УДК 004.9+574

В. Є. Копняк[1]
В. Б. Мокін[1]

ДОСЛІДЖЕННЯ ВПЛИВУ АНОМАЛЬНИХ ЗНАЧЕНЬ НА ТОЧНІСТЬ БУСТИНГОВИХ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ГЕТЕРОСКЕДАСТИЧНИХ ЧАСОВИХ РЯДІВ

[1]Вінницький національний технічний університет

Анотація

Досліджено задачу оцінювання впливу аномальних значень на точність прогнозування гетероскедастичних часових рядів бустинговими EGARCH-моделями з LSTM і без неї. Запропоновано підхід до локальної корекції аномального спостереження шляхом його заміни на сусіднє значення ряду перед навчанням моделей. На прикладі аналізу процесу поширення пилу Сахари в атмосферному повітрі за даними Української мережі громадського моніторингу якості повітря «Eco City» показано, що усунення одного екстремального значення дозволило підвищити коефіцієнт детермінації (R^2) прогнозу з 0,97 до 0,99, що свідчить про високу чутливість моделі до аномальних спостережень та доцільність їх попереднього оброблення.

Ключові слова: прогнозування часового ряду, гетероскедастичний процес, аномальні дані, бустинг, громадський моніторинг якості атмосферного повітря, місто Вінниця, EcoCity.

Abstract

The problem of assessing the impact of anomalous values on the forecasting accuracy of heteroscedastic time series using boosting EGARCH models with and without LSTM is investigated. An approach for the local correction of an anomalous observation by replacing it with a neighboring value in the series prior to model training is proposed. Using the analysis of the Sahara dust transport process in atmospheric air based on data from the Ukrainian public air quality monitoring network Eco City, it is shown that removing a single extreme value increased the coefficient of determination (R^2) of the forecast from 0.97 to 0.99. This result indicates the high sensitivity of the model to anomalous observations and confirms the feasibility of their preliminary preprocessing.

Keywords: time series forecasting, heteroscedastic process, anomalous data, boosting, public air quality monitoring, Vinnytsia city, Eco City.

Вступ

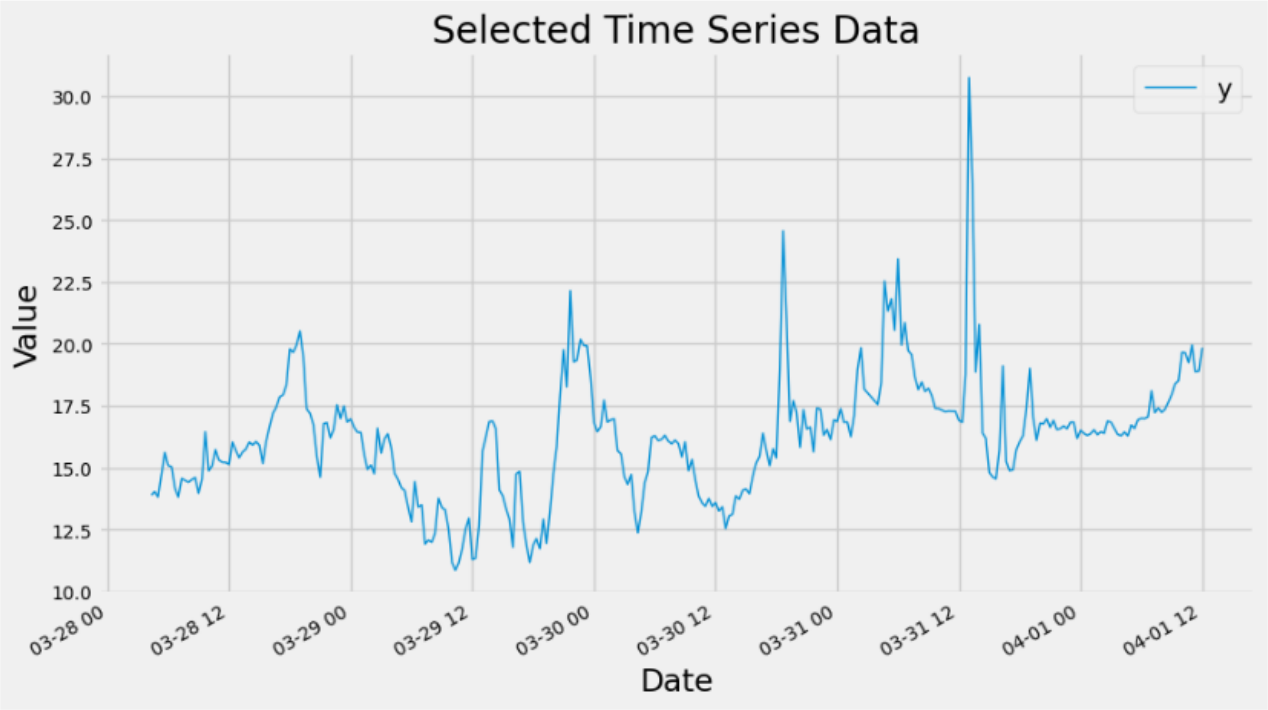
У роботі [1] авторами було запропоновано метод бустингу гетероскедастичних моделей та його наведено його успішне випробування на прикладі прогнозування концентрацій пилу Сахари в атмосферному повітрі м. Вінниці за даними громадського моніторингу мережі «Eco City» [2, 3]. Найкраща модель забезпечила точність 0.97 за метрикою R^2. Але на наведених у тій статті рисунках видно, що є явна аномалія – найбільше значення змінилось так стрімко, що модель не може його передбачити. Цікавим є якою було б значення метрики R^2 за цим методом у разі, якщо цю аномалію прибрати.

Мета дослідження – дослідити вплив усунення аномальних значень на точність бустингової моделі прогнозування гетероскедастичних часових рядів на прикладі даних Української мережі громадського моніторингу якості повітря «Eco City».

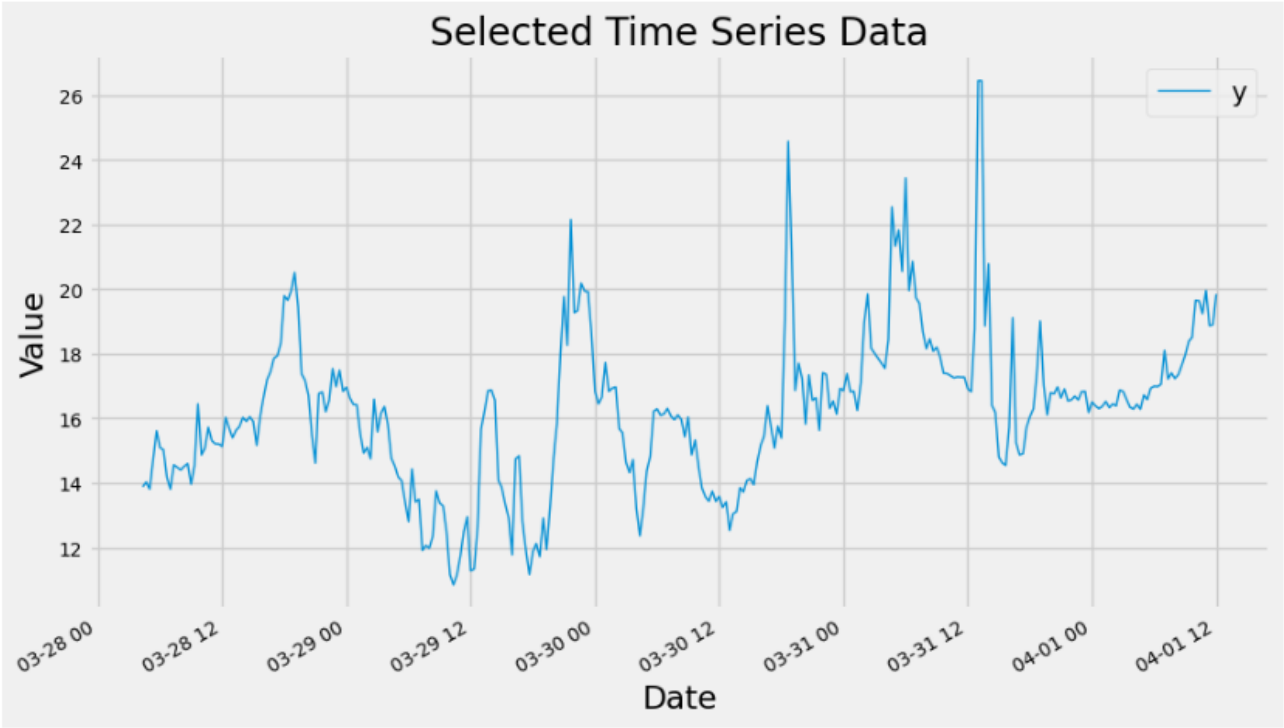
Результати дослідження

Як відомо, якісне передоброблення часто суттєво підвищує точність прогнозування часових рядів [4], особливо ефективним є видалення аномальних даних. У роботі [5] наведено багато прийомів і технологій виявлення та фільтрування аномалій. Однак, їх варто застосовувати з обережністю. Для задач моделювання гетероскедастичних процесів аномалії є дуже важливими, оскільки саме їх

наявність дозволяє виявити і діагностувати гетероскедастичність процесу. Тому пропонуємо прибрати тільки одне найбільше значення. Експерименти показали, що, якщо його просто видалити, замінивши на NaN, процес перестає бути гетероскедастичним – краще його на щось замінити. Якщо замінити на середнє, то це зіпсує його динаміку. Пропонуємо замінити на наступне, теж доволі велике, значення ряду (рис. 1).



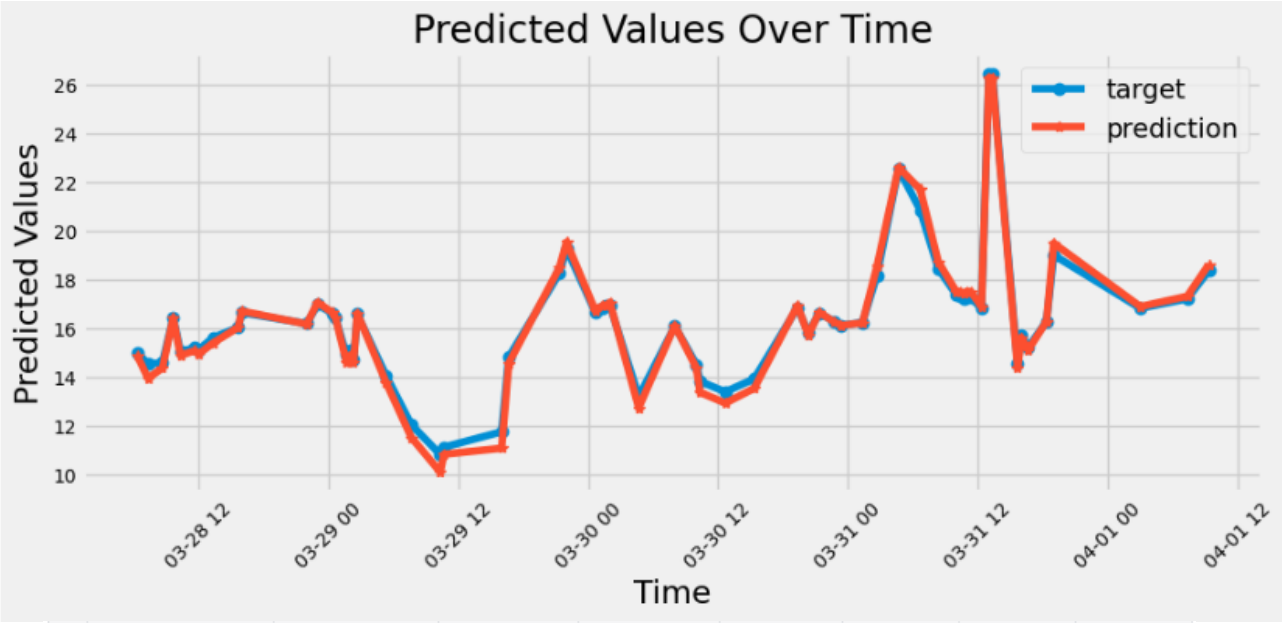
а)



б)

Рис. 1. Часовий ряд концентрації пилу «PM1» за даними Української мережі громадського моніторингу якості повітря «Eco City»: а) до передоброблення; б) після заміни найбільшого значення, яке мало місце о 13.00 31.03.2024 р., на наступне

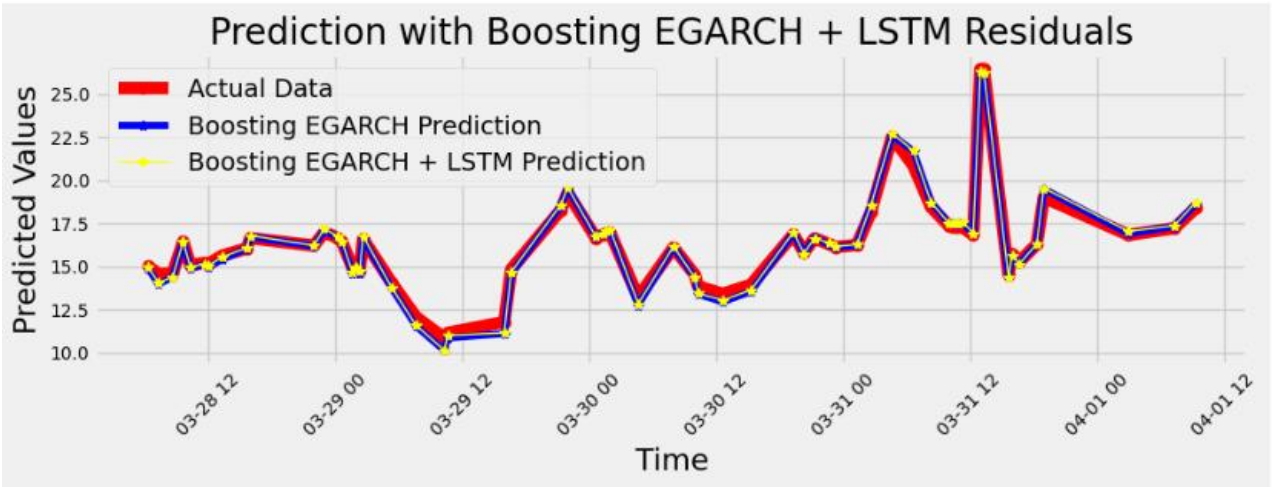
Була застосовано алгоритм зі статті [1] та відповідний Kaggle-ноутбук і побудована бустингова модель гетероскедастична EGARCH-модель. Результат прогнозування за нею наведено на рис. 2.



	name_model	opt_params_set	n_estimators	learning_rate	max_depth	rmse	r2_score	mape
0	Boosting EGARCH	r2_score	30	0.038	6	0.273818	0.989969	0.013163

Рисунок 2. Результат прогнозування за бустинговою гетероскедастичною моделлю з передобробленням найбільшої аномалії

А потім була застосована ще й LSTM-модель для оброблення залишків за алгоритмом із роботи [6], що дозволило ще підвищити точність (рис. 3).



а)

name_model	opt_params_set	n_estimators	learning_rate	max_depth	rmse	r2_score	mape
Boosting EGARCH + LSTM	r2_score	30	0.038	6	0.269038	0.990316	0.013109
Boosting EGARCH	r2_score	30	0.038	6	0.273818	0.989969	0.013163

б)

name_model	opt_params_set	n_estimators	learning_rate	max_depth	rmse	r2_score	mape
Boosting EGARCH + LSTM	r2_score	30	0.038	6	0.519675	0.970547	0.01562
Boosting EGARCH	r2_score	30	0.038	6	0.525981	0.969828	0.015924

в)

Рисунок 3. Результат прогнозування за бустинговою гетероскедастичною EGARCH моделлю з передобробленням найбільшої аномалії та обробленням залишків за допомогою LSTM: а) графіки, б) метрики моделей з передобробленням аномалії; в) метрики тих же моделей без передоброблення з роботи [7]

Як видно на рис. 3, передоброблення аномалії бустингової EGARCH-модель дозволяє підвищити коефіцієнт детермінації (R^2) прогнозу з 0,97 до 0,99, що свідчить про високу чутливість моделі до аномальних спостережень та доцільність їх попереднього оброблення.

Висновки

Робота присвячена дослідженню впливу передоброблення аномальних значень на точність прогнозування гетероскедастичних часових рядів бустинговою моделлю з використанням моделі LSTM для оброблення залишків і без неї. Запропоновано підхід до локальної корекції аномального спостереження шляхом його заміни на сусіднє значення ряду перед навчанням моделі. На прикладі аналізу процесу поширення пилу Сахари в атмосферному повітрі за даними Української мережі громадського моніторингу якості повітря «Eco City» показано, що усунення одного екстремального значення дозволило покращити усі метрики, наприклад коефіцієнт детермінації (R^2) прогнозу з використанням моделі EGARCH з LSTM зріс з 0,97 до 0,99, що свідчить про високу чутливість моделі до аномальних спостережень та доцільність їх попереднього оброблення.

Отримані у такий спосіб результати аналізу дозволять підвищити точність прогнозування гетероскедастичних процесів зі значними аномаліями.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Копняк В. Є., Мокін В. Б., Жуков С. О., Варчук І. В., Скринник Т. В., Метод бустингу гетероскедастичних моделей для прогнозування концентрацій пилу Сахари в атмосферному повітрі України, *Наукові праці ВНТУ* [Електронний ресурс]. Вип. 2, Лип 2024. Режим доступу: doi https://doi.org/10.31649/2307-5376-2024-2-28-38
2. «Eco City» Громадський моніторинг стану якості повітря, 2025 [Електронний ресурс] – Режим доступу до ресурсу: https://reborn.eco-city.org.ua/
3. Mokin Vitalii, Shmundiak Dmytro, Kopniak Volodymyr. Air Quality Monitoring from EcoCity, May 2024, Kaggle Dataset. URL: https://www.kaggle.com/datasets/vbmokin/air-quality-monitoring-from-ecocity
4. Наука про дані: машинне навчання та інтелектуальний аналіз даних : електронний навчальний посібник комбінованого (локального та мережевого) використання [Електронний ресурс] / В. Б. Мокін, М. В. Дратований – Вінниця : ВНТУ, 2024. – 258 с. – Режим доступу: https://docs.vntu.edu.ua/card.php?id=8163.
5. Системний аналіз стану атмосферного повітря регіону, з урахуванням впливу аномалій : монографія / Д. О. Шмундяк, В. Б. Мокін, Є. М. Крижановський. – Вінниця : ВНТУ, 2025 - 169 с.
6. Копняк В. Є. Мокін В. Б. Порівняння моделей LSTM та GRU для прогнозування залишків у бустинговій гетероскедастичній EGARCH-моделі. Міжнародна науково-практична інтернет-конференція «Молодь в науці: дослідження, проблеми, перспективи», 15-16 червня 2025 року (МН-2025). https://conferences.vntu.edu.ua/index.php/mn/mn2025/paper/view/25752/21234
7. Копняк В. Ю., Мокін В. Б. Структурний аналіз процесу поширення пилу Сахари в атмосферному повітрі міста Вінниці. LV Всеукраїнська науково-технічна конференція підрозділів Вінницького національного технічного університету: Науково-технічна конференція факультету інтелектуальних інформаційних технологій та автоматизації, Вінниця, 24-27 березня 2026 року. https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2026/paper/view/28839/23511

Копняк Володимир Євгенович — аспірант кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, vkopnyak@gmail.com

Мокін Віталій Борисович – д-р техн. наук, проф., завідувач кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, e-mail: vbmokin@vntu.edu.ua

Kopniak Volodymyr Y. – Postgraduate student of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, vkopnyak@gmail.com

Mokin Vitalii B. **–** Dr. Tech. Sciences, Prof., Head of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: vbmokin@vntu.edu.ua