Wrocław, 06.12.2024

Projekt

Metody analizy i eksploracji danych

Ekstrakcja wiedzy z danych

„Badanie mikroklimatów na podstawie danych meteorologicznych i topograficznych”

Autorzy: Bartosz Palmer 260346

Nataniel Jargiło 252888

Termin zajęć: Wtorek 17:05

Prowadzący: dr inż. Agata Migalska

1. Wstęp

W niniejszym sprawozdaniu przeprowadzono analizę danych meteorologicznych oraz topograficznych dla stacji synoptycznych IMGW w Polsce, mającą na celu identyfikację mikroklimatów oraz anomalii klimatycznych.

Wykorzystano metody klasteryzacji, takie jak K-Means, DBSCAN, oraz hierarchiczną, zarówno dla pełnego zestawu danych, jak i dla wybranych miesięcy reprezentujących pory roku.

Dodatkowo zidentyfikowano stacje regularnie występujące w tych samych klastrach oraz te należące do klastrów odstających, co pozwoliło na wyciągnięcie wniosków dotyczących wpływu topografii na lokalne warunki klimatyczne.

Przeprowadzono również identyfikację anomalii, czyli stacji które były umieszczone w klastrach odstających (o małej ilości elementów i odległych od głównej chmury punktów).

2. Klasteryzacja na pełnym zestawie danych

W ramach analizy wykorzystano kilka metod klasteryzacji, aby wybrać tę, która najlepiej odpowiada specyfice problemu i pozwala na najbardziej czytelne grupowanie stacji meteorologicznych. Do klasteryzacji w tym etapie wykorzystano dane obejmujące każdą stację meteorologiczną z osobna dla każdego miesiąca.

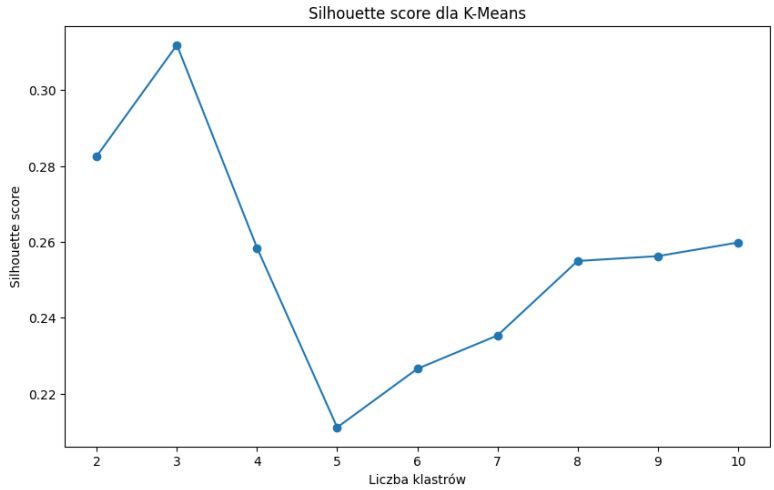
Przed przystąpieniem do procesu klasteryzacji, dane zostały przekształcone za pomocą metody PCA. Celem tego kroku było zredukowanie wymiarowości danych i wyznaczenie dwóch głównych składowych, które najlepiej wyjaśniają zmienność w zestawie danych. Dzięki temu możliwe było lepsze zobrazowanie wyników klasteryzacji na wykresach dwuwymiarowych.

Dla zredukowanych danych utworzono wizualizacje wyników dla każdej zastosowanej metody klasteryzacji. Wykorzystano zarówno algorytm K-Means, który grupuje dane poprzez minimalizowanie różnic wewnątrz klastrów, jak i algorytm DBSCAN, który pozwala na identyfikację klastrów o różnej gęstości oraz wykrywanie punktów odstających. Utworzone wykresy dla każdej z tych metod pozwoliły na graficzne przedstawienie wyników oraz analizę ich jakości.

2.1. K-Means

Przed przystąpieniem do klasteryzacji metodą K-Means przeprowadzono analizę mającą na celu określenie optymalnej liczby klastrów. W tym celu wykorzystano wykres zależności współczynnika Silhouette od liczby klastrów. Współczynnik Silhouette mierzy, jak dobrze dane punkty zostały przypisane do swoich klastrów, porównując średnią odległość punktu do punktów w tym samym klastrze z odległością do punktów w najbliższym sąsiednim klastrze.

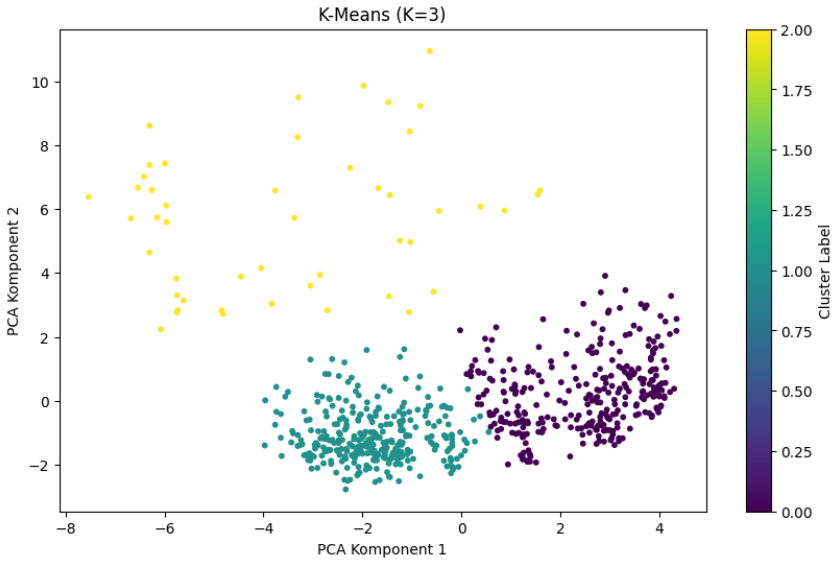
Na wykresie przedstawionym na Rysunku 1. można zauważyć, że optymalna liczba klastrów wynosi 3. Wartość współczynnika Silhouette dla tej liczby klastrów osiąga lokalne maksimum, co oznacza, że punkty są najlepiej skupione wewnątrz klastrów, a jednocześnie dobrze oddzielone od innych klastrów.



Rys. 1. Wykres Silhouette score od ilości klastrów dla metody K-Means

Na podstawie wyznaczonej optymalnej liczby klastrów (K=3) przeprowadzono klasteryzację metodą K-Means. Na rysunku 2. można zaobserwować, że stacje meteorologiczne dla poszczególnych miesięcy zostały podzielone na trzy wyraźne grupy punktów, które nie nachodzą na siebie. Taki podział wskazuje na istnienie charakterystycznych cech meteorologicznych dla każdej z grup.

Dla uzyskanych wyników klasteryzacji wartość współczynnika Silhouette wyniosła 0.32.



Rys. 2. Klasteryzacja K-Means

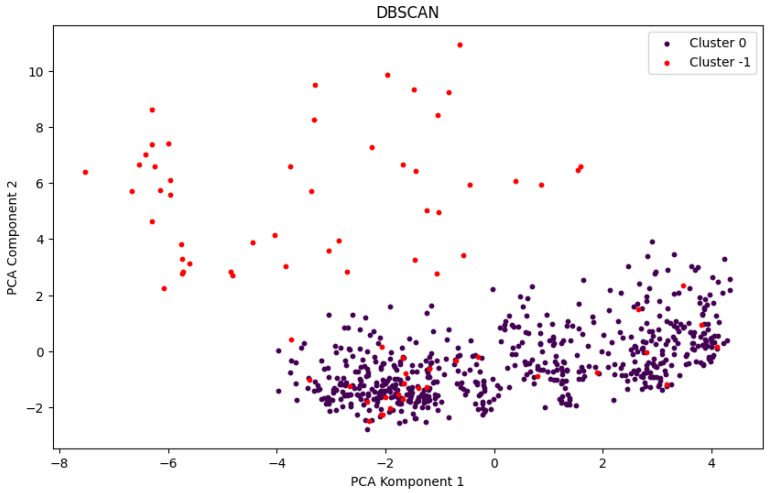
2.2. DBSCAN

W ramach analizy przeprowadzono klasteryzację za pomocą metody DBSCAN, wykorzystując automatyczny dobór optymalnych parametrów. Wartość parametru epsilon była przeszukiwana w zakresie [0.5, 4] z krokiem co 0.1, natomiast parametr MinPts przyjmował wartości ze zbioru {3, 5, 7, 10}. Dla każdego zestawu parametrów obliczano wartość Silhouette score, aby wyznaczyć kombinację zapewniającą najlepsze dopasowanie punktów do klastrów.

Optymalne parametry wyznaczone w tym procesie to:

- Epsilon = 3.8
- MinPts = 10

Na rysunku 3. przedstawiono uzyskane wyniki klasteryzacji dla tych parametrów. Wyraźnie widoczne są zidentyfikowane klastry oraz punkty uznane za szum (oznaczone kolorem czerwonym). Wartość współczynnika Silhouette score dla tej klasteryzacji wyniosła 0.41.

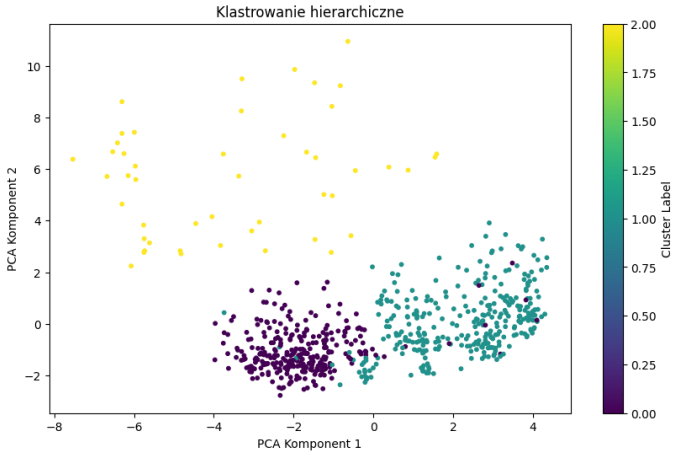


Rys. 3. Klasteryzacja DBSCAN

2.3. Klasteryzacja hierarchiczna

W celu przeprowadzenia klasteryzacji hierarchicznej zastosowano metodę aglomeracyjną, polegającą na iteracyjnym łączeniu punktów w grupy w zależności od ich podobieństwa. Proces ten pozwala na tworzenie klastrów w sposób hierarchiczny.

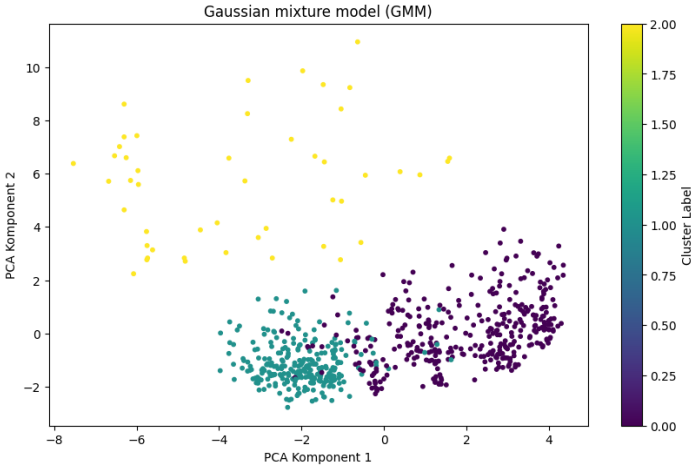
Na rysunku 4. przedstawiono wyniki klasteryzacji hierarchicznej. Widoczne są utworzone klastry, które grupują stacje meteorologiczne według ich podobieństw w zestawach danych. Dla tej klasteryzacji wartość Silhouette score wyniosła 0.28.



Rys. 4. Klasteryzacja hierarchiczna

2.4. GMM

Podobnie jak w przypadku klasteryzacji hierarchicznej, wykorzystano metodę GMM w celach porównawczych. Na rysunku 5 przedstawiono wyniki klasyfikacji uzyskane przy użyciu tej metody. Dla klasteryzacji GMM wartość Silhouette score wyniosła 0.27.



Rys. 5. Klasteryzacja GMM

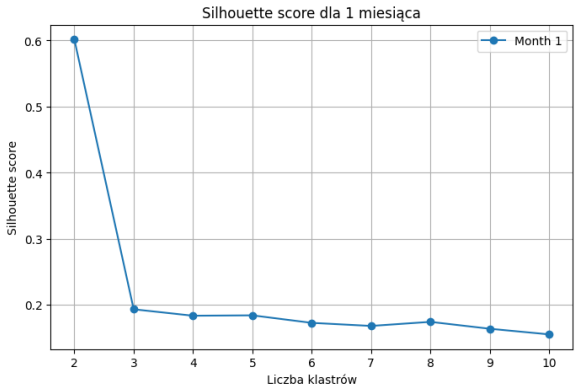
3. Klasteryzacja dla wybranych miesięcy

W ramach analizy sezonowej przeprowadzono klasteryzację metodą K-means dla czterech wybranych miesięcy: stycznia, kwietnia, lipca oraz listopada, reprezentujących odpowiednio zimę, wiosnę, lato i jesień. Dla każdego z tych miesięcy oddzielnie wyznaczono optymalną liczbę klastrów za pomocą analizy współczynnika Silhouette, a następnie przeprowadzono klasteryzację.

Proces ten pozwolił na zbadanie różnic w strukturze grupowania stacji meteorologicznych w zależności od warunków sezonowych. Otrzymane wyniki wskazują, że liczba optymalnych klastrów różni się w zależności od miesiąca, co odzwierciedla zmienność charakterystyk meteorologicznych w ciągu roku.

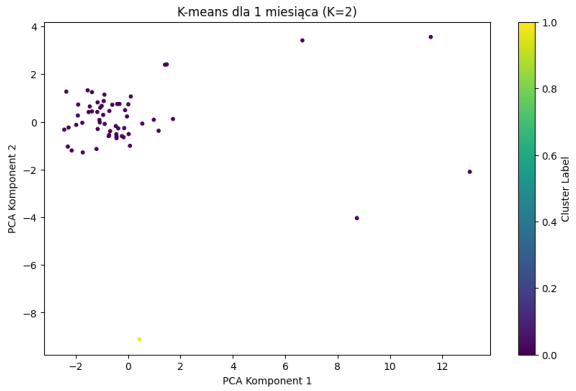
- Miesiąc = 1

Na rysunku 6 widać, że optymalna liczba klastrów dla miesiąca stycznia wynosi 2.



Rys. 6. Wykres Silhouette score od ilości klastrów dla metody K-Means - styczeń

Na rysunku 7. przedstawiono wyniki klasteryzacji dla miesiąca stycznia, gdzie większość obserwacji została przypisana do klastra 0, tworząc zwartą chmurę punktów. Jedna obserwacja została przypisana do klastra 1, a kilka wartości w klastrze 0 wykazuje cechy odstające.



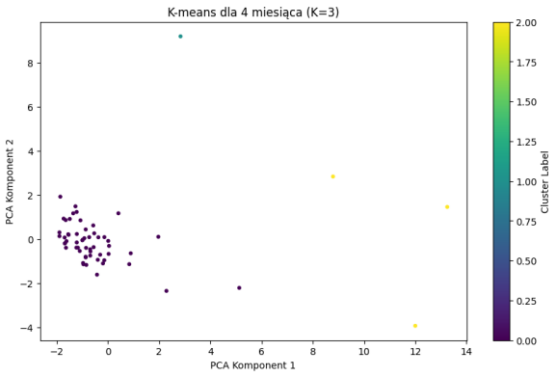
Rys. 7. Klasteryzacja K-Means - styczeń

- Miesiąc = 4

Na rysunku 8 widać, że optymalna liczba klastrów dla miesiąca kwietnia wynosi 3.

Rys. 8. Wykres Silhouette score od ilości klastrów dla metody K-Means - kwiecień

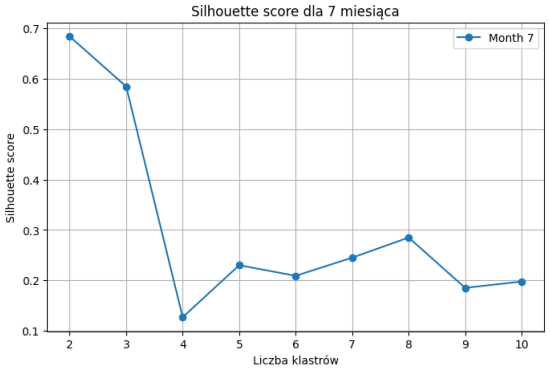
Na rysunku 9. widoczne są wyniki klasteryzacji dla miesiąca kwietnia, gdzie większość danych znajduje się w klastrze 0, tworząc wyraźną chmurę punktów. Zaobserwowano 3 wartości odstające przypisane do klastra 2 oraz jedną wartość odstającą należącą do klastra 1.



Rys. 9. Klasteryzacja K-Means - kwiecień

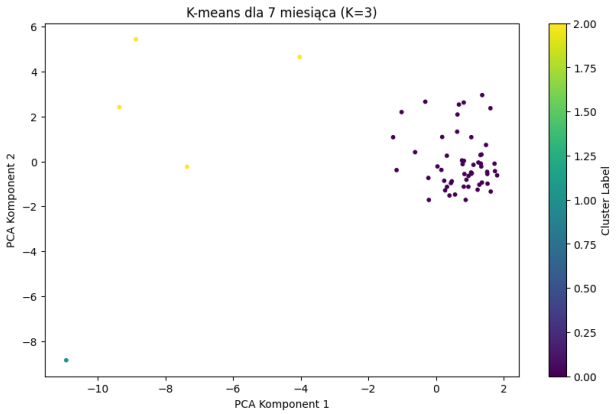
- Miesiąc = 7

Na rysunku 10 możemy zauważyć, że optymalna liczba klastrów dla miesiąca lipca wynosi 2 lub 3. W tym przypadku do klasteryzacji dla lipca wykorzystano K=3.



Rys. 10. Wykres Silhouette score od ilości klastrów dla metody K-Means - lipiec

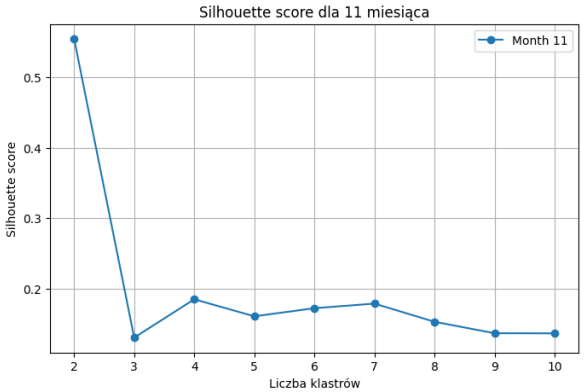
Na rysunku 11. zaprezentowano wyniki klasteryzacji dla miesiąca lipca, z dominującą chmurą punktów klastra 0. Dodatkowo zaobserwowano 4 wartości odstające przypisane do klastra 2 oraz jedną wartość odstającą przypisaną do klastra 1.



Rys. 11. Klasteryzacja K-Means - lipiec

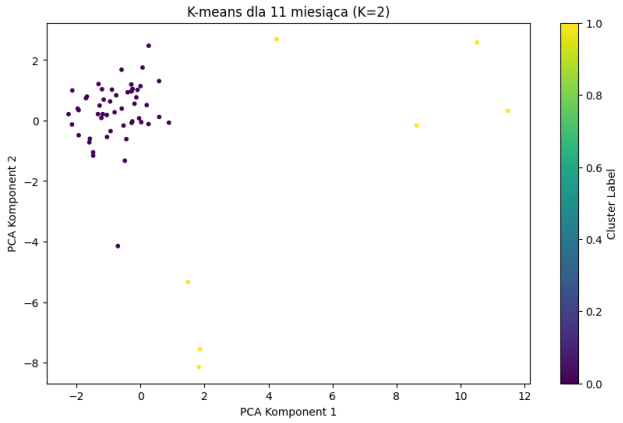
- Miesiąc = 11

Na rysunku 12 zauważamy, że optymalna liczba klastrów dla miesiąca listopada wynosi 2.



Rys. 12. Wykres Silhouette score od ilości klastrów dla metody K-Means - listopad

Na rysunku 13. przedstawiono wyniki klasteryzacji dla miesiąca listopada, gdzie główną grupę tworzy chmura punktów klastra 0. Zidentyfikowano także 7 wartości odstających przypisanych do klastra 2.



Rys. 13. Klasteryzacja K-Means – listopad

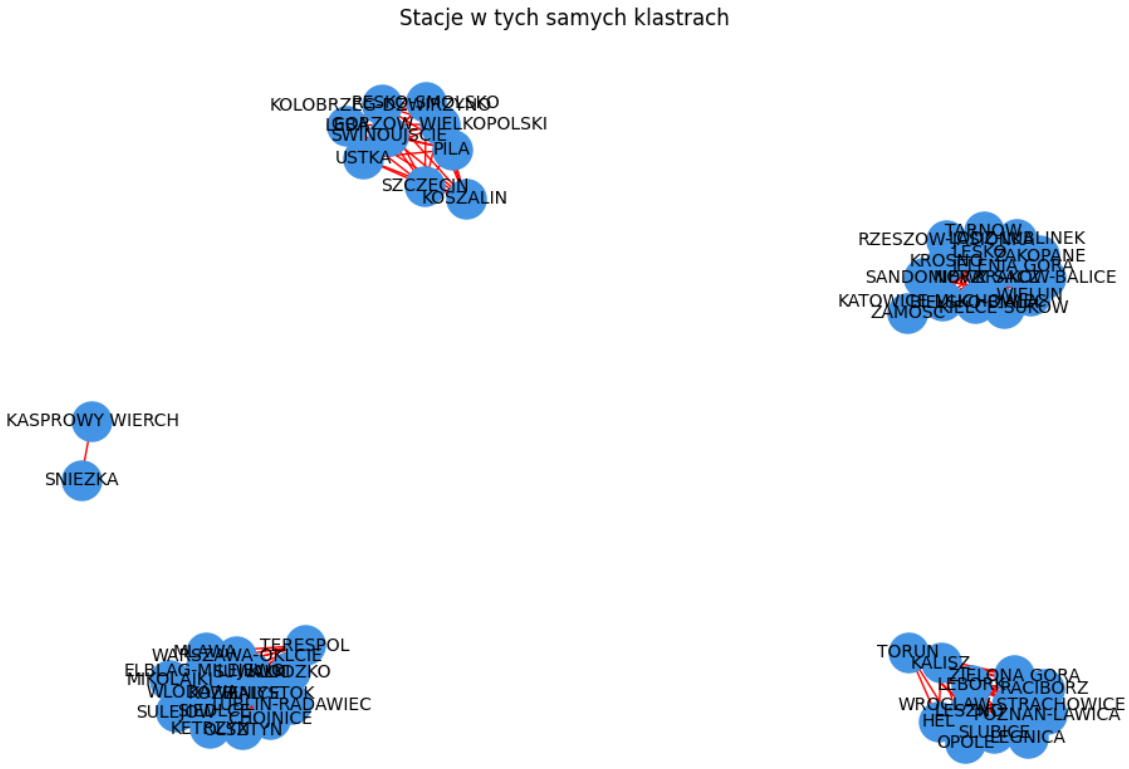
3.1. Części wspólne klastrów

Przeprowadzono analizę mającą na celu identyfikację stacji meteorologicznych, które regularnie występowały w tych samych klastrach w różnych miesiącach. Najpierw dla każdego miesiąca wyznaczono pary stacji należących do tych samych klastrów, zapisując je w postaci zbiorów. Następnie wyznaczono przecięcie tych zbiorów, aby określić pary stacji, które były w tych samych klastrach we wszystkich analizowanych miesiącach.

W kolejnym kroku zidentyfikowano unikalne stacje wchodzące w skład tych par, a następnie powiązano je z ich nazwami. W wyniku tego uzyskano zestawienie stacji meteorologicznych, które wykazywały regularność w przynależności do tych samych klastrów w analizowanych miesiącach.

- Zbiór 1 (Północna Polska): Świnoujście, Piła, Szczecin, Ustka, Resko-Smolsko, Kołobrzeg-Dźwirzyno, Gorzów Wielkopolski, Koszalin, Łeba
- Zbiór 2 (Północno-centralna Polska): Olsztyn, Elbląg-Milejewo, Mikołajki, Siedlce, Kętrzyn, Kozienice, Chojnice, Mława, Lublin-Radawiec, Warszawa-Okęcie, Białystok, Suwałki, Kłodzko, Terespol, Włodawa, Sulejów
- Zbiór 3 (Południowo-centralna Polska): Lębork, Poznań-Ławica, Zielona Góra, Wrocław-Strachowice, Hel, Toruń, Racibórz, Opole, Leszno, Legnica, Słubice, Kalisz
- Zbiór 4 (Południowa Polska): Katowice-Muchowiec, Krosno, Łódź-Lublinek, Wieluń, Nowy Sącz, Kielce-Suków, Tarnów, Sandomierz, Lesko, Jelenia Góra, Rzeszów-Jasionka, Zakopane, Kraków-Balice, Zamość, Bielsko-Biała
- Zbiór 5 (Stacje wysokogórskie): Kasprowy Wierch, Śnieżka

Dodatkowo zbiory te zostały zwizualizowane na grafie przedstawionym na rysunku 14. Węzły grafu reprezentują poszczególne stacje, a krawędzie wskazują powiązania między stacjami w ramach wspólnych klastrów.



Rys. 14. Graf zbiorów stacji meteorologicznych

4. Wykrywanie anomalii

W ramach wykrywania anomalii przeprowadzono identyfikację stacji meteorologicznych, które zostały przypisane do klastrów będących odstającymi w wybranych miesiącach: styczeń, kwiecień, lipiec i listopad. Klastry odstające zdefiniowano jako te inne niż klaster główny (klaster 0). W każdym miesiącu zidentyfikowano stacje meteorologiczne, które wykazywały nietypowe zachowanie w porównaniu do pozostałych stacji. Analiza anomalii pozwala na zidentyfikowanie obszarów, w których specyficzne warunki lokalne mogą znacząco różnić się od ogólnych trendów.

- Styczeń

W klastrze odstającym (klaster 1) znalazła się stacja KOŁO, co wskazuje na jej nietypowy charakter w tym miesiącu w porównaniu z pozostałymi stacjami.

- Kwiecień

W klastrze odstającym zidentyfikowano dwie grupy stacji:

  - Klaster 2: KASPROWY WIERCH, HALA GĄSIENICOWA, ŚNIEŻKA. Stacje te są zlokalizowane w wysokich partiach wysokogórskich, co może wyjaśniać ich odmienny charakter.
  - Klaster 1: KOŁO, które również w tym miesiącu zostało uznane za odstającą stację.

- Lipiec
    - Klaster odstający (klaster 2) obejmował stacje górskie: ZAKOPANE, KASPROWY WIERCH, HALA GĄSIENICOWA i ŚNIEŻKA, co jest spójne z obserwacjami z kwietnia.
    - W klastrze 1 ponownie znalazła się stacja KOŁO, która po raz kolejny wykazała nietypowe zachowanie.
- Listopad

Klaster odstający (klaster 1) był bardziej rozbudowany i obejmował stacje:

- Górskie: ZAKOPANE, KASPROWY WIERCH, HALA GĄSIENICOWA i ŚNIEŻKA.
- Równinne: CZĘSTOCHOWA, PŁOCK oraz GDAŃSK-ŚWIBNO. Ich obecność w klastrze odstającym może sugerować nietypowe warunki lokalne w tych obszarach w listopadzie.

5. Wnioski

Analiza klasteryzacji pozwoliła na wyróżnienie grup stacji meteorologicznych charakteryzujących się podobnymi warunkami klimatycznymi w różnych porach roku. Regularność przynależności stacji do tych samych klastrów wskazuje na istotny wpływ lokalnej topografii i położenia geograficznego na mikroklimaty. Przykładowo, stacje w terenach górskich, jak Kasprowy Wierch i Śnieżka, wyróżniały się na tle innych stacji.

Wyniki klasteryzacji dla wybranych miesięcy (styczeń, kwiecień, lipiec, listopad) ukazały sezonowe różnice w warunkach klimatycznych. Stacje takie jak Koło, Zakopane, Kasprowy Wierch, Hala Gąsienicowa i Śnieżka regularnie występowały w klastrach odstających, co może wskazywać na unikalne warunki klimatyczne związane z ich położeniem geograficznym lub wysokością. Szczególnie Koło wyróżniało się jako stacja występująca w klastrach odstających w prawie każdym z analizowanych miesięcy, co może sugerować występowanie mikroklimatu w tym regionie dodatkowo biorąc pod uwagę położenie tego regionu w centralnej Polsce, co może nie być intuicyjnym położeniem na występowanie mikroklimatu.

Zidentyfikowane grupy stacji w stałych klastrach wskazują na wyraźny podział geograficzny – regiony nadmorskie, górskie, nizinne oraz środkową Polskę. Wyniki te potwierdzają, że topografia terenu znacząco wpływa na kształtowanie się lokalnych warunków meteorologicznych.