

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
Імені ІГОРЯ СІКОРСЬКОГО»
НАВЧАЛЬНИЙ-НАУКОВО ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра інформаційної безпеки

До захисту допущено
Завідувач кафедри
_____ **Дмитро ЛАНДЕ**
(підпис)
«___» _____ 2024р.

Дипломна робота
на здобуття ступення бакалавра
за освітньо-професійною програмою «Системи, технології та математичні
методи кібербезпеки»
спеціальності: 125 «Кібербезпека»

на тему: Розробка алгоритму виявлення аномалій в наборах даних для попередження кібератак з застосуванням вейвлет-аналізу

Виконав: здобувач вищої освіти IV курсу, групи ФБ-02
Замрій Денис Сергійович

(підпис)

Керівник: професор, завідувач кафедри ІБ,
Ланде Дмитро Володимирович

(підпис)

Рецензент: доцент кафедри ММЗІ, канд. фіз.-мат. наук,
Хмельницький Микола Олексійович

(підпис)

Засвідчую, що у цій дипломній роботі немає
запозичень з праць інших авторів без
відповідальних посилань.
Здобувач вищої освіти _____

Київ – 2024 року

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
Імені ІГОРЯ СІКОРСЬКОГО»
НАВЧАЛЬНИЙ-НАУКОВО ФІЗИКО-ТЕХНІЧНИЙ ІНСТИТУТ
Кафедра інформаційної безпеки

Рівень вищої освіти – перший (бакалаврський)

Спеціальність – 125 «Кібербезпека»

Освітньо-професійна програма «Системи, технології та математичні методи кібербезпеки»

ЗАТВЕРДЖУЮ
Завідувач кафедри

_____ Дмитро ЛАНДЕ
(підпис)
«___» _____ 2024р.

ЗАВДАННЯ
на дипломну роботу здобувачу вищої освіти
Замрій Денис Сергійович

1. Тема роботи: Розробка алгоритму виявлення аномалій в наборах даних для попередження кібератак з застосуванням вейвлет-аналізу.

Керівник роботи : професор, завідувач кафедри ІБ, Ланде Д.В.

Затверджені наказом по університету від «31» травня 2024р. № 2251-с

2. Термін подання здобувачем вищої освіти роботи 07 червня 2024р.

3. Вихідні дані: практична реалізація та тестування розробленого алгоритму.

4. Зміст роботи: застосування вейвлет-аналізу для виявлення аномалій у часових рядах, що містять дані про кібератаки, включаючи збір та попередню обробку даних, побудову часових рядів, та застосування вейвлет-перетворення для виявлення аномалій.

5. Перелік ілюстративного матеріалу (із зазначенням плакатів, презентацій, тощо) :
презентація на захист дипломної роботи.

6. Дата видачі завдання: 21 вересня 2023р.

Календарний план

№ з/п	Назва етапів виконання дипломної роботи	Термін виконання етапів дипломної роботи	Примітка
1	Обговорення та визначення теми дипломної роботи	30.09.2023	виконано
2	Формування мети та завдання роботи	15.10.2023-29.12.2023	виконано
3	Ознайомлення з літературою	30.12.2023-04.02.2024	виконано
4	Написання першого та другого розділів	05.02.2024-11.04.2024	виконано
5	Розробка алгоритму для виявлення аномалій	12.04.2024-19.05.2024	виконано
6	Проходження переддипломної практики	15.04.2024 - 19.05.2024	виконано
7	Написання третього розділу диплому та висновків	20.05.2024 – 02.06.2024	виконано
8	Створення презентації до передзахисту	03.06.2024 – 05.06.2024	виконано
9	Передзахист дипломної роботи	07.06.2024	виконано
10	Доопрацювання дипломної роботи	07.06.24-19.06.24	виконано
11	Захист дипломної роботи	20.06.2024	виконано

Здобувач вищої освіти
(підпис)

_____ Денис ЗАМРІЙ
(власне ім'я, ПРІЗВИЩЕ)

Керівник роботи
(підпис)

_____ Дмитро ЛАНДЕ
(власне ім'я, ПРІЗВИЩЕ)

РЕФЕРАТ

Дипломна робота має обсяг 46 сторінок, містить 17 рисунків, 2 таблиці, 2 додатки, 10 джерел посилань.

Зловмисники щоденно вдосконалюють свої методи для здійснення кібератак, що значно ускладнює їх своєчасне виявлення. З початком повномасштабного вторгнення в Україну у лютому 2022 року кількість кібератак стрімко зросла. Найбільш поширеними типами атак були DDos-атаки, які спрямовані на відмову в обслуговуванні, шкідливе програмне забезпечення, таке як віруси, трояни. Також зловмисники використовували фішингові електронні листи, для того щоб заволодіти обліковими даними користувачів.

В загальному, хакери використовували широкий спектр кібератак, для того щоб завдати максимальної шкоди Україні в умовах повномасштабної війни.

Окрім технічно складних атак, зловмисники також використовували методи соціальної інженерії, для того, щоб ввести українських користувачів в оману. Фішингові електронні листи та підроблені веб-сайти, які були замасковані під державні організації, підштовхували людей до переходу за шкідливим посиланням. Через недостатній досвід більшості користувачів, хакери з легкістю отримували доступ до корпоративних або особистих даних.

Таким чином, обмежена обізнаність та навички кібергігієни серед частини населення України стала вразливістю, яку активно експлуатували хакери. Підвищення рівня цифрової грамотності та навчання безпечної поведінки в інтернеті могли б значно ускладнити успіхи подібних соціальних інженерних атак.

В умовах війни, українські користувачі стикалися з великим обсягом інформації про те, як захиститися від кібератак, проте, в умовах інформаційного перенавантаження та постійного стресу, не встигали ознайомитися з цими рекомендаціями завчасно. Що в свою чергу ускладнювало зусилля з протидії

кіберзагрозам та робило певну частину користувачів більш вразливими перед тактиками зловмисників.

Ефективним рішенням цієї проблеми може бути виявлення аномалій у великих наборах даних та використання вейвлет-аналізу.

Метою роботи є розробка алгоритму виявлення аномалій у наборах великих даних для попередження кібератак із використанням вейвлет-аналізу.

Об'єктом дослідження є власне дані, які включають в себе інформацію про кібератаки.

Ключові слова: ВЕЙВЛЕТ-АНАЛІЗ, АНОМАЛІЇ, НАБОРИ ВЕЛИКИХ ДАНИХ, КІБЕРАТАКИ, ЧАСОВІ РЯДИ.

ABSTRACT

The thesis is 46 pages long, contains 17 figures, 2 tables, 2 appendices, and 10 references.

Attackers are improving their methods of cyberattacks on a daily basis, which makes them much more difficult to detect in a timely manner. Since the beginning of the full-scale invasion of Ukraine in February 2022, the number of cyberattacks has increased dramatically. The most common types of attacks were DDos attacks aimed at denial of service, malicious software such as viruses and trojans. Attackers also used phishing emails to take over user credentials. In general, hackers used a wide range of cyberattacks to inflict maximum damage on Ukraine in a full-scale war.

In addition to technically sophisticated attacks, the attackers also used social engineering techniques to mislead Ukrainian users. Phishing emails and fake websites disguised as government organizations encouraged people to click on malicious links. Due to the lack of experience of most users, hackers easily gained access to corporate or personal data.

Thus, the limited awareness and cyber hygiene skills of a part of the Ukrainian population became a vulnerability that was actively exploited by hackers. Increasing the level of digital literacy and teaching safe behavior on the Internet could significantly complicate the success of such social engineering attacks.

During the war, Ukrainian users were exposed to a large amount of information on how to protect themselves from cyberattacks, but due to information overload and constant stress, they did not have time to read these recommendations in advance. This, in turn, complicated efforts to counter cyber threats and made some users more vulnerable to attackers' tactics. An effective solution to this problem can be to detect anomalies in large data sets and use wavelet analysis.

The aim of the study is to develop an algorithm for detecting anomalies in big data sets to prevent cyberattacks using wavelet analysis.

The object of research is the data itself, which includes information about cyberattacks.

Keywords: WAVELET ANALYSIS, ANOMALIES, BIG DATA SETS, CYBERATTACKS, TIME SERIES.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ.....	9
ВСТУП.....	10
1 ТЕОРЕТИЧНІ ВІДОМОСТІ ЧАСОВИХ РЯДІВ	11
1.1 Часові ряди	11
1.2 Дослідження та аналіз трендів у часових рядах	13
1.3 Виявлення сезонності та циклічності у часових рядах	15
ВИСНОВКИ ДО РОЗДІЛУ 1	18
2 ВЕЙВЛЕТ-АНАЛІЗ.....	19
2.1 Теоретичні основи вейвлет-перетворення	19
2.2 Застосування вейвлет-аналізу для аналізу часових рядів	21
2.3 Методи виявлення аномалій на основі вейвлет-аналізу	23
2.4 Вибір оптимальних вейвлет-функцій та рівнів декомпозиції.....	24
ВИСНОВКИ ДО РОЗДІЛУ 2.....	27
3 РЕАЛІЗАЦІЯ АЛГОРИТМУ ВИЯВЛЕННЯ АНОМАЛІЙ	28
3.1 Збір та попередня обробка даних	28
3.2 Побудова часових рядів	31
3.3 Застосування вейвлет-аналізу.....	35
ВИСНОВКИ ДО РОЗДІЛУ 3	37
ВИСНОВКИ	38
ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ	39
ДОДАТКИ	40
ДОДАТОК А ЛІСТИНГ ПРОГРАМНОГО КОДУ.....	41
ДОДАТОК Б ПОЛЯ В НАБОРАХ ДАНИХ	46

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ПЗ – Програмне Забезпечення

ACF - Автокореляція

CWT – Безперервне вейвлет перетворення

OSINT - Open Source Intelligence

ВСТУП

Актуальність роботи. Згідно зі статистикою, кількість кібератак в Україні значно зросла з початком повномасштабного вторгнення в лютому 2022 року. Зловмисники активно використовували різноманітні методи атак, такі як DDoS-атаки, шкідливе ПЗ та фішингові електронні листи, для завдання максимальної шкоди. Тому потрібно підвищувати рівень цивільної грамотності українців та вживанням інших заходів для протидії кіберзагрозам в умовах війни.

Тому виникає необхідність створити та впровадити алгоритм, який буде допомагати відслідковувати активність кібератак в певний період часу.

Метою роботи є створення алгоритму, який буде полегшувати виявлення та протидію атакам зловмисників, що в свою чергу допоможе більше захистити дані користувачів.

Об'єктом дослідження є процес набору даних, що містять інформацію про кібератаки, з різних джерел.

Предмет дослідження – методи виявлення аномалій у наборах великих даних з використанням вейвлет-аналізу.

Методи дослідження. У процесі досліджень використовувались: літературний аналіз джерел за обраною темою бакалаврської дипломної роботи, аналіз механізмів виявлення аномалій, дослідження існуючих атак зловмисників.

Новизна отриманих результатів.

1. Подальшого розвитку отримав метод виявлення аномалій у наборах великих даних з використанням вейвлет-аналізу, який, на відміну від існуючих застосовується для аналізу даних у кібербезпеці.

Практична цінність отриманих результатів. Практична цінність одержаних результатів полягає в тому, що на основі отриманих в бакалаврській дипломній роботі теоретичних положень запропоновано методи та розроблено алгоритми для виявлення аномалій у наборах великих даних.

1 ТЕОРЕТИЧНІ ВІДОМОСТІ ЧАСОВИХ РЯДІВ

1.1 Часові ряди

У сучасних дослідженнях властивостей складних систем, у тому числі експериментальних, широко використовується підхід, заснований на аналізі сигналів від цих систем. Це особливо вірно, коли важко або неможливо описати досліджуваний процес математично, але ми можемо виміряти деякі характеристики цього процесу. Тому аналіз систем, у тому числі експериментальних, часто проводиться за допомогою обробки сигналів. Цей метод дослідження часто називають реконструкцією динамічних систем, а відповідний розділ теорії динамічних систем відомий як аналіз часових рядів.

Часовий ряд — це послідовність значень деякої змінної (або змінних), які реєструються безперервно або через певні проміжки часу. Однак варто зазначити, що наявність лише часових рядів обмежує наше розуміння досліджуваної системи, оскільки не забезпечує повного розв'язку математичних рівнянь, що описують цю систему. Такі обмеження роблять метод реконструкції складним завданням [1].

На рисунку 1.1 зображено приклад часового ряду.

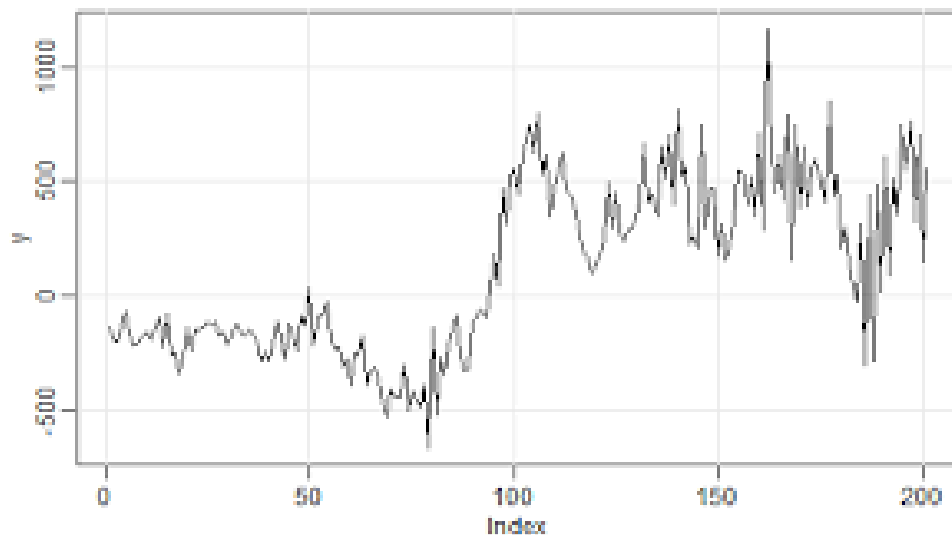


Рисунок 1.1 – Приклад часового ряду

Аналіз часових рядів – це специфічний підхід до аналізу послідовності даних, зібраних за певний період часу. Це вимагає послідовних інтервалів запису даних, що дозволяє зрозуміти динаміку та зв'язки між точками даних. Цей підхід відрізняється від простого збору даних у випадкові або періодичні моменти часу. Аналіз часових рядів дозволяє визначити закономірності, тенденції, сезонність та інші характеристики системи, які неможливо виявити за допомогою інших методів аналізу даних [2].

Одним із головних аспектів аналізу часових рядів є те, що він дозволяє вам вивчати, як змінні змінюються з часом. Час є важливою змінною, оскільки він показує, як дані змінюються в різні моменти часу, і відображає кінцеві результати. Це надає додаткову інформацію та дозволяє встановлювати зв'язки між даними.

Для ефективного аналізу часових рядів зазвичай потрібна велика кількість точок даних, щоб забезпечити послідовність і надійність результатів. Велика кількість даних забезпечує репрезентативний розмір вибірки та дозволяє ідентифікувати тенденції, викиди та сезонні коливання. Крім того, дані часових рядів можна використовувати для прогнозування майбутніх значень на основі історичних даних.

Таким чином, аналіз часових рядів допомагає виявити динаміку даних, виявити закономірності та залежності та зробити прогнози на основі історичних даних.

Аналіз часових рядів існує дві основні задачі, такі як ідентифікація та прогноз. Завдання ідентифікації в аналізі часових рядів дозволяє отримати відповіді про системні параметри, які сформували цей часовий ряд, зокрема, розмірність вбудовування, кореляції, ентропію тощо. Розмірність інвестицій визначає мінімальну кількість динамічних змінних, які однозначно описують досліджуваний процес.

Натомість завдання прогнозування має на меті передбачити майбутні значення вимірюваних характеристик досліджуваного об'єкта за наявними спостереженнями, тобто скласти прогноз на певний період часу. В даний час розроблено та обґрунтовано кілька різних методів прогнозування, які можна розділити на два основних класи: локальні та глобальні.

Часовий ряд може містити умовні значення, які, як правило, не мають самостійного значення, а використовуються для подальших розрахунків.

За кількістю показників динамічні ряди поділяються на комплексні та ізольовані. Ізольований ряд містить лише один показник, а комплексний ряд складається із системи узагальнюючих показників.

Залежно від рівномірності інтервалів між рівнями ряду часові ряди поділяються на повні та неповні. Повний ряд має рівні інтервали між датами або інтервалами, а неповний ряд має нерівні інтервали.

Рівні часових рядів можуть бути виражені як абсолютні, відносні та середні значення. Основними є часові ряди абсолютних значень, а похідними є часові ряди відносних і середніх значень. Часові ряди абсолютних значень дають більш повну картину динаміки процесу чи явища, а часові ряди відносних значень можуть відображати зростання чи зниження показників, зміну частки показників у загальній сукупності чи зміну інтенсивності окремих явищ.

1.2 Дослідження та аналіз трендів у часових рядах

Аналіз трендів часових рядів є важливою складовою економічного аналізу та прогнозування.

Він допомагає виявити основні тенденції розвитку процесів і явищ у часі, що дає змогу приймати зважені рішення та робити прогнози на майбутнє.

Одним з ключових аспектів аналізу трендів є визначення самого тренду. Тренд – це загальна тенденція зміни рівнів динамічного ряду протягом певного періоду.

Для виявлення тенденції використовуються різні методи, такі як аналітична апроксимація, згладжування, декомпозиція тощо.

На рисунку 1.2 відображено можливість графіку приймати різні форми, такі як лінійна, експоненціальна, логарифмічна або поліноміальна.

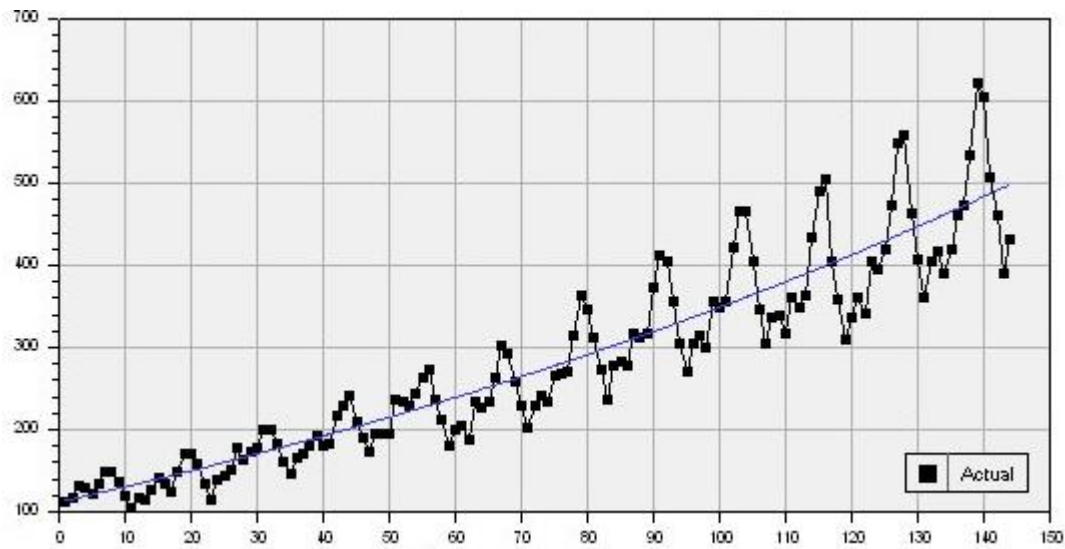


Рисунок 1.2 – Наявність тренду на часовому ряді

Аналітична апроксимація є одним із методів, який використовується для ідентифікації та опису тенденції рівнів часових рядів. Цей метод передбачає побудову математичної моделі, яка найкраще відповідає динаміці даних. Залежно від характеру даних, це може бути лінійна або нелінійна регресія, яка дозволяє аналізувати та прогнозувати тенденцію в часовому ряді.

Для усунення випадкових коливань часових рядів і виділення основної тенденції використовується метод згладжування. На рисунку 1.3 метод усуває шум і випадкові відхилення, що допомагає краще сприймати основну тенденцію. До таких методів згладжування належать ковзне середнє, експоненціальне згладжування, адаптивне ковзне середнє та інші. Вони допомагають створювати більш стабільні та зрозумілі зображення трендів, що полегшує аналіз і прогнозування даних.

Таким чином, методи аналітичної апроксимації та згладжування є ефективними для ідентифікації та аналізу тенденцій у часових рядах. Вони дозволяють побудувати

математичну модель і усунути випадкові коливання, що допоможе зрозуміти основну тенденцію та зробити прогноз на основі історичних даних.

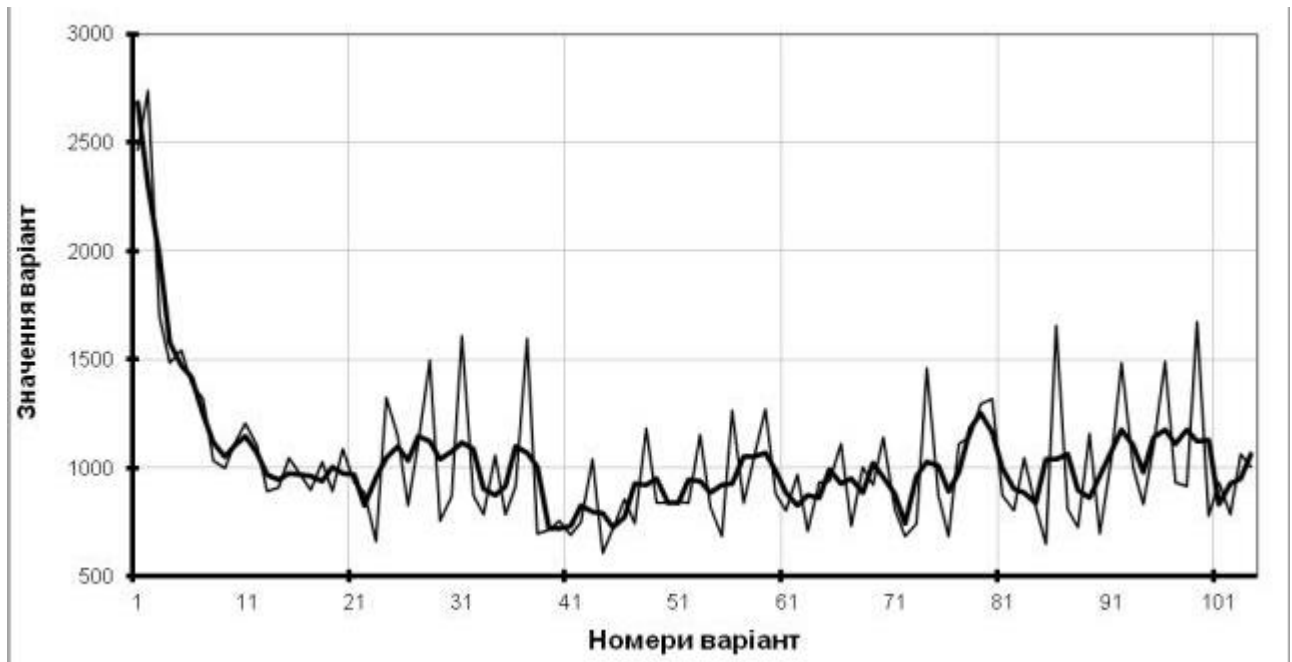


Рисунок 1.3 – Згладжування часового ряду

Декомпозиція часового ряду — це процес розкладання ряду на складові елементи: тренд, сезонність, циклічність і випадковість. Це дозволяє більш точно аналізувати та прогнозувати поведінку часового ряду.

Аналіз трендів часових рядів широко використовується в різних галузях. Він допомагає виявити закономірності та тенденції розвитку процесів, що важливо для прийняття ефективних управлінських рішень та розробки стратегій розвитку [3].

1.3 Виявлення сезонності та циклічності у часових рядах

Виявлення сезонності та циклічності в часових рядах є важливим аспектом аналізу часових рядів, оскільки це допомагає краще зрозуміти закономірності та структуру даних. Сезонність - це регулярна і повторювана зміна в часовому ряді, яка пов'язана з певними сезонами або періодами часу [4]. На рисунку 1.4 відображено

можливість кібератак збільшуватись під час святкових періодів або в певний період року.

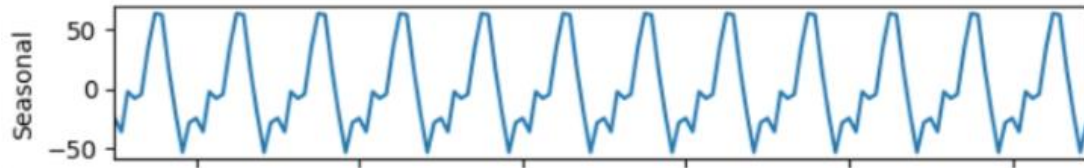


Рисунок 1.4 – Сезонність часового ряду

Для виявлення сезонності використовуються різні методи, такі як автокореляційний аналіз, індекси сезонності, декомпозиція часових рядів тощо.

Автокореляційний аналіз допомагає визначити ступінь залежності між рівнями ряду, розділених певним інтервалом часу, що може вказувати на наявність сезонності.

На рисунку 1.5 відображено індекси сезонності – це коефіцієнти, які показують відхилення сезонних рівнів від середнього значення ряду.

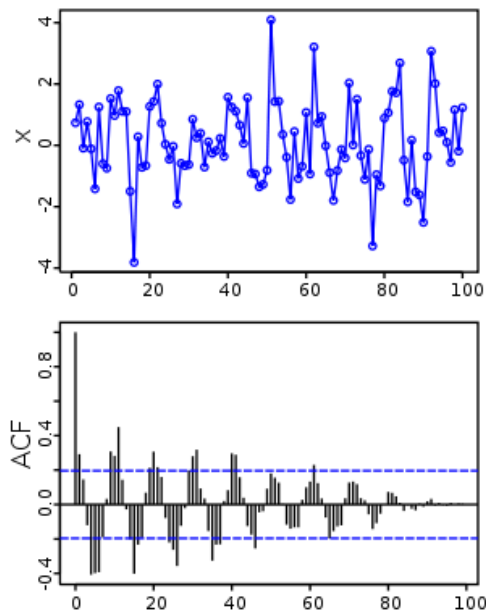


Рисунок 1.5 – Автокореляція функції

Декомпозиція часових рядів дозволяє розкласти ряд на складові елементи: тренд, сезонність, циклічність і випадковість, що полегшує ідентифікацію сезонних коливань.

Циклічність — це триваліші та менш регулярні коливання в часовому ряді, пов'язані з економічними, політичними чи соціальними факторами. Циклічність може мати різну тривалість і амплітуду, і її важко виявити безпосередньо з часового ряду. Для аналізу циклічності використовуються такі методи, як декомпозиція часових рядів, автокореляційний аналіз і спектральний аналіз.

Виявлення сезонності та циклічності у часових рядах допомагає краще зрозуміти закономірності та структуру даних, що є важливим для прийняття ефективних управлінських рішень та розробки стратегій розвитку [5].

ВИСНОВКИ ДО РОЗДІЛУ 1

Аналіз часових рядів є важливим для вивчення та розуміння динаміки різних показників за певний період. Виявлення основних класифікаційних ознак часових рядів дозволяє більш ефективно підбирати методи їх дослідження та інтерпретації отриманих результатів.

Зокрема, поділ рядів на умовні, складні/ізолювані, повні/неповні, а також абсолютні, відносні та середні ряди дає змогу всебічно охарактеризувати досліджуваний процес чи явище в його динаміці. Найбільш повну характеристику дають ряди абсолютних величин, а ряди відносних і середніх показників дозволяють виявити зміни інтенсивності, темпів зростання/зниження, структурних зрушень тощо.

Поглиблений аналіз часових рядів з урахуванням їх класифікаційних ознак є важливою передумовою для прийняття обґрунтованих рішень і розробки ефективних стратегій алгоритмів для прогнозування кіберзагроз.

2 ВЕЙВЛЕТ-АНАЛІЗ

2.1 Теоретичні основи вейвлет-перетворення

Вейвлет-перетворення є потужним математичним інструментом, який широко використовується для аналізу даних. На відміну від класичного перетворення Фур'є, вейвлети дозволяють одночасно отримувати інформацію як про частотний, так і про часовий склад сигналу.

Це особливо корисно для аналізу нестационарних сигналів, які містять різкі зміни або локалізовані події. Традиційне перетворення Фур'є не здатне ефективно обробляти такі характеристики, оскільки воно надає лише інформацію про частоту, втрачаючи інформацію про час.

Короткочасне перетворення Фур'є вирішує цю проблему, але має нижчу частотну роздільну здатність. Вейвлет-перетворення забезпечує хороший баланс між роздільною здатністю частоти та часу, що робить його ефективним інструментом для аналізу складних сигналів в обох областях одночасно.

Це дозволяє глибше зрозуміти фізичні процеси, які породжують такі нестационарні сигнали, що робить вейвлети корисними в багатьох областях, від обробки сигналів до аналізу зображень.

Безперервне вейвлет-перетворення визначається як згортка сигналу $x(t)$ як:

$$W(c, \tau) = \int_{-\infty}^{\infty} x(t) \psi_{c, \tau}^*(t) dt \quad (2.1)$$

де $w(S, \tau)$ представляє вейвлет-коефіцієнт, який відповідає частоті зі шкалою S і часом τ .

CWT схоже на узгодження шаблону для кореляції між сигналом та попередньо визначеною формою хвилі, як це відбувається у випадку узгодженого фільтра.

Перевага CWT над узгодженим фільтром полягає в особливій властивості вейвлет-перетворення, яке розкладає сигнал як у частотній, так і в часовій областях.

Безперервне вейвлет-перетворення створює велику складність та надмірність, оскільки воно аналізує сигнал на дуже великій кількості частот. Для вирішення цієї

проблеми було введено дискретне вейвлет-перетворення, яке дозволяє аналізувати та розкласти сигнал лише на певні дискретні значення.

На рисунку 2.1 зображено приклад вейвлет-перетворення.

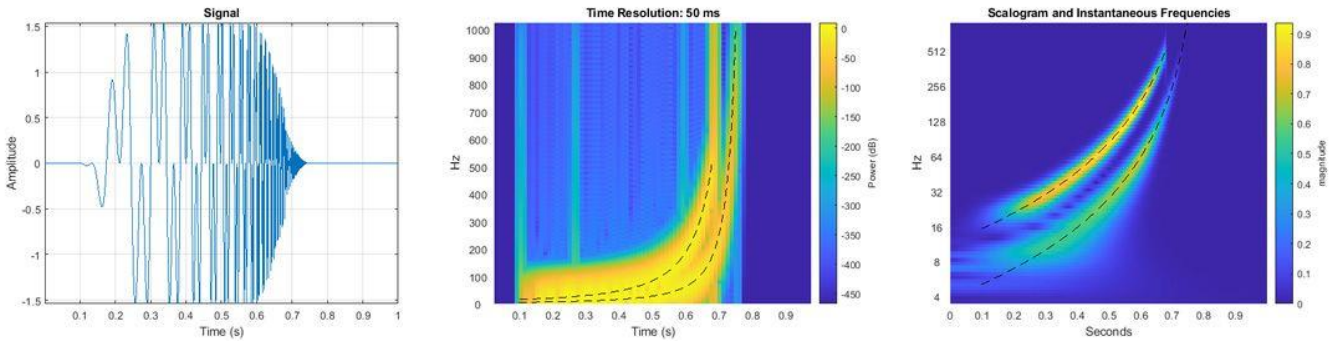


Рисунок 2.1 – Приклад вейвлет-перетворення

Перевагою використання вейвлет-аналізу є забезпечення локалізації в часовому просторі, що дозволяє точно виявити аномалії частотному діапазоні. Вейвлет-аналіз найбільш ефективний у випадках, коли потрібно виявити локальні особливості поведінки процесу, що досліджується. За допомогою вейвлет-аналізу можна ефективно адаптуватися до сигналів змінної частоти, за допомогою вейвлетів [6].

Вейвлети – це короткі хвилі з нулевим інтегральним значенням і з відображенням по осі незалежної змінної, здатних до зсуву по цій осі. Технологія використання вейвлетів дозволяє виявляти одиничні та нерегулярні «сплески», різкі зміни значень кількісних показників у різні періоди часу.

Часовий ряд можна описати за допомогою кривої, її ж можна розкласти на суму гармонічних коливань різної частоти та амплітуди. Коливання з низькою частотою відповідають за плавні зміни у значеннях часового ряду, а високочастотні коливання - за короткі. Чим більше змінюється значення часового ряду на певному масштабі, тим більшою є амплітуда компонентів на відповідній частоті. Таким чином, часовий ряд, що вивчається, можна розглядати у частотно-часовому просторі, який залежить від обох цих параметрів - часу та частоти.

Вейвлет-перетворення сигналу є узагальненням ряду Фур'є за системою базисних функцій, і побудовані з вихідного вейвлету $\psi(t)$, який має певні властивості завдяки операцій зсуву в часі (b) і зміни часового масштабу (f)

$$\Psi_{ab} = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) \quad (2.2)$$

Для того, щоб початкова функція стала вейвлетом, вона повинна володіти потрібно знати ознаки якими вона має володіти.

Обмеженість

$$\|\psi\|^2 = \int_{-\infty}^{\infty} |\psi(t)|^2 dt < \infty \quad (2.3)$$

Локалізація, адже вейвлет-перетворення використовує локалізовану вихідну функцію як в часі так і в частоті, для цього мають виконуватися умови:

$$|\psi(t)| \leq C(1+|t|)^{-1-\epsilon} \quad (2.4)$$

$$|f_{\psi}(\omega)| \leq C(1+|\omega|)^{-1-\epsilon} \quad (2.5)$$

Нульове середнє. Графік функції має бути знакозмінним навколо нуля на осі часу і мати нульову площу:

$$\int_{-\infty}^{\infty} |\psi(t)| dt = 0 \quad (2.6)$$

Проте, інколи для цього потрібен не тільки нульовий момент дорівнював нулю, а й інші n моментів:

$$\int_{-\infty}^{\infty} t^n |\psi(t)| dt = 0 \quad (2.7)$$

2.2 Застосування вейвлет-аналізу для аналізу часових рядів

Вейвлети можуть бути використані, головним чином, для аналізу часових рядів, надаючи більш точніших часово-частотних характеристик. У свою чергу, аналіз частотно-часових характеристик може використовуватися в самому моделюванні за допомогою гармонічної моделі, допомагаючи у виборі порядку регресії.

На другому етапі аналіз часових рядів допоможе визначити оптимальне перетворення часового ряду. Результати перетворення можуть бути використані для вилучення інформації, пов'язаної з частотою, і зменшити складність вхідного простору, але також можуть бути використані для виконання багаторозв'язного моделювання.

Багатомасштабний характер вейвлет-аналізу дозволяє розкласти часовий ряд на компоненти різних масштабів (частот). Що дає можливість вивчати поведінку ряду на різних часових горизонтах: від високочастотних, короткочасних коливань до низькочастотних, довгострокових трендів. Цей багатомасштабний підхід особливо корисний для розуміння складних динамічних процесів, які проявляються в різних часових масштабах [7].

На відміну від аналізу Фур'є, який надає інформацію про частотний склад сигналу, але втрачає інформацію про час, вейвлет-аналіз зберігає часову локалізацію ознак. Це дає змогу виявляти локальні зміни, аномалії, стрибки та інші нестационарні події в часовому ряді. Такий підхід дозволяє краще зрозуміти внутрішню структуру та динаміку часових рядів.

Вейвлет-аналіз може виявити приховані періодичні компоненти в часових рядах, навіть якщо вони замасковані через певні проміжки часу. Це досягається завдяки можливості аналізувати ряди в різних масштабах одночасно. Виявлення таких прихованих періодів може бути корисним для прогнозування, розуміння циклічності процесів.

Вейвлет-перетворення можна використовувати для фільтрації шумових компонентів у часовому ряду. Вибіркова вибірка окремих вейвлет-коефіцієнтів дозволяє виділити та реконструювати потрібні компоненти сигналу, наприклад, тренд, сезонні коливання. Такі операції згладжування та фільтрації можуть бути корисними для підготовки даних для подальшого аналізу чи моделювання.

Вейвлет-аналіз добре підходить для вивчення нестационарних, нелінійних процесів, які змінюють свої статистичні характеристики з часом. На відміну від

аналізу Фур'є, вейвлети здатні слідкувати за еволюцією частотного спектру з часом. Це дозволяє виявляти та аналізувати нестационарні особливості, такі як зміни дисперсії, поява/зникнення періодичності [8].

Загалом вейвлет-аналіз є надзвичайно гнучким інструментом, який дозволяє виявляти та характеризувати різноманітні структурні особливості часових рядів на різних часових масштабах.

2.3 Методи виявлення аномалій на основі вейвлет-аналізу

Методи виявлення аномалій на основі вейвлет-аналізу включають аналіз вейвлет-коефіцієнтів у різних масштабах. Високі значення коефіцієнта в певних масштабах можуть вказувати на наявність аномалій у даних. Такий підхід дозволяє визначити локальні зміни та сигналізувати про особливості, які можуть бути викликані аномаліями.

Інший підхід полягає у використанні порогових механізмів для виявлення аномалій. У цьому випадку вейвлет-коефіцієнти порівнюються з певним порогом, і якщо значення коефіцієнта перевищує цей поріг, то відповідна частина сигналу вважається аномальною. Порогові механізми можуть бути м'якими або жорсткими залежно від того, як обробляються коефіцієнти, що перевищують поріг [9].

Також можна використовувати методи виявлення аномалій, засновані на порівнянні вейвлет-перетворень двох або більше сигналів. Наприклад, можна порівняти вейвлет-коефіцієнти сигналу з опорним сигналом або з коефіцієнтами інших пов'язаних сигналів. Якщо відмінності між коефіцієнтами перевищують певний рівень, можна зробити висновок про наявність аномалії.

Крім того, вейвлет-аналіз можна використовувати для побудови моделей даних і прогнозування їх поведінки. У цьому випадку аномалії можна виявити шляхом порівняння прогнозованих значень з фактичними даними. Якщо відхилення між прогнозами та фактичними даними перевищує певний рівень, то можна припустити, що дані містять аномалію [10].

2.4 Вибір оптимальних вейвлет-функцій та рівнів декомпозиції

Вибір вейвлет-функцій є важливим етапом при проведенні вейвлет-аналізу часових рядів. Різні вейвлет-функції мають свої характерні властивості та особливості, що обумовлюють їх придатність для дослідження конкретних типів сигналів.

2.4.1 Wavelet Mexican Hat

Mexican Hat Wavelet, також відома як хвильова функція Ріккера, є типом хвильової функції, який широко використовується в обробці сигналів і аналізі зображень. Названа вона так через характерну форму, яка нагадує мексиканський капелюшок (див. рис. 2.2).

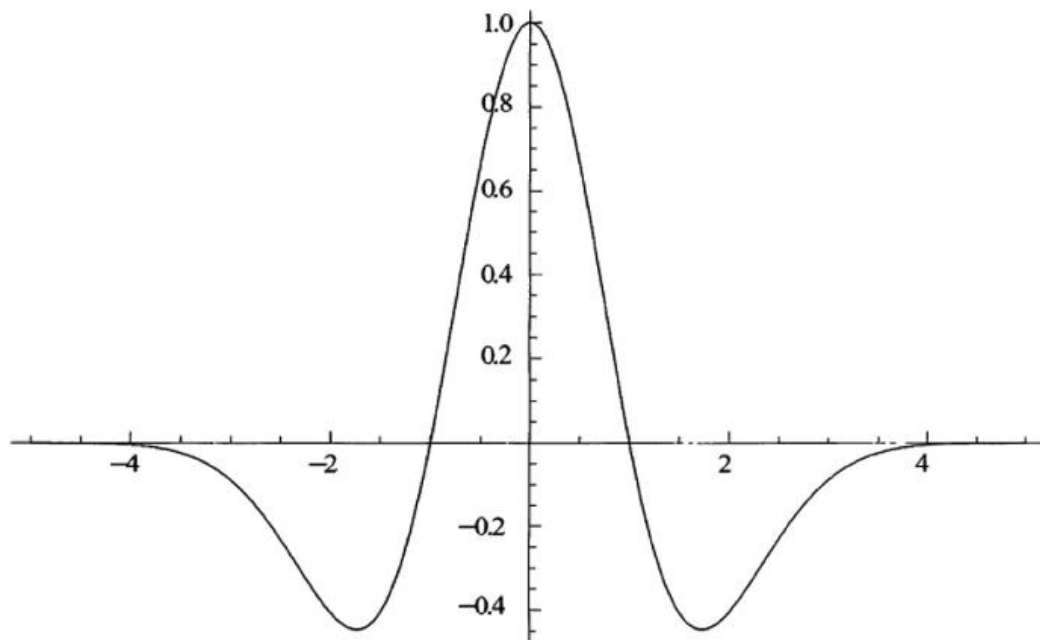


Рисунок 2.2 – Mexican Hat Wavelet

Mexican Hat визначається як друга похідна функції Гауса. Математично це можна записати так:

$$\psi(t) = (1 - t^2) * \exp\left(\frac{-t^2}{2}\right) \quad (2.8)$$

де t представляє часову або просторову змінну. Ця хвильова функція має нульове середнє, симетричну форму відносно осі y з додатними та від'ємними частинами та обмежену підтримку, тобто вона відрізняється від нуля лише на кінцевому інтервалі. Її часто використовують як смуговий фільтр для виявлення особливостей у певних масштабах.

Також ця функція може бути дуже корисною для виявлення та аналізу моделей та аномалій у часових рядах даних щодо кількості кібератак протягом певного періоду. Ця функція дозволяє здійснювати багатомасштабний аналіз, виявляючи характерні риси на різних часових масштабах.

2.4.2 Wavelet Morle

Вейвлет Морлі може бути корисним інструментом в області кібербезпеки, особливо для аналізу часових рядів даних, пов'язаних з кібератаками (див. рис. 2.3).

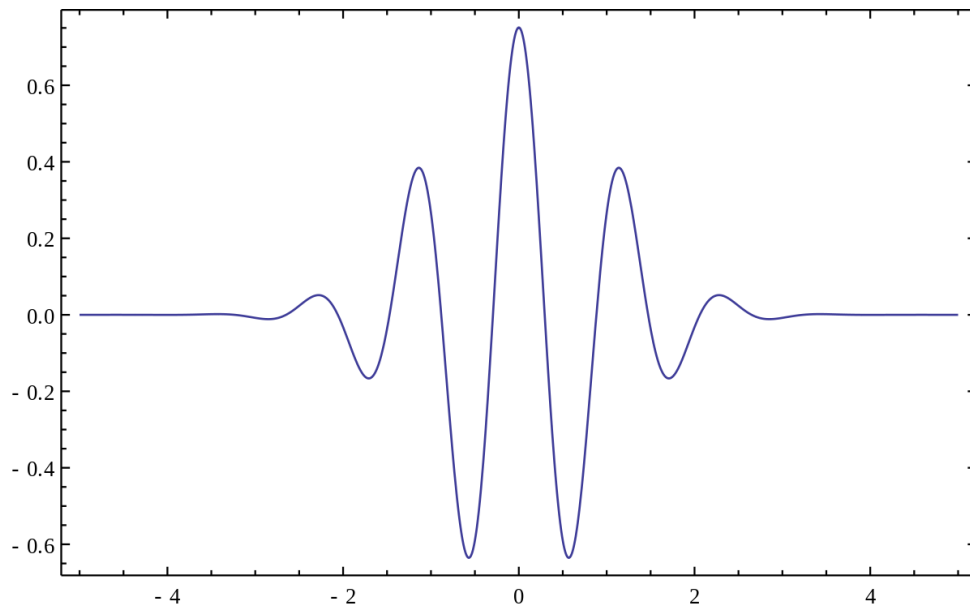


Рисунок 2.3 – Wavelet Morle

Він дозволяє проводити багатомасштабний аналіз даних про кількість кібератак за певний період, що допомагає виявляти характерні піки, тренди та аномалії на різних

часових масштабах, які можуть вказувати на появу нових видів загроз або зміну активності злоумисників.

Вейвлет Морлі також може використовуватись як смуговий фільтр для виділення специфічних частотних компонентів у даних про кібератаки, що дозволяє фокусуватись на певних масштабах або діапазонах частот, які є найбільш значущими для аналізу кіберзагроз. Характеристики, отримані з використанням вейвлета Морлі, можуть бути використані як входи в моделі машинного навчання для прогнозування та виявлення кіберзагроз, оскільки багатомасштабний аналіз забезпечує більш інформативні ознаки для таких моделей.

2.4.3 Порівняння вейвлетів Mexh та Morle

У таблиці 2.1 наведено порівняльні характеристики для вейвлетів Mexican Hat та Morle.

Таблиця 2.1 – Порівняння вейвлетів Mexh та Morle

Характеристика	Mexh	Morle
Локалізація	Добре локалізований в часовій області, але менше в частотній	Більше локалізований в частотній області, менш локалізований в часовій
Симетрія	Симетричний	Асиметричний
Швидкість спадання	Швидке спадання	Повільне спадання
Застосування	Виявлення локальних особливостей сигналу	Виявлення частотно-часових особливостей

Загалом, вейвлет Mexh є більш локалізований в часовій області і підходить для виявлення локальних особливостей сигналу, а вейвлет Morle, краще підходить для аналізу сигналів з гармонічною структурою та частотно-часових характеристик.

ВИСНОВКИ ДО РОЗДІЛУ 2

Вейвлет-аналіз є надзвичайно ефективним методом виявлення та аналізу аномалій у широкому діапазоні даних завдяки своїм унікальним властивостям. Інваріантність масштабу дозволяє аналізувати аномалії різного розміру, від локальних до глобальних. Тимчасова локалізація забезпечує точне визначення моменту виникнення аномалії, на відміну від аналізу Фур'є. Частотна локалізація дозволяє відокремити аномалії, викликані різними причинами, які проявляються на різних частотах. Ці переваги роблять вейвлет-аналіз потужним інструментом для виявлення, дослідження та розуміння аномалій у різноманітних типах даних – фінансових, біомедичних, геофізичних та інших. Завдяки своїм унікальним характеристикам вейвлет-аналіз виявився одним із найефективніших методів аналізу даних і виявлення аномалій.

3 РЕАЛІЗАЦІЯ АЛГОРИТМУ ВИЯВЛЕННЯ АНОМАЛІЙ

3.1 Збір та попередня обробка даних

Правильно зібрана та оброблена інформація – ключ до якісного проведення аналізу, та отримання потрібних результатів. Одним з можливих способів зібрати дані є налаштування автоматичного парсингу даних з сайтів, де може міститися інформація про будь-які кібератаки.

Проте з такою інформацією не завжди зручно працювати, тому перше, що потрібно зробити – це оформити дані в зручний для аналізу вигляд. Чудовим рішенням є використання JSON-формату та регулярних виразів, для цього була використана вбудована в Python бібліотека `re`, яка є потужним інструментом для роботи з регулярними виразами, що дає можливість робити різні операції з текстом.

Структура JSON дозволяє чітко розділити дані на потрібні нам елементи, наприклад ‘title’, ‘content’, ‘date’, ‘source’. Це значно допомагає при подальшому аналізі.

На рисунку 3.1 зображена структура оброблених даних.

Тепер ми маємо можливість аналізувати ті дані, які нам потрібно. Проте це ще тільки початковий етап, і дані потрібно очистити шуму, та інформації яка не впливає на подальший аналіз.

```
{
  "title": "Хакери з РФ хотіли проникнути в мережі НАТО і армій країн Східної Європи",
  "content": "Тетяна Кушнерик Сьогодні, 05:44 glavcom.ua\nАтаку організувала ,
  ,група хакерів під назвою Coldriver\nфото з відкритих ,
  ,джерел\nКиберзловмисники атакували сайт НАТО\nЗафіксована спроба ,
  ,російських хакерів проникнути в мережі НАТО і збройні сили деяких ,
  ,східноєвропейських країн. Про це повідомляє звіт групи аналізу загроз ,
  ,Google.\nЗазначається, що група хакерів, що базується в РФ, під назвою ,
  ,Coldriver (ще називається Calisto) проводила атаки, націлені на збройні ,
  ,сили кількох країн Східної Європи, а також на Центр передового досвіду ,
  ,НАТО.\nЦі кампанії були відправлені з використанням нещодавно створених ,
  ,облікових записів Gmail на облікові записи, відмінні від Google, тому ,
  ,ймовірність успіху цих кампаній невідома", - йдеться у звіті.\nУ Google ,
  ,не виявили під час цих атак жодного скомпрометованого облікового запису ,
  ,Gmail.\nЯк повідомили в групі аналізу загроз про шкідливу кіберактивність, ,
  ,перед тим організація Coldriver (Calisto) \nзапустила кампанії з фішингу ,
  ,облікових даних, націлені на кілька НУО та аналітичних центрів, що ,
  ,базуються в США, збройних сил балканської країни та оборонного підрядника, ,
  ,який базується в Україні".\nНагадаємо, за місяць Росія здійснила понад 3 ,
  ,тис. кібератак на українські сервіси.\n",
  "date": "Главком 2022.03.31 06:05",
  "source": "https://glavcom
  .ua/techno/telecom/hakeri-z-rf-hotili-proniknuti-v-merezhi-nato-i-armiy
  -krajin-shidnoji-jevropi-834381.html"
}
```

Рисунок 3.1 – структура даних для аналізу.

Також можна зберігати дані в форматі .csv для того, щоб порівняти швидкість обробки файлу і розміри файлу.

Таблиця 3.1 – Порівняння JSON та CSV

Назва файлу	Швидкість обробки	Розмір
data.json	0.294 с	19500 КБ
data.csv	0.332	19003 КБ

Як бачимо, Json-файл має більшу швидкість обробки даних, незважаючи на більший розмір.

На рисунку 3.2 зображено кінцевий результат обробки даних.

Date	Title	Content	Source
2022-03-31 06:05:00	Хакери з РФ хотіли проникнути в мережі НАТО і ...	Тетяна Кушнерик Сьогодні, 05:44 glavcom.ua\nАт...	https://glavcom.ua/techno/telecom/hakeri-z-rf-...
2022-03-30 15:37:00	30.03.2022 15:02 Українців попереджають про ел...	Українцям почали надходити листи на електронну...	https://glavpost.ua/news/ukrainsiv-poperedzha...
2022-03-29 16:49:00	Зафіксована кібератака на державні органи Укра...	29 марта 2022 г., 16:35\nКібервійна триває, і ...	https://ko.com.ua/zafiksovana_kiberataka_na_de...
2022-03-29 14:12:00	Чому інтернет та зв'язок в Україні ворогу не в...	2 години ago\nПопри потужні кібератаки та фізи...	https://techtoday.in.ua/news/chomu-internet-ta...
2022-03-29 00:31:00	Фахівці Держспецзв'язку нейтралізували масован...	Київ, Вівторок 29 березня 2022 00:30\nФото: ма...	https://www.rbc.ua/ukr/news/spetsialisty-gossp...

Рисунок 3.2 – Кінцевий результат обробки даних

В результаті обробки даних, зображено таблицю, яка містить інформацію про різні події або записи, що стосуються кіберзагроз в період з лютого 2022 року по липень 2022 року.

3.2 Побудова часових рядів

Як нам вже відомо, побудова часових рядів розпочинається з визначення залежних змінних, а саме часу від кількості. Після цього будується власне часовий ряд, і відображається на графіку. На рисунку 3.3 можна побачити залежність кількості публікацій від часу. Дані для аналізу було взято за період від початку лютого 2022 року до кінця Липня 2022.

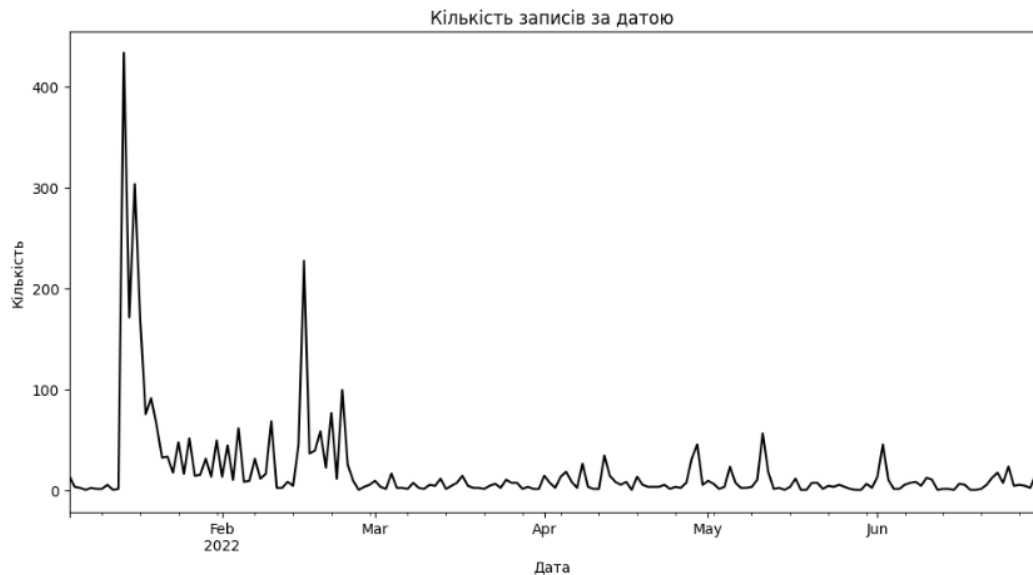


Рисунок 3.3 – Часовий ряд отриманий в результаті дослідження

Проаналізувавши графік, можна побачити яка кількість новин публікувалася в період двох кварталів, після початку повномасштабного вторгнення в Україну. На ньому чітко видно, що в період з 22.02.2022 до кінця березня 2022 року спостерігається найвища активність, саме на цей період припадала найбільша кількість кібератак. Далі бачимо різкий спад публікацій новин, що свідчить про зменшення інтенсивності кібератак.

Таким чином, виставивши поріг кількості публікацій, можна побачити локальні максимуми часового ряду, що може свідчити про наявність певних аномалій (див. рис. 3.4).

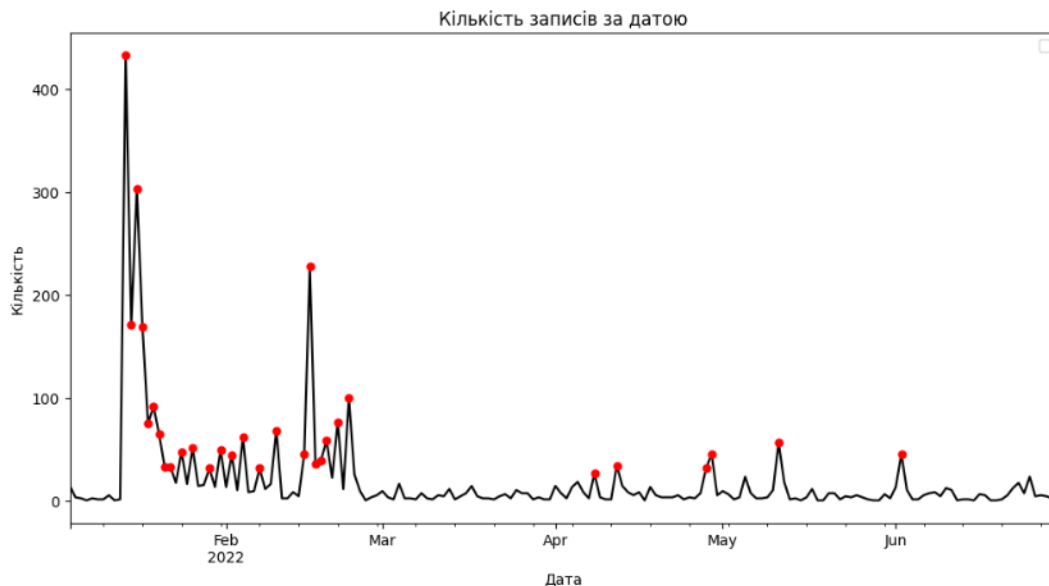


Рисунок 3.4 – Локальні максимуми

Для того, щоб проаналізувати, наскільки сильно значення в одній точці часу корелюють із значеннями в іншій точці, побудуємо графік автокореляції, це допоможе виявити наявність сезонності або циклічності в даних (див. рис. 3.5).

При побудові графіка автокореляції, по осі X відкладаються лаги (відстані в часі), а по осі Y — значення коефіцієнта автокореляції. Цей коефіцієнт може коливатися від -1 до 1, де 1 означає повної позитивної кореляції, -1 — повної негативної кореляції, а 0 — відсутність кореляції.

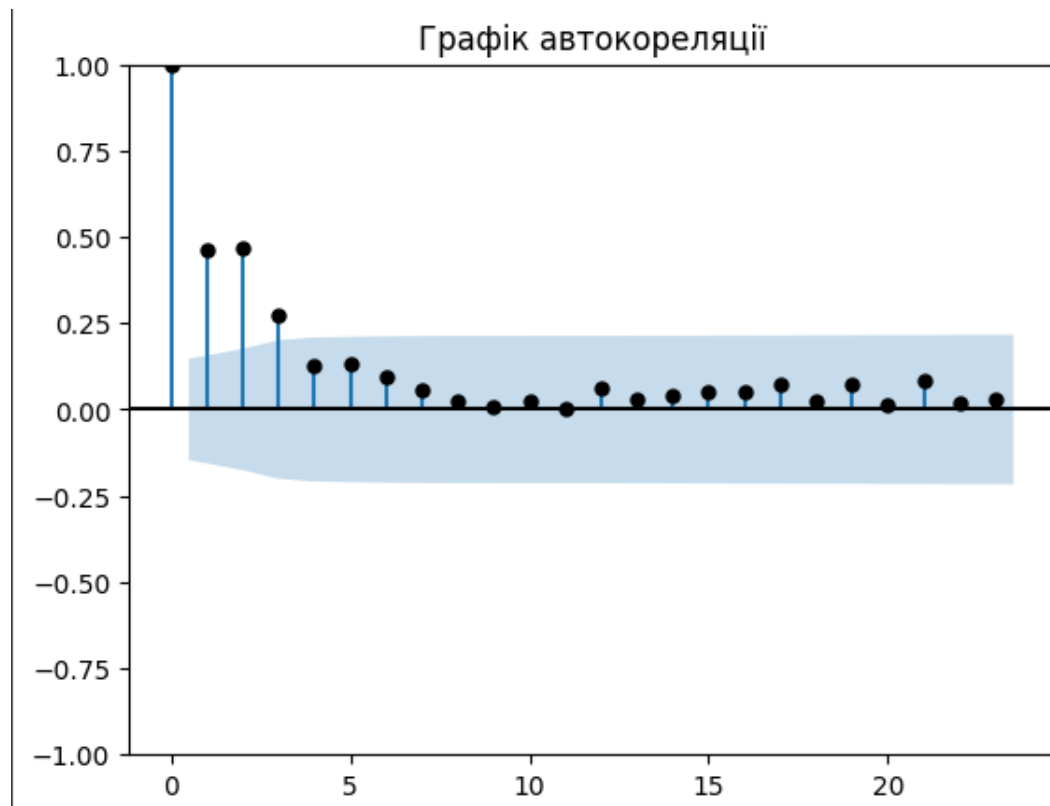


Рисунок 3.5 – Графік автокореляцій між точками часу

Автокореляційний аналіз, який показав відсутність значущої кореляції, адже майже всі точки знаходяться в межах довірчого інтервалу, а ті точки які знаходяться поза межами цієї зони можуть вказувати на статистично значущу кореляцію.

Ще один спосіб виявити періодичні компоненти в часовому ряді, це побудувати графік спектральної щільності (див. рис. 3.6). Вивчення графіка спектральної щільності є важливим для розуміння структури часового ряду і може допомогти виявити не лише сезонність, але й інші періодичні ефекти, які можуть бути важливими для прогнозування та аналізу даних. Це дозволяє краще зрозуміти динаміку даних і врахувати ці ефекти при побудові моделей прогнозування. А також виконати декомпозицію часового ряду, щоб краще розуміти структуру даних (див рис 3.7).

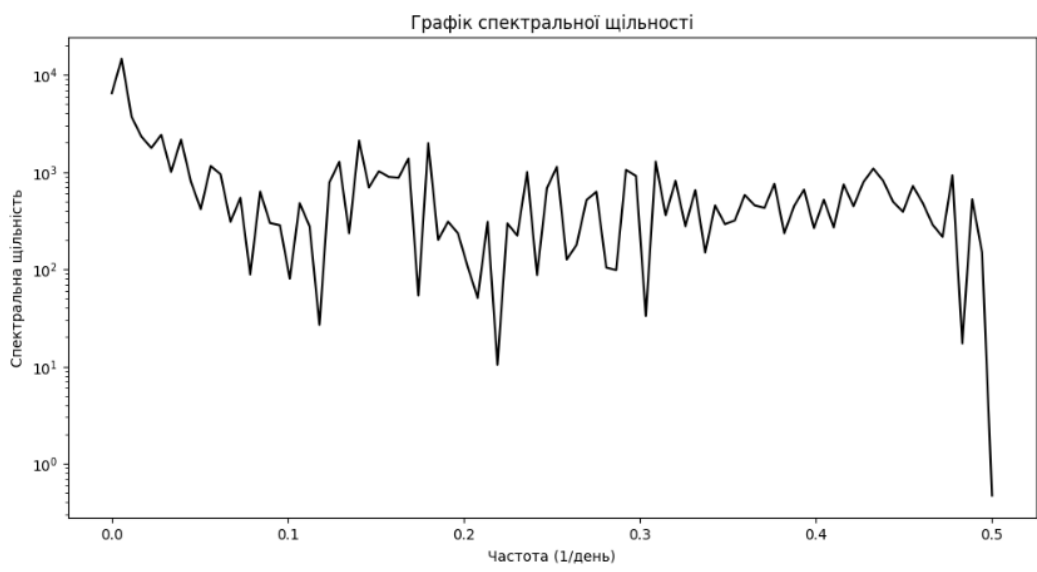


Рисунок 3.6 – Графік спектральної щільності

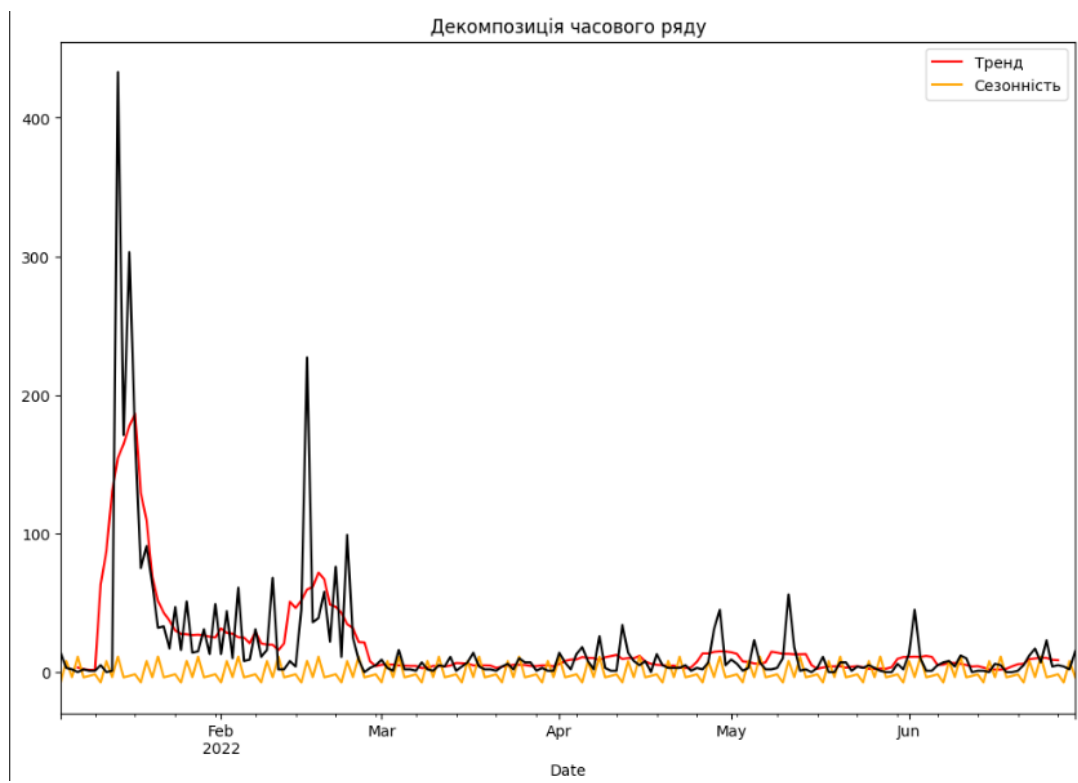


Рисунок 3.7 – Декомпозиція часового ряду

Проаналізувавши графік, можемо побачити, що лінія тренду відображає загальну направленість зміни значень у часовому ряді.

Лінія сезонності(жовтий колір) показує регулярні та повторювані коливання в даних, які відбуваються з певною періодичністю.

3.3 Застосування вейвлет-аналізу

Після аналізу часових рядів, включаючи методи спектрального аналізу та автокореляції, наступним логічним кроком є застосування вейвлет-перетворення. Вейвлет-аналіз дозволяє дослідити часовий ряд у частотно-часовій області, що суттєво розширює можливості аналізу, порівняно з традиційними методами.

Для практичної реалізації вейвлет-перетворення необхідно визначити діапазон частот, які будуть досліджуватись. Це можна зробити на основі попередньо отриманих результатів аналізу часового ряду. Зокрема, піки в періодограмі та значення автокореляційної функції дають розуміння про домінуючі частоти, які присутні в даних.

Далі буде представлено результати вейвлет-аналізу, проведеного на основі обраного діапазону масштабів, та їх інтерпретація для глибшого розуміння динаміки досліджуваного процесу.

На рисунку 3.8 зображено скейлограма часового ряду, отримана за допомогою вейвлет-перетворення з використанням вейвлетів Mexican Hat, та Morle (див. рис. 3.9).

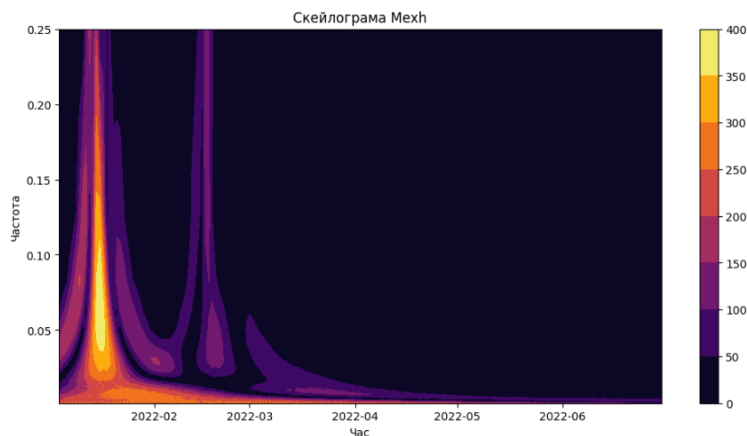


Рисунок 3.8 – Результат вейвлет-аналізу з використанням вейвлету Mexican Hat

На скейлограмі, побудованій для вейвлету Mexican Hat можна спостерігати аномальні сплески, які були в період з початку лютого 2022 року по кінець липня 2022 року. Найбільше виділяється період від початку лютого до кінця березня. Що свідчить про дуже велику кількість публікацій, щодо кібератак на Україну. Далі ми вже не бачимо сильних змін на скейлограмі, проте це не означає, що в цей період не було публікацій, просто вони не настільки часті, як в попередньому кварталі.

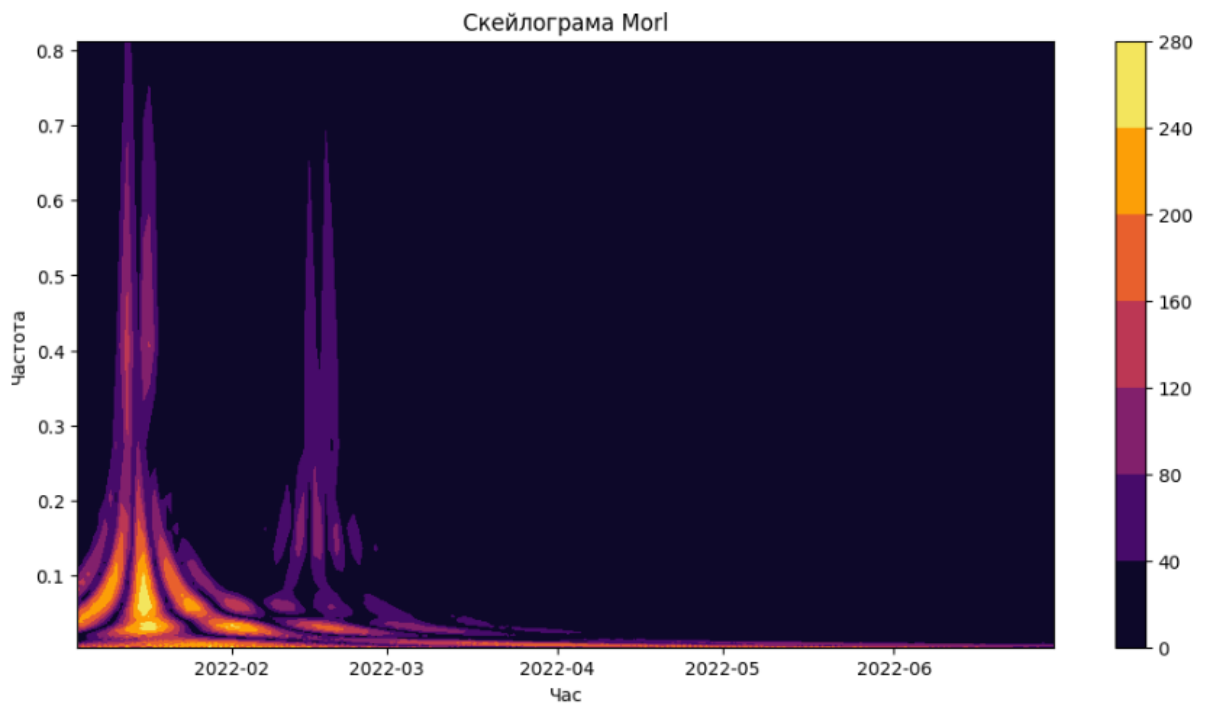


Рисунок 3.9– Результат вейвлет-аналізу з використанням вейвлету Morle

На скейлограмі побудованій для вейвлету Морле ми бачимо те ж саме, що і на попередній, але через те, що вейвлет Морле забезпечує вищу роздільну здатність, ніж вейвлет Mexican Hat ми можемо побачити більш чіткішу структуру сплесків.

ВИСНОВКИ ДО РОЗДІЛУ 3

У третьому розділі було описано практичну реалізацію даного дослідження. Це дозволило перевірити ефективність підходу на реальних даних.

За допомогою вейвлет-скейлограм вдалося виявити аномальну активність, що може вказувати на виникнення нових або підвищення активності існуючих загроз.

Ці результати підтверджують високу ефективність вейвлет-аналізу для виявлення аномалій та потенційних загроз у реальних системах.

Крім того, використання вейвлет-аналізу дозволяє краще зрозуміти природу виявлених аномалій та їх зв'язок з іншими характеристиками системи. Ця здатність забезпечити глибше розуміння проблем робить вейвлет-аналіз надзвичайно цінним інструментом для комплексного вивчення та аналізу складних систем

ВИСНОВКИ

У ході дослідження було розроблено алгоритм виявлення аномалій для попередження кібератак з застосуванням вейвлет-аналізу.

Проведено аналіз новинних повідомлень про кіберзагрози, та виявлено, що вейвлет-аналіз дозволяє ефективно виявити аномальні зміни в характері та частоті публікацій, які можуть вказувати на наявність нових або посилення існуючих кіберзагроз.

Порівняльний аналіз продемонстрував, що запропонований підхід, який поєднує вейвлет-аналіз та аналіз часових рядів, значно перевищує за точністю традиційні методи, засновані лише на статистичних характеристиках новинних публікацій.

Додатково проведено аналіз часових рядів новинних публікацій про кіберзагрози. Виявлено наявність сезонних та циклічних закономірностей, а також точки структурних зламів у часових рядах, які можуть вказувати на значущі зміни в характері кіберзагроз.

Комбіноване використання вейвлет-аналізу та аналізу часових рядів показало високу ефективність у виявленні аномалій, пов'язаних з кіберзагрозами, з мінімальною кількістю хибних спрацювань.

Розроблено алгоритм виявлення аномалій на основі вейвлет аналізу новинних даних. Алгоритм дозволяє виявляти різкі сплески публікацій про кіберзагрози, аномальні зміни в частотних характеристиках новин, а також інші форми аномальної активності, пов'язаної з кіберзагрозами.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Д.В.Ланде, І.Ю.Субач, Ю.Є.Бояринова. Основи теорії і практики інтелектуального аналізу даних у сфері кібербезпеки. — Київ : Електронні текстові дані, 2018.
2. А.Т.Яровий, Є.М.Страхов. Аналіз часових рядів.. — Одеса : Освіта України, 2019.
3. Time Series Analysis: Definition, Types, Techniques, and When It's Used. [Електронний ресурс] – Режим доступу до ресурсу: <http://surl.li/uoaiie>.
4. Adam Hayes. What Is a Time Series and How Is It Used to Analyze Data?
5. Раєвська О. В., Чанкіна І. В. Аналіз та прогнозування рядів динаміки. — Харків, 2014
6. Wavelet transformation. [Електронний ресурс] – Режим доступу до ресурсу: <http://surl.li/uoaim>.
7. Wavelets and Time Series Modeling, 2021. [Електронний ресурс] – Режим доступу до ресурсу: <http://surl.li/uoaiio>.
8. Michael P.Bishop, Zhaohui Chi. Technology-Driven Geomorphology Geospatial Data Science.
9. Daniel W. Wavelets. Statistical Methods in the Atmospheric Sciences. — 2019.
10. Christopher Schölzel. Introduction to wavelet theory. [] – Режим доступу до ресурсу: <http://surl.li/uoais>.

ДОДАТКИ

ДОДАТОК А

ЛІСТИНГ ПРОГРАМНОГО КОДУ

Файл Wavelet.ipynb

```
import pywt
import numpy as np
import pandas as pd
from scipy import signal
import matplotlib.pyplot as plt
from statsmodels.graphics.tsaplots import plot_acf
from statsmodels.tsa.seasonal import seasonal_decompose

df = pd.read_csv('/content/drive/MyDrive/dipl/data.csv')
df.head()
print(df['Date'])

df['Date'] = df['Date'].str.replace(r'^0-9', '', regex=True)
df['Date'] = df['Date'].str.strip()
df['Date'] = pd.to_datetime(df['Date'], format='%Y%m%d%H%M', errors='coerce')
df['Date'] = df['Date'].fillna(pd.to_datetime(df['Date'], format='%Y%m%d',
errors='coerce'))
df['Date'] = df['Date'].fillna(pd.to_datetime(df['Date'], format='%Y%m%d%H',
errors='coerce'))
df = df[df['Date'].notna()]
df['Date'].to_csv('/content/drive/MyDrive/dipl/date_data.csv', index=False, header=True)
df = df.set_index('Date')
```

```
plt.figure(figsize=(12, 6))
df.resample('D').size().plot(color='black')
plt.title('Кількість записів за датою')
plt.xlabel('Дата')
plt.ylabel('Кількість')
plt.show()
df.head()
```

```
plt.figure(figsize=(12, 6))
daily_counts = df.resample('D').size()
daily_counts.plot(color='black')
anomalies = daily_counts[daily_counts > 25]
plt.plot(anomalies.index, anomalies, 'ro', markersize=5, label='Аномалії', color='red')
plt.title('Кількість записів за датою')
plt.xlabel('Дата')
plt.ylabel('Кількість')
plt.legend()
plt.show()
```

```
plot_acf(daily_counts, color='black')
plt.title('Графік автокореляції')
plt.show()
```

```
freqs, psd = signal.welch(daily_counts, fs=1)
plt.figure(figsize=(12, 6))
plt.semilogy(freqs, psd, color='black')
plt.title('Графік спектральної щільності')
```

```
plt.xlabel('Частота (1/день)')
plt.ylabel('Спектральна щільність')
plt.show()
```

```
result = seasonal_decompose(daily_counts, model='additive', period=7)
plt.figure(figsize=(12, 8))
result.trend.plot(color='blue', label='Тренд')
result.seasonal.plot(color='orange', label='Сезонність')
result.resid.plot(color='green', label='Залишки')
plt.title('Декомпозиція часового ряду')
plt.legend()
plt.show()
```

```
time_series = df.resample('D').size()
coef, freqs = pywt.cwt(time_series, np.arange(1, len(time_series)), 'mexh')
plt.figure(figsize=(12, 6))
plt.contourf(time_series.index, freqs, abs(coef))
plt.colorbar()
plt.title('Скейлограма Mexh')
plt.xlabel('Час')
plt.ylabel('Частота')
plt.show()
```

```
coef, freqs = pywt.cwt(time_series, np.arange(1, len(time_series)), 'morl')
plt.figure(figsize=(12, 6))
plt.contourf(time_series.index, freqs, abs(coef))
```

```
plt.colorbar()
plt.title('Скейлограма Morl')
plt.xlabel('Час')
plt.ylabel('Частота')
plt.show()
```

Файл json.py

```
import json
with open('data.txt', 'r', encoding='utf-8') as file:
    text = file.read()
    articles = text.split('***')
    data = []
    for article in articles:
        lines = article.strip().split('\n')
        title = lines[0]
        content = '\n'.join(lines[1:-2])
        link = lines[-2]
        source_date = lines[-1]
        data.append({
            'title': title,
            'content': content,
            'date': link,
            'source': source_date,
        })

with open('output.json', 'w', encoding='utf-8') as file:
    json.dump(data, file, ensure_ascii=False, indent=4)
```

Файл csv.py

```
import json
```

```
import csv
```

```
with open('output.json', 'r', encoding='utf-8') as f:
```

```
    data = json.load(f)
```

```
with open('data.csv', 'w', newline="", encoding='utf-8') as csvfile:
```

```
    writer = csv.writer(csvfile)
```

```
    writer.writerow(['Date', 'Title', 'Content', 'Source'])
```

```
    for item in data:
```

```
        writer.writerow([item['date'], item['title'], item['content'], item['source']])
```

ДОДАТОК Б ПОЛЯ В НАБОРАХ ДАНИХ

Date

Title

Content

Source