропущених інтервалів на результати прогнозу використаємо лінійну та поліноміальну модель. Для побудування прогнозу на основі обраних моделей використаємо MS Excel (Рис. 3).





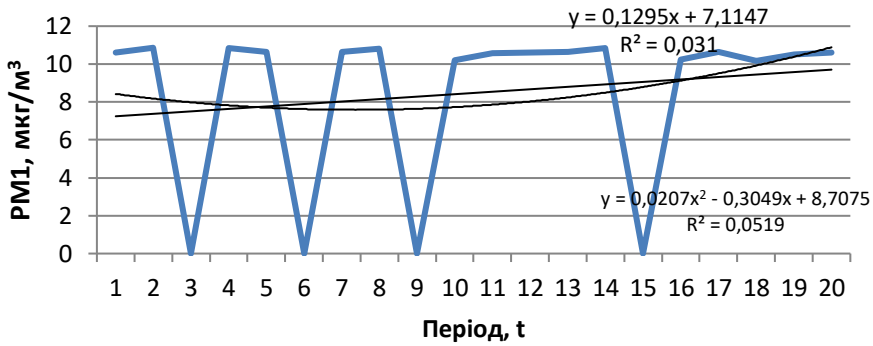


Рисунок 3. – Прогнозування на основі поліноміальної та лінійної моделі для повного ряду даних, ряду з порожніми значеннями, та ряду з нульовими значеннями

Як видно з рисунку 3, використання необроблених вхідних даних для аналізу чи прогнозування може призводити до отримання результатів з низькою достовірністю або прогностичною придатністю.

Для того щоб заповнити дані за певні періоди спочатку використаємо метод ковзної середньої, тобто використаємо середнє значення за найближчі періоди для періоду згладжування m:

$$\hat{x}_i = \frac{1}{p} \sum_{j=t-m}^{t+m} x_j \tag{1}$$

Також відсутні дані можна замінити методом екстраполяції існуючих:

$$\hat{x}_i = \begin{cases} x_{i-1}, \text{ якщо } x = 0, \\ x_i, \text{ якщо } x > 0. \end{cases} \tag{2}$$

Розглянемо використання методу простої ковзної середньої для заповнення відсутніх елементів ряду даних.

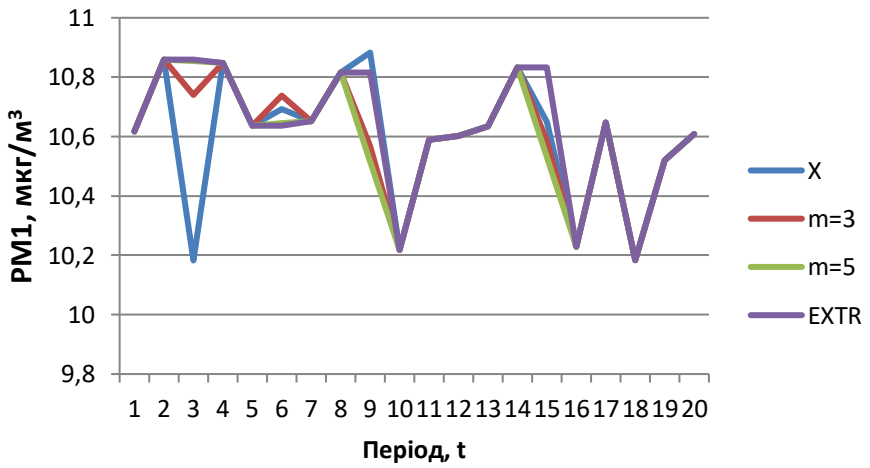


Рисунок 4. – Використання простої ковзної середньої для заповнення відсутніх даних часового ряду (період згладжування m=3, m=5) та екстраполяції

Таблиця 2. – Основні статистичні характеристики досліджуваного часового ряду (X) та ряду отриманого методом згладжування (m=3, m=5) та методом екстраполяції (EXTR)

	X	m=3	m=5	EXTR
Середнє	10,59493	10,60622	10,60185	10,63172
Дисперсія	0,048268	0,036987	0,038935	0,042445
Стандартне відхилення	0,219701	0,19232	0,197319	0,206021

Як видно з результатів розрахунку основних статистичних характеристик досліджуваного часового ряду (X), ряду отриманого методом згладжування (m=3, m=5) та методом екстраполяції (EXTR), наведених у таблиці 2, навіть прості методи згладжування дають більш достовірні результати у порівнянні з екстраполяцією. Тим не менш, методи згладжування хоча й можуть бути достатньо ефективно використані з метою усунення пропусків або некоректних нульових значень, однак не дозволяють врахувати наявність тенденції у часовому ряді і можливість її прояву саме у пропущеному інтервалі.

З метою врахування тенденції часового ряду при заповненні ряду використаємо прогнозування (3), реверсивне прогнозування (4) для локальних інтервалів на основі лінійної та поліноміальної моделі:

$$\hat{x}_i = \begin{cases} a_0 + a_1 x_{i-1} \vee a_0 + a_1 x_{i-1} + a_2 x_{i-1}{}^2, \ \textit{якщо} \ x = 0, \\ x_i, \ \textit{якщо} \ \ x > 0. \end{cases} \tag{3}$$

$$\hat{x}_i = \begin{cases} a_0 + a_1 x_{i+1} \vee a_0 + a_1 x_{i+1} + a_2 x_{i+1}{}^2, \ \textit{якщо} \ x = 0, \\ x_i, \ \textit{якщо} \ \ x > 0. \end{cases} \tag{4}$$

Для лінійної моделі, використаємо наступні параметри (Рис. 1.) :

$a_0 = 10,702$

$a_1 = -0,0102$

Для поліноміальної моделі, відповідно:

$a_0 = 10,625$

$a_1 = 0,0107$

$a_2 = -0,001$

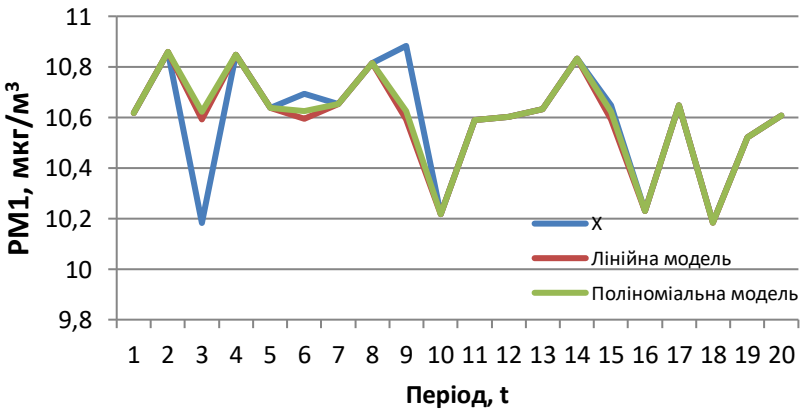


Рисунок 5. – Використання локального прогнозування для заповнення відсутніх даних часового ряду (лінійна та поліноміальна модель)

Таблиця 3. – Основні статистичні характеристики досліджуваного часового ряду (X) та ряду отриманого шляхом заповнення порожніх значень методом локального прогнозування для заповнення відсутніх даних часового ряду (лінійна та поліноміальна модель)

	X	Лінійна модель	Поліноміальна модель
Середнє	10,59298848	10,59940492	10,59298848
Дисперсія	0,034990411	0,035142256	0,034990411
Стандартне відхилення	0,18705724	0,18746268	0,18705724

Для побудування реверсивного прогнозу виконаємо транспонування початкового ряду даних та розрахуємо параметри моделей:

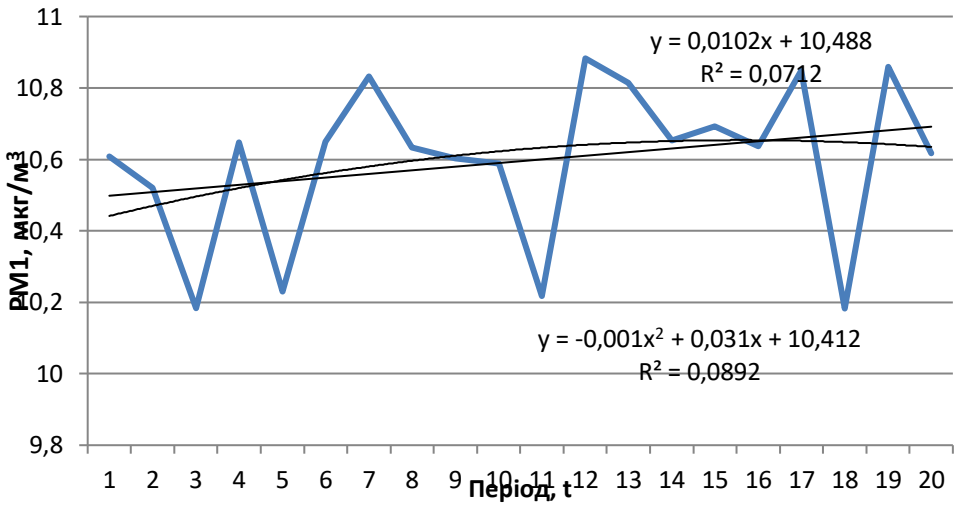


Рисунок 6. – Розрахунок параметрів лінійної та поліноміальної моделі для реверсивного прогнозування

Таким чином, для лінійної моделі, використаємо наступні параметри:

$a_0 = 10,488$

$a_1 = 0,0102$

Для поліноміальної моделі, відповідно:

$a_0 = 10,412$

$a_1 = 0,031$

$a_2 = -0,001$

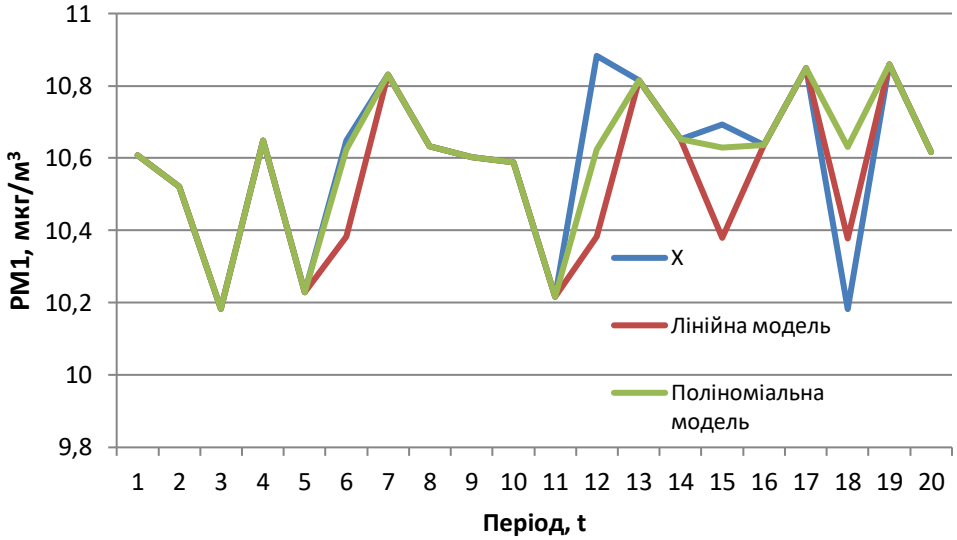


Рисунок 7. – Використання локального реверсивного прогнозування для заповнення відсутніх даних часового ряду (лінійна та поліноміальна модель)

Таблиця 4. – Основні статистичні характеристики досліджуваного часового ряду (X) та ряду отриманого шляхом заповнення порожніх значень методом локального реверсивного прогнозування для заповнення відсутніх даних часового ряду (лінійна та поліноміальна модель)

	X	Лінійна модель	Поліноміальна модель
Середнє	10,59493	10,5508	10,6
Дисперсія	0,048268	0,042197	0,035174
Стандартне відхилення	0,219701	0,205418	0,187548

Для досліджуваного прикладу, згідно з результатами отриманими з таблиці 3 та таблиці 4, більш точні результати при заповненні порожніх значень ряду дозволяє отримати метод локального прогнозування. В той же час, в деяких випадках використання реверсивного прогнозування є єдиним можливим способом отримання відсутніх значень. Наприклад, якщо порожніми є одні з перших спостережень, виявлення чіткої тенденції на основі декількох попередніх спостережень або при їх повній відсутності не дає достовірних результатів. У таких випадках, коли реверсивне прогнозування дає менш точні результати у порівнянні із прогнозуванням на основі попередніх значень, і при цьому відсутні початкові значення ряду, можливе використання комбінованого методу. Тобто, для частини порожніх значень використовується прогнозування на основі попередніх значень, а там де це є необхідним – на основі послідуючих значень, тобто реверсивне прогнозування.

Висновки

Використані у дослідженні методи згладжування за допомогою простої ковзної середньої та екстраполяції дозволяють підвищити якість вхідних даних, та, як результат, підвищити прогностичну якість отриманих прогнозних моделей. В той же час, застосування локальних прогнозів для заповнення пропущених значень дозволяє отримати найбільш точні результати замість відсутніх даних, і, як результат, підвищити прогностичну якість розроблюваних прогнозних моделей. Точність результатів отриманих замість відсутніх даних перевіряється розрахунком основних статистичних показників ряду з пустими значеннями та повного ряду. Розрахунок параметрів моделей прогнозування для заповнення пустих інтервалів може здійснюватись на основі попередніх даних або тенденції всього ряду.

Отримані в даному дослідженні результати можуть бути використані у подальшому для заповнення неповних рядів при аналізі даних станцій екологічного моніторингу або інших рядів даних, що використовуються для прогнозування або аналітичних розрахунків. Вибір моделі прогнозування або аналітичного методу, за допомогою яких здійснюється згладжування та заповнення порожніх значень даних може залежати від загальної тенденції часового ряду.

Список використаних джерел

1. Bogolyubov V.M., Golub B.L. Information-analytical system for assessing the state of atmospheric air / Sustainable development — 21st century. Discussions 2021: collective monograph / National University - Kyiv-Mohyla Academy / edited by Prof. Khlobistova E.V. — Kyiv, 2021. — 469 p. ISBN: 978-617-7668-22-9.
2. Bogoliubov V.M. and other. Optimization of the Structure of Atmospheric Air Monitoring System https://chmnu.edu.ua/wpcontent/uploads/MONOGRAPH-2.pdf
3. Khan, S.; Anjum, R.; Raza, S.T.; Ahmed Bazai, N.; Ihtisham, M. Technologies for Municipal Solid Waste Management: Current Status, Challenges, and Future Perspectives. Chemosphere 2022, 288, 132403.
4. Andeobu, L.; Wibowo, S.; Grandhi, S. A Systematic Review of E-waste Generation and Environmental Management of Asia Pacific Countries. Int. J. Environ. Res. Public Health 2021, 18, 9051.
5. Ma, S.; Zhou, C.; Chi, C.; Liu, Y.; Yang, G. Estimating Physical Composition of Municipal Solid Waste in China by Applying Artificial Neural Network Method. Environ. Sci. Technol. 2020, 54, 9609–9617.
6. Lin, K.; Zhao, Y.; Tian, L.; Zhao, C.; Zhang, M.; Zhou, T. Estimation of Municipal Solid Waste Amount Based on One-Dimension C N Network and Long Short-Term Memory with Attention Mechanism Model: A Case Study of Shanghai. Sci. Total Environ. 2021, 791, 148088.
7. Sharma, M.; Joshi, S.; Kannan, D.; Govindan, K.; Singh, R.; Purohit, H. Internet of Things (IoT) Adoption Barriers of Smart cities' Waste Management: An Indian Context. J. Clean. Prod. 2020, 270, 122047.
8. Jassim, M.S.; Coskuner, G.; Zontul, M. Comparative Performance Analysis of Support Vector Regression and Artificial Neural Network for Prediction of Municipal Solid Waste Generation. Waste Manage. Res. 2022, 40, 195–204.
9. Lin, K.; Zhao, Y.; Kuo, J.-H.; Deng, H.; Cui, F.; Zhang, Z.; Zhang, M.; Zhao, C.; Gao, X.; Zhou, T.; et al. Toward smarter management and recovery of municipal solid waste: A critical review on deep learning approaches. J. Clean. Prod. 2022, 346, 130943.
10. Fasano, F.; Addante, A.S.; Valenzano, B.; Scannicchio, G. Variables Influencing per Capita Production, Separate Collection, and Costs of Municipal Solid Waste in the Apulia Region (Italy): An Experience of Deep Learning. Int. J. Environ. Res. Public Health 2021, 18, 752.

11. Hussain, A.; Draz, U.; Ali, T.; Tariq, S.; Irfan, M.; Glowacz, A.; Antonino Daviu, J.A.; Yasin, S.; Rahman, S. Waste Management and Prediction of Air Pollutants Using IoT and Machine Learning Approach. Energies 2020, 13, 3930.

12. Ihsanullah, I.; Alam, G.; Jamal, A.; Shaik, F. Recent Advances in Applications of Artificial Intelligence in Solid Waste Management: A Review. Chemosphere 2022, 309, 136631.

13. Hettiarachchi, H.; Meegoda, J.N.; Ryu, S. Organic Waste Buyback as a Viable Method to Enhance Sustainable Municipal Solid Waste Management in Developing Countries. Int. J. Environ. Res. Public Health 2018, 15, 2483.

14. Rahman, M.W.; Islam, R.; Hasan, A.; Bithi, N.I.; Hasan, M.M.; Rahman, M.M. Intelligent Waste Management System Using Deep Learning with IoT. J. King Saud Univ.-Comput. Inf. Sci. 2022, 34, 2072–2087.

15. Mookkaiah, S.S.; Thangavelu, G.; Hebbar, R.; Haldar, N.; Singh, H. Design and Development of Smart Internet of Things–based Solid Waste Management System Using Computer Vision. Environ. Sci. Pollut. Res. 2022, 29, 64871–64885.

16. Nowakowski, P.; Pamuła, T. Application of Deep Learning Object Classifier to Improve E-waste Collection Planning. Waste Manag. 2020, 109, 1–9.

17. Niu, D.; Wu, F.; Dai, S.; He, S.; Wu, B. Detection of Long-Term Effect in Forecasting Municipal Solid Waste Using a Long Short-Term Memory Neural Network. J. Clean. Prod. 2021, 290, 125187.

18. Сайт моніторингу рівня забруднення атмосферного повітря у місті Київ https://www.saveecobot.com/maps/kyiv

19. Woodward, W. A., Gray, H. L. & Elliott, A. C. (2012), *Applied Time Series Analysis*, CRC Press

20. Needleman, Saul B. & Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology*.