

УДК 519.866

DOI: 10.31673/2786-8362.2025.025360

Бутко І.М., д.т.н.; Голубенко О.І., к.т.н.;
Коваленко С.М., к.ф.-м.н.; Маковейчук О.М., д.т.н.

ПРОГНОСТИЧНЕ МОДЕЛЮВАННЯ В SMART CITY НА ОСНОВІ ARIMA-МОДЕЛЕЙ

Butko I.M., Golubenko O.I., Kovalenko S.M., Makoveichuk O.M. Predictive modeling in Smart Cities based on ARIMA models. Accurate time series forecasting is a cornerstone of smart city management, enabling informed decision-making across energy systems, transportation networks, environmental monitoring, and public safety. This study investigates the application of ARIMA models for urban time series prediction, focusing on both standard ARIMA and residual-corrected ARIMA variants. Residual correction allows the capture of latent structural patterns and systematic deviations not fully addressed by the base ARIMA model, thereby enhancing forecast accuracy. Model performance is rigorously evaluated using a comprehensive set of metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), symmetric Mean Absolute Percentage Error (sMAPE), Mean Absolute Percentage Error (MAPE), and Mean Absolute Scaled Error (MASE), complemented by a multi-criteria ranking procedure to identify the optimal model configuration. The results demonstrate that residual-corrected ARIMA consistently outperforms standard ARIMA for short- and medium-term forecasting tasks, particularly for variables such as air temperature, which exhibit complex temporal dynamics. The findings underscore the practical relevance of ARIMA-based forecasting as a reliable tool for data-driven decision support in smart city infrastructures.

Keywords: time series, forecasting, ARIMA model, ARIMA with residual correction, model selection criteria

Бутко І.М., Голубенко О.І., Коваленко С.М., Маковейчук О.М. Прогнозоване моделювання в Smart City на основі ARIMA-моделей. У статті досліджується застосування ARIMA-моделей для прогнозування часових рядів. Розглянуто базові моделі ARIMA та їх модифікації з корекцією залишків, які дозволяють враховувати додаткові структурні закономірності у часових рядах та підвищувати точність прогнозу. Для оцінки ефективності моделей використано комплекс метрик точності — MAE, RMSE, sMAPE, MAPE та MASE — а також багатокритеріальне ранжування моделей для визначення оптимальної конфігурації. Результати дослідження підтверджують доцільність застосування ARIMA з корекцією залишків для коротко- та середньострокового прогнозування. Практичне значення роботи полягає у підвищенні точності прогнозів та підтримці прийняття рішень в інформаційно-орієнтованому міському середовищі.

Ключові слова: часові ряди, прогнозування, модель ARIMA, модель ARIMA з залишками, критерії вибору моделі

Вступ

Постановка завдання. Сучасна концепція «розумного міста» (*Smart City*) ґрунтується на інтеграції цифрових технологій, систем штучного інтелекту та аналітики даних у всі аспекти міського управління. Основна мета такої інтеграції – підвищення ефективності функціонування міської інфраструктури, раціональне використання ресурсів і поліпшення якості життя громадян. Кожен із цих процесів характеризується часовими закономірностями, трендами та періодичними коливаннями, що робить прогнозування ключовим інструментом у прийнятті рішень.

У системах *Smart City* часові ряди формуються практично у кожній підсистемі: від споживання електроенергії, обсягів транспортування води та кількості пасажирів у громадському транспорті – до рівня забруднення повітря, трафіку автомобілів, показників шумового навантаження чи запитів до муніципальних сервісів. У цьому контексті *прогнозування часових рядів* посідає центральне місце, оскільки саме воно забезпечує можливість *передбачати майбутні події та динаміку процесів* у міському середовищі на основі аналізу історичних даних [1-2].

Аналіз останніх досліджень. Прогнозування часових рядів охоплює широкий спектр підходів – від класичних статистичних моделей до сучасних методів глибинного навчання.

Кожен клас моделей ґрунтується на власній математичній парадигмі, що відображає різні уявлення про структуру та закономірності часових даних.

Однією з базових моделей лінійного прогнозування є авторегресивна інтегрована модель ковзного середнього – *ARIMA* (Autoregressive Integrated Moving Average) [3]. Вона описує три основні явища у часових рядах:

- авторегресія – залежність поточного значення від попередніх спостережень;
- інтеграція – усунення нестационарності ряду шляхом його диференціювання (differencing), тобто перетворення до різницево-стаціонарного часового ряду певного порядку, надалі цей порядок будемо називати порядком інтеграції;

- ковзне середнє – залежність від похибок попередніх прогнозів.

Формально модель *ARIMA* порядку (p, d, q) описується рівнянням:

$$\varphi_p(B)(1 - B)^d y_t = \theta_q(B)\varepsilon_t, \quad (1)$$

де y_t – спостережене значення часового ряду у момент часу t ;

$\varphi_p(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p$ – авторегресивний поліном порядку p ;

$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$ – поліном порядку q для ковзного середнього;

B – оператор зсуву назад у на один відлік по часу $B y_t = y_{t-1}$;

d – порядок інтеграції;

ε_t – випадкова складова (білий шум).

Компактність структури *ARIMA* забезпечує її високу ефективність у виявленні лінійних часових кореляцій. Схематичне зображення моделі наведено на рис. 1.



Рис. 1. Структурна схема моделі *ARIMA* (відтворено за [3])

Розширенням цієї моделі є *сезонна модель SARIMA* (Seasonal *ARIMA*) [4], що дає змогу враховувати регулярні сезонні коливання. Модель *SARIMA* порядку (P, D, Q) визначається рівнянням:

$$\Phi_P(B^S)\varphi_p(B)(1 - B)^d(1 - B^S)^D y_t = \Theta_Q(B^S)\theta_q(B)\varepsilon_t, \quad (2)$$

де S – довжина сезонного циклу (наприклад, $S = 12$ для по-місячних даних із річною сезонністю), $\Phi_P(B^S)$ і $\Theta_Q(B^S)$ – сезонні поліноми авторегресії порядку P та ковзного середнього порядку Q відповідно, а $(1 - B^S)^D$ – оператор сезонного диференціювання порядку D .

Головна перевага *SARIMA* над *ARIMA* полягає у здатності явно враховувати періодичність. Це критично важливо для часових рядів із регулярними повторюваними структурами – наприклад, місячні продажі, квартальні фінансові показники або сезонні кліматичні коливання [5, 6].

Модель *ARIMA* є ефективним інструментом для моделювання лінійних залежностей у часових рядах. Проте реальні дані часто містять залишкові структурні закономірності, які не повністю пояснюються первинною *ARIMA*. Для підвищення точності прогнозу застосовується *корекція залишків*, яка передбачає повторне моделювання похибок первинного прогнозу [4; 7].

Нехай \hat{y}_t^{base} – прогноз, отриманий за базовою моделлю без корекції:

$$\hat{y}_t^{\text{base}} = \text{ARIMA}_{\text{base}}(y_{t-1}, y_{t-2}, \dots; p, d, q, P, D, Q, S), \quad (3)$$

де y_t – фактичні значення часового ряду;

p, d, q – порядок авторегресії, інтеграції та ковзного середнього;

P, D, Q – сезонні параметри;

S – довжина сезонного циклу.

Визначимо тепер залишки як:

$$\hat{\varepsilon}_t = y_t - \hat{y}_t^{\text{base}}. \quad (4)$$

У загальному випадку залишки можуть містити власну структуру автокореляції та сезонності. Для їх моделювання використовується *додаткова ARIMA* з порядком, який може відрізнятись від базової моделі:

$$\hat{\varepsilon}_t = \text{ARIMA}_{\text{resid}}(\hat{\varepsilon}_{t-1}, \hat{\varepsilon}_{t-2}, \dots; p_r, d_r, q_r, P_r, D_r, Q_r) + \eta_t, \quad (5)$$

де p_r, d_r, q_r – порядки авторегресії, інтеграції та ковзного середнього для моделі залишків; P_r, D_r, Q_r – сезонні параметри; η_t – новий випадковий шум.

Цей підхід дозволяє вловлювати додаткові закономірності, які не були враховані базовою ARIMA, включаючи локальні коливання та сезонні ефекти.

Остаточний скоригований прогноз \hat{y}_t^{corr} формується як сума базового прогнозу та прогнозу залишків:

$$\hat{y}_t^{\text{corr}} = \hat{y}_t^{\text{base}} + \hat{\varepsilon}_t. \quad (6)$$

Застосування такої моделі є особливо ефективним у системах *Smart City*, де часові ряди характеризуються багаторівневими динамічними процесами – добовими і сезонними коливаннями енергоспоживання, змінним транспортним навантаженням, навантаженням на міську інфраструктуру та іншими системними впливами. У таких випадках корекція залишків дозволяє значно підвищити точність прогнозу та стабільність прийнятих рішень у реальному часі [8;9].

Метою роботи є практична оцінка ефективності застосування ARIMA-моделей, включно з моделями з корекцією залишків, для прогнозування часових рядів на основі реальних даних, а також у формулювання та обґрунтування критеріїв вибору оптимальної моделі для забезпечення високої точності та надійності прогнозів.

Виклад основного матеріалу дослідження

Гарним прикладом є прогнозування температури повітря, оскільки воно безпосередньо впливає на міське планування, управління енергоспоживанням та забезпечення громадської безпеки. Прогнози температури сприяють підтримці екологічного моніторингу, регулюванню транспортних потоків та реалізації заходів, пов'язаних із охороною здоров'я, що робить їх ключовим елементом інформаційно-орієнтованої міської інфраструктури [10].

У зв'язку з цим для дослідження моделей ARIMA було обрано завдання прогнозування температури повітря у м. Київ, Україна. Враховуючи певні обмеження щодо доступу до актуальної інформації через її чутливий характер, дослідження було проведено на спрощеній задачі з використанням даних за період 2012–2017 рр., частини набору даних Kaggle, що охоплює період 1881–2017 [11]. У наданому наборі даних інформація фіксувалася щодня; проте для цілей даного дослідження достатньо розглядати середні тижневі значення температури. Таким чином, сформовано набір даних із 304 записів середніх місячних температур для Києва за період від 1 січня 2012 року до 22 жовтня 2017 року. Для подальшого аналізу набір даних було розділено на дві частини: навчальну вибірку з 209 записів (від 1 січня 2012 року до 27 грудня 2015 року) та тестову вибірку з 95 записів (від 3 січня 2016 року до 22 жовтня 2017 року). Графік вихідних даних зображено на рис. 2.

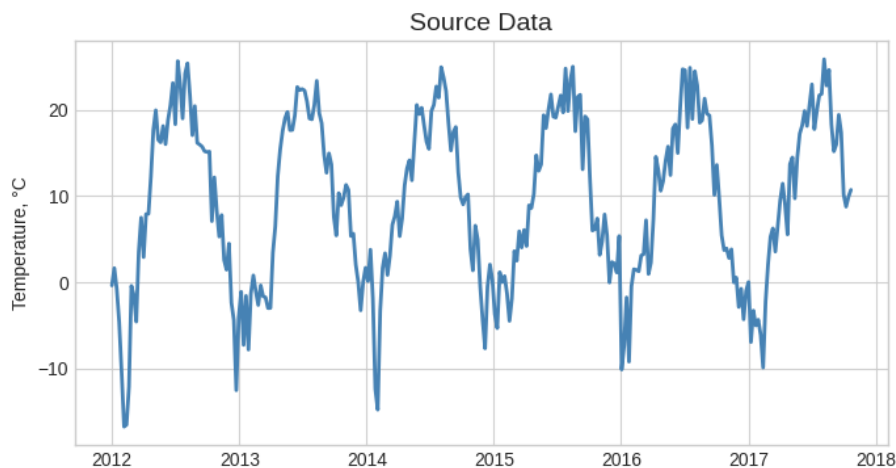


Рис. 2. Вихідні дані, що використовувалися для дослідження

Для оцінки якості даних було побудовано гістограму їх розподілу (рис. 3). Це дає змогу візуально проаналізувати форму розподілу, виявити можливу асиметрію, наявність викидів або сезонних аномалій [7; 12].

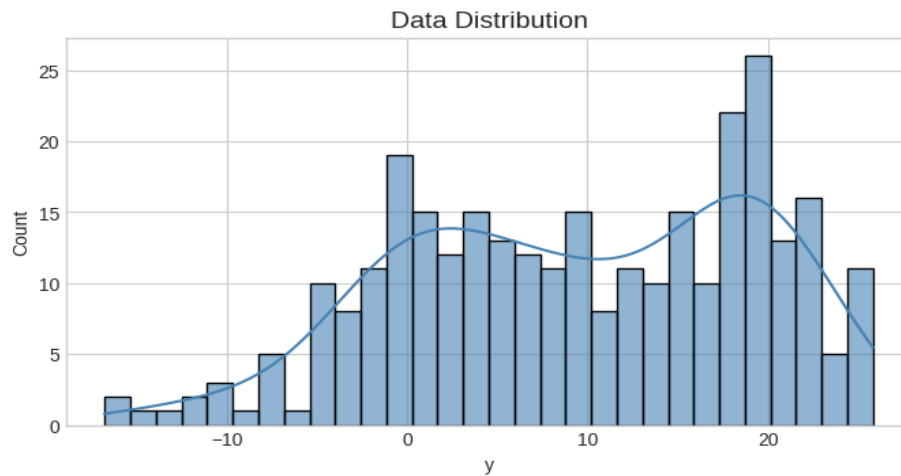


Рис. 3. Гістограма розподілу вихідних даних

Гістограма також дозволяє перевірити наближеність даних до нормального розподілу. У нашому випадку спостерігається бімодальний характер розподілу, що свідчить про виражену сезонність температурних коливань і наявність двох домінуючих режимів у динаміці ряду.

Для оцінки точності прогнозів у даному дослідженні застосовується комплекс метрик, що враховує як абсолютні, так і відносні відхилення прогнозу \hat{y}_t від фактичних значень y_t :

MAE (Mean Absolute Error) визначає середнє абсолютне відхилення:

$$\text{MAE} = \frac{1}{n} \sum_{t=1}^n |y_t - \hat{y}_t|, \quad (7)$$

RMSE (Root Mean Squared Error) підкреслює вплив великих похибок:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \hat{y}_t)^2}, \quad (8)$$

sMAPE (Symmetric Mean Absolute Percentage Error) відображає симетричну відносну похибку:

$$\text{sMAPE} = 100 \cdot \frac{1}{n} \sum_{t=1}^n \frac{2 |y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|}, \quad (9)$$

MAPE (Mean Absolute Percentage Error) дає середнє відносне відхилення у відсотках, із захистом від ділення на нуль:

$$\text{MAPE} = 100 \cdot \frac{1}{n} \sum_{t=1}^n \frac{|y_t - \hat{y}_t|}{|y_t| + 1}, \quad (10)$$

Нарешті, **MASE** (Mean Absolute Scaled Error) дозволяє порівнювати точність прогнозу з найвним методом, масштабуючи похибку відносно середньої різниці послідовних спостережень:

$$\text{MASE} = \frac{\text{MAE}}{\frac{1}{n-1} \sum_{t=2}^n |y_t - y_{t-1}|}. \quad (11)$$

У процесі оцінювання моделей ARIMA для прогнозування температури повітря використано *багатокритеріальний підхід* до вибору найкращої моделі. Для кожної

конфігурації параметрів (p, d, q) обчислювалися основні метрики точності прогнозу – MAE, RMSE, sMAPE, MAPE та MASE.

Оскільки жодна окрема метрика не відображає повною мірою якості моделі, було застосовано метод ранжування. Для кожної метрики визначався ранг моделі: менше значення метрики відповідало кращому рангу. Далі ранги по всіх метриках сумувалися для кожної моделі, утворюючи інтегральний показник якості [13; 14].

Нехай $R_{i,j}$ – ранг i -ї моделі за j -ю метрикою ($j = 1, 2, \dots, m$), де m – кількість обраних метрик. Тоді сумарний ранг моделі визначається як

$$S_i = \sum_{j=1}^m R_{i,j}, \quad (12)$$

а найкраща модель вибирається за мінімальним сумарним рангом:

$$i_{\text{best}} = \arg \min_i S_i. \quad (13)$$

Для підбору оптимальної конфігурації ARIMA у дослідженні було проведено тестування моделей у такому діапазоні параметрів:

- p (порядок авторегресії, AR) від 0 до 2,
- q (порядок ковзного середнього, MA) від 0 до 1,
- d (порядок інтеграції) від 0 до 2.

Загалом формувалися всі комбінації параметрів (p, d, q) у визначеному діапазоні, за винятком трійки $(0,0,0)$, оскільки вона не задає ARIMA-модель. Для кожної конфігурації ARIMA обчислювалися метрики точності MAE, RMSE, sMAPE, MAPE та MASE за формулами (7)–(11). Крім того, відповідно до виразів (12)–(13), визначався сумарний ранг моделі, який дозволяє оцінити її стабільність та ефективність за всіма критеріями одночасно. Отримані результати узагальнено у Табл. 1.

Зауважимо, що частина конфігурацій моделей ARIMA не змогла бути оцінена через обмеження щодо включення трендових компонентів у моделі з інтеграцією ($d > 0$) або сезонною інтеграцією ($D > 0$). Зокрема, у моделях, де порядок інтеграції перевищує нуль, трендові терми нижчого порядку не можуть бути використані, оскільки вони автоматично усуваються операцією різниці. Через це деякі моделі, наприклад ARIMA(0,1,0), ARIMA(0,1,1) та ARIMA(0,1,2), не були обчислені.

Таблиця 1

Порівняння моделей ARIMA за метриками точності та сумарним рангом

index	MAE	RMSE	sMAPE	MAPE	MASE	Total Rank (S_i)
ARIMA(0, 0, 1)	2,97	3,75	55,43	200,76	0,97	5
ARIMA(0, 0, 2)	3,05	3,91	56,48	222,41	0,99	10
ARIMA(1, 0, 2)	4,66	5,80	67,45	439,34	1,52	15
ARIMA(2, 0, 2)	4,76	5,90	67,97	446,80	1,55	20
ARIMA(1, 0, 1)	5,07	6,22	69,59	472,34	1,65	25
ARIMA(2, 0, 0)	5,11	6,27	69,74	481,06	1,66	30
ARIMA(1, 0, 0)	5,17	6,34	70,03	487,10	1,68	35
ARIMA(2, 0, 1)	5,34	6,53	70,86	504,35	1,74	40

Для побудови базового прогнозу моделі ARIMA використовувалася конфігурація $seasonal_order = (1,0,1,52)$ та $trend = 'ct'$, що дало змогу врахувати сезонні коливання з періодом 52 тижні і трендову складову даних. Найкращою за результатами багатокритеріальної оцінки виявилася модель ARIMA(0,0,1) (рис. 4).

На рис. 5 подано графік залишків базового прогнозу для цієї моделі, який відображає відхилення прогнозованих значень від фактичних. Для зазначеної конфігурації отримано такі значення метрик точності: MAE = 2.973, RMSE = 3.751, sMAPE = 55.43 %, MAPE = 200.76 %, MASE = 0.968. Сума рангів моделі становила 5.0, що свідчить про стабільно добрі результати за всіма критеріями.

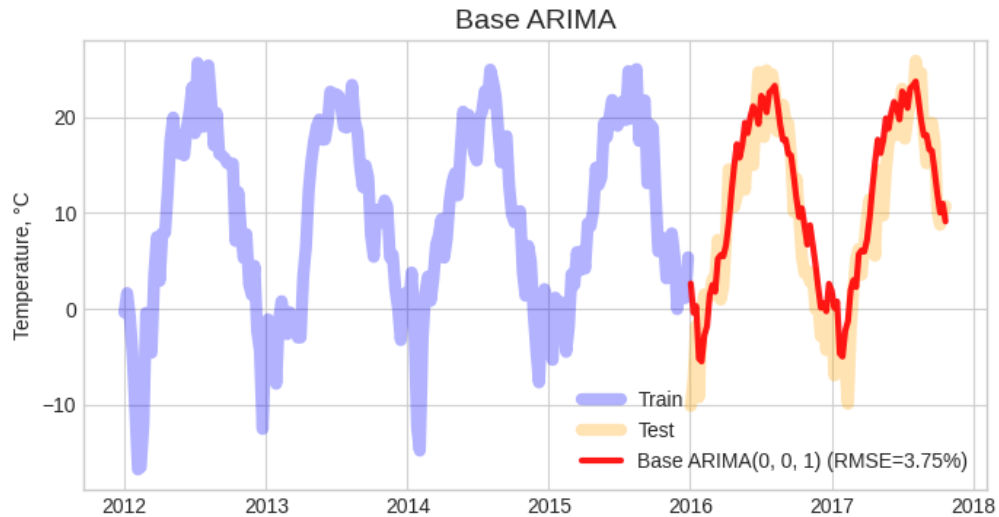


Рис. 4. Базовий прогноз температури повітря за найкращою моделлю ARIMA(0,0,1)

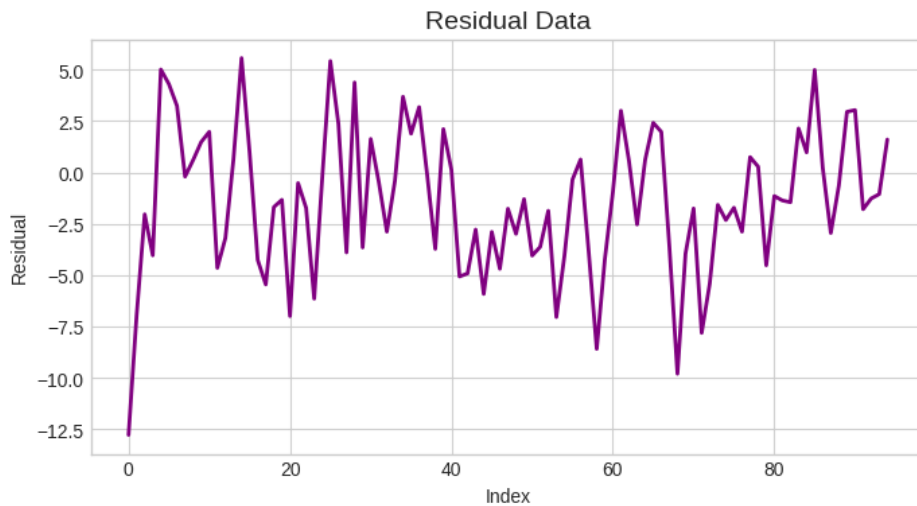


Рис. 5. Залишки базового прогнозу ARIMA(0,0,1)

Для оцінки якості побудованої моделі та перевірки адекватності прогнозу було виконано аналіз залишків. На рис. 6 наведено гістограму розподілу залишків, яка дозволяє візуально оцінити їхню форму, симетрію та наявність відхилень від нормального розподілу. Як видно, залишки мають майже симетричний розподіл навколо нуля, що свідчить про відсутність систематичних похибок у моделі.

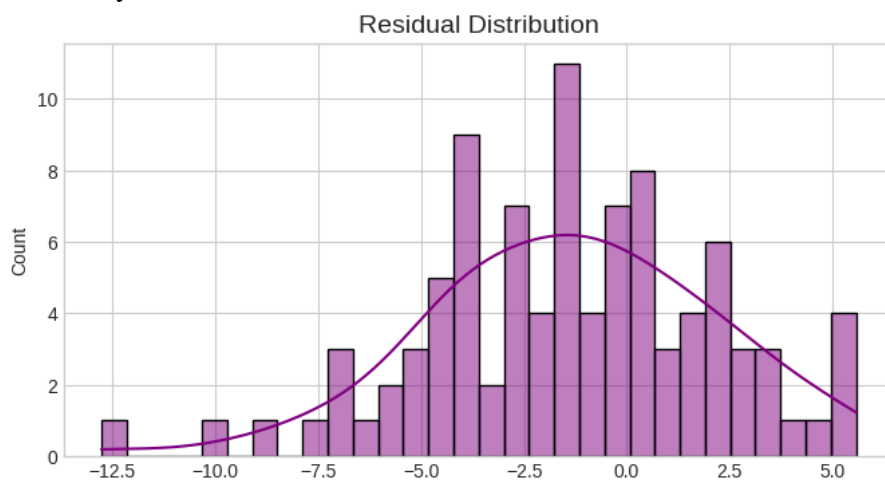


Рис. 6. Гістограма розподілу залишків

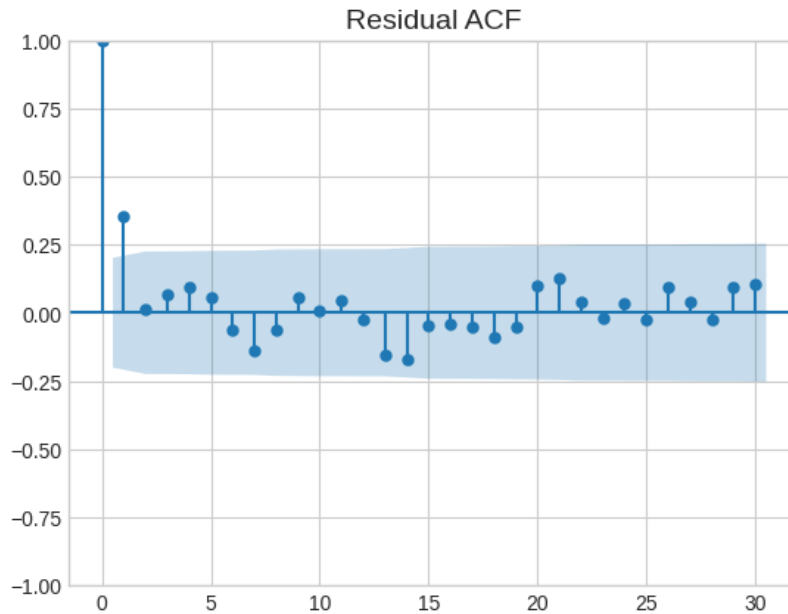


Рис. 7. Автокореляційна функція залишків

На рис. 7 подано графік автокореляційної функції (ACF) залишків. Відсутність статистично значущих автокореляцій за межами довірчого інтервалу підтверджує, що залишки можна вважати випадковими і неструктурованими, тобто модель **ARIMA(0,0,1)** адекватно описує динаміку даних.

На рис. 8 представлено уточнений прогноз, отриманий після моделювання та корекції залишків базової ARIMA-моделі. Додавання прогнозу залишків дозволяє врахувати додаткові залежності та локальні коливання у даних, що підвищує точність прогнозу та зменшує систематичні відхилення від фактичних значень. Графік демонструє покращену відповідність прогнозу реальним спостереженням порівняно з базовим прогнозом (рис. 4).

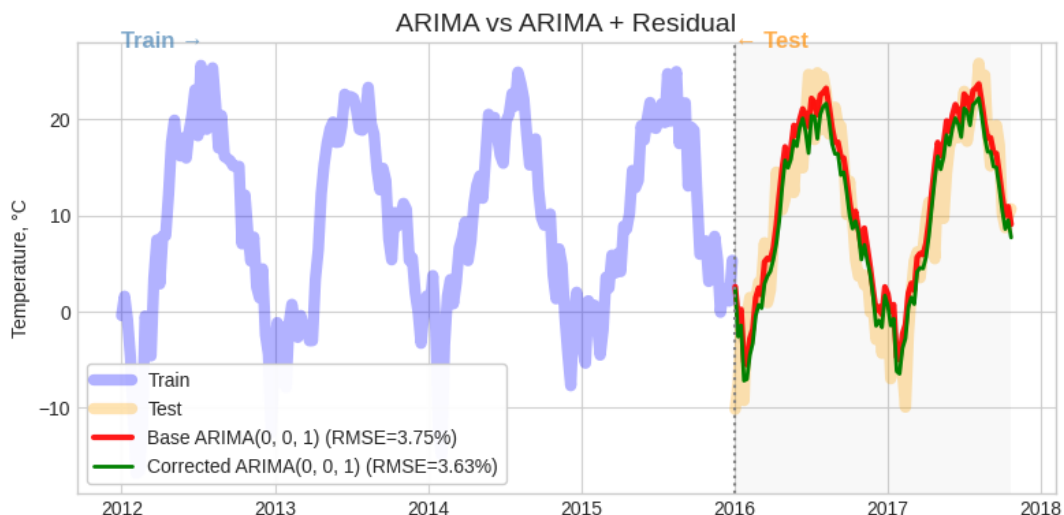


Рис. 8. Прогноз після корекції залишків базової ARIMA-моделі

Моделювання виконувалося в середовищі Google Colab [15] із використанням мови Python та стандартних бібліотек для роботи з часовими рядами та статистичним моделюванням, таких як pandas [16], numpy [17], statsmodels [18] та seaborn [19].

Висновки

У роботі досліджено застосування ARIMA-моделей для прогнозування середньомісячної температури повітря у м. Київ. Показано, що базові ARIMA-моделі дозволяють моделювати

сезонні коливання та трендову складову, тоді як корекція залишків за допомогою додаткової ARIMA-моделі підвищує точність прогнозу та зменшує систематичні похибки.

Найкращою виявилася модель ARIMA(0,0,1) із сезонним порядком (1,0,1,52) та трендовою складовою, яка продемонструвала стабільно добрі результати за усіма метриками точності та сумарним ранжуванням.

Рекомендації для подальших досліджень:

1. Розширити аналіз на інші метеорологічні параметри та часові ряди Smart City (наприклад, трафік, енергоспоживання, рівень забруднення повітря).
2. Порівняти ефективність ARIMA з іншими підходами прогнозування, зокрема LSTM та Prophet, для виявлення більш складних залежностей.
3. Розглянути використання багатофакторних моделей та інтеграцію зовнішніх даних (погодні умови, свята, соціальні події) для підвищення точності прогнозів.

Список використаної літератури:

1. Zhang Z., Ren S., Qian X., Duffield N. Towards Invariant Time Series Forecasting in Smart Cities // arXiv preprint arXiv:2405.05430. 2024. URL: <https://arxiv.org/abs/2405.05430>.
2. Liang S. Predicting short-term urban bike sharing demand in a smart city // Computers, Environment and Urban Systems. 2026. Vol. 88. P. 101703. URL: <https://doi.org/10.1016/j.compenvurbsys.2021.101703>.
3. Box G.E.P., Jenkins G.M. Time Series Analysis: Forecasting and Control. Prentice Hall, 1970.
4. Hyndman R.J., Athanasopoulos G. Forecasting: Principles and Practice. OTexts, 2021. URL: <https://otexts.com/fpp3/>.
5. Yavuz V.S. Forecasting monthly rainfall and temperature patterns in Van Province, Türkiye using ARIMA and SARIMA models // Journal of Water and Climate Change. 2025.
6. GeeksforGeeks. ARIMA vs. SARIMA model. URL: <https://www.geeksforgeeks.org/machine-learning/arima-vs-sarima-model/>.
7. Machine Learning Mastery. Model residual errors: Correct time series forecasts in Python. URL: <https://machinelearningmastery.com/model-residual-errors-correct-time-series-forecasts-python/>.
8. ARIMA-model of Ukrainian Macroeconomic Indicators Forecasting // In: Emergence of public development: financial and legal aspects: monograph. Coventry: Agenda Publishing House, 2019. P. 213-221.
9. Application of residual modification approach in seasonal ARIMA for electricity demand forecasting: A case study of China / Y. Wang et al. Energy Policy. 2012. Vol. 48. P. 284–294. URL: <https://doi.org/10.1016/j.enpol.2012.05.026>.
10. Real-Time Temperature Prediction Models for Smart City Applications. URL: https://www.researchgate.net/publication/388194104_Real-Time_Temperature_Prediction_Models_for_Smart_City_Applications.
11. Kyiv, Ukraine Weather Statistics Dataset. URL: <https://www.kaggle.com/datasets/thedevastator/kyiv-ukraine-weather-statistics>
12. APXML. ARIMA Model Diagnostics. URL: https://apxml.com/courses/time-series-analysis-forecasting/chapter-4-arima-models-forecasting/arima-model-diagnostics?utm_source=chatgpt.com.
13. Bhutani K. Evaluating Time Series Models for Real-World Forecasting: A Practical Comparison. URL: <https://medium.com/%40karanbhutani477/evaluating-time-series-models-for-real-world-forecasting-a-practical-comparison-5c9622618715>.
14. Khudov H., Makoveychuk O., Butko I., Khizhnyak I. A model for prediction of geospatial data in systems for processing geospatial information // Systems of Arms and Military Equipment. – 2021. № 2(66). С. 123-128. URL: <https://doi.org/10.30748/soivt.2021.66.16>.
15. Google Colab. URL: <https://colab.google/>.
16. Pandas. URL: <https://pandas.pydata.org/>.

17. NumPy. URL: <https://numpy.org/>.
18. Statsmodels. URL: <https://www.statsmodels.org/stable/index.html>.
19. Seaborn. URL: <https://seaborn.pydata.org/>.

Автори статті

Бутко Ігор – доктор технічних наук, професор, Заклад вищої освіти «Міжнародний науково-технічний університет імені академіка Юрія Бугая», Київ, Україна.

ORCID: 0000-0002-2859-0351

Голубенко Олександр – кандидат технічних наук, доцент, Заклад вищої освіти «Міжнародний науково-технічний університет імені академіка Юрія Бугая», Київ, Україна.

ORCID: 0000-0002-1776-5160

Коваленко Сергій – кандидат фізико-математичних наук, доцент, Заклад вищої освіти «Міжнародний науково-технічний університет імені академіка Юрія Бугая», Київ, Україна.

ORCID: 0000-0002-8315-1589

Маковейчук Олександр – доктор технічних наук, професор, Заклад вищої освіти «Міжнародний науково-технічний університет імені академіка Юрія Бугая», Київ, Україна.

ORCID: 0000-0003-4425-016X

Authors of the article

Butko Ihor – Doctor of Sciences (technical), Professor, Higher Education Institution “Academician Yuriy Bugay International Scientific and Technical University”, Kyiv, Ukraine.

ORCID: 0000-0002-2859-0351

Golubenko Oleksandr – Candidate of Sciences (technical), Associate Professor, Higher Education Institution “Academician Yuriy Bugay International Scientific and Technical University”, Kyiv, Ukraine.

ORCID: 0000-0002-1776-5160

Kovalenko Serhii – Candidate of Sciences (physics and mathematics), Associate Professor, Higher Education Institution “Academician Yuriy Bugay International Scientific and Technical University”, Kyiv, Ukraine.

ORCID: 0000-0002-8315-1589

Makoveichuk Oleksandr – Doctor of Sciences (technical), Professor, Higher Education Institution “Academician Yuriy Bugay International Scientific and Technical University”, Kyiv, Ukraine.

ORCID: 0000-0003-4425-016X