Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет інформаційно-аналітичних технологій та менеджменту
(повна назва)

Кафедра прикладної математики
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Аналіз, комп'ютерне моделювання та стійке оцінювання
статистичних розподілів у потоках даних
(тема)

Виконав:
студент _2_ курсу, групи _САУм-20-1
Белявський Д.О.
(прізвище, ініціали)

Спеціальність
124 Системний аналіз
(код і повна назва спеціальності)
Освітня програма
Системний аналіз і управління
(повна назва освітньої програми)

Керівник доц. Кобзєв В.Г.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри ПМ _____ Тевяшев А.Д.
(підпис) (прізвище, ініціали)

2021 р.

Харківський національний університет радіоелектроніки

Факультет інформаційно-аналітичних технологій та менеджменту
Кафедра _прикладної математики
Рівень вищої освіти _ другий (магістерський)
Спеціальність _124 Системний аналіз
(код і повна назва)
Освітня програма _ Системний аналіз і управління
(повна назва)

ЗАТВЕРДЖУЮ:

Зав. кафедри ПМ _____
(підпис)
“____” _____ 2021 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Студентові Белявському Денису Олександровичу
(прізвище, ім’я, по батькові)

1. Тема роботи Аналіз, комп'ютерне моделювання та стійке оцінювання статистичних розподілів у потоках даних.

затверджена наказом по університету від 05 листопада 2021 р. № 1642 Ст

2. Термін подання студентом роботи до екзаменаційної комісії 10 грудня 2021 р.

3. Вихідні дані до роботи потік даних при невідомих степенях порушень початкових припущень про їх статистичний розподіл

4. Перелік питань, що потрібно опрацювати в роботі
1. Системний аналіз проблеми стійкого оцінювання параметрів статистичного розподілу
2. Вибір і обґрунтування методу розв’язання
3. Програмна реалізація
4. Результати обчислювального експерименту
5. Аналіз можливих застосувань

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (слайдів)

1. Актуальність теми роботи
2. Постановка задачі
3. Системний аналіз проблеми
4. Метод чисельного аналізу
5. Результати обчислювального експерименту

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Термін виконання етапів роботи	Примітка
1	Підбір та вивчення технічної літератури за темою роботи	8 – 14 листопада 2021 р.	виконано
2	Вибір та обґрунтування методу	15 – 21 листопада 2021 р.	виконано
3	Розробка алгоритму і програми	22 – 28 листопада 2021 р.	виконано
4	Проведення аналітичних досліджень та розрахунків	29 листопада – 5 грудня 2021 р.	виконано
5	Робота над текстом пояснювальної записки	6 – 9 грудня 2021 р.	виконано
6	Представлення роботи на рецензію в ЕК	10 грудня 2021 р.	виконано

Дата видачі завдання _8_ листопада 2021 р.

Студент _____
(підпис)

Керівник роботи _____ ___доц. Кобзєв В.Г.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 67 с., 14 табл.,16 рис., 1 дод.,13 джерел.

РОБАСТНІСТЬ, ПАРАМЕТРИ РОЗПОДІЛУ, СТІЙКІ ОЦІНКИ, ПОТОКИ ДАНИХ, М-ОЦІНКИ, L-ОЦІНКИ.

Об’єкт дослідження – потік даних.

Мета роботи – дослідження та аналіз методів робастного оцінювання параметрів розподілів у потоках даних. Основними завданнями досліджень є побудова, дослідження і застосування робастних підходів в потоках даних.

Методи дослідження – для вирішення поставлених завдань використовувався апарат теорії ймовірностей, математичної статистики, математичного аналізу, обчислювальної математики, статистичного моделювання, стійкі оцінки параметрів статистичних розподілів.

Проведено огляд основних підходів теорії робастного оцінювання параметрів до потоків даних. Проведено дослідження стійких методів оцінювання розподілів у потоці даних. Проведено теоретичні дослідження локально-стійких методів оцінювання параметрів в потоці даних з засміченням.

ABSTRACT

Introductory note: 67 pages, 14 tables, 16 figures, 1 appendix, 13 sources.

ROBUSTNESS, DISTRIBUTION PARAMETERS, SUSTAINABLE EVALUATIONS, DATA FLOWS**,** L-ESTIMATE, M-ESTIMATE.

Object of research – data flow.

The purpose of the work – the main purpose of the work is the further development of methods for robust estimation of distribution parameters in data streams. The main objectives of research are the construction, research and application of robust approaches in data flows.

Research methods – to solve the tasks used the apparatus of probability theory, mathematical statistics, mathematical analysis, computational mathematics, statistical modeling, stable estimates of the parameters of statistical distributions.

The main approaches of the theory of robust estimation of parameters to data streams are reviewed. The research of stable methods of estimation of parameters in a data stream is carried out. Theoretical researches of locally stable methods of estimation of parameters in a data stream with clogging are carried out.

ЗМІСТ

ВСТУП

Актуальність теми. У міру того як все більше пристроїв підключається до Інтернету, спостерігається постійне зростання обсягів даних, які збираються, обробляються і аналізуються для більш глибокого розуміння. На відміну від традиційної аналітики, обробка потоку даних дозволяє в реальному часі ідентифікувати системні збої та прогнозувати зміни активів на основі даних що надходять. У таких випадках кожна кінцева точка діє як джерело даних, яке послідовно генерує потоки даних, які потім приймаються та обробляються для створення бізнес-аналітики. Данні використовуються для оптимізації мережі та обслуговування підключених пристроїв в інтернет мережах, транспортних системах, інтелектуальних автомобілях, банківських системах та аналогічних програмах обробки у реальному часі. Концепція потоків даних може бути простежена, до програм, запропонованих Дугласом Макілроєм [1] для зв'язування макросів, які були реалізовані в 1964 як "файли зв'язку" в системі поділу часу Дартмута і інтегровані в Unix. операційну систему 1972 року. Це з'єднання даних між двома процесами, засноване на принципі FIFO. Принцип потоків даних тепер можна знайти у більшості сучасних мов програмування.

Потік даних являє собою послідовність кодованих у цифровому вигляді когерентних сигналів що використовуються для передачі або отримання інформації, яка знаходиться в процесі передачі. Тобто потік даних – це набір інформації, яку було отримано від постачальника даних. Він містить необроблені дані, зібрані з поведінки користувачів у браузері на веб-сайтах, на яких розміщений спеціальне програмне забезпечення. Основними постачальниками потоків даних є компанії, що займаються інформаційними технологіями.

Потік даних майже завжди здійснюється при впливі деяких перешкод, які, незважаючи на прагнення звести їх до мінімуму, вони ніколи не можуть бути повністю усунені. Таким чином, маємо справу не з детермінованими, а з випадковими величинами. У багатьох випадках вимірювані величини є випадковими по своїй природі.

Обробка потоків даних базується на певних припущеннях про моделі їх формування і зміни у часі. Багато уваги цим питанням приділено у роботі [2].

Необхідність застосування апарату математичної статистики при обробці потоку даних, де випадковою складовою не можна знехтувати очевидна а також очевидна і необхідність подальшого розвитку.

Проблемі стійкого оцінювання присвячено багато робіт [2 – 8]. Зараз існує цілий напрямок в теорії оцінювання, який вивчає методи, стійкі до тих чи інших відхилень від модельних припущень. Збірна назва для таких методів і відповідних оцінок параметрів - робастні. Розвиток теорії робастності пов'язано з подальшим удосконаленням застосовуваних статистичних моделей, що описують реальну ситуацію.

У зв'язку з тим, що теорія робастності не завершена і знаходиться в стадії активного розвитку, А. І. Орлов в своїй роботі [3] назвав напрямок, пов'язаний з побудовою робастних процедур статистичного аналізу, однією з головних «Точок зростання» прикладної статистики. Дана робота присвячена питанню розробки і дослідження робастних методів оцінювання параметрів в потоці даних.

Задача стійкого оцінювання статистичних розподілів у потоках даних розглядається як задача стійкого оцінювання основних параметрів цих розподілів.

Мета і завдання кваліфікаційної роботи. Метою кваліфікаційної роботи є дослідження та аналіз методів робастного оцінювання параметрів розподілів у потоках даних. Для досягнення поставленої мети необхідно виконати наступні завдання:

– провести огляд і аналіз сучасного стану задачі «Аналізу, комп'ютерного моделювання та стійкого оцінювання статистичних розподілів у потоках даних»;

– провести стійке оцінювання параметрів статистичного розподілу в потоках даних ;

– провести порівняльний аналіз стійких оцінок.

Об’єктом дослідження є потік даних.

Предметом дослідження є стійке оцінювання статистичних розподілів у потоках даних

Методи дослідження. У кваліфікаційній роботі використовуються апарат теорії ймовірностей, математичної статистики, математичного аналізу, обчислювальної математики, статистичного моделювання, стійкі оцінки параметрів статистичних розподілів.

1 СИСТЕМНИЙ АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ТА ПОСТАНОВКА ЗАДАЧ ДОСЛІДЖЕННЯ

1.1 Системний аналіз проблеми стійкого оцінювання параметрів статистичних розподілів та постановка задач дослідження

1.1.1 Вербальна модель системи

Об’єкт аналізу – потік даних.

Предмет аналізу – ефективність використання стійких оцінок параметрів статистичних розподілів у потоці даних.

Точка зору: дослідник.

Ціль роботи: подальший розвиток дослідження та аналізу методів робастного оцінювання параметрів розподілів у потоці даних. Основними завданнями досліджень є побудова і застосування робастних підходів для потоків даних.

На вхід системи поступає потік даних при невідомих степенях порушень початкових припущень про їх статистичний розподіл. На виході – порівняльний аналіз результатів застосування стійких оцінок до набору вимірювань при різній степені порушень початкових припущень. Механізми системи – дослідник та методи прикладного статистичного аналізу.

1.1.2 Морфологічний опис системи

Морфологічний опис задачі аналізу включає як невід’ємну частину розгляд поняття зовнішнього середовища.

Зовнішнє середовище – сукупність всіх об’єктів за межами границі системи, зміна властивостей яких має вплив на систему, а також тих об’єктів, чиї властивості змінюються в результаті поведінки системи.

Для опису функціонування моделі із зовнішнім середовищем на рисунку

1.1 представлена модель типу «чорний ящик».

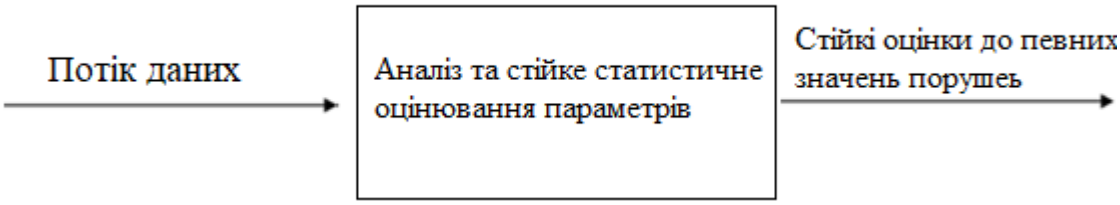

Рисунок 1.1 – Модель системи типу «чорний ящик»

Модель, що представлена як «чорний ящик», включає моделі границі, модель зовнішнього середовища, входи і виходи.

Виходи системи – це цільовий продукт системи. Входи системи – це частіше всього ресурсний вплив зовнішнього середовища. «Чорний ящик» здійснює перетворення входів у виходи системи.

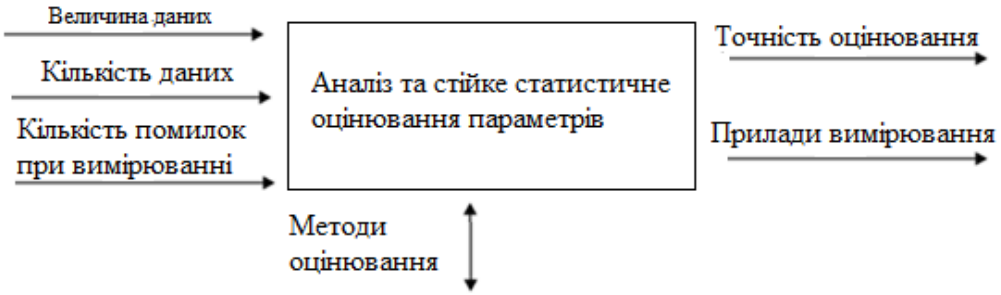

Рисунок 1.2 – Модель системи типу «зовнішнє середовище»

1.1.3 Функціональна модель системи

Для формулювання логіки та взаємодії процесів при вирішені задачі визначення та аналізу параметрів використовується засоби графічної нотації бізнес-процесів IDEF0, що дозволяє наочно побачити ієрархію основних елементів в побудові системи вирішення задачі. IDEF0 може використовуватися для моделювання широкого спектру автоматизованих та неавтоматичних систем. Для нових систем його можна використовувати спочатку для визначення вимог та

специфікації функцій, а потім для розробки реалізації, що відповідає вимогам та виконує функції. Для існуючих систем IDEF0 можна використовувати для аналізу функцій, які виконує система, та для запису механізмів (засобів), за допомогою яких вони виконуються. Результатом застосування IDEF0 до системи є модель, що складається з ієрархічної серії діаграм, тексту пов'язаних один з одним. Двома основними компонентами моделювання є функції та об'єкти, які пов'язують ці функції (представлені стрілками).

Стандарт IDEF0 представляє організацію як набір модулів, тут існує правило – найбільш важлива функція знаходиться у верхньому лівому кутку, крім того, існують правила сторін:

– стрілка входу завжди приходить в ліву кромку активності;

– стрілка управління – в верхню кромку;

– стрілка механізму – нижня кромка;

– стрілка виходу – права кромка.

Верхнім рівнем моделювання системи в IDEF0 є рівень визначення контексту, тобто найбільш абстрактного рівня опису системи в цілому. В контекст входить визначення суб’єкта моделювання, цілі і точка зору на модель. Для даної задачі:

– точка зору – дослідник;

– суб’єкт – спотворенні дані;

– ціль – порівняльний аналіз результатів застосування стійких оцінок до набору значень вимірювань при порушеннях попередніх припущень про їх статистичний розподіл.

Контекстна діаграма IDEF0 зображує функціонування системи в цілому (рис. 1.3). В результаті виконання роботи очікується отримати порівняльний аналіз результатів застосування стійких оцінок до набору вимірювань при різній степені порушень початкових припущень.

Для деталізації контекстної діаграми виконується декомпозиція системи. Декомпозиція роботи зображена на рисунку 1.3.

Процес розділено на три задачі:

– збір даних;

– аналіз та стійке статистичне оцінювання;

– представлення результату.

Рисунок 1.3 – Контекстна діаграма IDEF0 (рівень А – 0) задачі аналізу та стійкого оцінювання

Після декомпозиції контекстної діаграми проводиться декомпозиція кожного великого фрагмента системи на більш дрібні і так далі, до досягнення потрібного рівня подробиці опису.

IDEF3 є стандартом системного аналізу при документуванні технологічних процесів та операцій, що регламентуються на підприємстві. Він надає засоби наочного дослідження і моделювання сценаріїв. IDEF3 широко застосовується при розробці інформаційних систем. Для створення комп’ютерного графічного зображення використовується інструмент BpWin візуального моделювання бізнеспроцесів. Система описується як упорядкована послідовність подій з одночасним описом об'єктів, що мають відношення до процесу, що моделюється. Тобто IDEF3 спосіб опису процесів з використанням структурованого методу, що дозволяє експерту в предметній області уявити стан речей як упорядковану послідовність подій з одночасним описом об'єктів, що мають безпосереднє відношення до процесу. IDEF3 є технологією, що добре пристосована для збору даних, потрібних для проведення структурного аналізу системи. На відміну від більшості технологій моделювання процесів, IDEF3 не має жорст-

ких синтаксичних або семантичних обмежень, які роблять незручним опис неповних або нецілісних систем. Крім того, автор моделі (системний аналітик) позбавлений необхідності змішувати свої власні припущення про функціонування системи з експертними твердженнями з метою заповнення прогалин в описі предметної області. IDEF3 також можна використовувати як метод проектування процесів. IDEF3моделювання органічно доповнює традиційне моделювання з використанням стандарту методології IDEF0. В даний час воно набуває все більшого поширення як цілком життєздатний шлях побудови моделей проектованих систем для подальшого аналізу імітаційними методами. Імітаційне тестування часто використовують для оцінки експлуатаційних якостей системи, що розробляється.

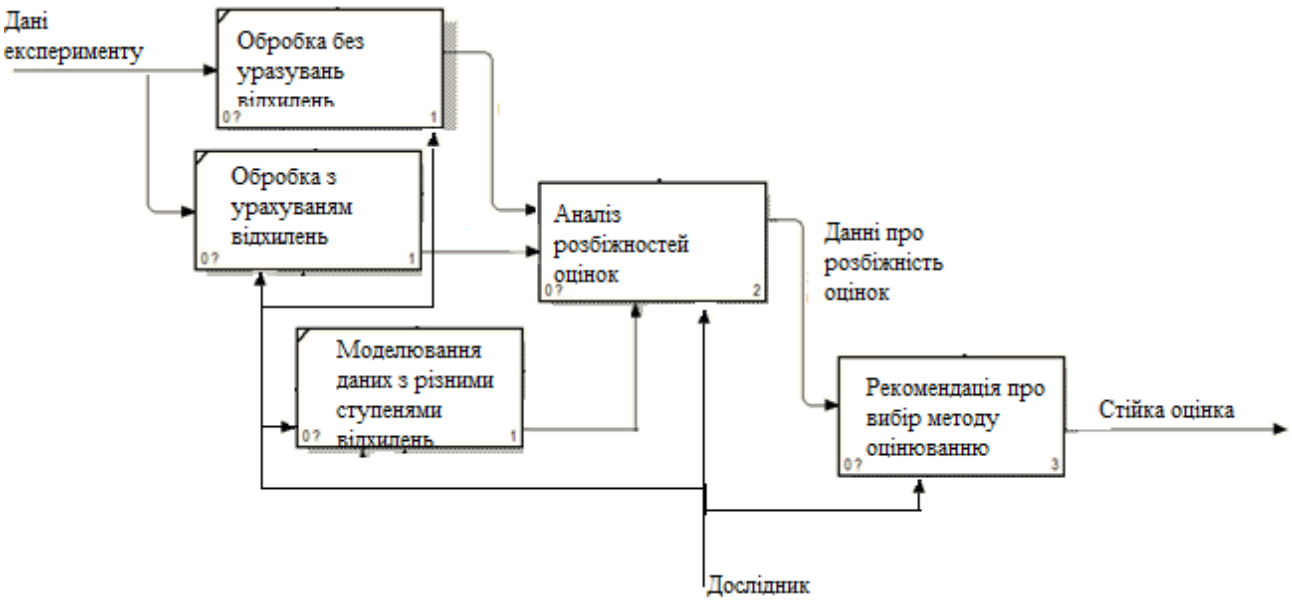

Рисунок 1.4 – Діаграма декомпозиції (рівень А0) задачі аналізу та стійкого оцінювання

В аналізі стійких оцінок проводитися порівняльний аналіз результатів застосування стійких оцінок до потоку даних при різних степенях порушень попередніх припущень про їх статистичний розподіл, шляхом порівняння їх відхилень відносно вибірки, що не містить відхилень.

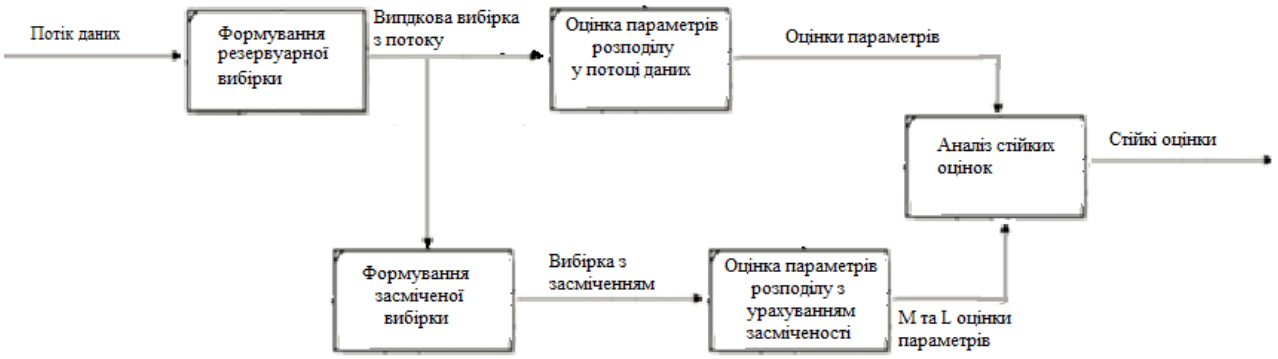

Рисунок 1.5 – Декомпозиція оцінювання параметрі за допомогою M і L оцінок (рівень A1)

1.1.4 Інформаційна модель

Інформаційні моделі відображають різні типи систем об'єктів, в яких реалізуються різні структури взаємодії і взаємозв'язку між елементами системи.

Діаграма потоку даних (DFD) є способом представлення розробника системи потоків даних в системі. Кожна діаграма є елементом ієрархії. Вона підлягає уточненню шляхом деталізації процесів та потоків даних. Удосконалене подання процесу може бути виконано на іншій схемі потоку даних, яка підрозділяє цей процес на під процеси. Засоби методології DFD дозволяють відображати джерела і призначення даних, описати процеси і групи даних. Діаграма потоку даних зв'язують одну функцію з іншого, і ефективно використовуються для опису процесів при впровадженні процесного підходу до створення систем. Нотація DFD дозволяє описувати потоки не тільки інформаційні потоки, але і матеріальні – потоки документів і ресурсів.

Діаграма дерева вузлів показує ієрархію робіт в моделі і дозволяє розглянути всю модель цілком. Діаграма дерева вузлів для задачі, що розглядається показана на рисунку 1.6.

Рисунок 1.6 – Діаграма дерева вузлів

1.2 Аналіз сценаріїв вирішення проблеми стійкого оцінювання параметрів статистичних розподілів в потоках даних

1.2.1 Модель аналізу проблеми

Для вирішення задачі оцінювання параметрів повинна бути визначена математична модель с точки зору застосування різних алгоритмів її використання з подальшим їх порівнянням з точки зору точності висновків та особливостей програмної реалізації. Результатом стане визначення найбільш відповідного алгоритму, який може бути застосований для вирішення даної задачі.

В якості критеріїв, які можуть бути застосовані для оцінки якості очікуваного результату, можна вказати наступні:

– ресурсоємність (К1);

– величина похибок при стійкому оцінюванні (К2);

– час роботи прогарами для певного потоку даних(К3);

– складність застосування (К4).

Розглянемо ці критерії та виділимо частини, що будуть порівнюватися у запропонованих альтернатив.

Порівнюючи ресурсоємність альтернатив, маємо на увазі ресурси ЕОМ, які будуть необхідні для успішного закінчення роботи алгоритмів без технічних помилок: об’єм оперативної пам’яті, машинний час виконання розрахунків.

Порівнюючи альтернативи на величину похибки, перевага буде віддана тому алгоритму, який має найменшу похибку при підрахунку.

При порівнянні складності застосування алгоритмів, перевага буде віддана найпростішому способу оцінювання параметрів.

При порівнянні альтернативи на час роботи, перевага віддаватиметься алгоритму, який матиме мінімальний сумарний час роботи.

Будуть розглянуті такі альтернативи:

– метод максимальної правдоподібності (А1);

– L-оцінки (А2);

– мінімаксні оцінки (А3).

Оцінка максимальної правдоподібності є популярним статистичним методом, який використовується для створення статистичної моделі на основі даних і забезпечення оцінки параметрів моделі. Метод максимальної правдоподібності відповідає багатьом відомим методам оцінки в області статистики. Для фіксованого набору даних і базової ймовірнісної моделі, використовуючи метод максимальної правдоподібності, ми отримаємо значення параметрів моделі, які роблять дані «ближчими» до реальних. Оцінка максимальної правдоподібності дає унікальний і простий спосіб визначити рішення в разі нормального розподілу.

Мінімаксні оцінки, як правило, узгоджуються зі здоровим глуздом. Хоча мінімаксна оцінка може виявитися гіршою, ніж інші оцінки у великій області параметричного простору, гідністю її є те, що верхня величина похибки для неї явно не гірша, ніж будь-яка інша оцінка. Оскільки мінімаксна оцінка дерева може виявитися досить тривалою процедурою, було витрачено багато зусиль для підвищення її ефективності, що увінчалися відомим успіхом.

L-оцінка параметрів зсуву і масштабу, обчислення яких базується на значеннях вибіркових квантилів. Найскладнішою операцією при обчисленні таких оцінок є сортування наявної вибірки по зростанню з метою визначення вибіркових квантилів спостережуваного закону. Як і оцінки максимальної правдоподібності по групованим спостереженнями, дані оцінки є робастними. Робастність цих оцінок підтверджує вид функції впливу Хампеля, яка для L-оцінок є ступінчастою, обмеженою по абсолютній величині. Такий вид функцій впливу говорить про те, що присутність у вибірці аномальних спостережень не матиме приводити до різкої зміни L-оцінок. Коефіцієнти в лінійної комбінації L-оцінок визначаються тим, які і скільки квантилів використовуються при побудові оцінок. Від вибору квантилів при побудові оцінок залежать і асимптотичні властивості одержуваних L-оцінок.

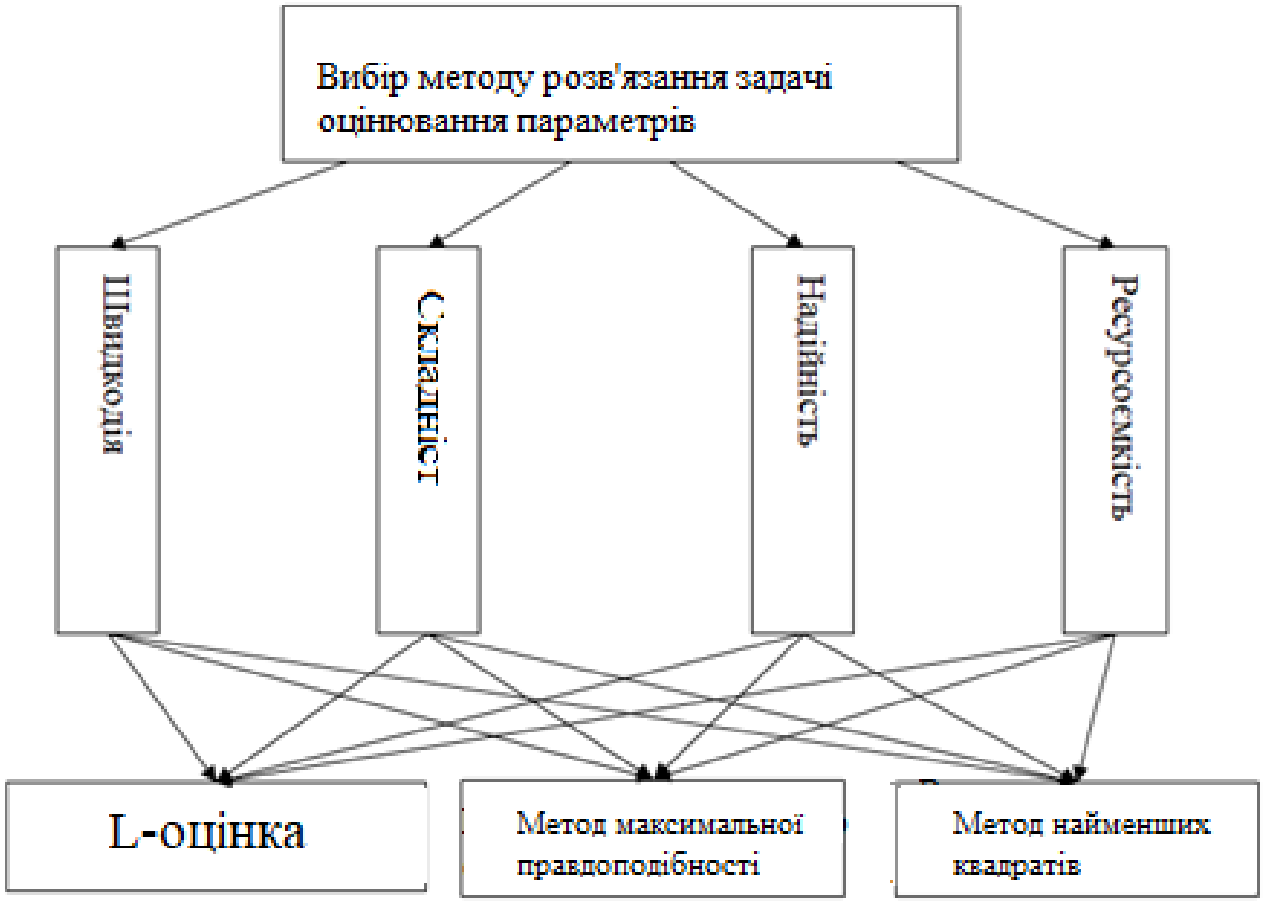

Рисунок 1.7 – Ієрархічна структура задачі вибору методу

1.2.2 Оцінювання вектору пріоритетів незадоволеностей методом аналізу ієрархій

Необхідно сформувати матрицю попарних порівнянь для елементів першого рівня ієрархії (табл. 1.1, 1.2), тобто матрицю попарних порівнянь важливості критеріїв. Шляхом порівняння методів оцінювання параметрів отримано інформацію про важливість критеріїв і складена матриця попарних порівнянь критеріїв.

Таблиця 1.1 – Матриця попарних порівнянь для елементів першого рівня ієрархії

	$К_1$	$К_2$	$К_3$	$К_4$
$К_1$	1	4	6	8
$К_2$	1/4	1	8	9
$К_3$	1/6	1/8	1	3
$К_4$	1/8	1/9	1/3	1

Таблиця 1.2 – Вектор локальних пріоритетів критеріїв

	$К_1$	$К_2$	$К_3$	$К_4$	Вектор пріоритетів
$К_1$	1	4	6	8	0,489
$К_2$	1/4	1	8	9	0,368
$К_3$	1/6	1/8	1	3	0,076
$К_4$	1/8	1/9	1/3	1	0,067

Щоб знайти необхідний індексу узгодженості треба знайти суми усіх елементів матриці по стовпцях

$$y_1 = 1,541;\ y_2 = 5,236;\ y_3 = 15,5;\ y_4 = 21,0.$$

Тоді знайдемо

$$\lambda_{\max} \approx 1,541 \cdot 0,489 + 5,236 \cdot 0,368 + 15,5 \cdot 0,076 + 20 \cdot 0,067 = 5,136;$$

де і індекс узгодженості

$$CI^K = \frac{5,136 - 4}{4 - 1} = 0,399.$$

Така як матриця попарних порівнянь критеріїв це матриця четвертого порядку, тоді $RI^K = 0.9$, і

$$CR^K = \frac{CI^K}{RI^K} = \frac{0,399}{0,9} = 0,443.$$

Так як співвідношення узгодженості доволі близько до 0,1, то можна сказати що матриця попарних порівнянь критеріїв побудована вірно і не має помилок. Побудуємо матриці попарних порівнянь альтернатив К (табл. 1.3–1.6) Виходячи з цього вектор локальних пріоритетів критеріїв щодо проблеми вибору буде дорівнювати:

$$\vec{p}^K = \left(0,0489;\ 0,368;\ 0,076;\ 0,067\right)^T.$$

Таблиця 1.3 – Матриця попарних порівнянь альтернатив $К_1$

$К_1$	$А_1$	$А_2$	$А_3$
$А_1$	1	3	8
$А_2$	1/3	1	6
$А_3$	1/8	1/6	1

Таблиця 1.4 – Матриця попарних порівнянь альтернатив $К_2$

$К_2$	$А_1$	$А_2$	$А_3$
$А_1$	1	4	7
$А_2$	1/4	1	5
$А_3$	1/7	1/5	1

Таблиця 1.5 – Матриця попарних порівнянь альтернатив $К_3$

К3	А1	А2	А3
А1	1	3	9
А2	1/3	1	7
А3	1/9	1/7	1

Таблиця 1.6 – Матриця попарних порівнянь альтернатив $К_4$

$К_4$	$А_1$	$А_2$	$А_3$
$А_1$	1	4	6
$А_2$	1/4	1	8
$А_3$	1/6	1/8	1

Необхідно знайти для кожної матриці вектори локальних пріоритетів що дорівнюють:

$$\vec{p}_1^{\,A} = \begin{pmatrix} 0,491 \\ 0,433 \\ 0,075 \end{pmatrix},\ \vec{p}_2^{\,A} = \begin{pmatrix} 0,544 \\ 0,370 \\ 0,085 \end{pmatrix},\ \vec{p}_3^{\,A} = \begin{pmatrix} 0,494 \\ 0,440 \\ 0,064 \end{pmatrix},\ \vec{p}_4^{\,A} = \begin{pmatrix} 0,555 \\ 0,389 \\ 0,054 \end{pmatrix}.$$

Через те що матриця попарних порівнянь альтернатив – це матриці тре-

тього порядку, то $RI^A = 0.9$. Індекси узгодженості і відносин узгодженості для матриць парних порівнянь альтернатив за кожним критерієм будуть дорівнювати:

$$CI^A_{K1} = 0,0081257,\ CI^A_{K2} = 0,02573241;$$

$$CI^A_{K3} = 0,0085412, CI^A_{K4} = 0,04457;$$

$$CR^A_{K1} = 0,021415,\ CR^A_{K2} = 0,046457,;$$

$$CR^A_{K3} = 0,0114536,\ CR^A_{K4} = 0,06767.$$

Також необхідно розрахувати вектор глобальних пріоритетів альтернатив і вектор глобальних пріоритетів дорівнює

Для цього необхідно скласти матрицю: P^A

$$P^A = \begin{pmatrix} 0,491 & 0,544 & 0,494 & 0,555 \\ 0,433 & 0,370 & 0,440 & 0,389 \\ 0,075 & 0,085 & 0,064 & 0,054 \end{pmatrix};$$

$$\vec{p} = \begin{pmatrix} 0,491 & 0,544 & 0,494 & 0,555 \\ 0,433 & 0,370 & 0,440 & 0,389 \\ 0,075 & 0,085 & 0,064 & 0,054 \end{pmatrix} \cdot \begin{pmatrix} 0,626 \\ 0,242 \\ 0,080 \\ 0,050 \end{pmatrix} = \begin{pmatrix} 0,413 \\ 0,410 \\ 0,102 \end{pmatrix}.$$

Тоді індекс узгодженості і відношення узгодженості для всієї ієрархії буде дорівнювати:

$$CI = CI^K + \vec{p}^K,\ \overrightarrow{CI^A} = 0,13454,$$

$$RI = RI^K + RI^A = 1,554,$$

$$CR = \frac{CI}{RI} = 0,097878.$$

Максимальна компонента вектору глобальних пріоритетів відповідає першій альтернативі, методу максимальної правдоподібності але і друга альтернатива (L-оцінки) має гарні показники, тому проведемо їх порівняльний аналіз.

1.3 Змістовна та формальна постановка задачі

1.3.1 Змістовна постановка задачі

Системи потоків даних в з'явилися давно, і протягом багатьох років залишалися вотчиною апаратних систем. У таких системах часто невиконання певних норм загрожує життю людині. Але за останні десять років з'явилися і стали швидко поширюються системи потокової обробки. Приклади потокової обробки зустрічаються всюди: соціальні мережі, ігри, розумні міста, інтелектуальні вимірювальні пристрої, ваша нова пральна машина - список можна продовжувати довго. Найпростішим прикладом може слугувати потік даних в операційних системах

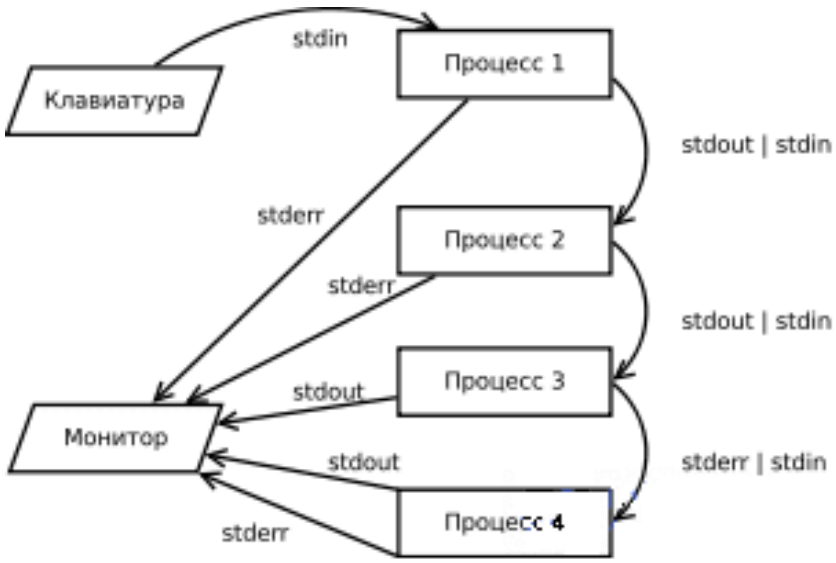

Рисунок 1.8 – Приклад ланцюга процесів, що спілкуються за допомогою потоків даних.

Коли навіть від найменшої частки даних може залежати чиєсь життя необхідно правильно аналізувати та оцінювати параметри статистичного розподілу в потоках даних і відокремлювати дійсну інформацію від похибки при вимірюванні або передачі. Для цього можна використовувати так звану надійну статистику.

Існують різні визначення "надійної статистики". Строго кажучи, надійна статистика є стійкою до помилок в результатах, отриманих відхиленнями від припущень (наприклад, нормальності). Це означає, що якщо припущення будуть виконані лише приблизно, надійний оцінювач все одно матиме розумну ефективність та досить низьке зміщення, а також буде асимптотично незміщеним, тобто має зміщення, яке має тенденцію до 0, оскільки розмір вибірки має тенденцію до нескінченності.

Одним з найважливіших випадків є надійність розподілу. Класичні статистичні процедури, як правило, чутливі до "тривалості" (наприклад, коли розподіл даних має довші хвости, ніж передбачається нормальний розподіл). Це означає, що вони будуть сильно залежати від наявності викидів в даних, і вироблені ними оцінки можуть бути сильно спотворені, якщо в даних є екстремальні викиди, в порівнянні з тим, чим вони були б, якби викиди не були включені до дані.

Навпаки, більш надійні оцінки, які не так чутливі до спотворень розподілу також стійкі до присутності викидів. Таким чином, в контексті надійної статистики розподілений надійні та стійкі до викидів є фактично синонімами.

При розгляді того, наскільки стійка оцінка до наявності викидів, корисно перевірити, що відбувається, коли екстремальний викид додається в набір даних, і перевірити, що відбувається, коли екстремальний викид замінює один з існуючих точок даних, а потім розглянути ефект багаторазового додавання або заміни.

На змістовному рівні задачу дослідження можна сформулювати наступнім чином: заданий певний потік даних необхідно провести аналіз та стійке оцінювання параметрів статистичного розподілу в цьому потоці. Застосувати та

визначити найбільш стійкі оцінки до наявності викидів в потоці даних коли екстремальний викид додається в набір інформації.

1.3.2 Формальна постановка задачі

Нехай задано патерн односторонньої взаємодії. Цей патерн найчастіше зустрічається в ситуаціях, коли система, що відправляє запит, не потребує відповіді. Іноді його ще називають «Вистрілив і забув». У деяких випадках цей патерн має очевидні переваги і, можливо, є єдиним способом взаємодії між клієнтом і службою. Він схожий на патерни запит-відповіді запит-підтвердження в тому сенсі, що повідомлення передається від клієнта службі. А основна відмінність полягає в тому, що служба не відправляє ніякої відповіді. В інших патернах клієнт знає, що запит був отриманий і якось оброблений, а тут невідомо навіть, чи дійшов запит до служби.

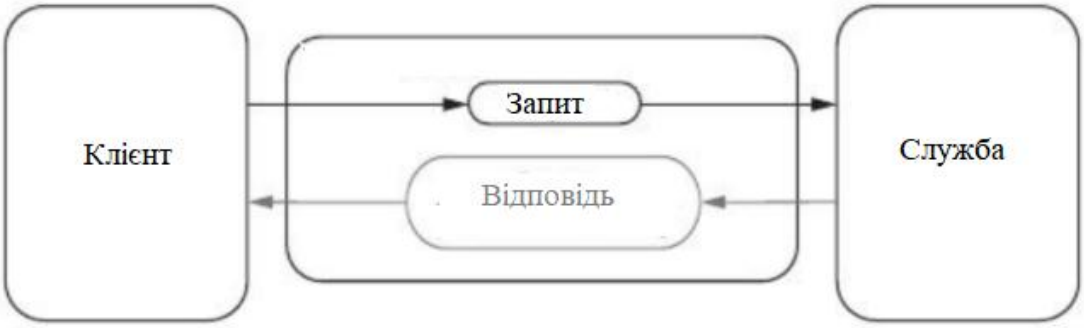

Рисунок 1.9 – Патерн односторонньої взаємодії

Вихідними даними (даними що надсилає нам клієнт) є спостереження випадкової величини. Нехай $x_1,...,x_m$ – спостереження випадкової величини ξ, розподіленої з щільністю $f(x,\theta)$, де $x \in X \subseteq R$ і невідомий параметр $\theta \in \Theta \subseteq R$. Визначимо оцінку параметра в цій моделі як

$$\hat{\theta} = \arg\min_{\theta} \sum_{i=1}^{m} M(x_i, \theta), \tag{1.1}$$

де $M(x,\theta): X \times \Theta \to R$ – неперервна, диференційована майже всюди функція, звана функцією втрат. Диференціюючи суму в вираженні (1.1) по параметру і прирівнюючи похідну нулю, отримуємо оцінне рівняння

$$\sum_{i=1}^{m} \psi(x_i, \hat{\theta}) = 0; \tag{1.2}$$

розв'язок якого використовується для знаходження оцінки (2.12). Функція ψ в рівнянні (1.2)

$$\psi(x_i, \theta) = \frac{\partial}{\partial \theta} M(x, \theta); \tag{1.3}$$

називається оціночної функцією (вона визначена з точністю до параметра θ).

Наприклад, оцінка максимальної правдоподібності (ОМП) [5, 6] визначається як результат максимізації функції правдоподібності

$$L(\theta) = \prod_{i=1}^{m} f(x_i, \theta); \tag{1.4}$$

На підставі (1.4) можна записати

$$\hat{\theta}_{ОМП} = \arg\max_{\theta} L(\theta) = \arg\min_{\theta}[-\ln L(\theta)] = \arg\min_{\theta} \sum_{i=1}^{m} [-\ln f(x_i, \theta)]. \tag{1.5}$$

Порівнюючи (1.1) і (1.5), отримуємо вираз для функції втрат ОМП

$$M_{ОМП}(x,\theta) = -\ln f(x,\theta); \qquad (1.6)$$

звідси, з точністю до $c = c(\theta)$ яка не залежить від x величини,

$$\psi_{ОМП}(x_i,\theta) = \frac{\partial}{\partial\theta}\ln f(x,\theta); \qquad (1.7)$$

Оцінка, яка визначається виразами (1.1) або (1.2), була названа Хьюбером *M*-оцінкою, була доведена спроможність і асимптотична нормальність при (не строго) опуклій функції втрат [4]. Крім *M*-оцінок виділені *L*-оцінки, які є лінійними комбінаціями порядкових статистик, і R-оцінки, одержувані в рангових умовах. Жакель [5] довів асимптотичну еквівалентність цих трьох типів оцінок, з яких *M*-оцінки найбільш зручні для аналізу. Пфанзагль [6] назвав їх оцінками мінімуму контрасту.

Необхідно проаналізувати однорідність потоку даних, обрати та застосувати оцінки параметрів статистичного розподілу, стійкі до можливих відхилень від класичних припущень про його вигляд, а також встановити вплив таких відхилень на потік даних.

1.4 Постановка задач дослідження

Ми живемо в світі, який все сильніше орієнтований на сьогодення: це і соціальні мережі, і магазини, що відстежують переміщення покупців по проходах, і датчики, що реагують на зміни в навколишньому середовищі. Необхідність застосування стійких оцінок для потоку даних різко підвищилася.

Класичні методи статистики в більшості своїй мають підвищену чутливість до вихідних передумов статистичної моделі, прийнятої при обробці даних експерименту. При вирішенні прикладних задач неминуче виникають відхилення від вихідних передумов моделі, і застосування стандартних методів в цих

умовах може виявитися мало ефективним і часто призводить до істотних викривлень статистичних висновків. У зв'язку з цим виникає необхідність побудови нових, нетрадиційних методів обробки інформації, стійких (або робастних) до можливих відхилень характеристик реальних даних від передбачуваних.

Базуючись на системному аналізі задачі та вхідних даних до роботи можна сформулювати задачу дослідження:

– дослідити моделі потоку даних, які використовуються в сучасному світі;

– дослідити методи оцінювання параметрів до потоку даних, стійкі до найбільш частих видів відхилень;

– розробити програмний продукт, що дозволить автоматизувати процес моделювання та вирішити описану прикладну задачу.

2 ВИБІР ТА ОБҐРУНТУВАННЯ МЕТОДУ РОЗВ’ЯЗАННЯ

2.1 Потік даних

Потік даних – це набір інформації, отриманої від постачальника даних. Він містить необроблені дані, які були зібрані в результаті обробки певних спостережень. Основними постачальниками потоків даних є компанії, що займаються інформаційними технологіями. Наприклад такі компанії як Google, Яндекс, SpaceX, Microsoft і т.д.

Візьмемо за потік даних зміни курсу гривні до євро за весь час до сьогоднішнього дня. Ця інформація є загально доступною і її можна знайти за посиланням [6]. Нам небхідно знадобився статистичний аналіз даних в режимі реального часу. Дані надходять швидко, і їх надходження ніколи не припиняється, отже і всі дані в пам'яті компютера не можуть розміститися. Можливе рішення - зробити випадкову вибірку потоку даних, що надходять.

Традиційно організації збирають дані, зберігають їх у сховищах даних та обробляють їх партіями. Це заощаджує мізерні обчислювальні потужності. За останні роки структура даних та технології обробки сильно змінилися. IoT (Інтернет речей від англ. internet of things) представив широкий спектр датчиків, що генерують потокові дані. Кредитні картки та фінансові операції в Інтернеті також генерують дані у реальному часі, які необхідно аналізувати та перевіряти. Веб-браузери створюють онлайн-транзакції та журнали активності. Потокова передача даних та потокова обробка невід’ємні частини цієї структури. Без них не можливо ані передати дані ані перевірити їх на надійність.

2.2 Стійкі оцінки в потоці даних

Відразу провести стійку оцінку всього потоку даних неможливо через постійне додавання інформації до нього. Часто потрібно зробити випадкову вибі-

рку з потоку для того щоби провести його аналіз. Припустимо, що ми розробляємо юридичну мережу, і наші сервери кожну хвилину отримують мільйони запитів на показ даних. Це чудово, але далі - більше, і нам знадобиться статистичний аналіз показу даних в режимі реального часу. На перший погляд, нічого складного, але при найближчому розгляді ми розуміємо, що дані надходять швидко, надходження ніколи не припиняється, і всі дані в пам'яті комп’ютера не поміщаються. Можливе рішення - зробити вибірку потоку даних, що надходять.

Типовий підхід до вирішення проблеми - резервуарна вибірка [7]. Ідея полягає в тому, щоб зберігати заздалегідь визначене число значень з потоку (резервуар) і при надходженні нового приймати ймовірнісне рішення: помістити його в резервуар або використовувати в якості випадкового прикладу. Саме по вибірці буде проводитися аналіз та стійке оцінювання параметрів у потоці даних, кожний раз після заміни елемента буферної вибірки на елемент потоку. Аналіз та стійке оцінювання проводилося за допомогою методу максимальної правдоподібності та L-оцінок.

2.3 Метод максимальної правдоподібності

Найбільш поширеним методом точкових оцінок параметрів є метод максимальної правдоподібності. Цей метод вперше був запропонований Р. Фішером.

Нехай є вибірка $x_1, x_2, ..., x_n$ з генеральної сукупності з невідомою теоретичною функцією розподілу $F_x(x)$, що належить відомому однопараметричному сімейству $F_x(x,\theta)$. Функція невідомого параметра θ

$$L(x_1, x_2, ..., x_n, \theta) = f(x_1, \theta) f(x_2, \theta) .... f(x_n, \theta)$$

називається функцією правдоподібності. Тут $f(x,\theta)$ – щільність розподілу випад-

кової величини X при неперервному розподілі, а в разі дискретного розподілу $f(x,\theta) = P\{X = x;\theta\}$. Чудова властивість функцій правдоподібності полягає в тому що вони як би вбирають в себе всю інформацію, яка дається вибіркою щодо параметра θ. Функція правдоподібності по суті не що інше, як ймовірність (в безперервному випадку щільність розподілу) отримати саме ту вибірку $x_1, x_2, ..., x_n$, яку б ми реально мали, якби значення невідомого параметра дорівнювало θ. Природно, тому в якості оцінки невідомого параметра θ вибрати θ^*, що доставляє найбільше значення функції правдоподібності $L(x_1, x_2, ...x_n, \theta)$. Оцінкою максимальної правдоподібності називається таке значення θ^*, для якого

$$L(x_1, x_2, ..., x_n, \theta) = \max L(x_1, x_2, ..., x_n, \theta). \qquad (2.1)$$

На практиці [1] використовують не саму функцію правдоподібності, а її логарифм $\ln L(x_1, x_2, ..., x_n, \theta)$.

Використовуючи необхідну і достатню умову екстремуму функції, оцінка максимальної правдоподібності θ^* може бути знайдена наступними діями:

а) знайти похідну $\frac{\partial}{\partial\theta}\ln L(x_1, x_2, ..., x_n, \theta)$, прирівняти її до нуля і знайти корінь θ^* рівняння правдоподібності;

б) знайти другу похідну $\frac{\partial^2}{\partial\theta^2}\ln L(x_1, x_2, ..., x_n, \theta)$ і, якщо при $\theta = \theta^*$ друга похідна від'ємна, то θ^* – оцінка максимальної правдоподібності невідомого параметра θ.

Для використання методу максимальної правдоподібності необхідно, згідно з [1], щоб функція правдоподібності була диференційованою. Оцінку θ^* слід шукати серед значень θ, що задовольняють рівнянню правдоподібності або належать границі області допустимих значень θ. Для найбільш важливих, з практичної точки зору, сімейств $F_x(x;\theta)$ рівняння правдоподібності має єдине рішення θ^*. Це вирішення і є оцінкою максимальної правдоподібності.

Метод максимальної правдоподібності до цього моменту був викладений для випадку оцінки одного параметра θ [10]. Природно, що все вищесказане поширюється і на випадок оцінки *k* невідомих параметрів $\theta_1, \theta_2, ..., \theta_n$.

Переваги методу максимальної правдоподібності:

– для випадку оцінки одного параметра оцінка максимальної правдоподібності θ^* завжди буде спроможною;

– при великих обсягах вибірки *n* розподіл оцінки максимальної правдоподібності θ^* можна наближено вважати нормальним із середнім θ і дисперсією $\frac{1}{nI(\theta)}$, де $I(\theta)$ – інформація Фішера. Оцінка θ^* буде асимптотично ефективною в тому сенсі, що не існує іншої асимптотично нормальної оцінки, що має меншу дисперсію.

Якщо існує ефективна оцінка невідомого параметра $\overline{\theta^E}$, то вона є оцінкою максимальної правдоподібності θ^*.

У загальному випадку оцінка максимальної правдоподібності може бути не тільки неефективною, але і зміщеною [1]. Однак ця зміщеність не має істотного значення і може бути виправлена, наприклад, помноженням на відповідний множник.

Так як заздалегідь в нашому випадку неможливо визначити які данні будуть надходити, припустимо що щільність вибірки має нормальний розподіл. Нормальний розподіл випадкової величини визначається як

$$f(x;\mu;\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x_i-\mu)^2}{2\sigma^2}\right\};$$

має два параметра, середнє μ і дисперсію σ, які слід оцінити.

Функція правдоподібності має вид

$$L(x;\mu;\sigma^2)=\frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}.$$

Після логарифмування отримаємо:

$$\ln L=-\frac{n}{2}(\ln\sigma^2+\ln(2\pi))-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\ .$$

Щоб знайти параметри μ і σ часткові похідні за цими параметрами необхідно прирівняти нулю і розв'язати відповідну систему рівнянь [5]:

$$\begin{cases}\dfrac{\partial\ln L}{\partial\mu}=\dfrac{1}{\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu)=0;\\ \dfrac{\partial\ln L}{\partial\sigma^2}=\dfrac{1}{2\sigma^2}\displaystyle\sum_{i=1}^{n}(x_i-\mu)^2-\dfrac{n}{2\sigma^2}=0.\end{cases}$$

З першого рівняння отримаємо:

$$\sum_{i=1}^{n}(x_i-\mu)=\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\mu=\sum_{i=1}^{n}(x_i)-n\mu=0\,,$$

звідки

$$\mu=\frac{1}{n}\sum_{i=1}^{n}x_i\ .$$

$$\bar{\mu}_{ммп}=\overline{X}\ .$$

З другого рівняння після скорочення і підстановки отримаємо:

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\overline{X})^2-n=0.$$

Звідки отримаємо:

$$\sigma^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{X})^2.$$

Тобто

$$\sigma^2{}_{ммп}=s_x{}^2.$$

Таким чином, оцінками за методом максимальної правдоподібності математичного сподівання і дисперсії випадкової величини X, що має нормальний розподіл, є відповідно вибіркове середнє $\overline{X}$ і вибіркова дисперсія $s_x{}^2$. Оцінки за методом моментів і методом максимальної правдоподібності для середнього і дисперсії співпадають, але тільки для випадкової величини X, що має нормальний розподіл.

Оцінки максимальної правдоподібності, як правило є асимптотично ефективними [11]. Основний недолік цього методу пов'язаний с труднощами розрахунку оцінок, а також і те, що для побудови оцінок і забезпечення їх властивостями необхідно знати закон розподілу випадкової величини, що у багатьох випадках практично неможливо.

2.4 Оптимальні L-оцінки параметрів зсуву і масштабу розподілів за вибірковими квантилями

L-оцінки параметрів зсуву μ і масштабу σ використовуються в тих випа-

дках, коли закон розподілу повністю визначається тільки цими параметрами з функцією розподілу $F\left(\frac{x-\mu}{\sigma}\right)$ і функцією щільності $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$.

L-оцінки параметрів розподілів, що формуються як лінійні комбінації порядкових статистик або вибіркових квантилів, володіють двома важливими для широкого практичного застосування якостями: надзвичайною простотою обчислень і дуже хорошими властивостями робастності. Найскладнішою операцією при обчисленні таких оцінок є сортування наявної вибірки по зростанню (формування варіаційного ряду) з метою визначення вибіркових квантилів спостережуваного закону. Значення вибіркових квантилів $x_{(i)}$ (табл. 2.1) визначають при розбитті області визначення випадкової величини (розмаху вибірки) на інтервали, величини яких пропорційні можливостям P_i попадання в інтервал при асимптотично оптимальному групуванні: число влучень в інтервал вибирається рівним nP_i, де n – обсяг вибірки, і k-кількість інтервалів.

Таблиця 2.1 – Коефіцієнти параметра зсуву. Нормальний розподіл.

k	γ_1	γ_2	γ_3	γ_4	γ_5	γ_6	γ_7	γ_8
3	0,50000	0,50000						
4	0,22437	0,55125	0,22437					
5	0,10857	0,39142	0,39142	0,10857				
6	0,06781	0,23406	0,39624	0,23406	0,06781			
7	0,04318	0,14193	0,31488	0,31488	0,14193	0,04318		
8	0,02987	0,09690	0,21693	0,31257	0,21693	0,09690	0,029871	
9	0,02154	0,06810	0,14860	0,26173	0,26173	0,14860	0,068108	0,02154

В табл. 2.2 отримана оптимальна кількість інтервалів групування та границі цих інтервалів. На рисунку 2.1 наведені центровані щодо істинного значення σ_0 щільності оцінок $\tilde{\sigma}$ параметра нормального закону при обсязі вибірки

$n = 500$ та різному числі $k - 1$ використовуваних вибіркових квантилів для випадку одночасного оцінювання двох параметрів закону. Вибірки нормального закону генерувалися з параметрами $\mu_0 = 0$ і $\sigma_0 = 1$.

Таблиця 2.2 – Рекомендовані значення кількості інтервалів при оцінюванні за вибіркою параметрів нормального закону залежно від обсягу вибірки

Обсяг вибірки n	Число інтервалів k	Обсяг вибірки n	Обсяг вибірки n
≤37	3	390-650	≤10
36-60	≤4	510-850	≤11
67-111	≤5	640-1070	≤12
100-167	≤6	810-1350	≤13
152-254	≤7	1000-1670	≤14
213-355	8	1200+	≤15
294-490	9		

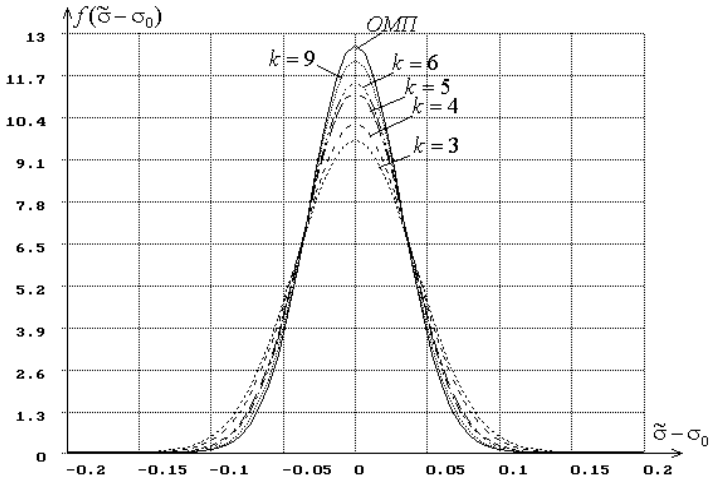

Рисунок 2.1 – Щільності розподілу L-оцінок $\tilde{\sigma}$ при $n = 500$ залежно від k

Оцінювання невідомого μ при відомому σ здійснюється за формулою

$$\mu = \alpha_0 \sigma + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_{k-1} x_{k-1}. \qquad (2.2)$$

Оцінювання невідомого σ при відомому μ здійснюється за формулою

$$\sigma = \beta_0\sigma + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-1} x_{k-1}. \quad (2.3)$$

При оцінюванні відразу двох параметрів використовуються співвідношення:

$$\tilde{\mu} = \gamma_1\hat{x}_1 + \gamma_2\hat{x}_2 + ... + \gamma_{k-1}\hat{x}_{k-1}, \quad (2.4)$$

$$\tilde{\sigma} = \nu_1\hat{x}_1 + \nu_2\hat{x}_2 + ... + \nu_{k-1}\hat{x}_{k-1}. \quad (2.5)$$

Значення $x_{(i)}$, які фігурують в формулах (2.3), (2.2), (2.4) і (2.5) слід вибирати з умови

$$X(\lfloor nI \rfloor) \le \hat{x}_i \le X(\lfloor nI \rfloor + 1),$$

де $X_{(i)}$ – члени варіаційного ряду $X_{(1)} \le X_{(2)} \le .... \le X_{(n)}$, побудованого по вихідній вибірці;

$P^i = \sum_{j=1}^{i} P_j$ – означає цілу частину числа;

P_j – вибираються з відповідного рядка таблиці оптимальних ймовірностей [4]. Наприклад, в якості $x_{(i)}$ можуть бути взяті середні значення між відповідними (табл. 2.1), сусідніми членами варіаційного ряду.

Значення коефіцієнтів для формул (2.3), (2.2), (2.4) і (2.5) що відповідають конкретним законам розподілів, вибираються з (таблиці 2.1).

Рівні квантилів для побудови L-оцінок напряму пов’язані з кількістю рівноймовірних інтервалів, на які розбивається простір вибіркових значень (у таблицях позначено *k*) і є кратними величинам 1/*k*. Такі рівні квантилів можна застосовувати не тільки для нормального розподілу, а також для будь-якого роз-

поділу, симетричного відносно своєї медіани (у тому числі розподілу Коші).

Розподіли порядкових статистик з номерами, що є рівновіддаленими від країв упорядкованої вибірки кінцевого об’єму N є симетричними. Точкою перетину щільностей розподілів таких порядкових статистик навіть з несиметричних неперервних сукупностей є медіана.

3 ПРОГРАМНА РЕАЛІЗАЦІЯ

3.1 Система комп’ютерної алгебри Wolfram Mathematica

Wolfram Mathematica 10 (зазвичай званий Mathematica) – це сучасна технічна обчислювальна система, що охоплює більшість областей технічних обчислень, включаючи нейронні мережі, машинне навчання, обробку зображень, геометрію, науку про дані, візуалізації та інші. Система використовується в багатьох технічних, наукових, інженерних, математичних і обчислювальних областях. Мова Wolfram це мова програмування, який використовується в Mathematica.

Основна ідея мови Wolfram Language – це забезпечення високого рівня результативності програміста шляхом автоматизації наскільки це можливо, операцій, і включення, безпосередньо в мову, якомога більшого обсягу функціональних можливостей.

Мова Wolfram Language виділяється тим, що безпосередньо інтегрує концепції з навколишнього світу і уявлення реально існуючих у ньому цілісних самостійних одиниць. Символьна природа мови робить її ідеальним для написання високорівневих сценаріїв використання зовнішніх систем та мов, природним чином значно покращуючи притаманні вихідній системі інтерфейси. Символьна природа мови Wolfram Language та її інтеграція з обчислюваними документами роблять його ідеальним вибором для метапрограмування та здійснення символьної обробки коду.

Система Mathematica має майже 5000 вбудованих функцій, що покривають всі області технічних розрахунків — всі вони ретельно інтегровані для ідеальної спільної роботи, і всі вони включені в повністю інтегровану систему Mathematica. Система Mathematica використовує Wolfram Notebook Interface, який дозволяє організувати все, що Ви робите, в багатий змістом документ, який включає текст, здійсненний код, динамічну графіку, інтерфейс і багато іншого. слів, мова Wolfram Language винятково просто читати, використовува-

ти та вивчати. Система Mathematica побудована з метою надання можливостей промислової потужності, з міцними ефективними алгоритмами у всіх галузях, здатними вирішувати великомасштабні завдання з паралелізмом, обчисленнями на графічних процесорах та багато інших. Супер функції, мета-алгоритми, система Mathematica надає прогресивне високорівневе середовище з максимальним рівнем автоматизації, що дозволяє бути найпродуктивнішими.

3.2 Алгоритм розв'язання задачі аналізу, та стійкого оцінювання параметрів в потоках даних

На вхід програми подається потік даних з невідомими степенями відхилень. Після цього генерується вибірка розміром *n* яка заповнюється першими *n* элементами потоку даних. Після того як вибірка була згенерована потік даних не зупиняється, а отже і вибірка повинна змінюватися відповідно до даних що надходять. Для цього кожний елемент що надходить має імовірність (0,9) замінити вже існуючий в вибірці. Для визначення імовірності, в програмі генерується за допомогою генератора рівномірно розподілених чисел випадкове число від 0 до 1, якщо воно більше 0.5 то зміни відбуваються і навпаки. Так само і визначається елемент вибірки що буде змінюватися, тобто за допомогою генератора рівномірно розподілених чисел генерується видаткове натуральне число від 1 до n наприклад 4 і тоді необхідно виконати заміну четвертого елемента попередньої буферної вибірки на елемент потоку, що щойно надійшов. Після формування нової вибірки проводиться аналіз та стійке оцінювання параметрів статистичного розподілу за допомогою M та L оцінок.

M оцінки вираховуються за формулами

$$\hat{\mu}_M = \frac{1}{n}\sum_{i=1}^{n} x_i ;$$

$$\sigma_M{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2 = s_x{}^2 .$$

Знайдемо L-оцінки.

В припущенні, що вибірка належить нормальному закону, знайдемо оптимальні L-оцінки його параметрів. Число інтервалів вибираємо максимально можливим. З огляду на обмеження $\min_i nP_i > 1$, намагаємося, щоб, по крайній мірі, виконувалося нерівність $nP_i > 3/5$. Рівні квантилів (табл. 2.1) знайдено з припущення, що вибірка має симетричний розподіл і належить нормальному закону. А точне значення взято з експерименту [9]. Розбиваючи впорядковану вибірку на інтервали пропорційно nP_i, знаходимо граничні точки інтервалів як середні значення між спостереженнями, які потрапили в суміжні інтервали:

$$\hat{x}_{(1)} = (X_{(3)} + X_{(4)})/2;$$

$$\hat{x}_{(2)} = (X_{(20)} + X_{(21)})/2;$$

$$\ldots\ldots\ldots$$

$$\hat{x}_{(k)} = (X_{(n-6)} + X_{(n-7)})/2.$$

Використовуючи коефіцієнти таблиць, наведених в [8], знаходимо оцінку параметра зсуву

$$\hat{\mu}_L = \gamma_1(\hat{x}_{(n)} + \hat{x}_{(1)}) + \gamma_2(\hat{x}_{(n-1)} + \hat{x}_{(2)}) + \ldots + \gamma_k(\hat{x}_{(n-n/2)} + \hat{x}_{(n-(n/2-1))});$$

використовуючи коефіцієнти таблиць, наведених в [8], – оцінку параметра масштабу

$$\hat{\sigma}_L = \nu_1(\hat{x}_{(n)} - \hat{x}_{(1)}) + \nu_1(\hat{x}_{(n-1)} - \hat{x}_{(2)}) + .. + \nu_k(\hat{x}_{(n-7)} - \hat{x}_{(n-6)});$$

Після проведення стійкого оцінювання знову формується нова вибірка з новими даними і про нове стійке оцінювання. Цей цикл закінчується лише тоді коли дані перестають надходити.

3.3 Опис програми

Спочатку задається потік даних з файлу за допомогою команди

$$Imort["Data/elements.xls"],$$

це дозволяє зчитувати інформацію в реальному часі і не втрачати час на перенесення даних власноруч.

Після цього задається перші *k* елементів що будуть входити до нашої початкової вибірки. Коли вибірка заповнена і надходить *n*-й елемент, ми поміщаємо його в резервуар з ймовірністю *k/ n*, де *k*-розмір резервуара, а *n* - номер оброблюваного елемента даних. Для вирішення про те, чи помістити в резервуар *n*-й елемент, в програмі генерується за допомогою генератора рівномірно розподілених чисел випадкове число від 0 до 1 за допомогою команди *RandomReal*. Якщо воно менше або дорівнює *k/n*, то поміщаємо елемент в резервуар, замінюючи один випадковий елемент попередньої буферної вибірки на елемент потоку, що щойно надійшов. Це задається за допомогою команд:

$$a = k/(n);$$
$$If\left[a >= RandomReal[1],\right.$$
$$\{q = RandomInteger[\{1,n\}];$$
$$Part[mat,q] = d.$$

Після генерування вибірки проводиться застосування стійких М- оцінок

за допомогою команди:

$$X = N[Total[R] / Length[R]],$$

де задається сума всіх елементів в вибірці і ділиться на довжину вибірки. Так знаходимо математичне очікування. Після чого визначено дисперсію використовуючи команду:

$$X1 = \left(Part\left[T,3\right] + Part\left[T,4\right]\right) / 2;$$

Також для того щоб в кінці роботи програми визначити середнє значення математичного очікування та дисперсії кожне значення буде додаватися до попереднього, за допомогою циклу:

$$wer = X + wer;$$

де X це значення математичного очікування. Усі вхідні данні будуть записані та проаналізовані додатково для подальшого аналізу. Після чого серед усіх оцінок знаходимо середнє квадратичне значення математичного очікування і дисперсії за допомогою команд:

$$Msr = wer / L;$$
$$Qsr = qsr / L;$$

де L це загальна кількість циклів що виконала програма, а *wer* і *qsr* – сума усіх значень математичне очікування та дисперсія відповідно. Після того як потік даних закінчується цикл зупиняється. Вже з відомим потоком даних ми можемо задати новий з іншим степенем засмічення і порівняти стійкі оцінки.

Для L-оцінок кількість інтервалів і рівні квантилів були визначенні зазда-

легідь з припущення що вибірка має симетричний і нормальний розподіл. Розбиваючи впорядковану вибірку на інтервали пропорційну nP_i при k, знаходимо граничні точки інтервалів як середні значення між спостереженнями, які потрапили в суміжні інтервали:

$$X1 = (Part[T,17] + Part[T,18]) / 2;$$
$$X2 = (Part[T,33] + Part[T,34]) / 2;$$
$$X3 = (Part[T,45] + Part[T,46]) / 2;$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$X6 = (Part[T,91] + Part[T,92]) / 2;$$

Після чого знаходимо оцінку параметра зсуву:

$$M2 = 0,04318 * (X1 + X6) + 0,141936 * (X2 + X5) + 0,314884 * (X3 + X4),$$

а також оцінку параметра масштабу:

$$Q2 = 0,095717 * (X6 - X1) + 0,186279 * (X5 - X2) + 0,136715 * (X4 - X3).$$

Зазначимо, що рівні квантилів різні для кожної величини вибірки. Тут задано лише для приклад. Звісно що знайдені стійкі оцінки необхідно вивести на екран, це робиться за допомогою команди:

$$Print["\backslash n \;\; Моценки = ", Msr, "\backslash n \;\; Q=", Qsr, "\backslash n \;\; Loценки = ", M2, "\backslash n \;\; Q=", Q2].$$

4 РЕЗУЛЬТАТИ ОБЧИСЛЮВАЛЬНОГО ЕКСПЕРИМЕНТУ

Вхідними даними є потік інформації про зміну курсу гривні до долару, взятий з загально доступного сайту [12].

Для аналізу потоку даних були використані вибірки розміром 50,100,1000. Після цього був проведений аналіз за допомогою методу максимальної правдоподібності і L-оцінок.

Знайдемо значення асиметрії.

Значення асиметрії може бути позитивним, нульовим, негативним чи невизначеним. Для унімодального розподілу негативний перекіс зазвичай свідчить про те, що хвіст перебуває у лівій частині розподілу, а позитивний перекіс вказує, що хвіст перебуває справа. Якщо один хвіст довгий, а інший товстий, перекіс не підпорядковується простому правилу. Наприклад, нульове значення означає, що хвости з обох боків від середнього в цілому врівноважуються; це вірно для симетричного розподілу, але може також бути вірно і для асиметричного розподілу, коли один хвіст довгий і тонкий, а інший короткий, але товстий:

$$\mu_0 = 1 \;\; \mu_1 = 0;$$

$$\mu_2 = a_3 - 2a_1^2;$$

$$\mu_3 = a_3 - 3a_2a_1 + 2a_1^3;$$

$$\mu_4 = a_4 - 4a_3a_1 + 6a_2a^2{}_1 - 3a_1^4;$$

$$A = \frac{\mu_3}{\sigma^3[\mathrm{X}]} = -0.2253;$$

σ необхідно оцінювати за вибіркою.

Коефіцієнт асиметрії має значення менше нуля, це означає що лівий хвіст розподілу довший за лівий.

Знайдемо ексцес в потоці даних щоб побачити до якого розподілу він на-

лежить.

Ексцес для будь-якого одновимірного нормального розподілу дорівнює 3. Зазвичай ексцес розподілу порівнюють із цим значенням. Розподіли з ексцесом менше 3 є плоскими, хоча це не означає, що розподіл є «плоским», як іноді стверджують. Швидше це означає, що розподіл виробляє менші і менше екстремальних викидів, ніж нормальний розподіл. Розподіли з ексцесом більше 3-х називаються лептокуртичними [9]. Також звичайною практикою є використання скоригованої версії ексцесу Пірсона, надлишкового ексцесу, який є ексцесом мінус 3, для порівняння зі стандартним нормальним розподілом. Деякі автори [10] використовують термін «ексцес» для позначення надлишкового ексцесу.

$$E = \frac{\mu_4}{\sigma^4[\mathrm{X}]} - 3 = 2{,}5216.$$

Для нормального закону розподілу $\mu_4 = 3\sigma^4$.

Це значить що розподіл є «плоским», тобто це означає, що розподіл виробляє менші і менше екстремальних викидів, ніж нормальний розподіл.

Для подальшого експерименту обираєм засмічення що задається за допомогою формування буферної вибірки і замінуємо її частку певних даних, що мають інший закон розподілу. У якості розподілу засмічення обрано розподіл Коші де щільність ймовірностей задається як:

$$f(x) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x - x_0}{\gamma}\right)^2\right]};$$

а функція розподілу має вигляд:

$$F(x) = \frac{1}{\pi} arctg\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2},$$

де $x_0 \in R$ — параметр зсуву, а $\gamma > 0$ — параметр масштабу.

Модель вивчення засмічення має вид:

$$F(x) = (1-\varepsilon)F_0(x) + \varepsilon H(x).$$

Ця модель є також моделлю Тьюкі – Хубера [13]. Вона показує, що з близькою до 1 ймовірністю, а саме, з ймовірністю $(1-\varepsilon)$ спостереження беруться з сукупності з функцією розподілу $F_0(x)$ яка передбачається, що володіє "добрими" властивостями. Наприклад, вона має відомий статистику вид (хоча б з точністю до параметрів), у неї існують всі моменти, і т.д. Але з малою вірогідністю ε з'являються спостереження з сукупності з "поганим" розподілом, узяті з розподілу Коші, яке не має математичного сподівання, що різко виділяються аномальні спостереження, тобто викиди.

Проведемо порівняльній аналіз стійких М оцінок та L оцінок, що формуються як лінійні комбінації порядкових статистик або вибіркових квантилів, при різних величинах буферної вибірки і при різних степенях засміченості. Порівняємо отримані значення математичного очікування та дисперсії.

Проведемо порівняльний аналіз L_1 та L_2 для вибірки розміром 100 без засмічення (табл. 4.1). та при засміченні (табл. 4.2).

Число інтервалів вибираємо для L1 =6 а для L2=4. З огляду на обмеження $\min_i nP_i > 1$, намагаємося, щоб, по крайній мірі, виконувалося нерівність $nP_i > 3/5$. Рівні квантилів знайдено з припущення, що вибірка має симетричний розподіл і належить нормальному закону. А точне значення взято з (табл. 2.1).

Розбиваючи впорядковану вибірку на інтервали пропорційною nP_i, знаходимо граничні точки інтервалів як середні значення між спостереженнями, які потрапили в суміжні інтервали. Для L_1

$$\hat{x}_{(1)} = (X_{(15)} + X_{(16)})/2; \hat{x}_{(2)} = (X_{(30)} + X_{(31)})/2;$$

$$\hat{x}_{(3)} = (X_{(45)} + X_{(46)})/2;\ \hat{x}_{(4)} = (X_{(60)} + X_{(61)})/2;$$
$$\hat{x}_{(5)} = (X_{(75)} + X_{(76)})/2;\ \hat{x}_{(6)} = (X_{(94)} + X_{(95)})/2;$$

для L_2

$$\hat{x}_{(1)} = (X_{(12)} + X_{(13)})/2;\ \hat{x}_{(2)} = (X_{(30)} + X_{(31)})/2;$$
$$\hat{x}_{(3)} = (X_{(55)} + X_{(56)})/2;\ \hat{x}_{(4)} = (X_{(88)} + X_{(89)})/2;$$

для L_3

$$\hat{x}_{(1)} = (X_{(4)} + X_{(98)})/2\ \hat{x}_{(2)} = (X_{(6)} + X_{(96)})/2;$$
$$\hat{x}_{(3)} = (X_{(12)} + X_{(88)})/2\ \hat{x}_{(4)} = (X_{(18)} + X_{(82)})/2;$$
$$\hat{x}_{(5)} = (X_{(26)} + X_{(76)})/2\ \hat{x}_{(6)} = (X_{(30)} + X_{(76)})/2;$$
$$\hat{x}_{(7)} = (X_{(36)} + X_{(70)})/2\ \hat{x}_{(8)} = (X_{(42)} + X_{(64)})/2;$$
$$\hat{x}_{(9)} = (X_{(48)} + X_{(58)})/2;\ \hat{x}_{(10)} = (X_{(53)} + X_{(54)})/2.$$

Використовуючи коефіцієнти таблиць, наведених в [9], знаходимо оцінку параметра зсуву. Для L_1

$$\hat{\mu}_{L1} = 0,2243(\hat{x}_{(1)} + \hat{x}_{(6)}) + 0,5512(\hat{x}_{(2)} + \hat{x}_{(5)}) + 0,2243(\hat{x}_{(3)} + \hat{x}_{(4)}),$$

для L_2

$$\hat{\mu}_{L2} = 0,500(\hat{x}_{(1)} + \hat{x}_{(4)}) + 0,500\hat{x}_{(2)} + \hat{x}_{(3)}),$$

для L_3

$$\hat{\mu}_{L3} = 0,067815(\hat{x}_{(1)} + \hat{x}_{(10)}) + 0,234061(\hat{x}_{(2)} + \hat{x}_{(9)}) + 0,396249(\hat{x}_{(3)} + \hat{x}_{(8)}) +$$

$$+0,234061(\hat{x}_{(4)}+\hat{x}_{(7)})+0,067815(\hat{x}_{(5)}+\hat{x}_{(6)});$$

використовуючи коефіцієнти таблиць, наведених в [9]. Для L_1

$$\hat{\sigma}_{L1}=-0,3614(\hat{x}_{(6)}-\hat{x}_{(1)})+0(\hat{x}_{(5)}-\hat{x}_{(2)})+0,36142(\hat{x}_{(4)}-\hat{x}_{(3)});$$

для L_2

$$\hat{\sigma}_{L2}=-0,4502(\hat{x}_{(4)}-\hat{x}_{(1)})+0,4502(\hat{x}_{(3)}-\hat{x}_{(2)});$$

для L_3

$$\hat{\sigma}_{L3}=-0,140732(\hat{x}_{(1)}+\hat{x}_{(10)})+-0,235892(\hat{x}_{(2)}+\hat{x}_{(9)})+0(\hat{x}_{(3)}+\hat{x}_{(8)})+$$
$$+0,235892(\hat{x}_{(4)}+\hat{x}_{(7)})+0,140732(\hat{x}_{(5)}+\hat{x}_{(6)});$$

Таблиця 4.1 – Порівняльний аналіз L_1, L_2 і L_3

	L1	L2	L3
M	3069,61	3092,72	3076,77
S^2	743,14	763,14	783,14
A	-0,2797	-0,3233	-0,2853
E	2,00335	2,0123	2,1609

З результатів виходить що доцільно використовувати оцінку L_1.

Таблиця 4.2 – Порівняльний аналіз L_1, L_2 і L_3 до засмічених даних

	L1	*L2*	*L3*
M	3309	3084,35	3076,77
S^2	673,85	675,4	783,14
A	-0,2559	-0,3217	-0,2253
E	2,745	2,0539	2,2909

Отже, точність а відповідно і вибір L оцінок залежить від величини вибірки що буде оцінюватись, звідси виходить що L_1 доцільно використовувати при оцінці вибірки розміром 100, L_2 при вибірці розміром 50, і L_3 при 1000.

Знайдемо значення асиметрії, ексцесу і математичного очікування для M і L оцінок (табл. 4.3).

Таблиця 4.3 – Стійке оцінювання в потоці даних

	M			*L*		
	A	*E*	*M*	*A*	*E*	*M*
k=50	-0,286857	2,15827	3075,94	-0,2497	2,1384	3085
k=100	-0,2934	1,9899	3076,733	-0,3022	2,04204	3090,88
k =1000	-0,5112	2,0107	3066,66	-0,4073	2,17143	3089,3

Таблиця 4.4 – M і L оцінок в потоці даних з засміченням 0,1

	M-оцінки			L-оцінки		
	A	*E*	*M*	*A*	*E*	*M*
k=50	-0,0739	3,0173	3234,01	-0,3309	2,3014	3101,6
k=100	-0,8956	2,9709	3223,24	-0,132	2,5548	3611,27
k =1000	-0,610	2,1254	3245,56	-0,5798	1,9651	3151,81

Проведемо порівняємо застосування стійких M і L оцінок до потоку даних без засмічення (табл. 4.5).

Таблиця 4.5 – Стійкі оцінки в потоці даних

	Розмір вибірки	k=50		k=100		k=1000	
		M	S^2	M	S^2	M	S^2
M-оцінки	Мінімальне значення оцінки	2639,53	768,97	2674,84	767,581	2971,17	450,54
	Максимальне значення оцінки	3485,95	794,5	3389,42	795,4	3147,13	456,54
	Середнє значення оцінки	3077,9	783,14	3077,02	785,284	3052,84	554,35
L-оцінки	Мінімальне значення оцінки	2622,54	767,77	2244,23	770,97	3063,01	379,36
	Максимальне значення оцінки	3467,42	7963	3167,75	793,5	3092,01	386,9
	Середнє значення оцінки	3085,44	784,54	2938,55	786,24	3074,84	383,14

Зі збільшенням величини буферної вибірки стійкі оцінки мають значно меншу ефективність. Це значить що величину буферної вибірки слід формувати залежно від очікуваної кількості даних, що буде надходити.

Також слід порівняти застосування стійких оцінок при рівні засміченості даних 0,1 (табл. 4.6) щоб порівняти якість стійких оцінок.

Найкращий результат оцінювання при великій кількості засмічення має метод L-оцінок. Цей метод оцінювання призводить до меншого зсуву у порівнянні з методом максимальної правдоподібності.

Таблиця 4.6 – Оцінки математичного очікування при замічені 0,1

	Розмір вибірки	k=50		k=100		k=1000	
		M	S^2	M	S^2	M	S^2
M-оцінки	Мінімальне значення оцінки	2583,71	766,97	2708,25	779,97	3120,92	394
	Максимальне значення оцінки	3297,8	805,5	3584,62	810,5	3328,25	408,5
	Середнє значення оцінки	3209,82	793,14	3211,09	785,35	3231,97	403,14
L-оцінки	Мінімальне значення оцінки	2698,07	767,77	2770,7	767,77	3143,18	380,15
	Максимальне значення оцінки	3509,32	810,3	3456,57	853,23	3155,88	386,36
	Середнє значення оцінки	3090,86	796,54	3026,23	784,54	3172,1	384,54

Знайдемо довірчий інтервал в потоці даних. Довірчим називається інтервал I_γ, у який із заданою довірчою ймовірністю (надійністю) γ потрапляють значення параметра θ. Імовірність γ обирається близькою до 1, в нашому випадку γ=0,95. Іноді розглядають односторонні довірчі інтервали: відповідно верхній (вигляду $\theta < t_2$) і нижній (вигляду $t_1 < \theta$):

$$\left(\bar{x} - t_{kp}\frac{s}{\sqrt{n}}; \bar{x} + t_{kp}\frac{s}{\sqrt{n}}\right);$$

$$\varepsilon = t_{кр}\frac{s}{\sqrt{n}} = 1{,}96\frac{215}{\sqrt{3650}} = 7{,}02;$$

$$(3076{,}77 - 7.02; 3076{,}77 + 7{,}02);$$

$$(3069{,}75;\ 3083{,}79).$$

З ймовірністю 0.95 можна стверджувати, що середнє значення під час вибірки більшого обсягу не вийде за межі знайденого інтервалу.

Побудуємо графік математичного очікування після стійкого оцінювання потоку даних, що дозволить детальніше його оцінити (рис. 4.1).

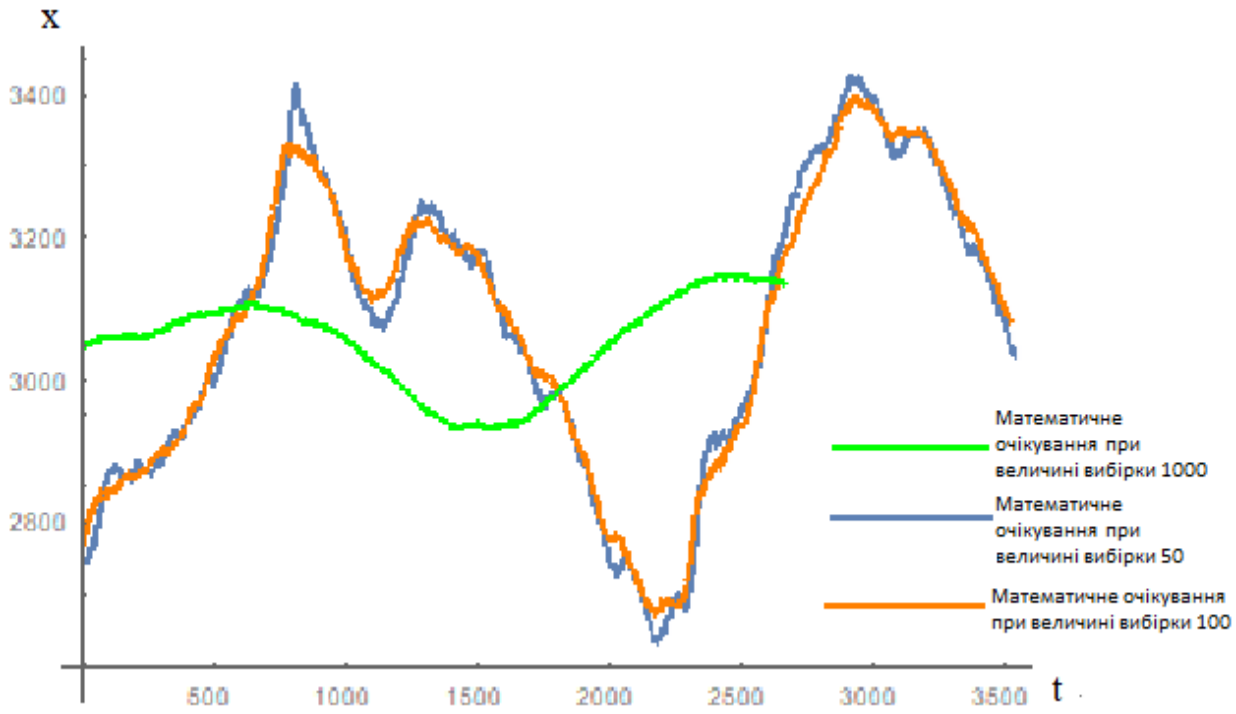

Рисунок 4.1 – Оцінки математичного очікування в потоці даних за допомогою М-оцінок

На графіку ми бачио три кольори де зелений це потік даних при обробці з вибіркою величина якої 1000, помаранчевий це величина 100 і синій 50. Зрозуміло, що при великому обсязі вибірки програмній аналіз данних почнеться значно пізніше, так як на заповнення цієї вибірки необхідний певний час відпорвідно і дані будуть значно відставати. Також їм необхідний певний час для заповнення вибірки, і її оцінювання що приводить до того що коли будуть отримані стійкі оцінки вони будуть вже “застарілими” і необхідність в них відпаде.

Найбільш точна інформація поступає при стійкому оцінюванні даних з величиною вибірки 50. Тобто при невеликому потоці даних краще

використовувати вибірки невеликого розміру, і збільшувати їх відповідно до величини даних що надходять.

Для більш детального і точного аналізу параметрів статистичного розподілу в потоках даних, було проведено додаткові дослідження з відомим степенем відхилень (рис. 4.2).

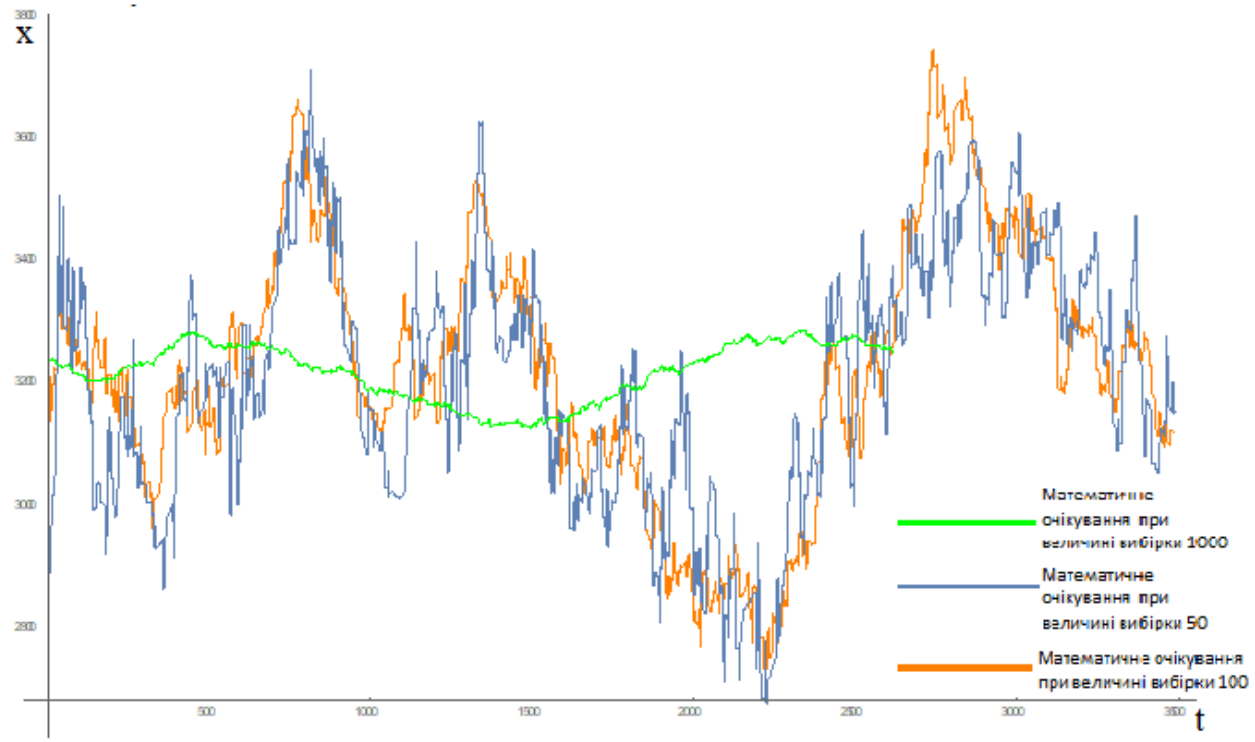

Рисунок 4.2 – Оцінки математичного очікування в потоці даних за допомогою М-оцінок зі степенем засмічення 0,1

На графіку можна побачити що при використанні вибірки розміром 1000 математичне очікування майже не змінилися на відміну від інших, це дозволяє сказати що ця вибірка є більш стійкою до викидів, але вона менш детально зображує зміни в потоці, саме це дозволяє з впевненістю використовувати її в потоці з дуже великою кількістю даних.

Також побудуємо графік математичного очікуванн для L-оцінок, що дозволить детальніше його оцінити (рис. 4.3).

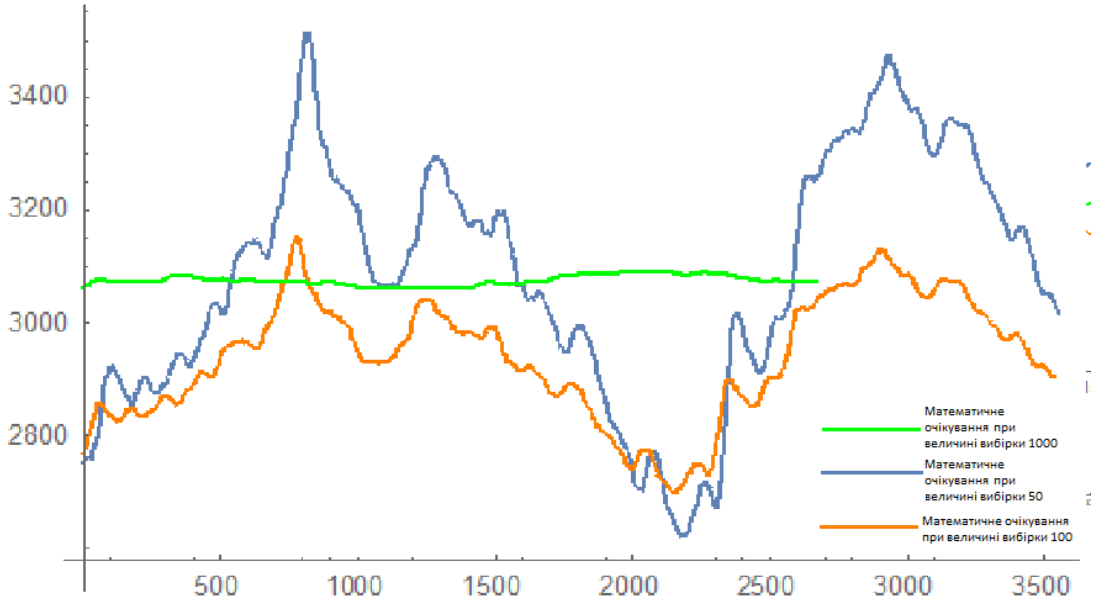

Рисунок 4.3 – Оцінки математичного в потоці даних L-оцінки

На графіку зображено три кольори де зелений це потік даних при обробці з вибіркою величина якої 1000, помаранчевий це величина 100 і синій 50. Зрозуміло, що при великому обсязі вибіркі програмній аналіз данних почнеться значно пізніше, так як на заповнення цієї вибірки необхідний певний час і дані будуть трішки відставати.

Також можливо помітит що найбільш точна інформація поступає при обрці даних з величиною вибірки 50. Тобто при невиликому потоці даних краще використовувати вибірки невеликого розміру, і збільшувати їх відповідно до величини даних що надходять.

Для більш детального і точного аналізу параметрів статистичного розподілу в потоках даних, було проведено додаткові дослідження з відомим степенем відхилень. Величин було зміно так само як і в випадку з методом максимальної імовірності, і отримана такі дані (рис. 4.4).

Графік свідчить, що L-оцінки не чутливі до відхилень. При великому обсязі вибірки в 1000 лементів L-оцінки показують середнье значення величин в потку даних, це свідчить про те що як і в випадку з M-оцінками краще використовувати вибірки не великого обсягу.

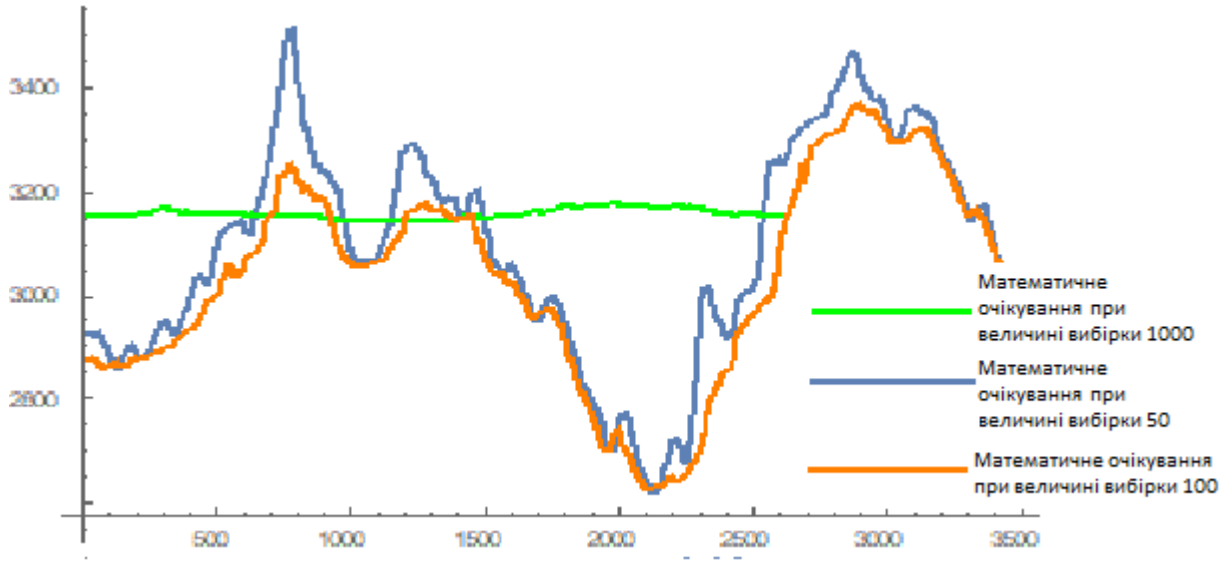

Рисунок 4.4 – Оцінки математичного очікування в потоці даних з степенем засмічення 0,1 за допомого L-оцінки

Тепер порівняємо застосування стійких оцінок (рис. 4.5), при обсязі вибірки 50. Для цього побудуємо графік стійких M і L оцінок до потоку даних, щоб побачити відхилення.

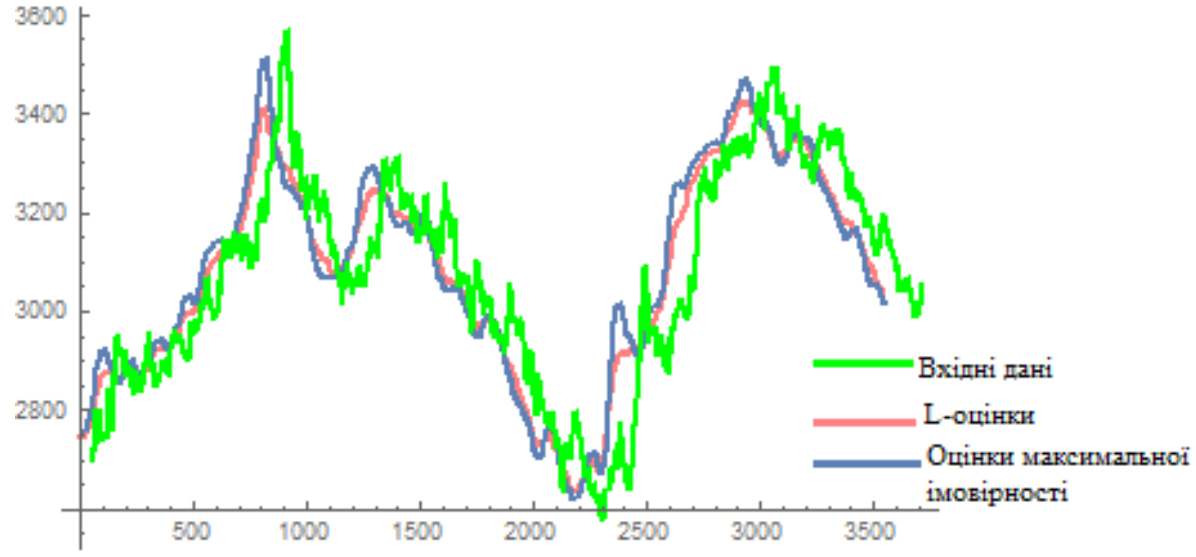

Рисунок 4.5 – Порівняння застосування стійких оцінок, при обсязі вибірки 50

На рис 4.5 зображено зеленим кольором вхідні дані, синім після оцінювання методом максимальної імовірності і помаранчевим після використання L-

оцінок. Як видно найкращий результат отримано при використанні L-оцінок. Стійкі оцінки відстають від потоку даних через необхідність формування вибірки і подальшого її оцінювання.

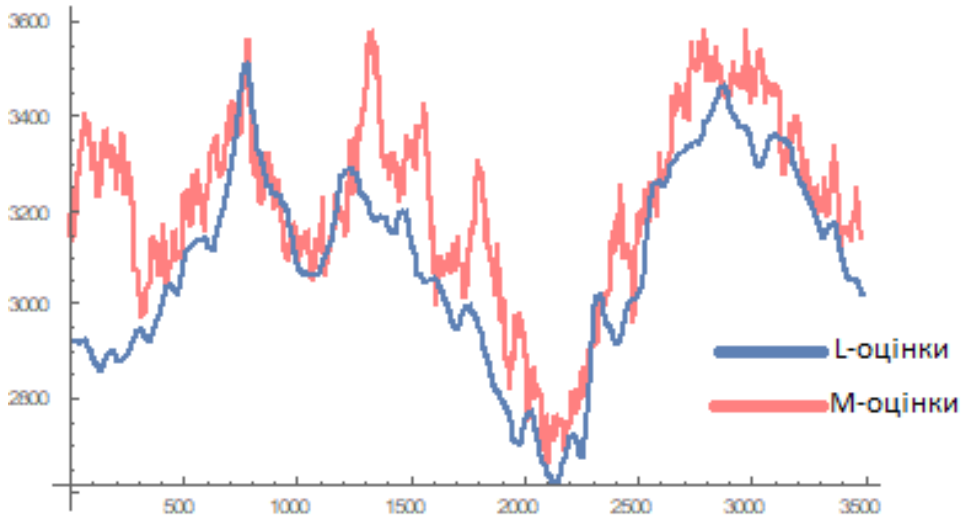

Рисунок 4.6 – Порівняння застосування стійких оцінок, при обсязі вибірки 50 зі засміченням 0,1

З графіку (рис. 4.6) зрозуміло що L оцінки є більш стійкими до засмічення в потоці даних що робить доцільне їх використання. L-оцінки мають велику кількість переваг. Застосування готових таблиць [9] ймовірностей попадання в інтервал, що відповідають асимптотически оптимальному групуванню, і формул, що спираються на обчислені таблиці коефіцієнтів, робить процес обчислення цих оцінок дуже простим. Не потрібне спеціальне програмне забезпечення. Виключаючи процес формування варіаційного ряду, який елементарно реалізується сортуванням вибірки в будь-якій електронній таблиці, всі обчислення обмежуються десятком арифметичних операцій. Розглянуті методи дають хороші результати безпосередньо як оцінки від-повідних характеристик розподілів, так і в інших завданнях математичної статистики.

5 АНАЛІЗ МОЖЛИВИХ ЗАСТОСУВАНЬ

Дані оточують нас скрізь, і кожен день в мережі з'являються нові джерела даних. Якщо ви поки не зустрічалися зі створенням систем обробки даних в реальному часі, то зіткнетеся в найближчому майбутньому. З ростом Інтернету і очікуваних клієнтів значення потокової передачі даних і потокової обробки зростає. Персональні монітори здоров'я та домашні системи безпеки - два приклади джерел потокової передачі даних. Система домашньої безпеки включає в себе кілька датчиків руху для спостереження за різними частинами будинку. Ці датчики генерують потік даних, неперервно передаючи їх в інфраструктуру обробки, яка відстежує будь-яку нежданну активність або в режимі реального часу, або зберігають дані для аналізу, щоб їх було складніше виявити в подальшому. Монітори здоров'я - ще один приклад джерел потокової передачі даних, включаючи монітори серця, артеріального тиску чи кисню. Ці пристрої неперервно генерують дані. Своєчасний аналіз цих даних крайнє важливий, так як від цього може залежати безпека людини. Обробка цих даних у режимі реального часу є одночасно проблемою та можливістю для організацій.

При традиційній обробці даних часто зберігаються у великих обсягах в сховищах даних. Вартість цих систем зберігання та обладнання часто є тягарем для організацій. При потоковій обробці дані не зберігаються у великих обсягах, тому системи обробки мають менші витрати на устаткування.

Аналіз ті стійке оцінювання потоків даних в реальному часі дозволяє організаціям неперервно контролювати свою бізнес-екосистему. Дані інформують організації про можливі порушення безпеки, виробничі проблеми, незадоволеність клієнтів, фінансові кризи або неминуче порушення суспільного іміджу. Завдяки неперервній потоковій передачі та обробці даних організації можуть уникнути таких проблем.

Завдяки обробці даних у реальному часі організації можуть заздалегідь вирішувати можливі проблеми до того, як вони матеріалізуються. Це дає їм час та перевагу над конкурентами. Потокова передача та обробка даних також під-

вищує задоволеність клієнтів, оскільки їхні проблеми можна вирішувати в реальному часу. Завдяки неперервній обробці даних у реальному часі немає затримок, викликаних зберіганням даних на складах, що очікують на обробку.

Дана робота може бути використана для вирішення проблеми аналізу та стійкого оцінювання параметрів в потоках даних таких як наприклад зміни на фондовому ринку, курсу валюти, кількість хворих

Цей напрямок має ще багато перспективних реалізацій вже на цей час, та має великий потенціал до ще більшого розширення можливостей.

ВИСНОВКИ

У роботі проведений системний аналіз проблеми стійкого оцінювання параметрів статистичного розподілу в потоках даних. Проведено моделювання потоку даних у чистому вигляді та на основі вибраної моделі засмічення. Розглянуті оцінки максимальної правдоподібності (M-оцінки) та оцінки на основі комбінацій порядкових статистик (L-оцінки).

В результаті виконання роботи була розроблена програма, що реалізує аналіз і стійке оцінювання параметрів статистичного розподілу в потоках даних за допомогою M та L-оцінок. Дану програму можна використовувати для стійкого оцінювання параметрів статистичного розподілу для будь якого потоку даних що робить доцільним її використання не лише в даній роботі, а і в інших сферах послуг.

Детально розглянуто методи для задачі стійкого оцінювання параметрів статистичного розподілу в потоках даних. Розроблено алгоритм роботи та архітектуру програми. Для розробки продукту було обрано Wolfram|Alpha 10. Наведений аналіз роботи програми та практичні рекомендації по роботі з програмою.

Дана робота може бути використана для вирішення проблеми аналізу та стійкого оцінювання параметрів в потоках даних таких як наприклад зміни на фондовому ринку, курсу валюти, кількість хворих. Така аналітика зростаючих обсягів даних в режимі, близькому до реального часу, може забезпечити ключову перевагу для бізнесу в конкурентних галузях, що швидко розвиваються.

Визначено що залежно від імовірного потоку даних слід змінювати і розмір вибірки по якій проводиться оцінка. Так при невеликому потоці даних слід брати вибірку не великого розміру, а при його збільшенні збільшувати відповідно і вибірку.

Найточніші оцінки було отримано при використанні вибірки найменшого розміру 50, а найбільш спотворенні оцінки при 1000. Це означає що для більш точного оцінювання краще використовувати вибірки невеликого розміру, але це

так само призводить до більшого завантаження обчислювальної системи. Тому необхідно обирати вибірки розміром найбільш підходящих до вашого потоку даних та апаратної системи що буде використовуватись.

Пропоновані оптимальні L-оцінки параметрів зсуву і масштабу по вибірковим квантилям є найкращими в своєму класі. Застосування готових таблиць ймовірностей попадання в інтервал, відповідних асимптотично оптимальному групуванню, і формул, що спираються на обчислені таблиці коефіцієнтів, робить процес обчислення цих оцінок дуже простим. Не потрібно складного спеціального програмного забезпечення. Виключаючи процес формування варіаційного ряду, який елементарно реалізується сортуванням вибірки в будь-якій електронній таблиці, все обчислення обмежуються десятком арифметичних операцій. Вони стійкі до наявності аномальних помилок вимірювань, до малих відхилень від вихідних припущень про вид спостережуваного закону розподілу. Це дозволяє використовувати L-оцінки в процедурах параметричного відбраковування спостережень.

Використання L-оцінок не викликає проблем при подальшій перевірці адекватності отриманої моделі, так як в можна скористатися критеріями згоди типу Пірсона і відношення правдоподібності. Застосування готових таблиць ймовірностей попадання в інтервал, відповідних асимптотично оптимальному групуванню, з одного боку, робить елементарною процедуру обчислення статистики, з іншого, – забезпечує максимальну потужність проти близьких альтернатив.

Все вищесказане дозволяє рекомендувати використання L-оцінок для оперативного аналізу потоку даних. Доцільність застосування L-оцінок визначається сукупністю 2-х достоїнств: простотою обчислень і робастністю. Природно, ці рекомендації не виключають можливості застосування на наступних етапах аналізу більш ефективних оцінок і більш потужних критеріїв.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Douglas McIlroy. URL : https://en.wikipedia.org/wiki/Douglas_McIlroy (дата звернення: 10.10.2021).

2. Прискорене навчання нейронної мережі ADALINE за наявності стаціонарного корельованого шуму Інформаційні системи та технології: праці 10-ї Міжнародної науково-технічної конференції, Харків-Одеса, 13-19 вересня 2021 року / О. Безсонов, О. Руденко, Н. Сердюк [та ін.] // Харків: ХНУРЕ, 2021. С. 28-32.

3. Орлов А. И. Прикладная статистика. Москва : Экзамен, 2004. 483 с.

4. Peter J. Huber. Robust Estimation of a Location Parameter. Ann. Math. Statist. Institute of Mathematical Statistics : March, 1964. p. 68.

5. Волкова В. Н., Денисов А. А. Основы теории систем и системного анализа. Санкт-Петербург : Изд-во СПбГТУ, 1997. 510 с.

6. Саати Т. Принятие решений. Метод анализа иерархий. Москва : Радио и связь, 1993. 278 с.

7. Эндрю Дж. Пселтис. Потоковая обработка данных. Конвейер реального времени. Москва : ДМК Пресс, 2018, - 218 с.

8. Лемешко Б. Ю., Чимитова Е. В. Построение оптимальных L-оценок параметров сдвига и масштаба распределений по выборочным квантилям // Сибирский журнал индустриальной математики. 2001. Т. 4, № 2. С. 166–183.

9. Шуленин В. П. Робастные методы математической статистики. Томск : Изд-во НТЛ, 2016. – 260 с.

10. Дэйвид Г. Порядковые статистики. Москва : Наука, 1979. 336 с.

11. Грибкова Н. В., Егоров В. А. В робастных оценках параметра сдвига, являющихся линейными комбинациями порядковых статистик // Вестник ЛГУ, 1978. № 13. С. 24-57.

12. The official exchange rates of the national currency of Ukraine The National Bank. Of Ukraine URL : https://bank.gov.ua/en/markets/exchangerate-chart?startDate=2015-03-20&endDate=2021-12-03 (дата звернення: 10.11.2021).

13. Використання характеристики зразка закону розподілу даних у за-

ваннях машинного навчання Л. Кириченко, О. Пічугіна, В. Кобзєв [та ін.] // Інформаційні системи та технології: праці 10-ї Міжнародної науково-технічної конференції. Харків-Одеса, 13-19 вересня 2021 року. Харків : ХНУРЕ, 2021. С. 208- 212.