KLASTERYZACJA

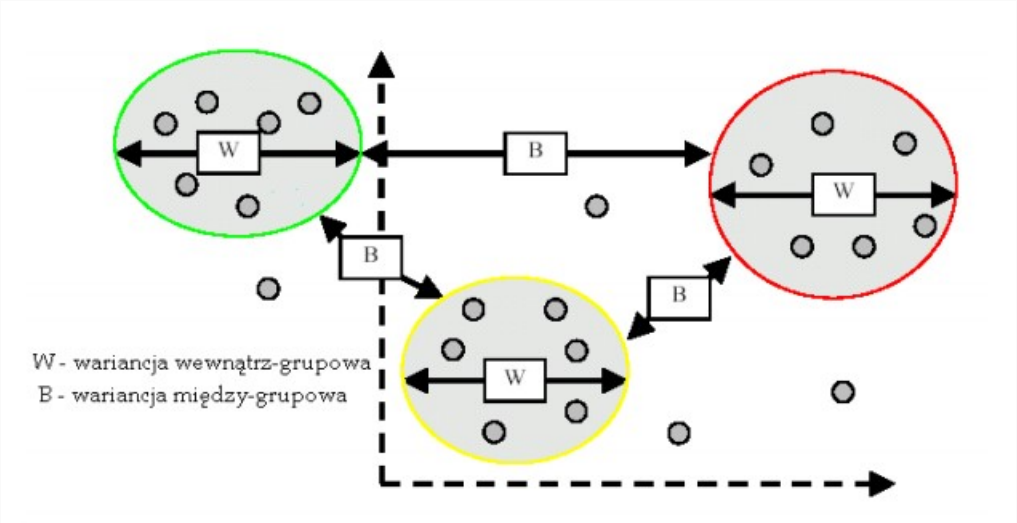
Paweł Buglewicz, **Krzysztof Cieśla**, **Justyna Olczak**

20.12.2016

Politechnika Krakowska

Klasteryzacja

Klasteryzacja (grupowanie) jest jedną z metod nienadzorowanej (bez dostępnej a priori wiedzy) analizy danych. Głównym celem klasteryzacji jest podział rozpatrywanego zbioru obiektów na grupy (klastry), w ten sposób, aby każda z grup była możliwie jednorodna (tzn. zawierała elementy podobne do siebie), a jednocześnie poszczególne klastry były jak najbardziej zróżnicowane między sobą (rys. 1).

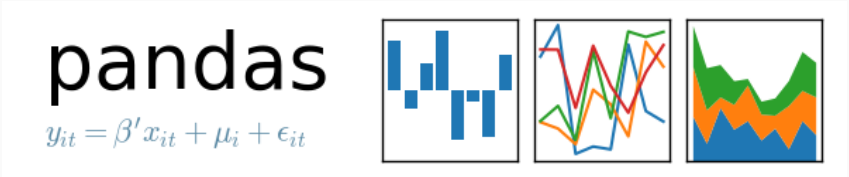


Rys. 1: Idea klasteryzacji. Dążymy do takiego podziału zbioru danych aby wariancja wewnątrz-grupowa (w każdym z klastrów była możliwie mała) a jednocześnie wariancja między-grupowa możliwie duża.

Co nam daje klasteryzacja?

Klasteryzacji używa się w celu zredukowania złożoności wielowymiarowego problemu. Pozwala ona zauważyć podział, który mógłby być niewidoczny dla badacza bez dostępu do komputera (który sprawia, że czas znalezienia podziału jest znacznie skrócony). Gdy nastąpił już podział na grupy można opracować wiele prostych modeli dla każdej z osobna, zamiast jednego, bardzo złożonego modelu.

Wykorzystane oprogramowanie



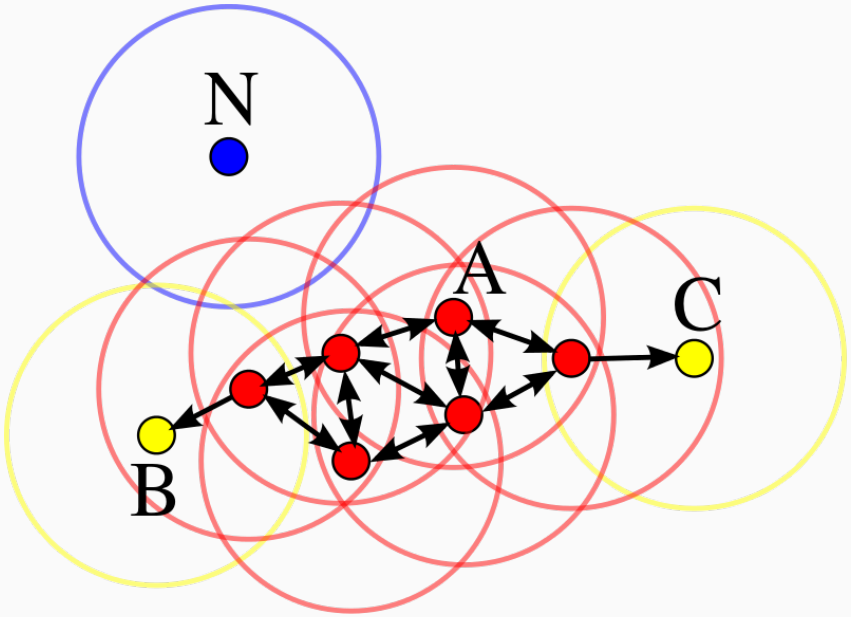
Metody klasteryzacji - przykłady

Method name	Parameters	Scalability	Usecase	Geometry (metric used)
K-Means	number of clusters	Very large `n samples`, medium `n_clusters` with MiniBatch code	General-purpose, even cluster size, flat geometry, not too many clusters	Distances between points
Affinity propagation	damping, sample preference	Not scalable with n_samples	Many clusters, uneven cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Mean-shift	bandwidth	Not scalable with `n_samples`	Many clusters, uneven cluster size, non-flat geometry	Distances between points
Spectral clustering	number of clusters	Medium `n samples`, small `n_clusters`	Few clusters, even cluster size, non-flat geometry	Graph distance (e.g. nearest-neighbor graph)
Ward hierarchical clustering	number of clusters	Large `n samples` and `n_clusters`	Many clusters, possibly connectivity constraints	Distances between points
Agglomerative clustering	number of clusters, linkage type, distance	Large `n samples` and `n_clusters`	Many clusters, possibly connectivity constraints, non Euclidean distances	Any pairwise distance
DBSCAN	neighborhood size	Very large `n samples`, medium `n_clusters`	Non-flat geometry, uneven cluster sizes	Distances between nearest points
Gaussian mixtures	many	Not scalable	Flat geometry, good for density estimation	Mahalanobis distances to centers
Birch	branching factor, threshold, optional global clusterer.	Large `n clusters` and `n_samples`	Large dataset, outlier removal, data reduction.	Euclidean distance between points

Rys. 2: Porównanie algorytmów grupowania w scikit-learn.

DBSCAN

- **D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise - Martin Ester, Hans-Peter Kriegel, Jörg Sander i Xiaowei Xu (1996).
- DBSCAN jest jedną z najprostszych, najpowszechniejszych oraz najszybszych metod klaseryzacji. Polega ona na szukaniu klastrów jako obszarów o zwiększonej gęstości, oddzielonch obszarami o mniejszej gęstości.



DBSCAN

Zalety:

- Jako parametru początkowego nie wymaga liczby spodziewanych klastrów w danych.
- Wymaga tylko dwóch parametrów początkowych:
  - Minimalna liczba punktów w sąsiedźtwie
  - Maksymalny promień sąsiedźtwa
- Potrafi wyszukać klastry nawet o bardzo skomplikowanym kształcie.

Wady:

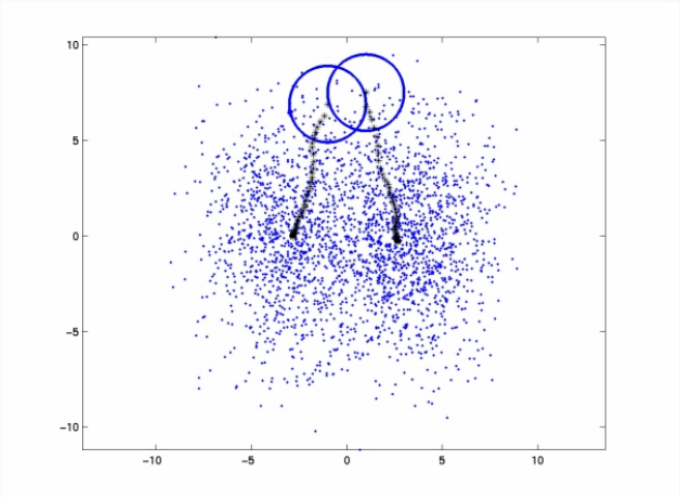
- Nieprzydadny dla wielowymiarowych danych (pomiar odległości między punktami).
- Nie działa jeśli różnice w gęstości są zbyt duże.

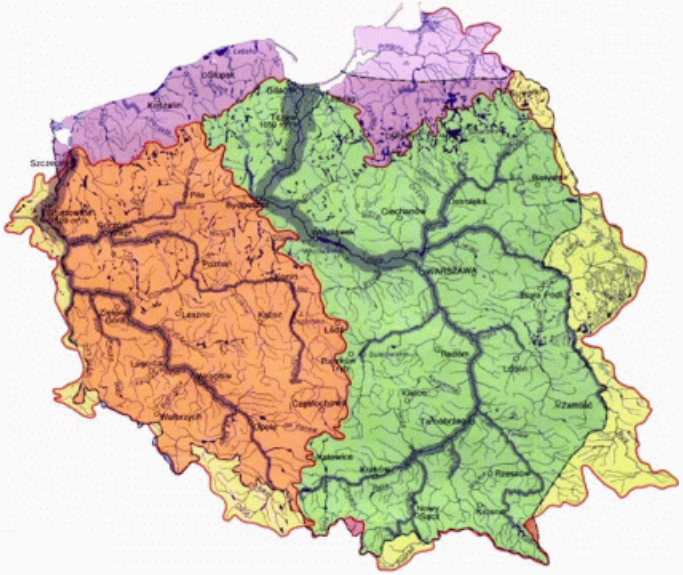
Mean Shift

Algorytm poszukuje klastrów, których centra znajdują się w atraktorach danego układu. Atraktorem w tym wypadku nazywamy punkt, do którego docieramy, podążając od każdego punktu w kierunku, w którym gradient jest największy. Ten krok nazywamy *Średnim przesunięciem - mean shift*.

Dla danych dyskretnych musi być stosowany odpowiedni estymator gradientu.

Dla wzrokowców - jest to analogiczne do zlewisk rzek





Mean Shift

Zalety:

- Można zastosować go do dowolnej liczby wymiarów.
- Kształt klastra może być dowolny.

Wady:

- Czuły na wybór parametrów początkowych.
- Przy dużej liczbie próbek algorytm znacznie zwalnia.

Affinity propagation

Metoda affinity propagation szuka klastrów poprzez wymianę informacji między punktami. Algorytm określa podobieństwo między parami punktów (poprzez to, jaką informację wymieniły), jednocześnie je modyfikując (zmienia się odległość tych punktów w przestrzeni wielowymiarowej). Po wielu krokach, jeśli punkty są odpowiednio podobne, czyli odpowiednio bliskie, stają się jednym klastrem. Osobne klastry nie są podobne do siebie i się ”odpychają”.
Cechy punktów są opisywane za pomocą macierzy – wartości liczbowych.

Affinity propagation

Zalety:

- Jako parametr początkowy nie jest potrzebne zadanie liczby klastrów
- Może być stosowany w dowolnej liczbie wymiarów
- Stabilny.

Wady:

- Wrażliwy na szumy.
- Trudny wybór parametrów.
- Powolny - złożoność $O(N^2)$.

Wykorzystane dane

Dane wykorzystane przez nas do klasteryzacji zebrane zostały poprzez 27 pytaniową ankietę, zawierającą odpowiedzi w postaci liczb całkowitych.

`https://goo.gl/forms/aDAEtAKZTxXPf5j32`

Dodatkowo wykorzystaliśmy podobne dane, które zebrane zostały w Portugalii.

`https://www.kaggle.com/uciml/student-alcohol-consumption`

W naszej „analizie” zwróciliśmy uwagę na współczynniki:

- Homogeneity - h - każdy klaster zawiera tylko dane z jednego zestawu
- Completness - c - wszystkie dane z jednego zestawu znajdują się w tej samej klasie
- V-measure - v - średnia harmoniczna h oraz c

Warunkowa entropia klas przy chwilowym (w danym kroku algorytmu) przydzieleniu do klastrów:

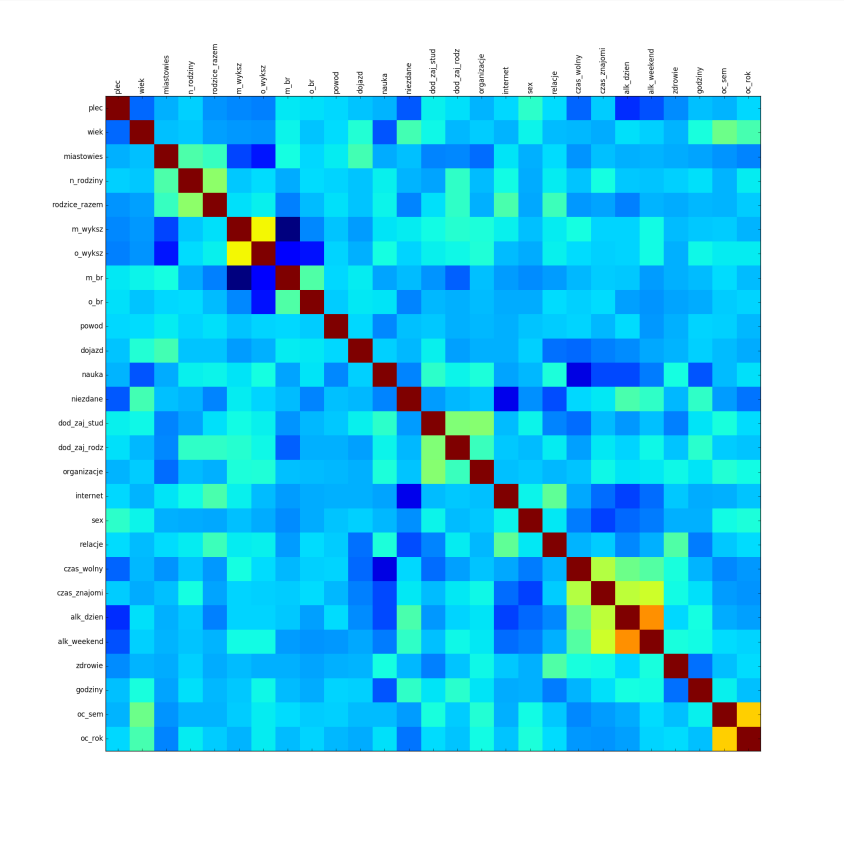
$$H(C|K) = -\sum_{c=1}^{|C|}\sum_{k=1}^{|K|}\frac{n_{c,k}}{n}\cdot log\left(\frac{n_{c,k}}{n_k}\right) \tag{1}$$

Entropia klas:

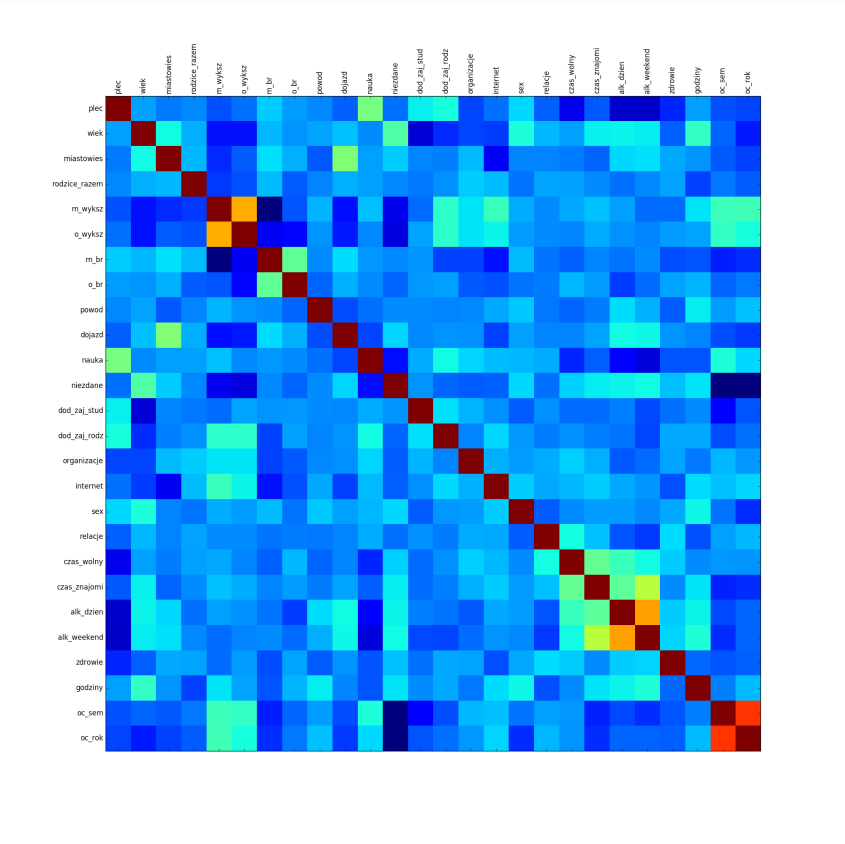
$$H(C) = -\sum_{c=1}^{|C|}\frac{n_c}{n}\cdot log\left(\frac{n_c}{n}\right) \tag{2}$$

$$h = 1 - \frac{H(C|K)}{H(C)} \qquad c = 1 - \frac{H(K|C)}{H(K)} \qquad v = 2\cdot\frac{h\cdot c}{h+c} \tag{3}$$

Wykresy korelacji

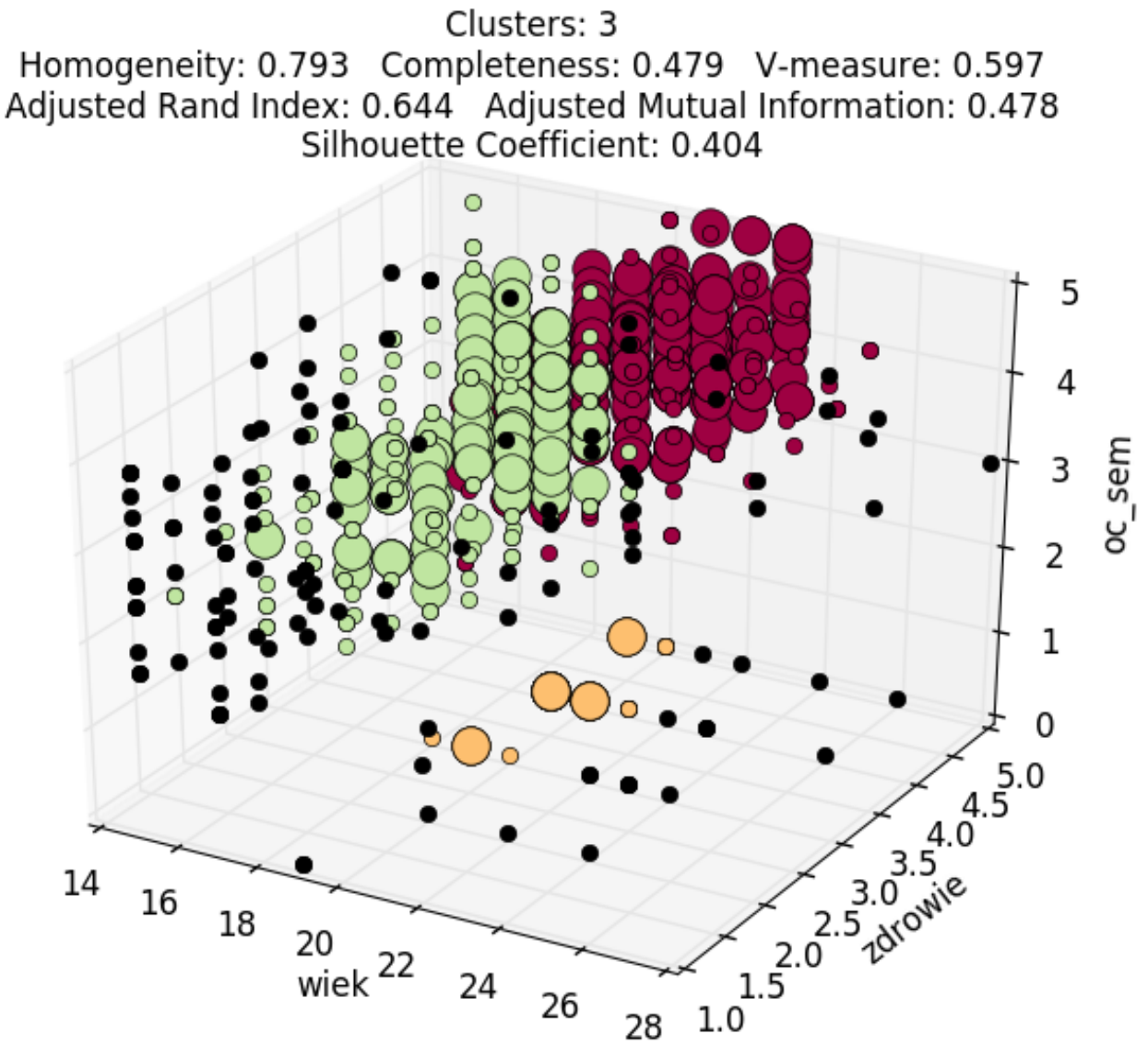


Rys. 3: Wykres korelacji dla naszych danych

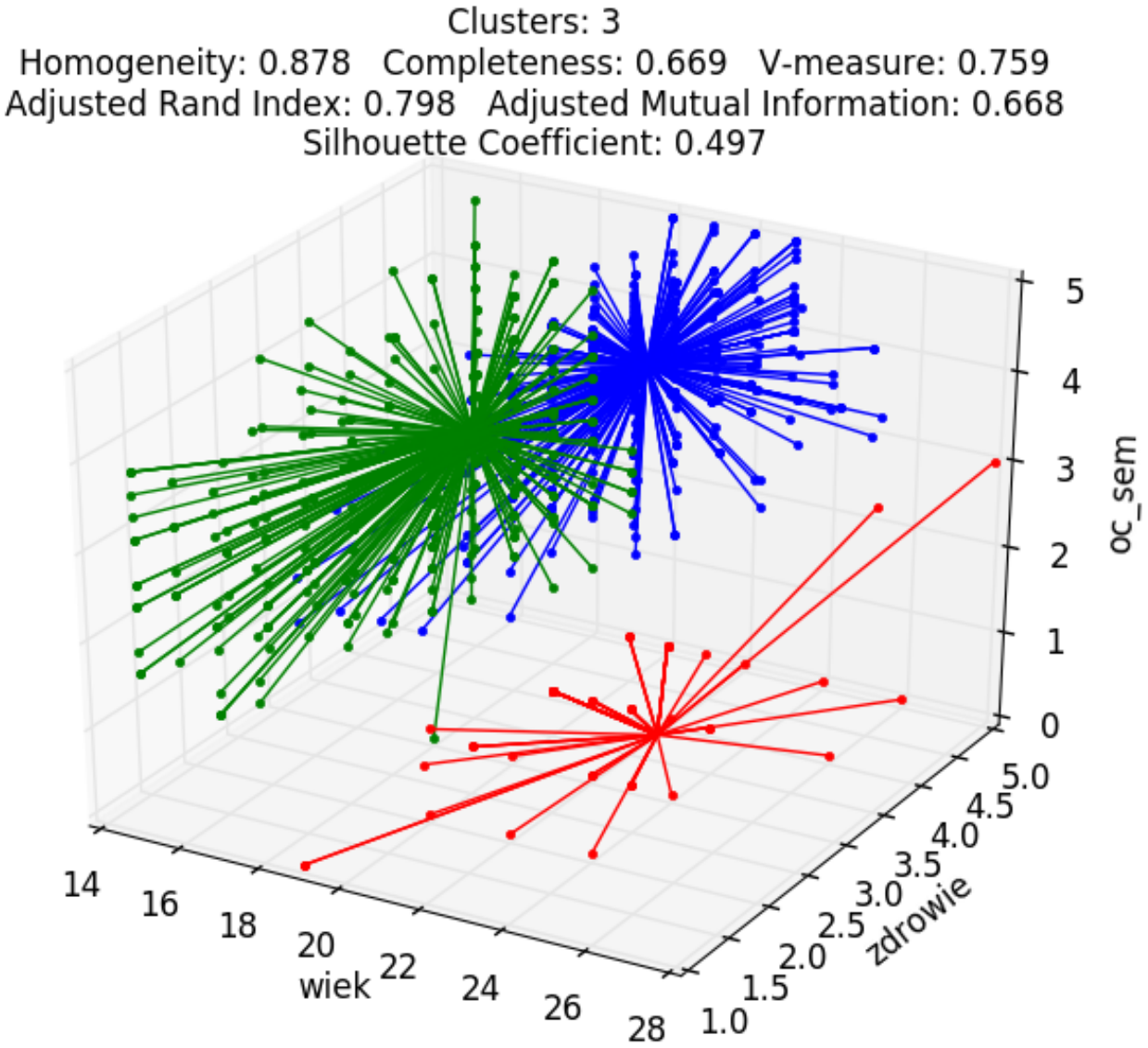


Rys. 4: Wykres korelacji dla danych z Portugalii

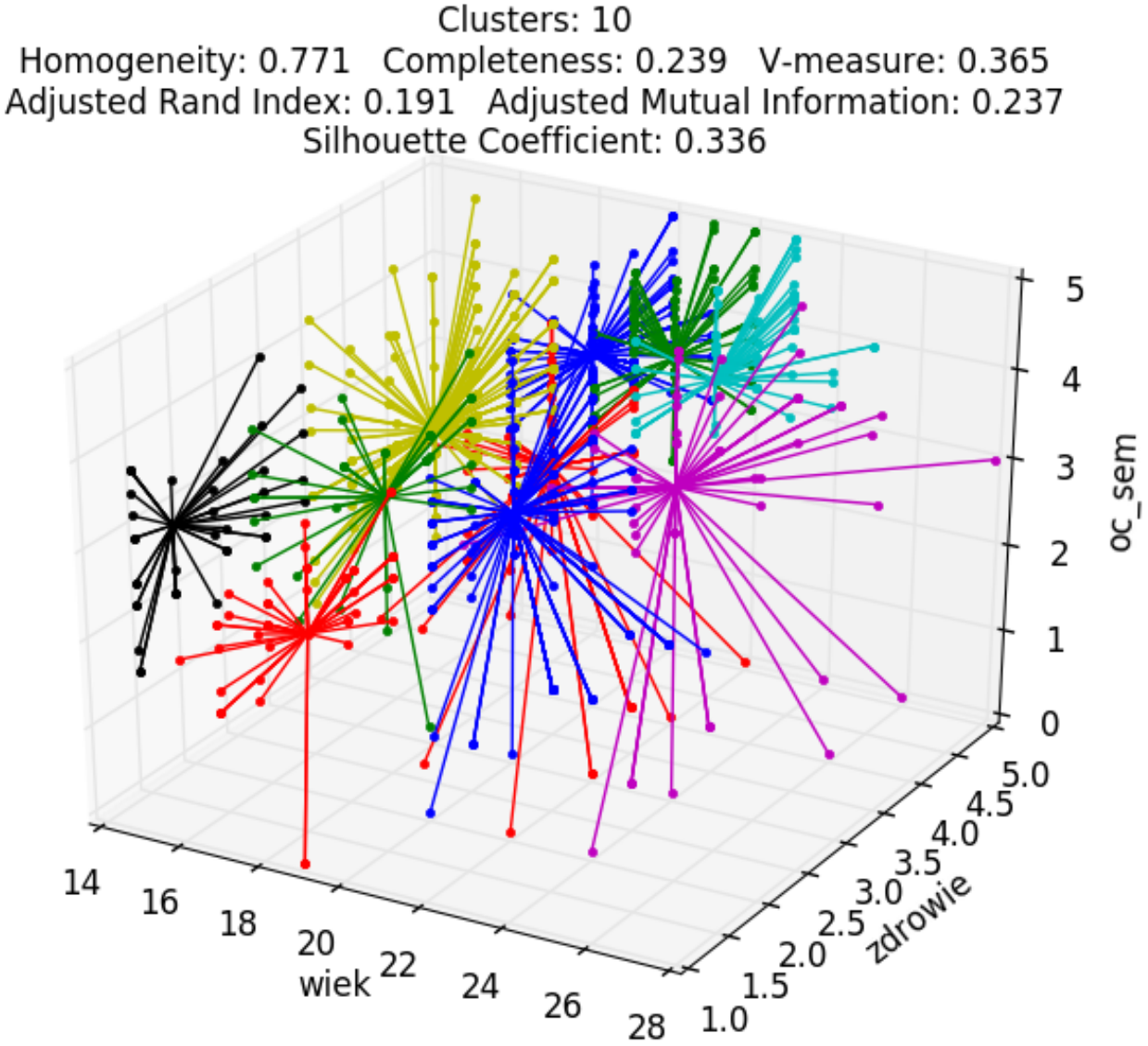
DBSCAN


Clusters: 3
Homogeneity: 0.793 Completeness: 0.479 V-measure: 0.597
Adjusted Rand Index: 0.644 Adjusted Mutual Information: 0.478
Silhouette Coefficient: 0.404
wiek
zdrowie
oc_sem

Mean shift


Clusters: 3
Homogeneity: 0.878 Completeness: 0.669 V-measure: 0.759
Adjusted Rand Index: 0.798 Adjusted Mutual Information: 0.668
Silhouette Coefficient: 0.497
wiek
zdrowie
oc_sem

Affinity propagation



Dziękujemy za uwagę!