Д.О. Шмундяк

ПОРІВНЯЛЬНИЙ АНАЛІЗ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ПОКАЗНИКА СТАНУ АТМОСФЕРНОГО ПОВІТРЯ

Вінницький національний технічний університет, Україна

Періодичні часові ряди зустрічаються в багатьох аспектах нашого життя. Прикладами періодичних часових рядів є показники атмосферного повітря, показники фінансових ринків, метеорологічні параметри, тощо. Через це аналіз та прогнозування періодичних часових рядів є розповсюдженим та досить цікавим науковим напрямком. Одними з основних проблем при моделюванні періодичних часових рядів є визначення параметрів сезонності цього ряду, а також ідентифікація та усунення аномальних значень, які можуть суттєво впливати на точність прогнозування. У даній роботі проведено порівняний аналіз розроблених моделей та підходів прогнозування показнику стану атмосферного повітря за реальними даними мережі громадського моніторингу якості атмосферного повітря EcoCity. Наведено опис методу ідентифікації параметрів сезонності часового ряду, що базується на декомпозиції ряду. Наведено підхід пошуку локальних у часі аномалій в часовому ряду за виділення окремих півхвиль цього ряду. Результати роботи описаних моделей застосовано для прогнозування показника пилу PM2.5 однієї з станцій моніторингу якості повітря у Вінницькій області. Для автоматизації процесу прогнозування застосовувалась мова програмування Python, а сам програмний код реалізовано у системі Kaggle – веб-платформи від компанії Google для інженерів машинного навчання. Для прогнозування використовувалась модель для роботи з часовими рядами Prophet. Було наведено порівняльну таблицю точності прогнозу моделі Prophet з налаштуваннями за замовчуванням та з різними комбінаціями наборів аномальних значень та налаштувань сезонності. Дослідження та аналіз показав, що застосування як комбінації, так і окремо кожної розробленої моделі дозволяє зменшити помилку прогнозу для показника якості атмосферного повітря. У порівнянні з точністю роботи моделі Prophet з параметрами за замовчуванням, для найкращого з варіантів вдалось зменшити значення помилки за показником MAE на 30%, а за показником RMSE – на 21%. Це продемонструвало, що дані методи є ефективними для аналізу та прогнозування часових рядів, в тому числі безпосередньо часових рядів показників стану атмосферного повітря.

Ключові слова: аналіз часових рядів, аномалії часових рядів, машинне навчання, сезонна декомпозиція, гармоніки ряду Фур'є, Prophet, якість атмосферного повітря, EcoCity.

Вступ

В науковій літературі згадується досить велика кількість методів та підходів щодо прогнозування періодичних часових рядів. Широкого застосування мають підходи з використанням непромережених моделей як LSTM. Багато дослідників використовують моделі на основі авторегресії та проінтегрованого ковзного середнього (АРПКС – англ. «ARIMA») [1, 2]. Відносно новим та досить потужним інструментом також є модель Facebook Prophet (FB Prophet) [3, 4]. Кожна з моделей здатна ефективно використовуватися для аналізу та прогнозування періодичних часових рядів, але наявність в даних аномальних значень зазвичай призводить до погіршення точності цих прогнозів і тому вимагає попередньої обробки з ціллю ідентифікувати та усунуті ці аномалії. Також, виходячи з власного попереднього досвіду роботи з періодичними часовими рядами показників стану довкілля, відомо , що ці показники зазвичай мають певну циклічність, спричиненою різними природніми факторами та забрудненням. Попередні наукові дослідження були спрямовані на розробку моделей та підходів, які дозволяють виконувати пошук локальних у часі аномалій часового ряду та ідентифікувати параметри сезонності, що в подальшому може використовуватися для побудови ефективних моделей для роботи з часовими рядами. В попередніх роботах за мету ставилось питання апроксимації часового ряду, а тому доцільно спробувати застосувати ці моделі для прогнозу значень часового ряді і оцінити їх точність.

Метою роботи є визначення ефективності використання розроблених моделей прогнозування показника стану атмосферного повітря та порівняння з іншими методами прогнозування періодичних часових рядів.

Дані для дослідження

Одним з ключових елементів дослідження є наявність якісних даних для аналізу та прогнозування. В даному дослідженні використовувались дані моніторингу якості атмосферного повітря від мережі громадського моніторингу EcoCity. Мережа EcoCity складається з великої кількості станцій, які збирають інформацію по певних показниках якості повітря та передають її на сервери мережі. Як результат, користувачі веб-порталу EcoCity можуть майже в режимі реального часу спостерігати за значеннями відповідних показників у відповідній точці. Інтерфейс веб-порталу EcoCity зображено на рис. 1.

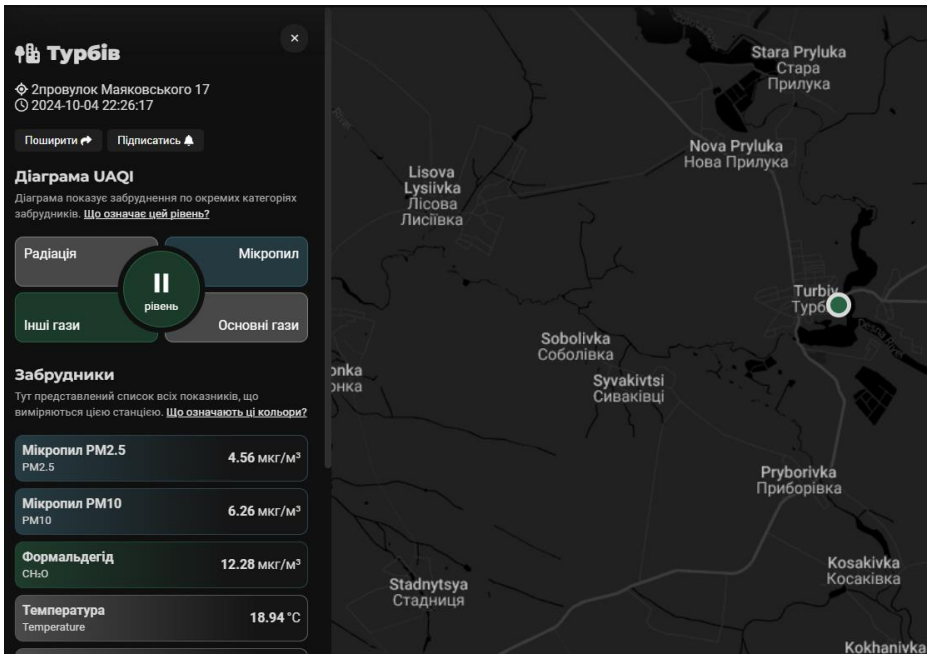

Рис. 1. Сторінка веб-порталу EcoCity

Безпосередньо в цьому дослідженні використовувались щоденні дані показника пилу PM2.5 (мікроскопічні тверді частинки), отримані з однієї з станцій моніторингу у Вінницькій області за період 2021-2024 рр. Графік зміни значень цього показника в часі наведено на рис. 2.

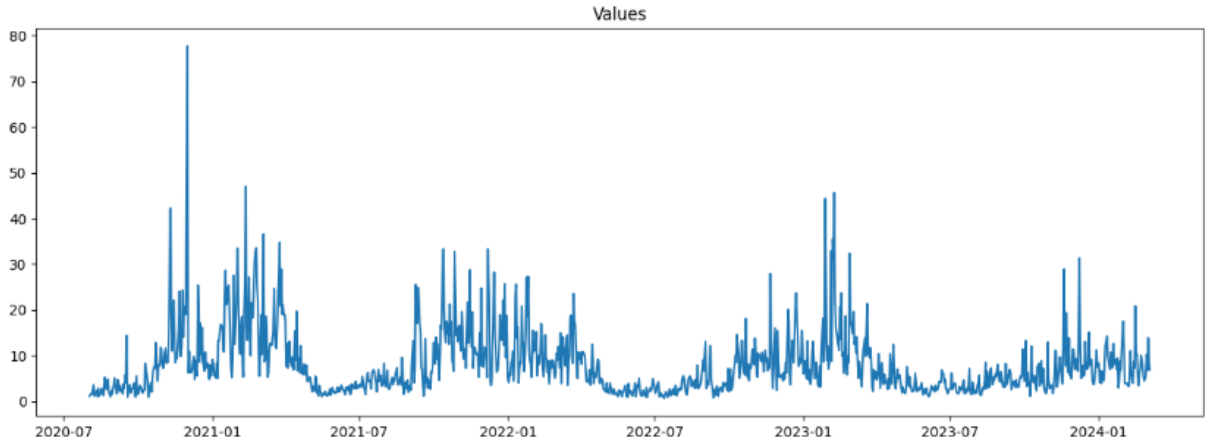

Рис. 2. Дані показника пилу PM2.5 за 2021-2024 рр.

Метод ідентифікації параметрів сезонності

У роботі [5] було розроблено метод ідентифікації параметрів гармонік та аномалій періодичного часового ряду на основі адаптивної декомпозиції. Суть методу полягає в побудові так званої

«декомпозиційної кривої» - графік усіх варіантів відношень амплітуд $S_Y(P)$ з періодом P від 1 до 50% від кількості усіх значень ряду [5, 6]:

$$S_Y(P) = \frac{S(P)}{Y}, P = 1,...0.5N, \quad (1)$$

де N – кількість значень ряду.

Після додаткового згладжування кривої, для неї виконується пошук локальних максимумів – точок кривої, де зростання $S_Y(P)$ змінюється на спадання. Отримані точки є значеннями різних періодів сезонності ряду, які в подальшому використовуються для синтезу декількох налаштувань сезонності з використанням прийомів, які детально описані в згаданій роботі [5].
Застосуємо даний метод для даних показника пилу PM2.5 та знайдемо значення періодів для нашого часового ряду. Згладжена «декомпозиційна крива» та її локальні максимуми зображено на рис. 3.

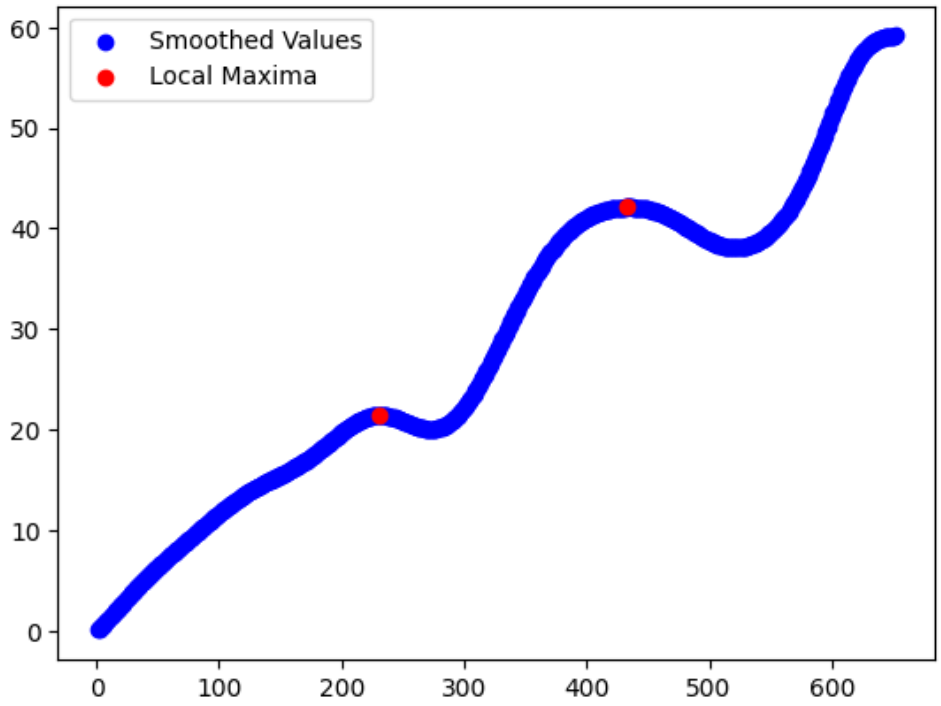

Рис. 3. Декомпозиційна крива часового ряду та локальні максимуми цієї кривої

Пошук локальних максимумів для отриманої кривої дозволив знайти два значення періоду: $P_1 = 230, P_2 = 433$ діб.

Метод пошуку локальних аномалій

В наукової роботі [7] було запропоновано метод пошуку локальних аномалій за рахунок декомпозиції часового ряду на півхвилі. Суть даного методу полягала в гіпотезі, що часовий ряд складається з декількох окремих фрагментів, кожен з яких виділений локальними мінімумами тренду цього ряду. Замість пошуку аномалій усього ряду загалом, було запропоновано розглядати кожен з фрагментів окремо і відповідно застосовувати методи пошуку аномалій на цьому відрізку. Результатом ж роботи методу була сукупність усіх ідентифікованих аномальних значень. В рамках цих досліджень вдалось підтвердити гіпотезу та довести, що подібний підхід дозволяє покращити точність ідентифікації аномальних значень.
Застосуємо описаний метод для показника пилу PM2.5 та виділимо окремі фрагменти часового ряду. Результат використання методу та поділ часового ряду на фрагменти наведено на рис. 4.

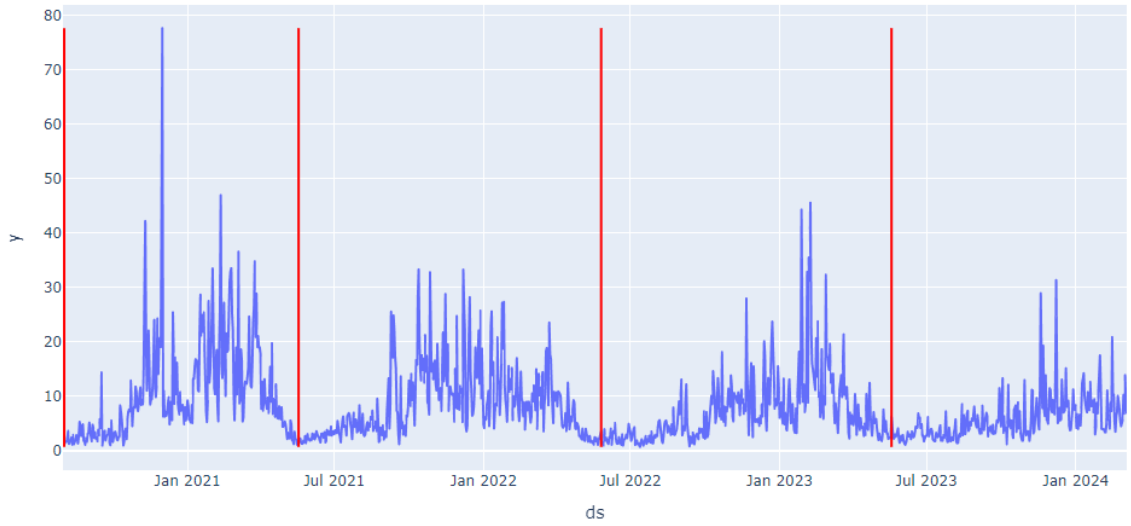

Рис. 4. Фрагменти часового ряду показника пилу PM2.5

На кожному з фрагментів застосуємо метод пошуку аномалій «коефіцієнт локального відхилення» (англ. «Local outlier factor» або «LOF») – алгоритм, який дозволяє знайти локальну густину точок в наборі [8, 9]. Метод використовується для визначення наскільки кожен об'єкт набору даних відрізняється від своїх найближчих сусідів і тим самим дозволяє ідентифікувати аномальні значення. Даний метод гарно себе зарекомендував в попередніх дослідженнях [7], які також були пов’язані з даними показників якості атмосферного повітря. Загальний набір аномальних значень часового ряду показника пилу PM2.5 зображено на рис. 5.

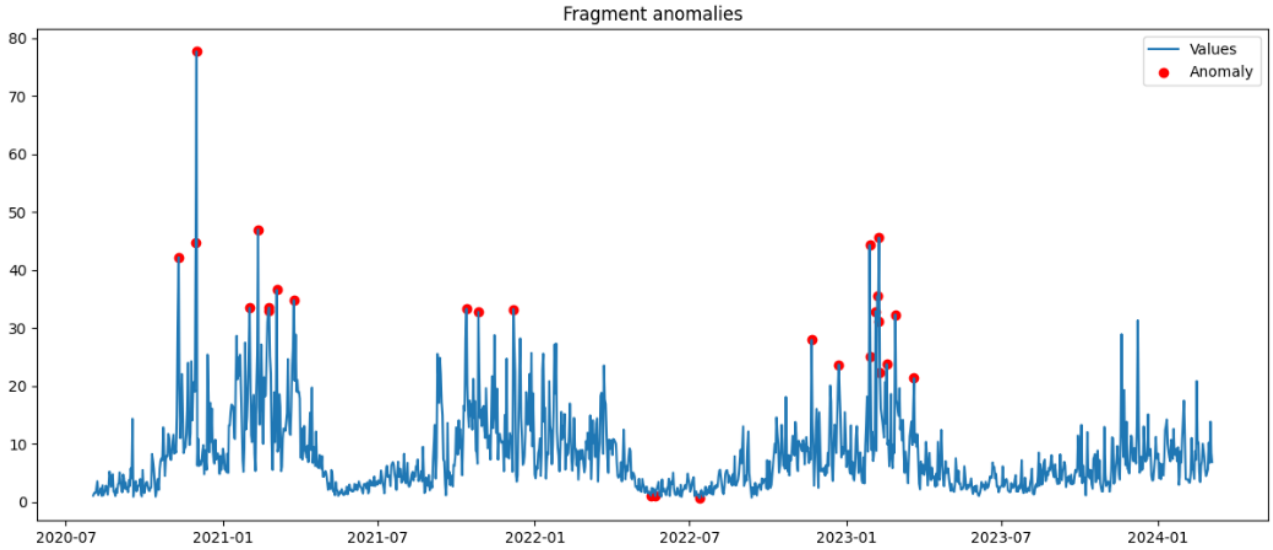

Рис. 5. Ідентифіковані аномалії часового ряду показника пилу PM2.5

Порівняння точності прогнозування

Для порівняльного аналізу ефективності розроблених методів для прогнозування часового ряду, пропонується побудова моделі Prophet та її застосування для часового ряду показника пилу PM2.5 з декількома варіантами конфігурації:

– Модель без використання вбудованих параметрів сезонності (англ. «Without default seasonality»);

– Модель з використанням вбудованих параметрів сезонності (англ. «Default seasonality»);

– Модель з використанням власних налаштувань сезонності за рахунок використання розробленого методу ідентифікації параметрів сезонності (англ. «Custom seasonality»);

– Модель з використанням вбудованих параметрів сезонності та аномаліями, знайденими звичайним методом (англ. «Series anomalies + default seasonality»);
– Модель з використанням вбудованих параметрів сезонності та аномалій, знайдених методом пошуку локальних аномалій (англ. «Fragment anomalies + default seasonality»);
– Модель з використанням власних налаштувань параметрів сезонності та аномаліями, знайденими звичайним методом (англ. «Series anomalies + custom seasonality»);
– Модель з використанням власних налаштувань параметрів сезонності та аномалій, знайдених методом пошуку локальних аномалій (англ. «Fragment anomalies + default seasonality»).
Для кожного варіанту конфігурації створюється окрема модель Prophet та тренується на даних часового ряду показника пилу PM2.5. Після цього кожною з моделей робиться прогноз на 7 наступних днів та порівнюються отримані значення прогнозу з реальними показниками часового ряду, використовуючи метрику MAE та RMSE [10]. Таблицю з результатами застосування цих метрик наведено на рис. 6.

	name	mae	rmse	mape
0	Without default seasonality	7.300028	8.724954	0.498087
1	Default seasonality	7.331930	8.725626	0.506034
2	Custom seasonality	5.444506	7.056464	0.452396
3	Series anomalies + default seasonality	7.293572	8.687440	0.504247
4	Fragment anomalies + default seasonality	7.128705	8.491353	0.495288
5	Series anomalies + custom seasonality	5.419272	7.038344	0.451271
6	Fragment anomalies + custom seasonality	5.080210	6.829684	0.440533

Рис. 6. Результати порівняльного аналізу точності прогнозування з використанням розроблених методів

Як можна побачити на рисунку, використання власних параметрів сезонності дає меншу помилку прогнозування у порівнянні з вбудованими параметрами сезонності моделі Prophet. В той же час, використання пошуку локальних аномалій також дозволяє зменшити помилку прогнозування у порівнянні з звичайним підходом до пошуку аномалій, коли метод застосовується для усього ряду. Найкращий ж результат показав варіант з комбінуванням локальних аномалій та використання власних параметрів сезонності. У порівнянні з точністю роботи моделі Prophet з параметрами за замовчуванням, вдалось зменшити помилки прогнозування за метрикою MAE на 30%, а за метрикою RMSE – на 21%.

. . . . . .

Висновки

У роботі наведено короткий опис двох раніше розроблених методів, які базуються на пошуку аномалій: пошук аномалій в декомпозиційній кривій та пошук аномалій в фрагментах даних часового ряду. Проведений порівняльний аналіз продемонстрував, що метод ідентифікації параметрів сезонності та метод пошуку локальних у часі аномалій часового ряду дозволяють покращити точність прогнозування періодичного часового ряду показника стану атмосферного повітря. Тестування точності прогнозу з та без використання згаданих методів проводилось на реальних даних моніторингу якості атмосферного повітря від мережі громадського моніторингу EcoCity. В дослідженні використовувались дані показника пилу PM2.5 за 2021-2024 рр, отриманих з станції, розташованої у смт Турбів Вінницькій області. Одночасне застосування обох методів дозволило зменшити похибку передбачення у порівнянні з роботою моделі Prophet з параметрами за замовчування, а саме за метрикою MAE на 30%, а за метрикою RMSE – на 21%. Це доводить, що ці методи дійсно є ефективними.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] Shumway, Robert & Stoffer, David. (2011). Time Series Analysis and Its Applications With R Examples. 10.1007/978-1-4419-7865-3.
[2] Terence C. Mills, Chapter 3 - ARMA Models for Stationary Time Series, Editor(s): Terence C. Mills, Applied Time Series Analysis, Academic Press, 2019, Pages 31-56, ISBN 9780128131176, https://doi.org/10.1016/B978-0-12-813117-6.00003-X.
[3] Dr. Shikha Gaur. "Global forecasting of COVID-19 USING ARIMA BASED FB-PROPHET". *International Journal of Engineering Applied Sciences and Technology*, 2020 Vol. 5, Issue 2, ISSN No. 2455-2143, Pages 463-467.
[4] Sean J Taylor, Benjamin Letham, Forecasting at scale, PeerJ Preprints, 5, 2017, https://doi.org/10.7287/peerj.preprints.3190v2
[5] Д. О. Шмундяк і В. Б. Мокін, «Метод ідентифікації параметрів гармонік та аномалій періодичного часового ряду на основі адаптивної декомпозиції», Вісник ВПІ, вип. 6, с. 46–56, Груд. 2023.
[6] Dmytro Shmundiak and Vitalii Mokin, "Adaptive decomposition for harmonics and anomalies" Kaggle Notebook. [Electronic resource]. Available: https://www.kaggle.com/code/dimashmundiak/adaptive-decomposition-for-harmonics-and-anomalies. Accessed:05.10.2024.
[7] Д. О. Шмундяк і В. Є. Копняк, «Метод ідентифікації локальних аномалій значень показників стану довкілля з використанням декомпозиції на півхвилі», *Вісник ВПІ*, вип. 1, с. 88–100, Лют. 2024.
[8] Omar, Salima & Ngadi, Md & Jebur, Hamid & Benqdara, Salima. (2013). Machine Learning Techniques for Anomaly Detection: An Overview. International Journal of Computer Applications. 79. 10.5120/13715-1478.
[9] Markus M. Breunig, Hans-Peter Kriegel, Raymond T. Ng, and Jörg Sander. 2000. LOF: identifying density-based local outliers. SIGMOD Rec. 29, 2 (June 2000), 93–104. https://doi.org/10.1145/335191.335388
[10] Sklearn. API Reference – Available: https://scikit-learn.org/stable/modules/classes.html. Accessed: 05.10.2024

Шмундяк Дмитро Олександрович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: dimashmund@gmail.com

Вінницький національний технічний університет, Україна

D. O. Shmundiak

Comparative Analysis of Forecasting Models of the Air Condition Indicator

Vinnytsia National Technical University, Ukraine

Periodic time series have many applications in our lives. Examples of periodic time series are air quality indicators, financial market indicators, meteorological parameters, etc. Because of this, the analysis and forecasting of periodic time series is a widespread and interesting scientific topic. One of the main problems in the analysis of periodic time series is the determination of the parameters of the seasonality of this series and the identification and elimination of abnormal values that can significantly affect the accuracy of data forecasting. In this work, a comparative analysis of the previously developed models and approaches for air quality indicator forecasting is given. The research is based on real data from the EcoCity public air quality monitoring network. A brief description of the method of identifying the seasonality parameters of the time series based on the decomposition of this time series and an approach to finding local anomalies in a time series based on the results of series decomposition is given. The results of the described models were used to forecast the PM2.5 dust index of one of the air quality monitoring stations in the Vinnytsia region. The Python programming language was used to automate the forecasting process, and the program code itself was implemented in the Kaggle system, a web platform from Google for machine learning engineers. The Prophet time series model was used for forecasting. A comparison table of the forecast accuracy of the Prophet model with default settings and with custom configuration based on the data from developed models and approaches was provided. The study and analysis showed that using both developed methods helps to reduce the forecasting error for the air quality indicator. Compared to the accuracy of the Prophet model with the default parameters, it was possible to reduce the MAE error value by 30% and the RMSE by 21%. This proves that these methods are effective for the analysis and forecasting of time series, including time series of air quality indicators.

Keywords: time series data analysis, anomalies of time series data, machine learning, seasonal decomposition, Fourier series harmonics, Prophet, air quality, EcoCity.

Shmundiak Dmytro Oleksandrovych – Post-graduate Student of the Chair of System Analysis and Information Technologies, e-mail: dimashmund@gmail.com**.**