w wykładzie wykorzystano:

1. materiały dydaktyczne przygotowane w ramach projektu
   Opracowanie programów nauczania na odległość na kierunku studiów wyższych – Informatyka
   http://wazniak.mimuw.edu.pl
2. *Internetowy Podręcznik Statystyki*
   http://www.statsoft.pl/textbook/stathome.html

Analiza Skupień

Cluster analysis

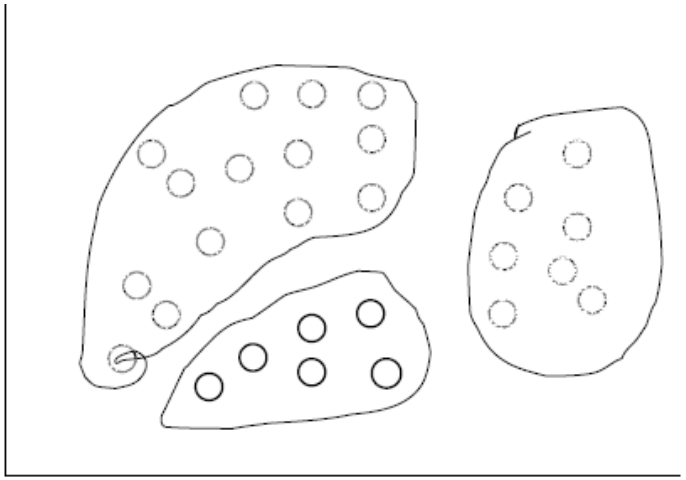
Krzysztof Regulski,
WIMiIP, KISiM,
regulski@agh.edu.pl
B5, pok. 408

Rodzaje modeli:

- **metoda k-średnich (iteracyjno-optymalizacyjne),**
- metody hierarchiczne,
- sieci Kohonena,
- **grupowanie probabilistyczne - algorytm EM**
- algorytm BIRCH,
- grupowanie oparte na gęstości

Grupowanie

Znajdź *„naturalne"* pogrupowanie obiektów w oparciu o ich wartości



zastosowania grupowania:

» grupowanie dokumentów

» grupowanie klientów

» segmentacja rynku

» Kompresja obrazów

» Odkrywanie oszustw

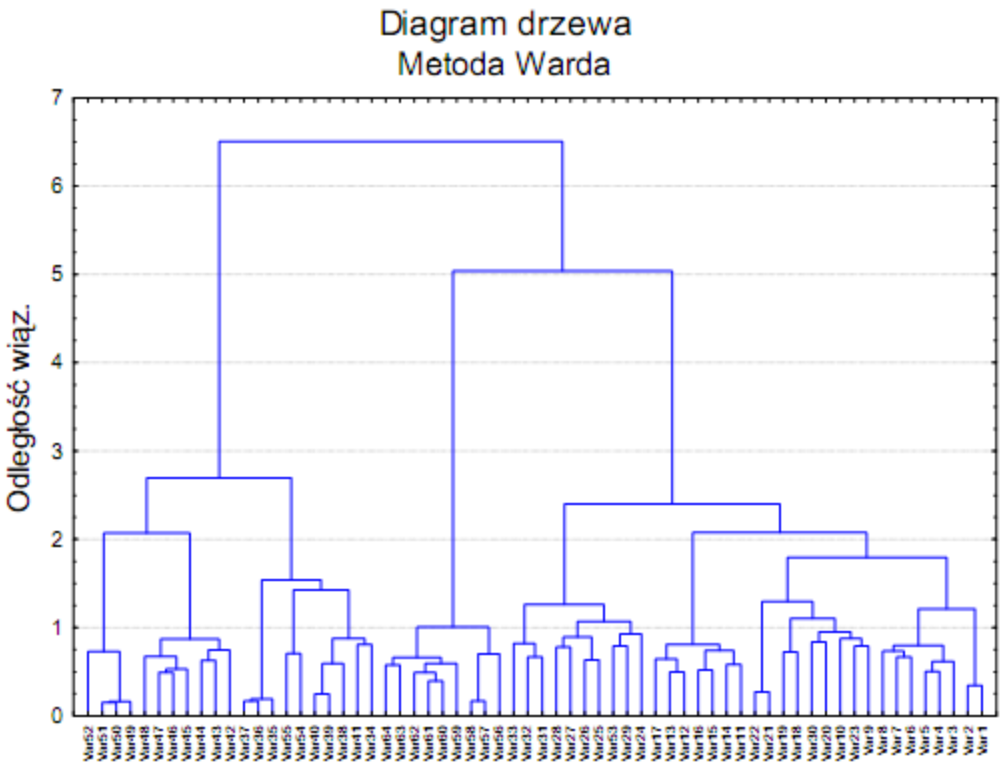
Grupowanie

- *Grupowanie (klastrowanie)* - obejmuje metody znajdowania skończonych zbiorów klastrów obiektów posiadających podobne cechy.
- W przeciwieństwie do metod klasyfikacji i predykcji, podział na klastry nie jest znany *a-priori*, lecz jest celem metod grupowania.
- Metody te grupują obiekty w klasy w taki sposób, aby *maksymalizować podobieństwo wewnątrzklasowe obiektów i minimalizować podobieństwo pomiędzy klasami obiektów*.
- W przeciwieństwie do klasyfikacji wzorcowej (analizy z nauczycielem), polegającej na przyporządkowywaniu przypadków do jednej z określonych klas, tu klasy nie są znane ani w żaden sposób scharakteryzowane przed przystąpieniem do analizy.

Sformułowanie problemu

– Grupowanie może dotyczyć zarówno obiektów rzeczywistych (np. pacjentów, sekwencji DNA, dokumenty tekstowe), jak również obiektów abstrakcyjnych (sekwencja dostępów do stron WWW, grafy reprezentujące dokumenty XML, itp.).

– Grupowanie jest jedną z najstarszych i najbardziej popularnych metod eksploracji danych (1939).



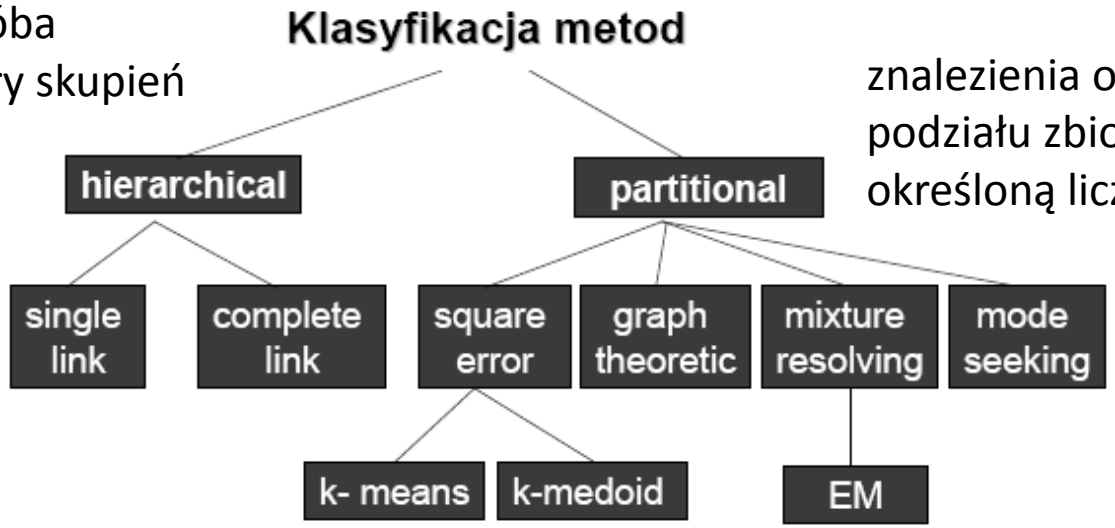
Czym jest klaster?

- Zbiór obiektów, które są „podobne”.
- Zbiór obiektów, takich, że odległość pomiędzy dwoma dowolnymi obiektami należącymi do klastra jest mniejsza aniżeli odległość pomiędzy dowolnym obiektem należącym do klastra i dowolnym obiektem nie należącym do tego klastra.
- Spójny obszar przestrzeni wielowymiarowej, charakteryzujący się dużą gęstością występowania obiektów.

Klasyfikacja metod

hierarchiczna próba odkrycia struktury skupień

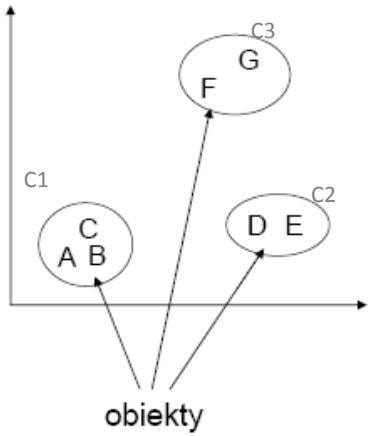
znalezienia optymalnego podziału zbioru przykładów na określoną liczbę skupień

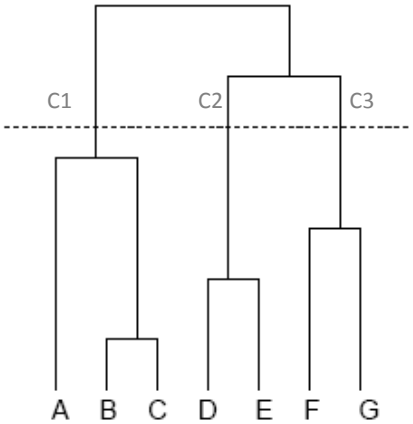


- Pierwsza grupa algorytmów konstruuje klastry sekwencyjnie wykorzystując cechy obiektów, druga konstruuje klastry wykorzystując jednocześnie wszystkie cechy (atrybuty) obiektów.
- Metody hierarchiczne generują zagnieżdżoną sekwencję podziałów zbiorów obiektów w procesie grupowania
- Metody z iteracyjno-optymalizacyjne generują tylko jeden podział (partycję) zbioru obiektów w dowolnym momencie procesu grupowania
- Algorytmy oparte na gęstościach (ang. *density-based algorithms*), które dzielą zbiory przykładów korzystając z modelu probabilistycznego dla bazowych skupień.

Metody grupowania hierarchicznego

Metody grupowania hierarchicznego

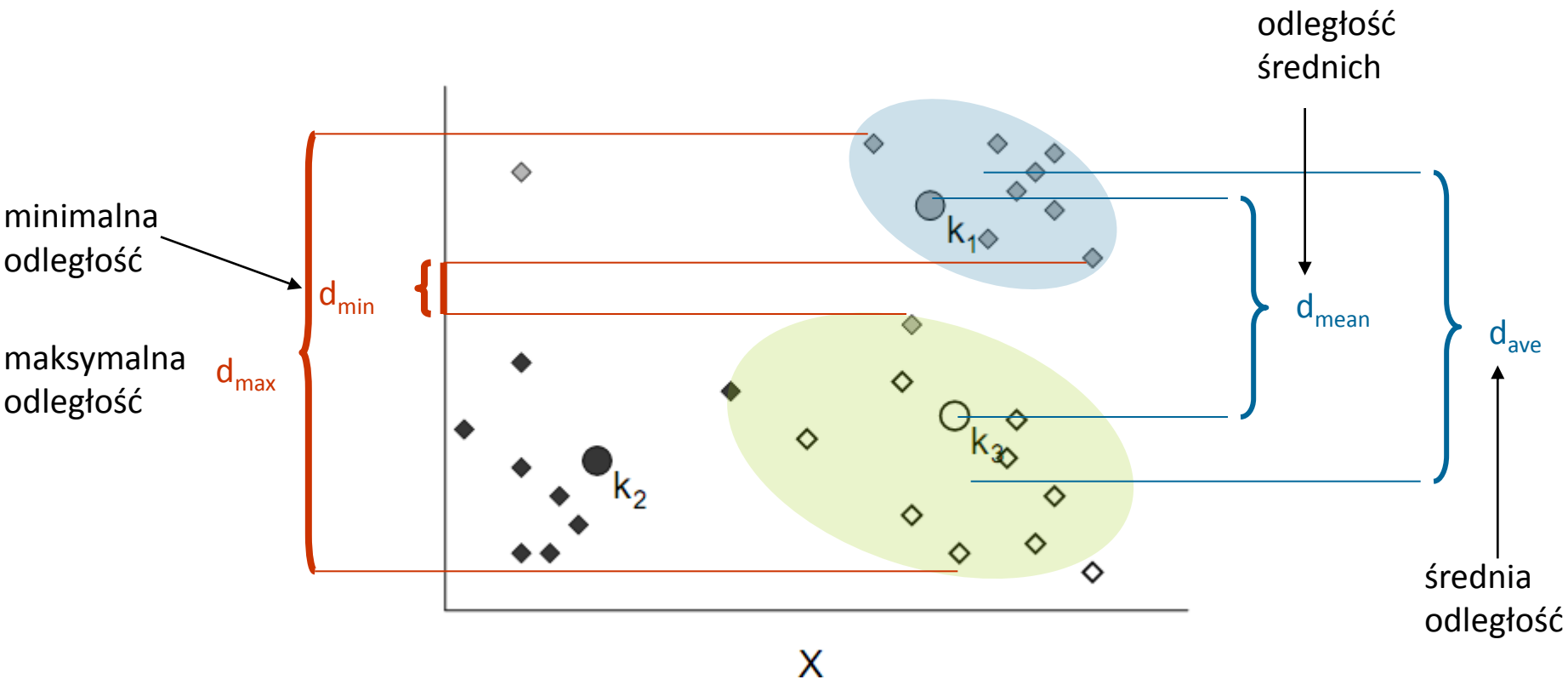




dendrogram

- Metoda grupowania hierarchicznego polega na sekwencyjnym grupowaniu obiektów - drzewo klastrów (tzw. *dendrogram*)
- Początkowo, wszystkie obiekty A, B, ... G należą do osobnych klastrów. Następnie, w kolejnych krokach, klastry są łączone w większe klastry (łączymy B i C, D i E, oraz F i G, następnie, A łączymy z klastrem zawierającym obiekty B i C, itd.).
- Proces łączenia klastrów jest kontynuowany tak długo, aż liczba uzyskanych klastrów nie osiągnie zadanej liczby klastrów. Graficznie, na dendrogramie, warunek stopu (tj. zadana liczba klastrów) przedstawia linia pozioma przecinająca dendrogram.
- C1={A, B, C}, C2={D, E} oraz C3={F, G}.

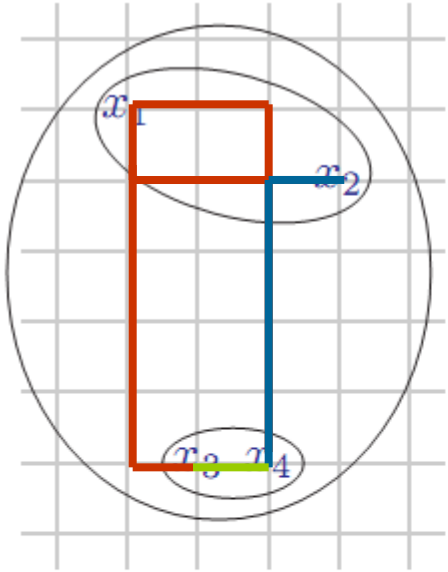
Odległość klastrów



Hierarchiczny aglomeracyjny algorytm grupowania

1. Umieść każdy obiekt w osobnym klastrze. Skonstruuj macierz przyległości zawierającą odległości pomiędzy każdą parą klastrów
2. Korzystając z macierzy przyległości znajdź najbliższą parę klastrów. Połącz znalezione klastry tworząc nowy klaster. Uaktualnij macierz przyległości po operacji połączenia
3. Jeżeli wszystkie obiekty należą do jednego klastra, zakończ procedurę grupowania, w przeciwnym razie przejdź do kroku 2

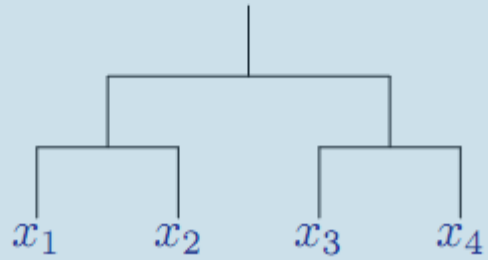
Algorytm aglomeracyjnej analizy skupień



Odległości taksówkowe, najbliższe sąsiedztwo:

$\{x_2\}$	$\{x_3\}$	$\{x_4\}$	$d(_,_)$
4	6	7	$\{x_1\}$
	6	5	$\{x_2\}$
		1	$\{x_3\}$

macierz przyległości

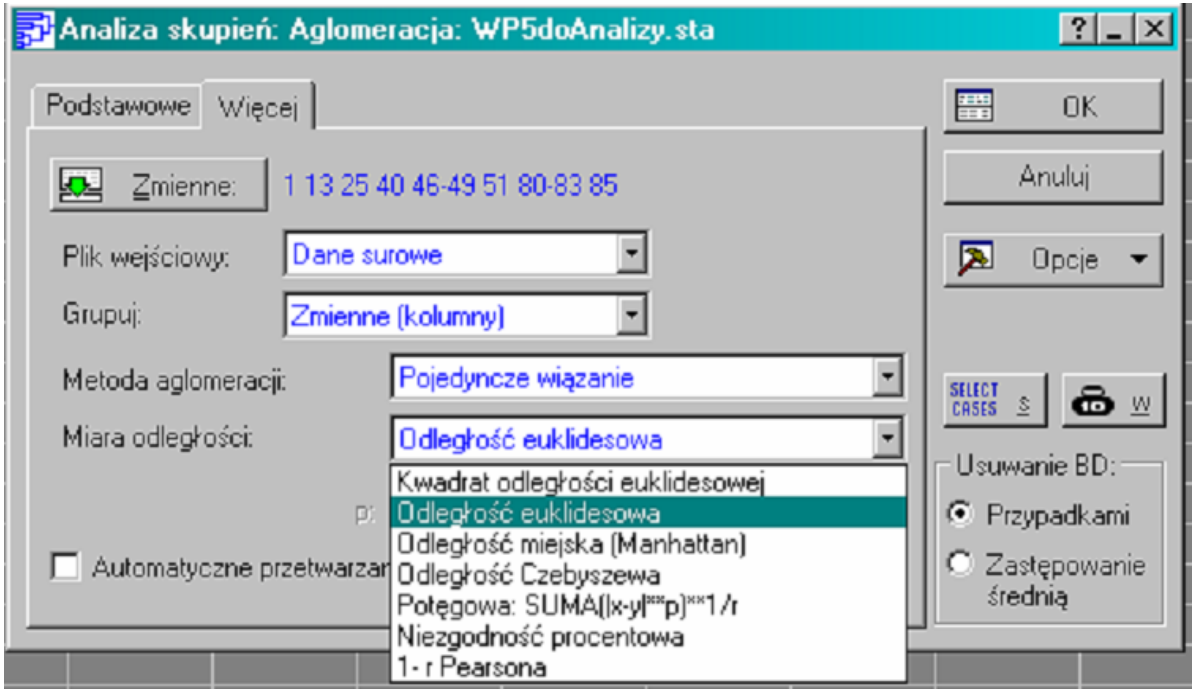


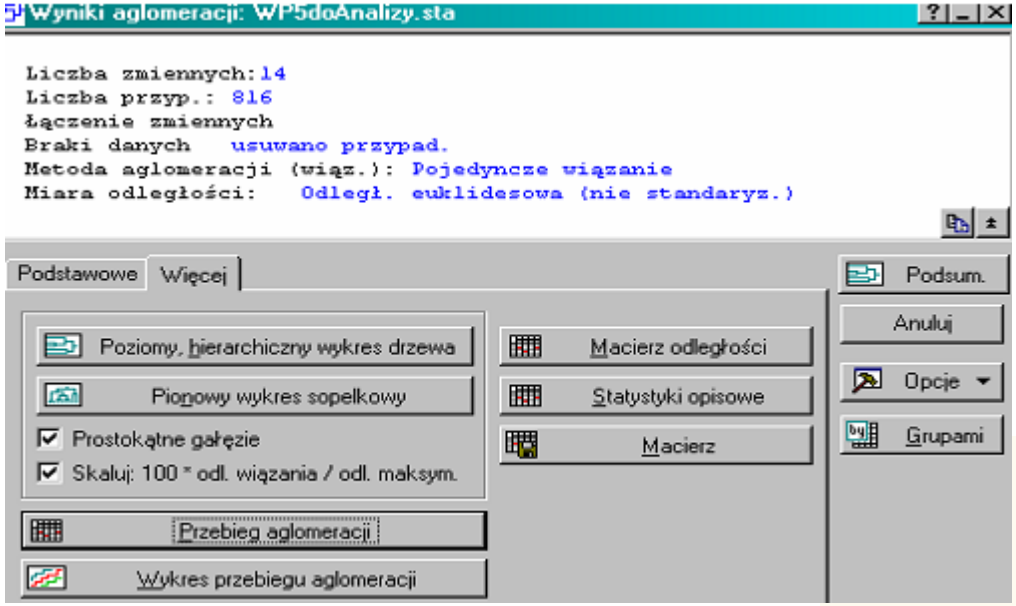
dendrogram

dane dotyczące przebiegu procesu wytopu surówki w wielkim piecu

Korelacje (WP5doAnalizy.sta)
Oznaczone wsp. korelacji są istotne z p < ,05000
N=816 (Braki danych usuwano przypadkami)

Zmienna	Sur_zawSi	Sur_zawMn	CO_zawWGgazieGardzielowym	CO2_zawWGazieGardzielowym	H2_zawWGazieGardzielowym	TemperaturaGazuGardzielowego	TemperaturaSurowki
Spiekllosc	0,03	-0,03	-0,04	-0,01	-0,04	-0,03	0,06
Grulilosc	-0,01	-0,00	0,06	-0,05	0,01	0,02	-0,14
Topnikllosc	-0,05	0,00	-0,05	0,19	-0,05	-0,06	0,07
Koksllosc	-0,12	-0,01	0,01	0,01	0,02	0,00	0,00
DmuchWilgotny_ilosc	0,11	-0,02	-0,09	-0,03	0,03	-0,12	0,07
DmuchWilgotny_wilgotnosc	0,39	-0,07	-0,22	-0,01	0,43	0,21	0,19
DmuchWilgotny_temperatura	-0,29	0,03	0,12	0,02	-0,18	-0,15	-0,00
WilgotnoscPowietrza	-0,20	0,00	0,08	0,11	0,02	-0,00	-0,09




Wyniki aglomeracji: WP5doAnalizy.sta
Liczba zmiennych: 14
Liczba przyp.: 816
Łączenie zmiennych
Braki danych usuwano przypad.
Metoda aglomeracji (wiąz.): Pojedyncze wiązanie
Miara odległości: Odległ. euklidesowa (nie standaryz.)
Podstawowe
Więcej
Podsum.
Anuluj
Opcje
Grupami
Poziomy, hierarchiczny wykres drzewa
Pionowy wykres sopelkowy
Prostokątne gałęzie
Skaluj: 100 * odl. wiązania / odl. maksym.
Macierz odległości
Statystyki opisowe
Macierz
Przebieg aglomeracji
Wykres przebiegu aglomeracji

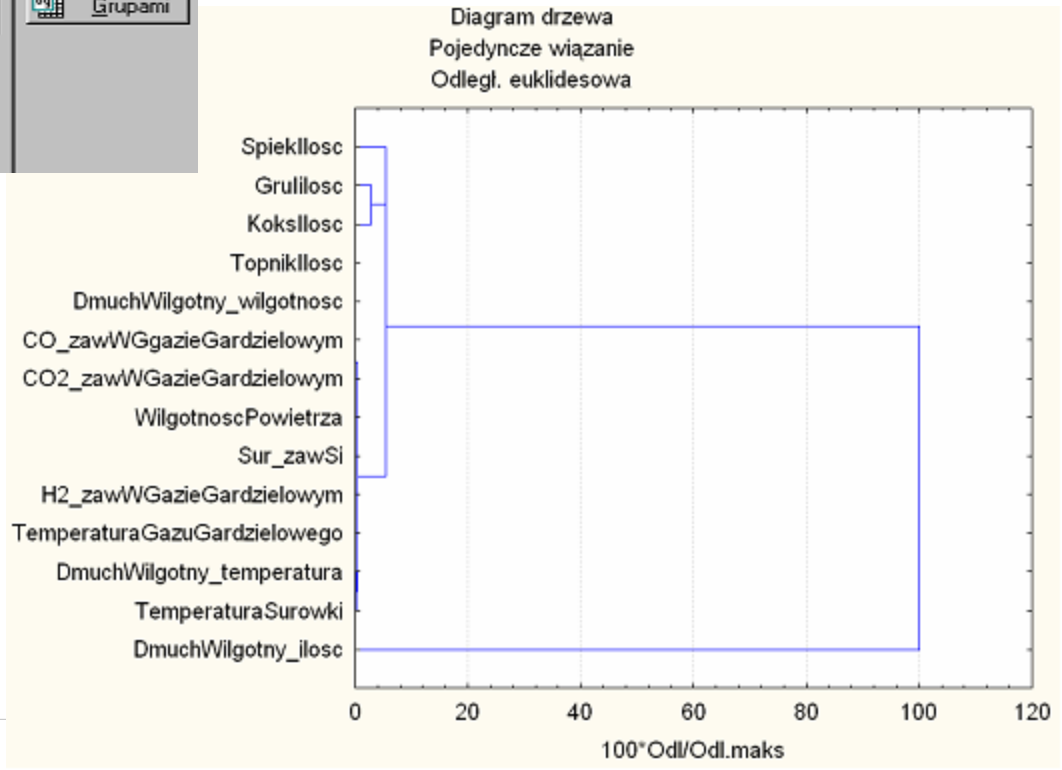


Diagram drzewa
Pojedyncze wiązanie
Odległ. euklidesowa
SpiekIlosc
GruliIosc
KoksIlosc
TopnikIlosc
DmuchWilgotny_wilgotnosc
CO_zawWGgazieGardzielowym
CO2_zawWGazieGardzielowym
WilgotnoscPowietrza
Sur_zawSi
H2_zawWGazieGardzielowym
TemperaturaGazuGardzielowego
DmuchWilgotny_temperatura
TemperaturaSurowki
DmuchWilgotny_ilosc
0
20
40
60
80
100
120
100*Odl/Odl.maks

Macierz odległości pomiędzy łączonymi zmiennymi

Odległ. euklidesowa (WP5doAnalizy.sta)

Zmienna	Spiekilosc	Grulilosc	Topnikilosc	Koksilosc	DmuchWilgotny_ilosc	DmuchWilgotny_wilgotnosc	DmuchWilgotny_temperatura	WilgotnoscPowietrza	Sur_zawSi	CO_zawWGgazieGardzielowym	CO2_zawWGazieGardzielowym
Spiekilosc	0	258292	633134	336187	4705663	634571	607306	634957	635073	634443	634547
Grulilosc	258292	0	411980	129473	4928425	413472	386525	413852	413967	413343	413446
Topnikilosc	633134	411980	0	301228	5332274	4371	26220	4520	4571	4324	4355
Koksilosc	336187	129473	301228	0	5032618	302749	275366	303138	303255	302622	302726
DmuchWilgotny_ilosc	4705663	4928425	5332274	5032618	0	5333784	5306419	5334173	5334290	5333658	5333761
DmuchWilgotny_wilgotnosc	634571	413472	4371	302749	5333784	0	27459	452	555	268	232
DmuchWilgotny_temperatura	607306	386525	26220	275366	5306419	27459	0	27844	27961	27327	27431
WilgotnoscPowietrza	634957	413852	4520	303138	5334173	452	27844	0	119	518	413
Sur_zawSi	635073	413967	4571	303255	5334290	555	27961	119	0	635	530
CO_zawWGgazieGardzielowym	634443	413343	4324	302622	5333658	268	27327	518	635	0	109
CO2_zawWGazieGardzielowym	634547	413446	4355	302726	5333761	232	27431	413	530	109	0
H2_zawWGazieGardzielowym	635063	413957	4567	303244	5334279	543	27950	110	23	624	520
TemperaturaGazuGardzielowego	630640	409669	5178	298799	5329850	4360	23577	4730	4838	4256	4354
TemperaturaSurowki	593287	372692	40221	261290	5292345	41562	14157	41951	42068	41434	41539

- Najmniejsza odległość występuje pomiędzy zawartością krzemu w surówce a zawartością wodoru w gazie gardzielowym
- Istnieje powiązanie pomiędzy zawartością CO i CO_2 w gazie gardzielowym
- Następne grupy tworzą wilgotność powietrza z Si w surówce i H_2 w gazie gardzielowym.

…mediany

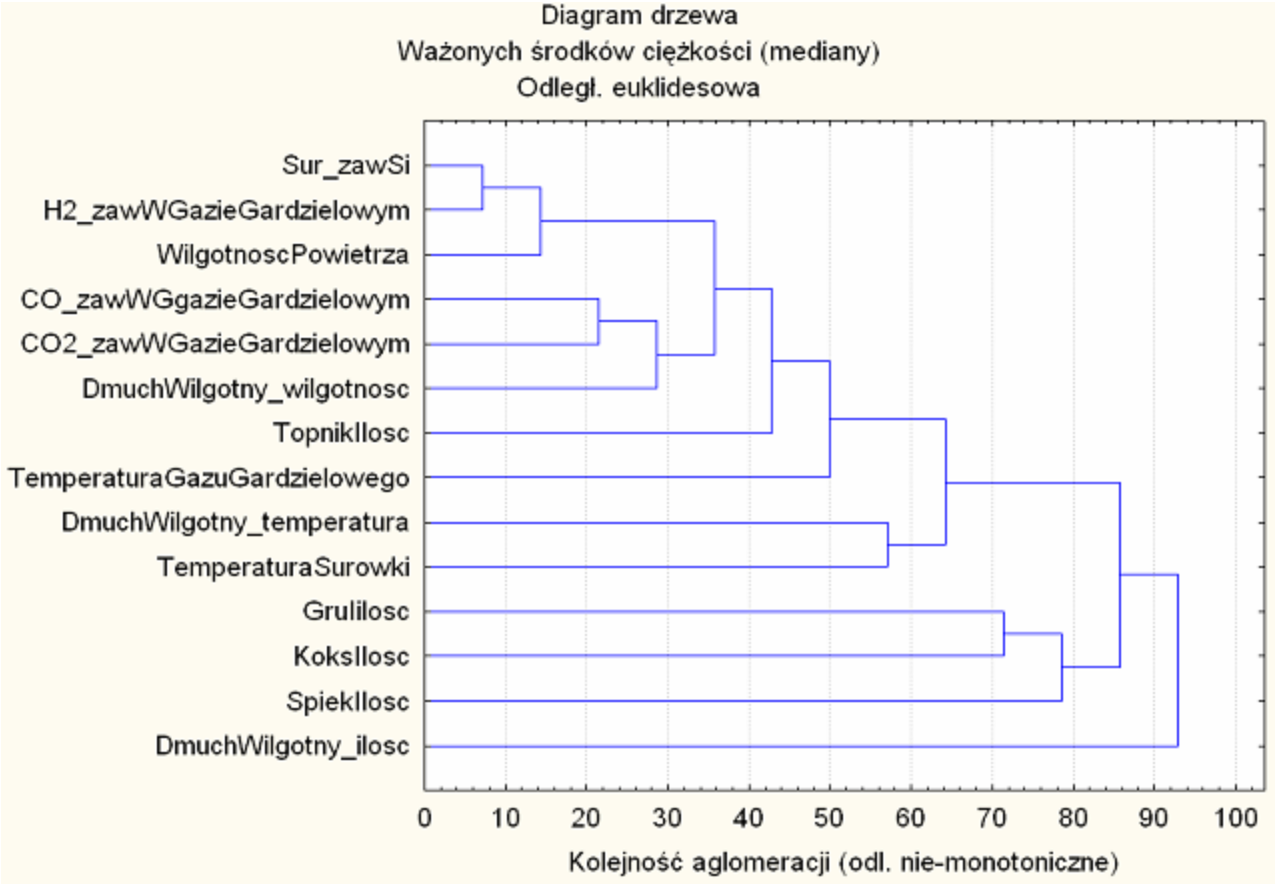
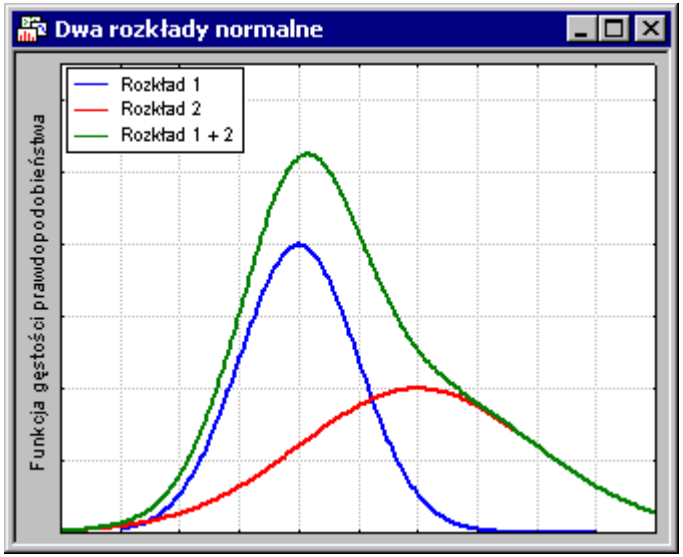


Diagram drzewa
Ważonych środków ciężkości (mediany)
Odległ. euklidesowa
Sur_zawSi
H2_zawWGazieGardzielowym
WilgotnoscPowietrza
CO_zawWGgazieGardzielowym
CO2_zawWGazieGardzielowym
DmuchWilgotny_wilgotnosc
TopnikIlosc
TemperaturaGazuGardzielowego
DmuchWilgotny_temperatura
TemperaturaSurowki
GruliIosc
KoksIlosc
SpiekIlosc
DmuchWilgotny_ilosc
0
10
20
30
40
50
60
70
80
90
100
Kolejność aglomeracji (odl. nie-monotoniczne)

Probabilistyczny algorytm EM

Probabilistyczny algorytm EM
Expectation Maximisation

- Przykład: obserwujemy dużą próbę pomiarów jednej zmiennej ilościowej.
- Zamiast patrzyć tylko na odległość, uwzględniamy dodatkowo informację o rozkładzie przykładów.



- Zamiast przypisywać definitywnie przykład do grupy estymuje prawdopodobieństwo takiego przynależenia.
- różne rozkłady, jak np. rozkład normalny , logarytmiczno-normalny czy Poissona . Możemy także wybrać różne rozkłady dla różnych zmiennych i stąd, wyznaczać grupy z mieszanin różnych typów rozkładów.

Algorytm EM

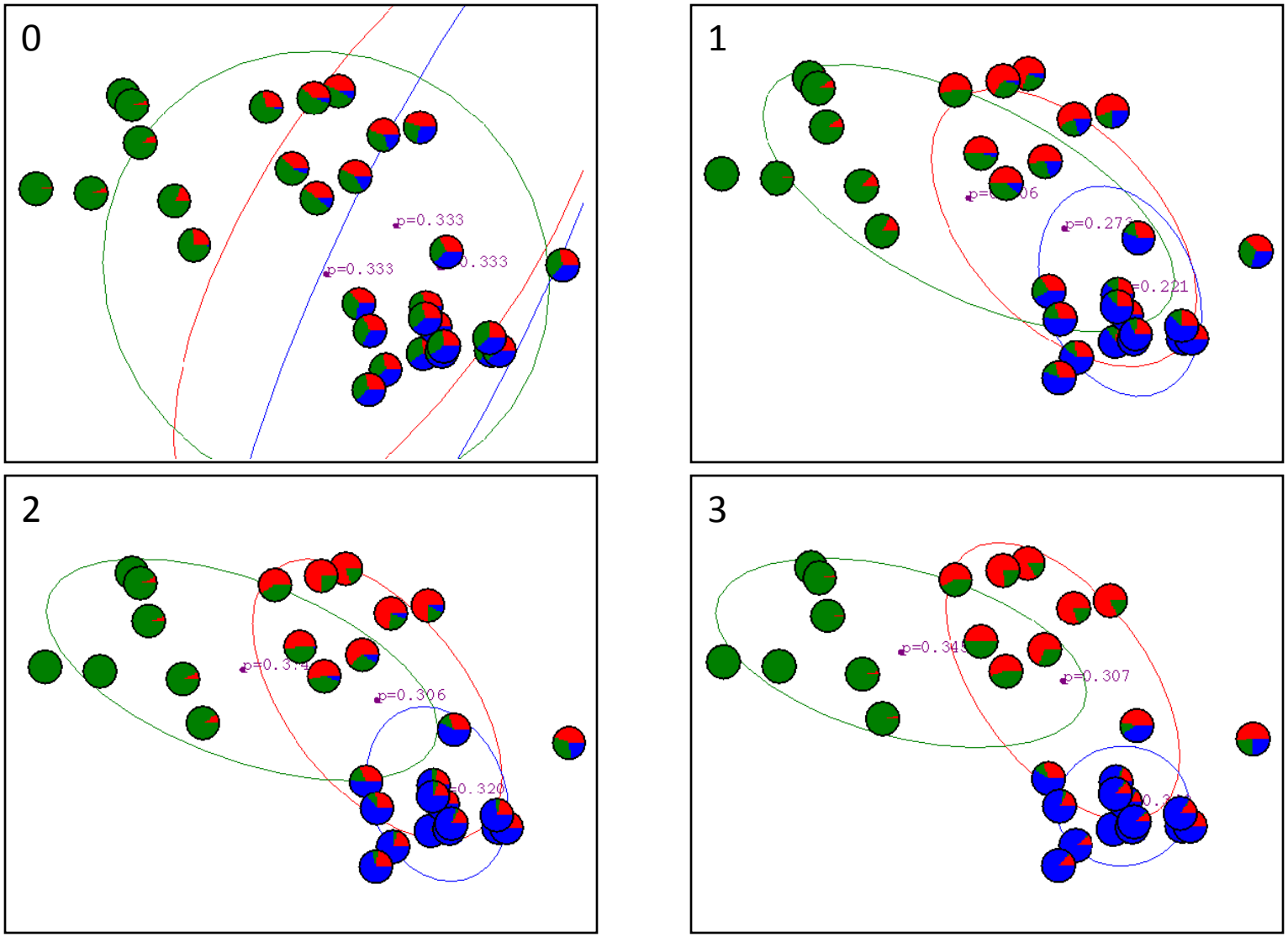
- **Zmienne jakościowe.** Implementacja algorytmu *EM* potrafi korzystać ze zmiennych jakościowych. Najpierw losowo przydziela prawdopodobieństwa (wagi) każdej z klas (kategorii), w każdym ze skupień. W kolejnych iteracjach prawdopodobieństwa są poprawiane tak, by zmaksymalizować wiarygodność danych przy podanej ilości skupień.
- **Prawdopodobieństwa klasyfikacyjne zamiast klasyfikacji.** Wyniki analizy skupień metodą *EM* są inne niż obliczone metodą *k*-średnich. Ta ostatnia wyznacza skupienia. Algorytm *EM* nie wyznacza przyporządkowania obserwacji do klas lecz prawdopodobieństwa klasyfikacyjne.

Algorytm EM

– Metoda składa się z dwu kroków wykonywanych na przemian tak długo, aż pomiędzy kolejnymi przebiegami nie dochodzi do zauważalnej poprawy.

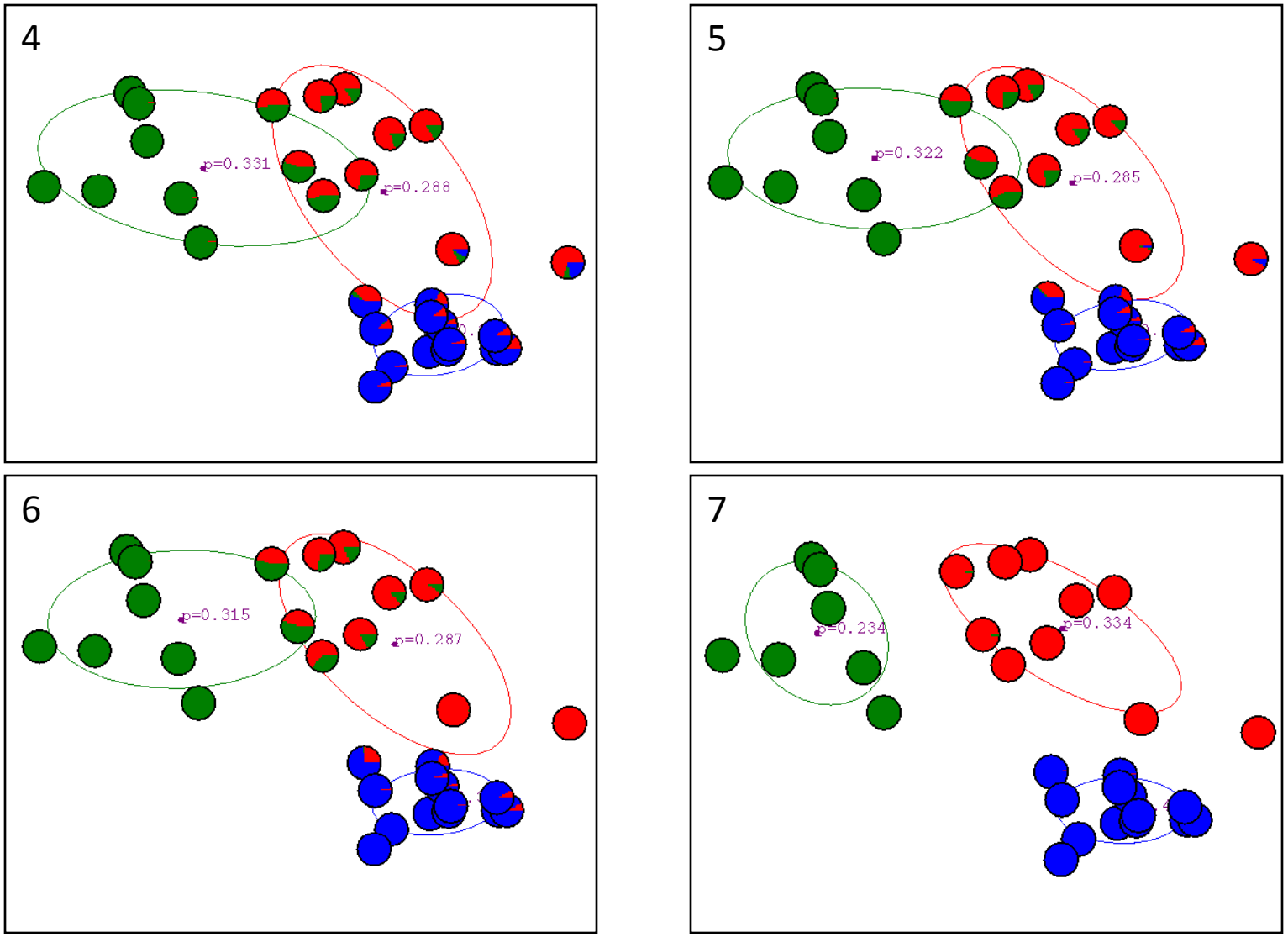
1. **Estymacja** (*expectation*). Dla aktualnego, wyestymowanego układu parametrów rozkładu przykładów dokonaj przypisania przykładom prawdopodobieństwa przynależenia do grup.
2. **Maksymalizuj** - Zamień aktualne parametry rozkładu na takie, które prowadzą do modelu bardzie zgodnego z danymi (rozkładem przykładów). W tym celu wykorzystaj prawdopodobieństwa przynależenia do grup z kroku 1.

Gaussian mixture model



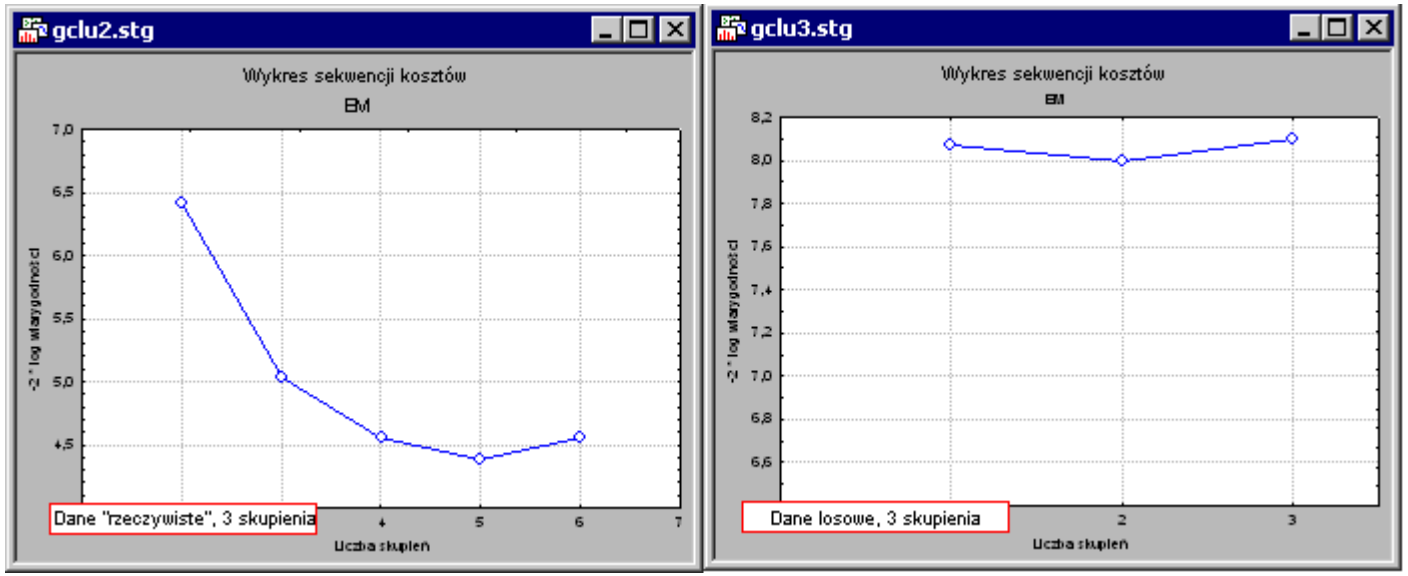
http://www.autonlab.org/tutorials/gmm.html

Gaussian mixture model



Jak dobrać *k*?

- Podobnie jak w przypadku metody k-średnich, problem dotyczy liczby skupień
- W praktyce analityk nie ma zazwyczaj pojęcia ile skupień jest w próbie. Algorytm *v-krotnego sprawdzianu krzyżowego* – automatycznie wyznacza liczbę skupień danych.



Metody iteracyjno-optymalizacyjne

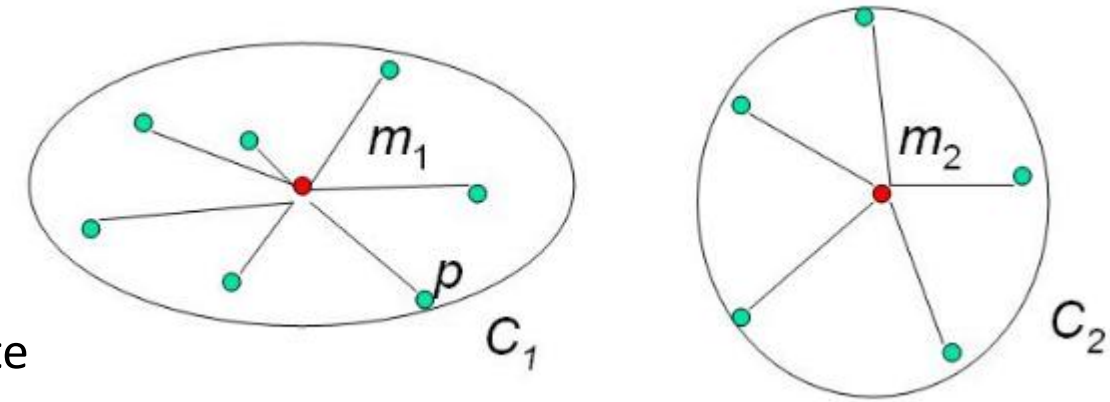
Metody iteracyjno-optymalizacyjne

- Iteracyjno-optymalizacyjne metody grupowania tworzą jeden podział zbioru obiektów (partycję) w miejsce hierarchicznej struktury podziałów
- Metody iteracyjno-optymalizacyjne realokują obiekty pomiędzy klastrami optymalizując funkcję kryterialną zdefiniowaną lokalnie (na podzbiorze obiektów) lub globalnie (na całym zbiorze obiektów)
- Przeszukanie całej przestrzeni wszystkich możliwych podziałów zbioru obiektów pomiędzy k klastrów jest, praktycznie, nie realizowalne
- W praktyce, algorytm grupowanie jest uruchamiany kilkakrotnie, dla różnych podziałów początkowych, a następnie, najlepszy z uzyskanych podziałów jest przyjmowany jako wynik procesu grupowania

Funkcje kryterialne: Suma błędów kwadratowych

Suma błędów kwadratowych:

$$E = \sum_{i=1}^{k} \sum_{p \in C_i} dist(p, m_i)^2$$



Klastry powinny być zwarte

Klastry powinny być maksymalnie rozłączne

odchylenie wewnątrzklastrowe:

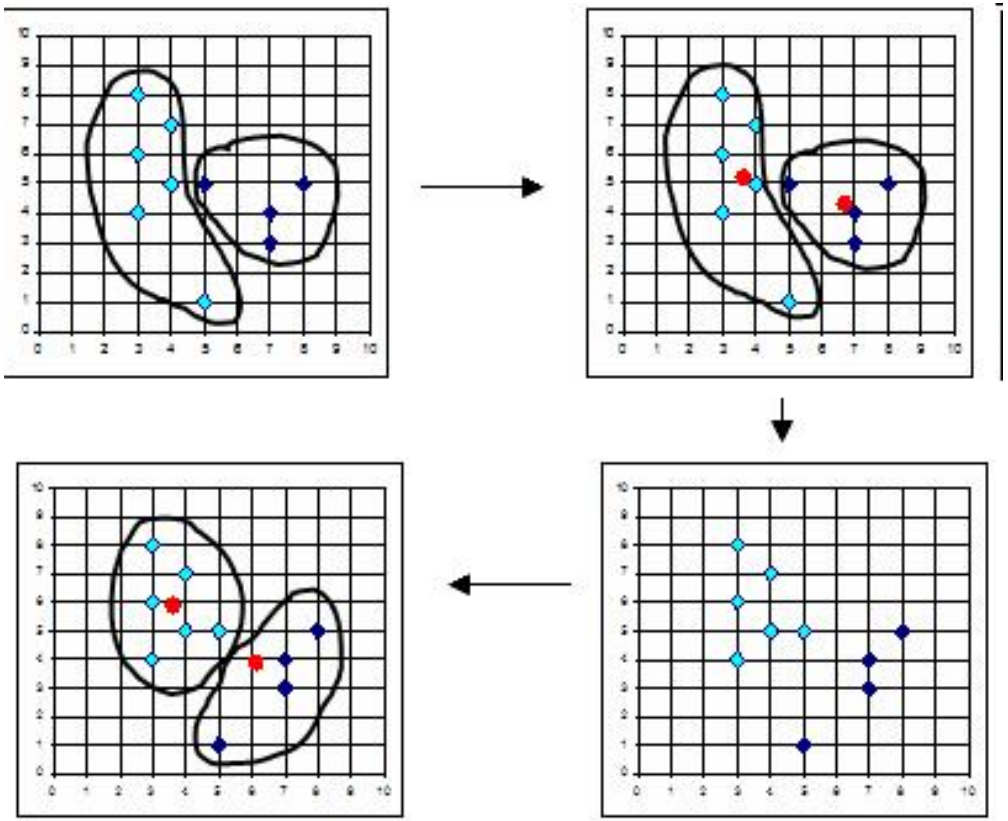
dla każdego obiektu należącego do klastra obliczamy odległość tego punktu do najbliższego obiektu w tym klastrze, i bierzemy max z tych odległości

Algorytm K-means

- idea algorytmu *K-means (k-średnich)* - rozpoczyna się od losowo wybranego grupowania punktów, następnie ponownie przypisuje się punkty tak, aby otrzymać największy wzrost (spadek) w funkcji oceny, po czym przelicza się zaktualizowane skupienia, po raz kolejny przypisuje się punkty i tak dalej aż do momentu, w którym nie ma już żadnych zmian w funkcji oceny lub w składzie skupień.
- To zachłanne podejście ma tę zaletę, że jest proste i gwarantuje otrzymanie co najmniej lokalnego maksimum (minimum) funkcji oceny.
- Osiągnięcie „optimum" globalnego podziału obiektów wymaga przeanalizowania wszystkich możliwych podziałów zbioru n obiektów pomiędzy k klastrów

Algorytm K-means

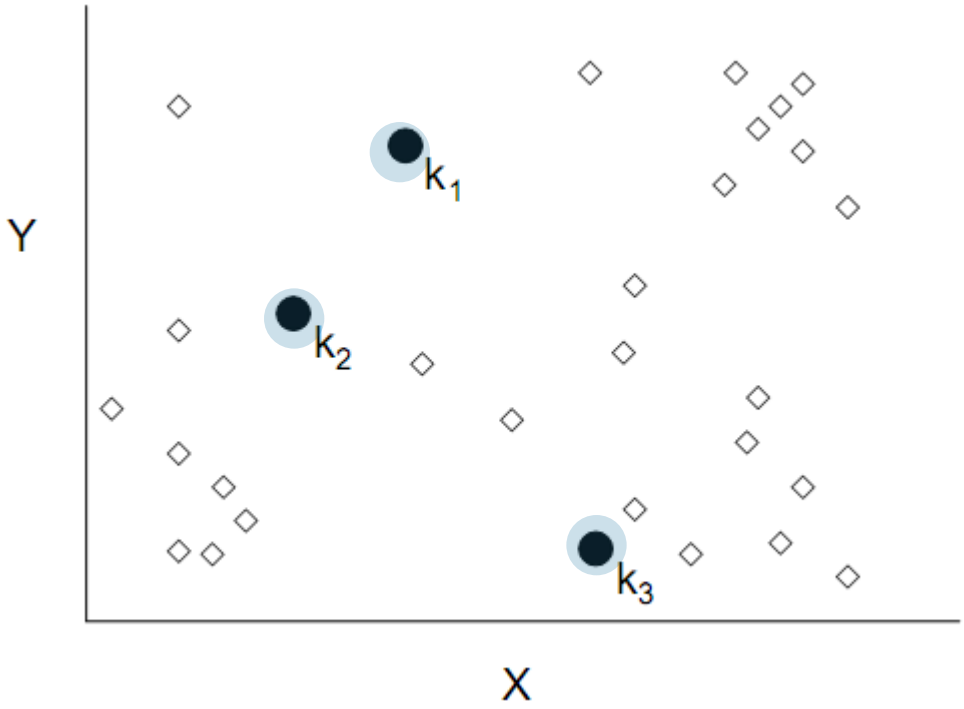
while *są zmiany w skupieniach* C_k **do**
- Podziel zbiór danych na dowolnych k rozłącznych zbiorów;
- Dla każdej grupy wyznacz środek ciężkości;
- Podziel przykłady do najbliższego środka;



krok 1

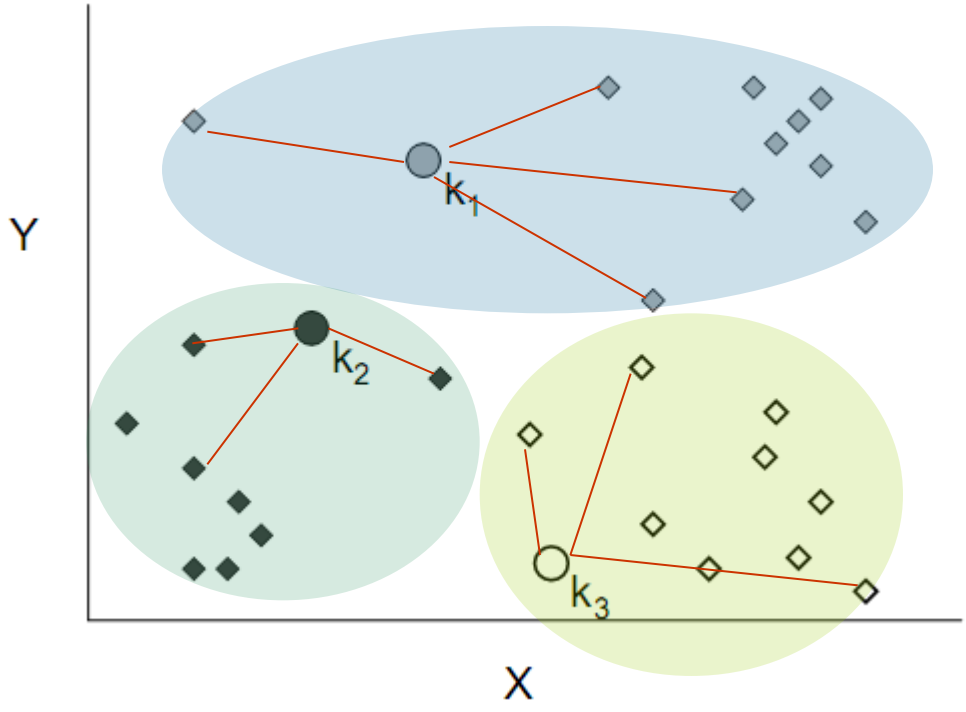
Założenie: k = 3

wybierz 3 początkowe środki klastrów (losowo)



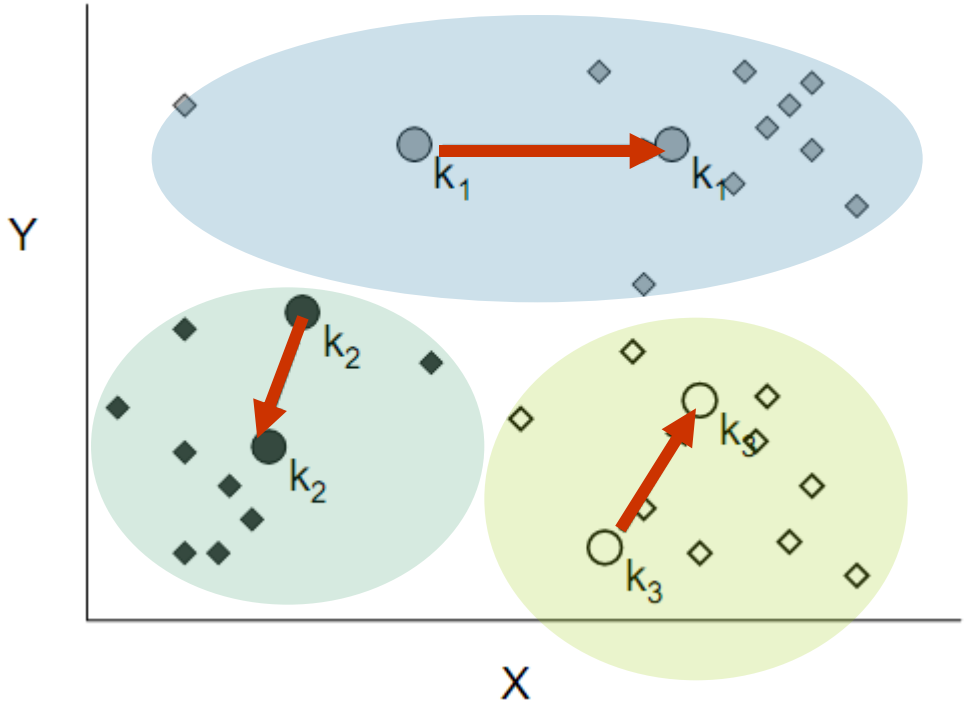
krok 2

– Przydziel każdy obiekt do klastra w oparciu o najmniejszą odległość obiektu od środka klastra



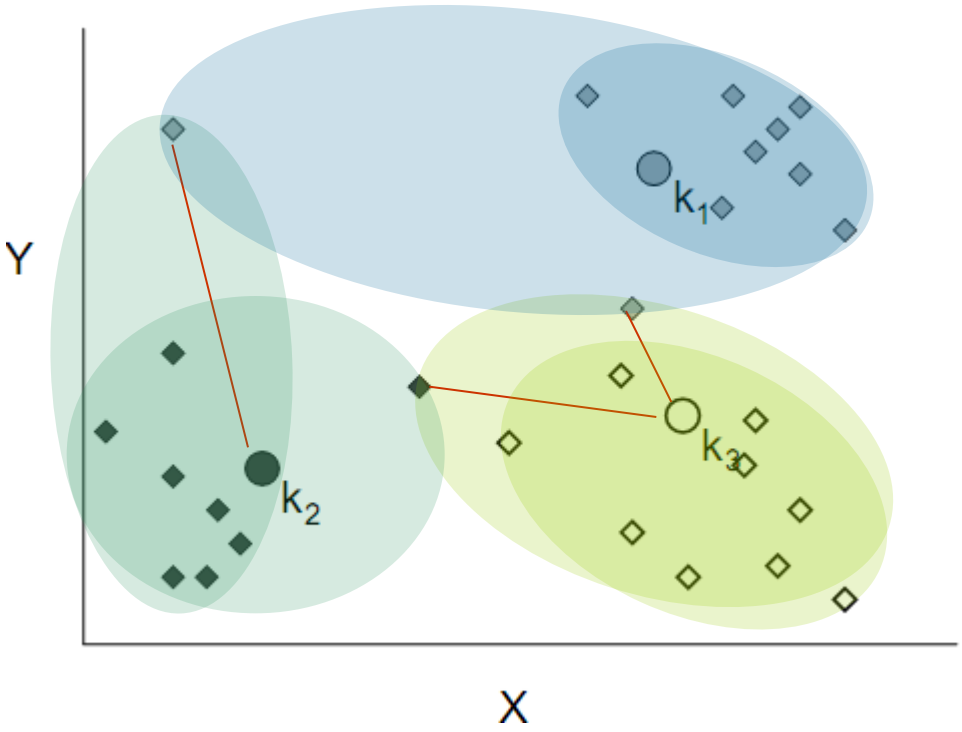
krok 3

– Uaktualnij środki (średnie) wszystkich klastrów



krok 4

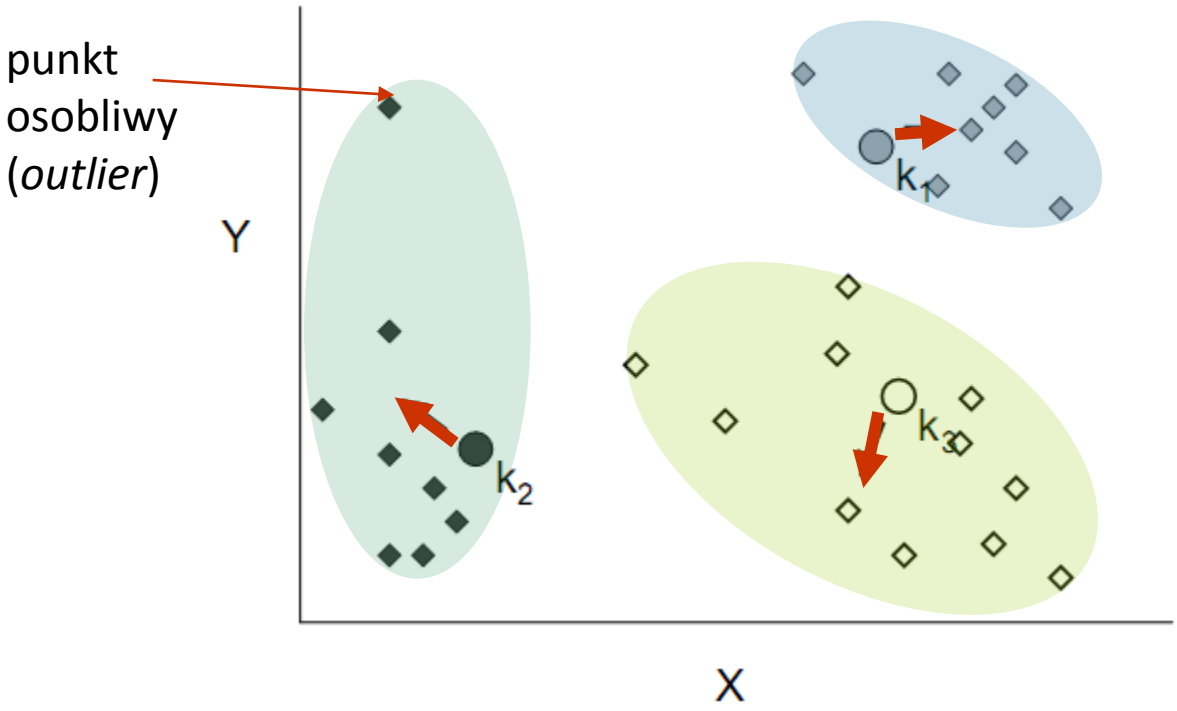
– Realokuj obiekty do najbliższych klastrów



krok 4b

– Oblicz nowe średnie klastrów…

Algorytm bardzo czuły na dane zaszumione lub dane zawierające punkty osobliwe, gdyż punkty takie w istotny sposób wpływają na średnie klastrów powodując ich zniekształcenie



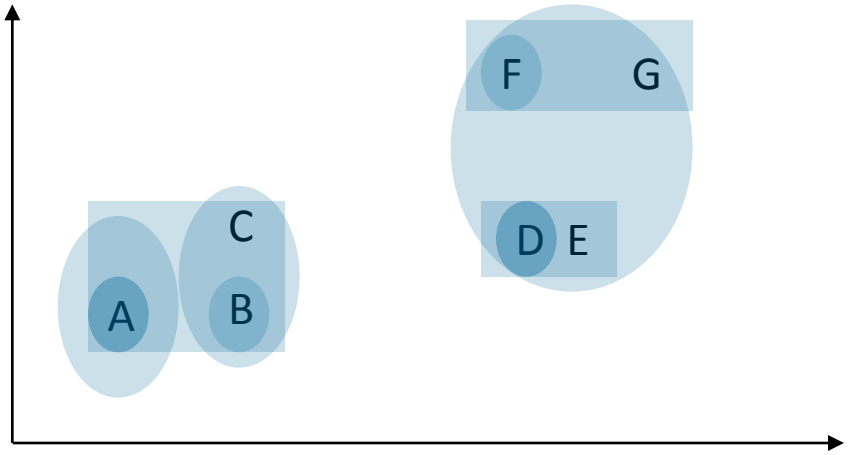
… i wracamy do kroku realokacji obiektów. Dla każdego obiektu następuje weryfikacja, czy obiekt ten podlega realokacji. Jeżeli żaden z obiektów nie wymaga realokacji następuje zakończenie działania algorytmu.

Algorytm k-średnich

- Wynik działania algorytmu (tj. ostateczny podział obiektów pomiędzy klastrami) silnie zależy od początkowego podziału obiektów
- Algorytm może „wpaść” w optimum lokalne
- W celu zwiększenia szansy znalezienia optimum globalnego należy kilkakrotnie uruchomić algorytm dla różnych podziałów początkowych

Algorytm k-średnich

- Przyjmijmy, że środkami klastrów są obiekty A,B, i D, wokół których konstruowane są klastry
- Lepszy (z mniejszym błędem średniokwadratowym) podział uzyskamy przyjmując początkowe środki klastrów A, D, i F (prostokąty)



Algorytm k-medoidów

- W celu uodpornienia algorytmu k-średnich na występowanie punktów osobliwych, zamiast średnich klastrów jako środków tych klastrów, możemy przyjąć medoidy – najbardziej centralnie ulokowane punkty klastrów:
  - Średnia zbioru 1, 3, 5, 7, 9 wynosi 5
  - Średnia zbioru 1, 3, 5, 7, 1009 wynosi 205
  - Medoid zbioru 1, 3, 5, 7, 1009 wynosi 5

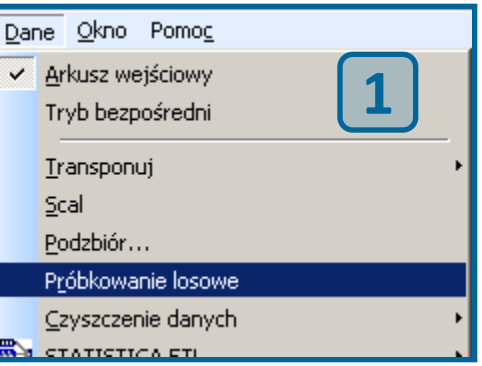
Inne metody grupowania

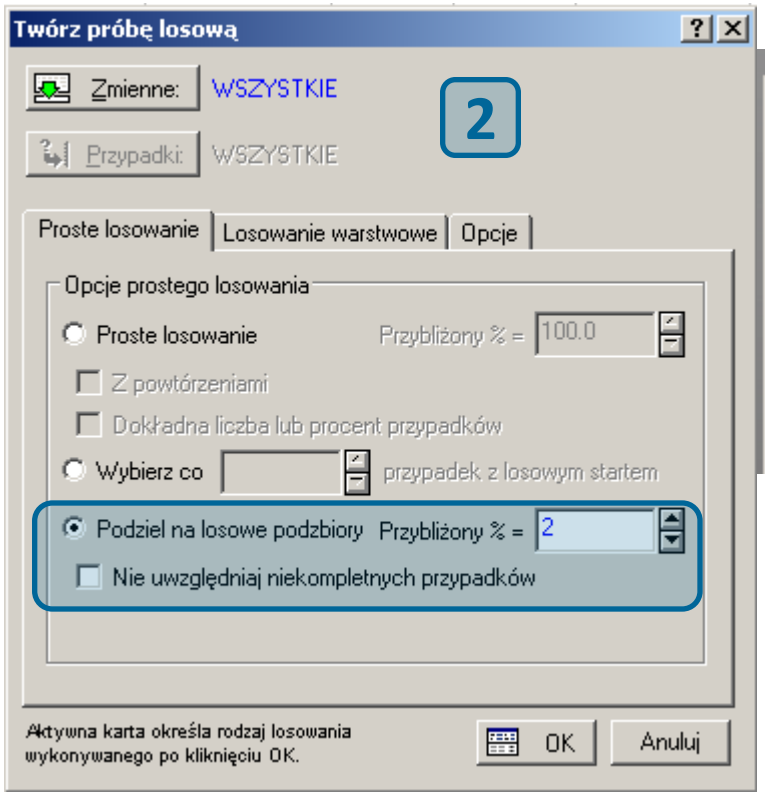
- **Metody oparte o analizę gęstości**
  - » dany klaster jest rozszerzany o obiekty należące do jego sąsiedztwa, pod warunkiem, że gęstość obiektów w danym sąsiedztwie przekracza zadaną wartość progową (algorytmy DBSCAN, OPTICS, DENCLUE)
  - » Ta grupa metod była rozwijana, głównie, w odniesieniu do zastosowań w dziedzinie rozpoznawania obrazów.
- **Metody oparte o strukturę gridową**
  - » przestrzeń obiektów jest dzielona na skończona liczbę komórek, które tworzą strukturę gridu; cały proces grupowania jest wykonywany na tej strukturze (algorytmy STING, CLIQUE)
  - » Cechą i zasadniczą zaletą tej grupy metod jest skalowalność i możliwość analizy dużych wolumenów danych wielowymiarowych.

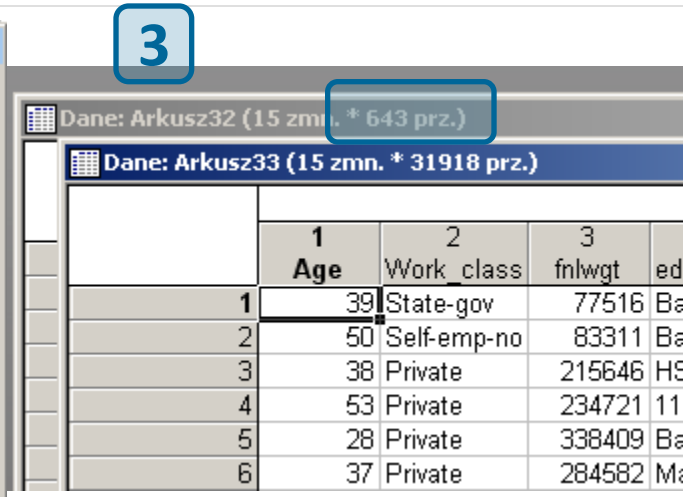
Przykład w STATISTICA

Adult.sta

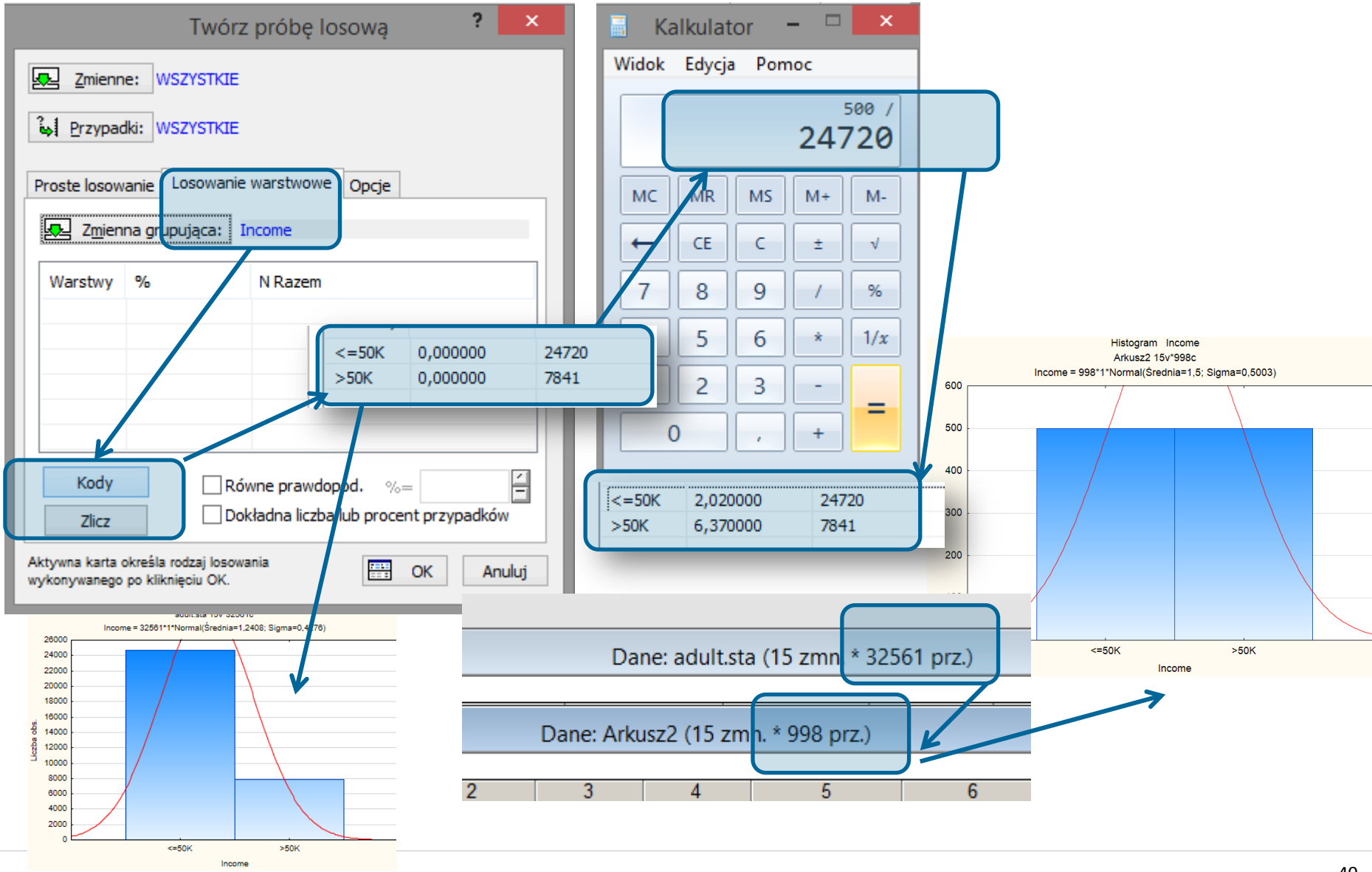
Przykład w STATISTICA

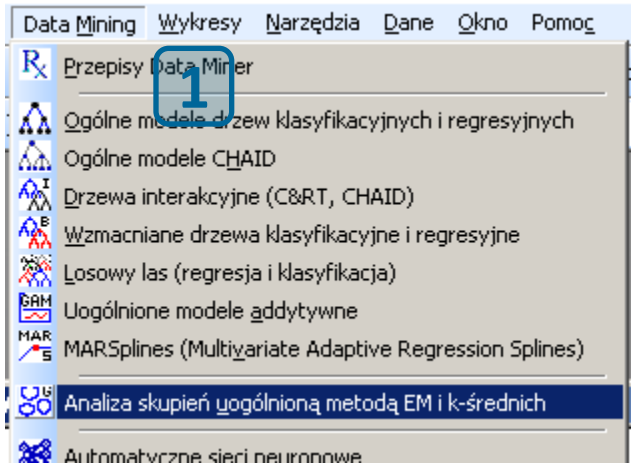


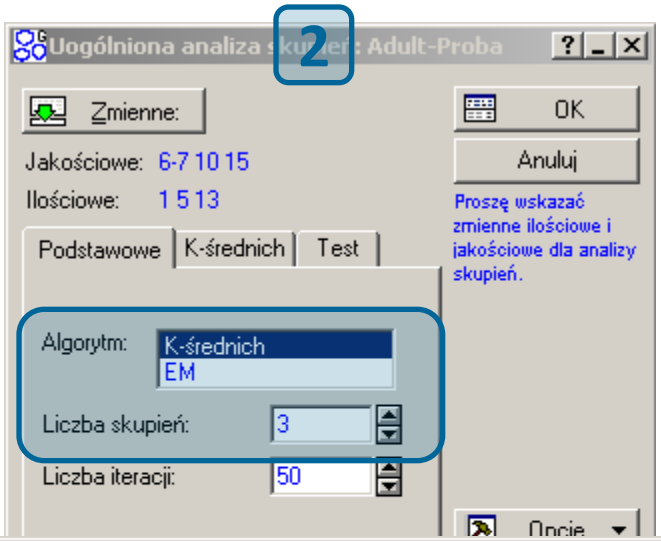




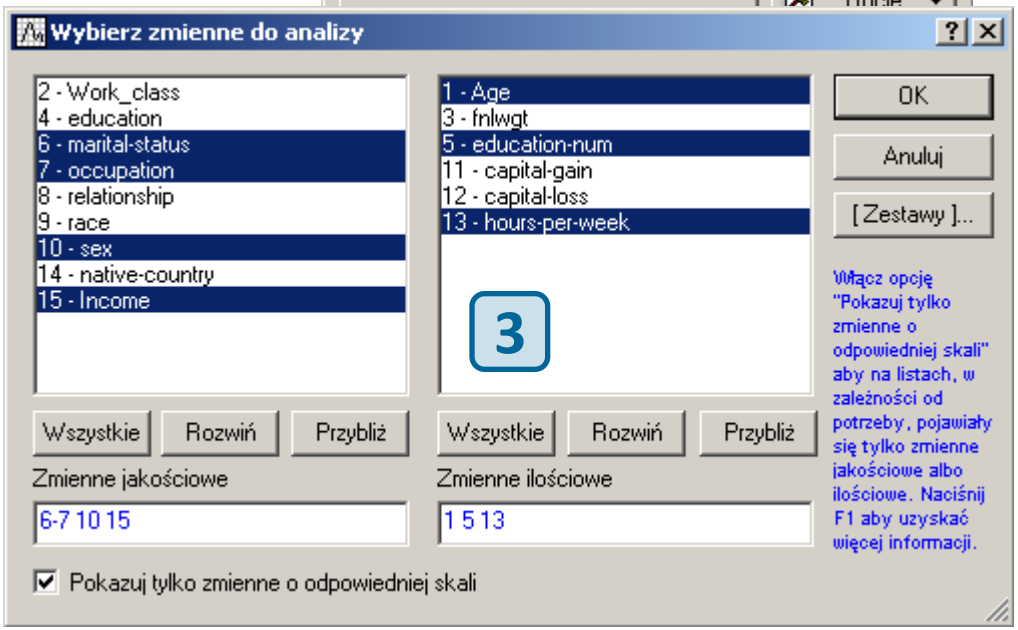
Losowanie warstwowe

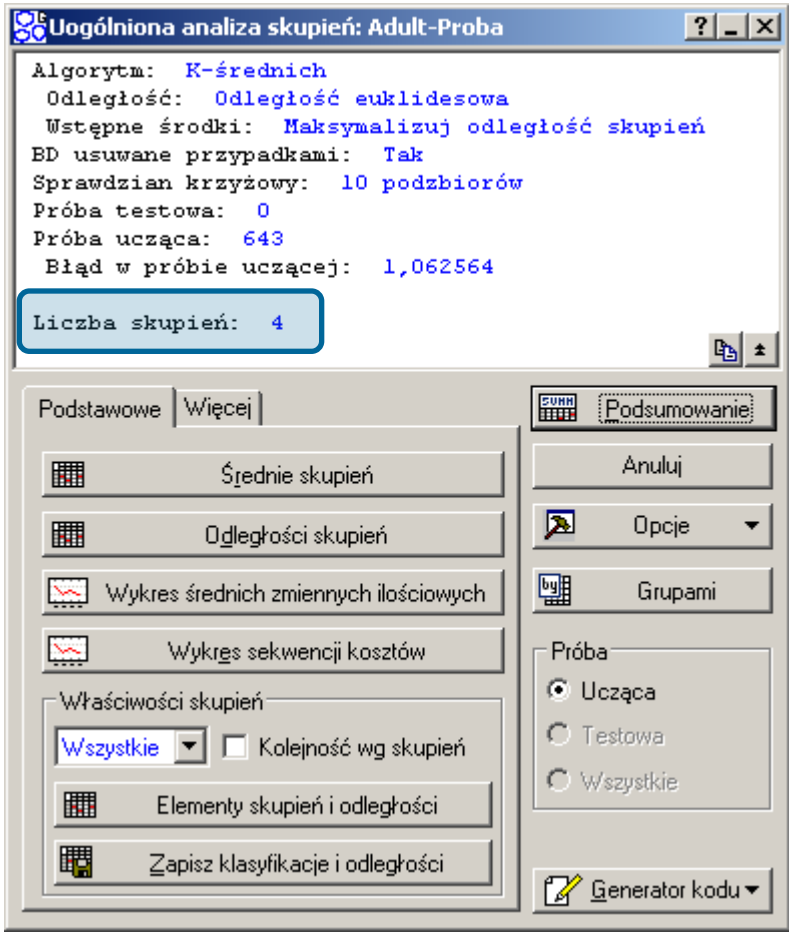


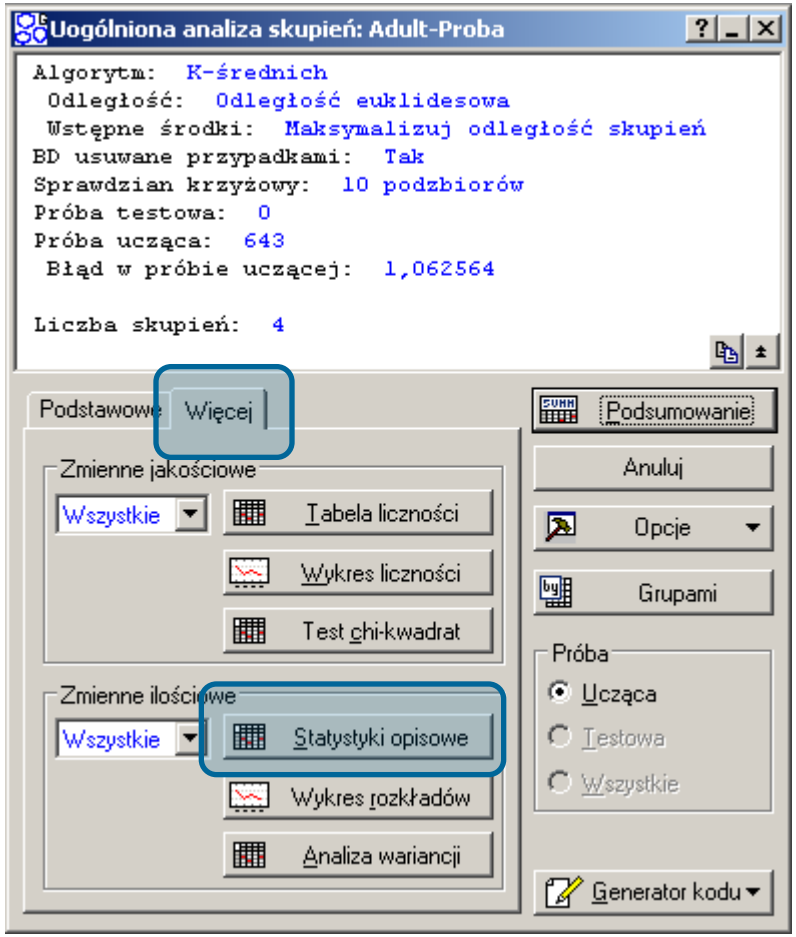

Data Mining
Wykresy
Narzędzia
Dane
Okno
Pomoc
Przepisy Data Miner
Ogólne modele drzew klasyfikacyjnych i regresyjnych
Ogólne modele CHAID
Drzewa interakcyjne (C&RT, CHAID)
Wzmacniane drzewa klasyfikacyjne i regresyjne
Losowy las (regresja i klasyfikacja)
Uogólnione modele addytywne
MARSplines (Multivariate Adaptive Regression Splines)
Analiza skupień uogólnioną metodą EM i k-średnich
Automatyczne sieci neuronowe
1


Uogólniona analiza skupień: Adult-Proba
2
Zmienne:
OK
Anuluj
Jakościowe: 6-7 10 15
Ilościowe: 1 5 13
Proszę wskazać zmienne ilościowe i jakościowe dla analizy skupień.
Podstawowe
K-średnich
Test
Algorytm:
K-średnich
EM
Liczba skupień:
3
Liczba iteracji:
50
Opcje

Uogólniona analiza skupień: Adult-Proba
Zmienne:
OK
Anuluj
Jakościowe: 6-7 10 15
Ilościowe: 1 5 13
4
Proszę wskazać zmienne ilościowe i jakościowe dla analizy skupień.
Podstawowe
K-średnich
Test
Próba testowa
Nie
V-krotny sprawdzian krzyżowy
Liczba prób (zbiorów):
10
Jądro generatora liczb losowych:
1
Opcje
Otwórz dane
Minimalna liczba skupień:
2
Maksymalna liczba skupień:
25
Minimalny spadek, %:
5,
Braki danych usuwane przypadkami


Wybierz zmienne do analizy
2 - Work_class
4 - education
6 - marital-status
7 - occupation
8 - relationship
9 - race
10 - sex
14 - native-country
15 - Income
1 - Age
3 - fnlwgt
5 - education-num
11 - capital-gain
12 - capital-loss
13 - hours-per-week
OK
Anuluj
[Zestawy]...
3
Wszystkie
Rozwiń
Przybliż
Zmienne jakościowe
6-7 10 15
Zmienne ilościowe
1 5 13
Włącz opcję "Pokazuj tylko zmienne o odpowiedniej skali" aby na listach, w zależności od potrzeby, pojawiały się tylko zmienne jakościowe albo ilościowe. Naciśnij F1 aby uzyskać więcej informacji.
Pokazuj tylko zmienne o odpowiedniej skali


Uogólniona analiza skupień: Adult-Proba
Algorytm: K-średnich
Odległość: Odległość euklidesowa
Wstępne środki: Maksymalizuj odległość skupień
BD usuwane przypadkami: Tak
Sprawdzian krzyżowy: 10 podzbiorów
Próba testowa: 0
Próba ucząca: 643
Błąd w próbie uczącej: 1,062564
Liczba skupień: 4
Podstawowe
Więcej
Podsumowanie
Średnie skupień
Anuluj
Odległości skupień
Opcje
Wykres średnich zmiennych ilościowych
Grupami
Wykres sekwencji kosztów
Próba
Właściwości skupień
Ucząca
Wszystkie
Kolejność wg skupień
Testowa
Elementy skupień i odległości
Wszystkie
Zapisz klasyfikacje i odległości
Generator kodu


Uogólniona analiza skupień: Adult-Proba
Algorytm: K-średnich
Odległość: Odległość euklidesowa
Wstępne środki: Maksymalizuj odległość skupień
BD usuwane przypadkami: Tak
Sprawdzian krzyżowy: 10 podzbiorów
Próba testowa: 0
Próba ucząca: 643
Błąd w próbie uczącej: 1,062564
Liczba skupień: 4
Podstawowe
Więcej
Podsumowanie
Zmienne jakościowe
Anuluj
Wszystkie
Tabela liczności
Opcje
Wykres liczności
Grupami
Test chi-kwadrat
Próba
Zmienne ilościowe
Ucząca
Wszystkie
Statystyki opisowe
Testowa
Wykres rozkładów
Wszystkie
Analiza wariancji
Generator kodu

Statystyki opisowe skupień

Statystyki zmiennej ilościowej Age (Adult-Proba)
Liczba skupień: 4
Całkowita liczba przypadków uczących: 643

	Skupienie 1	Skupienie 2	Skupienie 3	Skupienie 4	Ogółem
Minimum	17,00000	22,00000	18,00000	21,00000	17,00000
Maksimum	73,00000	90,00000	80,00000	90,00000	90,00000
Średnia	30,30736	45,48299	40,13253	45,07071	38,58631
Odchylenie std.	11,92972	11,76502	13,12113	12,85417	13,94951

– Wniosek 1: skupienie 1. to ludzie młodsi (30lat), skupienia 2 i 4 to starsi (45 lat), skupienie 3 (40lat)

Statystyki zmiennej ilościowej education-num (Adult-Proba)
Liczba skupień: 4
Całkowita liczba przypadków uczących: 643

	Skupienie 1	**Skupienie 2**	Skupienie 3	Skupienie 4	Ogółem
Minimum	1,00000	6,00000	2,00000	3,00000	1,00000
Maksimum	15,00000	16,00000	14,00000	14,00000	16,00000
Średnia	9,74459	11,87755	9,19880	9,35354	10,03110
Odchylenie std.	2,26690	2,23269	2,26494	2,33567	2,48716

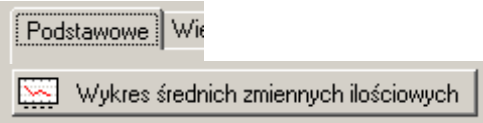
– Wniosek 2: skupienie 2. to ludzie lepiej wykształceni niż skupienia 1., 3. i 4.

Statystyki zmiennej ilościowej hours-per-week (Adult-Proba)
Liczba skupień: 4
Całkowita liczba przypadków uczących: 643

	Skupienie 1	Skupienie 2	Skupienie 3	Skupienie 4	Ogółem
Minimum	4,00000	10,00000	5,00000	4,00000	4,00000
Maksimum	99,00000	84,00000	99,00000	99,00000	99,00000
Średnia	35,60606	45,14286	43,72892	41,85859	40,84603
Odchylenie std.	12,51350	11,40115	13,48732	11,09974	12,94728

– Wniosek 3: najmniej pracują osoby ze skupienia 1, najciężej ze skupienia 2.; skupienia 3 i 4 pracują normalnie – ok. 40h/tydzień

Co już wiemy?



Skupienie 1:

- młodzi (30lat);
- średnio wykształceni;
- pracujący stosunkowo mało;

Skupienie 2:

- w średnim wieku (ok. 45lat);
- bardzo dobrze wykształceni;
- pracujący dużo;

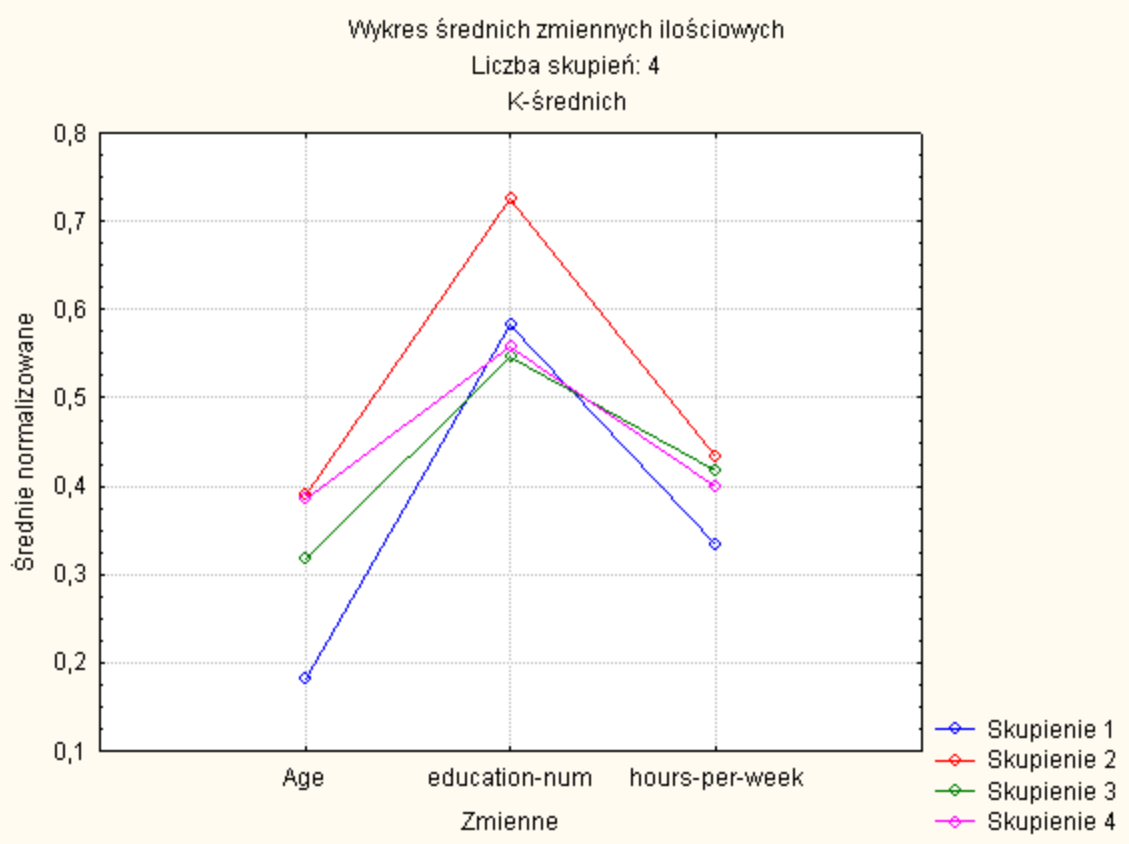
Skupienie 3:

- w wieku ok. 40 lat;
- słabo wykształceni;
- pracujący stosunkowo dużo;

Skupienie 4:

- w średnim wieku (ok. 45lat);
- raczej słabo wykształceni;
- pracujący ok. 41h/tyg

kto jest najlepszym klientem?



Podstawowe | Więcej

Średnie skupień

Średnie skupień (metoda k-średnich) (Adult-Proba)
Liczba skupień: 4
Całkowita liczba przypadków uczących: 643

Skupienie	marital-status	occupation	sex	Income	Age	education-num	hours-per-week	Liczba przypadków	Procent (%)
1	Never-married	Other-service	Female	<=50K	30,30736	9,74459	35,60606	231	35,92535
2	Married-civ-spouse	Prof-specialty	Male	>50K	45,48299	11,87755	45,14286	147	22,86159
3	Married-civ-spouse	Sales	Male	<=50K	40,13253	9,19880	43,72892	166	25,81649
4	Divorced	Craft-repair	Male	<=50K	45,07071	9,35354	41,85859	99	15,39658

Tabela liczności dla zmiennej jakościowej: occupation (Adult-Prob
Liczba skupień: 4
Całkowita liczba przypadków uczących: 643

	Skupienie 1	**Skupienie 2**	Skupienie 3	Skupienie 4	Razem
Other-service	52	1	8	0	61
Sales	15	9	45	2	71
Craft-repair	3	17	15	46	81
Machine-op-inspct	9	2	15	4	30
Farming-fishing	2	3	14	2	21
Prof-specialty	22	47	2	8	79
Tech-support	7	9	2	2	20
Adm-clerical	39	9	11	11	70
Handlers-cleaners	16	5	8	2	31
?	23	6	10	6	45
Transport-moving	13	6	17	5	41
Armed-Forces	0	0	1	0	1
Exec-managerial	28	25	18	7	76
Protective-serv	2	8	0	4	14
Priv-house-serv	2	0	0	0	2

Najlepiej zarabia skupienie 2 – kim ONI są?

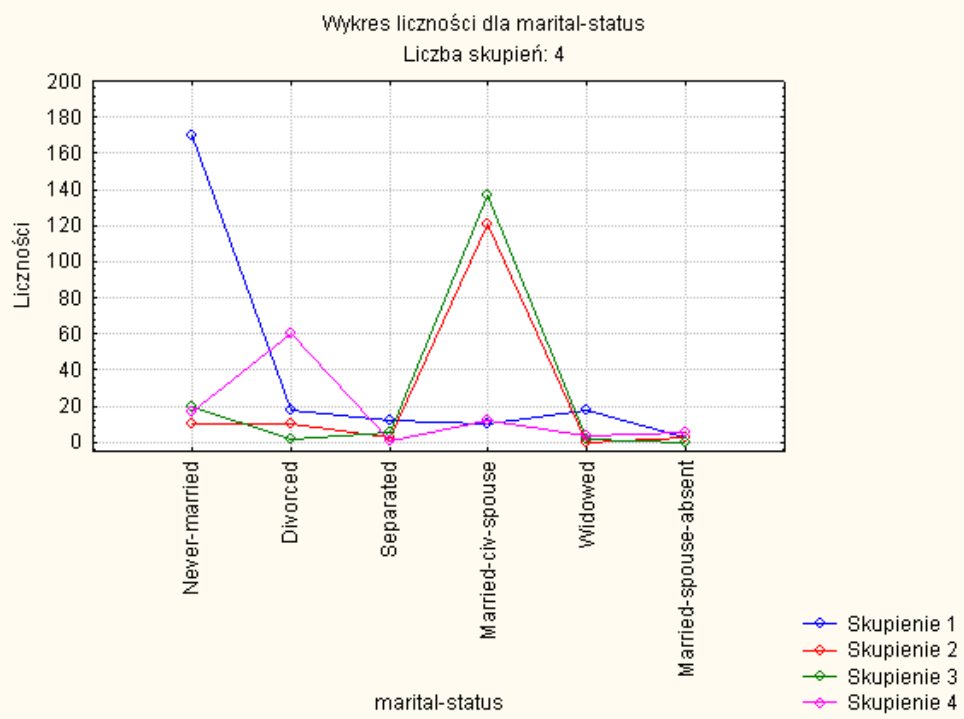
Skupienie 2:

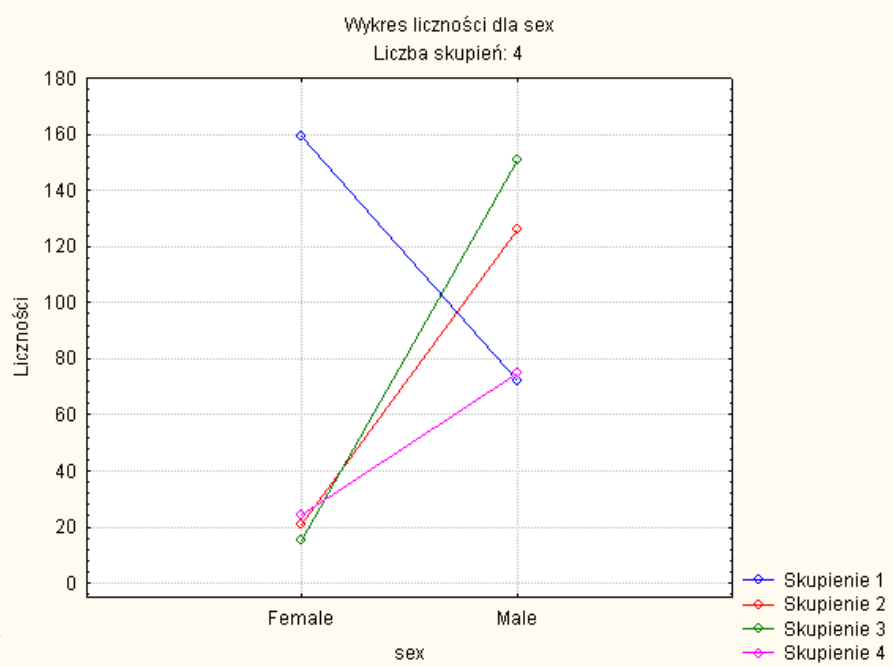
- w średnim wieku (ok. 45lat);
- bardzo dobrze wykształceni;
- pracujący dużo;

Co nowego o skupieniu 2?

Skupienie 2 to:

- ludzie średnim wieku (ok. 45lat);
- bardzo dobrze wykształceni;
- pracujący dużo;
- w małżeństwie;
- zawód: specjalista/kierownik
- mężczyzna

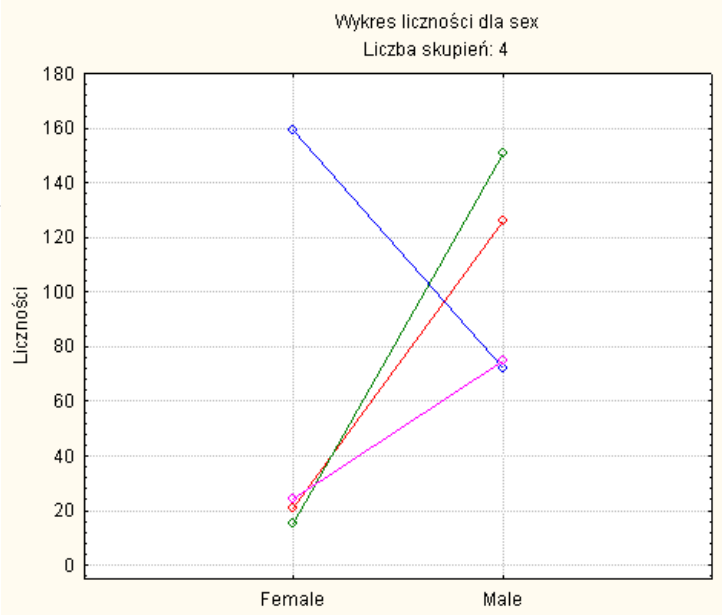


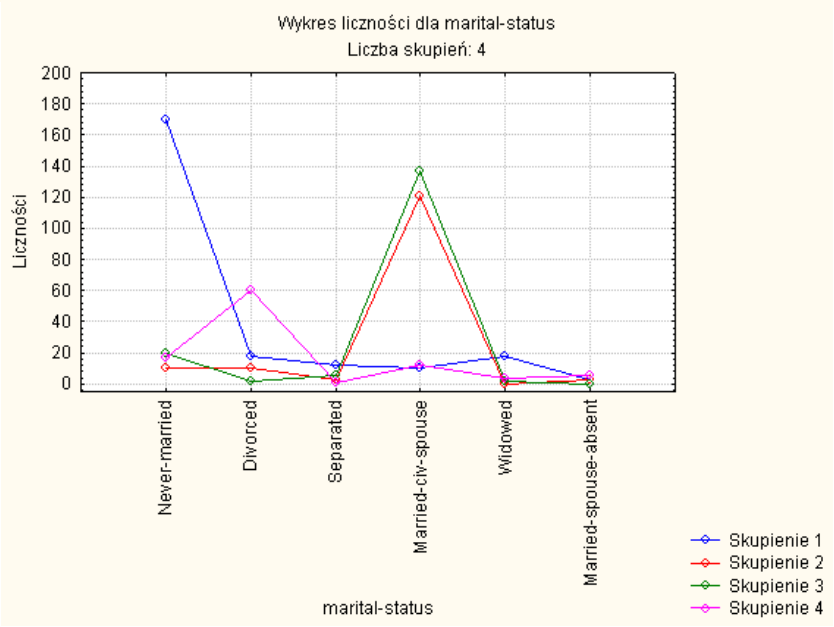


Co z kobietami?

Kobiety w naszej próbie zdominowały skupienie 1:

- są młode (30lat);
- średnio wykształcone;
- pracujące stosunkowo mało (dzieci?);
- niezamężne;
- pracują w usługach;
- prawie wszystkie zarabiają poniżej 50 000$;

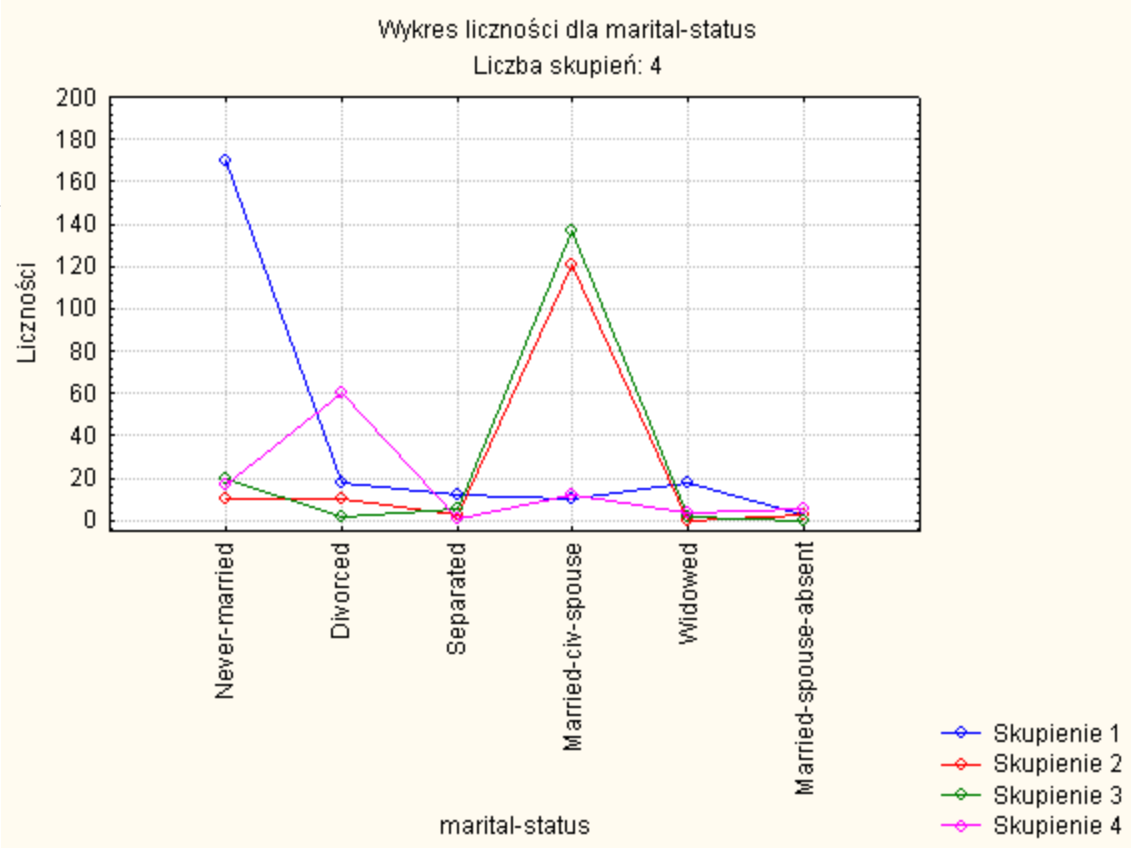




Kim są pozostali?

Skupienie 3:

- w wieku ok. 40 lat;
- słabo wykształceni;
- pracujący stosunkowo dużo;
- mężczyźni;
- żonaci;
- głównie sprzedawcy;
- zarabiają poniżej 50 000$



Skupienie 4:

- w średnim wieku (ok. 45lat); raczej słabo wykształceni; pracujący ok. 41h/tyg;
- kobiety i mężczyźni po rozwodzie,
- rzemieślnicy, fachowcy…
- zarabiający poniżej 50 000$

Kto się najbardziej różni?

- Najbardziej rozróżnialne są skupienia 1. i 2. – czyli młodzi specjaliści i panny na wydaniu
- Widać to również w zarobkach.
- Po ślubie różnice się wyrównują…

	Standaryzowana odległość między centroidami k-średnich Liczba skupień: 4			
	Skupienie 1	Skupienie 2	Skupienie 3	Skupienie 4
Skupienie 1	0,000000	2,018295	1,739756	1,745254
Skupienie 2	2,018295	0,000000	1,427405	1,740558
Skupienie 3	1,739756	1,427405	0,000000	1,416005
Skupienie 4	1,745254	1,740558	1,416005	0,000000