НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО

Факультет інформатики та обчислювальної техніки
(назва факультету, інституту)

Кафедра автоматизованих систем обробки інформації і управління
(назва кафедри)

"На правах рукопису"
УДК 004.93

«До захисту допущено»
В.о.завідувача кафедри

_____ О.А.Павлов
(підпис) (ініціали, прізвище)

“ ” 20 *19* р.

МАГІСТЕРСЬКА ДИСЕРТАЦІЯ

на здобуття ступеня магістра

за спеціальністю *126 Інформаційні системи та технології*
(код та назва спеціальності)

ОПП *Інформаційні управляючі системи та технології*
(код та назва спеціалізації)

на тему: *Виявлення аномалій в часових рядах довільної природи*

Виконав: студент *VI* курсу групи ІС-82мп
(шифр групи)

Логвинчук Андрій Ігорович _____
(прізвище, ім'я, по батькові) (підпис)

Науковий керівник *доц., к.т.н., Баклан І. В.* _____
(посада, науковий ступінь, вчене звання, прізвище та ініціали) (підпис)

Консультант *доц., к.т.н. Жданова О.Г.* _____
(науковий ступінь, вчене звання, прізвище, ініціали) (підпис)

Рецензент *доц., к.т.н. Верба О. А.* _____
(посада, науковий ступінь, вчене звання, прізвище та ініціали) (підпис)

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студент _____
(підпис)

Київ – 2019

РЕФЕРАТ

Магістерська дисертація: 79 с., 21 рис., 31 табл., 34 джерела, 1 додаток.

Актуальність. На сьогоднішній день виявлення аномалій є однією із головних причин виконання аналізу даних. Можливість виявляти рідкісні та нетипові показники та події широко застосовується у найрізноманітніших сферах людської діяльності: інженернотехнічній, фінансово-економічній, медичній та інших.

Із подальшим розвитком інтернету речей, потреба у автоматизованих системах моніторингу та прийняття рішень, здатних вчасно розпізнати збої або помилки в роботі різного роду пристроїв та інфраструктури, та недопустити небажаних наслідків, буде тільки зростати. Саме тому надзвичайно важливою сьогодні є розробка ефективних алгоритмів виявлення аномалій.

Зв'язок роботи з науковими програмами, планами, темами. Робота виконувалась на кафедрі автоматизованих систем обробки інформації та управління Національного технічного університету України «Київський політехнічний інститут ім. Ігоря Сікорського» в рамках теми «Інтелектуальні методи програмування, моделювання і прогнозування з використанням ймовірністного і лінгвістичних підходів» (№ ДР 0117U000926).

Мета дослідження - роботи є підвищення швидкості прийняття рішень в автоматизованих системах управління за рахунок розробки алгоритмів передбачення аномалій у процесах, представлених часовими рядами.

Для досягнення мети необхідно виконати наступні **завдання**:

- виконати огляд існуючих методів та алгоритмів виявлення аномалій у часових рядах;
- здійснити порівняльний аналіз зазначених методів і алгоритмів;
- формалізувати задачу виявлення аномалії у часовому ряді;
- розробити алгоритм виявлення аномалій у часових рядах на основі лінгвістичних моделей;
- експериментально порівняти розроблений алгоритм з існуючими;
- виконати аналіз отриманих результатів.

Об'єкт дослідження – аномалії у процесах, представлених часовими рядами.

Предмет дослідження – методи та алгоритми виялвлення аномалій у часових рядах.

Методи дослідження, застосовані у даній роботі, базуються на методах машинного навчання та експертних оцінок. Серед інших, використані структурні, синтаксичні та лінгвістичні методи.

Наукова новизна одержаних результатів полягає у використанні методів лінгвістичного моделювання для виявлення аномалій у часових рядах, яке раніше не мало подібного застосовування.

Публікації. Матеріали роботи опубліковані в міжнародному журналі Slovac International Scientific Journal та у збірнику тез III всеукраїнської науково-практичної конференції молодих вчених та студентів «Інформаційні системи та технології управління» (ІСТУ-2019).

ВИЯВЛЕННЯ АНОМАЛІЙ, ЧАСОВІ РЯДИ, КЛАСИФІКАЦІЯ, ПОШУК ПАТТЕРНІВ

ABSTRACT

Master thesis: 79 pages, 21 figures, 31 tables, 34 sources.

Topicality. As for today, anomaly detection is one the main drivers for data analysis. The ability to detect rare and atypical characteristics and events is widely used in different domains: engineering, economics, health care and others.

With further development of Internet of Things, the demand for automated systems for monitoring and decision making, capable of early detection of faults and errors in critical infrastructure, and that will be able to prevent unwanted effects, will only grow. This is the reason why development of effective anomaly detection algorithms is so important.

Aim of the work is to increase decision making speed in automated control systems achieved with the development of effective anomaly detection algorithms.

To achieve the goal, the following **objectives** must be completed:

– do an overview of existing methods and algorithms for time series anomaly detection;

– conduct a comparison analysis of mentioned methods and algorithms; – formalize the anomaly detection problem for time series;

– develop the anomaly detection algorithm based on linguistic models; – compare the developed algorithm to existing experimentally;

– analyze the results.

Objects of research are anomalies in processes represented with time series.

Subjects of research are methods and algorithms for anomaly detection in time series.

Methods of research used in this work are based on machine learning methods and the method of expert evaluation. Among others, structural, linguistic and syntax methods are used.

Scientific novelty. The scientific novelty of this result consists on the usage of linguistic modelling methods which haven’t been used for this purpose previously.

Publications. The work materials were published in the Slovac International Scientific Journal and in “Information Systems and Management Technologies” conference thesis compilation.

ANOMALY DETECTION, TIME SERIES, CLASSIFICATION, PATTERN MATCHING

ЗМІСТ

ВСТУП

Виявлення аномалії, як одна із форм аналітики даних, з кожним роком знаходить все більше і більше застосувань у найрізноманітніших сферах людської діяльності. Так, із зростанням сектору IoT та глобальної інтеграції, все частіше постає потреба у інструментах для здійснення моніторингу кібернетичних систем, засобах реагування та усунення наслідків у разі збоїв та помилок у них. Виявлення вторгнень, несанкціонованого доступу, несправностей у критично важливих системах безпеки та системах управління інфраструктурою є одними із пріорітетних задач у сучасному світі інформаційних технологій.

Виявлення аномалії – це задача пошуку шаблонів у даних, які не відповідають очікуваній поведінці. Аномалії – це невідповідні спостереження, викиди, що суперечать природі досліджуваного процесу.

Важливість виявлення аномалії пов'язана з тим, що аномалії в даних несуть, часто, критичну інформацію про можливу загрозу приватності, конфіденційності та, навіть, життю і здоров'ю людини. Наприклад, аномальна схема трафіку в комп'ютерній мережі могла б означати, що зламаний комп’ютер посилає конфіденційні дані до несанкціонованого пункту призначення. Аномальне зображення МРТ може вказувати на наявність злоякісної пухлини. Аномалії в даних про транзакцію кредитної картки можуть вказувати на викрадення картки чи посвідчення особи, а аномальні показники з датчиків космісного корабля можуть свідчити про його несправність.

Методи виявлення аномалій, що базуються на навчанні застосовуються в багатьох прикладних областях, включаючи інформаційну безпеку, біоінформатику, автомобільну індустрію, астрономію та інші [1].

Проблема виявлення аномалій була досліджена в численних наукових областях та прикладних сферах. Багато підходів та технік було спеціально

розроблено для застосування у певних предметних областях, тоді як деякі є більш узагальненими.

В загальному ж, більшість підходів до виявлення аномалій є вузькоспеціалізованими до можуть застосовуватись лише в певних умовах із накладанням численних обмежень. Досі не існує такого універсального методу, який би дозволяв виконувати детекцію аномалій на довільних даних, без спостереження та з достатньою точністю [2]. Серед відомих причин цього можна виділити наступні: ті дані, які є нормальними на даний момент, можуть вважатися аномалією в майбутньому (і навпаки), іноді границя між нормальними та аномальними значеннями дуже нечітка, також дуже часто існує дефіцит даних для тренування і валідації моделей.

Ще одним недоліком багатьох існуючих алгоритмів є неможливість виявляти аномалії в потокових даних у реальному часі, багато з них можуть працювати тільки із статистичними історичними даними, тобто аналізувати характер перебігу явищ та процесів, які вже відбулись, маючи достатню кількість знань про хід аналогічних процесів.

На сьогоднішній день усі алгоритми виявлення аномалій можна поділити на 3 групи [3]: алгоритми виявлення аномалій без нагляду дослідника, виявлення аномалій із наглядом дослідника, а також комбіновані підходи.

Метою даної роботи є підвищення швидкості виявлення аномалій, в порівнянні з алгоритмами, що застосовуються в автоматизованих системах управління, за рахунок використання лінгвістичного моделювання та синтаксичного підходу до аналізу часових рядів.

Досягнення мети базується на розробці оригінальних математичних методів та алгоритмів на базі алгоритмів прогнозування та виявлення аномалій в історичних даних, а також їх реалізації у вигляді програмної системи.

Об’єктом даного дослідження є визначення аномальних значень, що не відповідають природі процесу, що досліджується, та прогнозування появи аномалій у майбутньому, грунтуючись на історичній поведінці даного процесу.

1 ОГЛЯД СУЧАСНИХ МЕТОДІВ ВИЯВЛЕННЯ АНОМАЛІЙ У ЧАСОВИХ РЯДАХ

1.1 Визначення часового ряду

Практично будь-який процес можна описати за допомогою математичної функції, яка встановлює відповідність між точками з множини значень деякого показника, що характеризує стан процесу та моментами часу, у які ці значення було зафіксовано.

Часовий ряд (також «ряд динаміки», «динамічний ряд») - це послідовність значень деякого показника, що впорядкована у хронологічному порядку. В англомовній літературі використовують термін "time series". Елементами ряду, або його рівнями, називаються окремі спостереження. Кожен елемент ряду відповідає деякому моменту у часі. Рівні ряду можуть бути виміряні як через однакові інтервали часу, так і у довільні моменти. Порядок розташування рівнів є істотною характеристикою ряду і не може змінюватися довільно. Іноді кожному моменту часу приводять у відповідність декілька значень різних показників досліджуваного об’єкта. Тоді говорять про багатовимірний часовий ряд [4].

Ряди можуть бути моментними або інтервальними, в залежності від характеру часового параметра. Моментні ряди характеризують значення показника у заданий момент часу, наприклад курс валют станом на певну дату, температура повітря, силу електричного струму в мережі тощо.

Інтервальні ряди показують динаміку зміни значень моментного ряду. Так, наприклад інтервальні ряди показують, на скільки змінилось значення показника з моменту попереднього вимірювання. Проводячи аналогії із поняттям функції в математичному аналізі, можна сказати, що інтервальні ряди є похідними моментних рядів.

Використовуючи метод акумульованих сум, можна із будь якого інтервального ряду отримати моментний. З іншого боку, підсумовуючи значення рівнів моментного ряду, ми отримаємо величину, що не має фізичного змісту. Однак, поділивши цю суму на кількість вимірів у досліджуваному періоді можна отримати статистично значуще число, а саме – середнє значення ряду.

Ще один вид класифікації рядів – за відсутністю деяких рівнів ряду - на повні й неповн. У неповних часових рядах можуть бути пропущені занчення, що відповідають деяким моментам часу. Рівні рядів динаміки можуть бути абсолютними, відносними або середніми значеннями певних показників. Якщо вони є не величинами, вимірюваними

безпосередньо, а похідними від них – середніми, відносними й т.д., то відповідні ряди називають похідними [4].

Рівні ряду зазвичай є статистично незалежними, а також не є однаково розподіленими.

Існує два основні напрямки обробки даних, представлених часовими рядами – прогнозування та аналіз історичних даних.

Прогнозування часових рядів – це використання моделі для передбачення майбутніх значень на основі раніше спостережуваних значень. Регресійний аналіз часто використовується, щоб перевірити теорії про те, що поточні значення одного або декількох незалежних часових рядів впливають на поточне значення іншого часового ряду.

Серед інших методів, які застовують для прогнозування часових рядів можна назвати методи Гольта, Гольта-Вінтерса, а також статистичні моджелі типу ARIMA та ін. Прогнозування ряду і методи виявлення аномалій часто застосовуються разом. Це дозволяє не лише виявляти викиди в так званих історичних даних (даних відомих на момент аналізу), а довзоляє виявляти аномалій у прогнозованих даних. Цей факт є дуже важлим, оскільки дозволяє завчасно прийняти рішення та вжити відповідних заходів з недопущення аномальних відхилень.

Процес аналізу часового ряду полягає у визначенні його статистичних характеристик, таких як середнє значення, середньоквадратичне відхилення, характер розподілу рівнів ряду та ін. Крім того важлим є завдання пошуку прихованих закономірностей. Виявлення структури часового ряду необхідно для того, щоб побудувати математичну модель того явища, яке є джерелом аналізованого часового ряду. При аналізі часового ряду використовують наступні моделі:

- екстраполяційні;
- наближення за допомогою кривих (поліноміальна регресія);
- апроксимація функцією;
- моделі Маркова.

Стохастична модель часового ряду, як правило, відображатиме той факт, що спостереження, близькі між собою у часі, будуть тісніше пов'язані, ніж спостереження далі. Крім того, моделі часових рядів часто використовують природне одностороннє впорядкування часу, так що значення для даного періоду будуть виражатися як похідні певним чином від минулих значень, а не від майбутніх значень. На плакаті «Класифікація методів аналізу часових рядів» Додатку А перелічено деякі методи аналізу часових рядів.

Для ефективного аналізу часового ряду важливо правильно обрати інтервал між сусідніми членами. Зручно використовувати сталий інтервал між спостереженнями, хоча це

не завжди можливо. Так, наприклад, курс валют не визначається у святкові та вихідні дні, отже члени цього ряду не будуть рівновіддаленими. Якщо взяти занадто великий інтервал – можна втрати інформацію про суттєві особливості ряду, якщо ж частота спостережень буде надмірно великою, то це ускладнить аналіз ряду (оскільки призведе до збільшення кількості розрахунків), а також може внести випадкові шуми.

Основними завданнями дослідження часових рядів є [4]:

- виокремлення та опис основних характерних особливостей ряду;
- підбір статистичної моделі, що найкращим у певному розумінні способом відображає ряд;
- прогнозування майбутніх значень показників, що утворюють ряд, за попередніми спостереженнями;
- підготовка рекомендацій з управління процесом, що породжує досліджуваний часовий ряд.

При аналізі часових рядів зазвичай здійснюється такі кроки [4]:

- графічне подання й попередній аналіз поведінки часового ряду;
- виокремлення і видалення закономірних складових ряду (тренду, сезонних та циклічних компонент);
- виявлення і видалення низько- та високочастотних складових (фільтрація);
- дослідження випадкової складової часового ряду, що залишилася після видалення вищезазначених компонент;
- побудова і перевірка адекватності моделі випадкової складової;
- побудова загальної моделі досліджуваного ряду;
- дослідження отриманої моделі і прогнозування майбутньої поведінки процеса, що вивчається;
- вивчення взаємодії між різними часовими рядами, що характеризують певну систему або процес.

Важливою умовою аналізу є порівнянність рівнів ряду. Інформація про ряд має бути достатньо повною. Зокрема, для рядів, що містять сезонну складову, залежно від застосовуваних методів математичної обробки, потрібна інформація щодо проміжку, який перевищує 3–6 повних циклів. При побудові регресійних моделей необхідно мати ряди,

довжина яких у кілька разів перевищує кількість параметрів, що визначаються. Елементи рядів динаміки можуть мітстити аномальні значення (викиди). Часто їх спричиняють помилки під час збору та обробки інформації. Однак вони також можуть бути викликані вливом зовнішніх факторів.

1.2 Визначення поняття аномалії

У перекладі з грецької, аномалія - неправильність, відхилення від норми. При застосуванні у аналізі часових рядів під поняттям "аномалія" розуміється наявність значень, які значно виділяються із загальної закономірності, тобто такі значення, зокрема значення, які не зумовлені тенденцією, циклічною природою ряду або сезонними сплесками.

В якості яскравого прикладу аномалії у часовому ряді можна привести графік ціни акцій компанії Microsoft у з 2006 по 2018 роки (рисунок 1.1):

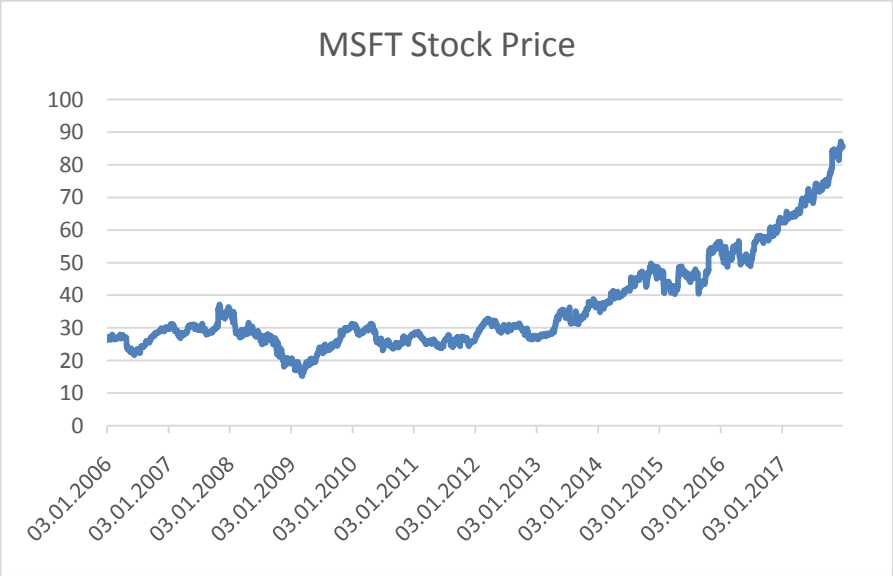


Рисунок 1.1 – Ціни на акції Microsoft

Можна спостерігати різкий "провал" графіку, що розпочинається у 2008 році, який пояснюється початком світової фінансової-економічної кризи.

1.3 Класифікація методів виявлення аномалій, що застосовуються до часових рядів

Основні підходи до виявлення аномалій поділяють у часових рядах на 3 великі групи: ті, що базуються на відстанях між точками, на щільності або на рангах [5].

- точки, що є більш віддаленими від інших вважаються аномальними;
- точки, що розміщуються в регіонах з низькою щільністю вважаються аномальними;
- найбільш аномальними є точки, чиї найближчі сусіди знаходяться по сусідству з іншими аномальними точками.

Для всіх трьох підходів, дані за своєю природої можуть споглядані, частково споглядані або не споглядані.

У випадку спогляданих даних, мітки класифікації є відомими для деякої множини "тренувальних" даних, тож всі порівння і визначення відстаней здійснюється по відношенню до таких тренувальних даних.

У не спогляданому випадку, такі мітки не відомі, тож визначення відстаней і порівння здійснюються по відношенню до всього набору даних.

В останньому випадку, мітки відомі для деякої частини даних, але не для більшості. Наприклад, за наявності деякої нової категорії аномалій, алгоритм із частковим наглядом може спробувати визначити які інші підозрілі випадки можуть відноситись до однієї ктаегорії. Алгоритми часто працюють у декілька фаз, із ранньою фазою, під час якої здійснюється призначення попередніх міток нерозміченим даним.

Алгоритм виялвення аномалій без нагляду повинен відповідати наступним вимогам [6]:

а) патерни нормальної поведінки мають визначатись динамічно та не повинні вимагати попереднього тренувального або довідкового набору даних для її визначення;

б) дані, що виходять за рамки норми повинні визначатись ефективно навіть якщо розподіл даних невідомий

в) алгоритм повинен мати можливість застосування його у різних доменах без необхідності знати принципи та основоположення дзаної предметної області.

Основними алгоритмами, що застосовуються для вирішення проблеми виявлення аномалій є методи кластеризації k-найближчих сусідів (k-NN, англ. k-Nearest Neighbors), k-Means, динамічні довірчі мережі Байеса, приховані моделі Маркова (ПММ), а також різноманітні рекурентні нейронні мережі, зокрема нейронні мережі з довгою короткотривалою пам'яттю (LSTM). Класифікація методів наведена на плакаті «Класифікація методів виявлення аномалій» Додатку А. Дані підходи є актуальними для задачі виявлення та передбачення аномалій, оскільки (відповідно деякі з них) застосовуються для виконання різноманітних класифікацій, деякі добре підходять для опису не детермінованих процесів.

1.3.1 Методи із застосуванням кластеризації

Кластеризація може базуватись на обчислення схожості або відстані; ці два підходи відрізняються, проте вчасто приводять до одного й того ж кінцевого результату, оскільки міри відстані та схожості мають сильну від’ємну кореляцію. Кластеризація на основі відстаней грунтується на ідеї про те, що точки дані з одного кластера не сильно віддалені одна від одної, в той час точки із різних кластерів мають між собою значну відстань. При здійсненні кластеризації на основі схожості вважають, що точки які подібні одна до одної повинні належати до одного кластеру, так як зі зменшенням відстані між ним схожість збільшується.

При використанні методів, що базуються на кластеризації, зазвичай вважається, що дані знахожяться в окресленому мультивимірному просторі, та що міра схожості або відстані була заздалегідь обрана.

Кластеризація методом k найближчих сусідів (k-Nearest-Neighbours clustering)

k-NN – метод k найближчих сусідів:

- застосовується для автоматичної класифікації об’єктів;
- об’єкту присвоюється той клас, який є найбільш поширеним серед його сусідів.

В рамках теми наукової роботи може застосовуватись при роботі з граничними значеннями, наприклад, чи відносити дану точку до класу нормальних або аномальних. Незважаючи на те, що метод загалом є простим та ефективним, він має певні обмеження, а саме у випадку незбалансованості класів. Оскільки нормальних точок буде значно більше ніж аномальних, алгоритм матиме тенденцію частіше класифікувати точку як нормальну. Щоб компенсувати це, необхідно вводити ваговий коефіцієнт. Також цей алгоритм покладається на наявність великого тренувального набору даних.

Кластерицазція методом k-середніх (англ. k-Means clustering)

Кластеризація методом k-середніх – це метод квантизації векторів, який початково застосовувався для обробки сигналів, а наразі знайшов широке застосування в кластерному аналізі да дата майнінгу. Метою цього алгоритму є розподіл n спостережень на k кластерів, в яких кожне спостереження (елемент) належить кластеру з найближчим середнім значенням, що слугує прототипом кластера. Метод k середніх мінімізує відхилення всередині кластера (квадратичні Евклідові відстані), але не звичайні евклідові відстані.

Даний алгоритм також пов’язаний із методом k-NN, згаданим вище популярним підходом з області машинного навчання. Застосування класифікатора 1 найближчого зусіда дає змогу розподіляти нові спостереження в існуючі кластери.

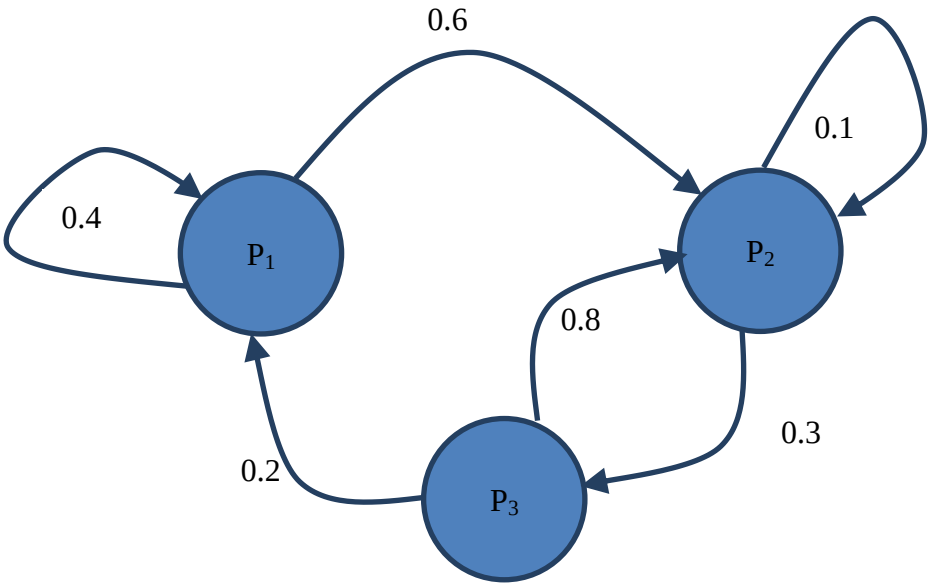
В алгоритмах глибокого навчання метод k-середніх іноді застосовують не за прямим призначенням (класифікація розбивкою на кластери), а для створення так званих фільтрів (ядер згортки, словників). Наприклад, для розпізнавання зображень в алгоритм k-середніх подають невеликі випадкові частини зображень навчальної вибірки у вигляді лінійного вектора, кожен елемент якого кодує колір своєї точки. Кількість кластерів k задається великою. Навчений метод k-середніх за певних умов визначає при цьому центри кластерів (центроїди), які представляють собою зручні базиси, на які можна розкласти будь вхідне зображення. Такі центроїди надалі використовують в якості фільтрів, наприклад для згорткової нейронної мережі в якості ядер згортки або інших аналогічних систем машинного зору. Таким чином здійснюється навчання без учителя за допомогою методу k-середніх.

1.3.2 Методи із використанням моделей Маркова

ПММ – приховані моделі Маркова, це статистичні моделі, в яких система моделюється як марковський процес із станами, що не спостерігаються (тобто прихованими). Вони добре відомі своїм застосуванням у навчанні із закріпленям та розпізнаванні часових патернів. Останнє робить можливим їх застосування до вирішення задачі виявлення аномалій. Процеси Маркова – це абстракції послідовності подій у певній системі, де ймовірність виникнення якої-небудь події залежить лише від стану системи у даний момент часу, тоді як факт виникнення якоїсь події переводить систему у інший стан. Марковські моделі часто зображають у вигляді матриць переходів, які містять ймовірності переходів між станами або у вигляді графів (рисунок 1.2).

Рисунок 1.2 – Граф Марковсього процесу

Макрковські процеси можуть бути як дискретними, тобто такими, у яких



переходу між станами відбуваються у фіксовані відрізки часу, та неперервними. У не перервних марковських замість ймовірностей переходу до наступного стану вказують інтенсивніості переходів таким способом, що в кожен момент часу система може перейти в наступний стан з певною ймовірністю, або ж залищистись у цьому ж стані надалі.

Зважаючи на це визначення, велику кількість процесів із різних областей можна вважати марковськими і застосовувати до них ПММ, щоб виявляти аномалії. Очевидним недоліком можна вважати те, що в реальному житті існують процеси, які не можуть не відповідають критеріям марковських, що унеможливлює застосування цього підходу [7, 8].

У сучасних методах виявлення аномалій, моделі Маркова широко застосовуються завдяки хз простій імплементації та низькій кількості параметрів. Однак, властивості короткитривалої пам’яті моделей Маркова низьких порядків ігнорує взажмодію даних, а у випадку моделей вищих порядків довготривала пам’ять робить їх менше очевидними що погіршує надійність моделі. Крім того, обидві моделі не здатні успішно описувати послідовності, що змінюються з деяким трендом [9].

В деяких роботах [10] пропонується використання динамічних марковських моделей. Такі підходи застосовують рухоме вікно, в якому визначається модель Маркова вищого порядку, щоб збалансувати властивості коротко- і довготривалою пам’яті та слідувати тренду послідовності. Крім того, пропонується підхід із заміною аномальних ділянок з метою запобігання їх впливу на результуючу модель.

Мережі Байеса – це ймовірнісні графові моделі, що представляють набір випадкових змінних та їхніх залежностей за допомогою орієнтованого ациклічного графа [11]. Баєсові мережі, що моделюють серію змінних називаються динамічними і можуть застосовуватись для виявлення аномалій, оскільки мають наступні властивості:

- однаково працюють як з дискретними так і з неперервними даними;
- підтримують моделі великої розмірності;
- допускають часткову відсутність даних (як протягом навчання, так і протягом виявлення/передбачення аномалій);
- одна й та ж модель може містити дані без відношенні до часу, так і часові.

Застосовуються так звані алгоритми локалізації конфліктів. Початкова мережа розбивається на підмножини із накладанням, для кожної з яких будується спеціальна функція ціни. Шукаючи максимум цієї функції ми знайдемо ділянку, яка з великою ймовірністю є джерелом аномалії. Це дозволяє звузити коло початкового пошуку і далі застосовувати інший метод для точного виявлення аномалій. Недоліком можна вважати низьку точність, тобто даний підхід може використовуватись тільки як допоміжний.

1.3.3 Методи із застосуванням нейронних мереж

Long Short-term memory networks – нейронні мережі з довгою короткостроковою пам’яттю є підвидом рекурентних нейронних мереж. РНМ – це нейронні мережі, зв’язки між вузлами яких формують орієнтовані ациклічні графи, таким чином, що дає змогу РНН використовувати свій внутрішній стан, щоб обробляти послідовності вхідних даних. РНН можуть вивчати залежності у послідовностях протягом тривалих проміжків часу.

LSTM часто застосовуються у таких сферах, як розпізнавання та синтез мови, розпізнавання рукописного вводу, автоматизовані системи управління та інші. РНН із довгою короткостроковою пам’яттю можуть застосовуватись для виявлення аномалій як без спостереження, потребуючи тільки задання порогових значень похибки, так і зі спостереженням. В останньому випадку необхідний набір даних із достатнім числом відмічених аномалій для тренування.

Застосування LSTM для виявлення аномалій полягає у передбаченні кількох значень з майбутнього, що відповідають характеру процесу, та порівняння їх із реальними значеннями. У разі їх невідповідності можна стверджувати про наявність аномалії. Щоб краще передати часову структуру послідовності, зазвичай виконується передбачення кількох значень наперед, що здійснюється у різні моменти часу в минулому.

1.3.4 Огляд інших методів виявлення аномалій

У відкритому доступі можна знайти численні публікації, присвячені алгоритмам виявлення аномалії. Однією з таких робіт є [12]. Дана робота не спрямована на розробку або поліпшення якого-небудь алгоритму, а є агрегацією з детальним описом різноманітних варіантів задач, що тим чи іншим, способом покладаються на виявлення аномалій, а також детальну класифікацію існуючих підходів та практик їх вирішення. У ній розглядаються різноманітні аспекти виявлення аномалій, оскільки кожне специфічне формулювання проблеми визначається численими факторами, такими як природа вхідних даних, наявність або відсутність поміток в даних а також обмеженнями та вимогами, які накладаються предметною областю застосування. У джерелі [13] структурований огляд на застосування виявлення аномалій із охопленням багатьох областей дослідження та застосування. Крім того, виділяються 2 нові категорії технік виявлення аномалій, а також вводиться поняття комплексних аномалій.

Одним з нових підходів і методів, що розглядаються у [14], є спектральні техніки виявлення аномалій, що базуються на припущенні, що деякі дані можуть бути вбудовані у підпростір з меншою кількістю вимірів, де нормальні та аномальні дані матимуть різкі та очевидні відмінності.

Іншим новим підходом, розглянутим у [15] є техніка виявлення аномалій, що базується на теорії інформації, із застосуванням теоретичних мір, таких як складність Колмогорова, ентропія, відносна ентропія, та ін. Припускається, що аномалії в сирих даних створюють закономірні порушення в інформації, що переноситься цими даними.

Темою публікації [16] є виявлення аномалій в різноманітних часових рядах, тобто даних представлених послідовністю значень у певний момент часу. Автори дають вичерпну класифікацію існуючих форматів вхідних даних, серед яких: одновимірні часові серії, потокові дані, просторово-часові дані та мережеві дані. Також наводяться характерні риси даних з різних областей, таких як біологія, астрономія, економіка та ін.

У роботі [17] пропонуються до застосування алгоритми k-найближчих-сусідів, довірчі мережі Байеса, приховані моделі Маркова, а також рекурентні нейронні мережі з довгою короткостроковою пам'яттю.

Метою дослідження є підвищення швидкості прийняття рішень в автоматизованих системах управління за рахунок розробки алгоритму передбачення аномалій у процесах, представлених часовими рядами.

Для досягення мети необхідно виконати наступні завдання:

- здійснити огляд і порівняльний аналіз алгоритмів виявлення аномалій у часових рядах;
- формалізувати задачу виявлення аномалій у часових рядах;
- розробити алгоритм виявлення аномалій на основі лінгвістичних моделей;
- експериментально дослідити його ефективність та надійність;
- проаналізувати отримані результати.

1.4 Висновки до розділу 1

В даному розділі було наведено основні поняття, що відносяться до розгляданої предметної області. Було здійснено огляд літератури за темою дисертації. В працях зазначається, що існує три основні групи алгоритмів виявлення аномалій, кожна з них має свої переваги і недоліки. До переваг алгоритмів з учителем можна віднести точність та швидкість, недолік – необхідно готувати набір даних для навчання. Алгоритми без учителя дещо програють у точності, і мають значно більшу обчислювальну складність, проте є автономними і не потребують додаткових підготовок і втручань, а одразу можуть працювати із початковими даними.

Зважаючи на наведені твердження, не існує алгоритму який би поєднував у собі кращі риси обох підходів і при цьому не мав би значних недоліків та обмежень.

2 ЗАСТОСУВАННЯ ЛІНГВІСТИЧНОГО МОДЕЛЮВАННЯ ДО ВИЯВЛЕННЯ АНОМАЛІЙ У ЧАСОВИХ РЯДАХ

2.1 Принции лінгвістичного моделювання

Лінгвістичний метод виявлення аномалій полягає у застосуванні лінгвістичного моделювання процесу. Термін «лінгвістичне моделювання» бере свій початок з робіт американського науковця китайського походження King-Sun Fu. В його роботах були застосовані синоніми — «структурний підхід», «синтаксичний підхід», «лінгвістичний підхід» [21].

Головною метою лінгвістичного моделювання є перетворення числових рядів, експериментальних, багатовимірних даних до лінгвістичних послідовностей та виведення за ними формальної граматики мови відповідного характеру для вирішення наступного спектру проблем: аналіз та прогнозування часових рядів, розпізнавання образів різноманітної природи, автентифікація користувача за його рухами, розпізнавання емоційного стану оператора, діагностика хвороб опорно-рухового апарату операторів складних технічних систем на ранніх стадіях захворювання [22].

Лінгвістичне моделювання базується на трьох основних підходах: структурний підхід та математична лінгвістика, інтервальні обчислення та робастні методи, сучасні методи ймовірнісного моделювання. В основі лінгвістичного моделювання лежить лема існування ізоморфізма відворення чисельних даних до лінгвістичних послідовностей, на основі яких може бути побудована мова. Висновок, який злідує з цього – факт існування унікальної мови, яка виражається послідовністю чисел [22].

Ідея структурного підходу полягає у заміні числових значень часового ряду на символи наперед визначеного алфавіту та подальшого виведення правил граматики деякої мови, за якими утворюєтья даний лінгвістичний ланцюг.

Після того, як стає відомою граматика, що відповідає нормальному, не аномальному ряду, ми можемо співставляти її із граматикою побудованою на інших значеннях цього ж ряду. Якщо буде помічено значні відхилення, можна сказати, що ці значення є аномальними.

Граматика, що виводиться, є лінгвістичною моделлю і являє собою ніщо інше, як матрицю переходів між станами дискретного марковського процесу. Даний марковський процес є дискретним, тому що інтервали часу між спостереженнями у ряді є рівними. Кожен символ алфавіту у лінгвістичному ланцюгу відповідає стану процесу у даний момент часу [23]. У кожному новому рівні ряду система переходить у інший стан із певною ймовірністю.

Оскільки, алфавіт є скінченним та відомим, усі можливі варіанти переходів між станами можна представити у вигляді матриці передування G, що має розмірність $N \times N$, де N – потужність множини символів алфавіту $|A|$. Кожен елемент матриці $g_{i,j}$ відповідає частоті (тобто, ймовірності) переходу із стану i у стан j. Дані ймовірності визначаються статистично. Як вже було зазначено, дана матриця є матрицею переходів марковського процесу і водночас – статистичним представленням граматики деякої мови j. Дані ймовірності визначаються статистично. Як вже було зазначено, дана матриця є матрицею переходів марковського процесу і водночас – статистичним представленням граматики деякої мови L, що відповідає представленому часовим рядом процесу.

2.2 N-грами як один із видів лінгвістичних моделей

У комп’ютерній лінгвістиці та теорії ймовірності, N-грами (англ. N-gram) – це суміжні послідовності з n елементів для даного зразка тексту або мови. Ці елементи можуть бути фонемами, звуками, літерами або словами в залежності від області застосування. У таблиці 2.1 наведено приклади використання N-грам у різних сферах.

Таблиця 2.1 – Приклади використання N-грам

Область	Одиниця	Зразок послідовності	1-грам	2-грам	3-грам
Порядок Марковської моделі					
Секвенування білків	амінокислота	Cys-Gly-Leu-Ser-Trp	Cys, Gly, Leu, Ser, Trp	Cys-Gly, Gly-Leu, Leu-Ser, Ser-Trp	Cys-Gly-Leu, Gly-Leu-Ser, Leu-Ser-Trp
Генетика	основна пара	AGCTTCGA	A, G, C, T, T, C, G, A	AG, GC, CT, TT, TC, CG, GA	AGC, GCT, CTT, TTC, TCG, CGA
Комп'ютерна лінгвістика	літера	to_be_or_not_to_be	t, o, _, b, e, _, o, r, _, n, o, t, _, t, o, _, b, e	to, o_, _b, be, e_, _o, or, r_, _n, no, ot, t_, _t, to, o_, _b, be	to_, o_b, _be, be_, e_o, _or, or_, r_n, _no, not, ot_, t_t, _to, to_, o_b, _be
Комп'ютерна лінгвістика	слово	to be or not to be	to, be, or, not, to, be	to be, be or, or not, not to, to be	to be or, be or not, or not to, not to be

N-грам моделі відносяться до імовірнісних мовних моделей для прогнозування наступних елементів в послідовності у формі моделі Маркова порядку $(n-1)$. Вони широко застосовуються в теорії ймовірності, комп'ютерній лінгвістиці, біології та у алгоритмах компресії даних. Дві головних переваги N-грам моделей це їх простота та здатність до масштабування – із збільшенням n модель може зберігати більший контекст із добре зрозумілим балансом між пам'яттю, що використовується, та часом виконання обчисленнь.

N-грам модель прогнозує елемент x_i деякої послідовності на основі елементів, що передують йому $x_{i-(n-1)}, \dots, x_{i-1}$. В термінах ймовірності, це позначається $P(x_i \mid x_{i-(n-1)}, \dots, x_{i-1})$. При застосуванні для моделювання мови, незалежні припущення здійснюються по відношенню до кожного слова. Таким

чином, кожне слово залежить від $n-1$ попередніх. Ця модель Маркова використовується як апроксимація справжіньої мови, що лежить в її основі. Це припущення є дуже важливим, оскільки воно значно спрощує проблему оцінки мовної моделі. На додаток до цього, для натуральної мови властиво групувати слова, які до неї не належать, у груп поруч одне з одним.

Варто зазначити, що в простій мовній моделі, ймовірність знаходження слова на певній позиції, обумовлена деякою кількістю попередніх слів (одним для бі-грам моделі, двома для три-гра і т. д.), модже бути описана як така, що слідує категорійному розподілу.

На практиці, розподіли ймовірностей згладжуються за допомгою різноманітних методів (зокрема, методом лінійної інтерполяції), а словам, яких не було в навчальних даних присвоюють ненульові ймовірності.

2.3 Загальна схема побудови лінгвістичної моделі

Нехай дано часовий ряд $X = \{x_1, x_2, x_3, \dots, x_m\}$, рівні якого виміряні через однакові проміжки часу. На першому етапі необхідно визначитись із алфавітом та виконати інтервалізацію ряду. Оберемо в якості алфавіту множину малих латинських літер – $A = \{a, b, c, \dots, z\}$, $N = |A| = 26$. Кількість символів алфавіту визначає кількість інтервалів, по яких треба розподілити значення рівнів ряду. Зручно включати до алфавіту як великі, так і малі літери та використовувати їх для позначення додатних та від'ємних значень відповідно.

Для того, щоб подувати інтервали необхідно визначити X_{min} та X_{max}. Визначаємо крок інтервалу – $step = (X_{max} - X_{\min})/N$. Розбиваємо інтервал $[X_{min}; X_{max}]$ на N частин [23]:

$$[X_{min}; X_{max}] = [X_{min}; X_{min} + step], [X_{min} + step; X_{min} + 2 \times step], \dots,$$
$$[X_{min} + j \times step; X_{min} + (j+1) \times step], \dots, [X_{min} + (N-2) \times step; X_{max}]$$

Поставимо у відповідність кожному інтервалу літеру із алфавіту:

$$a = [X_{min}; X_{min} + step],$$
$$b = [X_{min} + step; X_{min} + 2 \times step],$$
$$\dots,$$

$$k = [X_{min} + j \times step; X_{min} + (j+1) \times step],$$
$$\dots,$$
$$z = [X_{min} + (N-2) \times step; X_{max}].$$

Далі виконується підстановка еквівалентних символьних значень замість числових у вихідний ряд, за наступним принципом: якщо рівень x_i потрапляє до інтервалу j $[X_{min} + j \times step; X_{min} + (j+1) \times step]$, то замінюємо значення x_i на літеру, яка відповідає цьому інтервалу.

Таким чином, ряд $X = \{x_1, x_2, x_3, \dots, x_m\}$ перетворюється на ряд $L = \{l_1, l_2, l_3, \dots, l_m\}$, де $l_i \in A$.

Виконавши перетворення, потрібно побудувати матрицю передування. Для цього виконується підрахунок кількості входжень усіх символів, що передують символу, вказаному у рядку. Результат підрахунку представлено на рисунку 2.1.

	a	b	c	...	z	Всього
a	14	3	12	...	2	43
b	1	3	2	...	1	15
c	5	1	4	...	3	32
...	...	...	...	...	...	22
z	0	2	3	...	5	16

Рисунок 2.1 – Кількість входжень кожного символа на відповідній позиції у послідовності

Після підрахунку необхідно визначити відповідну ймовірність знаходження кожного символу на даній позицій. Для цього необхідно для кожного рядка розділити значення кожного елемента на суму елементів у рядку, тобто на загальну кількість символів, що зустрічаються після даного. Матриця частот зображена на рисунку 2.2.

	a	b	c	...	z	Сума
a	0.32	0.069	0.27	...	0.046	1
b	0.067	0.2	0.13	...	0.067	1
c	0,156	0.031	0.125	...	0.093	1
...	...	...	...	...	...	1
z	0	0.125	0.188	...	0.313	1

Рисунок 2.2 – Частота розміщень кожного символа на відповідній позиції у послідовності

В результаті отримуємо матрицю із ймовірностями переходу станів марковського процесу, яка і є нашою лінгвістичною моделлю.

Іноді, для побудови моделі доцільно використовувати не оригінальний ряд, а його похідні – першу, другу різниці [25]. Використання похідних дає змогу виявити неочевидні закономірності, які важко помітити у вихідному ряді.

Першою різницею ряду $X = \{x_1, x_2, x_3, \dots, x_m\}$ є ряд $X^1 = \{x_2 - x_1, x_3 - x_2, \dots, x_m - x_{m-1}\}$, тобто ряд, який складається із попарних різниць сусідніх елементів. Друга різниця X^2 отримується шляхом застосування тих самих операцій по відношенню до членів ряду X^1.

Наявність аномалії встановлюється шляхом порівняння двох моделей - для навчального набору даних, у якому аномалії гарантовано відсутні, та для фактичних даних, у яких ми шукаємо аномалію. Для оцінки подібності застосовуються різноманітні метрики, а також методи експертних оцінок [26].

2.4 Два способи виявлення аномалій за допомогою лінгвістичних моделей

Оскільки синтаксичний метод базується на порівнянні моделей лвінгвстичних послідовностей, можна виділити 2 різні способи його застосування.

2.4.1 Виявлення факту наявності аномалій

Перший підхід покладається на створення моделі, що характеризує нормальний перебіг процесу. Для цього із ряду виключаються всі аномальні ділянки, або одна обирається одна з ділянку ряду, що не містить відхилень від нормаи – таким чином утворюється так званий референтний ряд. Для нього виконується побудова лінгвістичної моделі.

Для того щоб визначити наявність аномалій у досліджуваному ряді потрібно для нього також побудувати лінгвістичну модель, а потім порівняти обидві моделі між собою [27]. Для цього вводиться рівень відстані, перебільшення якого можна вважати аномалією. Визначатися він повинен експертами, які виставляють свої експертні оцінки для цього рівня. В якості критеріїв порівняння можна використовувати як статистичні характеристики – середнє квадратичне (англ. RMS – root mean square), так і метрики відстаней, що застосовуються до порівняння лінгвістичних послідовностей – відстані Хемінга, Левентшейна та ін. Спершу визначається еталонне значення критерія подібності $\varepsilon_{\text{ет}}$ – відстань між моделями референтної та аномальної ділянок ряду. Після цього за допомогою рухомого вікна будуються моделі для ділянок досліджуваного ряду, що мають довжину близьку до референтної. Якщо розмір порівнюваних ділянок значно відрізнятиметься – граматика одного із рядів буде частково невизначеною, що може внести значні похибки при оцінюванні подібності моделей. Для кожної ділянки, обмеженої вікном, виконується порівняння з референтною моделлю і обчислюється $\varepsilon_{\text{факт}}$ – фактична відстань між граматиками досліджуваного та еталонного рядів. У випадку, якщо $\varepsilon_{\text{ет}} \gg \varepsilon_{\text{факт}}$, можна вважати, що аномалії відстутні. Якщо ж обидва значення

мають однаковий порядок і відрізняються на незначну величину – у данній ділянця наявна аномалій.

2.4.2 Пошук конкретного шаблону аномальної поведінки

Другий спосіб полягає у побудові моделей для аномальних ділянок ряду. Таким чином вивчається характер конкретних слесків, або викидів, - патерни аномальної поведінки. Ця задача дуже подібна до задачі розпізнавання образів [28, 29]. Маючи достатній набір даних для тренування можна побудувати модель, що точно визначатиме викиди певного типу. Такі моделі можуть знайти застосування у випадках, коли типові аномалії повторюються з деякою періодичність і, зазвичай, спричинені одними й тіми ж зовнішніми або внутрішніми (процеси із зворотним зв’язком) факторами. Існує можливість використовувати одночасно не одну, а декілька моделей, і таким чином визначати різні типи аномалій у ряді, що досліджується.

Варто зазначити, що обидва підходи в значній мірі залежать від експертних оцінок рівнів подібності моделей. Зазвичай, для обробки експертних оцінок застосовують різноманітні методи багатокритеріального прийняття рішень – метод аналізу ієрархій (метод Сааті) MAI, метод аналітичних мереж (англ. ANP – analytic network processes), метод рандомізації вільних показників (англ. AIRM – aggregated indices randomization method) та інші.

2.5 Критерії подібності лінгвістичних моделей

Важливою скаладовою лінгвістичного підходу до виявлення аномалій у часових рядах є критерій, за яким оцінюється подібність двох моделей. Саме від вибору критеріїв залежить можливість застосування лінгвістичного підходу до аналізу часових рядів різної природи.

2.5.1 Метрики схожості текстів

Оскільки однією із головних складових синтаскичного методу є перетворення числових рядів до лінгвістичних послідовностей – рядків (англ. strings), доцільно припустити, що для їх порівняння можна застосовувати алгоритми визначення подібностей у текстах. До таких відносяться відстань Геммінга, відстань Левенштейна, схожість Джаро – Вінклера, та інші [30].

Алгоритм найдовшої спільної субпослідовності бере до уваги схожість між двома рядками, що грунтується на довжині послідовностей сусідніх символів, що існують в обох рядках.

Відстань Левенштейна визначає відстань між двома рядками шляхом підраховування мінімальної кількості операцій, необхідних для перетворення одного рядка в інший, де під операціями перетворення мають на увазі вставку, видалення, заміну або транспозицію (перестановку) двох суміжних символів [31].

Метод Джаро базується на визначенні кількості та порядку спільних символів у двох рядках, а його модифікація – метод Джаро – Вінклера враховує також спільні перфікси двох рядків.

Алгоритм Нідлмана-Вунша [30] є прикладом алгоритму із застосуванням динамічного програмування і був вперше застосований для порівняння біологічних послідовностей (білків, амінокислот, тощо). Він виконує глобальне вирівнювання послідовності, щоб знайти найкраще співпадіння серед двох двох символьних послідовностей.

Далі буде розглянуто деякі особливості, переваги та недоліки перераховиних вище метрик.

Відстань Геммінга

Геммінгова відстань характеризує число позицій, в яких відповідні символи двох слів однакової довжини відрізняються. В більш загальному випадку, відстань Геммінга використовується для рядків однакової довжини і слугує метрикою відмінності (функцією, що визначає відстань у метричному

просторі) об'єктів однакової розмірності. Метод був запропонований Р. Геммінгом для визначення міри відмінності між кодовими комбінаціями (двійковими векторами) у векторному просторі кодових послідовностей. Геммінгова відстань є частинним випадком метрики Міньковського.

Визначення відстані Геммінга застосовується, наприклад, при кодуванні кодом Грея в логічних пристроях, де необхідно виключити так звані «логічні гонки» - відстань між сусідніми кодами завжди рівна одиниці, тобто кожного разу змінюється тільки один біт числа. Множина слів однакової довжини утворює, так званий, метричний простір, для кожної пари елементів якого визначено Геммінгову відстань $d(x, y)$, що задовольняє аксіомам [29]:

а) $d(x, y) = 0 \Leftrightarrow x = y$ (аксіома тотожності);

б) $d(x, y) = d(y, x)$ (аксіома симетрії);

в) $d(x, z) \leq d(x, y) + d(y, z)$ (аксіома трикутника).

Із аксіом також слідує, що відстань завжди є невід'ємною ($d(x, y) \geq 0$), а також те, що вона завжди менша за довжину слів у символах ($d(x, y) \leq n$).

Типовим прикладом метричного простору трибітовий бінарний куб (рисунок 2.4), усі вершини якого, що з'єднані ребром, мають між собою відстань рівну одиниці.

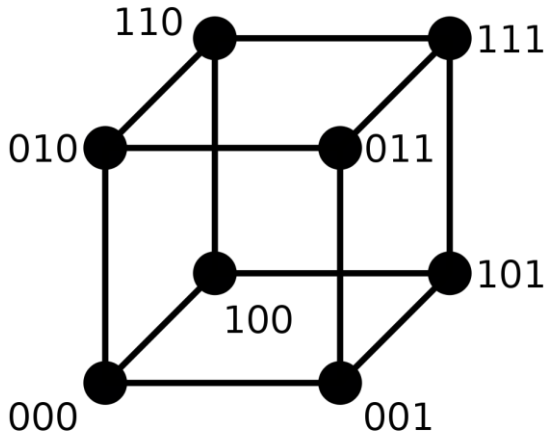


Рисунок 2.4 – Метричний простір

Перевагою цієї метрики є простота реалізації та швидкодія – зі зростанням довжини слів час роботи алгоритму зростає лінійно.

До критичних **недоліків** варто віднести:

- неможливість порівняння послідовностей різної довжини – кількість елементів в обого послідовностях має бути строго однаковою;
- алгоритм враховує співпадіння символів лише на однакових відповідних позиціях, у той час як одні й тіж граматичні ланцюги можуть зустрічатись у різних частинах послідовностей і при цьому не бути аномаліями.

Зважаючи на це, даний алгоритм не є доцільним застосовувати для визначення подібності двох лінгвістичних послідовностей.

Відстань Левенштейна (алгоритм Левенштейна, відстань редагування)

У теорії інформації так комп'ютерній лінгвістиці – метрика подібності двох лінгвістичних послідовностей (текстів). Значення цієї відстані виражається у мінімальній кількості операцій вставки, заміни і видалення, які необхідно виконати, щоб перетворити один текст в інший. Метод отримав свою назву на честь свого автора – В. Й. Левенштейна. Метод широко застосовується для виправлення помилок у словах (в пошукових системах, базах даних, при вводі тексту та автоматичному розпізнаванні відсканованого тексту або мови), порівняння текстових файлів та в біонформатиці для порівняння генів, хромосом та білків.

В загальному випадку, операції редагування мають різну ціну:

$w(a, b)$ – ціна заміни символа a на b;

$w(\varepsilon, b)$ – ціна вставки символа b;

$w(a, \varepsilon)$ – ціна видалення символа a.

Зокрема, застосовується правило трикутника – якщо дві послідовні операції можна замінити одною, то це не погіршує загальну ціну (наприклад, заміна символу x на y, а потім y на z не є кращим, ніж одразу виконати заміну з x на z).

Якщо S_1 та S_2 – два рядки, що мають довжину M та N відповідно, утворені з деякого алфавіту, тоді відстань редагування Левенштейна $d(S_1, S_2)$

можна обчислити за наступною рекурентною формулою [32] $d(S_1, S_2) = D(M, N)$, де:

$$D(i,j) = \begin{cases} 0, & \text{якщо } i = 0, j = 0; \\ i, & \text{якщо } j = 0, i > 0; \\ j, & \text{якщо } i = 0, j > 0; \\ \min \begin{pmatrix} D(i, j-1) + 1, \\ D(i-1, j) + 1, \\ D(i-1, j-1) + m(S_1[i], S_2[j]) \end{pmatrix}, & \text{якщо } i > 0, j > 0. \end{cases} \quad (2.1)$$

Тут $m(a,\ b)$ дорівнює нулю, якщо $a = b$, та одниці в іншому випадку; $\min(a, b, c)$ повертає найменший із аргументів, i та j – індекси символів у рядках. Пошук відстані полягає в обчисленні можливих варіантів замін, вставок і видалень та вибору наймешного значення на даному кроці. Матриця порівняння зображена на рисунку 2.5.

		P	O	L	Y	N	O	M	I	A	L
	0	1	2	3	4	5	6	7	8	9	10
E	1	1	2	3	4	5	6	7	8	9	10
X	2	2	2	3	4	5	6	7	8	9	10
P	3	2	3	3	4	5	6	7	8	9	10
O	4	3	2	3	4	5	5	6	7	8	9
N	5	4	3	3	4	4	5	6	7	8	9
E	6	5	4	4	4	5	5	6	7	8	9
N	7	6	5	5	5	4	5	6	7	8	9
T	8	7	6	6	6	5	5	6	7	8	9
I	9	8	7	7	7	6	6	6	6	7	8
A	10	9	8	8	8	7	7	7	7	6	7
L	11	10	9	8	9	8	8	8	8	7	6

Рисунок 2.5 – Приклад матриці порівняння для слів POLYNOMIAL та EXPONENTIAL

Удосконалений лгоритм Вагнера – Фішера для пошуку накоротшої відстані D потребує $M \times N$ пам'яті та має обчислювальну складність $O(n^2)$.

Перевагою даного метода є можливість порівнювати часові ряди у вигляді текстів, без необхідності побудови матриць передуваня.

Серед **недоліків** метода можна виділити наступні:

– потребує багато пам'яті та має відносно низьку швидкодію;
– при перестановці місцями слів або чистин слів отримують відносно великі відстані;
– відстані між абсолютно різними короткими словами виявляються незначними, у той час як відстань між довгими, але дуже схожими послідовностями, є досить великою.

З огляду на ці факти, не можна рекомендувати використання цього методу для порівняння лінгвістичних послідовностей, отриманих шляхом перетворення чисельних часових рядів, у зв'язку із їх значною довжиною.

Схожість Джаро – Вінклера

Критерій схожості двох послідовностей Дажаро – Вінклера також застосовується в інформатиці та статистиці. Як і у випадку з відстанню Левенштейна, відстань М. Джаро d_j – це кількість односимвольних перетвореннь, які необхідно виконати, щоб перетворити одне слово в інше. У. Вінклер запропонував використовувати поняття схожості, замість відмінності. Таким чином, схожість Джаро – Вінклера визначається як $d_w = 1 - d_j$, тобто чим менша відстань – тим більша подібність.

Відстань Джаро визначається для двох рядків S_1 та S_2 за формулою [33]:

$$d_j = \begin{cases} 0, & \text{якщо } m = 0; \\ \frac{1}{3}\left(\frac{m}{|S_1|} + \frac{m}{|S_2|} + \frac{m-t}{m}\right), & \text{якщо } m \neq 0, \end{cases} \tag{2.2}$$

де:

$|S_i|$ – довжина рядка S_i;

m – кількість символів, що співпадають;

t – половина кількості трапспозицій (перестановок).

Кількість перестановок визначається кількістю символів, що співпадають, але відрізняються порядковими номерами.

Відстань Джаро – Вінклера використовує масштабний коефіцієнт p, що сприяє збільшенню рейтингу рядків, що співпадають від початку до деякої довжини l, що називається префіксом. Для тих самих рядків відстань визначається формулою $d_w = d_j + lp(1 - d_j)$.

Перевагою метода Джаро – Вінклера є врахування співпадіння символів починаючи від початку послідовності до певного моменту. Це має позитивний ефект при оцінці подібності послідовностей (слів та фраз) у натуральних мовах.

Недоліком в контескті застосування відстані Джаро до порівняння часових рядів є те, що у послідовності символів однакові субпослідовності можуть знаходитись на різних позиціях. У такому випадку оцінка схожості префіксів не дає значного позитивного ефекту і відстані є завищеними, що ускладнює прийняття рішення щодо того, чи є дана послідовність аномальною.

2.5.2 Використання кореневого середньоквадратичного як міри подібності лінгвістичних моделей

З огляду на недоліки перелічених вище метрик подібності символьних послідовностей постає необхідність розробки критерію порівняння двох ймовірнісних лінгвістичних моделей, що представлені у вигляді матриць переходжів марковскього процесу.

Необхідно різницю матриць P_1 та P_2, що представляють порівнювані моделі, із деякою матрицею P' третьої моделі. Необхідно підрахувати суму різниць всіх відповідних елементів двох матриць та порівняти отримані для обох моделей числа між собою. При цьому, оскільки елементами матриць є частоти (або ймовірності), з якими зустрічаються ті чи інші елементи алфавіту у лвінгвстичній послідовності, ми маємо справу лише з невід'ємними числами.

$$\varepsilon = \sum_{i=1}^{N} \sum_{j=1}^{N} (p_{i,j} - p'_{i,j}), \tag{2.3}$$

де N – розмірність алфавіту (кількість елементів матриць);

$p_{i,j}$ – елемент матриці однієї з порівнюваних моделей P;

$p'_{i,j}$ – елемент матриці третьої моделі.

Проблемою цього підходу є те, що його результати можуть бути легко спотворені за рахунок накопичення похибки при обробці чисел з плаваючою комою. Крім того, ще одним значним недоліком є те, що даний спосіб не дає можливості розрізняти велику кількість незначних відхилень від однієї вираженої аномалії.

Більш досконалим підходом буде застосування кореневого середньоквадратичного. У такому разі формула 2.1 набуває наступного вигляду:

$$\varepsilon = \frac{1}{N}\sqrt{\sum_{i=1}^{N}\sum_{j=1}^{N}(p_{i,j} - p'_{i,j})^2}. \quad (2.4)$$

Дана формула нівелює основні недоліки, а саме – від'ємні різниці відповідних елементів підносяться до квадрату, таким чином вплив має абсолютне значення різниці, при чому чим більша складова різниця елементів двох матриць – тим більшу вагу вона матиме в кінцевому результаті. Операція ділення на кількість елементів матриці $N \times N$ винесена за знак кореня і грає роль нормалізуючого коефіцієнта [34], таким чином зменшуючи діапазон можливих значень ε.

2.6 Висновки до розділу 2

В даному розділі були розглянуті принципи лінгвістичного моделювання, а також методи вирішення супутніх йому задач. Серед них варто виділити:

- процес диференціювання ряду (для переходу від нестаціоноарного ряду до стаціонарного);
- розбиття ряду на інтервали;
- перехід від числової послідовності до лінгвістичної;
- створення ймовірнісної мовної моделі для отриманого лінгвістичного ряду;

– порівняння та оцінка подібності двох лінгвістичних моделей.

Було розглянуто можливість порівняння рядів у вигляді символьних послідовностей із використанням різних метрик подібності рядків (відстань Левенштейна, відстань Геммінга, схожість Джаро – Вінклера), та порівняннь ймовірнісних мовних моделей.

Розроблено метод лінгвістичного моделювання процесу, представленого рядом динаміки, на основі моделей N-грам першого порядку.

Наведено два можливі підходи до виявлення аномалій за допомогою лінгвістичних моделей, а саме пошук будь-яких відхилень від норми у даному часовому ряді та пошук конкретного патерну поведінки у послідовності.

Запропоновано стастичний критерій подібності ймовірнісних мовних моделей, що базується на використанні кореневого середньоквадратичного.

3 РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНИХ ДОСЛІДЖЕНЬ

Для доведення ефективності розробленого алгоритму було проведено ряд експериментів. В ході експериментів було здійснено аналіз часових рядів із фінансової сфери за допомогою забезпечення, що є програмною реалізацією розроблених методів.

Крім того було здійснено дослідження швидкодії розробленого алгоритму.

3.1 Опис програмного забезпечення

3.1.1 Вимоги до програмного забезпечення

Під час виконання завдань даної роботи було розроблено програмне забезпечення для виконання лінгвістичного моделювання процесів, представлених часовим рядом.

Перед розроблюваним програмним забезбеченням для виконання лінгвістичного моделювння були поставлені наступні вимоги:

- програмне забезпечення повинно давати можливість працювати з будь-яким часовим рядом у форматі CSV;
- користувач повинен мати можливість задавати кількість інтервалів для лінгвістичного перетворення;
- програмне забезпечення повинно мати зручний та зрозумілий інтерфейс користувача.

3.1.2 Засоби розробки

Розробка програмного забезпечення виконувалась із використанням мови програмування Java. Java – це кросплатформенна об’єктно-орієнтована мова програмування, що розробляється корпорацією Oracle. На вибір цієї мови вплинув ряд факторів, а саме:

- доступність – компілятор та середовище виконання розповсюджуються безкоштовно;

- велика спільнота розробників та числені навчальні ресурси;
- наявність великої кількості сторонніх бібліотек, зокрема для роботи з CSV файлами;
- можливість запуску програм, написаних цією мовою у різних операційних системах.

Розробка виконувалась у інтегрованому середовищі розробки IntelliJ IDEA. Воно включає до свого складу текстовий редактор, компілятор, систему збірки, а також систему контролю версій.

3.1.3 Опис програмної реалізації

Програмна реалізація алгоритму складається з трьох основних класів: класу Model, ArrayUtils та RangeMap. Плакат «UML-діаграма класів» Додатку А ілюструє відношення між компонентами програми.

Клас SeriesUtils містить універсальні методи для пошуку мінімального, максимального значень часового ряду.

Клас Model є головним класом, який є абстракцією лінгвістичної моделі. В ньому містяться методи для диферинціювання та інтервалізації вхідних рядів, побудови алфавіту, трансформації числового ряду до лінгвістичного, визначення статичтичних характеристик моделі, а також оцінки двох моделей та визначення ступеню їх подібності.

Клас RangeMap реалізує деревовидну структру, яка застосовується при інтервалізації ряду. Вона дає змогу встановлювати відповідність між діапазонами числових значень та елементами алфавіту і забезпечує ефективний пошук елементів за логарифмічний час $O(\log_2 n)$.

Точкою входу програми є клас Main, який забезпечує зчитування вхідних даних, що вводяться користувачем, а саме: розмір алфавіту, шлях до CSV файлу із даними, тривалість періоду, кількість періодів та зсув від початку часового ряду, які будуть використані для побудови моделі.

3.2 Методика експерименту

Експериментальні дослідження здійснювались із застосуванням описаного вище програмного забезпечення. В якості експериментальних даних використовувався набір часових рядів біржового індексу цінних паперів 30 найбільших американскьих компаній – Dow Jones Industrial Average.

Кожна серія складається із 3019 рівнів, що охоплюють період з січня 2006 року по грудень 2017 року. Для більшості часових рядів із цього корпусу характерною особливістю є від'ємна динаміка цін на акції у 2008 – 2009 роках, пов'язана із світовою фінансово-економічною кризою. За винятком цього періоду, усі ряди є нестаціонарними із чітко вираженим трендом зростання.

Методика експерименту полягає у створенні лінгвістичної моделі для двох періодів – нормального та аномального, визначення ступеня їх подібності, а потім порівняння референтної моделі з моделлю побудованою для відповіних річних періодів цього ж ряду починаючи з 2010 року. Перший період, для якого створюється референсна модель може тривати, наприклад, від початку 2006 до початку 2008 року, тобто складатися з 503 спостережень і відображає характер динаміки ціни цінних паперів для даної компанії. Другий період – триватиме з січня 2008 по кінець грудня 2009 року, складається 505 спостережень та містить аномально від'ємну динаміку. В залежності від особливостей ряду, границі періодів можуть змінюватись в індивідуальному порядку.

Для даних моделей обчислюється характеристика подібності $\varepsilon_{\text{ет}}$ за формулою (2.4). Для даних з 2010 по 2017 рік будуються моделі, кожна з яких охоплює період 2 роки. По суті, це є рухоме вікно із кроком 1 ріки та шириною – 2 роки. Для кожної з отриманих дворічних моделей обчислюється $\varepsilon_{\text{факт}}$ за тією ж формулою (2.4). Схема експерименту наведена у Додатку А на плакаті «Схема експерименту».

Варто зазначити, що модель будується не для оригінального ряду, а для його першої різниці.

3.3 Аналіз часових рядів за допомогою лінгвістичних моделей

3.3.1 Дослідження ряду Google

Першим рядом для дослідження було обрано ряд під назвою *GOOGL_2006-01-01_to_2018-01-01* із вище згаданого архіву.

На рисунку 3.1 зображено графік ціни на акції для компанії Google Inc.

Рисунок 3.1 – Ціна акцій Google Inc.

Для даного ряду очевидним є негативний тренд у 2008 – 2009 року, тому саме цей період було обрано в якості аномального. Крім того, із графіка динаміки цін (рисунок 3.2) видно різкі сплески у 2014 – 2016 роках. Задачею методу було виявити ці відхилення.

Рисунок 3.2 – Динаміка цін акцій Google Inc.

В ході оцінки моделей було отримано результати, наведені у таблиці 3.1.

Таблиця 3.1 – Результати оцінки моделей для ряду GOOGL

$\varepsilon_{ет}$	$\varepsilon_{9,10}$	$\varepsilon_{10,11}$	$\varepsilon_{11,12}$	$\varepsilon_{12,13}$	$\varepsilon_{13,14}$	$\varepsilon_{14,15}$	$\varepsilon_{15,16}$	$\varepsilon_{16,17}$	$\varepsilon_{17,18}$
0.0649	0.4256	0.0633	0.0648	0.0609	**0.0782**	0.0499	**0.0858**	0.0373	0.0474

Із результатів утаблиці 3.1 видно, що значеня $\varepsilon_{13,14}$ та $\varepsilon_{15,16}$, за 2013 – 2014 та 2015 – 2016 роки відповідно, перевищують значеня $\varepsilon_{ет}$ таким чином, можна вважати що періоди 2013 – 2014 та 2015 – 2016 є аномальними. Значення за 2011 рік виявилось близьким до пороговорого, проте не перевищило його, отже повністю відповідає характеру досліджуваного процесу.

3.3.2 Дослідження ряду Apple

Наступни дослідженим рядом є ряд *AAPL_2006-01-01_to_2018-01-01*. Він відображає зміну ціни акцій корпорації Apple Inc.

Графік на рисунку 3.3 відображає характер зміни ціни цінних паперів Apple.

Рисунок 3.3 – Ціна акцій Apple Inc.

На відміну від попереднього ряду, падіння ціни у 2008 – 2009 роках не є яскраво вираженим, проте починаючи с 2012 року динаміка росту помітно збальшується, а потім у 2014 так само стрімко спадає. Ще одне значне падіння можна спостерігати у 2015 – 2016 роках. В якості аномального періоду було використано 2012 – 2014 роки.

Перша різниця ряду зображена на рисунку 3.4.

Рисунок 3.4 – Динаміка цін акцій Apple Inc.

Після оцінки моделей були отримані результати, наведені у таблиці 3.2.

Таблиця 3.2 – Результати оцінки моделей для ряду AAPL

$\varepsilon_{\text{ет}}$	$\varepsilon_{6,7}$	$\varepsilon_{7,8}$	$\varepsilon_{9,10}$	$\varepsilon_{10,11}$	$\varepsilon_{11,12}$	$\varepsilon_{14,15}$	$\varepsilon_{15,16}$	$\varepsilon_{16,17}$	$\varepsilon_{17,18}$
0.0586	0.0431	0.0371	0.0409	0.0558	0.0444	0.0453	**0.0606**	**0.0703**	0.0459

Резльтати з таблиці 3.2 показують, що період 2015 – 2017 років є аномальним, адже значення $\varepsilon_{15,16}$ та $\varepsilon_{16,17}$ перевищують величину $\varepsilon_{\text{ет}}$. Це відповідає тому, що можна побачити на рисунку 3.3.

3.3.3 Дослідження ряду Amazon

Ряд *AMZN_2006-01-01_to_2018-01-01* відображає ціну акцій Amazon.com. Для даного ряду характерною особливістю є те, що він не має яскраво виражених відхилень, а майже протягом всього періоду спостережень стабільно зростає. Даний часовий ряд зображений на рисунку 3.5.

Рисунок 3.5 – Ціна акцій Amazon.com

Перші значні відхилення помітні наприкінці 2015 року. Отже в якості аномального періоду візьмемо спостереження з 2015 по 2016 роки. В якості еталонного ряду оберемо період 2009 – 2010 років. Знайдемо першу похідну ряду AMZN. Вона зображена на рисунку 3.6.

Рисунок 3.6 – Динаміка цін акцій Amazon.com

Після побудови і оцінки моделей було отримано наступні результати, наведені у таблиці 3.3.

Таблиця 3.3 – Результати оцінки моделей для ряду AMZN

$\varepsilon_{ет}$	$\varepsilon_{6,7}$	$\varepsilon_{7,8}$	$\varepsilon_{8,9}$	$\varepsilon_{10,11}$	$\varepsilon_{11,12}$	$\varepsilon_{12,13}$	$\varepsilon_{13,14}$	$\varepsilon_{16,17}$	$\varepsilon_{17,18}$
0.0703	0.0467	0.0371	0.0409	0.0487	0.0459	0.058	0.0583	**0.0806**	0.0689

За результатами оцінки видно, що $\varepsilon_{16,17}$ перевищує значення $\varepsilon_{\text{ет}}$, обчислене для 2015 року, отже можна зробити висновок, що період з 2016 по 2017 роки також є аномальним, що підтверджується графіком на Рис.3.5.

3.3.4 Дослідження ряду IBM

Часовий ряд під назвою *IBM_2006-01-01_to_2018-01-01* відображає ціни на акцій компанії IBM. Часовий ряд зображено на рисунку 3.7. Характерними особливостями цього ряду є різке падіння ціни у 2008 і 2014 роках. В якості аномального періоду оберемо інтервал, коли значення ряду вперше втраили додатний приріст, а саме кінець 2007 року. Таким чином в якості еталонного періоду обираємо 2006 – 2007 роки, а в якості аномального 2008 – 2009.

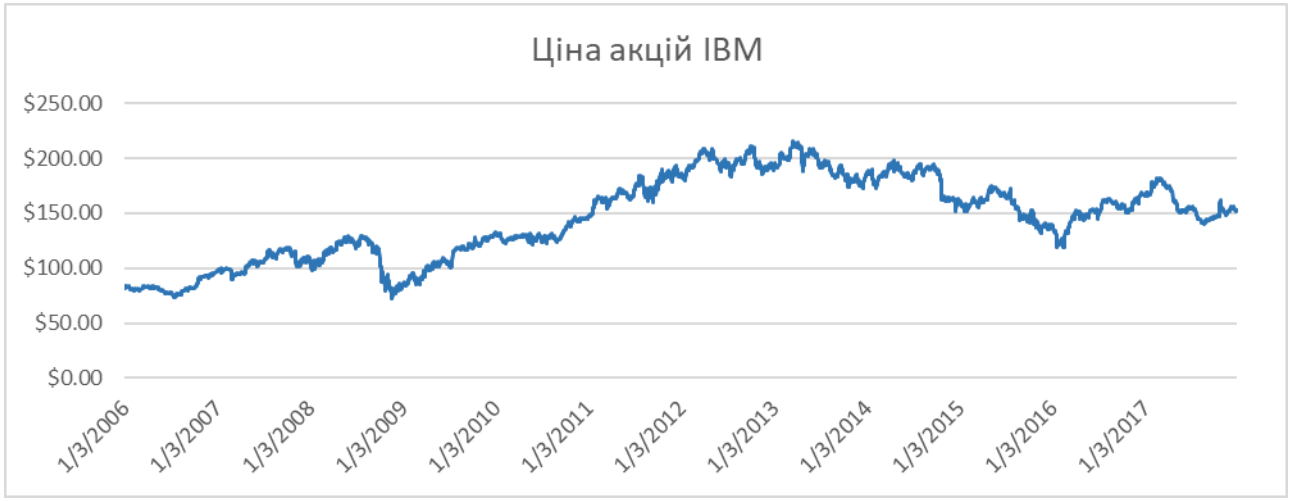


Рисунок 3.7 – Ціна акцій IBM

Перша різниця ряду зображена на рисунку 3.8.

Рисунок 3.8 – Динаміка цін акцій IBM

Результати моделювання та оцінки моделей наведені у таблиці 3.4.

Таблиця 3.4 – Результати оцінки моделей для ряду IBM

$\varepsilon_{\text{ет}}$	$\varepsilon_{9,10}$	$\varepsilon_{10,11}$	$\varepsilon_{11,12}$	$\varepsilon_{13,14}$	$\varepsilon_{14,15}$	$\varepsilon_{15,16}$	$\varepsilon_{16,17}$	$\varepsilon_{17,18}$
0.0579	0.0562	0.0383	0.0542	**0.0670**	**0.0733**	**0.0718**	**0.0589**	**0.0712**

З результатів оцінки моделей можна дійти висновку, що починаючи з 2013 року поведінка ряду стає непередбачуваною та абсолютно не відповідає характеру 2006 – 2008 років. Дійсно, оцінки всіх моделей починаючи з 2013 року перевищують еталонне значення. Цей висновок також підтверджується графіками на рисунках 3.7 та 3.8. На графіку ціни (рисунок 3.7) чітко видно негативний тренд незрозумілої природи, який з'являється у 2013 році та змінюється короткочасним зростанням, що триває протягом 2016 року. Серед усих періодів, що були визначені як аномальні, $\varepsilon_{16,17}$ має найменше значення і майже наближається до $\varepsilon_{\text{ет}}$, однак модель класифікувала його як викид.

3.3.5 Дослідження ряду Microsoft

Ряд *MSFT_2006-01-01_to_2018-01-01* містить дані про ціну акцій Microsoft. Графік ряду зображено на рисунку 3.9.

Рисунок 3.9 – Ціна акцій Microsoft

Характерною особливістю цього ряду є дуже повільна динаміка змін. Крім того, майже відсутні різкі стрибки значень. Тим не менше, на графіку

видно від'ємний тренд в період з 2008 по 2009 рік. Також можна помітити коливання у 2015 році. Знайдемо першу похідну ряду, яка відображатиме динаміку зміни ціни. Вона зображена на рисунку 3.10.

Рисунок 3.10 – Динаміка цін акцій Microsoft

Побудуємо лінгвістичні моделі для відповідних періодів спостереження та здійснимо їх оцінку. Результати оцінки моделей для ряду MSFT наведено у таблиці 3.5.

Таблиця 3.5 – Результати оцінки моделей для ряду MSFT

$\varepsilon_{ет}$	$\varepsilon_{9,10}$	$\varepsilon_{10,11}$	$\varepsilon_{11,12}$	$\varepsilon_{13,14}$	$\varepsilon_{14,15}$	$\varepsilon_{15,16}$	$\varepsilon_{16,17}$	$\varepsilon_{17,18}$
0.0796	0.0748	0.0703	0.0711	0. 0743	0.0693	**0.0941**	**0.0921**	0.0712

З результатів моделювання та оцінки моделей видно, що для 2015 та 2016 років $\varepsilon_{15,16}$ та $\varepsilon_{16,17}$ перевищують порогове значення, що свідчить про наявність аномалії. Це підтверджується графіком на рисунку 3.9.

3.3.6 Дослідження ряду Wallmart

Часовий ряд *WMT_2006-01-01_to_2018-01-01* відповідає даним про акції компанії Wallmart. Для даного ряду, згідно з графіком на рисунку 3.11, яскраво

вираженою аномалією є період 2015 – 2016 років. В якості періоду з адекватними даними оберемо 2006 – 2007 роки.

Рисунок 3.11 – Ціна акцій Wallmart

Знайдемо першу різницю ряду (рисунок 3.12).

Рисунок 3.12 – Динаміка цін акцій Wallmart

Результат моделювання та оцінки моделей для ряду WMT наведено у таблиці 3.6.

Таблиця 3.6 Результати оцінки моделі для ряду WMT

$\varepsilon_{ет}$	$\varepsilon_{8,9}$	$\varepsilon_{9,10}$	$\varepsilon_{10,11}$	$\varepsilon_{11,12}$	$\varepsilon_{13,14}$	$\varepsilon_{14,15}$	$\varepsilon_{16,17}$	$\varepsilon_{17,18}$
0.0705	0.0696	0.0358	0.0383	0.0342	0.0338	0.0358	0.0521	0.0696

За результатами моделювання, у ряді WMT не було виявлено аномалій, оскільки всі значення ε виявились меншини за $\varepsilon_{ет}$. Значення $\varepsilon_{8,9}$ було близьким до того, щоб бути класифікованим як аномалія, проте не перебільшило порогового значення. Дійсно, з графіка на рисунку 3.11 видно, що окрім вище зазначеного періоду 2015 – 2016 років ряд містить від'ємний тренд у 2008 році, отже даний результат можна вважати похибкою моделі.

3.3.7 Дослідження ряду Nike

Часовий ряд *NKE_2006-01-01_to_2018-01-01* відображає характер змін ціни акцій компанії Nike. Даний ряд містить значні коливання у 2016 – 2018 роках, це видно на рисунку 3.13.

Рисунок 3.13 – Ціна акцій Nike

В якості аномального періоду візьмемо 2015 – 2017 роки. В якості еталонних даних візьмемо 2006 – 2007 роки. Знайдемо першу різницю ряду (рисунок 3.14).

Рисунок 3.14 – Динаміка цін на акції Nike

Результати моделювання та оінцки моделей наведено у таблиці 3.7.

Таблиця 3.7 Результати оцінки моделі для ряду NKE

$\varepsilon_{\text{ет}}$	$\varepsilon_{7,8}$	$\varepsilon_{8,9}$	$\varepsilon_{9,10}$	$\varepsilon_{10,11}$	$\varepsilon_{11,12}$	$\varepsilon_{12,13}$	$\varepsilon_{13,14}$	$\varepsilon_{17,18}$
0.0621	0.0330	0.0345	0.0366	0.0484	0.0408	0.0358	0.0586	**0.0663**

З результатів моделювання видно, що єдина значна аномалія була виявлена у період 2017 – 2018 року. Це підтвердужє графік на рисунку 3.13. Зведені результати дослідження для всіх рядів наведені на плакаті «Результати експериментів» Додатку А.

3.4 Оцінка швидкодії алгоритму

Важливою характеристикою будь якого алгоритму є його швидкодія. Характеристикою швидкодії алгоритму є його обчислювальна складність. Зазвичай вона виржається вигляді залежності часу обчислень до розміру вхідних даних.

Для розробленого алгоритму лінгвістичного моделювання було здійснено виміри швидкодії, результат зображено на рисунку 3.15.

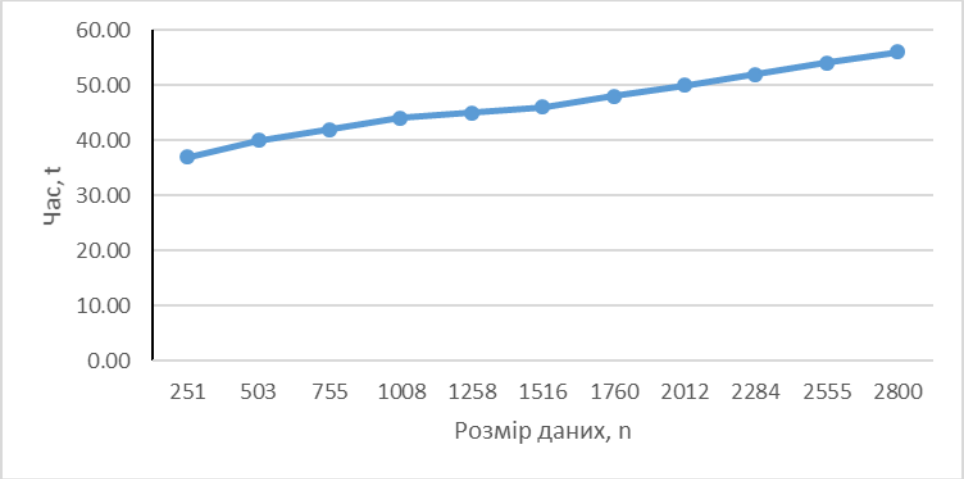


Рисунок 3.15 – Залежність часу виконання обчислень від обсягу вхідних даних

Звідси можна зробити висновок, що час роботи алгоритму зростає лінійно $(O(n))$ із збільшенням обсягу вхідних даних. Це значно краще, ніж у популярних алгоритмів на основі кластеризації $(O(n^2))$, методів із застосуванням перетворення Фур'є $\left(O(nlog_2 n)\right)$ та нейронних мереж.

3.3 Висновки до розділу 3

Даний розділ було присвячено експериментальним дослідженням ефективності розробленого алгортму лінгвістичного моделювання. На початку розділу була описана архітектура та основні складові програмної реалізації алгоритму. Описано методику експерименту та характерні особливості експериментальних даних.

Було проведено 7 експериментів із часовими рядами фінансово-економічних показників. В ході кожного експерименту здійснювався аналіз особливостей ряду, обирались інтервали ряду для побудови лінгвістичної моделі. Оцінка моделей здійснювалась шляхом порівняння річних моделей ряду із еталонною моделлю. В результаті були обчислені коефіцієнти подібності моделей, які наведені в таблицях після опису ходу експерименту. За результатами експериментів зроблено висновки про ефективність даного методу виявлення аномалій у часових рядах, а саме:

– у 5 з 7 експериментів було виявлено всі аномальні ділянки;
– в одному експерименті єдина аномалія у ряді не була виявлена;

– в одному експерименті модель помилково класифікувала аномалію.

Таким чином, можна стверджувати, що метод виявлення аномалій на основі лінгвістичного моделювання часового ряду є досить ефективним. Крім того, було досліджено швидкодію алгоритму реалізованого у програмному забезпеченні – час обчислень зростає лінійно із збільшенням обсягу вхідних даних. Цей результат є кращим, ніж у популярних алгоритмів, що базуються на методах кластеризації, застосуванні перетворення Фур’є або методах, що використовують нейронні мережі.

4 РОЗРОБЛЕННЯ СТАРТАП-ПРОЕКТУ

Метою даного розділу є проведення маркетингового аналізу стартап проекту для визначення принципової можливості його ринкового впровадження та можливих напрямів реалізації цього впровадження.

4.1 Опис ідеї проекту

Опис ідеї стартап-проекту наведено у таблиці 4.1.

Таблиця 4.1 – Опис ідеї стартап-проекту

Зміст ідеї	**Напрямки застосування**	**Вигоди для користувача**
Розробка програмної платформи, що дозволяє користовучу додавати різноманітні пристрої та задавати шаблони аномальної поведінки з метою подальшого виявлення такої поведінки та попередження власника пристрою	1. Моніторинг параметрів приміщення	Дозволяє контролювати параметри пристроїв, надсилає сигнал тривоги у разі, коли параметри виходять за межі норми
	2. Віддалений контроль пристроїв	Дозволяє вмикати/вимикати пристрої
	3. Автоматизація пристроїв	Автоматизація, вмикання і вимикання пристроїв за сценаріями

У таблиці 4.2 наведено сильні, слабкі та нейтральні характеристики ідеї проекту.

Таблиця 4.2 – Визначення сильних, слабких та нейтральних характеристик ідеї проекту

№ п/п	Техніко-економічні характеристики ідеї	(потенційні) товари/концепції конкурентів				W (слабка сторона)	N (нейтральна сторона)	S (сильна сторона)
		Мій проект	Конкурент1	Конкурент2	Конкурент3			
1	Форма використання	Веб + мобільний телефон	Веб	Мобільний клієнт	Веб + мобільний телефон			+
2	Собівартість	Низька	Низька	Низька	Висока		+	
3	Наявність інтернету	-	+	+	+			+
4	Крос-платформеність	+/-	+	-	+	+		

4.2 Технологічний аудит ідеї проекту

Технологічну здійсненність ідей проекту наведено у таблиці 4.3

Таблиця 4.3 – Технологічна здійсненність ідеї проекту

№ п/п	Ідея проекту	Технології її реалізації	Наявність технологій	Доступність технологій
1	Створення веб-додатку	JavaScript	Наявна	Платна, недоступна
		Java	Наявна	Безкоштовна, доступна
		Spring MVC	Наявна	Безкоштовна, доступна
2	Створення мобільного додатку	Android	Наявна	Безкоштовна, доступна
		iOS	Наявна	Платна, недоступна
3	Протокол взаємодії	MQTT	Наявна	Безкоштовна, доступна
		REST	Наявна	Безкоштовна, доступна
Обрана технологія реалізації ідеї проекту: Створення веб-додатку – Java тому що члени команди мають досвід роботи а також через поширеність технологій та простоті розробки, мобільного – Android бо безкоштовний а також пристроїв на цієї ОС значно більше, протокол взаємодії – MQTT, бо потребує менше ресурсів				

4.3 Аналіз ринкових можливостей запуску стартап-проекту

Попередню характеристику потенційного ринку стартап-проекту наведено у таблиці 4.4.

Таблиця 4.4 – Попередня характеристика потенційного ринку стартап-проекту

№ п / п	Показники стану ринку (найменування)	Характеристика
1	Кількість головних гравців, од	3
2	Загальний обсяг продаж, грн/ум.од	5000 грн./ум.од
3	Динаміка ринку (якісна оцінка)	Зростає
4	Наявність обмежень для входу (вказати характер обмежень)	Немає
5	Специфічні вимоги до стандартизації та сертифікації	Немає
6	Середня норма рентабельності в галузі (або по ринку), %	R = (3000000 * 100) / (1000000 * 12) = 25%

Характеристику потенційних клієнтів стартап-проекту наведено у таблиці 4.5.

Таблиця 4.5 – Характеристика потенційних клієнтів стартап-проекту

№ п/п	Потреба, що формує ринок	Цільова аудиторія (цільові сегменти ринку)	Відмінності у поведінці різних потенційних цільових груп клієнтів	Вимоги споживачів до товару
1.	Необхідність побудови розумної системи IoT у себе в дома	Люди з технічною освітою, які прагнуть оптимізувати своє життя	Різні розміри об’єктів, різні мобільні пристрої, різні сервіси інтеграції	Наявність веб інтерфейсу, віддаленого керування, програми для мобільного

Фактори загроз наведено у таблиці 4.6.

Таблиця 4.6 – Фактори загроз

№ п/п	**Фактор**	**Зміст загрози**	**Можлива реакція компанії**
1.	Конкуренція	Вихід на ринок великої компанії	1) Вихід з ринку 2) Запропонувати великій компанії поглинути себе 3) Передбачити додаткові переваги власного ПЗ для того, щоб повідомити про них саме після виходу міжнародної компанії на ринок
2.	Зміна потреб користувачів	Користувачам необхідне програмне забезпечення з іншим функціоналом	1) Передбачити можливість додавання нового функціоналу до створюваного ПЗ

Фактори можливостей наведено у таблиці 4.7.

Таблиця 4.7 – Фактори можливостей

№ п/п	**Фактор**	**Зміст можливості**	**Можлива реакція компанії**
1.	Зростання можливостей потенційних покупців	Ріст зацікавленості до продукту серед інших груп користувачів з різним рівнем технічної грамотності	Додати підказки, інструкції та демонстрації роботи системи
2.	Зниження довіри до конкурента 3	У ПЗ конкурента №3 нещодавно була знайдена помилка, завдяки чому вдалося отримати контроль над системою третій особі	При виході на ринок звертати увагу покупців на безпеку нашого ПЗ

У таблиці 4.8 наведено ступеневий аналіз конкуренції на ринку.

Таблиця 4.8 – Ступеневий аналіз конкуренції на ринку

Особливості конкурентного середовища	**В чому проявляється дана характеристика**	**Вплив на діяльність підприємства (можливі дії компанії, щоб бути конкурентоспроможною)**
1. Вказати тип конкуренції - досконала	Існує 3 фірми-конкурентки на ринку	Врахувати ціни конкурентних компаній на початкових етапах створення бізнесу, реклама (вказати на конкретні переваги перед конкурентами)
2. За рівнем конкурентної боротьби - міжнародний	Компаній – з інших країни	Додати можливість вибору мови ПЗ, щоб легше було у майбутньому вийти на міжнародний ринок
3. За галузевою ознакою - внутрішньогалузева	Конкуренти мають ПЗ, яке використовується лише всередині даної галузі	Створити основу ПЗ таким чином, щоб можна було легко переробити дане ПЗ для використання у інших галузях
4. Конкуренція за видами товарів: - товарно-видова	Види товарів є однаковими, а саме – програмне забезпечення	Створити ПЗ, враховуючи недоліки конкурентів
5. За характером конкурентних переваг - нецінова	Вдосконалення технології створення ПЗ, щоб собівартість була нижчою	Використання менш дорогих технологій для розробки, ніж використовують конкуренти
6. За інтенсивністю - марочна	Бренди присутні	Активна реклама, яка вказує на переваги саме даного рішення, натякаючи на недоліки конкурентів

У таблиці 4.9 наведено аналіз конкуренції в галузі за М. Портером.

Таблиця 4.9 – Аналіз конкуренції в галузі за М. Портером

Складові аналізу	**Прямі конкуренти в галузі**	**Потенційні конкуренти**	**Постачальники**	**Клієнти**	**Товари-замінники**
	Навести перелік прямих конкурентів	**Визначити бар'єри входження в ринок**	**Визначити фактори сили постачальників**	**Визначити фактори сили споживачів**	**Фактори загроз з боку замінників**
Висновки:	Існує 3 конкуренти на ринку. Найбільш схожим за виконанням є конкурент 3, так як його рішення також представлене у вигляді веб-додатку та мобільному додатку.	Є конкуренти, є можливість виходу на ринок	Постачальників багато, тому можна вважати, що вони не диктують умови на ринку	Важливим для користувача є наявність веб та мобільного додатку а також безпека системи	Товари-замінники можуть використати більш дешеву технологію створення ПЗ та зменшити собівартість товару.

У таблиці 4.10 наведено обґрунтування факторів конкурентоспроможності.

Таблиця 4.10 – Обґрунтування факторів конкурентоспроможності

№ п/п	**Фактор конкурентоспроможності**	**Обґрунтування (наведення чинників, що роблять фактор для порівняння конкурентних проектів значущим)**
1.	Простота інтерфейсу користувача	Простота роботи х програмою спрощує роботу користувачеві,щоробитьїїбільш комфортною та зручною
2.	Наявність додатків під різні платформи	Дозволяє працювати з системою з різних пристроїв у зручній формі

У таблиці 4.11 наведено порівняльний аналіз сильних та слабких сторін.

Таблиця 4.11 – Порівняльний аналіз сильних та слабких сторін

№ п/п	**Фактор конкурентоспроможності**	**Бали 1-20**	**Рейтинг товарів-конкурентів у порівнянні**						
			-3	**-2**	**-1**		**1**	**2**	**3**
1.	Простота інтерфейсу користувача	20				+			
2.	Наявність додатків під різні платформи	15			+				

У таблиці 4.12 наведено SWOT- аналіз стартап-проекту.

Таблиця 4.12 – SWOT- аналіз стартап-проекту

Сильні сторони: Простота інтерфейсу користувача, захищеність, наявність веб та мобільного додатку	Слабкі сторони: Відсутність IOS версії та розкрученості бренду
Можливості: у конкурента 3 виявлена проблема із безпекою ПЗ, зацікавленість продуктом ширших груп споживачів	Загрози: конкуренція, зміна потреб користувачів

У таблиці 4.13 наведено альтернативи ринкового впровадження стартап-проекту.

Таблиця 4.13 – Альтернативи ринкового впровадження стартап-проекту

№ п/п	**Альтернатива (орієнтовний комплекс заходів) ринкової поведінки**	**Ймовірність отримання ресурсів**	**Строки реалізації**
1.	Використання MQTT для взаємодій пристроїв контролю параметрів	80%	1 місяць
2.	Використання RESTfull API для взаємодій пристроїв контролю параметрів	60%	2 місяці

Обираємо альтернативу 1, тому що вона має більшу ймовірність отримання ресурсів та менший час реалізації.

4.4 Розроблення ринкової стратегії проекту

У таблиці 4.14 наведено вибір цільових груп потенційних споживачів.

Таблиця 4.14 – Вибір цільових груп потенційних споживачів

№ п/п	Опис профілю цільової групи потенційних клієнтів	Готовність споживачів сприйняти продукт	Орієнтовний попит в межах цільової групи (сегменту)	Інтенсивність конкуренції в сегменті	Простота входу у сегмент
1.	Люди з технічною освітою, що мають схильність та ентузіазм до покращення навколишнього середовища	Критичним є інтеграція з іншими сервісами, можливість використання сторонніх пристроїв	Контроль параметрів приміщення, автоматизація процесів	Існує 3 конкуренти, які надають схожі рішення. До того ж – лише 1 конкурент надає веб та мобільний додаток	У сегмент увійти просто, необхідно лише надати можливість до самостійної зміни деяких компонентів
2.	Люди, які мають достатньо коштів і небайдужі до впровадження нових рішень	Критичним є простій та зрозумілий інтерфейс	Контроль параметрів приміщень		Маючи простий та зрозумілий інтерфейс, вийти на ринок не складно
3.	Люди, які хочуть економити енергоносії	Критичним є енергоефективність	Автоматизація процесів, енергозбереження		Складно, необхідні складні механізми енергоефективності

Які цільові групи обрано: групи 1 та 2.

У таблиці 4.15 наведено визначення базової стратегії розвитку.

Таблиця 4.15 – Визначення базової стратегії розвитку

№ п/п	Обрана альтернатива розвитку проекту	Стратегія охоплення ринку	Ключові конкурентоспроможні позиції відповідно до обраної альтернативи	Базова стратегія розвитку*
1.	Використання MQTT для взаємодій пристроїв контролю параметрів	Ринкове позиціювання	Простота використання, пришвидшення роботи, крос-платформеність	Диференціації

У таблиці 4.16 наведено визначення базової стратегії конкурентної поведінки.

Таблиця 4.16 – Визначення базової стратегії конкурентної поведінки

№ п/п	Чи є проект «першопрохідцем» на ринку?	Чи буде компанія шукати нових споживачів, або забирати існуючих у конкурентів?	Чи буде компанія копіювати основні характеристики товару конкурента, і які?	Стратегія конкурентної поведінки*
1.	Ні	Так	Буде, наявність як веб так і мобільного додатку	Зайняття конкурентної ніші

У таблиці 4.17 наведено визначення стратегії позиціонування.

Таблиця 4.17 – Визначення стратегії позиціонування

№ п/п	Вимоги до товару цільової аудиторії	Базова стратегія розвитку	Ключові конкурентоспроможні позиції власного стартап-проекту	Вибір асоціацій, які мають сформувати комплексну позицію власного проекту (три ключових)
1.	Простота інтерфейсу, можливість кастомізації системи	Диференціації	Простота користувацького інтерфейсу, що дозволяє спростити роботу з системою, можливість кастомізації, що вимагає наявність можливості роботи з сторонніми системами та пристроями	Кастомізація, простота, гнучкість

4.5 Розроблення маркетингової програми стартап-проекту

У таблиці 4.18 наведено визначення ключових переваг концепції потенційного товару.

Таблиця 4.18 – Визначення ключових переваг концепції потенційного товару

№ п/п	Потреба	Вигода, яку пропонує товар	Ключові переваги перед конкурентами (існуючі або такі, що потрібно створити
1.	Кастомізація	Можливість інтеграції з іншими системами	Можливість роботи з модулями інших систем та пристроями
2.	Простота інтерфейсу	Простота та зручність ПЗ	Користувачам не потрібно замислюватися над тим як працювати з системою

У таблиці 4.19 наведено опис трьох рівнів моделі товару.

Таблиця 4.19 – Опис трьох рівнів моделі товару

Рівні товару	**Сутність та складові**		
I. Товар за задумом	Об’єкт дозволяє користувачам налаштувати параметри моніторингу стану приміщення, керувати віддалено пристроями		
II. Товар у реальному виконанні	Властивості/характеристики	М/Нм	Вр/Тх /Тл/Е/Ор
	1. Наявність веб додатку 2. Наявність мобільного додатку 3. Інтеграція з іншими сервісами 4. Простота використання	-	-
	Якість: згідно до стандарту ISO 29119 буде проведено тестування		
	Маркування присутнє.		
	Моя компанія. «ACIoT»		
III. Товар із підкріпленням	Відсутня підтримка до продажу		
	Постійна підтримка для користувачів після продажу		
За рахунок чого потенційний товар буде захищено від копіювання: ноу-хау.			

У таблиці 4.20 наведено визначення меж встановлення ціни.

Таблиця 4.20 – Визначення меж встановлення ціни

№ п/п	**Рівень цін на товари-замінники**	**Рівень цін на товари-аналоги**	**Рівень доходів цільової групи споживачів**	**Верхня та нижня межі встановлення ціни на товар/послугу**
1	6000	7000	200000	5000

У таблиці 4.21 наведено формування системи збуту.

Таблиця 4.21 – Формування системи збуту

№ п/п	**Специфіка закупівельної поведінки цільових клієнтів**	**Функції збуту, які має виконувати постачальник товару**	**Глибина каналу збуту**	**Оптимальна система збуту**
1.	Купують пристрої, ПЗ йде у комплекті	Продаж	0(напряму), 1(через одного посередника)	Власна та через посередників

У таблиці 4.22 наведено концепція маркетингових комунікацій.

Таблиця 4.22 – Концепція маркетингових комунікацій

№ п/п	**Специфіка поведінки цільових клієнтів**	**Канали комунікацій, якими користуютьс я цільові клієнти**	**Ключові позиції, обрані для позиціонування**	**Завдання рекламного повідомлення**	**Концепція рекламного звернення**
1.	Замовлення через інтернет. Використан ня пристроїв на різних платформах	Інтернет	Можливість кастомізацій, наявність веб та мобільного додатку, простота інтерфейсу	Показати переваги ПЗ, у тому числі і перед конкурентами	Демо-ролик із використ-тання

4.6 Висновки до розділу 4

Згідно до проведених досліджень:

- існує можливість ринкової комерціалізації проекту;
- існують перспективи впровадження з огляду на потенційні групи клієнтів, бар’єри входження не є високими, проект має дві значні переваги перед конкурентами;
- необхідно реалізувати веб-додаток, мобільний додаток та хаб з MQTT;
- подальша імплементація є доцільною.

ВИСНОВКИ

У даній магістерській дисертації було здійснено огляд існуючих методів та алгоритмів виявлення аномалій у даних, що представлені у вигляді часових рядів. Було проведено порівняльний аналіз даних методів та алгоритмів, виявлено їх переваги та недоліки.

Формалізовано задачу виявлення аномалій у часовому ряді, сформульовано математичну постановку.

Було розроблено алгоритм виявлення аномалій, що грунтується на використанні синтаксичних та структурних методів із використанням ймовірнісних лінгвістичних моделей. Алгоритм дає змогу виявляти аномалії двома способами – виявляти сам факт наявності аномалій у ряді або шукати конкретний шаблон аномальниї поведінки. Це досягається за рахунок вибору відповідних інтервалів тренувальних даних.

В ході розробки алгоритму було розглянуто 4 різні метрики оцінки лінгвістичних моделей. Беручи до уваги їхні сильні та слабкі сторони було обрано кореневе середньоквадратичне як найбільш прийнятний підхід.

Було розроблено програмне забезпечення для виконання лінгвістичного моделювання на основі часового ряду. Розробка даного програмного забезпечення здійснювалась із використанням мови програмування Java та інтегрованого середовища IntelliJ IDEA. Програмне забезпечення має інтерфейс командного рядка та надає можливість зміни часового ряду, алфавіту, його потужності, періодів, а також передбачає виведення на друк (файл, екран) лінгвістичного ланцюга і результатів оцінки моделі. Отримані моделі можна використовувати також для прогнозування подальших рівнів ряду.

Було здійснено 7 експериментів із зазначеною вище програмною реалізацією алгоритму пошуку аномалій на основі лінгвістичного моделювання часового ряду. Експерименти підтвердили ефективність методу, оскільки для 5 з 7 досліджуваних рядів були точно виявлені всі аномалії. В одному з

експериментів алгоритм помилково класифікував нормальні дані як аномалію. Ще в одноу з експериментів одна із аномалій не була виявлена.

Було досліджено швидкоді розробленого алгоритму. Вимірювання показали, що даний алгоритм добре масштабується – із збільшенням кількості вхідних даних час обчислень зростає лінійно. Це пояснюється тим, що потужність алфавіту, яка визначає розмір лвінгвістичної моделі, є параметром алгоритму і відома заздалегідь. Таким чином, на час виконання впливає тільки кількість рівнів часового ряду, що досліджується.

ПЕРЕЛІК ПОСИЛАНЬ

1. Chandola V., Banerjee A., Kumar V., 2009. Anomaly Detection : A Survey. ACM Computing Surveys.
2. Gupta M., Gao J., Aggarwal C.C. and Han, J., 2014. Outlier detection for temporal data: A survey. IEEE Transactions on Knowledge and Data Engineering, 26(9), pp. 2250-2267.
3. Hodge V. J., Austin J., 2004. A Survey of Outlier Detection Methodologies.
4. Прогнозування та аналіз часових рядів. Методичні вказівки до практичних занять та самостійної роботи студентів спеціальності 051 «Економіка» // Укл: Юрченко М. Є. – Чернігів: ЧНТУ. 2018. – с/ 5-15.
5. Mehrotra G., Mohan K., Huang H., 2017. Anomaly Detection Principles and Algorithms
6. Patcha A., Park J., 2007. An overview of anomaly detection techniques: Existing solutions and latest technological trends.
7. Steinwart I., Hush D., Scovel C., 2005. A Classification Framework for Anomaly Detection. C.C. Aggarwal, Outlier Analysis (Springer Science & Business Media, New York, 2013)
8. C.C. Aggarwal, C.K. Reddy, Data Clustering: Algorithms and Applications(CRC Press, Boca Raton, 2013)
9. H. Akaike, “A new look at the statistical model identification.” IEEE Trans. Automatic Control 19(6), pp. 716-723 (1974).
10. F. Angiulli, C. Pizzuti, “Fast outlier detection in high dimensional spaces,” in Principles of Data Mining and Knowledge Discovery (Springer, New York, 2002), pp. 15-27.
11. D. Asteriou, S.G. Hall, Applied Econometrics (Palgrave Macmillan, New York, 2011).
12. R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval (Addison-Wesley Longman Publishing, Boston, 1999)
13. D.J. Berndt, J. Clifford, “Using dynamic time warping to find patterns in time series,” in AAAI
14. Working Notes of the Knowledge Discovery in Databases Workshop, pp. 359–370, 1994
15. J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms (Kluwer Academic Publishers, Norwell, 1981).
16. Chawla S., Chandola V., 2011, Anomaly Detection: A Tutorial.
17. Angiulli F., Pizzuti C., 2002. Fast outlier detection in high dimensional spaces. European Conference on Principles of Data Mining and Knowledge Discovery, pp. 15-27.

18. Goldstein M., Uchida S., 2016. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. PloS one, 11(4), p.e0152173.

19. Kannan R., Woo H., Aggarwal C.C. and Park, H., 2017, June. Outlier detection for text data. In Proceedings of the 2017 SIAM International Conference on Data Mining, pp. 489-497. Society for Industrial and Applied Mathematics.

20. Kishan G. M., Chilukuri K. M., Huang H., Anomaly Detection Principles and Algorithms. Terrorism, Sercurity and Computaion, 2017 – 217 p.

21. Лінгвістичне моделювання (математичне моделювання) – Режим доступу: https://uk.wikipedia.org/wiki/Лінгвістичне_моделювання_(математичне_моделювання).

22. Логвинчук А. І. Застосування лінгвістичного моделювання до вирішення задачі пошуку аномалій / І. В. Баклан // Матеріали ІІІ всеукраїнської науково-практичної конференції молодих вчених та студентів «Інформаційні системи та технології управління» (ІСТУ-2019) – м. Київ.: НТУУ «КПІ ім. Ігоря Сікорського», 20-22 листопада 2019 р. – с. 65-67.

23. Lohvynchuk A., Baklan I. Linguistic approach for a time series anomaly detection – Slovac International Scientific Journal. – 2019. – №35, Vol. 1. – pp. 16-18

24. Баклан І. В. Аналіз поведінки економічних часових рядів з використанням структурних підходів. Сборник МКММ-2006. – Херсон: ХГТУ, 2006.

25. Баклан І. В. Лінгвістичне моделювання: основи, методи, деякі прикладні аспекти. Систем. технології. – 2011. – № 3. – с. 10-19.

26. Баклан І. В. Структурний підхід до розпізнавання образів у системах безпеки. Національна безпека України: стан, кризові явища та шляхи їх подолання. Міжнародна науково-практична конференція (Київ, 7-8 грудня 2005 р.). Збірка наукових праць. – К.: Національна академія управління – Центр перспективних соціальних досліджень, 2005. – с. 375-380.

27. Нарасимхан Р. Лингвистический подход к распознаванию образов. Автоматический анализ сложных изображений. — М.: Мир, 1969.

28. Дуда Р., Харт П., Распознавание образов и анализ сцен. – М.: Мир, 1976. – с. 53.

29. Орлов А. И. Теория принятия решений: учебник. – М.: Экзамен, 2006. – с. 274.

30. Cohen, W. A comparison of string distance metrics for name-matching tasks – KDD Workshop on Data Cleaning and Object Consolidation: journal. – 2003. – Vol.3. – p. 73.

31. Відстань Геммінга – Режим доступу: https://uk.wikipedia.org/wiki/Відстань_Геммінга.

32. Відстань Левенштейна – Режим доступу: https://uk.wikipedia.org/wiki/Відстань_Левенштейна.

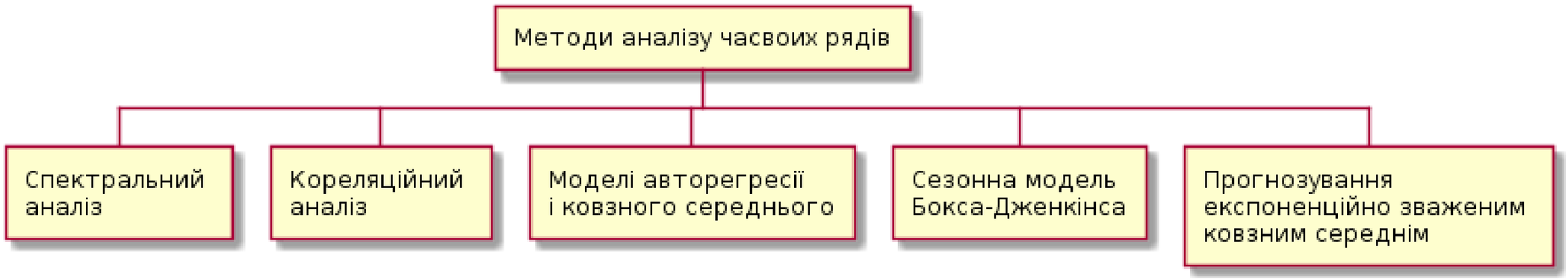
33. Jaro, M. Advantages in record linkage methodology as applied to the 1985 cenus of Tampa Florida – Journal of the American Statistical Association. – 1989 – p. 74.

34. Мышкис А. Д. Элементы теории математических моделей. — 3-е изд., испр. –М.: КомКнига, 2007. – с. 145.

ДОДАТОК А

Графічний матеріал

Класифікація методів аналізу часових рядів



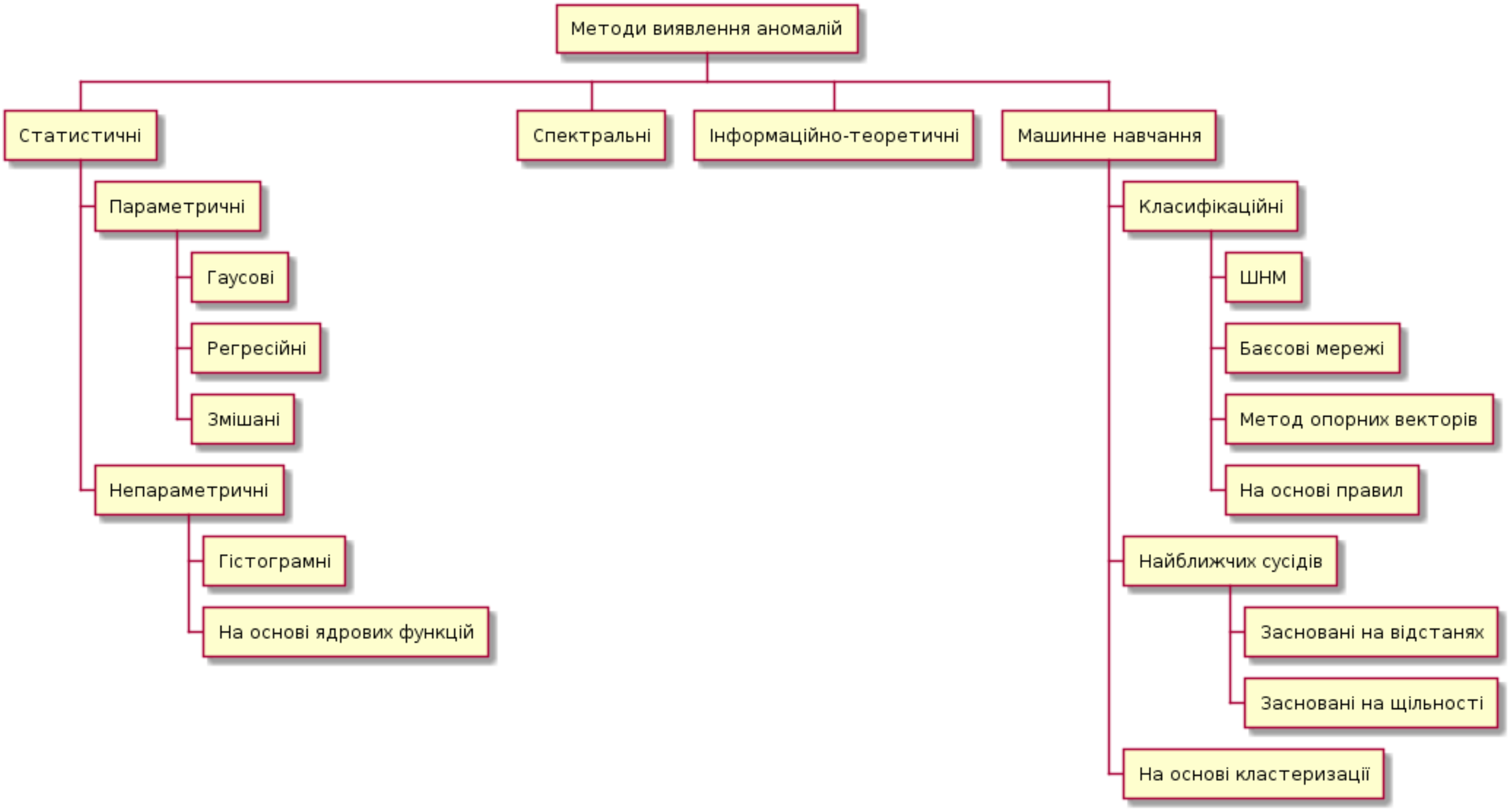
Демонстраційний плакат до магістерської дисертації

на тему «Виявлення аномалій в часових рядах довільної природи»

Виконав студент гр. ІС-82мп Логвинчук А. І.

Керівник Баклан І. В.

Класифікація методів виявлення аномалій



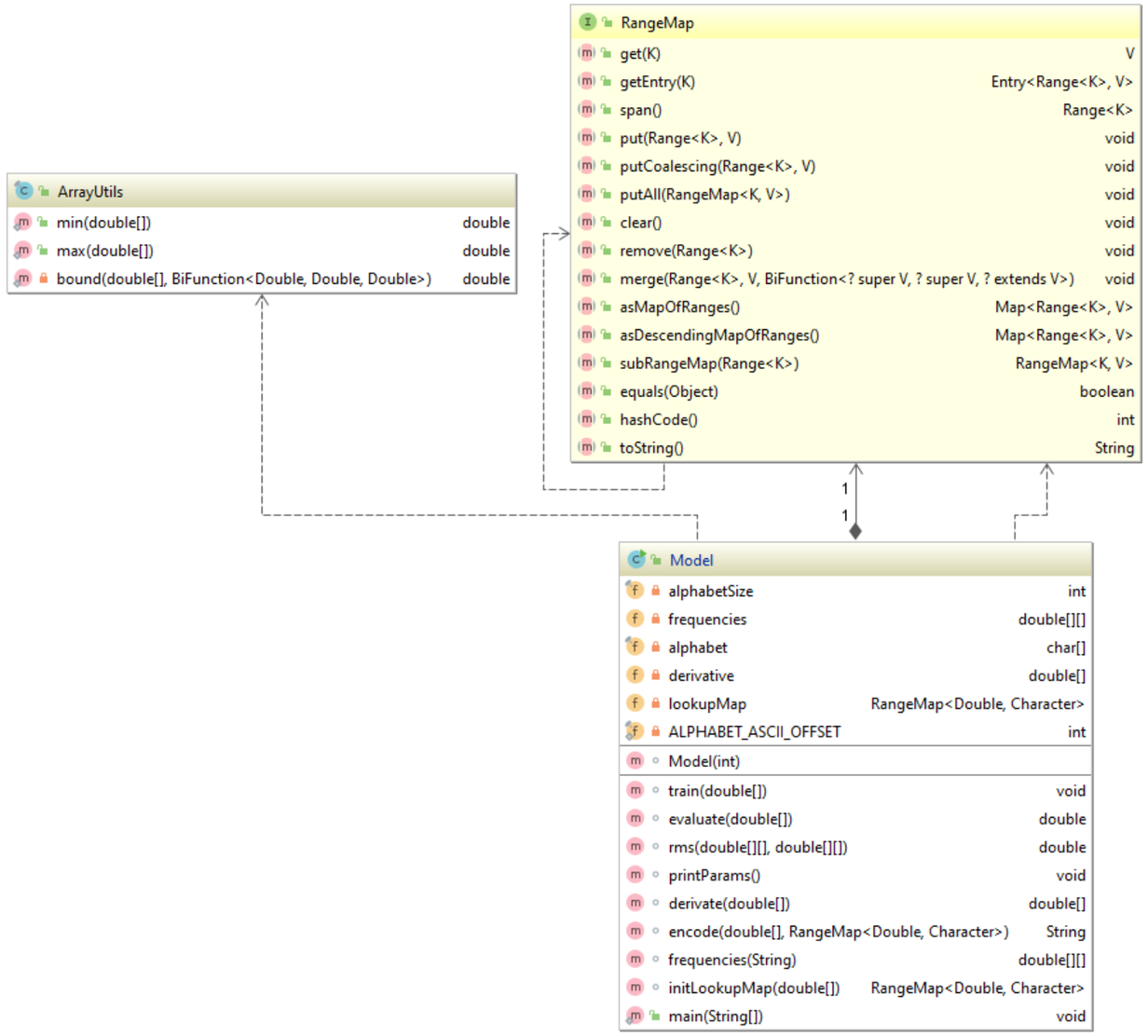
Демонстраційний плакат до магістерської дисертації

на тему «Виявлення аномалій в часових рядах довільної природи»

Виконав студент гр. ІС-82мп Логвинчук А. І.

Керівник Баклан І. В.

UML-діаграма класів



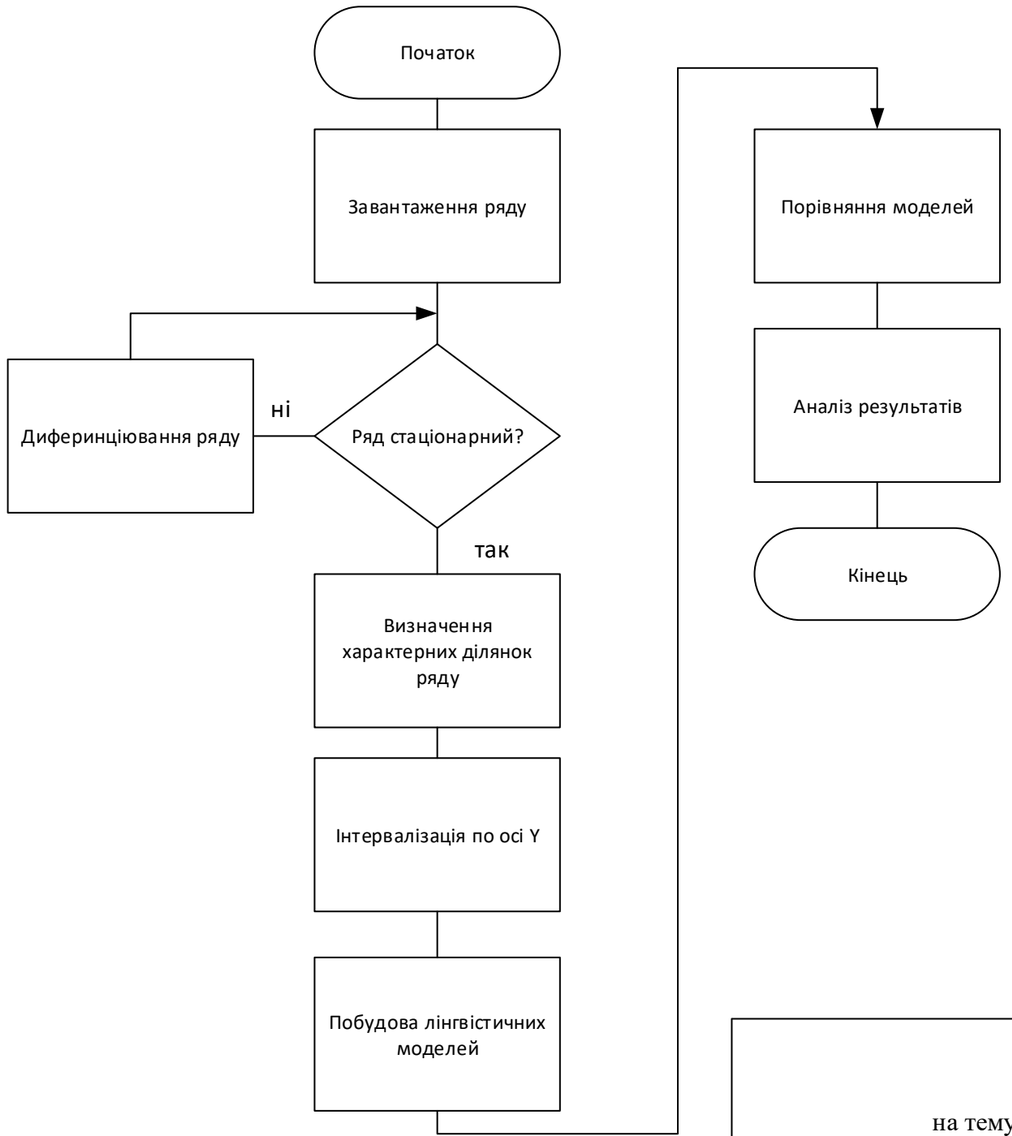
Демонстраційний плакат до магістерської дисертації

на тему «Виявлення аномалій в часових рядах довільної природи»

Виконав студент гр. ІС-82мп Логвинчук А. І.

Керівник Баклан І. В.

Схема експерименту



Демонстраційний плакат до магістерської дисертації

на тему «Виявлення аномалій в часових рядах довільної природи»

Виконав студент гр. ІС-82мп Логвинчук А. І.

Керівник Баклан І. В.

Результати експериментів

Ряд	$\varepsilon_{\text{ет}}$	$\varepsilon_{6,7}$	$\varepsilon_{7,8}$	$\varepsilon_{8,9}$	$\varepsilon_{9,10}$	$\varepsilon_{10,11}$	$\varepsilon_{11,12}$	$\varepsilon_{12,13}$	$\varepsilon_{13,14}$	$\varepsilon_{14,15}$	$\varepsilon_{15,16}$	$\varepsilon_{16,17}$	$\varepsilon_{17,18}$
GOOGL	0.0649	-	-	-	0.4256	0.0633	0.0648	0.0609	**0.0782**	0.0499	**0.0858**	0.0373	0.0474
AAPL	0.0586	0.0431	0.0371	-	0.0409	0.0558	0.0444	-	-	0.0453	**0.0606**	**0.0703**	0.0459
AMZN	0.0703	0.0467	0.0371	0.0409	-	0.0487	0.0459	0.058	0.0583	-	-	**0.0806**	0.0689
IBM	0.0579	-	-	-	0.0562	0.0383	0.0542	0.0568	**0.0670**	**0.0733**	**0.0718**	**0.0589**	**0.0712**
MSFT	0.0796	0.784	-	-	0.0748	0.0703	0.0711	-	0. 0743	0.0693	**0.0941**	**0.0921**	0.0712
WMT	0.0705	-	-	0.0696	0.0358	0.0383	0.0342	-	0.0338	0.0358	0.0521	0.0687	0.0696
NKE	0.0621	-	0.0330	0.0345	0.0366	0.0484	0.0408	0.4256	0.0586	0.0558	-	-	**0.0663**

Демонстраційний плакат до магістерської дисертації

на тему «Виявлення аномалій в часових рядах довільної природи»

Виконав студент гр. ІС-82мп Логвинчук А. І.

Керівник Баклан І. В.