ISSN 2707-1898 (print)

Український журнал інформаційних технологій
Ukrainian Journal of Information Technology
http://science.lpnu.ua/uk/ujit
https://doi.org/10.23939/ujit2024.02.049
Article received 15.10.2024 p.
Article accepted 19.11.2024 p.
UDC 004.94

Correspondence author
O. F. Shevchuk
shevchuk177@gmail.com

А. А. Яровий, О. Ф. Шевчук, А. В. Козловський, Ю. М. Паночишин, С. В. Сімончук

Вінницький національний технічний університет, Вінниця, Україна

ВИКОРИСТАННЯ ARIMA МОДЕЛЕЙ ДЛЯ ПРОГНОЗУВАННЯ ЗАГАЛЬНОГО РІВНЯ ЗЛОЧИННОСТІ В УКРАЇНІ

Прогнозування рівня злочинності – важливий аспект розроблення стратегій сталого соціально-економічного розвитку правової держави. Особливої значущості точне прогнозування набуває в умовах економічної нестабільності та геополітичних криз, характерних для України. У статті досліджено проблеми побудови та використання авторегресійних моделей інтегрованого ковзного середнього (ARIMA) для прогнозування загальної кількості злочинів, вчинених на території України. Розрахунки показали, що часовий ряд злочинності (1990–2023 рр.) демонструє ознаки спадного тренду, є нестаціонарним і містить аномальні значення кількості злочинів у 2003, 2013 та 2020 рр. Використання методу інтегрування даних, із взяттям перших різниць між спостереженнями, призводить до втрати автокореляційної структури, яка була притаманна загальному ряду злочинності. Як наслідок, початкову модель ARIMA (1, 0, 0) побудовано на підставі неперетворених вхідних даних. Точність цієї моделі (MAPE = 8,61 %) виявилася вищою порівняно з моделлю, отриманою за методом експоненційного згладжування (MAPE = 9,38 %). Логарифмування часового ряду злочинності та згладжування аномальних рівнів сприяли підвищенню прогностичної валідності, що дало змогу моделі ARIMA врахувати додаткову автокореляцію, уникнувши необхідності введення компоненти ковзного середнього. В результаті модель ARIMA (2, 0, 0) показала найвищу точність (MAPE = 7,04 %) за найменшої складності, що підтверджують результати визначення інформаційних критеріїв. Крім того, модель успішно пройшла перевірку на стійкість за допомогою методу перехресної валідації з вилученням одного спостереження. Прогнозні оцінки, побудовані на основі усіх розглянутих ARIMA моделей, вказують на подальше зростання загального рівня злочинності в Україні, яке розпочалося у 2021 р. після тривалого періоду зниження.

Ключові слова: моделювання, автокореляція, аномальні значення, крос-валідація, кількість вчинених злочинів.

Вступ / Introduction

Прогнозування рівня злочинності є одним із ключових чинників у розробленні стратегій сталого соціально-економічного розвитку будь-якої правової держави. Ефективне прогнозування дає змогу вчасно реагувати на потенційні загрози, оптимізувати роботу правоохоронних органів та забезпечувати стабільність у суспільстві, що особливо актуально для країн зі складними соціально-економічною та політичною ситуацією.

Саме в таких надзвичайних та перехідних умовах опинилася Україна після здобуття незалежності. В її економічному просторі відбувалися складні, довготривалі та болючі для суспільства трансформаційні процеси, які додатково підсилювалися загальною політичною нестабільністю. Високий рівень безробіття, неодноразова девальвація національної валюти, політичні кризи, спровоковані рф, протестні настрої у суспільстві, Помаранчева революція (2004–2005 рр.), Революція гідності (2013–2014 рр.), військова агресія рф, окупація частин Донецької та Луганської областей, анексія Криму, а також карантинні обмеження через пандемію COVID-19 – всі ці фактори істотно впливали на формування рівня злочинності в Україні та одночасно підвищували її загальний рівень невизначеності.

Отже, перераховані чинники не лише підкреслюють актуальність теми дослідження, але й вказують на необхідність ґрунтовного статистичного оцінювання історичних даних із застосуванням сучасних методів побудови надійних прогнозних моделей.

Об'єкт дослідження – процес побудови авторегресійних моделей ковзного середнього (ARIMA) для прогнозування загального рівня злочинності.

Предмет дослідження – методи та підходи до побудови ARIMA моделей прогнозування рівня злочинності, а також способи підвищення їхньої точності та прогностичної валідності

Мета роботи – розроблення та побудова автокореляційної моделі інтегрованого ковзного середнього для прогнозування загального рівня злочинності в Україні.

Для досягнення зазначеної мети визначено такі основні *завдання дослідження*:

1) дослідити часовий ряд злочинності на стаціонарність, наявність тренду та аномальних рівнів;

2) побудувати ARIMA модель злочинності із оптимально підібраними параметрами;

3) визначити можливі способи підвищення прогностичної валідності та точності побудованої ARIMA моделі.

Матеріали і методи дослідження. *Емпіричні дані.* Дослідження здійснено на підставі офіційної статистичної інформації про кількість злочинів, вчинених на території України (y_t) у 1990–2023 рр. [14, 15].

Методологічною основою дослідження є математичний апарат кореляційно-регресійного аналізу. Зокрема, в роботі використано спеціалізовані параметричні тести Дікі – Фуллера, Манна – Кендалла та Ірвіна для аналізу часових рядів; загальну методологію побудови ARIMA-моделей для прогнозування часових рядів; тести Льюнга – Бокса (*Q*-тест), Жарка – Бера (*JB*-тест) та критерії Акаіке (AIC) і Байєса (BIC) для оцінювання адекватності та якості побудованих моделей.

Програмні інструменти та бібліотеки. Прогнозування та аналіз часового ряду злочинності виконано за допомогою мови програмування *Python* в інтерактивному середовищі *Jupyter Notebook* із використанням спеціалізованих бібліотек. Для роботи з масивами даних та опрацювання табличної інформації використано бібліотеки *numpy* та *pandas*. Візуалізацію результатів забезпечено бібліотекою *matplotlib*. Для побудови та аналізу ARIMA-моделей застосовано бібліотеки *statsmodels, pmdarima* та *scikit-learn*.

Аналіз останніх досліджень і публікацій. Аналіз літературних джерел показує, що авторегресійні моделі інтегрованого ковзного середнього (AutoRegressive Integrated Moving Average, ARIMA) є ефективним інструментом для короткострокового прогнозування часових рядів, оскільки вони здатні виявляти приховані структури та закономірності, що можуть бути неочевидними без попередньої обробки даних [1]. Ці моделі нині широко застосовують у різних галузях науки та техніки, зокрема для прогнозування метеоданих [2]; характеристик комп'ютерних мереж [3]; грошової вартості автомобілів [4]; обсягів прямих іноземних інвестицій [5]; динаміки зовнішньої торгівлі України [6] тощо.

Проте для прогнозування рівня злочинності в Україні ARIMA моделі вітчизняні науковці практично не використовували. Натомість у зарубіжній літературі цей напрям наукових досліджень доволі популярний [7–13]. Зокрема, у роботах [7–9] автори підкреслюють значно вищу прогностичну валідність саме моделей ARIMA порівняно з методами експоненційного згладжування. Такі висновки зроблено на основі аналізу часових рядів про рівень злочинності в Лондоні [7], майнових злочинів у одному з міст Китаю [8], кримінальних злочинів проти життя та здоров'я особи у регіоні Мванза, Танзанія [9].

Достатньо високу точність прогнозування забезпечили й ARIMA моделі, побудовані за статистичними даними про кількість злочинів у Сан-Франциско [10] та Чикаго [11]. До того ж за їх допомогою вдалося виявити щомісячні, щотижневі та щоденні сезонні закономірності у показниках злочинності.

Відзначимо також й результати досліджень, викладені у роботі [12]. Автори цієї статті проаналізували можливість використання моделей ARIMA для прогнозування різних типів злочинності серед неповнолітніх осіб в Індії, виділяючи різні вікові категорії. Параметри моделей вибирали як вручну, так і за допомогою автоматичного підбору через *Auto ARIMA*, після чого їх порівнювали. У результаті виявилося, що для трьох вікових категорій за п'ятьма основними видами злочинів найкращих результатів було досягнуто саме завдяки ручному коригуванню параметрів ARIMA, а не автоматичним підбором Auto ARIMA. Це, зокрема, підкреслює важливість експертного підходу під час налаштування моделей для підвищення точності прогнозування.

Результати дослідження та їх обговорення / Research results and their discussion

Особливості застосування моделей ARIMA (p, d, q). Модель ARIMA (*p, d, q*) – лінійна модель, яка підходить для роботи зі стохастичними рядами. Вона поєднує авторегресійну модель AR(*p*) та модель ковзного середнього MA(*q*) [1]. До того ж для забезпечення умови стаціонарності застосовується й попереднє інтегрування часового ряду I(*d*). Зазвичай ARIMA модель подається у такій формі:

$$\left(1-\phi_1 L-\phi_2 L^2-\ldots-\phi_p L^p\right)\left(1-L\right)^d y_t = \\ =\left(1+\theta_1 L+\theta_2 L^2+\ldots+\theta_q L^q\right)\varepsilon_t \qquad (1)$$

або за відсутності інтегрування ($d = 0$)

$$y_t = a + \sum_{i=1}^{p}\phi_i\ y_{t-1} + \sum_{j=1}^{q}\theta_j\ y_{t-j} + \varepsilon_t\ ,$$

де y_t – значення часового ряду в час t; ϕ_t – коефіцієнти авторегресії; p – порядок авторегресії; L – оператор лагу; d – порядок інтегрування; θ_t – коефіцієнти ковзної середньої; q – порядок ковзної середньої; ε_t – випадковий шум.

Проте важливо зазначити, що для адекватного опису часових рядів та підвищення точності прогнозування на основі ARIMA-моделі необхідно врахувати певні умови щодо її використання.

Стаціонарність. Часовий ряд має бути стаціонарним. Це означає, що такі характеристики, як середнє, дисперсія та автокореляція, повинні бути сталими та не змінюватися із часом.

Відсутність сезонності. ARIMA підходить для серій з трендами, але непридатна для даних із вираженою сезонною компонентою.

Лінійність. ARIMA – лінійна модель, тому вона найкраще працює з лінійними залежностями між спостереженнями.

Випадковість залишків. Після моделювання залишки (помилки прогнозування) повинні бути випадковими і не автокорельованими.

Кількість спостережень. Для коректного оцінювання параметрів моделі ARIMA необхідна достатня кількість спостережень. Рекомендовано мати не менше ніж 50 спостережень, хоча точна кількість варіюється.

Повнота даних. Часовий ряд має містити повний набір даних без відсутніх значень. Пропуски можуть вплинути на оцінку параметрів моделі.

Аналіз часового ряду кількості вчинених на території України злочинів

Насамперед зазначимо, що моделювання здійснено на основі 34 спостережень часового ряду злочинності (1990–2023 рр.). Така мінімально допустима кількість для використання моделі ARIMA. Це потребує ретельнішого аналізу історичних даних, додаткового оцінювання стабільності моделі та інтерпретації результатів прогнозування.

Попередній візуальний аналіз часового ряду кількості вчинених злочинів на території України (y_t) вказує на його певну коливальну структуру із незначним спадним трендом (рис. 1). Тому на початковому етапі, для підтвердження зроблених припущень, ряд злочинності досліджено на стаціонарність за тестом Дікі–Фуллера [16], наявність тренду за тестом Манна–Кендалла [17], а також наявність аномальних рівнів за методом Ірвіна [18]. Значення критерію Ірвіна (λ_t) розраховано за формулою

$$\lambda_t = \frac{y_t - y_{t-1}}{\sigma_y} \qquad (2)$$

та порівняно з критичною точкою (рис. 1)

Виконавши розрахунки, гіпотезу про стаціонарність часового ряду ми відхилили, тоді як гіпотеза про існування статистично значущого тренду у даних злочинності, навпаки, підтвердилась. Також було підтверджено й виявлену у роботі [19] аномальну зміну кількості злочинів у 2003, 2013 та 2020 рр. (рис. 1), яка корелює із суспільно-політичними кризами в Україні. А саме: аномальне збільшення кількості злочинів (2003 р. та 2013 р.) відповідає зростанню протестних настроїв у суспільстві, що передували Помаранчевій революції (2004–2005 рр.) та Революції гідності (2013–2014 рр.). Натомість аномально низькі показники злочинності у 2020 р. зумовлені запровадженням обмежувальних карантинних заходів, спричинених пандемією COVID-19.

Варто також зазначити, що виявлені аномалії можуть істотно вплинути на результати моделювання, зокрема на точність побудованих прогнозів. Тому в подальших розрахунках важливо враховувати їх наявність, особливо під час оцінювання можливих напрямів підвищення прогностичної валідності отриманої моделі.

Автокореляційна функція (ACF) та часткова автокореляційна функція (PACF)

Автокореляційна та часткова автокореляційна функції – важливі інструменти аналізу часових рядів, зокрема під час побудови ARIMA-моделей. Вони допомагають зрозуміти, як значення ряду корелюють між собою на різних затримках (лагах). Отже, проаналізувавши їх, можна вибрати найоптимальніші значення параметрів p та q.

Проте, оскільки в нашому випадку часовий ряд злочинності має ознаки тренду та є нестаціонарним, на початковому етапі ми додатково проінтегрували його, взявши перші різниці між спостереженнями ($d = 1$), що дало змогу отримати стаціонарний інтегрований ряд злочинності. На рис. 2 побудовано ACF та PACF для основного та інтегрованого рядів злочинності.

Аналіз даних, наведених на рис. 2, показує, що для інтегрованого ряду злочинності як ACF, так і PACF не мають істотних значень для жодного з лагів. Це означає, що після інтегрування ряд злочинності не лише став стаціонарним, але й втратив автокореляцію, наявну в початкових даних. Отже, отримані після інтегрування дані є випадковими і не містять статистично значущої структури, необхідної для моделювання та прогнозування. У такому випадку будувати ARIMA недоцільно, оскільки відсутня стійка залежність між спостереженнями.

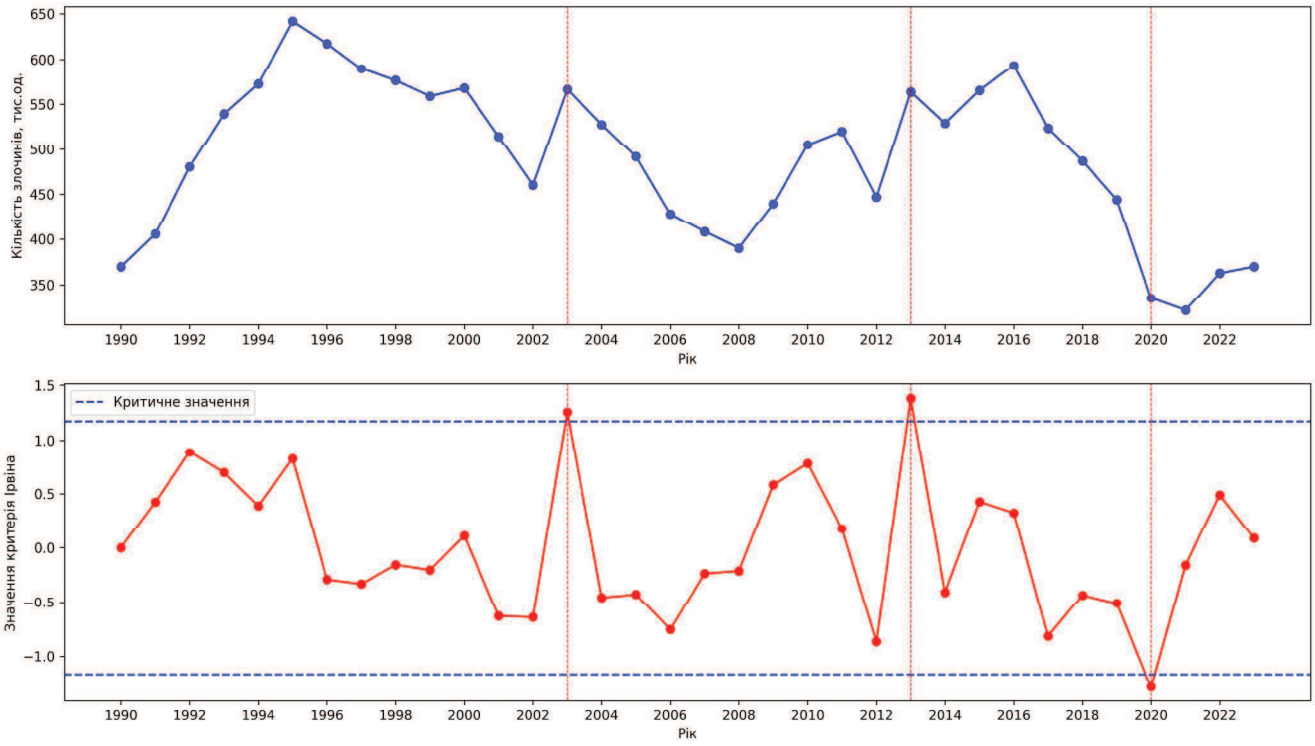

Рис. 1. Часовий ряд кількості злочинів, вчинених на території України, та відповідна динаміка зміни, розрахована за критерієм Ірвіна (2) / The time series of the number of committed crimes in Ukraine and the corresponding dynamic changes, calculated using the Irwin criterion (2)

Пунктиром відзначено роки, у яких спостерігалися аномальні зміни злочинності.

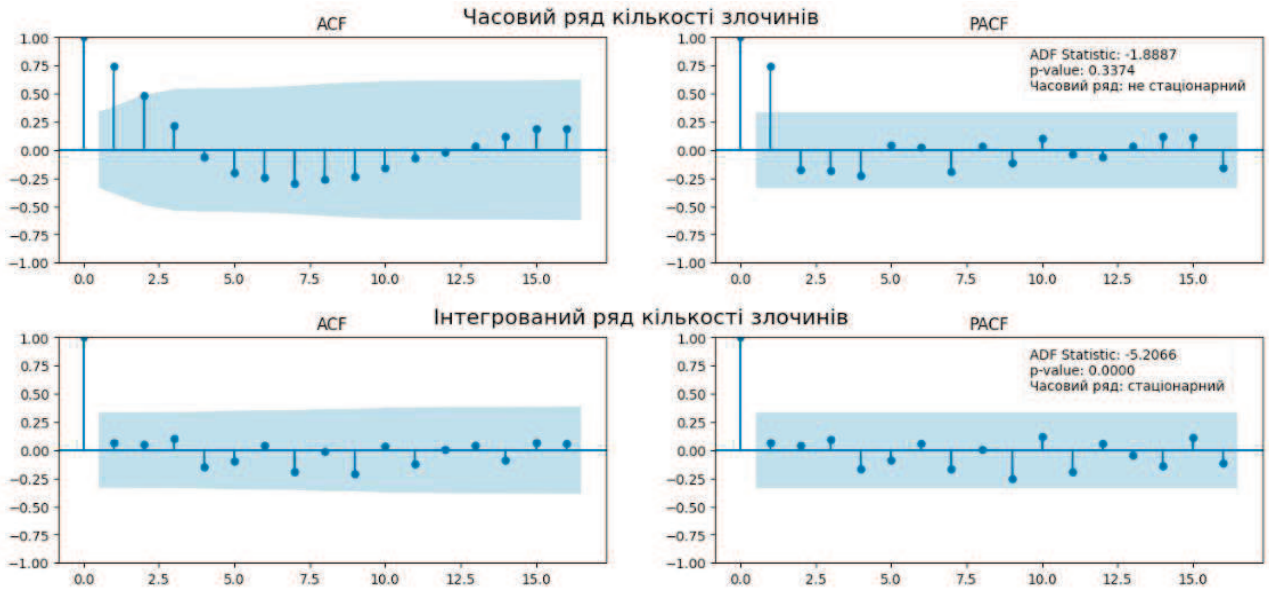

Рис. 2. Автокореляційна функція (ACF) та часткова автокореляційна функція (PACF) початкового та інтегрованого ряду кількості вчинених на території України злочинів / Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) of the original and integrated series of the number of committed crimes in Ukraine

Побудовано за допомогою функцій plot_acf та plot_pacf бібліотеки statsmodels.

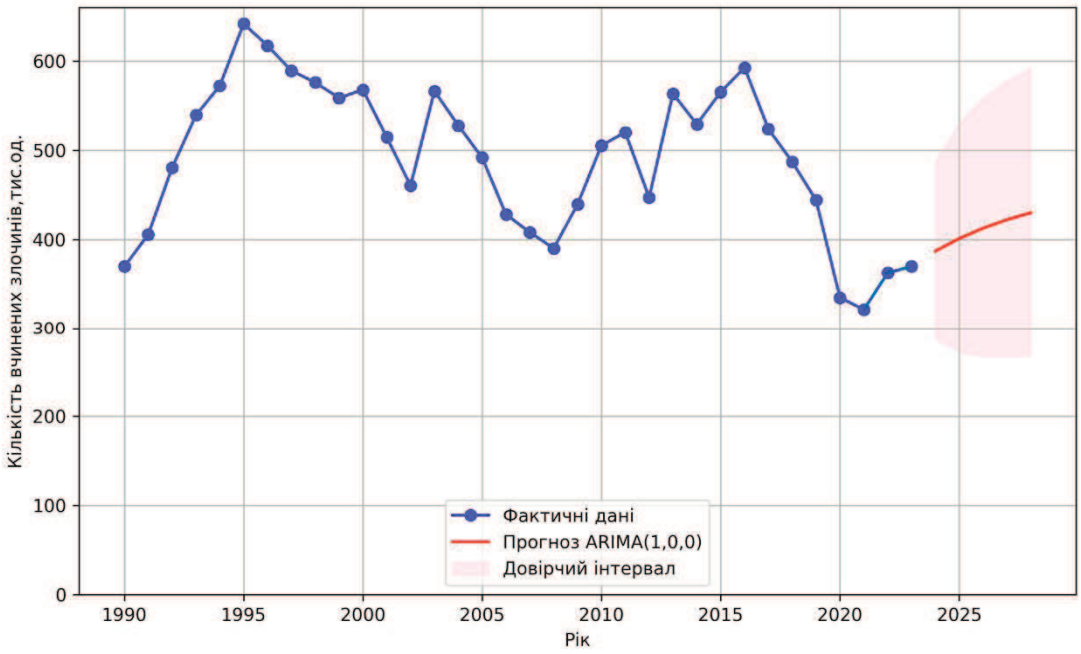

Рис. 3. Динаміка (1990–2023 рр.) та п'ятирічний прогноз (2024–2028 рр.) кількості вчинених на території України злочинів, побудований на основі моделі ARIMA(1, 0, 0) / Dynamics (1990–2023) and five-year forecast (2024–2028) of the number of committed crimes in Ukraine, based on the ARIMA(1, 0, 0) model

Однак для початкового часового ряду злочинності спостерігається висока автокореляція на першому лагу як для ACF, так і для PACF (рис. 2), що вказує на наявність структури в даних. Це дає змогу застосувати модель ARIMA для початкового ряду без попереднього інтегрування.

Побудова ARIMA моделі на основі часового ряду загальної кількості вчинених злочинів

Для автоматичного підбирання параметрів (*p, d, q*) ARIMA моделі використано функцію *auto_arima* з бібліотеки *pmdarima*. Ця функція здатна враховувати наявний тренд та автоматично вибирати оптимальні параметри для побудови моделі, що сприяє підвищенню точності прогнозування.

У результаті виконання функції *auto_arima* найкращою вибраною моделлю виявилася ARIMA (1, 0, 0). Тобто в цьому випадку використовується лише перший порядок авторегресії, і немає потреби в інтегруванні часового ряду, що підтверджує попередні висновки, зроблені на основі аналізу ACF та PACF (рис. 2). Основні параметри цієї моделі та прогноз на наступні п'ять років подано у табл. 1 та зображено графічно на рис. 3.

Як видно з наведених у табл. 1 даних, побудована модель виявилася адекватною. До того ж оцінка коефі-

цієнта авторегресії, на відміну від вільного члена, є статистично значущою та достатньо істотною. Спеціалізовані статистичні тести Льюнга – Бокса та Жарка – Бера вказують на відсутність істотної автокореляції у залишках моделі ($P(Q) = 0{,}46$) та підтверджують їхній нормальний розподіл ($P(JB) = 0{,}81$).

З метою порівняння різних методів прогнозування розглянуто також альтернативну модель на основі експоненційного згладжування. Для її побудови використано функцію *ExponentialSmoothing* з модуля *statsmodels.tsa.holtwinters* з автоматичним підбиранням оптимальних параметрів. У результаті розрахунків виявилося, що середня абсолютна процентна помилка цієї моделі (MAPE = 9,38 %) дещо вища за аналогічний показник попередньої моделі ARIMA (1, 0, 0) (MAPE = 8,61 %). А це означає, що модель ARIMA виявилася точнішою для прогнозування порівняно з моделлю експоненційного згладжування, про що зазначено також у інших роботах [7], [8], [9].

Покращення прогностичної валідності ARIMA моделі

Оскільки попереднє інтегрування даних призводило до повного зникнення автокореляції, як альтернативний підхід, здатний підвищити прогностичну валідність моделі, застосовано метод логарифмування. Цей метод дає змогу стабілізувати дисперсію, зберігши основну інформацію про структуру даних, яка втрачається під час інтегрування. Крім того, логарифмування допомагає також частково усунути нерівномірність коливань у часі, зменшити вплив значних викидів та виявити приховані закономірності.

Подальше використання функції *auto_arima* до логарифмованих значень часового ряду ($\ln y_t$) визначило модель ARIMA(2, 0, 1) як найкращу. Отже, запропонований підхід дав змогу підвищити загальний рівень структурованості даних, завдяки чому отримана модель вже враховує не одну, а дві лагові автокореляції, а також одну компоненту ковзного середнього. До того ж розрахунки (табл. 1) вказують й на високий рівень статистичної значущості оцінок обох коефіцієнтів авторегресії ϕ_1 та ϕ_2 ($P < |z| = 0{,}000$), на відміну від оцінок коефіцієнта ковзного середнього θ_1 ($P < |z| = 0{,}081$) та вільного члена α ($P < |z| = 0{,}198$). Зазначимо також, що зросла й точність побудованої моделі (MAPE = 7,47 %) порівняно з попередньою розглянутою ARIMA(1, 0, 0), середня абсолютна процентна помилка для якої становила 8,61%.

Водночас варто зазначити та врахувати у подальших розрахунках низьку статистичну значущість коефіцієнта ковзного середнього. Це, найімовірніше, пов'язано з раніше виявленими аномальними стрибками рівня злочинності у 2003, 2013 та 2020 роках. Щоб усунути їхній вплив на результати моделювання, показники за ці роки згладжено із заміною фактичних значень на середні значення попереднього та наступного періодів. Як виявилося, після такого коригування не потрібне загальне згладжування динамічного ряду методом ковзного середнього, оскільки тепер найоптимальнішою моделлю вже є ARIMA(2, 0, 0).

Статистичну оцінку параметрів цієї моделі та її основні характеристики наведено у табл. 1. Як бачимо, практично за усіма метриками ARIMA (2, 0, 0) демонструє найкращі результати: найвищу точність прогнозування (MAPE) та найнижчі значення критеріїв оптимальності (AIC (критерій Акаіке), BIC (критерій Байєса) та HQIC (критерій Ханнана – Квіна)). А це означає, що порівняно з іншими альтернативними моделями вона дає найточніші прогнози за мінімальної складності. Крім того, всі параметри цієї моделі, разом із вільним членом, виявилися статистично значущими.

Отже, найоптимальніша модель для прогнозування загальної кількості вчинених злочинів, з урахуванням попереднього згладжування аномальних рівнів та логарифмування даних, має такий вигляд:

$$\ln(y_t^*) = 1{,}4 + 1{,}26 \ln(y_{t-1}^*) - 0{,}487 \ln(y_{t-2}^*) + \varepsilon_t , \qquad (3)$$

де y_t^* – часовий ряд зі згладженими значеннями злочинності у 2003, 2013 та 2020 роках.

Крос-валідація моделі ARIMA(2, 0, 0)

Для перевірки стійкості та узагальнювальної здатності побудованої моделі здійснено також її крос-валідацію. Для цього, з огляду на обмежену кількість спостережень, застосовано метод перехресної валідації з вилученням одного спостереження (Leave-One-Out Cross-Validation (LOOCV)). Це різновид крос-валідації, в якому на кожній ітерації модель навчається на всіх спостереженнях, окрім одного, а потім на ньому її тестують. Отже, LOOCV використовує майже всі доступні дані для навчання, забезпечуючи стабільніші та надійніші оцінки моделі. Проте, з огляду на високу обчислювальну складність, її застосування виправдане переважно для малих вибірок, коли інші методи крос-валідації можуть бути менш ефективними або недоцільними.

Розраховані за допомогою функції LeaveOneOut (бібліотека scikit-learn) середні коефіцієнти авторегресійних компонентів ($\bar{\phi}_1 = 1{,}246$ та $\bar{\phi}_2 = -0{,}491$) практично збігаються із оцінками їхніх значень, отриманими під час побудови моделі ARIMA(2, 0, 0). Це свідчить про високий рівень стабільності моделі та відсутність надмірної адаптації до окремих спостережень, що підкреслює її узагальнювальну здатність. Отже, після коригування аномальних показників злочинності (2003, 2013 та 2020 рр.), інші коливання часового ряду критично не впливають на модель, що, своєю чергою, вказує й на відсутність значущих викидів або аномалій, які могли б істотно вплинути на результати моделювання.

Наостанок зазначимо, що прогнозні оцінки за всіма розглянутими моделями (табл. 1, рис. 3, рис. 4) є доволі невтішними та вказують на подальше зростання показників злочинності, яке розпочалося у 2021 р. після тривалого періоду зниження.

Обговорення результатів дослідження. З наведених результатів очевидно, що побудовані ARIMA-моделі, незважаючи на обмежений обсяг історичних даних, демонструють вищу точність прогнозування загального рівня злочинності порівняно з моделями експоненційного згладжування. Цей висновок узгоджується із дослідженнями інших авторів, які отримали аналогічні результати під час аналізу та розроблення прогнозних моделей рівня злочинності для окремих територіальних одиниць та різних типів злочинів [7], [8], [9]. Крім того, підвищення прогностичної валідності ARIMA-моделей можна досягти завдяки не лише налаштуванню її параметрів, але й попередньому обробленню часового ряду. До того ж таке оброблення має вихо-

дити за межі класичного методу диференціювання. Отже, звичайне використання функції *auto_arima* для автоматичного підбирання параметрів ARIMA-моделі не завжди є достатнім та оптимальним рішенням. Це, зокрема, підкреслюють й автори роботи [12].

Наукова новизна отриманих результатів дослідження – удосконалено ARIMA-модель для прогнозування загального рівня злочинності в Україні, яка відрізняється від інших застосуванням методу логарифмування початкових даних за згладжених аномальних рівнів, що підвищує її прогностичну валідність.

Практична значущість результатів дослідження – підвищено точність прогнозування загального рівня злочинності в Україні, що сприятиме ефективнішому плануванню превентивних заходів та ресурсів правоохоронних органів.

Табл. 1. Основні метрики моделей ARIMA та п'ятирічний прогноз загальної кількості злочинів, побудований на їхній основі / Key metrics of ARIMA models and the five-year forecast of total crime, based on these models

Часовий ряд	y_t		$\ln(y_t)$		$\ln(y_t^*)$	
Модель	ARIMA(1,0,0)		ARIMA(2,0,1)		ARIMA(2,0,0)	
Показник	coef. (std. err)	$P > \|z\|$	coef. (std. err)	$P > \|z\|$	coef. (std. err)	$P > \|z\|$
α	83,42 (73,47)	0,256	0,950 (0,737)	0,198	1,400 (0,589)	0,017
ϕ_1	0,821 (0,154)	0,000	1,568 (0,274)	0,000	1,260 (0,150)	0,000
ϕ_2	–	–	–0,722 (0,187)	0,000	–0,487 (0,171)	0,004
θ_1	–	–	–0,662 (0,380)	0,081	–	–
σ	2591,1 (694,7)	0,000	0,0091 (0,002)	0,000	0,0067 (0,002)	0,006
AIC	370,8		–48,25		–63,85	
BIC	375,4		–40,62		–57,74	
HQIC	372,4		–45,65		–61,76	
P(Q)	0,46		0,71		0,80	
P(JB)	0,81		0,66		0,44	
MAPE	8,61 %		7,47 %		7,04 %	
Рік	Точковий та інтервальний прогнози кількості злочинів, тис. од.					
2024	387,1 (287,3; 486.9)		411,4 (341,2; 495,9)		394,6 (335.9; 463,5)	
2025	401,2 (272,1; 530,3)		454,7 (353,3; 585,1)		424,1 (327,3; 549,2)	
2026	412,7 (267,2; 558,3)		492,6 (370,7; 654,5)		449,7 (328,7; 615,3)	
2027	422,2 (266,5; 577,9)		519,7 (386,6; 698,6)		467,7 (333,8; 655,2)	
2028	430,0 (267,9; 592,1)		533,4 (395,9; 718,5)		477,5 (338,4; 673,7)	

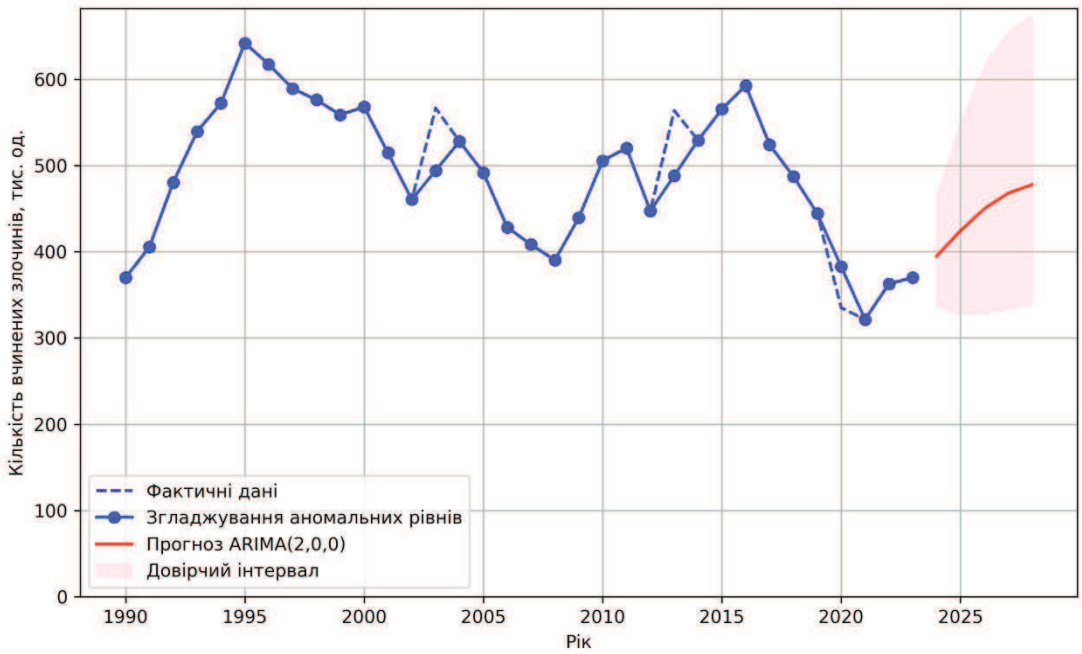

Рис. 4. Динаміка (1990–2023 рр.) та п'ятирічний прогноз (2024–2028 рр.) кількості вчинених на території України злочинів, побудований на основі моделі ARIMA(2,0,0) із використанням методу логарифмування та попереднього згладжування аномальних рівнів / Dynamics (1990–2023) and five-year forecast (2024–2028) of the number of committed crimes in Ukraine, based on the ARIMA(2, 0, 0) model, using logarithmic transformation and preliminary smoothing of anomalous levels

Висновки / Conclusions

Досліджуваний часовий ряд загальної кількості вчинених на території України злочинів є нестаціонарним, має ознаки спадного тренду та містить аномальні зміни у рівнях злочинності.

Побудовані результати виконаного аналізу моделі ARIMA ураховують одну (для початкового ряду) або дві (у разі логарифмування часового ряду) авторегресійні компоненти та забезпечують вищу точність прогнозування ($MAPE_1$ = 8,61 %, $MAPE_2$ = 7,47 %, $MAPE_3$ = 7,04 %) порівняно з моделлю експоненційного згладжування (MAPE = 9,38 %).

Логарифмування початкових даних та коригування аномальних показників злочинності підвищили прогностичну валідність і стійкість ARIMA моделі, що підтверджено інформаційними критеріями та методом перехресної валідації.

References

1. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control* (5th ed.). Hoboken, NJ, USA: John Wiley & Sons Inc. https://doi.org/10.1111/jtsa.12194
2. Dzendzeliuk, O., Kostiv, L., & Rabyk, V. (2013). Building ARIMA models of time series for weather data forecasting in R programming language. *Electronics and Information Technologies*, 3, 211–219. http://nbuv.gov.ua/UJRN/Telt_2013_3_24
3. Strilets, V. Ye., & Doroshenko, M. I. (2022). Analysis and forecasting of computer network characteristics. *Bulletin of V. N. Karazin Kharkiv National University. Mathematical Modeling. Information Technologies. Automated Control Systems*, 55, 49–57. https://doi.org/10.26565/2304–6201–2022–55–05
4. Marchuk, D. K., Kravchenko, S. M., Levchenko, A. Yu., & Lezhnov, I. Ya. (2023). Using time series for forecasting the monetary value of cars. *Scientific Notes of the V. I. Vernadsky Taurida National University. Series: Technical Sciences*, 34(73), 119–125.
5. Masliy, V. V., & Berezka, K. M. (2017). Selection and evaluation of ARIMA models for forecasting foreign direct investment. *Scientific Bulletin of the International Humanitarian University. Series: Economics and Management*, 24(2), 115–119. http://nbuv.gov.ua/UJRN/Nvmgu_eim_2017_24(2)__26
6. Dziubanovska, N. V., & Liashenko, O. M. (2018). Application of ARIMA models for forecasting the dynamics of Ukraine's foreign trade. *Black Sea Economic Studies*, 35(1), 142–147.
7. Islam, K., & Raza, A. (2020). Forecasting crime using ARIMA model. *arXiv*. http://arxiv.org/abs/2003.08006
8. Chen, P., Yuan, H., & Shu, X. (2008). Forecasting crime using the ARIMA model. In *Proceedings of the 5th International Conference on Fuzzy Systems and Knowledge Discovery* (pp. 627–630). https://doi.org/10.1109/FSKD.2008.222
9. Salati, L., & Majige, S. (2022). Forecasting criminal offenses against persons using time series models: A case study of Mwanza Region. *Asian Journal of Multidisciplinary Research & Review*, 3(2), 61–77. https://doi.org/10.55662/AJMRR.2022.3202
10. Lu, Y. (2023). Crime prediction utilizing ARIMA model. *BCP Business & Management*, 38, 410–418. https://doi.org/10.54691/bcpbm.v38i.3721
11. Vijayarani, S., Suganya, E., & Navya, C. (2021). Crime analysis and prediction using enhanced ARIMA model. *International Journal of Research Publication and Reviews*, 2, 257–266. https://www.ijrpr.com/uploads/V2ISSUE1/IJRPR153.pdf
12. Jain, H., & Patel, R. (2024). Analysis & forecasting of juvenile crime using variance threshold and time series algorithm. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042–024–19780–x
13. Triana, Y. S., & Retnowardhani, A. (2019). Enhance interval width of crime forecasting with ARIMA model-fuzzy alpha cut. *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, 17(3), 1193–1201. https://doi.org/10.12928/TELKOMNIKA.v17i3.12233
14. State Statistics Service of Ukraine (2024). Demographic and social statistics. Retrieved from https://www.ukrstat.gov.ua (accessed on 01.09.2024).
15. National Police of Ukraine (2024). Annual reports. Retrieved from https://www.npu.gov.ua/diyalnist/zvitnist/richni–zviti (accessed on 01.09.2024).

16. Said, S. E., & Dickey, D. A. (1984). Testing for unit roots in autoregressive-moving average models of unknown order. *Biometrika*, 71, 599–607. https://doi.org/10.1093/biomet/71.3.599
17. Mann, H. B. (1945). Non-parametric tests against trend. *Econometrica*, 13(3), 245–259. http://dx.doi.org/10.2307/1907187
18. Irwin, J. O. (1925). On a criterion for the rejection of outlying observations. *Biometrika*, 17(3–4), 238–250. https://doi.org/10.2307/2332079
19. Shevchuk, O. F. (2023). Statistical analysis of the dynamics of crimes committed in Ukraine in 1990–2020. *Current Issues in Modern Science*, 2(8), 268–279. https://doi.org/10.52058/2786–6300–2023–2(8)–268–27

A. A. Yarovyi, O. F. Shevchuk, A. V. Kozlovskyi, Yu. M. Panochyshyn, S. V. Simonchuk
Vinnytsia National Technical University, Vinnytsia, Ukraine

USING ARIMA MODELS FOR FORECASTING OF OVERALL CRIME RATE IN UKRAINE

Crime rate forecasting is a critical element in the development of strategies for sustainable socio-economic growth in a rule-of-law state. Accurate forecasting becomes particularly important in times of economic instability and geopolitical crises, as is the case in Ukraine. This article explores the problem of constructing and applying autoregressive integrated moving average (ARIMA) models to predict the total number of crimes committed in Ukraine. The statistical analysis of the crime time series was conducted using the Python programming language, utilizing specialized libraries such as numpy, pandas, matplotlib, statsmodels, pmdarima, and scikit-learn. The calculations indicate that the crime time series (1990–2023) demonstrates a declining trend, is non-stationary, and contains anomalous values in crime rates in 2003, 2013, and 2020, correlating with socio-political crises in Ukraine. Specifically, the anomalous increases in crime rates (in 2003 and 2013) align with heightened public unrest preceding the Orange Revolution (2004–2005) and the Revolution of Dignity (2013–2014). In contrast, the unusually low crime rates observed in 2020 are attributed to restrictive quarantine measures implemented due to the COVID–19 pandemic. The use of data integration by taking the first differences between observations resulted in the loss of autocorrelation structure inherent in the overall crime series. Consequently, the initial ARIMA (1, 0, 0) model was built based on the untransformed input data. The accuracy of this model was higher compared (MAPE = 8.61 %) to the model obtained using the exponential smoothing method (MAPE = 9.38 %). Logarithmic transformation of the crime time series and smoothing of anomalous levels enhanced the predictive validity, allowing the ARIMA model to account for additional autocorrelation while avoiding the need for a moving average component. As a result, the ARIMA (2, 0, 0) model demonstrated the highest accuracy (MAPE = 7.04 %) with minimal complexity, as confirmed by information criteria results. Furthermore, the model successfully passed robustness testing using the cross-validation method with the exclusion of a single observation. The forecasted estimates, derived from all the examined ARIMA models, indicate a continued increase in the overall crime rate in Ukraine, which began in 2021 following a prolonged period of decline.

Keywords: modeling, autocorrelation, anomalous values, cross–validation, crime count.

Інформація про авторів:
Яровий Андрій Анатолійович, д-р техн. наук, професор, завідувач кафедри комп'ютерних наук.
Email: a.yarovyy@vntu.edu.ua; https://orcid.org/0000–0002–6668–2425
Шевчук Олександр Федорович, канд. фіз.-мат. наук, доцент, доцент кафедри комп'ютерних наук.
Email: shevchuk177@gmail.com; https://orcid.org/0000–0002–8600–0700
Козловський Андрій Володимирович, канд. техн. наук, доцент, доцент кафедри комп'ютерних наук.
Email: akozlovskyi@vntu.edu.ua; https://orcid.org/ 0000–0001–9697–1511
Паночишин Юрій Миколайович, канд. техн. наук, доцент, доцент кафедри комп'ютерних наук.
Email: y.panochyshyn@vntu.edu.ua; https://orcid.org/0000–0003–1546–3422
Сімончук Сергій Володимирович, асистент кафедри комп'ютерних наук. **Email**: sergii.simonchuk@vntu.edu.ua; https://orcid.org/0009–0000–7295–5357

Цитування за ДСТУ: Яровий А. А., Шевчук О. Ф., Козловський А. В., Паночишин Ю. М., Сімончук С. В. Використання ARIMA моделей для прогнозування загального рівня злочинності в Україні. *Український журнал інформаційних технологій*. 2024, т. 6(2), С. 49–56.
Citation APA: Yarovyi, A. A., Shevchuk, O. F., Kozlovskyi, A. V., Panochyshyn, Yu. M., & Simonchuk, S. V. (2024). Using ARIMA models for forecasting of overall crime rate in Ukraine. *Ukrainian Journal of Information Technology,* 6(2), 49–56. https://doi.org/10.23939/ujit2024.02.049