

МАГІСТЕРСЬКА РОБОТА

МР. ІІМ - 40.00.00.000 ПЗ

Група ІІМ-24-2

Николайчук Ярослав

2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Николайчук Ярослав Анатолійович

(прізвище, ім'я, по батькові)

УДК 004.9
(індекс)

МАГІСТЕРСЬКА РОБОТА

Моделі та методи побудови фреймворків генерації специфікованих

візуалізацій

(назва роботи)

Інженерія програмного забезпечення

(назва освітньої програми)

121 - Інженерія програмного забезпечення

(шифр і назва спеціальності)

Николайчук Я.А.

(підпис, ініціали та прізвище здобувача освітнього ступеня)

Науковий керівник

Яцишин Микола Миколайович, к.т.н., доцент

(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Допущено до захисту

Завідувач кафедри

доц.

Бандура В.В.

(посада) (підпис) (дата) (ініціали та прізвище)

Нормоконтроль

доц.

Вовк Р.Б.

(посада) (підпис) (дата) (ініціали та прізвище)

Робота містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело

Івано-Франківськ – 2025

Івано-Франківський національний технічний університет нафти і газу

Факультет інформаційних технологій

Кафедра інженерії програмного забезпечення

Освітній рівень магістр

Спеціальність 121 – Інженерія програмного забезпечення

ЗАТВЕРДЖУЮ:

Зав. кафедрою

ІПЗ

доц.

В.В. Бандура

“ 04 ” вересня 2025 р.

ЗАВДАННЯ

НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Николайчуку Ярославу Анатолійовичу

(прізвище, ім'я, по-батькові)

1. Тема магістерської роботи “ Моделі та методи побудови фреймворків генерації специфікованих візуалізацій ”

керівник проекту (роботи) Яцишин М.М., к.т.н., доцент

затверджені наказом закладу вищої освіти від “ 05 ” листопада 2025 р. № 695/7

2. Строк подання студентом проекту (роботи) 15 грудня 2025 р.

3. Вихідні дані до проекту (роботи) Концепції, формальні моделі і методи побудови інформаційних та програмних технологій генерації візуалізацій

4. Зміст розрахунково - пояснювальної записки(перелік питань, які потрібно розробити)

1. Аналіз предметної області візуалізації даних та процесів генерації специфікованих візуалізацій

2. Концептуальна модель процесу візуалізації даних

3. Дослідження моделей та методів генерації специфікованих візуалізацій на основі сприйняття

4. Імплементация моделей побудови фреймворку генерації специфікованих візуалізацій

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслень)

1. Вісім представлень подібного набору даних з різною кількістю кластерів (рис. 1.1)

2. Приклад згладжування ліній для зашумлених даних (рис. 1.2)

3. Відтворення типів візуальних кодувань (рис. 1.3)

4. Відтворення рейтингу візуальних кодувань (рис. 1.4)

5. Процес візуалізації даних (рис. 1.6)

6. Консультанти розділів проекту (роботи)

Розділ	Консультант	Підпис, дата
Перевірка на плагіат	доц., к.т.н. Вовк Р.Б.	

7. Дата видачі завдання 04 вересня 2025 р.

Керівник

_____ (підпис)

Завдання прийняв до виконання

_____ (підпис)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Назви етапів магістерської роботи	Строк виконання етапів роботи	Примітка
1	Підбір і вивчення літератури по темі магістерської роботи	15.09.2025	виконано
2	Аналіз предметної області візуалізації даних та процесів генерації специфікованих візуалізацій	01.10.2025	виконано
3	Концептуальна модель процесу візуалізації даних	17.10.2025	виконано
4	Дослідження моделей та методів генерації специфікованих візуалізацій на основі сприйняття	02.11.2025	виконано
5	Імплементация моделей побудови фреймворку генерації специфікованих візуалізацій	19.11.2025	виконано
6	Аналіз обчислювальної ефективності моделі оптимізації	02.12.2025	виконано
7	Затвердження пояснювальної записки роботи завідувачем кафедри	15.12.2025	виконано

Студент – магістр

_____ (підпис)

Керівник роботи

_____ (підпис)

АНОТАЦІЯ

Магістерська робота: 79 с., 19 рис., 3 табл., 39 джерел.

Тема: Моделі та методи побудови фреймворків генерації специфікованих візуалізацій

Мета роботи: розробка моделей та методів побудови фреймворку генерації специфікованих візуалізацій, що ґрунтуються на когнітивних механізмах сприйняття, властивостях даних та аналітичних завданнях користувача.

Об'єкт дослідження - процеси візуального представлення та інтерпретації даних у системах аналізу інформації.

Предмет дослідження - моделі, методи та алгоритми побудови фреймворків генерації специфікованих візуалізацій на основі когнітивного сприйняття та структурно-аналітичних характеристик даних.

Результати дослідження

В роботі створено інтерактивний алгоритмічний фреймворк, здатний оптимізувати дизайн візуалізації відповідно до характеристик даних у режимі реального часу

Висновок

Досліджено та запропоновано методологію поєднання підвибірки даних, проекції та вилучення особливостей для цілей оптимізації дизайну візуальних представлень.

ВІЗУАЛІЗАЦІЯ ДАНИХ, ВІЗУАЛЬНЕ КОДУВАННЯ, КОГНІТИВНЕ СПРИЙНЯТТЯ, ОПТИМІЗАЦІЯ ДИЗАЙНУ, КЛАСТЕРИЗАЦІЯ, ДІАГРАМА РОЗСПОВАННЯ, ПОРОГОВИЙ ГРАФІК, АДАПТИВНІ ВІЗУАЛІЗАЦІЇ

ABSTRACT

Master Thesis: 79 pp., 19 fig., 3 tab., 39 sources.

Topic: Models and methods for building frameworks for generating specified visualizations

Purpose of the work: development of models and methods for building a framework for generating specified visualizations based on cognitive mechanisms of perception, data properties and analytical tasks of the user.

The object of the study is the processes of visual representation and interpretation of data in information analysis systems.

The subject of the study is models, methods and algorithms for building frameworks for generating specified visualizations based on cognitive perception and structural and analytical characteristics of data.

Research results

The work has created an interactive algorithmic framework capable of optimizing visualization design according to data characteristics in real time

Conclusion

A methodology for combining data subsampling, projection and feature extraction for the purposes of optimizing the design of visual representations has been studied and proposed.

DATA VISUALIZATION, VISUAL CODING, COGNITIVE PERCEPTION, DESIGN OPTIMIZATION, CLUSTERIZATION, SCATTER PLOT, THRESHOLD PLOT, ADAPTIVE VISUALIZATIONS

ЗМІСТ

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ	9
ВСТУП	10
РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ВІЗУАЛІЗАЦІЇ ДАНИХ ТА ПРОЦЕСІВ ГЕНЕРАЦІЇ СПЕЦИФІКОВАНИХ ВІЗУАЛІЗАЦІЙ	14
1.1. Візуалізація даних як засіб пізнання та підвищення ефективності аналізу	14
1.1.1. Роль та функціональність візуалізації	14
1.1.2. Вплив дизайну візуалізації на продуктивність	15
1.2. Структура оптимізації дизайну візуалізації на основі завдань та сприйняття	16
1.3. Фундаментальні концепції сприйняття у візуалізації даних	20
1.3.1. Графічне сприйняття	20
1.3.2. Моделювання сприйняття візуалізацій. Закон Вебера-Фехнера та концепція “ледве помітної різниці”	21
1.4. Оптимізація дизайну візуалізації на основі сприйняття	24
1.4.1. Концептуальна модель процесу візуалізації даних	24
Висновки до розділу	30
РОЗДІЛ 2. ДОСЛІДЖЕННЯ МОДЕЛЕЙ ТА МЕТОДІВ ГЕНЕРАЦІЇ СПЕЦИФІКОВАНИХ ВІЗУАЛІЗАЦІЙ НА ОСНОВІ СПРИЙНЯТТЯ	31
2.1. Таксономія ефективності візуалізації на основі взаємозв'язку низькорівневих завдань та візуальних кодувань	31
2.1.1. Таксономія низькорівневих завдань	31
2.1.2. Таксономія візуального кодування	33
2.1.3. Категоризація візуалізацій	34
2.2. Візуальне кодування та його вплив	35
2.3. Візуальне кодування та ефективність пошуку	39
2.3.1. Просторова позиція, форма та розмір	39

2.3.2. Ефективність візуалізацій у завданні фільтрації	40
2.4. Ефективність візуалізацій у кластеризації даних	42
2.4.1. Візуальне кодування та кластеризація	42
2.4.2. Ефективність візуального кодування у кластеризації	44
2.5. Оптимізація дизайну діаграми розсіювання на основі завдання кластеризації.....	45
Висновки до розділу	48
РОЗДІЛ 3. ІМПЛЕМЕНТАЦІЯ МОДЕЛЕЙ ТА АЛГОРИТМІВ ПОБУДОВИ ФРЕЙМВОРКУ ГЕНЕРАЦІЇ СПЕЦИФІКОВАНИХ ВІЗУАЛІЗАЦІЙ	49
3.1. Методологія оптимізації дизайну діаграми розсіювання	49
3.1.1. Оптимізація дизайну та завдання кластеризації	49
3.1.2. Огляд моделі	49
3.1.3. Підвибірка даних	51
3.2. Використання концепції порогового графіка для оцінки кластеризації при візуалізації	53
3.2.1. Пороговий графік	53
3.2.2. Порогова смуга графіка як помітність структури кластера	57
3.3. Розробка та імплементація інтерактивної моделі оптимізації	59
3.3.1. Вибір наборів даних	59
3.3.2. Етап проєкції.....	59
3.3.3. Вилучення особливостей	61
3.4. Аналіз обчислювальної ефективності моделі оптимізації	62
3.4.1. Час обчислення для підвибірки даних	62
3.4.2. Час Обчислення для проєкції та вилучення особливостей	63
3.4. Валідація моделі.....	64
3.5. Оцінка якості результатів та корисності інтерфейсу	69
Висновки до розділу	72
ВИСНОВКИ.....	73
ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ	76

ПЕРЕЛІК УМОВНИХ СКОРОЧЕНЬ

AUC - Area Under Curve

PS - Point Size

OP - Point Opacity

SR - Sampling Rate

t-SNEt - distributed Stochastic Neighbor Embedding

HIT - Human Intelligence Task

SS - Somewhat Similar

DS - Dissimilar

R - Reference (використовується для позначення еталонної діаграми)

H - High (використовується для позначення високої помітності)

M - Medium (використовується для позначення середньої помітності)

L - Low (використовується для позначення низької помітності)

ВСТУП

Актуальність теми.

Візуалізація даних стала одним із ключових інструментів сучасної аналітики та досліджень, забезпечуючи можливість опрацювання складних масивів інформації через зрозумілі графічні форми. У міру зростання обсягів даних та ускладнення структур аналітичних задач зростає потреба у побудові таких візуалізацій, які не лише відображають дані, а й оптимально підтримують когнітивні процеси користувача. Традиційні засоби візуалізації, хоча й мають широкий спектр застосувань, часто не гарантують достатнього рівня ефективності у специфічних задачах, зокрема тоді, коли аналітичні рішення потребують максимально точної та швидкої інтерпретації структур у даних.

Наукові дослідження у галузі візуального сприйняття демонструють, що ефективність графічного представлення значною мірою залежить від відповідності між типом завдання, властивостями даних та обраними візуальними кодуваннями. Отже, виникає потреба у створенні нових моделей та методів, які забезпечують автоматизоване або інтерактивне формування специфікованих візуалізацій під конкретні аналітичні цілі. У сучасних інтелектуальних системах аналітики стає особливо важливим врахування психофізичних механізмів сприйняття — таких як закон Вебера–Фехнера та концепція помітності відмінностей, — що дозволяють оптимізувати дизайн не лише з точки зору естетики, але й когнітивної ефективності.

У цьому контексті актуальним є створення фреймворків, здатних адаптивно генерувати візуалізації на основі характеристик даних, завдань та моделей людського сприйняття. Дослідження таких фреймворків потребує теоретичного обґрунтування, розробки алгоритмічних компонентів, оцінювання їх ефективності та практичної перевірки у реальних сценаріях аналізу даних. Представлена магістерська робота виконує комплекс цих завдань, пропонуючи системний підхід до побудови моделей та методів

генерації специфікованих візуалізацій, що базуються на інтеграції когнітивних принципів та алгоритмічних технік.

Актуальність роботи зумовлена стрімким зростанням обсягів даних та потребою користувачів у високоефективних інструментах для їх інтерпретації. Сучасні підходи до візуалізації часто не враховують психологічні межі сприйняття людини, що призводить до втрати важливої інформації або перекручення структури даних. У багатьох випадках вибір типу візуалізації здійснюється інтуїтивно або за допомогою шаблонів, що не гарантує оптимальності для конкретного аналітичного завдання.

Додаткову важливість набуває завдання адаптації візуалізацій до низькорівневих дій користувача, таких як виявлення трендів, кластеризація, фільтрація або пошук аномалій. Дослідження останніх років свідчать, що ефективність виконання таких завдань залежить від вибору візуального кодування, а параметри візуальних елементів (розмір, колір, позиція) мають непрямий, але суттєвий вплив на швидкість і точність інтерпретації.

Наукова спільнота активно працює над удосконаленням інструментів візуалізації, але питання адаптивного та когнітивно-обґрунтованого формування візуалізацій залишається недостатньо розробленим. Особливо це стосується інтерактивних фреймворків, що мають працювати в режимі реального часу та підлаштовуватися під особливості даних і користувацькі цілі.

Таким чином, дослідження моделі побудови фреймворків генерації специфікованих візуалізацій є своєчасним і важливим кроком для розвитку інтелектуальних систем аналізу даних та підвищення ефективності аналітичних процесів у різних галузях.

Метою роботи є розробка моделей та методів побудови фреймворку генерації специфікованих візуалізацій, що ґрунтуються на когнітивних механізмах сприйняття, властивостях даних та аналітичних завданнях користувача.

Об'єкт дослідження - процеси візуального представлення та інтерпретації даних у системах аналізу інформації.

Предмет дослідження - моделі, методи та алгоритми побудови фреймворків генерації специфікованих візуалізацій на основі когнітивного сприйняття та структурно-аналітичних характеристик даних.

Завдання дослідження:

1. Проаналізувати предметну область візуалізації даних та визначити фактори, що впливають на ефективність графічних представлень.
2. Сформувати таксономію візуальних кодувань та низькорівневих аналітичних завдань користувача.
3. Дослідити вплив різних типів візуального кодування на точність і швидкість виконання завдань пошуку, фільтрації та кластеризації.
4. Розробити модель оптимізації дизайну діаграми розсіювання на основі особливостей даних та механізмів сприйняття.
5. Імплементувати алгоритмічний фреймворк генерації специфікованих візуалізацій з використанням порогових графіків та методів вилучення особливостей.

Методи дослідження

- аналітичний метод для вивчення теоретичних основ візуалізації та моделей сприйняття;
- методи таксономічного аналізу для систематизації візуальних кодувань і завдань;
- алгоритмічні методи обробки та підвибірки даних;
- методи багатовимірної проєкції та вилучення особливостей;
- методи моделювання порогового графіка;
- експериментальні методи для валідації ефективності моделі/

Наукова новизна отриманих результатів

Сформовано цілісну модель адаптивної генерації специфікованих візуалізацій, яка інтегрує когнітивні принципи сприйняття та алгоритмічні механізми оптимізації. Запропоновано розширену таксономію візуальних

кодувань та низькорівневих аналітичних завдань, що є основою для побудови адаптивних фреймворків.

Практичне значення отриманих результатів

Результати роботи можуть бути застосовані у побудові інтелектуальних систем аналізу даних, зокрема в аналітичних платформах, системах моніторингу, інструментах бізнес-аналітики та наукових дослідженнях. Розроблений фреймворк дозволяє автоматизувати створення ефективних візуалізацій, підвищувати точність виконання аналітичних завдань та зменшувати когнітивне навантаження на користувачів.

Структура магістерської роботи. Представлена робота складається зі вступу, трьох розділів та висновків. Загальний обсяг роботи становить 79 сторінок, і містить 29 рисунків, 3 таблиці, перелік використаних джерел із 39 позицій.

РОЗДІЛ 1. АНАЛІЗ ПРЕДМЕТНОЇ ОБЛАСТІ ВІЗУАЛІЗАЦІЇ ДАНИХ ТА ПРОЦЕСІВ ГЕНЕРАЦІЇ СПЕЦИФІКОВАНИХ ВІЗУАЛІЗАЦІЙ

1.1. Візуалізація даних як засіб пізнання та підвищення ефективності аналізу

1.1.1. Роль та функціональність візуалізації

Образність як інструмент пізнання має фундаментальну цінність, оскільки сприяє виявленню неочікуваних або латентних закономірностей у даних. У сучасному контексті, де дані розглядаються як ключовий стратегічний ресурс ("паливо майбутнього"), інструменти візуалізації даних є критично важливими для трансляції інформації та структур, які приховані в емпіричних масивах. Візуалізація стала незамінним елементом у середовищі, що інтенсивно використовує дані (data-driven world), оскільки вона забезпечує покращення людського розуміння та процесів прийняття рішень шляхом розкриття внутрішніх закономірностей та структури інформації.

Візуалізація даних (ВД) є інтегральним компонентом майже в усіх наукових та аналітичних дисциплінах. Вона підтримує динамічне дослідження даних, ідентифікацію патернів, а також характеристику розподілів у різноманітних наборах даних. Діапазон застосування охоплює масиви від малих до великих, від простих до складних, від одновимірних до багатовимірних, включаючи часові, просторові та інтенсивні кількісні дані, а також часові ряди.

Ефективна візуалізація має забезпечувати чітке та об'єктивне представлення інформації, що міститься у вихідних даних. Це необхідно для підвищення точності користувачів при виконанні візуальних аналітичних завдань та експресивної комунікації на основі даних. Проте, існуючі інструменти інформаційної візуалізації наразі демонструють обмежені можливості щодо оптимізації та забезпечення ефективного дизайну візуалізацій.

1.1.2. Вплив дизайну візуалізації на продуктивність

Спосіб кодування даних у візуальній формі істотно впливає на продуктивність користувачів при виконанні аналітичних завдань. Цей вплив зумовлений двома основними аспектами:

- Візуальні канали. Вибір параметрів візуального кодування, таких як колір, тип символу та розмір.
- Аспекти даних. Характеристики самого набору даних, наприклад, кількість точок даних або частота дискретизації (у випадках субдискретизації).

Візуальний дизайн та типологія діаграм можуть кардинально модифікувати інтерпретаційні висновки, зроблені суб'єктами на основі даних. Наприклад, в медичній візуалізації (як-от аналіз кровотоку), вибір візуальних каналів (колір чи тип позначки) безпосередньо впливає на формулювання медичних рекомендацій. Таким чином, настанови щодо ефективного дизайну ВД є контекстно-залежними від обраних візуальних каналів, типів діаграм та специфіки аналітичних завдань.

Ефективні візуалізації оптимізують розуміння даних, використовуючи принципи візуального сприйняття. Наприклад, розмір позначок у діаграмах розсіювання є більш ефективним для представлення кількісних даних, тоді як колір може бути використаний для кодування категоріальних змінних. Важливо, що ефективність візуалізації варіюється залежно від виконуваного завдання; наприклад, для виявлення кластерів порівняно з викидами, параметр непрозорості (opacity) істотно впливає на видимість даних.

Дизайн ВД безпосередньо спрямований на підвищення інтерпретаційної здатності користувача та точності виконання візуальних завдань. Оцінка впливу візуального дизайну на судження користувача здійснюється методами, які включають дослідження людського розуміння та сприйняття.

Ключові емпіричні результати досліджень графічного сприйняття слугують основою для розробки керівних принципів дизайну шляхом

вимірювання точності при використанні різних візуальних кодувань даних у кількісних завданнях. Більшість існуючих настанов зосереджуються або на низькорівневих, специфічних завданнях, або на оптимізації самого дизайну візуалізації. Оптимізація дизайну через візуальне кодування та аспекти даних сприяє покращенню продуктивності користувачів у низькорівневих візуальних завданнях, інтерпретації даних та комунікації.

Проте, сучасний стан оптимізації дизайну часто є відокремленим для кожного типу візуалізації (наприклад, стовпчасті діаграми, кругові діаграми, діаграми розсіювання). Вивчення ефективності ВД, як правило, відбувається за ізольованими аспектами, що підкреслює критичну потребу в інтегрованому розумінні взаємодії цих факторів для розробки посправжньому оптимізованих візуалізацій. Сучасні дослідження рекомендацій щодо дизайну концентруються на ефективності ВД, яка є залежною від візуальних каналів, типів діаграм або візуальних завдань, але вивчаються вони незалежно. Ефективність ВД з погляду вибору дизайну та завдання сприяє розумінню наборів даних аналітиками. Однак встановлено, що вибір дизайну та візуальне судження є взаємопов'язаними, а ефективність не є одномірною величиною. Це формує значну дослідницьку прогалину щодо взаємодії цих факторів для створення оптимізованих візуалізацій. Відповідно, розробка структур, які одночасно враховують дизайнерські рішення у контексті людського сприйняття та виконане аналітичне завдання, є ключовим напрямком для максимізації ефективності дизайну візуалізації.

1.2. Структура оптимізації дизайну візуалізації на основі завдань та сприйняття

Розробка структури дизайну на перетині візуального кодування та низькорівневих аналітичних завдань спрямована на усунення розриву між загальними перцептивними настановами та їхнім практичним застосуванням.

Метою є надання оптимізованого дизайну візуалізації для конкретних завдань, що підвищує якість та достовірність прийняття рішень. Така стратегія забезпечує дизайнерів об'єктивними рекомендаціями.

Ця робота відрізняється від попередніх досліджень тим, що пропонує інтегровану структуру, яка досліджує ефективність дизайну візуалізації відповідно до конкретного завдання, на противагу індивідуальній оцінці ефективності окремих елементів візуального дизайну.

Дослідження інтегрує перцептивні настанови стосовно візуального кодування (наприклад, колір, розмір) та аспектів даних (наприклад, обсяг, частота дискретизації) у дизайн візуалізації. Ці параметри далі вводяться у модель структури, метою якої є представлення оптимального дизайну. Ця структура пропонує:

- Параметри для ключових факторів дизайну.
- Перцептивну модель, засновану на впливі цих факторів.
- Аналітичний рейтинг для трансформації даних у контексті визначених візуальних завдань.

На високому рівні абстракції, певні типи візуалізацій можуть демонструвати вищу ефективність або точність для специфічних візуальних завдань та наборів даних. Дослідження останнього десятиліття підтвердили, що не всі візуалізації є універсальними для кожного завдання [4]. Наприклад, кореляційний аналіз є особливо ефективним при використанні діаграм розсіювання (scatter plots) [3], а ефективність вибору візуального кодування варіюється залежно від продуктивності візуалізації при виконанні завдань [4].

Як ілюструє рисунок 1.1, сприйняття кількості кластерів у наборі даних може змінюватися залежно від варіації таких факторів, як розмір точки, непрозорість, кількість точок даних або характеристики розподілу даних. Надання оптимального рішення щодо вибору дизайну діаграми розсіювання для завдання кластеризації має вирішальне значення для оптимізації ефективної комунікації даними.

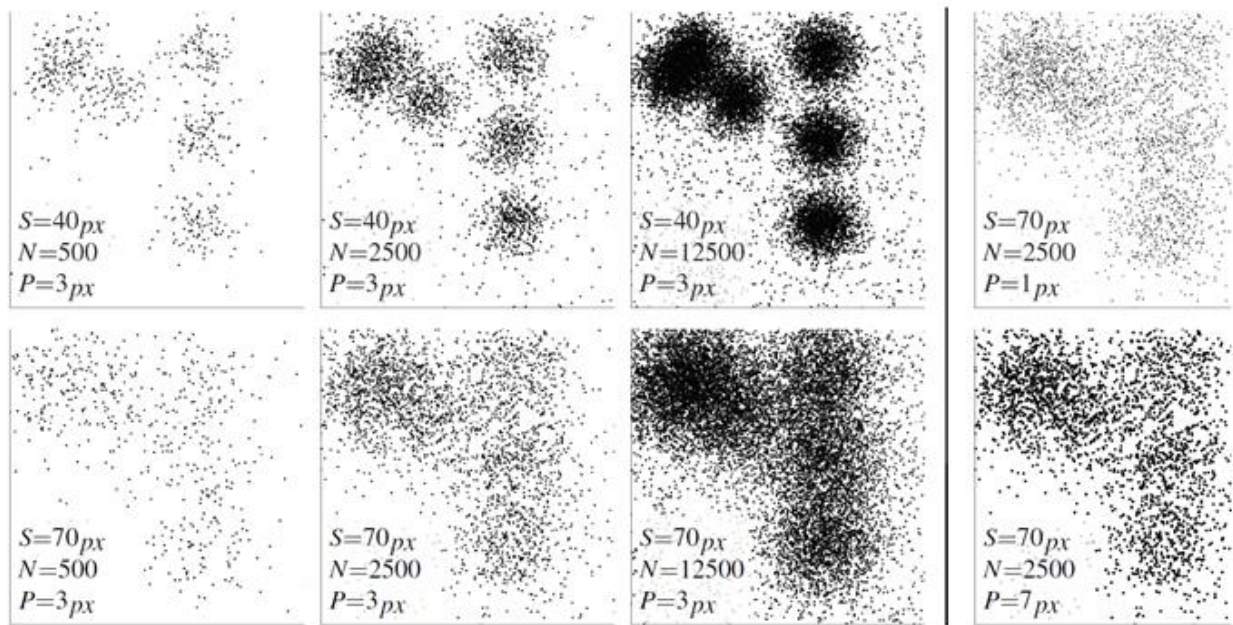


Рис. 1.1. Вісім представлень подібного набору даних з різною кількістю кластерів

В даному випадку, кластер — це групування подібних об'єктів або точок даних. У цьому випадку всі вісім представлень подібного набору даних з різним розміром точок, кількістю точок даних та розміром розподілу даних дають різне сприйняття кількості кластерів.

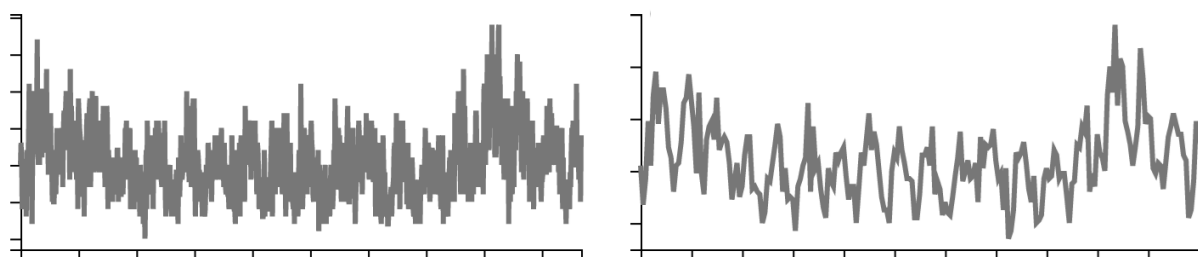


Рис. 1.2. Приклад згладжування ліній для зашумлених даних

Аналогічно, для покращення продуктивності завдання може бути застосована оптимізація візуального представлення, наприклад, згладжування ліній для зашумлених даних (рисунок 1.2). Однак вибір

трансформації даних або представлення візуалізації вимагає подальшої емпіричної оцінки для визначення оптимального підходу, зокрема:

- визначення найбільш ефективного методу згладжування для певного аналітичного завдання.

- визначення дизайну, наприклад, горизонт-графіка, що забезпечує найвищу точність оцінки.

Це дослідження спрямоване на інтеграцію візуального кодування, аналітичних завдань та візуалізації для розробки орієнтованих на завдання настанов для ефективної комунікації даними. Дослідження охоплює два основні напрямки:

- розробка структур, оптимізованих для завдань, які генерують візуалізації, що підтримують ретельне наукове дослідження для конкретного аналітичного завдання.

Оптимізовані структури пропонують набір візуальних кодувань та пов'язаних параметрів (наприклад, непрозорість/розмір символів). Включення факторів, що керують вибором та аспектом даних (наприклад, кількість точок даних, частота дискретизації). Інтеграція методів трансформації даних (наприклад, згладжування ліній).

- Проведення оцінки за допомогою досліджень користувачів, зокрема із залученням учасників через краудсорсингові платформи. Результати будуть квантифіковані для підвищення ефективності комунікації даними у широко використовуваних візуалізаціях.

Оскільки кінцевою метою є покращення ефективності дизайну візуалізації, критично важливо встановити вплив різних візуальних кодувань, аспектів даних та трансформації даних на дизайн та продуктивність завдань. Застосування моделей дозволяє дизайнерам зменшити двозначність у даних і, як наслідок, мінімізувати ймовірність неправильної інтерпретації (наприклад, запобігаючи надто розрідженій або перенасиченій візуалізації для завдань кластеризації). Крім того, стратегічна трансформація даних може призвести до покращення візуального дизайну, що підвищує точність продуктивності

користувачів у низькорівневих завданнях (наприклад, згладжування на лінійних графіках для усунення шуму).

1.3. Фундаментальні концепції сприйняття у візуалізації даних

Нижченаведений розділ містить стислий огляд фундаментальних концепцій графічного сприйняття, психофізичних ефектів та когнітивних/перцептивних упереджень, які є центральними для цієї дисертації. Детальніше вивчення цих концепцій рекомендовано здійснювати через звернення до першоджерел, зазначених у посиланнях.

1.3.1. Графічне сприйняття

Візуалізації даних кодують інформацію, використовуючи різні візуальні кодування (або візуальні канали, візуальні властивості), які включають позицію, довжину, кут, площу, об'єм, затінення, напрямок, кривизну та колір (рисунок 1.3).

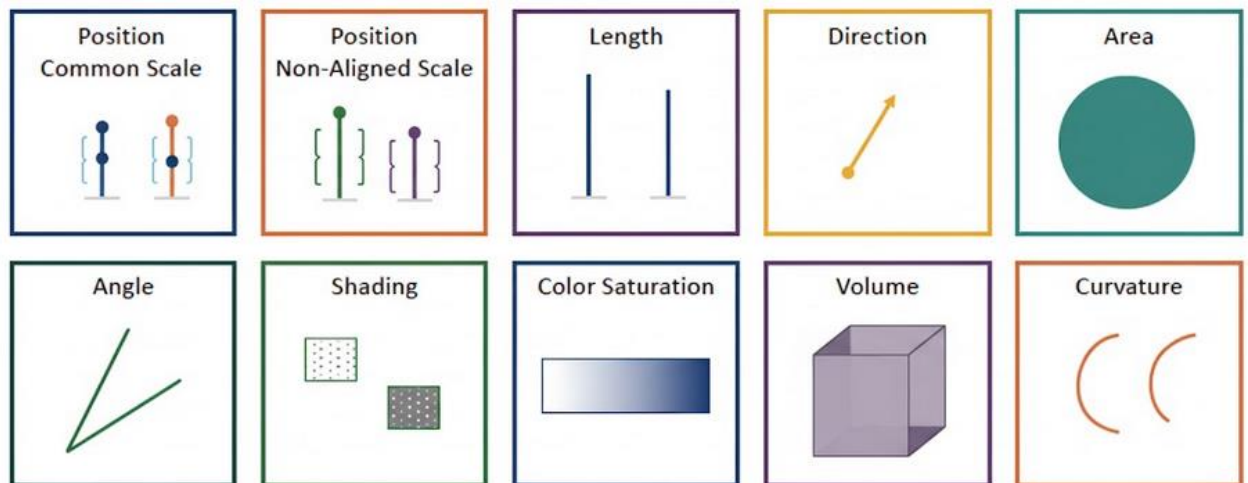


Рис. 1.3. Відтворення типів візуальних кодувань

Термін графічне сприйняття — візуальне декодування інформації, закодованої на графіках, — був вперше введений в [5]. Вони емпірично ранжували візуальні кодування за їхньою перцептивною точністю,

вимірюючи величину суджень для визначення їхньої точності. Дослідження показало, що деякі кодування (наприклад, позиція) є більш перцептивно точними, ніж інші (наприклад, площа або насиченість кольору).

Розуміння ролі сприйняття у виборі візуальних кодувань є критично важливим для розробників візуалізацій, оскільки дисципліна візуалізації інтегрувала знання та теорії сприйняття у візуальний дизайн. Більшість настанов щодо ефективного дизайну візуалізації походить із галузі графічного сприйняття — вивчення того, наскільки точно глядачі можуть інтерпретувати різні кодування даних. В [6] створено комплексний рейтинг візуальних кодувань за типом даних (рисунок 1.4), який був згодом підтверджений та уточнений у численних дослідженнях [7, 9].

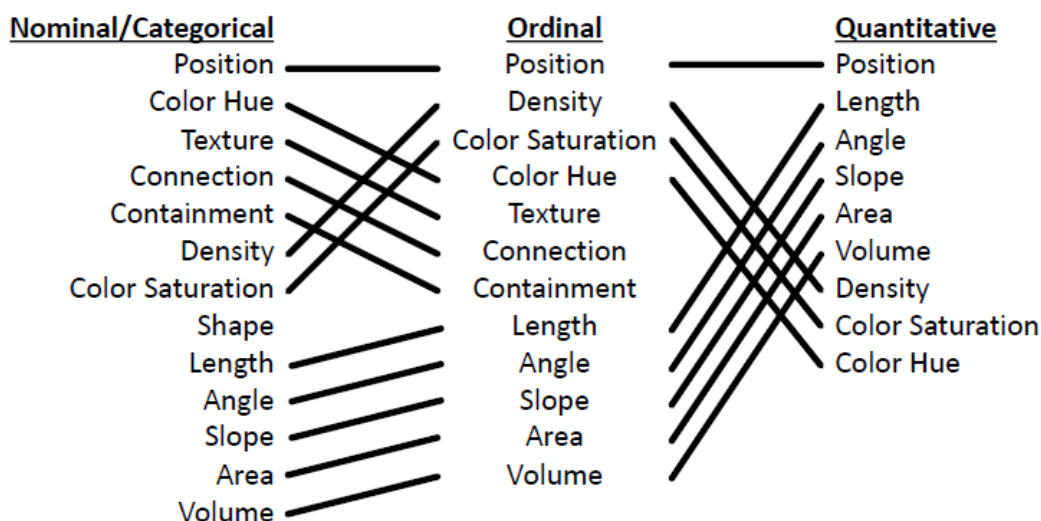


Рис. 1.4. Відтворення рейтингу візуальних кодувань

1.3.2. Моделювання сприйняття візуалізацій. Закон Вебера-Фехнера та концепція “ледве помітної різниці”

Психофізика — це сукупність методів, які кількісно досліджують взаємозв'язки між фізичними стимулами та суб'єктивним відчуттям сприйняття, яке вони викликають [10]. Становлення психології як наукової дисципліни значною мірою пов'язане із застосуванням методів вимірювання психічних подій. Центральними концепціями психофізики є сенсорні порогови:

- Абсолютний поріг (поріг стимулу) - найменша інтенсивність стимулу, необхідна для виникнення відчуття.

- Диференціальний поріг - мінімальна зміна стимулу, необхідна для створення ледве помітної різниці (Just Noticeable Difference, JND) у відчутті.

Закон Вебера-Фехнера є ключовим у психофізиці і моделює, як люди сприймають певні характеристики у візуалізації. Закон стверджує, що зміна стимулу, яка буде ледве помітною (ΔI), є постійним співвідношенням до початкової величини стимулу (I). Це співвідношення ($\Delta I/I = \text{const}$) відоме як частка Вебера. Цей закон емпірично підтверджений для дискримінації ваги, візуальної та тональної дискримінації.

Ледве помітна різниця (JND), або поріг різниці, визначається як мінімальний рівень зміни стимуляції, який суб'єкт може надійно виявити, зазвичай у 50% випадків. JND використовується в перцептивних дослідженнях для визначення мінімальної необхідної корекції стимулу, щоб людина могла достовірно виявити зміну.

Взаємозв'язок між JND та стимулом формалізується законом Вебера:

$$dP = k \left(\frac{dI}{I} \right)$$

де dP — диференціальна зміна сприйняття; k — частка Вебера; dI — диференціальна зміна стимулу; I — фактична інтенсивність стимулу. Застосування JND у поєднанні з психофізичними оцінками є цінним інструментом для вимірювання людських суджень щодо ефективності візуальних кодувань або для вдосконалення дизайну.

При оцінці візуалізацій вкрай важливим є розуміння впливу когнітивного та перцептивного упередження на експериментальні результати.

Перцептивні упередження — це систематичні помилки, які виникають на рівні сенсорного сприйняття візуалізації та/або пов'язаних завдань.

Приклади таких упереджень, що досліджуються у візуалізації, включають ілюзії кластеризації та упередження праймінгу. Наприклад, ілюзія кластеризації виникає, коли суб'єкти недооцінюють дисперсію у невеликих наборах випадкових даних. Інші досліджувані перцептивні феномени включають сліпоту до змін (change blindness) [12] та вплив фотореалізму на сприйняття віртуальних об'єктів. Перцептивні упередження також можуть впливати на усвідомлення користувачем невизначеностей, що, своєю чергою, впливає на довіру, а також можуть бути кількісно визначені для порівняння перебільшення, викликаного оманливими представленнями.

Когнітивні упередження відрізняються від перцептивних тим, що вони зберігаються навіть за умови коректної перцептивної обробки інформації. Ці упередження досліджуються як у когнітивній науці, так і в контексті візуальних аналітичних інструментів та прийняття рішень.

Ефект привабливості (Attraction Effect) - це когнітивне упередження, при якому наявність нерелевантної альтернативи (приманки) впливає на вибір між двома основними варіантами, надаючи перевагу одному з них. У візуалізації цей ефект може спотворювати завдання прийняття рішень на діаграмах розсіювання [14].

	Боб	Аліса	(Єва)
освіта	★★★★★	★★	★★
боротьба зі злочинністю	★★	★★★★★	★★★★

Рис. 1.5. Приклад ефекту привабливості при виборі гіпотетичних кандидатів на виборах

На рисунку 1.5 приведено приклад ефекту привабливості при виборі гіпотетичних кандидатів на виборах де введення "приманки" (Єви) упереджує вибір на користь одного з основних кандидатів (Аліси).

Ефект якоря (Anchoring Effect) - тип когнітивного упередження, при якому початковий стимул або інформація впливає на подальше судження людини у процесі прийняття рішень. Подібні ефекти впорядкування (ordering effects) описують, як послідовність подання інформації може впливати на сприйнятий розмір ефекту.

Когнітивні упередження, такі як якор та праймінг, свідчать, що процес прийняття рішень залежить не лише від поточних візуальних особливостей, але й від попередньої системи відліку.

1.4. Оптимізація дизайну візуалізації на основі сприйняття

Оптимізація дизайну візуалізації є однією з ключових тем у галузі взаємодії людини з комп'ютером та спільноті візуалізації. Цей процес визначається як вибір візуальних кодувань, що забезпечують максимальну помітність та ясність даних, що, своєю чергою, підвищує точність оцінки та судження при виконанні аналітичних завдань. У науковій літературі концепція оптимізації візуального дизайну ретельно вивчалася під різними термінологічними назвами.

1.4.1. Концептуальна модель процесу візуалізації даних

Вибір та рекомендації щодо дизайну становлять критично важливий елемент для забезпечення ефективної візуалізації.

В роботі [15] детально розглянуто людське сприйняття у контексті дизайну інформаційної візуалізації. Метою роботи було узагальнення наслідків досліджень сприйняття та надання експліцитних керівних принципів дизайну.

Процес візуалізації даних (ВД) може бути концептуалізований як ітеративна система, що складається з чотирьох основних етапів, об'єднаних численними циклами зворотного зв'язку (рис. 1.6):

- Збір та зберігання даних - етап первинного отримання та архівування емпіричних даних.

- Передобробка даних (preprocessing) - обчислювальний етап, призначений для трансформації сирих даних у формат, який є придатним для візуального кодування та людського розуміння.

- Візуальне представлення - апаратні та програмні компоненти, що включають дисплейне обладнання та графічні алгоритми, які генерують зображення на екрані.

- Сприйняття та когнітивна система (перцептор) - людський оператор, який здійснює інтерпретацію візуального представлення.

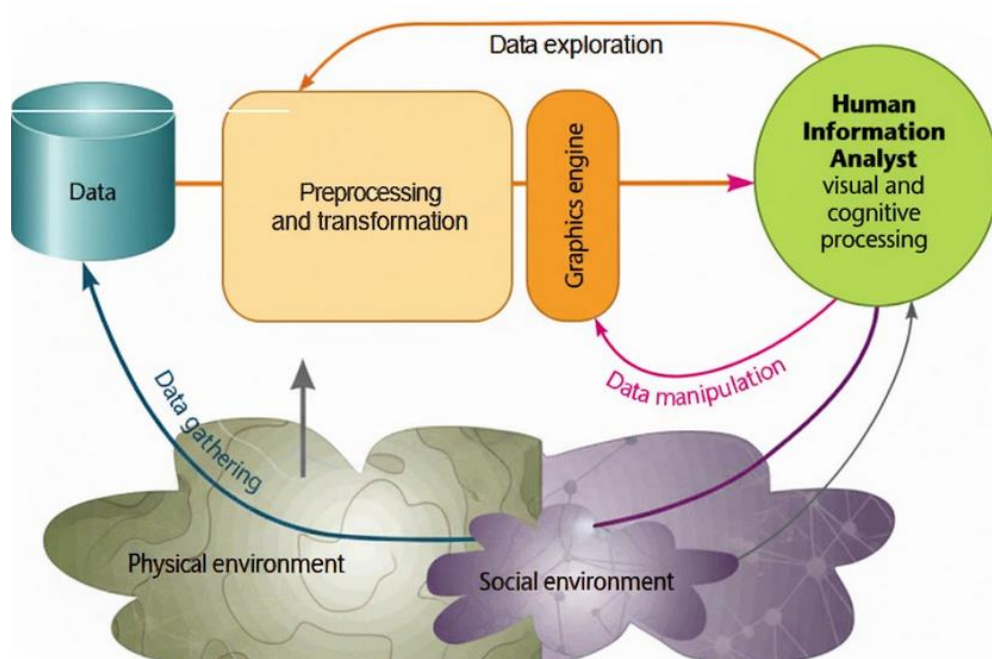


Рис. 1.6. Процес візуалізації даних

Система візуалізації даних підтримується кількома контурами зворотного зв'язку, які забезпечують її динамічний та ітеративний характер:

1. Контур збору даних.

Найдовший цикл, який передбачає можливість того, що аналітик (наприклад, науковець або біржовий аналітик) може вирішити зібрати додаткові дані для подальшого дослідження виявлених закономірностей або

гіпотез. У цьому контурі задіяне як фізичне середовище (як джерело даних), так і соціальне середовище (яке комплексно визначає, що саме збирається і як інтерпретується).

2. Контур перед обробки.

Контролює обчислювальну трансформацію, що передує візуалізації. Аналітик може застосувати специфічні перетворення до даних (наприклад, нормалізацію, агрегацію), вважаючи, що це сприятиме виявленню прихованого сенсу в даних.

3. Контур візуальної взаємодії.

Безпосередньо стосується інтерактивності візуалізації. Наприклад, у 3D-візуалізації даних дослідник може змінити точку спостереження (*vantage point*) для кращого розуміння просторових структур, або ж використати пристрої введення (наприклад, мишу) для інтерактивного вибору діапазонів параметрів, що представляють найбільший інтерес.

Основна увага в даному дослідженні зосереджена на даних, людському сприйнятті та аналітичних завданнях, до яких застосовується візуалізація. Обговорення алгоритмів обмежене лише тими аспектами, що мають пряме відношення до перцептивних процесів. Комп'ютер розглядається, з певними застереженнями, як універсальний інструмент для створення інтерактивної графіки. Це припущення означає, що після визначення оптимального способу візуалізації даних для конкретного завдання, вважається можливим створення відповідних алгоритмів для генерації необхідних зображень.

Ключове наукове питання полягає у визначенні найбільш ефективного методу трансформації даних у форму, яка є зрозумілою для людини, з метою оптимізації прийняття рішень. Мета цього дослідження полягає у створенні концептуальної основи, яка дозволить формалізувати твердження про "візуальну ефективність" або "природність" візуалізацій у вигляді емпірично перевірюваних прогнозів.

В [16] акцентували увагу на ролі уваги та візуальної пам'яті у сприйнятті візуалізацій, підкреслюючи, як характер візуального досвіду

користувачів впливає на точність їхніх суджень. Проект VisGuides [17] є прикладом веб-форуму, створеного для агрегації практичних знань щодо настанов з візуалізації та збору зворотного зв'язку щодо існуючих дизайнів.

Форум VisGuides функціонує як спеціалізована платформа для агрегації, обговорення та систематизації настанов (guidelines) щодо візуалізації даних. Він збирає керівні принципи, посилання на відповідні наукові дослідження, приклади, підтверджувальні та суперечливі емпіричні докази, а також професійні думки та запитання. Загальна схема різних компонентів форуму представлена на рисунку 1.7.

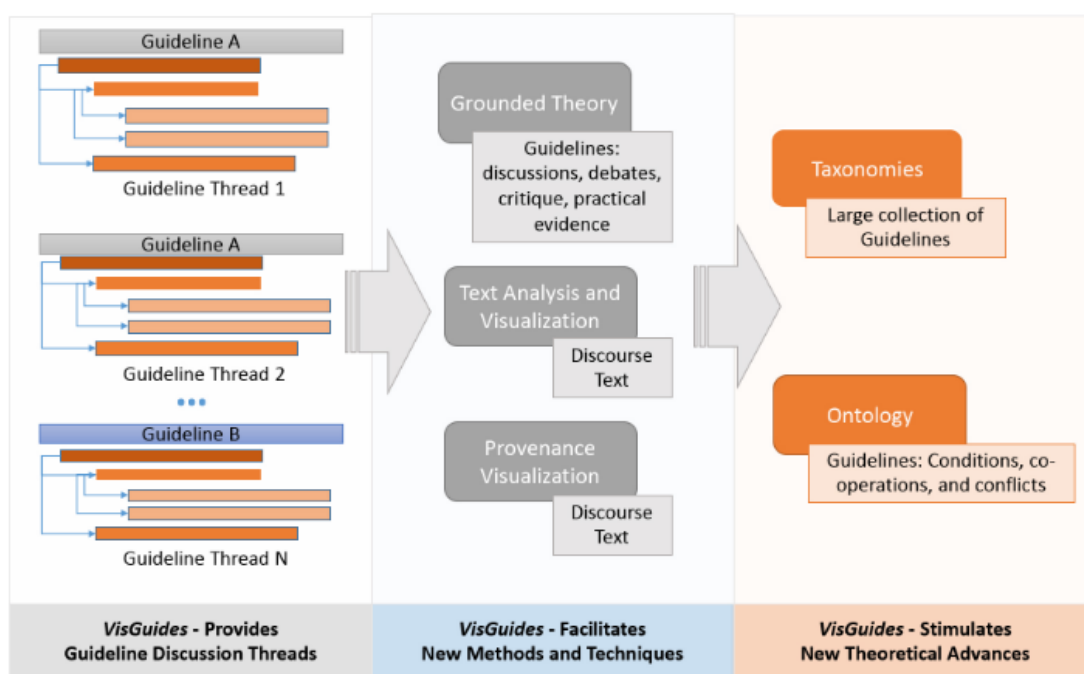


Рис. 1.7. Схема, що відображає компоненти VisGuides

Ліворуч на рис. 1.7 — приклад різних гілок обговорення настанов. Настанова А є прикладом настанови з більш ніж однією активною гілкою. Настанови, дискусії та докази асимілюються для побудови таксономій і, зрештою, онтології настанов та концепцій візуалізації.

Настанова у VisGuides може існувати як окрема гілка обговорення (single thread), охоплювати множинні гілки або бути результатом дискусії, що ґрунтується на існуючому обговоренні (рис. 1.6 ліворуч). Критичною

вимогою є емпіричне або теоретичне обґрунтування тверджень за допомогою прикладів та наукових посилань.

Для початкового наповнення форуму було зібрано 20 настанов з академічної літератури та шляхом консультацій з фахівцями у спільноті візуалізації. Наразі форум містить шість ключових настанов, кожна з яких організована в окрему гілку обговорення:

- Ефект "розкидання" (Blow-apart effect)

Використання кольорової карти "веселка" (Rainbow colormap usage).

- Принцип коефіцієнта брехні (Lie-factor guideline).

- Дія без взаємодії (Action-without-interaction).

- Коефіцієнт чорнила даних (Data-ink ratio).

Принцип відповідності (Correspondence principle).

Типовий приклад формулювання настанови виглядає так: "Не використовуйте ефекти розкидання". Далі в гілці обговорення наводиться першоджерело (наприклад, [17]), а користувачі можуть додавати уточнення контексту використання та ставити питання, як-от: "Чи можу я використовувати ефект розкидання як зворотний зв'язок для інтерфейсу зі зв'язаними поданнями (linked-view interface)? В рамках інформаційної панелі [...] Колірний канал був використаний для позначення [...]. Дослідник візуалізації сказав, що я не повинен цього робити, оскільки це ефект розкидання. Але те, що він запропонував, здається настільки ж проблематичним (наприклад, потовщення контуру, зміна розміру, анімація тощо). Будь ласка, допоможіть."

Зібрані настанови, дискусії та докази проходять процес асиміляції (інтеграції) з метою побудови таксономій. Кінцевою метою є розробка онтології керівних принципів та концепцій візуалізації.

На додаток до загальних керівних принципів, значна кількість досліджень зосереджена на розробці надійних структур для оптимізації дизайну.

- Перцептивне обґрунтування дизайну – це структура для міркувань про сприйняття дизайнів візуалізацій пропонує інтеграцію технік із науки про зір. Це ґрунтується на двох ключових тезах:

- Розглядає глядача та систему візуалізації як єдину інтегровану систему.

- Постулює концепцію оптимальної візуалізації. Дослідження порушує фундаментальні питання, зокрема: Як найкраще виміряти продуктивність даної візуалізації? Чи можливо визначити оптимальність її дизайну?

В роботі [17] представили простір дизайну експериментальних методів для емпіричного дослідження перцептивних процесів, залучених у перегляд візуалізацій. Це спрямовано на формування настанов щодо дизайну візуалізації. Надаючи спільний простір дизайну та лексикон, ця робота полегшує емпіричні дослідження, сприяючи створенню інноваційних досліджень та просуванню наукового розуміння вибору дизайну та оцінок у візуалізації.

На противагу цим підходам, дане дослідження пропонує систематизований огляд досліджень, заснованих на сприйнятті, класифікуючи роботи, які оцінювали ефективність різних візуалізацій у контексті широкого діапазону завдань. На відміну від попередніх підходів, ця робота характеризує ефективність візуалізації як двовимірну метрику, що функціонує на перетині візуального дизайну та низькорівневих завдань.

Ця робота відрізняється від попередніх досліджень розробкою структури, яка досліджує ефективність дизайну візуалізації у прив'язці до завдань, а не лише індивідуальну ефективність візуального дизайну.

Дослідження сприйняття має вирішальне значення, оскільки воно зосереджується на тому, як візуальна система інтерпретує низькорівневі абстрактні особливості (наприклад, колір, позицію) і як точність та продуктивність змінюються залежно від їхнього вибору дизайну. В наступному розділі буде надано огляд досліджень візуалізації, заснованих на сприйнятті, які вимірювали ефективність візуального дизайну.

Висновки до розділу

У першому розділі проведено аналіз предметної області візуалізації даних, що дозволило сформулювати цілісне розуміння ролі візуалізації у процесах аналітичного мислення. Розглянуто ключові функції та когнітивні механізми, які впливають на ефективність сприйняття та інтерпретації графічних представлень. Показано, що дизайн візуалізації суттєво визначає швидкість, точність та якість аналітичних рішень користувача. Дослідження сутності графічного сприйняття, включаючи закон Вебера–Фехнера та концепцію «ледве помітної різниці», підтвердило необхідність врахування психофізичних особливостей людини у побудові візуальних моделей. Розкрито структуру оптимізації дизайну на основі взаємодії між типом завдання та характеристиками сприйняття, що дозволило сформулювати відповідну концептуальну модель процесу візуалізації. Детальний аналіз різних форм візуальних кодувань та їх впливу на когнітивні процеси дав змогу визначити критерії ефективності відображення даних. У підсумку встановлено, що оптимізація дизайну візуалізацій повинна ґрунтуватися на інтеграції даних, завдань користувача та механізмів візуального сприйняття.

РОЗДІЛ 2. ДОСЛІДЖЕННЯ МОДЕЛЕЙ ТА МЕТОДІВ ГЕНЕРАЦІЇ СПЕЦИФІКОВАНИХ ВІЗУАЛІЗАЦІЙ НА ОСНОВІ СПРИЙНЯТТЯ

2.1. Таксономія ефективності візуалізації на основі взаємозв'язку низькорівневих завдань та візуальних кодувань

У науковій літературі метод візуального кодування даних традиційно розглядається як основний компонент візуалізації. Проте аналітичне завдання є не менш, якщо не більш, важливим фактором. Численні оціночні дослідження підтверджують, що ефективність візуалізації є залежною від завдання [19]. Значний обсяг досліджень спрямовано на визначення перцептивно оптимальних представлень даних для специфічних низькорівневих завдань. Оскільки переважна більшість досліджень у сфері візуалізації використовувала конкретне низькорівневе завдання як основну мету або як компонент оцінки, ця категорія була обрана як основний критерій таксономії. Крім того, більшість публікацій розглядають низькорівневі завдання лише в контексті обмеженої підмножини візуальних кодувань та/або типів візуалізацій. Відповідно, обговорення низькорівневих завдань далі поділяється за типами візуального кодування та візуалізації.

2.1.1. Таксономія низькорівневих завдань


Для формування основи огляду було розглянуто дві існуючі таксономії завдань. Спочатку була розглянута таксономія абстрагованих завдань [21], яка є таксономією вищого рівня (включає як перцептивні, так і когнітивні аспекти). Через акцент дослідження на сприйнятті, і необхідність мінімізувати когнітивне міркування в рамках завдання, ця таксономія виявилася менш придатною.



Натомість, була обрана таксономія низькорівневих завдань [10], оскільки визначені в ній завдання вимагають меншого рівня когнітивного міркування про дані. На її основі було прийнято десять низькорівневих

завдань (отримання значення, фільтрація, обчислення похідного значення, знаходження екстремуму, сортування, визначення діапазону, характеристика розподілу, знаходження аномалій, кластеризація та кореляція), до яких було додано одне похідне завдання — порівняння. Це утворило сукупно одинадцять груп низькорівневих завдань (таблиця 2.1). Кожне розглянуте дослідження було віднесено до однієї або кількох цих груп.

Таблиця 2.1.

Візуалізація таксономії для низькорівневих завдань

Завдання	Іконка	Опис
Отримати значення		Знайти атрибути для заданого набору конкретних випадків.
Фільтрація		Знайти випадки даних, що задовольняють задані умови на значеннях атрибутів.
Обчислення похідного значення		Обчислити агреговане числове представлення заданого набору випадків даних.
Знаходження екстремуму		Знайти випадки даних з екстремальним значенням атрибута в межах його діапазону.
Визначення діапазону		Знайти діапазон значень у наборі даних для атрибута інтересу.
Сортування		Ранжувати заданий набір випадків даних відповідно до певної ординальної метрики.
Знаходження аномалій		Ідентифікувати будь-які аномалії (наприклад, статистичні викиди) у заданому наборі випадків даних.
Характеристика розподілу		Охарактеризувати розподіл значень кількісного атрибута по набору випадків даних.
Кластеризація		Знайти кластери подібних значень атрибутів у заданому наборі випадків даних.

Завдання	Іконка	Опис
Кореляція		Визначити корисні відношення між значеннями двох атрибутів у наборі випадків даних.
Порівняння		Порівняти будь-які атрибути в межах та між відношеннями заданого набору випадків даних.




2.1.2. Таксономія візуального кодування

Візуальні кодування — це властивості, які використовуються для кодування даних у візуалізації, включаючи такі канали, як позиція, довжина, кут, площа, об'єм, затінення, напрямок, кривизна та колір (див. рис. 1.3). Терміни "графічне кодування", "візуальний канал", "візуальні кодування" та "візуальні властивості" часто використовуються як взаємозамінні.

Роботи з візуального кодування ідентифікували численні окремі канали. З метою забезпечення керованості таксономії, візуальні кодування були об'єднані у дві основні категорії: просторові кодування та колірні кодування, кожна з яких має дві підкатегорії (таблиця 2.2).

Таблиця 2.2.

Таксономія візуального кодування

Категорія	Підкатегорія	Візуалізація	Опис
Просторові кодування	Просторова позиція та форма		Кодування, що стосуються компоновання (позиція) або форми (напрямок/кут, кривизна).
	Просторова розмірність		Кодування, де розмір об'єктів (довжина, площа, об'єм) кодує дані.
Колірні кодування	Колірний відтінок		Варіації кольору (у побутовому сенсі), що зазвичай стосуються категорійних кольорів або колірних карт.

Категорія	Підкатегорія	Візуалізація	Опис
	Інтенсивність кольору		Кодування даних через інтенсивність кольорового представлення (насиченість, світність, затінення або непрозорість), об'єднані через їх тісний взаємозв'язок.

Просторові кодування охоплюють кодування, пов'язані з позицією, розміром (довжиною, площею або об'ємом), напрямком та формою. Дослідження показали, що просторові кодування є особливо ефективними для кількісних даних.

Колірні кодування включають властивості кольору, такі як відтінок, насиченість, світність та непрозорість. Психофізичні експерименти продемонстрували, що колір може бути використаний для представлення впорядкування, категорії, кількості та судження про сприйняті відмінності у візуалізації.

2.1.3. Категоризація візуалізацій

Остання категорія цієї таксономії — це тип візуалізації. Більшість проаналізованих статей вивчали стандартні типи візуалізацій, часто з варіаціями в їх дизайні. Шляхом вилучення та об'єднання пов'язаних типів візуалізацій було сформовано десять основних категорій та одна категорія "інші", представлені на рисунку 2.1.

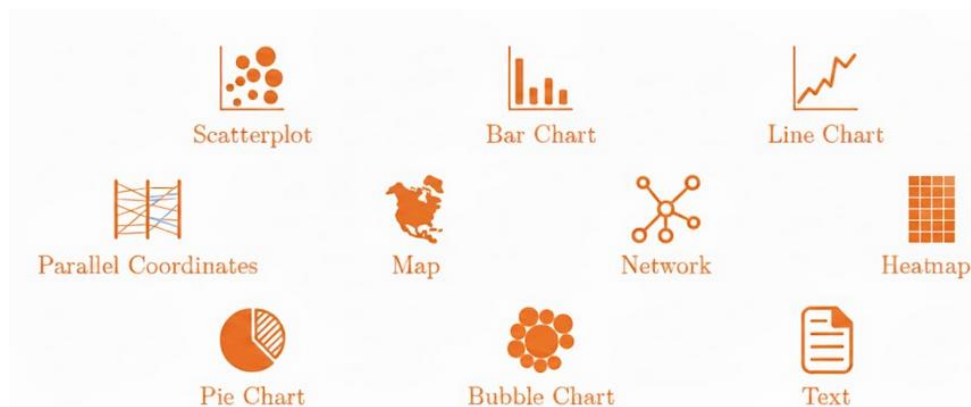


Рис. 2.1. Перелік категорій візуалізацій, використаних для таксономії

Метою даного огляду було систематизувати та структурувати наявні дослідження таким чином, щоб забезпечити генерацію ефективних візуальних дизайнів, обґрунтованих принципами графічного сприйняття. Проте, аналіз виявив, що більшість емпіричних досліджень у цій галузі мають обмежене охоплення, що проявляється у:

- Вузькій вибірці візуальних кодувань та типів візуалізацій.
- Обмеженій кількості аналітичних завдань.
- Невеликому розмірі та низькій різноманітності тестованих наборів даних.

Крім того, більшість висновків досліджень не підлягають систематичній реплікації або зовнішній валідації, за винятком невеликої кількості винятків.

2.2. Візуальне кодування та його вплив

Традиційно, візуальні канали, такі як позиція вздовж загальних та невіривняних шкал, демонструють вищу точність суджень порівняно з довжиною, напрямком та кутом у простих перцептивних завданнях.

Проте, недавні дослідження розширюють це розуміння, вказуючи на складні взаємодії:

- У візуалізаціях, заснованих на позиції (наприклад, діаграми розсіювання), кодування додаткових атрибутів розміром вимагало більше часу для отримання значень, ніж кодування кольором.

- При цьому точність учасників парадоксально покращувалася при використанні розміру.

- Це спостереження призвело до гіпотези, що варіації розміру позначок можуть заважати декодуванню кількісного значення через позиційні канали.

- Було також показано, що орієнтація діаграми впливає на продуктивність: діаграми з фасетками по осі у виконувалися у 0.9 разів швидше, ніж по осі x.

Таким чином, розуміння того, як варіації в дизайні візуалізації впливають на продуктивність користувачів у завданні отримання значення, знаходиться на стадії розвитку.

Кілька робіт оцінювали здатність суб'єктів ідентифікувати значення за допомогою колірних візуалізацій. Продуктивність у визначенні або ідентифікації значення залежить не лише від кольору, але й від інших візуальних властивостей, таких як рух або компоновання [18].

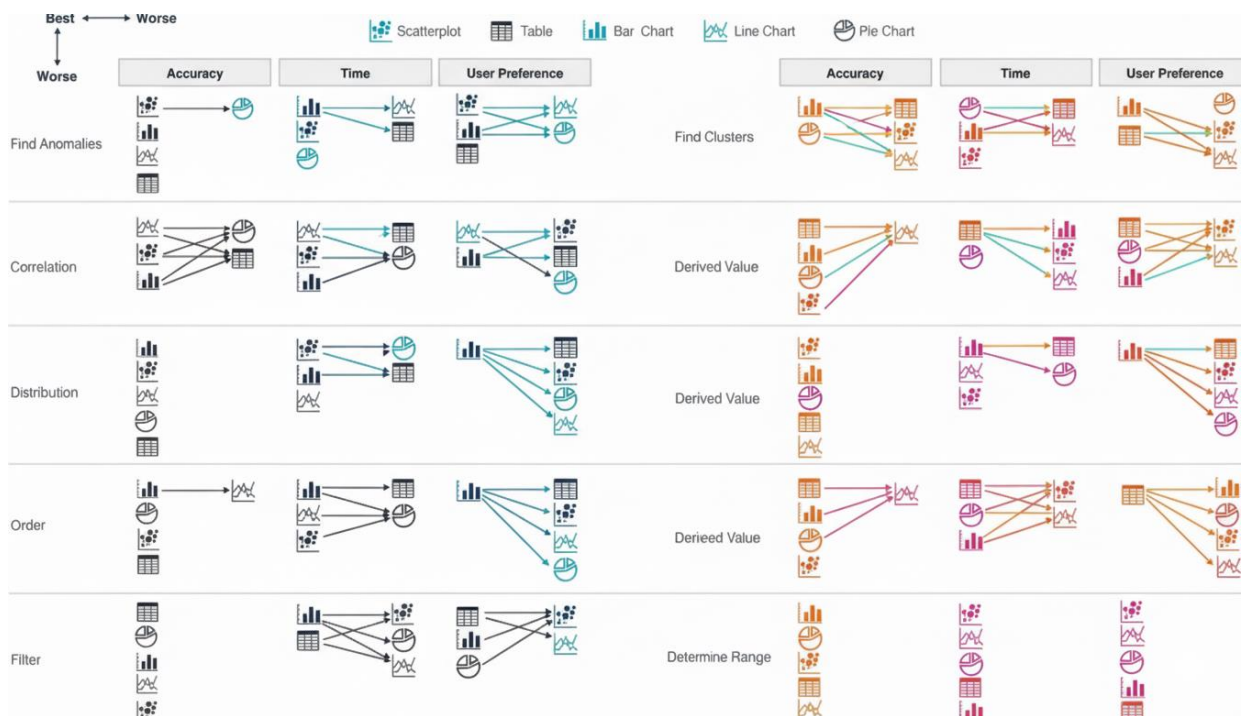


Рис. 2.2. Емпіричне дослідження оцінки попарного відношення між різними типами візуалізацій у контексті аналітичних завдань

На рис. 2.2. стрілки позначають, що вихідний елемент (джерело) демонструє статистично значущу перевагу над цільовим елементом за показниками ефективності (метриками продуктивності).

Були проведені дослідження, зосереджені на застосуванні кольору для зниження когнітивного навантаження під час пошуку цілей або зчитування значень. Проблема колірної сліпоти (дальтонізму) була вивчена у контексті автоматизованого дизайну колірних карт. Експерименти, що порівнювали

колірні градієнти, згенеровані автоматизованою системою, з тими, що були розроблені вручну, показали зіставну продуктивність глядачів у точному зчитуванні значень на діаграмах розсіювання, теплових картах та хороплетних картах.

Звідси впливає необхідність розумної обережності при розробці візуалізацій для покращення доступності (наприклад, для користувачів із дальтонізмом або низьким зором), що забезпечує неупереджену реакцію користувача та мінімізує складність завдання.

Дослідження підтверджують, що світлість (luminance) символу точки даних є впливовим фактором для візуальних завдань, включаючи локалізацію та ідентифікацію значень у розріджених діаграмах розсіювання на білому тлі. Цей експеримент визначив, що світлість може бути змодельована як комбінація двох протилежних степеневих функцій для визначення розрізняваності (discriminability). Непрозорість (opacity), яка пов'язана зі світлістю, використовувалася як конструктивний фактор у дизайні діаграм розсіювання для пом'якшення надмірного накладання при читанні та ідентифікації значень.

Порівняльне дослідження, що оцінювало п'ять типів візуалізацій (таблична візуалізація, діаграми розсіювання, стовпчасті діаграми, лінійні діаграми та кругові діаграми) для малих наборів даних у контексті десяти завдань [10], показало, що для завдання отримання значення:

- Таблична візуалізація перевершила всі інші за точністю, часом та перевагами користувачів.

- Стовпчасті та кругові діаграми показали хороші результати за точністю та часом.

- Діаграми розсіювання були ефективними лише за точністю.

Подібний аналіз, який порівнював багатовимірні візуалізації (паралельні координати, матриця діаграм розсіювання та таблична візуалізація), виявив, що паралельні координати забезпечили найвищу точність для завдання отримання значення. В обох дослідженнях табличні

візуалізації дозволяли швидше та точніше приймати рішення, що пояснюється високою знайомістю користувачів з табличним форматом.

Діаграма розсіювання (Scatter Plot) є високоефективними для ідентифікації та зчитування значень на основі різних візуальних кодувань. Наприклад, їх забарвлення додатковими категоріальними даними виявилось високоефективним для порівняння окремих даних.

Стовпчаста діаграма (Bar Chart) полегшує ідентифікацію кількісних даних, і дизайнерські рішення суттєво впливають на продуктивність. Використання піктограм у стовпчастих діаграмах знижує продуктивність у завданні отримання значення. Існує потреба в прямому кодуванні абсолютних значень та їхніх відношень у стовпчастих або точкових діаграмах. Естетичні прикраси можуть покращувати запам'ятовуваність, але потенційно знижують продуктивність у цьому завданні.

У візуалізаціях, що містять текст, оптимізація типографіки (розташування міток, вибір шрифту, розміру, ваги, кольору, орієнтації, інтервалу) є критичною для читабельності та, відповідно, зрозумілості візуалізації. Неправильне використання типографічних параметрів може викликати упередження, що впливають на точність користувача.

Дослідження масштабованості сприйняття на великих дисплеях показало, що використання більшого дисплея для карт не призводить до збільшення часу або зниження точності виконання завдання отримання значення [23].

Дослідження демонструють, що ефективність користувачів у завданні отримання значення значно варіюється залежно від вибору візуалізації та її дизайну. Попередні дані свідчать про те, що таблична візуалізація та стовпчасті діаграми є найбільш оптимальними, коли точність є пріоритетною. Хоча естетичні прикраси можуть підвищити запам'ятовуваність, вони повинні використовуватися з обережністю через їхній потенційний негативний вплив на продуктивність у завданні зчитування значень.

2.3. Візуальне кодування та ефективність пошуку

2.3.1. Просторова позиція, форма та розмір

Ефект виділення (Popout), спричинений змінами у певних візуальних кодуваннях, критично важливий для привернення уваги користувача до релевантних елементів. Дослідження ефективності виділення для ідентифікації цілі в діаграмах розсіювання з використанням кольорового відтінку, форми, світності, миготіння, руху та розміру показало значні відмінності у продуктивності за метриками сприйнятого успіху, видимості та точності:

- Форма вимагала значних зусиль і демонструвала низьку продуктивність.
- Рух показав мінімальні зусилля та високу продуктивність.
- Інші чотири канали продемонстрували пряму залежність між інтенсивністю кодування та сприйнятою видимістю/точністю.

Експеримент, що досліджував пошук цілі в сітковій матриці з використанням розміру позначки, обсягу набору та кольору [21], запропонував настанови щодо дизайну на основі групування та розміру символів:

- Пошук елемента був швидшим, коли кольори були просторово згруповані.
- Обсяг набору символів мав незначний вплив на час пошуку у згрупованих відображеннях, але продуктивність сповільнювалася для випадкових відображень.

Оптимізація колірною кодування є необхідною для підвищення ефективності візуалізації категоріальних даних. Хоча категоріальні кольори легко розрізняються на великих колірних смугах, їхня видимість зменшується, коли вони застосовуються до карт або інших графіків із варіативними розмірами елементів.

Видимість класу (Class Visibility) є важливою метрикою, що відображає, як колір і просторовий розподіл кожного класу впливають на його перцептивну інтенсивність для візуальної системи людини [160].

2.3.2. Ефективність візуалізацій у завданні фільтрації

Завдання фільтрації підкреслює необхідність явного та прямого кодування числових відмінностей між значеннями даних.

У дослідженні [24] для малих наборів даних, табличні візуалізації та стовпчасті діаграми показали високу ефективність за метриками точності, часу виконання та переваг користувачів (рис. 2.3).

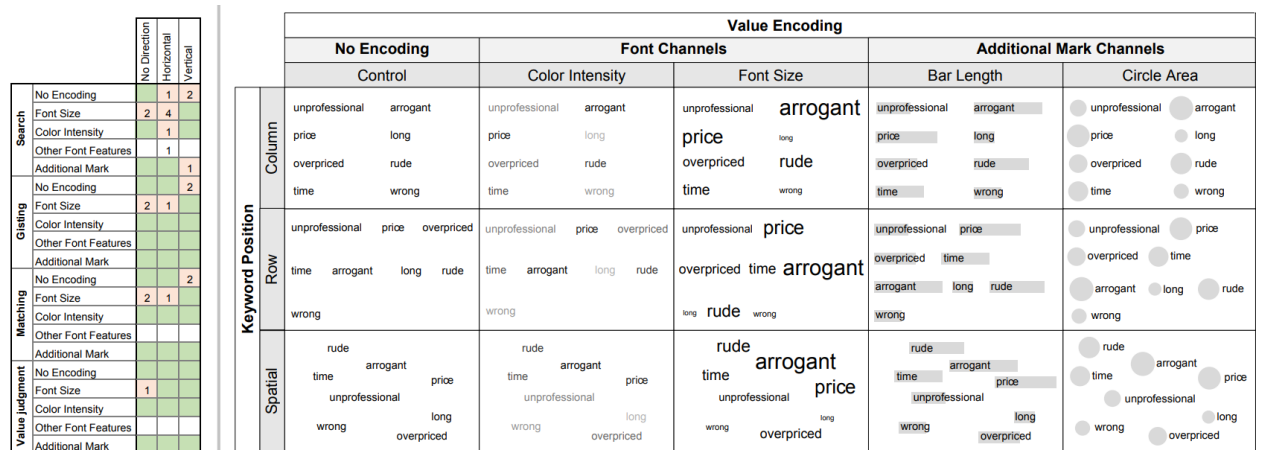


Рис. 2.3. Приклади візуалізацій тексту, згенерованих візуальними кодуваннями та компоуваннями

Різні представлення впливають на продуктивність людей у вилученні інформації з текстових візуалізацій, зокрема хмар слів. Це резюме простору дизайну демонструє приклади візуалізацій, створених за допомогою різних візуальних кодувань та компоувань. У експериментах із пошуком цільового зображення в наборі зображень було показано, що як затримка (latency), так і складність завдання відіграють значну роль у поведінці пошуку.

Експеримент із використанням паралельних координат (Parallel Coordinates) для оцінки зручності використання у завданні фільтрації із застосуванням відстеження погляду (eye-tracking) показав, що користувачі

демонстрували тривалий час фіксації для підтвердження інтерпретації результатів [26]. Дослідження також показало, що вплив затримки в інтерактивній системі є швидше поступовим, ніж бінарним.

Карти використовуються для представлення структур та відносин у просторових даних. Дослідження [21, 23] зосереджені на оцінці людського сприйняття в ідентифікації просторових візерунків та структур. Видимість класу на картах залежить від вибору колірної кодування та просторового розподілу

Оцінка картограм за чотирма типами (суцільні, несучільні, прямокутні та Дорлінга) показала:

- Картограми, що зберігають відносну позицію регіонів, полегшують швидший пошук.

- Карти Дорлінга та прямокутні карти забезпечують вищу точність.

Оптимальне налаштування типографічних параметрів (шрифту, розміру, ваги, кольору, орієнтації) визначає читабельність тексту та, відповідно, зрозумілість візуалізації. У завданні пошуку ключового слова в хмарах слів, розмір шрифту та колір виявилися більш ефективними кодуваннями, ніж додаткові позначки (смуги або кола) на ключових словах (рис. 2.3).

Просторове та стовпцеве компонування ключових слів перевершувало рядкову версію у пошуку. Алфавітне впорядкування було найшвидшим для пошуку. Дослідження підтвердили, що слова з більшими розмірами шрифту частіше згадувалися, підкреслюючи вплив типографії на пам'ять.

Продуктивність користувачів у завданні фільтрації варіюється залежно від дизайну та вибору візуалізації. Табличні візуалізації та стовпчасті діаграми показали надійну ефективність за ключовими метриками. Пошук цілей (як ключового слова в тексті, так і значення в діаграмі) виграє від стратегічного використання кодувань розміру шрифту/символу та кольору, а також алфавітного впорядкування для текстових даних.

2.4. Ефективність візуалізацій у кластеризації даних

Завдання кластеризації зосереджене на ідентифікації груп подібних елементів у заданому наборі даних. Кластеризація та сегментація точок даних розкривають внутрішні характеристики набору, дозволяючи дизайнерам та практикам візуалізації глибше досліджувати дані. У контексті інформаційної візуалізації кластеризація визначається як високорівнева характеристика даних — здатність ідентифікувати групи подібних об'єктів.

2.4.1. Візуальне кодування та кластеризація

Розмір символу (позначки) є ключовим візуальним кодуванням, що впливає на щільність та концентрацію кластерів точок. Він має прямий вплив на здатність ідентифікувати кластери, і було продемонстровано, що розрізняюваність є залежною від завдання.

Роздільність (Separability) між символами або їхніми групами є важливим фактором. Форма позначки суттєво впливає на сприйняття як розміру, так і кольору, причому роздільність між цими трьома кодуваннями функціонує асиметрично. Оскільки концентрація щільності впливає на сприйняття кластерів, збільшення розміру точок даних веде до зростання концентрації та щільності. Існує математична модель для представлення сприйняття кластерів у розподілі точок, припускаючи, що сприйняття значно залежить від зміни концентрації та щільності.

Дослідження перцептивної оптимізації діаграм розсіювання, яке вивчало стандартні параметри дизайну (включаючи розмір позначки, непрозорість та співвідношення сторін), продемонструвало, що ефективний вибір цих змінних посилює розділення класів.

Колір є ще одним важливим візуальним кодуванням у діаграмах розсіювання, що впливає на завдання сегментації, кластеризації та групування. Було встановлено, що сприйняття кольорової різниці змінюється обернено пропорційно до діаметра позначки (див. рисунок 2.4). Крім того,

форма символу значно впливає на те, наскільки добре ми сприймаємо різницю кольорів.

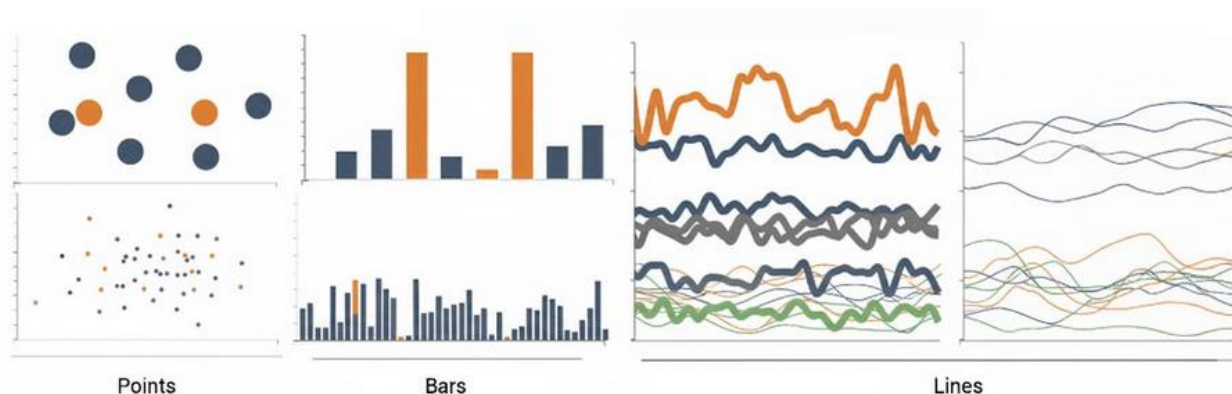


Рис. 2.4. Сприйнята різниця кольорів змінюється обернено пропорційно до розміру позначки та має складну залежність від її форми

Оптимізований вибір кольорів допомагає користувачам ефективно розуміти роздільність у багатокласових діаграмах розсіювання. Деякі методи використовують призначення кольору для оптимізації сприйняття роздільності класів, враховуючи фактори, пов'язані з щільністю (наприклад, просторове відношення, ступінь перекриття точок і кластерів, а також колір фону), чого неможливо досягти за допомогою стандартного кольорового картографування.

Зменшення непрозорості (opacity) позначок може пом'якшити надмірне накладання (overplotting), сприяючи виконанню різних завдань візуальної аналітики, зокрема сприйняттю та ідентифікації кластерів, зберігаючи при цьому просторову інформацію. Різні рівні непрозорості сприяють посиленню розділення класів: низька непрозорість корисна для оцінки щільності у великих даних, але ускладнює локалізацію викидів. Однак, ці рішення мають обмеження, наприклад, коли непрозорість нижче доступної точності або точки мають мінімально можливий розмір.

Ефекти праймінгу та якорування можуть спотворювати процес прийняття рішень користувачем. Визначення роздільності двох кластерів залежить не лише від їхньої фізичної відстані, але й від попередньо

побачених стимулів. Хоча ці ефекти можуть бути випадково спричинені явищем хромостереопсису [8] і повинні досліджуватися на монохромних кольорах, вони підкреслюють, як повторний вплив візуалізації впливає на інтерпретацію.

2.4.2. Ефективність візуального кодування у кластеризації

У дослідженні [24] для малих наборів даних було виявлено, що стовпчасті діаграми та кругові діаграми перевершують таблиці, діаграми розсіювання та лінійні діаграми в завданнях кластеризації (див. рис. 2.2). Висока продуктивність кругових діаграм у сприйнятті кластерів може бути пояснена їхньою ефективністю у сприйнятті пропорційним судженням через відношення "частина-до-цілого".

Діаграми розсіювання широко використовуються для візуалізації даних з метою виявлення патернів, таких як сегментація класів або кластери. Численні дослідження зосереджувалися на перцептивному розділенні кластерів для ідентифікації впливу різних факторів дизайну та візуальних кодувань. Якісна оцінка заходів роздільності кластерів запропонувала таксономію з чотирьох факторів, що впливають на сприйняття розділення: масштаб, відстань точок, форма та позиція.

Для мінімізації впливу людського судження, оцінили 15 сучасних показників роздільності класів, використовуючи людську "еталонну істину" (human ground truth) як вхідні дані для машинного навчання. Подальші роботи, такі як [26] (використовує глибоке навчання для емуляції рішень людини щодо кластеризації) та техніка Scagnostics [28] (фокусується на ідентифікації патернів, включаючи кластери), є алгоритмічно орієнтованими, і більшість їхніх оцінок не враховувала візуальні канали.

Паралельні координати є корисними для різноманітних завдань, включаючи кластеризацію в реальних додатках. Дослідження, що вимірювало результати навчання учасників із застосуванням кластеризації як

основного завдання, показало більш привабливий досвід для інтерактивних паралельних координат порівняно зі статичними.

Були розроблені паралельні координати з покращеною орієнтацією [23] для великих наборів даних з метою посилення відображення базової структури даних та покращення розрізнюваності кластерів. Залишається відкритим питання про те, що краще для ідентифікації кластерів: діаграми розсіювання чи паралельні координати. Одне дослідження показало, що паралельні координати краще відображають фактичну форму кластерів.

Алгоритми компонування графів оптимізують візуальні характеристики візуальних кодувань для створення інтуїтивно зрозумілих візуалізацій. Дослідження компонування графів показало, що користувачі досягають більшої впевненості та вищої загальної точності завдання під час взаємодії, швидко коригуючи візуальні кодування для відображення кластерів.

Хоча більшість досліджень кластеризації зосереджені на діаграмах розсіювання (ефективні для біваріантних даних), для багатовимірних даних можуть бути використані паралельні координати та матриці діаграм розсіювання. Ключові фактори, що впливають на сприйняття кластерів, включають розмір та форму символів, інтенсивність кольору (непрозорість) та просторову концентрацію. Ефективний вибір цих візуальних кодувань має вирішальне значення для оптимізації роздільності класів.

2.5. Оптимізація дизайну діаграми розсіювання на основі завдання кластеризації

Діаграми розсіювання (scatterplot) є інтуїтивно зрозумілим і широко застосовуваним інструментом візуалізації, призначеним для виявлення взаємозв'язків та закономірностей між кількісними змінними, включаючи кореляції, кластери та дослідження викидів. Діаграми розсіювання кодують двовимірні (біваріативні) дані за допомогою різних методів. Численні

дослідження продемонстрували, що візуальне кодування та аспекти даних (наприклад, обсяг) суттєво впливають на виконання візуальних завдань.

Оптимізація дизайну, що базується на виборі візуального кодування (наприклад, розміру точок або непрозорості) та аспектів даних (наприклад, кількості точок або частоти дискретизації), підвищує продуктивність користувачів у низькорівневих візуальних завданнях, покращуючи інтерпретацію та комунікацію даних. Вибір дизайнерських рішень, таких як графічне кодування або властивості даних, безпосередньо впливає на якість прийняття рішень для низькорівневих завдань, зокрема кластеризації. Кластеризація визначається як виявлення випадків, коли патерни в даних формують чіткі групи. Ефективний візуальний дизайн посилює розуміння даних, використовуючи принципи візуального сприйняття. Отже, створення структур, які одночасно враховують сприйняття візуальних кодувань та аналітичне завдання, дозволяє досягти оптимізації дизайну візуалізації для максимізації ефективності. На даний час відчувається брак уніфікованої структури, здатної надати оптимізовані дизайнерські рішення для діаграм розсіювання, які підвищують ясність даних або ефективність завдання, зокрема, ефективне сприйняття кластерів, із залученням візуальних кодувань та аспектів даних.

Надмірне накладання (overplotting) у діаграмі розсіювання прямо впливає на візуальну щільність відображення і може приховувати базові патерни даних. Візуальні дизайнери контролюють такі фактори, як розмір точок та непрозорість. Водночас розподіл даних є внутрішньою властивістю, над якою дизайнер не має прямого контролю. Контроль над кількістю точок даних обмежується переважно застосуванням підвибірки (subsampling). Підвибірка, поряд зі зменшенням розміру позначок або непрозорості, зазвичай використовується для зниження кластеризації, викликані надмірним накладанням. Оптимальна оптимізація дизайну повинна інтегрувати ці детермінанти надмірного накладання (кількість точок/частоту

дискретизації) і пропонувати дизайнеру варіанти рішень для дизайну діаграми розсіювання.

Дана робота пропонує рішення для оптимізації дизайну діаграм розсіювання, заповнюючи прогалину між теоретичними настановами та практичним застосуванням. Ми інтегруємо перцептивні настанови щодо візуального кодування та аспектів даних у діаграмі розсіювання і представляємо модель оптимізації, керовану користувачем. Ця модель пропонує оптимальний дизайн діаграми розсіювання, надаючи параметри для факторів, що впливають на візуальну щільність та техніку зменшення безладу: розмір точок даних, непрозорість, кількість точок даних та алгоритм дискретизації. Запропоноване оптимальне рішення зосереджується на завданні кластеризації і може бути масштабоване на інші низькорівневі завдання.

Інтерактивний інструмент, розроблений у цьому розділі, використовує застосування порогового графіка (threshold graph), який обчислюється на основі візуальної щільності для оптимізації дизайнерських рішень. Конверс моделі працює наступним чином:

- Дані піддаються підвибірці.
- Візуалізується діаграма розсіювання.
- Генерується пороговий графік для розмитих зображень діаграм розсіювання на основі візуальної щільності.
- Структури кластерів для заданих діаграм ранжуються для надання оптимального рішення.
- Дизайн із найвищим рейтингом з моделі вважається оптимальним вибором дизайну, що відображає помітну структуру кластерів.

Модель оптимізації, керовану користувачем, та інструмент динамічного вибору параметрів для оптимізації дизайну діаграм розсіювання на основі даних. Валідацію моделі, яка може бути використана як проксі для помітності структури кластерів. Створення інструменту веб-інтерфейсу з відкритим кодом для динамічного ранжування помітності структури

кластерів із застосуванням аспектів даних (алгоритм підвибірки та частота дискретизації) та візуальних кодувань.

Висновки до розділу

Другий розділ присвячено систематизації моделей та методів генерації специфікованих візуалізацій, орієнтованих на низькорівневі аналітичні завдання. У межах розділу розроблено розширену таксономію візуальних кодувань і завдань, що забезпечує формалізацію процесу вибору оптимального графічного представлення. Досліджено вплив окремих візуальних характеристик, таких як просторова позиція, розмір і форма, на швидкість та точність виконання користувачем завдань пошуку, фільтрації та кластеризації. Показано, що різні типи кодувань мають різний рівень ефективності залежно від структури даних і специфіки аналітичної задачі. Особливу увагу приділено задачі кластеризації, для якої визначено ключові фактори, що впливають на помітність та відокремленість груп даних у двовимірному просторі. На основі цих досліджень обґрунтовано підходи до оптимізації дизайну діаграми розсіювання, що базуються на сприйнятті та когнітивних закономірностях. Загалом результати розділу підтвердили, що систематизація і формалізація властивостей візуального кодування є критично важливими для побудови адаптивних фреймворків генерації візуалізацій.

РОЗДІЛ 3. ІМПЛЕМЕНТАЦІЯ МОДЕЛЕЙ ТА АЛГОРИТМІВ ПОБУДОВИ ФРЕЙМВОРКУ ГЕНЕРАЦІЇ СПЕЦИФІКОВАНИХ ВІЗУАЛІЗАЦІЙ

3.1. Методологія оптимізації дизайну діаграми розсіювання

3.1.1. Оптимізація дизайну та завдання кластеризації

Оптимізація дизайну візуалізації є неперервним науковим напрямом, спрямованим на підвищення її ефективності. Оптимізований дизайн передбачає вибір найкращого варіанту з множини можливих для забезпечення ефективної передачі варіацій даних та їхніх структур. В контексті візуалізації даних, такий дизайн є критично важливим для забезпечення неупередженості та достовірності висновків.

Основна мета даного дослідження полягає у розробці оптимізованого вибору дизайну для діаграм розсіювання, що максимізує помітність структури кластерів. Запропонована модель інтегрує стратегії боротьби з надмірним накладанням (*overplotting*) через підвибірку даних, застосування розмиття (*blurring*), та використання базових візуальних кодувань, таких як розмір точки та непрозорість. У цьому розділі представлено обговорення методів вибірки даних, візуального кодування, розмиття, порогового графіка (*threshold graph*) та їхньої імплементації в запропонованій моделі.

3.1.2. Огляд моделі

Запропонована модель оптимізації забезпечує інтерактивний вибір оптимального дизайну за допомогою автоматизованої, керованої користувачем параметризації, яка ґрунтується на пороговому графіку. Фундаментальна ідея полягає у використанні моделі як проксі-метрики для кількісної оцінки помітності структури кластерів залежно від дизайнерських факторів діаграми розсіювання. Оптимізована параметризація включає аспекти даних (кількість точок даних) та візуальні кодування (розмір точок та

непрозорість). Мета полягає у досягненні оптимізації параметрів, керованої користувачем, що покращує якість даних і сприяє кращому виконанню завдань прийняття рішень, зокрема кластеризації. Вихідні дані інтерактивної моделі — це ранжований набір діаграм розсіювання в порядку спадання, що відповідає оцінці помітності кластерів.

Абстрактний огляд моделі оптимізації ілюструється на рисунку 3.1 і включає такі етапи: 1) Набір даних, 2) Проекція, 3) Вилучення особливостей, 4) Оптимізований дизайн.

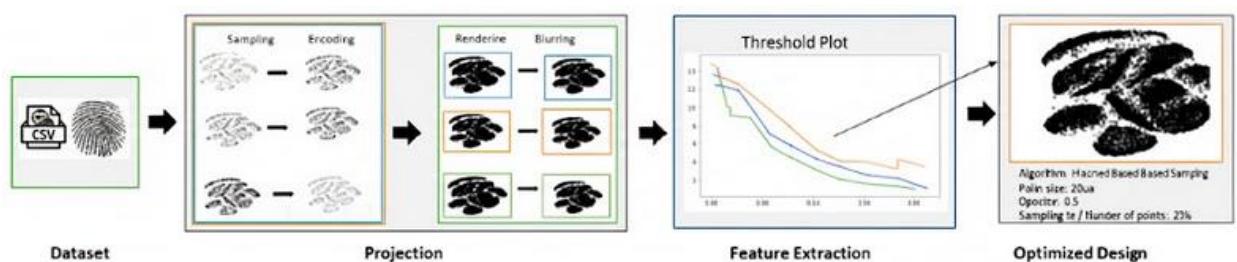


Рис. 3.1. Ілюстрація моделі оптимізації вибору дизайну

Модель на рис. 3.1 охоплює чотири етапи: введення набору даних, проєкції, вилучення особливостей з використанням порогового графіка, і, нарешті, оптимізований вибір дизайну. Етап проєкції деталізовано за допомогою 14 різних технік та частот вибірки, візуального кодування, рендерингу діаграми розсіювання та розмиття із застосуванням методу Гаусса.

Набір даних (1). Вхідними даними є вісім прикладів діаграм розсіювання з вираженим надмірним накладанням.

Проєкція (2). Цей етап охоплює кілька підетапів:

- Вибірка. Набір даних піддається підвибірці з використанням різних алгоритмів для зменшення надмірного накладання.

- Точки кодуються з використанням визначених значень розміру та непрозорості.

- Візуалізація - здійснюється рендеринг діаграм розсіювання.

- Розмиття. Зображення діаграми розсіювання розмиваються для приховування візуального шуму.

Вилучення особливостей (3). Обчислюється візуальна щільність діаграми розсіювання. Цей показник використовується в моделі для розрахунку та побудови порогових графіків.

Оптимізований дизайн (4). Смуги порогового графіка використовуються для ідентифікації помітної структури кластерів, що визначає оптимізований вибір дизайну.

3.1.3. Підвибірка даних

Підвибірка (Subsampling) є ключовою стратегією для зменшення візуального безладу (visual clutter) шляхом редукції кількості точок даних при збереженні основної структури даних. Процес передбачає вибір підмножини точок, яка є репрезентативною для всього оригінального набору. Методи підвибірки організовано за таксономією, що ґрунтується на властивостях, які зберігаються алгоритмом: просторове розділення, відносна візуальна щільність, викиди та випадковість.

Випадкова вибірка є класичним методом однорідної вибірки, що легко реалізується та не вимагає попередніх знань про набір даних. Вона використовує рівномірну стратегію, де всі зразки вибираються з однаковою ймовірністю. Хоча метод зберігає інтенсивність різниці щільності, він може призвести до втрати локальних або глобальних трендів та важливих викидів, особливо у випадках надзвичайно високої щільності, де необхідне агресивне співвідношення вибірки.

Методи цієї категорії спрямовані на забезпечення кращого просторового розділення між класами/кластерами. Сюди входять:

- Вибірка за синім шумом (Blue Noise Sampling) - створює рівномірно розподілені зразки в просторі, що є важливим для ідентифікації базових патернів.

- Вибірка за найдалшою точкою (Farthest Point Sampling) - ітеративно вибирає зразки з максимальною мінімальною відстанню до раніше обраних, забезпечуючи оптимальне просторове розділення.

- Багатокласова вибірка за синім шумом - розширення методу для підтримки властивостей синього шуму в кожному класі.

Головною особливістю є підтримка просторового розподілу, але ці алгоритми є часовитратними, оскільки час вибірки зростає пропорційно до кількості точок даних.

Методи цієї категорії прагнуть зберегти відносну візуальну щільність у щільних і розріджених регіонах, запобігаючи втраті інформації в останніх. Алгоритми зберігають властивості, пов'язані зі щільністю.

Збереження щільності формалізовано як:

$$\sum_{i=1}^{n_i} x_i P(\bar{x}_j) = K n_i$$

де n_i — розмір групи, $P(\bar{x}_j)$ — ймовірність вибірки точки, а K — константа.

Відносна щільність даних між двома регіонами Ω_A та Ω_B (з однаковою площею) визначається функцією ϕ :

$$\phi(D^0(\Omega_A), D^0(\Omega_B)) = \begin{cases} 1 & \text{якщо } (D^0(\Omega_A) > D^0(\Omega_B)) \\ 0 & \text{якщо } (D^0(\Omega_A) = D^0(\Omega_B)) \\ -1 & \text{якщо } (D^0(\Omega_A) \leq D^0(\Omega_B)) \end{cases}$$

Ключові методи:

1. Вибірка, що враховує щільність (Density-biased sampling) - ймовірно перевибіркує розріджені регіони та недооцінює щільні.

2. Вибірка SVD (Singular Value Decomposition) - вибирає зразки з найбільшою кореляцією з верхніми базовими векторами.

3. Нерівномірна вибірка. Націлена на збереження або збільшення різниці щільності регіонів.

Ці підходи можуть призводити до значної втрати інформації (наприклад, рекурсивна вибірка на основі поділу) або бути часовитратними.

Метою збереження викидів є підвищення ймовірності включення викидів у вибірку. Точки даних у регіонах з низькою щільністю часто вважаються викидами. Але збереження викидів може спотворювати відносні щільності даних, оскільки більше точок вибирається в регіонах з низькою щільністю.

Новітні підходи (наприклад, Z-послідовності, багатокласова вибірка за синім шумом) спрямовані на збалансоване вирішення надмірного накладання, зберігаючи як викиди, так і відносні щільності, але вони можуть створювати оманливі представлення у регіонах перетину класів.

3.2. Використання концепції порогового графіка для оцінки кластеризації при візуалізації

3.2.1. Пороговий графік

Пороговий графік, заснований на дереві злиття (merge tree) щільності, є моделлю, керованою щільністю, яка спрямована на пряму ідентифікацію відносної візуальної щільності, при якій користувачі здатні розрізняти кластери. Ця модель ґрунтується на обчисленні дерева з'єднання (join tree) скалярного поля.

Пороговий графік використовується для обчислення порогового значення (ρ), при якому кількість кластерів у заданій діаграмі розсіювання (X) стає сприйнятною для користувача.

Однак, його пряме застосування є недостатньо обмеженою оптимізацією, оскільки воно не враховує помітність (salience) цих кластерів і, принаймні, вимагає попереднього визначення користувачем кількості кластерів.

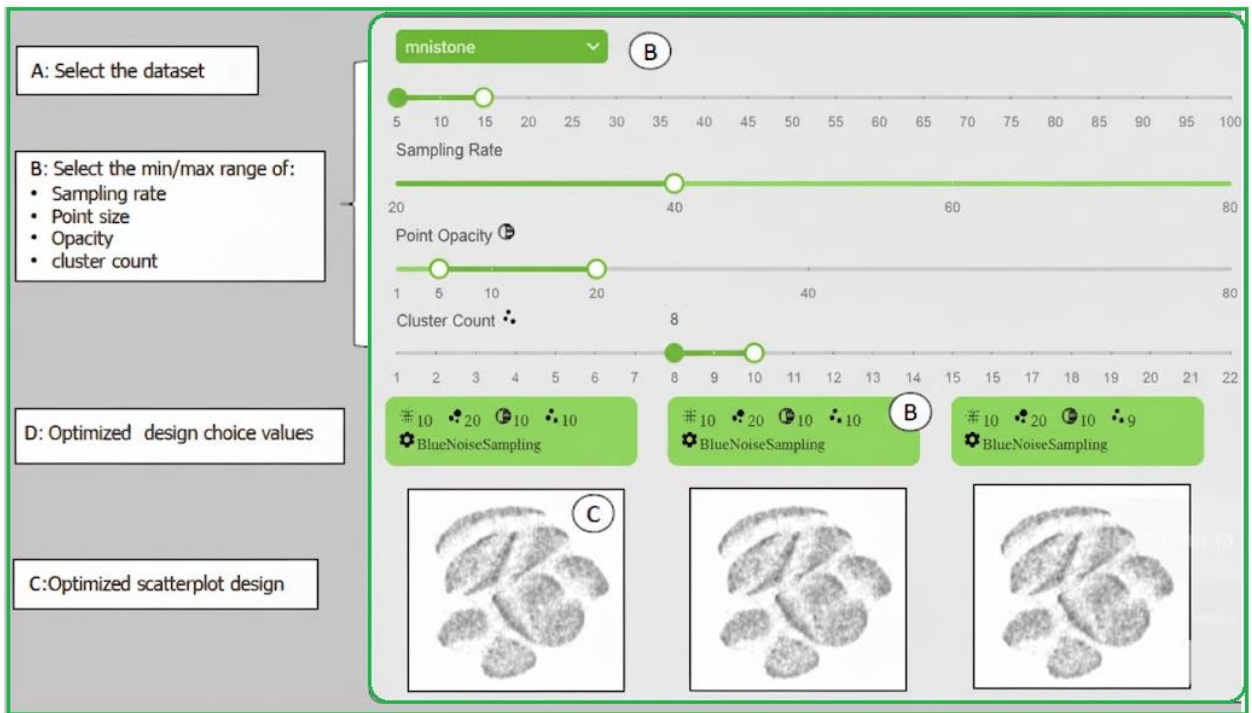


Рис. 3.2. Демонстрація інтерактивної моделі, керованої користувачем

Рис. 3.2 ілюструє етапи інтерактивної моделі для отримання оптимальних виборів дизайну. (А) Користувач обирає один із восьми наборів даних. (В) Користувач визначає мінімальний та максимальний діапазон параметрів, включаючи частоту дискретизації (sampling rate), розмір точок, непрозорість точок та кількість кластерів. (С) На основі вибору модель представляє оптимальний дизайн діаграми розсіювання у порядку спадання їхнього значення помітності (saliency value). (D) Кожна діаграма розсіювання представляє набір оптимальних параметрів дизайну

Таким чином, у цій моделі структура кластеризації кодується як функція щільності в дереві злиття. Для визначення кількості сприйнятих кластерів ми генеруємо пороговий графік стійкості (persistence threshold plot). Для заданого дерева злиття кожен компонент має свій поріг (ρ) (рис. 3.2). Стійкість вимірює відносний масштаб особливості (наприклад, відносну зміну щільності), а не її абсолютне значення, що є ключовим для ідентифікації значущих кластерів.

Автоматична параметризація моделі оптимізації використовує пороговий графік для керування дизайнерськими рішеннями. Пороговий

графік застосовує модель дерева злиття на основі щільності з покращеними функціями приховування шуму та викидів. Ми постулюємо, що порогове значення на пороговому графіку (рисунок 3.3), обчислене на візуальній щільності діаграми розсіювання, може слугувати проксі-метрикою для людського сприйняття помітності структури кластера (рисунок 3.4).

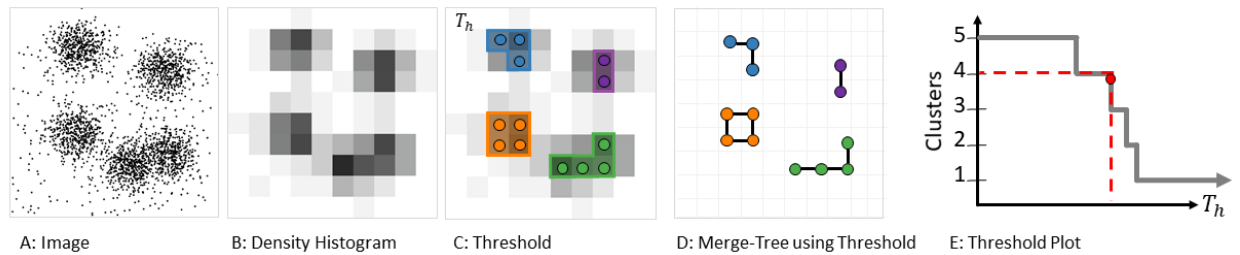


Рис. 3.3. Демонстрація генерації порогового графіка з використанням візуальної щільності

Використовуючи зображення діаграми розсіювання та модель на основі щільності:

А) на вхідні дані розмитого зображення діаграми розсіювання

В) обчислюється гистограма щільності.

С) Простір аналізується при різних значеннях щільності, і компоненти вилучаються.

Д) Відстеження компонентів уздовж значень щільності призводить до створення дерева злиття (merge tree), яке позначає топологічні події вузлами.

Е) Створюється пороговий графік (threshold plot) для компонента щільності, заснованого на дереві злиття. Горизонтальна вісь представляє поріг (ρ), тоді як вертикальна вісь показує кількість кластерів. Червона лінія ілюструє, як поріг може бути вилучений із заданої кількості кластерів і навпаки.

Чітка структура кластера ідентифікується як більш помітна за умови меншої кількості шуму та викидів у розподілі даних. Значення порогових смуг (threshold bars) аналізуються за їхнім максимумом для вибору найбільш

помітної структури кластера. Модель відображає дев'ять оптимальних варіантів для кожної ітерації.

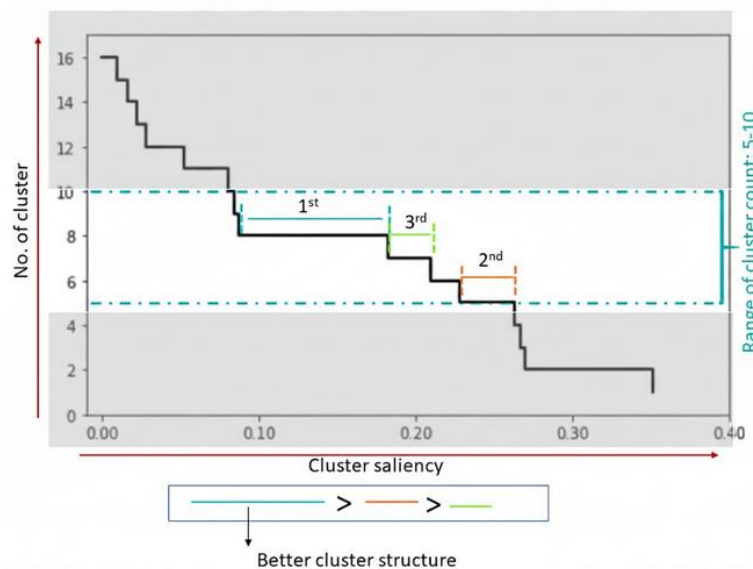


Рис. 3.4. Пороговий графік для помітності структури кластера

На рис. 3.4 горизонтальна вісь представляє поріг (ρ), тоді як вертикальна вісь показує кількість кластерів. Смуга діапазону відображає ранжування помітності кластера при заданому порозі. Помітна структура кластера може бути ідентифікована з діапазону кластера. Як продемонстровано, існують три порогові смуги, але найдовша/найширша смуга представляє найбільш помітну структуру кластера.

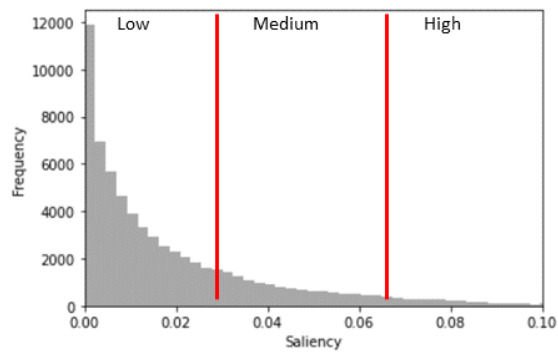
Для демонстрації моделі розроблено веб-інтерфейс, що дозволяє інтерактивно обирати оптимальний дизайн діаграми розсіювання на основі чіткої помітності структури кластера. Оптимізований дизайн визначається через набір параметрів: кількість точок (частота дискретизації), розмір точок, непрозорість точок та методи дискретизації. В роботі використовуються вісім наборів даних.

Інтерфейс моделі надає результати у формі ранжування діаграм розсіювання за їхньою помітністю структури кластера, декодуючи це ранжування в оптимізований вибір параметрів дизайну: значення розміру точок, непрозорості, частоти дискретизації та методу підвибірки.

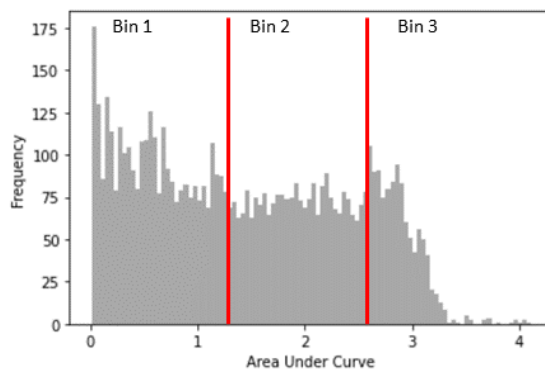
3.2.2. Порогова смуга графіка як помітність структури кластера

Довжина крокової смуги порогового графіка (рис. 3.3 Е) використовується як значення помітності для заданої кількості кластерів, що слугує для подальшої ідентифікації помітної структури кластера (рис. 3.4). Помітна структура кластера може бути ідентифікована з діапазону значень кількості кластерів. Наша мета — вибрати дизайн, що надає найбільш чітку структуру кластера.

На рисунку 3.4 спостерігаються три видатні порогові смуги з різною довжиною, які можна класифікувати як високу, середню та низьку помітність. Для інтеграції цієї техніки була побудована гістограма (рис. 3.5 а) для значень помітності для кожного набору даних, з подальшим бінінгом (розподілом за категоріями) на високу (В), середню (С) та низьку (Н) помітність. Наприклад, для MNIST: 0.0–0.33 (Н), 0.34–0.66 (С), > 0.66 (В).



а) Розподіл помітності



б) Розподіл значень площі під кривою

Рис. 3.5. Бінінг для розподілу помітності та площі під кривою (Area Under Curve, AUC)

Дві візуально подібні діаграми розсіювання, що потенційно представляють схожу структуру кластерів, можуть бути перцептивно подібними. У цій моделі дві діаграми розсіювання вважаються перцептивно подібними, якщо їхні порогові графіки схожі. Для кількісної оцінки перцептивної подібності структури кластерів використовується міра різниці площі під кривою (AUC) (рис. 3.6), яка вимірює та сумує різницю між пороговими графіками на кожній позиції значення кластера.

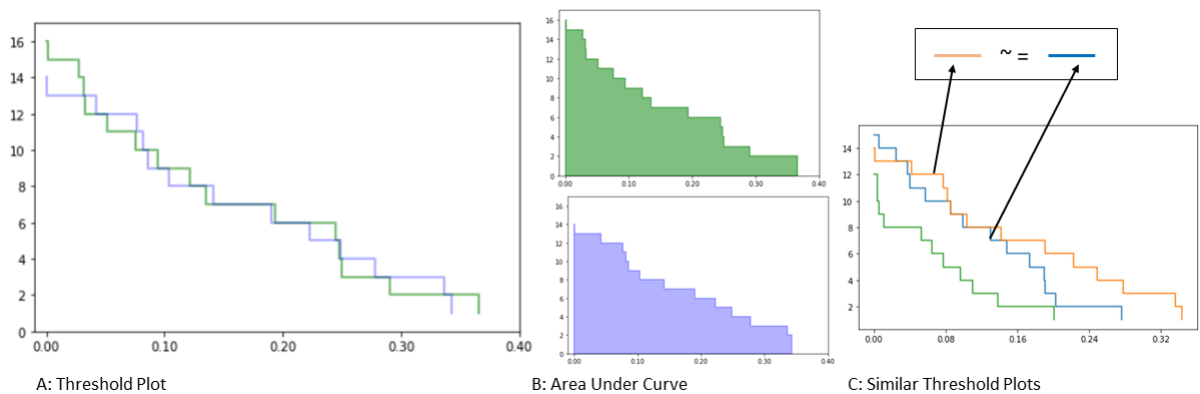


Рис. 3.6. Порогові графіки, що демонструють: подібні патерни є перцептивно схожим

Формула для обчислення ℓ_1 норми різниці між двома пороговими графіками X та Y :

$$\ell_1(X, Y) = \sum_{i=1}^n |x_i - y_i|$$

Значення різниці AUC застосовується для визначення рівня подібності між двома структурами кластерів. Для інтеграції цієї техніки побудовано гістограму (рис. 3.5 б) значень AUC для кожного набору даних із бінінгом розподілу на три категорії. Наприклад, для MNIST: 0.0–1.1 (Бін 1), 1.2–2.5 (Бін 2), > 2.5 (Бін 3). Це дозволяє визначити три критерії подібності:

- Подібні (SR) - діаграми розсіювання належать до одного біну.

- Дещо подібні (SS) - діаграми розсіювання належать до сусідніх бінів (наприклад, 1 і 2, або 2 і 3).
- Неподібні (DS) - 0 діаграми розсіювання належать до бінів 1 і 3.

3.3. Розробка та імплементація інтерактивної моделі оптимізації

Далі представлена розробка інтерактивного інструменту для демонстрації оптимізації візуальних кодувань та аспектів даних на реальних наборах. Ця робота вирішує проблему надмірного накладання (*overplotting*) у діаграмах розсіювання, використовуючи різні методи підвибірки (*sampling*) у поєднанні з розміром точок та непрозорістю. Мета — забезпечити кращу помітність (*saliency*) для чіткої структури кластера, при цьому значення помітності структури кластера обчислюється на основі дизайнерських факторів за допомогою порогового графіка.

Тут детально описуються етапи, проілюстровані в огляді моделі на рисунку 3.1.

3.3.1. Вибір наборів даних

Для забезпечення надійності результатів моделі було відібрано вісім репрезентативних наборів даних із попередніх досліджень у сфері візуалізації. Оскільки більшість із них є багатовимірними (*high-dimensional*), їх спочатку було трансформовано у двовимірні (2D) дані за допомогою алгоритму t-SNE та нормалізовано до діапазону $[0,1] \times [0,1]$. Набори даних класифіковані за характеристиками:

- Кластерно-орієнтовані (6 наборів),
- 3 вигнутими смугами (2 набори).

3.3.2. Етап проєкції

Цей етап деталізує візуалізацію зображень діаграм розсіювання з 2D вхідних даних, включно з механізмом зменшення надмірного накладання.

Набори даних піддавалися підвибірці з використанням 14 різних алгоритмів, які класифіковані за чотирма категоріями:

1. Випадкова вибірка (Random sampling).
2. З упередженням щільності (Density-biased sampling), нерівномірна (Non-uniform), SVD-заснована, багатовидова за Z-послідовністю (Multivariate Z-sequence), рекурсивна на основі поділу (Recursive partition-based).
3. З упередженням викидів (Outlier-biased) - на основі щільності, випадкова, на основі хеш-таблиці, за синім шумом.
4. Збереження просторового розділення - багатокласова за синім шумом (Multi-class Blue Noise), за синім шумом, за найдалшою точкою (Farthest Point), за Z-послідовністю.

Кожен із цих алгоритмів зберігає одну або кілька особливостей: візуальну щільність, відносну щільність класів, викиди. Для оцінки продуктивності моделі були обрані частоти дискретизації (ЧД) (Sampling Rate, SR) в діапазоні 0.05–0.95 з інтервалом 0.05, що генерує 19 наборів підвибіркових файлів для кожного набору даних. Вибір ЧД змінює кількість точок на екрані, що, як відомо, впливає на візуальні завдання, такі як усереднення та кластеризація.

Дані кодуються як позначки точок (кола). У дослідженні використовуються два ключові візуальні кодування:

- Розмір точки (РТ) (Point Size, PS): {20px, 40px, 60px, 80px}.
- Непрозорість Точки (НТ) (Point Opacity, OP): {1%, 5%, 10%, 20%, 40%, 80%}.

Діаграми розсіювання візуалізуються з використанням комбінацій ЧД×РТ×НТ для всіх наборів даних. Вибрані параметри візуалізації:

- Розміри зображення (X×Y): [700px×700px].
- Розмір/площа точки даних (РТ): {20px, 40px, 60px, 80px}.
- Непрозорість точки даних (НТ): {1%, 5%, 10%, 20%, 40%, 80%}.
- Частота дискретизації (ЧД): Діапазон {5%–95%} з інтервалами {5%}.

Для зменшення шуму та викидів, які маскують структуру кластера, було застосовано гаусове розмиття до візуалізованих зображень діаграм розсіювання. Пороговий графік обчислюється не на вихідному, а на розмитому зображенні для покращення помітності структури кластера.

3.3.3. Вилучення особливостей

Основною особливістю моделі є візуальна щільність. Для моделювання на візуальній щільності використовується структура даних топологічного аналізу даних — дерево злиття (merge tree). Ця модель прагне безпосередньо ідентифікувати ту відносну візуальну щільність, при якій користувачі здатні розрізнити структури кластерів, ґрунтуючись на дереві з'єднання скалярного поля. Моделі намагаються захопити людське сприйняття кластерів, враховуючи візуальну щільність, при якій розподіли кластерів зливаються, щоб сприйматися як єдине ціле.

Для візуалізації структури кластеризації, згенерованої моделями, використовуються порогові графіки (рисунок 3.3), які відображають візуальну помітність проти кількості видимих кластерів. Пороговий графік демонструє рівень помітності для заданого дизайну діаграми розсіювання на основі факторів: алгоритмів вибірки, частоти дискретизації, розміру точок та їхньої непрозорості.

Оптимізованим дизайном вважається той, що має максимальне порогове значення серед усіх порогових значень діаграм розсіювання, згенерованих у роботі.

Кінцеві результати являють собою ранжований порядок дизайну діаграми розсіювання, заснований на помітності структури кластера. Важливо відзначити, що багато дизайнів можуть демонструвати подібні або перцептивно схожі значення помітності. Модель пропонує метод для вибору найкращої помітної діаграми розсіювання або комбінації факторів для заданої кількості кластерів або набору даних.

3.4. Аналіз обчислювальної ефективності моделі оптимізації

Модель оптимізації застосовує 14 алгоритмів підвибірки до восьми наборів даних із високим рівнем надмірного накладання. Для оцінки операційної та обчислювальної складності проведено аналітичний та обчислювальний аналіз часу виконання операцій.

3.4.1. Час обчислення для підвибірки даних

Вибрані вісім наборів даних мають змінну кількість точок, що варіюється від $N=4177$ до $N=70000$ (MNIST). Кількість точок для ключових наборів даних: MNIST ($N=70000$), умовне технічне обслуговування ($N=10000$), одяг ($N=26569$), краудсорсингове картографування ($N=10845$), епілептичні напади ($N=11500$), швейцарський рулет 2D ($N=8000$), швейцарський рулет 3D ($N=10000$), абалон ($N=4177$).

Деякі алгоритми вибірки, зокрема вибірка за синім шумом, є часовитратними для підвибірки великих наборів. Для аналізу було зафіксовано час завершення підвибірки для кожного з восьми наборів даних із застосуванням 14 алгоритмів при частотах дискретизації (ЧД) від 5% до 95% з інтервалом 5%. Оскільки час виконання був схожим для більшості наборів даних, результати детально обговорюються лише для набору MNIST.

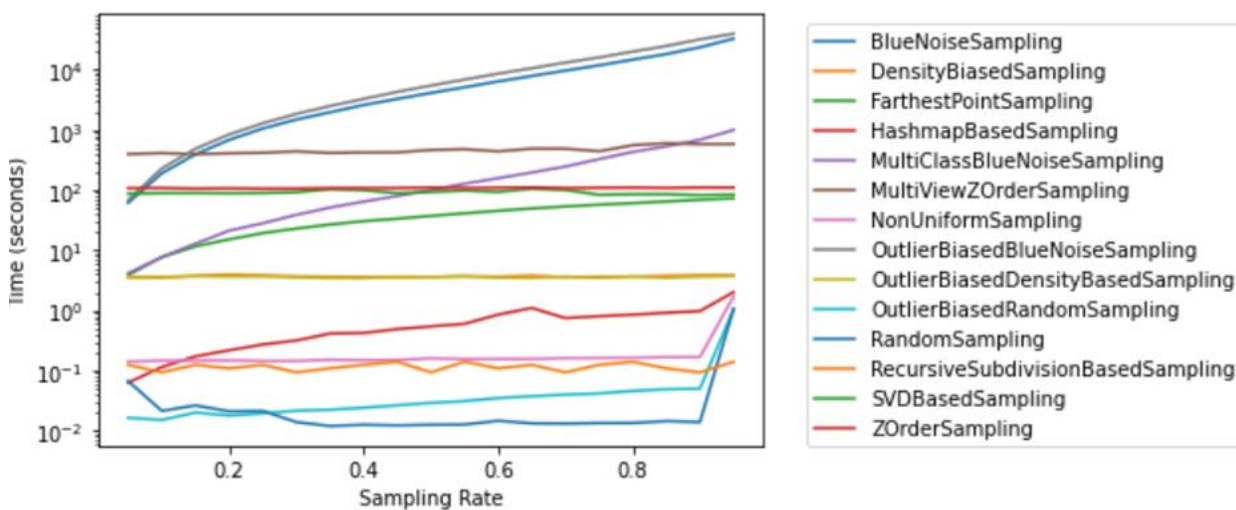


Рис. 3.7. Аналіз часу підвибірки даних для 14 алгоритмів вибірки

Рисунок 3.7 ілюструє час завершення підвибірки для набору даних MNIST для 14 алгоритмів. Продуктивність підвибірки можна класифікувати за часом завершення:

- Низький час: випадкова вибірка, випадкова вибірка з упередженням викидів.
- Середній час: вибірка на основі SVD, вибірка за найдальшою точкою.
- Високий час: вибірка за синім шумом, вибірка за синім шумом з упередженням викидів.

Крім того, деякі алгоритми (наприклад, нерівномірна вибірка, вибірка з упередженням щільності викидів, випадкова вибірка) демонструють рівномірний час завершення незалежно від частоти дискретизації. Натомість час завершення для алгоритмів, таких як вибірка за синім шумом та вибірка за синім шумом з упередженням викидів, суттєво зростає зі збільшенням частоти дискретизації.

3.4.2. Час Обчислення для проєкції та вилучення особливостей

Після етапу підвибірки слідує етапи візуалізації, розмиття та вилучення особливостей за допомогою порогового графіка, які також є обчислювально інтенсивними.

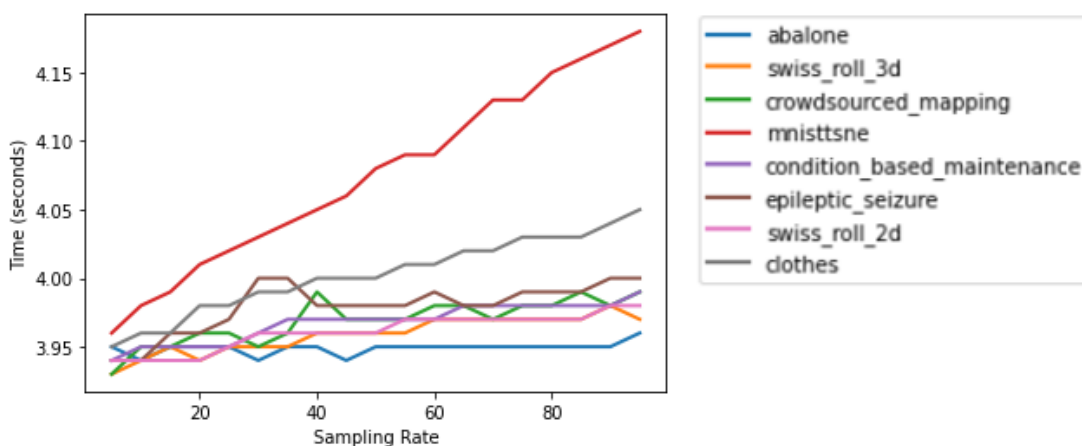


Рис. 3.8. Аналіз часу для проєкції та вилучення особливостей (використано вісім наборів даних)

Обчислений та зафіксований час охоплює повний набір операцій для кожного зображення діаграми розсіювання. Повний набір складається з 14 (алгоритмів вибірки) \times 19 (частот дискретизації) \times 4 (значення розміру точок) \times 5 (значення непрозорості точок).

Як показано на рисунку 3.8, обчислений час є пропорційним до кількості точок даних у кожному наборі. Таким чином, обчислювальні витрати на етапи проєкції та вилучення особливостей лінійно залежать від розміру обробленого набору даних.

Максимальна кількість точок у первинному списку наборів даних становить 70000 (MNIST). Для подальшої оцінки масштабованості моделі на більших обсягах даних було обрано набір BitcoinHeist [4], що містить приблизно 3 мільйони точок даних.

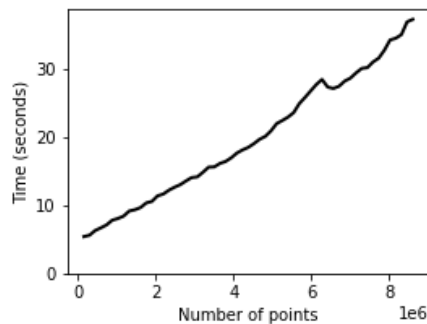


Рис. 3.9. Аналіз часу масштабованості моделі

Було обчислено та зафіксовано час обчислення для етапів проєкції та вилучення особливостей для цього великого набору даних (як описано на рис. 3.1). Тенденція, відображена на рис. 3.9, демонструє лінійну залежність часу обчислення від кількості точок. Це свідчить про те, що масштабованість запропонованої моделі є лінійною щодо розміру вхідного набору даних.

3.4. Валідація моделі

Для емпіричної валідації результатів запропонованої моделі та детального дослідження продуктивності дизайну діаграм розсіювання на

реальних даних було розроблено та проведено дослідження на платформі Amazon Mechanical Turk (АМТ).

Дослідження базується на двох основних гіпотезах, пов'язаних із застосуванням порогового графіка як проксі-метрики для людського сприйняття.

Гіпотеза_1 (Г1). Чим більша довжина порогової смуги, тим вищою є помітність структури кластера. Пороговий графік використовується як проксі для оцінки того, які дизайни діаграм розсіювання забезпечують більш помітну структуру. Довжина смуги обчислюється на основі візуальної щільності, яка, у свою чергу, залежить від змінних дизайну: алгоритму та частоти вибірки, розміру точок та непрозорості. Ми очікуємо, що вища довжина смуги корелює з кращою продуктивністю сприйняття кластерів.

Гіпотеза_2 (Г2). Діаграми розсіювання зі схожими патернами порогових графіків є перцептивно подібними і мають схожу структуру кластера. Діаграми розсіювання з подібною формою порогових графіків мають схожу візуальну щільність та візуальне розділення. Таким чином, візуальне та перцептивне порівняння помітності кластерів у таких діаграмах буде складним і двозначним, що може бути виміряно за допомогою показника різниці площі під кривою (AUC).

Із восьми початкових наборів даних було вибрано шість.

Зображення діаграм розсіювання візуалізувалися з використанням таких параметрів:

- Розміри стимулів ($X \times Y$): [700px × 700px].
- Розмір точки (РТ): {20px, 40px, 60px, 80px}.
- Непрозорість точки (НТ): {1%, 5%, 10%, 20%, 40%, 80%}.
- Частота дискретизації (ЧД): У діапазоні {5%–95%} з інтервалом 5%.

На основі гіпотез Г1 та Г2 було розроблено два типи завдань:

31. Порівняння помітності (валідація Г1). Запитання: Яка з показаних діаграм розсіювання має чіткішу структуру кластера?

Процедура наступна. Користувачеві демонструються два дизайни діаграм розсіювання. Кожна діаграма має значення помітності, обчислене моделлю. Відповідь учасника оцінюється як 1, якщо обрано діаграму з вищим значенням помітності (цільова), і 0 в іншому випадку. Гіпотеза полягає в тому, що користувач обере діаграму з вищою помітністю, підтверджуючи, що пороговий графік є дійсним проксі для людського сприйняття.

32. Порівняння перцептивної подібності (валідація Г2). Запитання: Яка з діаграм розсіювання має більш подібну структуру кластера до еталонної діаграми розсіювання?

Процедура наступна. Користувачеві демонструються еталонна діаграма та дві інші діаграми розсіювання (варіанти А та В). Учасник обирає дизайн, найбільш схожий на еталонний. Відповідь оцінюється як 1, якщо обрано діаграму з меншою різницею AUC від еталонної (цільова), і 0 в іншому випадку. Це перевіряє припущення, що подібні патерни порогових кроків відповідають перцептивно подібним структурам.

Стимули генерувалися відповідно до визначених бінів помітності (для 31) та перцептивної подібності (для 32).

Завдання 31. Використовувалися три біни помітності: висока (В), середня (С), низька (Н). Було сформовано шість комбінацій для порівняння: В×В, В×С, В×Н, С×Н, С×С, Н×Н. Кожен набір даних (D=6) був використаний для двох випадково вибраних комбінацій, що дало 12 стимулів для 31.

Завдання 32. Використовувалися три категорії перцептивної подібності на основі бінів AUC: подібні (П) (Similar, SR), дещо подібні (ДП) (Somewhat Similar, SS) та неподібні (НП) (Dissimilar, DS). комбінації порівняння були:

- Варіант А (П до Еталона) проти Варіант В (НП до Еталона)
- Варіант А (П до Еталона) проти Варіант В (ДП до Еталона)
- Варіант А (ДП до Еталона) проти Варіант В (НП до Еталона).

Кожен набір даних (D=6) був використаний для однієї випадково вибраної комбінації, що дало 6 стимулів для 32. Загалом, кожен учасник отримав 18 стимулів (12 31 та 6 32) у випадковому порядку.

Для аналізу результатів використано змішані моделі біноміального логарифмічного відношення шансів (mixed-effects binomial logistic regression). Моделі підганялися за допомогою пакета glmmTMB у R.

Результати підтверджують Г1: довжина порогової смуги суттєво впливає на сприйняття чіткості структури кластера. Учасники значно частіше обирали цільову діаграму розсіювання як чіткішу, коли різниця в довжинах порогових смуг між двома діаграмами була великою. Цей ефект був особливо вираженим, коли значення довжини порогової смуги цільової діаграми розсіювання було високим.

Таблиця 3.1.

Параметри моделі для завдання 1

Змінна	Оцінка (β)	SE	95% CI
Перетин	0.71	0.26	[0.20, 1.23]
Довжина цільового	0.27	0.10	[0.07, 0.48]
Довжина порівняння	-0.69	0.19	[-1.07, -0.31]
SD (Перетин: Учасник)	0.29	-	[0.09, 0.94]
SD (Перетин: Набір даних)	0.23	-	[0.08, 0.64]

Таблиця 3.2.

Прогнози моделі для завдання 1

Довжина цільового (приклад)	Довжина порівняння (приклад)	Різниця	Ймовірність	SE	95% CI
Низька [0.15]	Низька [0.15]	0.00	0.66	0.06	[0.54, 0.76]
Середня [0.50]	Середня [0.50]	0.00	0.62	0.06	[0.51, 0.73]
Висока [1.50]	Висока [1.50]	0.00	0.52	0.07	[0.38, 0.66]
Середня [0.50]	Низька [0.15]	0.35	0.68	0.05	[0.56, 0.77]
Висока [1.50]	Середня [0.50]	1.00	0.68	0.05	[0.57, 0.78]
Висока [1.50]	Низька [0.15]	1.35	0.73	0.05	[0.62, 0.82]

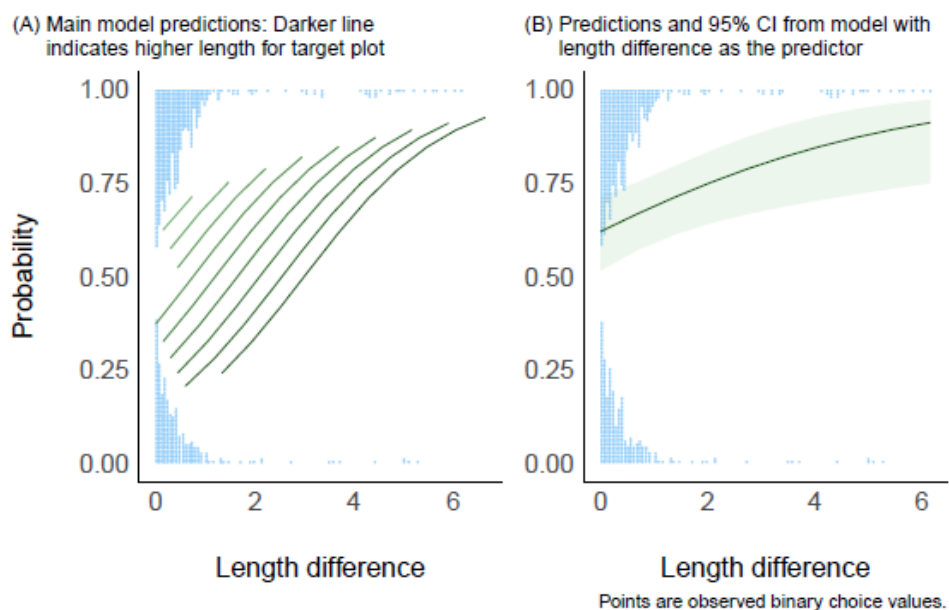


Рис. 3.10. Аналіз дослідження користувачів для завдання 1

Результати підтверджують Г2: різниці AUC суттєво впливають на сприйняття подібності. Учасники значно частіше обирали цільову діаграму розсіювання (з меншою різницею AUC від еталонної) як більш подібну, коли різниця в значеннях AUC між двома діаграмами була великою. На відміну від 31, цей ефект не змінювався суттєво залежно від абсолютних рівнів AUC цільової діаграми розсіювання.

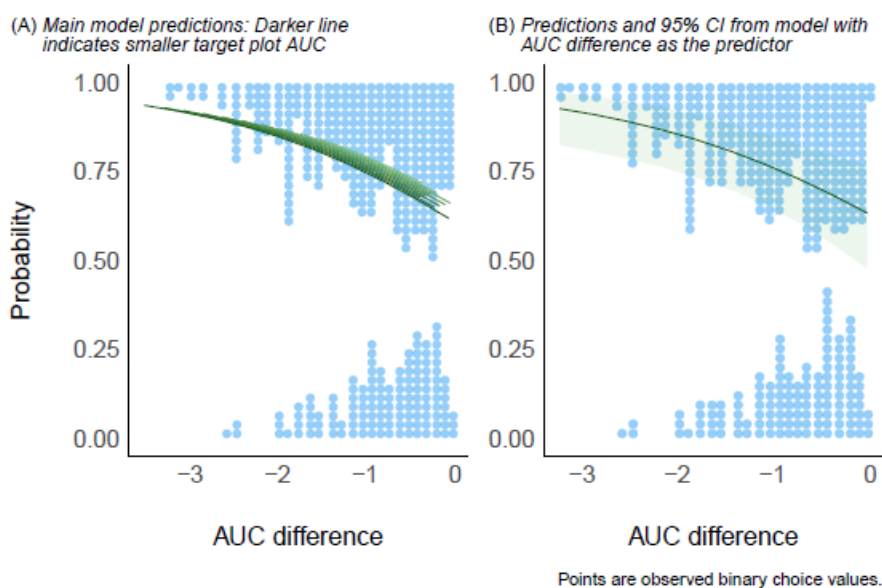


Рис. 3.11. Аналіз дослідження користувачів для завдання 2

3.5. Оцінка якості результатів та корисності інтерфейсу

Для оцінки якості результатів, корисності та ефективності розробленого інтерактивного інтерфейсу та моделі автоматичної параметризації було проведено кейс-стаді з порівнянням функціональності автоматичної параметризації, керованої користувачем, та інтерфейсу ручної параметризації.

Як ілюструє рисунок 3.12, ручна оптимізація вимагала значно більшої кількості взаємодій (де взаємодія визначалася як вибір значень факторів для кожної ітерації) для досягнення результату. Водночас, з точки зору часу, різниця між режимами не була значною. Тривалість залучення в автоматичній моделі пояснюється необхідністю перегляду дев'яти ранжованих оптимальних дизайнів. Крім того, було встановлено, що результати, обрані учасниками в автоматичному режимі, відповідали смугам помітності, згенерованим моделлю.

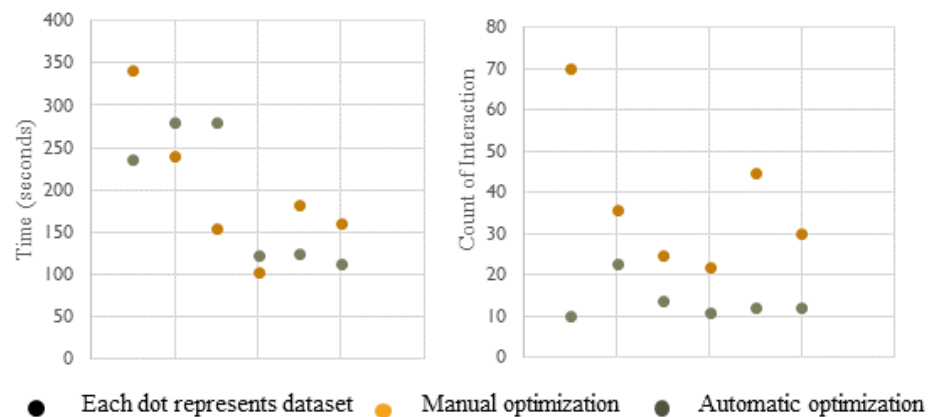


Рис. 3.12. Продуктивність учасників кейс-стаді за часом та кількістю взаємодій

Розроблений інтерфейс моделі функціонує як структура, керована даними, що дозволяє порівнювати та спостерігати, як патерни даних реагують на візуальні кодування (розмір точок, непрозорість) та аспекти

даних (частота дискретизації). Чіткість даних для виконання користувацьких завдань залежить від внутрішніх особливостей набору даних.

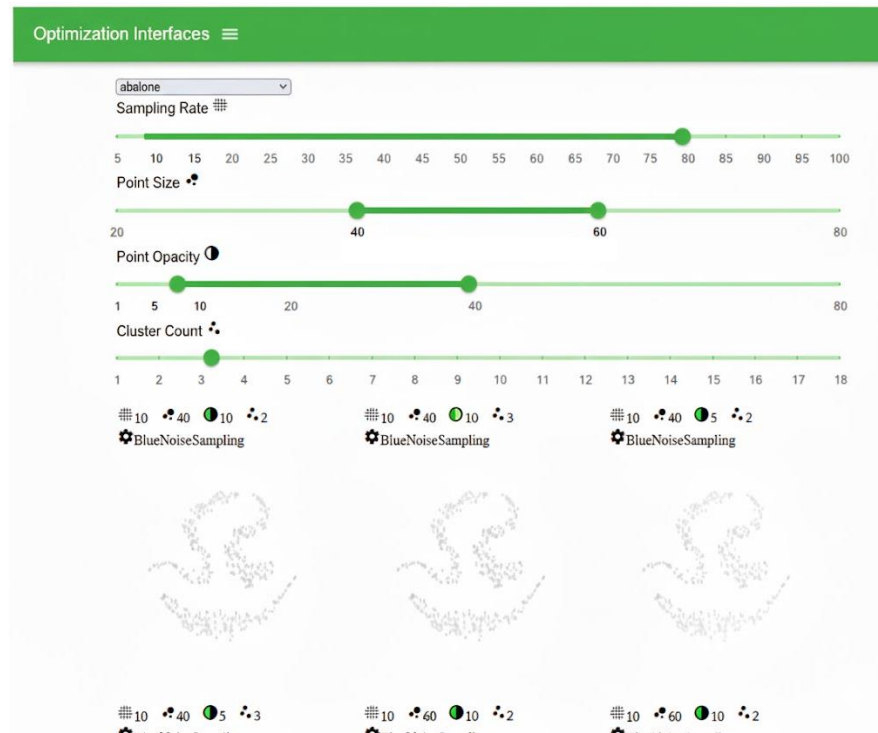


Рис. 3.13. Інтерфейс користувача із вибраним набором даних “abalone” для візуалізації

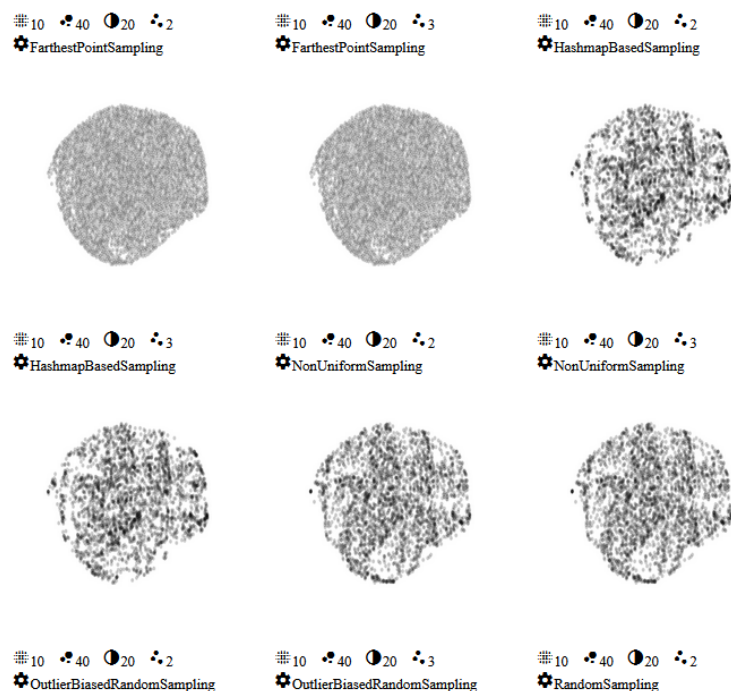


Рис. 3.14. Приклади візуалізацій для набору даних “credit_card”

Модель демонструє це через ранжування дизайнів діаграми розсіювання за помітністю структури кластера, використовуючи комбінації частоти дискретизації, розміру точок, непрозорості та методів вибірки. Вихід інтерактивної моделі — це набір діаграм, впорядкованих за спаданням значення помітності кластера, де кожне значення кодує оптимальні параметри дизайну.

Метою розділу є пропозиція оптимізованого дизайну візуалізації для підвищення ефективності виконання завдань. Теоретичні моделі сприйняття стверджують, що близькість, кількість, концентрація та зміна щільності впливають на сприйняття кластера. Пороговий графік обчислюється на основі оцінки візуальної щільності. Довжина порогових смуг, яка відображає помітність кількості кластерів за заданої щільності, збільшується для більш чіткої структури. Найдовші порогові смуги ідентифікуються як індикатори найбільш помітної структури кластера.

Внутрішні властивості набору даних, такі як розподіл, можуть впливати на візуалізацію, призводячи до двозначності, навіть при оптимальному виборі дизайну. У моделі продемонстровано, що порогові графіки, які мають подібні кроки/структуру та близькі в просторі, представляють перцептивно подібні діаграми розсіювання та структуру кластера. Ідентифікація помітної структури кластера залишається двозначною для таких наборів даних, і оптимізація дизайну має мінімальний вплив на їхню візуалізацію.

Інтерфейс оптимізації, керований користувачем, надає дев'ять конфігурацій діаграм розсіювання, ранжованих за помітністю. Оптимальний вибір дизайну полягає у виборі найкращої конфігурації з цього набору підходящих варіантів.

Підвибірка даних є першим часозатратним етапом. Алгоритми, що зберігають просторове розділення (наприклад, синій шум), вимагають більше часу зі збільшенням частоти дискретизації, тоді як інші (наприклад, випадкова вибірка) демонструють сталий час виконання.

Отже, запропоновано інструмент автоматичної оптимізації, керований користувачем, для оптимізації факторів дизайну діаграм розсіювання з метою підвищення помітності структури кластера. Пропонований інтерактивний інструмент використовує структуру даних дерева злиття (merge tree) для керування рішеннями щодо дизайну на основі частоти дискретизації, алгоритмів вибірки, розміру символів та непрозорості. Результати були емпірично валідовані, що дозволило сформулювати вказівки, які можуть бути розширені практиками та дизайнерами на інші завдання, пов'язані з діаграмами розсіювання.

Висновки до розділу

У третьому розділі було реалізовано модель і алгоритмічний фреймворк генерації специфікованих візуалізацій, побудований на основі концепцій, розроблених у попередніх розділах. Запропоновано методологію оптимізації діаграми розсіювання, що враховує особливості даних та показники помітності кластерних структур. Використання підвибірки, проєкції та вилучення особливостей дало змогу створити ефективний алгоритмічний конвеєр для оцінки якості графічного представлення. Пороговий графік та концепція порогової смуги були застосовані як інструменти вимірювання ступеня видимості кластерів, що забезпечило кількісну оцінку ефективності візуалізації. Розроблена інтерактивна модель оптимізації дозволяє динамічно змінювати параметри візуального кодування, забезпечуючи адаптацію під конкретну конфігурацію даних та потреби користувача. Проведений аналіз обчислювальної ефективності показав, що запропоновані алгоритми можуть бути застосовані в умовах інтерактивного аналізу. Результати валідації підтвердили практичну корисність фреймворку та його здатність покращувати точність та зручність аналізу даних, виконуваного користувачами.

ВИСНОВКИ

У магістерській роботі здійснено дослідження моделей, методів та підходів до побудови фреймворку генерації специфікованих візуалізацій, у якому інтегровано теоретичні засади візуального сприйняття, сучасні таксономії візуалізації даних та алгоритмічні компоненти для оптимізації дизайну графічних представлень. Проведене дослідження дало змогу сформуванню науково обґрунтовану методологію побудови таких візуалізацій, яка орієнтована на підвищення якості аналітичної взаємодії користувача з даними.

У першому розділі проведено ґрунтовний аналіз предметної області візуалізації даних. Показано, що візуалізація є не лише засобом представлення даних, а й ключовим когнітивним механізмом, який дозволяє підвищити ефективність аналітичної діяльності. На основі аналізу ролі дизайну, структурних особливостей візуального представлення та механізмів людського сприйняття встановлено, що якість візуалізації визначається не лише вибором графічного методу, а й відповідністю дизайну низькорівневим завданням користувача. Значну увагу приділено фундаментальним законам сприйняття — зокрема, закону Вебера–Фехнера та концепції “ледве помітної різниці”, — які визначають межі помітності змін у графічних кодуваннях. Це дозволило обґрунтувати модель дизайну візуалізації як оптимізаційний процес, що враховує як сприйняття, так і характер даних та аналітичні наміри користувача. Узагальнення результатів розділу дало підстави для формування концептуальної моделі процесу візуалізації, яка виступає теоретичним фундаментом розробленого фреймворку.

У другому розділі сформовано системну модель взаємозв’язку між низькорівневими завданнями користувачів, візуальними кодуваннями та характеристиками ефективності візуалізацій. Запропоновано розширену таксономію візуального кодування та низькорівневих завдань, яка дозволяє формалізувати процес вибору оптимального графічного представлення під

конкретні аналітичні дії. Досліджено вплив різних видів візуального кодування — позиції, форми, розміру, кольору тощо — на швидкість та точність виконання завдань з пошуку, фільтрації та кластеризації. Особлива увага зосереджена на задачі кластеризації, оскільки саме вона є найбільш чутливою до властивостей графічних елементів та щільності розташування точок на діаграмі розсіювання. Отримані результати показали, що ефективність кластеризації, виконуваної користувачем на основі візуалізації, може бути підвищена шляхом адаптивної оптимізації візуальних параметрів. Це дало змогу обґрунтувати модель оптимізації дизайну, що була реалізована у третьому розділі.

У третьому розділі запропоновано та імплементовано модель та алгоритмічний фреймворк генерації специфікованих візуалізацій, у центрі якого — оптимізація діаграми розсіювання для задачі кластеризації. Було здійснено підвибірку даних, побудовано механізм проекції та розроблено процедуру вилучення особливостей для оцінки видимості кластерних структур. Для оцінювання якості кластеризації при візуалізації запропоновано використання порогового графіка та концепції порогової смуги як метрики помітності структур. На основі цих інструментів створено інтерактивну модель оптимізації, що адаптивно змінює візуальні параметри відповідно до властивостей даних та особливостей сприйняття користувача. Проведений аналіз обчислювальної ефективності показав, що запропонований підхід є придатним для інтерактивних систем, забезпечуючи прийнятні часові витрати на етапах підвибірки, проекції та аналізу особливостей.

Результати валідації та оцінки якості інтерфейсу засвідчили, що розроблений фреймворк забезпечує підвищення точності сприйняття кластерних структур та покращує загальну зручність роботи з даними. Користувачі демонстрували кращі показники у виконанні завдань аналізу, що підтверджує ефективність використання оптимізованих візуальних параметрів і адаптивних механізмів генерації описаних у роботі.

Загалом дослідження доводить, що інтеграція моделей графічного сприйняття, алгоритмів оцінки помітності структур та адаптивних методів оптимізації є перспективною стратегією розвитку сучасних фреймворків візуалізації даних. Запропонована методологія дозволяє будувати візуалізації, які не тільки коректно відображають дані, а й відповідають когнітивним особливостям користувачів, тим самим підвищуючи ефективність аналітичних процесів.

ПЕРЕЛІК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. VisGuides: A Forum for Discussing Visualization Guidelines A. Dieh // <https://diglib.eg.org/server/api/core/bitstreams/bd482977-6276-4f16-a6df-5d3beee5887c/content>
2. Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387), 531–554.
3. Ware, C. (2004). *Information Visualization: Perception for Design*. Morgan Kaufmann, San Francisco.
4. Healey, C. G., & Enns, J. T. (2012). Attention and visual memory in visualization and computer graphics. *IEEE Transactions on Visualization and Computer Graphics*, 18(7), 1170–1188.
5. Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2), 110–141.
6. Card, S. K., Mackinlay, J., & Shneiderman, B. (1999). *Readings in Information Visualization: Using Vision to Think*. Morgan Kaufmann, San Francisco.
7. Bertin, J. (2023). *Semiology of Graphics*. University of Wisconsin Press, Madison.
8. Buja, A., Cook, D., & Swayne, D. F. (1996). Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics*, 5(1), 78–99.
9. Keim, D. A. (2002). Information visualization and visual data mining. *IEEE Transactions on Visualization and Computer Graphics*, 8(1), 1–8.
10. Amar, R., Eagan, J., & Stasko, J. (2005). Low-level components of analytic activity in information visualization. *Proceedings of IEEE InfoVis*, IEEE Press, 111–117.
11. Chen, C. (2005). Top 10 unsolved information visualization problems. *IEEE Computer Graphics and Applications*, 25(4), 12–16.

12. Heer, J., & Bostock, M. (2010). Declarative language design for interactive visualization. *ACM UIST Proceedings*, ACM Press, New York, 507–518.
13. Brehmer, M., & Munzner, T. (2013). A multi-level typology of abstract visualization tasks. *IEEE Transactions on Visualization and Computer Graphics*, 19(12), 2376–2385.
14. Schulz, H. (2011). A systematic survey of visual analytics methodologies. *Computer Graphics Forum*, 30(8), 2221–2237.
15. Tory, M., & Möller, T. (2004). Human factors in visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 10(1), 72–84.
16. Lam, H. et al. (2012). Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9), 1520–1536.
17. Jänicke, H., Scheuermann, G., & Chen, M. (2015). Visual abstraction and exploration of multivariate data. *Computer Graphics Forum*, 34(3), 241–260.
18. Correll, M., & Heer, J. (2017). Surprise! Re-evaluating implicit assumptions in visual data analysis. *IEEE VIS Proceedings*, IEEE Press, 1–10.
19. Kosara, R. (2016). An empire built on sand: Reexamining what we think we know about visualization. *IEEE VIS Workshop on Evaluation*, 1–8.
20. Baudel, T. (2004). From information visualization to direct manipulation: Extending the perspective wall to a multi-scale timeline. *ACM CHI*, ACM Press, 112–119.
21. Kindlmann, G., & Scheidegger, C. (2014). An algebraic process for visualization design. *IEEE Transactions on Visualization and Computer Graphics*, 20(12), 2181–2190.
22. Dörk, M., Carpendale, S., & Williamson, C. (2011). The information flaneur: A fresh look at information seeking. *ACM CHI*, ACM Press, 1215–1224.
23. Shneiderman, B. (1996). The eyes have it: A task by data type taxonomy for information visualizations. *IEEE Visual Languages Conference*, IEEE Press, 336–343.

24. Sedlmair, M., Meyer, M., & Munzner, T. (2012). Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12), 2431–2440.
25. Tomaselli, G., & Schulz, H. (2020). Quantitative evaluation of cluster visualizations. *Computer Graphics Forum*, 39(3), 251–263.
26. Ziegler, H., Klein, H., & Fellner, D. (2002). Visual exploration of time-series data with time spirals. *Proceedings of IEEE Visualization*, IEEE Press, 175–182.
27. Ankerst, M. et al. (1999). Visual classification: A technique for exploratory data analysis. *ACM KDD*, ACM Press, 392–396.
28. Van Der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
29. McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
30. Liu, S. et al. (2014). Visualizing high-dimensional data: Advances and challenges. *IEEE Computer Graphics and Applications*, 34(6), 36–46.
31. Pineo, C., & Ware, C. (2014). Data visualization for cluster analysis. *IEEE VIS*, IEEE Press, 1–8.
32. Gleicher, M. (2013). Explaining visual comparison. *IEEE Computer Graphics and Applications*, 33(4), 70–79.
33. Ham, F., & Han, J. (2004). Motion-based visual data exploration. *IEEE Visualization Conference*, 425–432.
34. Liu, J., & Stasko, J. (2010). Mental models, visual reasoning and interaction in information visualization: A top-down perspective. *IEEE VAST*, IEEE Press, 19–26.
35. Kindermann, J., & Rieder, A. (2019). Evaluation of cluster separability in scatter plots. *Computer Graphics Forum*, 38(3), 401–412.
36. Holzinger, A. (2018). From machine learning to explainable AI: Challenges and opportunities. *IEEE ICDH*, IEEE Press, 55–60.

37. Rensink, R. A. (2014). On the prospects for a science of visualization. In *Information Visualization* (pp. 147–175). Springer, London.
38. Vilar, G., & Ware, C. (2017). An evaluation of visual cluster separation measures. *IEEE Transactions on Visualization and Computer Graphics*, 23(9), 2137–2146.
39. Ellis, G., & Dix, A. (2007). A taxonomy of clutter reduction for information visualisation. *IEEE Transactions on Visualization and Computer Graphics*, 13(6), 1216–1223.