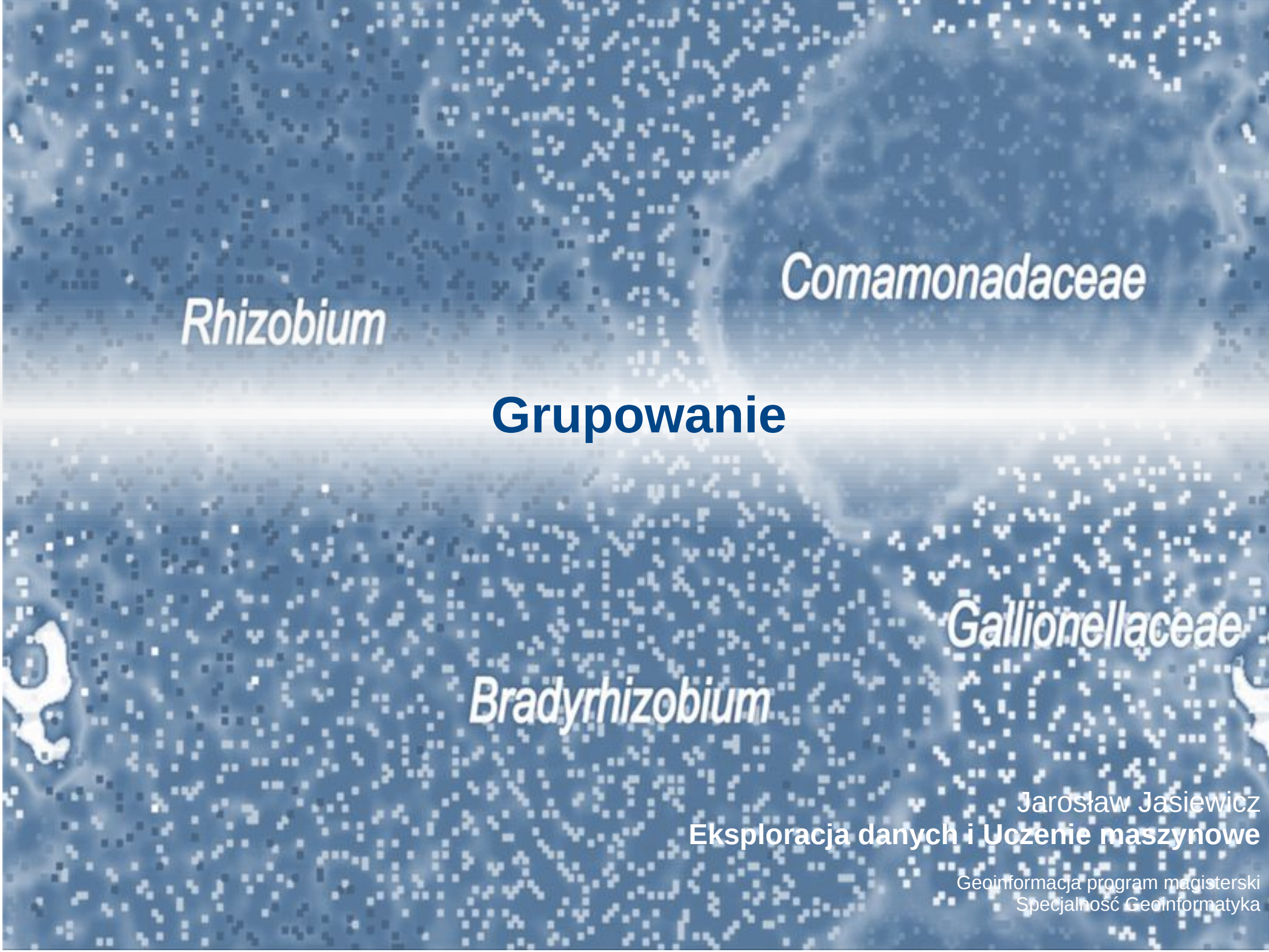


Grupowanie

Jarosław Jasiewicz

Eksploracja danych i Uczenie maszynowe

Geoinformacja program magisterski
Specjalność Geoinformatyka

Dlaczego klasyfikujemy dane

- Klasyfikacja danych ma na celu zredukowanie złożoności danych. Zamiast wielu obiektów, każdy opisany kilku-kilkunastoma parametrami mamy kilka klas, a każdy obiekt jest przypisany do jednej (niekiedy więcej) klas
- **Klasyfikacje nadzorowane** polegają na przypisaniu nowego obiektu do już istniejącego zestawu klas
- **Klasyfikacje nienadzorowane** mają na celu wykrycie w danych ukrytych, nieoczywistych struktur. Krokiem w klasyfikacji danych jest grupowanie lub analiza skupień (*clustering*)
- Problem terminologiczny: w języku angielskim termin *classify data* jest niejednoznaczny. Może oznaczać zarówno proces klasyfikacji jak i **utajniania** danych.

Grupowanie/Analiza skupień

- Analiza skupień to proces przypisywania obiektów do niezdefiniowanych a priori grup na podstawie analizy struktury danych
- Obiekty w skupieniach wykazują tendencję do wzajemnego podobieństwa, a obiekty w różnych skupieniach wykazują tendencję do niepodobieństwa
- Podstawą analizy skupień jest koncepcja niepodobieństwa pomiędzy obiektami

Koncepcja niepodobieństwa

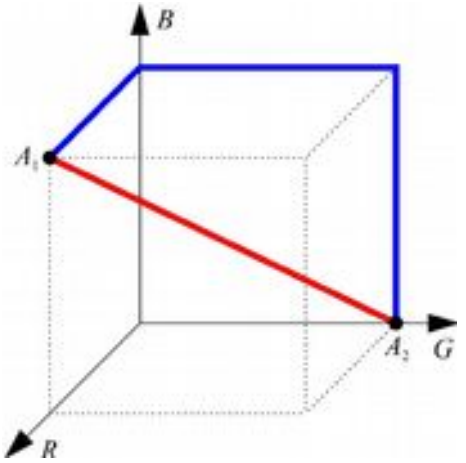
- Koncepcja podobieństwa/niepodobieństwa jest kluczowa dla analizy skupień, tak aby podobne obiekty były klasyfikowane do tych samych skupień, a niepodobne do różnych
- Pojęcie intuicyjnie zrozumiałe, ale trudne do wyrażenia matematycznie
- Niepodobieństwo jest proste to wyrażenia pomiędzy obiektami opisanymi jako punkty w przestrzeni dwu- lub trójwymiarowej poprzez pojęcie **odległości**

Koncepcja odległości

- W przestrzeni fizycznej odległość pomiędzy dwoma punktami to najkrótsza droga pomiędzy dwoma punktami
- Przy założeniu braku przeszkód – odległość liczymy jako najkrótszą możliwą odległość w przestrzeni – generalizowaną jako odległość euklidesową (*euclidean*)
- Przy istnieniu przeszkód – odległość liczymy jako najkrótszą możliwą drogę pomiędzy dwoma punktami – generalizowaną jako odległość miejską (*manhattan*).
- Odległość może być liczona w przestrzeni lub na dowolnej płaszczyźnie (np. na sferze jako tzw wielkie koło)

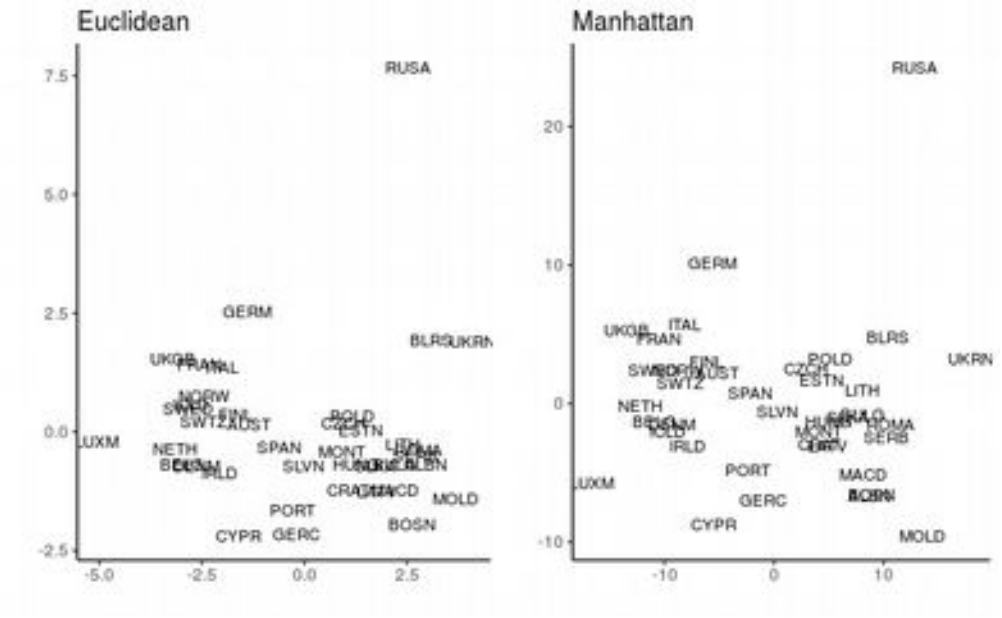
Odległość euklidesowa i miejska

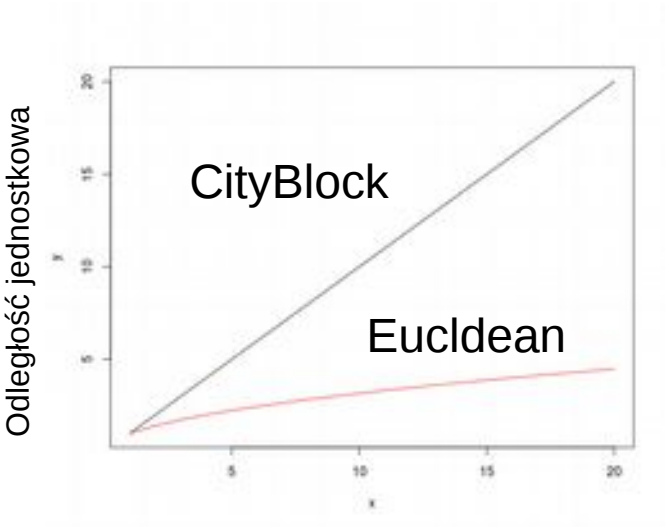
- Odległość euklidesowa pomiędzy dwoma punktami to długość linii łącząca te dwa punkty
  - Dla dwóch wymiarów: $d_{eucl}=\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$
  - Postać ogólna: $d_{eucl}=\sqrt{\sum (x_i-y_i)^2}$
- Odległość miejska pomiędzy dwoma punktami to suma odległości w każdym z wymiarów z osobna
  - Dla dwóch wymiarów: $d_{manh}=|x_2-x_1|+|y_2-y_1|$
  - Postać ogólna $d_{manh}=\sum |x_i-y_i|$



Niepodobieństwo a przekleństwo wymiarowości

- Cechy miary euklidesowej powodują że wraz ze wzrostem wymiarowości wpływ kolejnych wymiarów jest coraz mniejszy, przy dużej liczbie wymiarów należy rozważyć stosowanie odległości miejskiej





Metryka

- Przestrzeń metryczna to przestrzeń w której odległości pomiędzy wszystkimi obiektami są zdefiniowane, Zbór wszystkich odległości zwane są **metryką zbioru**
- Koncepcja metryki jest generalizacją odległości euklidesowa w 3-wymiarowej przestrzeni euklidesowej
- Każda metryka musi spełniać następujące aksjomaty:
  - Nieujemność: $d(a,b) \geq 0$
  - Identyczność: $d(a,b)==0 <=> a==b$
  - Symetria: $d(a,b) == d(b,a)$
  - Nierówność trójkątna $d(a,b) \leq d(a,c)+d(b,c)$

Wektory i Normy

- Definiując obiekt jako zbiór cech opisanych wartościami możemy przedstawić go jako punkt w wielowymiarowej (n-wymiarowej) przestrzeni
- Norma to funkcja, która przypisuje długość do wektora wyznaczonego przez ten punkt (stąd **wektor cech**)
- W przestrzeni n-wymiarowej norma Euklidesowa (L2) to najkrótsza odległość pomiędzy początkiem okładu a punktem:

  Gdzie x_1 ... x_2 kolejne wymiary

$$\|x\|_2 := \sqrt{x_1^2 + x_2^2 + \dots x_n^2}$$

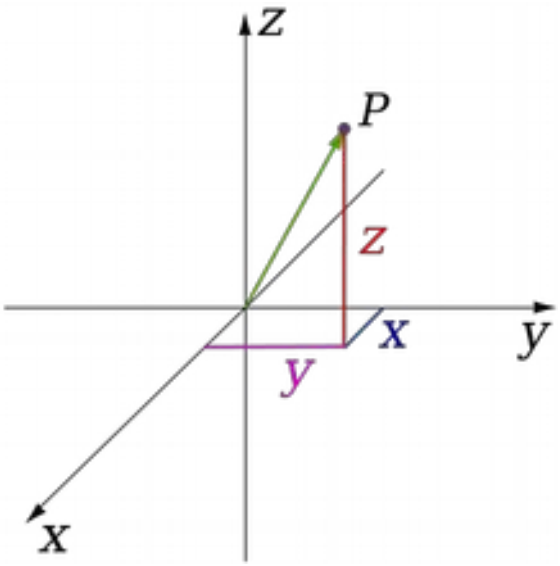
- Norma miejska - Manhattan (L1) suma współrzędnych

$$\|x\|_1 := x_1 + x_2 + \dots x_n$$

- P-norma

$$\|x\|_p := \sum \left(|x_1{}^p| + |x_2{}^p| + \dots |x_n{}^p| \right)^{1/p}$$

$$\|x\|_{inf} := max\left(x_1 + x_2 + \dots x_n\right)$$



Odległość a niepodobieństwo

- Jeżeli w analizie danych obiekty opisane są poprzez ich wektory cech, niepodobieństwo pomiędzy nimi utożsamia się z odległością – metryką euklidesową. W żargonie termin *distance* używa się jako synonimu niepodobieństwa
- Nie wszystkie miary niepodobieństwa dają się wyrazić jako odległość
- W Spatial Data Science pojęcie odległości jest ambiwalentne: odległość pomiędzy obiektami to odległość w przestrzeni geograficznej czy niepodobieństwo?
- Bezpieczne terminy:

Przestrzeń Geograficzna i kartezjańska	Odległość *Distance*	Bliskość *Proximity*
Przestrzeń Informacyjna	Niepodobieństwo *Dissimilarity*	Podobieństwo *Similarity*

Inne miary niepodobieństwa

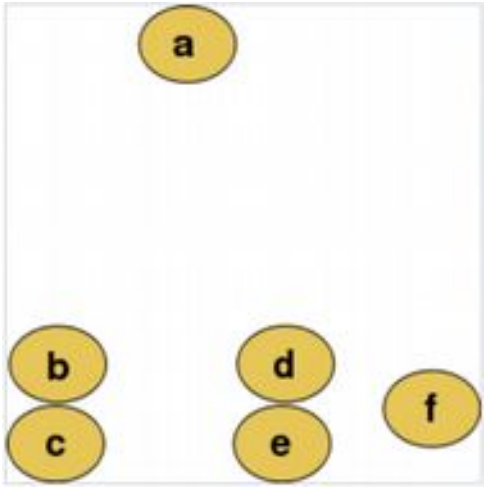
- Istnieje ponad 200 miar podobieństwa i niepodobieństwa.
- Dobór miar zależy od:
  - Ilości wymiarów
  - Rodzaju atrybutów (komplementarne, binarne itp.)
  - Rodzaju problemu
- Źródłem miar jest:
  - Norma
  - Przecięcie zbiorów
  - Ilość informacji (entropia)
  - Iloczyn skalarny
  - Test statystyczny
- Jeżeli wartość miary niepodobieństwa może być interpretowana ilościowo, mówimy że miara ma semantykę (np. odległość euklidesowa da się interpretować ilościowo)

Wybrane miary

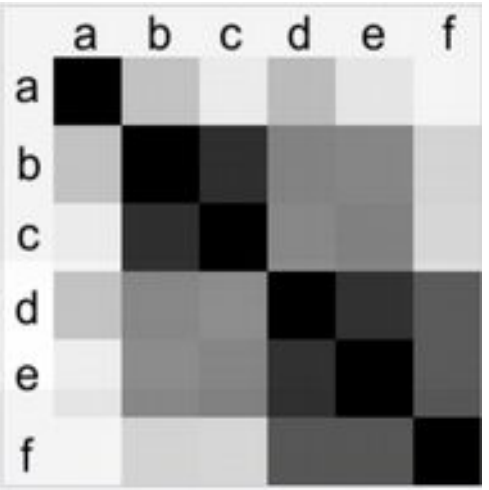
Miara	Rodzaj	Opis	Wzór
Współczynnik korelacji	S	Współczynnik korelacji Pearsona	
Kosinusowa	D	Kąt pomiędzy wektorami o niezerowej długości. Pokazuje orientację a nie natężenie. Stosuje się do wielowymiarowych zbiorów np. tekstów	$cosine(x,y)=\frac{\sum x_i, y_i}{\sqrt{\sum X^2}\sqrt{\sum Y^2}}$
Mahalanobis	D	Miara dostosowuje się do liniowej kombinacji wymiarów, określa ile odchyleń standardowych jest obiekt od średniej dla każdego z wymiarów	$d_{mah}=\sqrt{(X-Y)S^{-1}(X-Y)^T}$
Canberra	D	Standaryzowana [0,1] odmiana odległości miejskiej	$d_{canb}=\sum\frac{\lvert x_i-y_i\rvert}{(x_i+y_i)}$
Trójkątna	D	Standaryzowana [0,1] odmiana odległości euklidesowej, wysoka zgodność z JSD	$d_{tri}=\sqrt{\frac{1}{2}\sum\frac{(x_i-y_i)^2}{(x_i+y_i)}}$
Jensen-Shannon	D	Współdzielona ilość informacji dla dwóch rozkładów zmiennej kategoryzowanej, entropia wzajemna	$d_{jsd}=\sqrt{H(\frac{X+Y}{2})-\frac{1}{2}[H(X)+H(Y)]}$
Jaccard	S	Miara wielkości przecięcia dwóch zbiorów, stosowana dla atrybutów binarnych	$s_{jacard}=\sum\frac{\lvert X\cdot Y\rvert}{X}+Y-\lvert X\cdot Y\rvert$
Ružicka	D	Miara niezgodności rozkładów	$s_{roz}=\frac{\sum min(X,Y)}{\sum max(X,Y)}$

Macierz niepodobieństwa

- Macierz niepodobieństwa – zestawienie każdy z każdym wartości niepodobieństwa pomiędzy obiektami. W praktyce macierz dwu- wymiarowa, na przekątnej wartości 0 (aksjomat identyczności) i symetryczna (aksjomat symetryczności metryki)
- W przypadku nawet niewielkich zbiorów macierze przedstawia się w formie wizualizacji graficznej zamiast zbioru liczb
  - Skalowanie wielowymiarowe
  - Mapy ciepła
  - Grafy

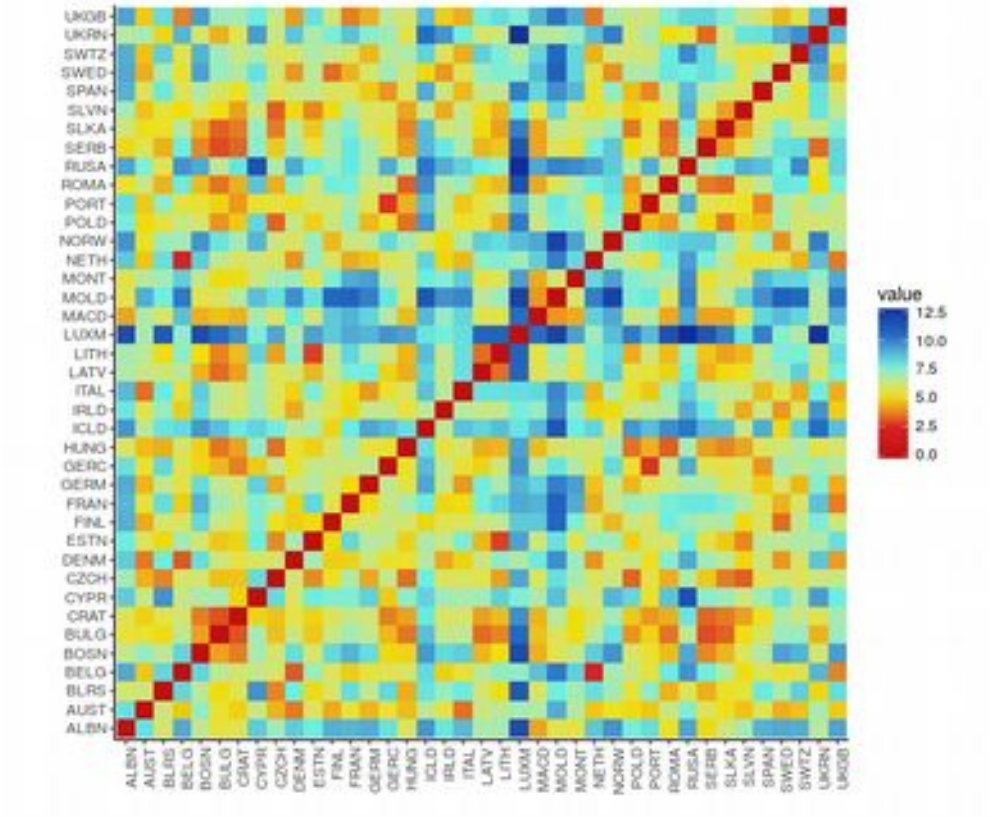


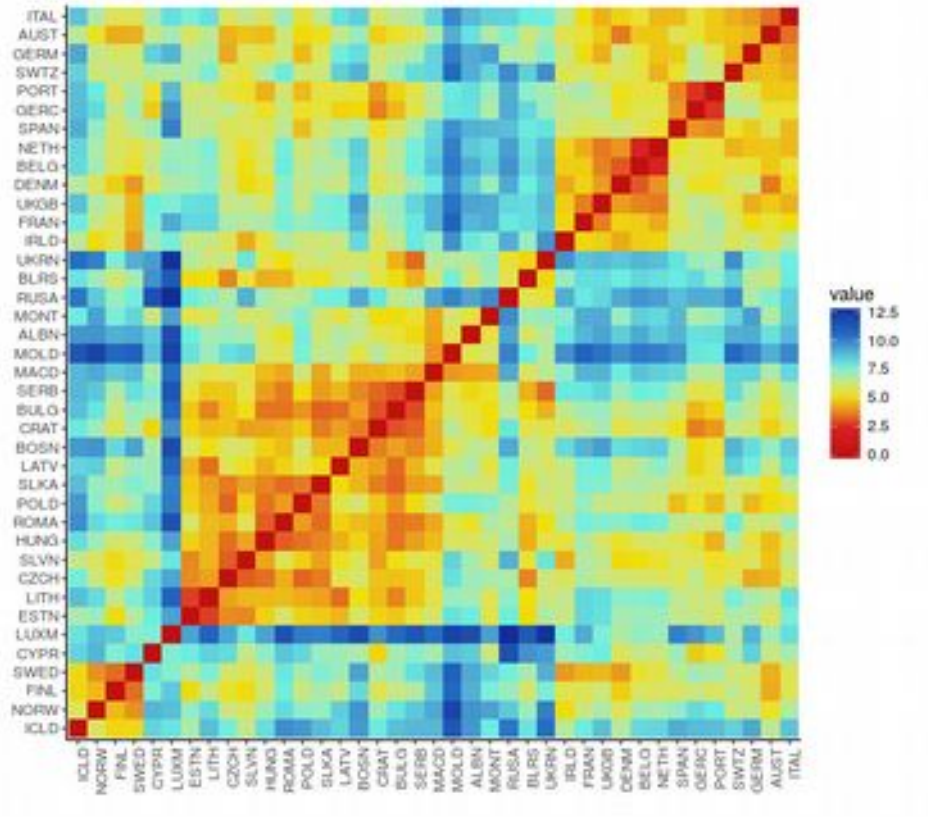
	a	b	c	d	e	f
a	0	184	222	177	216	231
b	184	0	45	123	128	200
c	222	45	0	129	121	203
d	177	123	129	0	46	83
e	216	128	121	46	0	83
f	231	200	203	83	83	0



Mapy ciepła

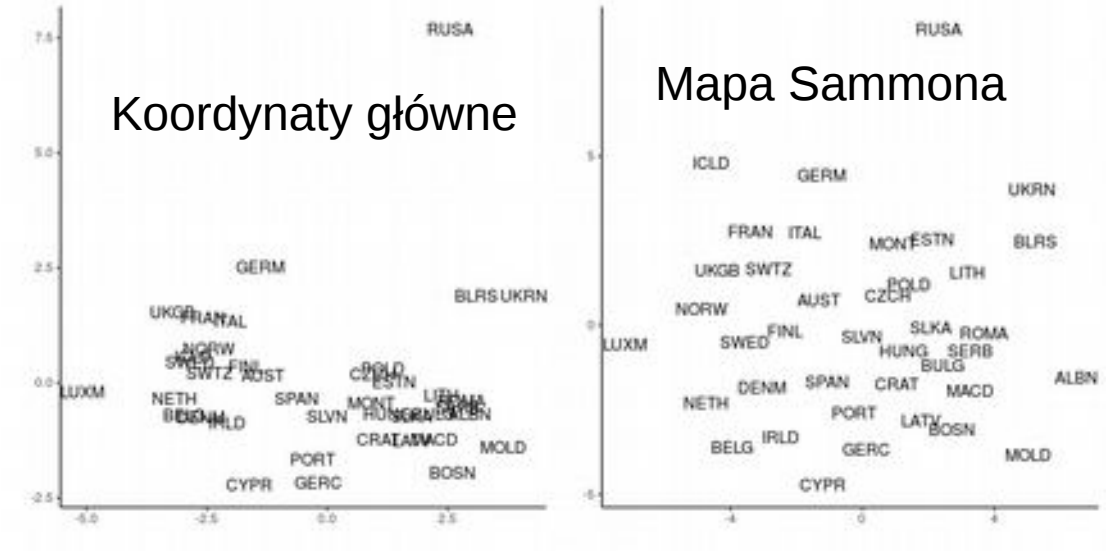
- Mapy ciepła to wizualizacja macierzy gdzie niepodobieństwo wyrażone jest kolorem. Uporządkowanie mapy wg niepodobieństwa ciepła pozwala wykryć struktury w danych

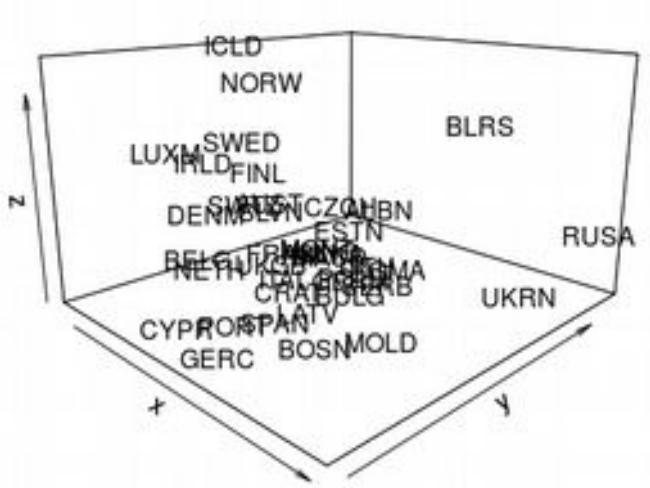




Skalowanie wielowymiarowe

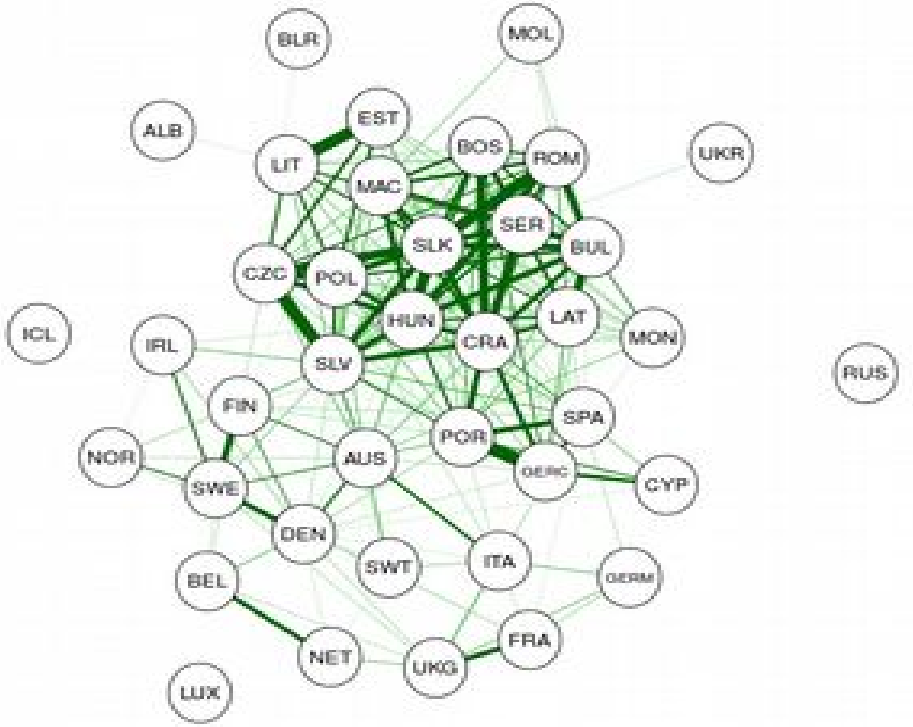
- To forma prezentowania niepodobieństwa pomiędzy obiektami poprzez rzutowanie ich do przestrzeni 2-u lub trójwymiarowej, w taki sposób aby minimalizować różnice niepodobieństwa pomiędzy wartościami z oryginalnej wielowymiarowej przestrzeni a nowej przestrzeni zredukowanej
- Skalowanie wielowymiarowe stosuje się również do konwersji pomiędzy atrybutami komplementarnymi a wektorami cech





Grafy

- Grafy to forma prezentacji (nie)podobieństwa w formie obiektów rozmieszczonych w przestrzeni (wierzchołki albo węzły) a łączących je linii (krawędzie), których waga reprezentuje podobieństwo między obiektami

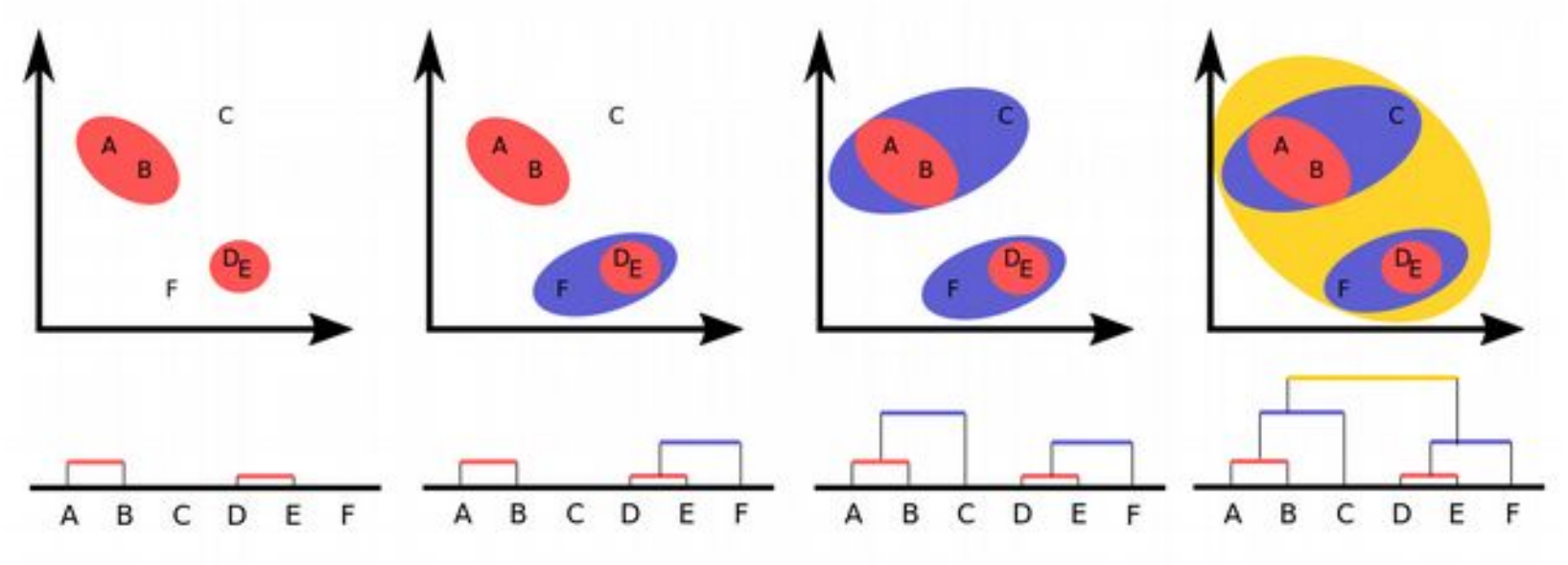


Algorytmy grupowania

- Grupowanie to podział zbioru danych na grupy rozłączne i wewnętrznie spójne
- Stosuje się kilka różnych metod
  - **Metody hierarchiczne**
  - **Metody rozdzielające (partitioning)**
  - **Metody rozmyte**
  - **Grupowanie probablistyczne**
  - Metody gęstościowe
  - Ulepszony hierarchiczny (BIRCH)

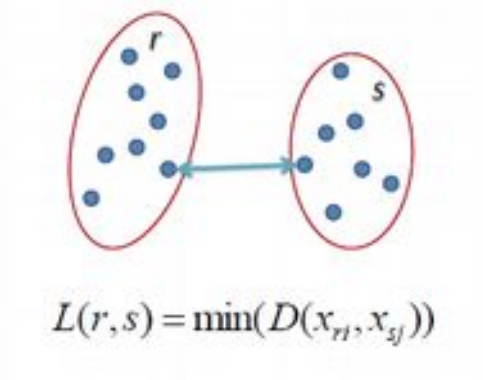
Grupowanie hierarchiczne

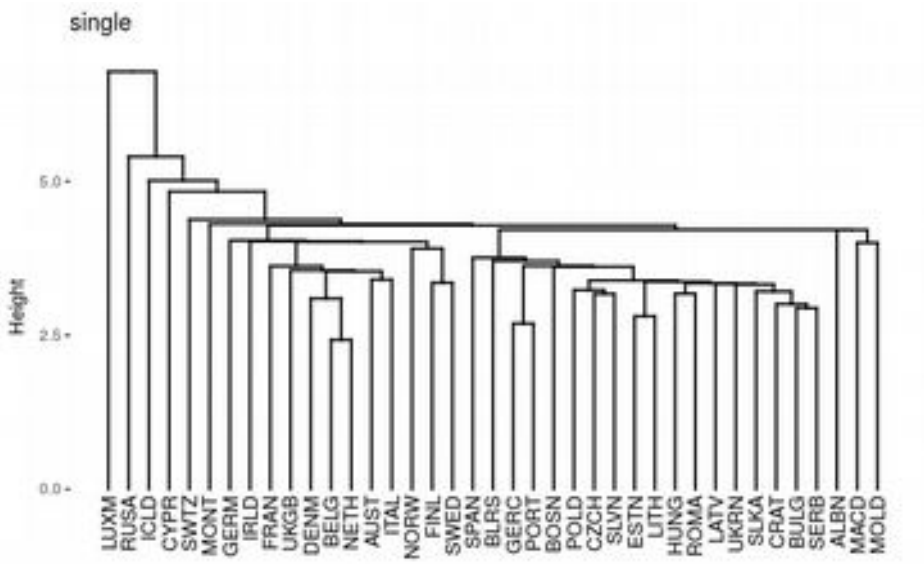
- Metoda analizy skupień, której celem jest zbudowanie hierarchii grup
- Stosuje metody aglomeracyjne lub dzielące (rzadziej)
- Algorytm zachłanny, szybki ale niepotymalny globalnie
- Nadaje się do małych zbiorów danych, gdzie struktura (hierarchia) jest ważniejsza niż same skupienia
- Strategie łączenia: pojedyncze, całkowite, średnie



Pojedyncze łączenie

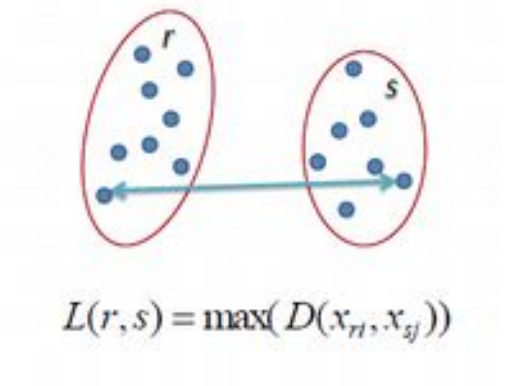
- Jako pierwsze zostaną połączone dwa obiekty o najmniejszym niepodobieństwie
- W następnych krokach będą łączone te obiekty lub grupy, gdzie niepodobieństwo pomiędzy dwoma **najbardziej podobnymi** obiektami **jest najmniejsze**
- Metoda wykrywania obiektów odstających

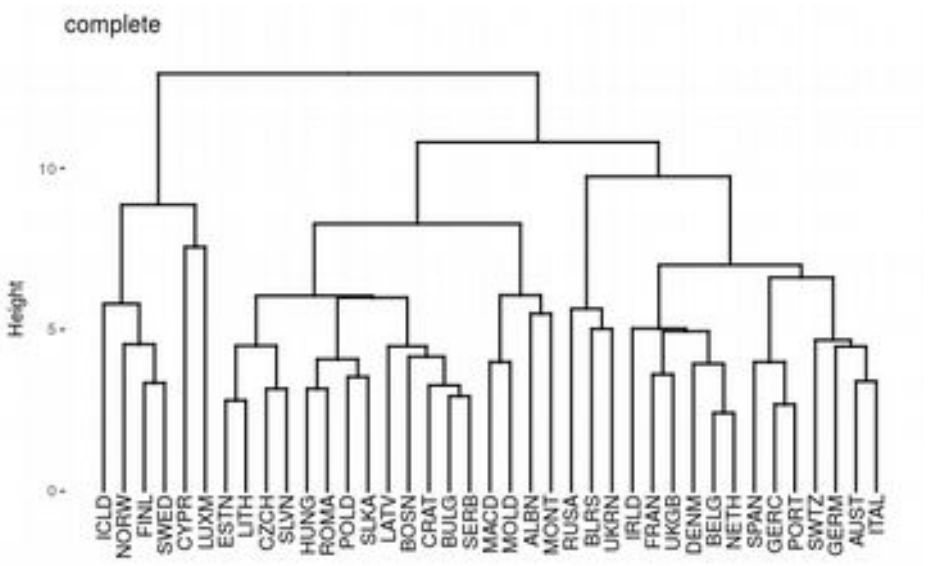




Całkowite łączenie

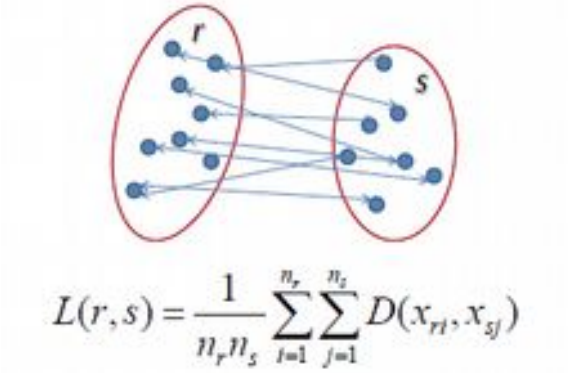
- Jako pierwsze zostaną połączone dwa obiekty o najmniejszym niepodobieństwie
- W następnych krokach będą łączone te obiekty lub grupy, gdzie niepodobieństwo pomiędzy dwoma **najmniej podobnymi** obiektami **jest najmniejsze**
- Klasyczna metoda budowania hierarchii, nie wykrywa obiektów odstających



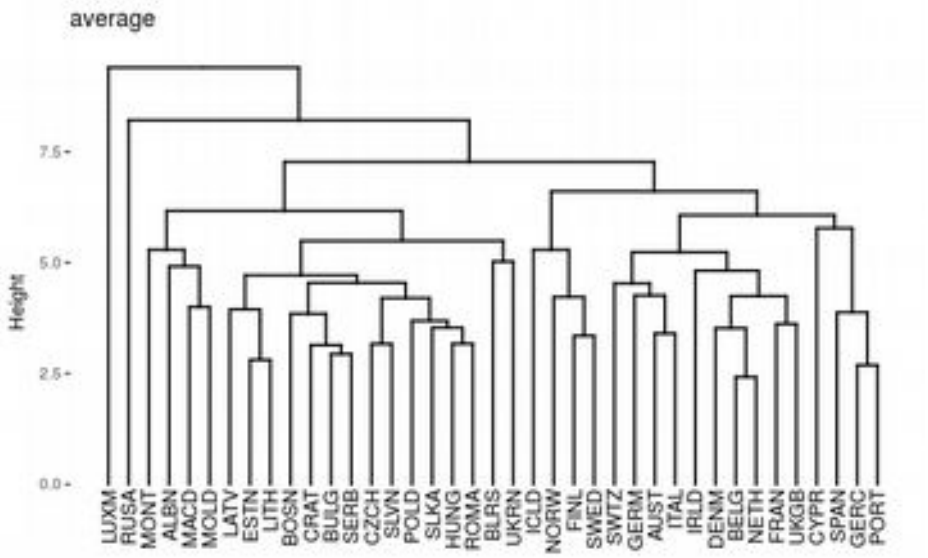


Średnie Łączenie

- Jako pierwsze zostaną połączone dwa obiekty o najmniejszym niepodobieństwie
- W następnych krokach będą łączone te obiekty lub grupy, gdzie **średnie** niepodobieństwo pomiędzy obiektami **jest najmniejsze**
- Klasyczna metoda budowania hierarchii, z wykrywaniem obiektów odstających

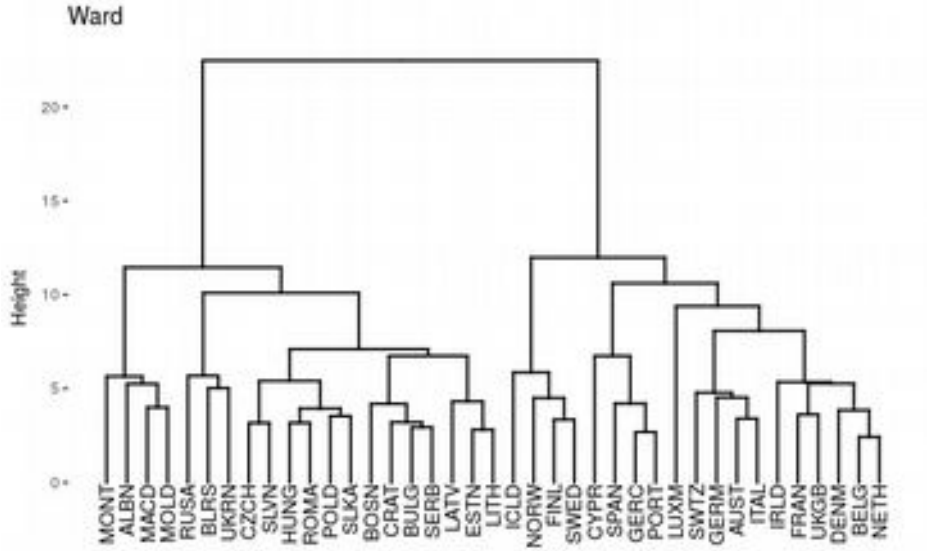


$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$



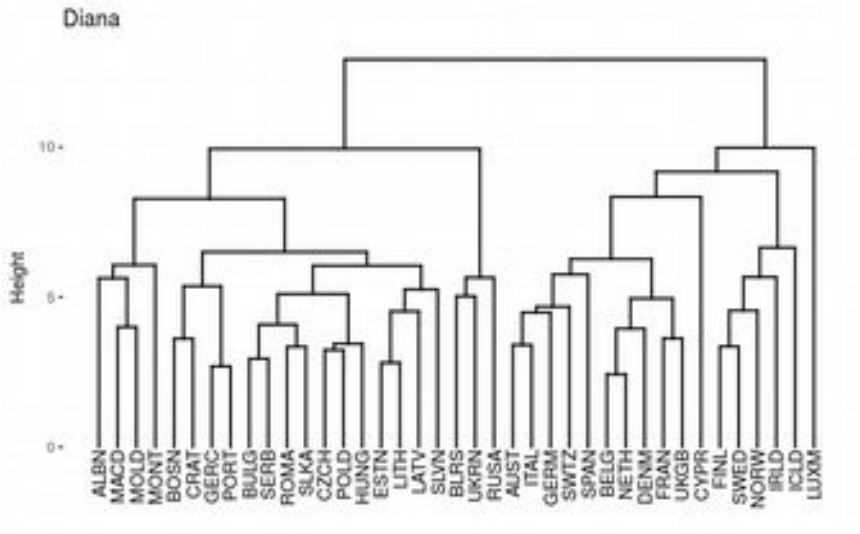
Grupowanie hierarchiczne Warda

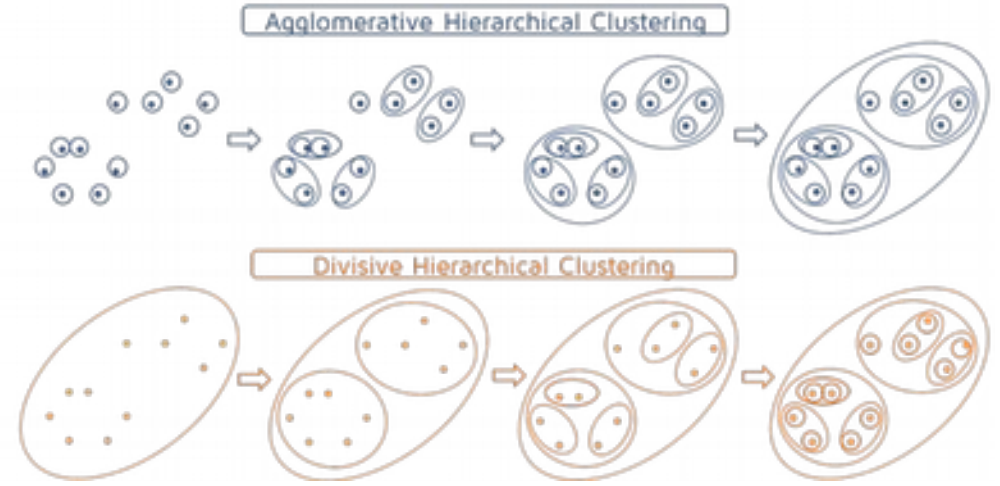
- Minimalizacja wariancji wewnątrz skupień, maksymalizacja wariancji między skupieniami
- Nie buduje rzeczywistych hierarchii ale pozwala określić naturalną liczbę skupień
- Nie wykrywa obiektów odstających



Metoda hierarchicznego rozdzielania

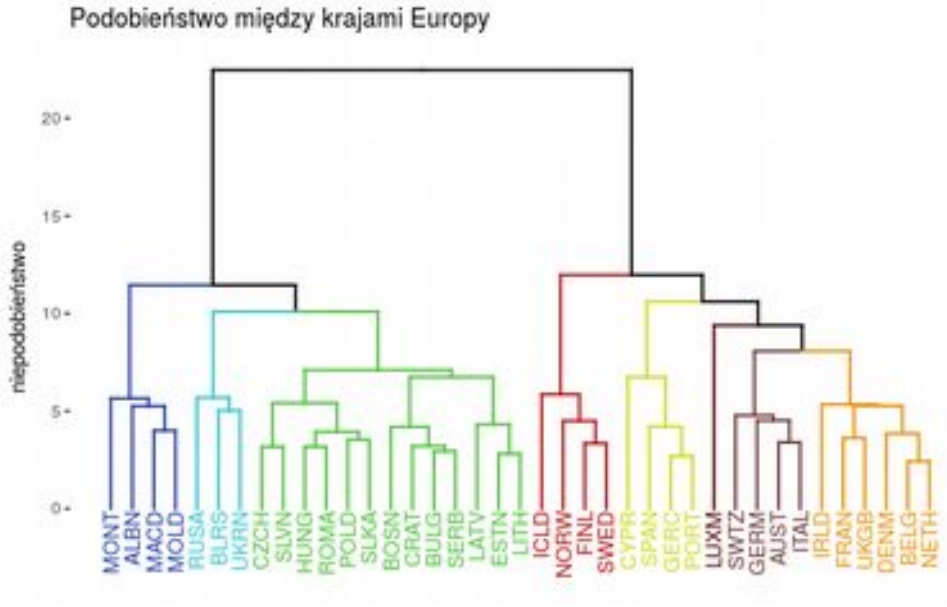
- Rozpoczyna od jednego skupienia obejmującego wszystkie obiekty
- Rozdziela skupienie tak aby maxymalizować wariancję między nimi
- Kontynuuje proces aż do końca
- Nie wykrywa obiektów odstających

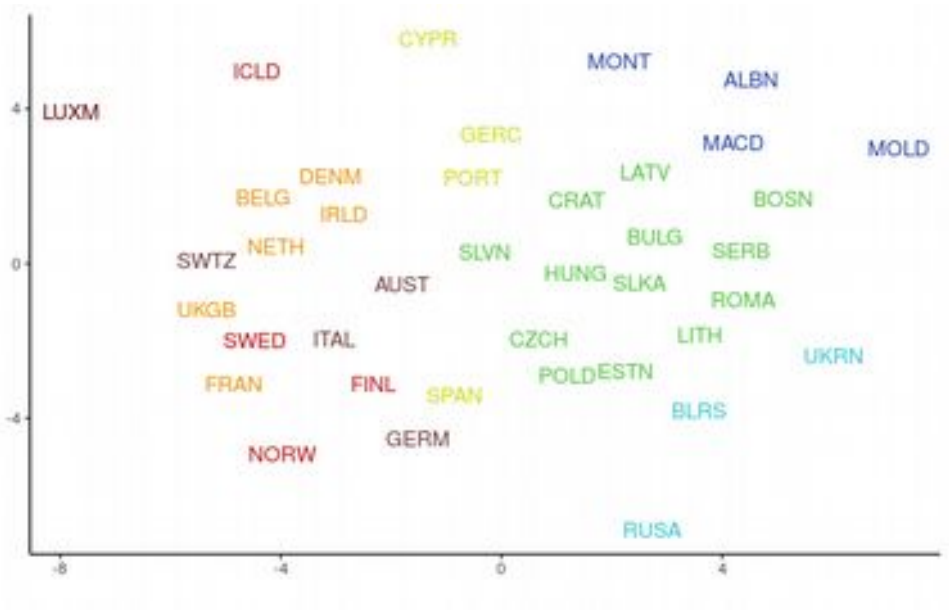




Zalety i wady metod hierarchicznych

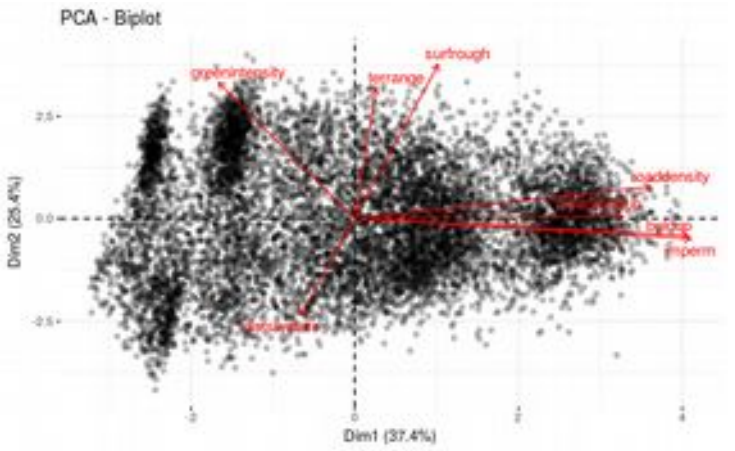
- Zalety
  - Szybki algorytm
  - Deterministyczny algorytm (powtarzalność wyników)
  - Buduje intuicyjnie zrozumiałą hierarchię
- Wady
  - Algorytm zachłanny, optymalizowany na poziomie kroku a nie całości wyniku
  - Każda decyzja nie może być zmieniona
  - Skupienia rozmieszczone są liniowo, tracimy informację o relacjach pomiędzy skupieniami



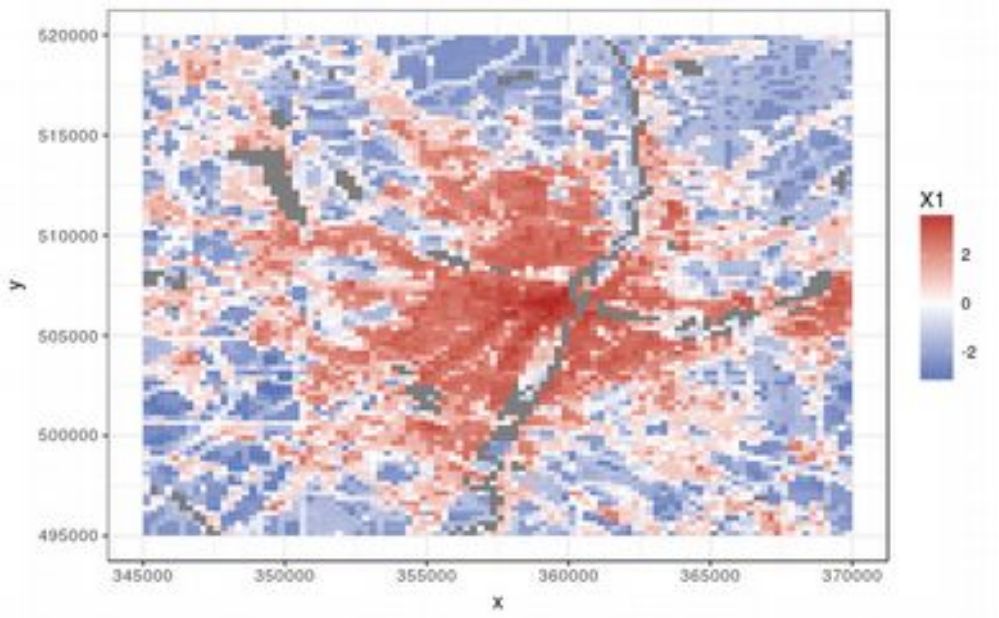


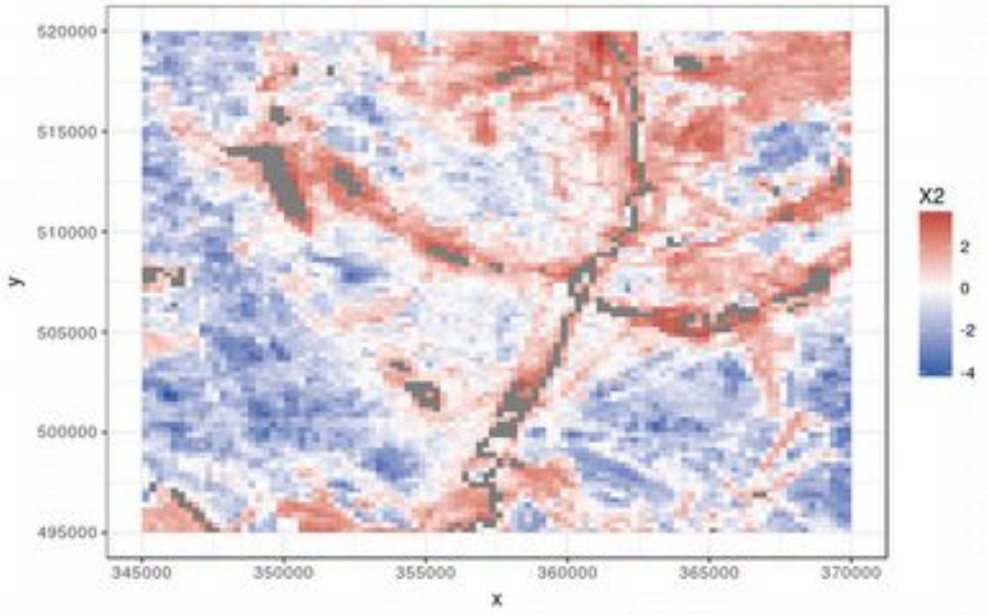
Metody partycjonujące

- K-średnie
- K-medoidy
- Propagacja afiniczności
- Rozmyte k-średnie



Jako przykład zostaną użyte dane z poprzedniego wykładu

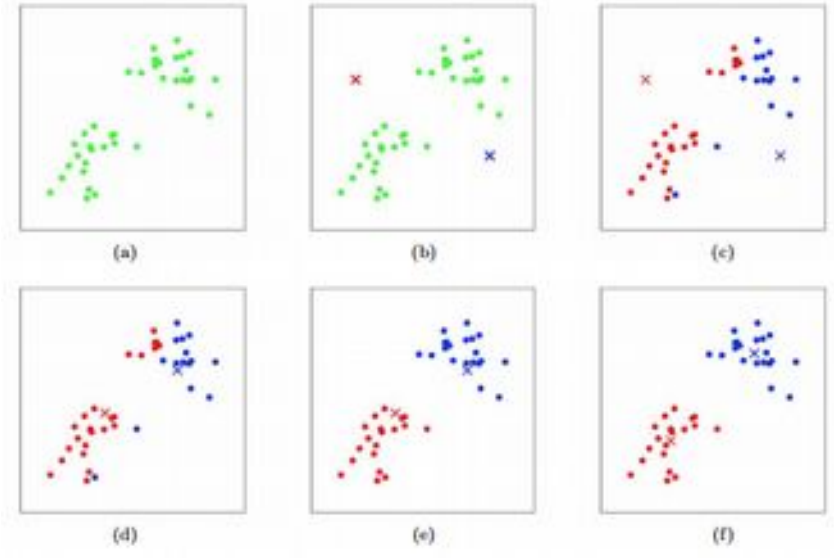


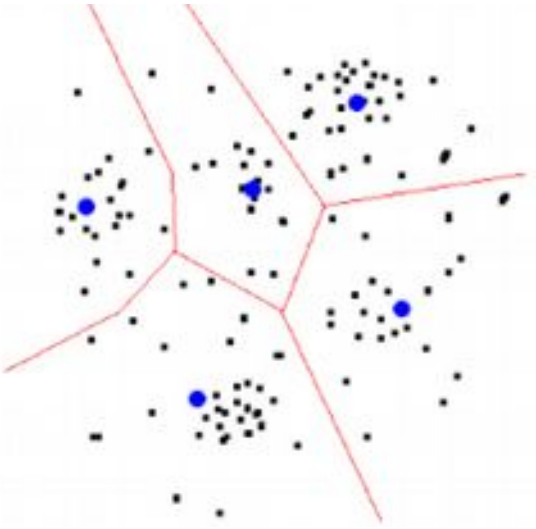


Metoda k-średnich

1) Algorytm stochastyczny rozpoczyna losowo położonymi punktami (centroidami)
2) Przypisuje obiekty do centroidów na zasadzie minimalnego niepodobieństwa
3) Wyznacza nową lokalizację na podstawie zasięgu skupienia
4) Powtarza (2) aż do momentu gdy położenie centroidów nie zmieni się

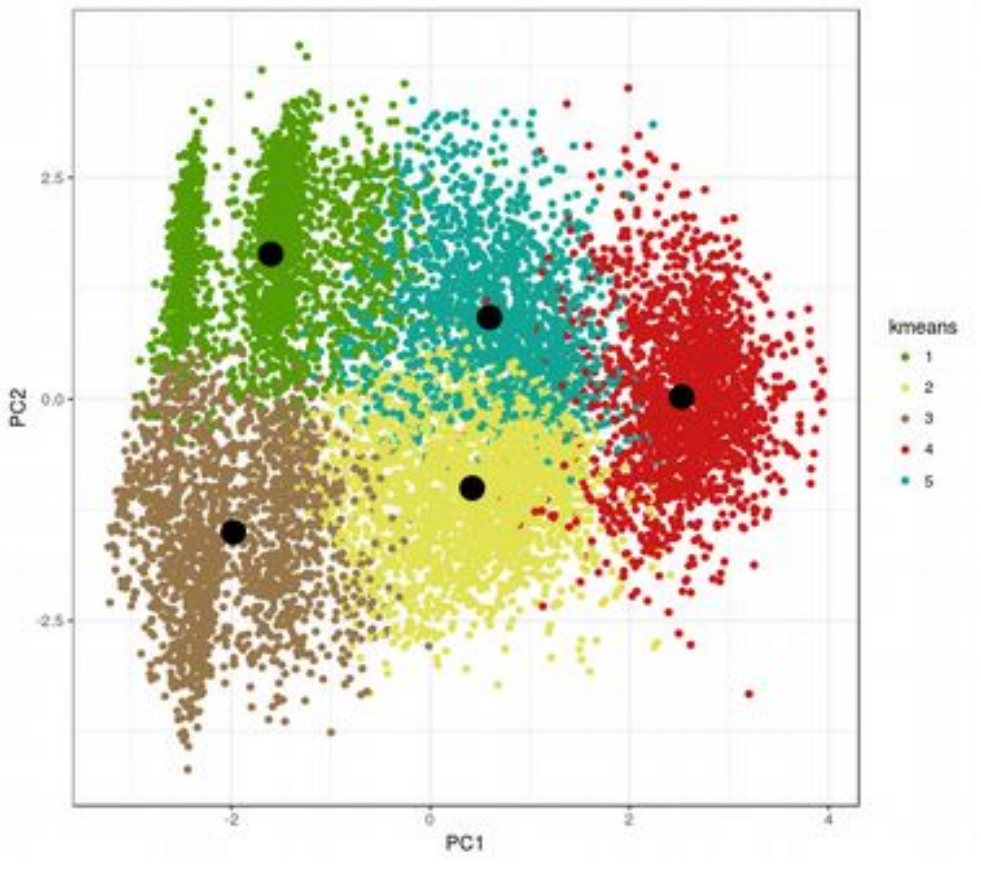
- Ze względu na duży wpływ początkowej konfiguracji algorytm rozpoczyna proces wielokrotnie, wybierając najbardziej powtarzalne wyniki

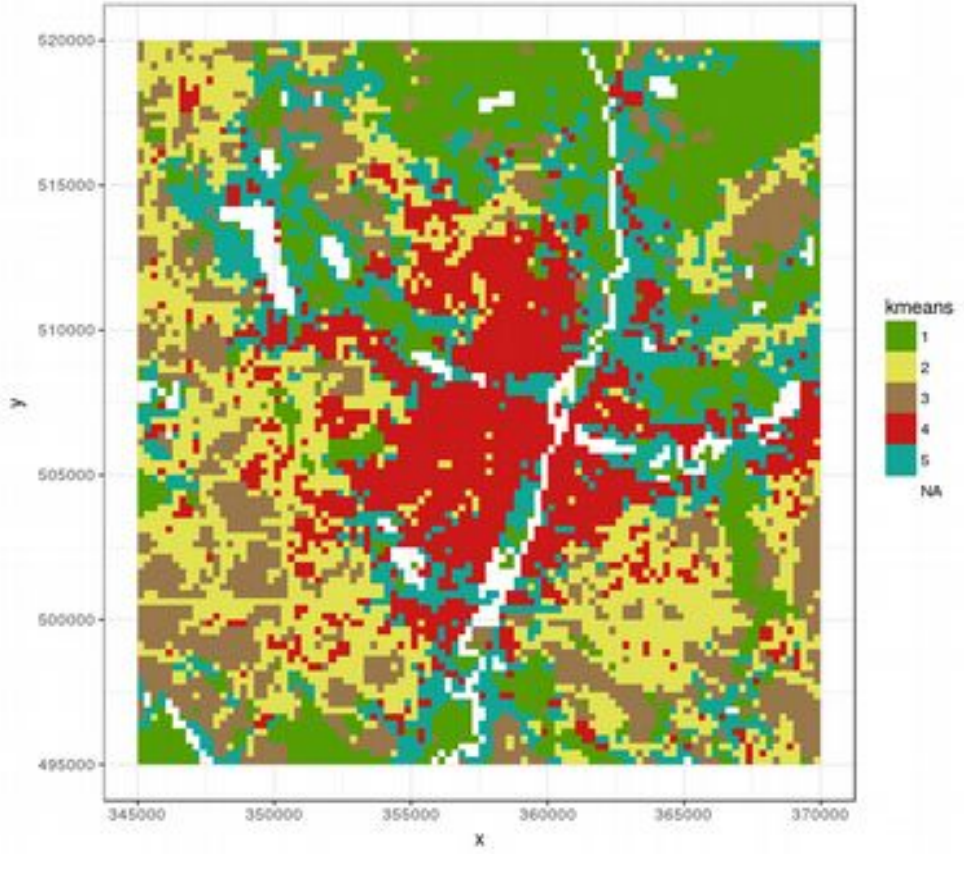




Zastosowanie dla danych geoprzestrzennych

- Metoda wymaga podania liczby skupień
- Wynik podziału jest zgodny z kryterium Voronoi
- Zagęszczenia w rozkładzie nie mają znaczenia dla procesu wyznaczania skupień
- Niepewność przynależności nie jest brana pod uwagę





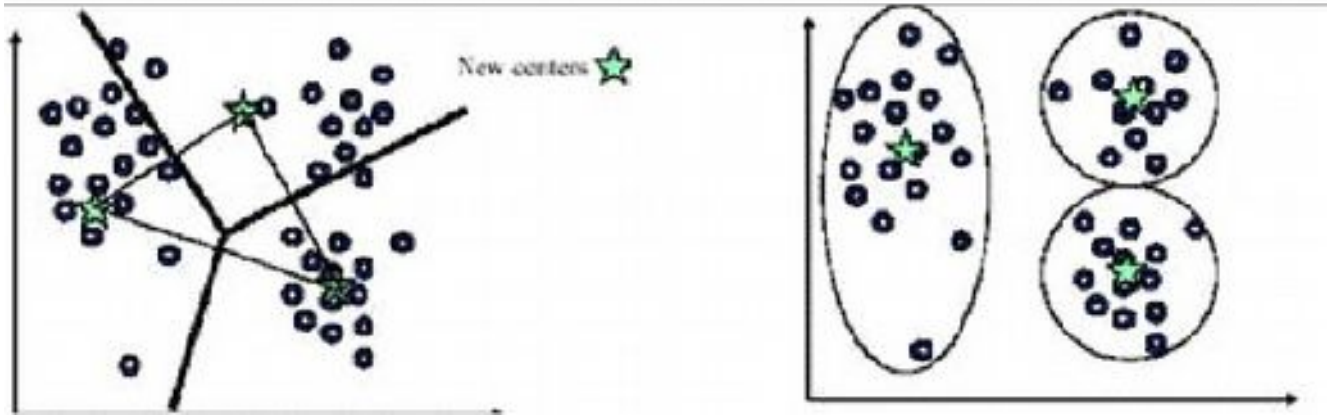
Metoda k-medoidów

- Algorytm podobny do k-średnich. Nie używa abstrakcyjnych centroidów ale rzeczywiste obiekty ze zbioru (medoidy)

1) Algorytm stochastyczny rozpoczyna losowo wybranymi obiektami (medoidami)

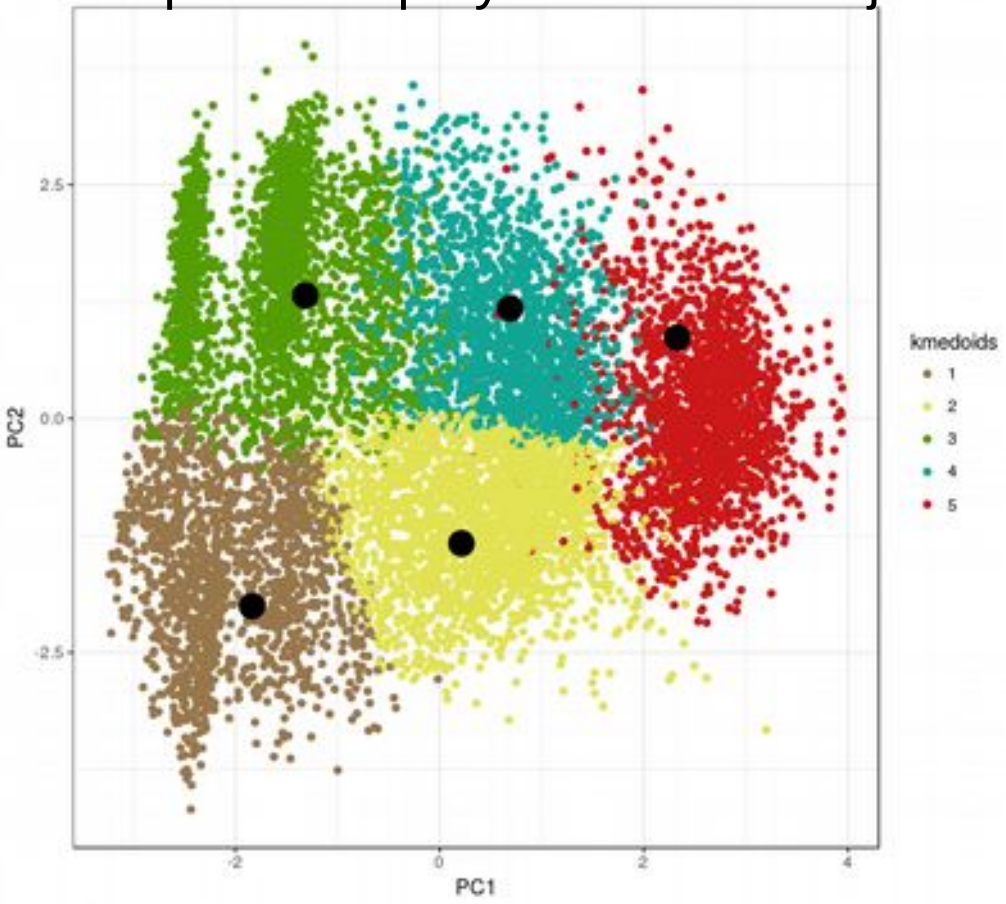
2) Pozostałe kroki jak w k-means

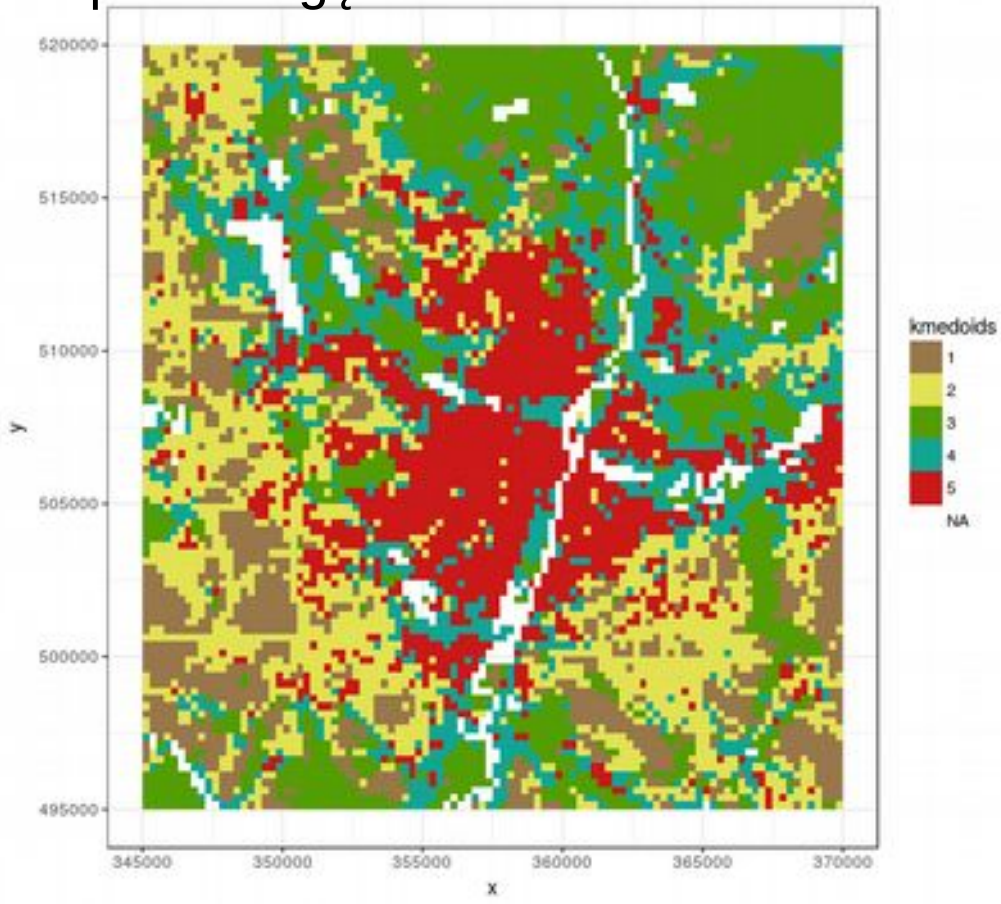
- Ze względu na duży wpływ początkowej konfiguracji algorytm rozpoczyna proces wielokrotnie, wybierając najbardziej powtarzalne wyniki
- W przeciwieństwie do kmeans dużo bardziej odporny na obiekty odstające – jeżeli występują



Zastosowanie dla danych geoprzestrzennych

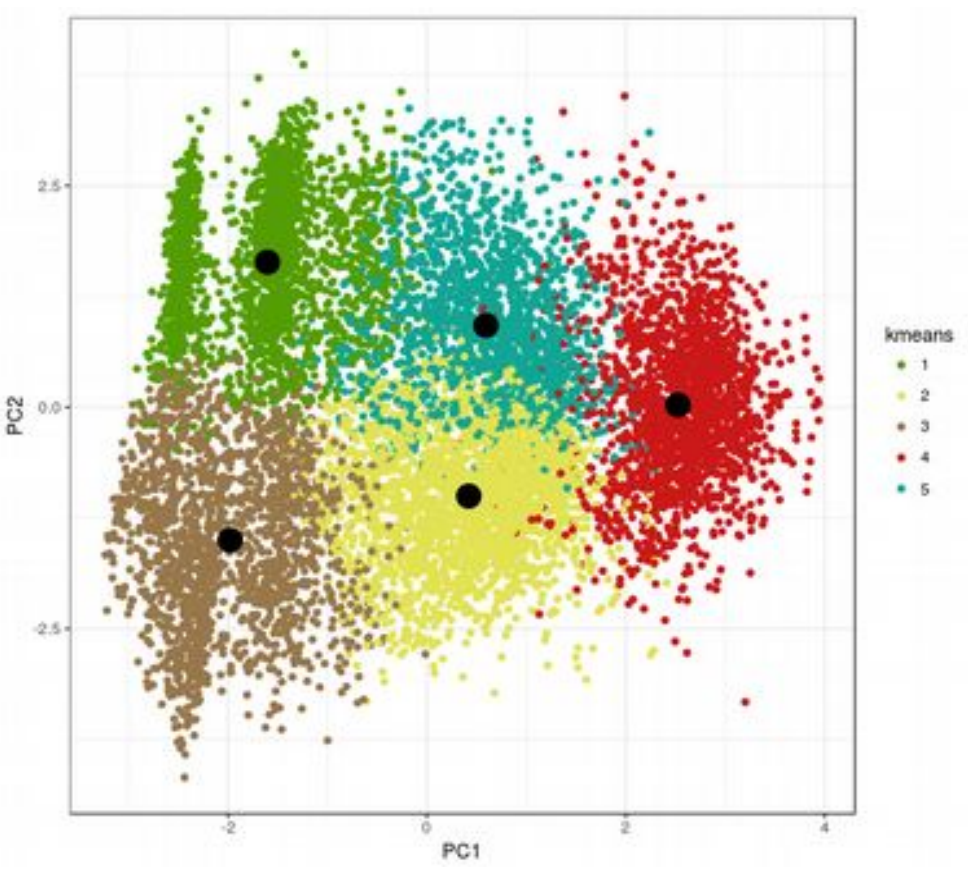
- Metoda wymaga podania liczby skupień
- Zagęszczenia w rozkładzie mają znaczenie dla wyznaczania skupień
- Mały wpływ obiektów odstających
- Niepewność przynależności nie jest brana pod uwagę

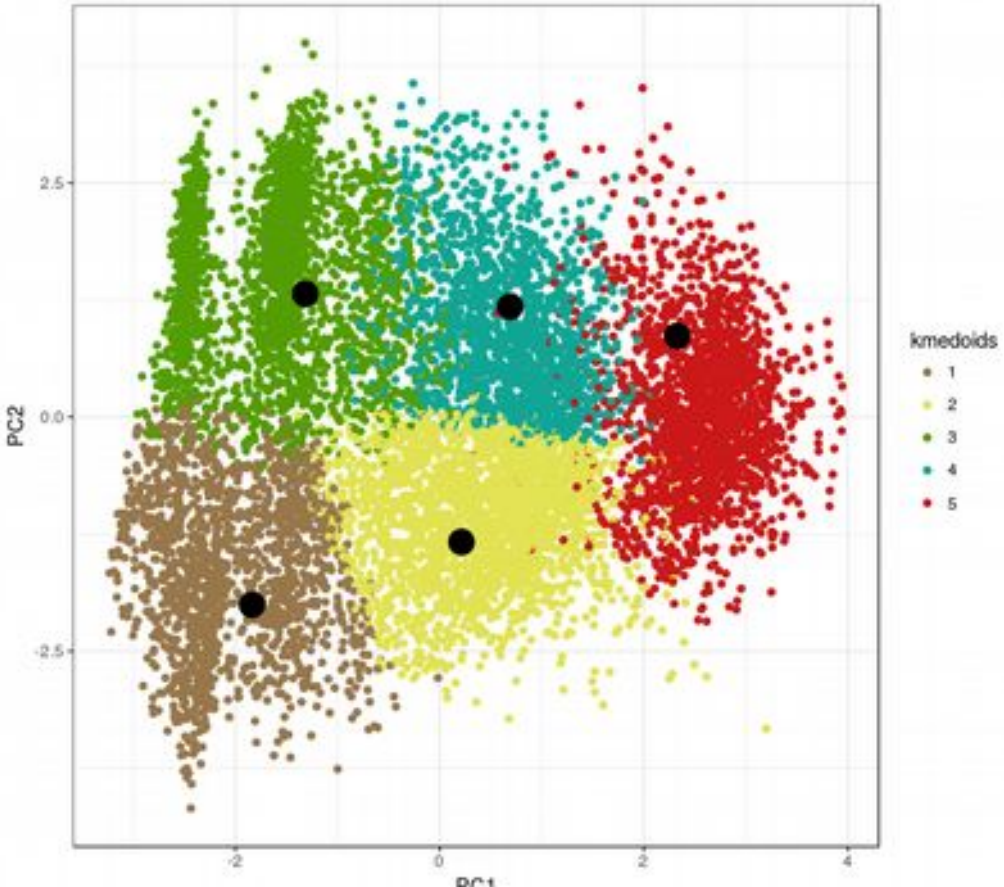




Porównanie skupień

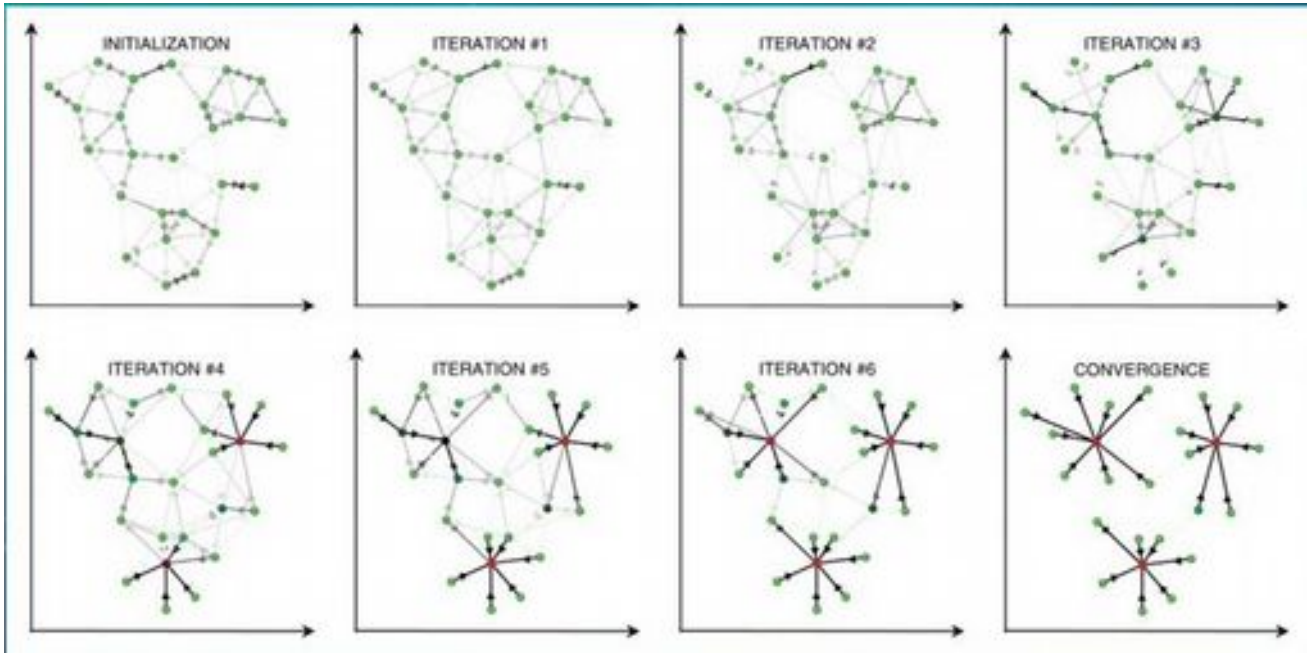
- Wyniki bardzo podobne
- K – medoids daje wyraźniejsze skupienia
- Obie metody zaliczane są do suboptymalnych tj wynik nie jest najlepszy z możliwych ale akceptowalny





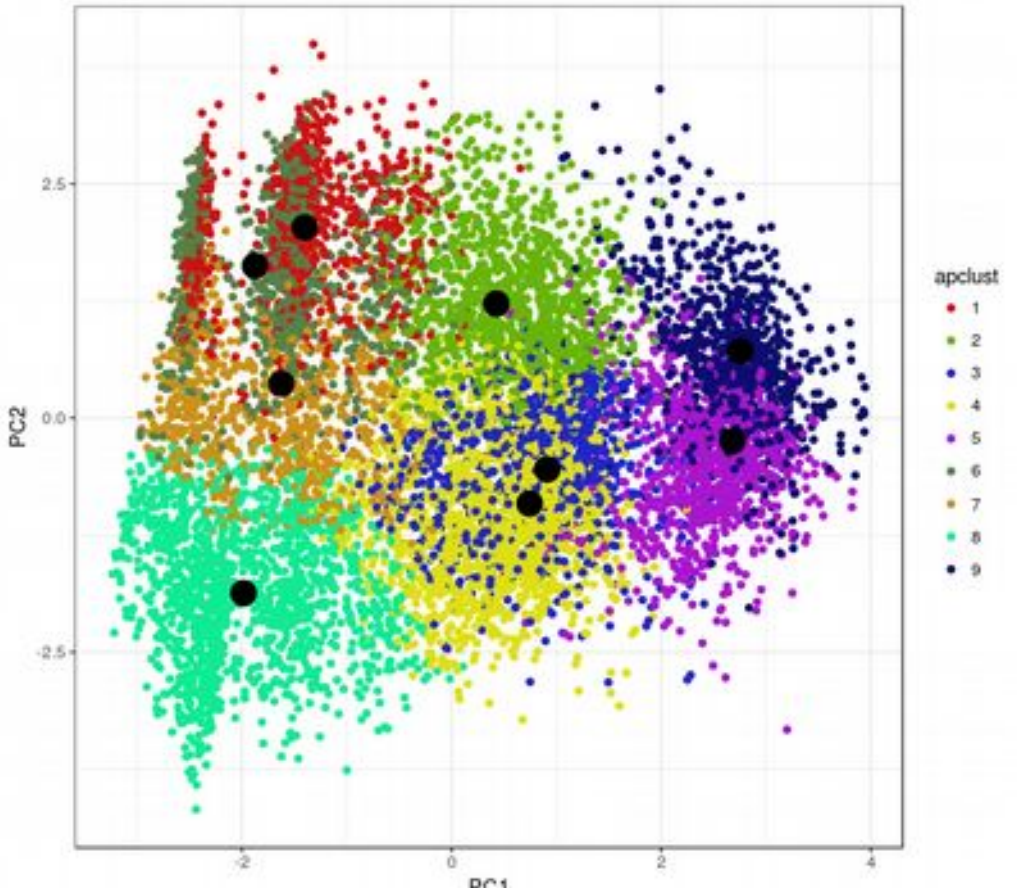
Propagacja powinowatości

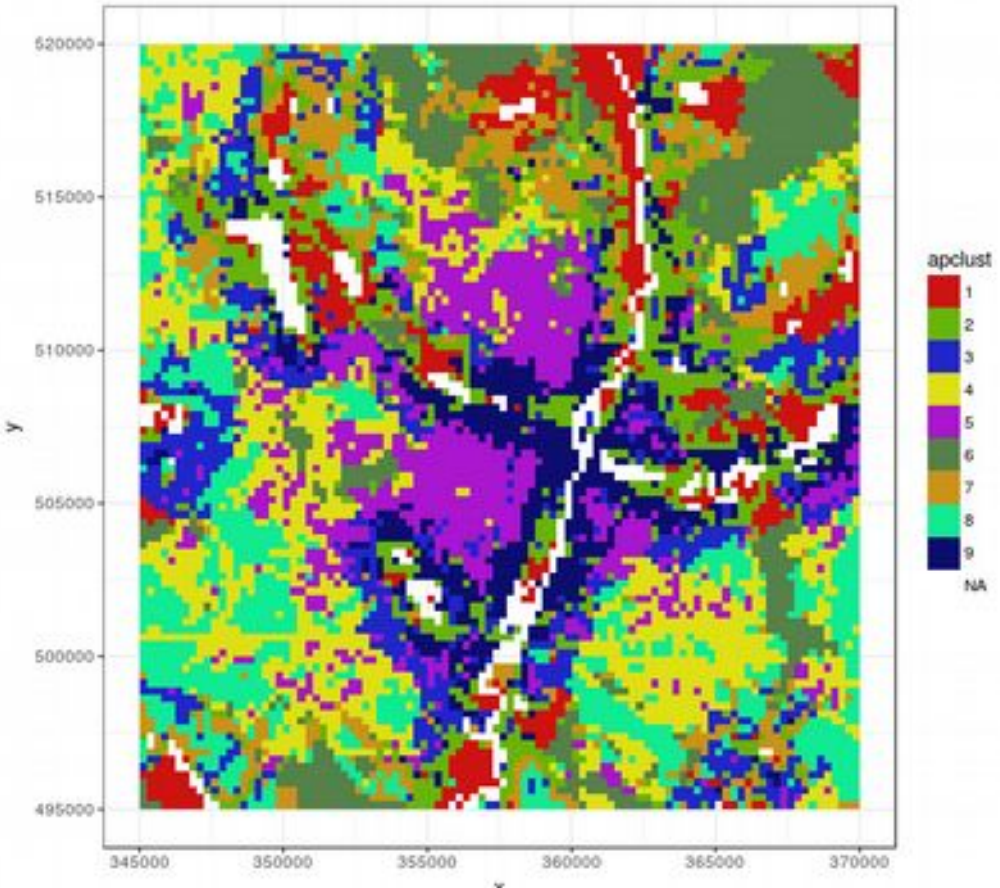
- Affinity propagation – metoda polegająca na iteracyjnym „przekazywaniu wiadomości” pomiędzy obiektami. Ma na celu wybór obiektów – egzemplarów
- Jako egzemlary wskazywane są te obiekty, które wykazują dodatni bilans pomiędzy byciem egzemplarem, a posiadaniem egzemplara
- Metoda wyszukiwania naturalnych liderów. Bardziej „pasuje do wszystkich” niż „mistrz w jednym”



Zastosowanie dla danych geoprzestrzennych

- Metoda nie wymaga podania liczby skupień jedynie kryterium selekcji egzemplarów
- Zagęszczenia w rozkładzie mają znaczenie dla wyznaczania skupień
- Obiekty odstające tworzą nowe skupienia
- Niepewność przynależności nie jest brana pod uwagę



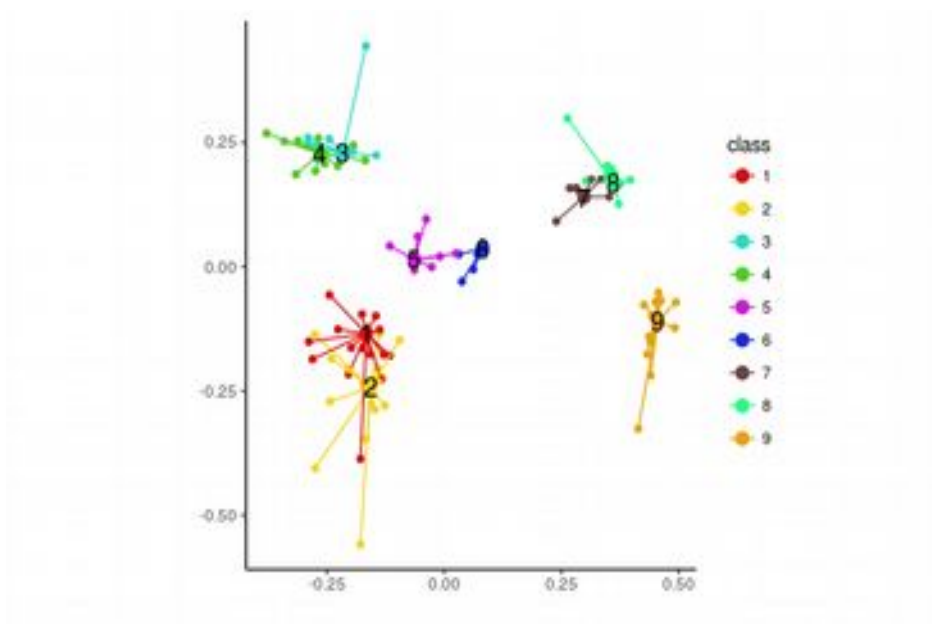


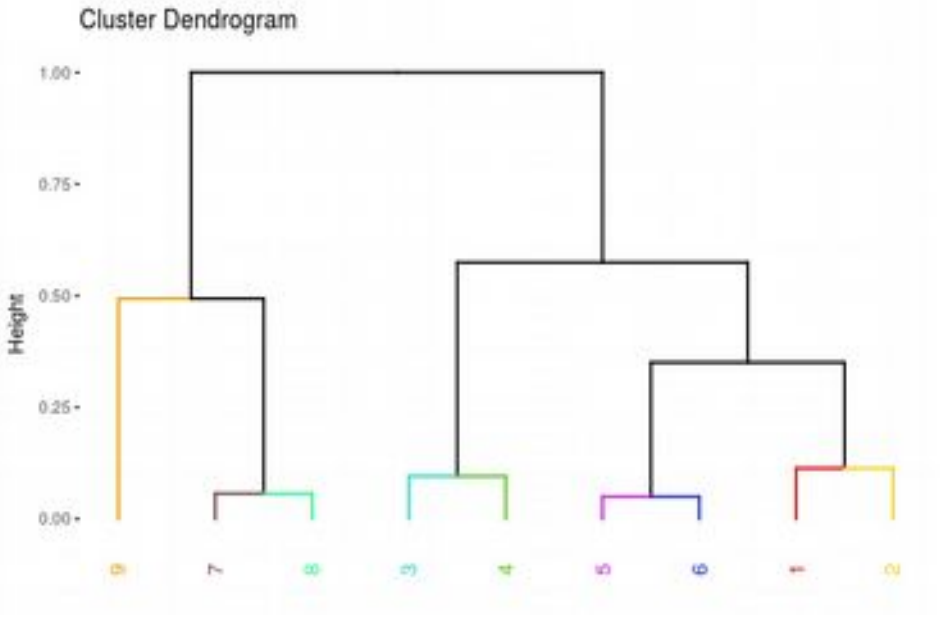
Centroid – Medoid - Egzemplar

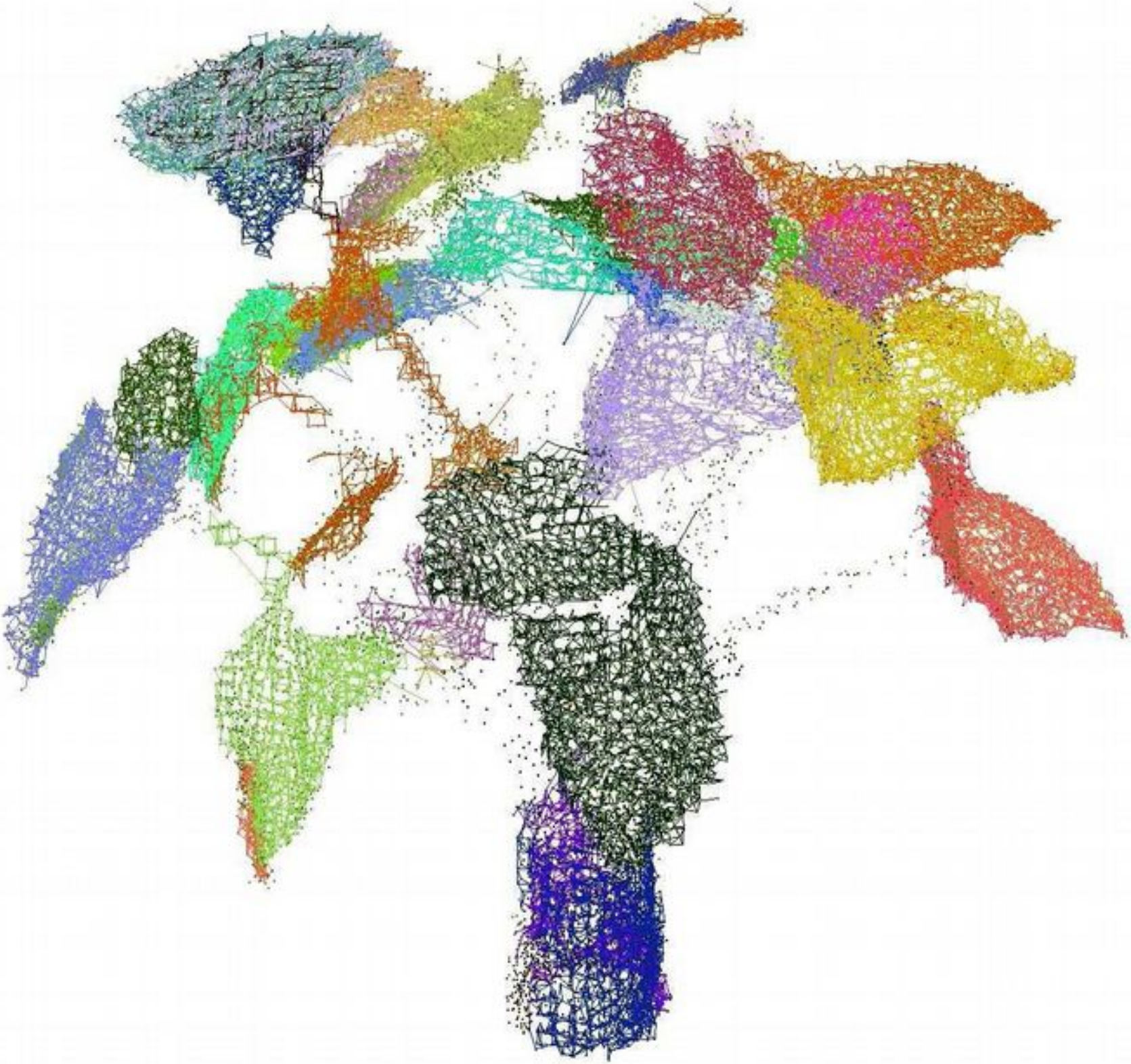
- **Centroid:** współrzędne w przestrzeni wielowymiarowej oznaczające geometryczny środek skupienia. Nie jest to fizyczny obiekt. Może być poza obszarem skupienia
- **Medoid:** obiekt najbardziej podobny do innych obiektów. Z reguły występuje w największym zagęszczeniu skupienia.
- **Egzemplar:** naturalny przedstawiciel skupienia, najbardziej reprezentatywny dla innych obiektów

Łączenie metod partycjonujące i hierarchicznych

- Metody partycjonujące i hierarchiczne można łączyć, ale nie wszystkie implementacje używają tej możliwości
- Metody hierarchiczne jako szybsze, ale mniej dokładne używa się do wstępnego podziału zbioru na skupienia, przed uruchomieniem metody k-means/k-medoids w celu uniknięcia losowej konfiguracji startowej
- W przypadku dużych zbiorów danych metody partycjonujące używa się do wyznaczenia dużej liczby małych zwartych skupień, a następnie małe skupienia łączy się w hierarchie

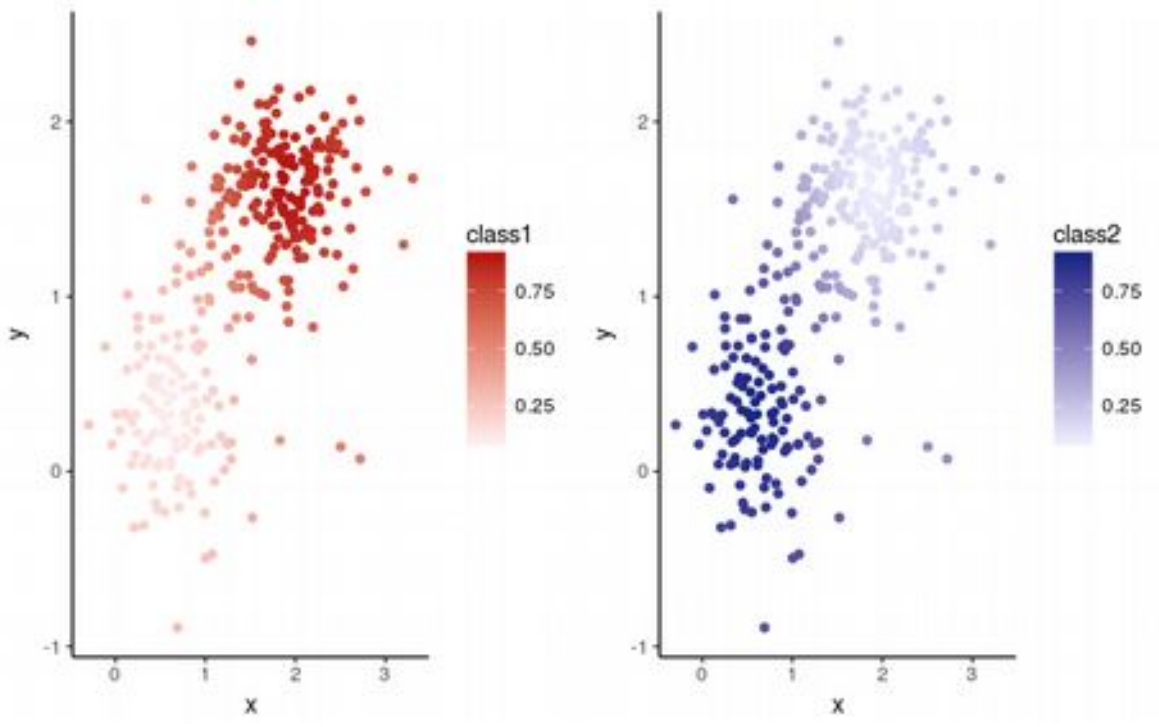






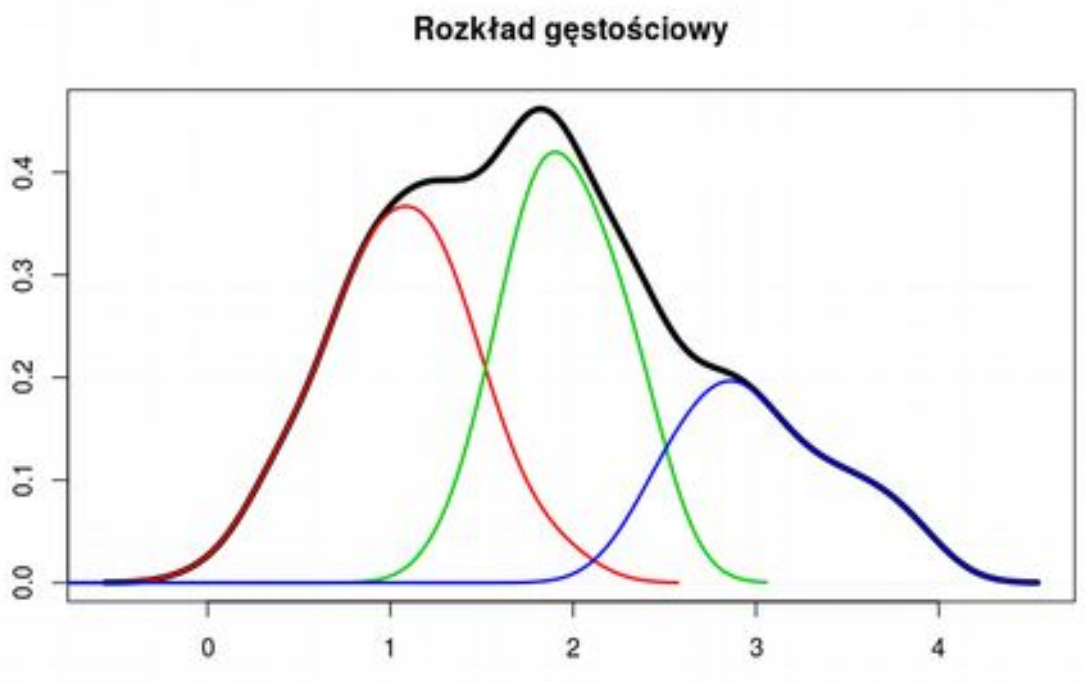
Metody rozmyte (k- means i k-medoids)

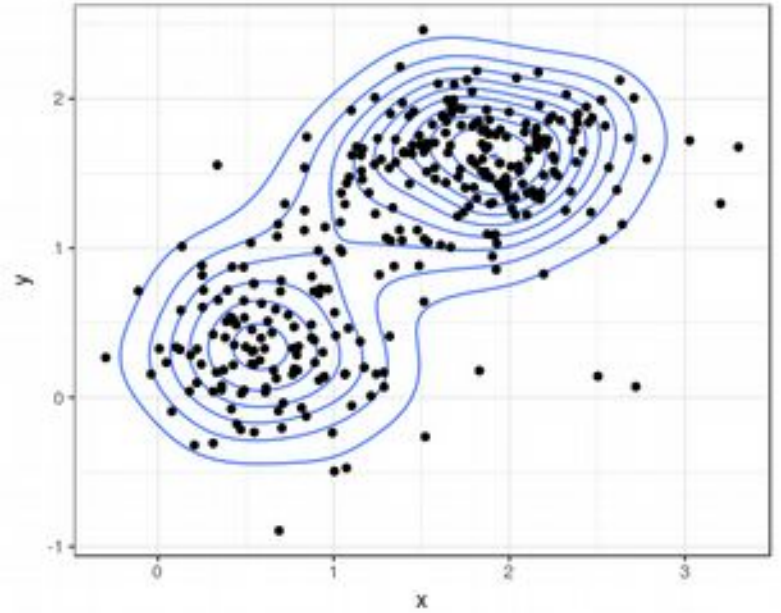
- Metoda bierze pod uwagę niepewność przynależności
- Każdy obiekt jest przypisywany do więcej niż jednej z klas
- zastosowaniem metod rozmytych jest sytuacja, gdy interesują nas jedynie wybrane skupienia i chcemy określić dla nich tolerancję przynależności kosztem innych skupień.
- Koncepcja krytykowana:
  - w ostateczności obiekt musi przynależeć to jakiegoś skupienia,
  - do określenia niepewności przynależności służą inne metody
  - metody rozmyte mają problemy z prawidłowym klasyfikowaniem punktów na obrzeżach.
  - problemy z wizualizacją skupień, gdyż wymagają osobnego diagramu dla każdego skupienia.



Metody probablistyczne

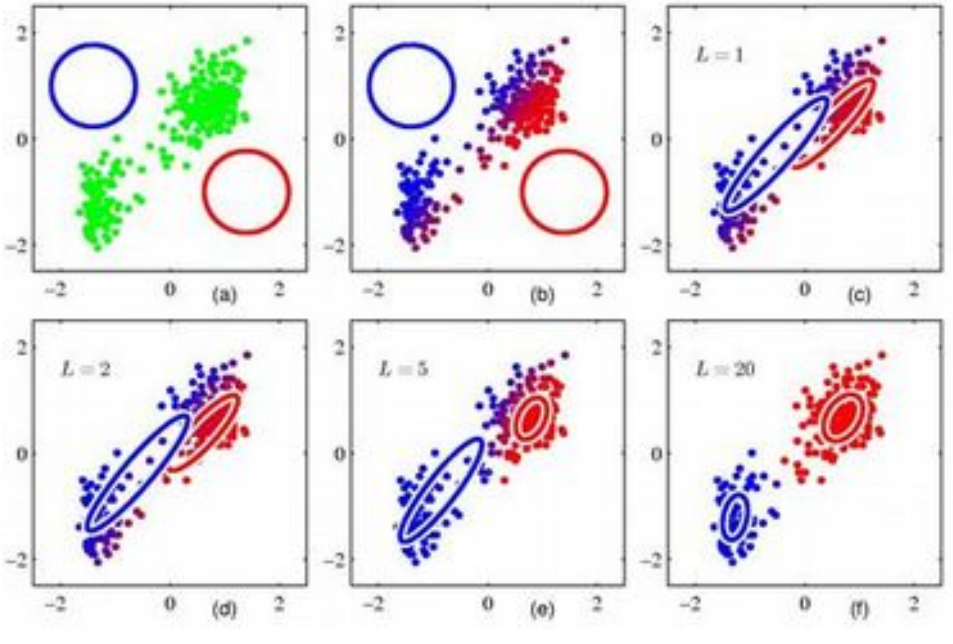
- Gaussowskie modele mieszane
- Jeżeli rozkład gęstości nie ma jakiejś konkretnej postaci można przyjąć założenie że jest sumą wielu rozkładów normalnych
- Znalezienie rozwiązania jest problemem optymalizacyjnym, szuka się optymalnej liczny skupeń oraz właściwych dla nich rozkładów

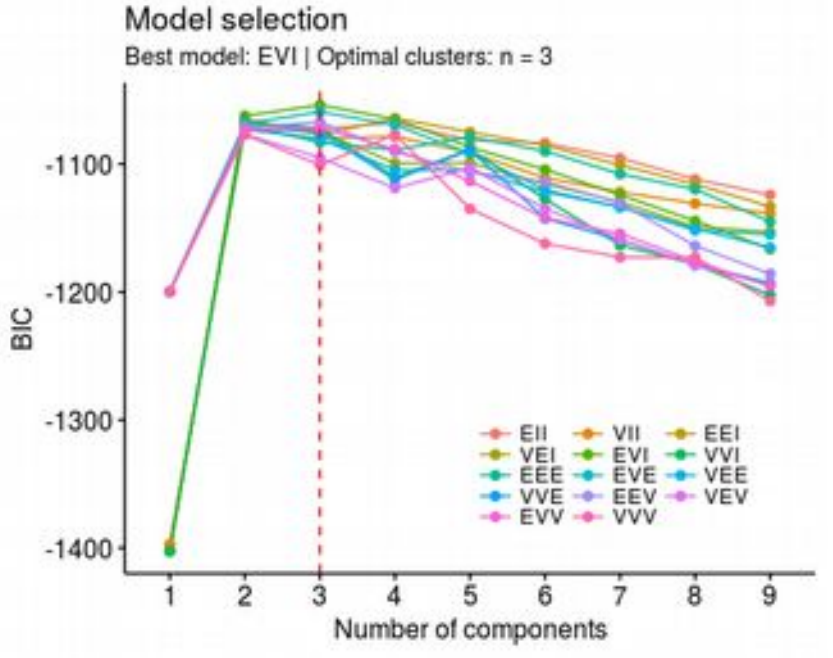




Expectation-maximisation

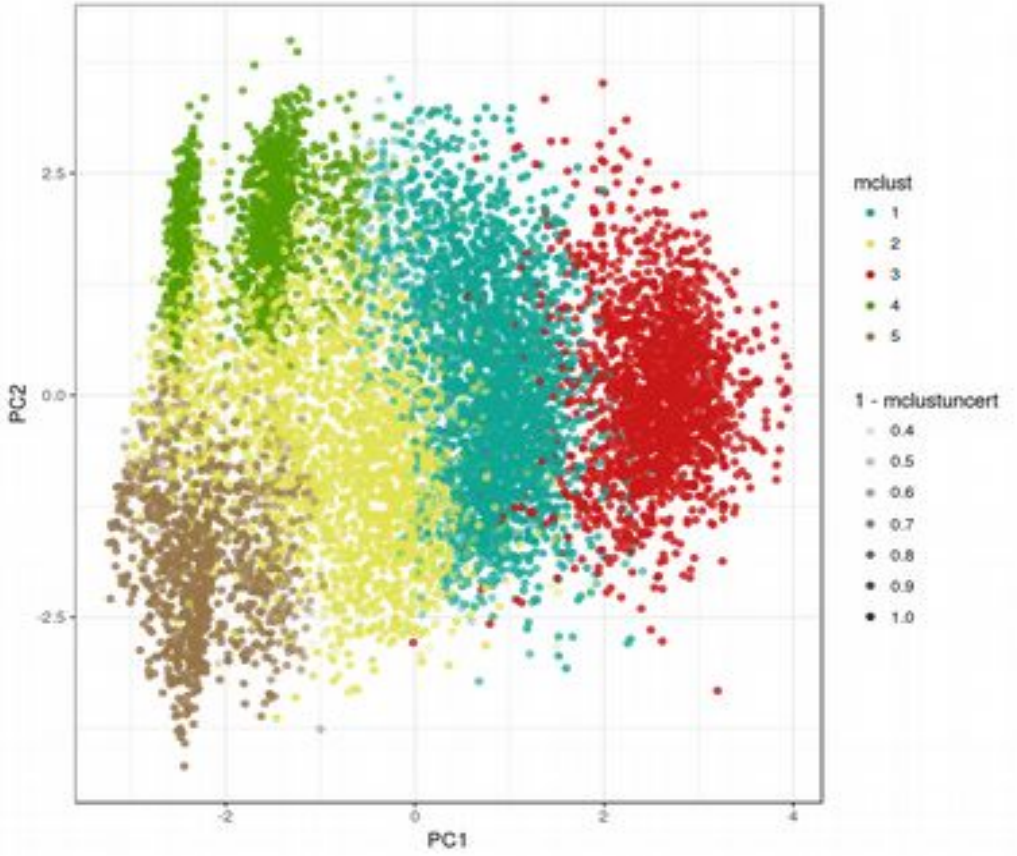
- Jest to proces iteracyjny, składający się z dwóch kroków: E (*expectation*) czyli znalezienia najlepszego rozkładu oraz M (*maximisation*) polegającego na uaktualnieniu parametrów modelu poprzez maksymalizację funkcji wiarygodności (*likelihood*)
- Wybór rozwiązania opiera się na minimalizacji parametru BIC (Bayes inf. criterion), które powinno być najmniejsze

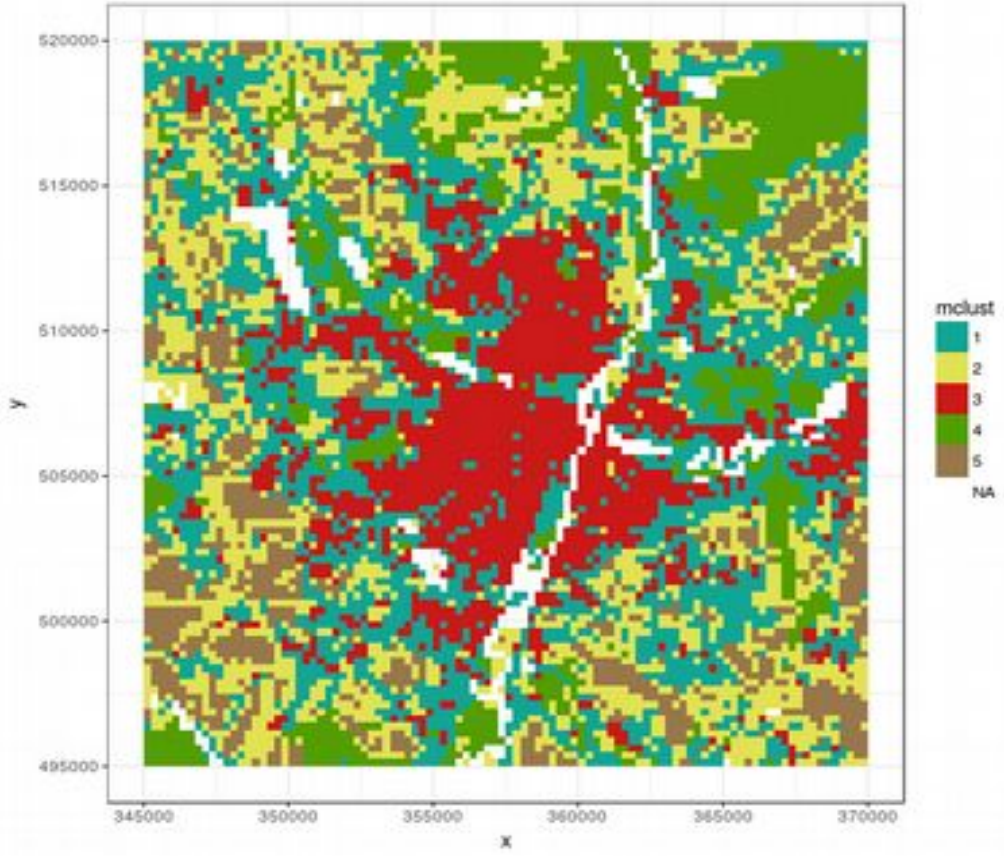


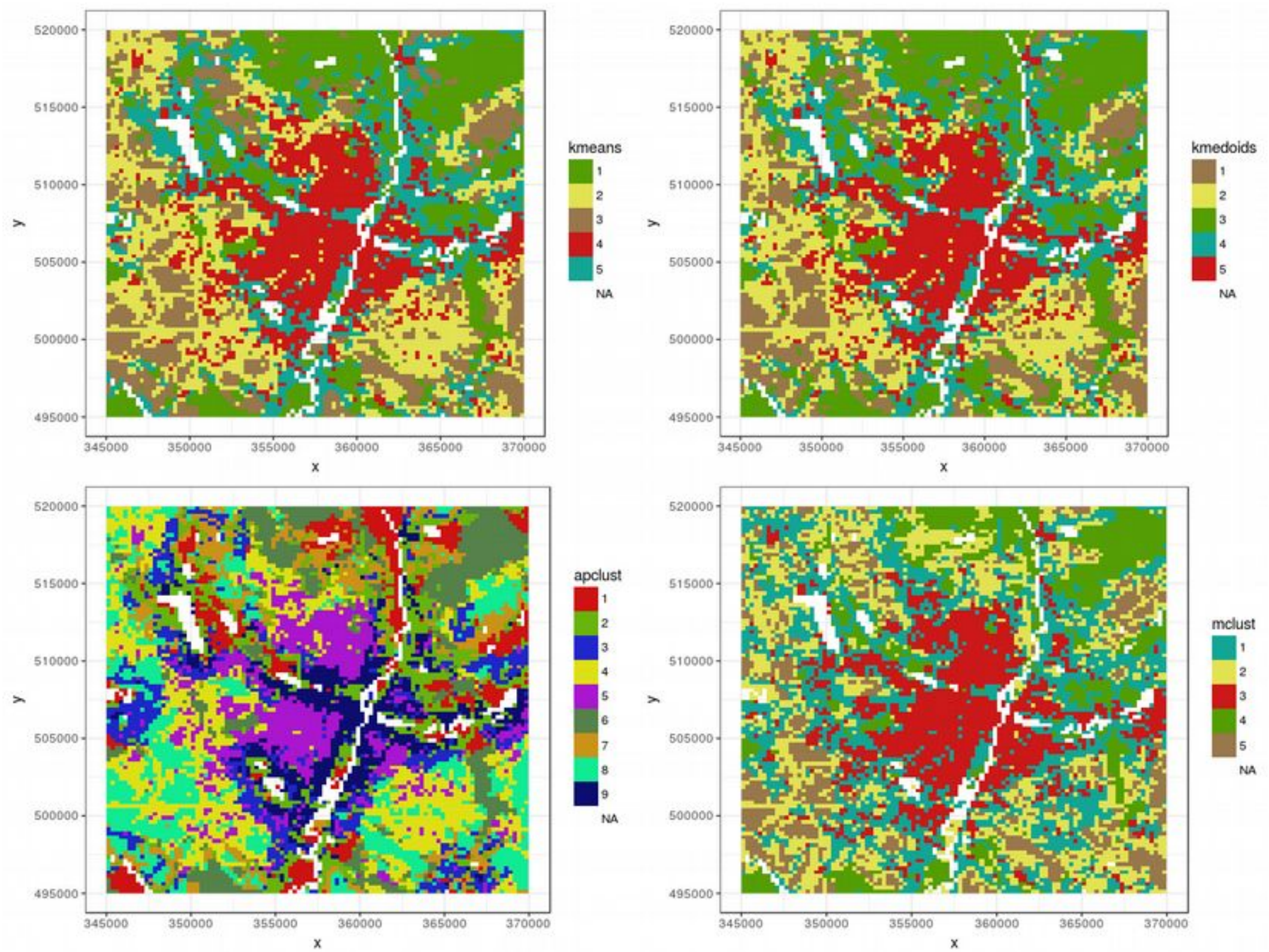


Zastosowanie dla danych geoprzestrzennych

- Zagęszczenia w rozkładzie mają duży wpływ na ostateczny wynik
- Niepewność jest brana pod uwagę
- Znikomy wpływ obiektów odstających
- Bardzo wolny czas obliczeń
- Wynik jest optymalny dla podzbioru danych





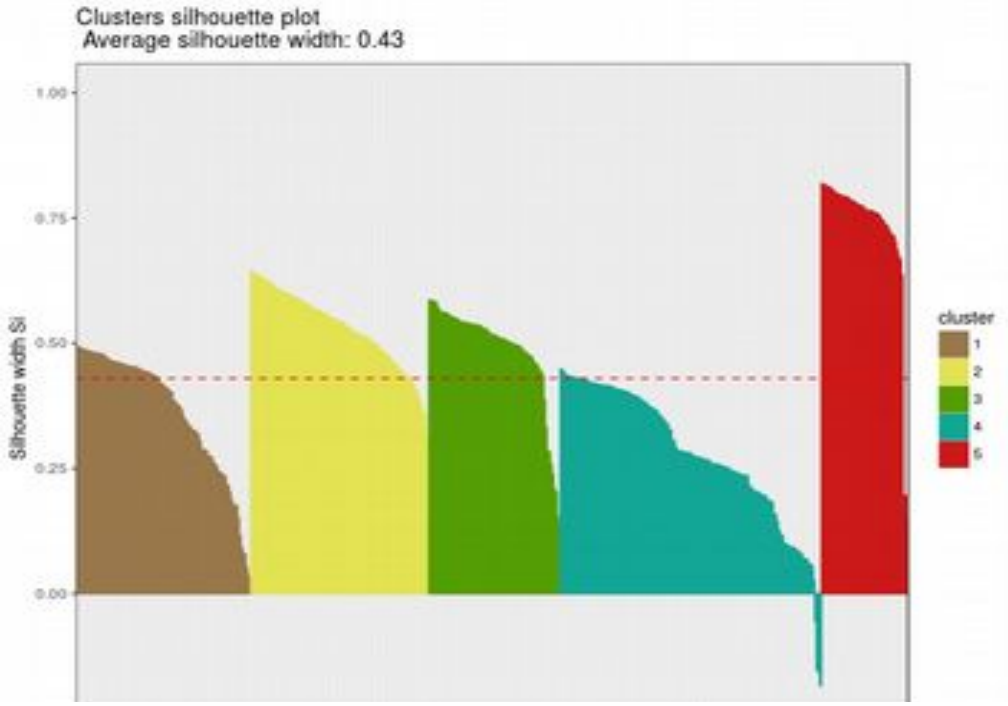

kmeans
1
2
3
4
5
NA
kmedoids
1
2
3
4
5
NA
apclust
1
2
3
4
5
6
7
8
9
NA
mclust
1
2
3
4
5
NA
y
x
520000
515000
510000
505000
500000
495000
345000
350000
355000
360000
365000
370000

Parametry skupień

- Ocena jakości skupień służy do określenia na ile prawidłowo dobrano liczbę skupień oraz do jakiego stopnia obiekty zostały zakwalifikowane do właściwych skupień
- Najpobularnijesze wskaźniki to:
  - **Zwartość** (*compactness*) - jak podobne względem siebie są obiekty w tym samym skupieniu – wzajemne średnie/maksymalne niepodobieństwo pomiędzy obiektami
  - **Oddzielność** (*separation*) – jak niepodobne są obiekty w różnych skupieniach: wzajemne minimalne/średnie niepodobieństwo obiektów w różnych skupieniach
  - **Łączność** (*conectivity*) – do jakiego stopnia obiekty położone blisko siebie znajdują się w tych samych skupieniach

Diagramy sylwetkowe

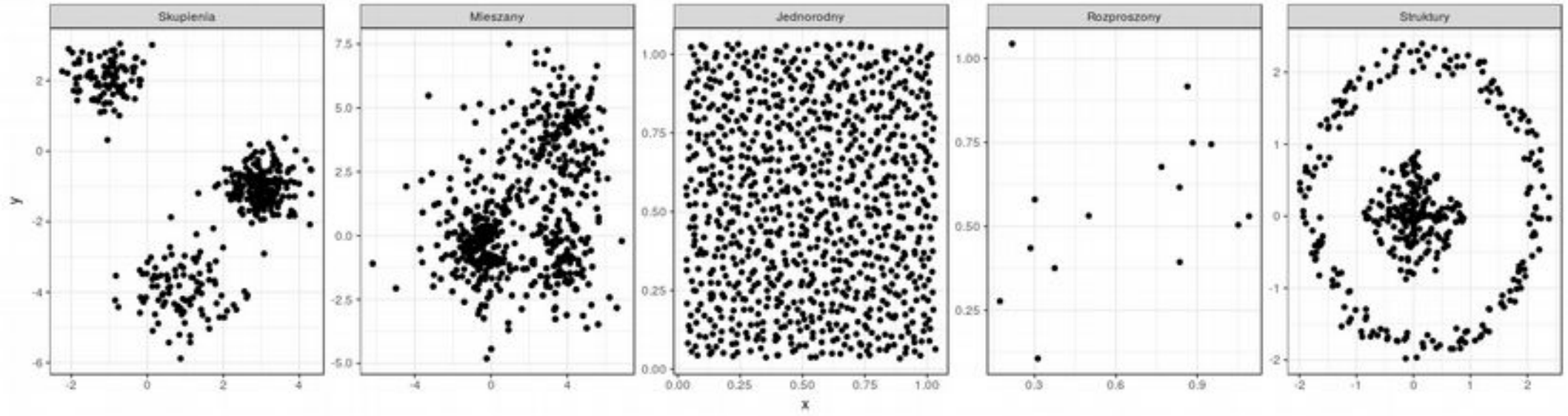
- **Sylwetki** – ocena w jak bardzo obiekty w skupieniu są podobne do pozostałych obiektów w skupieniu względem obiektów w innym (najbardziej podobnym) skupieniu, im większa wartość parametry sylwetki tym lepsze skupienia



- S – bliskie 1; dobre skupienie
- S – bliskie 0; przynależność niejasna
- S < 0 błędna przynależność, zmiana przynależności podniesie jakość skupień

Wybór algorytmu grupowania

- Brak jednolitej reguły



K-medoids
K-means
Ward

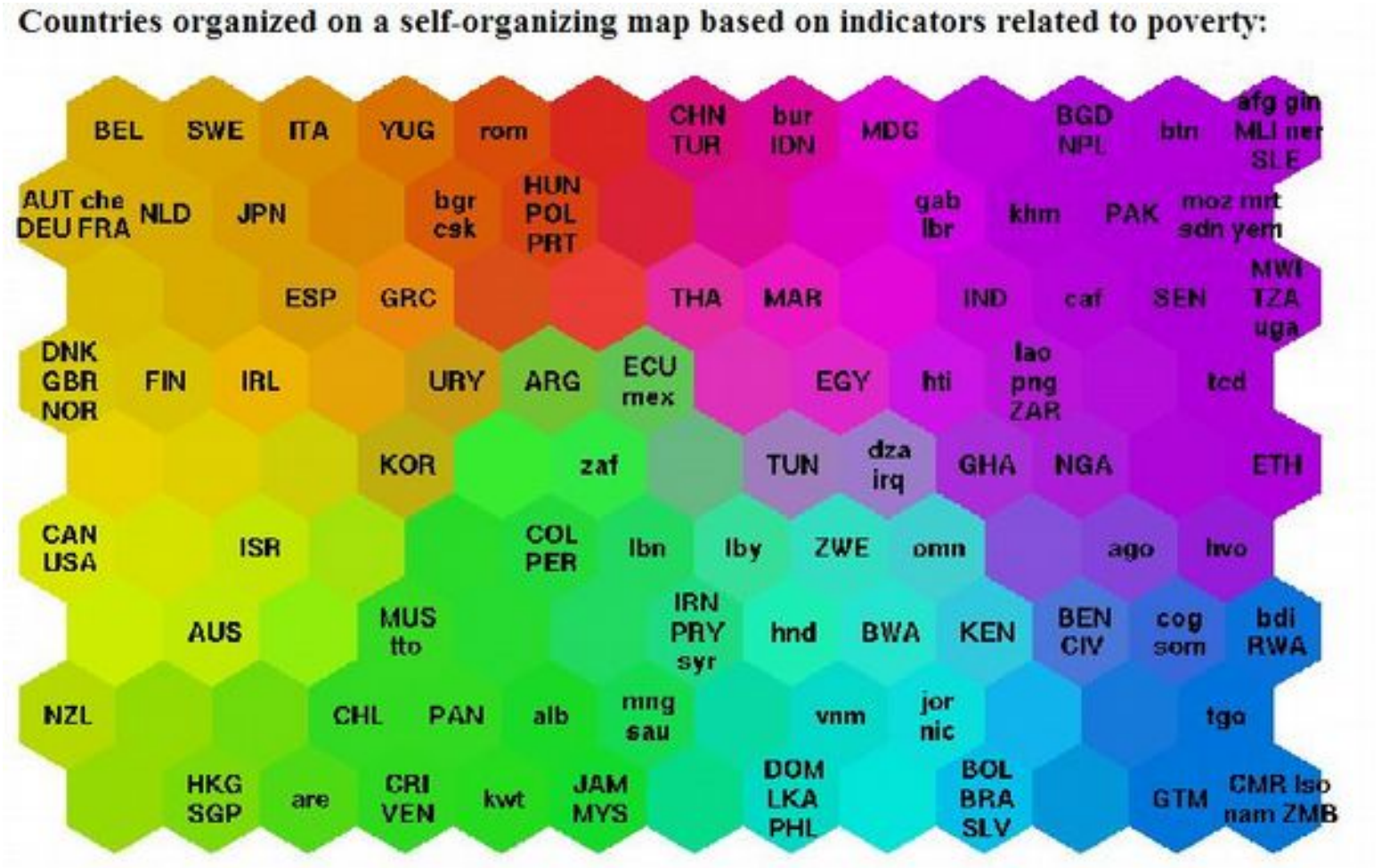
GMM

AP

Complete linkage

dbscan

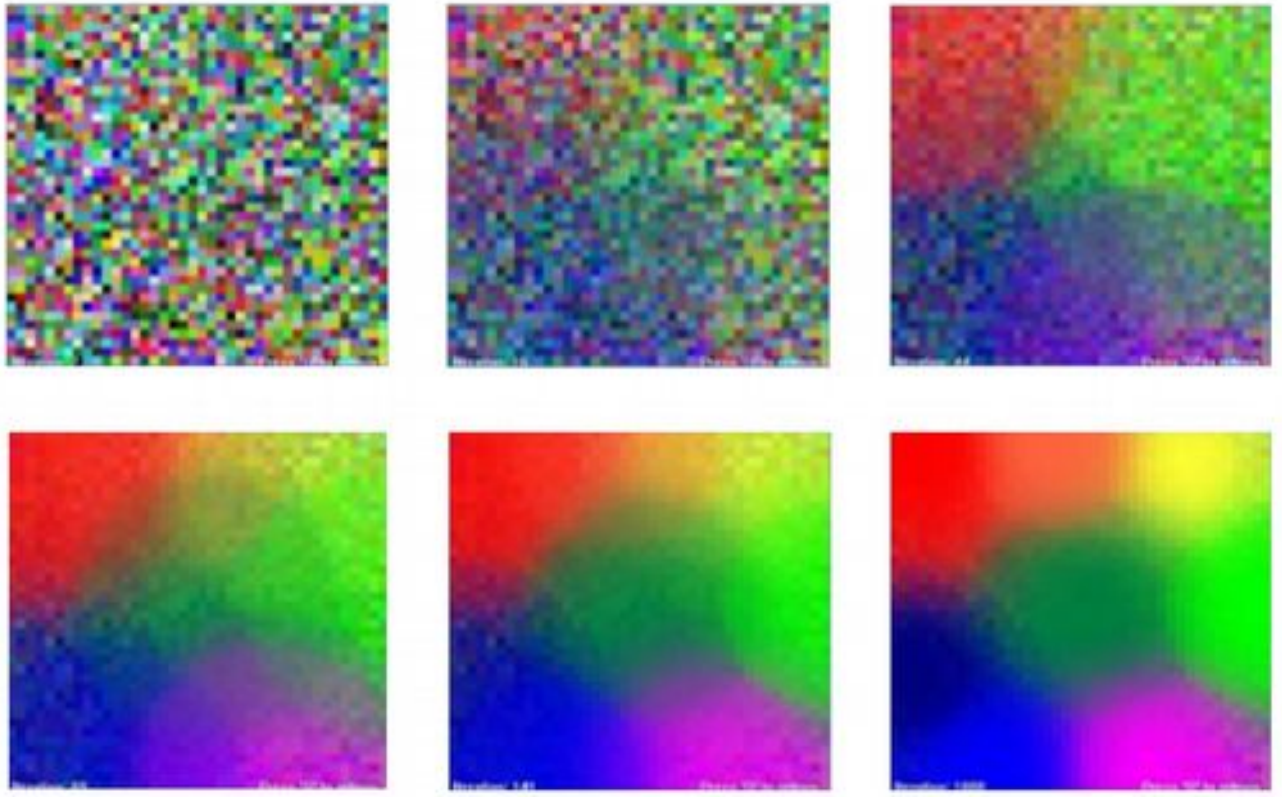
Samo-organizujące się mapy

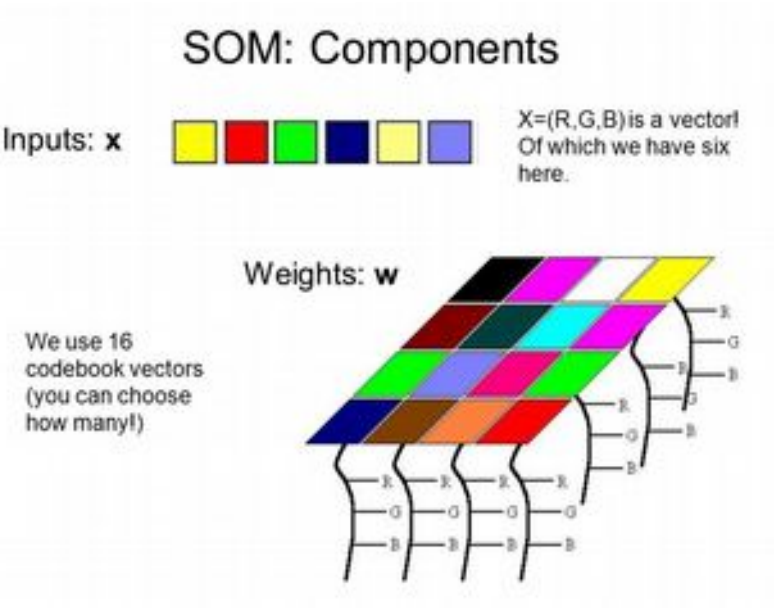


Koncepcja SOM

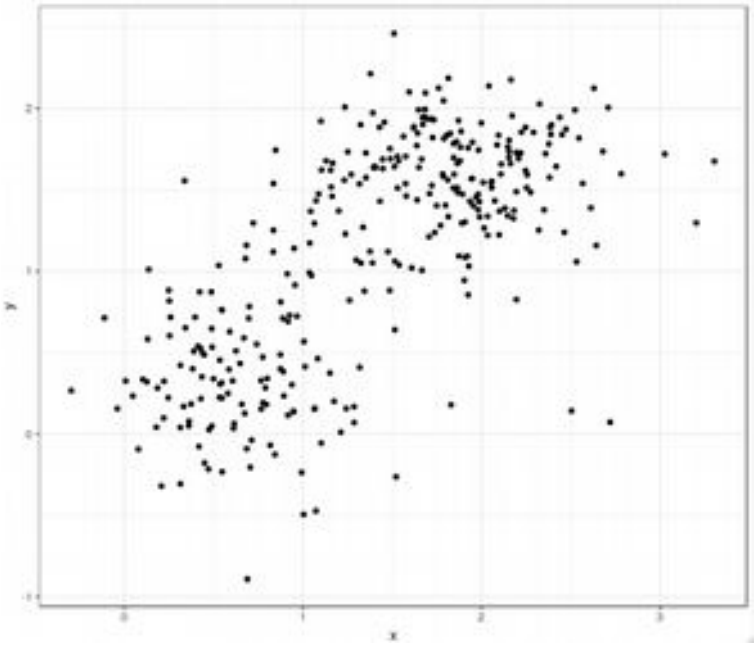
- Sieć neuronowa, narzędzie wizualizacji danych wielowymiarowych w postaci mapy topologicznej
- Metoda nienadzorowana, nie wymaga wzorców (w przeciwieństwie do klasycznych sieci neuronowych)
- Polega na obliczaniu odległości pomiędzy wektorami docelowymi a próbkami i przypisywaniu próbek do wektorów docelowych, jednocześnie je zmieniając. Stąd pojęcie „samoorganizujące”

Wektor kodowy

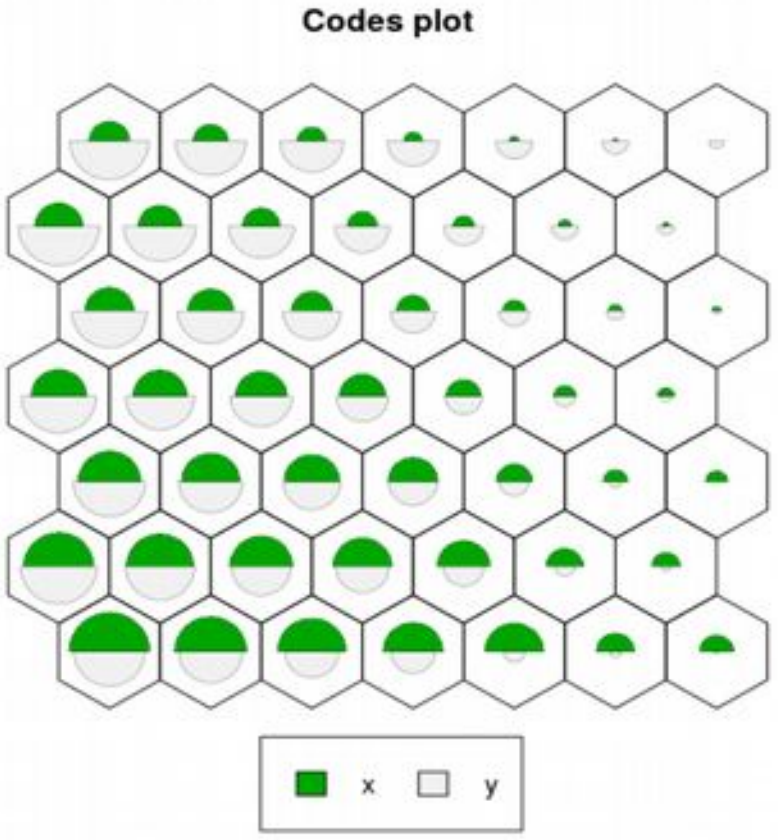


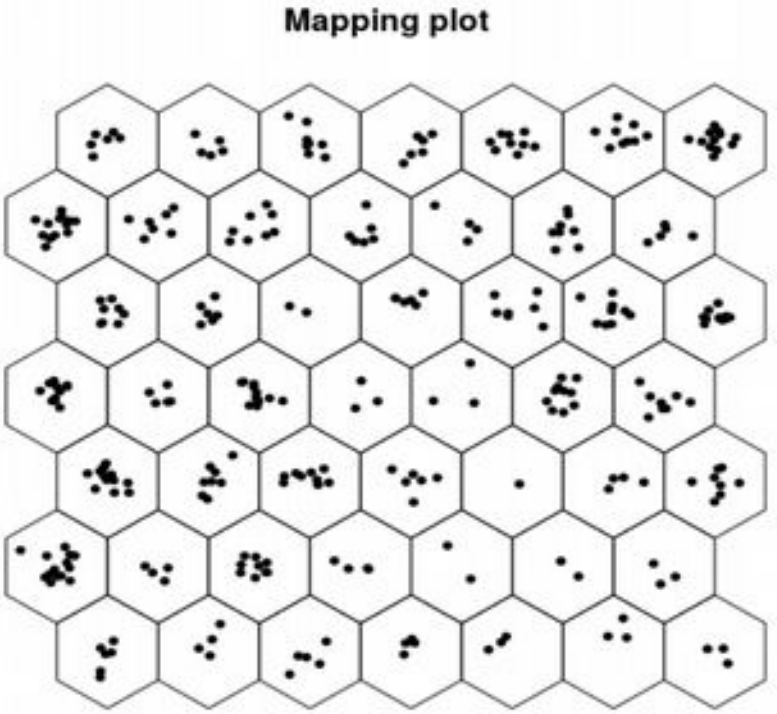


Porządkowanie kolorów RGB

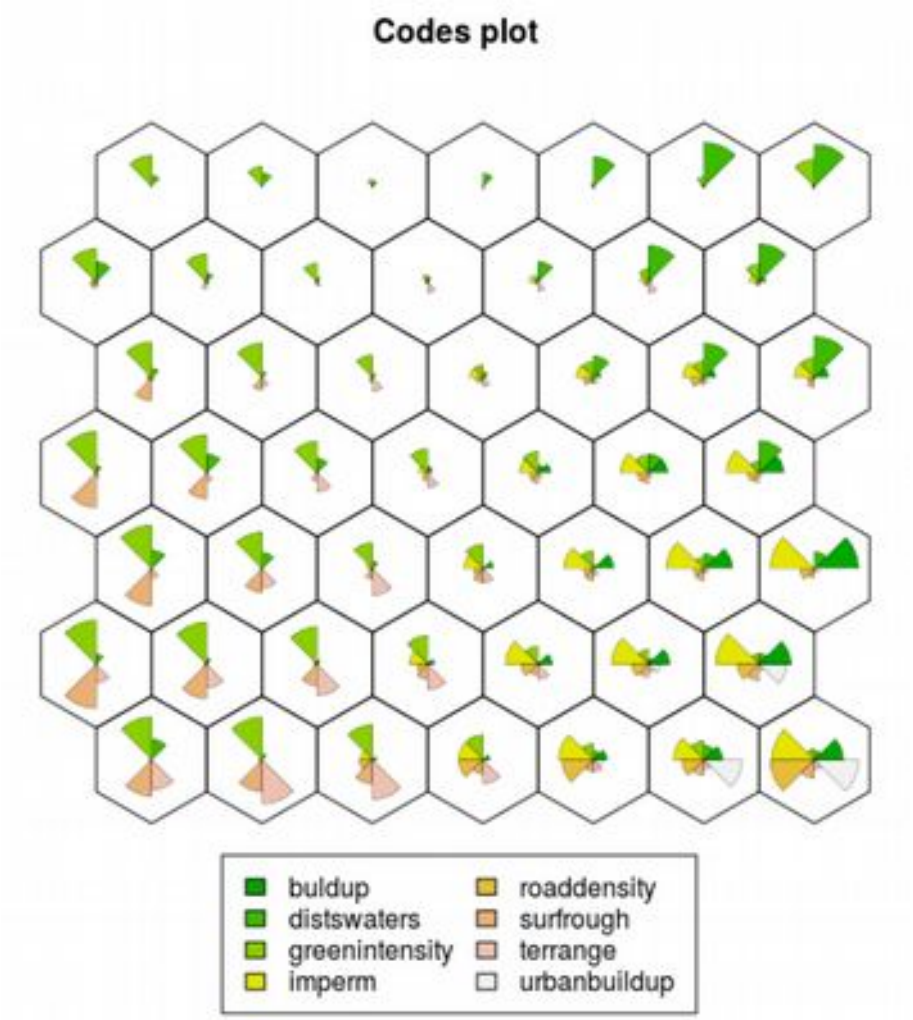


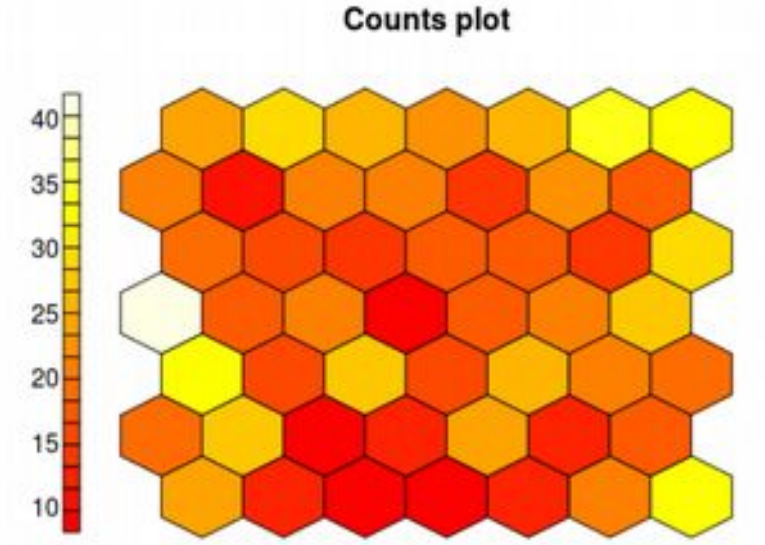
SOM jako narzędzie redukcji wymiarowości

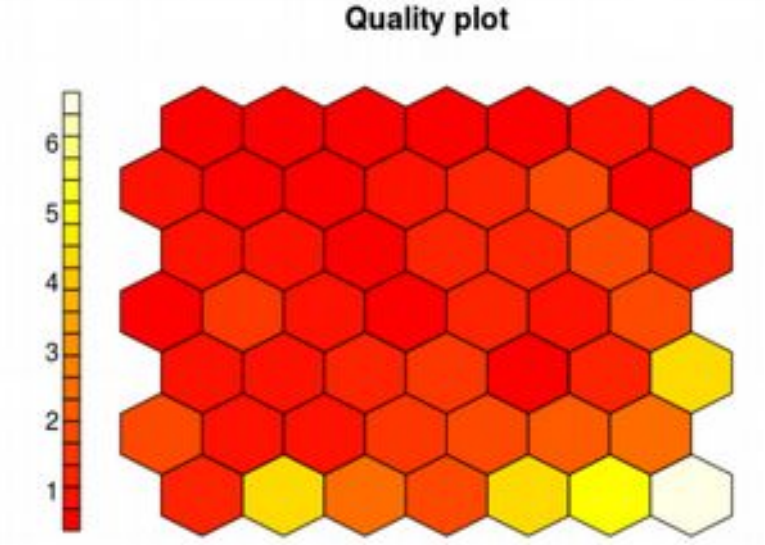




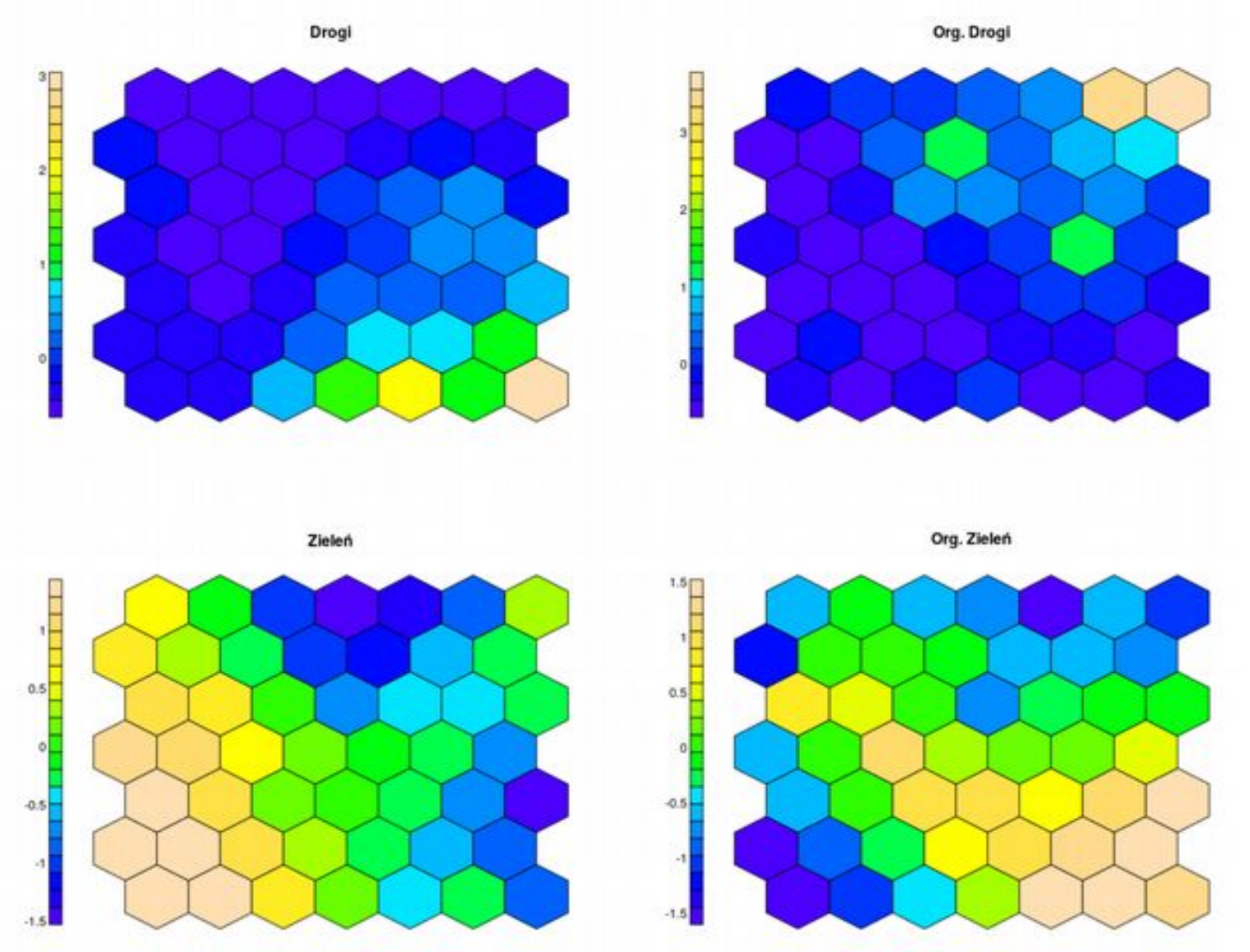
Dane miejskie



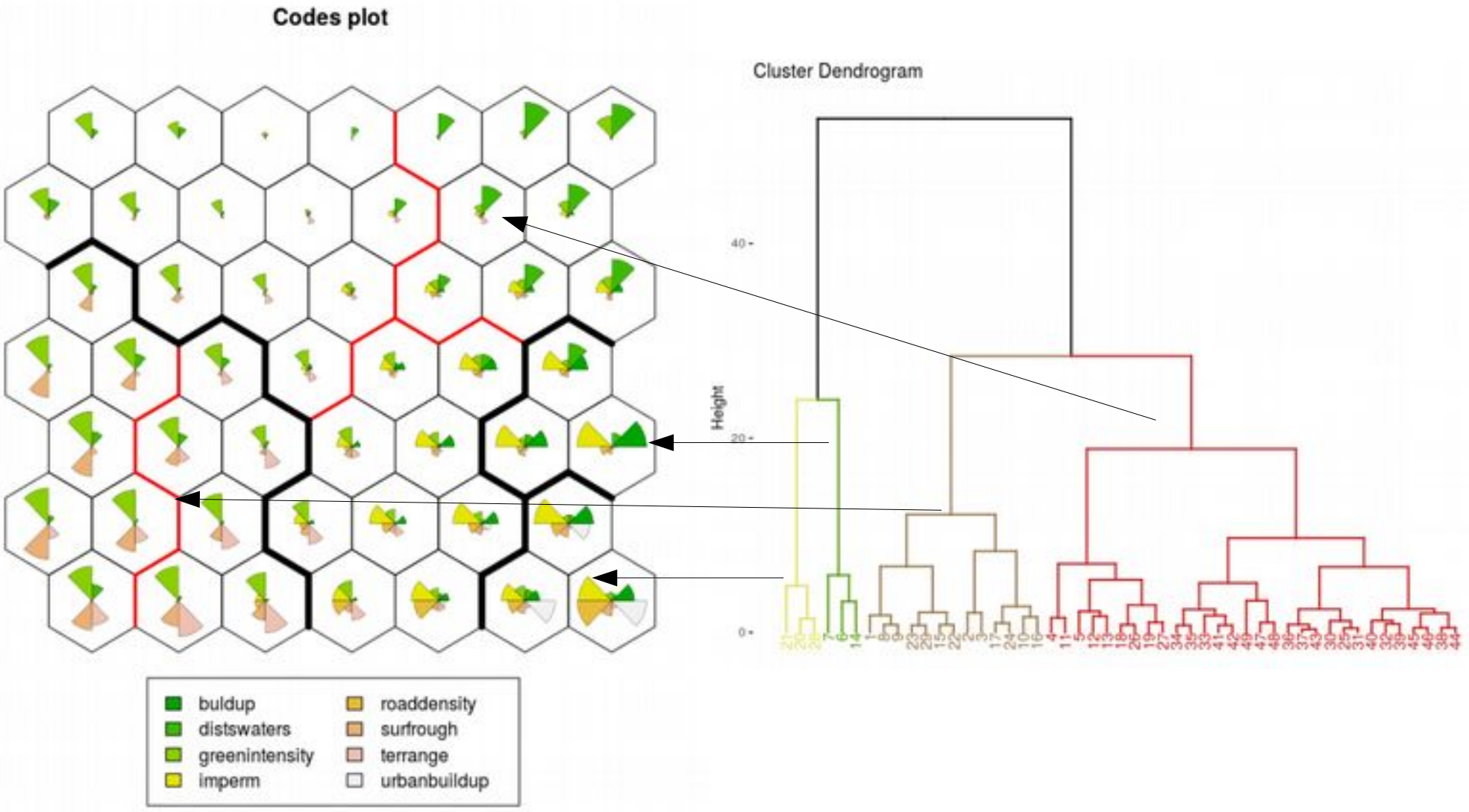




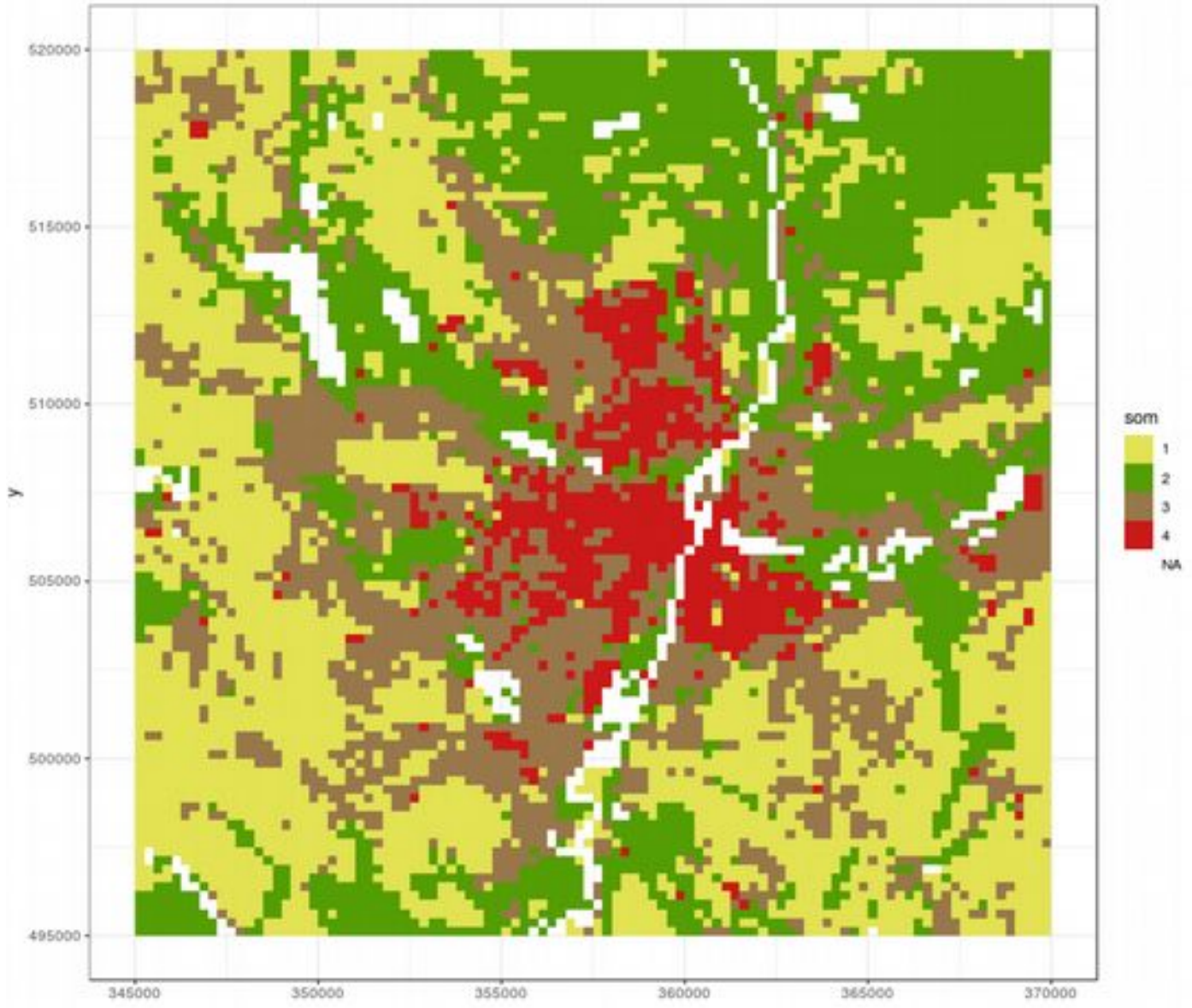
Codebook a dane oryginalne



SOM w klasyfikacjach nienadzorowanych



SOM a dane geoprzestrzenne



Grupowanie a klasyfikacja

- Grupowanie jest procesem budowania optymalnych skupień, proces klasyfikacji to nadawanie skupieniom znaczenia (*labeling*) *a posteriori*
- Optymalne skupienia nie muszą odpowiadać optymalnym klasom – klasy są pochodnymi badanego problemu: np. zróżnicowanie pokrywy roślinnej jest statystyczne większe; ale mniej istotne punktu widzenia człowieka niż np. zróżnicowanie pokrycia zabudowy
- W przypadku klasyfikacji danych uporządkowanych (np. geoprzestrzennych) do interpretacji klasy ma znaczenie nie tylko charakterystyka obiektów ale również ich położenie

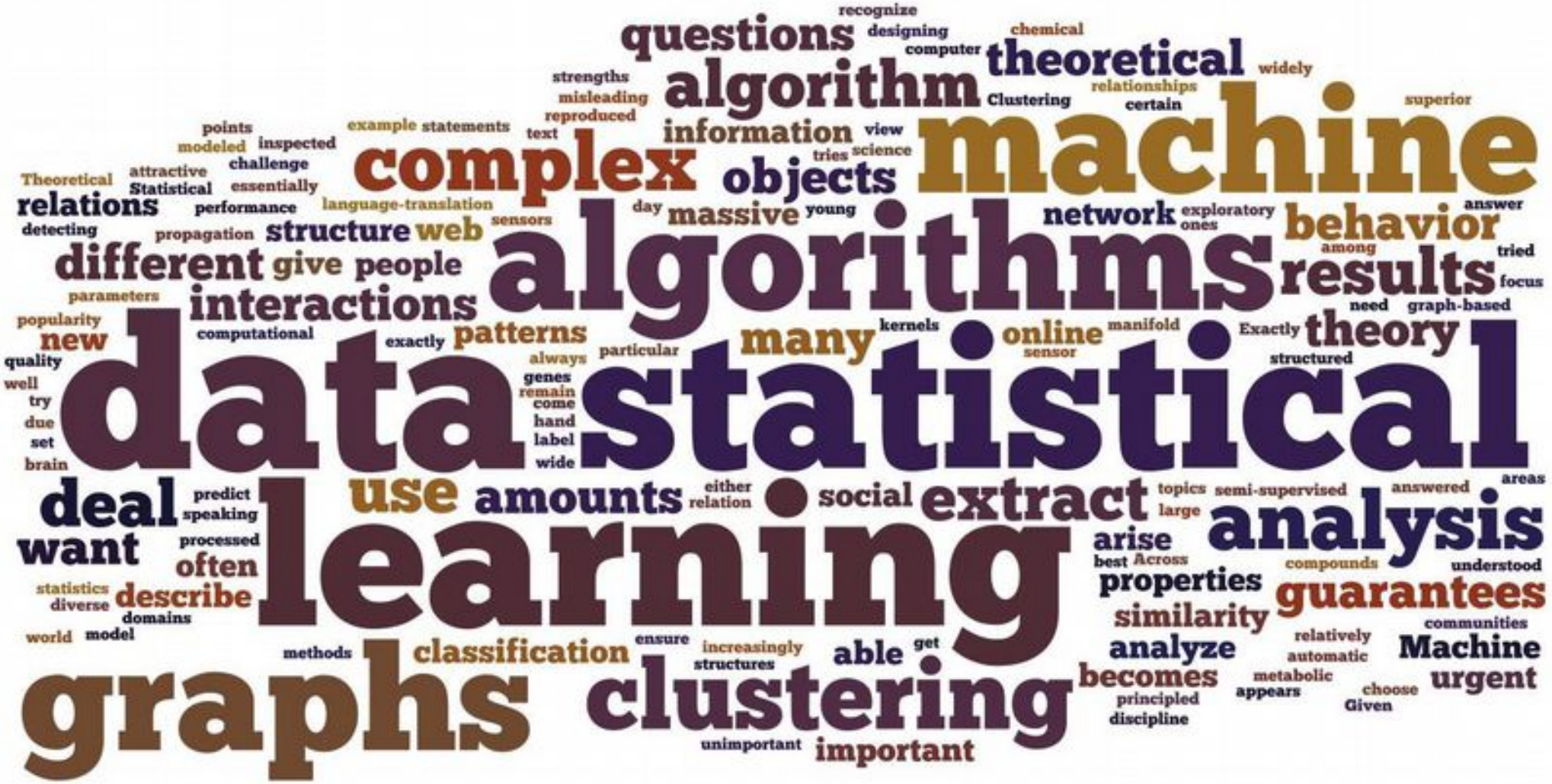
Nienadzorowana klasyfikacja bardzo dużych zbiorów danych

- Klasyfikacje nienadzorowane oparte o macierze niepodobieństwa działają dobrze na małych zbiorach danych (do 10 tyś obiektów)
- W dużych zbiorach proces klasyfikacji wykonujemy na podzbiorze, wyznaczając prototypy skupień a następnie pozostałe obiekty przypisywane są do zdefiniowanych prototypów
- W przypadku danych uporządkowanych (rastrowych) musimy dodatkowo usuwać wartości puste a następnie odtwarzać oryginalną uporządkowaną strukturę danych

	buildup	distswaters	greenintensity	imperm	roaddensity	surfrough	terrange	urbanbuildup	comp
1	0.0804	754.5159	0.3670942	0.3016	12682.1949	4.143160	7.449997	0.0199005	TRUE
2	0.0458	524.9283	0.3432195	0.3088	6039.1346	2.523248	11.080002	0.0000000	TRUE
3	0.0048	320.8700	0.3519326	0.0156	4251.2625	1.703604	12.139999	0.0000000	TRUE
4	0.0000	155.4437	0.5957925	0.0000	2182.2411	7.072344	4.520004	0.0000000	TRUE
5	0.0000	177.4489	0.5278172	0.0000	0.0000	1.589680	3.200005	0.0000000	TRUE
6	0.0884	710.9381	0.2966817	0.3436	9073.5875	4.032760	9.320000	0.0000000	TRUE
7	0.0932	462.3451	0.3695422	0.3268	9031.7141	4.124976	10.270004	0.0000000	TRUE
8	0.0648	215.7878	0.3925129	0.2836	1765.1961	3.178504	15.760002	0.0000000	TRUE
9	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	FALSE
10	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	FALSE
11	0.0204	700.7724	0.2787283	0.0556	328.3053	2.244912	5.559998	0.0000000	TRUE
12	0.1100	451.7021	0.3182577	0.4840	22840.3585	4.579864	8.089996	0.0000000	TRUE
13	0.0292	203.5503	0.3793118	0.2780	7842.2505	3.758920	14.220001	0.1506849	TRUE
14	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	FALSE
15	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	FALSE
16	0.0580	737.4005	0.3745020	0.2148	5880.5543	3.232536	3.500000	0.0000000	TRUE
17	0.0996	499.1948	0.3487019	0.3996	18703.4879	4.131880	11.220001	0.0000000	TRUE
18	0.0476	267.9737	0.4592403	0.3268	4704.3437	5.344536	9.580002	0.0000000	TRUE
19	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	FALSE
20	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	FALSE

Najczęstsze błędy w klasyfikacjach nienadzorowanych

- Brak transformacji danych (standaryzacji/normalizacji)
- Nieodpowiednie miary niepodobieństwa (nadużywanie metryki euklidesowej)
- Brak redukcji wymiarów i obiektów odstających
- Wymuszanie skupień w jednorodnych danych
- Stosowanie metod hierarchicznych do dużych zbiorów danych
- Sugerowanie się klasami *a priori* (zamiast metod nadzorowanych)


recognize
questions
designing
chemical
computer
theoretical
widely
strengths
misleading
reproduced
algorithm
Clustering
relationships
certain
superior
machine
points
example
statements
text
information
view
modeled
inspected
tries
science
challenge
complex
objects
Theoretical
attractive
Statistical
essentially
relations
performance
language-translation
day
massive
young
network
exploratory
ones
answer
behavior
detecting
propagation
structure
web
sensors
among
tried
different
give
people
algorithms
results
focus
parameters
interactions
need
graph-based
popularity
new
computational
exactly
patterns
many
kernels
online
manifold
Exactly
theory
quality
always
particular
sensor
structured
well
genes
data
statistical
try
remain
come
due
hand
set
label
brain
wide
areas
deal
predict
use
amounts
either
relation
social
extract
topics
semi-supervised
answered
speaking
large
analysis
want
processed
arise
often
best
Across
compounds
understood
properties
guarantees
statistics
diverse
describe
similarity
communities
domains
relatively
world
model
ensure
increasingly
analyze
Machine
methods
classification
able
get
structures
automatic
becomes
metabolic
urgent
graphs
clustering
principled
appears
choose
Given
discipline
unimportant
important