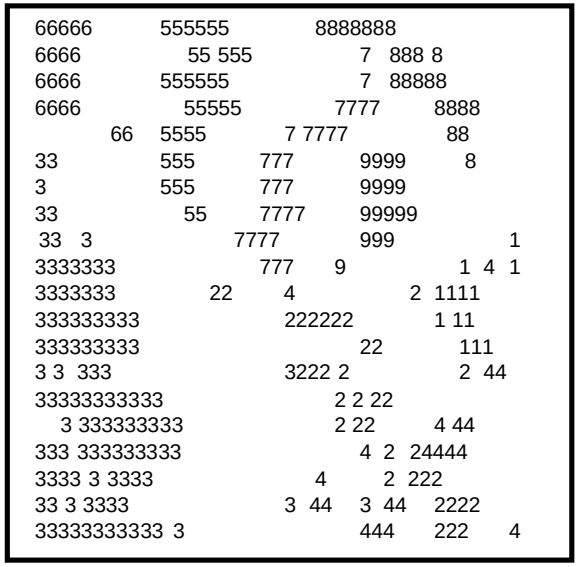
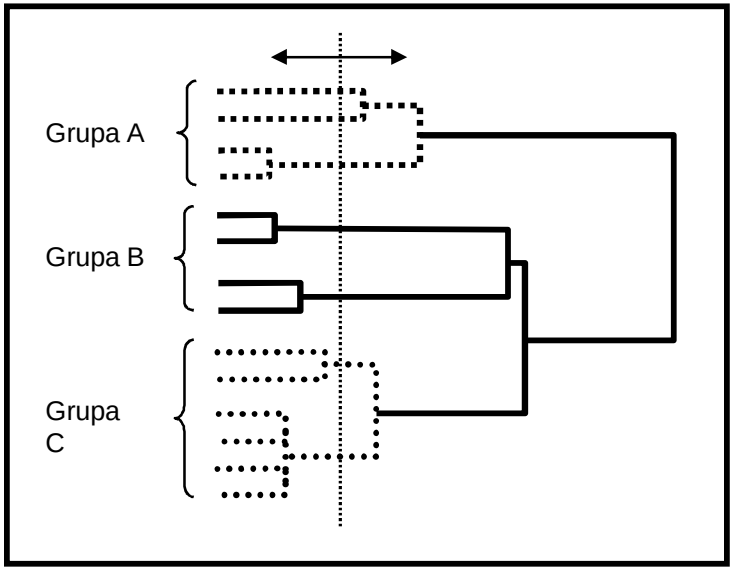
Grupiranje

Grupiranje objektov je pomembna obdelava kemijskih podatkov, ki jo **večinoma** uporabljamo v **začetnih** fazah pregledovanja in zbiranja informacij o kompleksnih objektih kot so **večkomponentne** analize, spektri, kemijske strukture, procesni vektorji, **časovne** vrste ipd. Grupiranje delimo na enonivojsko in **večnivojsko (hierarhično)** grupiranje.

Pri enonivojskem grupiranju gre **večinoma** za **določitev** števila grup, ki stestavljajo dano množico objektov in za **določitev** kateri grupi pripada vsak posamezni objekt. Na spodnjem primeru vidimo enonivojsko grupiranje objektov v devet **različnih** razredov (grup, skupkov, skupin ali klastrov). Vidimo da sta skupni 2 in 4 najbolj prepleteni in težko **ločljivi**,



Pri **večnivojskem (hierarhičnem)** grupiranju želimo **določiti** ne samo število skupin in objekte v njih, ampak tudi **natančnejšo** zgradbo posameznih skupin. Hkrati želimo skupine in njihove medsebojne podrobnosti oziroma razdalje kvantitativno opredeliti. **Če** na spodnji shemi skupkov premaknemo delitveno premico **(točkasta črta)** levo ali desno, tj., **če** spremenimo **odločitveni** nivo, lahko dobimo **večje** ali manjše število skupin.



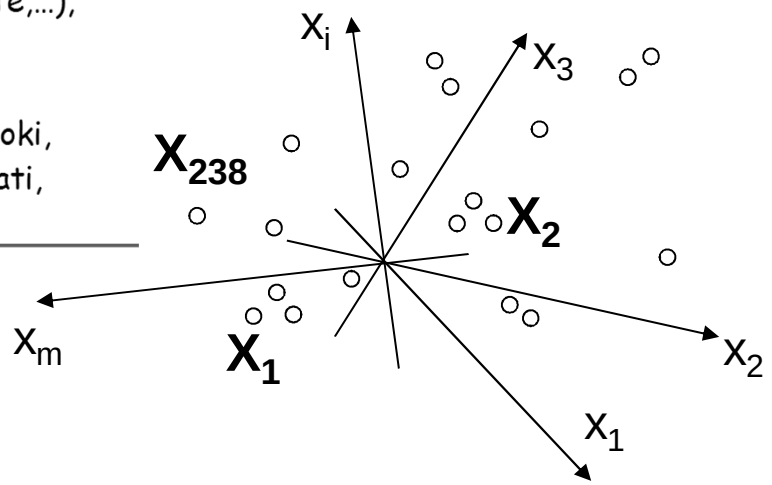
Grupiramo le objekte, ki so predstavljeni vsaj kot 2-dimenzionalni vektorji. Objekte, ki so opisani le z eno **količino** (1-dimenzionalne vektorje) grupiramo lahko le po njihovih vrednostih, kar navadno ne predstavlja posebnih težav.

Vsak kompleksen objekt X_i, npr. spekter, struktura, kompleksna analza, procesni vektor itd, je opisan z nizom podatkov in ga obravnavamo kot vektor ali kot **točko** v *m*-dimenzionalnem prostoru. Dimenzija merskega prostora je enaka *m,* t.j., številu spremenljivk s katerimi vsak vektor opišemo: $X_i = (x_{i1}, x_{i2}, \ldots x_{ij}, \ldots x_{im})$. Prostor meritev imenujemo tudi *m*-dimenzionalni merski prostor.

Primeri:

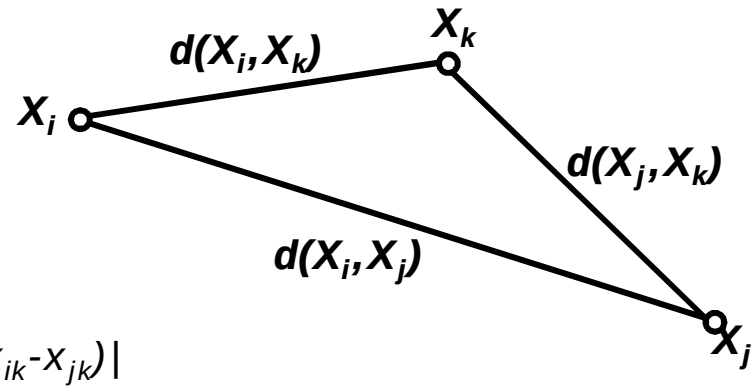
Objekt X_i	Komponente x_{ij}	Primeri
Spekter	intenzitete v spektru	absorbance, transmitanca, relativni premik, itd.
Analiza	koncentracije posameznih komponent	%, ‰, moli itd.
Struktura kemijskih spojin	deskriptorji in spektralne predstavitve	geometrijski (št. atomov, št. vezi, …), topolški (poti, sprehodi, invariante,…), elektrostatski (naboji, dipolini momenti…) itd.
Kemijski proces	procesni parametri	temperature, pH vrednosti, pretoki, koncentracije, čas, hitrosti, obrati, frekvence itd. …

Predstavitev točk v m-dimenzionalnem (merskem) *prostoru*



Da lahko objekte v merskem prostoru primerjamo in jih potem grupiramo, moramo poznati *metriko* prostora. Metriko vpeljemo v prostor z definicijo *razdalje*. Razdalja med dvema objektoma X_i in X_j, $d(X_i,X_j)$, mora biti definirana za toliko spremenljivk, kot jih zahteva merski prostor. Razdalja je vsaka funkcija *m*-spremenljivk, ki za *katerikoli* par X_i in X_j iz celotnega merskega prostora, **zadošča** naslednjim štirim (4) pogojem:

1. Razdalja med dvema objektoma je nenegativna: $d(X_i,X_j) \geqq 0$
2. Razdalja med dvema objektoma je enaka nič takrat in le takrat, ko je $X_i = X_j$: $d(X_i,X_j) = 0$, *če je* $X_i=X_j$
3. Razdalja med dvema objektoma je *komutativna*: $d(X_i,X_j) = d(X_j,X_i)$
4. Veljati mora trikotniško pravilo: $d(X_i,X_j) \leqq d(X_i,X_k) + d(X_j,X_k)$



Primeri za razdalje:

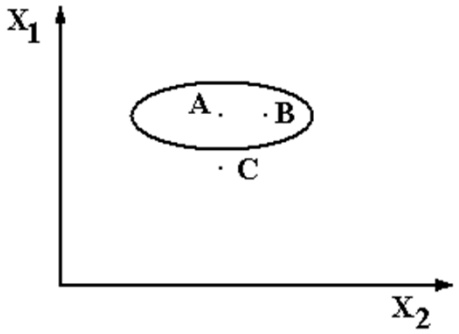
Manhattanska razdalja: $d(X_i,X_j) = \Sigma_k |(x_{ik}-x_{jk})|$

Euklidska razdalja: $d(X_i,X_j) = (\Sigma_k (x_{ik}-x_{jk})^2)^{1/2}$

Razdalja Minkowskega: $d(X_i,X_j) = (\Sigma_k |(x_{ik}-x_{jk})|^p)^{1/p}$

Mahalanobisova razdalja: $d(X_i,X_j) = (X_i-X_j)^T V^{-1} (X_i-X_j)$ v kateri je **V** matrika varianc

Mahalanobisova razdalja



Mahalanobisova razdalja je razdalja med skupino objektov predstavljeno s središčno točko A in objektoma B in C.

Razdalja AC je daljša kot AB, ker se slednja nahaja vzdolž glavne osi elipse.

Binarne spremenljivke

Binarne spremenljivke ponavadi zavzemajo vrednost 0, če je atribut odsoten in vrednost 1, če je prisoten.

Ujemanje vrednosti

$d_{ii',j} = 1$, če je $x_{i\,j} = x_{i',j}$

$d_{ii',j} = 0$, če je $x_{i\,j} \neq x_{i',j}$

$$D_{ii'} = \frac{1}{m}\sum_{j=1}^{m} d_{ii'j}$$

Jaccardova razdalj

$d_{ii',j} = 1$, če je $x_{i\,j} = x_{i',j} = 1$

$d_{ii',j} = 0$, če je $x_{i\,j} \neq x_{i',j}$

nedefiniran, če je $x_{i\,j} = x_{i',j} = 0$

Postopek grupiranje za N objektov, ki so predstavljeni v m-dimenzionalnem merskem prostoru kot vektorji $X_i = (x_{i1}, x_{i2}, .. x_{ij}, \ldots x_{im})$:

1. Izračun matrike razdalj $\boldsymbol{D}$ med vsemi objekti X_i. Matrika $\boldsymbol{D}$ ima dimenzijo NxN.

2. Določitev dveh objektov X_i in X_j ali dveh skupkov I in J, ki imata najmanjšo medsebojno razdaljo.

3. Izbris i-te in j-te vrstice ter i-tega in j-tega stolpca iz matrike razdalj $\boldsymbol{D}$. S tem dobimo novo matriko razdalj $\boldsymbol{D}^{nova}$, ki ima v primerjavi s staro matriko dimenzijo $(N-2)x(N-2)$

4. Združimo objekta X_i in X_j ali skupka I in J v novo skupino objektov K.

5. Izračunamo vse razdalje $d(X_i,K)$ ali $d(J,K)$ med novo skupino K in ostalimi objekti X_i ali skupinami J in dodamo izračunane razdalje v matriko $\boldsymbol{D}^{nova}$. S tem je matrika razdalj $\boldsymbol{D}^{nova}$ dobila eno dimenzijo več: $(N-1)x(N-1)$.

6. Če je dimenzija nove matrike razdalj večja kot 2, postopek nadaljujemo pri točki 2.

<u>Za postopek grupiranja potrebujemo definicijo **dveh** razdalj:</u>

1. razdaljo med posameznimi objekti X_i in X_j $d(X_i,X_j)$ in

2. razdaljo med dvema skupinama objektov I in J $d(I,J)$

Skupini I in J vsebujeta n_i in n_j objektov.
Če je v obeh skupinah le po en objekt mora biti $d(I,J) = d(X_i,X_j)$

Razdalje med skupinami objektov (skupki, grupami ali klastri) lahko izrazimo na več načinov. Najpogostejše razdalje med skupinami so:

- najmanjša razdalja, ki je možna med dvema objektoma iz obeh skupkov
- največja razdlaja, ki je možna med dvema objektoma iz obeh skupkov,
- razdalja med težišči obeh skupkov,
- povprečje vseh možnih razdalj med točkami iz obeh skupkov
- itd.

Vse zgoraj navedene razdalje in še nekaj drugih združuje Lance-Williamsova enačba, ki pravi, da je razdalja med poljubno skupino K, sestavljeno iz n_k objektov in skupino H, ki je nastala z združitvijo dveh skupkov I in J, ki sta imela n_i in n_j objektov enaka:

$$\mathbf{d(K,H) = d(K,(I,J)) = \alpha_1 d(K,I) + \alpha_2 d(K,J) + \beta d(I,J) + \gamma |d(K,I) - d(K,J)|}$$

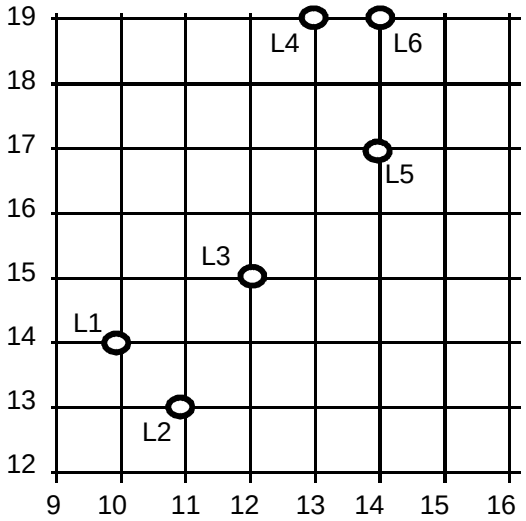
Koeficienti α_1, α_2, β in γ so za različne razdalje podani v spodnji tabeli

Razdalja	α_1	α_2	**b**	γ
Mimalna	½	½	0	- ½
Maksimalna	½	½	0	½
Povprečna	n_i/n_h	n_j/n_h	0	0
Uteženo povprečje	½	½	0	0
Centroidna	n_i/n_h	n_j/n_h	$-n_i n_j/n_h^2$	0
Mediana	½	½	$-(½)^2$	0
Wardova	$(n_i+n_k)/(n_i+n_j+n_k)$	$(n_j+n_k)/(n_i+n_j+n_k)$	$-n_k/(n_i+nj+n_k)$	0

Primer: Grupirajte šestih laboratorijev, ki so v istem neznanem vzorcu merili vsebnosti dveh nečistoč. ! Laboratorij je izmeril za prvi dodatek 10, za drugega pa 14 μg. Laboratorij in omenjeni dve vrednosti opišemo takole: $L_1(10,14)$. Analogno je ostalih 5 laboratorijev izmerilo naslednje koncentracij (vse v μg), $L_2(11,13)$, $L_3(12,15)$, $L_4(13,19)$, $L_5(14,17)$ in $L_6(14,19)$. Za razdaljo med laboratoriji uporabite uporabili Manhattansko razdaljo, kot razdaljo med skupinami laboratorijev po maksimalno razdaljo.

$$d_{Manh.}(X_i, X_j) = \sum_{k=1}^{m} \left| x_{ki} - x_{kj} \right|$$

$$D_{Max}(K,(I,J)) = \frac{1}{2}\left(D(K,I) + D(K,J) + \left|D(K,I) - D(K,J)\right|\right)$$



$D^1 =$

	L1	L2	L3	L4	L5	L6
L1	0	2	3	8	7	9
L2	2	0	3	8	7	9
L3	3	3	0	5	4	6
L4	8	8	5	0	3	1
L5	7	7	4	3	0	2
L6	9	9	6	1	2	0

$D^1 =$

	L1	L2	L3	L4	L5	L6
L1	0	2	3	8	7	9
L2	2	0	3	8	7	9
L3	3	3	0	5	4	6
L4	8	8	5	0	3	1
L5	7	7	4	3	0	2
L6	9	9	6	1	2	0

$D^2 =$

	L1	L2	L3	L5	(L4,L6)
L1	0	2	3	7	9
L2	2	0	3	7	9
L3	3	3	0	4	6
L5	7	7	4	0	3
(L4,L6)	9	9	6	3	0

⇨

	(L1,L2)	L3	L5	(L4,L6)
(L1,L2)	0	3	7	9
L3	3	0	4	6
L5	7	4	0	3
(L4,L6)	9	6	3	0

⇨

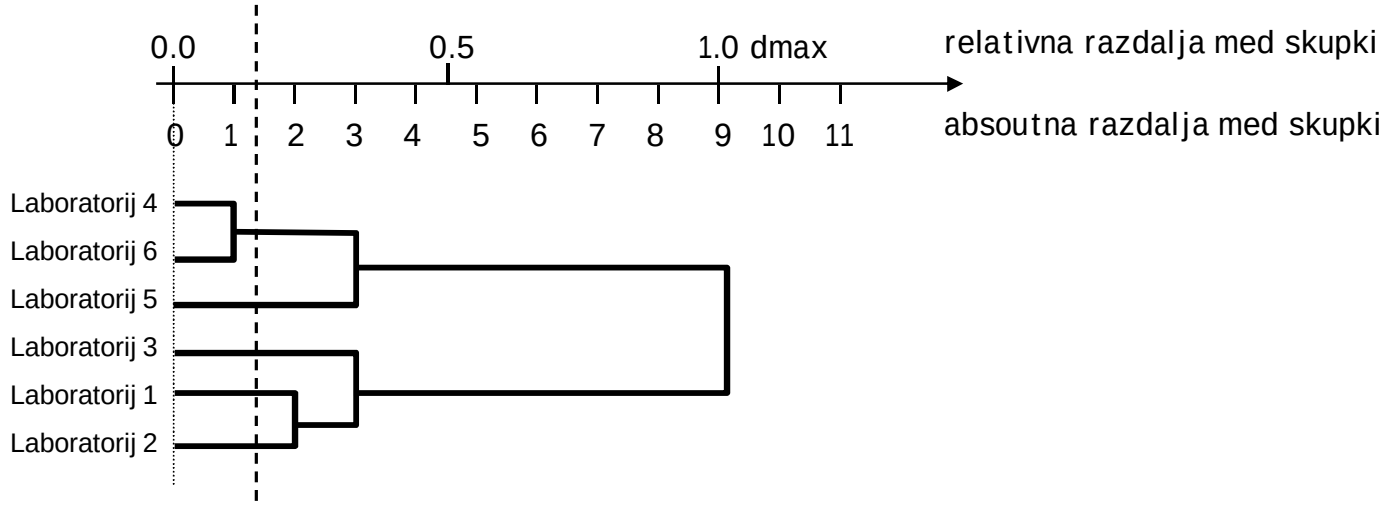
	(L1,L2)	L3	(L4…L6)
(L1,L2)	0	3	9
L3	3	0	6
(L4…L6)	9	6	0

⇩

$D^5 =$

	(L1…L3)	(L4…L6)
(L1…L3)	0	9
(L4…L6)	9	0

Rezultat grupiranja podamo v obliki dendrograma. Na abscisno os nanesemo objekte v skupkih tako, kot so se grupirali v postopku. Ustrezno razdaljo, pri kateri se dva objekta ali dva skupka grupirata, pa vzamemo iz trenutne matrike razdalj (modri krogi). Merilo na ordinatni osi je lahko absolutna razdalja, lahko pa tudi relativno merilo oddaljenosti, glede na največjo razdaljo med zadnjima dvema skupinama (v danem primeru 9).

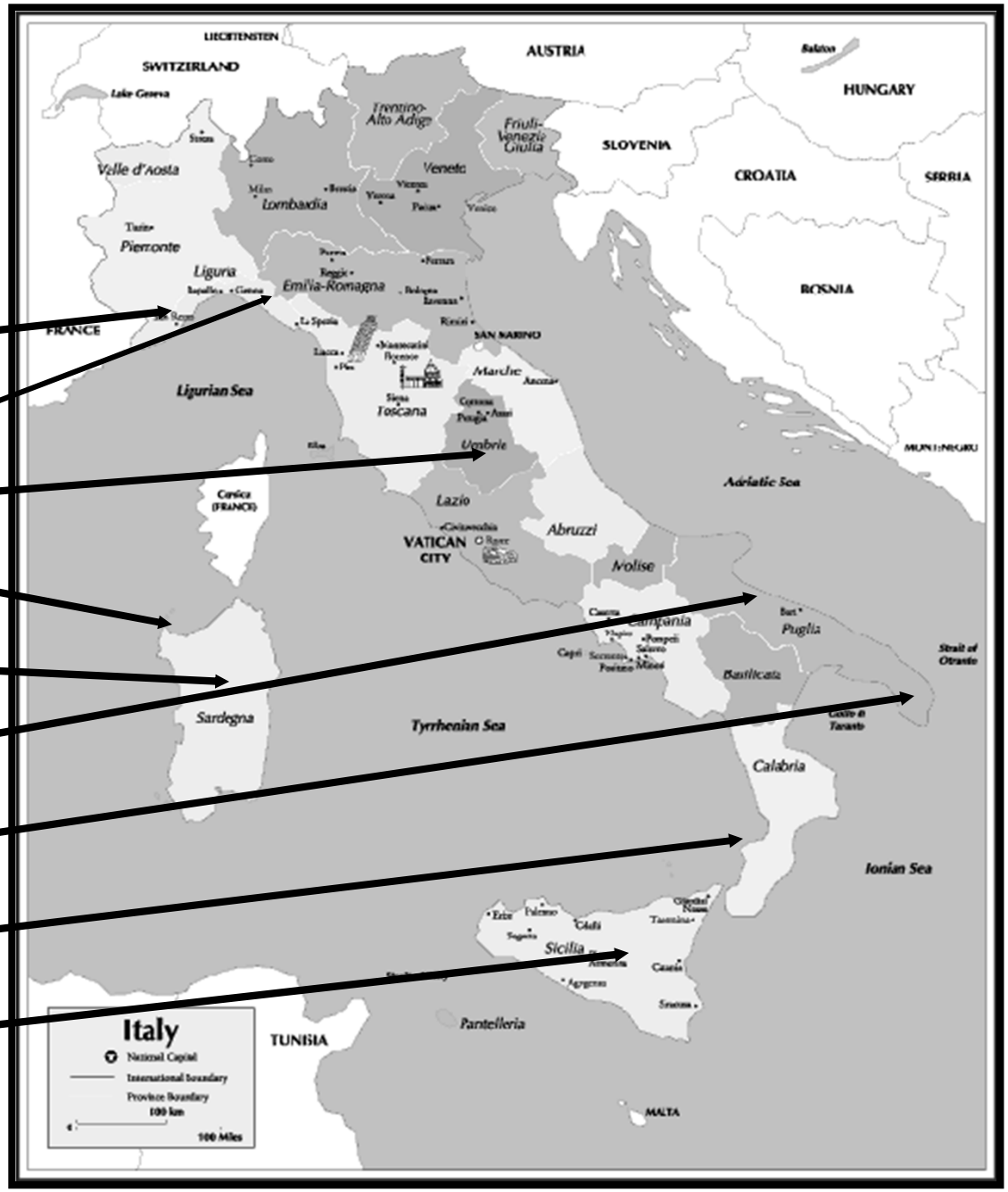


Položaj delitvene premice (črtkane vertikale) na zgornji sliki predstavlja prag (kriterij), ki odloča pri kateri razdalji je smiselno ločevati skupke. Ta kriterij povsem zavisi od uporabnika in od natančnosti spremenljivk s katerimi so opisani objekti, ki jih grupiramo. Če pomaknemo delitveno premico na levo pod vrednost absolutne razdalje 2, dobimo pet ali več skupkov, nad razdaljo 2 do vključno razdelje 3 imamo štiri, nad razdaljo 3 pa dva skupka.

Če imamo število in vrsto skupkov vnaprej znano in se objekti pri nekem določenem intervali položaja delitvene premice dejansko razporedijo v predvidene skupke, nam položaj premice in interval v katerem se mora ta nahajati, pove pomembno informacijo za napovedovanje kategorij vzorcev.

Pokrajine

- zahodna Ligurija (oker)
- vzhodna Ligurija (vijolična)
- Umbrija (svetlo zelena)
- obala Sardinije (rumena)
- notranja Sardinija (siva)
- severna Apulija (svetlo modra)
- južna Apulija (rdeča)
- Kalabrija (temno zelena)
- Sicilija (temno modra)

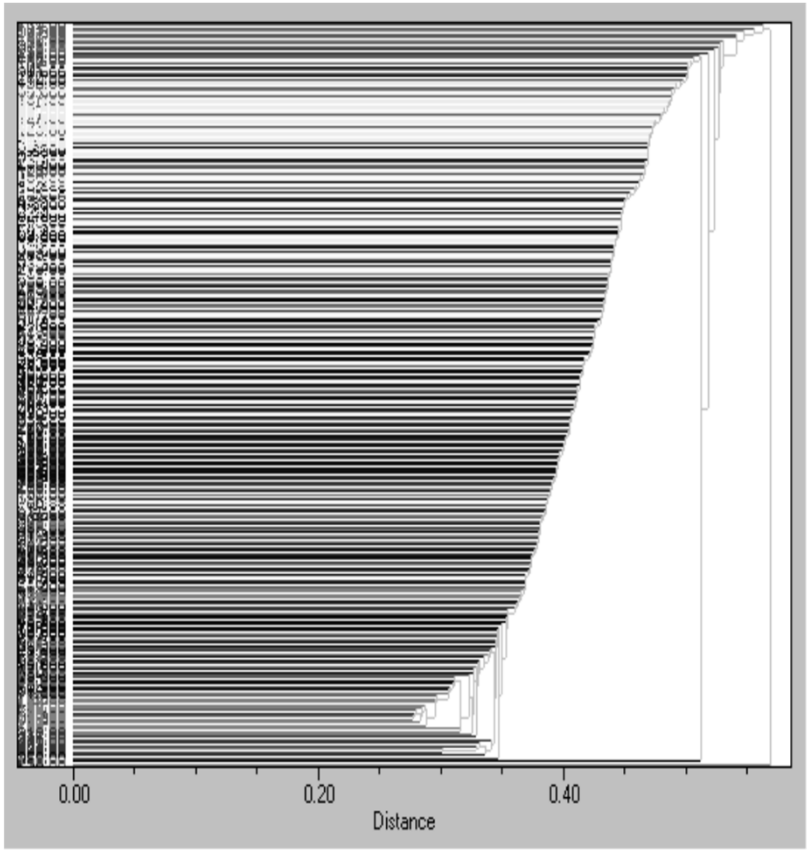


Vrste kislin

- Palmitinska
- Palmitotinska
- Stearinska
- Oleinska
- Arahidska
- Linolenska
- Eikosenionska

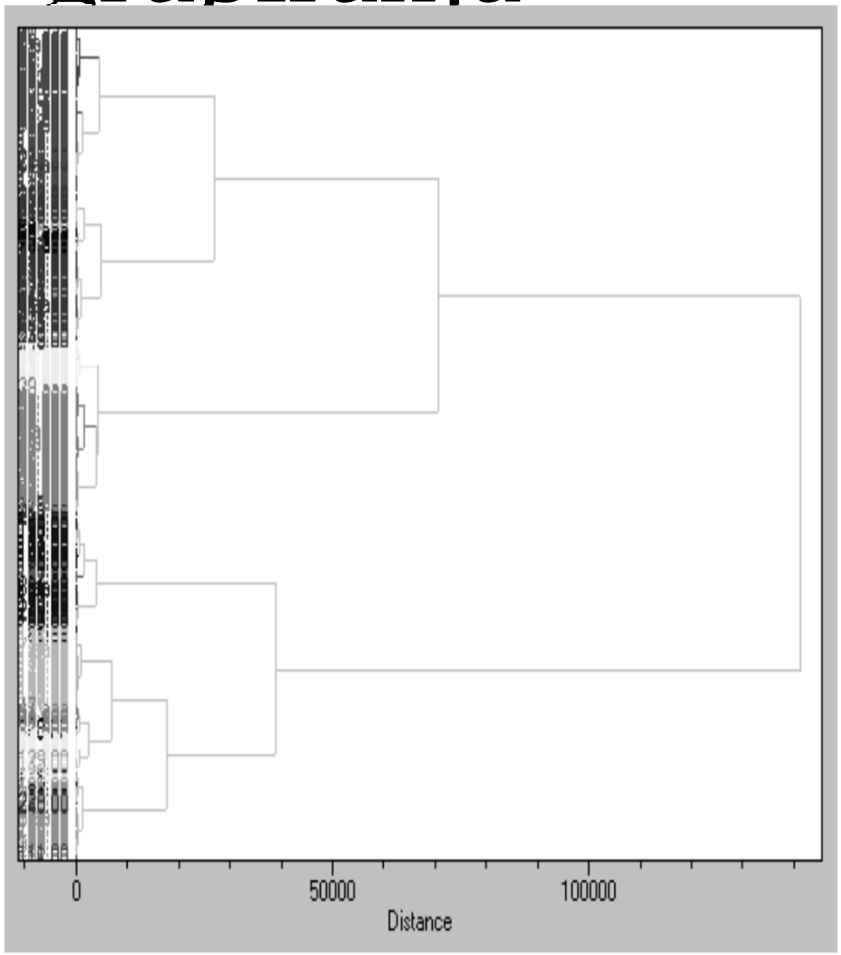
1. primer grupiranja

- povezave: single linkage
- razdalja: jaccard
- klasificiranje je neuspešno



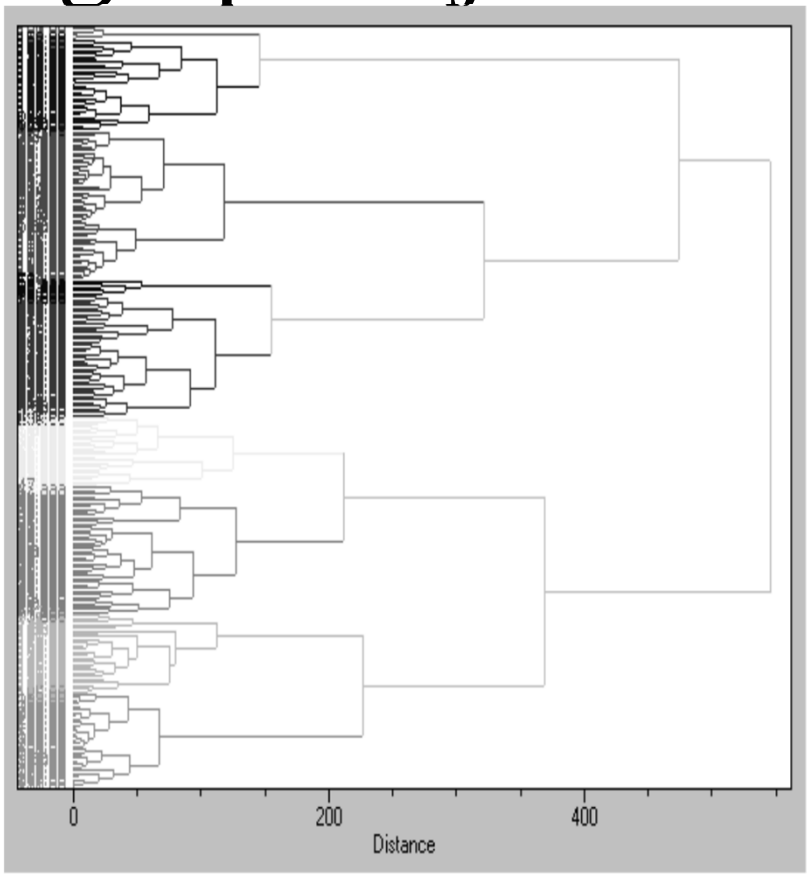
2. primer grupiranja

- Flexible strategy
- evklidska razdalja
- spremenimo parameter alfa na 0.5 pri razdalji 140
- pojavijo se ubežniki
- s spremembo preseka razdalje dobimo željeno število razredov



3. primer grupiranja

- vrsta povezave: flexible strategy
- vrsta razdalje: evklidska
- parameter alfa: 0.6
- presek pri razdalji 140: dobimo 9 klastrov brez ubežnikov



Linearna regresija in linarna regresija več faktorjev (Multivariate Linear Regression – MLR)

Za opis povezave med eno merljivo spremenljivko *x* (faktorjem) in odvisno spremenljivko *y* (odgovor ali response) potrebujemo matematično povezavo, ki ji pravimo model. V veliki večini primerov delamo z linearnimi modeli v obliki polinomov različnih redov ali stopenj (navadna enofaktorska linearna regresija):

Takoj je treba povedati, da polinome štejemo k linearnim modelom, ne glede na stopnje *p* (potence) faktorja *x*, s katerimi spremenlljivka *x* nastopa v polinomu. Polinomi so linearni glede na koeficiente b_i, ne pa glede na faktorje, ki stoje ob koeficientih b_i. Rešitev polinomskega modela, ob predpostavljeni obliki - stopnji polinoma, je določitev njegovih koeficientov b_i.
Tudi ko določamo parametre kalibracijske premice, je postopek enak: najprej izberemo obliko oziroma stopnjo polinoma (premica, *p=2*), nato pa določimo oba koeficienta (naklon *a* in odsek na ordinatni osi *b*).

V primeru, ko na odgovor *y* vpliva več faktorjev (x_1, x_2, …x_i,…x_m), moramo model opisati z večfaktorskim polinomom in reševati z večfaktorsko linearno regresijo (MLR):

$$y = b_0 + \underbrace{b_1 x_1 + b_2 x_2 + \ldots + b_m x_m}_{\text{Linearni členi}} + \underbrace{b_{m+1} x_1^2 + b_m x_2^2 + b_{m+2} x_1 x_2}_{\text{Kvadratni členi}} + \ldots + \underbrace{b_p x_i^s x_j^t x_k^v}_{\text{Členi višjih stopenj}} = \sum_{i=0}^{p} b_i u_i$$

Prosti člen

V polinomu so lahko ob koeficientih b_i nastopajo spremenljivke x_i, v poljubnih funkcijskih oblikah; npr. $\sin(x_1 + x_2)$ ali $\log(x_1 - x_2/2)$ in podobno. V vsakem primeru, le da je vsaka funkcijska oblika izračunljiva, ostane polinom linearen in metoda raševanja koeficientov b_i enaka za vse primere.

Reševanje sistema polinomskih enačb

Ne glede na to, ali gre za polinom v katerem nastopa en sam faktor z različnimi potencami, ali pa nastopa več faktorjev s poljubnimi potencami, se določanje koeficientov b_i, določa na enak način. Prvi pogoj za uspešno reševanje je, da moramo imeti vsaj en eksperiment več, kot je parametrov u_i. Število koeficientov b_i v polinomu označimo s črko p. Za določitev p koeficientov potrebujemo $p+1$ enačb (meritev). Izbor meritev na podlagi katerih izračunamo model, je zelo pomemben. (Glej poglavje o večnivojskih eksperimentalnih načrtih). Če je stopnja modelnega polinoma 2, 3 ali 4, potem moramo izbrati najmanj 3, 4, oziroma 5 nivojske delne eksperimentalne načrte.

Kot vidimo iz splošne enačbe polinoma, so merske spremenljivke x_i v njih lahko kombinirane v faktorje u_k. Ker vrednosti neodvisnih spremenljivk x_{ij} s katerimi naredimo eksperiment in določimo y_j,, izberemo sami, lahko vse faktorje u_{ij} izračunamo in potem so vse vrednosti v polinomskih enačbah, razen koeficientov b_i, znani. Pri MLR se razlika med faktorji in merljivimi spremenljivkami vidi razločno: fatorji u_k so sestavljeni iz spremenljivk x_i. Ne glede na njihovoe sestavljenost, bomo faktorje pisali s črko u_k (glej desno stran spodnje enačbe na prejšnji strani!). Sistem n enačb, ki sledijo iz meritev je naslednji:

$$\begin{aligned} y_1 &= b_0 + b_1 u_{11} + b_2 u_{12} + \ldots + b_p u_{1p} \\ y_2 &= b_0 + b_1 u_{21} + b_2 u_{22} + \ldots + b_p u_{2p} \\ &\ldots \\ y_n &= b_0 + b_1 u_{n1} + b_2 u_{n2} + \ldots + b_p u_{np} \end{aligned}$$

Sistem polinomskih enačb lahko najprej napišemo v razviti vektorsko-matrični obliki,

$$\begin{vmatrix} y_1 & y_2 & \ldots & y_n \end{vmatrix} = \begin{Vmatrix} 1 & u_{11} & u_{12} & \ldots & u_{1p} \\ 1 & u_{21} & u_{22} & \ldots & u_{2p} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & u_{n1} & u_{n2} & \ldots & u_{np} \end{Vmatrix} \begin{vmatrix} b_0 \\ b_1 \\ \ldots \\ b_p \end{vmatrix}$$

$$\vec{Y}^T = \| U \| \vec{B} \Longleftarrow$$

nato pa še v simbolni vektorski pisavi.

Reševanje vektorsko-matrične enačbe zahteva nekoliko več previdnosti kot reševanje skalarne enačbe. Množenje podatkovne matrike **U** z njeno transponirano obliko $\mathbf{U^T}$, potrebujemo zato, da naredimo kvadratno matriko, $\mathbf{U^TU}$, ki ima determinanto različno od nič. Inverzno matriko $(\mathbf{U^TU})^{-1}$ pa potrebujemo zato, ker je tak način v matrični algebri edina možnost za deljenje.

$||U||^T Y^T = ||U||^T ||U|| B$ — obe strani množimo na levi s transponirano matriko $\mathbf{U^T}$

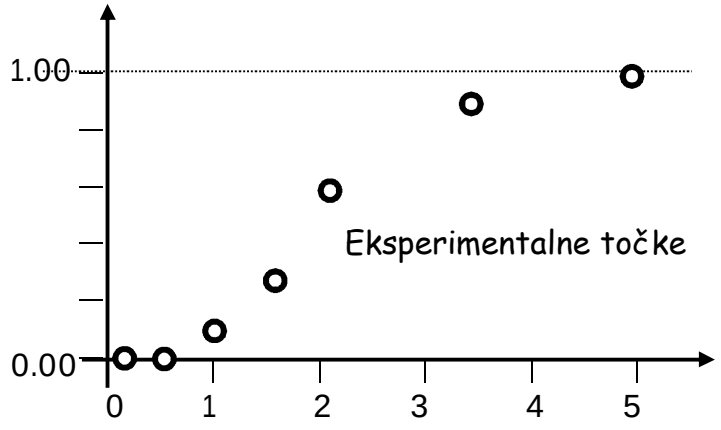
$(||U||^T ||U||)^{-1} ||U||^T Y^T = (||U||^T ||U||)^{-1} ||U||^T ||U|| B$ — obe strani množimo na levi z inverzno matriko $(\mathbf{U^TU})^{-1}$

$B = (||U||^T||U||)^{-1} ||U||^T Y^T$ — dobimo enačbo za izračun vseh neznanih koeficientov $b_0, b_1, b_2, \ldots b_p$

$\hat{Y}=||U||B = \underbrace{||U|| (||U||^T||U||)^{-1} ||U||^T}_{||H||} Y^T$ — Z matriko $\mathbf{H} = \mathbf{U(U^TU)^{-1}U^T}$, ki jo imenujemo tudi *Hat* matrika, lahko izračunamo modelne vrednosti v točkah na katerih smo modelirali (model postavljali)

Edini pogoj, ki ga imamo pri opisanem matričnem reševanu linearne enačbe je, da mora imeti matrika **U** več vrstic kot kolon (p > m) in da mora biti njena determinant od nič različna. Z drugimi besedami imeti moramo več eksperimentov, kot imamo v enačbi modela koeficientov b_i. Eksperimenti ne smejo biti identični.

Primer: Modeliranje s polinomom tretje stopnje



i	Odziv y_i	Eksperiment x_i	Izračunane x^2	vrednosti x^3
1	0.01	0.1	0.01	0.001
2	0.02	0.5	0.25	0.125
3	0.11	1.0	1.00	1.000
4	0.30	1.5	2.25	3.375
5	0.58	2.0	4.00	8.000
6	0.88	3.5	12.25	42.875
7	0.99	5.0	25.00	125.000

Primer za izračun transponirane (**U^T**) in inverzne matrike (**U^{-1}**)

Imamo sedem meritev x_i *(i=1…7)* sigmoidnega odziva y_i nekega inštrumenta in želimo ta odziv modelirati. Domnevamo, da bo zadostovala kubična enačba:

$$y = b_0 + b_1x + b_2x^2 + b_3x^3$$

V skladu s predpostavljeno kubično enačbo, napišemo sedem enačb, po eno za vsako meritev. Kot primer navajamo samo enačbo za četrto meritev :

$$0.21 = b_0 + 1.5\ b_1 + 2{,}25\ b_2 + 3.375\ b_3$$

Ustrezna vektorja *Y* in *B* sta naslednja: $\vec{Y} = \begin{vmatrix} 0.01 & 0.02 & 0.11 & 0.30 & 0.58 & 0.88 & 0.99 \end{vmatrix}$ $\vec{B} = \begin{vmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{vmatrix}$

Matrika **U** in njena transponirana oblika **U^T**:

$$U = \begin{Vmatrix} 1 & 0.1 & 0.01 & 0.001 \\ 1 & 0.5 & 0.25 & 0.125 \\ 1 & 1.0 & 1.00 & 1.000 \\ 1 & 1.5 & 2.25 & 3.375 \\ 1 & 2.0 & 4.00 & 8.000 \\ 1 & 3.5 & 12.25 & 42.875 \\ 1 & 5.0 & 25.00 & 125.00 \end{Vmatrix} \qquad U^T = \begin{Vmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0.1 & 0.5 & 1 & 1.5 & 2 & 3.5 & 5 \\ 0.01 & 0.25 & 1 & 2.25 & 4 & 12.25 & 25 \\ 0.001 & 0.125 & 1 & 3.375 & 8 & 42.875 & 125 \end{Vmatrix}$$

produkt **U^{T}U** in njegova inverzna matrika (**U^{T}U**)$^{-1}$:

$$U^TU = 10^2 \times \begin{Vmatrix} 0.07 & 0.14 & 0.45 & 1.80 \\ 0.14 & 0.45 & 1.80 & 7.97 \\ 0.45 & 1.80 & 7.97 & 3.69 \\ 1.80 & 7.97 & 3.69 & 17.54 \end{Vmatrix} \qquad (U^TU)^{-1} = \begin{Vmatrix} 1.055 & -1.684 & 0.686 & -0.077 \\ -1.684 & 3.996 & -1.900 & 0.235 \\ 0.686 & -1.900 & 0.990 & -0.129 \\ -0.077 & 0.235 & -0.129 & 0.017 \end{Vmatrix}$$

Za izračun koeficientov b_i, ki so komponente vektorja B, manjka še matrika (**U^{T}U**)$^{-1}$**U^T**.

(Računanje z matrikami je najenostavnejše z uporabo programa MatLab®),

$$(U^TU)^{-1}U^T = \begin{Vmatrix} 0.894 & 0.375 & -0.021 & -0.193 & -0.198 & 0.191 & -0.048 \\ -1.303 & -0.131 & 0.648 & 0.831 & 0.594 & -0.874 & 0.234 \\ 0.506 & -0.033 & -0.353 & -0.372 & -0.187 & 0.631 & -0.192 \\ -0.056 & 0.009 & 0.045 & 0.043 & 0.015 & -0.093 & 0.037 \end{Vmatrix}$$

Ko imamo izračunan $(\mathbf{U^TU})^{-1}\mathbf{U^T}$ lahko izračunamo koeficiente $B = (\mathbf{U^TU})^{-1}\mathbf{U^T}Y^T$

$$\vec{B} = (U^TU)^{-1}U^TY^T = \begin{Vmatrix} 0.894 & 0.375 & -0.021 & -0.193 & -0.198 & 0.191 & -0.048 \\ -1.303 & -0.131 & 0.648 & 0.831 & 0.594 & -0.874 & 0.234 \\ 0.506 & -0.033 & -0.353 & -0.372 & -0.187 & 0.631 & -0.192 \\ -0.056 & 0.009 & 0.045 & 0.043 & 0.015 & -0.093 & 0.037 \end{Vmatrix} \times \begin{vmatrix} 0.01 \\ 0.02 \\ 0.11 \\ 0.30 \\ 0.58 \\ 0.88 \\ 0.99 \end{vmatrix} = \begin{vmatrix} -0.038 \\ 0.112 \\ 0.111 \\ -0.019 \end{vmatrix}$$

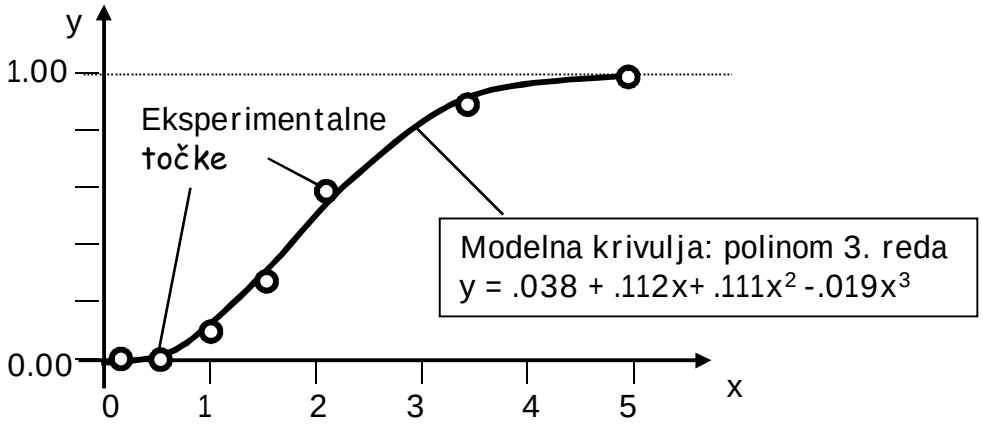
in s tem imamo določen končni modelni polinom 3. stopnje

$$\mathbf{Y_i^{model} = -0.038 + 0.112x_j + 0.111x_j^2 - 0.019x_j^3}$$

s katerim lahko izračunamo y_j^{model} za poljubno vrednost x_j:. To lahko storimo tako, da uporabimo gornjo enačbo, lahko pa tudi preko vektorske enačbe s tako imenovano "hat" matriko $\mathbf{H=U(U^TU)^{-1}U^T}$. Izračun poteka preko osnovne enačbe premice: $Y = \mathbf{U}B = \mathbf{U(U^TU)^{-1}U^T}Y^T$:

$$y^{model} = \hat{y} = [U(U^TU)^{-1}U^T]y^T = \begin{Vmatrix} 0.769 & 0.362 & 0.040 & -0.113 & -0.141 & 0.110 & -0.026 \\ 0.362 & 0.303 & 0.220 & 0.135 & 0.054 & -0.099 & 0.026 \\ 0.040 & 0.220 & 0.319 & 0.309 & 0.224 & -0.144 & 0.032 \\ -0.113 & 0.135 & 0.309 & 0.361 & 0.322 & -0.012 & -0.002 \\ -0.141 & 0.054 & 0.224 & 0.322 & 0.361 & 0.228 & -0.048 \\ 0.110 & -0.099 & -0.144 & -0.012 & 0.228 & 0.894 & 0.024 \\ -0.026 & 0.026 & 0.032 & -0.002 & -0.048 & 0.024 & 0.995 \end{Vmatrix} \times \begin{vmatrix} 0.01 \\ 0.02 \\ 0.11 \\ 0.30 \\ 0.58 \\ 0.88 \\ 0.99 \end{vmatrix} = \begin{vmatrix} -0.026 \\ 0.044 \\ 0.167 \\ 0.319 \\ 0.483 \\ 0.921 \\ 0.981 \end{vmatrix}$$

i	Eksperiment v točki x_i	Odziv y_i	Izračunani odziv y^{model}	Razlika d_i	d_i^2
1	0.1	0.01	-0.03	0.04	0.0016
2	0.5	0.02	0.04	-0.02	0.0004
3	1.0	0.11	0.17	-0.06	0.0036
4	1.5	0.30	0.32	-0.02	0.0004
5	2.0	0.58	0.48	0.10	0.0100
6	3.5	0.88	0.92	-0.04	0.0016
7	5.0	0.99	0.98	0.01	0.0001



Ko imamo vse eksperimentalne podatke (tudi ponovitve meritev v posameznih **točkah** x_j), **izračunan** model, lahko z ANOVO preverimo (testiramo, kako dober je dani model za naše podatke (glej poglavje o ANOVI).

Več faktorska linearna regresija (MLR)

V primerih, ko imamo namesto ene spremenljivke *x* <u>več različnih</u> spremenljivk x_1, x_2 …x_j … x_m (ponovitve meritev *j*-te **točke** so x_{ij}), je postopek **računanja** modela <u>povsem identičen</u> s postopkom modeliranja pri navadni regresiji, ki smo jo pravkar obravnavali. Najprej predpostavimo model, ki se nam zdi primeren. Npr.;

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_1x_2 + b_5x_2x_3 + b_6x_3^2 + b_7x_1x_2x_3$$

Nato naredimo matriko **U** iz vseh faktorjev (v tem primeru sedmih: x_1, x_2, x_3, x_1x_2, x_2x_3, x_3^2 in $x_1x_2x_3$) in s primernim številom meritev (v tem primeru najmanj osmih) pri katerih spreminjamo vse tri spremenljivke x_1, x_2 in x_3 , po možnosti po kakem eksperimentalnem **načrtu** (glej poglavje o delnih eksperimentalnih **načrtih**). Ko je matrika **U** narejena, je nadaljnji postopek **računanja** modela povsem **identičen** s prikazanim.

Modeliranje olivnih olj

- Dobimo linearni model oblike: $Y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_nx_n + e$, pri čimer je Y odvisna spremenljivka (odziv), $x_1 \ldots x_n$ so neodvisne spremenljivke (napovedne vrednosti), e je naključna napaka, $b_0 \ldots b_n$ pa so koeficienti regresije dobljeni iz podatkov

- MLR je uporabna za napoved odziva za različne spremenljivke ali za ugotavljanje zvez med odzivnimi in napovednimi spremenljivkami.

- Z MLR lahko tako povežemo izvor olja z različnimi faktorji, kot so temperatura, nadmorska višina, padavine,geografske širine, števila (pokrajine Italije od juga proti severu)…

- Podatke olivnih olj smo obdelali z Microsoft Excelom / tools / data analysis / regression in dobili koeficiente ($b_0 \ldots b_n$) in njihovo pomembnost (t vrednost)

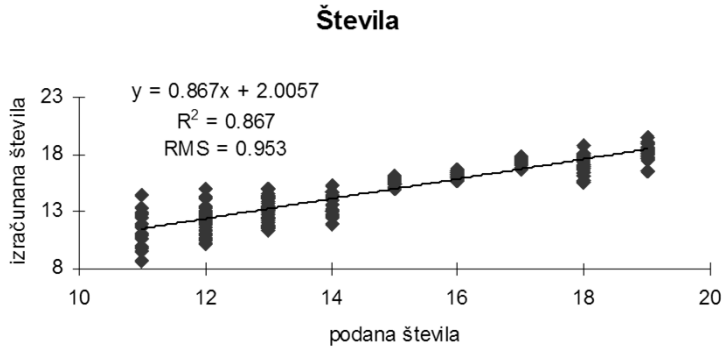
Regression Statistics	
Multiple R	0.931105
R Square	0.866956
Adjusted R	0.86254
Standard E	0.971143
Observatio	250

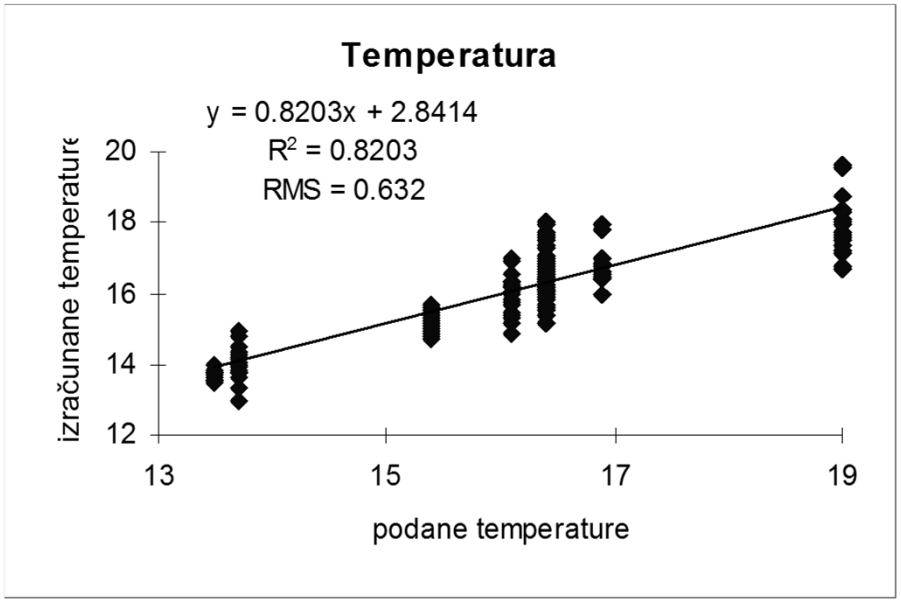
ANOVA

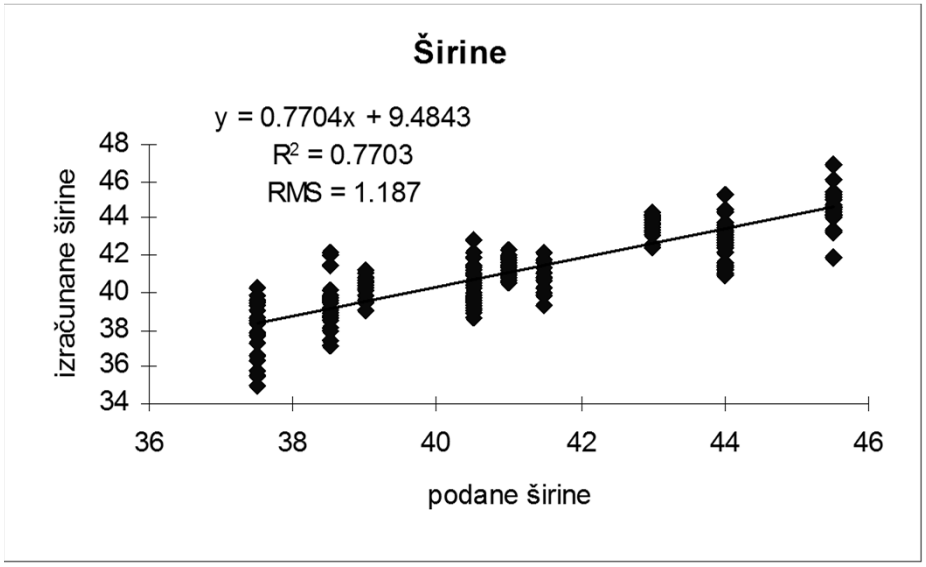
	df	*SS*	*MS*	*F*	*ignificance F*
Regressior	8	1481.108	185.1385	196.3045	5.5E-101
Residual	241	227.2917	0.943119		
Total	249	1708.4			

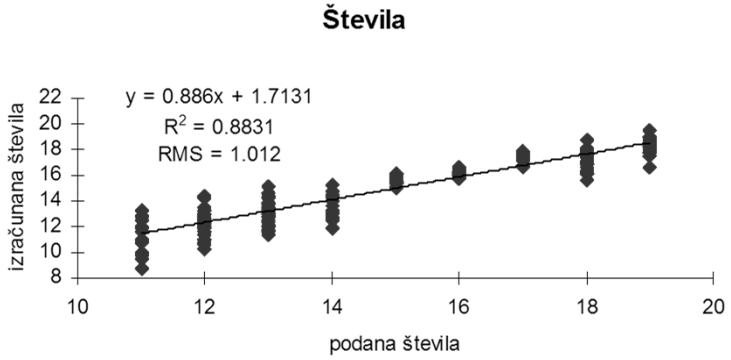
	Coefficients	*tandard Err*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*ower 095%*	*Jpper 095%*
Intercept	-13.1715	6.967895	-1.89031	0.059917	-26.8972	0.554293	-26.8972	0.554293
X Variable	0.111449	0.035477	3.14146	0.001891	0.041565	0.181334	0.041565	0.181334
X Variable	0.049589	0.009907	5.005709	1.07E-06	0.030075	0.069104	0.030075	0.069104
X Variable	0.035386	0.007342	4.819852	2.54E-06	0.020924	0.049849	0.020924	0.049849
X Variable	0.289	0.058972	4.900617	1.75E-06	0.172833	0.405167	0.172833	0.405167
X Variable	0.109104	0.02786	3.916107	0.000117	0.054223	0.163985	0.054223	0.163985
X Variable	-0.00885	0.00677	-1.3067	0.19256	-0.02218	0.00449	-0.02218	0.00449
X Variable	0.008214	0.004953	1.658429	0.098532	-0.00154	0.01797	-0.00154	0.01797
X Variable	-0.05732	0.004187	-13.6906	6.01E-32	-0.06557	-0.04907	-0.06557	-0.04907

Začetni grafi brez izločenih parametrov

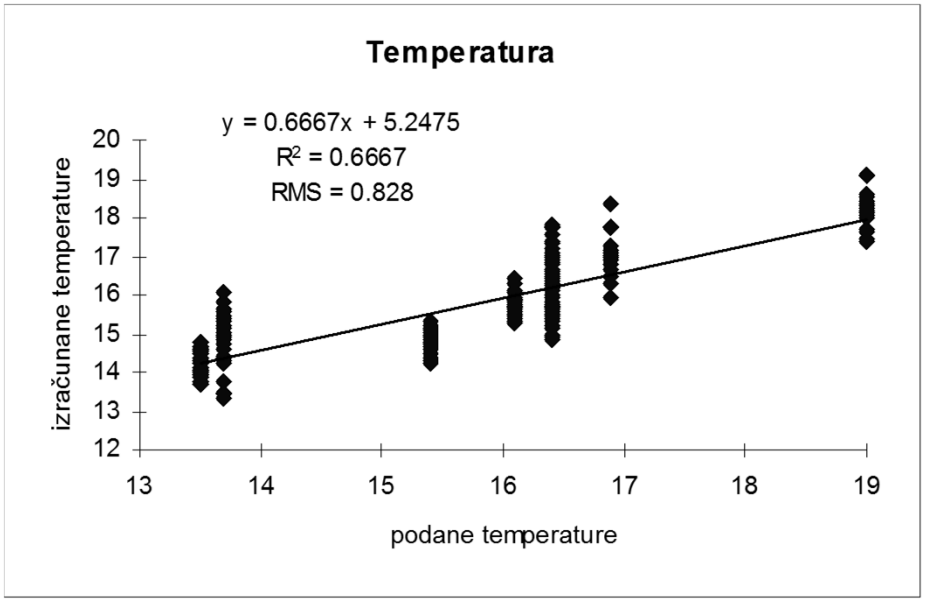


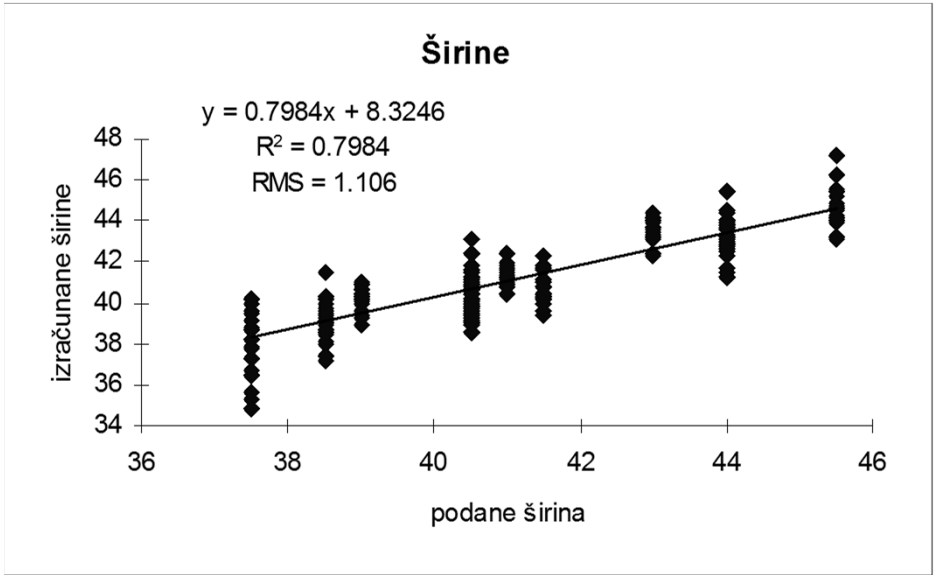


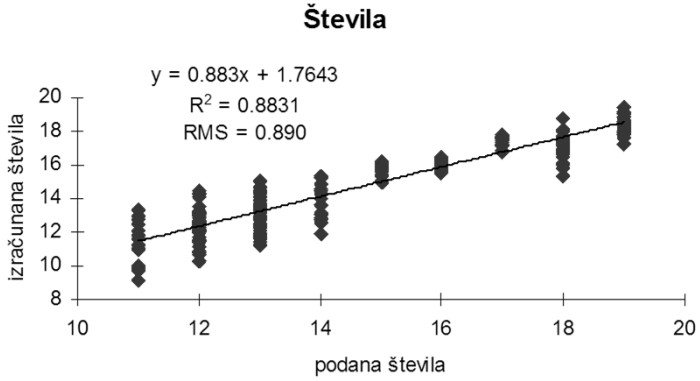




Na osnovi RMS izločena olja,
ki presegajo vrednost t*RMS







Na podlagi t testa izločene še maščobne kisline.