Харківський національний університет імені В.Н. Каразіна
Факультет математики і інформатики
Кафедра фундаментальної математики

Кваліфікаційна робота

освітньо-кваліфікаційний рівень: **магістр**

на тему «***Топологічний аналіз даних у дослідженні епідеміологічної інформації***»

Виконала: студентка групи М161 ІІ курсу (другий магістерський рівень), спеціальності 111 «Математика», освітньо-наукова програма «Математика»

Керівник: кандидат фіз.-мат. наук, Полякова Людмила Юріївна

кандидат фіз.-мат. наук, старший викладач кафедри фундаментальної математики, Петров Євген В'ячеславович

Рецензент: кандидат фіз.-мат. наук, Леонов Олександр

Харків - 2024 рік

Анотації

Шевцова В.В. Топологічний аналіз даних у дослідженні епідеміологічної інформації.

У зв'язку з зростанням кількості, різноманітності та розмірності доступних даних, виявлення та використання їх структури стали вагомою проблемою для аналізу даних. Аналіз топологічних даних (TDA) – це новий напрямок, який пропонує топологічні та геометричні інструменти для аналізу складних даних. Він включає математичні основи та обчислювальні методи, які можуть використовуватися як самостійно, так і в поєднанні з іншими методами аналізу даних та машинного навчання. У цій роботі методи топологічного аналізу даних (TDA) застосовуються для аналізу поширення COVID-19 США у початковій фазі пандемії з використанням даних із загальнодоступних джерел.

Shevtsova V.V. Topological data analysis in the study of epidemiological information.

Due to the increasing quantity, diversity, and dimensionality of available data, identifying and utilizing their structure has become a significant challenge for data analysis. Topological Data Analysis (TDA) is a novel approach that offers topological and geometric tools for analyzing complex data. It encompasses mathematical foundations and computational methods that can be used both independently and in conjunction with other data analysis and machine learning techniques. In this study, methods of TDA are applied to analyze the spread of COVID-19 in the United States during the early stages of the pandemic using data from publicly available sources.

Вступ

У світі дані стають невід'ємним ресурсом, що грає ключову роль розумінні і вирішенні складних проблем. Особливо в епідеміології, де ми маємо справу з багатовимірними та різноманітними даними — просторово-часовими, соціальними, економічними, демографічними, політичними та іншими факторами, які можуть вплинути на динаміку поширення пандемій. Аналіз цих даних ускладнений через велику кількість особливостей та взаємозв'язків.

Методи топологічного аналізу даних (TDA) дозволяють досліджувати складні та багатовимірні дані шляхом формування апроксимацій меншої розмірності, які зберігають структуру та інформацію про зв'язність у даних.

Включивши TDA в епідеміологічні дослідження, аналітики можуть вивчити геометричні та топологічні структури наборів даних, що лежать в основі, виявляючи приховані закономірності та взаємозв'язки, які можуть бути неочевидні тільки за допомогою звичайного статистичного аналізу, забезпечуючи більш цілісне розуміння динаміки захворювань, шляхів передачі та факторів ризику.

Метою TDA є забезпечення ґрунтовних математичних, статистичних та алгоритмічних підходів для вилучення та аналізу складних топологічних та геометричних структур, що лежать в основі даних, які часто представлені у вигляді хмари точок у метричному просторі.

У рамках цієї роботи планується розглянути існуючі методи TDA для аналізу даних, комплексно застосувати їх для дослідження поширення COVID-19 у кожному окрузі США, використовуючи дані з відкритих джерел. Крім цього, планується провести додатковий статистичний аналіз та використати методи машинного навчання для отримання більш стійких результатів. Мета цієї роботи не тільки виявити патерни у розповсюдженні захворювання, але й запропонувати

нові підходи до аналізу епідеміологічних даних із використанням методів дослідження TDA.

Аналіз зосереджено на ранній стадії пандемії. Ціль полягає в тому, щоб визначити, чому в деяких регіонах США пандемія поширювалася швидше після підтвердження першого випадку, ніж в інших, де спостерігалися набагато повільніші темпи. Розглядаються соціальні, економічні, демографічні, політичні та інші фактори, які можуть вплинути на схожість поширення пандемії.

Робота складається зі вступу, двох розділів, загальних висновків, списку використаної літератури (9). Зміст роботи висвітлено на 42 сторінках основного тексту і містить 17 рисунків.

1. Основи топологічного аналізу даних

Математичний формалізм, який був розроблений для включення геометричних і топологічних методів, має справу з хмарами точок, тобто кінцеві набори точок, оснащених функцією відстані. Хмари точок мають розглядатися як кінцеві зразки, взяті з геометричного об’єкта, можливо, з шумом. Ось деякі з ключових моментів, які виникають під час застосування цих геометричних методів для аналізу даних.

1.1. Схема топологічного аналізу даних

Припускається, що вхідні дані є скінченним набором точок, які надходять із поняттям відстані або подібності між ними. Ця відстань може бути викликана метрикою в навколишньому просторі (наприклад, евклідова метрика, коли дані вбудовані в $\mathbb{R}^d$) або постає як власна метрика, визначена попарною матрицею відстаней. Визначення метрики на даних зазвичай дається як вхідні дані або задається виходячи з цілей аналізу. Вибір метрики може мати вирішальне значення для виявлення цікавих топологічних і геометричних особливостей даних.

«Безперервна» форма будується поверх даних, щоб підкреслити базову топологію або геометрію. Це часто симпліціальний комплекс або вкладене сімейство симпліциальних комплексів, що називається фільтрацією, що відображає структуру даних у різних масштабах. Симпліціальні комплекси можна розглядати як багатовимірні узагальнення графів сусідів.

3. Топологічна або геометрична інформація витягується зі структур, побудованих поверх даних. Окрім ідентифікації цікавої топологічної/геометричної інформації та її візуалізації та інтерпретації, завдання на цьому етапі полягає в тому, щоб показати її релевантність, зокрема її стабільність щодо збурень або наявності шуму у вхідних даних.

З цією метою розуміння статистичної поведінки виведених ознак також є важливим питанням.

Витягнута топологічна та геометрична інформація надає нові сімейства ознак і дескрипторів даних. Їх можна використовувати для кращого розуміння даних, зокрема за допомогою візуалізації, або їх можна поєднувати з іншими функціями для подальшого аналізу та завдань машинного навчання. Цю інформацію також можна використовувати для розробки відповідних моделей аналізу даних і машинного навчання.

Аналіз топологічних даних і статистика

Статистичний підхід до TDA означає, що ми розглядаємо дані як такі, що генеруються з невідомого розподілу, але також що топологічні особливості, виведені за допомогою методів TDA, розглядаються як оцінки топологічних величин, що описують основний об’єкт. При цьому підході невідомий об'єкт зазвичай відповідає розподілу даних (або його частини). Основні цілі статистичного підходу до топологічного аналізу даних можна підсумувати як наступний перелік проблем:

- доведення узгодженості та дослідження швидкостей збіжності методів TDA.
- забезпечення довірчих областей для топологічних характеристик та обговорення значущості оцінених топологічних величин.
- вибір відповідних масштабів, на яких слід розглядати топологічний феномен, як функцію даних спостереження.
- робота з викидами та надання надійних методів для TDA.

1.2. Метричні простори

Оскільки топологічні та геометричні особливості зазвичай асоціюються з неперервними просторами, дані, представлені у вигляді кінцевих множин

спостережень, безпосередньо не виявляють жодної топологічної інформації як такої. Природний спосіб виділити деяку топологічну структуру з даних — це «з'єднати» точки даних, які знаходяться близько одна до одної, щоб продемонструвати глобальну безперервну форму, що лежить в основі даних.

Кількісна оцінка поняття близькості між точками даних зазвичай виконується за допомогою відстані (або міри відмінності), і часто виявляється зручним у TDA розглядати набори даних як дискретні метричні простори або як вибірка з метричних просторів [1]

Метричний простір (X, d)— це множина X із функцією $d: X \times X \to \mathbb{R}^+$, яка називається відстанню, якщо:

1) Для кожних $x, y \in X$, $d(x, y) \geq 0$.
2) Для кожного $x \in X$, $d(x, x) = 0$.
3) Для кожних $x, y \in X$ таких, що $x \neq y, d(x, y) > 0$.
4) Симетрія: для кожних $x, y \in X$, $d(x, y) = d(y, x)$.
5) Нерівність трикутника: для кожних $x, y, z \in X$, $d(x, y) + d(y, z) \geq d(x, z)$.

Це визначення іноді пом'якшується різними способами. Набір точок і функція відстані, які задовольняють усі перелічені вище умови, крім властивості 3 (тобто ми допускаємо, що дві різні точки x, y мають відстань 0), називається напівметрикою або псевдометрикою.

Деякі корисні функції відстані не задовольняють іншим властивостям, які з'являються у наведеному вище визначенні (або властивості 4, або властивості 5). Немає стандартних назв для цих послаблених визначень метрик, але їх іноді називають квазіметриками.

Для метричного простору (X, d) множина $\mathcal{K}(X)$ його компактних підмножин може бути наділена так званою відстанню Хаусдорфа.

Задані дві компактні підмножини $A, B \subseteq X$, відстань Хаусдорфа $d_H(A, B)$ між A і B визначається як найменше невід'ємне число δ таке, що для будь-якого $a \in A$ існує $b \in B$ таке, що $d(a, b) \leq \delta$, і для будь-якого $b \in B$ існує таке $a \in A$, що $d(a, b) \leq \delta$. Іншими словами, якщо для будь-якої компактної підмножини $C \subseteq X$ ми позначимо $d(., C): X \to \mathbb{R}^+$ функцію відстані до C, визначену $d(x, C) := inf_{c \in C} d(x, c)$ для будь-якого $x \in X$, то можна довести, що відстань Хаусдорфа між A і B визначається будь-якою з двох наступних рівностей:

$$d_H(A, B) = \max\left\{\sup_{b \in B} d(b, A), \sup_{a \in A} d(a, B)\right\}$$
$$= \sup_{x \in M} |d(x, A) - d(x, B)| = \|d(., A) - d(., B)\|_\infty$$

Базовим і класичним результатом є те, що відстань Хаусдорфа дійсно є відстанню на множині компактних підмножин метричного простору.

З точки зору TDA, це забезпечує зручний спосіб кількісної оцінки близькості між різними множинами даних, виданими з того самого зовнішнього метричного простору. Однак іноді трапляється так, що доводиться порівнювати множини даних, які не взяті з одного метричного простору.

Поняття відстані Хаусдорфа можна узагальнити для порівняння будь-якої пари компактних метричних просторів, породжуючи поняття відстані Громова–Хаусдорфа.

Два компактні метричні простори (X_1, d_1) і (X_2, d_2)є ізометричними, якщо існує бієкція $\phi: X_1 \to X_2$, яка зберігає відстані, тобто $d_2\big(\phi(x), \phi(y)\big) = d_1(x, y)$ для будь-яких $x, y \in X_1$.

Визначення 1. Відстань Громова–Хаусдорфа $d_{GH}(X_1, X_2)$ між двома компактними метричними просторами є нижньою величиною дійсних чисел $r \geq$

0 таких, що існує метричний простір (X, d) і два компактні підпростори C_1 і $C_2 \subset X$ які ізометричні X_1 і X_2 і такі, що $d_H(C_1, C_2) \leq r$.

З'єднання пар сусідніх точок даних ребрами призводить до стандартного поняття графа сусідів, за допомогою якого можна аналізувати зв'язність даних, наприклад, використовуючи алгоритми кластеризації. Щоб вийти за межі зв'язності, центральна ідея TDA полягає в тому, щоб побудувати еквіваленти графів сусідів більшої розмірності, використовуючи не лише з'єднання пари точок, але й вектори (k + 1) розмірності сусідніх точок даних. Отримані об'єкти, які називаються симпліціальними комплексами, дозволяють нам ідентифікувати нові топологічні особливості, такі як цикли, порожнечі та їхні аналоги з більшою розмірністю.

1.3. Геометричні та абстрактні симпліціальні комплекси

Симпліціальні комплекси можна розглядати як багатовимірне узагальнення графів. Це математичні об'єкти, які є як топологічними, так і комбінаторними, що робить їх особливо корисними для TDA.

Дана множина $\mathbb{X} = \{x_0, \dots, x_k\} \subset \mathbb{R}^d$ з $k + 1$ афінно незалежних точок, k-вимірний симплекс $\sigma = [x_0, \dots, x_k]$, натягнутий на $\mathbb{X}$, є опуклою оболонкою $\mathbb{X}$. Точки $\mathbb{X}$ називаються вершинами σ, а симплекси, охоплені підмножинами $\mathbb{X}$, називаються гранями σ. Геометричний симпліціальний комплекс K у $\mathbb{R}^d$ — це колекція симплексів, такий, що має місце наступне:

1) будь-яка грань симплекса K є симплексом K,
2) перетин будь-яких двох симплексів K є або порожнім, або спільною гранню обох.

Об’єднання симплексів K є підмножиною $\mathbb{R}^d$, що називається базисним простором K, який успадковує топологію $\mathbb{R}^d$. Отже, K також можна розглядати як топологічний простір через його базисний простір. Зауважте, що як тільки його вершини відомі, K повністю характеризується комбінаторним описом набору симплексів, які задовольняють деяким правилам інцидентності.

Дано множину V, абстрактний симпліціальний комплекс із множиною вершин V є множиною $\widetilde{K}$ скінченних підмножин V так, що елементи V належать $\widetilde{K}$ і для будь-якого $\sigma \in \widetilde{K}$ будь-яка підмножина σ належить $\widetilde{K}$. Елементи $\widetilde{K}$ називаються гранями або симплексами $\widetilde{K}$. Розмірність абстрактного симплексу є просто його потужністю мінус 1, а розмірність $\widetilde{K}$ є найбільшою розмірністю його симплексів.

Симпліціальні комплекси розмірності 1 є графами.

Комбінаторний опис будь-якого геометричного симпліціала K, очевидно, породжує абстрактний симпліціальний комплекс $\widetilde{K}$.

Справедливо і зворотне; можна завжди пов'язати з абстрактним симпліціальним комплексом $\widetilde{K}$ топологічний простір $|\widetilde{K}|$ таким чином, що якщо K є геометричним комплексом, комбінаторний опис якого такий самий, як $\widetilde{K}$, базовий простір K гомеоморфний $|\widetilde{K}|$. Таке K називається геометричною реалізацією $\widetilde{K}$. Як наслідок, абстрактні симпліціальні комплекси можна розглядати як топологічні простори, а геометричні комплекси можна розглядати як геометричні реалізації їх основної комбінаторної структури. Отже, можна розглядати симпліціальні комплекси одночасно як комбінаторні об’єкти, які добре підходять для ефективних обчислень, і як топологічні простори, з яких можна вивести топологічні властивості.

1.4. Побудова симпліціальних комплексів з даних

Маючи множину даних або, загалом, топологічний або метричний простір, існує багато способів побудови симпліціальних комплексів. Наведемо кілька класичних прикладів, які широко використовуються на практиці.

Перший приклад — безпосереднє розширення поняття α-сусіднього графа. Припустимо, що нам дано набір точок $\mathbb{X}$ у метричному просторі (X, d) і дійсне число $\alpha \geq 0$. Комплекс В'єторіса–Ріпса $Rips_\alpha(\mathbb{X})$ є множиною симплексів $[x_0, \dots, x_k]$ такий, що $d_{\mathbb{X}}(x_i, x_j) \leq \alpha$ для всіх (i, j), див. рис. 1.1. З визначення безпосередньо випливає, що це абстрактний симпліціальний комплекс. Однак, загалом, навіть коли $\mathbb{X}$ є скінченною підмножиною $\mathbb{R}^d$, $Rips_\alpha(\mathbb{X})$ не допускає геометричної реалізації в $\mathbb{R}^d$; зокрема, він може мати розмірність, більшу за d.

З комплексом В'єторіса–Ріпса тісно пов'язаний комплекс Чеха $Cech_\alpha(\mathbb{X})$, який визначається як набір симплексів $[x_0, \dots, x_k]$ таких, що $k+1$ замкнуті кулі $B(x_i, \alpha)$ мають непорожнє перехрестя, дивіться рис 1.1. Ці два комплекси пов'язані між собою:

$$Rips_\alpha(\mathbb{X}) \subseteq Cech_\alpha(\mathbb{X}) \subseteq Rips_{2\alpha}(\mathbb{X})$$

і що якщо $\mathbb{X} \subset \mathbb{R}^d$, то $Cech_\alpha(\mathbb{X})$ і $Rips_{2\alpha}(\mathbb{X})$ мають однаковий одновимірний скелетон, тобто однаковий набір вершин і ребер.

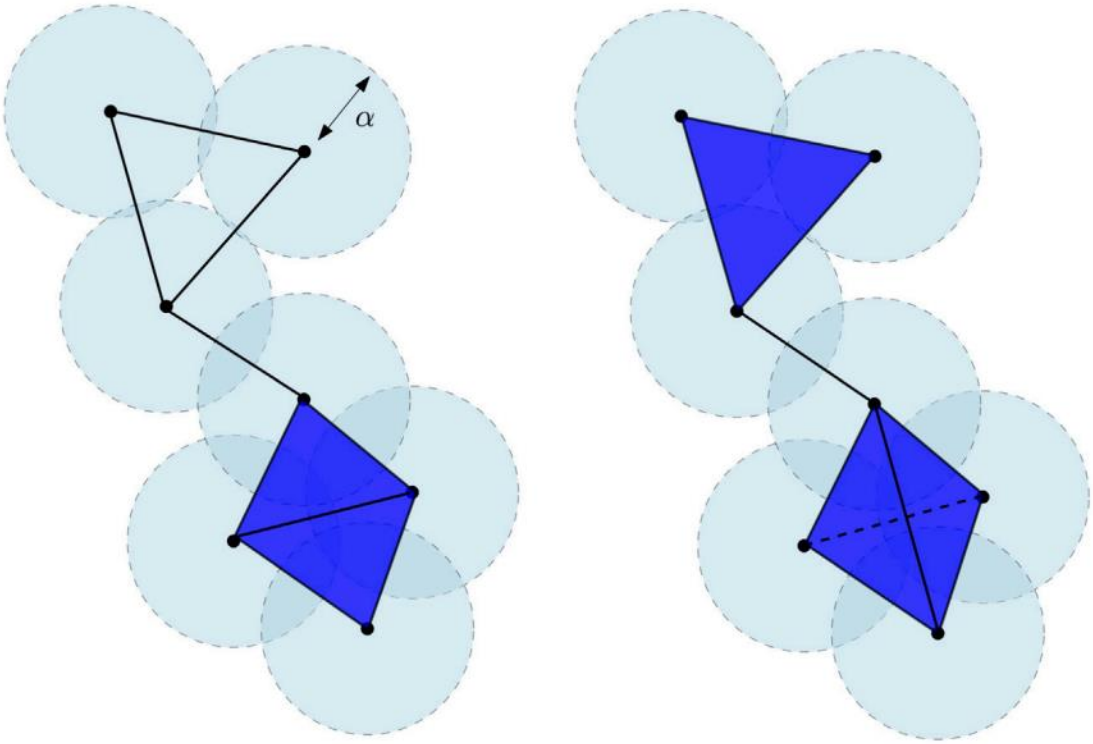


Рис. 1.1. Комплекс Чеха $Cech_\alpha(\mathbb{X})$ (ліворуч) і В'єторіса–Ріпса $Rips_{2\alpha}(\mathbb{X})$ (праворуч) [2].

Комплекс Чеха $Cech_\alpha(\mathbb{X})$ і комплекс В'єторіса–Ріпса $Rips_{2\alpha}(\mathbb{X})$ на рис. 1.1. із скінченної хмари точок у площині $\mathbb{R}^2$. Нижня частина $Cech_\alpha(\mathbb{X})$ є об'єднанням двох суміжних трикутників, тоді як нижня частина $Rips_{2\alpha}(\mathbb{X})$ є тетраедром, охопленим чотирма вершинами та всіма його гранями. Розмірність комплексу Чеха дорівнює 2. Розмірність комплексу Вієторіса–Ріпса дорівнює 3. Останній, таким чином, не вбудовано в $\mathbb{R}^2$

1.5. Теорема про нерв

Комплекс Чеха є окремим випадком сімейства комплексів, пов'язаних з покриттями. Дано покриття $\mathcal{U} = (U_i)_{i\in I}$ топологічного простору $\mathbb{M}$, тобто сімейство множин U_i таких, що $\mathbb{M} = \bigcup_{i\in I} U_i$, нерв $\mathcal{U}$ є абстрактним симпліціальним комплексом $C(\mathcal{U})$, вершинами якого є U_i та такий як

$\sigma = [U_{i_0}, \dots, U_{i_k}] \in C(\mathcal{U})$ тоді і тільки тоді, коли $\bigcap_{j=0}^{k} U_{i_j} \neq \emptyset$

Маючи покриття множини даних, де кожна множина покриття може бути, наприклад, локальним кластером або групою точок даних, що мають спільні властивості, його нерв забезпечує компактний і глобальний комбінаторний опис зв'язку між цими множинами через їх паттерни перетину (див. рис. 1.2).

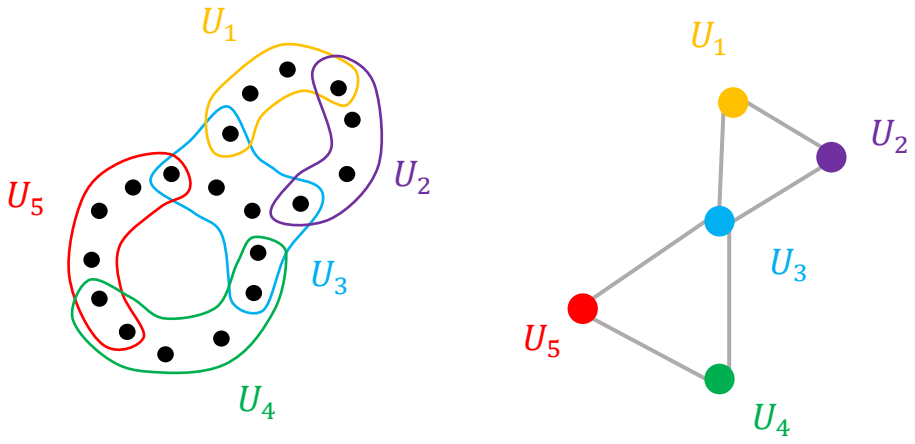


Рис. 1.2. Вибірка хмари точок на площині та покриття відкритих наборів для цієї хмари точок (ліворуч). Нервом цієї оболонки є фігура (справа). Вершини відповідають множині покриття, тоді як ребро відповідає непорожньому перетину між двома множинами покриття.

Фундаментальна теорема в алгебраїчній топології пов'язує, за деяких припущень, топологію нерва покриття з топологією об'єднання множин покриття. Для формулювання теореми нерва треба ввести кілька понять.

Два топологічні простори, X і Y, зазвичай вважаються однаковими з топологічної точки зору, якщо вони такими є гомеоморфні, тобто якщо існують два неперервних біективних відображення $f: X \to Y$ і $g: Y \to X$, такі що $f \circ g$ і $g \circ f$ є тотожним відображенням Y і X відповідно. У багатьох випадках вимога про те, щоб X і Y були гомеоморфними, виявляється занадто сильною вимогою, щоб

гарантувати, що X і Y мають однакові топологічні характеристики, що представляють інтерес для TDA. Два неперервних відображення $f_0, f_1: X \to Y$ називаються гомотопними, якщо існує неперервне відображення $H{:}\times [0,1] \to Y$ таке, що для будь-якого $x \in X,\ H(x,0) = f_0(x)$ і $H(x,1) = f_1(x)$. Тоді простори X і Y називаються гомотопічно еквівалентними, якщо існують два відображення, $f: X \to Y$ і $g: Y \to X$, так що $f \circ g$ і $g \circ f$ гомотопні тотожному відображенню Y і X відповідно. Тоді відображення f і g називаються гомотопічними еквівалентами. Поняття гомотопічної еквівалентності слабше, ніж поняття гомеоморфізму; якщо X і Y гомеоморфні, то вони, очевидно, гомотопно еквівалентні, але зворотне не вірно. Однак простори, які є гомотопічно еквівалентними, все ще мають багато спільних топологічних інваріантів; зокрема, вони мають однакову гомологію.

Простір називається стягуваним, якщо він гомотопічно еквівалентний точці. Основними прикладами стягувальних просторів є кулі та, в більш загальному випадку, опуклі множини в $\mathbb{R}^d$. Відкриті покриття, для яких усі елементи та їх перетини стягуються, мають чудову наступну властивість.

Теорема 1 (Теорема нерва). Нехай $\mathcal{U} = (U_i)_{i \in I}$ — покриття топологічного простору X відкритими множинами, так що перетин будь-якого піднабору U_i є або порожнім, або стягуваним. Тоді X і нерв $C(\mathcal{U})$ гомотопно еквівалентні.

Легко перевірити, що опуклі підмножини евклідових просторів стягуються. Як наслідок, якщо $\mathcal{U} = (U_i)_{i \in I}$ є набором опуклих підмножин $\mathbb{R}^d$, то $C(\mathcal{U})$ і $\bigcup_{i \in I} U_i$ гомотопічно еквівалентні. Зокрема, якщо $\mathbb{X}$ — множина точок у $\mathbb{R}^d$, то комплекс Чеха $Cech_\alpha(\mathbb{X})$ гомотопічно еквівалентний об'єднанню куль $\bigcup_{x \in \mathbb{X}} B(x, a)$.

Теорема нерва відіграє фундаментальну роль у TDA; вона забезпечує спосіб кодування топології неперервних просторів у абстрактні комбінаторні

структури, які добре підходять для розробки ефективних структур даних і алгоритмів.

1.6. Використання покриття та нервів для дослідження, аналізу та візуалізації даних: алгоритм Mapper

Використання покриття нерва як способу узагальнення, візуалізації та дослідження даних є природною ідеєю, яка породила так званий алгоритм Mapper.

Нехай $f: X \to \mathbb{R}^d$, $d \geq 1$, — неперервна дійсна функція, а $\mathcal{U} = (U_i)_{i \in I}$ — покриття $\mathbb{R}^d$. Pull-back покриття X, індуковане $(f, \mathcal{U})$, є сукупністю відкритих множин $\left(f^{-1}(U_i)\right)_{i \in I}$. Уточнений pull-back— це сукупність зв'язних компонент відкритих множин $f^{-1}(U_i)$, $i \in I$.

Ідея алгоритму Mapper полягає в тому, щоб, враховуючи множину даних $\mathbb{X}$ і дійсну функцію $f: X \to \mathbb{R}^d$, підсумувати $\mathbb{X}$ через нерв уточненого pull-back покриття $\mathcal{U}$ функції $f(x)$. Для добре підібраних покриттів $\mathcal{U}$ цей нерв є графіком, що забезпечує простий і зручний спосіб візуалізації зведення даних [3].

Вхідними даними алгоритму є датасет $\mathbb{X}$ з метрикою, функція $f: X \to \mathbb{R}^d$, яку називають ще фільтром або лінзою, покриття $\mathcal{U}$ функції $f(x)$.

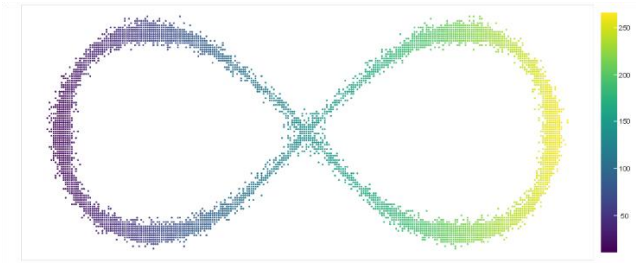
Для кожного $U_i \in \mathcal{U}$: розкладаємо $f^{-1}(U_i)$ на кластери $C_{U_i,1}, \dots, C_{U_i,k}$, застосовуючи обраний користувачем алгоритм кластерізації.

Результатом роботи алгоритму є симпліціальний комплекс; нерв (часто граф для добре підібраних покриттів, який легко візуалізувати), який складається (рис 1.3):

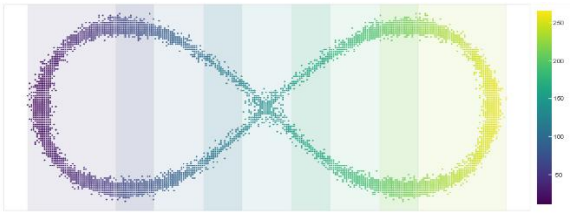
– вершин $V_{U_i,j}$, кожна предсталяє собою кластер $C_{U_i,j}$,

– ребро між $V_{U_i,j}$ і $V_{U_t,l}$, якщо $C_{U_i,j} \cap C_{U_t,l} \neq \emptyset$

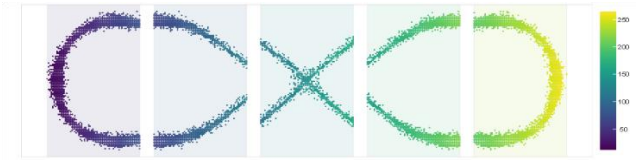
Крок 1: Функція фільтра: $f(x,y) = x$



Крок 2: розділити на інтервали, що перекриваються



Крок 3: кластеризація точок на кожному інтервалі



Вихідний граф

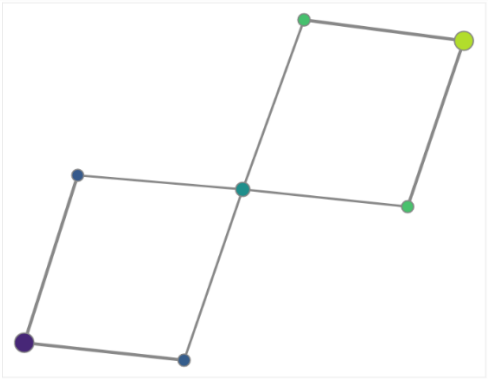


Рис.1.3. Приклад роботи алгоритму Mapper.

Вибір функції *f*

Вибір функції f, яку іноді називають функцією фільтра або лінзи, сильно залежить від особливостей даних, які треба дослідити. Граф або симпліціальний комплекс, який отримаємо як результат алгоритма Mapper, буде візуалізувати основні властивості топологічного простору через лінзу. Деякі функції, які можуть бути використані, як лінзи:

Centrality. Вказує на відстань точки від «центру» даних або наскільки добре точка відповідає «нормі», а не є викидом. Ця функція має один параметр p і обчислюється за формулою:

$$Centrality(x) = \begin{cases} \left(\frac{\sum_{i=1}^{N} d(x,x_i)^p}{N}\right)^{1/p}, \ if\ 1 \le p < +\infty \\ \max\limits_{i} d(x, x_i), \ if\ p = +\infty \end{cases},$$

де N – кількість точок.

Density. Оцінює щільність сусідніх точок навколо даної точки та обчислюється за формулою:

$$Gauss\_density(x) = \frac{1}{N(\sqrt{2\pi}\sigma)^n} \sum_{i=1}^{N} \exp\left(\frac{-d(x, x_i)^2}{2\sigma^2}\right),$$

де σ^2- параметр масштабу.

Методи зменшення розмірності. Вони можуть зменшити розмірність даних, зберігаючи при цьому їхню внутрішню структуру.

Зазвичай використовувані методи зменшення розмірності включають аналіз головних компонентів (PCA), багатовимірне масштабування (MDS), t-розподілене стохастичне вбудовування сусідів (t-SNE) та інші.

Перший алгоритм PCA полягає у лінійному перетворенні вихідного простору в підпростори меншої розмірності так, щоб даний проєкційний простір був ортогональним, а дисперсія була максимальною. Цей метод використовується для розкладання багатовимірного датасету на послідовні ортогональні компоненти, щоб середньоквадратична відстань між точками була максимальною.

Інший алгоритм, MDS, шукає низьковимірне представлення даних, де відстані точно відображають відстані в оригінальному високовимірному просторі [4]. MDS розміщує кожен об’єкт у просторі меншої розмірності таким чином, щоб відстані між об’єктами зберігалися якомога точніше. Цей метод являє собою тип нелінійного зменшення розмірності.

t-SNE — це метод нелінійного зменшення розмірності, який представляє кожен високовимірний об’єкт дво- або тривимірною точкою таким чином, що подібні об’єкти моделюються найближчими точками, тоді як різнорідні об’єкти моделюються віддаленими точками з високою ймовірністю [5]. Алгоритм t-SNE складається з двох основних етапів. По-перше, t-SNE створює розподіл ймовірностей за парами високовимірних об’єктів таким чином, що подібним об’єктам призначаються вищі ймовірності, тоді як різнорідним об’єктам призначаються менші ймовірності. По-друге, t-SNE визначає подібний розподіл ймовірностей за точками в низьковимірному просторі та мінімізує розбіжність Кульбака-Лейблера між двома розподілами.

Статистичні функції, такі як максимум, мінімум, середнє, медіана, дисперсія, ентропія вектора ознак для кожної точки датасету.

Геометричні проекції векторів ознак.

Data-driven лінзи. Лінзи побудовані на даних, які не були використані при побудові вхідного датасета. Наприклад, у клінічних дослідженнях проекцією може служити вік пацієнта.

Вибір покриття $\mathcal{U}$

Коли f є дійсною функцією, стандартним вибором є взяти в якості $\mathcal{U}$ множину інтервалів, що покриває $f(\mathbb{X})$. Кількість інтервалів іноді називають роздільною здатністю покриття, а відсоток перекриття між двома послідовними інтервалами називають перекриттям. Для розрахунку покриття ми можемо розділити лінзу на рівномірні інтервали, щоб отримати так зване рівномірне покриття. Цей підхід зручно використовувати, якщо значення лінзи більш-менш рівномірно розподілені по всьому діапазону значень. У разі нерівномірного розподілу краще збалансоване покриття, в якому кожен інтервал містить однакову кількість точок.

Можна використовувати кілька лінз або багатовимірну лінзу, яку можна отримати як комбінацію кількох одновимірних (рівномірних або збалансованих). Для створення такої комбінації проводиться ітерація по всім можливим парам інтервалів з двох лінз, і вибираються точки, що належать обом інтервалам одночасно (перетин інтервалів).

Якщо перекриття вибрано нижче 50%, то кожна точка покривається щонайбільше 2 відкритими множинами $\mathcal{U}$, а вихідний нерв є графом. Важливо зауважити, що вихідні дані Mapper дуже чутливі до вибору $\mathcal{U}$, і невеликі зміни роздільної здатності та параметрів перекриття можуть призвести до дуже великих змін вихідних даних, що робить метод дуже нестабільним (рис 1.4). Класична стратегія полягає в дослідженні певного діапазону параметрів і виборі тих, які, як виявилося, забезпечують найбільш інформативний та стабільний вихід з точки зору користувача.

5 інтервалів

10 інтервалів

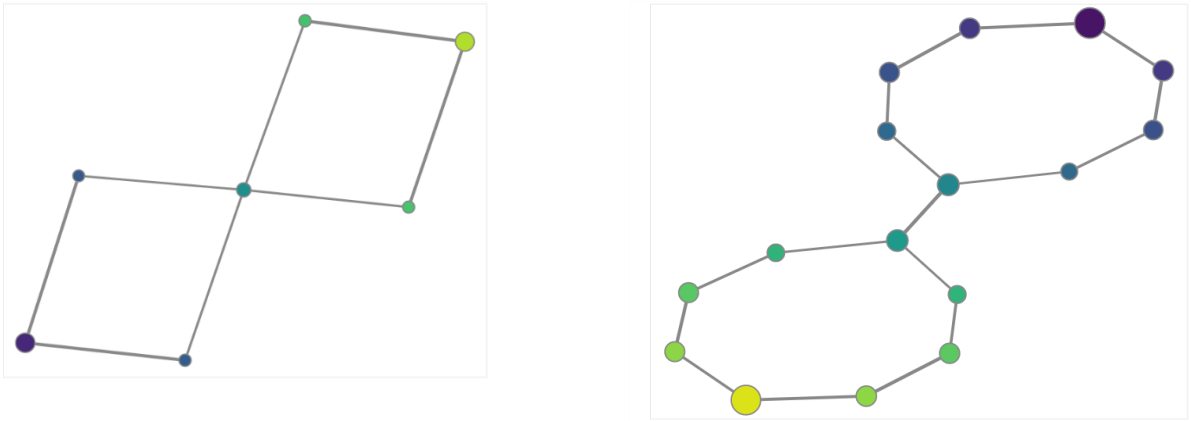


Рис 1.4. Приклад графів для різної кількості інтервалів одно і того ж датасета.

Вибір кластерів

Алгоритм Mapper вимагає кластеризації прообразу відкритих множин $U_i \in \mathcal{U}$. Існує дві стратегії для обчислення кластерів:

- Перша стратегія полягає у застосуванні алгоритму кластерізації, обраного користувачем, для кожного $U_i \in \mathcal{U}$ до прообразу $f^{-1}(U_i)$. Алгоритм кластеризації на виході повинен видавати множини точок вхідного датасета, що не перетинаються, так звана чітка кластеризація.
- Друга, більш глобальна стратегія полягає в побудові графа сусідів на вершинах множини X, наприклад, k-NN-графа або α-сусіднього графа, і для кожного $U_i \in \mathcal{U}$ та взяття компонент зв‘язності із множиною вершин $f^{-1}(U_i)$
-

1.7. α-сусідній граф з покриттям

Якщо у датасеті невелика кількість точок, то більш детально та інформативно буде побудувати α-сусідній граф або $Rips_{2\alpha}(\mathbb{X})$ для кожного елемента покриття. При цьому за рахунок перекриття можна досягти загальної зв'язності

Можна розглянути цю структуру як симпліційний комплекс $Rips_{2\alpha}(\mathbb{X})$, збудований на квазиметриці з урахуванням покриття $U_i \in \mathcal{U}$:

$$d'^{(x,y)} = \begin{cases} d(x,y), & if\ x,y \in U_i \\ +\propto, & overwise \end{cases}$$

При построении этой квазиметрики, если две точки x и y лежат в одном интервале, то расстояние между ними сохраняется, иначе меняется на $+\propto$ или константу, много большую максимального расстояния между точками в датасете.

При побудові цієї квазиметрики, якщо дві точки x і y лежать у одному інтервалі, то відстань з-поміж них зберігається, інакше змінюється на $+\propto$ чи константу, багато більшу максимальної відстані між точками в датасете.

При такому перетворенні може бути порушена властивість нерівності трикутника, наприклад, у випадку на рис 1.5.

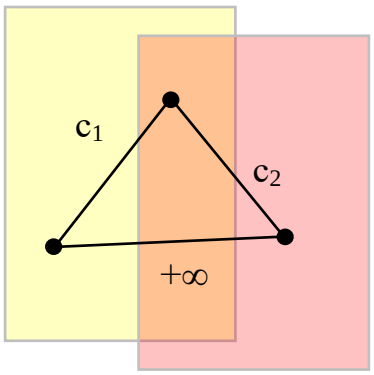


Рис 1.5. Приклад випадку порушення нерівності трикутника.

Если допустить, чтобы для каждого элемента покрытия можно было задать свой параметр α, то есть задано множество $\{\alpha_i\}$, то квазиметрику можно переписать в виде:

Якщо припустити, щоб для кожного елемента покриття можна було задати свій параметр α, тобто задана множина $\{\alpha_i\}$, то квазиметрику можна переписати у вигляді:

$$d''^{(x,y)} = \begin{cases} \dfrac{d(x,y)}{\max\limits_{x,y\in U_j}\{\alpha_j\}}, & if \bigcap\nolimits_{x,y\in U_j} U_j \neq \emptyset \\ +\propto, & overwise \end{cases}$$

Тобто, якщо пара точок належить лише одному інтервалу U_j, то відстань нормуємо на α_j, якщо пара точок належить відразу кільком інтервалам, то нормуємо на $\max\limits_{x,y\in U_j}\{\alpha_j\}$. Таким чином, для такої квазиметрики можна будувати $Rips_2(\mathbb{X})$.

Такий підхід можна застосовувати у випадку, якщо масштаб різний для різних елементів покриття.

Для побудови α-сусіднього графа з покриттям для множини точок $\mathbb{X}$ потужності N і функції фільтра $f: \mathbb{X} \rightarrow \mathbb{R}$, розрахуємо покриття: область значень

функції f розіб'ємо на множину інтервалів, що перекриваються. Це дасть два параметри, які можна використовувати для керування роздільною здатністю, а саме кількість інтервалів та перекриття між сусідніми інтервалами у відсотках.

Для побудови графа беремо точки даних як вершини графа. У кожному інтервалі з'єднуємо точки ребром, якщо відстань між точками менша за певний поріг. За рахунок інтервалів покриття, що перекриваються, граф може залишатися зв'язним.

У наступному розділі застосуємо цей підхід для аналізу розповсюдження пандемії COVID-19 у США у початковий період.

2. Аналіз поширення COVID-19 зі застосуванням TDA

Аналіз проводився на основі округів в США, що дало топологічну модель даних у формі графа, у якому кожен із 3142 вузлів відповідає одному округу, а два вузли з'єднані, якщо вони мають схожість. Набір даних включав ознаки, що відповідають кількості підтверджених випадків і смертей у кожному окрузі за визначений проміжок часу.

Основна мета — дослідити поширення пандемії з моменту її початку. Таким чином, за початкову точку інтервалу спостереження було взято перший підтверджений та зареєстрований випадок COVID-19 у кожному окрузі. Враховуючи швидкість, з якою просувалася пандемія, було критично важливо не лише вибрати відповідну початкову точку, але й обмежити аналіз, ретельно вибравши кінцеву точку часового інтервалу, щоб зробити його релевантним, не перевантажуючи модель. Стрибок у кількості випадків був значним, наприклад, з 451 до 330 384 у період з 8 березня по 5 квітня; отже, спостережуваний часовий інтервал був встановлений на рівні 39 днів.

Після того, як граф був побудований, реальні дані були використані для інтеграції в модель будь-яких предикторів, які могли бути відповідальними за схожість на ранній стадії поширення пандемії. У ході експерименту було використано понад 250 предикторів з різних загальнодоступних джерел. Крім того, був проведений статистичний аналіз виявлених закономірностей, щоб пояснити схожість у поширенні пандемії на основі предикторів.

2.1. Джерела даних

Дані були отримані з відкритого сховища даних, доступного на Github і названого Центром системної науки та інженерії (CSSE) при Університеті Джона Гопкінса [6], COVID-19 Data Repository. Зібрані дані включають зведені джерела даних, такі як Всесвітня організація охорони здоров'я, Європейський центр профілактики та контролю захворювань, Центри контролю та профілактики захворювань США тощо. У цьому наборі даних дані, пов'язані зі статистикою США, представлені на рівні штату чи округу/міста, включно з даними органів охорони здоров'я з усіх штатів і округів США, тоді як джерела даних за межами США агрегуються на рівні країни/регіону чи провінції. Поточне дослідження зосереджено на даних, які відповідають місцезнаходженням і кількості підтверджених випадків COVID-19 і смертей у всіх постраждалих округах США.

Дані, використані для встановлення результатів для моделі, були зібрані на рівні округів (адміністративних або політичних підрозділів штату в Сполучених Штатах) та еквівалентів округів (інших функціонально еквівалентних підрозділів під юрисдикцією США). Загалом отриманий набір даних включає дані, пов'язані з 3142 округами та еквівалентами округів. Вибраний набір даних включає кількість підтверджених випадків COVID-19 і кількість смертей, про які повідомляє кожен округ.

2.2. Вибір інтервалу часу спостереження та метріки

За початкову точку інтервалу спостереження було взято перший підтверджений та зареєстрований випадок COVID-19 у кожному окрузі. Багато округів не реєстрували нових випадків протягом кількох днів після того, як було повідомлено про перший випадок, тому день, коли було зареєстровано другий

підтверджений випадок, був прийнятий як день початку, щоб зменшити розрив у кількості днів, за які не збиралися нові дані.

Для аналізу необхідно було вибрати часовий інтервал однакової фіксованої довжини для кожного округу, оскільки це дозволило б побудувати граф за допомогою TDA на основі векторів, що представляє округа протягом конкретного проміжку часу.

Урядові та громадські органи охорони здоров'я Сполучених Штатів запровадили політику залишатися вдома, щоб запобігти поширенню COVID-19. Обмеження на перебування вдома прийнято в більшості штатів. Аналіз мав на меті дослідити, як COVID-19 поширювався в кожному окрузі на початку спалаху та протягом періоду перебування вдома, коли контакти між людьми були обмежені в більшості округів, а умови поширення захворювання в різних округах були настільки ж схожими, як можливо.

Для кожного округу було визначено, що часовий інтервал спостереження лежить у межах Start_day – другого підтвердженого випадку в окрузі, та End_day – дня звільнення від наказу про перебування вдома на основі розпоряджень по всьому штату (рис 2.1).

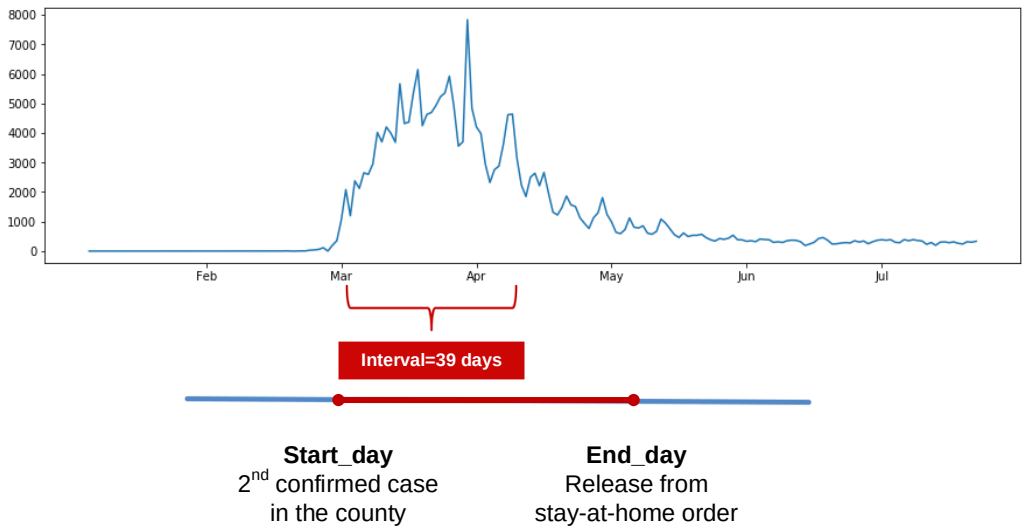


Рис. 2.1. Вибір інтервалу спостереження часового ряду.

У сценарії, коли між start_day і end_day у певному окрузі пройшло більше 39 днів, інтервал спостереження обмежувався 39-м днем після другого зареєстрованого випадку. У протилежному сценарії, коли не було достатньо даних між start_day і end_day для покриття 39-денного інтервалу, порожні дні заповнювалися нулями, а інтервал спостереження було розширено на додаткові 5 днів.

Вектор, створений для кожного округу, об'єднав 13 точок даних підтверджених випадків і 13 точок даних смертей, при цьому дані агрегувалися кожні три дні.

Аналіз проводився щодо 39-денного інтервалу спостереження. Таким чином, результати, вибрані в цьому дослідженні, включають 13 (39/3) нормалізовану кількість підтверджених випадків, логарифм загальної кількості підтверджених випадків, 13 (39/3) нормалізовану кількість випадків смерті та логарифми загальної кількості кількість смертей, розрахована для кожного округу. Отже, загальна кількість результатів для кожного округу становить 13 + 1 + 13 + 1 і дорівнює 28 (рис. 2.2).

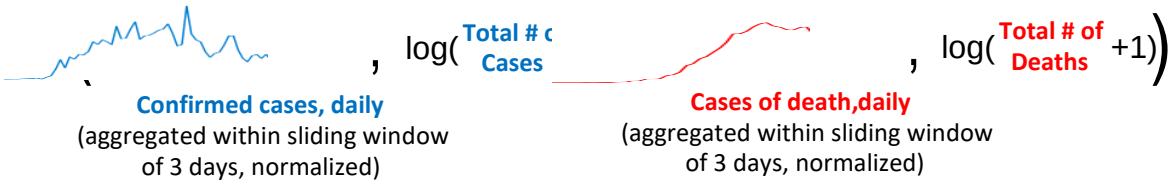


Рис. 2.2. Формування вектору ознак для округу.

Таким чином, відповідна інформація про поширення COVID-19 у кожному окрузі кодується 28-вимірним вектором $(x_1, \dots x_{14}, x_{15}, \dots, x_{28})$. Функція відстані на основі модифікованої евклідової відстані для вимірювання подібності поширення пандемії між округами:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_{14} - y_{14})^2} + \sqrt{(x_{15} - y_{15})^2 + \cdots + (x_{28} - y_{28})^2}$$

2.3. Побудова α-сусіднього графа з покриттям

Після цього, використовуючи двовимірну лінзу на основі Density та Centrality, був побудований результуючий α-сусідній граф з покриттям (рис. 2.3).

Centrality Density

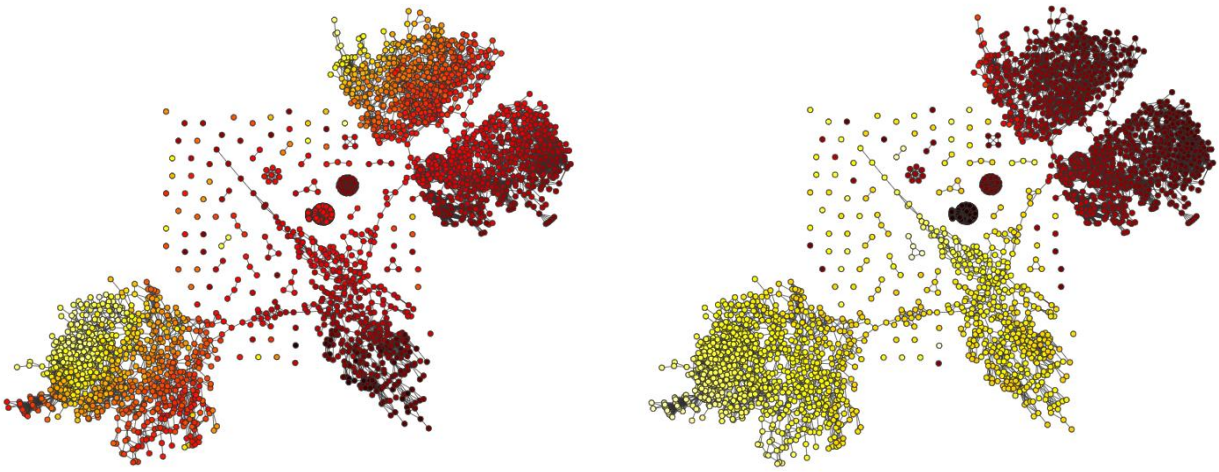
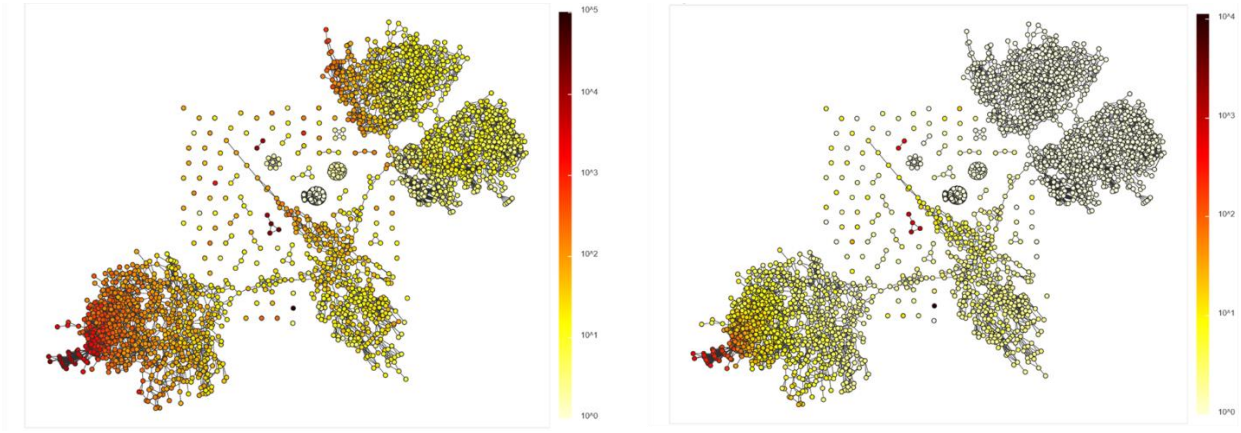


Рис. 2.3. α-сусідній граф з покриттям розповсюдження COVID-19 в 3142 округах США.

Основні особливості отриманого графу:

- Кожен вузол на графіку відповідає одному округу США (всього 3142 вузли).

- Результати включають кількість підтверджених випадків і смертей протягом 39-денного інтервалу спостереження від початку пандемії (рис. 2.4).
- Подібність округів з точки зору поширення пандемії вимірюється модифікованою функцією евклідової відстані.



а) б)

Рис. 2.4. Граф, розмальований за кількістю підтверджених випадків (а) і смертей (б)

2.4. Схеми аналізу топологічної моделі

Після виділення топологічної моделі даних дослідник візуально досліджує граф, щоб виявити цікаві підгрупи в даних. Ці підгрупи можна додатково вивчити, використовуючи стандартні статистичні методи для визначення предикторів, які можуть відповідати за подібність поширення пандемії, що спостерігається в межах визначеної підгрупи округів.

Дуже поширена ситуація в статистиці виникає, коли розподіл результату (або змінної відповіді) пов’язаний з одним або кількома предикторами (або пояснювальними змінними). Стандартним підходом, який використовують

дослідники для вивчення зв’язку між предиктором і результатом, є застосування відповідної статистичної моделі. Вибір моделі залежить від типів даних предиктора та результату (кількісний, бінарний, категорійний тощо) і часто передбачає додаткові припущення щодо розподілу результату. Щоб описати схожість у поширенні пандемії, ми інтегрували в нашу модель понад 250 предикторів, доступних із різноманітних загальнодоступних. У будь-який час додаткові предиктори можна легко інтегрувати в аналіз без будь-яких змін у топологічній моделі.

Інтерактивна візуалізація надає дослідникам можливість вручну виконувати візуальний огляд графіка, щоб визначити цікаві області. Наприклад, вузли, які утворюють вилоподібні структури або петлі, можуть представляти інтерес для подальших досліджень. Крім того, ізольовані компоненти або висококонцентровані групи вузлів, які утворюють спільноти, можуть вказувати на значущі зв’язки в наборі даних результатів. Виконуючи візуальний огляд, дослідник також може змінити колір графіка відповідно до медіанного значення результату або предиктора, вибраного з відповідних наборів даних. Використання кольорових кодів може підкреслити, чим підгрупа вузлів, представлена даною областю графа, може відрізнятися від решти вузлів.

Дослідник може вибрати будь-яку область графа, яка демонструє цікаві геометричні властивості, для подальшого статистичного аналізу. Після виконання статистичних тестів можна обчислити таблицю предикторів із відповідними *p*-значеннями, щоб визначити, чи відрізняється розподіл предикторів для вибраної підгрупи вузлів від розподілу решти вузлів. Якщо бажаний рівень значущості будь-якого предиктора виявляється статистично значущим, дослідник може побудувати гістограму, яка представляє нормалізовані частотні розподіли предиктора як для вузлів у вибраній області

графіка, так і для решти вузлів. Те саме можна зробити, порівнюючи між собою дві різні вибрані області вузлів.

З метою статистичного аналізу, проведеного для датасета, безперервні, змішані, бінарні та категоричні (небінарні) одновимірні предиктори були диференційовані відповідно до типу змінної. Безперервні предиктори перевіряли за допомогою стандартного двовибіркового тесту Манна–Уітні–Вілкоксона. Щоб перевірити статистичний зв’язок між двома вибірками в межах категоріальних даних, для бінарних і небінарних категоріальних змінних використовували точний критерій Фішера та критерій χ^2 відповідно.

2.5. Виділення ком'юніті на графі

Ключовою особливістю графа є структура ком'юніті. Зокрема, багато ребер з’єднують вершини в межах однієї ком'юніті, тоді як порівняно мало ребер з’єднує вершини між різними ком'юніті [7]. Можна вважати, що ці ком'юніті представляють незалежні структури в межах графа, і виявлення цих незалежних ком'юніті є однією з ключових цілей аналізу великих графів.

Метод Гірвана-Ньюмана

Алгоритм Гірвана-Ньюмана [8] намагається визначити ребра, які розташовані «між» деякими парами вузлів у графі. В алгоритмі обчислюється відстань між усіма парами вершин, тобто найкоротший шлях на основі ребра. Характеристика ребра edge betweenness — це кількість найкоротших шляхів між парами вершин, які проходять уздовж ребра.

Метод виявлення ком'юніті за допомогою алгоритму Гірвана-Ньюмена базується на обчисленні характеристики edge betweenness для всіх ребер графа. Спосіб включає в себе етапи видалення ребра, що має найвищу характеристику edge betweenness, і перерахунок характеристики edge betweenness для всіх ребер,

на які впливає видалення. Етапи повторюються, поки не залишиться ребер. Ребра, які мають найвищу характеристику edge betweenness, є найбільш «завантаженими». Видалення ребер з графа, які мають найвищі значення характеристики edge betweenness, призводить до того, що вершини потрапляють у ком'юніті.

Алгоритм Гірвана-Ньюмена застосовувався до різноманітних графів, наприклад, графів соціальних мереж людей і тварин, метаболічних графів, генних графів, графів, що представляють співпрацю між вченими та музикантами, тощо. Однак цей алгоритм потребує $O(m^2n)$ разів для графа з m ребрами та n вершинами. Через велику кількість часу, необхідного для виконання обчислень, використання алгоритму обмежене графами, які містять менше кількох тисяч вузлів. Крім того, алгоритм не показує, скільки ребер потрібно видалити, щоб забезпечити найоптимальніше виявлення ком'юніті.

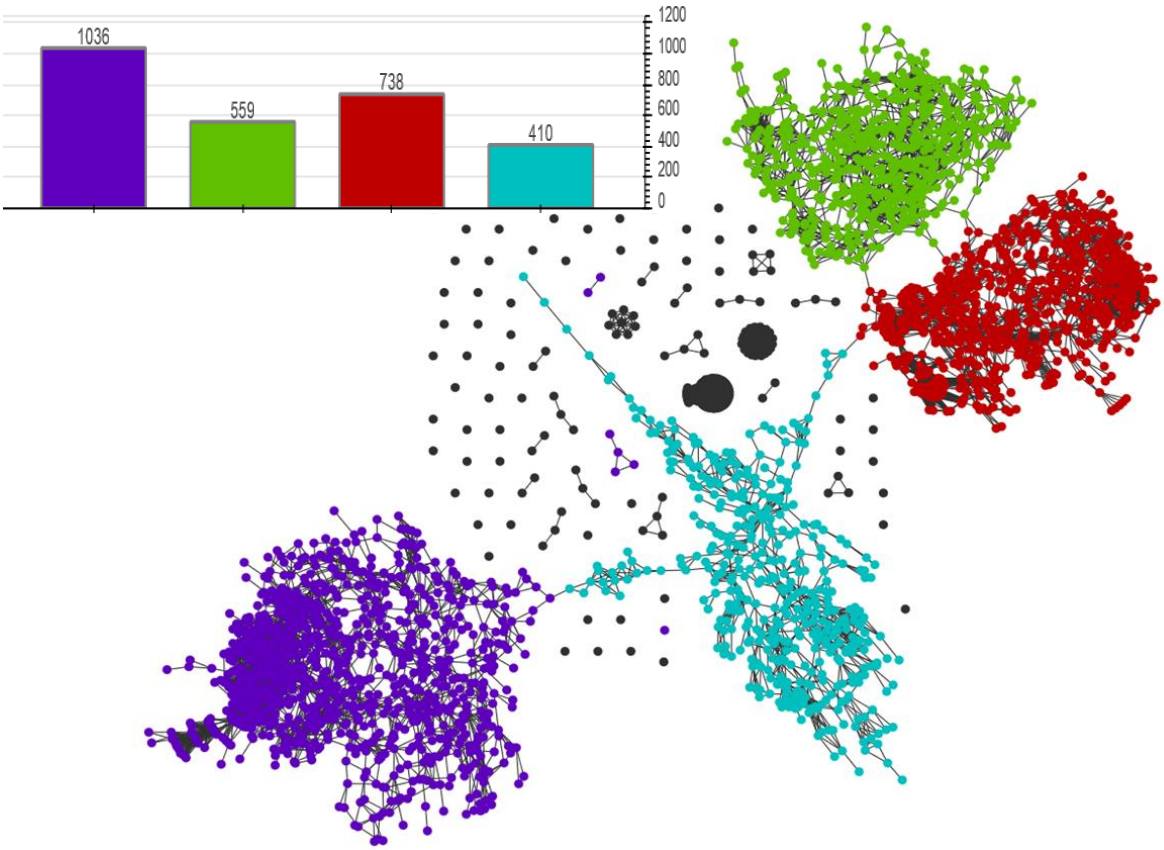


Рис. 2.5. Розбиття графа на ком'юніті методом Гірвана-Ньюмана

Дивлячись на геометричні властивості графа, отриманого з набору даних, який містить кількість підтверджених випадків і смертей протягом 39-денного інтервалу, ми можемо чітко виділити чотири основні регіони. За допомогою методу Гірвана-Ньюмана на графу було знайдено чотири громади, які мають більше схожості одна з одною, ніж з рештою округів (рис 2.5).

Метод перколяції кліки

Метод перколяції кліки [9] працює на основі припущення, що внутрішні ребра в межах ком'юніті утворюють k-кліку (тобто підграфи з k вершинами, у яких кожна пара вершин з’єднана ребром), а ребра, які лежать між ком'юніті, не можуть сформувати кліки.

Використання цього методу базується на припущенні, що якщо кліка може «рухатися» по графу, кліка опиниться в пастці всередині ком'юніті і не зможе пройти між двома ком'юніті через відсутність сполучних шляхів. У цьому методі одну кліку можна «перемістити» до іншої, якщо вони мають спільні всі вершини, крім однієї, а ком'юніті визначається як максимальний зв’язний підграф початкового графа, так що кожена вершина у цьому графі належить до деякої k-кліки, яка лежить повністю в підграфі.

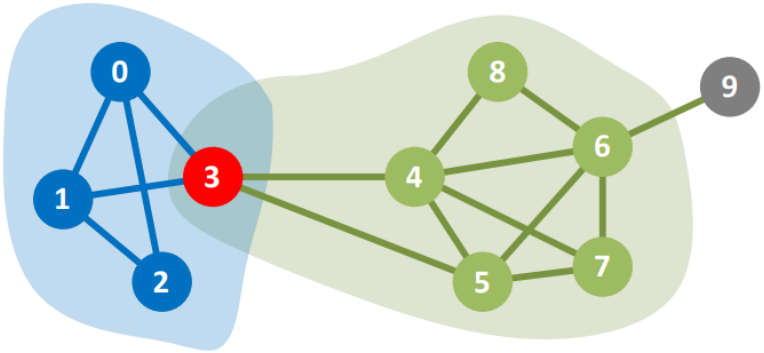


Рис. 2.6. Приклад виявлення ком'юніті, що накладаються, за допомогою перколяції 3-кліки на графу

Класичний метод перколяції кліки отримує значення k як вхідні дані та створює список усіх можливих ком'юніті як вихідні дані. Особливості методу включають здатність деяких вузлів належати до кількох ком'юніті, оскільки через ці вершини можуть проходити декілька k-клік, а також здатність деяких вершин бути поза ком'юніті, оскільки жодна з k-клік їх не містить.

Приклад методу перколяції 3-клік можна побачити на рис. 2.6. 3-кліка не може пройти через вузол 3, але синє ком'юніті та зелене ком'юніті перекриваються у вузлі 3.

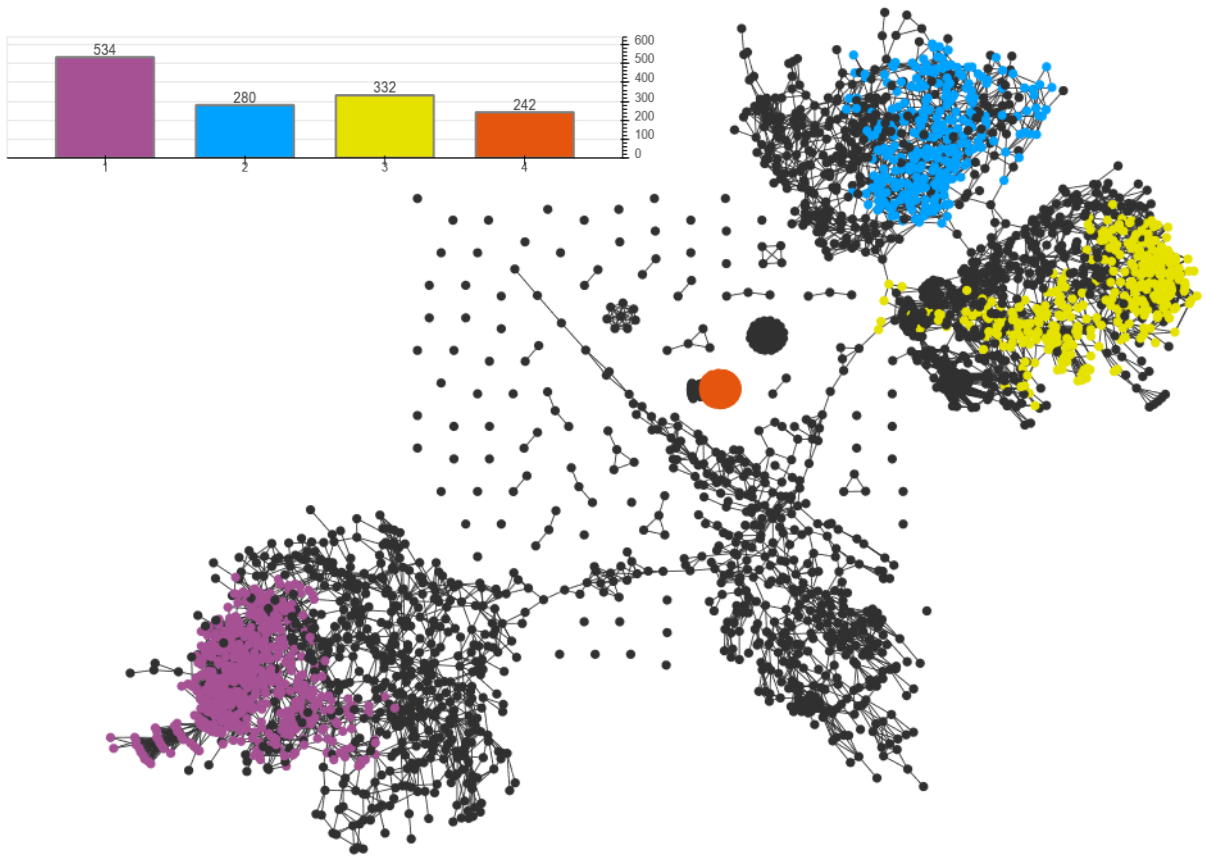


Рис. 2.7 Розбиття графа на ком'юніті методом перколяції клік. Чотири найбільших ком’юніті.

Хоча теоретично метод перколяції кліки вимагає обчислень, оскільки для виявлення максимальних клік потрібен час обробки, який експоненціально залежить від розміру графа, практичне застосування цього алгоритму до систем реального світу показало, що цей метод працює розумно швидко через обмежену кількість клік у реальних графах.

Метод перколяції кліки було застосовано до графа для виявлення ком'юніті, які неможливо ідентифікувати, якщо використовувати лише методи візуального дослідження. Рис. 2.7 ілюструє результат алгоритму пошуку ком'юніті методом перколяції для чотирьох найбільших ком'юніті. Зосереджемо аналіз на порівнянні синіх і жовтих ком'юніті, оскільки вони демонструють

подібні закономірності щодо кількості підтверджених випадків і кількості смертей (див. кольори вузлів на рис. 2.4).

2.6. Статистичний аналіз ком‘юніті на графі

Розглянемо два ком’юніті, отримані методом Гірвана-Ньюмана (рис. 2.5) фіолетове та зелене. Можно побачити, вони мають подібні моделі щодо кількості підтверджених випадків (рис. 2.4) порівняно з двома іншими ком’юніті (блакитне у центрі та червоне внизу праворуч). З іншого боку, ці два ком’юніті, фіолетове та зелене, мають різні закономірності щодо кількості смертей: зелене ком’юніті, у верхньому правому куті, має набагато менше смертей. Іншими словами, коли почалася пандемія, подібна картина поширення спостерігалася як у фіолетових, так і в зелених громадах. У той же час кількість смертей свідчить про те, що округи, розташовані в зеленому ком’юніті, ефективніше впоралися з вірусом, що значно знизило кількість летальних випадків, незважаючи на те, що у фіолетовому ком’юніті майже вдвічі більше округів, ніж у зеленому (1036 проти 559).

Після проведення статистичних тестів у цих двох спільнотах було розраховано таблицю предикторів із відповідними p-значеннями, щоб визначити, чи відрізняється розподіл предикторів для вибраної фіолетової підгрупи вузлів від розподілу вибраної підгрупи зелених вузлів. Якщо бажаний рівень значущості будь-якого предиктора виявився статистично значущим, ми змогли побудувати гістограму, що представляє нормалізовані частотні розподіли предиктора як для вузлів у вибраній фіолетовій області графіка, так і для вузлів із зеленої області.

Для цілей статистичного аналізу, проведеного для цього експерименту, безперервні, змішані, бінарні та категоричні (небінарні) одновимірні предиктори були диференційовані відповідно до типу змінної. Безперервні предиктори

перевіряли за допомогою стандартного двовибіркового тесту Манна–Уітні–Вілкоксона. Щоб перевірити статистичний зв'язок між двома вибірками в межах категоріальних даних, для бінарних і небінарних категоріальних змінних використовували точний критерій Фішера та критерій χ^2 відповідно.

Рис. 2.8 показує гістограму, яка порівнює популяцію в обох ком'юніті. Схоже, що фіолетова спільнота має більш густонаселені округи, включаючи міста та великі міські центри. Навпаки, зелене співтовариство виявилося з помірно населеними округами, переважно розташованими в сільській місцевості.

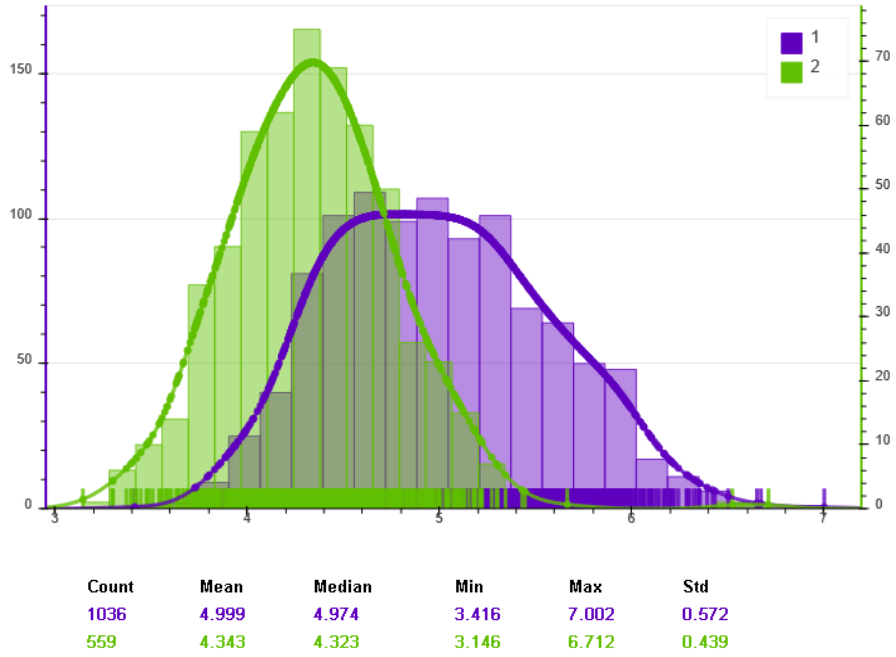
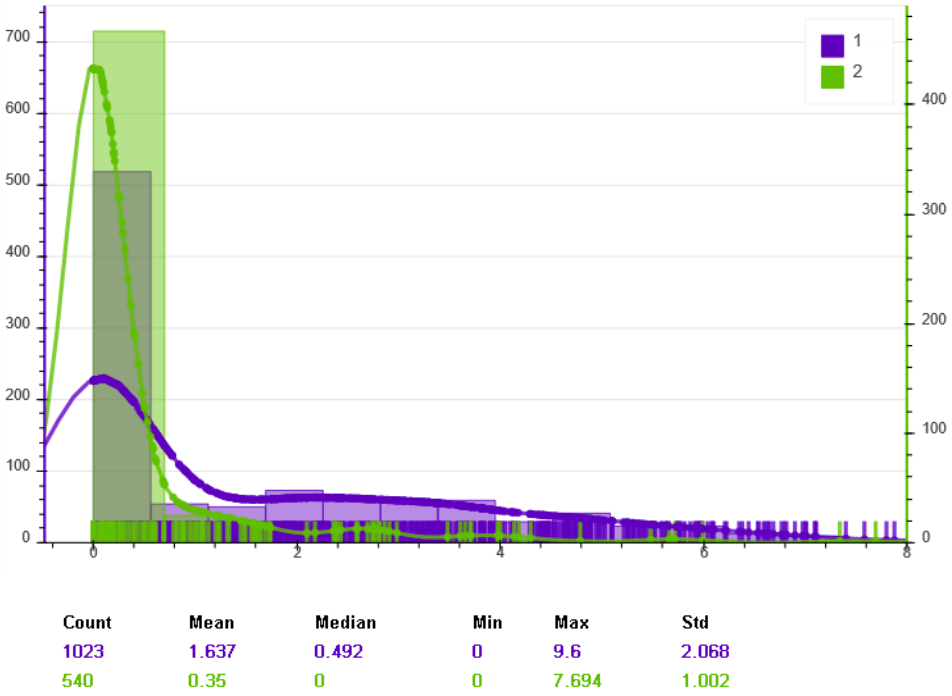


Рис. 2.8. Гістограма, що показує розподіл логарифма кількості населення в зелених і фіолетових громадах.

Іншим статистично значимим предиктором є система громадського транспорту, яка може бути причиною різниці в кількості смертей між зеленими та фіолетовими громадами. Ми порівняли ці громади на основі оцінки громадського транспорту (див. рис. 2.9). Оцінка громадського транспорту показує, наскільки добре місце обслуговується громадським транспортом.

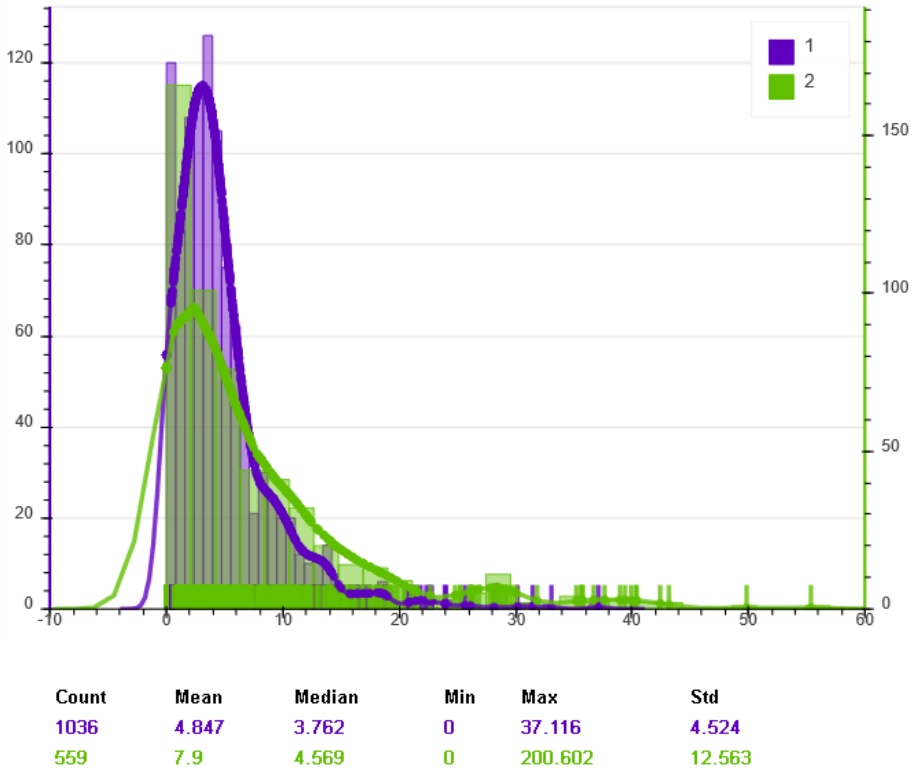
Виявилося, що фіолетова громада має набагато вищу оцінку громадського транспорту, ніж зелена.



Count	Mean	Median	Min	Max	Std
1023	1.637	0.492	0	9.6	2.068
540	0.35	0	0	7.694	1.002

Рис. 2.9. Гістограма показує оцінку громадського транспорту в зелених і фіолетових громадах

Інша група предикторів, які можуть бути відповідальними за різницю в кількості смертей у фіолетових і зелених громадах, стосується системи охорони здоров'я. На рис. 2.10 показана гістограма порівняння фіолетових і зелених спільнот. Зелена громада має вдвічі більше лікарень на 100 000 людей, ніж фіолетова. Це може бути однією з причин того, що округи зеленого співтовариства ефективніше впоралися зі спалахом пандемії, навіть незважаючи на те, що кількість випадків у них така ж, як і в округах фіолетового співтовариства.



Count	Mean	Median	Min	Max	Std
1036	4.847	3.762	0	37.116	4.524
559	7.9	4.569	0	200.602	12.563

Рис. 2.10. Гістограма показує кількість лікарень у зелених і фіолетових громадах.

Далі розглянемо ком'юніті, отримані методом перколяції кліки (рис. 2.7). Виявлення ком'юніті ґрунтується на виявленні підгруп щільно з'єднаних вузлів, де багато ребер з'єднують вузли однієї спільноти та порівняно мало ребер, що з'єднують вузли різних спільнот.

Ми зосереджуємо наш аналіз на порівнянні синіх і жовтих ком'юніті, оскільки вони демонструють подібні закономірності щодо кількості підтверджених випадків і кількості смертей (див. кольори вершин на рис. 2.4).

Першим статистично значущим предиктором для ком'юніті виявилося, що кількість днів, що минула між датою, коли було введено обмеження на перебування вдома, і днем початку інтервалу спостереження, тобто датою виявлення другого підтвердженого випадку в окрузі (рис. 2.11).

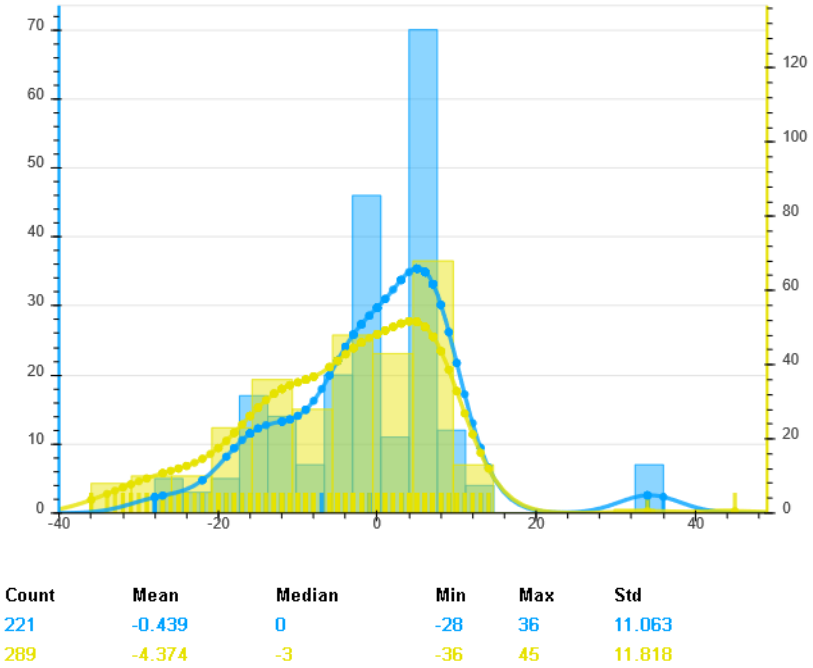
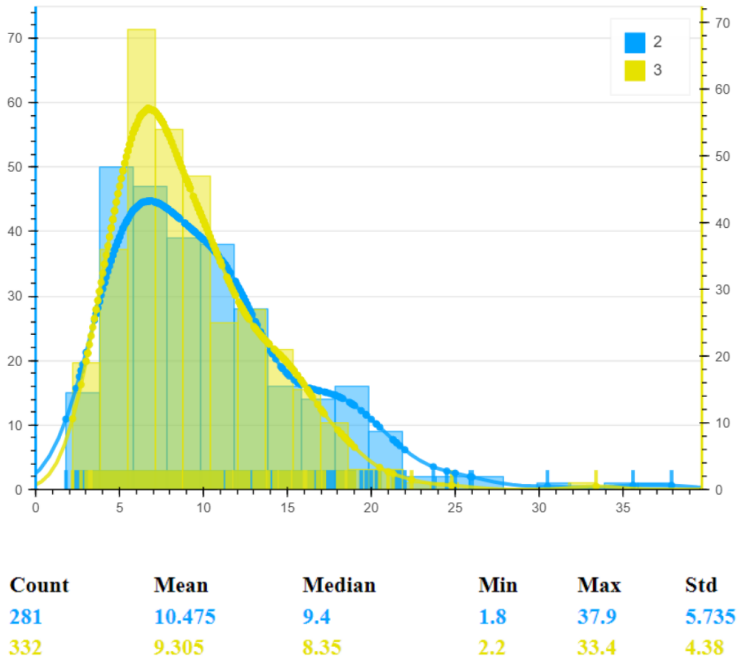


Рис. 2.11. Гістограма, що порівнює жовті та сині громади відповідно до кількості днів, що минули між датою введення обмеження на перебування вдома та датою початку виявлення підтверджених випадків у кожному окрузі.

Гістограма на рис. 2.12 показує відсоток населення без медичної страховки, дозволяє нам зробити висновок, що в синій громаді більше людей без страхового покриття.



Count	Mean	Median	Min	Max	Std
281	10.475	9.4	1.8	37.9	5.735
332	9.305	8.35	2.2	33.4	4.38

Рис. 2.12. Гістограма показує відсоток населення без медичного страхування в синіх і жовтих громадах

Статистичний аналіз даних вказав на різні предиктори, які можуть вплинути на поширення COVID-19 у США. До них належать географічні умови, щільність населення, громадський транспорт, міграція, кількість лікарень, кількість лікарів загальної практики, кількість днів, що минули між введенням обмеження на перебування вдома та датою початку виявлення. підтверджених випадків захворювання та багато інших. У будь-який час додаткові предиктори можна легко інтегрувати в модель, щоб розширити пошук параметрів, які можуть відповідати за схожість у поширенні пандемії, виявлену за допомогою TDA, алгоритмів машинного навчання та методів візуального дослідження.

Висновки

У цій роботі представлено огляд методів топологічного аналізу даних та застосування їх для дослідження поширення COVID-19 у кожному окрузі США. Була розроблена концепція α-сусіднього графа з покриттям та цей підхід був застосований до даних, вилучених з відкритих джерел. Аналіз виконувався за окремими округами шляхом вилучення з набору даних топологічної моделі, представленої у вигляді графа, на якому кожен із 3142 вузлів відповідає одному округу. Після побудови графа дані реального світу використовувалися для інтеграції в моделі прогностичних факторів, які могли бути відповідальними за схожість на ранній стадії поширення пандемії. У ході експерименту було використано понад 250 предикторів з різних загальнодоступних джерел. Експерименти з аналізу даних вказали на різні предиктори, які можуть вплинути на поширення COVID-19 у США. До них належать географічні умови, щільність населення, вплив міст, громадський транспорт, довжина шосе, міграція, кількість лікарень, кількість лікарів загальної практики, кількість днів, що минули між введенням обмеження на перебування вдома та датою початку виявлення. підтверджених випадків захворювання та багато інших. У будь-який час додаткові предиктори можна легко інтегрувати в модель, щоб розширити пошук параметрів, які можуть відповідати за схожість у поширенні пандемії, виявлену за допомогою TDA, алгоритмів машинного навчання та методів візуального дослідження.

Список використаних джерел

[1] Boissonnat, J.-D., Chazal, F., and Yvinec, M., «Geometric and Topological,» *Cambridge University Press,* т. 57, 2018.

[2] D. Sheffar, «Introductory Topological Data Analysis,» 2020.

[3] Singh, G.; Mémoli, F.; Carlsson, G., «Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition,» *Eurographics Symposium on Point-Based Graphics,* 2007.

[4] Borg, I.; Groenen, P., Modern Multidimensional Scaling: theory and applications (2nd ed.), New York: Springer-Verlag, 2005, pp. 207-212.

[5] Laurens van der Maaten; Geoffrey Hinton, «Visualizing Data using t-SNE,» *Journal of Machine Learning Research,* pp. 2579-2605, Nov 2008.

[6] «COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins,» [Онлайновий]. Available: https://github.com/CSSEGISandData/COVID-19.

[7] S. Fortunato, «Community detection in graphs,» *Physics Reports,* т. 486, pp. 75-174, 2010.

[8] M. Girvan, M. Newman, «Community structure in social and biological networks,» *Proceedings of the National ,* т. 99, № 12, pp. 7821-7826, 2002.

[9] G. Palla, I. Derenyi, I. Farkas, T. Vicsek, «Uncovering the overlapping community structure of complex,» *Nature,* т. 435, pp. 814-818, 2005.