

УДК 004.934

DOI: 10.32626/2308-5916.2025-28.28-36

Д. О. Зуй,**І. М. Кузьменко**, канд. техн. наук,**Т. В. Гевліч**

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського», м. Київ

ЗАСТОСУВАННЯ МОДЕЛЕЙ ЧАСОВИХ РЯДІВ ДЛЯ ПРОГНОЗУВАННЯ ПОГОДНИХ УМОВ

У роботі наведено оцінку моделей часових рядів ARIMA, SARIMA, SARIMAX та можливості їх застосування для прогнозування погодних умов, що є актуальним в умовах сучасних кліматичних змін.

Традиційні фізико-математичні моделі забезпечують високу точність, але вимагають значних обчислювальних ресурсів та не завжди оперативно реагують на змінні зовнішні фактори. Тому виникає необхідність у використанні інших методів прогнозування, які поєднують адаптивність, швидкодію та точність.

Наукова новизна роботи полягає у виконаному вперше порівняльному аналізі ефективності моделей ARIMA, SARIMA та SARIMAX для прогнозування погодних умов на основі відкритих наборів даних платформи Kaggle із використанням метрик оцінки точності RMSE та MAE.

Тому метою роботи є порівняння RMSE метрики моделей ARIMA, SARIMA, SARIMAX та апаратних ресурсів, потрібних для їх реалізації. Архітектура системи включає компоненти збору, очищення, опрацювання та візуалізації даних і реалізована на мові програмування Python з використанням бібліотек: Pandas, NumPy, Statsmodels, Pmdarima, Matplotlib, Seaborn, Plotly, Streamlit.

Для реалізації моделей опрацьовано набори відкритих даних з платформи Kaggle. Перед моделюванням проведено перевірку стаціонарності рядів даних за тестом Дікі-Фуллера за 10% рівня значущості. Для налаштування моделей застосовано середню абсолютну похибку (MAE) та для їх оцінки – корінь середньоквадратичної похибки (RMSE).

Результати показали, що модель ARIMA є ефективною для прогнозування нециклічних погодних даних, забезпечуючи RMSE близько 0,2...0,4, проте вимагає періоду тестування у понад 50 ітерацій. Модель SARIMA рекомендується при роботі з сезонними даними, забезпечуючи RMSE близько 0,8 після проходження близько 80 ітерацій. Модель SARIMAX рекомендується при роботі з сезонними даними, забезпечуючи RMSE близько 0,75 за 80 ітерацій за використання додаткових екзогенних змінних.

Представлені результати свідчать про перспективність моделей ARIMA, SARIMA та SARIMAX для прогнозування погодних умов.

Ключові слова: *прогнозування погоди, моделі часових рядів, ARIMA, SARIMA, SARIMAX.*

Вступ. Сучасне прогнозування погоди є важливим аспектом, оскільки несприятливі погодні умови можуть призвести до значних економічних збитків і загроз для здоров'я населення. Наприклад, екстремальні погодні явища, такі як сильні шторми, тривалі посухи та ін., ускладнюють традиційне прогнозування і потребують нових підходів. Існуючі фізичні моделі, засновані на математичних рівняннях і базових фізичних принципах, забезпечують певну точність, але мають обмеження в адаптації до нових даних і швидкому врахуванні складних багатофакторних змін [1, 3]. Методи часових рядів використовуються у прогнозуванні погоди, оскільки вони здатні швидко обробляти великі дані. Прогнозування погоди із застосуванням алгоритмів часових рядів та машинного навчання використовує історичні дані та зовнішні змінні в режимі реального часу, що допомагає створювати прогнози, здатні адаптуватися [4].

Складні погодні явища та їхні наслідки змушують фахівців розробляти інноваційні системи для прогнозування, що базуються на статистичних моделях аналізу часових рядів, таких як ARIMA (AutoRegressive Integrated Moving Average), SARIMA (Seasonal ARIMA) та SARIMAX (Seasonal ARIMA з додатковими змінними). Ці моделі дозволяють аналізувати погодні дані у форматі часових рядів, враховуючи як сезонність, так і екзогенні фактори, що позитивно впливає на загальну точність прогнозів [5].

Підбір параметрів у прогнозуванні часових рядів активно використовується у численних дослідженнях. Зокрема, ARIMA є однією з основних моделей для короткострокового прогнозування, але її можливості розширюються за рахунок моделей SARIMA та SARIMAX, які краще враховують сезонні коливання. Наприклад, SARIMA підвищує точність результатів для погодних даних, що мають виражену сезонність. Ця модель є особливо цінною для регіонів з яскраво вираженими кліматичними змінами між сезонами, оскільки вона дозволяє прогнозувати температуру, вологість та інші параметри з урахуванням сезонних циклів [2, 5]. Модель SARIMAX, додатково використовуючи екзогенні змінні, дозволяє інтегрувати зовнішні фактори, що робить прогнозування більш гнучким і точним для різних часових горизонтів [4].

Метою роботи є порівняння метрик моделей часових рядів та апаратних ресурсів, потрібних для їх реалізації на основі прогнозування погодних даних. Для досягнення мети виконуються наступні задачі: з

аналізу літератури встановити моделі аналізу числових рядів, розробити архітектуру та реалізувати систему тестування моделей, провести тестування і верифікацію моделей часових рядів з використанням RMSE метрики, встановити переваги та недоліки розглянутих моделей.

Традиційні підходи значною мірою базуються на чисельних моделях прогнозування погоди, які використовують рівняння математичної фізики для моделювання атмосферних процесів. Однак, ці моделі потребують високої обчислювальної потужності. З іншого боку, моделі часових рядів – це можливість аналізувати складні взаємозв'язки у даних та надавати точні прогнози з меншими обчислювальними витратами [6].

Існує багато платформ та інструментів, що використовують для прогнозування погоди. Наприклад, Google Weather API та OpenWeatherMap використовують наземні метеостанції, супутники для прогнозування погоди. Однак ці платформи потребують додаткової обробки даних. Платформа MeteoBlue використовує математичні моделі, що дозволяє робити точніші прогнози. Проте, вартість отриманих нею даних зростає [7].

Також існує ряд аналогів ARIMA-моделей, наприклад LSTM, GRU, що використовують нейронні архітектури для обробки часових рядів. Однак вони потребують більших апаратних ресурсів.

Наукова новизна роботи полягає у виконаному вперше порівняльному аналізі ефективності моделей ARIMA, SARIMA та SARIMAX для прогнозування погодних умов на основі відкритих наборів даних платформи *Kaggle* із використанням метрики оцінки точності RMSE. Для цього встановлено кількісні закономірності між кількістю ітерацій для налаштування моделей з використанням MAE та точністю прогнозу.

Архітектура системи. Архітектура системи прогнозування погодних умов побудована з використанням модульного підходу, що забезпечує гнучкість та масштабованість. Основні компоненти системи включають модулі, показані на рисунку 1.

Data Handler – модуль збору та обробки даних, що відповідає за збір даних з різних джерел. Дані збираються в реальному часі та проходять попередню обробку з використанням бібліотек Pandas та NumPy для видалення шумів, пропусків і аномалій. Для інтеграції даних використовується бібліотека Pandas, яка дозволяє працювати з великими обсягами інформації у вигляді таблиць. Також, для автоматичного оновлення даних виконуються запити до серверів погоди. Це дозволяє підтримувати актуальність інформації та забезпечувати безперервне тестування моделей і оновлення прогнозів.

Model Selector та Forecaster – модулі моделювання та прогнозування. Ці модулі є ядром системи, і здійснюють прогноз погоди на основі моделей ARIMA, SARIMA та SARIMAX, що навчаються на підготовлених даних з використанням бібліотек Scikit-learn, Statsmodels та Pmdarima. Тестування моделей з використанням бібліотек Scikit-learn, Statsmodels та Pmdarima дозволяє налаштувати параметри моделей.

Модуль візуалізації (Visualizer) з використанням бібліотек Matplotlib, Seaborn та Plotly створює статичні та інтерактивні графіки, що ілюструють результати моделювання. Для зберігання даних та результатів прогнозів використовується реляційна база даних PostgreSQL.

Інтерфейс користувача в системі реалізований за допомогою бібліотеки Streamlit, що дозволяє створювати зручний веб-додаток для взаємодії з системою.

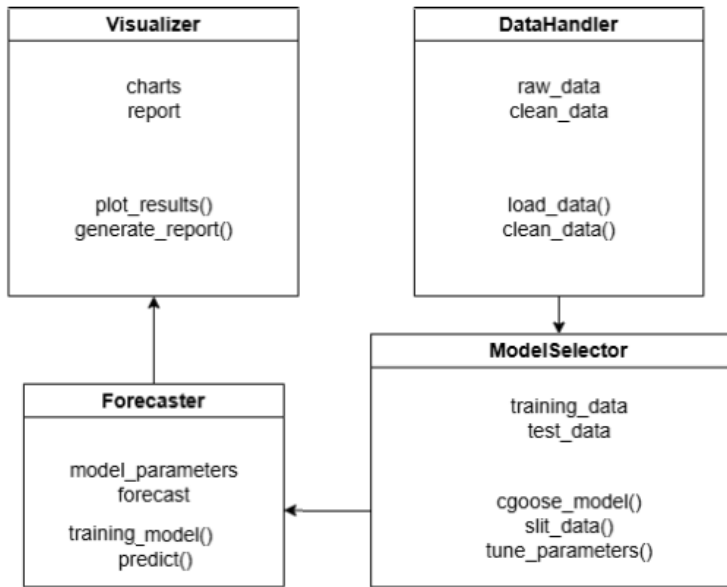


Рис. 1. Архітектура системи

Опис роботи моделей. Робота моделей вивчалася на наступному апаратному забезпеченні: 4 ядерний процесор з тактовою частотою 2,5 ГГц та 16 ГБ оперативної пам'яті.

Збір даних проведено з платформи Kaggle, зокрема набори Weather Events і Weather Forecasting використано для тренування і тестування моделей часових рядів. Ці набори даних містять інформацію про температуру, тиск, вологість, швидкість вітру, що є важливими для створення надійної моделі.

Етап збору даних включає тест Дікі-Фуллера для перевірки даних за рівнем значущості та відкидання нуль-гіпотези. За необхідності, дані відбиралися для проходження тесту, щоб забезпечити якість даних та надійність роботи моделей прогнозування.

Надалі виконувалось ітерація та оцінка моделей [7], моделі тестувалися та налаштовувалися для досягнення необхідної точності за допомогою наступних метрик – корінь середньоквадратичної похибки (*RMSE*) та середня абсолютна похибка (*MAE*), формули (1-2). Перша метрика (1) дозволяє кількісно оцінити результати прогнозування погодних умов. *MAE* (2) використана для налаштування власне роботи моделей – для порівняння результатів на кожній ітерації

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}, \quad (1)$$

де y_i – експериментальні дані з датасету, $\hat{y}_i$ – модельні передбачені значення, n – кількість спостережень.

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (2)$$

де y_i – модельні значення на попередній ітерації, $\hat{y}_i$ – модельні значення, n – загальна кількість ітерацій.

Набори відкритих даних з платформи *Kaggle* містили температуру, вологість, швидкість вітру, атмосферний тиск. В даній роботі для порівняння моделей розглядалися значення температури та тиску.

Надалі виконувалися ітерації параметрів моделей. В результаті ітерацій отримуються параметри моделей для реалізації системи прогнозування погоди, оскільки саме це визначає *RMSE* та *MAE* моделі.

Результати роботи моделей. Процес моделювання включає налаштування моделей на історичних погодних даних (набори відкритих даних з платформи *Kaggle*, зокрема набори *Weather Events* і *Weather Forecasting*) та їх тестування за *RMSE* для оцінки точності прогнозування. На рисунку 2 показано зміну *RMSE*, залежно від кількості ітерацій для підбору параметрів моделі, позначених *E*. З рисунка видно, що за перевищення близько 50 ітерацій відбувається стабілізація роботи моделі та зменшення *RMSE*. Швидкість налаштування моделі склала 7 хв за 43% завантаження апаратних ресурсів. Коефіцієнт детермінації для моделі 92%.

На рисунку 3 показано зміну *RMSE* залежно від кількості ітерацій *E*, під час налаштування моделі *SARIMA*, що демонструє ефективність моделювання сезонних компонентів на основі цієї моделі.

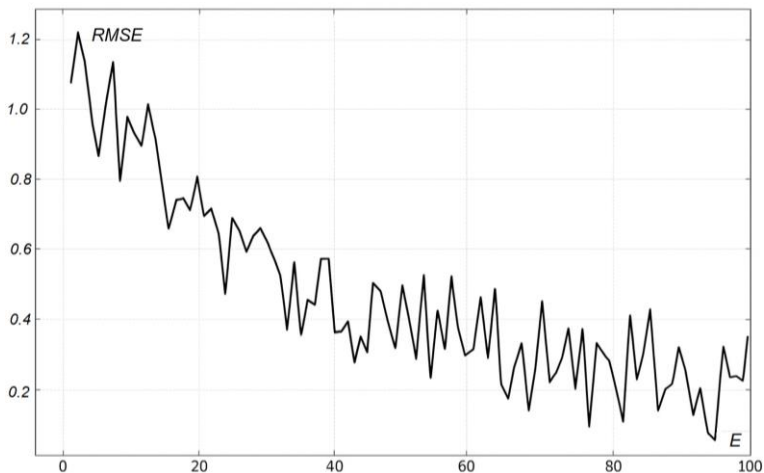


Рис. 2. Зменшення RMSE моделі ARIMA зі зростанням ітерацій

Як видно з рисунку 3, налаштування моделі SARIMA на 80 ітераціях забезпечує значення метрики $RMSE = 0.75...0.8$, що перевищує значення для моделі ARIMA. Швидкість налаштування моделі склала 12 хв за 63% завантаження апаратних ресурсів. Коефіцієнт детермінації для моделі 95%. На рисунку 3 також нанесено значення валідаційної точності, що перевірялася на 20% даних із датасету, які не ввійшли в налаштувальні. Як видно, в цьому випадку RMSE моделі зростає до 1.04 для 20 ітерацій.

На рисунку 4 показано зміну RMSE під час налаштування моделі SARIMAX, що підтверджує вплив екзогенних змінних на покращення точності.

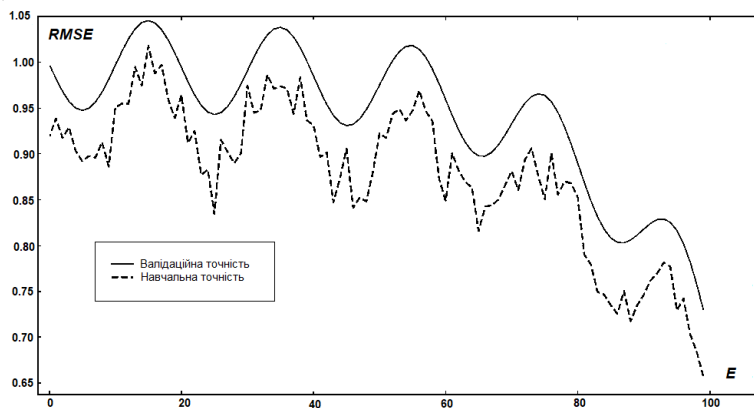


Рис. 3. Зменшення RMSE моделі SARIMA зі зростанням ітерацій

Для налаштування моделі SARIMAX використано атмосферний тиск, на відміну від попередніх рисунків 2 і 3, де для моделювання було обрано температуру повітря.

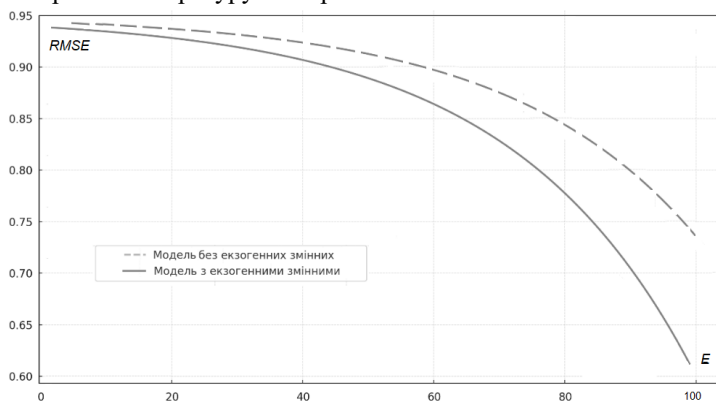


Рис. 4. Зменшення RMSE для моделі SARIMAX із врахуванням екзогенних змінних

Як видно з рисунку 4, налаштування моделі SARIMAX на 100 ітераціях із врахуванням впливу екзогенних змінних забезпечує $RMSE = 0.62...0.73$, що менше, ніж для моделі SARIMAX. Коефіцієнт детермінації для моделі SARIMAX склав 98%.

Швидкість налаштування моделі склала 22 хв за 80% завантаження апаратних ресурсів комп'ютера.

Висновки. В роботі вперше здійснено порівняльний аналіз ефективності моделей ARIMA, SARIMA та SARIMAX для прогнозування погодних умов на основі відкритих наборів даних платформи *Kaggle* із використанням метрик оцінки точності RMSE та MAE у єдиному програмному середовищі на мові *Python*.

Архітектура розробленої системи налаштування моделей ARIMA, SARIMA та SARIMAX містить модуль збору та обробки даних, модулі моделювання та прогнозування, модуль візуалізації, реалізовані з використанням сучасних бібліотек мови *Python*.

Моделювання погодних умов проведено на основі створеного застосунку, що використовує набори відкритих даних з платформи *Kaggle*, зокрема набори Weather Events і Weather Forecasting. Набори використано для оптимізації параметрів і верифікації моделей часових рядів.

Для оцінки ефективності моделей часових рядів ARIMA, SARIMA та SARIMAX використано корінь середньоквадратичної похибки (RMSE) для порівняння результатів моделювання з експериментальними даними. Також для ітераційного налаштування моделей обробовувалася метрика MAE.

Встановлено, що модель ARIMA потребує для налаштування більше 50 ітерацій на відміну від моделей SARIMA та SARIMAX, які рекомендується оптимізувати на 100 ітераціях. Також модель ARIMA має RMSE близько 0.2...0.4 за умови роботи з не циклічними даними на відміну від моделей SARIMA та SARIMAX, що мають RMSE близько 0.62...0.75, проте працюють з циклічними даними.

Отримано нові дані щодо взаємозв'язку точності прогнозу та апаратних ресурсів, необхідних для реалізації моделей ARIMA, SARIMA та SARIMAX, що дає змогу обґрунтовано обирати оптимальну модель залежно від обчислювальних можливостей та структури вхідних даних.

Список використаних джерел:

1. Зуй Д. О., Кузьменко І. М. Прогнозування погоди на основі методів машинного навчання та аналізу даних. *Збірник праць II Міжнародної науково-практичної конференції «Global Directions in Scientific Research and Technological Development»*. 2024. Р. 103-108.
2. Weather Events (2016-2022). *Kaggle*. URL: <https://www.kaggle.com/datasets/sobhanmoosavi/us-weather-events>.
3. Kontopoulou V. I., Panagopoulos A. D., Kakkos I., Matsopoulos G. K. A Review of ARIMA vs. Machine Learning Approaches for Time Series Forecasting in Data Driven Networks. *MDPI*. 2023.
4. Weather Forecasting. *Kaggle*. URL: <https://www.kaggle.com/datasets/riteshpatil8998/weather-forecasting>
5. Weather Forecasting Using Deep Learning And Seasonal Autoregressive Integrated Moving Average (SARIMA) Models. *IOSR Journals*. 2024.
6. De Gooijer J. G., Hyndman R. J. 25 Years of Time Series Forecasting. *International Journal of Forecasting*. 2016. Vol. 22. № 3. Р. 443-473.
7. Goodfellow I., Bengio Y., Courville A. Deep Learning. MIT Press, 2016. 775 p.

APPLICATION OF ARIMA, SARIMA, AND SARIMAX MODELS FOR WEATHER FORECASTING

The study presents an evaluation of time series models ARIMA, SARIMA, and SARIMAX, and assesses their applicability for weather forecasting, which is highly relevant under current climate change conditions.

The study presents an evaluation of the time series models ARIMA, SARIMA, and SARIMAX and their applicability for weather forecasting, which is particularly relevant under current climate change conditions.

Traditional physical and mathematical models provide high forecasting accuracy but require substantial computational resources and may not promptly adapt to rapidly changing external factors. Therefore, there is a growing need to employ alternative forecasting methods that combine adaptability, computational efficiency, and accuracy.

The scientific novelty of this research lies in the first-time comparative analysis of the performance and efficiency of ARIMA, SARIMA, and SARIMAX models for weather prediction based on open datasets from the

Kaggle platform, using Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) as evaluation metrics.

The purpose of the study is to compare the RMSE metrics of the ARIMA, SARIMA, and SARIMAX models and the computational resources required for their implementation. The system architecture includes modules for data collection, cleaning, processing, and visualization, implemented in the Python programming language using the following libraries: Pandas, NumPy, Statsmodels, Pmdarima, Matplotlib, Seaborn, Plotly, and Streamlit.

For model implementation, open-source datasets from the Kaggle platform were processed. Before modeling, the stationarity of the time series was verified using the Augmented Dickey–Fuller (ADF) test at a 10% significance level. The MAE metric was applied to tune the models, while RMSE was used for model evaluation.

The results show that the ARIMA model is effective for forecasting non-seasonal weather data, providing an RMSE of approximately 0.2-0.4, although it requires a testing phase of more than 50 iterations. The SARIMA model is recommended for seasonal datasets, achieving an RMSE of around 0.8 after approximately 80 iterations. The SARIMAX model, which incorporates exogenous variables, is also recommended for seasonal data, achieving an RMSE of about 0.75 over 80 iterations.

The presented findings demonstrate the potential and applicability of ARIMA, SARIMA, and SARIMAX models for weather forecasting tasks.

Key words: *weather forecasting, time series models, ARIMA, SARIMA, SARIMAX.*

Отримано: 11.07.2025