https://doi.org/10.31649/1997-9266-2025-183-6-139-148

УДК 004.9:504.3.054

В. Є. Копняк[1]
В. Б. Мокін[1]

МЕТОД СТРУКТУРНОГО АНАЛІЗУ ГЕТЕРОСКЕДАСТИЧНОГО ЧАСОВОГО РЯДУ З ВИКОРИСТАННЯМ БУСТИНГОВОЇ EGARCH-МОДЕЛІ

[1]Вінницький національний технічний університет

Розроблено комплексний метод структурного аналізу гетероскедастичного часового ряду, який ґрунтується на поєднанні авторської бустингової моделі умовного середнього та EGARCH-моделі умовної дисперсії. Метод спрямований на кількісну локалізацію та формальну ідентифікацію прихованих динамічних режимів процесу, викликаних дією значних зовнішніх збурювальних факторів. Запропоновано індекс модельної складності, що визначає міру структурного ускладнення процесу та відображає необхідність внесення додаткових бустингових корекцій у прогноз. Метод містить й мультишкальний аналіз рівнів впливу збурень, аналіз залежності між умовним середнім та умовною дисперсією, а також кластеризацію векторів рішень дерев в ансамблі, що дає можливість виявляти різні закономірності та режими процесу без використання додаткових ознак. Зазначено, що метод може бути поширений і на інші види подібних моделей гетероскедастичних процесів, які будуть використані замість EGARCH-складової моделі.

Запропоновано алгоритм виділення інтервалів підвищеної волатильності та підвищеної структурної складності, а також визначення їх статистичної значущості з використанням волатильнісних, ентропійних та економетричних метрик.

Продемонстровано практичне застосування методу на даних громадського моніторингу якості атмосферного повітря м. Вінниця за показником PM1. Отримані результати свідчать про високу точність прогнозу (коефіцієнт детермінації — 0,97) та дають змогу локалізувати короткочасні стохастичні збурення й тривалі структурні аномалії, що не ідентифікуються класичними моделями ARIMA, ARIMA-GARCH, Prophet, інтелектуальними багатофічерними моделями на основі результатів роботи методів Python-бібліотеки tsfresh чи інших методів формального генерування ознак. Узагальнений аналітичний висновок згенеровано автоматично з використанням великої мовної моделі. Доведено, що індекс структурної складності дозволяє виявляти аномалії у внутрішній структурі процесу, які не супроводжуються зростанням умовної дисперсії та відображають появу нових систематичних впливів або зміну динамічного механізму процесу.

Наведений приклад підтверджує ефективність методу для задач системного аналізу та створює основу для його подальшого застосування у системах підтримки прийняття рішень щодо оцінювання стану довкілля та стану інших складних систем.

Ключові слова: часові ряди, аналіз даних, системний аналіз, гетероскедастична модель, бустинг, інтелектуальна модель, генеративний штучний інтелект, якість атмосферного повітря.

Вступ

Аналіз часових рядів процесів у складних системах, зазвичай, зосереджується на дослідженні стаціонарності, сезонності, різних статистичних параметрах тощо. Проте важливіше вміти розкривати внутрішню структуру цих рядів, зокрема, з'ясовувати закономірності зміни в часі середнього значення та дисперсії ряду, які мають місце внаслідок впливу як попередніх значень ряду, так і зовнішніх збурювальних факторів, які змінюють режим функціонування системи та її процесів. Наприклад, у задачі аналізу, моделювання та прогнозування показників стану атмосферного повітря міста може мати вплив локальних збурень (значні викиди пересувних чи стаціонарних джерел), регіональних (викиди великого підприємства чи розподілених по місту котелень) або глобальних (хмара піску з Сахари, яка іноді долітає до Європи, навіть вже декілька разів діставалась до України [1]). Ці події можна розглядати як динамічні режими процесу у певних проміжках часу, які

статистично та структурно відрізняються від базової (фонової, звичайної) поведінки часового ряду. Їх своєчасне виявлення та правильна інтерпретація є ключовою умовою підвищення ефективності систем підтримки ухвалення управлінських рішень для екологічної безпеки міст. Ці режими є властивими й іншим видам часових рядів складних систем різної природи. Важливо зазначити, що йдеться про підвиди режимів, які саме обмежені певними інтервалами часу і властиві не усім значенням ряду. Іноді їх ще називають «локальні режими» [2], але в цій статті називатимемо їх просто «режими», щоб не вносити плутанину з режимами, утвореними внаслідок впливу локальних факторів.

Аналіз даних моніторингу стану атмосферного повітря у м. Вінниці за квітень–травень 2024 року, коли місто декілька разів накрила хмара піску з Сахари, дозволив діагностувати гетероскедастичність рядів показників концентрації пилу в повітрі (PM1, PM2.5, PM10) [1]. Але стандартні методи не дали змоги встановити, які саме структурні зміни відбулись у процесі, як вони проявлялися у різні моменти часу, ні якісно, ні кількісно, що ускладнює інтерпретацію виявлених закономірностей. Цікавою є саме можливість виявити приховані динамічні режими в одному часовому ряді, не аналізуючи ряди даних інших ознак. Досвід одного зі співавторів цієї роботи щодо моделювання та прогнозування кількості хворих на коронавірус в Україні протягом 2020—2022 років [3] показав, що іноді варто прогнозувати тільки один показник без урахування інших, особливо, якщо інші вимірюються менш точно, з різними запізненнями і помилками, їх взаємозв’язок є динамічним і має недостатньо вивчений характер.

Для аналізу часових рядів без додаткових ознак найчастіше застосовуються моделі ARIMA [4] або загальніші SARIMAX [5] та Prophet [3], [5], але вони орієнтовані на врахування тільки гомоскедастичних залишків і не розраховані на ряди з умовною гетероскедастичністю, відтак, не можуть виявляти динамічні режими, які характерні, передусім для гетероскедастичних часових рядів. Цих недоліків позбавлені спеціалізовані моделі ARCH та GARCH (узагальнена ARCH) [1], [6] — вони моделюють умовну дисперсію рядів, але не розкривають її структурних причин, виявляють волатильність, але не локалізують її сегментів, не мають засобів кількісного аналізу внутрішніх динамічних режимів. Ще популярним є використання бібліотеки tsfresh з генеруванням 1200 різних статистичних показників та відбором статистично значущих серед них [5], [7], які потім дозволяють використати їх в одній з відомих інтелектуальних моделей для аналізу і передбачення табличних даних, але цей підхід, по-перше, не моделює умовну дисперсію як окремий процес, не дозволяє виявити специфічну динамічну структуру самого ряду, оскільки виявлені показники важко пов’язуються з окремими моментами часу, а по-друге, більшість таких синтезованих параметрів важко піддаються інтерпретації. Аналогічно методи бібліотеки Merlion [8] та інші відомі методи формального обчислення ознак за різними алгоритмами теж не дозволяють аналізувати внутрішню структуру ряду та виявляти його динамічні режими. Більші можливості є в моделях ARIMA-GARCH [9], які добре працюють з умовним середнім і дисперсією та зі стрибками волатильності, але їм складно виявляти нелінійні залежності між вибірками даних, особливо внаслідок дії значних зовнішніх збурювальних факторів, та глибокі приховані закономірності, оскільки основний алгоритм оснований на лінійному поєднанні моделей.

Мета статті — розробити метод аналізу внутрішньої структури гетероскедастичного часового ряду для локалізації та кількісного оцінювання прихованих динамічних режимів процесу, що формуються під впливом значних зовнішніх збурювальних факторів.

Бустингова GARCH-модель гетероскедастичного часового ряду

У статті [1] авторами запропоновано бустингову GARCH-модель гетероскедастичного часового ряду. Зокрема, вибрано EGARCH-варіацію моделі гетероскедастичного процесу — це асиметричне розширення GARCH-моделі, яке дозволяє моделі по-різному реагувати на «негативні» та «позитивні» стрибки, а також гарантує невід’ємність дисперсії, завдяки операції логарифмування.

Для ряду даних X та цільової ознаки y на нульовій ітерації (час $t = 0$) використовується модель $f_0(X_T)$ — RandomForestRegressor (RDF) [1]:

$$f_0(X_T) = RFR(X_T, r[0]), \tag{1}$$

де $r[0] = y$ — початкові залишки.

На кожній наступній ітерації $t = 1...N-1$ [1]:

– навчаємо модель DecisionTreeRegressor (DTR)

$$f_t\left(X_T\right) = DTR\left(X_T, r^*\left[t\right]\right); \tag{2}$$

– обчислюємо прогнози

$$y_{pred}\left[t\right] = f_t\left(X_V\right) \tag{3}$$

та залишки

$$r\left[t+1\right] = r^*\left[t\right] - Ly_{pred}\left[t\right]; \tag{4}$$

– модель EGARCH (EG) дає умовну дисперсію залишків

$$\sigma^2\left[t+1\right] = EG\left(r\left[t+1\right]\right); \tag{5}$$

– обчислюємо кориговані залишки

$$r^*\left[t+1\right] = \frac{r\left[t+1\right]}{\sqrt{\sigma^2\left[t+1\right]+\xi}} = \Omega_{t+1}\left(\cdot\right), \tag{6}$$

де ξ — мала константа.

Фінальний прогноз — зважена сума передбачень на усіх ітераціях, в залежності від параметра L, який є швидкістю навчання (англ. «learning rate»):

$$y_\Sigma = \sum_{t=0}^{N-1} Ly_{pred}\left[t\right]. \tag{7}$$

Ця модель, як і модель ARIMA-GARCH, завдяки складовій у вигляді GARCH-моделі, добре працює з умовним середнім $\mu[t]$ і умовною дисперсією $\sigma[t]$ та зі стрибками волатильності, але, на відміну від ARIMA-GARCH, завдяки бустинговій архітектурі, дозволяє виявляти нелінійні закономірності та локальні режими, які можна кластеризувати та окремо проаналізувати. Отже, пропонується будувати метод структурного аналізу саме на основі цієї авторської бустингової EGARCH-моделі.

Позначимо Y_t випадкову величину, яка реально спостерігається і прогнозується формулою (7) у момент часу t, тоді стандартизовані залишки ζ_t бустингової гетероскедастичної моделі можна визначити так:

$$\zeta_t = r_t^* = \frac{Y_t - \hat{y}_\Sigma[t]}{\hat{\sigma}_t}, \tag{8}$$

де символ «капелюха» означає емпіричну оцінку значення величин, які обчислюються за формулами (5) та (7).

***Лема* 1.** Якщо модель правильно адаптована до усіх домінуючих процесів у складній системі з відомим розкидом значень, який є у тренувальній вибірці, то стандартизовані залишки ζ_t мають стандартизовану дисперсію, близьку до 1.

***Лема* 2.** Якщо у системі виникає новий домінуючий процес, який не відбувався під час навчання моделі, то з високою ймовірністю спостерігається або структурна зміна у послідовності $\hat{\sigma}_t$, або зміна розподілу ζ_t у формулі (8).

Якщо одночасні сплески $\hat{\sigma}_t$ спостерігаються на багатьох станціях спостережень стану складної системи, тоді це свідчить про можливі глобальні чи загальнорегіональні процеси збурення (хмара пилу з Сахари, дим від великих пожеж тощо), а якщо сплески спостерігаються на одній або декількох поряд розташованих станціях, тоді це може свідчити про локальне забруднення. Це дозволяє здійснювати просторово-часовий аналіз особливостей динаміки процесу.

Кожне дерево рішень — це набір «відтинків» часу, де модель додає корекцію до прогнозу, а ансамбль таких дерев — адаптивний розклад ряду на часові режими. Перші дерева, зазвичай, ловлять грубі, низькочастотні структури, а дерева, які формуються на пізніших ітераціях — дрібні, високочастотні патерни.

Вираз (7) можна проаналізувати на різній кількості K ітерацій (різній кількості дерев):

$$\hat{y}_\Sigma^K[t] = \sum_{i=1}^{K} Ly_{pred}^i[t]. \tag{9}$$

Це, фактично, дає «бустинговий спектральний розклад» (аналог елементів вейвлет-декомпозиції, але побудований даними). Можна показати, що тривалі збурення, в основному, «ловляться» першими деревами (малі значення K), а локальні, короткочасні збурення — на пізніших ітераціях (великі K). Це дає можливість виявити додаткові мультишкальні особливості процесу у системі. Такий вид аналізу назвемо мультишкальним.

Бустинг у такій моделі автоматично виконує кускову апроксимацію у часі, виділяє періоди, де попередній прогноз систематично помилявся, разом із EGARCH-частиною дає локальні режими і для середнього значення, і для дисперсії. Це дозволяє виділити певні структурні особливості процесу щодо його змін у часі.

Кожний момент часу t — це по суті, певна траєкторія «рішень» моделі на кожній ітерації. А це означає, що кожний момент часу можна подати як вектор індексів листків або вектор значень корекцій $v[t]$

$$v[t] = \left(f_0[t], f_1[t], \ldots, f_{N-1}[t] \right). \tag{10}$$

Побудовані у такий спосіб вектори (10) можна кластеризувати одним із відомих методів. У результаті можна буде отримати кластери часових режимів, які визначені не просто календарно, а за поведінкою моделі:

– кластер «звичайних міських днів»;

– кластер «локальних збурень»;

– кластер «епізодів глобальних збурень»;

– кластер «очікуваних добре передбачуваних днів» тощо.

Введемо індекс модельної (або структурної) складності процесу $C[t]$ як сумарний внесок бустингових корекцій для моменту t на основі формули (7), що відображає рівень локальної структурної аномальності

$$C[t] = \sum_{i=1}^{N-1} \left| Ly^{i}_{pred}[t] \right|. \tag{11}$$

Якщо значення $C[t]$ — мале, тобто базової моделі (першого дерева / RFR) вистачає, тоді це — «простий» режим; а якщо $C[t]$ — велике, тоді потрібно багато корекцій, що свідчить про складний режим та аномальну ситуацію.

Теорема. Якщо залишки $r[t]$ моделі є центрованими та мають обмежену дисперсію, а функції f_t мають скінченну норму та обмежену варіацію, тоді індекс складності $C[t]$ є монотонною мірою локальної структурних аномалій та задовольняє:

1. $C[t] \approx 0$ на ділянках, де процес є структурно однорідним і добре апроксимується базовою моделлю;

2. $C[t]$ зростає на ділянках, де мають місце локальні структурні зміни, зумовлені нелінійністю, різкими стрибками значень та/або зовнішніми впливами;

3. Для будь-яких двох моментів часу $t_1 < t_2$, якщо

$$\sum_{t} \left| r[t][t_1] \right| < \sum_{t} \left| r[t][t_2] \right|,$$

тоді $C[t_1] < C[t_2]$.

Доведення. З визначення $C[t]$ та властивостей бустингу випливає, що кожна корекція f_t мінімізує залишкову похибку на ітерації. За умов обмеженості $r[t]$ та нормування через EGARCH, сумарна величина корекцій є монотонною оцінкою локальної складності моделювання. Можна показати, що якщо на інтервалі процес змінює поведінку, то кількість корекцій, які необхідно внести, збільшується, і це відображається у зростанні $C[t]$.

Таким чином, модель (1)—(7), запропонована у роботі [1] авторами цієї статті, має переваги перед аналогами, оскільки дозволяє:

– визначати індекс складності процесу $C[t]$;

– виявляти додаткові мультишкальні особливості процесу $\hat{y}^{K}_{\Sigma}[t]$, тобто виявляти велико- та дрібномасштабні компоненти або глобальні/локальні впливи та їх характер зміни у часі;

– кластеризувати часові режими $v[t]$, виявляти їхні характерні типи та аналізувати можливі сценарії ризику;

– аналізувати динаміку середнього та дисперсії в явному вигляді як певний процес;

– аналізувати стрибки (чи, як це називати в економетрії — «шоки») як стохастичні процеси;

– моделювати нелінійну динаміку процесу;

– гарно інтерпретувати виявлені особливості.

Загалом, EGARCH-структура моделі не є принциповою. Це може бути й інша варіація, для якої варто реалізувати бустинговий алгоритм, подібний до (1)—(7). Наприклад, для моделювання різких забруднень атмосферного повітря може бути ефективнішою TGARCH-структура, яка краще враховує стрибки значень вгору та вниз, але коли підйоми є різкішими за падіння значень. А FIGARCH-структура (з довготривалою «пам'яттю» у дисперсії) є кращою, коли забруднення є регулярним, наприклад, сезонне незаконне спалювання залишків рослинності жителями міста.

Метод структурного аналізу

Пропонуємо будувати метод структурного аналізу, ґрунтуючись на таких гіпотезах:

H1. У часовому ряді існують короткотривалі режими високої волатильності, які не можуть бути виявлені стандартними економетричними чи статистичними моделями.

H2. Поведінка залишків бустингової моделі містить інформацію про структурну складність процесу, відмінну від стохастичної волатильності.

H3. Періоди, коли індекс модельної складності $C[t]$ зростає, відповідають зовнішнім збуренням або іншим аномальним впливам.

H4. Інтервали з підвищеною умовною дисперсією та індексом складності $C[t]$ можна формально локалізувати і кількісно оцінити.

Пропонований метод оснований на тому, що беруться різні параметри моделі, аналізується їхня базова статистика (медіана, квартилі) та виявляються діапазони, коли значення найбільше перевищують базовий рівень на певну величину. Математично, на прикладі деякого параметра $p[t]$ локалізацію інтервалів великих значень здійснюємо у такий спосіб:

– обчислюємо базову статистику — медіану m та медіану абсолютних відхилень або міжквартильний розмах s між третім і першим квартилем або 75-м і 25-м квантилями:

$$m(p) = median\big(p[t]\big),\ s(p) = Q_{0,75}\big(p[t]\big) - Q_{0,25}\big(p[t]\big);$$

– задаємо поріг β для значень, які вважаємо великими, наприклад,

$$\beta(p) = m(p) + ks(p), \quad k \in [1,5;\ 3];$$

– визначаємо множину точок високої складності

$$\Phi(p) = \big\{t : p[t] > \beta(p)\big\}.$$

Сусідні у часі точки з цієї множини $\Phi(p)$ об'єднуємо в інтервали значень. Потім для кожного такого інтервалу обчислюємо спеціальні метрики і порівнюємо їх зі значеннями на базовому режимі. Якщо нульова гіпотеза «немає різниці у значенні метрики» відхиляється на рівні значущості, наприклад $\alpha = 0{,}05$, тоді інтервал вважається структурно відмінним з погляду прогнозної складності. Усі такі інтервали об'єднуємо у нову множину. Будемо далі позначати результат роботи цього алгоритму для параметра p як $\Omega(p)$. Спеціальними метриками можуть бути відомі економетричні тести [10]—[14]:

– тест Дайболда–Маріяно (Diebold-Mariano – DM-тест), який порівнює похибки на інтервалах з високим та низьким значенням $C[t]$ та аналізує чи перевищує розраховане значення певний поріг;

– ентропія перестановок (Permutation Entropy — PE) — високі значення PE на інтервалах з великим $C[t]$ підтверджують їхню збільшену непередбачуваність;

– вибіркова ентропія або ентропія вибірки (Sample Entropy — SampEn) — демонструє рівень хаотичності, порівнюючи варіанти з високим і низьким $C[t]$;

– стійкість волатильності (Persistence of Volatility) — велике значення підтверджує структурний злам;

– індекс кластеризації волатильності (Volatility Clustering Index — VCI), оснований на пошуку кореляції залишків — зростає на інтервалах високої структурної складності.

Також, варто оцінювати відомі метрики MAE (Mean Absolute Error), MAPE (Mean Absolute Percentage Error), SMAPE (Symmetric Mean Absolute Percentage Error), RMSE (Root Mean Squared Error) тощо між прогнозними і реальними значеннями для усіх точок кожного інтервалу [5], [10].

Пропонуємо такий алгоритм застосування методу структурного аналізу:

Eman 1. Побудова та навчання складових ансамблевої моделі (1)—(7).

Eman 2. Обчислення $y_{pred}[t];\ \mu[t];\ r[t];\ r^*[t];\ \sigma[t];\ y_\Sigma[t];\ \zeta_t,\ \hat{y}_\Sigma^{K_i}[t];\ C[t]$ для всіх варіантів значень.

Eman 3. Аналіз гетероскедастичності процесу та оцінювання залежності (обчислення кореляції) між умовним середнім μ[t] і умовною дисперсією σ[t].

Eman 4. Аналіз інтервалів підвищеної волатильності та підвищеної складності.

Виявляються інтервали $\Omega(\sigma)$, $\Omega(C)$, визначаються їхні тривалості та межі. Для кожного такого інтервалу та для базового режиму визначаються середня волатильність, середня складність, середня ентропія, потім знаходяться їхні відношення. Якщо показники інтервалів $\Omega(\sigma)$, $\Omega(C)$ перевищують базові у 2 і більше разів, тоді вони вважаються статистично значущими.

Eman 5. Мультишкальний аналіз.

Вибрати набір значень $K_1 < K_2 < \ldots < K_M \le N$-1. Для кожного K_i (i = 1…N-1) з цієї послідовності та для кожного t обчислити

$$\hat{y}_{\Sigma}^{K_i}[t] = \sum_{k=0}^{K_i} L f_k[t].$$

Визначити різницеві компоненти, які можна вважати аналогом «шарів»

$$\Delta^i[t] = \hat{y}_{\Sigma}^{K_i}[t] - \hat{y}_{\Sigma}^{K_{i-1}}[t].$$

Після цього для кожного інтервалу $J_i \in \Omega(C)$ визначається «енергія» компоненти

$$E_j^i = \frac{1}{\left|J_j\right|} \sum_{t \in J_j} \left(\Delta^i[t]\right)^2.$$

Далі ця енергія порівнюється з базовою і виявляються інтервали, де вона є вищою за встановлений поріг. А потім аналізуються значення j, за яких вони є такими високими. Якщо ці значення — малі, тоді це — великий масштаб (глобальні впливи), а якщо — великі, тоді масштаб, навпаки — малий (локальний).

Eman 6. Аналіз взаємозв'язків між волатильністю, ентропією та індексом складності.

На кожному інтервалі $J_i \in \Omega(C)$ визначити кореляцію (можливо, з певним лагом) між різними комбінаціями пар показників $C[t]$, σ[t], PE[t], SampEn [t]. Проаналізувати варіанти виявленого наявного кореляційного зв'язку, зокрема й з певним лагом. У такий спосіб можна спробувати виявити, наприклад, коли система спочатку входить у режим підвищеної волатильності, а потім виникає потреба в ускладненні її моделі. Це дозволяє виявити причинно-наслідковий механізм ускладнення моделі на інтервалах значних структурних аномалій.

Eman 7. Кластеризація векторів дерев рішень.

Визначення векторів рішень $v[t]$, їх кластеризація одним з відомих методів (k-means++, агломеративна, HDBSCAN тощо). Потім для кожного кластера оцінюються згадані вище показники та обчислюються метрики для визначення статистичної значущості між цими кластерами. Також, варто проаналізувати їхні розміри та асиметрію між ними.

Eman 8. Формування висновків.

Все це можна автоматизувати у вигляді мультиагентної системи з подальшим аналізом з використанням великих мовних моделей (LLM) та автоматичним генеруванням висновків у вигляді природномовного тексту.

Приклад застосування методу

Продемонструємо приклад застосування запропонованого методу на тому ж прикладі, який розглядався у роботі [1] за даними «Кабінету дослідника якості повітря України» (https://archive.ecocity.org.ua) мережі громадського моніторингу атмосферного повітря EcoCity, до якого Вінницький національний технічний університет (ВНТУ) має авторизований доступ. Зокрема, у 2024 р. одним з адміністраторів Кабінету дослідника від ВНТУ, співавтором цієї статті В. Копняком здійснено імпорт даних спостережень у Вінницькій області за показником «PM1» (концентрація найбільш дрібнодисперсного пилу, який переноситься на великі відстані) за період з 25.03.2024 р. по 10.05.2024 р. та завантажено у Kaggle-датасет «Air Quality Monitoring from EcoCity» [15]. Розроблено Python-ноутбук на базі платформи Kaggle, який частково реалізує вищеописаний алгоритм методу, зокрема, автоматизовано етапи 1, 2 (частину показників), 3 (тільки кореляція між умовним середнім μ[t] і умовною дисперсією σ[t]), 4 (виявлення інтервалів значної підвищеної волатильності $\Omega(\sigma)$ та підвищеної структурної складності $\Omega(C)$, визначення їхніх тривалостей та меж, але без визначення

показників ентропії та оцінювання статистичної значущості за різними критеріями), інші етапи поки не автоматизувались. В результаті виконання Python-ноутбук усі обчислені значення збережено у різні CSV-файли та побудовані графіки і збережені у форматі PNG (рис. 1, 2).

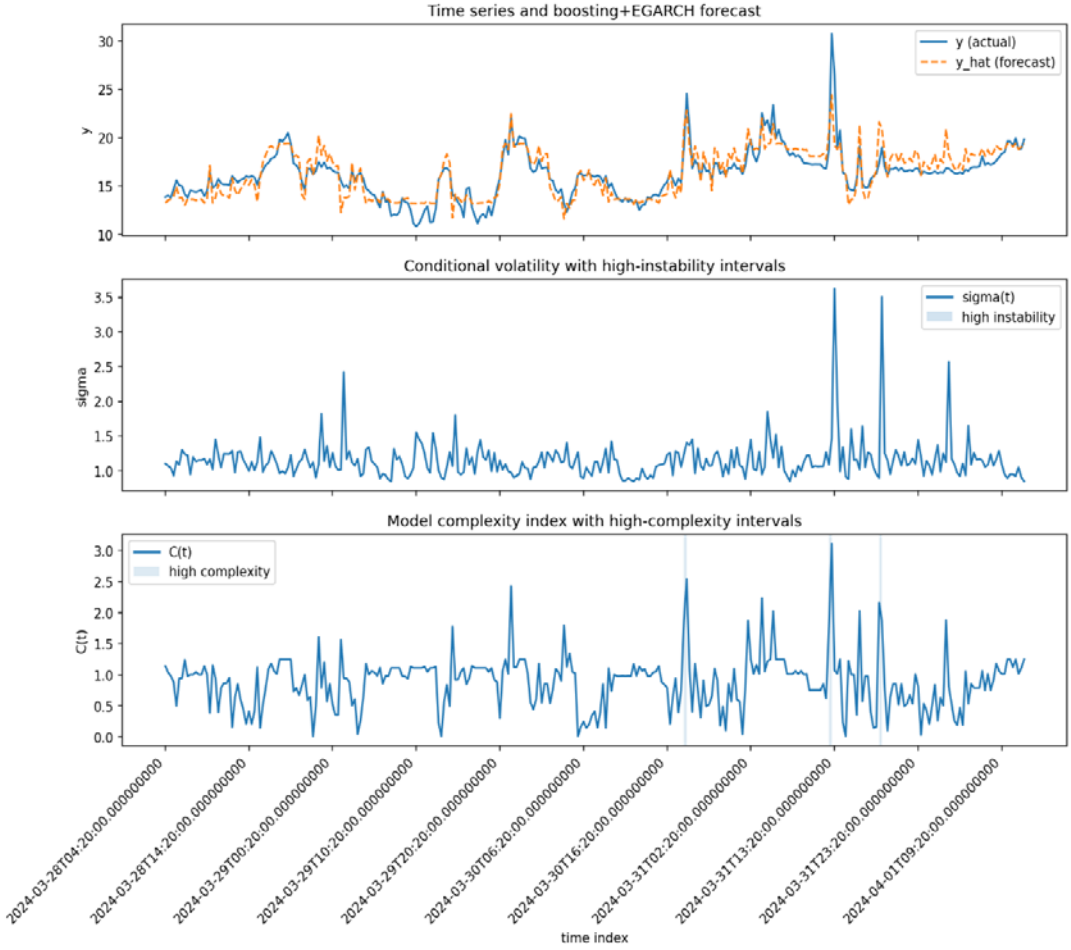

Рис. 1. Результати структурного аналізу часового ряду (окремі точки на графіках з'єднані лініями для кращої візуалізації): прогноз умовного середнього μ[*t*], умовна волатильність σ[*t*] та індекс модельної складності C[*t*] з локалізованими інтервалами підвищеної волатильності Ω(σ) та структурної складності Ω(*C*)

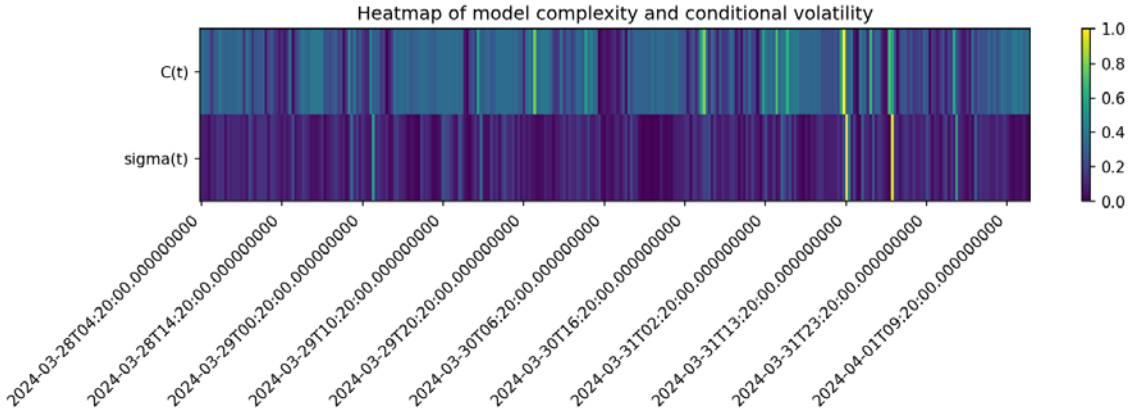

Рис. 2. Теплова карта нормалізованих значень σ[*t*] та *C*[*t*], яка синхронно відображає стохастичні та структурні зміни процесу і дозволяє візуально зіставити інтервали Ω(σ) і Ω(*C*)

Для автоматизації AI-генерування висновків на основі CSV-файлів та графічних файлів (рис. 1, 2), збережених у форматі PNG, використано велику мовну модель ChatGPT (OpenAI, GPT-5.1). Разом з цими файлами у промпті передано стислий опис методу і завдання виявити та лаконічно навести тільки основні патерни (закономірності), які видно у цих даних. Результат генерації подано на рис. 3.

Як випливає з рис. 3, бустингова EGARCH-модель (1)—(7) є адекватною та, як і в статті [1],

забезпечує високу точність прогнозування (метрика $R^2 = 0{,}97$), що підтверджує коректність моделювання і прогнозування умовного середнього $\mu[t]$ та умовної волатильності $\sigma[t]$.

Автоматизований аналіз структурних властивостей часового ряду $PM_{2.5}$ (м. Вінниця, 2024 р.)

1. **Точність прогнозу.**
   R^2 = **0.970**, RMSE = **0.495 µg/m³**, MAPE = **1.66%**, corr(y, ŷ) = **0.969**.
2. **Стохастична нестабільність σ(t).**
   Виявлено **два інтервали підвищеної волатильності**, кожен тривалістю 2 точки (≈ 40 хв):
   - **31.03, 13:00–13:20**, середнє σ(t) = **1.47 µg/m³**;
   - **31.03, 19:00–19:20**, середнє σ(t) = **3.50 µg/m³**.

   Обидва інтервали істотно перевищують **базовий рівень σ(t)**.
3. **Детермінована структурна складність C(t).**
   Ідентифіковано **три інтервали підвищеної складності**, розташовані у межах 29–31.03:
   - **Інтервал 1:** 29.03 (≈ 10:00–11:20), середнє C(t) = **2.24**, діапазон **2.10–2.35** (≈ 2.33 год);
   - **Інтервал 2:** 30.03 (≈ 07:40–09:00), середнє C(t) = **2.31**, діапазон **2.18–2.41** (≈ 2.33 год);
   - **Інтервал 3:** 31.03 (≈ 08:00–09:20), середнє C(t) = **2.44**, діапазон **2.30–2.52** (≈ 2.66 год).
4. **Режимні переходи.**
   Зміни між базовим та аномальними інтервалами спостерігаються з часовою роздільністю **20–40 хв**.
5. **Кореляційні властивості.**
   corr(µ(t), σ(t)) = **0.0808**, що свідчить про відсутність залежності між рівнем процесу та його волатильністю.

Рис. 3. Результат автоматичної генерації ключових висновків аналізу результатів роботи Python-програми (CSV-файли і графіки з рис. 1, 2 у форматі PNG) для реалізації методу структурного аналізу (OpenAI, ChatGPT 5.1, Nov. 2025)

Застосування алгоритму запропонованого методу дозволило здійснити структурний аналіз часового ряду та локалізувати окремі режими. На основі $\sigma[t]$ та нормованої залишкової волатильності виокремлено два інтервали підвищеної стохастичної волатильності, кожний з яких містить дві послідовні у часі точки з різким зростанням цієї волатильності. Обидва інтервали суттєво відхиляються від базового рівня $\sigma[t]$ і можуть інтерпретуватися як короткочасні стохастичні збурення, пов'язані з дією зовнішніх факторів. Інші аномальні значення $\sigma[t]$ не формують інтервалів і є одиничними аномаліями. Окремо за формулою (11) для індексу $C[t]$ виявлено три інтервали підвищеної структурної складності, локалізовані у межах 29—31.03.2024 р. Ці інтервали не супроводжуються пропорційним зростанням $\sigma[t]$, що свідчить про структурну складність, тобто про такі зміни у внутрішній організації процесу, коли модельна похибка зростає не через стохастичні стрибки значень дисперсії, а через ускладнення закономірностей, які модель повинна описувати. Це може означати появу нових систематичних чинників або їх накладення на фонові механізми формування концентрацій. Тривалість кожного з інтервалів підвищеної складності становить 7—8 точок, що відповідає приблизно 2,33…2,66 год, тобто вони є довшими за стохастичні збурення. Така відносно тривала стабільність високих значень $C[t]$ може свідчити про структурну аномалію процесу, ймовірно зумовлену тривалою дією зовнішнього фактора або зміною механізму перенесення й дисперсії частинок у міському атмосферному повітрі. Переходи між базовим режимом та аномальними інтервалами спостерігаються з кроком 20…40 хв.

Таким чином, отримані результати моделювання та структурного аналізу часового ряду концентрацій малих твердих частинок пилу (PM1) в атмосферному повітрі м. Вінниця у квітні–травні 2024 року свідчать про здатність розробленого методу забезпечувати глибоке розуміння динамічних властивостей процесу та його внутрішньої структури. Поєднання бустингової моделі для прогнозування умовного математичного сподівання та EGARCH-компоненти для моделювання умовної волатильності дало змогу отримати прогноз з високою точністю ($R^2 = 0{,}97$), а також сформувати додаткові структурні характеристики, які не доступні в рамках класичних економетричних підходів. По-перше, аналіз умовної волатильності $\sigma[t]$ дозволив чітко виокремити інтервали підвищеної стохастичної волатильності, що характеризуються різким зростанням дисперсії залишків і вказують на наявність інтенсивних зовнішніх впливів на систему. У досліджуваному часовому ряді такі режими фіксувалися двічі та супроводжувалися характерними піками волатильності, які можуть бути пов'язані з перенесенням пилу, раптовими змінами траєкторій повітряних мас або локальними викидами забруднювальних речовин. Для більшої точності результатів аналізу, тобто чи є ці зовнішні впливи локальними або глобальними, потрібний такий аналіз даних спостережень з усіх станцій міста за цей період за цим же показником. По-друге, індекс модельної складності $C[t]$ дійсно виявився інформативним індикатором структурних змін. Його локальні максимуми вказували на часові інтервали, коли процес ставав складнішим для апроксимації, навіть за високої прогнозної точності моделі. Це свідчить про появу в процесі нелінійних залежностей, локальних змін форми розподілу або короткочасних динамічних режимів, які не супроводжуються значним зростанням стохастичної дисперсії. Виявлено, що значна частина інтервалів високої складності не збігається з інтервалами високої волатильності, що підтверджує ефективність запропонованого методу та його здатність

фіксувати різну природу динамічних змін. По-третє, сформовані інтервали підвищеної стохастичної волатильності та високої складності дають змогу ідентифікувати режими функціонування екологічної системи, зокрема: базовий режим, режим підвищеного впливу зовнішніх збурень та комбінований режим. Візуалізація часових режимів у вигляді теплових карт показала чітку стратифікацію станів системи та підтвердила наявність короткочасних «аномальних вікон», які були б малопомітними у класичних регресійних моделях часових рядів, охарактеризованих вище.

Висновки

У роботі розроблено комплексний метод структурного аналізу гетероскедастичного часового ряду, заснований на поєднанні бустингової моделі умовного середнього та EGARCH-моделі умовної дисперсії. Метод, на додаток до типових підходів щодо аналізу даних, забезпечує кількісне виявлення локальних режимів процесу, оцінювання часової складності моделювання, визначення мультишкальних компонентів та виявлення шокових періодів дії зовнішніх збурень з підвищеною волатильністю. Статистична значущість структурних режимів підтверджується економетричними (DM-test), ентропійними (Permutation Entropy, Sample Entropy) та волатильнісними метриками (VCI, персистентність).

У прикладі застосування методу для часового ряду концентрацій пилу «PM1» в атмосферному повітрі м. Вінниці за даними громадського моніторингу у березні–травні 2024 р. досягнуто високу точність прогнозування умовного середнього значення з коефіцієнтом детермінації 0,97. У результаті застосування структурного аналізу за запропонованим методом кількісно локалізовано два інтервали підвищеної стохастичної волатильності тривалістю по 2 часові точки (приблизно по 40 хв. кожен), а також — три інтервали підвищеної структурної складності з тривалістю 7—8 точок, що відповідає приблизно 2,33…2,66 год. Виявлені інтервали високої складності не супроводжувалися пропорційним зростанням умовної дисперсії, що свідчить про наявність структурних змін у внутрішній динаміці процесу, які не фіксуються класичними волатильнісними моделями.

Розроблений метод може бути поширеним і на інші види моделей гетероскедастичних часових рядів, а не тільки на EGARCH, а саме: TGARCH, FIGARCH та інші після їх адаптування до ролі складової бустингової моделі на кшталт моделі (1)—(7).

Наукова новизна полягає у розробці нового методу, основаного на оригінальній авторській моделі, який, на відміну від наявних, дозволяє одночасно виявляти закономірності щодо динамічних режимів середнього, волатильності і структурної складності процесу у різних часових масштабах. Запропонований метод створює основу для подальшого розвитку інструментів системного аналізу складних систем різної природи та систем підтримки прийняття рішень.

Наведений приклад підтвердив ефективність розробленого методу у виявленні важливих закономірностей.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

[1] В. Є. Копняк, В. Б. Мокін, С. О. Жуков, І. В. Варчук, і Т. В. Скринник, «Метод бустингу гетероскедастичних моделей для прогнозування концентрацій пилу Сахари в атмосферному повітрі України,» *Наукові праці Наукові праці Вінницького національного технічного університету*, вип. 2, Лип. 2024. [Електронний ресурс]. https://doi.org/10.31649/2307-5376-2024-2-28-38 .

[2] J. D. Hamilton, “A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle,” *Econometrica*, vol. 57, no. 2, pp. 357-384, Mar. 1989. https://doi.org/10.2307/1912559 .

[3] В. Б. Мокін, А. В. Лосенко, і А. Р. Ящолт, «Інформаційна технологія аналізу та прогнозування кількості нових випадків захворювань на коронавірус SARS-CoV-2 в Україні на основі моделі Prophet,» *Вісник Вінницького політехнічного інституту*, № 5, с. 71-83, 2020. https://doi.org/10.31649/1997-9266-2020-152-5-71-83 .

[4] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: John Wiley & Sons, 2015, 712 p. ISBN: 978-1-118-67502-1.

[5] В. Б. Мокін, і М. В. Дратований, *Наука про дані: машинне навчання та інтелектуальний аналіз даних*, електр. навч. посіб. Вінниця, Україна: ВНТУ, 2024, 258 с. [Електронний ресурс]. Режим доступу: https://docs.vntu.edu.ua/card.php?id=8163.

[6] J. Brownlee, “How to Model Volatility with ARCH and GARCH for Time Series Forecasting in Python,” *Machine Learning Mastery*, Aug. 24, 2018. [Electronic resource]. Available: https://machinelearningmastery.com/develop-arch-and-garch-models-for-time-series-forecasting-in-python .

[7] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, “Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package),” *Neurocomputing*, vol. 307, pp. 72-77, 2018. https://doi.org/10.1016/j.neucom.2018.03.067 .

[8] A. Bhatnagar, P. Kassianik, C. Liu, et al., “Merlion: A Machine Learning Library for Time Series,” *arXiv preprint*, arXiv:2109.09265, 2021. https://doi.org/10.48550/arXiv.2109.09265 .

[9] Q. P. Duy, et al., “Estimating and forecasting bitcoin daily prices using ARIMA-GARCH models,” *Business Analyst Journal*, vol. 45, no. 1, pp. 11-23, 2024. https://doi.org/10.1108/BAJ-05-2024-0027.

[10] J. A. Mariño, M. E. Arrieta-Prieto, and S. A. Calderón, “Comparison between statistical models and machine learning for forecasting multivariate time series: An empirical approach,” *Communications in Statistics: Case Studies, Data Analysis and Applications*, vol. 11, no. 1, pp. 56-91, 2025. https://doi.org/10.1080/23737484.2025.2463905 .

[11] J. Olbryś, and E. Majewska, “Regularity in stock market indices within turbulence periods: The sample entropy approach,” *Entropy*, vol. 24, no. 7, p. 921, 2022. https://doi.org/10.3390/e24070921 .

[12] I. Chronopoulos, L. Giraitis, and G. Kapetanios, “Choosing between persistent and stationary volatility,” *The Annals of Statistics*, vol. 50, no. 6, pp. 3466-3483, Dec. 2022. https://doi.org/10.1214/22-AOS2236 .

[13] G. Bardas, N. Kefalakis, and J. Soldatos, “Indicators of External Disruptions in Supply Chains: A Framework for Early Detection and Resilience Planning,” in *2025 21st International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT)*. Lucca, Italy, 2025, pp. 1064-1071. https://doi.org/10.1109/DCOSS-IoT65416.2025.00160 .

[14] K. Shibano, R. Lin, and G. Mogi, “Volatility Reducing Effect by Introducing a Price Stabilization Agent on Cryptocurrencies Trading,” in *Proceedings of the 2020 2nd International Conference on Blockchain Technology (ICBCT ’20)*. New York, NY, USA: ACM, 2020, pp. 85-89. https://doi.org/10.1145/3390566.3391679 .

[15] V. Mokin, D. Shmundiak, and V. Kopniak, “Air Quality Monitoring from EcoCity,” *Kaggle Dataset*, May 2024. [Electronic resource]. Available: https://www.kaggle.com/datasets/vbmokin/air-quality-monitoring-from-ecocity .

Рекомендовано до друку кафедрою системного аналізу та інформаційних технологій ВНТУ

Стаття надійшла до редакції 5.12.2025

Копняк Володимир Євгенович — аспірант кафедри системного аналізу та інформаційних технологій, e-mail: vkopnyak@gmail.com ;

Мокін Віталій Борисович — д-р техн. наук, професор, завідувач кафедри системного аналізу та інформаційних технологій, e-mail: vbmokin@vntu.edu.ua .

Вінницький національний технічний університет, Вінниця

V. Ye. Kopniak[1]
V. B. Mokin[1]

Method of Structural Analysis of Heteroscedastic Time Series Using the Boosting EGARCH Model

[1]Vinnytsia National Technical University

A comprehensive method for the structural analysis of a heteroscedastic time series is developed in this study. The approach combines an original boosting-based model of the conditional mean with an EGARCH specification of the conditional variance. The method is designed to quantitatively localize and formally identify hidden dynamic modes of the process driven by substantial external perturbations. A model-complexity index is introduced to quantify the degree of structural complication in the process and to indicate when additional boosting-based corrections to the forecast are required.

The method also incorporates a multiscale assessment of disturbance-intensity levels, an analysis of the dependence between the conditional mean and conditional variance, and clustering of the decision vectors of trees within the ensemble. This enables the detection of distinct patterns and modes without relying on auxiliary features. It is noted that the method can be extended to other classes of heteroscedastic process models that may be used in place of the EGARCH component.

An algorithm is proposed for identifying intervals of elevated volatility and heightened structural complexity, as well as for assessing their statistical significance using volatility-, entropy-, and econometrics-based metrics.

The practical application of the method is demonstrated using public air quality monitoring data for the city of Vinnytsia, specifically the PM1 indicator. The results show high predictive accuracy (coefficient of determination — 0.97) and make it possible to localize short-term stochastic disturbances and long-term structural anomalies that remain undetected by classical ARIMA, ARIMA-GARCH, Prophet, or multifactor intelligent models based on feature-engineering techniques such as those provided by the Python library tsfresh. A generalized analytical conclusion was generated automatically with the assistance of a Large Language Model. The study shows that the proposed structural-complexity index can reveal anomalies within the internal structure of the process that are not accompanied by increases in conditional variance and that reflect the emergence of new systematic influences or changes in the process’s dynamic mechanism.

The example provided confirms the effectiveness of the method for systems-analysis tasks and forms a basis for its further use in decision-support systems for assessing environmental conditions and other complex systems.

Keywords: time series, data analysis, system analysis, heteroskedastic model, boosting, intelligent model, generative artificial intelligence, air quality.

Kopniak Volodymyr Ye. — Post-Graduate Student of the Chair of System Analysis and Information Technologies, e-mail: vkopnyak@gmail.com ;

Mokin Vitalii B. — Dr. Sc. (Eng.), Professor, Head of the Chair of System Analysis and Information Technologies, e-mail: vbmokin@vntu.edu.ua