Міністерство освіти і науки України

Національний університет «Києво-Могилянська академія»

Факультет інформатики

Кафедра математики

Кваліфікаційна робота

освітній ступінь – бакалавр

на тему: **«Методи виявлення аномалій для часових рядів»**

Виконав: студент 4-го року навчання освітньої програми «Прикладна математика»,
спеціальності 113 Прикладна математика

Огир Вадим Дмитрович

Керівник: Щестюк Н. Ю.,
кандидат фіз.-мат. наук, доцент

Рецензент: Івасюк Дмитро Ярославович, Ph. D.

Кваліфікаційна робота захищена
з оцінкою _____

Секретар ЕК _____
«____» _____ 20____ р.

Київ – 2022

Міністерство освіти і науки України

Національний університет «Києво-Могилянська академія»

Факультет інформатики

Кафедра математики

ЗАТВЕРДЖУЮ
Зав.кафедри математики,
проф., д.ф-м.н., Б. В. Олійник
_____
(підпис)
„____”_____2021 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ
на кваліфікаційну роботу

студенту 4-го курсу, факультету інформатики__
Огиру Вадиму Дмитровичу_____
Реалізувати, дослідити і порівняти методи визначення аномалій для часових рядів
Зміст ТЧ до кваліфікаційної роботи:

Зміст
Анотація
Вступ
1 Основні відомості про часові ряди та аномалії у них
2 Опис і реалізація методів визначення аномалій у часових рядах
3 Порівняння реалізованих методів
Висновки
Список використаних джерел

Дата видачі „___” _____ 2021 р. Керівник _____
(підпис)
Завдання отримав _____
(підпис)

Графік підготовки кваліфікаційної роботи до захисту

Графік узгоджено «_____» _____ 2021 р.

№ з/п	Перелік робіт	Термін	Підпис	Дата	Примітка
1.	Отримання теми кваліфікаційної роботи	16.10.2021			
2.	Ознайомлення з існуючою інформацією за темою курсової роботи	20.10.2021			
3.	Розробка плану та структури роботи	01.11.2021			
4.	Робота з науковою літературою	03.11.2021			
5.	Аналіз загальновідомих методів визначення аномалій, їх реалізація	01.01.2022			
6.	Реалізація модифікацій описаних алгоритмів для роботи у реальному часі	01.02.2022			
7.	Реалізація методу визначення аномалій з використанням показника Херста	01.04.2022			
8.	Виконання статистичного порівняння реалізованих методів на реальних даних	10.04.2022			
9.	Аналіз практичної частини, її корегування	15.04.2022			
10.	Початок написання текстової частини	20.04.2022			
11.	Подання проміжної версії текстової частини	10.05.2022			
12.	Остаточне завершення написання текстової частини роботи	20.06.2022			
13.	Створення презентації	22.06.2022			
14.	Захист кваліфікаційної роботи	05.07.2022			

ЗМІСТ

АНОТАЦІЯ 6

ВСТУП 7

РОЗДІЛ 1: ЧАСОВІ РЯДИ. ОСНОВНІ ВІДОМОСТІ 9

1.1 Визначення часового ряду і його складових 9

1.2 Числові характеристики часових рядів. Стаціонарність 12

1.3 Аномалії у часових рядах 15

1.4 Показник Херста 17

1.5 Приклади часових рядів 18

РОЗДІЛ 2: МЕТОДИ ВИЗНАЧЕННЯ АНОМАЛІЙ У ЧАСОВИХ РЯДАХ 20

2.1. Метод визначення аномалій на основі довірчого інтервалу даних компоненти залишків часового ряду і його модифікація для роботи у реальному часі 20

2.2 Метод визначення аномалій на основі довірчого інтервалу міжквартильного розмаху періодичних даних і його модифікація для роботи у реальному часі 22

2.3 Метод визначення аномалій на основі довірчого інтервалу значень перших різниць часового ряду і його модифікація для роботи у реальному часі 25

2.4 Метод визначення аномалій на основі прогнозування з використанням машинного навчання 28

2.5 Метод визначення аномалій на основі довірчого інтервалу значень показника Херста 29

РОЗДІЛ 3: ПРАКТИЧНЕ ДОСЛІДЖЕННЯ І ПОРІВНЯННЯ МЕТОДІВ ..35

3.1. Обрахунок історичної волатильності лог-прибутків 35

3.2 Порівняння методів визначення аномалій 37

ВИСНОВКИ.....43

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ 44

АНОТАЦІЯ

Роботу присвячено застосуванню і реалізації методів визначення аномалій у часових рядах, реалізації їх модифікацій для роботи у реальному часі, а також порівнянні на реальних даних. Також у роботі запропоновано метод визначення аномалій з використанням показника Херста – міри довготривалої пам’яті часового ряду. Отримані результати показали, що метод з використанням показника Херста здатний виявляти явні аномалії та аномальні підпослідовності.

ВСТУП

За останні роки стало значно простіше збирати величезні об’єми даних з різноманітних джерел, і навіть найменші компанії цим активно користуються, прагнучи покращити якість своїх сервісів та оцінити їх ефективність. Проте, працюючи з даними, ми маємо переконатися, що серед них немає так-званих аномалій (англ. outliers), оскільки в майбутньому вони можуть створити значну кількість проблем. Втім, варто зазначити, що аномалії бувають не лише шкідливими. Якщо у випадку, до прикладу, прогнозування нових даних, аномалії можуть значно погіршити отриманий результат, то у деяких сферах життєдіяльності виявлення аномалій є основною задачею. До прикладу, виявивши аномальні показники у серцебитті людини, можна зберегти їй життя. Аномалії у даних статистики мережевого трафіку допоможуть виявити кібератаку і вчасно зупинити її. Аномалії ж у сейсмологічних показниках можуть допомогти передбачити землетрус.

Наразі існує не так багато алгоритмів визначення аномалій, особливо ефективних, оскільки складно розробити універсальний алгоритм, який буде визначати, які дані насправді є аномальними, не враховуючи природу цих даних. Отож, метою роботи є огляд, реалізація і модифікація деяких методів виявлення аномалій у часових рядах, де під часовим рядом розуміється послідовність значень, впорядкована у хронологічному порядку, а також спроба застосування показника Херста для визначення аномалій.

Об’єктом дослідження є часові ряди та методи знаходження аномалій в них. Предметом дослідження є опис деяких методів визначення аномалій у часових рядах, реалізація їх модифікацій для роботи у реальному часі, а також їх порівняння на реальних даних.

Робота складається з трьох розділів, які містять підрозділи, висновків, а також списку використаних джерел.

Перший розділ містить детальний опис часових рядів, їх основних характеристик, а також аномалій. Зокрема, увагу присвячено опису показника Херста для часового ряду, який є основою одного з методів виявлення аномалій.

У другому розділі розглянуто методи визначення аномалій у часових рядах, а також виконана їх програмна реалізація. Також розглянуто і реалізовано модифікації деяких методів для роботи у реальному часі. Описано і реалізовано метод визначення аномалій з використанням показника Херста.

У третьому розділі виконано порівняння описаних методів виявлення аномалій у часових рядах на реальних даних.

РОЗДІЛ 1: ЧАСОВІ РЯДИ. ОСНОВНІ ВІДОМОСТІ

1.1 Визначення часового ряду і його складових

Часовий ряд (англ. time series) – це набір числових спостережень, проіндексованих і впорядкованих за часом:

$$\{X_t\}, \qquad t \in T$$

По суті, часовий ряд можна описати як набір векторів $x(t), t = 0,1,2$ ..., де t – час, що минув. В цьому випадку змінна $x(t)$ вважається випадковою величиною, а часовий ряд – одною з реалізацій випадкового процесу, яка ретельно спостерігається. Найчастіше часовий ряд є послідовністю, взятою на рівновіддалених точках в часі, які йдуть одна за одною – послідовність дискретного часу.

Часовий ряд може бути як дискретним (вимірювання в окремі моменти часу), так і неперервним (вимірювання в усі моменти часу). До прикладу, покази концентрації аргону в повітрі можуть бути описані неперервним часовим рядом, а ось чисельність населення Києва – лише дискретним. У дискретних часових рядах спостереження реєструють з однаковими інтервалами часу: щогодинний, щоденний, щотижневий тощо. Також найчастіше розглядають саме дискретні часові ряди, оскільки неперервні можуть бути зведеними до дискретних шляхом об’єднання і усереднення даних визначених часових інтервалів.

На поведінку часового ряду зазвичай впливають кілька його основних складових, які використовуються для його моделювання: тренд, сезонність, циклічність і залишки.

Тренд є корисним для прогнозування майбутньої поведінки ряду. Протягом тривалого періоду тренд показує, дані мають тенденцію збільшуватись чи зменшуватись. Звісно, різні проміжки часу можуть демонструвати різні тенденції, як зростання, так і спаду. Однак якщо брати до уваги весь ряд або ж велику його частину, загальний тренд має бути визначеним: висхідний,

стабільний, або ж спадаючий. Тренд у часовому ряді також називають довгостроковим рухом. Яскравий приклад тренду можемо побачити на графіку індексу фондового ринку – так склалось історично.

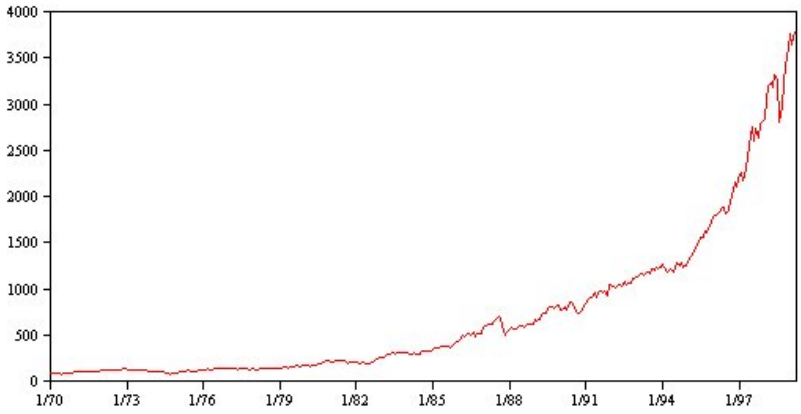

Рисунок 1.1 – Приклад тренду у часовому ряді

Сезонність у часових рядах – це наявність коливань, які відбуваються за певні регулярні інтервали, часто менші або рівні року. Сезонність може бути викликана різними факторами, як-от погодою або святами. Такі коливання можна порівняти з циклічними закономірностями, до прикладу – обсяги продажів морозива мають річну сезонність: збільшуються у літній період і зменшуються у зимній. Найчастіше зустрічається саме річна сезонність.

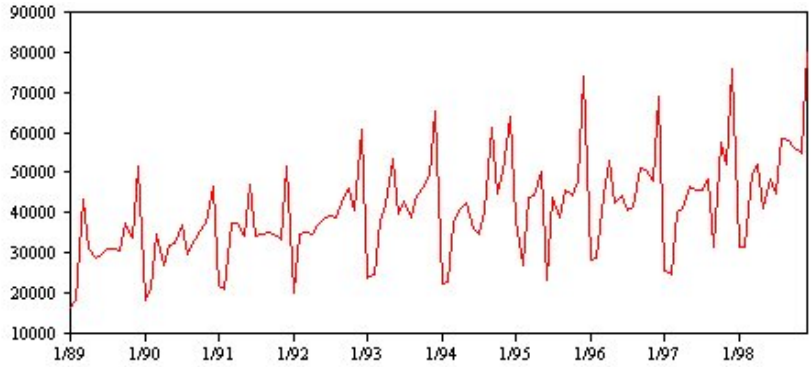

Рисунок 1.2 – Приклад сезонності у часовому ряді

Циклічність часового ряду описує середньострокові зміни ряду, що викликані циклічними обставинами – через нерегулярні інтервали. Тривалість циклу як правило більша, ніж сезону – два або більше років. Проте циклічні коливання можуть нести і сезонний характер, оскільки діяльність, до прикладу,

деяких галузей сільського господарства і економіки залежить від пори року. Таким чином, розділяють сезонну циклічність і несезонну. Прикладом несезонної циклічності можуть бути події, пов’язані з релігійними святами, дати яких за Григоріанським календарем змінюються з року в рік.

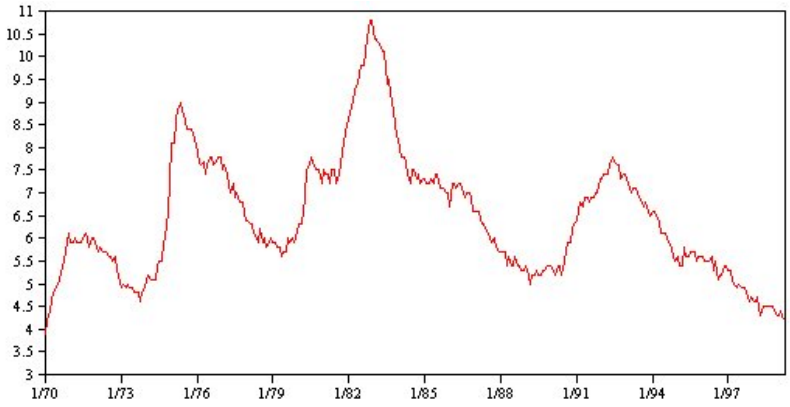

Рисунок 1.3 – Приклад несезонної циклічності у часовому ряді

Залишки у часових рядах – це випадкові коливання, які зумовлені непередбачуваними обставинами: війна, пандемія, землетрус тощо. Деякі часові ряди не містять тренду і сезонної компоненти, проте випадковий компонент є у будь-якому ряді. Залишковий компонент необхідний для того, щоб мати змогу включити його до моделі, яка описує часовий ряд. Простими словами, залишки – це те, що залишається після того, як з ряду вилучити тренд, сезонність і циклічну компоненти.

Таким чином, маючи вищеперераховані компоненти, будь-який часовий ряд можна описати таким рівнянням:

$$Y(t) = T(t) + S(t) + C(t) + I(t)$$

де $Y(t)$ – спостереження;

$T(t)$ – тренд;

$S(t)$ – сезонна компонента;

$C(t)$ – циклічна компонента;

$I(t)$ або $e(t)$ – залишки.

Така модель називається адитивною і використовується тоді, коли припускається, що компоненти ряду є незалежними, а також якщо амплітуда сезонних коливань не змінюється з часом.

Існує також мультиплікативна модель, яка заснована на припущенні, що компоненти ряду все ж не є незалежними і можуть впливати один на одного, зокрема якщо амплітуда сезонних коливань з часом змінюється. Така модель має вигляд:

$$Y(t) = T(t) \times S(t) \times C(t) \times I(t)$$

1.2 Числові характеристики часових рядів. Стаціонарність

Для роботи з часовими рядами використовують імовірнісно-статистичні моделі. Таким чином, часовий ряд ($\xi(t)$) розглядають як реалізацію випадкового процесу, який має такі основні характеристики:

1. $\mu(t) = E\big(\xi(t)\big)$ – тренд (математичне сподівання);
2. $D(t) = Var(t) = E\big(\xi - E_{\xi_t}\big)^2$ – дисперсія;
3. $\sigma(t) = \sqrt{D(t)}$ – середнє квадратичне відхилення;
4. $R(t_1, t_2) = cov\big(\xi(t_1), \xi(t_2)\big)$ – автоковаріація;
5. $\rho(t_1, t_2) = \frac{cov\left(\xi(t_1), \xi(t_2)\right)}{\sigma(t_1)\sigma(t_2)}$ – автокореляція.

Оскільки автокореляція і автоковаріація є досить важливими в аналізі часових рядів, зупинимось на цих поняттях детальніше. Спочатку розглянемо базові статистичні поняття.

Статистична коваріація – це дисперсія двох випадкових величин, тобто їх співвідношення. Коваріація показує, як змінюється одна випадкова величина при зміні іншої і приймає значення в інтервалі $(-\infty; \infty)$:

- Додатна коваріація означає прямий зв’язок між випадковими величинами. Це означає, що збільшення одної випадкової величини призведе до збільшення іншої, і чим більше значення – тим сильніше буде збільшення;

- Від’ємна коваріація означає, що збільшення одної випадкової величини призведе до зменшення іншої;
- Нульова коваріація означає відсутність зв’язку між випадковими величинами.

Статистична кореляція кількісно визначає силу зв’язку між двома змінними (випадковими процесами), тобто наскільки дві випадкові величини залежні одна від одної. Іншими словами, кореляція - це розширення коваріації – одинична міра, нормалізоване значення коваріації. Кореляція може приймати значення в інтервалі $[-1; 1]$.

Автоковаріаційна функція - це така функція, яка визначає коваріацію процесу із самим собою, зсуненим на деякий проміжок часу, що називають лагом. Простими словами, ця функція відповідає на питання, наскільки впливає деяке значення часового ряду у минулому (до прикладу, ξ_{n-2}) на деяке інше значення у майбутньому (ξ_n) враховуючи прямий зв’язок ($\xi_{n-2} \to \xi_n$) і всі непрямі ($\xi_{n-2} \to \xi_{n-1} \to \xi_n$). У випадку, коли лаг рівний 0, автоковаріація перетворюється на просту варіацію. Основні властивості автоковаріаційної функції:

1. $R(0) \geq 0$
2. $R(n) = R(-n)$
3. $R(0) \geq |R(n)|$

Автокореляційна ж функція (АКФ) – це така функція, яка дає кореляцію процесу із самим собою, зсуненим на деякий проміжок часу. Тобто, знову ж, це нормоване значення автоковаріації. У випадку, коли лаг рівний 0, автокореляційна функція, очікувано, рівна 1.

Також існує часткова автокореляційна функція (ЧАКФ). Вона дає відповідь на питання, як впливає деяке значення часового ряду у минулому на деяке інше значення у майбутньому лише напряму ($\xi_{n-2} \to \xi_n$), тобто після вилучення значень проміжних лагів. Ця функція дає більш «чисту» картину періодичних залежностей.

Всі випадкові процеси можна розділити на два види: стаціонарні і нестаціонарні. Стаціонарним часовим рядом називають випадковий процес y_t, статистичні характеристики якого не змінюються з часом, тобто $t \to t+T, y_t \to y_{t+1}$ при будь-якому T. У такому ряді:

- Немає тренду: $E(\xi_t) = E(\xi_{t-s}) = \mu = const$;
- Дисперсія незмінна: $E((\xi_t - \mu)^2) = E((\xi_{t-s} - \mu)^2) = \sigma^2 = const$;
- Автоковаріація незмінна і залежить лише від вибраних моментів часу t і $t-s$: $cov(\xi_t, \xi_{t-s}) = cov(\xi_{t-j}, \xi_{t-s-j}) = const$.

Відомим прикладом стаціонарного випадкового процесу є білий шум (ε_t), який має такі характеристики:

1. $E(\varepsilon_t) = 0$;
2. $E(\varepsilon_t, \varepsilon_{t-k}) = R(k) = 0, k \neq 0$;
3. $Var(\varepsilon_t) = \sigma^2$
4. $cov(\varepsilon_t, \varepsilon_{t-k}) = \begin{cases} 0, k \neq 0 \\ \sigma^2, k = 0 \end{cases}$

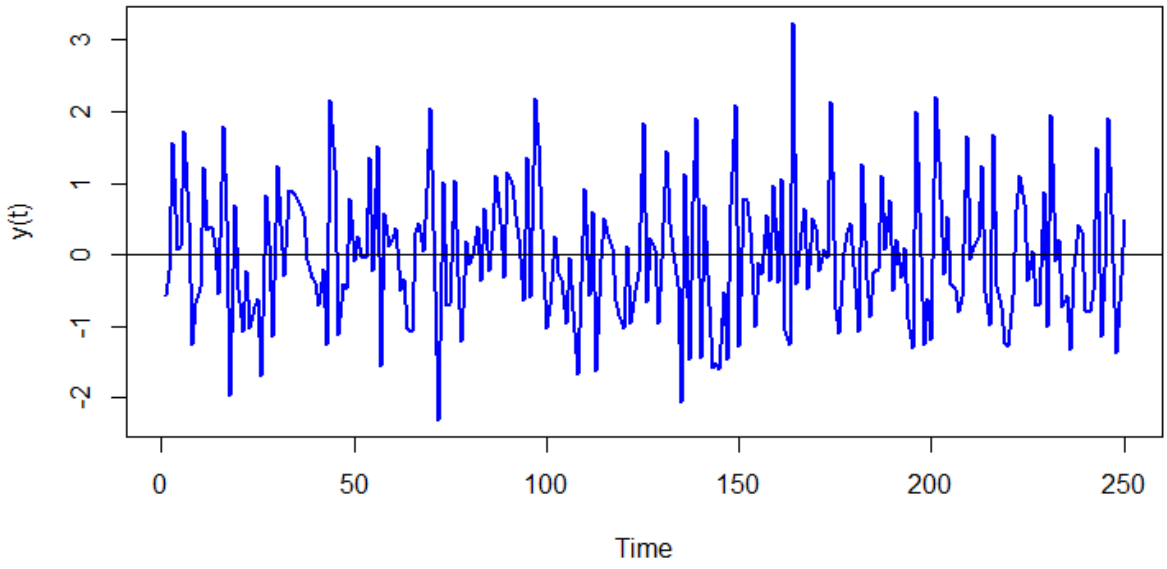

Рисунок 1.4 – Графічне представлення білого шуму

Графіки стаціонарних часових рядів нагадують насправді випадкові процеси.

В аналізі і прогнозуванні часових рядів найчастіше використовують саме стаціонарні ряди, оскільки вони позбавлені будь-яких «очікуваних» змін, як-от

постійно зростаюча коваріація, які можуть значно погіршити результати прогнозування.

Для зведення часових рядів до стаціонарних існує кілька методів, найбільш вживаний з яких – «log-difference». Цей метод з високою імовірністю допомагає позбутися тренду і сезонності у часовому ряді. Його суть полягає у тому, щоб поставити у відповідність кожному значенню ряду, починаючи з другого, різницю логарифмів цього значення і попереднього, тобто:

$$\Delta y(t) = \log\big(y(t)\big) - \log\left(y(t-1)\right)$$

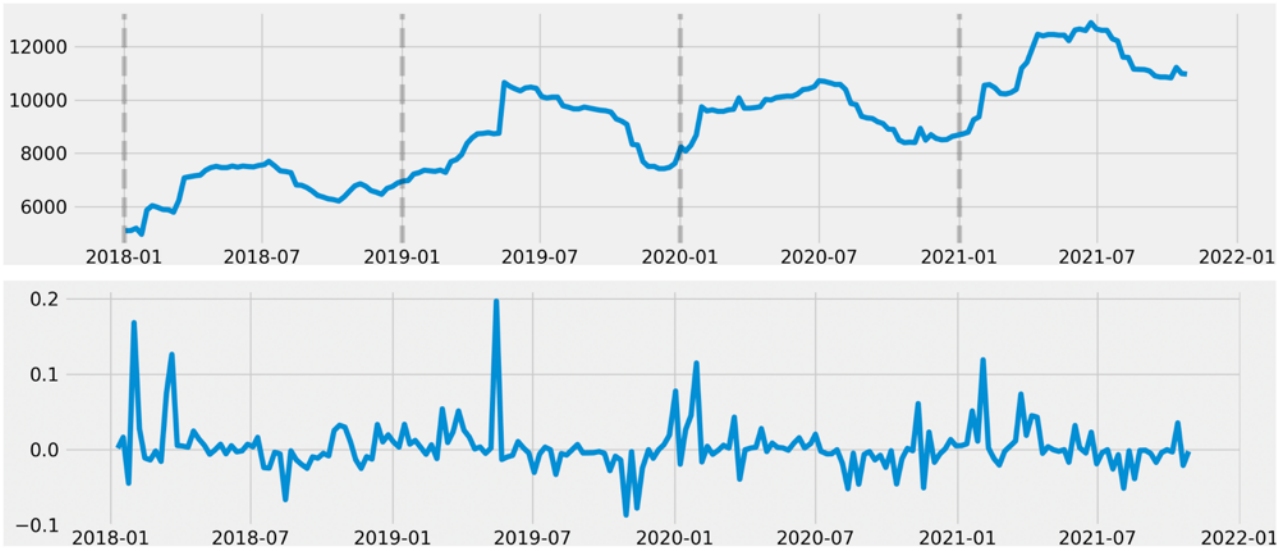

Рисунок 1.5 – Приклад перетворення часового ряду на стаціонарний з використанням методу перших різниць

1.3 Аномалії у часових рядах

Отож, як і у всіх наборах даних, не виключено, що аномалії можуть бути і в часових рядах. Більше того, про це можна стверджувати, оскільки зараз даних збирається набагато більше, а отже і аномалій також значно більше.

Спочатку маємо визначити, що таке аномалія (викид). Це спостереження, яке значно відхиляється від інших спостережень, що викликає підозри про те, що воно було створене іншим механізмом. Таким чином, можемо розглядати викиди як спостереження, що не відповідають очікуваній поведінці.

Як вже зазначалось раніше, є кілька причин шукати аномалії у даних: щоб запобігти їх значного впливу на подальший аналіз, зокрема – прогнозування нових даних, або ж щоб використати самі зібрані аномалії.

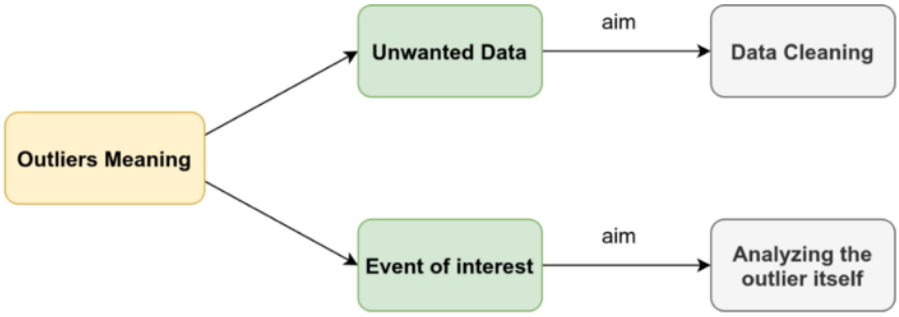

Рисунок 1.6 – Цілі пошуку аномалій у часових рядах

Аномалії у часових рядах діляться на два типи: точкові і аномальні підпослідовності. Точкова аномалія – це поодиноке аномальне спостереження серед набору очікуваних даних. Такі викиди часто досить просто виявити навіть за допомогою звичайної графічної візуалізації ряду.

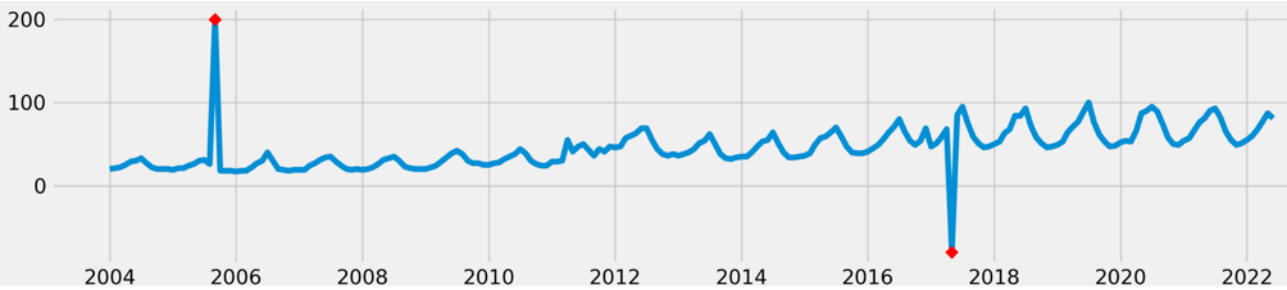

Рисунок 1.7 – Приклад точкових аномалій у часовому ряді

Аномальна підпослідовність – це набір послідовних аномальних спостережень. Аномалії цього виду виявляти значно складніше, оскільки потрібно аналізувати дані всього ряду на предмет повторюваних значень.

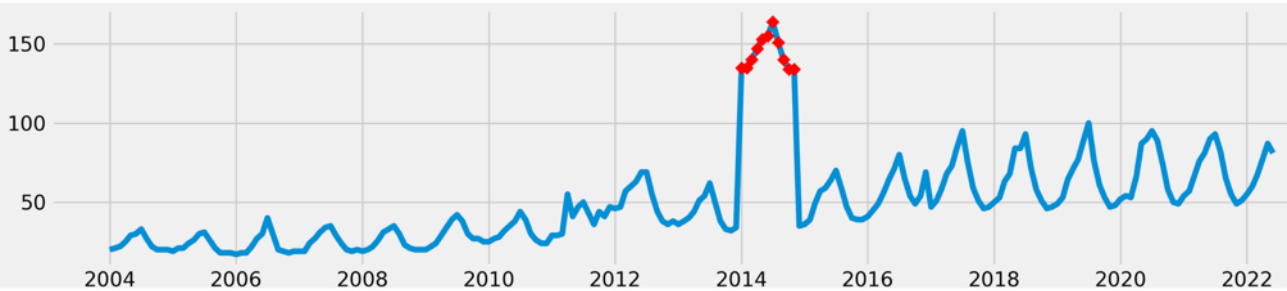

Рисунок 1.8 – Приклад аномальних підпослідовностей

Варто зауважити, що у сучасному світі, особливо в деяких критичних сферах, як-от медицина, виникла потреба у виявленні аномалій в реальному часі. Це також значно ускладнює цей процес, оскільки при обробці кожної нової точки ряду є можливість аналізувати лише дані з минулого, що, до прикладу, ускладнює виявлення аномальних підпослідовностей.

1.4 Показник Херста

Показник Херста ввів британський гідролог Гарольд Едвін Херст, який займався проектом греблі на річці Ніл. Для цього необхідно було оцінити приплив води. Спочатку вважалось, що приплив води – випадкова величина, стохастичний процес. Проте вивчивши дані розливу Нілу за останні 9 століть, гідролог знайшов закономірності. Було досліджено, що великі розливи змінювались ще більшими, а невеликі розливи – ще меншими. Одразу можемо помітити наявність циклів з неперіодичною тривалістю.

Для формування свого методу, Гарольд Херст взяв за основу роботу Ейнштейна про броунівський рух, в якій, по суті, описується модель випадкових блукань деякої частинки. Сутність теорії в тому, що відстань, яку проходить частинка, R, збільшується пропорційно квадратному кореню з часу T:

$$R = T^{0.5}$$

Якщо застосувати цю теорію на часові ряди, матимемо: розмах варіації, R, при більшій кількості випробувань буде дорівнювати кореню з кількості випробувань, T. Застосувавши алгоритм, який пізніше був названий R/S-аналізом, Херст розширив рівняння Ейнштейна, звідки вивів деякий показник, який пізніше було названо на його честь.

Тож показник Херста позначається H і приймає значення від 0 до 1, де:

- Значення від 0 до 0.5 вказують на відсутність стабільності у часовому ряді. Тренд у таких часових рядах змінюється швидко, без підпорядкування стохастичним законам. Це явище називають «рожевим шумом», воно зустрічається в процесах з турбулентністю;
- Значення 0.5 вказує на повністю стохастичну природу ряду, тобто на відсутність залежності між даними. Це явище називають білим шумом;
- Значення від 0.5 до 1 вказують на наявність стабільності у часовому ряді, позитивний тренд, а також на те, що ряд «має пам’ять» і легко прогнозується. Це явище називають «чорним шумом». Саме такі ряди наразі зустрічаються, до прикладу, на фінансових ринках.

Таким чином, Гарольд Херст довів, що чим більша затримка між двома однаковими парами значень у часовому ряді, тим менше значення приймає показник.

Отже, можемо стверджувати, що показник Херста несе багато інформації про часовий ряд, зокрема:

- персистентність/антиперсистентність;
- існування періодичних циклів;
- визначає вид шуму;
- сильну пам’ять.

У цій роботі, зокрема, маємо визначити, наскільки ефективно показник Херста допомагає у визначенні аномалій у часових рядах.

1.5 Приклади часових рядів

Часові ряди зустрічаються у багатьох сферах, таких як бізнес, економіка, промисловість, наука тощо. Для візуалізації їх, як правило, представляють графіком. Нижче наведено приклади двох різних часових рядів:

Рисунок 1.9 – Щомісячна популярність запиту «ice cream» у Google [8]

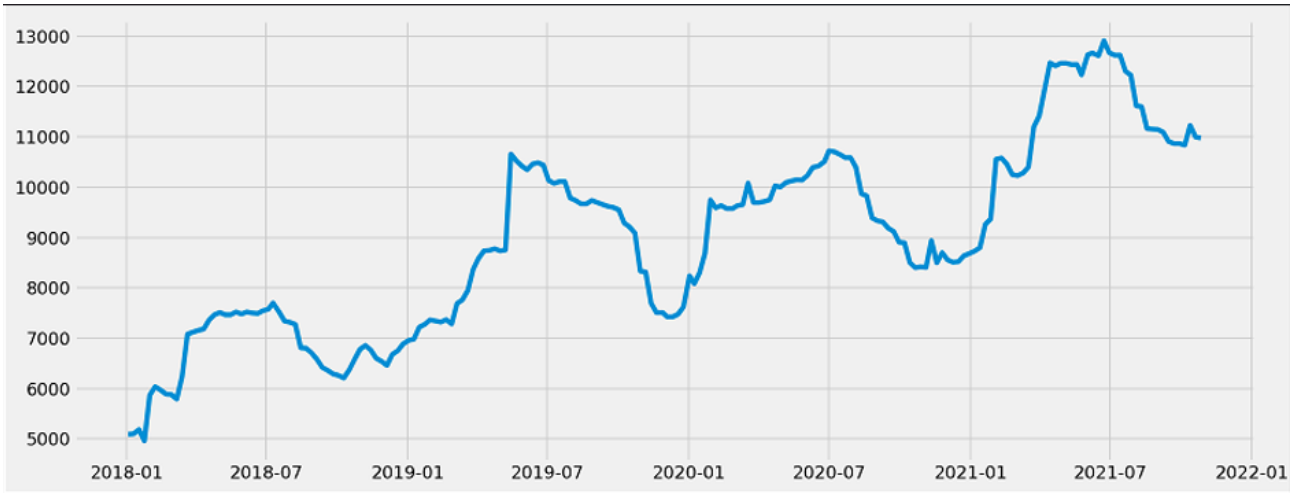

Рисунок 1.10 – Щотижнева середня ціна кілограму баранини у Монголії за період 2018-2021 [9]

Як бачимо, перший часовий ряд має чітко виражену річну сезонність: попит на морозиво зростає щорічно влітку і спадає взимку, а також майже відсутній тренд. Натомість на другому графіку бачимо чіткий зростаючий тренд, а також виражену річну сезонність. Насправді, навіть у цих даних є реальні аномалії, що ще раз підкреслює актуальність даної роботи. У наступному розділі розглянемо методи визначення аномалій у часових рядах.

РОЗДІЛ 2: МЕТОДИ ВИЗНАЧЕННЯ АНОМАЛІЙ У ЧАСОВИХ РЯДАХ

У цьому розділі розглянемо реалізації деяких методів виявлення аномалій та оцінимо їх ефективність за рядом критеріїв. Також реалізуємо і порівняємо модифікації цих методів для роботи у реальному часі, які позбавлені деяких обмежень і недоліків оригінальних версій.

Для початку, перерахуємо критерії оцінки методів визначення аномалій:

- Ефективність виявлення явних аномалій – такі аномалії можна побачити на графіку часового ряду навіть неозброєним оком;
- Ефективність виявлення неявних аномалій;
- Ефективність виявлення аномальних підпослідовностей;
- Можливість роботи у реальному часі;
- Необхідність виділення частини часового ряду для навчання алгоритму;
- Універсальність методу – незалежність від параметрів часового ряду.

2.1. Метод визначення аномалій на основі довірчого інтервалу даних компоненти залишків часового ряду і його модифікація для роботи у реальному часі

Суть цього методу полягає у розкладі часового ряду на компоненти (STL Decomposition) і використанні компоненти залишків. Оскільки інформація про явні викиди міститься у компоненті залишків, можемо провести визначення аномалій лише спираючись на неї. Тож побудувавши довірчий інтервал, до прикладу,

$$[\bar{x} - 3\sigma_{\bar{x}}; \bar{x} + 3\sigma_{\bar{x}}]$$

де $\bar{x}$ – середнє значення даних компоненти залишків;

σ – стандартне відхилення

можемо робити висновки: якщо значення залишків не належить даному інтервалу – є підстави вважати відповідне йому значення ряду аномальним.

Варто зазначити, що метод декомпозиції також надає можливість легко вилучити аномалії з вихідного ряду. Для цього достатньо замінити аномальне значення сумою відповідних значень компонент тренду і сезонності. А для більшої точності інтерпольованого значення можемо додати до нього білий шум.

Отож перевагами цього методу є проста реалізація та можливість вилучення викидів з вихідного ряду. Проте, у разі, коли ряд невеликого розміру містить явні аномалії, точність їх виявлення знижується через особливості розкладу – алгоритм, окрім власне аномалії, вважає аномаліями значення відповідного періоду у інших сезонах.

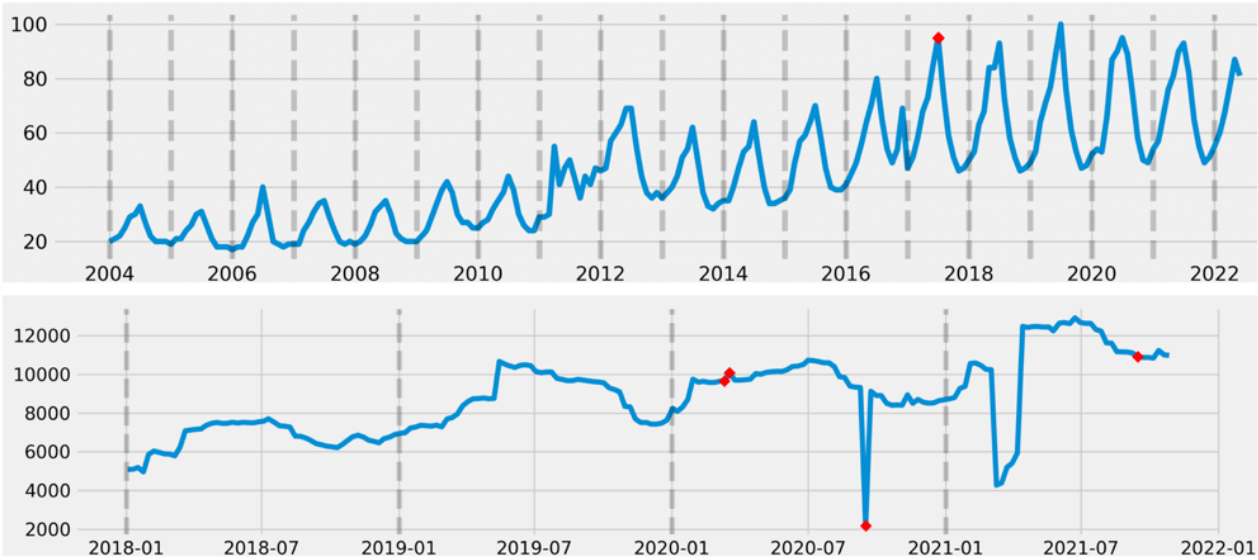

Рисунок 2.1 - Приклад роботи методу визначення аномалій на основі довірчого інтервалу даних компоненти залишків часового ряду при визначенні неявних аномалій, явних аномалій та аномальних підпослідовностей

Реалізована модифікація оригінального методу для роботи у реальному часі має підвищену ефективність, що зумовлено автоматичним вилученням знайдених аномалій. Алгоритм методу складається з кількох кроків. Спочатку виконується декомпозиція ряду і виділяється компонента залишків. Далі будується довірчий інтервал за даними цієї компоненти. Третій крок повторює перший, але цього разу – для часового ряду включно з новим значенням. Щоб

визначити, чи нове значення є аномальним, перевіряємо, чи входить останнє значення залишків нового часового ряду до побудованого інтервалу. І, нарешті, якщо значення справді аномальне – вилучаємо його описаним раніше способом.

Таким чином, ефективність виявлення явних і неявних аномалій, а також аномальних підпослідовностей є вищою, ніж у оригінальної версії, проте, згідно з наступним прикладом, метод виявляє не всі значення аномальних підпослідовностей. Також, очевидно, що алгоритму необхідна деяка частина часового ряду у якості затримки для здійснення точного розкладу.

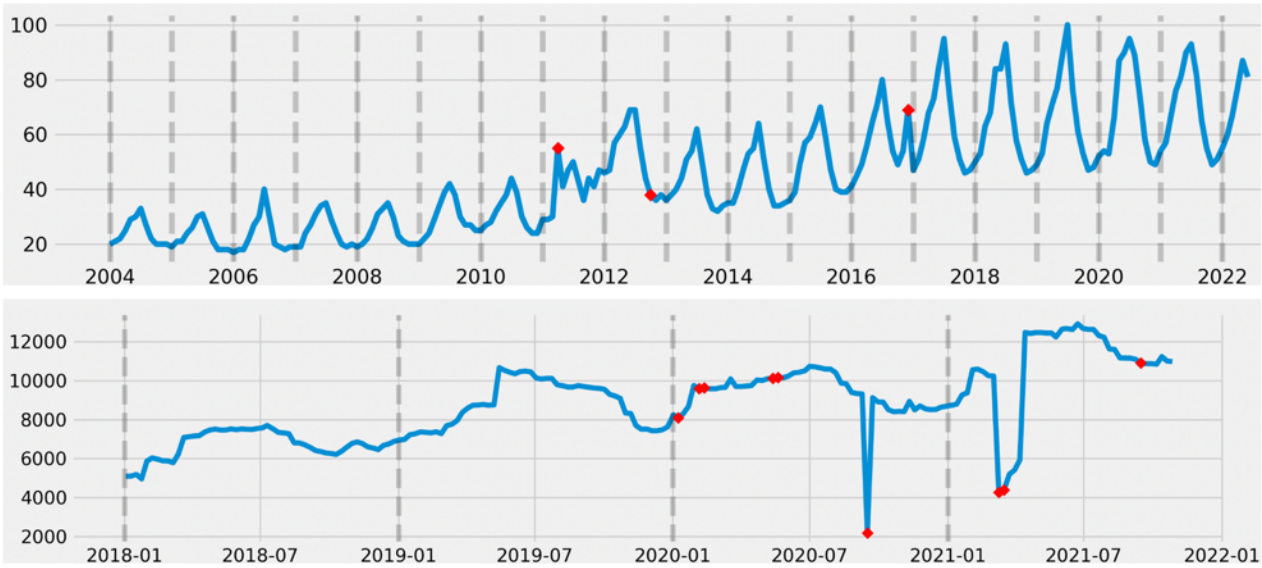

Рисунок 2.2 – Приклад роботи модифікації методу для роботи у реальному часі при визначенні неявних аномалій, явних аномалій та аномальних підпослідовностей

2.2 Метод визначення аномалій на основі довірчого інтервалу міжквартильного розмаху періодичних даних і його модифікація для роботи у реальному часі

Цей метод спирається на те, що часовий ряд має явну сезонну компоненту, і шукає аномалії шляхом порівняння значень кожного періоду у всіх сезонах. Наприклад, якщо ряд має річну сезонність, а січневе значення 2016 року є аномальним, порівнявши всі січневі значення ряду можемо легко це визначити.

Алгоритм цього методу складається з двох кроків. На першому кроці визначаємо, в якому саме періоді може бути аномалія. Для цього достатньо зібрати значення суми стандартних відхилень за кожен період ряду і побудувати довірчий інтервал на основі міжквартильного розмаху – різниці між значеннями 75-го і 25-го процентилів ряду:

$$[Q_{25} - 3 \times IQR; Q_{75} + 3 \times IQR]$$

де Q_{25}, Q_{75} – значення процентилів;
IQR – міжквартильний розмах

Періоди, що не входять в інтервал, вважатимемо такими, що містять принаймні одну аномалію. Другий крок полягає у знаходженні такого значення з кожного періоду, щоб стандартне відхилення підпослідовності з цим значенням було найбільшим. Отже, метод може знайти лише одне аномальне значення серед всіх значень періоду.

Виявилось, що ефективність методу залежить від розміру вихідного часового ряду. Якщо даних в ряді досить багато, аномальні значення просто «губляться» серед даних відповідних їх періодів. Таким чином, у порівнянні з іншими методами, ефективність при виявленні неявних аномалій і аномальних підпослідовностей є дуже низькою, проте метод здатен виявляти явні аномалії. Також метод визначення аномалій на основі довірчого інтервалу міжквартильного розмаху періодичних даних доволі обмежений у використанні, оскільки повністю спирається на сезонність, і, таким чином, залежить від характеристик ряду. Втім, варто зазначити, що метод також надає елегантну можливість вилучення аномалій шляхом їх заміни на середнє значення даних відповідних періодів.

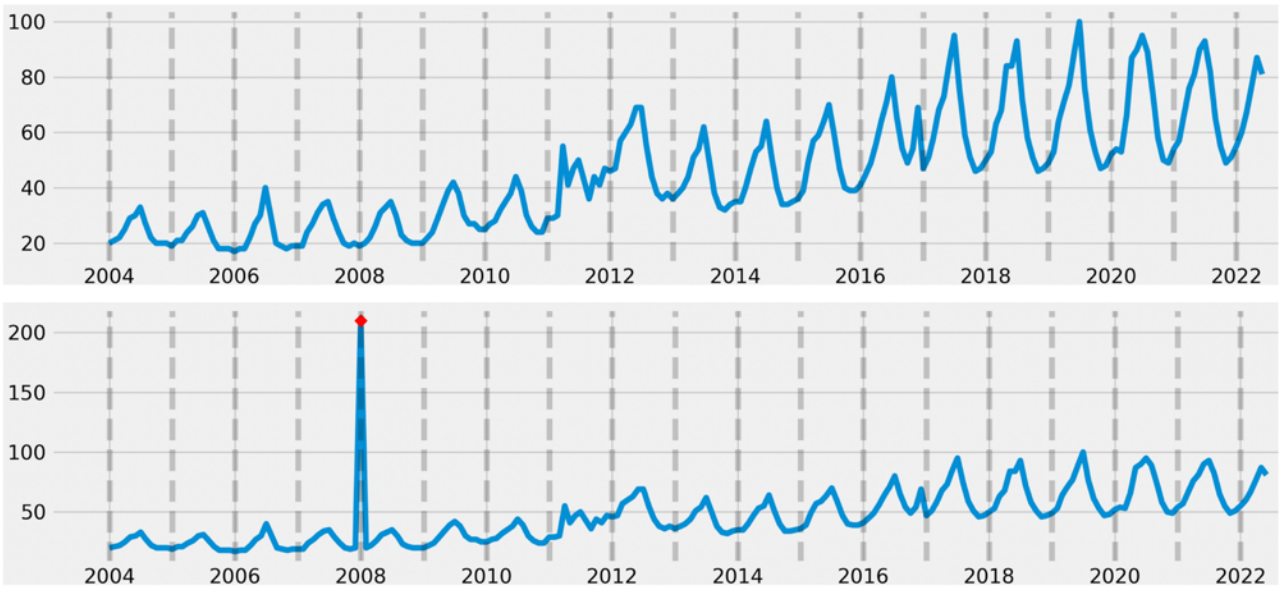

Рисунок 2.3 – Приклад роботи методу визначення аномалій на основі довірчого інтервалу міжквартильного розмаху періодичних даних при визначенні неявних аномалій, явних аномалій та аномальних підпослідовностей

Реалізація модифікації методу для роботи у реальному часі дещо відрізняється від інших. Справа в тім, що для коректної роботи алгоритму на кожній його ітерації маємо аналізувати не одне нове значення, а всі значення одного нового сезону, що робить метод значно більш точним у порівнянні з оригінальним, і здатність виявлення лише однієї аномалії серед значень одного періоду стає його перевагою. Таким чином, реалізована модифікація показує високу ефективність виявлення всіх видів аномалій, а також аномальних підпослідовностей, хоча, знову ж, потребує деякої частини часового ряду у якості затримки перед початком роботи для правильного визначення довірчого інтервалу.

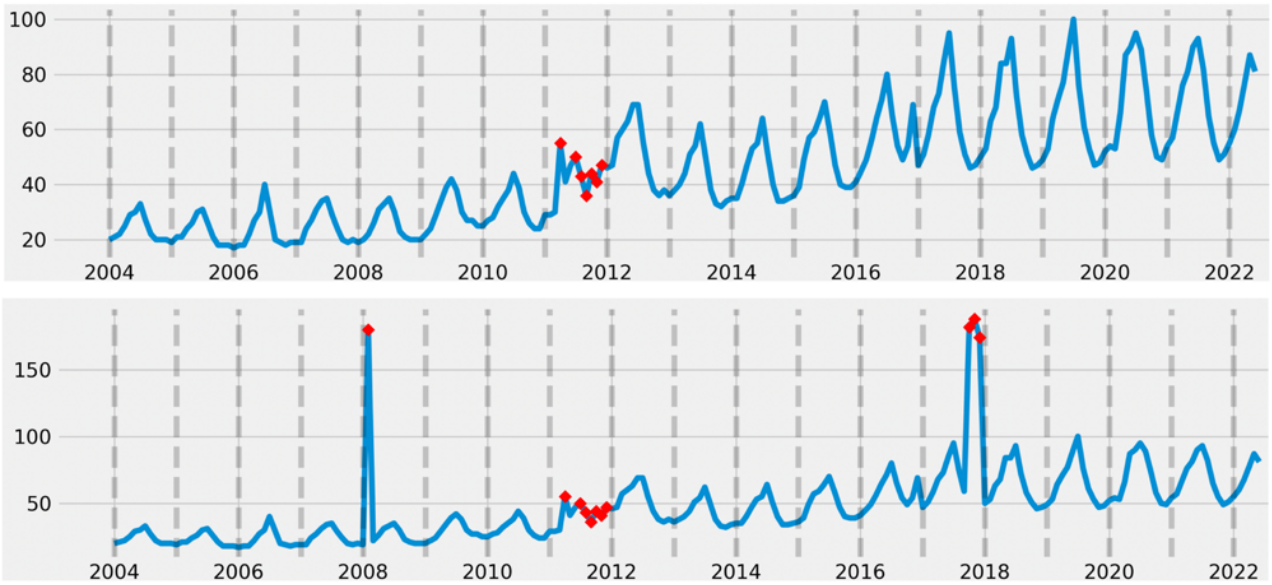

Рисунок 2.4 – Приклад роботи модифікації методу для роботи у реальному часі при визначенні неявних аномалій, явних аномалій та аномальних підпослідовностей

2.3 Метод визначення аномалій на основі довірчого інтервалу значень перших різниць часового ряду і його модифікація для роботи у реальному часі

Цей метод також є одним з найпростіших. Він спирається на стаціонарний ряд, отриманий з вихідного методом взяття перших різниць, таким чином підкреслюючи явні точкові аномалії. Залишається сформувати довірчий інтервал і точки, які не потрапили до нього, вважати аномальними. Втім, окрім аномалій, очевидно, що метод буде вважати аномальними і значення після них, оскільки, за визначенням методу перших різниць, різниця між аномальним значенням і звичайним буде великою як для самої аномалії, так і для наступної точки, якщо вона не є частиною аномальної підпослідовності.

Отже, у порівнянні з попередніми методами, ефективність виявлення явних аномалій цього методу є високою, проте через явні аномалії довірчий інтервал не здатен підкреслити неявні. Для виявлення аномальних підпослідовностей також необхідно модифікувати алгоритм, адже очевидно, що

різниця між значеннями аномальної підпослідовності буде мінімальною, і відповідні точки потраплять у довірчий інтервал. Перевагою ж методу є те, що його можна використовувати для виявлення аномалій у будь-якому часовому ряді, адже він не потребує налаштування під його характеристики.

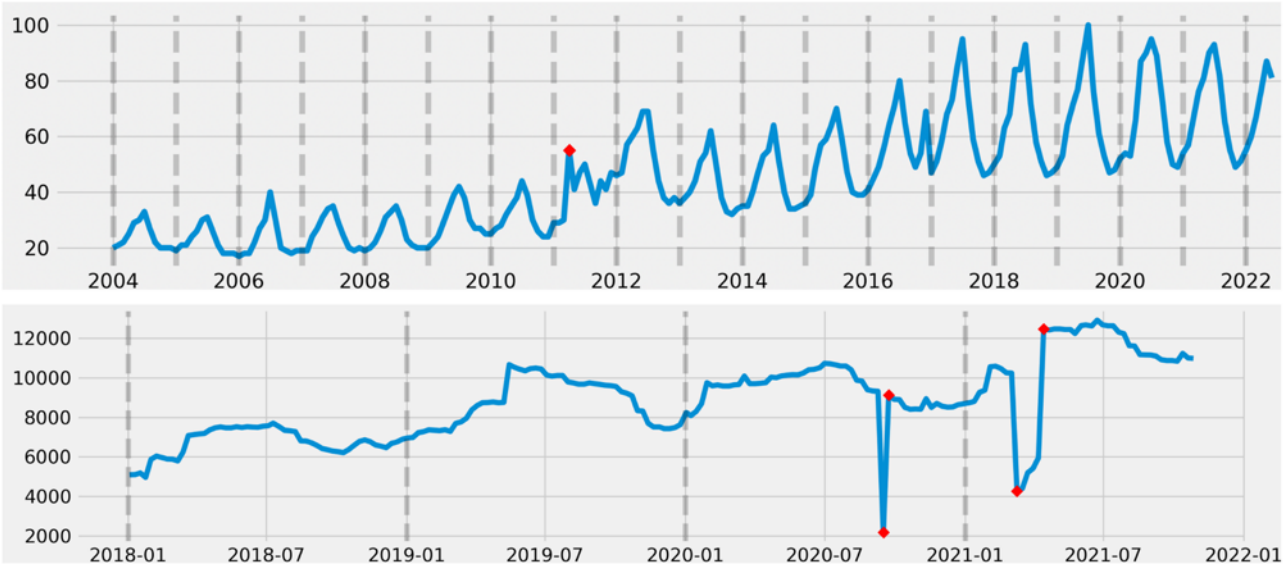

Рисунок 2.5 – Приклад роботи методу визначення аномалій на основі довірчого інтервалу значень перших різниць часового ряду при визначенні неявних аномалій, явних аномалій та аномальних підпослідовностей

Перерахованих недоліків оригінального методу можна уникнути, реалізувавши модифікацію методу для роботи у реальному часі. Тож, на кожному кроці алгоритму будується ряд перших різниць, для якого створюється довірчий інтервал, а далі перевіряється, чи входить різниця останнього і нового значень у цей інтервал. Далі, щоб уникнути проблеми з точками, що слідують одразу ж після аномальних, а також щоб зробити можливим визначення аномальних підпослідовностей, маємо вилучити знайдену аномалію.

На жаль, метод не надає простих способів вилучення аномалій. Найпростіше, що можемо зробити – продублювати останнє значення. Втім, у випадку невеликих неявних аномалій, а особливо, у випадку аномальних підпослідовностей – це може створити велику проблему при подальшій роботі

алгоритму, пов’язану з великими значеннями різниці, що також будуть вважатись аномальними:

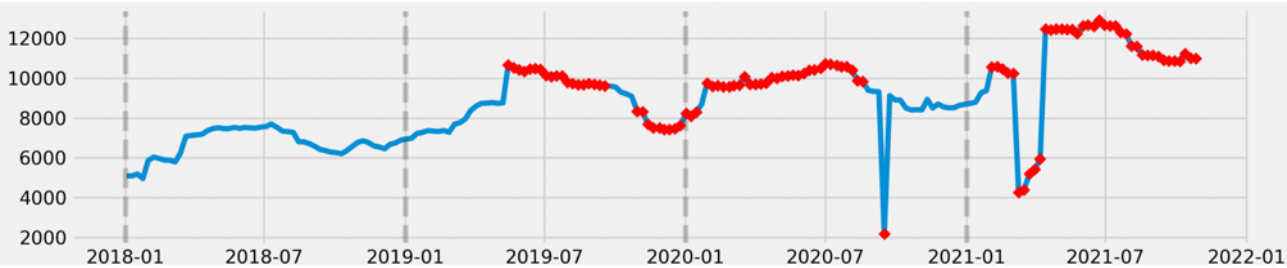

Рисунок 2.6 – Приклад роботи алгоритму при виконанні заміни аномальних значень останніми неаномальними

Щоб виправити цю ситуацію, я ввів додаткову логіку заміни аномальних значень. У випадку, якщо викид не входить в 3σ-інтервал, значення замінюється на середнє попереднього і поточного. Якщо ж аномальне значення не входить навіть в 5σ-інтервал – значення замінюється за такою формулою:

$$X_i = \frac{X_{i-1} + X_i}{2.2}$$

При цьому дільник бажано підбирати індивідуально для кожного конкретного ряду.

Отже, ефективність модифікації методу для роботи у реальному часі значно перевищує ефективність оригінального методу при виявленні явних аномалій і аномальних підпослідовностей. Алгоритму також необхідна деяка затримка перед початком роботи для правильного визначення довірчого інтервалу, проте вона може бути дещо меншою, ніж у інших методів, що працюють у реальному часі. Для збільшення точності роботи методу маємо покращити алгоритм вилучення аномальних значень, або ж підбирати коефіцієнти довірчого інтервалу індивідуально для кожного нового часового ряду.

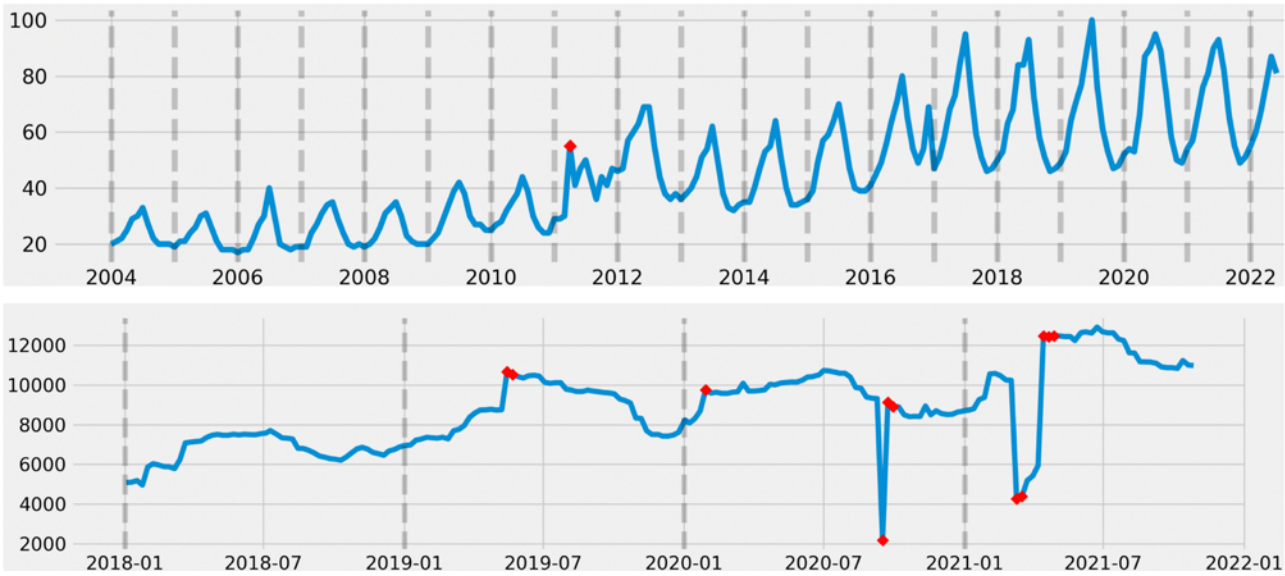

Рисунок 2.7 – Приклад роботи методу визначення аномалій на основі довірчого інтервалу значень перших різниць часового ряду у реальному часі при визначенні неявних аномалій, явних аномалій та аномальних підпослідовностей

2.4 Метод визначення аномалій на основі прогнозування з використанням машинного навчання

Насправді, найбільш ефективно використати можливості машинного навчання для виявлення аномалій можна виконуючи прогнозування. Тож суть цього методу полягає у тому, що алгоритм машинного навчання спочатку навчається на деякій частині часового ряду, а далі виконує прогнозування наступних значень. Висновок про те, чи є значення аномальним, робиться за тим фактом, чи входить значення у довірчий інтервал на основі прогнозованого значення. У програмній реалізації я використав бібліотеку Prophet [12] від розробників Facebook для прогнозування часових рядів за допомогою машинного навчання.

Тож метод визначення аномалій на основі прогнозування з використанням машинного навчання показав низьку ефективність виявлення неявних аномалій у порівнянні з іншими методами, що зумовлено невеликим розміром тестового часового ряду. Проте, з явними аномаліями і аномальними підпослідовностями

алгоритм впорався чудово. Також варто зазначити, що метод надає можливість простого вилучення аномальних значень шляхом їх заміни на прогнозовані.

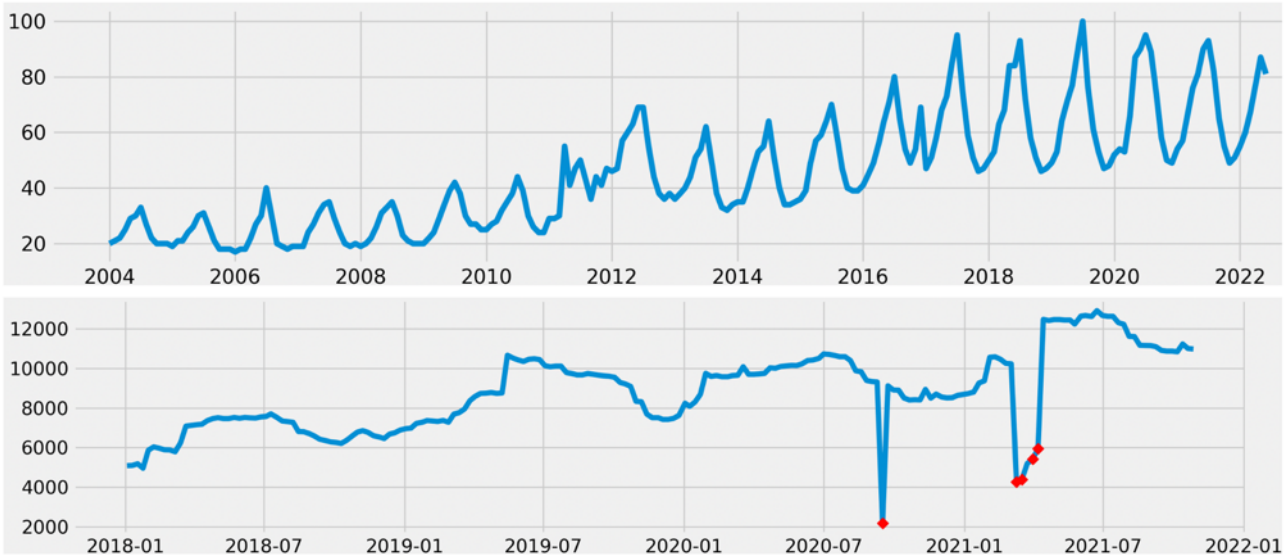

Рисунок 2.8 – Приклад роботи методу визначення аномалій на основі прогнозування з використанням машинного навчання при визначенні неявних аномалій, явних аномалій та аномальних підпослідовностей

2.5 Метод визначення аномалій на основі довірчого інтервалу значень показника Херста

Метою цієї роботи є також дослідження того, чи можна використовувати показник Херста часового ряду для виявлення аномалій, і якщо так – чи буде це ефективніше, ніж це роблять інші методи виявлення аномалій. Для цього спочатку маємо описати алгоритм знаходження показника Херста для часового ряду і реалізувати його програмно.

Тож нехай маємо часовий ряд $X = X_1, X_2, \dots, X_n$ розміру N. Також введемо A – деяку додатню константу, яку Гарольд Херст розрахував емпірично, і яка приблизно рівна $\frac{1}{2}$. Тоді алгоритм обрахунку показника Херста складатиметься з таких кроків:

1. Знайдемо середнє значення ряду $\bar{X}$:

$$\bar{X} = \frac{1}{N}\sum_{t=1}^{N} X_t$$

2. Створимо нормований ряд Y:

$$Y_t = X_t - \bar{X}$$

3. Обрахуємо середньоквадратичне відхилення нормованого ряду S:

$$S = \sqrt{\frac{1}{N}\sum_{t=1}^{N} Y_t}$$

4. Створимо ряд кумулятивних відхилень Z:

$$Z_t = \sum_{i=1}^{t} Y_i\,, \qquad t = 1,2,\dots,N$$

5. Обрахуємо ряд розмахів кумулятивних відхилень R:

$$R_t = \max\{Z_1, Z_2, \dots, Z_t\} - \min\{Z_1, Z_2, \dots, Z_t\}, \qquad t = 1,2,\dots,N$$

6. Пронормуємо ряд розмахів:

$$R/S$$

7. Знайдемо показник Херста H з рівняння:

$$\log(R/S) = Hlog(A) + Hlog(N)$$

Побудувавши графік залежності $\log(R/S)$ від $\log(N)$ як пряму $y = ax + b$, можемо визначити показник Херста H як тангенс кута нахилу отриманої прямої:

$$H = \frac{\log(R/S)}{\log(A) + \log(N)}$$

Насправді, цей алгоритм застосовується лише для часових рядів невеликого розміру. Для рядів, які містять тисячі значень, варто використовувати модифікацію описаного алгоритму [3]. Для цього маємо поділити часовий ряд на A рівних періодів розміру n та визнати R/S як середнє:

$$(R/S)_n = \frac{1}{A}\sum_{j=1}^{A}\left(R_j/S_j\right)_n$$

Таким чином, значення показника Херста буде дорівнювати тангенсу кута нахилу прямої лінійної регресії:

$$\log(R/S)_n = f(\log(n))$$

Програмна реалізація алгоритму обрахунку показника Херста для часового ряду засобами Python з використанням базових бібліотек матиме такий вигляд:

```
def get_hurst_exponent(time_series, lags=20):
    lags = range(2, lags)
    tau = [np.std(np.subtract(time_series[lag:], time_series[:-lag]))
for lag in lags]
    reg = np.polyfit(np.log(lags), np.log(tau), 1)
    return reg[0]
```

Приклад 2.1 – Реалізація функції обрахунку значень показника Херста для часового ряду засобами Python

Тут аргумент $lags$ відповідає за максимальний лаг при обрахунках. Якщо хочемо отримати значення показника Херста для довготривалої перспективи (враховуючи більше значень ряду одночасно) – маємо збільшити кількість лагів. Відповідно, чим більше значення максимального лагу – тим менше буде значення показника. Насправді, значення цього параметру найчастіше підбирається індивідуально для кожного конкретного часового ряду, зважаючи на його розмір і кількість періодів у сезоні.

Оскільки лише за одним значенням показника Херста для всього часового ряду неможливо визначити його аномалії, маємо побудувати ряд зі значень показника, кожне нове значення якого буде відповідати підряду вихідного ряду, що закінчується поточним елементом.

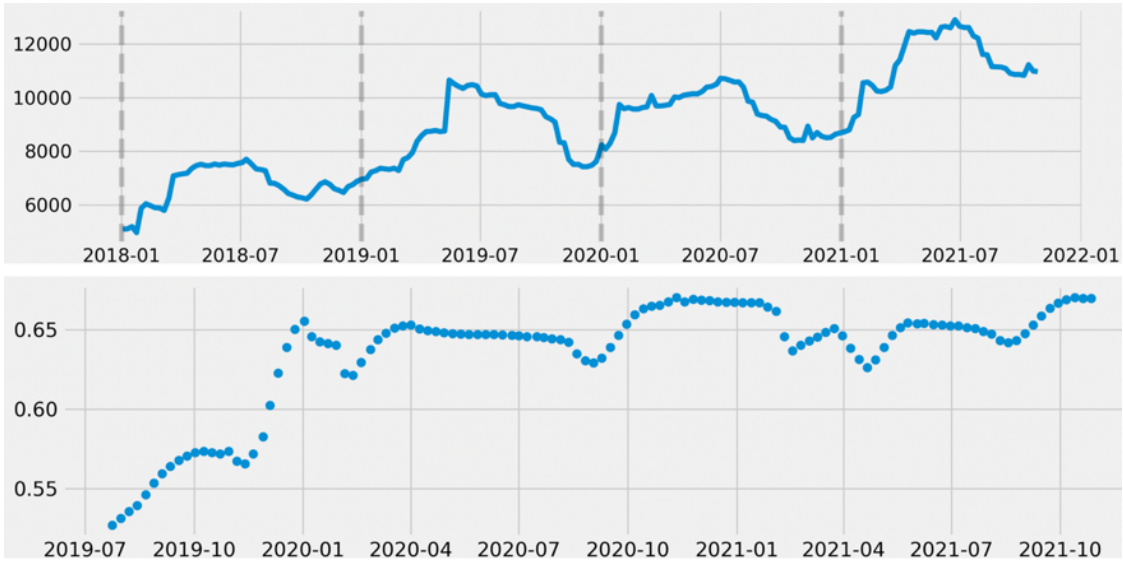

Рисунок 2.9 – Приклад обрахованого ряду значень показника Херста для часового ряду

Як бачимо, всі значення ряду показників Херста більші за $\frac{1}{2}$, що вказує на те, що часовий ряд зберігає свою поведінку, що відповідає дійсності. Тепер спробуємо побудувати ряд значень показника Херста для того ж часового ряду, але вже зі штучною явною аномалією.

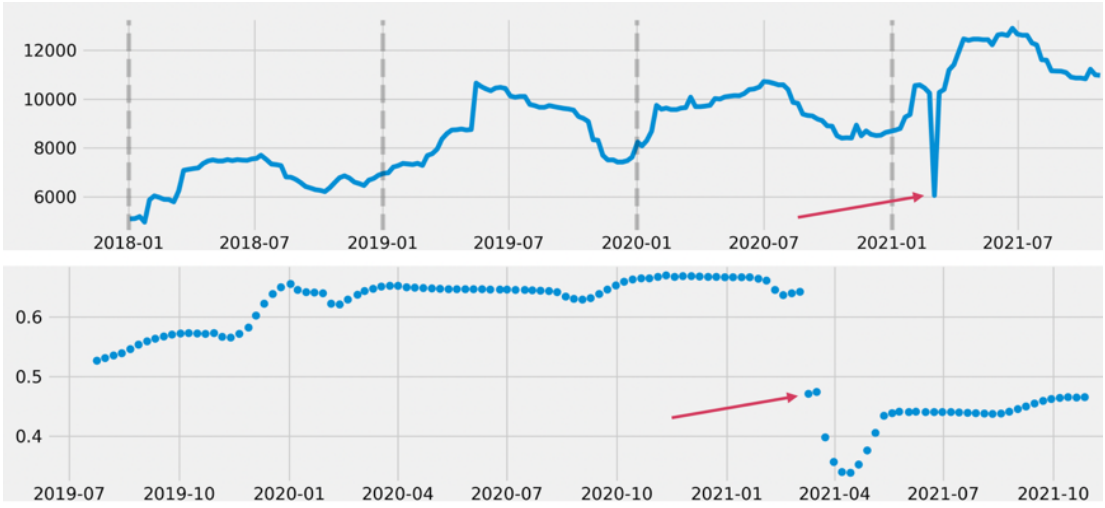

Рисунок 2.10 – Приклад обрахованого ряду значень показника Херста для часового ряду з явною аномалією

Аномалія виявилась настільки великою, що здатна була суттєво вплинути на значення показника Херста – після аномалії він став менший за $\frac{1}{2}$, що означає, що ряд змінив свою поведінку.

Отже, ми переконалися, що показник Херста доволі чутливий до аномалій у часовому ряді, принаймні явних. Втім, оскільки лише одна явна аномалія може кардинально вплинути на подальші значення показника Херста, має сенс реалізовувати метод визначення аномалій лише для роботи у реальному часі, при цьому одразу вилучаючи знайдені аномалії.

Отож, алгоритм методу складається з кількох кроків. На першому кроці будуємо довірчий інтервал для сформованого попередньо ряду з показників Херста. Далі додаємо до ряду нове значення і рахуємо показник вже для нього. На третьому кроці перевіряємо, чи входить нове значення показника до визначеного попередньо інтервалу, якщо так – вважаємо точку ряду, якій відповідає обраховане значення показника Херста, аномальною, і, нарешті, вилучаємо аномалію, щоб вона ні в якому разі не вплинула на подальші обрахунки показника. Звичайно, перед початком роботи методу необхідно надати йому достатньо велику частину часового ряду у якості затримки, щоб значення показника Херста були точними. Також маємо виділити ще певну затримку для створення ряду зі значень показника мінімального розміру, щоб можливо було побудувати довірчий інтервал.

Щоб проаналізувати ефективність методу, я поділив його на три складові:

- Обрахунок значення показника для даних ряду;
- Визначення того, чи є значення показника аномальним;
- Вилучення аномального значення.

Кожна з цих складових має багато варіантів реалізації. По-перше, існує декілька модифікацій алгоритму знаходження значень показника Херста, зокрема – модифікації для часових рядів різного розміру, які можуть повертати дещо різні значення показника. Також не варто забувати про визначення оптимальної кількості лагів. По-друге, в нашому випадку, створюючи з вихідного ряду ряд

значень показника Херста, ми лише намагаємось зробити аномалії більш явними, проте існує безліч методів саме їх виявлення, починаючи з найпростіших довірчих інтервалів, що були використані в реалізаціях методів, розглянутих в даній роботі раніше. І, нарешті, по-третє, є доволі багато способів заміни аномальних значень, наприклад – просте дублювання останнього значення ряду значно підвищує ефективність виявлення аномальних підпослідовностей, але знижує ефективність виявлення неявних аномалій. Таким чином, комбінації різних реалізацій перелічених кроків можуть показувати кардинально різні результати – і саме через це даний метод має великий потенціал розвитку. Власне, змінюючи реалізації описаних кроків можливо навіть підлаштовувати метод для виявлення аномалій саме визначеного типу.

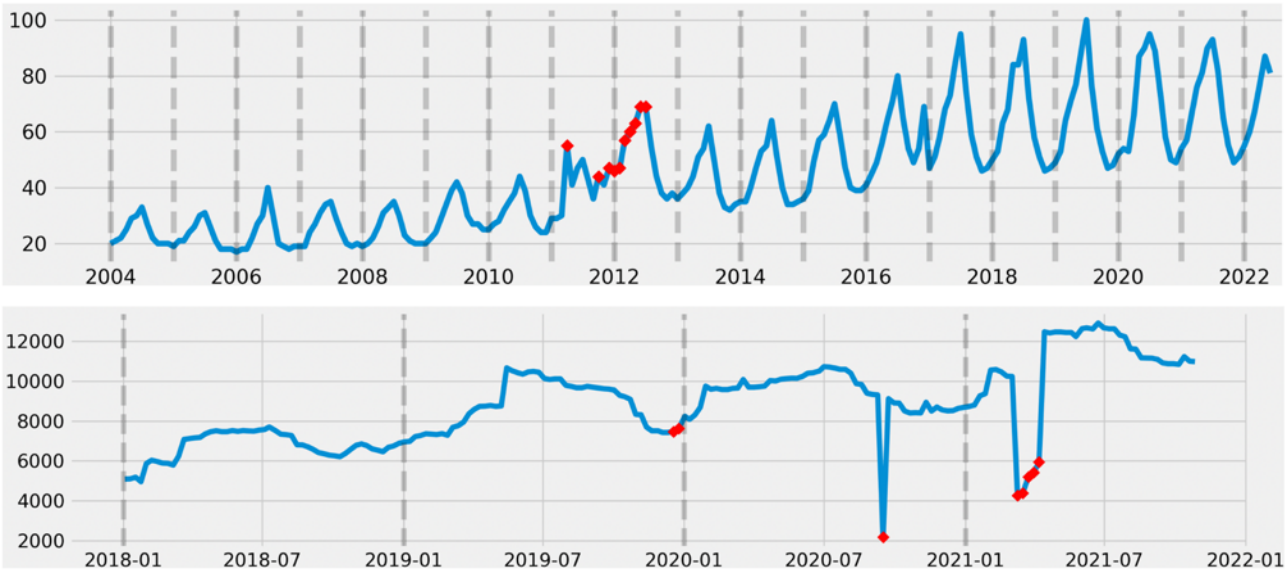

Рисунок 2.11 – Приклад роботи методу визначення аномалій на основі довірчого інтервалу значень показника Херста при визначенні неявних аномалій, явних аномалій та аномальних підпослідовностей

РОЗДІЛ 3: ПРАКТИЧНЕ ДОСЛІДЖЕННЯ І ПОРІВНЯННЯ МЕТОДІВ

Реалізувавши всі описані методи визначення аномалій, я оцінив ефективність кожного з них поодинці. Тепер настав час порівняти ці методи на прикладі реальних даних. І найбільш доступні та прості для розуміння дані, в яких можна шукати аномалії – це економічні дані: часовий ряд, що містить щоденну ціну акції компанії Google, до прикладу. Проте економісти і інвестори для оцінки дохідності акцій компаній і побудови своїх інвестиційних портфелів використовують деяку модифікацію оригінальних даних, що називається волатильністю лог-прибутків.

3.1. Обрахунок історичної волатильності лог-прибутків

Простими словами, волатильність – це мірило фінансового ризику, тобто міра коливання часових рядів – ціни акції компанії. У фінансовій статистиці волатильність – це показник, що характеризує тенденцію зміни ринкових цін і доходів впродовж певного часу. Історична ж волатильність – це міра варіації часового ряду минулих значень ринкової ціни. Оскільки для порівняння реалізованих методів визначення аномалій ми будемо використовувати історичні дані, надалі будемо працювати саме з історичною волатильністю.

Опишемо алгоритм розрахунку історичної волатильності лог-прибутків:

1. Нехай маємо часовий ряд $X = X_1, X_2, \dots, X_n$ щоденної ціни акції компанії Google за останні кілька років розміру N.
2. Спершу маємо обрахувати значення лог-прибутків R_t за формулою:

$$R_t = \ln\left(\frac{X_i}{X_{i-1}}\right) = \ln(X_i) - \ln(X_{i-1})$$

   По суті, отримаємо ряд перших різниць розміру $N - 1$.
3. Далі, оскільки волатильність – це стандартне відхилення, обрахуємо стандартне відхилення значень знайдених лог-прибутків. Проте, тут маємо ввести поняття вікна – кількості значень лог-прибутків, для яких

рахуватимемо стандартне відхилення. Зазвичай, в економіці, використовують такі значення: 21 день – це місяць, 63 дні – 3 місяці, 252 дні – один рік. Використаємо значення 21: обраховуватимемо волатильність даних за останній місяць. Тобто кожне значення нового ряду рахуватимемо за формулою:

$$Y_i = \sqrt{\frac{1}{20}\sum_{j=1}^{21}(R_i - \overline{R_{i-20} + \cdots + R_i})^2}$$

Отже, маємо ряд даних історичної волатильності щоденної ціни акції компанії Google за останні кілька років. Насправді, ще ні. Ми обрахували лише дані щоденної історичної волатильності. Сьогодні ж найчастіше використовується саме волатильність в річному обчисленні. Для цього маємо лише домножити отримані значення на $\sqrt{252}$, адже 252 – кількість днів у «торговому році». Також зазвичай отримані результати підносять у відсотках, але для нашої подальшої роботи можемо залишити їх абсолютними.

Програмна реалізація алгоритму обрахунку історичної волатильності лог-прибутків має наступний вигляд:

```
diff_data = np.log(data) - np.log(data.shift(1))
volatility = diff_data['Close']
             .rolling(21)
             .std().dropna()
             .apply(lambda x: x * np.sqrt(252))
```

Приклад 3.1 – Реалізація обрахунку історичної волатильності лог-прибутків засобами Python

Тепер спробуємо побудувати ряд історичної волатильності лог-прибутків історичних даних торгів акцій компанії Google.

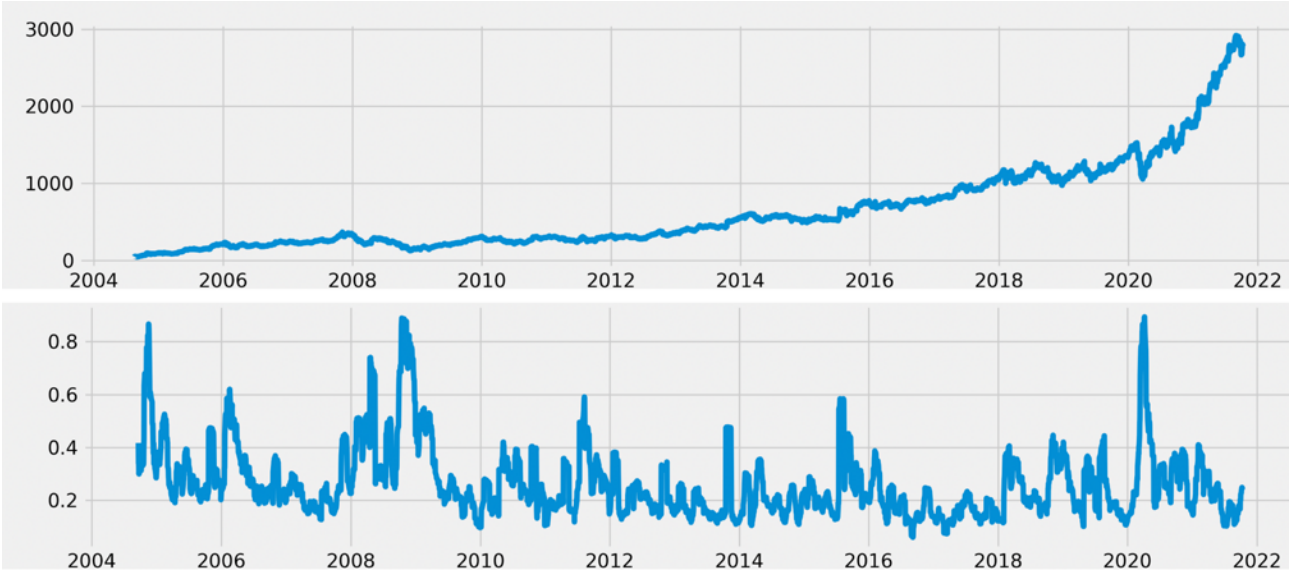

Рисунок 3.1 – Приклад обрахунку часового ряду значень історичної волатильності лог-прибутків для часового ряду ціни акції компанії Google

Як бачимо, значення волатильності дійсно відображають інтенсивність коливань ціни акції компанії Google, а що важливіше – набагато краще підходять для виявлення у них аномалій. Можемо бачити всі типи аномалій: точкові – явні і неявні, а також аномальні підпослідовності невеликого розміру.

3.2 Порівняння методів визначення аномалій

Отже, для порівняння описаних раніше методів виявлення аномалій у часових рядах я використав дані історичної волатильності, обраховані у попередньому підрозділі. Проте, для порівняння будемо використовувати дані, починаючи з 2010 року, адже протягом перших 6 років торгів акціями компанії їх ціна досить сильно коливалась, що негативно позначиться на якості виявлення аномалій.

Почнемо з методу визначення аномалій на основі довірчого інтервалу даних компоненти залишків часового ряду. У методу лише один параметр – кількість періодів у сезоні. Я обрав значення 21, адже це кількість торгових днів у одному календарному місяці. Як можемо бачити, метод виявив аномалії всіх

типів. Проте я схиляюсь до висновку, що точність їх виявлення є невисокою, адже, до прикладу, не всі значення явних аномальних підпослідовностей визначені, як аномальні. Для порівняння, кількість виявлених аномалій – 67.

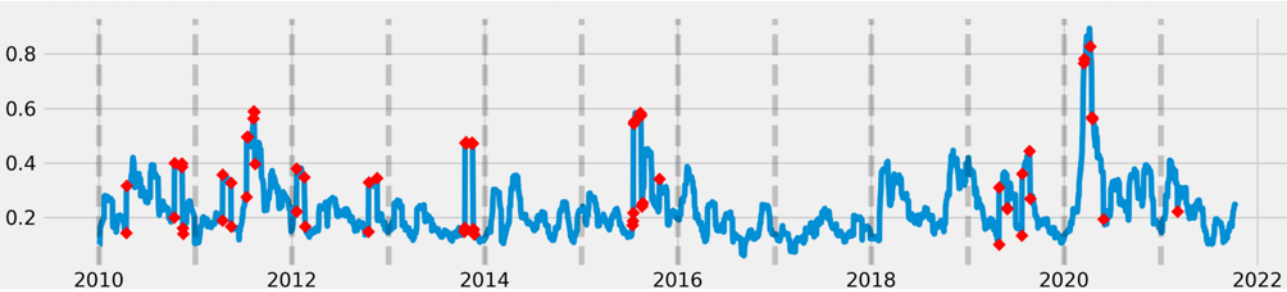

Рисунок 3.2 – Результат роботи методу визначення аномалій на основі довірчого інтервалу даних компоненти залишків часового ряду на реальних даних

Наступною до порівняння долучена модифікація методу, заснованого на довірчому інтервалі даних компоненти залишків часового ряду, для роботи у реальному часі. У неї два параметри: кількість періодів у сезоні і затримка перед початком роботи, їх значення я задав у розмірі 21 і 300 відповідно. Результат роботи модифікації майже не відрізняється від оригінального методу. Виявлено 51 аномальне значення. Зазначу, що червоною лінією на графіку позначено значення, з якого метод почав визначення аномалій.

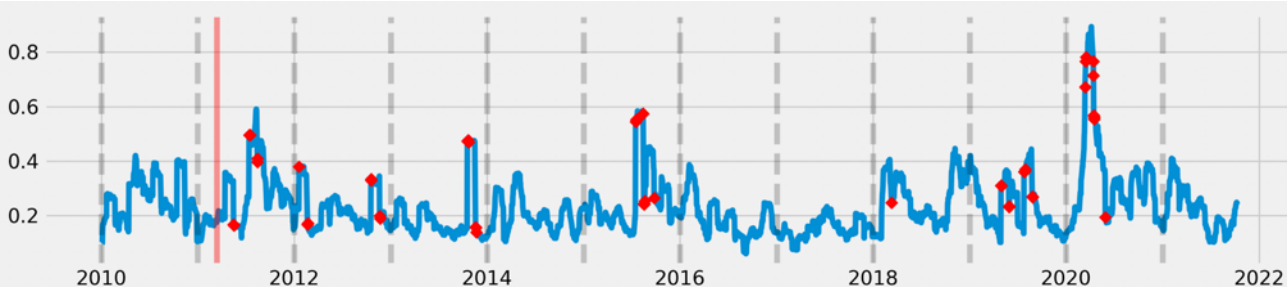

Рисунок 3.3 – Результат роботи модифікації методу на основі довірчого інтервалу даних компоненти залишків часового ряду для роботи у реальному часі на реальних даних

Рухаємось до методу, заснованого на довірчому інтервалі міжквартильного розмаху періодичних даних. Параметр у нього також один –кількість періодів у сезоні, який також заданий у розмірі 21. Бачимо найгірший результат серед усіх: метод, через свої особливості, не виявив жодної аномалії, що підтверджує попередній висновок про його вкрай низьку ефективність.

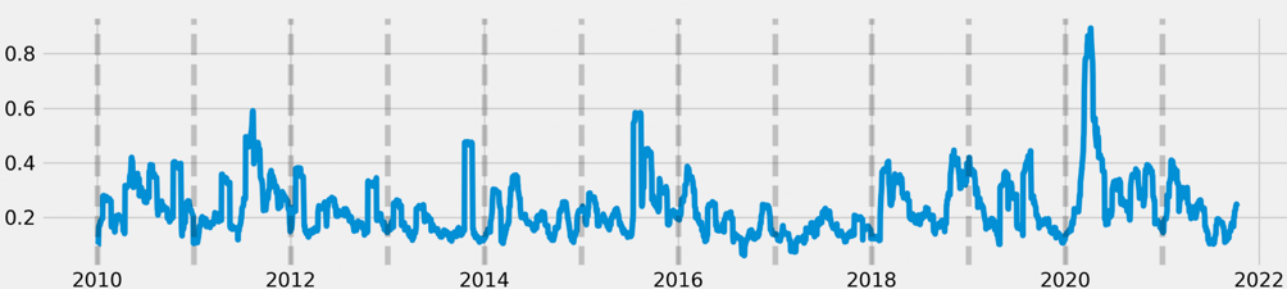

Рисунок 3.4 – Результат роботи методу визначення аномалій на основі довірчого інтервалу міжквартильного розмаху періодичних даних на реальних даних

Модифікація методу, заснованого на довірчому інтервалі міжквартильного розмаху періодичних даних, для роботи у реальному часі протестована з такими параметрами: кількість періодів у сезоні – 21, затримка – 14 сезонів. У порівнянні з оригінальною версією модифікація має значно вищу ефективність: виявлено кілька явних і неявних точкових аномалій, одну аномальну підпослідовність. На мою думку, метод показав вищу точність, ніж попередні, адже виявлені аномалії вже не повторюються регулярно і не мають певної послідовності. Варто зазначити, що метод, у порівнянні з попередніми, позначив усі значення останньої аномальної підпослідовності як викиди. Кількість виявлених аномальних значень – 26.

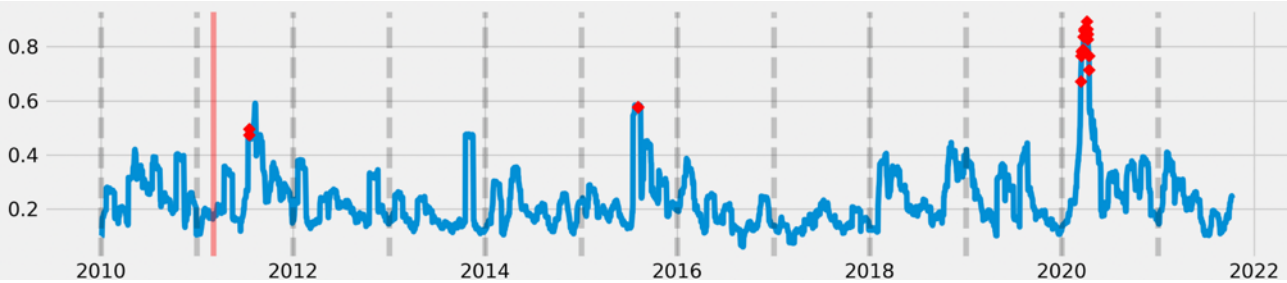

Рисунок 3.5 – Результат роботи модифікації методу на основі довірчого інтервалу міжквартильного розмаху періодичних даних у реальному часі на реальних даних

Метод визначення аномалій на основі довірчого інтервалу значень перших різниць часового ряду не має параметрів. Знову виявлені аномалії мають певний паттерн, що ставить під сумнів точність їх виявлення. Варто відмітити, що метод не знайшов жодної аномальної підпослідовності, включно з найбільш явною останньою. Кількість виявлених аномалій – 55.

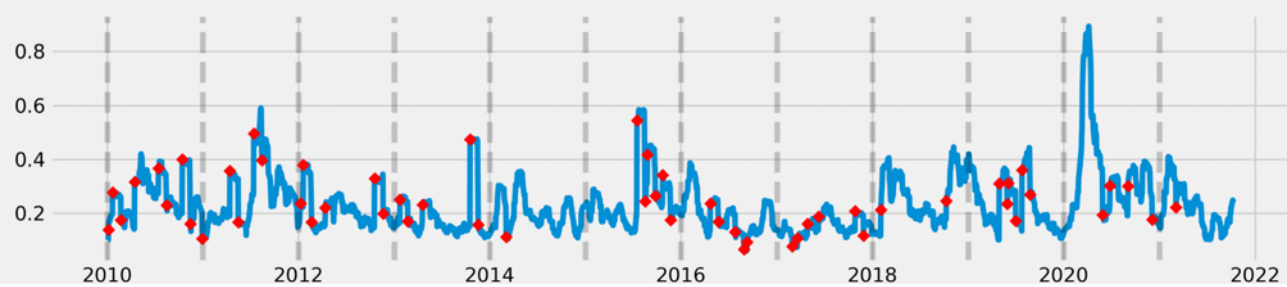

Рисунок 3.6 – Результат роботи методу визначення аномалій на основі довірчого інтервалу значень перших різниць часового ряду на реальних даних

Модифікація методу визначення аномалій на основі довірчого інтервалу значень перших різниць часового ряду у реальному часі протестована з затримкою у розмірі 300 значень. Ця модифікація показала результат практично ідентичний до оригінального методу. Кількість виявлених аномалій – 76.

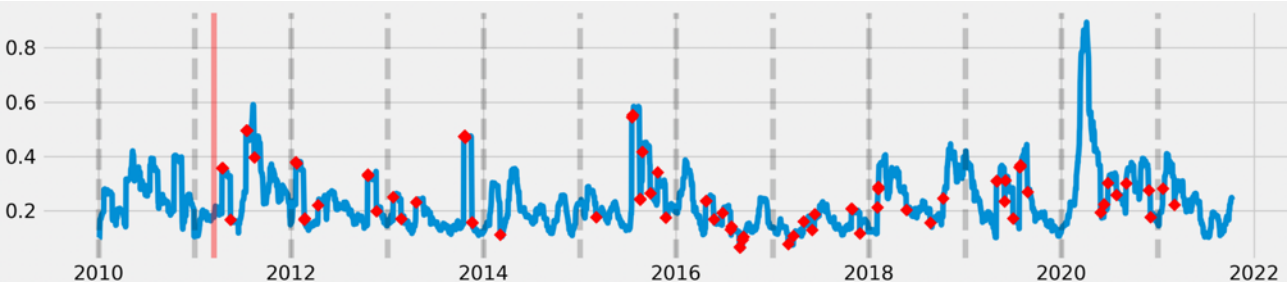

Рисунок 3.7 – Результат роботи модифікації методу на основі довірчого інтервалу значень перших різниць часового ряду для роботи у реальному часі на реальних даних

Метод визначення аномалій, заснований на прогнозуванні з використанням машинного навчання має один параметр – розмір тренувального ряду, тож я задав його у розмірі 300 значень. Як бачимо, поведінка цього методу відрізняється від попередніх. Позначено аномальними лише значення однієї аномальної підпослідовності. Втім, варто позначити якість виявлення значень цієї підпослідовності. Це робить роботу метод на основі машинного навчання схожою на роботу модифікації методу на основі довірчого інтервалу міжквартильного розмаху періодичних даних. Кількість виявлених аномальних значень – 22.

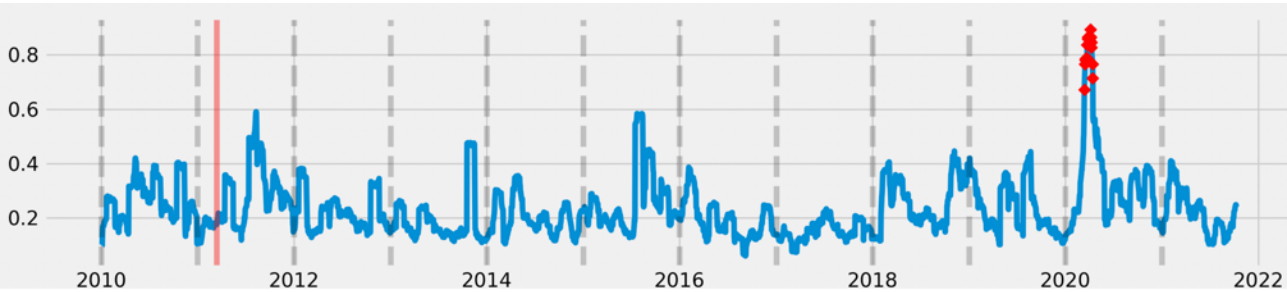

Рисунок 3.8 – Результат роботи методу визначення аномалій на основі прогнозування з використанням машинного навчання на реальних даних

І, нарешті, метод визначення аномалій на основі довірчого інтервалу значень показника Херста має два параметри: затримка – 500 значень, максимальна кількість лагів – 21. Результат методу вражає – виявлена до цього нова аномальна підпослідовність. Також доволі цікаво виявлено частину значень

останнього значного викиду і майже всі значення невеликих аномальних підпослідовностей протягом всього періоду. На мою думку, саме цей алгоритм показав найкращий результат. Кількість виявлених аномалій – 182.

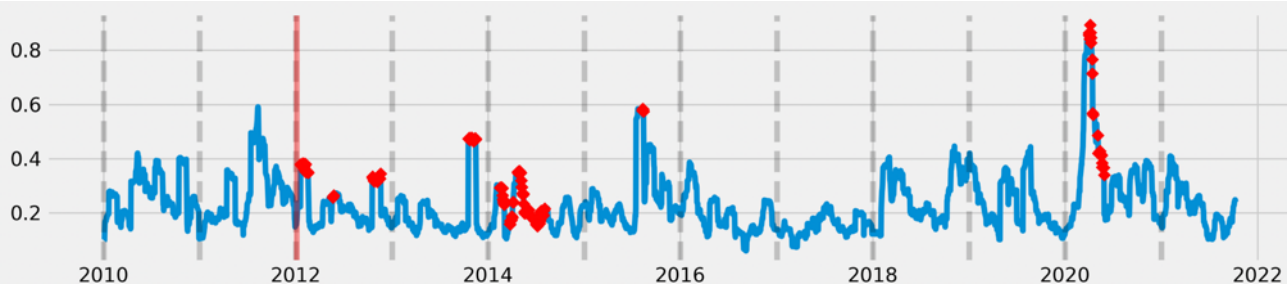

Рисунок 3.9 – Результат роботи методу визначення аномалій на основі довірчого інтервалу значень показника Херста на реальних даних

Отож, я хотів би виділити три методи, на які варто звернути увагу:

- Модифікація методу визначення аномалій на основі довірчого інтервалу міжквартильного розмаху періодичних даних для роботи у реальному часі;
- Метод визначення аномалій на основі прогнозування з використанням машинного навчання;
- Метод визначення аномалій на основі довірчого інтервалу значень показника Херста.

Ці методи, на мою думку, показали одні з найкращих результатів, і заслуговують подальшого, тепер вже статистичного, порівняння.

ВИСНОВКИ

1. У роботі було застосовано загальновідомі методи визначення аномалій до часових рядів: метод на основі довірчого інтервалу даних компоненти залишків, метод на основі довірчого інтервалу міжквартильного розмаху періодичних даних, метод на основі довірчого інтервалу значень перших різниць часового ряду, а також метод на основі прогнозування з використанням машинного навчання.
2. Було розроблено модифікації деяких методів для роботи у реальному часі. Знайдено їх переваги і недоліки у порівнянні з оригінальними версіями.
3. Запропоновано для визначення аномалій використати показник Херста – міру довготривалої пам’яті часового ряду.
4. Було з’ясовано, що найкраще себе показали саме модифікація методу на основі довірчого інтервалу міжквартильного розмаху періодичних даних для роботи у реальному часі, метод на основі прогнозування з використанням машинного навчання, а також запропонований метод визначення аномалій у часових рядах на основі довірчого інтервалу значень показника Херста.

Метод з використанням показника Херста має великий потенціал до розвитку через свою модульність і, разом із зазначеними методами, заслуговує додаткового статистичного порівняння.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Основы стохастической финансовой математики / А. Н. Ширяев : ФАЗИС, Москва, 1998.
2. The Analysis of Time Series: An Introduction. Sixth Edition / C. Chatfield : Chapman & Hall, 2004.
3. Застосування методу фрактального аналізу для визначення трендових характеристик числових рядів / Тур Г. І., Трунова О. В. : Вісник Чернігівського національного педагогічного університету. Режим доступу: http://nbuv.gov.ua/UJRN/VchdpuP_2015_125_61
4. Time Series Analysis and Its Applications With R Examples / Robert H/ Shumway, David S. Stoffer, 2004
5. Network Anomaly Detection Based on the Statistical Self-similarity Factor / Paweł Dymora, Mirosław Mazurek : Lecture Notes in Electrical Engineering 324(1):271-287, 2015. Режим доступу: https://www.researchgate.net/publication/283026441_Network_Anomaly_Detection_Based_on_the_Statistical_Self-similarity_Factor
6. Real-Time Time Series Anomaly Detection / Marco Cerliani, 2020. Режим доступу: https://towardsdatascience.com/real-time-time-series-anomaly-detection-981cf1e1ca13
7. Популярність запиту «ice cream» в Google, дані за період 2004-2022 рр. Режим доступу: https://trends.google.com/trends/explore?date=all&q=ice cream
8. Mongolia Meat Price Time Series. Weekly market prices of meat in Ulaanbaatar, Mongolia, from 2018-2021. Режим доступу: https://www.kaggle.com/datasets/robertritz/ub-meat-prices
9. Google Stock Price (All Time). Daily Google Stock Price, right from its IPO (19 Aug 2004). Режим доступу: https://www.kaggle.com/datasets/akpmpr/google-stock-price-all-time

10. Hurst Parameter Based Anomaly Detection for Intrusion Detection System / S. Yu, Pauline Koh, H. Kim : IEEE International Conference on Computer and Information Technology (CIT), 2016. Режим доступу: https://www.semanticscholar.org/paper/Hurst-Parameter-Based-Anomaly-Detection-for-System-Yu-Koh/d146fef6fc0ddd9aa204b11bbf08565c8e51c462
11. Бібліотека Prophet від розробників Facebook для прогнозування часових рядів з використанням машинного навчання. Режим доступу: https://facebook.github.io/prophet/