УДК 004.056.57

Н.С. ХАЛІЗЕВ, В.І. ДУБРОВІН, Л.Ю. ДЕЙНЕГА
Національний університет «Запорізька політехніка»

ПРОГНОЗУВАННЯ ПОГОДИ ЗА ДОПОМОГОЮ ЧАСОВИХ РЯДІВ

У сучасному світі точний та надійний прогноз погоди має вирішальне значення для різних сфер життя людини, від сільського господарства до транспорту, енергетичного сектору та туризму. Точний прогноз погоди дає змогу фермерам планувати сівозміну та обробку поля, а також допомагає підприємствам з керуванням енергетичними ресурсами (планування опалювального сезону, коливань споживання енергоресурсів тощо). У цьому контексті аналіз часових рядів набуває особливого значення, оскільки він дає змогу вивчити та передбачити зміни основних погодних показників (температура, тривалість дня, вологість тощо).

Часовий ряд – це послідовність точок даних, які з'являються в певному порядку протягом певного періоду часу [1]. У контексті погодних досліджень часові ряди являють собою дані про різні погодні показники (температура, вологість, атмосферний тиск, напрямок та швидкість повітря тощо) за певний період часу. Аналіз цих рядів дає змогу виявити тренди, сезонні коливання та інші закономірності, які можуть бути використані для прогнозування майбутніх значень.

Для розв'язку задачі прогнозування погоди, яка має яскраво виражену сезонність, у роботі було використано модель сезонної авторегресійної інтегрованої ковзної середнього (SARIMA). SARIMA є універсальною та широко використовуваною моделлю прогнозування часових рядів. Це розширення несезонної моделі ARIMA, призначеної для обробки даних із сезонними коливаннями [2].

Модель SARIMA характеризується параметрами (p, d, q)(P, D, Q)m, де p, d, q відповідають за несезонну частину моделі (авторегресія, різниця та ковзне середнє тренду відповідно), а P, D, Q – за сезонну частину. Параметр m визначає періодичність, тобто сезонність (наприклад, 12 для щомісячних даних з річною сезонністю) [2].

У ході дослідження було реалізовано отримання та обробку даних, перевірку даних на ознаки стаціонарності, пошук оптимальних параметрів моделі SARIMA та оцінювання точності результатів прогнозування.

Ключові слова: температура повітря, прогнозування часових рядів, SARIMA, сезонність, стаціонарність, аналіз часових рядів, пошук за сіткою, оцінка похибки.

N.S. KHALIZEV, V.I. DUBROVIN, L.YU. DEYNEGA
Zaporizhzhia Polytechnic National University

WEATHER FORECASTING USING TIME SERIES

In today's world, accurate and reliable weather forecasts are crucial for various areas of human life, from agriculture to transportation, the energy sector, and tourism. Accurate weather forecasts allow farmers to plan crop rotation and field cultivation, and help businesses manage energy resources (planning the heating season, fluctuations in energy consumption, etc.). In this context, time series analysis is of particular importance, as it allows you to study and predict changes in key weather indicators (temperature, day length, humidity, etc.).

A time series is a sequence of data points that appear in a certain order over a certain period of time [1]. In the context of weather research, time series are data on various weather indicators (temperature, humidity, atmospheric pressure, air direction and speed, etc.) over a certain period of time. Analyzing these series allows you to identify trends, seasonal fluctuations, and other patterns that can be used to predict future values.

To solve the problem of weather forecasting, which has a pronounced seasonality, this paper uses the Seasonal Autoregressive Integrated Moving Average (SARIMA) model. SARIMA is a versatile and widely used time series forecasting model. It is an extension of the non-seasonal ARIMA model designed to process data with seasonal fluctuations [2].

The SARIMA model is characterized by the parameters (p, d, q)(P, D, Q)m, where p, d, q are responsible for the non-seasonal part of the model (autoregression, difference and trend moving average, respectively), and P, D, Q are responsible for the seasonal part. The parameter m determines the periodicity, i.e. seasonality (for example, 12 for monthly data with annual seasonality) [2].

The study involved data acquisition and processing, checking the data for signs of stationarity, finding the optimal parameters of the SARIMA model, and evaluating the accuracy of the forecasting results.

Key words: air temperature, time series forecasting, SARIMA, seasonality, stationarity, time series analysis, grid search, error estimation.

https://doi.org/10.32782/mathematical-modelling/2024-7-2-22

Постановка проблеми

Атмосферні процеси вивчає наука метеорологія. Одним з основних її завдань є прогнозування погоди. Воно складається з трьох етапів: спостереження за погодою і збір інформації, обробка результатів спостережень, складання прогнозу.

Дані, зібрані метеорологами, передаються до центрів Всесвітньої служби погоди, де обробляються і аналізуються синоптиками [3].

Традиційні методи прогнозування часто потребують дуже великих обчислювальних потужностей. Наприклад, Британське Метеорологічне бюро (Met Office), придбало суперкомп'ютер вартістю 1,2 млрд фунтів стерлінгів [4]. Це призводить до великих витрат на обладнання, але не гарантує точного прогнозування погоди. Отже, існує необхідність розроблення та вдосконалення методів прогнозування погоди, які б мали враховувати складні та циклічні закономірності погодних явищ.

У цьому контексті використання методів аналізу часових рядів для прогнозування погоди набуває особливого значення. Часові ряди показують динаміку змін погодних показників з часом, і їх аналіз дає змогу виявити закономірності та тенденції, які можуть бути використані для прогнозування майбутніх погодних умов. Це допоможе знизити кількість обчислювальних ресурсів, потрібних для прогнозування погодних показників у майбутньому.

Аналіз останніх досліджень та публікацій

У статті [5] розглядається метод прогнозування температури повітря та вологості, базований на алгоритмі градієнтного бустингу. Дослідження включало розроблення та оцінювання моделі прогнозування, заснованої на бібліотеці XGBoost. Основною метою було передбачення погодних показників задля ефективного керування системою опалення, вентиляції та кондиціювання повітря (HVAC), спрямованого на зменшення споживання енергії.

У статті [6] описується попередня обробка даних, налаштування моделі LSTM (Довга короткочасна модель) та етап оцінки результатів прогнозування. В результаті прогнозування погодних параметрів, як-от температура, вологість, тривалість сонячного світла (години) і швидкості вітру, автор дійшов висновку, що модель LSTM підходить для прогнозування температури, але менш придатна для прогнозування даних про вологість повітря.

У статті [7] автор пропонує гібридну модель CEED-LSTM-ARIMA для прогнозування нестаціонарних часових рядів та порівнює її з моделями ARIMA та LSTM. Автор описує побудову гібридної моделі та порівнює точність прогнозування на основі даних часових рядів. Порівняно з ARIMA та LSTM, гібридна модель CEED-LSTM-ARIMA має вищу точність прогнозування.

Мета дослідження

Метою статті є дослідження процесу прогнозування погоди за допомогою даних часових рядів та використання моделі SARIMA (сезонної авторегресійної інтегрованої ковзної середньої). Дослідження прагне оцінити ефективність цього методу у вирішенні проблеми прогнозування сезонних та циклічних змін у погодних умовах.

Виклад основного матеріалу дослідження

Джерела даних та їх обробка

Дані отримані з використанням бібліотеки Meteostat Python. Серед джерел даних є національні метеорологічні служби, як-от Національне управління океанічних і атмосферних досліджень (NOAA) і Національна метеорологічна служба Німеччини (DWD) [8].

Дані містять записи про погодні умови, як-от середня, мінімальна та максимальна температура повітря, точка роси, загальна кількість опадів, напрямок та швидкість вітру.

https://doi.org/10.32782/mathematical-modelling/2024-7-2-22

Обробку даних у цьому дослідженні було здійснено методом вибірки даних, метою якого є обрання підмножини даних для навчання моделі. Вибірка даних дає нам змогу відкинути неправильні, нерелевантні дані з набору даних.

SARIMA-моделювання часових рядів

У цьому розділі описано запропоновану модель SARIMA та представлено процес вибору моделі. Першим кроком є отримання та аналіз даних часових рядів. Наступними кроками є пошук оптимальних параметрів моделі та прогнозування й оцінка моделі. Наступні розділи детально описують ці кроки.

Отримання та аналіз даних

Для отримання даних було використано дані проєкту Meteostats, який збирає та обробляє дані з національних метеорологічних служб Німеччини та США (DWD, NOAA).

Для отримання даних нам необхідно зазначити період, за який необхідні дані, та координати метеостанції (їх можна знайти на сайті проєкту Meteostat). У цьому разі було вибрано метеостанції біля Одеси та період з 1984 по 2022 рік.

Після отримання даних ми відокремлюємо дані про середню температуру повітря від інших та візуалізуємо отримані дані (рис. 1).

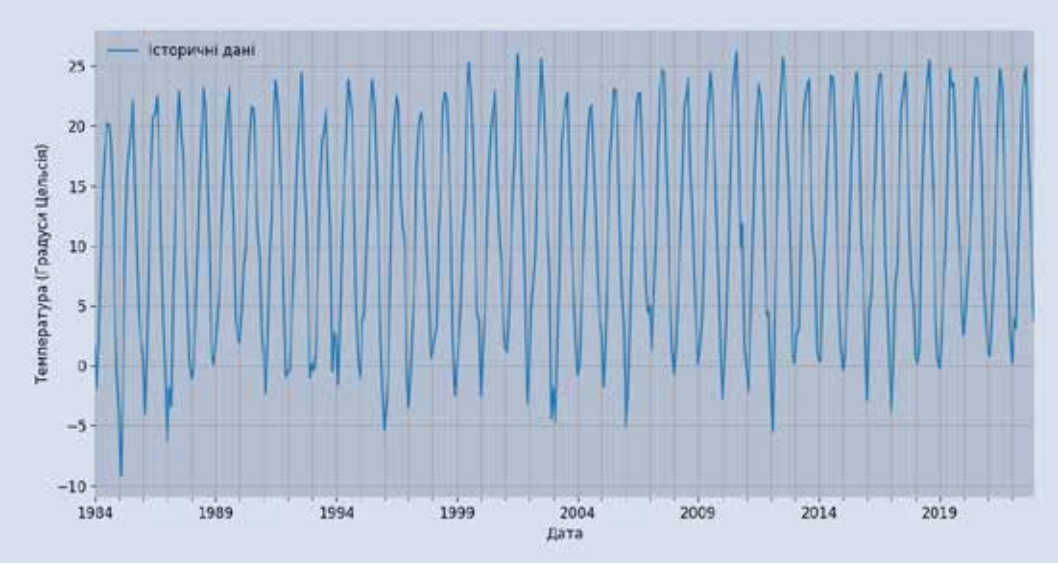

Рис. 1. Отримані дані

На рисунку можна чітко побачити сезонну закономірність даних, це очікувано, оскільки використовуються дані про температуру повітря.

Нам необхідно переконатися в сезонності цих даних, для цього можна використати функцію seasonal_decompose з пакета statsmodels. Припустимо, що дані мають сезонність у 12 місяців (рис. 2).

Результати підтверджують припущення про сезонність даних.

Для застосування моделі SARIMA важливо мати стаціонарний часовий ряд, тому для перевірки даних ми можемо використати розширений тест Дікі-Фулллера [9] (рис. 3).

Результат тесту нам говорить про те, що наш ряд є стаціонарним, тобто p-значення перебуває в межі (менше 0,05). Оскільки p-значення перебуває в межі, ми можемо відхилити нульову гіпотезу, тобто часовий ряд не має єдиного кореня, отже є стаціонарним. Для перевірки стаціонарності рекомендується використовувати декілька тестів, тому було проведено тест Квятковського-Філліпса-Шмідта-Шина [9] (рис. 4).

https://doi.org/10.32782/mathematical-modelling/2024-7-2-22

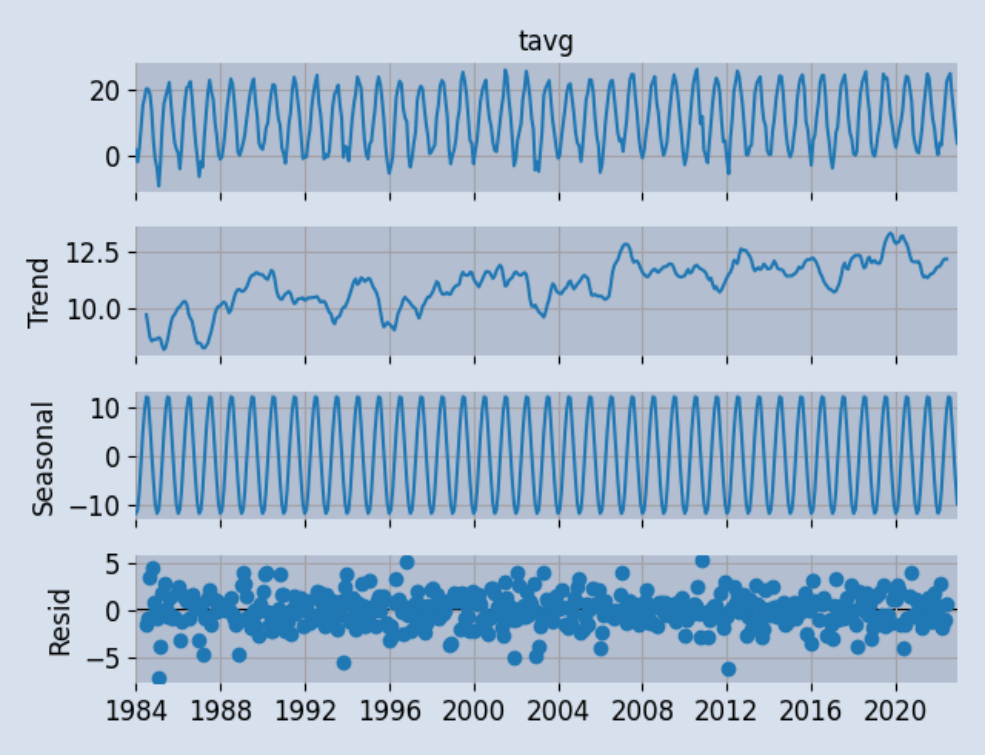

Рис. 2. Результат роботи функції

```
ADF Statistic: -3.536774
p-value: 0.007091
Critical Values:
        1%: -3.446
        5%: -2.868
        10%: -2.570
Stationary
```

Рис. 3. Результат розширеного тесту Дікі-Фуллера

```
KPSS Statistic: 0.450406
p-value: 0.055428
Critical Values:
        10%: 0.347
        5%: 0.463
        2.5%: 0.574
        1%: 0.739
Stationary
```

Рис. 4. Результат тесту Квятковського-Філліпса-Шмідта-Шина

Як можна побачити, за рівня значущості 0,05 ми не можемо відхилити нульову гіпотезу (p-значення більше 0,05), тобто ряд є стаціонарним за трендом та не потребує додаткового диференціювання (параметр d = 0).

Виходячи з результатів тестів, ми можемо стверджувати, що часовий ряд є стаціонарним [9].

https://doi.org/10.32782/mathematical-modelling/2024-7-2-22

Тому приступимо до ділення даних на тренувальну та тестову вибірки. Оскільки горизонт прогнозування був вибраний у 24 місяці, тобто 2 роки, ми відділяємо тестові дані, останні 24 місяці, від тренувальних (рис. 5).

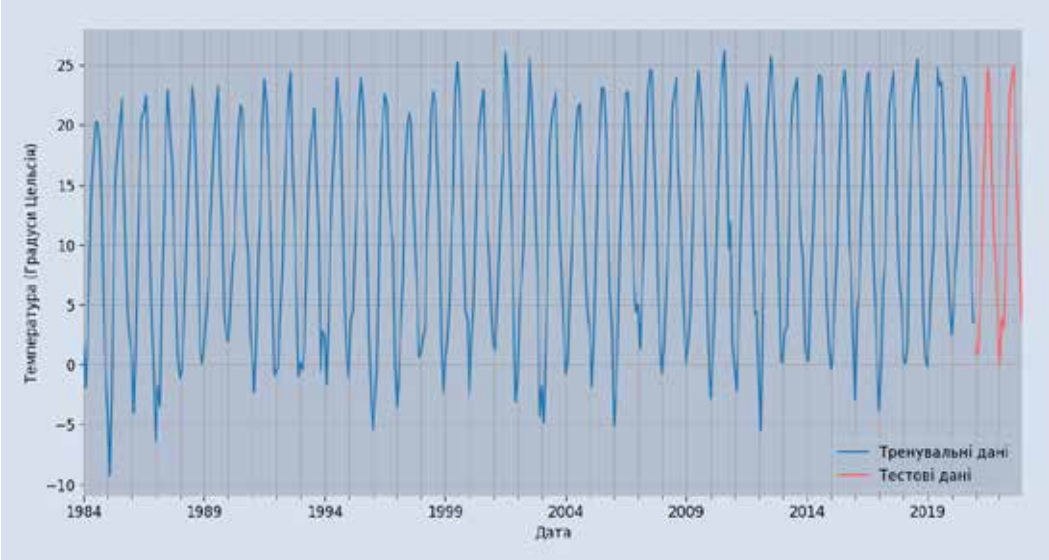

Рис. 5. Візуалізація тренувальних і тестових даних

На рис. 5 ми бачимо, що синім кольором виділені тренувальні дані, на яких модель буде навчатися, а пастельно-червоним – тестові.

Пошук оптимальних параметрів моделі

На цьому кроці вибираються параметри моделі SARIMA(p, d, q)(P, D, Q)$_m$. Модель потребує шести параметрів: p, d, q, P, D і Q. Значення m задано рівним 12, оскільки використовуються сезонні дані з періодом у 12 місяців. Для оцінки точності прогнозування моделі за заданими параметрами замість AIC (інформаційний критерій Акаіке) було вибрано значення середньої абсолютної похибки [10]. Пошук параметрів моделі виконується за допомогою пошуку за сіткою.

Згідно з результатами пошуку за сіткою, значення середньої абсолютної похибки (MAPE) у моделі з параметрами SARIMA (2, 0, 1) (3, 0, 3)$_{12}$ є найнижчими (рис. 6).

```
SARIMAX: (3, 0, 3) x (1, 0, 3, 12)  MAPE: 49.88210335816337
SARIMAX: (3, 0, 3) x (2, 0, 0, 12)  MAPE: 57.22690322657529
Failed: (3, 0, 3) x (2, 0, 1, 12)
LU decomposition error.
SARIMAX: (3, 0, 3) x (2, 0, 2, 12)  MAPE: 21.776133393725782
SARIMAX: (3, 0, 3) x (2, 0, 3, 12)  MAPE: 25.23461539041008
SARIMAX: (3, 0, 3) x (3, 0, 0, 12)  MAPE: 51.42796962521929
SARIMAX: (3, 0, 3) x (3, 0, 1, 12)  MAPE: 48.990490994897826
Failed: (3, 0, 3) x (3, 0, 2, 12)
LU decomposition error.
SARIMAX: (3, 0, 3) x (3, 0, 3, 12)  MAPE: 26.725329417708682
Best params: (2, 0, 1) x (3, 0, 3, 12) MAPE: 17.25177660286081
```

Рис. 6. Результати оцінки моделей

Тепер отримаємо результати діагностичного тесту моделі SARIMA (2, 0, 1) (3, 0, 3)12 (рис. 7).

https://doi.org/10.32782/mathematical-modelling/2024-7-2-22

SARIMAX Results

Dep. Variable:	tavg	No. Observations:	444
Model:	SARIMAX(2, 0, 1)x(3, 0, [1, 2, 3], 12)	Log Likelihood	-941.538
Date:	Tue, 21 Nov 2023	AIC	1903.076
Time:	12:36:12	BIC	1944.034
Sample:	01-01-1984 - 12-01-2020	HQIC	1919.228
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.2590	0.012	100.786	0.000	1.235	1.283
ar.L2	-0.2731	0.005	-50.916	0.000	-0.284	-0.263
ma.L1	-0.9361	0.021	-43.715	0.000	-0.978	-0.894
ar.S.L12	-0.7083	0.025	-28.387	0.000	-0.757	-0.659
ar.S.L24	0.7850	0.020	39.660	0.000	0.746	0.824
ar.S.L36	0.9229	0.029	31.671	0.000	0.866	0.980
ma.S.L12	0.7936	0.026	30.014	0.000	0.742	0.845
ma.S.L24	-0.6740	0.027	-25.135	0.000	-0.727	-0.621
ma.S.L36	-0.9235	0.024	-37.785	0.000	-0.971	-0.876
sigma2	3.5121	0.090	38.939	0.000	3.335	3.689

Ljung-Box (L1) (Q):	0.00	Jarque-Bera (JB):	13.63
Prob(Q):	0.96	Prob(JB):	0.00
Heteroskedasticity (H):	0.84	Skew:	-0.07
Prob(H) (two-sided):	0.29	Kurtosis:	3.85

Рис. 7. Результати діагностичного тесту моделі

Стовпчик “Coef” показує вагу (важливість) кожної ознаки. Всі значення P>|z| є нижчими за 0,05, що вказує на те, що ці коефіцієнти є статистично значущим [11].

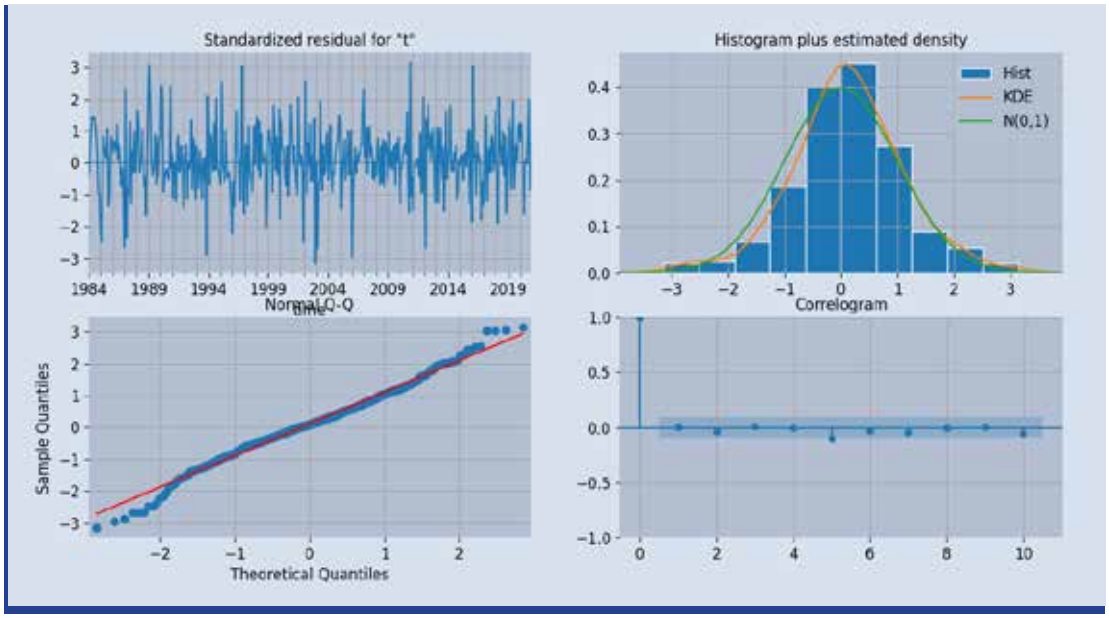

Рис. 8. Графіки залишків

Динаміка залишків у часі показана на рис. 8 (Standardized residual for “m”). Результати показують, що залишки не демонструють очевидної сезонності і є білим шумом.

На гістограмі (Histogram plus estimated density), ми бачимо, що помаранчева лінія KDE загалом слідує за зеленою N(0, 1), де N(0, 1) – позначення для нормального розподілу із

https://doi.org/10.32782/mathematical-modelling/2024-7-2-22

середнім 0 і стандартним відхиленням 1. Це означає, що залишки підпорядковуються нормальному розподілу N(0, 1).

qq-графік показує, що впорядкований розподіл залишків слідує лінійній тенденції вибірок з деякими відхиленнями вздовж кожного з хвостів. Виходячи з цього графіка, можемо сміливо припустити, що залишки розподілені нормально.

Прогнозування та оцінка моделі

Після отримання моделі для прогнозування часового ряду її можна використати для прогнозування. Далі ми отримаємо прогнози на наступні 24 місяці і порівняємо з історичними даними.

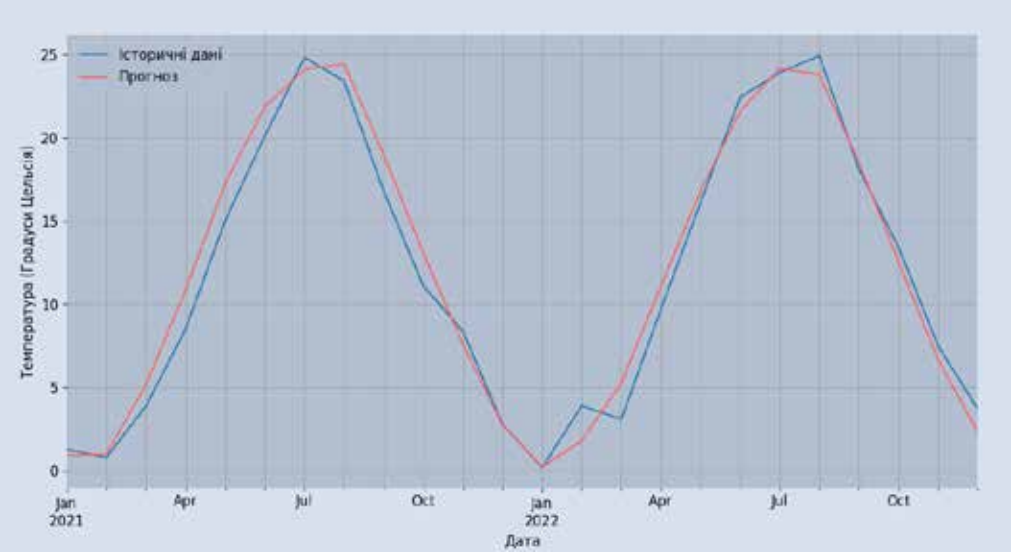

Рис. 9. Результати порівняння прогнозів з історичними даними

На рис. 9 наведено результати порівняння прогнозів з реальними історичними даними за період із січня 2021 року по грудень 2022 року: лінія пастельно-червоного кольору вказує на прогнозовані дані, синя – на історичні.

Далі нам необхідно порахувати похибки прогнозування. Це було реалізовано за допомогою функцій пакета sklearn.

```
The MSE is 1.8
The RMSE is 1.3
The MAPE is 17.3
```

Рис. 10. Похибки отриманих прогнозів

На рис. 10 наведено значення середньоквадратичної похибки (MSE), корінь із середньоквадратичної похибки (RMSE) та середньої абсолютної похибки у відсотках (MAPE) прогнозів порівняно з фактичним значеннями температури.

З рисунку видно, що середня абсолютна відносна похибка (MAPE) становить 17,3%, що свідчить про те, що в середньому прогноз не відповідає історичним значенням температури повітря на 17,3%. Це вказує на високий рівень точності прогнозування моделі SARIMA, враховуючи велику кількість факторів, які можуть впливати на температуру повітря (швидкість вітру, опади тощо).

https://doi.org/10.32782/mathematical-modelling/2024-7-2-22

Висновки

На підставі використання моделі прогнозування SARIMA можна дійти висновку, що запропоновані методи прогнозування мають достатньо високий рівень точності, це підтверджує порівняння тестової вибірки з результатами прогнозування моделі SARIMA.

Список використаної літератури

1. Hayes A. What Is a Time Series and How Is It Used to Analyze Data? URL: https://www.investopedia.com/terms/t/timeseries.asp (дата звернення: 19.11.2023).
2. SARIMA (Seasonal Autoregressive Integrated Moving Average) URL: https://www.geeksforgeeks.org/sarima-seasonal-autoregressive-integrated-moving-average (дата звернення: 19.11.2023).
3. Сикало Є.А. Атмосфера та клімат. Спостереження за погодою та її прогнозування. URL: https://subject.com.ua/geographic/zno_2017/47.html (дата звернення: 19.11.2023).
4. Міллер Н. Як роблять прогнози погоди і чому вони іноді не збуваються? *BBC News*. 2020. 22 лютого. URL: https://www.bbc.com/ukrainian/features-51545290 (дата звернення: 19.11.2023).
5. Ma X., Fang C., Ji J. Prediction of outdoor air temperature and humidity using Xgboost. *IOP Conference Series: Earth and Environmental Science*. 2020. Vol. 427. DOI: 10.1088/1755-1315/427/1/012013.
6. Nugraha Y.E., Ariawan I., Arifin W.A. Weather Forecast from Time Series Data using LSTM Algorithm. *Jurnal Teknologi Informasi dan Komunikasi*. 2023. Vol. 14. № 2. DOI: 10.51903/jtikp.v14i1.531.
7. Sun C., Nong Y., Chen Z., Liang D., Lu Y., Qin Y. The CEEMD-LSTM-ARIMA Model and Its Application in Time Series Prediction. *Journal of Physics: Conference Series*. 2021. Vol. 2179. DOI: 10.1088/1742-6596/2179/1/012012.
8. Meteostat. Introduction | Meteostat Developers. URL: https://dev.meteostat.net/guide.html (дата звернення: 19.11.2023).
9. Stationarity and detrending (ADF/KPSS). URL: https://www.statsmodels.org/dev/examples/notebooks/generated/stationarity_detrending_adf_kpss.html (дата звернення: 19.11.2023).
10. Hyndman R.J. Facts and fallacies of the AIC. URL: https://robjhyndman.com/hyndsight/aic (дата звернення: 19.11.2023).
11. Smigel L. How to Interpret ARIMA Results URL: https://analyzingalpha.com/interpret-arima-results (дата звернення: 19.11.2023).

References

1. Hayes, A. What Is a Time Series and How Is It Used to Analyze Data? [Official site]. Retrieved from https://www.investopedia.com/terms/t/timeseries.asp [in English].
2. SARIMA [Seasonal Autoregressive Integrated Moving Average]. Retrieved from https://www.geeksforgeeks.org/sarima-seasonal-autoregressive-integrated-moving-average [in English].
3. Sykalo, Ye. A. Atmosfera ta klimat. Sposterezhennia za pohodoiu ta yii prohnozuvannia [Atmosphere and Climate. Weather Observation and Forecasting]. Retrieved from https://subject.com.ua/geographic/zno_2017/47.html [in Ukrainian].
4. Miller, N (2020, February, 22). Yak robliat prohnozy pohody i chomu vony inodi ne zbuvaiutsia? [How are weather forecasts made, and why do they sometimes fail to come true]. *BBC News*. Retrieved from https://www.bbc.com/ukrainian/features-51545290 [in Ukrainian].
5. Ma, X., Fang, C., & Ji, J. (2020). «Prediction of outdoor air temperature and humidity using Xgboost». *IOP Conference Series: Earth and Environmental Science,* 427, 012013. DOI: 10.1088/1755-1315/427/1/012013 [in English].

https://doi.org/10.32782/mathematical-modelling/2024-7-2-22

6. Nugraha, Y.E., Ariawan, I., & Arifin, W.A. (2023). Weather Forecast from Time Series Data using LSTM Algorithm. *Jurnal Teknologi Informasi dan Komunikasi* [Journal of Information and Communication Technology]*,* 14 (2), 144–152. DOI: 10.51903/jtikp.v14i1.531 [in English].
7. Sun, C., Nong, Y., Chen, Z., Liang, D., Lu, Y., & Qin, Y. (2021). The CEEMD-LSTM-ARIMA Model and Its Application in Time Series Prediction. *Journal of Physics: Conference Series,* 2179, 012012. DOI: 10.1088/1742-6596/2179/1/012012 [in English].
8. Meteostat. Introduction | Meteostat Developers. (2022). [Official site]. Retrieved from https://dev.meteostat.net/guide.html [in English].
9. Stationarity and detrending (ADF/KPSS). Retrieved from https://www.statsmodels.org/dev/examples/notebooks/generated/stationarity_detrending_adf_kpss.html [in English].
10. Hyndman, R.J. Facts and fallacies of the AIC. Retrieved from https://robjhyndman.com/hyndsight/aic [in English].
11. Smigel, L. How to Interpret ARIMA. Results Retrieved from https://analyzingalpha.com/interpret-arima-results [in English].

Халізев Нікіта Сергійович – студент кафедри програмних засобів Національного університету «Запорізька політехніка». E-mail: nikitakhalizev@gmail.com, ORCID: 0009-0005-5316-2542.

Дубровін Валерій Іванович – професор кафедри програмних засобів Національного університету «Запорізька політехніка». E-mail: vdubrovin@gmail.com, ORCID: 0000-0002-0848-8202.

Дейнега Лариса Юріївна – старший викладач кафедри програмних засобів Національного університету «Запорізька політехніка». E-mail: deynega.larisa@gmail.com, ORCID: 0000-0003-0304-4327.

Khalizev Nikita Serhiiovych – Student at the Department of Software of the Zaporizhia Polytechnic National University. E-mail: nikitakhalizev@gmail.com, ORCID: 0009-0005-5316-2542.

Dubrovin Valery Ivanovych – Professor at the Department of Software of the Zaporizhia Polytechnic National University. E-mail: vdubrovin@gmail.com, ORCID: 0000-0002-0848-8202.

Deynega Larisa Yuriivna – Senior Lecturer at the Department of Software of the Zaporizhzhya Polytechnic National University. E-mail: deynega.larisa@gmail.com, ORCID: 0000-0003-0304-4327.

https://doi.org/10.32782/mathematical-modelling/2024-7-2-22