Analiza skupień (Cluster analysis)

Analiza skupień jest to podział zbioru obserwacji na podzbiory (tzw. *klastry)* tak, że obiekty (obserwacje) w tym samym klastrze były podobne (w pewnym sensie).
Jest to pojęcie z zakresu eksploracji danych oraz uczenia maszynowego, wywodzące się z szerszego pojęcia, jakim jest klasyfikacja bezwzorcowa (uczenie się bez nadzoru). Analizy są wykorzystywane do wykrywania struktury zebranych danych i dokonywania uogólnień, np. w analizie obrazu czy wyszukiwaniu informacji.

Wybrane cele grupowania są następujące:

- uzyskanie jednorodnych grup badanych obiektów, ułatwiających wyodrębnienie ich zasadniczych cech czy uzyskanie klasyfikacji obiektów typowych
- odkrycie nieznanej struktury analizowanych danych, a w konsekwencji klasyfikacja obiektów typowych
- zredukowanie dużej liczby danych pierwotnych do kilku podstawowych kategorii, które mogą być traktowane jako przedmioty dalszej analizy,
- porównywanie obiektów wielocechowych (tylko wskazanie innej grupy najbardziej podobnej do danej grupy).

Wybór konkretnej metody analizy skupień jest uwarunkowany źródłem danych (ich charakterem) oraz oczekiwaną postacią rezultatów (wymaganym rodzajem interpretacji uzyskanego wyniku).

Algorytmy analizy skupień dzieli się na kilka podstawowych kategorii:

- metody hierarchiczne (hierarchiczne analizy skupień) – algorytm tworzy dla zbioru obiektów hierarchię klasyfikacji (kolejnych grupowań/dzieleń zbioru), zaczynając od takiego podziału, w którym każdy obiekt stanowi samodzielne skupienie, a kończąc na podziale, w którym wszystkie obiekty należą do jednego skupienia (albo odwrotnie).
    - procedury aglomeracyjne (ang. *agglomerative*) – tworzą macierz podobieństwa klasyfikowanych obiektów, a następnie w kolejnych krokach łączą w skupienia obiekty (lub wcześniej utworzone grupy) najbardziej do siebie podobne,
    - procedury deglomeracyjne (ang. *divisive*) – zaczynają od skupienia obejmującego wszystkie obiekty, a następnie w kolejnych krokach dzielą je na mniejsze i bardziej jednorodne skupienia aż do momentu, gdy każdy obiekt stanowi samodzielne skupienie (rzadko używane)
- niehierarchiczne analizy skupień, np. grupa metod k-średnich (ang. *k-means*), w której grupowanie polega na wstępnym podzieleniu populacji na z góry założoną liczbę klas. Następnie uzyskany podział jest poprawiany w ten sposób, że niektóre elementy są przenoszone do innych klas, tak, aby uzyskać lepszy podział (minimalizują wariancję wewnątrz grup). Cały czas występuje taka sama liczba klas. Nie występuje wykres dendrogram.
    - losowy wybór środków (centroidów) klas (skupień),
    - przypisanie punktów do najbliższych centroidów,
    - wyliczenie nowych środków skupień,
    - powtarzanie algorytmu aż do osiągnięcia kryterium zbieżności (najczęściej jest to krok, w którym nie zmieniła się przynależność punktów do klas);
- metody rozmytej analizy skupień (ang. *fuzzy clustering*), np. c-średnich (*c-means*). Metody rozmytej analizy skupień mogą przydzielać element do więcej niż jednej kategorii (z ‘prawdopodobieństwem’ przynależności), co różni je od metod klasycznej analizy skupień, w których uzyskana klasyfikacja ma

charakter grupowania rozłącznego, którego wynikiem jest to, że każdy element należy do jednej i tylko jednej klasy.

- Biclustering (inaczej 2-way cluster analysis; analiza skupień działająca równocześnie na obiekty i cechy obiektów, tj wiersze i kolumny macierzy obserwacji). Występuje kilka rodzajów tej metody, uzależnionych od tego, co rozumiemy jako obiekty podobne pod względem obu cech. Na razie stosowana rzadko.

Hierarchiczna analiza skupień

Konkretna zależy od dwu wyborów: przyjętej miary odległości (określającej niepodobieństwo badanych **pojedynczych** obiektów) i metody skupień, aglomeracji (metody liczenia odległości między **grupami** obiektów)

Pomiar odległości

Ważnym krokiem w analizie skupień jest wybór metody pomiaru odległości, która określi stopień *podobieństwa /niepodobieństwa* dwóch elementów (lub grup elementów). Dla większości typów analizy skupień wystarczy określić odległość (stopień niepodobieństwa) pomiędzy pojedynczymi obiektami.

Jeżeli analizowane dane są jednowymiarowe, a badana cecha jest cechą ilościową, to odległość można określić jako różnicę wartości. np.

osoba	1	2	3	4	...
wzrost	152	165	170	180	...

macierz odległości

	1	2	3	4	...
1	*0*	7	18	28	...
2	7	*0*	5	15	...
3	18	5	*0*	10	...
4	28	15	10	*0*	...
...	...	...	...	...	

Analiza skupień stosowana jest również (**głównie**) do analiz wielowymiarowych, np.

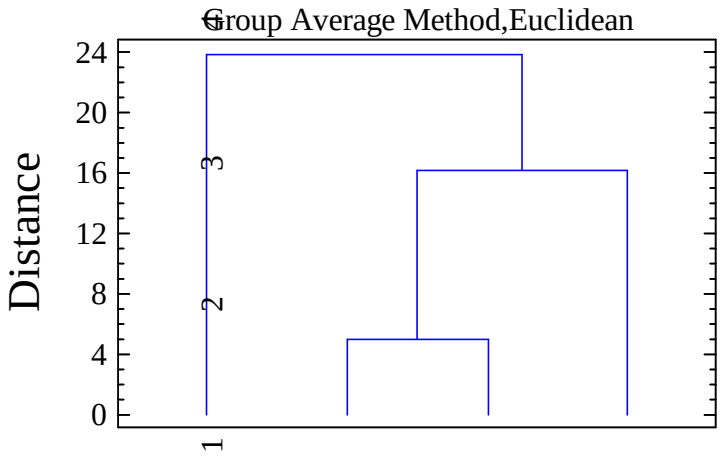
osoba	1	2	3	4	...
wzrost (cm)	152	165	170	180	...
waga (kg)	60	70	70	80	...

stosując odległość euklidesa między obserwacjami (niestandaryzowanymi) otrzymamy macierz odległości (niepodobieństwa):

	1	2	3	4	...
1	0,00	16,40	20,59	14,14	...
2	16,40	0,00	5,00	18,03	...
3	20,59	5,00	0,00	14,14	...
4	14,14	18,03	14,14	0,00	...
...	...	...	...	...	

Im większa liczba tym niepodobieństwo obiektów (osób pod względem wzrostu oraz wagi) większe. Im mniejsza liczba (bliższa 0) tym niepodobieństwo mniejsze (większe podobieństwo).

Różne cechy oznaczają zazwyczaj różne jednostki. Cecha o dużym zróżnicowaniu będzie odgrywać większą rolę, więc jeżeli wzrost wyrazimy w *cm* a nie w *m* to odgrywać będzie większą rolę, analogicznie, jeżeli wagę wyrazić w *g* a nie *kg* to będzie odgrywała o wiele większą rolę. Aby usunąć wpływ jednostek oraz wpływ zróżnicowania cechy można stosować standaryzację



Na osi Y znajduje się odległość między łączonymi skupieniami.

```
Clustering Method: Group Average
Distance Metric: Euclidean

          Clusters Combined                    Stage First Appears    Next
Stage   Cluster 1  Cluster 2    Coefficient   Cluster 1  Cluster 2   Stage
-------------------------------------------------------------------------
    1           2          3            5,0           0          0       2
```

odległość między {2} a {3} wynosi 5

```
    2           2          4        16,0849           1          0       3
```

odległość między skupieniem {2; 3} a {4} to średnia z odległości między wszystkimi parami w których jeden element jest z jednego a drugi z drugiego skupienia, tj {2} a {4} = 18,03 oraz {3} a {4} = 14,14, i wynosi 16,08

```
    3           1          2        23,8006           0          2       0
-------------------------------------------------------------------------------
```

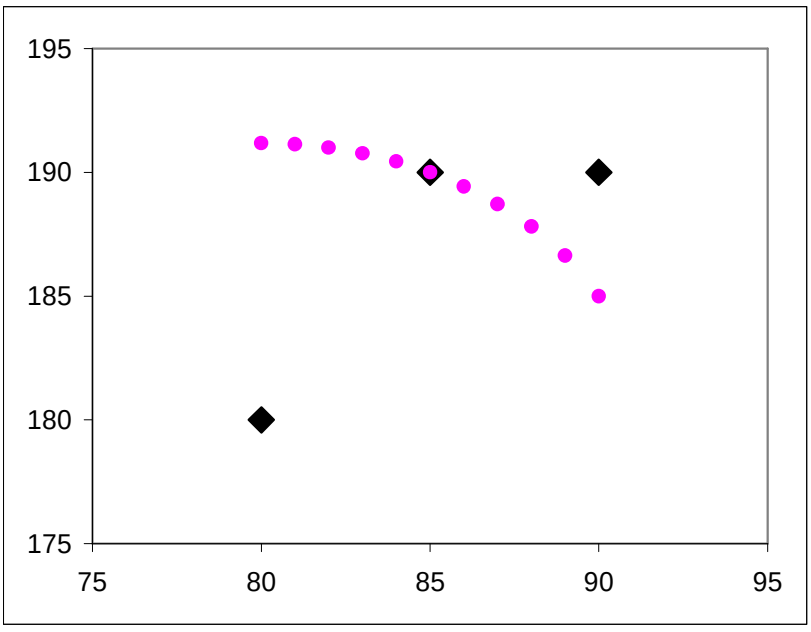
odległość między skupieniem {2; 3; 4} a {1} wynosi 23,80

Miary odległości dla cech ilościowych interwałowych

Miary niepodobieństwa (miary odległości) i miary podobieństwa

przykład: osoba x waży 80kg i ma 180cm wzrostu;
osoba y waży 85kg i ma 190cm wzrostu
osoba y waży 90kg i ma 190cm wzrostu

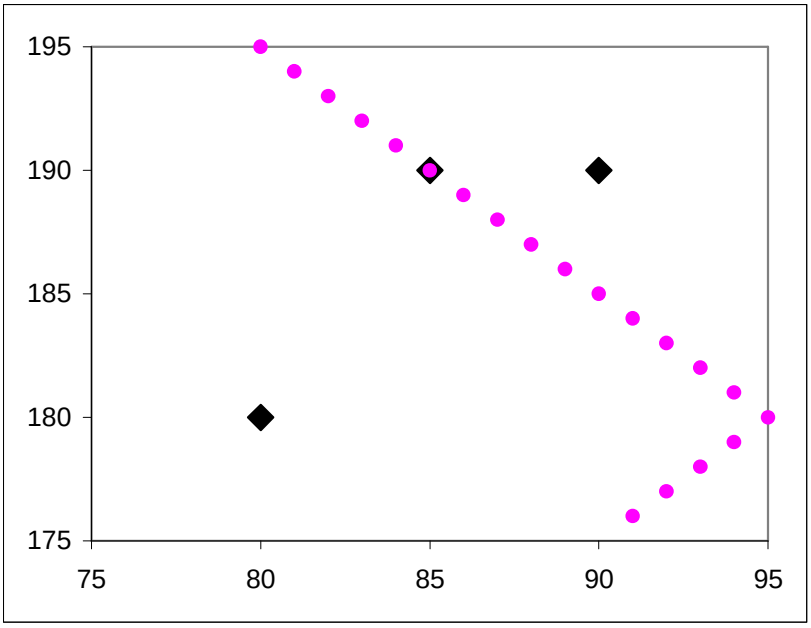
Odległość Euklidesa $E(x,y) = \{\Sigma_i (x_i - y_i)^2\}^{½}$
$E(x,y)=\{(80-85)^2+(180-190)^2\}^{½}=(25+100)^{½}=125^{½}$ =ok. 11,18
$E(x,y)=\{(80-90)^2+(180-190)^2\}^{½}=(100+100)^{½}=200^{½}$ =ok. 14,14
Jest to odległość w linii prostej między punktami.
Właściwości: wybór układu współrzędnych prostokątnych nie ma znaczenia dla pomiaru odległości. Stosowana bardzo często, szczególnie, gdy obserwowane jednostki miary są takie same.



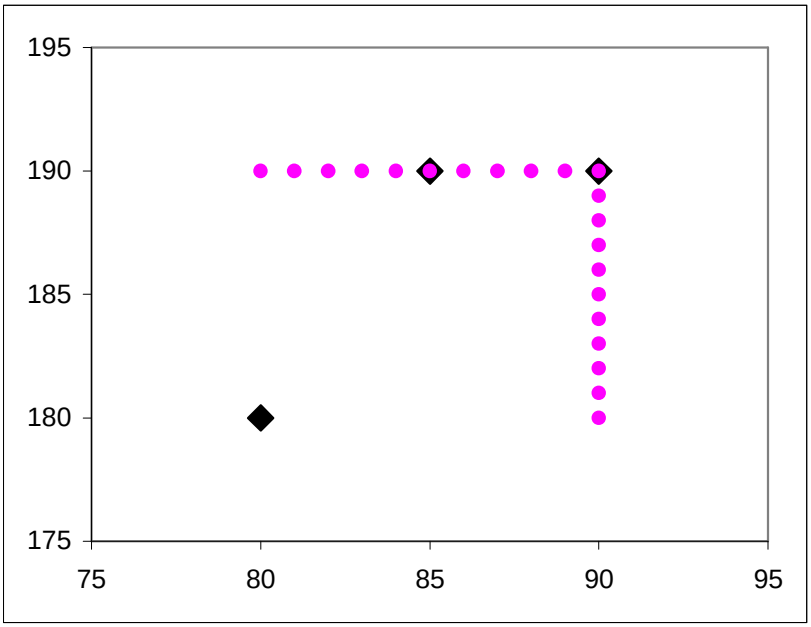
kwadrat odległości Euklidesa $E^2(x,y) = \{\Sigma_i (x_i - y_i)^2\}$
$E^2(x,y)=\{(80-85)^2+(180-190)^2\}=25+100=125$
$E^2(x,y)=\{(80-90)^2+(180-190)^2\}=100+100=200$
właściwości: wybór układu współrzędnych prostokątnych nie ma znaczenia dla pomiaru odległości. Stosowana bardzo

często, tam gdzie miara E. Przy większości sposobów grupowania (wyjątkiem metoda najbliższego sąsiada) nadaje większe znaczenie obiektom najbardziej niepodobnym.
(rys. jak dla E)

odległość miejska (city-block / manhattan distance)
$=\Sigma_i |x_i - y_i|$
$=|80-85|+|180-190|=5+10=15$
$=|80-90|+|180-190|=10+10=20$
Jest to suma różnic dla każdego kierunku układu współrzędnych
Właściwości: silnie związana z kierunkami osi układu współrzędnych. Nadaje nieco mniejsze znaczenie różnicy w tylko jednym kierunku osi współrzędnych przy podobieństwu w pozostałych kierunkach niż odległość E.



odległość Czebyszewa
$=\max_i |x_i - y_i|$
$=\max(|80-85|,|180-190|)=\max(5,10)=10$
$=\max(|80-90|,|180-190|)=\max(10,10)=10$
Jest to maksymalna różnica między punktami dla kierunków układu współrzędnych
Właściwości: silnie związana z kierunkami osi układu współrzędnych. Dla dwu obiektów nadaje większe znaczenie różnicy w jednym kierunku osi współrzędnych, w których różnica jest największa (różnica w pozostałych kierunkach osi układu współrzędnych nie ma znaczenia)



miary dla cechy o rozkładzie dwupunktowym (tj. 0/1)

Tabela łącznej liczebności (poprzez powtórzenia) występowania kombinacji cech 0/1

		cecha 2	
		1	0
cecha 1	1	a	b
	0	c	d

przykład:

cecha A	1	0	0	1
cecha B	1	0	0	0
cecha C	0	1	0	1

		cecha B	
		1	0
cecha A	1	1	1
	0	0	2

		cecha C	
		1	0
cecha A	1	1	1
	0	1	1

Manhattan, kwadrat Eulkidesa
E^2=b+c
odległość(A,B)=1 odległość(A,C)=2
Liczba przypadków, w których jest różnica między cechą 1 a cechą 2.
Właściwości: oba stany (0 i 1) traktowane są tak samo.

Miara podobieństwa genetycznego Jaccarda
Jaccard similarity =a/(a+b+c)
podobieństwo(A,B)=1/2 podobieństwo (A,C)=1/3
właściwości: równoczesne niewystępowanie cechy 1 i 2 dla danego powtórzenia nie stanowi podobieństwa między cechami, więc jest stosowana gdy stanem zwykłym, zazwyczaj występującym, jest jeden stan, oznaczony jako (0).
Jako miarę odległości stosuje się 1–Jaccard(A,B)=(b+c)/(a+b+c)
odległość(A,B)=1/2 odległość(A,C)=2/3

podobieństwo Dice'a (Sorensena, Dice's coefficient)
Dice's similarity measure = 2a/(2a+b+c)
podobieństwo(A,B)=2/3 podobieństwo (A,C)=2/4
właściwości: podobna do Jaccarda, ale większy nacisk na równoczesne występowanie cech $D = 2J / (1 + J)$ i $J = D / (2 – D)$.

metody aglomeracji

Metody aglomeracji (określenie odległości między grupami obiektów, gdy znane są odległości między pojedynczymi obiektami)

- Odległość średniego wiązania to średnia z odległości między jednym z obiektów grupy A a jednym z obiektów z grupy B. *The mean distance between elements of each cluster (also called average linkage clustering, used e.g. in UPGMA)*:

$$\frac{1}{|\mathcal{A}| \cdot |\mathcal{B}|} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{B}} d(x, y).$$

- metoda najdalszego sąsiada – odległość między grupami to odległość między najbardziej oddalonymi obiektami (jeden z grupy A a drugi z grupy B). *The maximum distance between elements of each cluster (also called complete linkage clustering)*:

  $\max\{\, d(x, y) : x \in \mathcal{A}, y \in \mathcal{B} \,\}.$
  Stosowana, gdy chcemy mieć ograniczenie na odległość dowolnych dwu elementów skupienia, tj wszystkie elementy są sobie bliższe niż pewna wartość krytyczna.

- metoda najbliższego sąsiada – odległość między grupami to odległość między najbliższymi obiektami (jeden z grupy A a drugi z grupy B). *The minimum distance between elements of each cluster (also called single-linkage clustering)*:

  $\min\{\, d(x, y) : x \in \mathcal{A}, y \in \mathcal{B} \,\}.$
  Stosowana, gdy chcemy mówić o „łańcuchach" podobnych obiektów, tj. można przejść od każdego obiektu w grupie do każdego innego poprzez kilka obiektów pośrednich leżących każdy blisko następnego.
  Właściwości: zazwyczaj generuje jedną dużą grupę obiektów centralnych i kilka małych podgrup.

- Metoda Warda. *The increase in variance for the cluster being merged (Ward's criterion)*. Podobna do średniego wiązania, ale z poprawką na wielkość grupy. Łączone są takie skupienia, aby dać najmniejszą łączną wariancję wewnątrzgrupową. Odległość określona jako suma kwadratów zmienności wewnątrzgrupowej połączonego klastra minus suma kwadratów zmienności wewnątrzgrupowej obu klastów /SS(A,B)–SS(A)–SS(B)/. Właściwości: zazwyczaj generuje grupy o podobnej liczebności.

$$\frac{(\text{wektor średni dla grupy A} - \text{wektor średni dla grupy B})^2}{1/\text{liczebność grupy A} + 1/\text{liczebność grupy B}}$$

Przykład dla danych dwuwymiarowych – wzrost i waga ludzi (dane poddane standaryzacji)

dane obserwowane

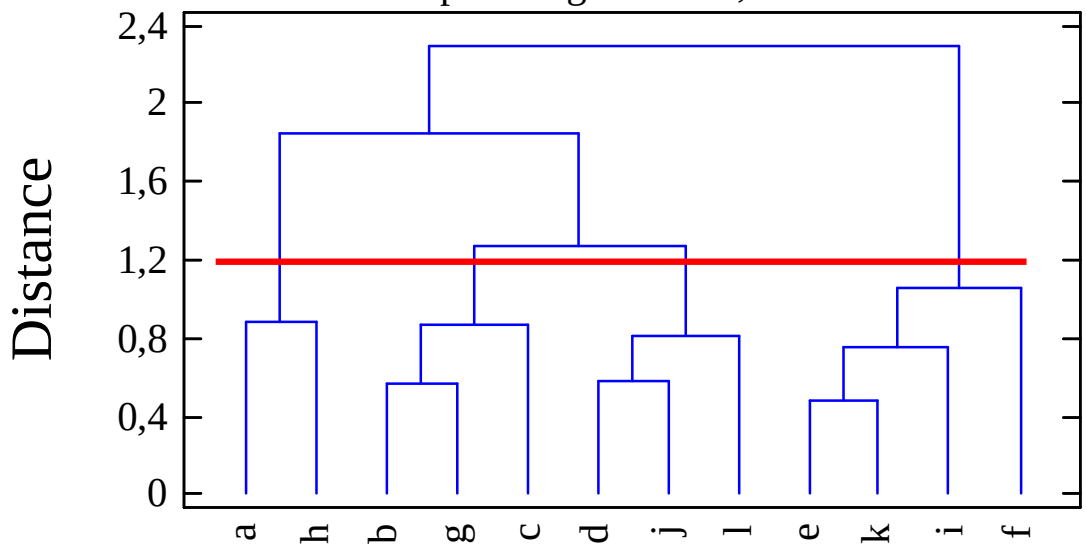
osoba	a	b	c	d	e	f	g	h	i	j	k	l
wzrost (cm)	154	162	171	178	184	192	168	159	175	179	182	184
waga (kg)	60	62	74	71	96	90	65	48	95	62	89	76

Średnia dla wzrostu wynosi 174 zaś odchylenie standardowe to 11,46. Dla wagi są to wartości 74 i 15,55 odpowiednio. Po standaryzacji dane są następujące:

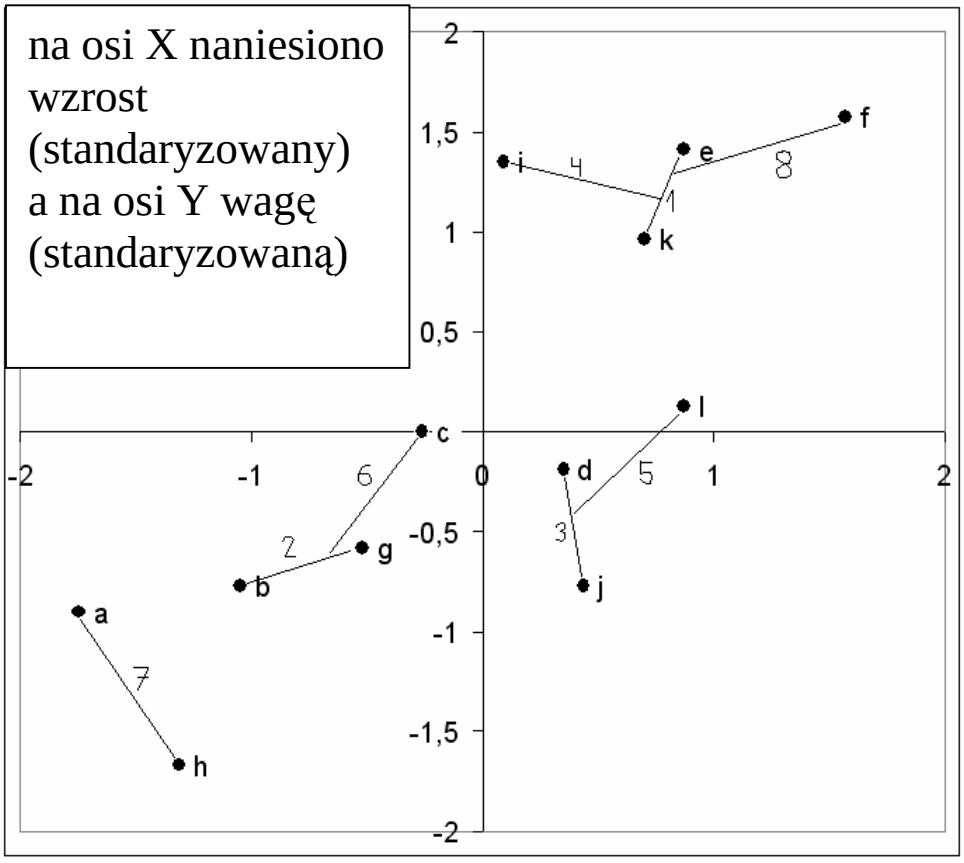
osoba	a	b	c	d	e	f	g	h	i	j	k	l
wzrost (w cm)	-1,75	-1,05	-0,26	0,35	0,87	1,57	-0,52	-1,31	0,09	0,44	0,70	0,87
waga (w kg)	-0,90	-0,77	0,00	-0,19	1,41	1,03	-0,58	-1,67	1,35	-0,77	0,96	0,13

Dendrogram

Group Average Method,Euclidean



na osi X naniesiono wzrost (standaryzowany) a na osi Y wagę (standaryzowaną)

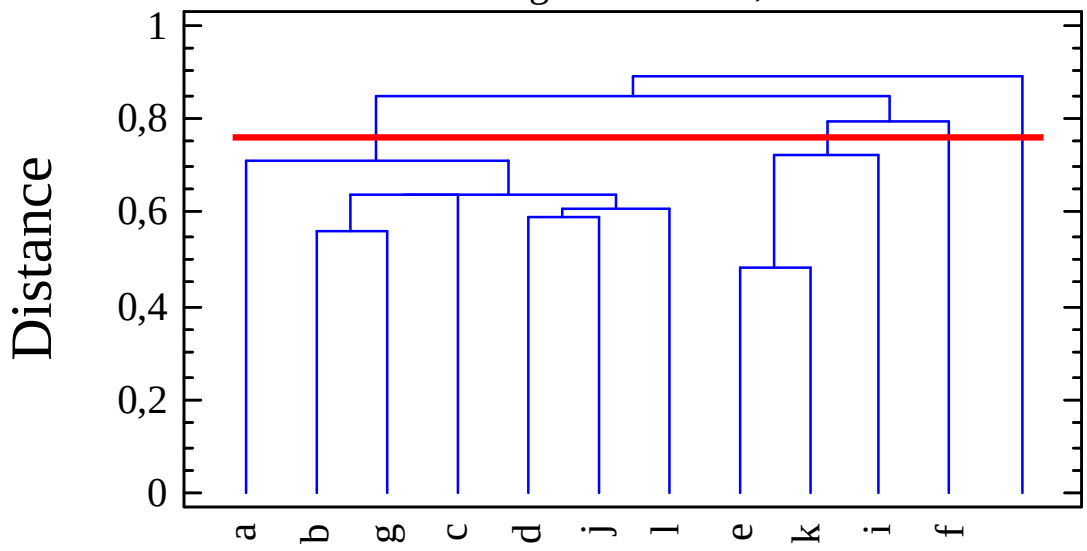


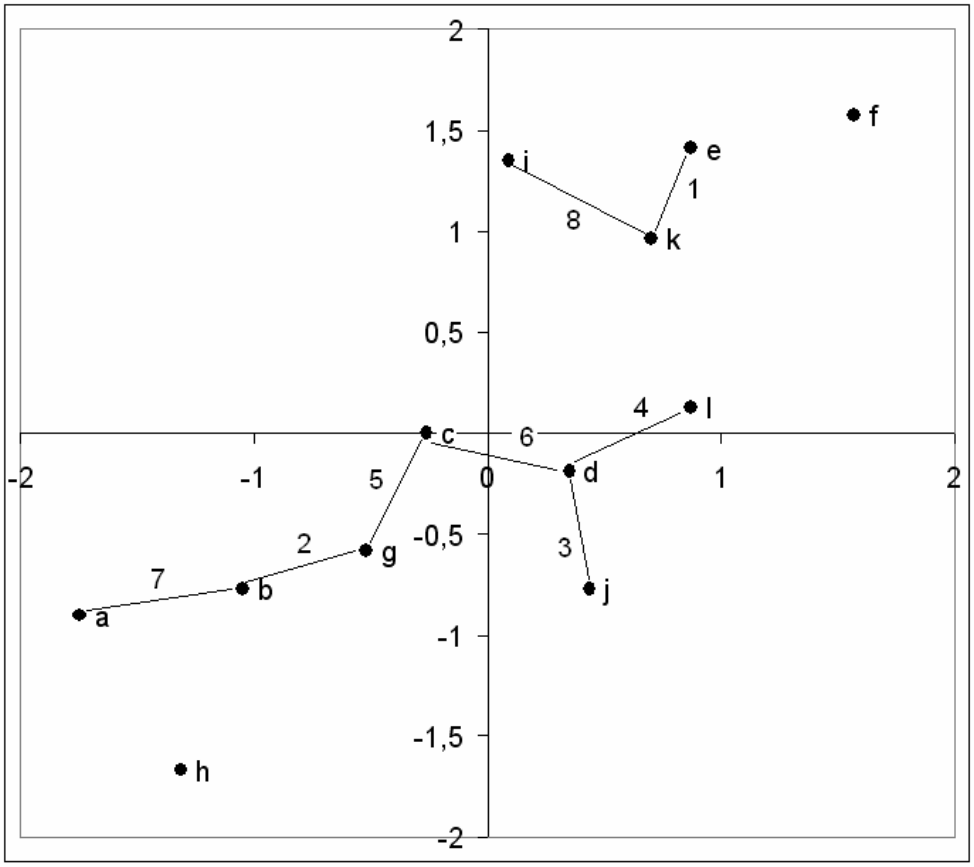
grupa {a,h} to osoby niskie i lekkie
grupa {b,g,h} to osoby odrobinę poniżej średniej zarówno pod względem wzrostu jak i wagi
grupa {d,j,l} to osoby wysokie ale chude.
grupa {k,e,i,f} to osoby wysokie i cięższe

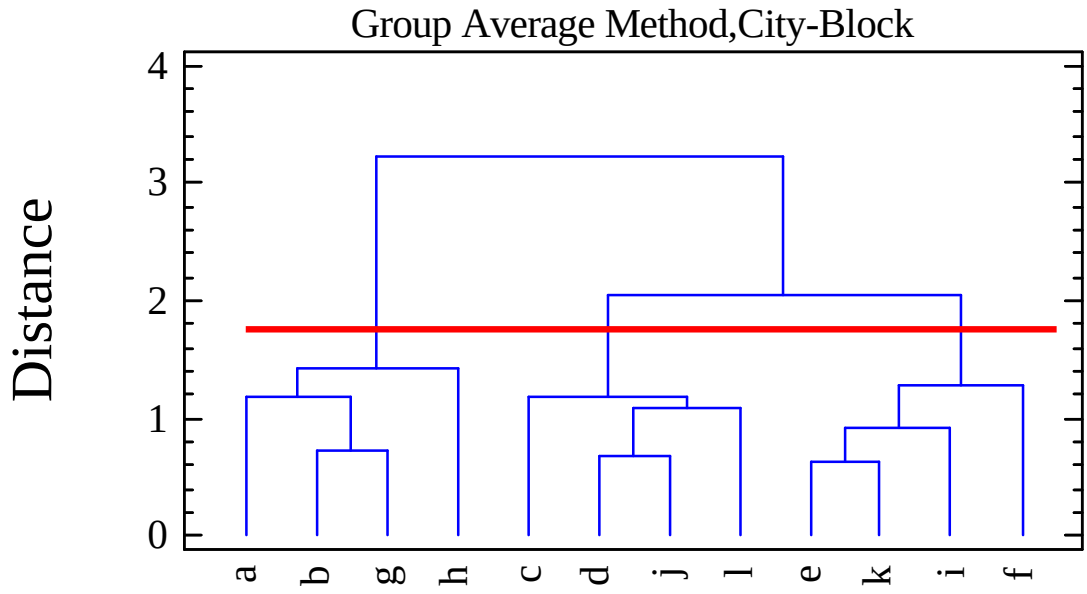
h

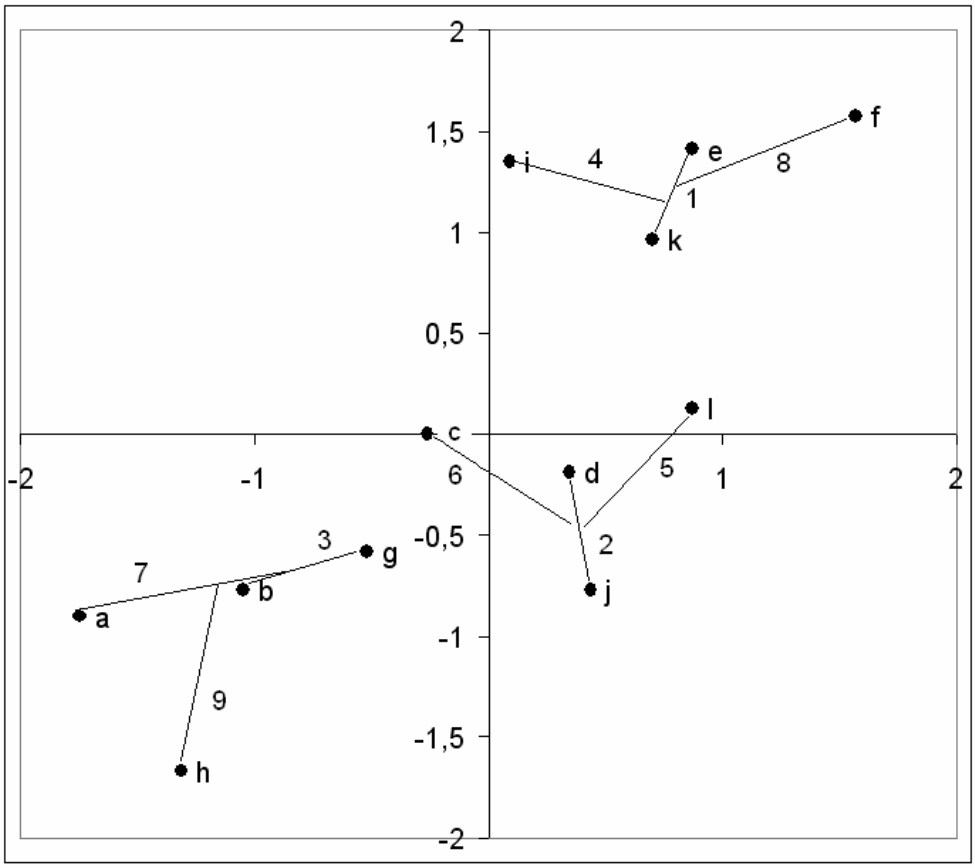
Dendrogram

Nearest Neighbor Method,Euclidean









At the first step, when each object represents its own cluster, the distances between those objects are defined by the chosen distance measure. However, once several objects have been linked together, how do we determine the distances between those new clusters? In other words, we need a linkage or amalgamation rule to determine when two clusters are sufficiently similar to be linked together. There are various possibilities: for example, we could link two clusters together when *any* two objects in the two clusters are closer together than the respective linkage distance. Put another way, we use the "nearest neighbors" across clusters to determine the distances between clusters; this method is called *single linkage*. This rule produces "stringy" types of clusters, that is, clusters "chained together" by only single objects that happen to be close together. Alternatively, we may use the neighbors across clusters that are furthest away from each other; this method is called *complete linkage*. There are numerous other linkage rules such as these that have been proposed.

Single linkage (nearest neighbor). As described above, in this method the distance between two clusters is determined by the distance of the two closest objects (nearest neighbors) in the different clusters. This rule will, in a sense, *string* objects together to form clusters, and the resulting clusters tend to represent long "chains."

Complete linkage (furthest neighbor). In this method, the distances between clusters are determined by the greatest distance between any two objects in the different clusters (i.e., by the "furthest neighbors"). This method usually performs quite well in cases when the objects actually form naturally distinct "clumps." If the clusters tend to be somehow elongated or of a "chain" type nature, then this method is inappropriate.

Unweighted pair-group average. In this method, the distance between two clusters is calculated as the average distance between all pairs of objects in the two different clusters. This method is also very efficient when the objects form natural distinct "clumps," however, it performs equally well with elongated, "chain" type clusters. Note that in their book, Sneath and Sokal (1973) introduced the abbreviation UPGMA to refer to this method as *unweighted pair-group method using arithmetic averages.*

Weighted pair-group average. This method is identical to the *unweighted pair-group average* method, except that in the computations, the size of the respective clusters (i.e., the number of objects contained in them) is used as a weight. Thus, this method (rather than the previous method) should be used when the cluster sizes are suspected to be greatly uneven. Note that in their book, Sneath and Sokal (1973) introduced the abbreviation *WPGMA* to refer to this method as *weighted pair-group method using arithmetic averages*.

Unweighted pair-group centroid. The *centroid* of a cluster is the average point in the multidimensional space defined by the dimensions. In a sense, it is the *center of gravity* for the respective cluster. In this method, the distance between two clusters is determined as the difference between centroids. Sneath and Sokal (1973) use the abbreviation *UPGMC* to refer to this method as *unweighted pair-group method using the centroid average*.

Weighted pair-group centroid (median). This method is identical to the previous one, except that weighting is introduced into the computations to take into consideration differences in cluster sizes (i.e., the number of objects contained in them). Thus, when there are (or we suspect there to be) considerable differences in cluster sizes, this method is preferable to the previous one. Sneath and Sokal (1973) use the abbreviation *WPGMC* to refer to this method as *weighted pair-group method using the centroid average*.

Ward's method. This method is distinct from all other methods because it uses an analysis of variance approach to evaluate the distances between clusters. In short, this method attempts to minimize the Sum of Squares (SS) of any two (hypothetical) clusters that can be formed at each step. Refer to Ward (1963) for details concerning this method. In general, this method is regarded as very efficient, however, it tends to create clusters of small size.