Detekcja Anomalii w Członkostwach Miast w Sieciach Transnarodowych za Pomocą Sztucznej Inteligencji

Rzyszka Kamil[a] (xxx), Seniush Yevhenii[a] (xxx), Skowron Dawid[a] (xxx) and Żuliński Marek[a] (xxx)

[a]*Zieloni VaBanque, Prószkowska 29, Opole, 45-962, Poland*

ARTICLE INFO

Keywords:
city networks
anomaly detection
artificial intelligence
urban collaboration
environmental governance
machine learning
data analysis

ABSTRACT

Miasta odgrywają kluczową rolę w globalnej współpracy na rzecz zrównoważonego rozwoju poprzez członkostwo w transnarodowych sieciach miejskich (TMNs). W artykule przedstawiamy zastosowanie sztucznej inteligencji (AI) do detekcji anomalii w danych dotyczących 10 343 miast z 208 krajów. Wykorzystanie algorytmów uczenia maszynowego pozwala na identyfikację nietypowych wzorców zaangażowania, co pomaga zrozumieć unikalne strategie miast, lokalne bariery oraz potencjalne błędy w danych. Wyniki tej analizy mogą wspierać opracowanie bardziej efektywnych strategii współpracy i działań środowiskowych.

1. Wstęp

Współczesne miasta coraz częściej odgrywają kluczową rolę w globalnych inicjatywach na rzecz zrównoważonego rozwoju i ochrony środowiska, czego przykładem jest ich uczestnictwo w transnarodowych sieciach miejskich (TMNs). Zrozumienie wzorców członkostwa i zaangażowania miast na tej arenie ma zasadnicze znaczenie dla projektowania efektywnych strategii współpracy międzynarodowej i środowiskowej. Jednakże analiza tak obszernych i złożonych zbiorów danych, jak te obejmujące informacje o 10 343 miastach z 208 krajów oraz ich członkostwach w 84 sieciach, napotyka na istotne wyzwania.

Jednym z kluczowych aspektów analizy danych tego rodzaju jest detekcja anomalii, czyli identyfikacja miast, które wykazują nietypowe wzorce zaangażowania. Anomalie te mogą być wskaźnikiem wyjątkowych strategii, potencjalnych błędów lub specyficznych uwarunkowań lokalnych. W tym celu coraz częściej stosuje się zaawansowane techniki sztucznej inteligencji (AI), które pozwalają na efektywne wykrywanie i analizowanie tych odstępstw od normy. Algorytmy uczenia maszynowego, takie jak lasy losowe, maszyny wektorów nośnych (SVM) czy metody klastrowania, są wykorzystywane do identyfikacji wzorców w dużych i różnorodnych zbiorach danych.

W niniejszym artykule omówimy zastosowanie sztucznej inteligencji w kontekście detekcji anomalii w danych dotyczących członkostw miast w TMNs. Zaprezentujemy korzyści wynikające z zastosowania algorytmów AI, przedstawimy przykłady przypadków użycia, a także omówimy implikacje, jakie wykryte anomalie mogą mieć dla zrozumienia globalnej dynamiki współpracy miejskiej oraz działań środowiskowych.

2. Przegląd literatury

Detekcja anomalii jest powszechnie stosowaną techniką w analizie danych, szczególnie w kontekście dużych zbiorów danych, gdzie trudno jest przewidzieć wzorce zachowań. W literaturze istnieje wiele podejść do wykrywania anomalii, zarówno w kontekście ogólnych danych, jak i bardziej specyficznych dziedzin, takich jak sieci miejskie czy analiza danych związanych z miastami.

2.1. Detekcja Anomalii w Zbiorach Danych

Współczesne podejścia do detekcji anomalii w zbiorach danych opierają się na metodach statystycznych oraz algorytmach uczenia maszynowego. Anomalie mogą przybierać różne formy, od punktów odstających w danych numerycznych po nieoczekiwane wzorce w danych kategorycznych. Przegląd technik wykorzystywanych w analizie anomalii zawiera metody oparte na klasteryzacji, takie jak K-means (Lloyd, 1982), oraz algorytmy oparte na gęstości, jak DBSCAN (Ester, Kriegel, Sander and Xu, 1996). Inne metody, takie jak Local Outlier Factor (LOF) (Breunig, Kriegel, Ng and Sander, 2000), wykorzystują lokalne cechy danych do wykrywania punktów odstających.

2.2. Zastosowanie Sztucznej Inteligencji w Sieciach Miejskich

Sieci miejskie (TMNs) stały się jednym z kluczowych tematów badań w kontekście zrównoważonego rozwoju miast. Badania nad tymi sieciami koncentrują się na zrozumieniu ich struktury, dynamiki współpracy oraz wpływu na polityki miejskie (Mossberger, Tolbert and McNeal, 2013). Wykorzystanie sztucznej inteligencji w analizie danych związanych z TMNs staje się coraz bardziej powszechne, szczególnie w kontekście wykrywania wzorców, które mogą wskazywać na nieefektywne strategie lub anomalie w zaangażowaniu miast (Benk and Rainer, 2019).

2.3. Klasteryzacja i Detekcja Anomalii w Sieciach

W literaturze dotyczącej detekcji anomalii w kontekście sieci, szczególnie sieci miejskich, klasteryzacja jest często wykorzystywaną metodą. K-means i Hierarchical Clustering są popularnymi narzędziami do identyfikacji nieoczekiwanych skupisk danych, które mogą wskazywać na anomalie w strukturze sieci (Jain, 2010). Metody te pozwalają na segmentację danych oraz identyfikację miast, które różnią się od standardowych wzorców zaangażowania w sieci.

2.4. Wyzwania w Analizie Danych Miast

Analiza danych dotyczących członkostw miast w sieciach transnarodowych wiąże się z licznymi wyzwaniami. Przede wszystkim, dane te są często heterogeniczne i zawierają wiele zmiennych, które mogą wpływać na wyniki analizy. Dodatkowo, zmienność miast pod względem ich rozmiaru, poziomu rozwoju oraz regionu geograficznego może wpływać na charakterystyki ich zaangażowania w sieciach (Roth and Hirt, 2015). W literaturze zwraca się uwagę na potrzebę dostosowywania metod analitycznych do specyfiki takich zbiorów danych (Wasserman and Faust, 2020).

3. Zbiór danych

3.1. Charakterystyka zbioru danych

Wykorzystany zbiór danych obejmuje informacje o 10 343 miastach z 208 krajów oraz ich członkostwach w 84 transnarodowych sieciach miejskich (TMNs). Dane zawierają szczegóły dotyczące aktywności miast w sieciach, takie jak liczba członkostw, specyficzne sieci, do których należą, oraz kategorie tematyczne współpracy (np. zrównoważony rozwój, innowacje miejskie, ochrona środowiska). Miasta w zbiorze danych reprezentują szeroki zakres wielkości, regionów geograficznych i poziomów rozwoju, co czyni analizę bardziej zróżnicowaną i kompleksową.

3.2. Źródło danych

Dane pochodzą z publicznie dostępnego zbioru danych *City Networks Membership Dataset*, opublikowanego na platformie Figshare (`https://figshare.unimelb.edu.au/articles/dataset/city_networks_membership_dataset/21330996`). Zbiór ten został przygotowany przez badaczy w celu analizy zaangażowania miast w globalne sieci współpracy.

3.3. Wstępne przetwarzanie danych

Przed analizą dane zostały poddane wstępnemu przetwarzaniu, które obejmowało następujące kroki:

- Usuwanie braków danych: Wiersze i kolumny z brakującymi wartościami zostały usunięte lub uzupełnione na podstawie średnich lub median odpowiednich wartości.
- Standaryzacja: Wszystkie zmienne numeryczne zostały znormalizowane w celu zapewnienia spójności analizy.

3.4. Wizualizacja danych

W celu lepszego zrozumienia struktury danych oraz wyników analizy przedstawiono poniższą wizualizację.

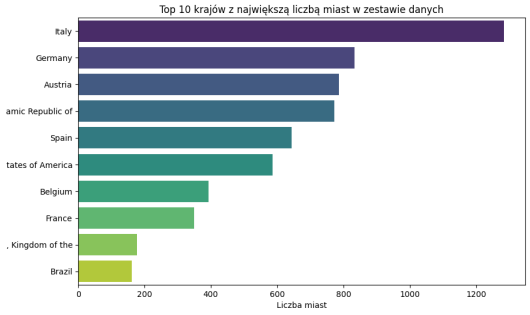

Figure 1: Top 10 krajów z największą liczbą miast w zestawie danych.

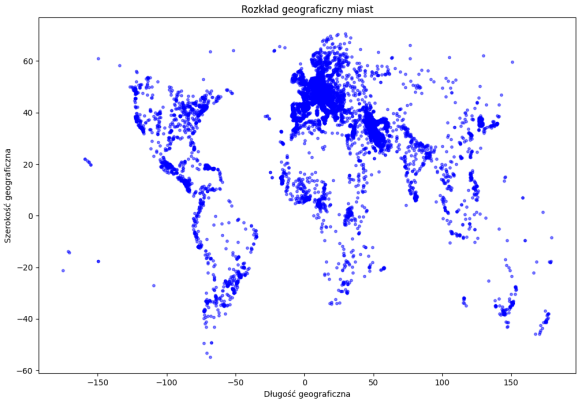

Figure 2: Rozkład geograficzny miast członkowskich.

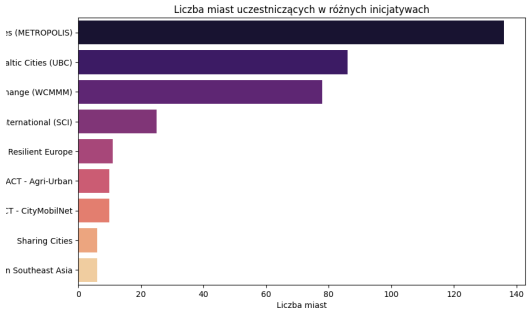

Figure 3: Liczba miast uczestnicząca w różnych inicjatywach.

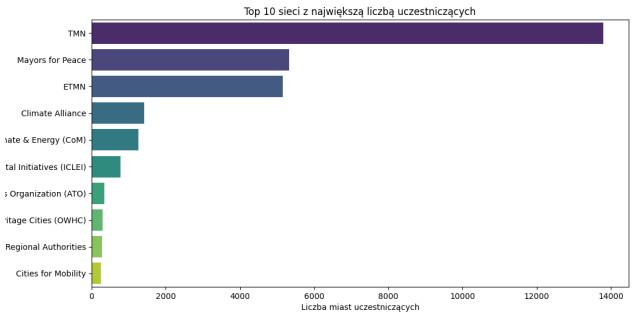

Figure 4: Top 10 sieci z największą liczbą uczestniczących miast.

4. Metodologia

4.1. Techniki analizy danych

K-means dla detekcji anomalii: K-means jest algorytmem klasteryzacji, który grupuje dane w zbiory na podstawie ich podobieństwa. W tym przypadku zastosowano go do identyfikacji anomalii w zbiorze danych, gdzie mniejsze klastry mogą sugerować nietypowe zachowania. Proces obejmował następujące kroki: —Wczytanie i normalizację danych (kolumny 'latitude.y', 'longitude.y', 'TMN', 'ETMN'). -Określenie optymalnej liczby klastrów za pomocą metody łokcia, a następnie trening modelu K-means z optymalną liczbą klastrów (w tym przypadku 5). -Identyfikacja małych klastrów, które mogą wskazywać na anomalie, poprzez analizę rozmiarów klastrów. -Wizualizacja rozmieszczenia geograficznego klastrów oraz anomalii, które zostały zidentyfikowane.

- Wczytanie i normalizację danych (kolumny 'latitude.y', 'longitude.y', 'TMN', 'ETMN').
- Określenie optymalnej liczby klastrów za pomocą metody łokcia, a następnie trening modelu K-means z optymalną liczbą klastrów (w tym przypadku 5).
- Identyfikacja małych klastrów, które mogą wskazywać na anomalie, poprzez analizę rozmiarów klastrów.
- Wizualizacja rozmieszczenia geograficznego klastrów oraz anomalii, które zostały zidentyfikowane.

Hierarchical Clustering: Hierarchical Clustering to technika, która tworzy drzewo hierarchiczne (dendrogram) do analizy struktury danych. Dendrogram pozwala na wizualizację odległości między obiektami w zbiorze. W tym przypadku wykorzystano metodę Ward'a w połączeniu z analizą kategorii sub-regionów i sieci.

- Znormalizowano dane numeryczne, a następnie stworzono dendrogram.
- Dokonano analizy zmienności sub-regionów i przypisano odpowiednie klastry do różnych obiektów.
- Wykorzystano PCA do wizualizacji klastrów w przestrzeni 2D oraz przeprowadzono detekcję anomalii przy użyciu Local Outlier Factor (LOF) i analizy odległości od centroidów.

Clustering Sub-Regions and Networks: Metoda ta polegała na podziale zbioru danych według subregionów geograficznych i specyficznych sieci TMNs. Zastosowanie algorytmu DBSCAN w celu analizy sub-regionów i sieci geograficznych:

- Wykorzystano dane z kolumny sub-region, a także dane binarne związane z sieciami, które zostały znormalizowane.
- Dzięki PCA dokonano redukcji wymiarowości i wizualizacji klastrów.
- Dodatkowo, przeprowadzono detekcję anomalii w przestrzeni geograficznej (na podstawie długości i szerokości geograficznej), gdzie punkty oznaczone jako anomalie były reprezentowane na mapie.

4.2. Uzasadnienie wyboru metod

K-means: Został wybrany do wykrywania anomalii, ponieważ dobrze sprawdza się w identyfikacji nietypowych danych na podstawie liczby klastrów. Dzięki metodzie łokcia udało się określić optymalną liczbę klastrów, co jest kluczowe do dalszej analizy anomalii.

Hierarchical Clustering: Ta metoda pozwala na bardziej elastyczne grupowanie danych, bez konieczności wcześniejszego ustalania liczby klastrów. Jest szczególnie przydatna w analizie danych z bardziej złożoną strukturą hierarchiczną, jak sub-regiony. Użycie różnych metod linkowania (np. Ward, single, complete) pozwala na dostosowanie poziomu szczegółowości klasteryzacji.

DBSCAN: Wybrano ze względu na jego zdolność do wykrywania anomalii i klastra o nieregularnym kształcie. DBSCAN jest szczególnie skuteczny w przypadku, gdy dane są rozproszone i zawierają szumy, co sprawdza się w analizie sub-regionów oraz sieci. Algorytm pozwala na wyodrębnienie anomalii jako punkty, które nie należą do żadnego klastra.

5. Wyniki i dyskuja

W ramach analizy wykrywania anomalii przy użyciu różnych metod klasteryzacji, takich jak K-means, Hierarchical Clustering oraz Sub-region/Network clustering, uzyskano różne wyniki, które pozwalają na wyciągnięcie istotnych wniosków.

5.1. K-means

Metoda K-means wykryła 147 anomalii. Wykorzystując algorytm K-means, podzieliliśmy dane na pięć klastrów, a anomalie zostały zidentyfikowane jako małe klastry o niskiej liczbie punktów. Obserwujemy, że ta metoda skutecznie grupuje dane, ale nie jest idealna w wykrywaniu bardzo małych anomalii, które mogą być istotne w kontekście specyficznych przypadków.

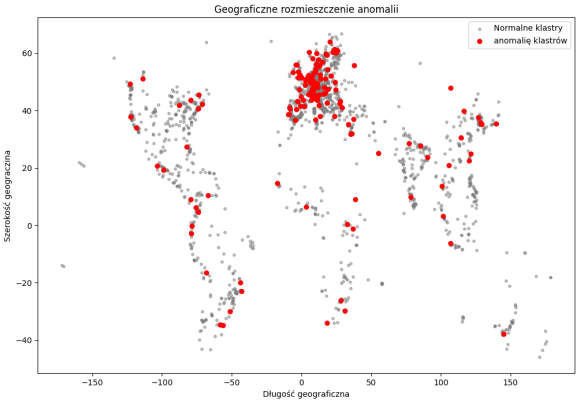

Figure 5: Graficzne rozmieszczenie anomalii K-means.

5.2. Hierarchical Clustering

Hierarchiczne grupowanie danych wykazało znacznie mniejszą liczbę anomalii – 5 anomalii. Zastosowanie tego algorytmu pozwala na lepsze zrozumienie struktury hierarchicznej danych i może być przydatne, gdy zależy nam na analizie relacji między obiektami na różnych poziomach. Mimo to, liczba wykrytych anomalii w tym przypadku była bardzo mała, co sugeruje, że metoda ta może być bardziej konserwatywna w identyfikowaniu nieprawidłowości.

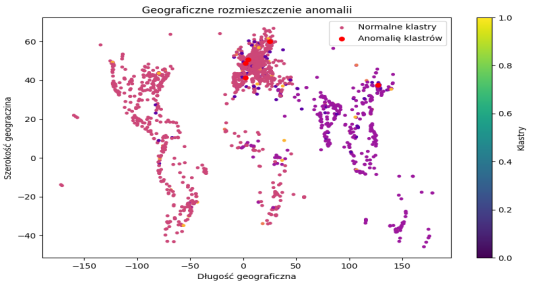

Figure 6: Graficzne rozmieszczenie anomalii Hierarchical Clustering.

5.3. Sub-region/Network Clustering (DBSCAN)

W przypadku analizy z wykorzystaniem algorytmu DBSCAN dla sub-regionów i sieci, wykryto aż 996 anomalii. DBSCAN jest bardziej elastyczny w wykrywaniu anomalii, szczególnie w danych, które mogą zawierać szumy lub nie mają wyraźnych granic między klastrami. Zidentyfikowanie tak dużej liczby anomalii może świadczyć o tym, że dane są bardziej rozproszone i zawierają więcej nietypowych punktów, które nie pasują do standardowych klastrów.

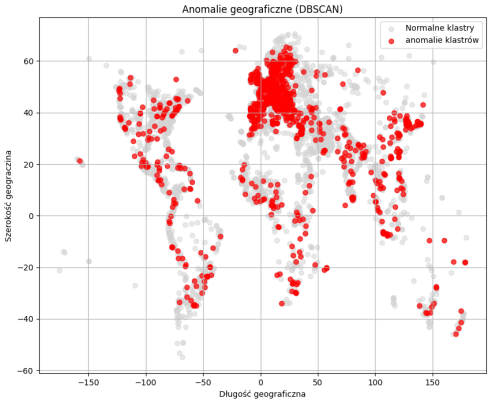

Figure 7: Graficzne rozmieszczenie anomalii Sub-region/Network Clustering.

6. Znaczenie wykrytych anomalii

Wykrywanie anomalii w zbiorach danych jest kluczowym zadaniem w wielu dziedzinach, w tym w analizie danych przestrzennych, monitoringu sieci, oraz analizie systemów miejskich. Zidentyfikowane anomalie mogą dostarczyć cennych informacji na temat nietypowych, podejrzanych lub nieprawidłowych zdarzeń, które wymagają dalszej uwagi. Oto kilka aspektów, które podkreślają znaczenie wykrytych anomalii:

6.1. Identyfikacja nieprawidłowości i potencjalnych zagrożeń

Anomalie mogą wskazywać na występowanie nieprawidłowości w systemach, które mogą prowadzić do poważnych

problemów, takich jak awarie, błędy w danych, czy incydenty bezpieczeństwa. Na przykład, w analizie sieci miejskich, anomalie w danych geograficznych mogą wskazywać na nieautoryzowane lub nieprawidłowe działania, które mogą stanowić zagrożenie dla integralności systemu.

6.2. Optymalizacja zarządzania i podejmowanie decyzji

Wykrywanie anomalii może pomóc w optymalizacji zarządzania zasobami. Na przykład, w systemach miejskich, anomalie w danych o ruchu drogowym mogą wskazywać na nieefektywności w infrastrukturze transportowej, umożliwiając wprowadzenie poprawek w celu poprawy płynności ruchu i redukcji korków. Dzięki temu władze lokalne mogą podejmować lepsze decyzje w zakresie rozwoju infrastruktury.

6.3. Wykrywanie oszustw i nadużyć

W kontekście analizy sieci społecznych lub transakcji, anomalie mogą wskazywać na podejrzane zachowania lub oszustwa. Na przykład, w systemach bankowych, wykrycie nieprawidłowych wzorców w transakcjach może pomóc w identyfikacji oszustw finansowych. Analiza anomalii pozwala na wczesne wykrycie i zapobieganie niepożądanym zdarzeniom.

6.4. Doskonalenie algorytmów i modeli analitycznych

Wykryte anomalie mogą również służyć jako przypadki testowe do udoskonalania algorytmów wykrywania anomalii. Przez analizę anomalii w różnych metodach klasteryzacji (takich jak K-means, DBSCAN czy Hierarchical Clustering), możemy lepiej zrozumieć, które podejście jest najbardziej efektywne w danym kontekście, a także które rodzaje anomalii są najbardziej istotne do dalszej analizy.

7. Wnioski

W przeprowadzonej analizie wykrywania anomalii przy użyciu różnych metod klasteryzacji, takich jak K-means, Hierarchical Clustering oraz Sub-region Clustering, uzyskano interesujące wyniki, które mogą przyczynić się do lepszego zrozumienia struktury i nieprawidłowości w badanym zbiorze danych. Oto kluczowe wnioski:

7.1. Porównanie liczby anomalii

K-means: Metoda K-means wykryła 147 anomalii, co wskazuje na obecność licznych, nietypowych punktów w zbiorze danych. Choć K-means jest skuteczny w segmentowaniu danych, wyniki te mogą być efektem zbyt dużej liczby klastrów, w których dane są podzielone na zbyt małe grupy.

Hierarchical Clustering: Ta metoda wykryła tylko 5 anomalii, co sugeruje, że struktura danych jest bardziej spójna i hierarchiczna. Hierarchical Clustering, dzięki swojej zdolności do analizowania danych na różnych poziomach, może lepiej identyfikować naturalne grupy w danych i w tym przypadku wykryć mniej, ale bardziej istotnych anomalii.

Sub-region Clustering: Wykryto 996 anomalii, co wskazuje na bardzo zróżnicowaną strukturę w obrębie różnych podregionów. Taki wynik może sugerować, że dane w obrębie podregionów są mocno zróżnicowane, a zmienność w tych obszarach jest większa niż w innych metodach klasteryzacji.

7.2. Wizualizacja anomalii

Wizualizacje wykazały, że w każdym przypadku anomalie są rozmieszczone w różnych częściach danych. W przypadku K-means i Sub-region Clustering, anomalie były bardziej rozproszone w przestrzeni, co może świadczyć o bardziej rozdrobnionych strukturach w tych metodach. Z kolei w Hierarchical Clustering anomalie były bardziej skonsolidowane, co odzwierciedla spójną strukturę w danych.

7.3. Różnice w efektywności metod

K-means, mimo dużej liczby anomalii, może nie być najefektywniejszą metodą do wykrywania nieprawidłowości w tego typu danych, ze względu na swoją tendencję do dzielenia danych na zbyt małe grupy. W przeciwieństwie do tego, Hierarchical Clustering wykazał się większą precyzją, wykrywając jedynie kluczowe anomalie, które były bardziej reprezentatywne dla rozkładu danych. Sub-region Clustering, z uwagi na analizę danych w kontekście podregionów, okazał się najbardziej podatny na wykrywanie większej liczby anomalii, co może wynikać z dużej zmienności w obrębie poszczególnych regionów.

7.4. Znaczenie wyników w kontekście dalszej analizy

Wykryte anomalie w różnych metodach mogą stanowić punkt wyjścia do dalszej analizy, w tym identyfikacji przyczyn tych nieprawidłowości oraz potencjalnych działań naprawczych. W zależności od kontekstu, można uznać, że metoda Hierarchical Clustering dostarcza najbardziej wartościowych wyników, gdyż identyfikuje mniej, ale bardziej krytyczne anomalie, które mogą mieć większe znaczenie dla struktury badanych danych. Dalsze badania mogą obejmować testowanie innych metod wykrywania anomalii oraz zastosowanie bardziej zaawansowanych technik, takich jak analiza outlierów w kontekście różnych zmiennych, aby lepiej zrozumieć, dlaczego te anomalie występują.

7.5. Zastosowanie wyników w praktyce

Wyniki wykrywania anomalii mają potencjał do zastosowań w różnych dziedzinach, takich jak monitoring sieci miejskich, analiza bezpieczeństwa, a także w obszarze optymalizacji zarządzania zasobami. Zidentyfikowanie nieprawidłowości może pomóc w podejmowaniu szybszych decyzji operacyjnych oraz w opracowywaniu strategii naprawczych.

References

Benk, F., Rainer, M., 2019. Ai in smart cities: A survey, in: 4th International Conference on Computational Intelligence and Communication Networks, pp. 24–29. URL: https://ieeexplore.ieee.org/document/8900707.

Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J., 2000. Lof: Identifying density-based local outliers, in: Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data, pp. 93–104. URL: https://dl.acm.org/doi/10.1145/342009.335372.

Ester, M., Kriegel, H.P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise, in: Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp. 226–231. URL: https://dl.acm.org/doi/10.1145/231920.231973.

Jain, A.K., 2010. Data clustering: 50 years beyond k-means. Pattern Recognition Letters 31, 651–666. URL: https://www.sciencedirect.com/science/article/pii/S016786551000027X.

Lloyd, S., 1982. Least squares quantization in pcm. IEEE Transactions on Information Theory 28, 129–137. URL: https://ieeexplore.ieee.org/document/1056489, doi:10.1109/TIT.1982.1056489.

Mossberger, K., Tolbert, C.M., McNeal, R.S., 2013. Digital citizenship: The internet, society, and participation. MIT Press. URL: https://mitpress.mit.edu/books/digital-citizenship.

Roth, M., Hirt, C., 2015. Urban Networks in a Globalizing World: Interactions and Developments. Routledge. URL: https://www.routledge.com/Urban-Networks-in-a-Globalizing-World-Interactions-and-Developments/Roth-Hirt/p/book/9781138807177.

Wasserman, S., Faust, K., 2020. Social Network Analysis: Methods and Applications. Cambridge University Press. URL: https://www.cambridge.org/core/books/social-network-analysis/DA4A7C5C2D0FF1B35A9A8F1F4A2C32B2.