Analiza skupień

Waldemar Wołyński, Tomasz Górecki

Wydział Matematyki i Informatyki UAM Poznań

6 marca 2013

Idea:

Analiza skupień jest narzędziem analizy danych służącym do grupowania n obiektów (jednostek) w K niepustych, rozłącznych i możliwie "jednorodnych" grup - skupień. Obiekty należące do danego skupienia powinny być "podobne" od siebie (używa się w tym celu różnych miar podobieństwa, a w zasadzie niepodobieństwa obiektów), a obiekty należące do różnych skupień powinny być z kolei możliwie mocno "niepodobne" do siebie.
Głównym celem tej analizy jest wykrycie z zbiorze danych, tzw. "naturalnych" skupień, czyli skupień, które dają się w sensowny sposób interpretować.

"Naiwne" rozwiązanie zagadnienia AS:

Krok 1: Wybieramy kryterium optymalnego podziału obiektów.

Krok 2: Ustalamy liczbę skupień K.

Krok 3: Sprawdzamy wszystkie możliwe podziały zbioru n obiektów na K podzbiorów i wybieramy optymalny.

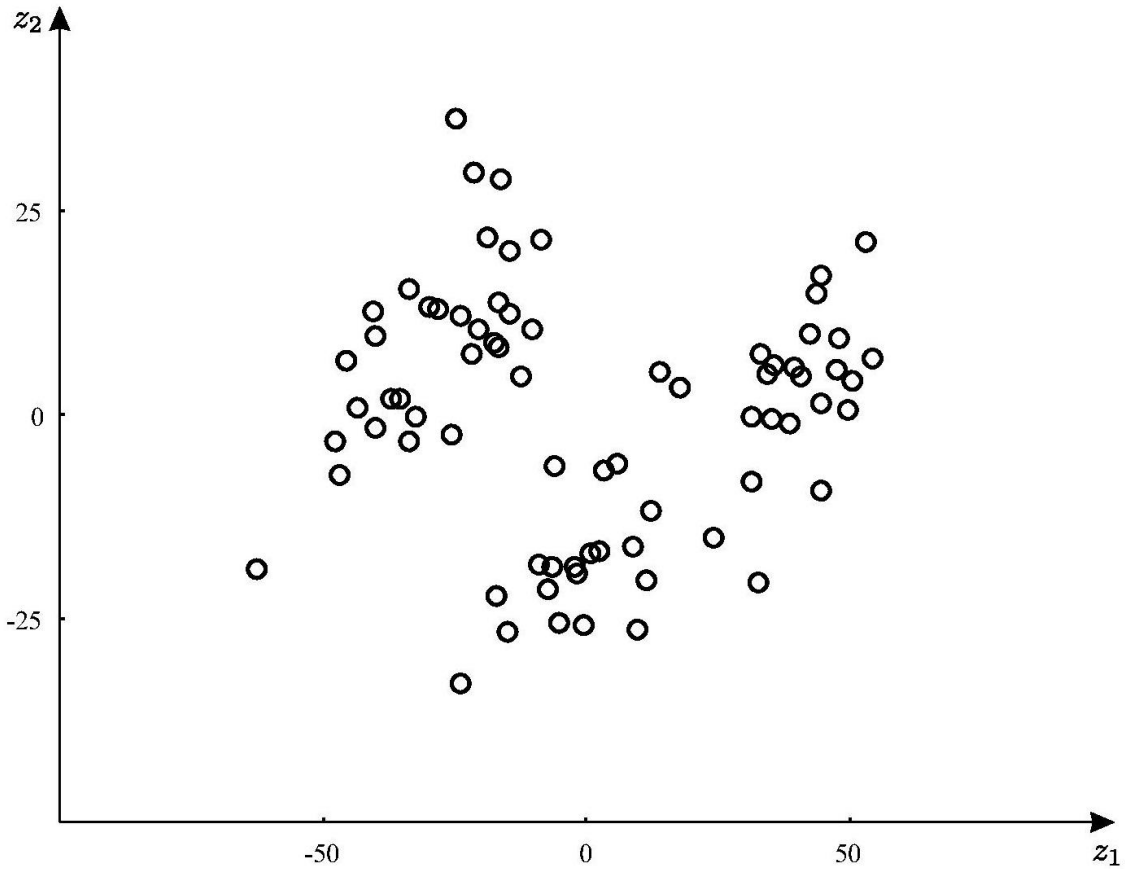
Ile jest wszystkich możliwych podziałów?

$$\frac{1}{K!}\sum_{k=1}^{K}(-1)^{K-k}\binom{K}{k}k^n.$$

Np. dla 100 obiektów i czterech skupień jest to liczba rzędu 10^{58}!!!

Obiekt (jednostka) opisana za pomocą wektora $\mathbf{p}$ cech

Przykład – Flea beetles



Dane pochodzą z obserwacji 6 cech na 74 okazach chrząszczy skaczących. Lubishew (1962).

Algorytmy hierarchiczne

Ideą algorytmów hierarchicznych jest wyznaczanie skupień poprzez łączenie (aglomerację) powstałych, w poprzednich krokach algorytmu, mniejszych skupień. Inne wersje tych algorytmów zamiast idei łączenia skupień, bazują na pomyśle ich dzielenia.

Algorytm aglomeracyjny

1. W pierwszym kroku każdy z obiektów tworzy oddzielne skupienie. Zatem skupień tych jest n.
2. Łączymy (wiążemy ze sobą) dwa najbardziej podobne do siebie skupienia, zmniejszając w ten sposób liczbę skupień o jeden.
3. Powtarzamy krok drugi do momentu uzyskania zadeklarowanej, końcowej liczby skupień K lub do połączenia wszystkich obiektów w jedno skupienie.

Miary niepodobieństwa obiektów

1. Odległość Minkowskiego:

$$\rho(\mathbf{x}_i, \mathbf{x}_j) = \left(\sum_{l=1}^{p} |x_{il} - x_{jl}|^q \right)^{1/q}, \quad q \geq 1.$$

2. Odległość Mahalanobisa:

$$\rho(\mathbf{x}_i, \mathbf{x}_j) = ((\mathbf{x}_i - \mathbf{x}_j)' \mathbf{S}^{-1} (\mathbf{x}_i - \mathbf{x}_j))^{1/2}.$$

3. Współczynnik podobieństwa Sneatha:

$$\rho(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{p} \sum_{l=1}^{p} I(x_{il} \neq x_{jl}).$$

Jeżeli dane są miary niepodobieństwa $\rho(x_{il}, x_{jl})$, $(l = 1, \ldots, p)$ oddzielnie dla każdej z p cech, to za *całkowitą miarę niepodobieństwa pomiędzy obiektami* możemy przyjąć kombinację wypukłą miar brzegowych postaci

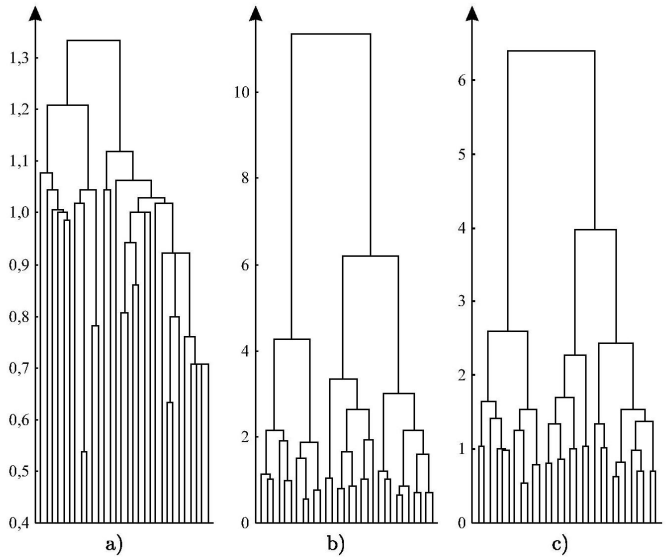
$$\rho(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{p} w_l^2 \rho(x_{il}, x_{jl}), \quad \sum_{l=1}^{p} w_l^2 = 1.$$

Metody wiązania skupień

1. Metoda pojedynczego wiązania (najbliższego sąsiedztwa). Miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako najmniejsza miara niepodobieństwa między dwoma obiektami należącymi do różnych skupień.
2. Metoda pełnego wiązania (najdalszego sąsiedztwa). Miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako największa miara niepodobieństwa między dwoma obiektami należącymi do różnych skupień.
3. Metoda średniego wiązania. Miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako średnia miara niepodobieństwa między wszystkimi parami obiektów należących do różnych skupień.
4. Metoda Warda. Miara niepodobieństwa pomiędzy dwoma skupieniami jest określona jako suma kwadratów odchyleń wewnątrz tych skupień.

Dendrogram

Graficzną ilustracją przebiegu aglomeracji jest wykres zwany dendrogramem. Jest to (binarne) drzewo którego węzły reprezentują skupienia, a liście pojedyncze obiekty. Liście umieszczone są na poziomie zerowym, pozostałe węzły drzewa umieszczone są na wysokości odpowiadającej mierze niepodobieństwa pomiędzy skupieniami reprezentowanymi przez węzły potomki.



a) metoda pojedynczego wiązania, b) metoda pełnego wiązania, c) metoda średniego wiązania.

Przyporządkowanie n obiektów do zadanej liczby skupień K, odbywa się niezależnie dla każdej wartości K - nie bazując na wyznaczonych wcześniej mniejszych lub większych skupieniach.
Najbardziej popularnym, niehierarchicznym algorytmem analizy skupień jest algorytm K-średnich. Główną ideą tego algorytmu jest taka alokacja obiektów, która minimalizuje zmienność wewnątrz powstałych skupień, a co za tym idzie maksymalizuje zmienność pomiędzy skupieniami.

Oznaczenia:

C_K – funkcja, która każdemu obiektowi (dokładnie jego numerowi), przyporządkowuje numer skupienia do którego jest on przyporządkowany (przy podziale na K skupień),

$W(C_K)$ – macierz zmienności wewnątrz skupień,

$B(C_K)$ – macierz zmienności pomiędzy skupieniami.

W algorytmie K-średnich minimalizujemy ślad macierzy zmienności wewnątrz skupień. Jeżeli C_K^* jest funkcją realizującą optymalny podział n obiektów na K skupień, to

$$C_K^* = \min_{C_K} \mathrm{tr}[W(C_K)] = \min_{C_K} \sum_{k=1}^{K} \sum_{C_K(i)=k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)'(\mathbf{x}_i - \bar{\mathbf{x}}_k).$$

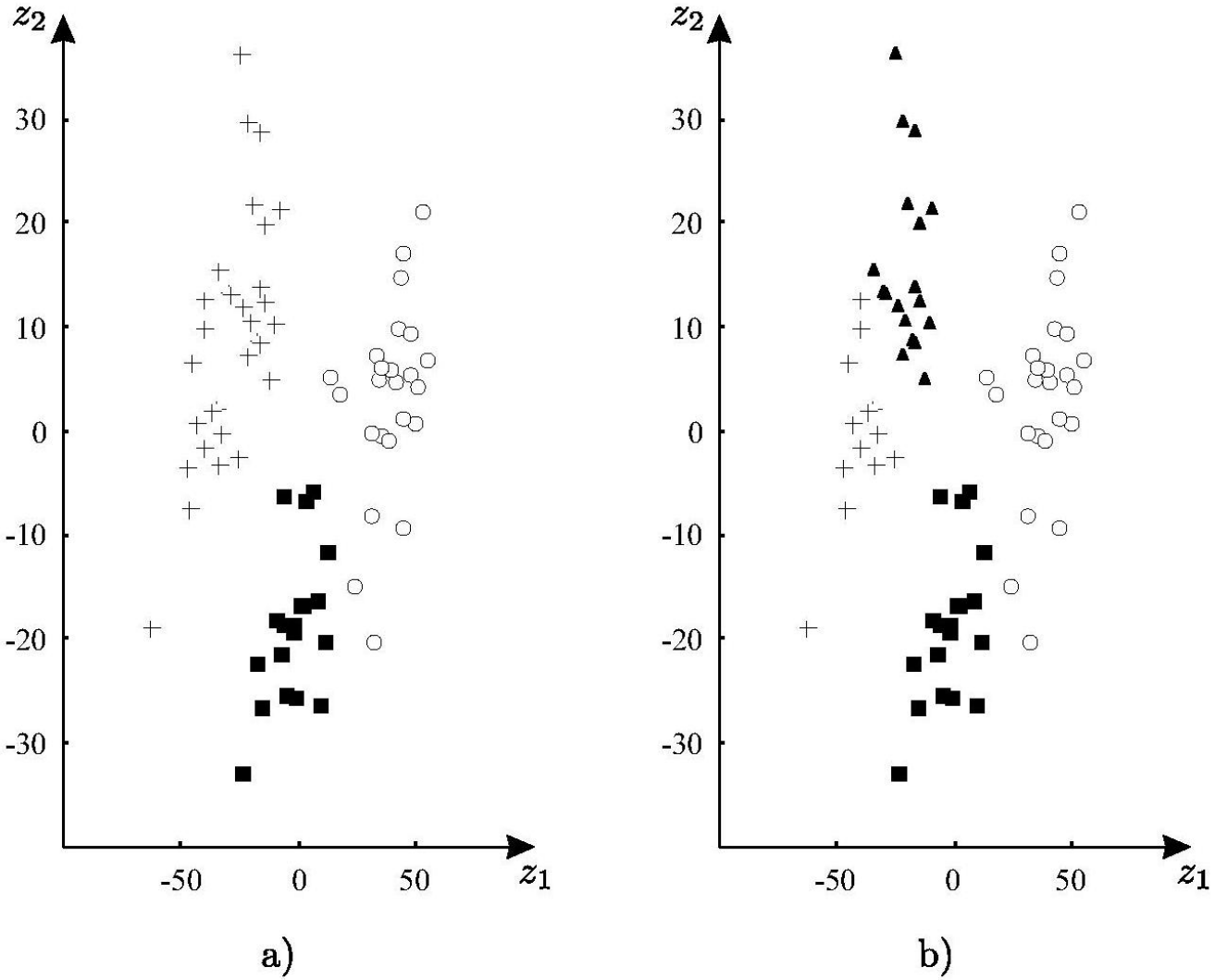
Algorytm K-średnich

1. W losowy sposób rozmieszczamy n obiektów w K skupieniach. Niech funkcja $C_K^{(1)}$ opisuje to rozmieszczenie.
2. Dla każdego z K skupień obliczamy wektory średnich $\bar{\mathbf{x}}_k$.
3. Rozmieszczamy ponownie obiekty w K skupieniach, w taki sposób że

$$C_K^{(l)}(i) = \arg \min_{1 \le k \le K} (\mathbf{x}_i - \bar{\mathbf{x}}_k)'(\mathbf{x}_i - \bar{\mathbf{x}}_k).$$

4. Powtarzamy kroki drugi i trzeci aż do momentu, gdy przyporządkowanie obiektów do skupień pozostanie niezmienione, tzn. aż do momentu, gdy $C_K^{(l)} = C_K^{(l-1)}$.

Skupienia wyznaczone metodą K-średnich, a) $K = 3$, b) $K = 4$

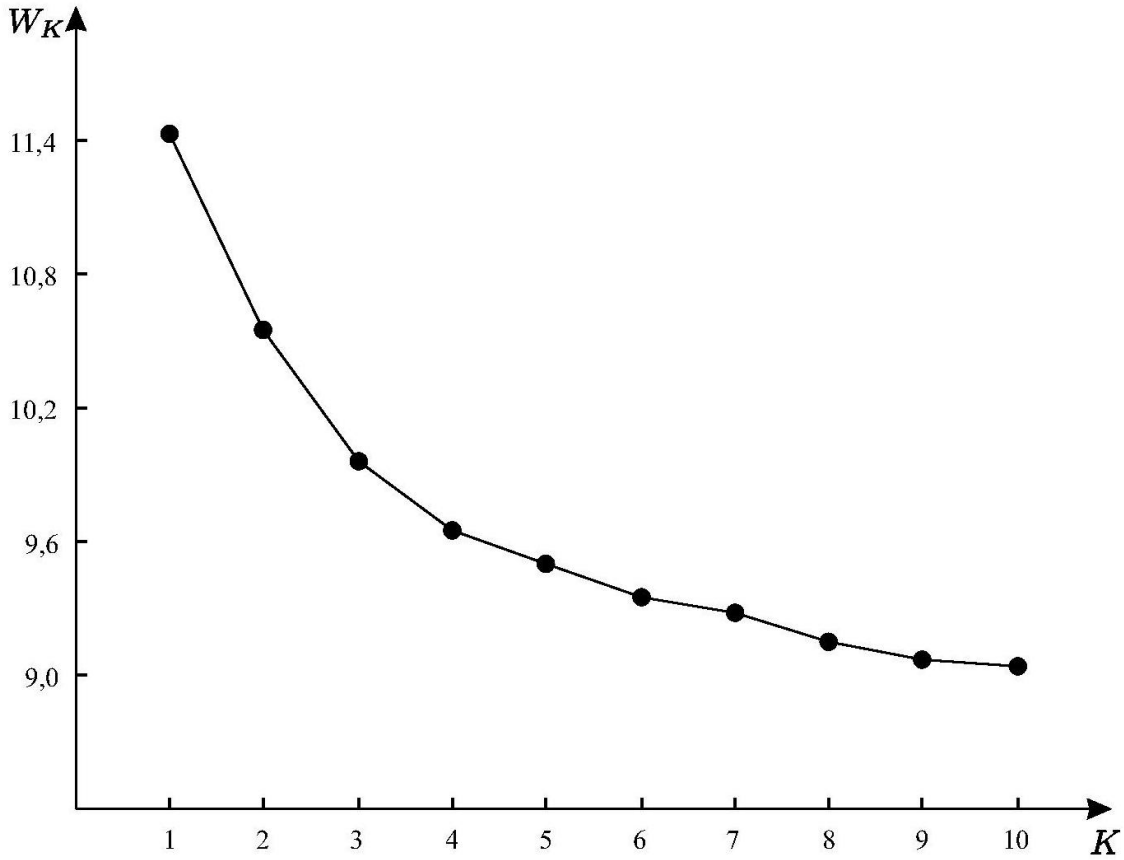


Inne algorytmy analizy skupień

1. Taksonomia wrocławska (algorytm ten jest równoważny algorytmowi aglomeracyjnemu opartemu na metodzie pojedynczego wiązania).
2. Algorytm K-medoid (odmiana algorytmu K-średnich dostosowana zwłaszcza do danych jakościowych).
3. EM-clustering (zakładamy, że rozkład prawdopodobieństwa analizowanych cech daje się opisać za pomocą rozkładu prawdopodobieństwa będącego mieszaniną K rozkładów odpowiadających podziałowi na K skupień).
4. Sieci samoorganizujące się (SOM).

Optymalna liczba skupień

Minimalizacja zmienności wewnątrz skupień.



Wartości $W_K = \log(\mathrm{tr}(W(C_K)))$ *dla metody K-średnich.*

Optymalna liczba skupień

- Indeks Calińskiego-Harabasza (1974):

$$CH(K) = \frac{\mathrm{tr}(B(C_K))/(K-1)}{\mathrm{tr}(W(C_K))/(n-K)}.$$

Optymalną wartość K dobieramy tak, aby zmaksymalizować indeks $CH(K)$.

Optymalna liczba skupień

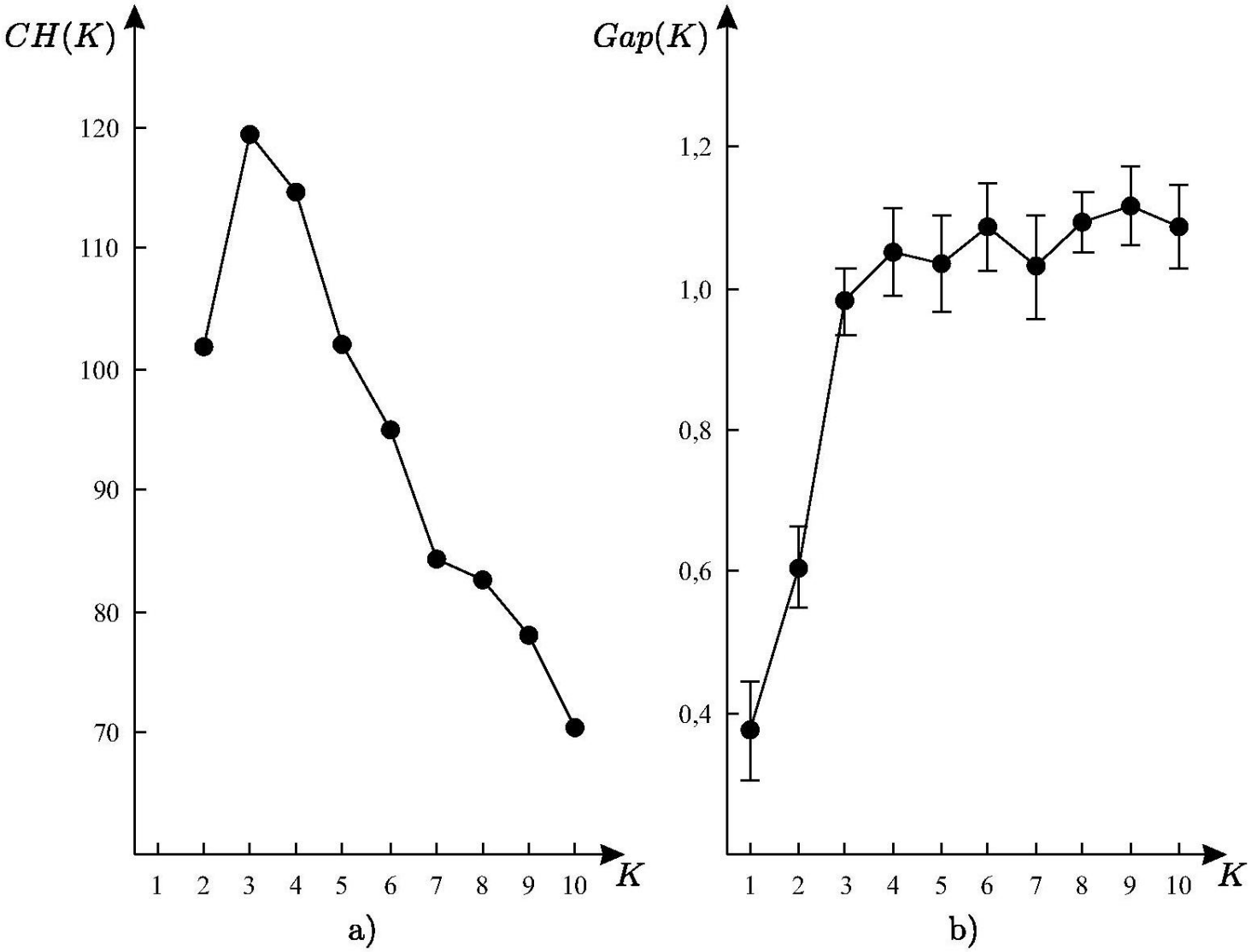
- Statystyka odstępu (Hastie, Tibshirani, Walther, 2001):

$$Gap(K) = W_K^* - W_K,$$

przy czym, w celu uzyskania wartości W_K^*, dla każdej z p-zmiennych generujemy n obserwacji z rozkładu jednostajnego na przedziale wyznaczonym przez zakres wartości tej zmiennej w pierwotnym zbiorze danych. Symulację tę powtarzamy B razy (zazwyczaj $B = 20$) i dla tak wyznaczonego, sztucznego zbioru danych obliczamy wartości W_K^b $(b = 1, 2, \ldots, B)$.
Niech W_K^* i s_K^* oznaczają średnią i odchylenie standardowe obliczone na podstawie wartości $W_K^1, \ldots, W_K^B$. Ponadto, niech $s_K = \sqrt{1 - (1/B)} s_K^*$.
Jako optymalną liczbę skupień przyjmujemy najmniejsze K dla którego

$$Gap(K) \geq Gap(K+1) - s_{K+1}.$$

Optymalna liczba skupień



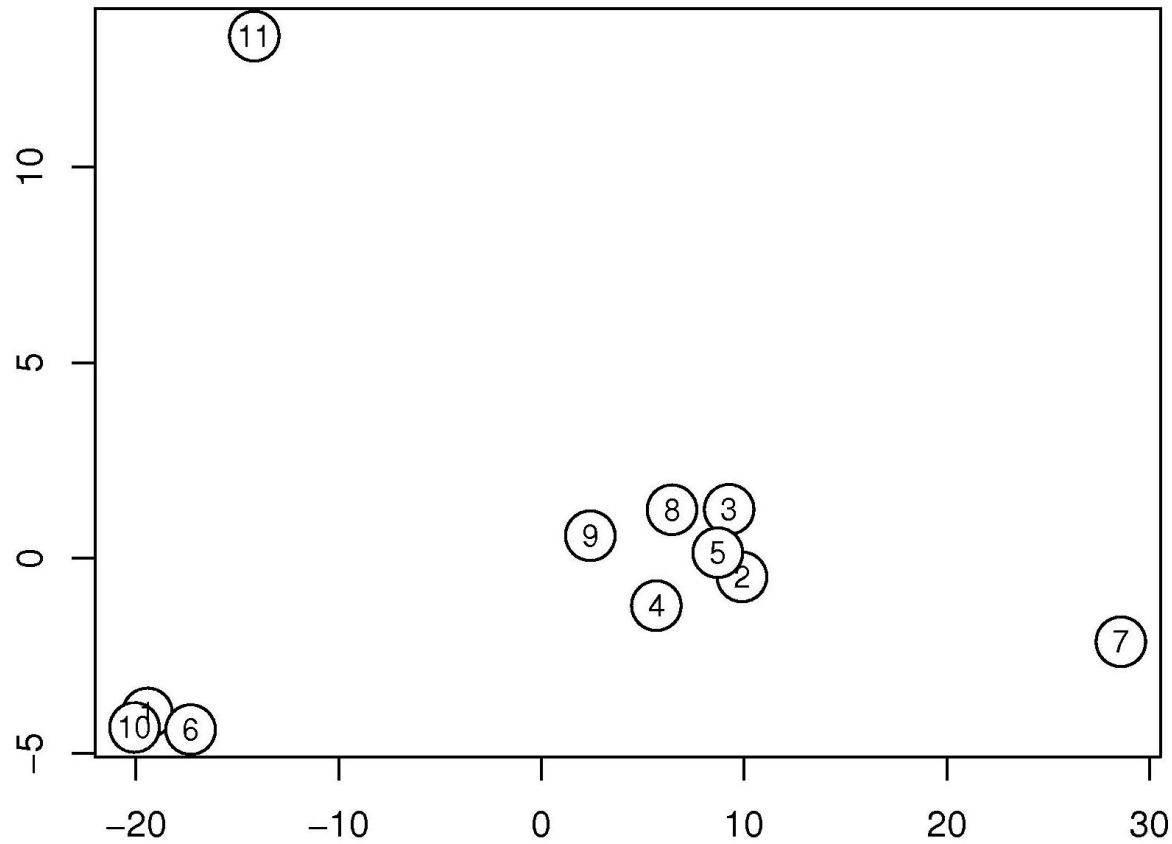
Optymalna liczba skupień

- Współczynnik zarysu (Kaufman, Rousseeuw, 1990).
- Indeks Hartigana (1975).
- Indeks Daviesa-Bouldina (1979).
- Indeks Krzanowskiego-Lai (1988).

Pakiet ClusterSim(R) autorstwa Marka Walesiaka i Andrzeja Dudka, pozwala na obliczenie 8 różnych indeksów związanych z wyznaczaniem optymalnej liczby skupień.

Analiza skupień dla populacji złożonych z obiektów (jednostek) opisanych za pomocą wektora $\mathbf{p}$ cech

Przykład – Słoneczniki



Dane pochodzą z badań hodowlanych nad rodami słonecznika prowadzonych w Stacji Hodowli Roślin IHAR w Borowie. Liczba rodów słonecznika - 11, liczba cech - 5.

Miary niepodobieństwa populacji

1. Odległość euklidesowa:

$$\rho(\pi_i, \pi_j) = ((\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)'(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j))^{1/2}.$$

2. Odległość Mahalanobisa:

$$\rho(\pi_i, \pi_j) = ((\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j))^{1/2}.$$

3. Odległość Bhattacharyya:

$$\rho(\pi_i, \pi_j) = \frac{1}{8}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)'\mathbf{S}^{-1}(\bar{\mathbf{x}}_j - \bar{\mathbf{x}}_j) + \frac{1}{2}\ln\left(\frac{|\mathbf{S}|}{|\mathbf{S}_i||\mathbf{S}_j|}\right),$$

gdzie

$$\mathbf{S} = \frac{\mathbf{S}_i + \mathbf{S}_j}{2}.$$

Taksonomia wrocławska

1. Rozpinamy na zbiorze n obiektów najkrótszy dendryt, zbudowany na bazie wybranej odległości (miary niepodobieństwa) pomiędzy obiektami.
2. Wydzielamy skupienia poprzez usunięcie najdłuższych jego krawędzi. Dokładnie, niech ρ_i oznacza wagę i-tej krawędzi dendrytu. Obliczamy średnią $\bar{\rho}$ i odchylenie standardowe s_ρ wag wszystkich jego krawędzi, a następnie usuwamy te z nich dla których

$$\rho_i > \bar{\rho} + cs_\rho,$$

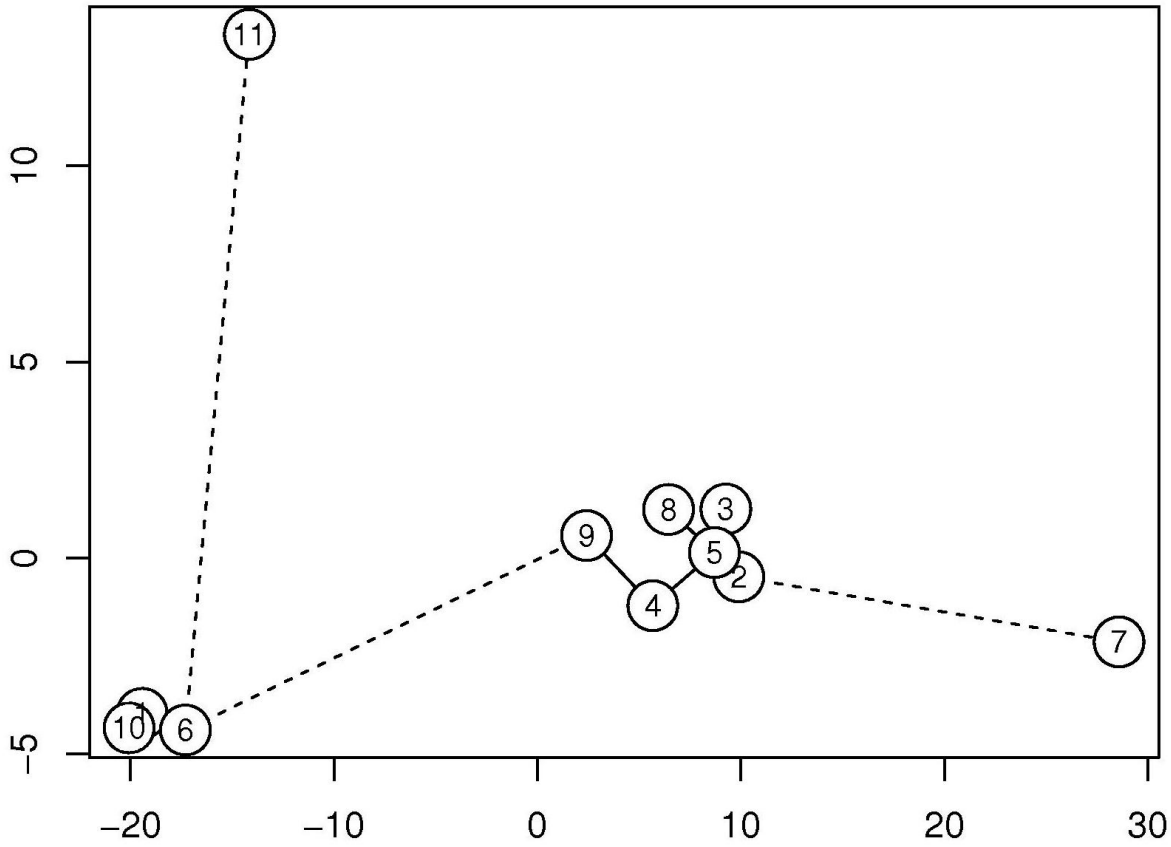
przy czym stałą c przyjmujemy zazwyczaj z przedziału $[1, 3]$.

Konstrukcja najkrótszego dendrytu

Algorytm Kruskala

1. Wybieramy krawędź o minimalnej wadze.
2. Z pozostałych krawędzi wybieramy tę o najmniejszej wadze, która nie prowadzi do cyklu (z krawędzi o jednakowych wagach wybieramy dowolną).
3. Powtarzamy krok drugi, aż do uzyskania najkrótszego dendrytu.

Skupienia wyznaczone metodą taksonomii wrocławskiej



Minimalny dendryt spinający. Usunięto krawędzie przyjmując c=1.25.

Progowa wartość odległości pomiędzy obiektami

Załóżmy, że $\pi_i \sim N_p(\boldsymbol{\mu}_i \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} > 0$.
Miarę niepodobieństwa populacji π_i oraz π_j definiujemy następująco:

$$\Delta_{ij}^2 = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$$

Jest to kwadrat odległości Mahalanobisa.

Oceną miary Δ_{ij}^2 jest wielkość $\rho(\pi_i, \pi_j)$ postaci

$$\rho(\pi_i, \pi_j) = (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j).$$

Progowa wartość odległości pomiędzy obiektami

Niech H_0 będzie hipotezą postaci:

$$H_0 : \ \Delta_{12} = \Delta_{23} = \cdots = \Delta_{K-1,K} = 0.$$

Hipotezę H_0 możemy zapisać jako przekrój hipotez

$$H_{ij} : \ \Delta_{ij} = 0, \ \ i, j = 1, 2, \ldots, K, \ \ j \neq i.$$

Hipotezy H_{ij} będziemy nazywać hipotezami implikowanymi przez hipotezę H_0. Jeżeli hipotezę H_0 odrzucimy, to możemy dokonać porównań wielokrotnych między k populacjami, tj. możemy zweryfikować $K(K-1)/2$ hipotez H_{ij} o braku istotności różnic między populacjami, przy czym miarą różnicy dwóch populacji jest ich odległość Δ_{ij}.
Do weryfikacji tych hipotez możemy zastosować jednoczesną procedurę testową podaną przez Gabriela (1968).

Progowa wartość odległości pomiędzy obiektami

Hipoteza $H_0: \ \Delta_{12} = \Delta_{23} = \cdots = \Delta_{K-1,K} = 0$ jest równoważna hipotezie $H_0': \ \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_K$.

Hipotezę H_0 możemy weryfikować za pomocą jednej z dwóch statystyk: Λ lub T^2.

Załóżmy, że użyjemy statystyki Λ.

- Jeżeli $m_E \geq p$ oraz $m_H \geq p$, to

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} \sim \Lambda_{p,m_H,m_E},$$

  gdzie $m_H = K-1$, $m_E = n-K$.
- Jeżeli $m_E \geq p > m_H$, to $\Lambda \sim \Lambda_{m_H,p,m_E+m_H-p}$.

Hipotezę H_0 odrzucamy wówczas, gdy $\Lambda \geq \Lambda_{p,m_H,m_E}(\alpha)$, gdzie $\mathrm{P}(\Lambda \geq \Lambda_{p,m_H,m_E}(\alpha)|H_0) = \alpha$.

Progowa wartość odległości pomiędzy obiektami

Załóżmy, że użyjemy statystyki T^2.

- Jeżeli $m_E \geq p$ oraz $m_H \geq p$, to

$$T^2 = \mathrm{tr}(\mathbf{H}\mathbf{E}^{-1}) \sim T^2_{p,m_H,m_E},$$

  gdzie $m_H = K - 1$, $m_E = n - K$.
- Jeżeli $m_E \geq p > m_H$, to $T^2 \sim T^2_{m_H,p,m_E+m_H-p}$.

Hipotezę H_0 odrzucamy wówczas, gdy $T^2 \geq T^2_{p,m_H,m_E}(\alpha)$, gdzie $\mathrm{P}(T^2 \geq T^2_{p,m_H,m_E}(\alpha)|H_0) = \alpha$.

Progowa wartość odległości pomiędzy obiektami

Z drugiej strony hipoteza $\Delta_{ij} = 0$ jest równoważna hipotezie $H_{ij}: \boldsymbol{\mu}_i = \boldsymbol{\mu}_j$, dla $i, j = 1, 2, \ldots, K$, $j \neq i$.

Hipotezę H_{ij} weryfikujemy za pomocą statystyki

$$T_{ij}^2 = \frac{n_i n_j}{n_i + n_j}(\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j)' \mathbf{S}^{-1} (\bar{\mathbf{x}}_i - \bar{\mathbf{x}}_j),$$

gdzie $\mathbf{S} = \frac{1}{m_E}\mathbf{E}$, $i, j = 1, 2, \ldots, K$, $j \neq i$.

W przypadku, gdy $K = 2$ i rozpatrujemy populacje π_i oraz π_j, zachodzą związki

$$\Lambda = \left(1 + \frac{1}{m_E} T_{ij}^2\right)^{-1} \quad \text{lub} \quad T_{ij}^2 = m_E \left(\frac{1}{\Lambda} - 1\right)$$

oraz

$$T^2 = \frac{1}{m_E} T_{ij}^2 \quad \text{lub} \quad T_{ij}^2 = m_E T^2.$$

Przykład - Słoneczniki

Wartości statystyk T^2_{ij}:

	π_2	π_3	π_4	π_5	π_6	π_7	π_8	π_9	π_{10}	π_{11}
π_1	282.4	303.7	264.1	282.9	6.9	513.6	243.4	245.0	1.6	486.3
π_2		14.1	11.9	3.0	282.2	125.5	9.1	20.1	316.1	523.6
π_3			10.3	4.6	294.0	128.5	11.7	9.3	337.2	456.5
π_4				6.0	254.7	136.1	19.5	8.6	295.7	535.9
π_5					279.2	130.4	6.7	9.6	316.7	488.9
π_6						462.6	248.4	244.1	5.2	532.7
π_7							170.1	191.1	535.7	934.2
π_8								11.0	276.4	395.5
π_9									278.3	414.3
π_{10}										519.1

Wspólna wartość krytyczna, na poziomie istotności $\alpha = 0.05$, dla wartości T^2_{ij} jest równa: dla statystyki Λ - 99.081, a dla statystyki T^2 - 65.636.

Zarówno dla kryterium Λ jak i T^2 uzyskujemy podział 11 rodów hodowlanych słonecznika na 4 skupienia:

I. 1,6,10

II. 2,3,4,5,8,9

III. 7

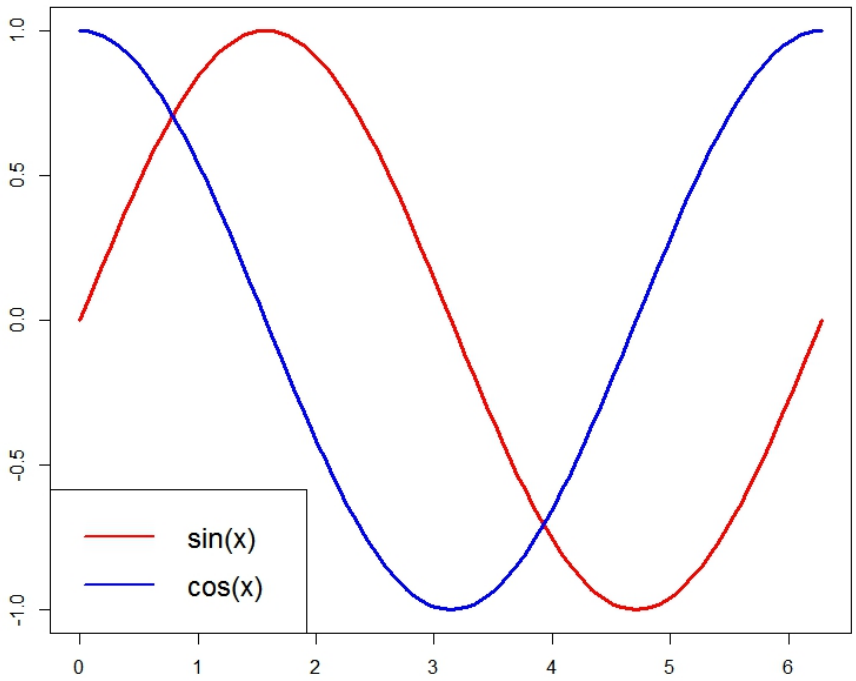
IV. 11

Analiza skupień dla obiektów (jednostek) opisanych za pomocą szeregu czasowego

Definicja szeregu czasowego

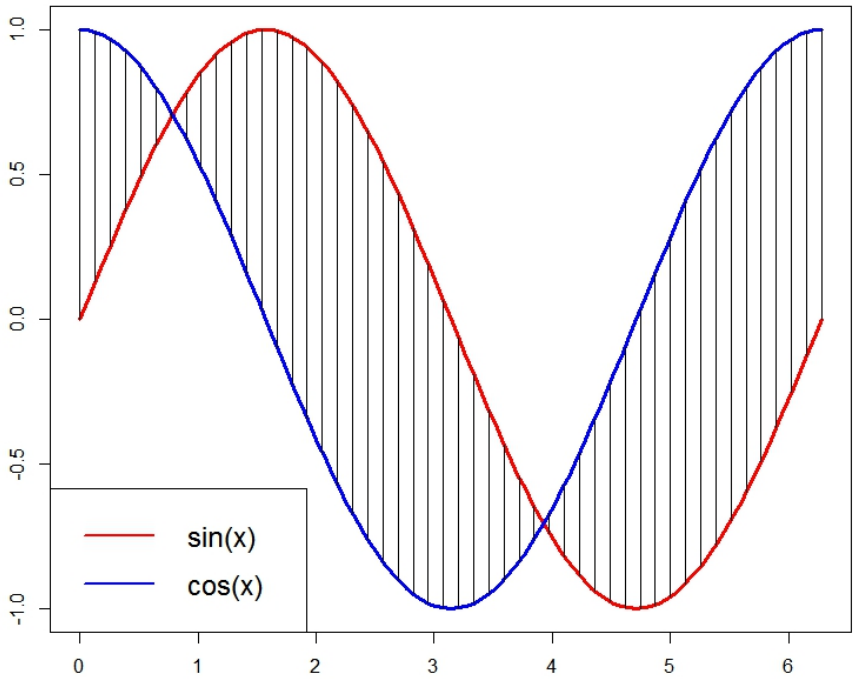
Szereg czasowy to sekwencja obserwacji, które uporządkowane są w czasie lub w przestrzeni (Box, Jenkins i Reisel, 2008). Dla prostoty i bez straty ogólności założymy, że czas jest dyskretny. Formalnie, szereg czasowy to sekwencja par $T = [(\boldsymbol{x}_1, t_1), (\boldsymbol{x}_2, t_2), \ldots, (\boldsymbol{x}_n, t_n)]$ $(t_1 < t_2 < \ldots < t_n)$, gdzie każdy $\boldsymbol{x}_i$ jest punktem w d-wymiarowej przestrzeni, a każdy moment czasowy t_i jest chwilą, w której zaszedł $\boldsymbol{x}_i$. Jeśli momenty czasowe dwóch szeregów są takie same, możemy je ominąć i rozważać jedynie sekwencje d-wymiarowych punktów. Taka reprezentacja jest nazywana surową. Liczba punktów n w szeregu czasowym jest nazywana jego długością. Na razie skupimy się na szeregach jednowymiarowych, które oznaczymy $x_i,\ i = 1, 2, \ldots, n$.

Odległość pomiędzy szeregami czasowymi



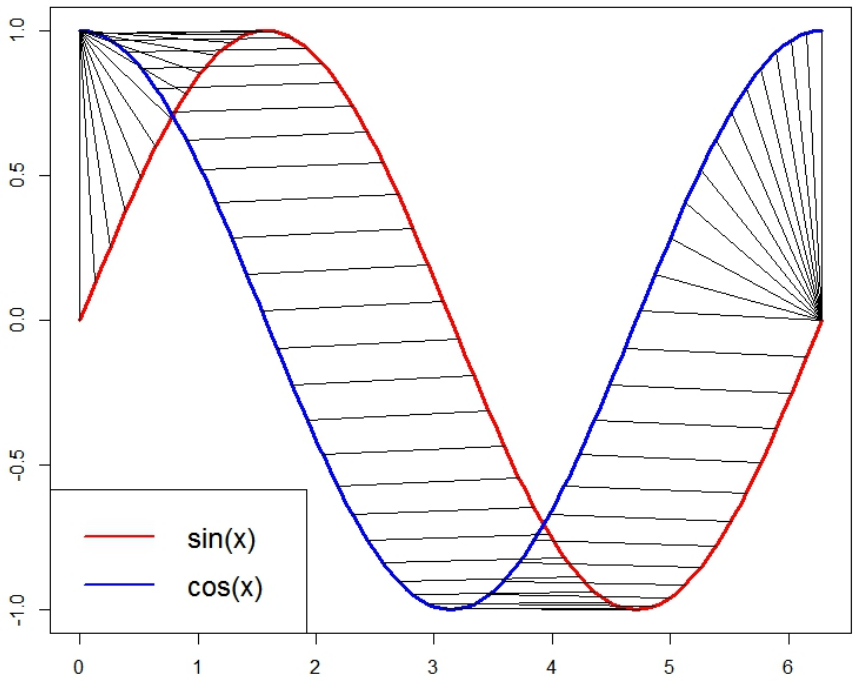
Jaka miara odległości jest najlepsza do porównania szeregów X oraz Y?

Odległość pomiędzy szeregami czasowymi



Miara odległości taksówkowej, czyli $d(X, Y) = \sum_{i=1}^{n} |x_i - y_i|$ oraz odległości euklidesowej, czyli $d(X, Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

Odległość pomiędzy szeregami czasowymi



Miara odległości wyznaczona za pomocą algorytmu DTW.

Algorytm DTW

DTW (ang. dynamic time warping) jest doskonale znaną techniką wykorzystywaną do znajdowania optymalnego „wyrównania" dwóch szeregów czasowych. Pierwotnie DTW było wykorzystywane do porównywania wzorców wymowy w automatycznym rozpoznawaniu mowy. Jest to metoda, która wyznacza odległość pomiędzy dwoma szeregami czasowymi, przy czym dopuszczamy pewne transformacje czasu. Aby znaleźć odległość DTW wpierw konstruujemy macierz, której element (i, j) odpowiada np. $d(x_i, y_j) = |x_i - y_j|$. Następnie poszukujemy minimalnej skumulowanej odległości przechodząc przez tę macierz. Odległość DTW odpowiada ścieżce o minimalnym koszcie:

$$\mathrm{DTW}(X, Y) = \min \sqrt{\sum_{k=1}^{K} w_k},$$

gdzie w_k jest elementem macierzy kosztów, który należy do ścieżki W.

Algorytm DTW

Ścieżkę tę konstruujemy przy trzech dodatkowych warunkach:

- $w_1 = (1, 1)$ oraz $w_K = (n, n)$ (warunki brzegowe, dopasowanie nie jest wykonane na fragmentach szeregów),
- Dla $w_k = (a, b)$ i $w_{k-1} = (a', b')$, $a - a' \leq 1$ i $b - b' \leq 1$ (ciągłość, żadne punkty nie są pomijane),
- Dla $w_k = (a, b)$ i $w_{k-1} = (a', b')$, $a - a' \geq 0$ i $b - b' \geq 0$ (monotoniczność, podobne fragmenty są łączone tylko raz).

Aby wyznaczyć taką ścieżkę używamy programowania dynamicznego, w którym wykorzystywane jest następujące równanie rekurencyjne:

$$\gamma(i, j) = d(x_i, y_j) + \min\{\gamma(i-1, j-1), \gamma(i-1, j), \gamma(i, j-1)\},$$

gdzie $d(x_i, y_j)$ jest odległością w danej komórce, a $\gamma(i, j)$ jest skumulowaną odległością $d(x_i, y_j)$ oraz minimum z trzech przyległych skumulowanych odległości.

Algorytm DTW – przykład

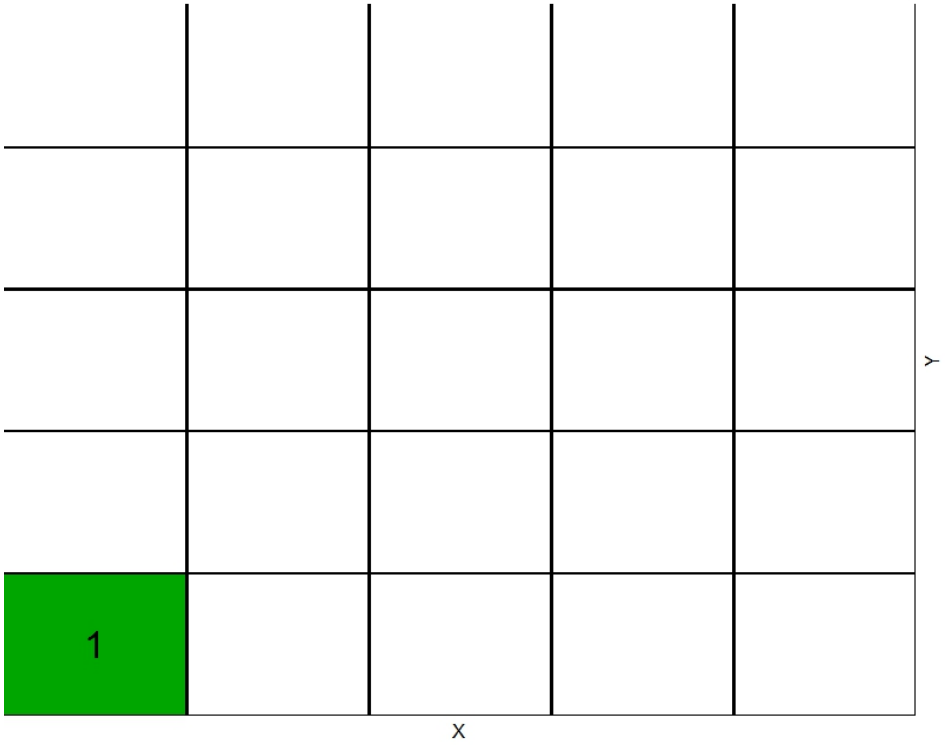
Rozważy dwa szeregi $X = (1, 2, 3, 4, 5)$ oraz $Y = (2, 4, 6, 8, 10)$. Skonstruujmy dla nich macierz kosztów $\boldsymbol{D}$. Ma ona postać:

$$\boldsymbol{D} = \underbrace{\left.\begin{bmatrix} 1 & 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 & 1 \\ 5 & 4 & 3 & 2 & 1 \\ 7 & 6 & 5 & 4 & 3 \\ 9 & 8 & 7 & 6 & 5 \end{bmatrix}\right\} Y}_{X}$$

Algorytm DTW – przykład

Następnie konstruujemy macierz kosztów skumulowanych. Wpierw wypełniamy lewy dolny róg tej macierzy. Mamy:

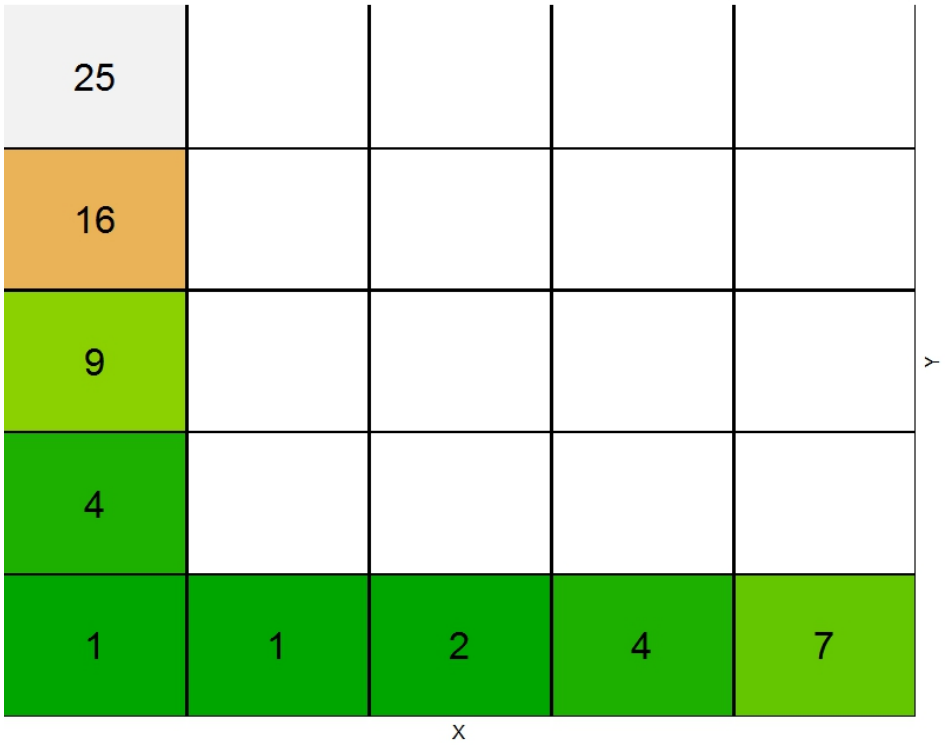
$$\gamma(1,1) = d(1,1) = 1.$$



Algorytm DTW – przykład

Następnie wypełniamy pierwszy wiersz i pierwszą kolumnę:

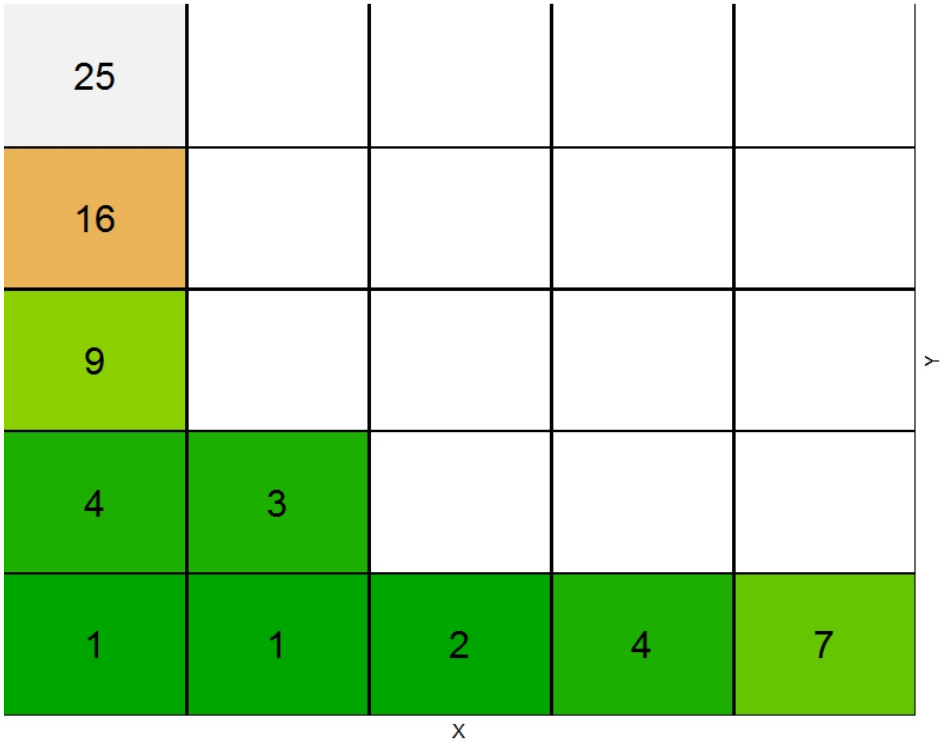
$$\gamma(1,j) = d(i,j) + \gamma(1,j-1),$$
$$\gamma(i,1) = d(i,j) + \gamma(i-1,1).$$



Algorytm DTW – przykład

Wyznaczamy:

$$\gamma(2,2) = d(2,2) + \min\{\gamma(1,1), \gamma(1,2), \gamma(2,1)\} =$$
$$= 2 + \min\{1,4,1\} = 3$$

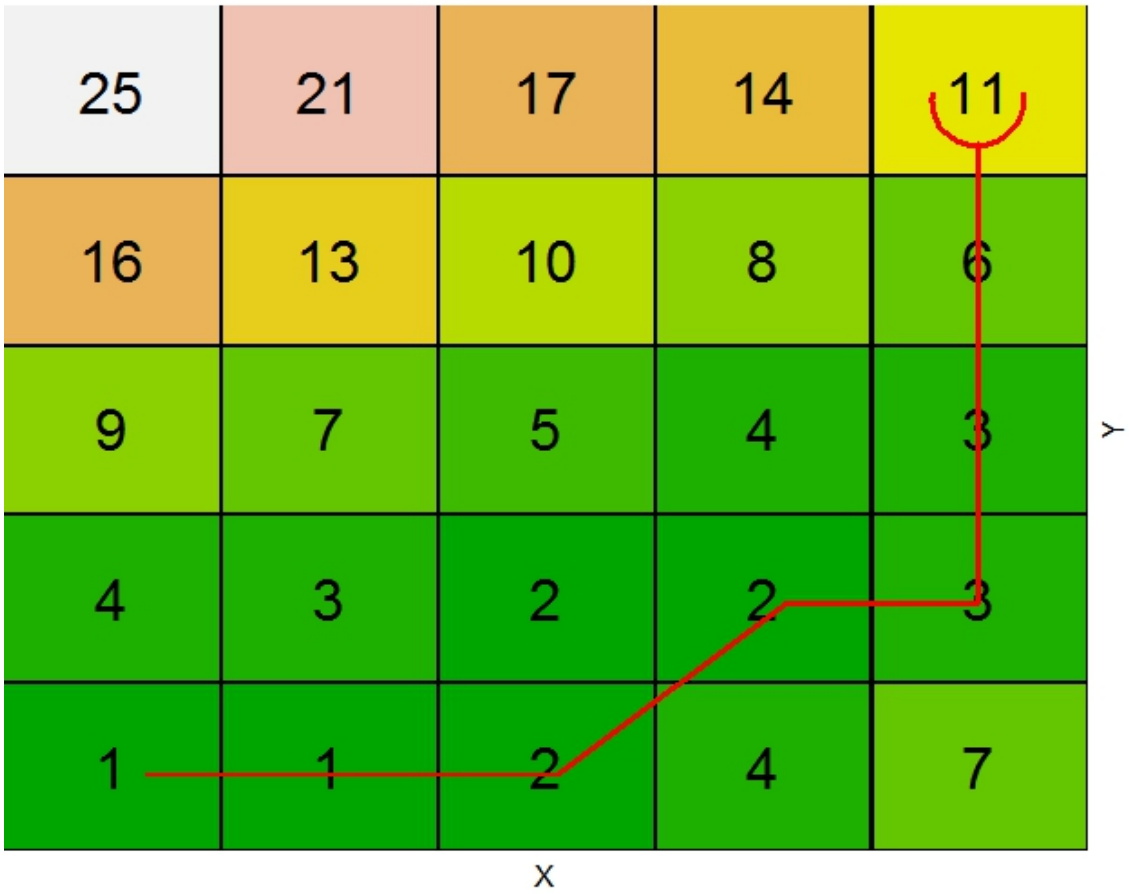


Algorytm DTW – przykład

Analogicznie wypełniamy resztę tablicy.

Algorytm DTW – przykład

Po wyznaczeniu całej macierzy Γ wyznaczamy optymalną ścieżkę z lewego dolnego rogu do prawego górnego.



Algorytm DTW – przykład

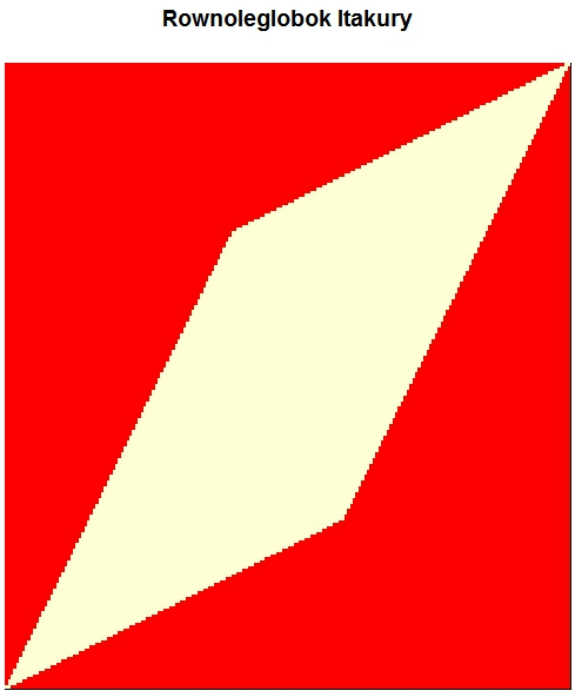
Ostatecznie otrzymujemy:

$$d(X, Y) = \sum_{i=1}^{5} |x_i - y_i| = 15,$$
$$\mathrm{DTW}(X, Y) = 11.$$

Algorytm DTW – okno

Często do wspomnianych wcześniej trzech warunków dodaje się jeszcze jeden, który mówi o tym, że dobra ścieżka nie może być zbyt daleko od przekątnej. Dwa najpopularniejsze to:

- Równoległobok Itakury,
- Pasmo Sakoe-Chiby.



Przykład

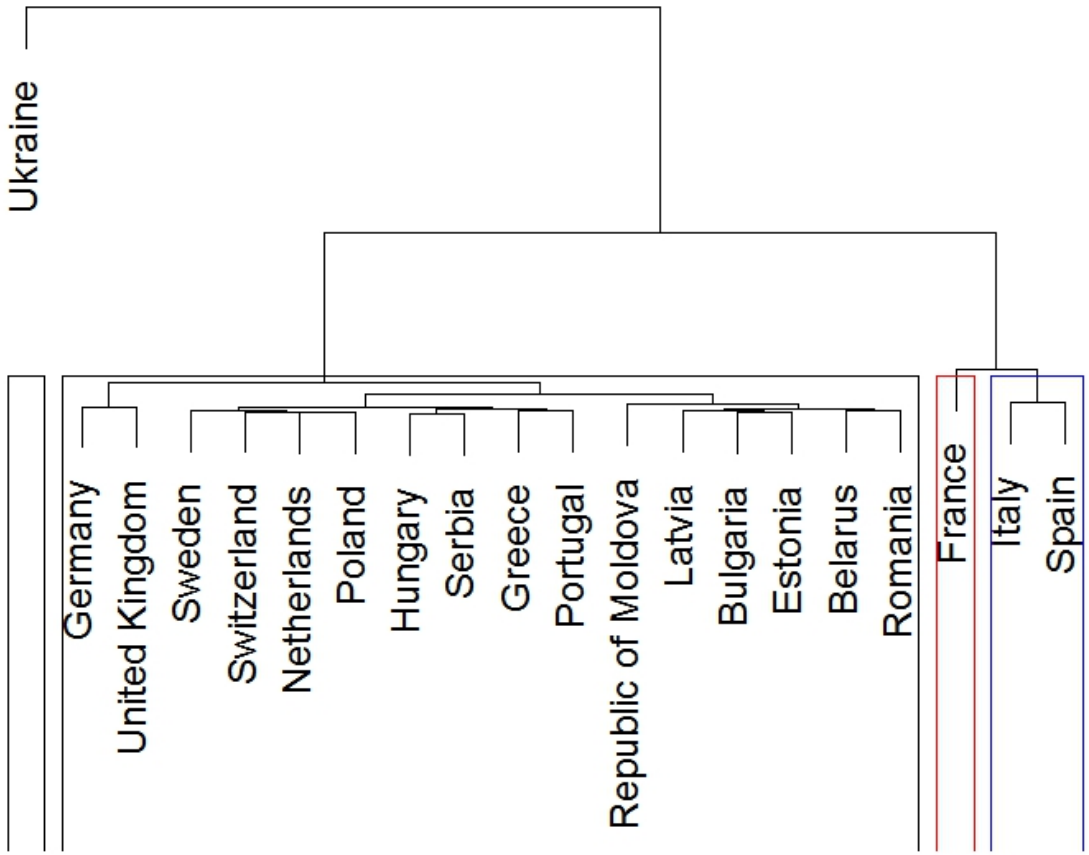
Dane dotyczą liczby zgonów z powodu AIDS w 20 krajach Europy w latach 1990-2011. Zostały zaczerpnięte z bazy: `http://www.aidsinfoonline.org/`. W przypadku danych dotyczących chorób zakaźnych często mamy do czynienia z sytuacją kiedy dwa szeregi mają podobna strukturę, ale są przesunięte w czasie (np. gdy szczyt umieralności (zapadalności) w danym obszarze występuje wcześniej/później niż w innym). W takich przypadkach odległość DTW jest bardziej odpowiednia niż odległość euklidesowa.

Przykład

Przeprowadzona została hierarchiczna analiza skupień wykorzystująca odległość euklidesową oraz odległość DTW.

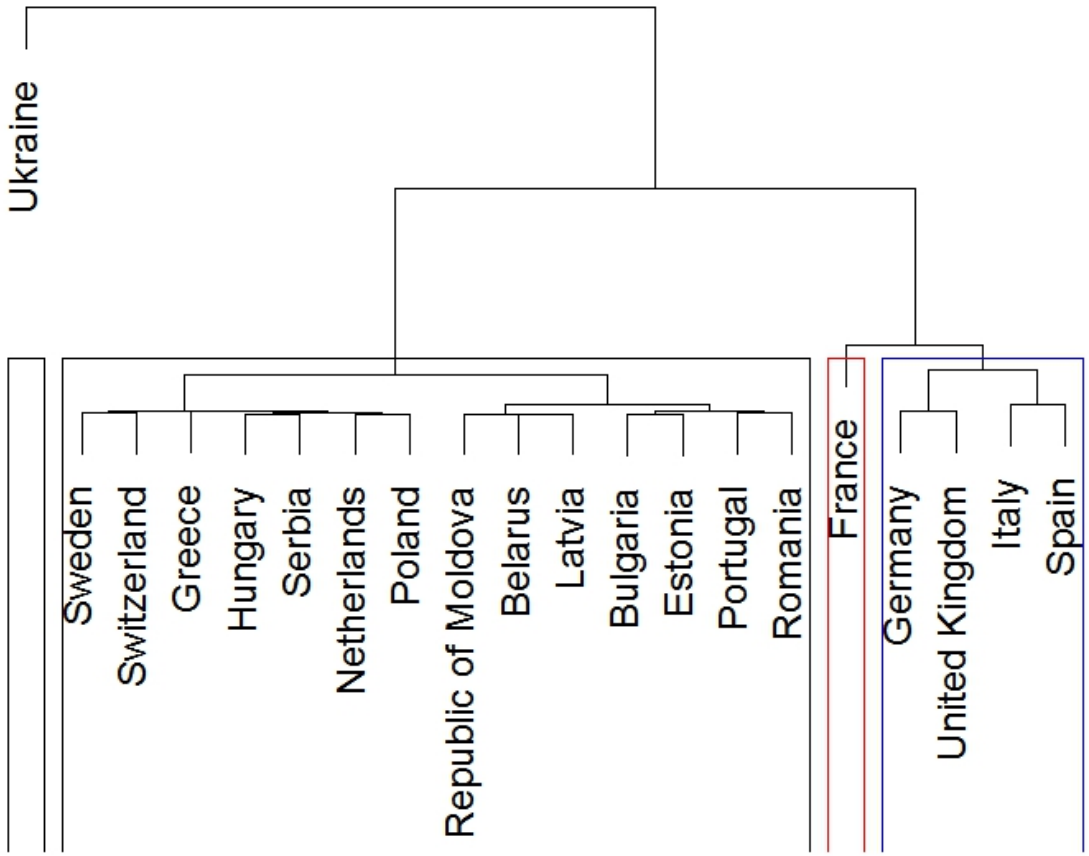
Przykład

Dendrogram dla odległości euklidesowej, metoda WARDA wiązania skupień.



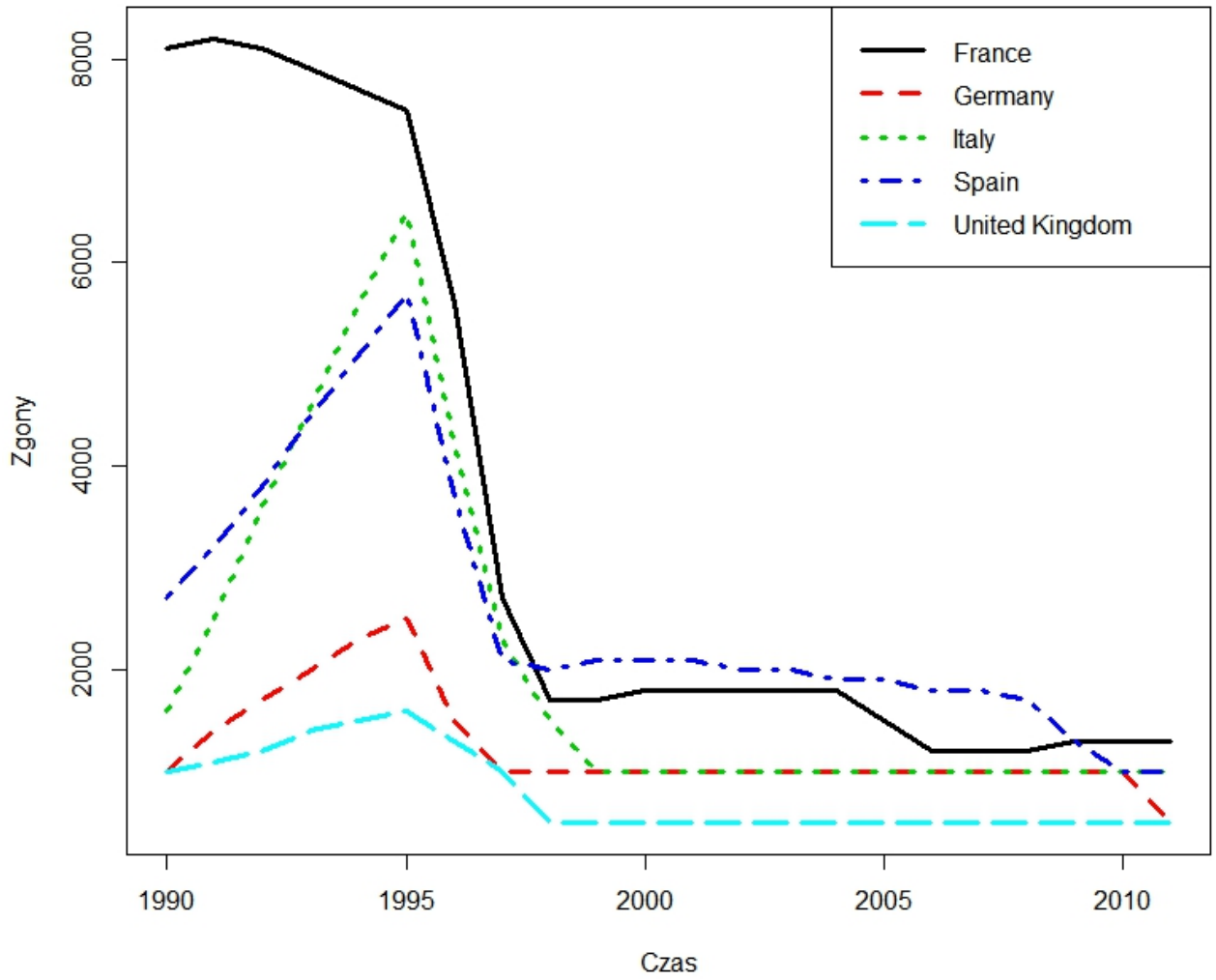
Przykład

Dendrogram dla odległości DTW, metoda WARDA wiązania skupień.



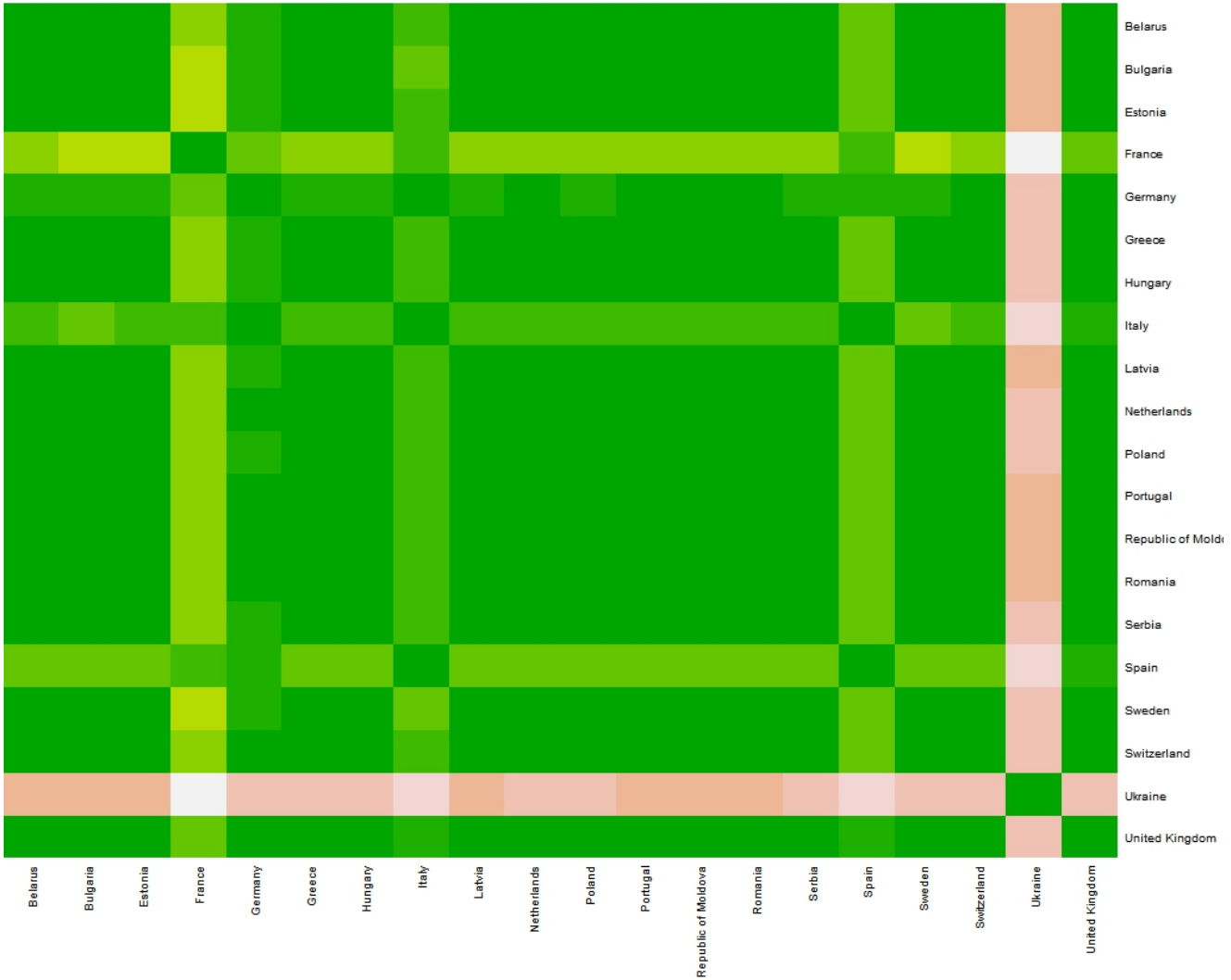
Przykład

Trajektorie szeregów czasowych dla pięciu wybranych krajów.



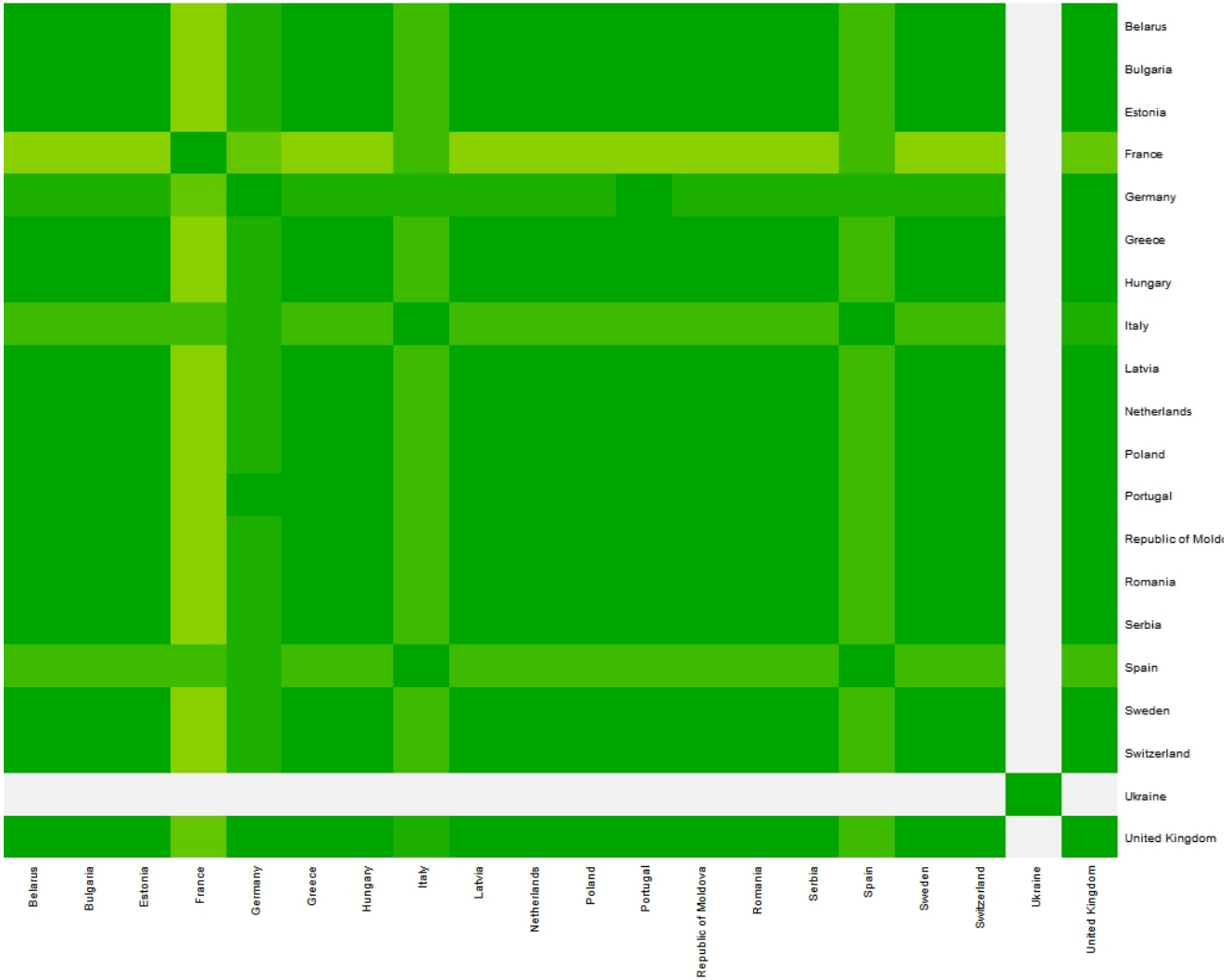
Przykład

Macierz ciepła dla odległości DTW.



Przykład

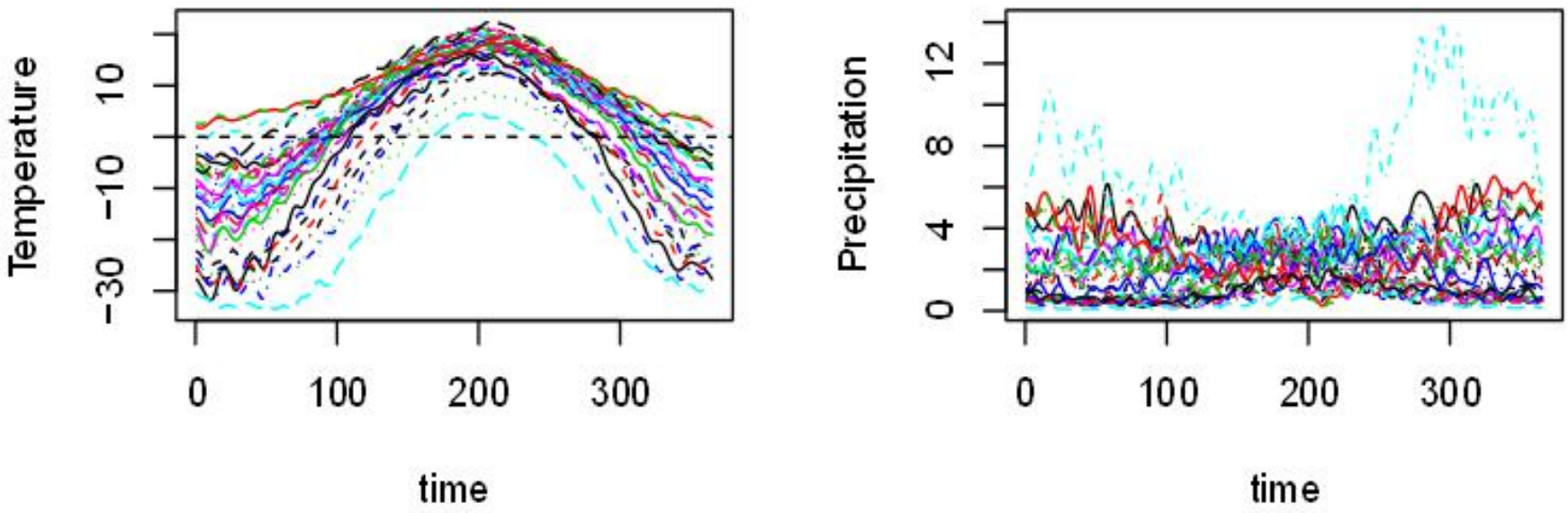
Macierz ciepła dla odległości euklidesowej.



- dtw(dtw) – odległość DTW pomiędzy dwoma szeregami,
- dtwDist(dtw) – macierz odległości DTW,
- dist(proxy) lub dist(stats) – macierz odległości,
- heatmap.2(gplots) – mapa ciepła.

Analiza skupień dla obiektów (jednostek) opisanych za pomocą wektora $\mathbf{p}$ *funkcji*

Przykład – Canadian weather



Dane pochodzą z 35 stacji meteorologicznych, z lat 1960-1994. Dostępne w pakiecie fda(R).

n niezależnych realizacji

$$\{\mathbf{x}_i(t),\ i = 1, 2, \ldots n,\ t \in [0, T]\}$$

p-wymiarowego procesu losowego

$$\mathbf{x}(t) = \begin{bmatrix} X_1(t) \\ \vdots \\ X_p(t) \end{bmatrix}.$$

Wielowymiarowe dane funkcjonalne

Konwersja danych dyskretnych $\{t_{ij}, \mathbf{x}_{ij} = \mathbf{x}_i(t_{ij})\}$, $i = 1, \ldots, n$, $j = 1, \ldots, J_i$ do danych funkcjonalnych.

Niech

$$x_{il}(t) = \sum_{k=1}^{m_l} c_{ilk}\phi_{lk}(t), \;\; i = 1, \ldots, n, \;\; l = 1, \ldots, p,$$

gdzie $\phi_{l1}, \ldots, \phi_{lm_l}$ są funkcjami bazowymi w $L_2([0, T])$.
Stąd

$$\mathbf{x}_i(t) = \mathbf{\Phi}(t)\mathbf{c}_i,$$

gdzie

$$\mathbf{\Phi}(t) = \begin{pmatrix} \phi_{11}(t) & \cdots & \phi_{1m_1}(t) & 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & \phi_{21}(t) & \cdots & \phi_{2m_2}(t) & 0 & 0 & \cdots & 0 \\ & \cdots & & & \cdots & & & & \cdots & \\ 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \phi_{p1}(t) & \cdots & \phi_{pm_p}(t) \end{pmatrix},$$

$$\mathbf{c}_i = (c_{i11}, \ldots, c_{i1m_1}, c_{i21}, \ldots, c_{i2m_2}, \ldots, c_{ip1}, \ldots, c_{ipm_p})'.$$

Niech

$$\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_{i1} \\ \vdots \\ \mathbf{x}_{iJ_i} \end{bmatrix}, \quad \mathbf{\Phi}_i = \begin{bmatrix} \mathbf{\Phi}(t_{i1}) \\ \vdots \\ \mathbf{\Phi}(t_{iJ_i}) \end{bmatrix}.$$

Współczynniki $\mathbf{c}_i$ szacujemy metodą najmniejszych kwadratów, tzn.

$$\hat{\mathbf{c}}_i = (\mathbf{\Phi}_i' \mathbf{\Phi}_i)^{-1} \mathbf{\Phi}_i \mathbf{x}_i.$$

Miary niepodobieństwa obiektów

$$\rho(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\int_0^T (\mathbf{x}_i(t) - \mathbf{x}_j(t))'(\mathbf{x}_i(t) - \mathbf{x}_j(t))dt},$$

$$= \sqrt{(\mathbf{c}_i - \mathbf{c}_j)'\mathbf{W}(\mathbf{c}_i - \mathbf{c}_j)},$$

gdzie

$$\mathbf{W} = \int_0^T \mathbf{\Phi}(t)'\mathbf{\Phi}(t)dt.$$

Uwaga: Odległości pomiędzy obiektami możemy również wyznaczyć wykorzystując pochodne $D\mathbf{x}_i(t)$.

Skupienia wyznaczone metodą K-średnich, $K = 4$.

Rysunki pochodzą z pracy: J. Jacques, C. Preda, Model-based clustering for multivariate functional data, Computational Statistics and Data Analysis, in press.

Skupienia wyznaczone metodą EM.

Rysunek pochodzi z pracy: J. Jacques, C. Preda, Model-based clustering for multivariate functional data, Computational Statistics and Data Analysis, in press.