

УДК 004.62

DOI: 10.31673/2786-8362.2025.011761

Беспала О.М.; Тимкова А.В.

ОПТИМІЗАЦІЯ ПРОГНОЗУВАННЯ ШЛЯХОМ УРАХУВАННЯ ПРИЧИННОГО ВПЛИВУ

Bespala O.M., Tymkova A.V. Optimization of forecasting through incorporation of causal influence. The article explores approaches to improving forecasting accuracy by incorporating causal influence between variables with time delays. The study uses Granger causality to detect temporal lags and adjust input variables accordingly. A method is proposed to enhance LSTM-model performance by integrating lag analysis into the training pipeline. Experimental evaluation shows that accounting for different lag times across variables significantly reduces forecasting error. The results confirm that causal-lag modeling improves robustness and reflects real-world dynamics in time series forecasting. The authors emphasize the importance of considering lagged interrelationships to form a more informative input space for the neural network. Experimental results indicate an improvement in forecasting accuracy when causally significant relationships are identified in advance.

Keywords: forecasting, time series, Granger causality, lag effect, prediction model, neural network

Беспала О.М., Тимкова А.В. Оптимізація прогнозування шляхом урахування причинного впливу. У статті досліджено підходи до врахування причинного впливу між змінними у задачах прогнозування. Основна увага приділяється методу причинності Грейнджера для виявлення часових затримок між змінними та їх подальшого застосування в побудові прогнозних моделей. Запропоновано підхід до покращення точності LSTM-моделі шляхом інтеграції аналізу часових затримок. Автори наголошують на важливості врахування лагових взаємозв'язків для формування більш інформативного вхідного простору нейромережі. Експериментальні результати свідчать про підвищення точності прогнозів за умови попереднього виявлення причинно-значущих зв'язків.

Ключові слова: прогнозування, часові ряди, причинність Грейнджера, лаговий вплив, нейронні мережі

Вступ

Одним із головних факторів вивчення та моніторингу складних систем є ефективний аналіз на основі даних спостережень. Багато таких інформаційних систем включають часові дані. Наприклад, на основі часових даних фінансові інформаційні системи (ІС) проводять аналіз фінансових ринків, прогнозують ціни акцій, валют і товарів, проводять обробку історичних даних фондового ринку [1]; енергетичні ІС, в тому числі і в режимі реального часу, здійснюють моніторинг та управління електромережами, аналізують їх навантаження та прогнозування споживання електроенергії, а IoT-платформи для "розумних" лічильників відстежують споживання електроенергії, води, газу [2, 3]; метеорологічні системи збирають і обробляють супутникові дані великих розмірностей [4, 5]; медичні ІС здійснюють моніторинг пацієнтів у реальному часі та аналіз серцевого ритму, тиску, температури, а з використанням бази даних епідеміологічного моніторингу проводиться аналіз і прогноз поширення хвороб; в області соціальних мереж та веб-аналітика особлива потреба зростає в аналізі динаміки пошукових запитів, переглядів та поведінки користувачів, а також популярності трендів та новин [6] тощо.

Оскільки спостерігається збільшення об'єму зібраних даних, в тому числі і часових, піднімається проблема вдосконалення методів використання часової інформації. У свою чергу, для більш ефективного аналізу системи виникає необхідність розуміння взаємодії між її компонентами та їхнього впливу як між собою, так і на загальну поведінку системи. Це критично важливо для прийняття рішень, оптимізації роботи системи, прогнозування її розвитку та управління ризиками.

Постановка завдання. Завдання полягає у вдосконаленні прогнозування за допомогою врахування часових лагів між змінними, що впливають на цільову функцію. Використовується метод причинності Грейнджера для виявлення та визначення цих лагів, що дозволяє з'ясувати,

як зміни в одних змінних впливають на інші в певний момент часу. Після виявлення лагів, їх слід застосувати для зміщення та коригування змінних з метою підвищення точності прогнозу.

Аналіз останніх досліджень. В огляді літератури представлено різні методи моделювання причинно-наслідкових зв'язків [7], які можна поділити на методи, що ґрунтуються на знаннях та на основі даних [8, 9]. Методи, що ґрунтуються на побудові моделі на основі знань, можуть мати перевагу в точності, проте вимагають глибокого розуміння процесу, заснованих на досвіді та фізичних моделях систем [10]. Оскільки сучасні системні архітектури стають складнішими, а кількість компонентів значно збільшується, процес формулювання точної моделі на основі досвіду та знань предметної області стає дуже складним, що означає, що більшість методів, заснованих на знаннях, мають деякі обмеження при застосуванні у великому масштабі. Саме тому, використання доступних даних для побудови моделей шляхом аналізу причинно-наслідкових зв'язків між змінними процесу стає більш актуальною задачею [11, 12], яка потребує системного підходу з можливістю працювати з часовими даними та обробляти одночасно великі об'єми інформації з часовою структурою.

Метою роботи є вдосконалення методу прогнозування за рахунок застосування методу Грейнджера для виявлення часових лагів між змінними, що впливають на цільову функцію. Задача полягає в тому, щоб знайти ці лаги, скоригувати змінні на їх основі та покращити точність прогностичної моделі, враховуючи причинно-наслідкові зв'язки між змінними.

Виклад основного матеріалу дослідження

У задачах прогнозування в контексті машинного навчання мета полягає у створенні моделі, яка здатна передбачити значення цільової змінної на основі наданих вхідних даних. Задача прогнозування є типовою для багатьох дисциплін, таких як економіка, фінанси, медицина, а також природничі науки, де часто необхідно передбачити значення, які залежать від багатьох змінних.

Нехай у нас є набір спостережень, де кожне спостереження складається з вектору вхідних змінних $x^i \in \mathbb{R}^m$, що представляють ознаки, і відповідного значення цільової змінної $y^i \in \mathbb{R}$. Таким чином, ми маємо набір даних:

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}, \quad (1)$$

де n – кількість спостережень. Задача прогнозування полягає у знаходженні функції $f(x, w)$, яка мапує вхідні дані x на прогнозоване значення y , де w – вектор параметрів, які визначають модель.

Математично це можна записати як:

$$\hat{y} = f(x, w), \quad (2)$$

де $f(x, w)$ – функція, яка залежить від вхідного вектору x та параметрів w , w – вектор параметрів, які потрібно навчити (ваги).

При цьому важливо, щоб ця функція була якомога точнішою для всіх спостережень, що відображає залежність між змінними. Для знаходження оптимальних значень параметрів w необхідно мінімізувати функцію втрат, яка вимірює точність моделі. Одним із найбільш використовуваних підходів є мінімізація середньоквадратичної помилки (MSE), яка визначається як середнє значення квадратів відмінностей між фактичними значеннями y_i та прогнозами $\hat{y}_i$:

$$L(w) = \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i, w))^2 \quad (3)$$

Функція втрат виражає відхилення прогнозів від реальних значень і служить основою для оцінки ефективності моделі. Метою оптимізації є знаходження таких значень параметрів w^* , які мінімізують функцію втрат:

$$w^* = \arg \min_w L(w) \quad (4)$$

Наприклад модель знайшла ваги, але відбулась затримка впливу на цільову змінну від одного з параметрів x . Це може значно погіршити точність прогнозування моделі. У такому випадку потрібно додатково відслідковувати зміщення в часі, та коригувати зміщення ваг w .

Задача прогнозування в машинному навчанні є надзвичайно важливою для численних застосувань, але у випадку, коли вплив вхідних змінних (наприклад, витрат на рекламу) на цільову змінну (наприклад, обсяг продажів) має різну віддаленість у часі (лаг), традиційні моделі, як лінійна регресія або інші моделі без врахування часу, можуть погано працювати. Це призводить до змінених результатів прогнозування та погіршення точності моделі.

Проблема різної віддаленості впливу змінних. У багатьох реальних задачах, як, наприклад, прогнозування обсягів продажів на основі витрат на рекламу, різні рекламні кампанії можуть мати різні часові лаги впливу. Це означає, що не всі змінні впливають на цільову змінну негайно, і їхній ефект може бути відстрочений на певний період часу. Якщо є кілька змінних, що мають різний час впливу на результат, то математична модель без врахування лагів може виглядати так:

$$y_i = f(x_{i,1}, x_{i,2}, \dots, x_{i,m}, w) + \varepsilon_i \quad (5)$$

де y_i – це цільова змінна (наприклад, обсяг продажів), $x_{i,j}$ – це вхідні змінні, що представляють різні фактори (наприклад, витрати на рекламу в поточному місяці, витрати на рекламу в попередньому місяці), w – вектор параметрів моделі, а ε_i – це похибка.

Змішування різних часових лагів в одній моделі може привести до перехресних кореляцій між змінними, що може погіршити здатність моделі правильно оцінювати залежність між змінними.

Вплив різних лагів на функцію втрат. Коли ми маємо кілька змінних з різними лагами, традиційна модель може мати змогу адаптуватися лише до певного виду впливу. Для кожного лагу потрібно розглядати окрему залежність, що можна записати як:

$$y_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \dots + w_m x_{i,m} + \varepsilon_i, \quad (6)$$

де кожен $x_{i,j}$ – це значення змінної, враховуючи її лаг. Традиційні методи оптимізації, які враховують всі змінні однаково, можуть привести до зміщення в оцінці параметрів w , якщо лаги мають різні часи впливу.

Функція втрат (наприклад, середньоквадратична помилка) в даному випадку виглядатиме так:

$$L(w) = \frac{1}{n} \sum_{i=1}^n \left(y_i - (w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \dots + w_m x_{i,m}) \right)^2 \quad (7)$$

Проте, якщо зміщення в часі (лаги) не враховані коректно, то точність прогнозу зменшується, оскільки різні лаги не будуть коректно враховані в моделі.

Метод Грейнджера для визначення лагів причинного впливу. Одним із способів вирішення цієї проблеми є використання методу причинності Грейнджера для виявлення і коригування лагів в часі між вхідними змінними та цільовою змінною [13]. Метод причинності Грейнджера дозволяє встановити, чи може змінна x_t передбачати значення змінної y_t в майбутньому на основі попередніх значень цих змінних.

Тест на причинність Грейнджера визначається як перевірка на наявність статистично значущої кореляції між ланцюговими значеннями двох часових рядів. Для двох змінних x_t та y_t тест Грейнджера виглядає так:

$$y_t = \alpha + \sum_{i=1}^p \beta_i y_{t-i} + \sum_{j=1}^q \gamma_j x_{t-j} + \varepsilon_t, \quad (8)$$

де p і q – це максимальні лаги для y і x відповідно, α – константа, β_i та γ_j – коефіцієнти, що визначають вплив минулих значень y і x на поточне значення y_t .

Якщо $\gamma_j \neq 0$, то змінна x_t має причинний вплив на y_t в часі з лагом j .

Якщо $\gamma_i = 0$ для всіх j , то змінна x_i не має причинного впливу на y_i .

Інтеграція цієї ідеї в модель прогнозування може відбуватись шляхом введення множини лагів для кожної змінної x_i . Таким чином, ми можемо створити модель, яка враховує різні лаги для різних змінних. Математична форма цієї моделі виглядатиме так:

$$y_t = w_0 + \sum_{i=1}^p w_i x_{t-i} + \varepsilon_t, \quad (9)$$

де x_{t-i} – це значення змінних на попередніх кроках часу (лаги), а w_i – ваги, які визначають вплив цих значень на прогноз.

У межах тестування запропонованого підходу було використано архітектуру спеціального типу нейронної мережі з довготривалою короткочасною пам'яттю (LSTM, Long Short-Term Memory), яка є однією з найефективніших моделей для обробки послідовних даних та часових рядів. Початкове навчання моделі проводилось на синтетичних даних, де всі змінні мали однакову часову віддаленість впливу на цільову змінну. У такому випадку модель демонструвала високу точність прогнозування, що підтверджувалося низьким значенням функції втрат (наприклад, середньоквадратичної помилки MSE).

Після цього було згенеровано новий набір даних, де ступінь часової віддаленості (лагу) між змінними та цільовою змінною був неоднаковим. Зокрема, деякі вхідні змінні починали впливати на результат із затримкою у кілька кроків, у той час як інші – миттєво. У цьому сценарії якість прогнозування значно погіршилася, що проявилось у збільшенні функції втрат $L(w)$, незважаючи на наявність тієї ж самої архітектури LSTM. Це свідчить про те, що базова модель не враховує асиметрію лагових впливів між змінними.

Щоб вирішити цю проблему, до моделі було інтегровано модуль попереднього аналізу лагових впливів, який базується на методі причинності Грейнджера. Цей модуль дозволяє автоматично виявляти, з яким лагом кожна змінна x_j має причинний вплив на цільову змінну y , шляхом статистичного тестування гіпотез для кожної пари змінна-цільова:

H0: x_j не причинно впливає на y з лагом l

H1: x_j причинно впливає на y з лагом l

Для цього було побудовано множину лагових моделей:

$$y_t = \sum_{j=1}^m \sum_{l=1}^L a_{j,l} x_{j,t-l} + \varepsilon_t, \quad (10)$$

та проведено тестування значущості коефіцієнтів $a_{j,l}$. Якщо певна змінна x_j виявилася значущою з лагом l , то вона включалась у модель LSTM як окремий часовий зріз із відповідним зсувом.

Для тестування запропонованого підходу було здійснено генерацію даних:

$x(t)$ – витрати на рекламу в день t ,

$u(t)$ – кількість відвідувачів у день t ,

вплив реклами має затримку (лаг) на певний період, тобто $x(t)$ впливає на $u(t+1)$.

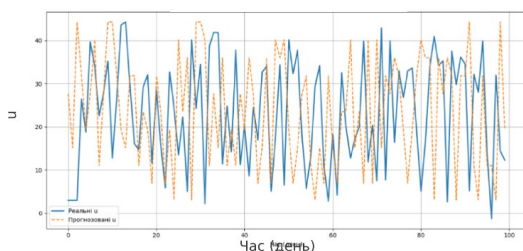


Рис. 1. Прогнозована модель 1 без врахування лагів

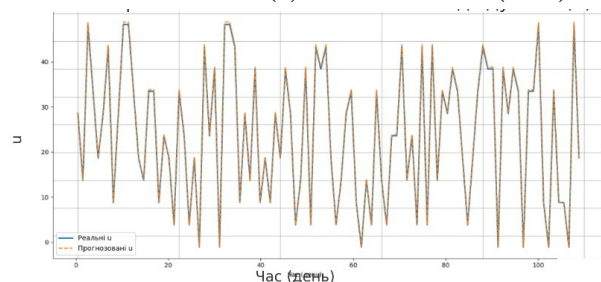


Рис. 2. Прогнозована модель 2 з урахуванням лагів

Після інтеграції модуля аналізу лагів модель була повторно протестована на тому ж наборі даних із затриманими впливами. У результаті було досягнуто суттєвого покращення якості прогнозу, що проявилось у зниженні функції втрат $L(w^*)$ у порівнянні з початковою LSTM-моделлю. На рисунку 1, 2 показано результати тестування моделей. В таблиці 1 продемонстровано відповідні показниками оцінки якості прогнозованих моделей.

Таблиця 1

Результатів оцінки якості прогнозованих моделей

	MAE $Y = x(t)$	MAE $Y = x(t+i)$
Модель 1	0.1908	15.84
Модель 2	0.1901	0.1900

Таким чином, включення попереднього лагового аналізу дозволяє адаптувати модель до структурної часової неоднорідності вхідних змінних і підвищує її робастність до зміщення впливу.

Висновки

Використання методу Грейнджера для виявлення часових лагів дозволяє більш точно моделювати вплив змінних на цільову функцію, що значно підвищує прогностичну здатність моделей. Зміщення змінних із врахуванням виявлених лагів дає змогу знизити помилки прогнозування, оскільки враховуються затримки в їх взаємодії. Причинно-наслідковий підхід, поєднаний з аналізом лагів, дозволяє точніше оцінювати динаміку змін і здійснювати прогнози, що відображають реальні процеси. Покращення точності прогнозування за рахунок виявлення лагів є важливим кроком у вдосконаленні моделей, особливо в задачах, де часові аспекти мають критичне значення для результатів.

Список використаної літератури:

1. Habibnia A., Etesami J., Kiyavash N. Modeling Systemic Risk: A Time-Varying Nonparametric Causal Inference Framework. SSRN. 2024. URL: <https://doi.org/10.2139/ssrn.4684230>.
2. Research on condition operation monitoring of power system based on supervisory control and data acquisition model / B. Li et al. *Alexandria Engineering Journal*. 2024. Vol. 99. P. 326–334. URL: <https://doi.org/10.1016/j.aej.2024.05.027>.
3. Open-Source Internet of Things-Based Supervisory Control and Data Acquisition System for Photovoltaic Monitoring and Control Using HTTP and TCP/IP Protocols / W. Khalid et al. *Energies*. 2024. Vol. 17, no. 16. P. 4083. URL: <https://doi.org/10.3390/en17164083>.
4. Lopes F. M., Dutra E., Boussetta S. Evaluation of Daily Temperature Extremes in the ECMWF Operational Weather Forecasts and ERA5 Reanalysis. *Atmosphere*. 2024. Vol. 15, no. 1. P. 93. URL: <https://doi.org/10.3390/atmos15010093>.
5. Siddiqi A. Value Analytics for Earth Observing Systems. *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, Athens, Greece, 7–12 July 2024. 2024. P. 3864–3867. URL: <https://doi.org/10.1109/igarss53475.2024.10642411>.
6. Liao S.-h., Widowati R., Lee C.-Y. Data mining analytics investigation on TikTok users' behaviors: social media app development. *Library Hi Tech*. 2022. URL: <https://doi.org/10.1108/lht-08-2022-0368>.
7. Chiang L. H., Braatz R. D. Process monitoring using causal map and multivariate statistics: fault detection and identification. *Chemometrics and Intelligent Laboratory Systems*. 2003. Vol. 65, no. 2. P. 159–178. URL: [https://doi.org/10.1016/s0169-7439\(02\)00140-5](https://doi.org/10.1016/s0169-7439(02)00140-5).
8. Nadim K., Ragab A., Ouali M.-S. Data-driven dynamic causality analysis of industrial systems using interpretable machine learning and process mining. *Journal of Intelligent Manufacturing*. 2022. URL: <https://doi.org/10.1007/s10845-021-01903-y>.

9. Беспала О.М. Інструментарій причинно-наслідкового висновку: огляд та перспективи. *Control Systems and Computers*. 2020. № 5 (289). С. 52–63. URL: <https://doi.org/10.15407/csc.2020.05.052>.
10. Deep understanding in industrial processes by complementing human expertise with interpretable patterns of machine learning / A. Ragab et al. *Expert Systems with Applications*. 2019. Vol. 122. P. 388–405. URL: <https://doi.org/10.1016/j.eswa.2019.01.011>.
11. Беспала О. М., Отрох С. І., Ружинський В. Г.. Моделювання спрямованого ациклічного графа для причинного висновку. *Наукові записки Державного університету телекомунікацій*. 2023. № 2 (1) С. 87-95 DOI: 10.31673/2518-7678.2023.020202.
12. Беспала О. М. Метод пошуку та оцінки впливу причинно-наслідкових зв'язків в системах прийняття рішень. *Вісник НТУ "ХПІ". Серія: Інформатика та моделювання*. 2020. № 2 (4). С. 59 – 72.
13. Granger C. W. J. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*. 1969. Vol. 37, no. 3. P. 424. URL: <https://doi.org/10.2307/1912791>.

Автори статті

Беспала Ольга – старший викладач, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.

ORCID: 0000-0003-3285-2585

Тимкова Анастасія – студент, Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Київ, Україна.

ORCID: 0009-0000-7619-3281

Authors of the article

Bespala Olha – senior lecturer, National Technical University of Ukraine “Ihor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine.

ORCID: 0000-0003-3285-2585

Tymkova Anastasiia – student, National Technical University of Ukraine “Ihor Sikorsky Kyiv Polytechnic Institute”, Kyiv, Ukraine.

ORCID: 0009-0000-7619-3281