Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет комп'ютерної інженерії та управління
(повна назва)

Кафедра електронних обчислювальних машин
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА

Пояснювальна записка

Рівень вищої освіти другий (магістерський)

Метод виявлення аномалій у мобільній мережі передачі даних

(тема)

Виконав:
здобувач 2 року навчання,
групи СПм-23-4
Олександр КОРОТЕЦЬКИЙ
(власне ім'я, прізвище)

Спеціальність
123 «Комп'ютерна інженерія»
(код і повна назва спеціальності)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма
Системне програмування
(повна назва освітньої програми)

Керівник: доц. Віталій МАРТОВИЦЬКИЙ
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ЕОМ _____ Андрій КОВАЛЕНКО
(підпис) (власне ім'я, прізвище)

2025 р.

Харківський національний університет радіоелектроніки

Факультет комп’ютерної інженерії та управління

Кафедра електронних обчислювальних машин

Рівень вищої освіти другий (магістерський)

Спеціальність 123 «Комп’ютерна інженерія»
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системне програмування
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
“____” _____ 20__ р.

ЗАВДАННЯ

НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Коротецькому Олександру Олексійовичу
(прізвище, ім’я, по батькові)

1. Тема роботи Метод виявлення аномалій у мобільній мережі передачі даних

затверджена наказом по університету від “ 21 ” квітня 2025 р. № 296 Ст

2. Термін подання здобувачем роботи до екзаменаційної комісії 16 червня 2025 р.

3. Вхідні дані до роботи Набір даних, що містить лючові параметри радіомережі (KPI)

4. Перелік питань, що потрібно опрацювати у роботі

Аналіз сучасних методів виявлення аномалій у мобільних мережах передачі даних

Вибір та обґрунтування критеріїв аномальності для мобільних мереж

Розробка математичної моделі для виявлення аномалій

Вибір та адаптація алгоритмів машинного навчання

Експериментальна перевірка ефективності методу

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій 15 слайдів

6. Консультанти розділів роботи (заповнюється за наявності консультантів згідно з наказом, зазначеним у п.1)

Найменування розділу	Консультант (посада, прізвище, ім'я, по батькові)	Позначка консультанта про виконання розділу	
		підпис	дата

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Огляд методів виявлення аномалій в	22.04.25-29.04.25	
	мобільних мережах		
2	Вибір та обгрунтування методики дослідження	30.04.25-05.05.25	
3	Вибір інструментальних засобів	06.05.25-09.05.25	
4	Розробка моделей	10.05.25-20.05.25	
5	Проведення експериментів	21.05.25-02.06.25	
6	Оформлення матеріалів кваліфікаційної роботи	03.06.25-05.06.25	
7	Подання кваліфікаційної роботи керівникові та	06.06.25-09.06.25	
	її попередній захист		
8	Подання кваліфікаційної роботи на рецензу-	10.06.25-12.06.25	
	вання		

Дата видачі завдання " 21 " квітня 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. Віталій МАРТОВИЦЬКИЙ
(підпис) (посада, власне ім'я, прізвище)

РЕФЕРАТ

Пояснювальна записка кваліфікаційної роботи: 68 с., 23 рис., 14 табл., 1 дод., 18 джерел.

АНОМАЛІЯ, МОБІЛЬНА МЕРЕЖА, ПЕРЕДАЧА ДАНИХ, АВТОЕНКОДЕР, МАШИННЕ НАВЧАННЯ, ВИЯВЛЕННЯ АНОМАЛІЙ, НАПІВКЕРОВАНЕ НАВЧАННЯ, СЕРЕДНЬОКВАДРАТИЧНА ПОМИЛКА.

Метою кваліфікаційної роботи є розробка ефективного методу виявлення аномалій у мобільній мережі передачі даних з використанням сучасних підходів машинного навчання, зокрема напівкерованих автоенкодерів.

У ході виконання кваліфікаційної роботи було здійснено попередню обробку даних мобільної мережі, зокрема очищення від аномалій та нормалізацію. Для побудови моделі використовувався набір даних без аномалій, розподілений на тренувальний, валідаційний та тестовий піднабори. На основі напівкерованого автоенкодера було створено модель, здатну виявляти відхилення у поведінці мережі за допомогою оцінки середньоквадратичної помилки. Після кількох епох тренування модель досягла стабільних показників точності. Проведено аналіз важливості змінних, що дало змогу визначити найбільш інформативні параметри – зокрема, змінні, пов’язані з часом доби, виявились важливішими за лічильники. Результати показали, що автоенкодер здатен ефективно виявляти нетипову активність у мобільній мережі, що може бути використано для підвищення її безпеки та якості обслуговування.

ABSTRACT

Master’s thesis: 68 pages, 23 figures, 14 tables, 1 appendices, 18 sources.

ANOMALY, MOBILE NETWORK, DATA TRANSMISSION, AUTOENCODER, MACHINE LEARNING, ANOMALY DETECTION, SEMI-SUPERVISED LEARNING, MEAN SQUARED ERROR.

The major goal of this thesis is to develop an effective method for detecting anomalies in a mobile data transmission network using modern machine learning approaches, particularly semi-supervised autoencoders.

During the implementation of the qualification work, preprocessing of mobile network data was performed, including anomaly removal and normalization. A clean dataset without anomalies was used to build the model, divided into training, validation, and test subsets. Based on a semi-supervised autoencoder, a model was developed to detect deviations in network behavior by evaluating the mean squared error. After several training epochs, the model reached stable accuracy levels. A variable importance analysis was conducted, allowing identification of the most informative parameters—specifically, time-of-day variables were found to be more important than counters. The results demonstrated that the autoencoder is capable of effectively detecting abnormal activity in the mobile network, which can be used to improve its security and service quality.

ЗМІСТ

СКОРОЧЕННЯ ТА УМОВНІ ПОЗНАКИ

AE – автоенкодер (англ., Autoencoder)

AN – аномалія (англ., Anomaly)

DNN – глибока нейронна мережа (англ., Deep Neural Network)

KPI – ключові показники ефективності (англ., Key Performance Indicators)

ML – машинне навчання (англ., Machine Learning)

MSE – середньоквадратична помилка (англ., Mean Squared Error)

QoS – якість обслуговування (англ., Quality of Service)

RAN – радіодоступна мережа (англ., Radio Access Network)

SSAE – напівкерований автоенкодер (англ., Semi-Supervised Autoencoder)

UE – кінцевий користувач (англ., User Equipment)

ВСТУП

Сучасні мобільні мережі передачі даних є критично важливою інфраструктурою, що забезпечує безперервний обмін інформацією для мільйонів користувачів, бізнесу та державних установ. З розвитком технологій 4G, 5G та зростанням кількості підключених пристроїв (у тому числі IoT) значно зросли обсяги трафіку, складність мережевої архітектури та вимоги до якості обслуговування. Водночас, це ускладнює завдання підтримки стабільності, надійності та безпеки мобільних мереж, зокрема – виявлення та локалізації аномальних подій у режимі реального часу.

Аномалії в мобільних мережах можуть проявлятись у вигляді раптових змін трафіку, затримок у доставці пакетів, падінь продуктивності базових станцій, порушення маршрутизації або спроб несанкціонованого доступу. Такі події можуть призводити до зниження якості обслуговування (QoS), втрати даних або повної недоступності сервісів для кінцевих користувачів. Тому надзвичайно важливо своєчасно виявляти подібні відхилення від норми, класифікувати їх та приймати відповідні заходи реагування.

У традиційних підходах до моніторингу мобільних мереж основний акцент робиться на граничні значення метрик та фіксовані правила. Однак такі методи часто неефективні в умовах динамічного навантаження, адаптивного поведінкового патерну трафіку та нових типів загроз. У зв’язку з цим все більше уваги приділяється інтелектуальним методам виявлення аномалій, зокрема тим, що базуються на аналізі часових рядів, машинному навчанні, кластеризації та нейронних мережах.

Метою цієї кваліфікаційної роботи є розробка та дослідження методу виявлення аномалій у мобільній мережі передачі даних, який дозволяє ефективно виявляти відхилення в поведінці мережі на основі аналізу її параметрів у реальному або наближеному до реального часі. У роботі буде досліджено архітектуру мобільної мережі, розглянуто існуючі підходи до аналізу трафіку та розроблено алгоритм, здатний адаптуватися до змін у середовищі мережі.

1 ТЕОРЕТИЧНІ ВІДОМОСТІ

1.1 Виявлення аномалій

Виявлення аномалій або виявлення викидів – це процес ідентифікації спостережень у наборі даних, які суттєво відрізняються від більшості інших спостережень. У [1] дається таке визначення викиду: "Спостереження, що відхиляється, або викид – це спостереження, яке помітно відхиляється від інших показників вибірки, в якій воно зустрічається". У [2] це визначення зроблено на крок далі: "Викид – це спостереження, яке настільки відхиляється від інших спостережень, що викликає підозру, що воно було згенероване за іншим механізмом". Саме цей "інший механізм", на який вказує відхилення, в кінцевому підсумку представляє інтерес. Цей "інший механізм" в контексті цього проекту потенційно може бути несправністю мережі або проблемою продуктивності мережі, яку необхідно вирішити, щоб запобігти погіршенню якості послуг. Альтернативною причиною наявності відхилення є помилка. Це може бути пов'язано з будь-якою кількістю помилок у процесі вимірювання, збору, обробки або зберігання даних

Існує велика кількість літератури, присвяченої виявленню аномалій. Огляд наведено в роботі в [3]. Там обговорюються різні підходи до виявлення аномалій, що застосовуються в різних середовищах. В роботі[4] автори зосереджуються на порівнянні різних алгоритмів виявлення аномалій без акценту на різні загальнодоступні багатовимірні набори даних. В роботах [3], і [4] обговорюють три основні аспекти. До них належать тип аномалії, доступні мітки даних і результати роботи методу виявлення аномалій. Розглянуто п'ять різних категорій аномалій. Це локальні, глобальні, точкові, колективні та контекстуальні. Ілюстрація локальної та глобальної аномалії показана на рисунку 1.1.

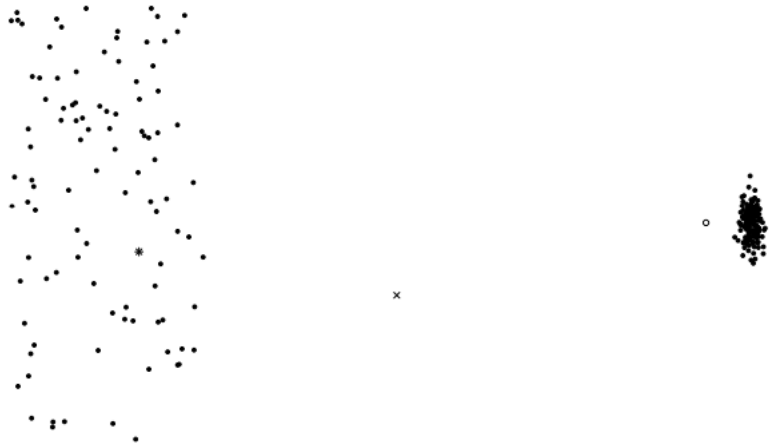


Рисунок 1.1 – Приклад глобального (x) та локального (o) викидів

Точка x знаходиться далеко від інших груп точок даних, що робить її глобальною аномалією. Точка o знаходиться близько до кластера праворуч на діаграмі, але досить далеко від локального кластера, щоб можна було визначити її як локальну аномалію. Як локальні, так і глобальні аномалії є прикладами точкових аномалій, оскільки їх можна відокремити від інших спостережень через їхню відстань від інших спостережень. Колективні аномалії представлені набором спостережень, які зазвичай не спостерігаються разом. Окремо ці спостереження не вважаються аномальними, але оскільки вони зазвичай не спостерігаються разом, то при їх спільному спостереженні вказується на аномальну подію. Контекстуальні аномалії – це приклади точок даних, які потрапляють в очікуваний діапазон, але не для цього конкретного типу спостережень. Наприклад, зовнішня температура коливається від 4 до 40 градусів за Цельсієм. Спостереження зі значенням 40 градусів може не бути відхиленням, якщо воно відбулося влітку, але якщо воно відбулося взимку, то це буде аномалією. У контексті часу спостереження є аномалією. Точка, позначена * в середині кластера зліва на рисунку 1.1, є прикладом контекстуальної аномалії. Вона полягає в тому, що всі спостереження є точками, а зоря відрізняється в контексті формою. Згідно з [4], більшість алгоритмів виявлення аномалій без спостережень призначені для виявлення точкових аномалій. За допомогою

цих підходів можна впоратися з контекстними та колективними аномаліями за допомогою інженерії ознак і включення контексту як ознаки набору даних.

Залежно від міток, наявних у наборі даних, можна застосовувати різні підходи до виявлення аномалій. Якщо є мітки, які вказують на те, чи є спостереження нормальним або аномальним, то можна використовувати підходи керованого навчання. Оскільки аномалії за визначенням рідкісні, підходи керованого навчання повинні враховувати незбалансовані набори даних. Якщо набір даних містить лише нормальні точки, можна використовувати напівкеровані підходи. У цьому випадку модель навчається ідентифікувати нормальну поведінку, а при появі аномального спостереження виявляється відхилення від норми. Нарешті, якщо мітки відсутні, то використовуються неконтрольовані методи виявлення аномалій. Остання категорія є найбільш поширеною через дефіцит маркованих наборів даних. Це дослідження підпадає під останню категорію, оскільки дані не містять міток.

Результатом роботи методів виявлення аномалій може бути або мітка, яка вказує на те, що спостереження є нормальним чи відхиленням, або оцінка, яка вказує на те, наскільки відхиленням є спостереження.

Підходи до навчання без вчителя можна розділити на основі базових внутрішніх характеристик набору даних, які використовуються для виявлення викидів. У роботі [5] визначено підходи на основі статистики, щільності, відстані та відхилення. Згідно з роботою [6], статистичні підходи припускають, що дані походять з певного розподілу. Випадкові дані ідентифікуються за допомогою тесту на розбіжності, при цьому застосовуються різні тести залежно від наявності параметрів розподілу, заздалегідь визначеної кількості випадкових даних та очікуваних типів випадкових даних. Більшість цих тестів на неузгодженість застосовуються лише для одновимірних наборів даних. Другий підхід, заснований на статистиці, називається підходом на основі глибини, де спостереженням присвоюється глибина в багатовимірному просторі, причому викиди, як

правило, мають меншу глибину. Підходи на основі щільності, включаючи Local Outlier Factor [7], використовують локальну щільність спостережень. Спостереження, які знаходяться в зоні з нижчою щільністю, ніж у навколишньому районі, визначаються як викиди. Кожному спостереженню присвоюється оцінка або фактор. Спостереження, яким присвоєно коефіцієнти, близькі до одиниці, вважаються нормальними, а викидам присвоюються значення, більші за одиницю. Чим вище значення, тим більшою мірою воно вважається відхиленням. Підходи на основі відстані використовують відносну відстань між спостереженнями. Вперше такий підхід було представлено в роботі [6], де викиди визначалися як спостереження, що мають найбільшу відстань до k-го найближчого сусіда. Підходи на основі відхилень ґрунтуються на ідентифікації викидів на основі характеристик спостережень.

1.2 Мобільне мережне середовище

Це дослідження проводилося в контексті оператора мобільного зв'язку,. Архітектура мережі відповідає специфікаціям 3GPP [8] і надає послуги голосового зв'язку та передачі даних понад 40 мільйонам абонентів [9]. Мережа надає послуги за допомогою різних технологій радіодоступу, включаючи UMTS, HSPA, HSPA+, LTE, а також технології фіксованого зв'язку, включаючи оптоволокно до будинку. Огляд базової 3GPP Access PLMN, що підтримує послуги з комутацією каналів і пакетною комутацією, згідно з [10], показано на рисунку 1.2. Це дослідження фокусується на даних, що складаються з двотижневих даних, згенерованих елементами SGW в мережі. Функції SGW згідно з [10] даними:

- локальна точка прив'язки мобільності для міжвузлового хендлінгу;
- точка прив'язки мобільності для мобільності між 3gpp;
- буферизація низхідних пакетів у режимі ecm-idle та ініціювання мережевих srp;

- законне перехоплення;
- маршрутизація та переадресація пакетів;
- маркування пакетів на транспортному рівні у висхідній та низхідній лініях;
- облік з деталізацією по користувачам і qci для міжоператорської тарифікації;
- звітування про події (зміна rat тощо) до pcrf;
- прив'язка носія висхідної та низхідної лінії до доступу 3gpp;
- перевірка прив'язки висхідної лінії зв'язку з відкиданням пакетів "неадекватного ul-трафіку";
- шлюз мобільного доступу (mag) функціонує, якщо використовується s5 або s8 на базі pmip;
- підтримка необхідних функцій для забезпечення роботи ланцюжка gtp/pmip.

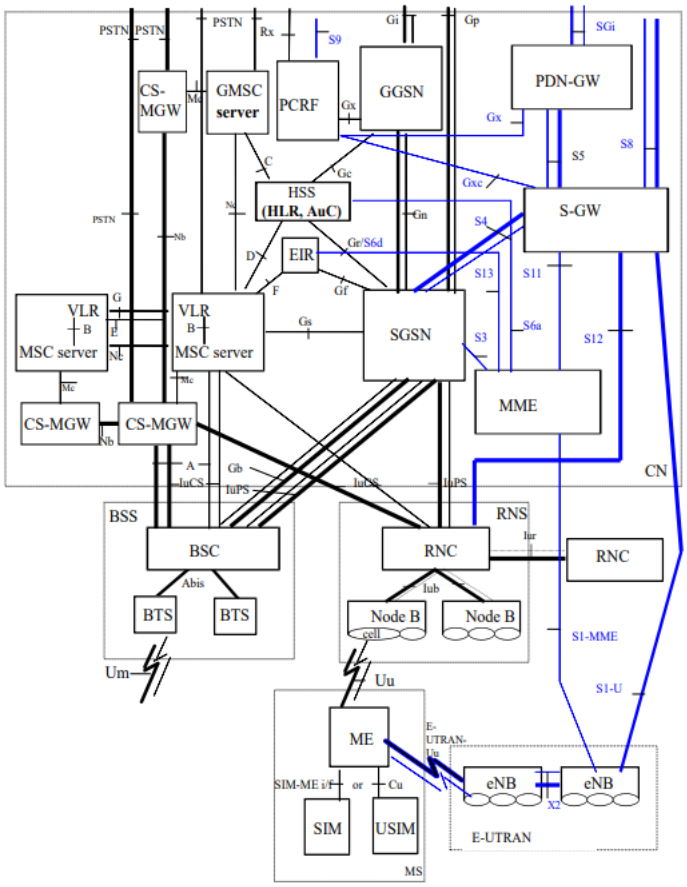


Рисунок 1.2 – Базова конфігурація мережі доступу 3GPP PLMN

Експлуатація та обслуговування мережі здійснюється за допомогою системи FCAPS. FCAPS – це абревіатура від несправності, конфігурації, обліку/адміністрування, продуктивності та безпеки. Сучасний підхід до моніторингу мережі на предмет виявлення несправностей або зниження продуктивності починається з визначення ключових показників ефективності (KPI), заснованих на знаннях про домен. Ці KPI можна згрупувати в 3 основні категорії [11], включаючи показники рівня успішності, такі як рівень успішності активації PDP, показники рівня відмов, такі як рівень обриву дзвінків, і нейтральні показники, такі як кількість одночасно підключених абонентів. Цей ручний процес визначення ключових показників ефективності починається з розуміння архітектури мережі, пропускної здатності елементів, стратегії резервування, функцій і процедур, які повинні виконуватися елементами мережі, потоків сигналів між елементами мережі, необхідних для виконання цих функцій, послуг, що надаються, і пов'язаних з ними моделей трафіку, а також життєвого циклу клієнта. Після визначення ключових показників ефективності, відповідні лічильники, необхідні для звітності за цими KPI, визначаються в даних, що генеруються мережевими елементами. Наступним кроком є визначення порогових значень для подачі сигналу тривоги за кожним KPI. Перевага цього підходу полягає в тому, що він адаптований до конкретної конфігурації розгорнутої мережі та використаного обладнання. Недоліком такого підходу є те, що не всі функції мережі активно відстежуються, залишаючи "білі плями". До того ж, мережа є динамічною, тому KPI, що відстежуються, а також пов'язані з ними порогові значення потребують оновлення. Обидва ці недоліки можна було б усунути, збільшивши розмір команди, яка виконує це завдання, але це є надто дорогим задоволенням. Таблиця 1.2 [9] показує, що мережа повинна підтримувати більше абонентів щороку, в той час як середній дохід на абонента щороку зменшується.

Таблиця 1.1 – ARPU та підрахунок абонентської бази

Рік	ARPU	Активні абоненти	Абоненти передачі даних	Абоненти IoT
2011	R183	22 880 000		
2012	R157	28 941 000		
2013	R129	30 348 000		
2014	R125	31 520 000	15 172 000	1 443 000
2015	R113	32 115 000	16 595 000	1 760 000
2016	R112	34 178 000	18 704 000	2 264 000
2017	R111	37 131 000	19 549 000	2 979 000
2018	R101	41 635 000	20 347 000	3 628 000

Як зазначено в [11], складність мобільних мереж зростає, що разом зі зниженням ARPU зумовлює необхідність зменшення людського навантаження, необхідного для виявлення несправностей та управління продуктивністю мережі. Таке зростання складності мережі можна частково пояснити наступним переліком факторів.

2 ОСОБЛИВОСТІ ІНЖИНІРИНГУ ТА ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

2.1 Джерело даних

У розділі описано джерело, формат, потік, агрегацію та зберігання даних, використаних у цьому дослідженні. На Рисунку 2.1 наведено огляд охоплених питань.

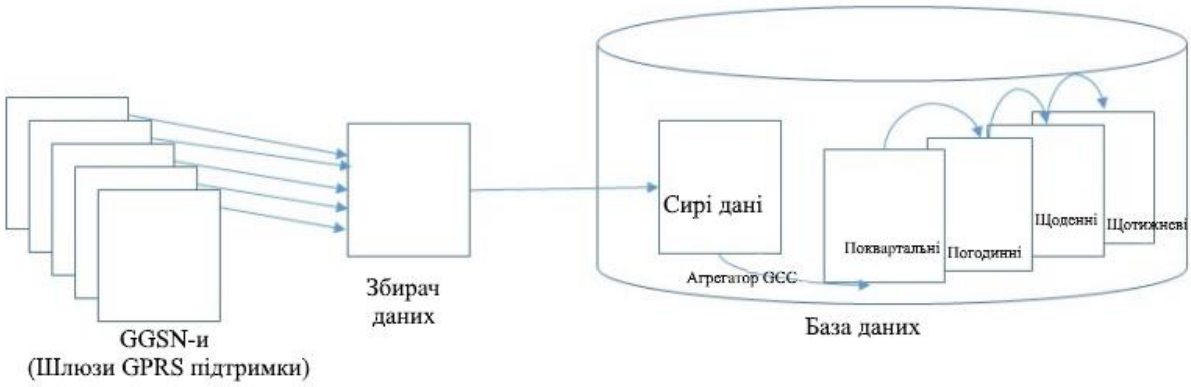


Рисунок 2.1 – Потік даних від генерації через GCC та агрегацію до зберігання

Це дослідження було необхідним для розуміння змісту даних, а також для виявлення потенційного пошкодження даних та проблем з повнотою даних, які самі по собі можуть бути представлені як аномалії. Джерелом даних є вузли GGSN/SGW/PGW мобільної мережі. Кожен вузол налаштований на вимірювання та реєстрацію діяльності, в якій він бере участь. Ці дані зберігаються локально у файлі, який називається "Bulkstats". Кожен вузол генерує новий файл "Bulkstats" кожні 15 хвилин, щоб зафіксувати дані, пов'язані з цим 15-хвилинним інтервалом. Кожен файл розділений на групи лічильників, які називаються схемами. Існує одна схема для кожної основної функції, що виконується кожним вузлом. Наприклад, схема "RADIUS" містить всі лічильники, пов'язані з функцією RADIUS, тоді

як схема "Diameter" містить лічильники, пов'язані з функцією Diameter. Кожна схема містить два типи даних. Перші називаються "Статистика", а другі – "Ключові змінні". Статистика може бути одного з трьох типів. Перший тип – "лічильник", який збільшується з кожною подією, що підраховуються, доки не буде досягнуто ліміту лічильника і лічильник не обнулиться до нуля. Прикладом цього може бути підрахунок кількості байт, що проходять через інтерфейс. Другий тип – "Вимірювач", який дає абсолютне значення в певний момент часу. Прикладом може бути кількість сеансів передачі даних, що підтримуються в певний момент часу. Останній тип – "інформація", який використовується для диференціації наборів статистичних даних. Прикладом може бути IP-адреса. Другим типом даних, що зберігаються в схемі, є "ключові змінні". Вони визначають виміри, з якими пов'язані "статистичні дані", і варіюються від схеми до схеми. У таблиці 2.1 перераховано кількість "статистичних даних" і "ключових змінних", пов'язаних з кожною схемою. Як "статистика", так і "ключові змінні" зберігаються в одному з 4 типів даних. Це int32, який є 32-бітним цілим числом, що переходить при 4 294 967 295, Int64, який переходить при 18 446 744 073 709 551 615, Float, який включає десяткові крапки, і, нарешті, String, який використовується для представлення символів.

Таблиця 2.1 – Підрахунок даних "статистика" та "ключова змінна"

Схема	Таблиці	Підсхеми	Ключові змінні	Статистика
1	2	3	4	5
APN	4	4	3	309
APNQCI	1	1	2	16
CARD	3	3	2	136
DCCA	2	2	4	26
DCCASCH	1	3	3	123
DIAUSCH	1	1	7	44

Продовження таблиці 2.1

1	2	3	4	5
DISCH	1	1	1	53
DPCA	1	1	4	41
ECS	12	12	0	1669
EGTPC	6	9	4	564
GTPC	5	5	6	276
GTPP	2	2	2	180
GTPU	2	2	4	91
IMSA	2	2	2	165
IPPOOL	1	1	8	180
LINK	1	5	1	7
P2PSCH	1	1	1	5
PDSN	15	22	7	2754
PES2A	2	2	5	488
PES2B	2	27	5	568
PES5S81	3	5	2	965
PGW	10	11	4	553
PORT	1	1	1	32
RADIUSGRP	1	1	7	78
RULEBASE	1	1	1	6
SGW	16	19	4	994

Процес ETL (вилучення, перетворення, завантаження) виконується колектором. Він збирає файли "Bulkstats" з вузлів GGSN/SGW/PGW і передає їх до функції бази даних. Функція бази даних завантажує дані в "сирі" таблиці, об'єднуючи всі схеми з окремих вузлів разом. Це призводить до створення таблиці або набору таблиць для кожної схеми. У таблиці 2. перелічено 26 схем.

У процесі завантаження до кожного рядка таблиці або таблиць додається ідентифікатор GGSN/SGW/PGW, географічний регіон, в якому знаходиться вузол, і час початку 15-хвилинного інтервалу. Функція бази

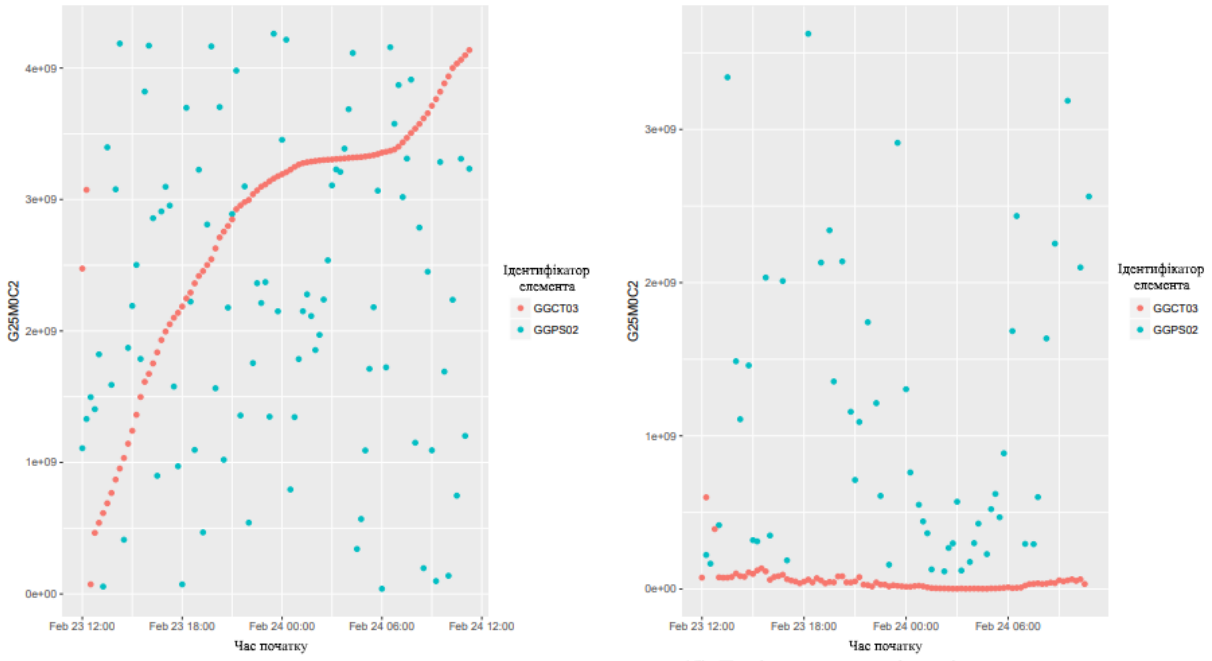
даних створює таблицю "QTR", пов'язану з кожною таблицею "RAW", шляхом корекції статистики типу "Лічильник". Цей процес передбачає віднімання значення лічильника попереднього інтервалу від поточного інтервалу для отримання абсолютного значення, пов'язаного з цим 15-хвилинним періодом. Статистичні дані типу "Gauge" копіюються безпосередньо з таблиць "Raw" до таблиць "QTR". Потім функція "База даних" відповідає за агрегування даних у часовій області в погодинні, добові, тижневі, місячні та річні таблиці. Нарешті, функція бази даних відповідає за видалення старих даних. Періоди зберігання для кожного типу таблиць наведено в таблиці 2.2.

Таблиця 2.2 – Періоди зберігання даних

Тип таблиці	Період зберігання
RAW	14 днів
15 хвилин	14 днів
Погодинна	90 днів
Щоденна	400 днів
Щотижнева	2 роки
Щомісячна	5 років
Щорічна	10 років

Дослідження потоку даних показує, що існує декілька аномалій, які виникають в даних внаслідок проблем з обробкою даних. Ці аномалії описані нижче, починаючи з генерації даних на вузлах GGSN/SGW/PGW і закінчуючи агрегуванням даних у часовій області. При генерації "статистики" типу "лічильник" очікується, що значення лічильника буде обнулятися, коли воно перевищить максимальне значення, яке може бути представлене описаними вище типами даних. Було виявлено два сценарії, коли це не так. Перший випадок виникає, коли процес, відповідальний за керування лічильниками, перезапускається. Коли це відбувається, всі

прирости лічильників за відповідний 15-хвилинний інтервал обнуляються. Це призводить до того, що лічильники обнуляються частіше. Частота повторень варіюється від вузла до вузла. Вузли з більшим трафіком мають більшу кількість випадків. Другий випадок, коли перекидання відбувається занадто часто – це випадки, коли тип даних, що використовується для лічильника, недостатньо великий, щоб відобразити збільшення статистики. На рисунку 2.2 показано приклад того, що відбувається з лічильником G25M0C2. На підрисунку a показано значення для GGCT03 як монотонно зростаючу функцію в необроблених статистичних даних для всіх точок, окрім перших 3, які зазнали перекидання. На підрисунку b показано, як необроблені дані скориговано за лічильником (GCC), що фактично показує значення лічильника для GGCT03 (за винятком перших 3 точок даних). Однак GGPS02 має більшу інтенсивність руху, що призводить до перекидання лічильника кожні 15 хвилин. На підрисунку b показано пошкоджені дані для GGPS02 після GCC.



(а) Сирі дані

(б) Дані, скориговані за лічильником.

Рисунок 2.2 – Вихідні дані, що показують перекидання на лічильнику G25M0C2

Наступна група потенційних аномалій пов'язана з процесом ETL. Спостерігалися випадки, коли функція колектора переривалася. Це призводить до відсутності даних на всіх або деяких вузлах. Неузгодженість у часі між колектором і базою даних призводить до прикладів, коли статистичні дані подвійно враховуються або відсутні для підмножини вузлів. Неправильне визначення "статистики" у функції бази даних призводить до пошкодження даних під час створення таблиці QTR. Були виявлені випадки, коли деякі статистичні дані типу "Вимірник" розглядалися як тип "Лічильник", а деякі типу "Лічильник" розглядалися як тип "Вимірник". Нарешті, до деяких лічильників було застосовано неправильне правило агрегації для агрегацій часового домену. Наприклад, лічильник, що вимірює відсоток, слід агрегувати за допомогою функції максимуму або середнього значення, а не функції суми, і так само лічильник, що вимірює кількість пакетів, слід агрегувати за допомогою суми, а не середнього або максимуму.

З метою зменшення впливу пошкодження даних на подальші методи виявлення аномалій були зроблені наступні кроки. Було перевірено та виправлено визначення типів лічильників. Таймери, що контролюють завантаження бази даних з колекторів, були скориговані для забезпечення відсутності пропущеної або дубльованої статистики. Використання неправильних типів даних, що призводило до перенесення даних, було зареєстровано у постачальника обладнання (виправлення будуть доступні лише в наступному випуску програмного забезпечення). Мета полягала в тому, щоб усунути решту аномалій у потоці даних шляхом їх фільтрації. Це означало, що таблиці QTR повинні були стати основою дослідження, оскільки таблиці з вищим рівнем агрегації в часовому діапазоні вже агрегували відсутні та пошкоджені дані з чистими даними.

Дослідження фокусується на даних, що містяться у схемі SGW.

2.2 Опис даних зі схеми SGW

Схема SGW містить 994 ознаки "статистичних даних", чотири ключові змінні, а також доданий час початку та розміри елементів SGW. Вибірка даних була отримана за 14-денний інтервал з 2018-02-24 09:00 до 2018-03-10 08:15 і складалася з 28161 спостережень. Ключові змінні включають VPN ID, VPN NAME, SERV ID та SERV NAME. У таблиці 2.3 показано їх розподіл для кожного вузла GGSN/PGW/SGW. Кожен вузол має лише один ідентифікатор VPN і один ідентифікатор SERV, які відповідають відповідно IMEHI VPN для "SGW" і IMEHI SERV для "SGW SVC". Таким чином, ключові змінні не вказують на жодний підвимір і можуть бути проігноровані.

Таблиця 2.3 – Схема SGW ключові ознаки

SGW_ID	VPN_ID	VPN_NAME	SERV_ID	SERV_NAME
1	2	3	4	5
CSGNMT01	6	SGW	8	SGW_SVC
GGCF01	4	SGW	9	SGW_SVC
GGCF02	5	SGW	9	SGW_SVC
GGCT01	14	SGW	9	SGW_SVC
GGCT03	7	SGW	9	SGW_SVC
GGCT04	3	SGW	9	SGW_SVC
GGDM01	4	SGW	9	SGW_SVC
GGDM02	3	SGW	9	SGW_SVC
GGDN01	5	SGW	9	SGW_SVC
GGDN02	9	SGW	9	SGW_SVC
GGDN03	3	SGW	9	SGW_SVC
GGJF01	4	SGW	9	SGW_SVC
GGMT01	20	SGW	11	SGW_SVC
GGMT03	7	SGW	9	SGW_SVC
GGMT04	7	SGW	9	SGW_SVC

Продовження таблиці 2.3

1	2	3	4	5
GGPR01	4	SGW	9	SGW_SVC
GGPS02	18	SGW	9	SGW_SVC
GGPS03	7	SGW	9	SGW_SVC
GGPS04	5	SGW	9	SGW_SVC
NFV1-GGMD01	3	SGW	9	SGW_SVC
NFV1-GGPR02	3	SGW	9	SGW_SVC

Вузли SGW розгорнуті в 4 регіонах, як показано в таблиці 2.4

Таблиця 2.4 – Регіональне розташування вузлів GGSN/PGW/SGW

CTN	DBN	JHB	PTA
GGCF01	GGDM01	GGJF01	GGPR01
GGCF02	GGDM02	GGMT01	GGPS02
GGCT01	GGDN01	GGMT03	GGPS03
GGCT03	GGDN02	GGMT04	GGPS04
GGCT04	GGDN03		NFV1-GGPR02
	NFV1-GGMD01		

Для всього подальшого аналізу даних і побудови моделі використовували програму R. Для очищення даних SGW були використані наступні процеси. Спочатку були видалені змінні, що складаються лише з нульових значень. Потім були вилучені змінні з високим відсотком невідповідних значень. Потім були вилучені спостереження з нульовим значенням. Поріг, що використовувався для відсотка не зазначених даних для кожної змінної, впливав на те, скільки спостережень було вилучено на останньому кроці. Цей поріг було обрано для того, щоб мінімізувати кількість даних, вилучених з початкового набору даних. Нарешті, було помічено, що деякі вузли не підтримують функцію SGW. Спостереження для цих вузлів містили лише нулі і були видалені.

З початкових 994 змінних 636 містили лише нулі і залишило 358. Решта змінних складалися лише з позитивних цілих чисел та пропусків. У таблиці 2.5 показано кількість змінних з однаковою кількістю N/A.

Таблиця 2.5 – Кількість змінних за кількістю пропущених значень (N/A)

Variable Count	N/A Count
670	0
6	1
17	2
28	3
11	4
28	5
2	6
5	7
24	8
93	9
104	10
3	11
1	81
1	137
1	139
1	584
1	8075
1	8997

Наприклад, було 670 змінних з 0 н.д. і 1 змінна з 8997 спостереженнями, що містила N/A. Всі змінні, що містили більше ніж 11 N/A, були вилучені з набору даних. Це залишило 349 з початкових 994 змінних. Після видалення спостережень, що містять N/A, залишилося 28148 з початкових 28161 спостережень. 9 вузлів, які не підтримували функцію

SGW, було видалено, залишилося 12 вузлів. Після видалення цих вузлів залишилося 16070 спостережень. Таблиця 2.6 показує, що в остаточному наборі даних SGW кількість спостережень на вузлах збалансована.

Таблиця 2.6 – Спостереження за GGSN/SGW/PGW в остаточних даних SGW

SGW_ID	Observations
GGCF01	1342
GGCT03	1342
GGCT04	1342
GGDM01	1337
GGDN01	1342
GGDN02	1337
GGJF01	1341
GGMT03	1342
GGMT04	1332
GGPR01	1336

Очікується, що активність, яку підтримує функція SGW, буде періодичною, відповідно до активності мобільних абонентів. Змінна G19M4C49 представляє байти низхідного каналу зв'язку на інтерфейсі S1U SGW. На рисунках 2.3, 2.4 та 2.5 показані наступні характеристики обсягів даних, що підтримуються SGW.

Обсяги трафіку мають добовий цикл, з найнижчою точкою рано вранці і найвищою – о 21:00.

Обсяги трафіку, що підтримуються різними вузлами SGW, різняться, причому найбільше навантаження припадає на GGPS04, а найменше - на GGDM01.

Обсяги трафіку в перший день місяця вищі, ніж в останній день місяця.

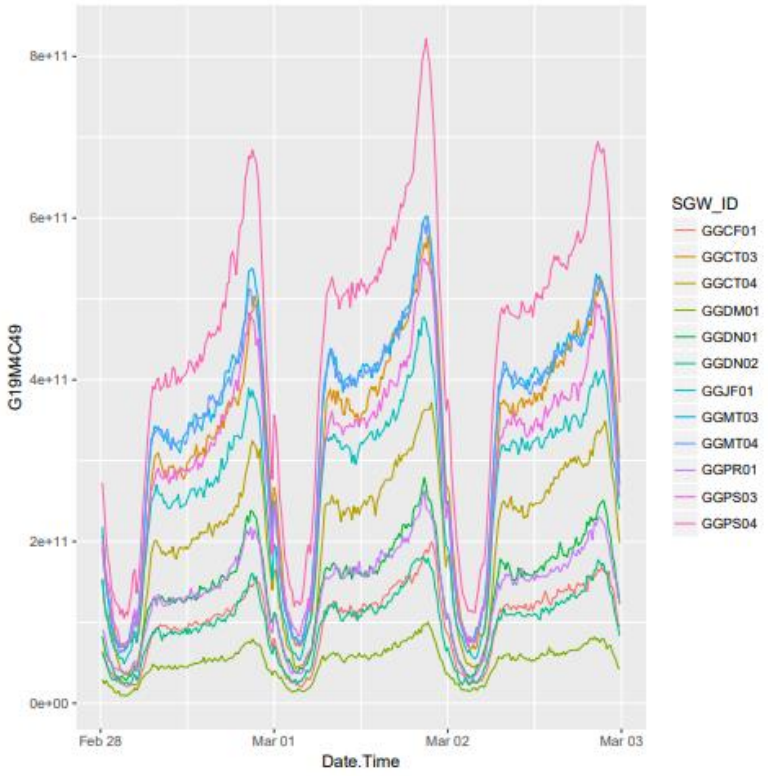


Рисунок 2.3 – Байти в нисхідному каналі на інтерфейсі S1U SGW

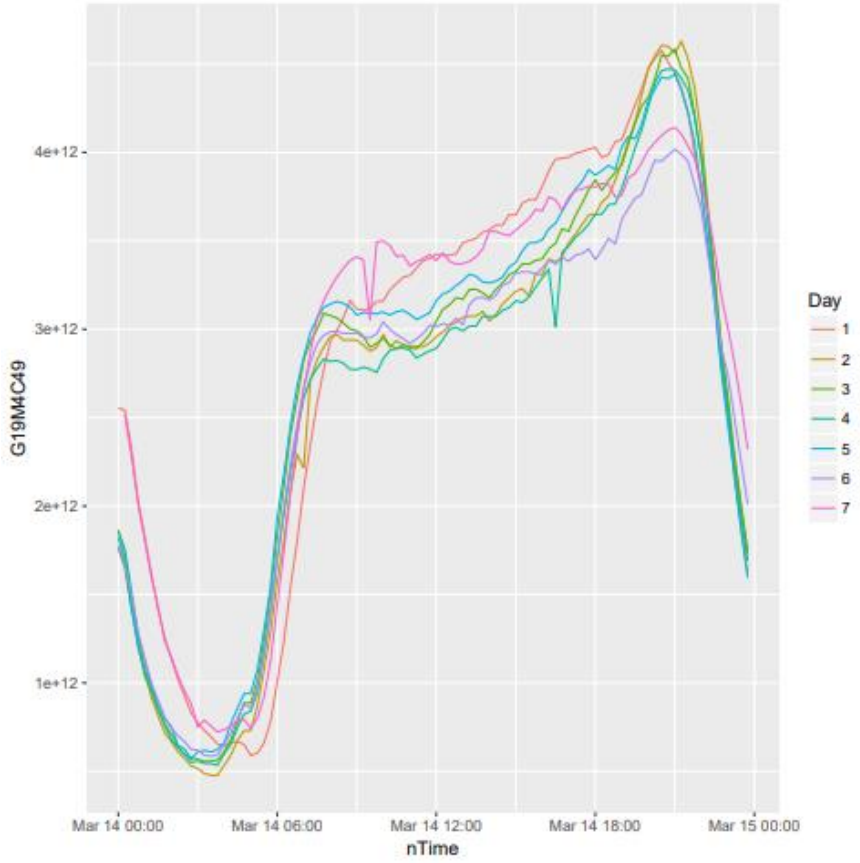


Рисунок 2.4 – Порівняння днів тижня

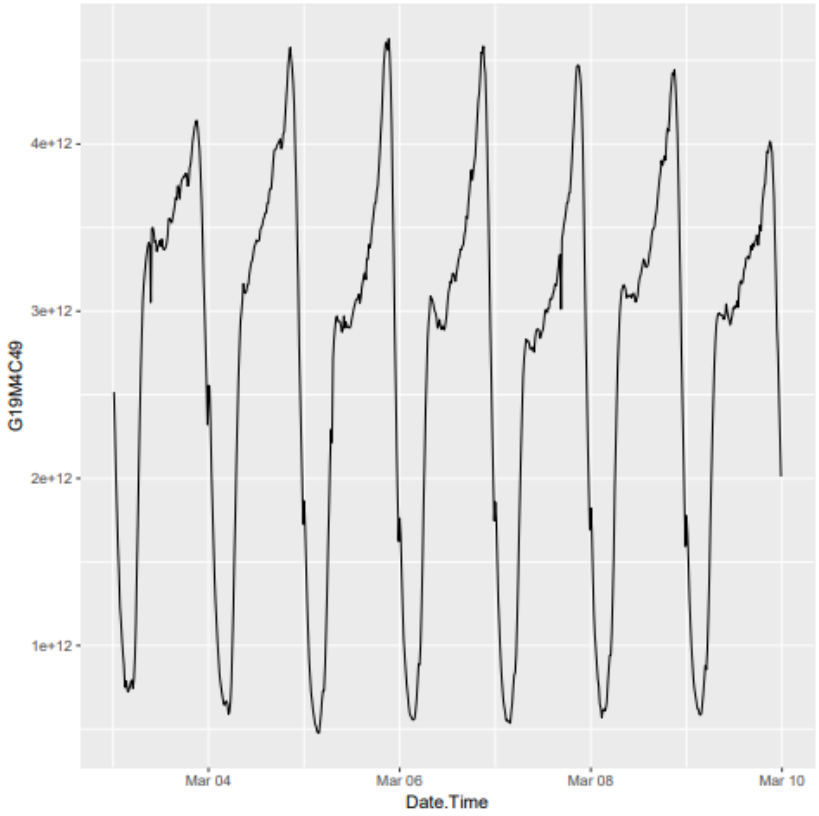


Рисунок 2.5 – Загальний обсяг байт S1U- за 7 днів

Інтенсивність руху в суботу та неділю залишається вищою до ранку порівняно з буднями.

У неділю інтенсивність руху вранці довше залишається нижчою, ніж в інші 6 днів тижня.

Різниця в інтенсивності руху між ранком і пізнім вечором найменша в суботу.

На рисунку 2.3 видно два провали на 7-й та 4-й день. Вони спричинені відсутністю спостережень для GGMT04.

Оскільки змінні є вимірами різних видів діяльності, що підтримуються вузлом, очікується, що між ними буде високий рівень кореляції. Наприклад, рано вранці, коли більшість абонентів сплять, дуже мало абонентів відправляють або отримують дані, це означає, що кількість відправлених пакетів, байт відправлених, отриманих пакетів, байт отриманих буде нижчою для всіх 9 груп індикаторів класу якості обслуговування. Це означає, що

всі 36 лічильників, які вимірюють ці показники, матимуть низькі значення. Коли люди почнуть прокидатися і використовувати дані, всі ці лічильники почнуть синхронно зростати. Оскільки в кореляційній матриці занадто багато змінних, вона представлена в растровому форматі, де кореляції, що дорівнюють 1, показані чорним кольором, а -1 - білим. Все, що знаходиться між ними, показано різними відтінками сірого. Рисунок 2.6 показує високий рівень кореляції.

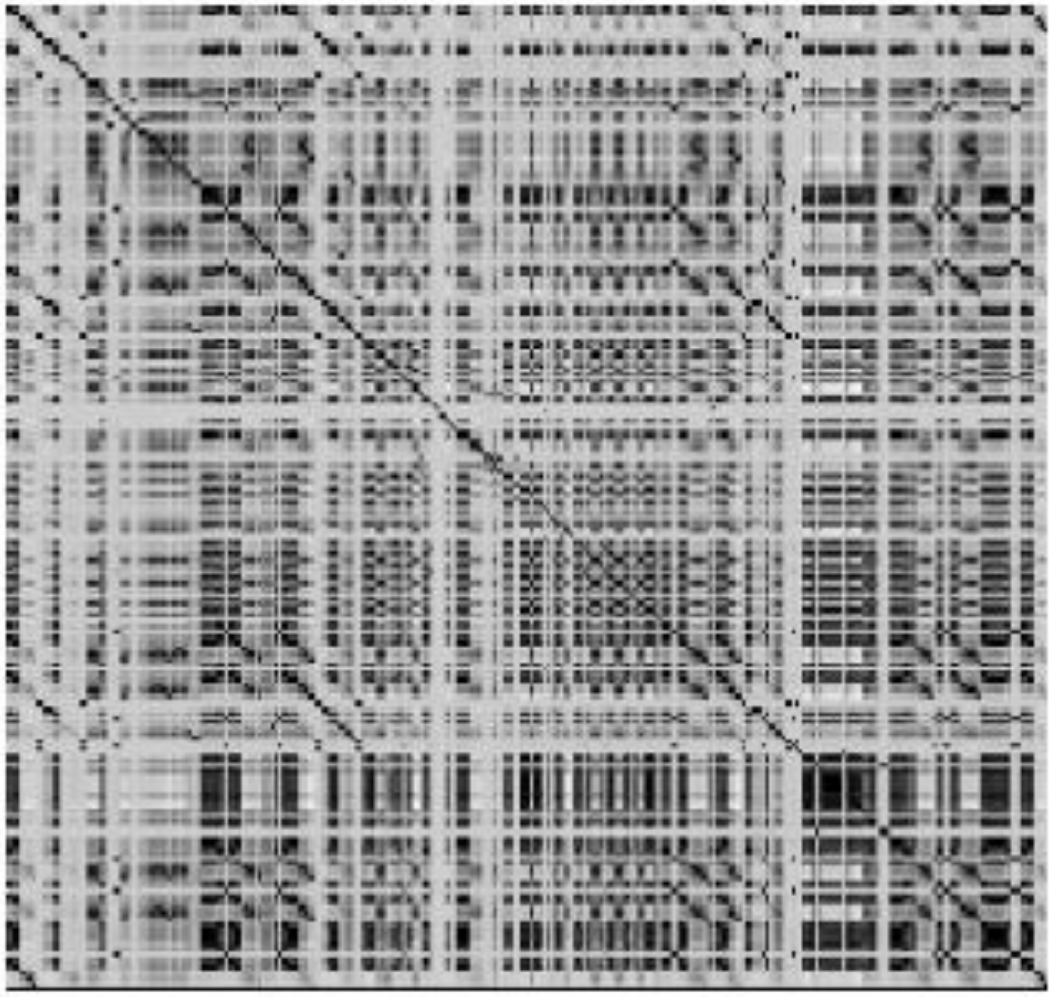
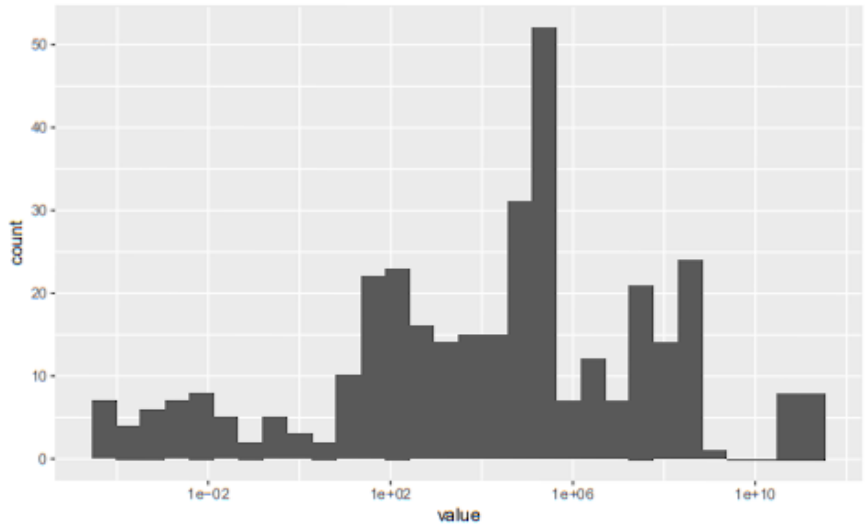
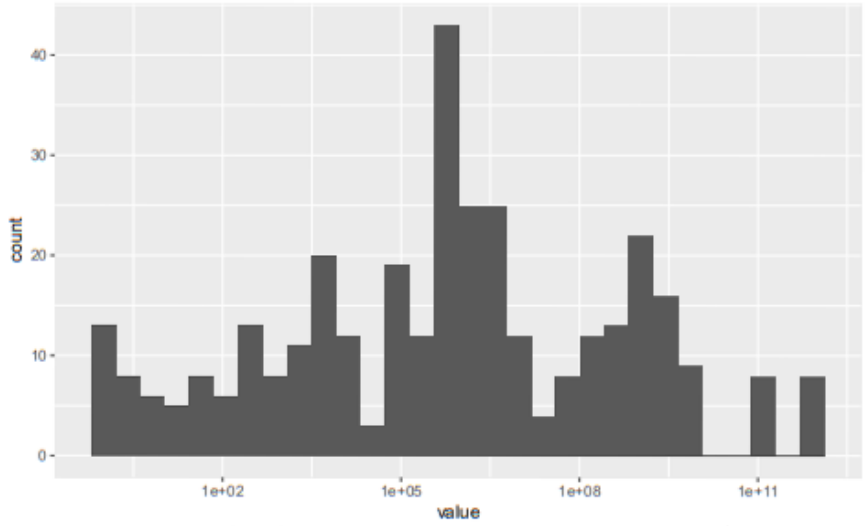


Рисунок 2.6 – Потік даних від генерації через GCC та агрегацію до зберігання

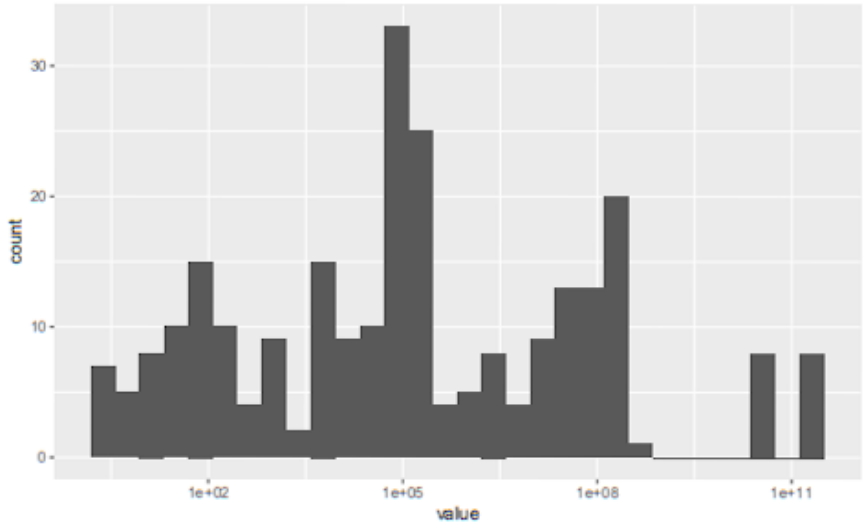
На рисунках 2.7 та 2.8 показано зведені дані змінних SGW.



a) Гістограма середніх значень змінних SGW

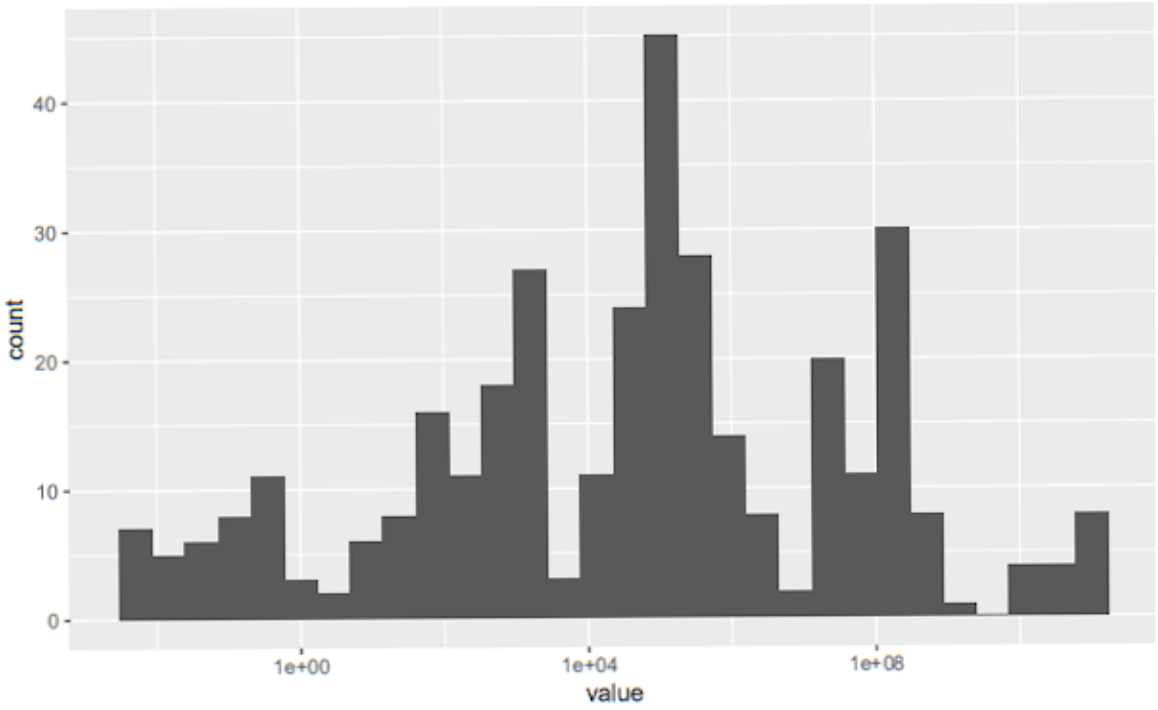


b) Гістограма максимальних значень змінних SGW

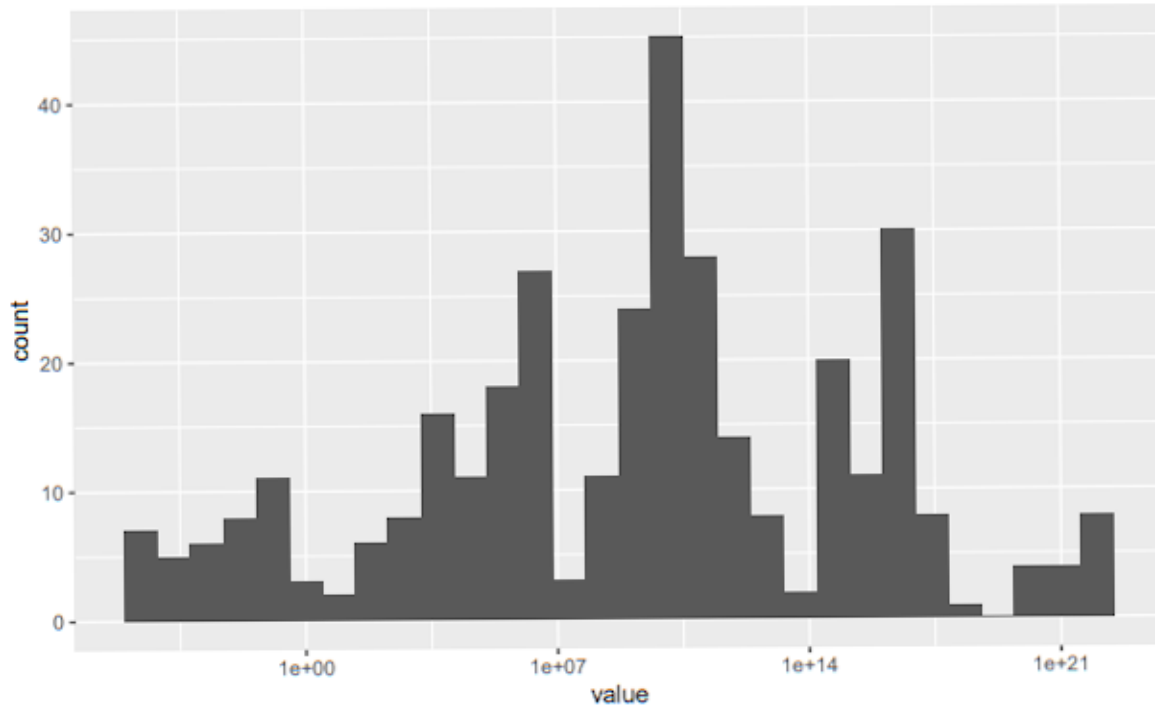


c) Гістограма медіанних значень змінних SGW

Рисунок 2.7 – Узагальнення даних SGW



a) Гістограма стандартних відхилень змінних SGW



b) Гістограма дисперсій змінних SGW

Рисунок 2.8 – Узагальнення даних по SGW

Розподіл значень не є рівномірним, з однією групою низьких значень, що домінує в результатах.

3 МЕТОДИ ВИЯВЛЕННЯ АНОМАЛІЙ

Застосований підхід полягав у підготовці набору даних таким чином, щоб його можна було використовувати в напівконтрольованому навчанні. Для цього потрібно було виявити та видалити викиди з вихідних даних. Для цього спочатку використовується аналіз головних компонент (PCA), щоб зменшити кількість змінних. Потім ці дані подаються до 4 методів виявлення аномалій, а саме: K-найближчих сусідів, однокласових машин опорних векторів, локального фактору викидів на основі щільності та багатовимірного гаусівського розподілу. Результати використовуються для виявлення та видалення спостережень, які, найімовірніше, є аномаліями у наборі вихідних даних. Отриманий набір даних формує вхідні дані для підходу напівкерованого навчання.

3.1 Зменшення розмірності

Як показано на рисунку 3.1, дані SGW є високо корельованими і, отже, виграють від зменшення розмірності за допомогою PCA. Вперше метод PCA було представлено в роботі [12]. Цей підхід усуває надмірність даних шляхом створення нових незалежних змінних (головних компонент), які є лінійними комбінаціями вихідних змінних. Перша головна компонента пояснює найбільшу варіацію в даних, а кожна наступна головна компонента пояснює все менше і менше. Мета полягає в тому, щоб зменшити кількість вимірів/змінних, які використовуються в наступних методах виявлення аномалій, що застосовуються для видалення аномалій з даних як підготовка до підходу напівкерованого навчання. Крім того, вплив "прокляття розмірності" буде зменшено в подальшому аналізі.

Варіація, що пояснюється першими 3 головними компонентами даних SGW, становить 8,55%, 5,52% і 3,12% відповідно, що разом становить 17,2%.

Перші 60 головних компонент пояснюють 81,67% варіації, а перші 100 – 95,23%.

Графік перших двох головних компонент на Рисунку 3.1 показує потенційні кластери на основі часу доби, причому спостереження, що відбуваються в один і той самий час доби, згруповані разом. Рисунок 3.2 показує, що спостереження, згенеровані кожним вузлом, також утворюють кластери.

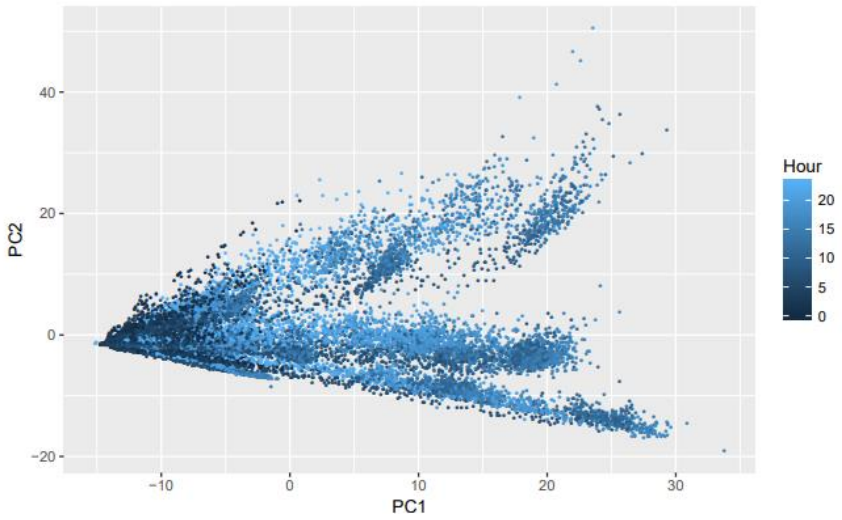


Рисунок 3.1 – Перша та друга головні компоненти

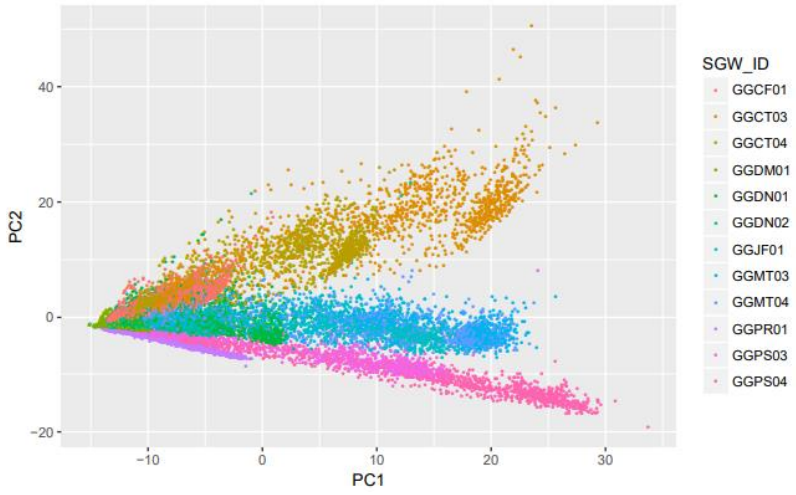


Рисунок 3.2 – Перша та друга головні компоненти на вузол

При подальшому розбитті даних за вузлами та днями тижня, починають проявлятися викиди на основі перших двох головних компонент. Приклади можна побачити на перший день для GGCT04 і на третій день для GGDN01.

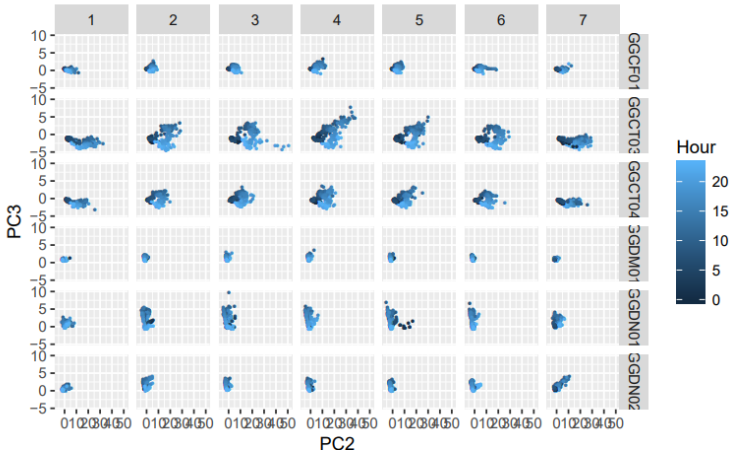


Рисунок 3.3 – Перша та друга головні компоненти на один вузол за добу

Зосередившись на одному з вузлів, GGJF01, і поділ за днем тижня і годиною доби, можна побачити більше кластерів, заснованих на цих вимірах.

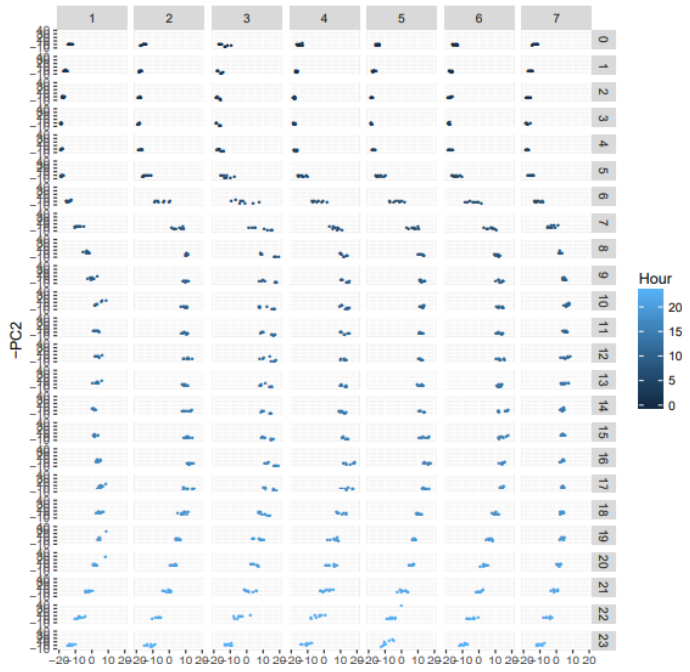


Рисунок 3.4 – Перша та друга головні компоненти для вузла GGJF01 поділеного за днем тижня і годиною доби

Рисунок 3.5 знову показує ті ж самі результати для GGJF01, але замість того, щоб ґрунтуватися на оригінальних даних PCA, які були зроблені на всьому наборі даних, що складається з усіх вузлів, він показує результат окремого PCA, зробленого на підмножині цих даних, що складається тільки зі спостережень GGJF01.

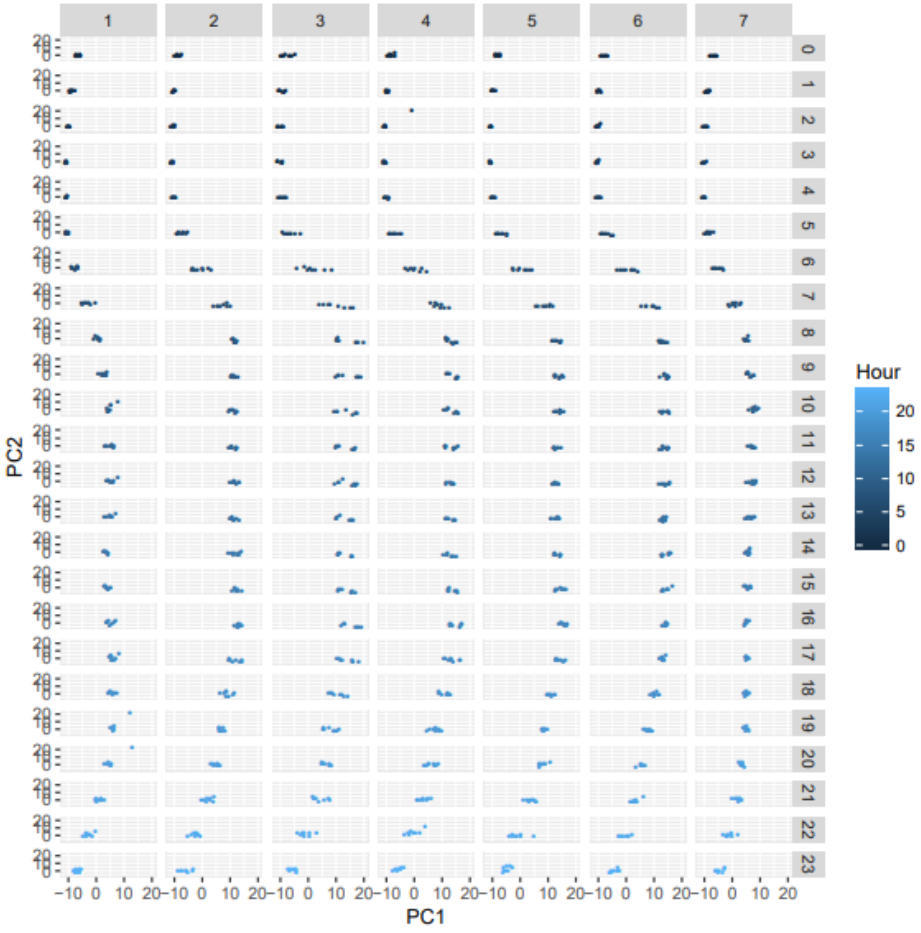


Рисунок 3.5 – Перша та друга головні компоненти на основі лише даних GGJF01, розподілених за день

Очевидно, що результати схожі з подібними кластерами і викидами, але не ідентичні. Наприклад, обидва результати показують викиди в перший день для годин 19 і 20, але це менш очевидно в оригінальних даних PCA. Оригінальні дані PCA не показують аномалії, очевидної на 2-й годині 4-го дня.

3.2 К-Найближчі сусіди

В роботі [6] представлено методи виявлення викидів обмежувалися статистичними підходами. Альтернативний підхід на основі відстані запропоновано в [6]. Метою було подолання обмежень статистичних підходів, які включали їхню нездатність підтримувати набори даних високої розмірності та необхідність мати розуміння характеру розподілу набору даних. В роботі [6] дається таке визначення викиду: "Об'єкт O у наборі даних T є викидом DB(p, D) (на основі відстані), якщо принаймні частка p об'єктів у T лежить на відстані, більшій за відстань D від O". Алгоритм виявлення викидів бере кожне спостереження і підраховує кількість сусідів на відстані D, якщо їх менше, ніж T - pT, то воно класифікується як викид. Це призводить до складності алгоритму порядку $O(kN^2)$. Двома недоліками цього підходу є те, що не передбачено ранжування викидів, а також необхідність для користувача визначати відстань D методом спроб і помилок. Ці недоліки були розглянуті в роботі [13].

Для цього було запропоновано обчислити відстань до k-го найвіддаленішого сусіда, а потім оголосити x точок з найбільшою відстанню викидами. Відстані викидів до k-го сусіда були використані для ранжування рівня кожного викиду. Альтернатива використанню відстані до k-го найближчого сусіда запропонована в [5], де оцінка викиду, яка називається "вагою", розраховується як середня відстань до k найближчих сусідів.

Ці підходи KNN ідентифікують глобальні аномалії, але не локальні. Вибір k впливає на результати, причому низькі значення можуть призвести до того, що оцінка щільності буде потенційно ненадійною, а високі значення можуть призвести до того, що оцінка щільності буде занадто грубою [4]. Згідно з [4], не існує формалізованого підходу для визначення найкращого значення k, тому досліджується і порівнюється діапазон значень від 10 до 50. "Сила цих алгоритмів полягає в тому, що вони за своєю суттю є неконтрольованими і мають інтуїтивно зрозумілі критерії для виявлення

викидів. Їхні обмеження включають квадратичну обчислювальну складність і можливу некоректність при обробці даних великої розмірності". [13]

Перший KNN-аналіз, проведений для набору даних SGW, був виконаний на перших 100 головних компонентах (що пояснюють 95% варіації в даних). Було використано як метод k-го найближчого сусіда, так і ваговий підхід. Було використано п'ять різних значень k: 10, 20, 30, 40 і 50. У таблиці 3.1 наведено порівняння 100 найбільших аномалій, виявлених кожним з десяти методів. Аномалії, виявлені за допомогою 10 різних підходів, є схожими, середній показник подібності становить 92,5%. Найбільша різниця між виявленими аномаліями спостерігається між підходом, що використовує вагу 10 найближчих сусідів, та підходом, що використовує вагу 50-го найближчого сусіда. Тут 82 з 100 аномалій були спільними для обох підходів. Кількість унікальних аномалій, виявлених усіма десятьма підходами, становила 124.

Таблиця 3.1 – Порівняння варіацій KNN, що показує відповідну кількість викидів на пару

	50th NN	40th NN	30th NN	20th NN	10th NN	Weight of 50	Weight of 40	Weight of 30	Weight of 20	Weight of 10
1	2	3	4	5	6	7	8	9	10	11
50th NN	100	99	98	95	84	95	94	92	87	82
40th NN	99	100	99	96	85	96	95	93	88	83
30th NN	98	99	100	97	86	97	96	94	89	84
20th NN	95	96	97	100	89	98	97	97	92	86
10th NN	84	85	86	89	100	89	90	92	94	90
Weight of 50	95	96	97	98	89	100	99	97	92	87
Weight of 40	94	95	96	97	90	99	100	98	93	88

Продовження таблиці 3.1

1	2	3	4	5	6	7	8	9	10	11
Weight of 30	92	93	94	97	92	97	98	100	95	89
Weight of 20	87	88	89	92	94	92	93	95	100	94
Weight of 10	82	83	84	86	90	87	88	89	94	100

Щоб перейти від виявлення контекстних аномалій до виявлення точкових аномалій, набір даних було розбито на підмножини. Було досліджено різні комбінації підмножин. Перші підмножини були засновані на вузлах. Тут значення KNN розраховувалися для кожного вузла окремо. Було використано кожен з десяти підходів KNN, і 100 найкращих результатів кожного з цих підходів дали 112 унікальних аномалій. Матриця, що показує кількість спільних аномалій, виявлених для кожної пари підходів KNN, наведена в таблиці 3.2. Аномалії, виявлені за допомогою 10 різних підходів, є більш схожими, ніж перший підхід: середня схожість становить 95,6%. Найбільша різниця між виявленими аномаліями була між підходом, що використовує вагу 10 найближчих сусідів, та підходом, що використовує вагу 50-го найближчого сусіда. Тут 89 зі 100 аномалій були спільними для обох підходів.

Таблиця 3.2 – Порівняння варіацій KNN, що показує відповідну кількість викидів на пару на основі групування вузлів

	50th NN	40th NN	30th NN	20th NN	10th NN	Weight of 50	Weight of 40	Weight of 30	Weight of 20	Weight of 10
1	2	3	4	5	6	7	8	9	10	11
50th NN	100	97	97	95	95	95	95	95	94	89
40th NN	97	100	99	97	97	96	96	96	95	90
30th NN	97	99	100	98	98	97	97	97	96	91

Продовження таблиці 3.2

1	2	3	4	5	6	7	8	9	10	11
20th NN	95	97	98	100	98	97	97	97	96	91
10th NN	95	97	98	98	100	98	98	99	98	93
Weight of 50	95	96	97	97	98	100	100	99	99	94
Weight of 40	95	96	97	97	98	100	100	99	99	94
Weight of 30	95	96	97	97	99	99	99	99	99	94
Weight of 20	94	95	96	96	98	99	99	99	100	95
Weight of 10	89	90	91	91	93	94	94	94	95	100

Підхід з використанням групування на основі вузлів дозволив виявити 4 аномалії, які не були ідентифіковані в першому підході, що розглядав всі точки даних разом. Таблиця 3.3 показує ці 4 нові спостереження, а також початковий та новий середній рейтинг за методом KNN.

Таблиця 3.3 – Спостереження, що потрапили до топ-100 за результатами групування на основі вузлів

Спостереження	Оригінальний KNN-рейтинг	KNN-рейтинг на основі вузлів
641	155,5	121,6
5007	136,5	111,6
5261	120,6	97,5
5462	134,3	119,2

Таблиця 3.4 показує середні рейтинги, отримані за допомогою підходу на основі вузлів та оригінального негрупового підходу, для всіх спостережень, для яких оригінальний підхід отримав перші 100 місць, а підхід на основі вузлів – ні.

Таблиця 3.4 – Спостереження, що не потрапили до 100 найкращих за результатами групування на основі вузлів

Спостереження	Оригінальний KNN-рейтинг	KNN-рейтинг на основі вузлів
1431	116,6	111,7
4238	150,3	170
4262	141,9	162,2
5272	94,5	112,3
5891	99,1	113
5952	108,6	131,3
6203	123,3	141,6
6371	114,9	125,9
8497	130,1	141,5
9666	118	122,7
11471	115,7	135,9
12723	116,4	135,6
12783	101,9	120,5
15242	113,8	118,7

Другі підмножини, що аналізувалися, базувалися на вузлі та часі доби. Вимір часу доби представляв собою 15-хвилинний інтервал, пов'язаний зі спостереженням. Тут значення KNN обчислювалися для кожного вузла та часу доби окремо. Кількість спостережень у кожній підмножині становила лише 14. З цієї причини кількість найближчих сусідів було змінено з 10, 20, 30, 40 і 50 на 2, 4, 6, 8 і 10. Було використано кожен з десяти підходів KNN, і 100 найкращих результатів кожного з цих підходів дали 104 унікальні аномалії.

3.3 Однокласова машина опорних векторів (OCSVM)

Машини опорних векторів (SVM) розроблені як бінарні класифікатори, що використовуються для класифікації даних на один-два можливих класи. В основі SVM-класифікатора лежить розділова гіперплощина. Відстань, на якій знаходиться спостереження від гіперплощини, дає уявлення про достовірність отриманої класифікації. Чим далі спостереження від гіперплощини, тим більша ймовірність того, що класифікація є правильною.

Перші дослідження, що заклали основу для методу опорних векторів (SVM), були проведені Вапніком та Лернером у роботі [14], де було запропоновано концепцію класифікатора з максимальним зазором. Цей метод визначає гіперплощину, яка розділяє два лінійно роздільні класи з максимально можливим зазором. Оскільки положення роздільної гіперплощини залежить від окремих спостережень, існує ризик перенавчання.

У роботі [15] було запроваджено концепцію опорно-векторної мережі, яка усуває вимогу до ідеальної лінійної роздільності класів. Це стало можливим завдяки введенню поняття "м'якого" зазору, що дозволяє деяким навчальним спостереженням знаходитись на неправильній стороні зазору або навіть по інший бік гіперплощини. Для таких спостережень вводяться змінні ослаблення, а ступінь їхнього впливу регулюється спеціальним параметром. Чим вища значення цього параметра, тим більше модель толерантна до порушень зазору.

У роботі [16] було представлено власне метод опорних векторів, який розширив можливості підходу на нелінійні випадки. Це досягається шляхом розширення простору ознак за рахунок введення нелінійних компонентів з використанням ядерних функцій, що забезпечує обчислювальну ефективність методу.

Ці три класифікатори належать до методів навчання з учителем. У дослідженні [17] було запропоновано однокласовий метод опорних векторів

(One-Class SVM), який переносить цей підхід у область навчання без учителя. Тут точки даних відокремлюються від початку координат у просторі ознак за допомогою гіперплощини. Параметр регулювання ν визначає "м'який" зазор, що дозволяє коректно обробляти аномалії. Він задає верхню межу частки викидів та нижню межу відсотка опорних векторів. Кожній точці присвоюється значення рішення: від'ємні значення вказують на спостереження з боку початку координат, а додатні – на інші. Величина значення пропорційна відстані до гіперплощини, тому викидам присвоюються значні від'ємні значення.

У роботі [18] було запропоновано альтернативний підхід One-Class SVM, де замість гіперплощини використовується гіперсфера. Однак, як зазначено в [13], One-Class SVM чутливий до викидів, які можуть зміщувати межу прийняття рішень. Для зменшення цього впливу в [13] запроваджено два методи:

- надійний One-Class SVM: дозволяє спостереженням, далеким від центроїда даних, мати великі значення ослаблення, зміщуючи межу рішень у бік "нормальних" точок;

- η One-Class SVM: вводиться змінна η, яка оцінює "нормальність" точки. Значення $\eta = 0$ позначає викиди, які виключаються з процесу навчання, щоб межа формувалася лише на основі нормальних точок.

У цьому дослідженні використовується One-Class SVM з [17]. Його гіперпараметри:

- ν (як описано вище);

- ядро (з власними параметрами, напр., γ для радіального базисного ядра).

Реалізація здійснена через інтерфейс до libsvm у пакеті R e1071. Оскільки дані не мітили, пошук гіперпараметрів проводився не через точність класифікації, а через аналіз розподілу значень рішень. Досліджувалися комбінації з:

- радіальним ядром ($\gamma \in \{0.01, 0.05, 0.1, 0.5, 1\}$);

- $\nu \in \{0.01, 0.05, 0.1, 0.5, 1\}$.

Найкращий результат (рис. 3.6а) досягнуто при $\nu = 0.1$ та $\gamma = 0.01$, де більшість спостережень мали значення > 0, а викиди — значні від'ємні значення. Процес включав три ітерації підбору параметрів.

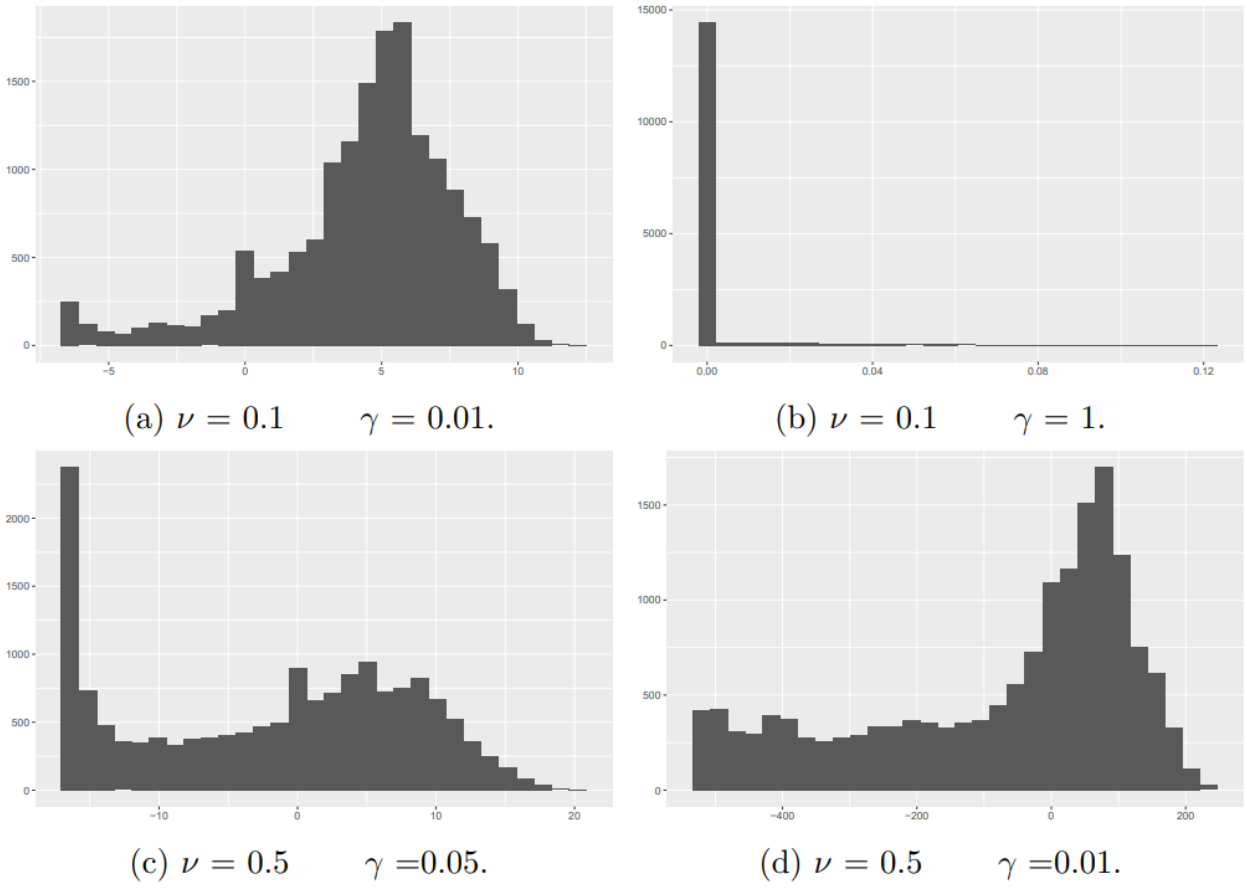


Рисунок 3.6 – Розподіл результатів першого пошуку за гіперпараметром

Другий набір гіперпараметрів, які досліджувалися, продовжував використовувати ядро у вигляді радіальної базисної функції з $\gamma \in \{0.001, 0.005, 0.01, 0.02\}$ у всіх комбінаціях з $\nu \in \{0.07, 0.1, 0.15\}$. Найбільш перспективні результати представлені на рисунку 2.16. Найкраще критерії відбору задовольняли результати, отримані з $\nu \in \{0.1, 0.15\}$ та $\gamma = 0.001$.

Остання ітерація дослідження передбачала використання ядра у вигляді радіальної базисної функції з $\gamma \in \{0.0007, 0.002, 0.0055\}$ у всіх комбінаціях з $\nu \in \{0.13, 0.18, 0.2\}$. Найкращі результати представлені на рисунку 3.7.

Найбільш відповідними критеріям вибору виявилися результати з ν ∈ {0.18, 0.13} та γ = 0.002 (підрисунки a та c). Вони демонстрували чітку групу викидів на найбільш негативному кінці розподілу. Топ-100 викидів, виявлених цими двома підходами, майже повністю збігалися – 98 із них були спільними.

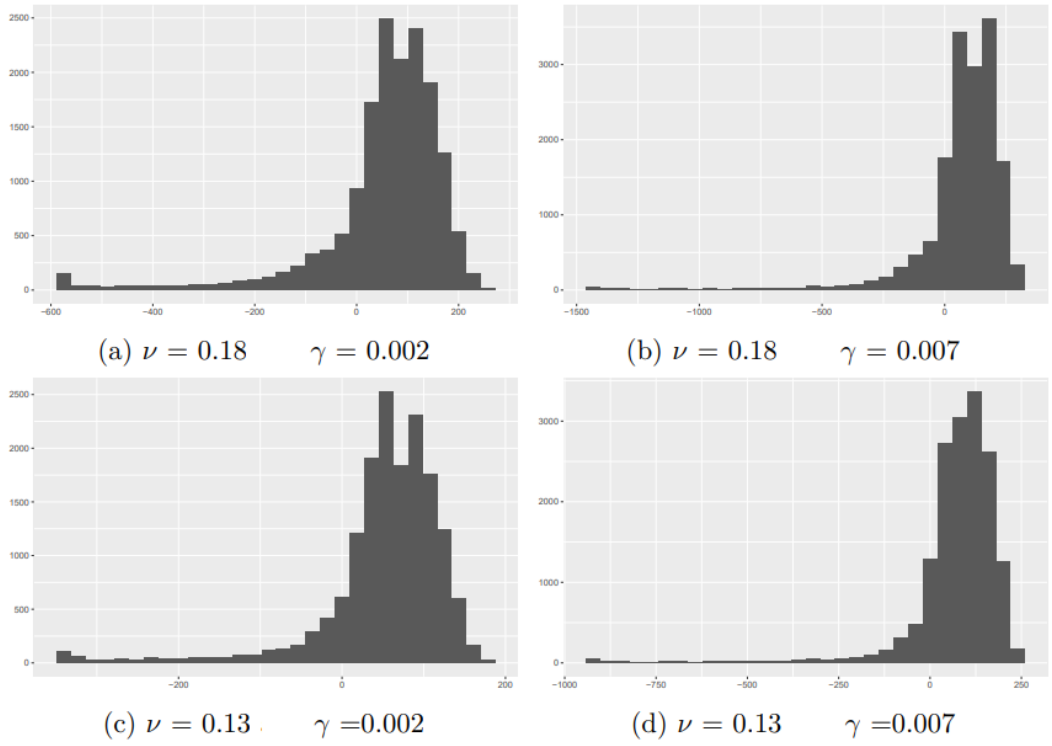


Рисунок 3.7 – Результати третього підбору параметрів

Як і у випадку з методом K-найближчих сусідів, аномалії необхідно розглядати в контексті. Для переходу від контекстного до точкового виявлення аномалій набір даних було розділено на підмножини. Перші підмножини базувалися на вузлах. Для кожного вузла окремо було розраховано значення рішень SVM. Використовувалися ті самі гіперпараметри, що й у останній ітерації: γ ∈ {0.0007, 0.002, 0.0055} у всіх комбінаціях з ν ∈ {0.13, 0.18, 0.2}. Гістограми значень рішень SVM для кожної комбінації показані на рисунку 3.8 (вісь y масштабована за допомогою квадратного кореня).

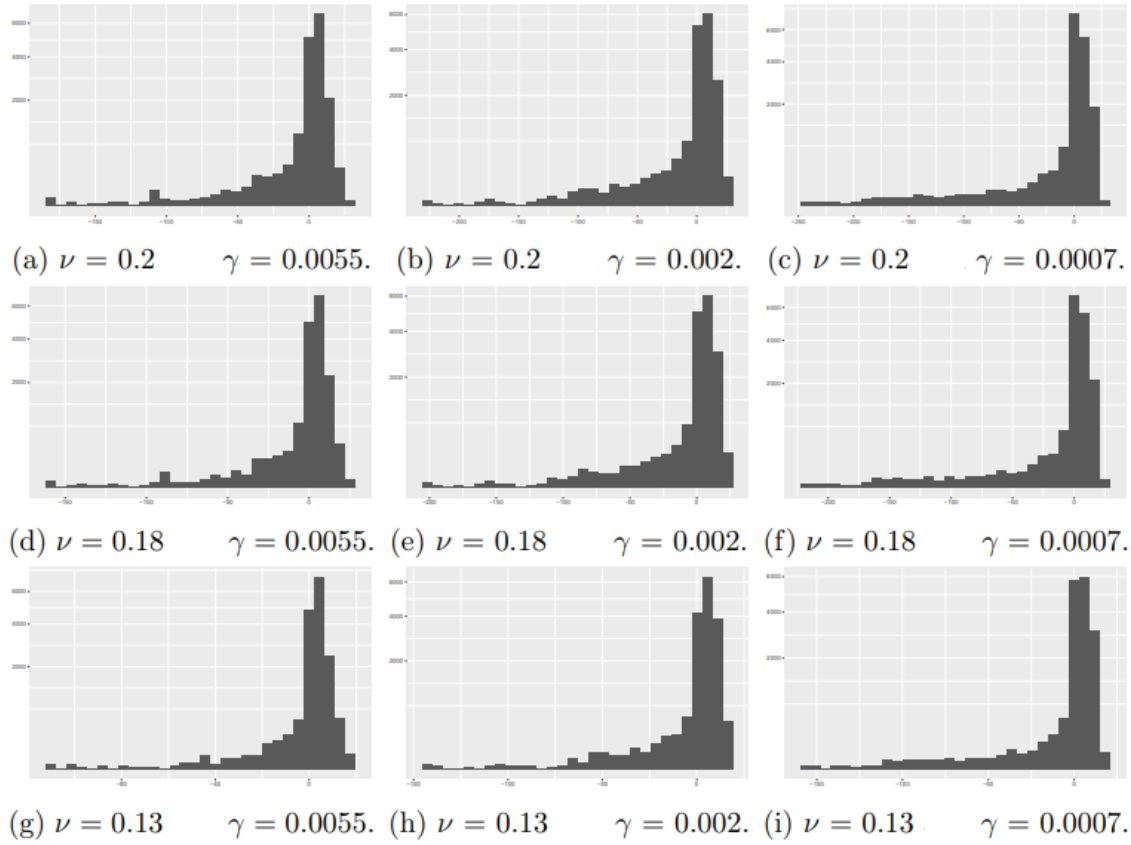


Рисунок 3.8 – Отримані розподіли при першому пошуку на основі вузлів

Комбінація γ = 0.0007 та ν = 0.2 забезпечила найкраще відокремлення викидів від інших спостережень. Як видно з таблиці 2.15, ця комбінація також дала 75 спільних викидів із топ-100 у порівнянні з аналізом для всієї групи (ν = 0.13, γ = 0.002).

Таблиця 3.4 – Кількість спільних викидів з оригінальними результатами OCSVM (без розбиття на підмножини)

	Збіги у топ100	Збіги у топ 200	Збіги у топ300
ν = 0.13 , γ = 0.0007	68	167	241
ν = 0.13 , γ = 0.002	46	120	164
ν = 0.13 , γ = 0.0055	9	38	108
ν = 0.18 , γ = 0.0007	73	174	248
ν = 0.18 , γ = 0.002	48	120	164
ν = 0.18 , γ = 0.0055	9	38	91
ν = 0.2 , γ = 0.0007	75	176	248

Було перевірено додаткові значення гіперпараметрів на основі найперспективніших результатів. Досліджувалися $\gamma \in \{0.0001, 0.0002, 0.0003\}$ у всіх комбінаціях з $\nu \in \{0.3, 0.4, 0.5\}$. Гістограми значень рішень OCSVM для цих комбінацій показані на рисунку 3.9 (вісь y масштабована за допомогою квадратного кореня). Це призвело до ще кращого відокремлення викидів порівняно з результатами на рисунку 3.8.

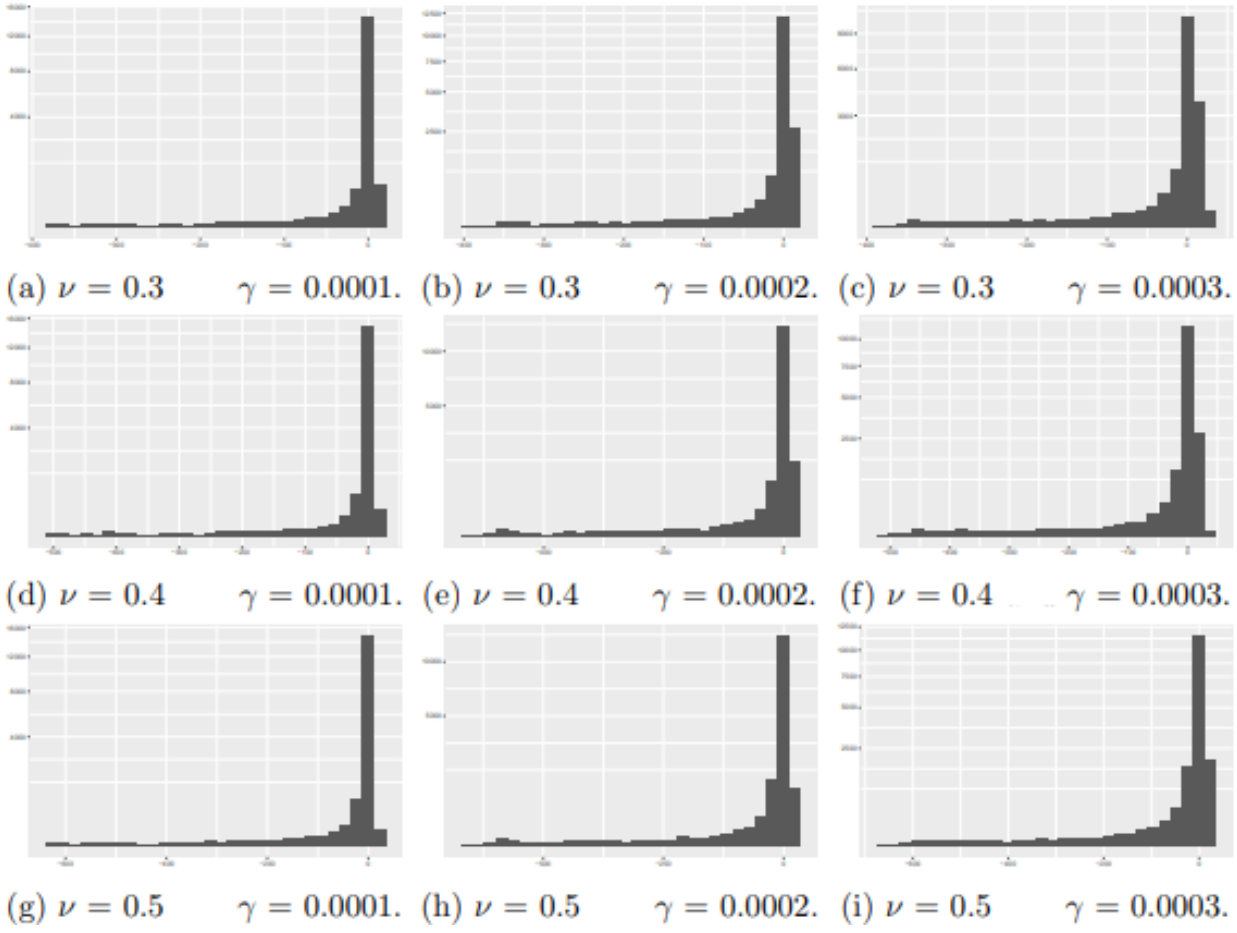


Рисунок 3.9 – Результати другого пошуку за вузлам

Ці гіперпараметри забезпечили дуже схожий топ викидів із результатами аналізу OCSVM для не розділених даних ($\nu = 0.13$, $\gamma = 0.002$). Порівняння наведено в таблиці 3.5.

Таблиця 3.5 – Кількість спільних викидів у топі за результатами другого пошуку порівняно з базовою OCSVM-моделлю (без розбиття)

Параметри	Спільні для топ-100	Спільні для топ-200	Спільні для топ-300
$\nu = 0.3, \gamma = 0.0001$	92	189	271
$\nu = 0.3, \gamma = 0.0002$	91	189	270
$\nu = 0.3, \gamma = 0.0003$	90	186	265
$\nu = 0.4, \gamma = 0.0001$	92	190	274
$\nu = 0.4, \gamma = 0.0002$	92	189	272
$\nu = 0.4, \gamma = 0.0003$	91	187	268
$\nu = 0.5, \gamma = 0.0001$	92	190	277
$\nu = 0.5, \gamma = 0.0001$	92	190	274
$\nu = 0.5, \gamma = 0.0003$	91	187	272

3.4 Локальний фактор відхилення на основі щільності (LOF)

Наступний спосіб виявлення викидів було запропоновано у [7]. Він має дві основні цілі: знаходити локальні викиди (точки, що є аномальними у своєму локальному оточенні) та призначати їм значення (Фактор локальних викидів, LOF), яке показує ступінь їх аномальності.

Є два типи викидів: глобальний (позначено "x") та локальний (позначено "o"). Метод KNN виявляє лише глобальні викиди. Не існує такого значення k, при якому KNN знайде локальний викид "o", не помилково не позначивши точки з рідкісного кластера також як викиди. Метод LOF спеціально розроблений для виявлення таких точок, як "o".

Ступінь аномальності точки визначається її LOF:

- чим ближче LOF до 1, тим нормальніша точка;
- чим вище LOF, тим більш аномальною є точка.

Визначення LOF починається з поняття k-відстані (k-distance(p)) – відстані до k-го найближчого сусіда точки p. Оточення точки p (Nk-distance(p)(p)) включає всі точки, які знаходяться не далі, ніж її k-й сусід.

Далі визначається досяжна відстань точки p: вона дорівнює або реальній відстані від o, якщо p знаходиться далі, ніж k-відстань o, або самій k-відстані o, якщо p ближче. Це дозволяє зменшити статистичні коливання для точок поблизу o.

Разом із параметром MinPts (кількість найближчих сусідів для визначення локального оточення) ці поняття використовуються для визначення локальної густини досяжності точки p (рівняння 3.1). Нарешті, рівняння 3.2 визначає LOF точки p.

$$lrd_{MinPts} = \frac{|N_{MinPts}(p)|}{\sum_{o \in N_{MinPts}(p)} reach - dist_{MinPts}(p,o)}, \tag{3.1}$$

$$LOF_{MinPts} = \frac{\sum_{o \in N_{MinPts}(p)} \frac{lrd_{MinPts}(o)}{lrd_{MinPts}(p)}}{|N_{MinPts}(p)|}, \tag{3.2}$$

щільність локальної досяжності об'єкта p та

$$reach - dist(p,o) = \max\{k - distance(o), d(p,o)\}. \tag{3.3}$$

Досяжна відстань об'єкта p від об'єкта o.

Для обчислення значень LOF (фактор локальних викидів) для нашого набору даних SGW було використано функцію `lof` з пакету DBSCAN в R. Спочатку ми працювали з усіма даними разом, без розбиття на підмножини. Було перевірено діапазон значень k (кількість сусідів) від 5 до 250 з кроком 5. На рисунку 3.10 показано розподіли для 9 обраних значень k. Оскільки дані не мітили міток, вибір k не міг ґрунтуватися на точності класифікації. Як і

для методу SVM, ми вибирали оптимальні параметри, орієнтуючись на розподіли, де викиди найчіткіше відокремлювалися від основної маси даних. Усі перевірені значення k дали схожі результати без суттєвих відмінностей. Для виявлення відмінностей ми порівняли топ-100 викидів для кожного значення k з результатами, отриманими при інших значеннях k.

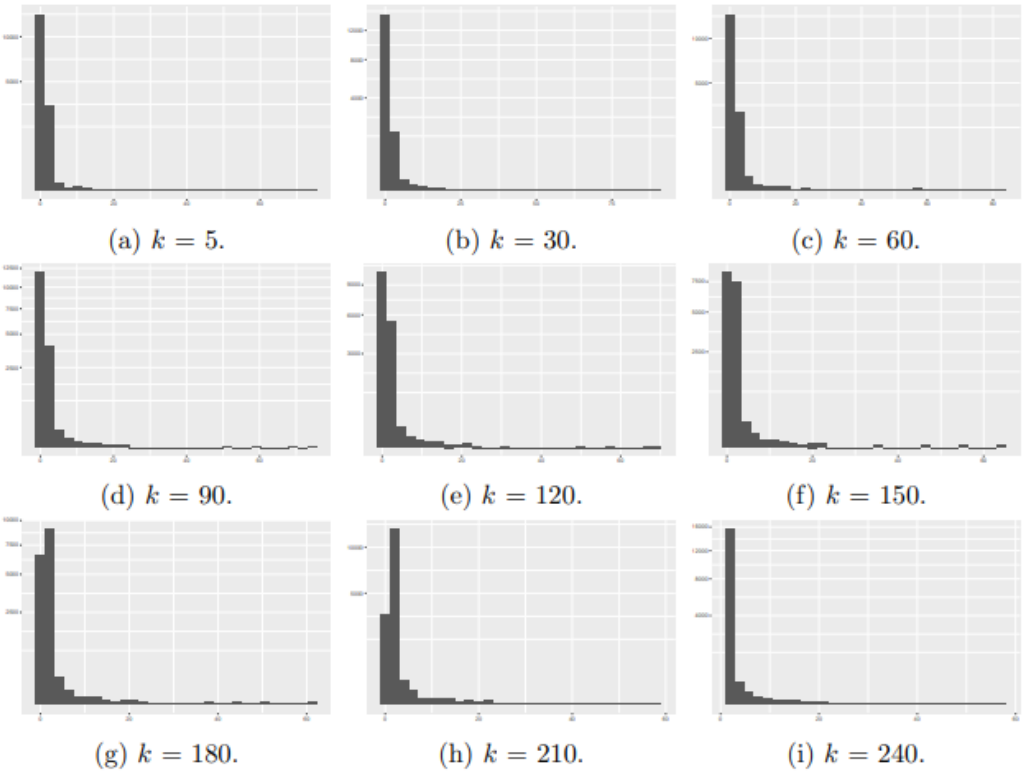


Рисунок 3.10 – Результуючі розподіли значень LOF

Як і в методах k-найближчих сусідів та однокласового SVM, аномалії також потребують розгляду в контексті. Для переходу від контекстного до точкового виявлення аномалій набір даних розділили на підмножини за вузлами. Для кожного вузла окремо обчислили значення LOF.

Використовували той самий набір значень k (від 5 до 250 з кроком 5), що й для аналізу без розбиття. Як і в тому підході, отримані розподіли не дали можливості вибрати оптимальне k, оскільки всі вони були дуже схожими. Деякі з цих розподілів показані на рисунку 3.11.

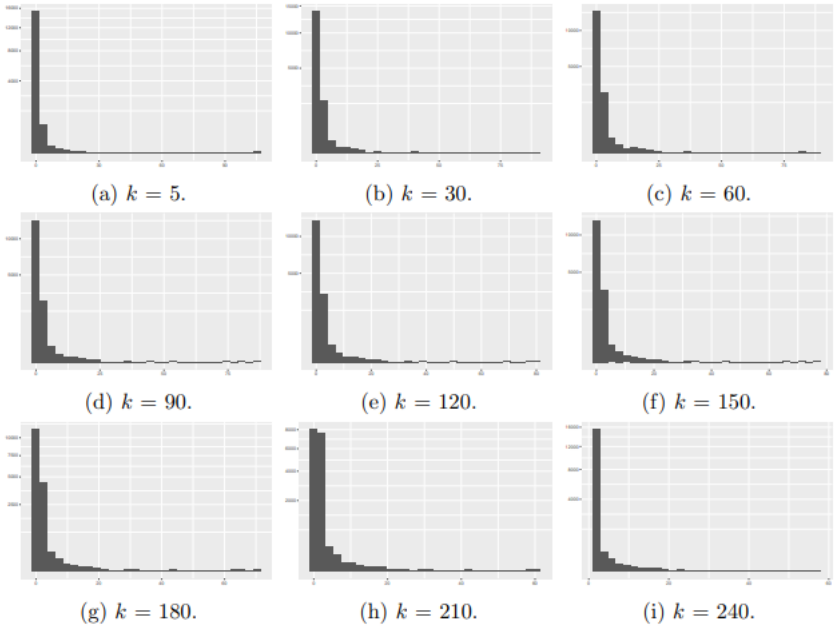


Рисунок 3.11 – Результуючі розподіли LOF після групування за вузлами

Для підходу, що базується на вузлах/часі, не використовували метод LOF, оскільки у кожній підмножині було лише 12 спостережень, а мінімально рекомендоване значення k для LOF становить 10.

3.5 Багатовимірний гаусівський розподіл

Останній підхід використовує параметричну модель — багатовимірний нормальний розподіл для виявлення аномалій у наборі даних SGW. Багатовимірний нормальний розподіл описується функцією щільності ймовірності, визначеною у рівнянні 3.4.:

$$p(x;\mu,\Sigma) = (2\pi)^{-\frac{k}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right). \quad (3.4)$$

Тут середнє значення кожної змінної представлено у векторі μ, а k позначає кількість змінних. Σ визначає матрицю коваріацій спостережень, яка відображає кореляції між ознаками.

Обчислення Σ вимагає, щоб кількість спостережень перевищувала кількість ознак і щоб не було дубльованих ознак, інакше матриця Σ не буде оберненою, і рівняння 3.4 не можна буде обчислити. Після того як функція щільності ймовірності обчислена, можна визначити ймовірність кожного спостереження для кожної точки в наборі даних. Спостереження з найнижчими ймовірностями позначаються як аномальні.

Як і в інших підходах, описаних вище, спочатку аналізувався весь набір даних, а потім його було розділено на піднабори для кожного вузла, і кожен з них аналізувався окремо. Із 100 найбільш аномальних спостережень, виявлених цими двома підходами, лише 42 збігалися. Піднабори за вузлом/часом не аналізувалися, оскільки така група містила лише 13 спостережень. Оскільки кількість змінних перевищує кількість спостережень, матриця Σ не є оберненою, і функцію щільності ймовірності неможливо було обчислити.

3.6 Створення набору даних без аномалій

У цьому розділі описано, як інформація про аномалії, отримана за допомогою 10 різних методів виявлення аномалій, описаних вище, використовується для видалення аномалій із початкового сирого набору даних. Метою є створення набору даних, який є максимально вільним від аномалій, для подальшого напівконтрольованого навчання. Напівконтрольований підхід до навчання передбачає використання автоенкодера, який після навчання на даних без аномалій зможе відтворювати нормальні спостереження, тоді як ненормальні спостереження він позначатиме як аномальні. На рисунку 3.12 ця схема підготовки даних зображена графічно.

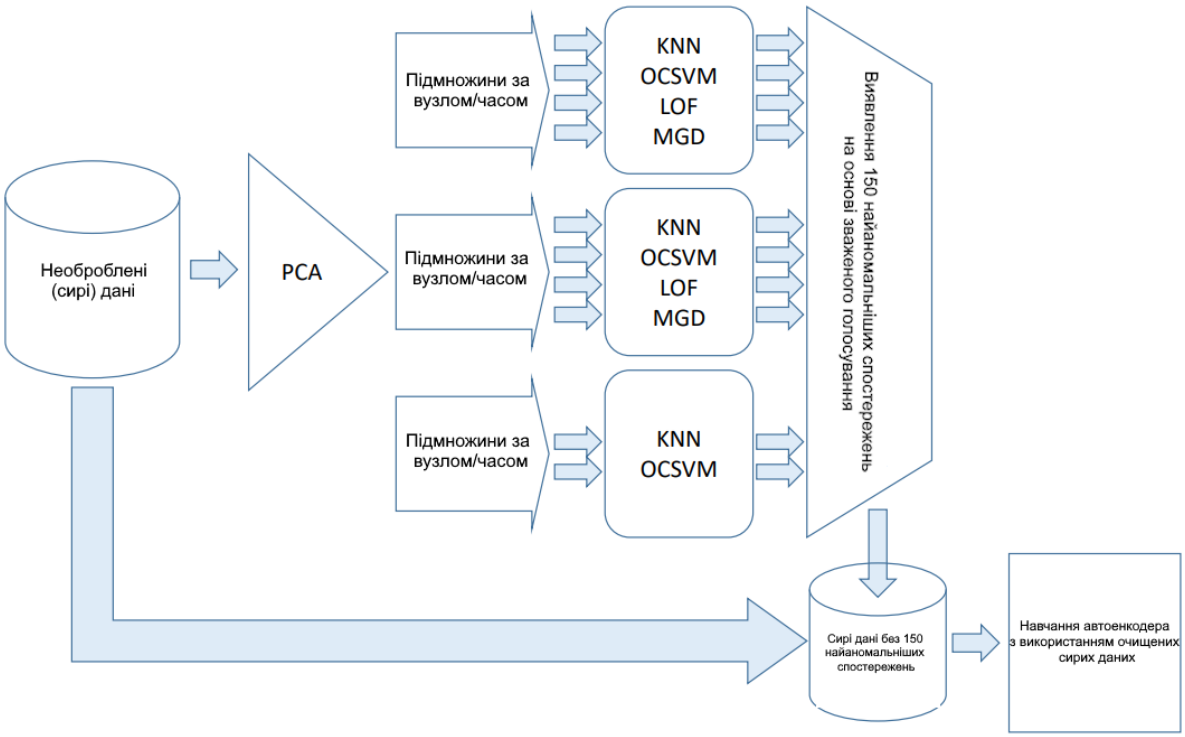
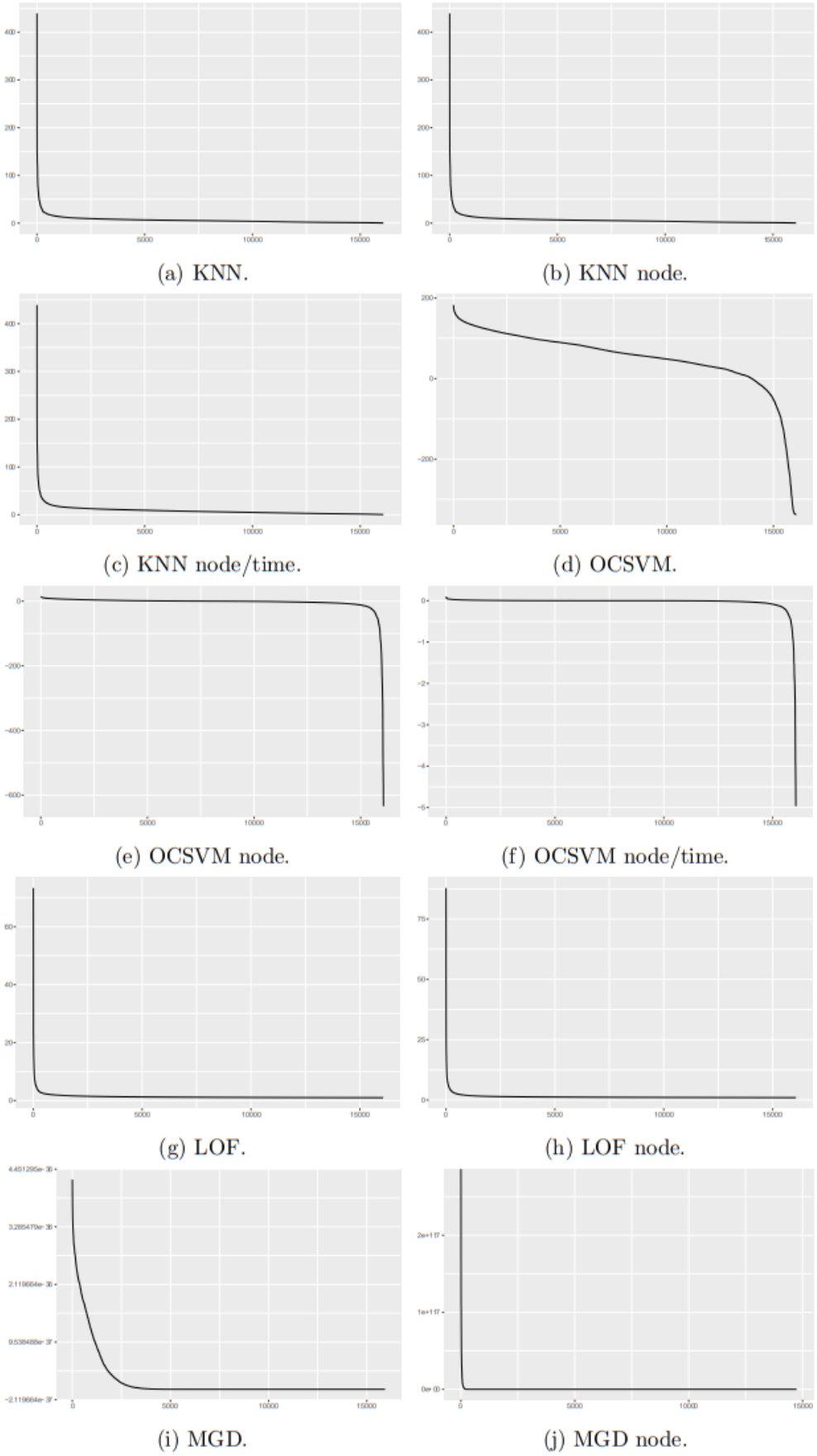


Рисунок 3.12 – Процес підготовки даних, видалення аномалій для навчання автокодера

Першим кроком є визначення кількості аномальних спостережень, які потрібно вилучити з набору даних. Потрібно знайти баланс між видаленням занадто великої кількості спостережень з ризиком видалити також і нормальні дані та видаленням занадто малої кількості, залишаючи при цьому аномалії. Обраний підхід полягав у сортуванні спостережень на основі результатів кожного з 10 методів. Ці значення аномалій були нанесені на графік, як показано на рисунку 3.13. KNN, KNN по вузлу, KNN по вузлу/часу, OCSVM по вузлу, OCSVM по вузлу/часу, LOF і LOF по вузлу – усі продемонстрували чітку точку в області близько 150 спостережень. OCSVM і MGD були менш однозначними, при цьому MGD показав точку на рівні 2500. На основі цього аналізу було вирішено вважати 150 точок аномальними.



(a) KNN. (b) KNN node.

(c) KNN node/time. (d) OCSVM.

(e) OCSVM node. (f) OCSVM node/time.

(g) LOF. (h) LOF node.

(i) MGD. (j) MGD node.

Рисунок 3.13 – Впорядковані значення факторів аномальності за кожним підходом

Наступним кроком було визначити, які 150 спостережень слід видалити. Першим етапом у цьому процесі було створення десяти окремих списків спостережень, по одному для кожного з використаних підходів. Кожен список був упорядкований від найбільш аномальних до найменш аномальних відповідно до десяти різних методів. Далі кожному спостереженню в кожному зі списків було присвоєно ранг – 1 отримувало найбільш аномальне спостереження, а 16070 – найменш аномальне. Потім 10 рангів для кожного спостереження сумувалися, щоб отримати підсумковий рейтинг. Таким чином, спостереження, які отримали низькі ранги у всіх десяти методах, мали низький загальний підсумковий рейтинг. 150 спостережень із найнижчими підсумковими рангами були вилучені з початкового сирого набору даних. У таблиці 3.6 наведено відсоток спостережень із топ-150 кожного методу, які увійшли до остаточного списку аномалій.

Метод, який виявив найбільшу кількість спостережень, що увійшли до остаточного списку аномалій, – це MGD (98,7%), а найменше – MGD node (50,7%). Три підходи KNN разом виявили найбільшу кількість аномалій. Отриманий набір даних став основою для навчання фінальної моделі виявлення аномалій. Після вилучення аномальних спостережень із початкового набору даних були додані інженерні ознаки, необхідні для контекстуалізації аномалій.

Остаточний набір даних складався з 28011 спостережень із 1001 змінною. Ці змінні включали SGW, який згенерував спостереження, регіон розташування SGW, часову позначку (дата/час), годину доби, час доби, хвилину доби, день тижня, а також 996 лічильників SGW.

Таблиця 3.6 – Кількість спостережень із топ-150 кожного методу, які увійшли до остаточного списку аномалій

Метод	Відсоток у фінальній вибірці
KNN.Weight50	0,847
KNN.Node.Weight50	0,86
KNN.Node.Time.Weight8	0,8
OCSVM.ν=0.13, γ=0.002	0,818
OCSVM.Node.ν=0.5, γ=1e-04	0,8
OCSVM.Node.Time.ν=0.5, γ=1e-04	0,8
LOF.K100	0,687
LOF.Node.K100	0,647
MGD.Node	0,507
MGD	0,987

3.7 Автокодер для виявлення аномалій

Наступним кроком у створенні системи виявлення аномалій є використання набору даних без аномалій, створення якого описано вище, для створення напівкерованого автокодера.

Набір даних було розподілено на тестовий, валідаційний та тренувальний набори, кожен з яких складав 10%, 10% та 80% від початкового набору даних відповідно. Після однієї епохи середньоквадратична помилка (Mean Square Error) автоенкодера на валідаційному та тренувальному наборах знизилась до 0,001, а після 4,8 епох обидва значення дещо зменшилися до 0,0008. 665 змінних не мали значущості, що було очікувано через велику кількість змінних, які містили лише нулі. Решта 493 змінні мали вузький діапазон процентного значення важливості, при цьому найважливішою змінною була «Time.4:30» із відсотком 0,2393%, а найменш важливою – лічильник G19M4C8 із важливістю 0,1749%. Ці відсотки є дуже низькими, що свідчить про те, що

всі змінні вносять вклад у модель. У середньому групою змінних із найнижчою середньою важливістю були лічильники, а змінна, що містить час доби, мала найвищу важливість. Середні відсоткові значення для різних груп наведені в таблиці 3.7.

Таблиця 3.7 – Середній відсоток важливості змінних за групами змінних

Група змінних	Середня важливість змінної (%)
SGW ID	22.5
Час доби	22.3
Година доби	21.8
День тижня	21.2
Регіон	20.5
Хвилина доби	19.8
Лічильник	19.4

Процес оцінки його ефективності полягав у пропусканні нового сирого набору даних, отриманого з мобільної мережі, через автоенкодер. Вихідні аномалії ідентифікувалися за високими помилками відтворення (reconstruction errors). Ці аномалії порівнювалися з «ground truth» — реальними подіями, що сталися в мобільній мережі. Для контекстуального аналізу відзначаються зміни в мережі, що відбулися протягом 5 місяців між першим та цим екстрактом даних, а саме: по-перше, загальний обсяг трафіку, який підтримують SGW, зріс на 27%. По-друге, старі SGW були замінені новими, доданими до мережі.

Новий набір даних було пропущено через автоенкодер, після чого було проаналізовано помилку відтворення (MSE) для кожного спостереження. Середнє значення MSE для всіх спостережень становило 0,032, при цьому найвищим було 96,57978, а найнижчим – 0,000154722.

ВИСНОВКИ

В кваліфікаційній роботі було представлено підхід для виявлення аномалій у високовимірних даних SGW у мобільному мережевому середовищі. Цей підхід доповнює існуючу модельну методику, яка виявляє деградації у заздалегідь визначених ключових показниках ефективності (KPI). Підхід виявлення аномалій заповнює «сліпу пляму», що залишилась у поточній методиці, охоплюючи всі лічильники. Було продемонстровано здатність цього підходу виявляти аномалії, що впливають на багатьох абонентів і лічильників,

Аномалії, виділені цим підходом, лише вказують на зміни від цієї базової лінії. Зміни можуть відображати як погіршення, так і покращення роботи мережі. Для визначення цього та з’ясування деталей змін потрібне подальше дослідження мобільної мережі. На жаль, чорний ящик кінцевого автоенкодера не надає користувачу допомоги у таких дослідженнях.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Peña, Daniel. "Detecting outliers and influential and sensitive observations in linear regression." *Springer Handbook of Engineering Statistics* (2023): 605-619.

2. Čampulová, Martina, Roman Čampula, and Jan Holešovský. "An R package for identification of outliers in environmental time series data." *Environmental Modelling & Software* 155 (2022): 105435.

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

4. ANGIULLI, Fabrizio; PIZZUTI, Clara. Fast outlier detection in high dimensional spaces. In: *European conference on principles of data mining and knowledge discovery*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2002. p. 15-27.

5. Li, K., Gao, X., Jia, X., Xue, B., Fu, S., Liu, Z., ... & Huang, Z. (2022). Detection of local and clustered outliers based on the density–distance decision graph. *Engineering Applications of Artificial Intelligence*, *110*, 104719.

6. Bletskan D. I., Glukhov K. E., Frolova V. V. Electronic structure of 2H-SnSe2: ab initio modeling and comparison with experiment. Semiconductor Physics Quantum Electronics & Optoelectronics. 2016. Vol. 19, No 1. P. 98–108.

7. Corain, M., Garza, P., & Asudeh, A. (2021, April). Dbscout: A density-based method for scalable outlier detection in very large datasets. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (pp. 37-48). IEEE.

8. Zelalem Jembre, Yalew, et al. "Mobile broadband performance evaluation: analysis of national reports." *Electronics* 11.3 (2022): 485.

9. Granà, Fabrizio, and Libero Mario Mari. "The Case of Vodacom Group." *Integrated Reporting: Concepts and Cases that Redefine Corporate Accountability*. Cham: Springer International Publishing, 2013. 237-253.

10. 3GPP. "Technical specification Group services and system Aspects." *Release* 15 (2018).

11. Nováczki, S. (2013, March). An improved anomaly detection and diagnosis framework for mobile network operators. In *2013 9th international conference on the design of reliable communication networks (drcn)* (pp. 234-241). IEEE.

12. Szilágyi, P., & Nováczki, S. (2012). An automatic detection and diagnosis framework for mobile communication systems. *IEEE transactions on Network and Service Management*, *9*(2), 184-197.

13. Babaie, Tahereh, et al. "A unified approach to network anomaly detection." *2014 IEEE International Conference on Big Data (Big Data)*. IEEE, 2014.

14. Ali, W. A., Manasa, K. N., Bendechache, M., Fadhel Aljunaid, M., & Sandhya, P. (2020). A review of current machine learning approaches for anomaly detection in network traffic. *Journal of Telecommunications and the Digital Economy*, *8*(4), 64-95.

15. Zuo, Yuan, et al. "An intelligent anomaly detection scheme for micro-services architectures with temporal and spatial data analysis." *IEEE Transactions on Cognitive Communications and Networking* 6.2 (2020): 548-561.

16. Burgueño, J., de-la-Bandera, I., Mendoza, J., Palacios, D., Morillas, C., & Barco, R. (2020). Online anomaly detection system for mobile networks. *Sensors*, *20*(24), 7232.

17. Wagner, Cynthia, et al. "Machine learning approach for ip-flow record anomaly detection." *NETWORKING 2011: 10th International IFIP TC 6 Networking Conference, Valencia, Spain, May 9-13, 2011, Proceedings, Part I 10*. Springer Berlin Heidelberg, 2011.

18. МАРТОВИЦЬКИЙ, В., СВИРИДОВ, А., АВДЄЄВ, О., ГУДЗИНСЬКИЙ, І., & КОРОТЕЦЬКИЙ, О. ДОСЛІДЖЕННЯ МЕТОДІВ ВИЯВЛЕННЯ АНОМАЛІЙ У API ЖУРНАЛАХ. Вісник Херсонського національного технічного університету, 2(1 (92)), С. 142-148.