Univerzitet u Novom Sadu

Prirodno-matematički fakultet

Departman za matematiku i informatiku

Miloš Banjanin

Analiza sličnosti podataka

-Master rad-

Mentor: dr Zagorka Lozanov-Crvenković

Novi Sad, 2016

Sadržaj

1 Uvod

"There is no
statistical tool that is as powerful as a well-chosen graph!"

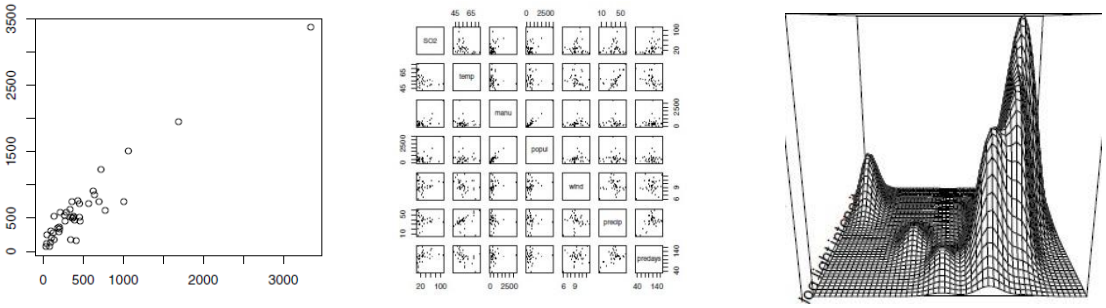
Ideja prikazivanja stvari u što prostijem obliku je odavno poznata ljudima, a kako je većina podataka dobijenih iz raznih istraživanja, upitnika i anketa multivarijantna, i u statistici se ta ideja koristi kako bi se veliki broj podataka predstavio pomoću manjeg broja podataka, koji je mnogo razumljiviji i koji se na jednostavniji način može predstaviti grafički.

Poznato je da grafička prezentacija numeričkih rezultata ima mnoštvo prednosti u odnosu tabelarnu, I to ne samo u stvaranju interesa I privlačenju pažnje čitaoca. Grafikoni podataka vizuelno prikazuju izmerene podatke kombinujući tačke, linije, koordinatni sistem, brojeve, simbole I boje. Procenjuje se da se svake godine odštampa između 900 milijardi (9×10^{11}) I 2 triliona(2×10^{12}) statističkih grafika. Pretpostavlja se da je jedan od razloga ovolike popularnosti grafičke prezentacije što ona često predstavlja motor za otkrivanje neočekivanog; ljudski vizuelni sistem je veoma moćan u otkrivanju šablona. Neke od prednosti korišćenja grafičkih metoda:

- U poređenju sa drugim tipovima prezentacija, dobro dizajnirani grafikoni imaju više efekta u privlačenju pažnje publike
- Vizuelne veze prikazane kao grafikoni I dijagrami su lakše shvaćeni I lakše se pamte
- Korišćenje grafikona čuva vreme, s obzirom da osnovna značenja velikih merenja mogu biti vizualizovana na prvi pogled.

Tokom poslednje dve decenije razvijen je velik broj novih metoda za prikazivanje podataka grafički. Ove metode pomažu u pronalaženju autlajera (eng.outliers), šablona, dijagnostifikovanju modela itd. Grafički prikaz bi trebalo da pomogne oko sređivanja velikog broja podataka I da smanji kognitivni napor potreban da se napravi poređenje.

Da bi željeni cilj bio postignut, vrlo često je potreban veliki broj takvih grafikona I danas se koriste kompjuteri za to, iz razloga jer su kompjuteri brzi I pouzdani u izvođenju tih operacija.



Slika 1.1: Primeri 2D i 3D grafika

Ovaj rad će se baviti dvema tehnikama koje nam pomažu pri obradi velikog broja podataka. U prvom delu rada biće reči o multidimenzionalnom skaliranju (MDS). Jedan od problema sa velikim skupovima multivarijantnih podataka je taj što jednostavno ima puno promenljivih da bi primenili neke od poznatih grafičkih tehnika. Problem mnogo promenljvih je ponekad poznat kao *kletva dimenzionalnosti.* Multidimenzionalno skaliranje je (vidi u [3]) jedna od multivarijantnih statističkih tehnika pomoću koje se pronalazi skup tačaka malih dimenzija koji najbolje aproksimira visoko dimenzionalnu konfiguraciju podataka, predstavljenu početnom matricom bliskosti. Multidimenzionalno skaliranje otkriva strukturu skupa podataka, crtajući tačke u jednoj, dve ili tri dimenzije, ukoliko je to moguće. Udaljenost između tačaka se podudara sa posmatranim sličnostima koliko god je to moguće. Na početku poglavlja biće izloženo kako se dobija matrica bliskosti iz skupa podataka, kakvi tipovi podataka mogu biti korišćeni u MDS-u, kao i matematički pristup multidimenzionalnom skaliranju. U nastavku rada govoriće se o dve metode, klasičnom multidimenzionalnom skaliranju i ne-metričkom multidimenzionalnom skaliranju.

U drugom delu rada govoriće se o klaster (eng. cluster) analizi. Tokom života ljudi svakodnevno klasifikuju stvari. Jedna od osnovnih sposobnosti živih bića jeste grupisanje sličnih objekata. Od davnina ljudi su morali da razlikuju otrovne biljke od hranljivih, pitome životinje od divljih, a danas mnoge bolesti, da bi se lečile i razumele, moraju se klasifikovati.

Slika 1.2: Jedan od načina sortiranja stvari

Klasifikacija može predstavljati odgovarajući metod za organizovanje velikog skupa podataka, tako da se skup može lakše razumeti. Ukoliko podatke, na odgovarajući način možemo sažeti u mali broj grupa objekata, tada sami nazivi grupa objekata mogu bliže opisati podatke u grupama. Pravilnom primenom i dobrim razumevanjem kako multidimenzionalnog skaliranja tako i klaster analize, velika količina podataka se može lako obrađivati, i potom koristiti u razne svrhe. Od marketinga, koji iziskuje obradu velikog broja podataka zbog targetiranja, pa sve do medicine gde je veoma važno imati prave klasifikacije i podatake o samim grupama lekova ili bolesti, i upravo na tome će biti baziran ovaj rad, kako bi bliže objasnio ova dva tipa statističke analize.

Vrlo često se dešava da se određeni objekti nalaze u više grupa. Na primer, ljudi se mogu klasifikovati po polu, ali i po krvnoj grupi, tako da će svaka klasifikacija dati različite rezultate. U ovom radu biće objašnjeni pojmovi odstojanja, kao što su Euklidsko, Menhetn (eng.Manhattan), odstojanje Minkovskog (eng.Minkowski) i još neka. Neke od tehinka klaster analize koje će biti obrazložene su:

- hijerarhijske metode (poređenje aglomerativnih metoda i metoda deljenja)
- nehijerarskijske metode (metoda k-sredina)

Na samom kraju biće sumirani rezultati rada.

2 Istorija nastanka multivarijantne analize

Multivarijantna analiza se prvi put pojavljuje krajem 19. veka, u radovima Francisa Galtona[1] i Karla Pirsona[2], tokom proučavanja odnosa između potomaka i roditelja, kao i pri razvijanju koeficijenta korelacije. Nakon toga, početkom 20. veka, Čarls Spirman[3] polaže temelje faktorske analize, tokom istraživanja testa koeficijenta inteligencije (IQ). U naredne dve decenije, Spirmanov rad su nastavila dva naučnika, Hoteling[4] i Turstone[5].

U samom početku, teret izračuvanja velikog broja aritmetičkih aplikacija multivarijantnih metoda je bio matematički i uglavnom deo linearne algebre, i stoga su postojala velika ograničenja. Dolaskom modernih računara koji preuzimaju posao računanja od ljudi, povećava se samo interesovanje za postojeće metode i to dovodi do pronalaska nekih novih metoda multivarijantne analize. Kako su personalni računari postajali sve brži, a statistički softveri pristupačniji, metoda multivarijantne analize su uspešno primenjivani na sve veći broj podataka. Naročitu primenu, su imali u genetici, astronomiji, u poslednje vreme u marketingu i obrađivanju slike.

[1]Francis Galton (1822-1911), engleski matematičar i statističar

[2]Karl Pearson (1857-1936), britanski matematičar i statističar, poseban doprinos dao poljima biometrike i meteorologije. Osnivač prvog departmana statistike na University College London

[3]Charles Spearman (1863-1945), engleski psiholog, poznat po svom doprinosu statistici, kao pionir faktorske analize i po Spirmanovom koeficijentu korelacije

[4]Harold Hotelling (1895-1973), američki statisičar i ekonomski teoretičar, poznat po Hotelingovom zakonu i Hotelingovoj T-kvadratnoj distribuciji u statistici

[5]Louis Leon Thurstone (1887-1955), američki naučnik u poljima psihometrike i psihofizike. Dao velik doprinos faktorskoj analizi

3 Multivarijantni podaci

3.1 Udaljenosti

Za tehnike kao što su multidimenzionalno skaliranje i klaster analiza, koncept distance (udaljenosti) izmedju objekata u skupu podataka je od velike važnosti. Za određivanje blizine između objekata koriste se mere bliskosti, kao što su razlika, udaljenost ili sličnost među objektima. Dva objekta su bliža jedan drugome, kada je razlika ili udaljenost među njima mala, a sličnost velika. Pretpostavimo da imamo promenljivu sa dva objekta, objekat *i* i objekaat *j.*

Za meru bliskosti d_{ij} kažemo da predstavlja *meru razlike* objekata i i j ako zadovoljava sledeće:

1. $d_{ij} > 0$, ako se objekti i i j razlikuju, a $d_{ij} = 0$, samo ako su objekti identični (uslov nenegativnosti);
2. $d_{ij} = d_{ji}$ (uslov simetričnosti);
3. $d_{ij} \leq d_{ik} + d_{kj}$ za sve objekte i, j i k (uslov triangularnosti).

Za meru bliskosti d_{ij} kažemo da predstavlja *meru sličnosti* objekata i i j ako zadovoljava sledeće:

1. $0 \leq d_{ij} \leq 1$, za sve objekte i i j (uslov normalnosti);
2. $d_{ij} = 1$, samo ako su objekti i i j identični;
3. $d_{ij} = d_{ji}$, (uslov simetričnosti).

Razlikuju se mere bliskosti za promjenljive sa neprekidnim i promjenljive sa kategoričkim vrednostima. Pitanje je kako bi se mogla da izmeriti udaljenost između ova dva objekta? Najčešći metod koji se koristi je Euklidska distanca.

Definicija 1. Euklidsko odstojanje je definisano kao

$$d_{ij} = \sqrt{\sum_{k=1}^{q} \left(x_{ik} - x_{jk}\right)^2}$$

gde su x_{ik} i x_{jk} za $k = 1, \dots, q$ vrednosti promenljive za objekte *i* i *j,* respektivno. Euklidsko odstojanje predstavlja dužinu najkraće linije koja spaja tačke *i* i *j*. Ovo odstojanje je poznato kao i l_2 norma.

Definicija 2. Menhetn odstojanje je odstojanje definisano kao

$$d_{ij} = \sum_{k=1}^{q} |x_{ik} - x_{jk}|$$

i predstavlja odstojanje gradskog bloka (poznato onim čitaocima koji su šetali ulicama New York-a). Ovo odstojanje predstavlja l_1 normu.

Definicija 3. Odstojanje Minkowski je predstavljeno kao

$$d_{ij} = \left(\sum_{k=1}^{q} (x_{ik} - x_{jk})^r \right)^{\frac{1}{r}}, r \geq 1$$

Primetimo da su Menhetn i Euklidsko odstojanje specijalni slučajevi odstojanja Minkovskog za $r = 1$ i $r = 2$, respektivno.

3.2 Matrica sličnosti

Multivarijantni podaci se pojavljuju kada istraživači zapisuju vrednosti nekoliko slučajnih promenljivih koje istražuju, i koje se sastoje od više "objekata", što dovodi do višedimenzionalne (multidimenzionalne) observacije za svaku promenljivu. Ovakvi podaci se skupljaju u mnogim oblastima, i mnogi podaci sa kojima se susrećemo svakodnevno, su multivarijantni. Većina multivarijntnih podataka se predstavalja u pravougaonom obliku, gde elementi svake vrste odgovaraju vrednostima jednog objekta promenljivih u skupu podataka, a elementi kolone odgovaraju vrednostima jedne promenljive. Neka je n broj različitih objekata i neka su razlike (sličnosti) između objekata *i* i *j* dati sa x_{ij}.

Definicija 1. Polazna matrica (matrica podataka) **X** je pravougaona matrica formata $n \times q$

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1q} \\ x_{21} & x_{22} & \cdots & x_{2q} \\ \vdots & \vdots & \ddots & \\ x_{n1} & x_{n2} & \cdots & x_{nq} \end{bmatrix}$$

gde je n broj objekata, a q je broj promenljivih. Dakle, x_{ij} predstavlja vrednost j-te promenljive i-tog objekta. Suprotno od posmatranih multivarijantnih podataka, kod univarijantnih podataka podaci q promenljivih su predstavljeni kao slučajne promenljive $X_1, \dots, X_q$.

Definicija 2. Matrica udaljenosti, sličnosti ili razlike je $n \times n$ matrica

$$D = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

dobijena iz matrice **X**.

4 Multidimenzionalno skaliranje

"*A picture is worth more than a thousand words"*

Multidimenzionalno skaliranje (MDS) je jedna od nekoliko multivarijantnih tehnika koja ima za cilj, da otkrije strukturu skupa podataka, crtajući tačke u malom broju dimezija (najčešće u jednoj ili dve dimenzije). Tačke na grafiku bi trebalo da što bolje predstavljaju Euklidske distance posmatranih objekata. MDS tehnika se ne primenjuje direktno na početnu matricu podataka **X**, nego na matricu udaljenosti **D** (često se ova matrica naziva i matrica sličnosti ili razlike). Ukoliko se radi o nekim subjektivnim ocenama, kao što je na primer ocenjivanje koliko su neke dve boje slične, ne može se reći da je neko dao pogrešnu ocenu, i tada se dobija matrica sličnosti tj. razlike. Često se koristi i izraz bliskost podataka, kojim se obuhvataju i sličnosti i razlike.

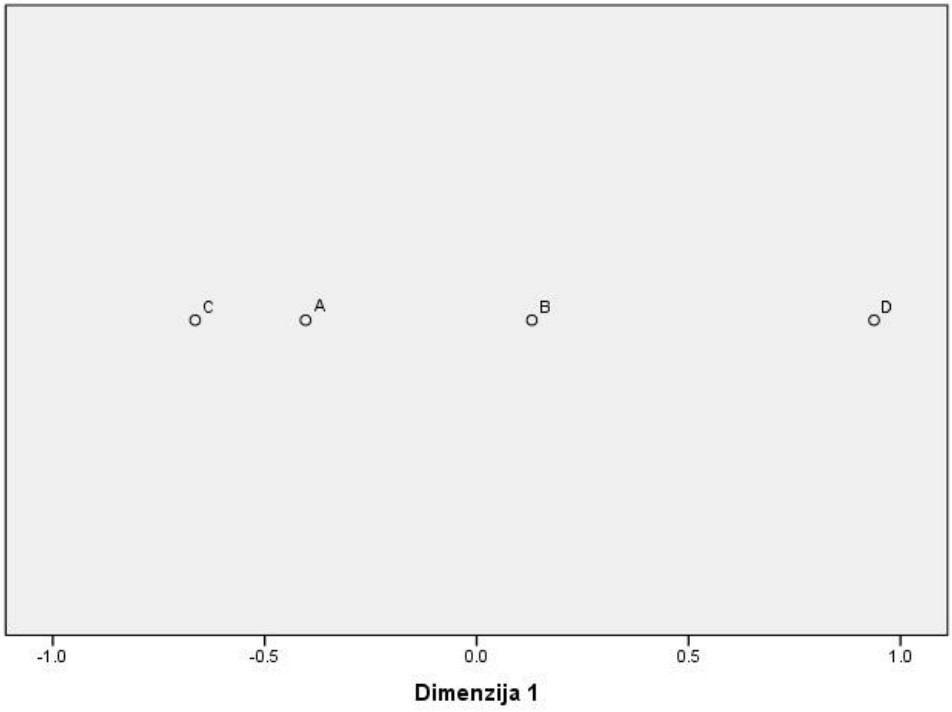
Osnovna ideja može biti predstavljena primerom iz geografije. Pretpostavimo da su date udaljenosti između gradova. Mogući zadatak bi bio da se rekonstruiše dvodimenzionalna mapa iz koje su dobijene te udaljenosti. Jedan od načina koji se može primeniti, jeste da se te tačke fizički pomeraju na papiru dok se ne dobije zadovoljavajuća „mapa". Ovaj proces je moguć samo ukoliko imamo veoma mali skup podataka, i očigledno je da to iziskuje mnogo vremena. Proces koji ovo radi automatski je multidimenzionalno skaliranje. Treba napomenuti da nismo ograničeni na crtanje samo u dve dimenzije, već u onoliko koliko korisnik želi. Vratimo se na naš primer udaljenosti gradova, i obratimo pažnju kako se naš primer razlikuje od tipičnog MDS-a. Kao prvo, reč „udaljenost" je veoma jednosmislena u ovom slučaju (bilo da se radi miljama ili kilometrima), dok kod MDS-a ta reč može predstavaljati subjektivan osećaj onoga koji istražuje date podatke (u ovom slučaju reč udaljenost će se pre odnositi na sličnost ili razliku nego stvarnu udaljenost između podataka). Drugo, opšte je poznato da se gradovi mogu predstaviti u dve dimenzije (ignorisati zakrivljenost Zemljine lopte), a u MDS-u će biti poznato vrlo malo, koliko je dimenzija potrebno da bi makar i približno bile predstavljene udaljenosti između posmatranih objekata. Čak i ukoliko se ispostavi da je potrebno više od dve dimenzije, najlakši način da se vide te tačke je ukoliko su nacrtane u dve dimenzije.

Primer 1. Označimo sa A, B, C i D četiri grada. Udaljenosti gradova su date u tabeli (u stotinama kilometara)

	A	B	C	D
A	.			
B	6.000	.		
C	5.000	7.000	.	
D	9.000	7.000	10.000	.

Tabela 4.1: Primer udaljenosti 4 grada

Koristeći MDS ili pomeranje gradova ručno, moguće je udaljenosti predstaviti tačno u jednoj dimenziji. Jedno od mogućih rešenja je dato sa



Slika 4.1: Jednodimenzionalna konfiguracija četiri grada dobijena korišćenjem klasičnog MDS-a

Svaka MDS analiza se počinje sa matricom udaljenosti **D**. Izbor između udaljenosti i bliskosti nije od tolikog značaja u klaster analizi, dok je u MDS-u udaljenost primarni koncept. To znači da iako se počne sa matricom sličnosti ili razlike, verovatno će se morati pretvoriti u matricu udaljenosti.

Kao što je rečeno, MDS, se koristi kako bi se utvrdilo da li se matrica udaljenosti može prikazati pomoću grafika u malom broju dimenzija, tako da udaljenosti na grafiku približno reprezentuju stvarne distance, tj. matricu distanci $\{x_{ij}\}$, što znači da dva objekta najbliža jedan drugome prema matrici udaljenosti, moraju da budu i najbliži jedan drugome na grafiku. Prostorni prikaz matrice sličnosti sastoji se od skupa od n m-dimenzionalnih koordinata, tako da svaka predstavlja jedan od n objekata. Potrebne koordinate se uglavnom nalaze minimiziranjem neke mere fitovanja između distanci. Uglavnom (ali ne i stalno) distance su Euklidske. Idealno je ukoliko je broj dimenzija, m, mali, na primer dve ili tri dimenzije, tako da se to lako može nacrtati. MDS je u osnovi tehnika koja smanjuje skup podataka (eng.data reduction technique), i koja za cilj ima da nađe niskodimenzionalni skup podataka koja aproksimira visokodimenzionalnu konfiguraciju predstavljenu u početnoj matrici sličnosti. U ovom radu govoriće se o dvema tehnikama multidimenzionalnog skaliranja, *klasično multidimenzionalno skaliranje (CMDS)* i *ne-metričko multidimenzionalno skaliranje(OMDS).* Kod klasičnog multidimenzionalnog skaliranja udaljenosti na mapi će biti u istoj metrici (skali merenja) kao i originalni x_{ij}. Suprotno, recimo u društvenim naukama, vrlo često će vrednosti x_{ij} biti interpretirane u ordinalnom smislu, kao što je slučaj sa subjektivnim upoređivanjem objekata. U tom slučaju govori se o ne-metričkom ili ordinalnom multidimenzionalnom skaliranju.

4.1 Klasično multidimenzionalno skaliranje

Prvo, kao i sve MDS tehnike, klasično skaliranje pokušava da predstavi matricu sličnosti pomoću prostog geometrijskog modela. Videli smo kako se od matrice podataka dobija matrica udaljenosti. Ovde postoji obrnut problem, treba otkriti matricu podataka iz početne matrice udaljenosti. Takav model se sastoji od skupa tačaka $x_1, x_2, \dots, x_n$ u m dimenzija, gde svaka tačka predstavlja jednu jedinicu merenja i meru udaljenosti između parova tačaka. Cilj MDS-a je (videti u [2]) da odredi i dimenzionalnost, m, i n m-dimenzionalnih koordinata $x_1, x_2, \dots, x_n$, tako da nam model daje „dobar fit“ posmatranih sličnosti. Koliko je neki model dobro fitovan, tj. koliko dobro se sličnosti predstavljaju stvarne udaljenosti, se saznaje pomoću nekih numerčkih indeksa o kojima će biti reči nešto kasnije. Postavlja se pitanje kako odrediti m, i vrednosti koordinata $x_1, x_2, \dots, x_n$ iz posmatrane matrice sličnosti. Primetimo da ne postoji jedinstven skup vrednosti koordinata koje predstavljaju udaljenosti, jer pomeranjem cele konfiguracije, ili rotacijom, udaljenosti će ostati nepromenjene. Drugim rečima, ne može se jedinstveno odrediti ni lokacija ni orijentacija konfiguracija, tj. mogu se dobiti dva ista rešenja koja na grafiku izgledaju potpuno drugačije, prostom zamenom x i y osa.

4.1.1 Tehnički detalji MDS-a

Za početak uzmimo da je matrica sličnosti sa kojom će se raditi, matrica Euklidskih udaljenosti, **D**, dobijena iz $n \times q$ matrice podataka, **X**. U prethodnom delu rada, pokazano je kako da dobiti Euklidske udaljenosti iz **X**. Klasično multidimenzionalno skaliranje se bavi obrnutim problemom, ako su date udaljenosti, kako da naći **X**? Pretpostavimo da je **X** poznato i pogledajmo $n \times n$ unutrašnji proizvod matrica, **B**

$$B = XX^T. \tag{4.1}$$

Elementi matrice **B** su

$$b_{ij} = \sum_{k=1}^{q} x_{ik}x_{jk.} \tag{4.2}$$

Lako je uočiti da kvadrirane Euklidske distance između vrsta matrice **X** mogu biti napisane preko elemenata matrice **B** kao

$$d_{ij}^2 = b_{ii} + b_{jj} - 2b_{ij}. \tag{4.3}$$

Ako se elementi matrice **B** mogu zapisati preko elemenata matrice **D**, onda se tražene vrednosti koordinata mogu dobiti faktorisanjem matrice **B** kao u (4.1). Ukoliko nije uvedeno dodatno ograničenje, ne postoji jedinstveno rešenje. To ograničenje obično se dobija kada centar tačaka $\bar{x}$, postavimo na koordinatni početak, tako da je $\sum_{i=1}^{n} x_{ik} = 0$ za sve $k = 1,2, \dots, m$. Ova ograničenja i formula (4.2) govore da suma činilaca u svakoj vrsti matrice **B** mora biti nula. Ukoliko sumiramo (4.2) po *i* i po *j* i na kraju po oba zajedno dolazi se do sledeće jednačine

$$\sum_{i=1}^{n} d_{ij}^2 = T + nb_{jj},$$

$$\sum_{j=1}^{n} d_{ij}^2 = T + nb_{ii},$$

$$\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2 = 2nT,$$

gde je $T = \sum_{i=1}^{n} b_{ii}$ trag matrice **B**. Elementi matrice **B** sada mogu biti predstavljeni pomoću kvadrata Euklidskih distanci kao

$$b_{ij} = -\frac{1}{2}\left(d_{ij}^2 - d_{i.}^2 - d_{.j}^2 + d_{..}^2\right)$$

gde je

$$d_{i.}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2,$$

$$d_{.j}^2 = \frac{1}{n}\sum_{j=1}^{n} d_{ij}^2,$$

$$d_{..}^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}^2,$$

Kako su sada izvedeni elementi matrice **B** i prikazani preko Euklidskih distanci, ostaje samo da se faktorišu kako bi se dobile vrednosti koordinata. Zbog SVD dekompozicije, **B** se može zapisati kao

$$B = V\Lambda V^T,$$

gde je $\boldsymbol{\Lambda} = diag(\lambda_1, \dots, \lambda_n)$ diagonalna matrica karakterističnih korena matrice **B**, a $\mathbf{V}= (V_1, \dots, \mathrm{V}_n)$ odgovarajuća matrica karakterističnih vektora, normalizovanih tako da je suma kvadrata njihovih elemenata jednaka jedinici, tj. $V_i V_i^T = 1$. Za karakteristične korene uzimamo da važi $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Kada je **D** dobijeno iz $n \times q$ matrice punog ranga, onda je matrica **B** ranga q pa je poslednjih $n - q$ karakterističnih korena jednako nuli. Sada **B** može biti zapisano kao

$$B = V_1 \Lambda_1 V_1^T,$$

gde je V_1 sadrži prvih q karakterističnih vektora i Λ_1 q ne-nula karakterističnih korena. Tražene vrednosti koordinata su

$$X = V_1 \Lambda_1^{\frac{1}{2}},$$

gde je $\boldsymbol{\Lambda_1}^{\frac{1}{2}} = diag(\lambda_1^{\frac{1}{2}}, \dots, \lambda_q^{\frac{1}{2}})$

Korišćenje svih q dimenzija dovodi do kompletnog rekonstruisanja originalnih Euklidskih udaljenosti. Prikaz u m-dimenzija koji najbolje fituje podatke, dat je sa m karakterističnih vektora **B** koji odgovaraju m najvećim karakterističnim korenima. Koliko je dobar m-dimenzionalan prikaz određujemo pomoću pomoćne funkcije

$$P_m = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i}.$$

Vrednosti P_m reda 0.8 i više, nedvosmisleno govore da su podaci dobro fitovani.

Kada posmatrana matrica nije matrica Euklidskih udaljenosti, matrica **B** nije pozitivno definitna. Tada će neki od karakterističnih korena biti negativni, a neke od koordinata će biti kompleksni brojevi. Adekvatnost rešenja može biti ocenjena korišćenjem jednog od sledeća dva kriterijuma

$$P_m^{(1)} = \frac{\sum_{i=1}^{m} |\lambda_i|}{\sum_{i=1}^{n} |\lambda_i|},$$

$$P_m^{(2)} = \frac{\sum_{i=1}^{m} {\lambda_i}^2}{\sum_{i=1}^{n} {\lambda_i}^2}.$$

Ponovo tražimo vrednosti iznad 0.8 da bi imali "dobar" fit. Ukoliko matrica **B** ima značajan broj velikih negativnih karakterističnih korena, trebalo bi koristiti neku drugu metodu, na primer ne-metričko (ordinarno) multidimenzionalno skaliranje.

4.2 Ordinarno (ne-metričko) skaliranje

U nekim psihologijskim radovima i u istraživanjima tržišta, matrice sličnosti se dobijaju iz postavljanja pitanja ispitanicima i bazirano na njihovim procenama o sličnosti i razlikama objekata, kreiraju se matrice. Prilikom skupljanja takvih podataka, može se steći utisak da su ispitanici u mogućnosti da daju samo „ordinalne" ocene; na primer, kada se porede razne nijanse neke boje ili ako bi se poredilo više boja, sa sigurnošću se mogu dobiti odgovori da je neka boja svetlija od druge, ali ukoliko bi pitanje bilo, koliko je tačno svetlija jedna boja od druge, odgovorima bi sigurno nedostajalo pouzdanosti. Takva razmatranja su 1960-ih, dovela do metode multidimenzionalnog skaliranja koja koristi samo rank (dva objekta su poređana u istom redosledu) sličnosti da bi napravili prostorni prikaz. Drugim rečima, metod je invarijantan na monotone transformacije posmatrane matrice sličnosti; na primer, dobijene koordinate će ostati iste ako su numeričke vrednosti posmatrane matrice promenjene, ali je rank ostao isti. Ovaj metod je prvi predstavio Kruškal*.

*Joseph Kruskal (1928-2010) je bio američki matematičar, statističar i psihometričar. Poznat kao jedan od prvih naučnika koji su radili na MDS-u

Veoma često, stvarne distance x_{ij} nisu mnogo značajne, nego njihova vrednost u odnosu na to koliko su daleko od drugih objekata. Ovo je posebno tačno kada su x_{ij} rezultat eksperimenta gde su ispitanici upitani da daju njihovo subjektivno mišljenje na udaljenosti između objekata. Tada, x_{ij} mogu biti interpretirani samo u ordinarnom obliku. Kod ordinarnog MDS-a, cilj je da nađemo konfiguraciju tako da su d_{ij} istog ranka (poređani u istom redosledu) kao i originalni x_{ij}. Na primer, ako je udaljenost između objekata 1 i 5, petog ranka među x_{ij}-ovima, onda bi trebala da bude petog ranka i u MDS konfiguraciji. Kod ordinalnog MDS-a mi konstruišemo fitovane distance, često nazivane *dispariteti,* $\widehat{d_{ij}}$. Razlike dobijamo iz d_{ij}, tako da su istog ranka kao i x_{ij}(za razlike) ili obrnutog ranka (za sličnosti). Recimo da su $\widehat{d_{ij}}$ "uglađene" (eng.smoothed) verzije d_{ij}. Posmatrane razlike, x_{ij}, su rangirane od najmanje do najveće

$$x_{i_1 j_1} < x_{i_2 j_2} < \cdots < x_{i_N j_N}$$

gde je $N = n(n-1)/2$ tako da

$$\hat{d}_{i_1 j_1} < \hat{d}_{i_2 j_2} < \cdots < \hat{d}_{i_N j_N}$$

Ovo postižemo metodom nazvanom, metoda najmanjih kvadrata monotone regresije (reč monotono se odnosi na to, da je regresiona kriva ili neopadajuća ili nerastuća). Prilikom crtanja d_{ij} i x_{ij} (videti u [2]), cilj je da dobiti monotonu krivu (takvu da su linije koje spajaju susedne tačke ravne/rastuće ako su x_{ij} razlike ili ravne/opadajuće ako su x_{ij} sličnosti). Ako su d_{ij} i x_{ij} istog ranka, onda će graf pokazivati takvu monotonu krivu gde nije potrebno uglađivanje. Uglavnom to nije slučaj i potrebno je neko prilagođavanje.

Cilj monotone regresije je da fituje monotonu krivu ka tačkama (d_{ij}, x_{ij}), istovremeno tražeći da suma kvadrata vertikalnih devijacija bude što manja. (Videti u [3]) Tačka na monotonoj krivi, $\widehat{d_{ij}}$, je fitovana ili predviđena vrednost za d_{ij} iz monotone regresije. Kada hoćemo da ocenimo koliko nam je fit dobar, gledamo koliko su blizu udaljenosti d_{ij} od dispariteta $\widehat{d_{ij}}$.

Definicija 1. Stres funkcija u ordinalnom MDS je

$$Stress - 1 = \sqrt{\frac{\sum_{i<j}\left(d_{ij} - \widehat{d_{ij}}\right)^2}{\sum_{i<j} d_{ij}^2}}.$$

Ovo je takođe poznato kao i Kruskalov stres, tipa 1, koji ćemo nazivati jednostavno stres. Optimum je određen minimiziranjem ove mere stres-a. Minimizacija stres funkcije je kompleksan problem, pa MDS programi koriste iterativne numerčke algoritme kako bi našli matricu **X** za koju je Stress-1 minimalan.

4.3 Procena "fit-a" i odabir broja dimenzija

Postoje mnogi načini procene koliko je MDS rešenje dobro fitovano. Jedna metoda uključuje poređenje dobijenih vrednosti Stres-a sa reperom iz Tabele 4.2. Ove vrednosti je prvi napravio Kruskal (1964) i one su bazirane pre na empirijskom iskustvu nego na teoretskim kriterijumima. Ovo uvek treba koristiti sa određenom dozom fleksibilnosti i paziti koliko ustvari to rešenje dobro interpretira rešenje koje tražimo.

Stress-1(Kruskal-ov tip)	**Procena fita**
0.20	Slab
0.05	Dobar
0.00	Odličan

Tabela 4.2: Ocene koliko je fit dobar

Drugi metod koji se može koristiti da za odabir broja dimenzija, jeste da korišćenje dvodimenzionalnog grafika, gde će na jednoj osi biti imati vrednosti Stres-a, a na drugoj broj dimenzija. Kako broj dimenzija raste, tako se vrednost Stres-a smanjuje, ali postoji mala začkoljica, a to je, koliko je isplativo povećavati dimenzije, a da se ne izgubi na interpretabilnosti rešenja. U dobijenim graficima, biće tražen „lakat“ (eng.elbow), koji predstavlja tačku gde povećanje broja dimenzija ima malo efekta na smanjenje Stres-a. Ova metoda se u praksi pokazala kao veoma dobra.

Slika 4.2: Primer “lakta” koji pokazuje da je dvodimenzionalan grafik dovoljan za prikaz podataka

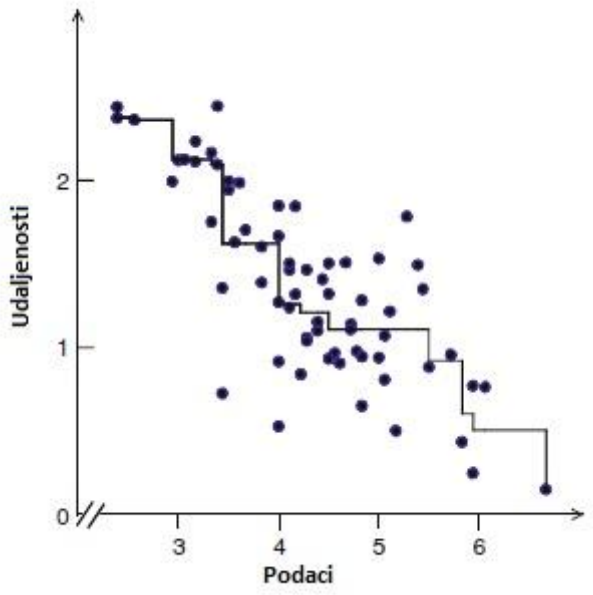
Postoje i drugi korisni grafici. U slučaju ordinarnog skaliranja, oni uključuju sve parove d_{ij}, $\widehat{d_{ij}}$, x_{ij} koju mogu biti proučavani kako bi izmerili fit MDS rešenja.

i. Nacrtati d_{ij} i $\widehat{d_{ij}}$. Ukoliko rešenje MDS dobro fituje, grafik će pokazati dobru linearnu vezu, sa nagibom od 45 stepeni, i veoma malo rasipanje oko linije. Ukoliko je bilo potrebno da “ugladimo” d_{ij} kako bi dobili $\widehat{d_{ij}}$, onda bi oni trebali biti skoro istog ranka i približne vrednosti
ii. Nacrtati d_{ij} i x_{ij}. Ukoliko rešenje dobro fituje, d_{ij} i x_{ij} bi trebalo da budu približno istog (ili obrnutog) ranka, i grafik bi trebalo da pokazuje monotonu krivu, ili rastuću (za razlike) ili opadajuću (za sličnosti).

iii. Nacrtati $\widehat{d_{ij}}$ i x_{ij}. $\widehat{d_{ij}}$ su "uglađene" verzije d_{ij} konstruisane da imaju isti rank kao i x_{ij} (za razlike) ili obrnuti rank (za sličnosti). Ukooliko je potreban veliki broj uglađivanja za dobijanje monotone krive (u slučaju da slabog fitovanja), na grafiku će biti prikazani veliki broj horizontalnih koraka gde je uglađivanje bilo potrebno. Kada je fit dobar, imaćemo mali broj horizontalnih koraka.

Kao što je navedeno, MDS traži vrednosti koordinata *n* tačaka u q dimenzija ($n \times q$ matrica koordinata **X**) čije distance predstavljaju sličnosti, približno koliko god je to moguće. Kao što je dosad viđeno, predstavljeno je nekoliko metoda kako izabrati dimenzionalnost za rešenje MDS-a. Unidimenzionalno skaliranje je slučaj kada je $q = 1$, i treba biti obazriv, zato što je moguće da MDS algoritimi završe u lokalnom minimumu, koji je daleko od globalnog. Jedan od pristupa kako odrediti dimenzionalnost, jeste naći rešenja za više dimenzija, recimo od dve do šest, i kao što je već rečeno, nacrtati vrednosti Stres-a i vrednosti tj. broj dimenzija na drugoj osi. Kakogod, jedan od najvažnijih kriterijuma za odabir broja dimenzija je jednostavno baziran na interpretabilnosti grafika. Zbog toga, velika većina MDS rešenja su predstavljena dvodimenzionalnim i ponekad trodimenzionalnim graficima.

Alternativno, može se kreirati funkcija gubitka koja će pokazati koliko je podataka izgubljeno u MDS prikazu. Geometrijski, vertikalna distanca između regresione linije i tačke daje grešku odgovarajuće udaljenosti u MDS prikazu. Gubitak informacija u MDS prikazu može se meriti kao suma reziduala.



Slika 4.3: Primer Shepard-ovog diagrama. Vertikalne udaljenosti između tačaka i reg.linije predstavljaju grešku odgovarajuće udaljenosti i MDS rešenja

Definicija 1. Stres funkcija (eng. Stress) je minimizirana funkcija gubitka po Kruskal-u.

$$Stress = \sqrt{\sum_{i<j} \left(d_{ij} - x_{ij}\right)^2} \tag{1}$$

Iako bi najjednostavnija forma funkcije gubitka bila formula iz (1), hoćemo da Stres funkcija bude standardizovana i nevezana za jedinice mere, pa stoga koristimo normalizovanu stres funkciju.

Definicija 2. Normalizovan stres je

$$Stress = \sqrt{\frac{\sum_{i<j} \left(d_{ij} - x_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}}. \tag{2}$$

Treba naglasiti da postoje i drugi načini računanja Stres-a, a jedan od njih je da d_{ij} iz jednačine (2) u imeniocu zameni sa x_{ij}. Vrednosti Stres funkcije koje su blizu nuli, govore da MDS rešenje dobro fituje originalne x_{ij}. Perfektno MDS rešenje ima $Stress = 0$. Ovo se dešava (videti u [7]) kada distance MDS rešenja predstavljaju podatke tačno. Još jedan od pokazatelja da rešenje nije dobro, jeste ako je Stres za neke proizvoljne vrednosti manji od Stres-a za vrednosti koje su izabrane. Evaluacija vrednosti Stres-a je kompleksan problem, i prilikom izučavanja treba obratiti pažnju na više parametara:

- Broja tačaka (n). Što veće n, veći je očekivani Stres.
- Dimenzionalnost MDS rešenja (m). Što veće m, manji je očekivani Stres (veći broj dimenzija daje veću slobodu za optimalno pozicioniranje tačaka)
- Greška u podacima. Što veći šum u podacima, veći je očekivani Stres.
- Udeo podataka koji fale. Što više podataka fali, lakše je naći MDS rešenje sa malim Stres-om.

4.4. Razne greške u MDS-u

Postoje mnoge greške koje korisnici prave prilikom korišćenja MDS-a, od konceptualne nejasnoće, preko korišćenja MDS-a za pogrešan tip podataka, ili korišćenje MDS programa koji imaju loše podešene parametre, do pogrešnog tumačenja MDS rešenja. U ovom delu rada govoriće se o nekim najčešćim greškama koje se javljaju.

Korišćenje pojma MDS previše uopšteno

Pojam multidimenzionalno skaliranje je obično rezervisano za modele diskutovane u ovom radu, i još par koji nisu spomenuti (zbog opširnosti rada). Kakogod, mnogi nazivaju bilo koju tehniku koja daje vizualni prikaz u malom broju dimenzija kao proceduru „multidimenzionalno skaliranje“, što svakako ne može biti tačno. Naime, svaka od ovih tehnika se razlikuje po tome šta pokazuje na grafiku, i kako bi grafici trebali biti interpretirani. Zbog ovakvog korišćenja pojma MDS-a, jasne razlike između ovih tehnika mogu nestati, što dovodi do konfuzije i pogrešne interpretacije. Zbog toga, najbolje bi bilo korisiti pojam multidimenzionalno skaliranje za modele koji prikazuju sličnosti između objekata koje posmatramo, pomoću udaljenosti između tačaka u malom broju dimenzija i ne koristiti pojam ni za šta drugo.

Korišćenje pojma udaljenosti previše široko

Na primer, mnoge knjige i publikacije, razlike (eng.dissimilarity data) nazivaju „udaljenostima“. Ovo može dovesti do konfuzije o svrsi i samom procesu MDS-a. Šta MDS uvek radi, jeste da predstavlja sličnosti, precizno koliko god je moguće, kao udaljenosti. Zbog ovoga, udaljenosti se javljaju u samom modelu MDS, a mnogo ređe u samim podacima. Štaviše, te distance su specijalne distance, kao na primer Minkovski, ili neki specijalan slučaj Minkovski udaljenosti(Menhetn ili Euklidska udaljenost).

Sličnosti su distance, ako i samo ako, zadovoljavaju aksiome predstavljene na početku ovog rada. U mnogo slučajeva se ne mogu svi aksiomi testirati. Uglavnom je to zbog nedostataka svih podakata potrebnih za test. Na primer, retko će biti slučaj da istraživač prikupi podatke o sličnosti između *i* i *j,* i takođe između *j* i *i,* pa stoga simetričnost ne može biti proverena. U mnogim aplikacijama date sličnosti mogu biti konvertovane u vrednosti koje ne narušavaju mogućnost testiranja aksioma udaljenosti. Zaista, jedna od hipoteza koju testira MDS je da li podaci mogu biti dopustljivo transformisani u (neki od) Minkovski udaljenosti u *m*-dimenzionalnom prostoru.

Korišćenje premalo iteracija

Skoro svi softverski programi za rešavanje MDS problema imaju podešene parametre. Obično, iteracije optimizacije algoritama se završavaju prerano, i to se dešava zato što je proces konvergirao u lokalni minimum. Ovakvo prerano zaustavljanje programa je prouzrokovano postavljanje kriterijuma previše defanzivno. Mnogi programi imaju podešeno za maksimalan broj iteracija 100 ili čak i manje, i ovo vuče korene kada je izvođenje ovakvih operacija bilo skupo i sporo. Iteracije se takođe zaustavljaju ako se Stres ne smanjuje barem za 0.005 po iteraciji. Kakogod, može biti pokazano da veoma mala smanjenja Stresa ne znače uvek da će sve tačke ostati fiksirane u daljim iteracijama. Zbog ovoga se preporučuje da se ova podešavanja uvek promene kako bi programi radili duže. Umesto dosadašnjih 50, mogu se zahtevati 500 ili čak 1000 iteracija. Kriterijum za konvergenciju može biti podešen na 0.000001. Jedina "mana" ovakvog pristupa je što će program raditi koju sekundu duže nego inače.

Korišćenje pogrešne startne konfiguracije

Svi MDS programi automatski generišu „racionalnu" startnu konfiguraciju ako korisnik sam ne unese eksterne podatke, za neku drugu startnu poziciju. Jedna od najvećih zabluda je što korisnici često smatraju da će konfiguracija koja je automatski podešena u programu uvek voditi ka optimalnom MDS rešenju. Treba napomenuti da nijedna startna konfiguracija, racionalna ili postavljena od strane korisnika, ne garantuje najbolje moguće krajnje rešenje, i zbog toga bi uvek trebalo proveriti i neke alternativne konfiguracije pre nego što se odlučimo za neko MDS rešenje. Na primer, može se koristiti nasumično odabrana startna konfiguracija. Ukoliko postavimo taj broj na recimo 1000, program će tražiti 1000 različitih startnih konfiguracija, i na kraju nam izbaciti onu sa najmanjim Stres-om.

Ne raditi ništa da se izbegne lokalni minimum

MDS rešenje se gotovo uvek nalazi kroz serije malih pomeranja tačaka, tako da se vrednost Stres-a smanjuje. Algoritmi koji se danas koriste za računanje takvih iteracija, garantuju nalaženje lokalnog minimuma. Problem je u tome što početne konfiguracije moraju da počnu sa nekom početnom konfiguracijom, i ako je ona pogrešna, mogu završiti u pogrešnom lokalnom minimumu (ako postoji). MDS uvek nastoji da nađe lokalni minimum sa najmanjim mogućim stresom, tj. *globalni minimum*. Korisnici MDS algoritma mogu doprineti pronalsku ovog globalnog minimuma obraćajući pažnju na sledeće probleme:

- Dobra startna konfiguracija je najbolji način da se izbegne lokalni minimum.
- Ukoliko je moguće, odlično bi bilo koristiti nekoliko različitih početnih konfiguracija. Kako su moderni MDS programi veoma brzi, ponavljanje skaliranja sa velikim brojem

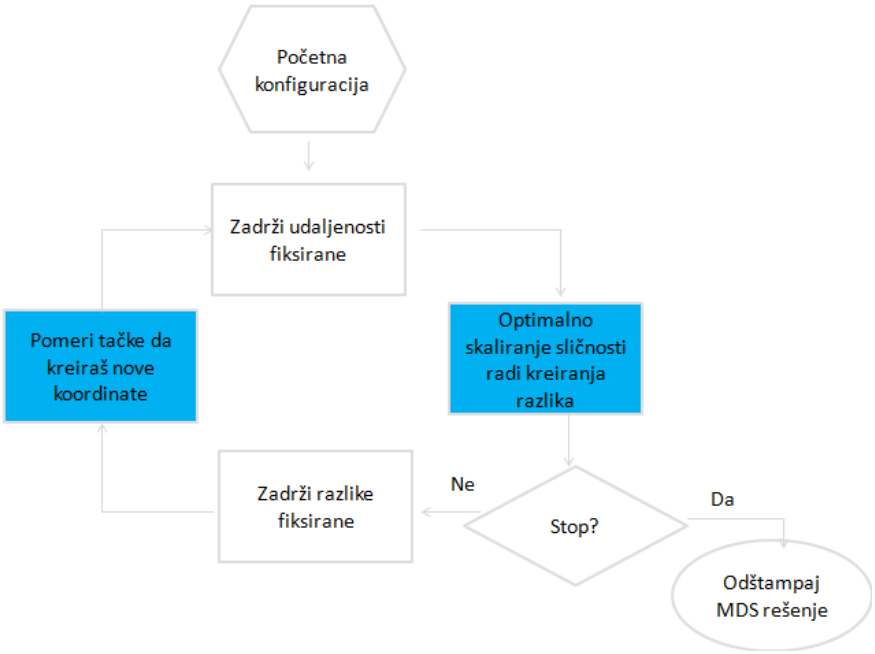
različitih početnih konfiguracija ne iziskuje mnogo vremena i napora(recimo 1000 ili više)

- Menhetn distance povećavaju rizik da algoritam završi u lokalnom minimumu. Programi koji se najčešće koriste su posebno osetljivi po ovom pitanju. Zato postoje posebni programi koji su optimizovani za ovakve udaljenosti, ali uglavnom zahtevaju znanje eksperta za korišćenje istih.
- Lokalni minimumi su karakteristični za jednodimenzionalni MDS. Standardni programi skoro nikad ne nađu globalni minimum. Ukoliko je baš potrebno tražiti jednodimenzionalni MDS, najbolje bi bilo korisiti program namenjen isključivo tome, što opet može iziskivati napor.

Evaluacija Stres-a mehanički

Veoma česta greška koja se pravi prilikom analize MDS rešenja, je da se rešenje odbacuje previše brzo ukoliko je vrednost Stres-a „prevelika“. Kakogod, vrednost Stresa-a je samo tehnički indeks, koji uglavnom služi kao kriterijum za algoritme MDS programa. Stres, šta više je suštinski „slep“ (Guttman 1977), i ne govori ništa o kompatabilnosti teorije sa MDS konfiguracijom, ili bilo šta o interpretabilnosti. Stres je *sumativni* index za sve sličnosti. On ne govori koliko dobro su određene vrednosti predstavljene u datom MDS prostoru. Pretpostavimo da imamo konfiguraciju gde nisu sve sličnosti predstavljene jednako dobro, i pretpostavimo da imamo jedan outlajer(eng.outlier), koji kvari Stres, zato što je veoma daleko od regresione linije. Dalje, postavlja se pitanje koliko dobro je pojedinačan objekat predstavljen u MDS konfiguraciji. Ovo se meri funkcijom Stres po tački (eng. Stress per point-SPP), koji se dobija kao srednja vrednost kvadrata grešaka za svaku tačku. Jednostavan način da se nosimo sa tačkama koje loše fituju, je da ih jednostavno izbacimo iz analize. Ovo je popularan pristup, baziran na tome da te tačke imaju specijalan odnos prema ostalim tačkama, i da zahtevaju dodatna razmatranja. Neko od rešenja je povećavanje dimenzionalnosti, tako da se date tačke mogu pomeriti u dodatan prostor i formirati nove distance. U svakom slučaju, odbijanje ili prihvatanje MDS prikaza na osnovu Stresa može biti previše jednostavno. Recimo ukoliko bi povećali broj dimenzija sa dve na tri dimenzije, Stres bi se smanjio. Ako se nastavi u istom smeru, Stres bi se i dalje smanjivao. Za podatke koji imaju komponente sa puno šuma, niskodimenzionalno MDS rešenje može imati veliku vrednost Stres, ali opet biti bolje nego visokodimenzionalno rešenje sa manjom vrednosti za Stres. U tom slučaju, nisko-dimenzionlano rešenje može biti efikasno prilikom otkrivanja prave strukture podataka.

4.5 Iterativni MDS algoritam



Slika 4.4: Princip iterativnog MDS algoritma

Iterativne metodu su mnogo fleksibilnije nego klasični MDS. One pronalaze Stres-optimalnu MDS konfiguraciju, i radeći to, reskaliraju podatke optimalno sa nekim datim ograničenjima. S druge strane, oni ne garantuju da će uvek naći globalni minimum, zato što ta mala poboljšanja koja program pravi, mogu završiti u lokalnom minimumu. Iterativni MDS algoritmi postupaju u dve faze (Slika 4.4). U svakoj fazi jedan set parametara (udaljenosti ili razlike, respektivno) se uzimaju kao fiksne, dok se druge menjaju tako da se vrednost Stres-a smanjuje:

1. Razlike su fiksirane; tačke u MDS prostoru su pomerene($\boldsymbol{X}_t$ je promenjen u $\boldsymbol{X}_{t+1}$) tako da $\boldsymbol{X}_{t+1}$ minimizira Stres funkciju.
2. MDS konfiguracija, $\boldsymbol{X}$ je fiksirana; razlike su reskalirane tako da je Stres funkcija minimizarana (optimalno skaliranje)

Posle t faza, ovaj ping-pong proces više ne smanjuje Stres vrednost više od neke fiksne vrednosti (npr., 0.0005), algoritam se zaustavlja na $\boldsymbol{X}_t$, i to se uzima kao optimalno rešenje.

U *Fazi 1*, se javlja težak matematički problem, sa $n \cdot m$ nepoznatih parametara, vrednosti $\boldsymbol{X}$. Za rešavanje ovog problema razvijeni su mnogi algoritmi.

Faza 2 postavlja relativno lak problem. U zavisnosti od tipa MDS-a, treba rešiti problem sa nekom od regresija (linearnom, monotonom…). Ovi problemi su čisto matematički problemi. Korisnici MDS programa ne treba da brinu o tome. To je kao kada se vozi auto; (vidi u [7]) vozači treba da znaju kako da upravljaju autom, ne moraju da brinu šta se dešava ispod haube prilikom same vožnje. Tako i korisnici MDS programa, moraju da snadbeju program potrebnim podacima kako bi dobili optimalna rešenja. Kao što je već rečeno, važno je izabrati dobru startnu konfiguraciju . Svi MDS programi, nude nekoliko alternativa koje korisnici mogu probati, i videti da li svi vode ka istom rešenju. Ipak je uvek bolje aktivno uticati na početnu konfiguraciju, nego prepustiti to MDS programu. Često, dobra opcija je uzimanje startne konfiguracije bazirane na teoretiskim temeljima. U zavisnosti od MDS programa koji se koristi, razne „tehničke“ opcije su ponuđene korisniku, i u zavisnosti od toga šta korisnik odabere, imaće velik uticaj na krajnje MDS rešenje, zato što to uglavnom onemogućuje program da završi iteracije, iako bi Stres mogao biti još manji. Kao što je već pomenuto, maksimalan broj iteracija i numerčki kriterijum za konvergenciju su podešene od strane korisnika, ali uobičajena podešavanja su zbog istorijskih razloga podešene previše defanzivno kako bi se iteracije završile što ranije. Korisnik bi trebalo da podesi ove parametre, tako da bude onoliko iteracija koliko je potrebno da se Stres smanji. U ovom radu korišćen je softverski paket *Proxscal (SPSS).*

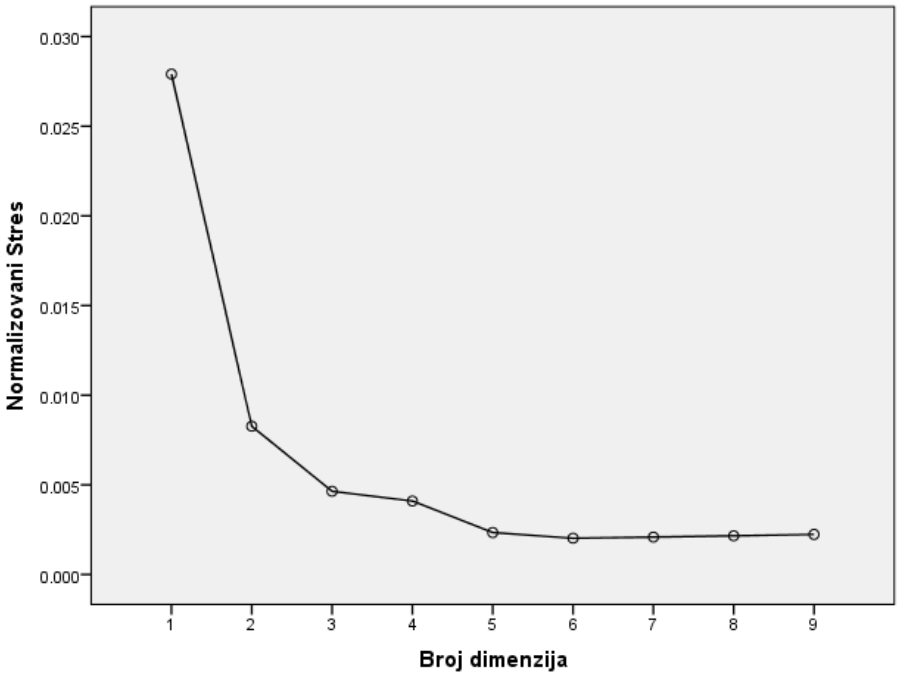
4.6 Primeri

Primer 1. U sledećem primeru, biće pokazano kako multidimenzionalno skaliranje funkcioniše, i biće prikazano malo više detalja nego u narednim primerima kako bi čitalac imao bolju sliku o tome šta se dešava u samom procesu. U Tabeli 4.3 date su udaljenosti nekih većih gradova u Srbiji.

	Novi_Sad	Beograd	Kragujevac	Nis	Krusevac	Subotica	Zrenjanin	Cacak	Novi_Pazar	Leskovac
Novi_Sad	.									
Beograd	94.000	.								
Kragujevac	223.000	140.000	.							
Nis	270.000	238.000	151.000	.						
Krusevac	284.000	195.000	107.000	74.000	.					
Subotica	95.000	189.000	335.000	433.000	389.000	.				
Zrenjanin	51.000	71.000	210.000	302.000	270.000	146.000	.			
Cacak	230.000	139.000	59.000	187.000	98.000	332.000	214.000	.		
Novi_Pazar	242.000	187.000	160.000	172.000	124.000	482.000	372.000	130.000	.	
Leskovac	368.000	276.000	188.000	43.000	128.000	470.000	340.000	224.000	196.000	.

Tabela 4.3: Udaljenost gradova u Srbiji (u km), izvor Here maps

Pre nego što počnemo sa primenom MDS, naći ćemo koliko dimenzija je potrebno kako bi podaci iz tabele bili predstavljeni što je bolje moguće.



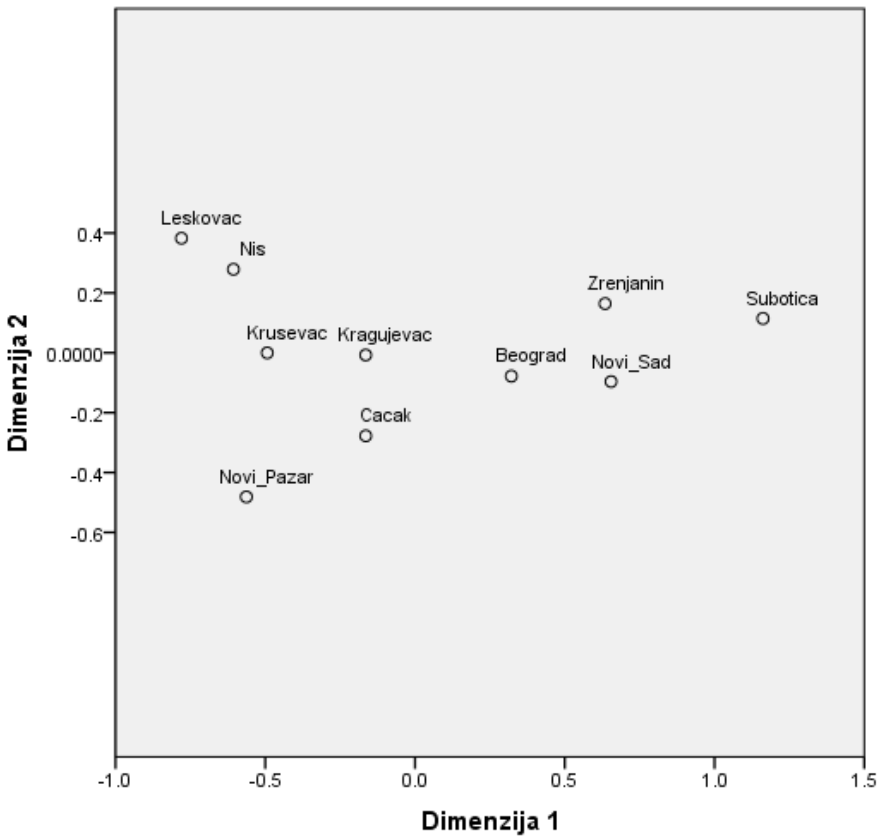
Slika 4.5: Slika „Scree plot-a" na kome se traži "lakat"

Na osnovu Slike 4.5 vidi se da su dve dimenzije sasvim dovoljne, kako bi udaljenosti između gradova bile prikazane što je preciznije. (Zbog lakoće čitanja i razumevanja, grafici bi trebali da budu prikazani u što manje dimenzija).

Iteracija	Normalizovani Stress	Poboljšanje
0	.21295[a]	
1	.01180	.20115
2	.01033	.00147
3	.00955	.00078
4	.00906	.00050
5	.00873	.00033
6	.00850	.00023
7	.00834	.00016
8	.00823	.00012
9	.00814	.00009[b]

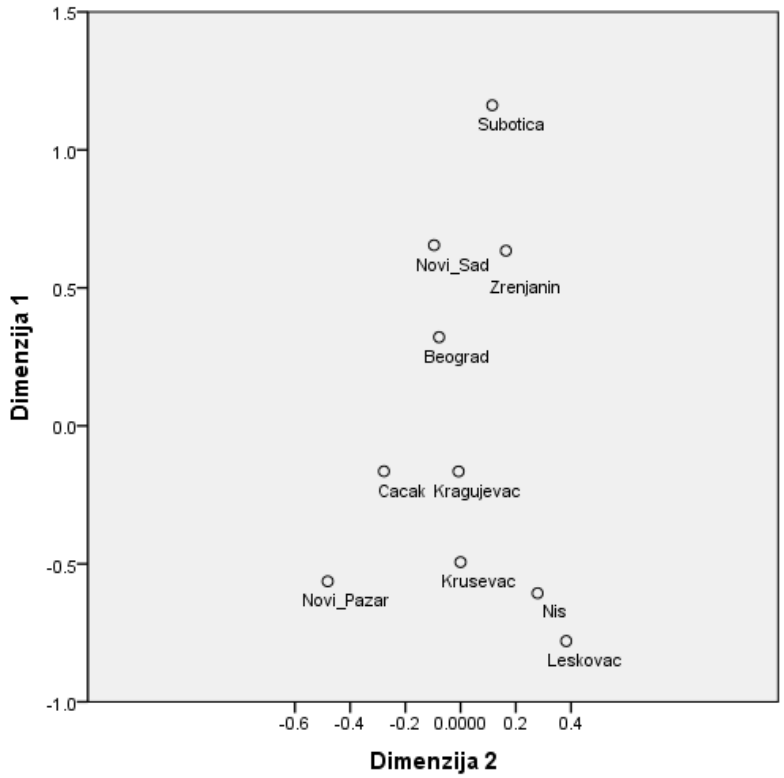
Tabela 4.4: Istorija iteracija

U gornjoj tabeli se vidi kolika je bila vrednost Stres funkcije na početku, i sva poboljšanja do kraja, tj. poslednje iteracije.



Slika 4.6: Dvodimenzionalan prikaz udaljenosti između gradova

U radu je navedeno da rotacija osa ne utiče na samo rešenje problema. Na Slici 4.6 se vidi inicijalno rešenje. Međutim, ukoliko se ose rotiraju, kao što je prikazano na Slici 4.7, dobijena sliku je prepoznatljiva kao mapa Srbije, što će i biti konačno rešenje.



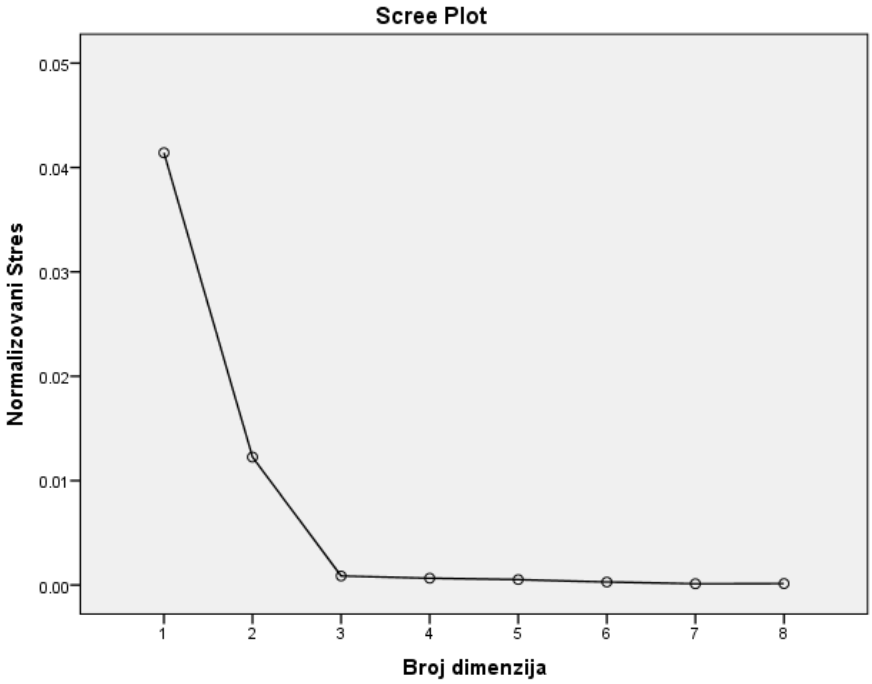
Slika 4.7: Rotiran dvodimenzionalan prikaz udaljenosti između gradova

Primer 2. Objektivno poređenje 9 sportova, gde je 1 predstavlja skoro isti, a 9 potpuno različiti.

	fudbal	tenis	kosarka	rukomet	vaterpolo	odbojka	atletika	triatlon	biciklizam
fudbal	.								
tenis	7.000	.							
kosarka	6.000	8.000	.						
rukomet	5.000	8.000	5.000	.					
vaterpolo	8.000	8.000	8.000	7.000	.				
odbojka	6.000	6.000	6.000	7.000	8.000	.			
atletika	5.000	6.000	5.000	6.000	9.000	6.000	.		
triatlon	8.000	7.000	8.000	8.000	6.000	8.000	4.000	.	
biciklizam	9.000	9.000	9.000	9.000	9.000	9.000	8.000	5.000	.

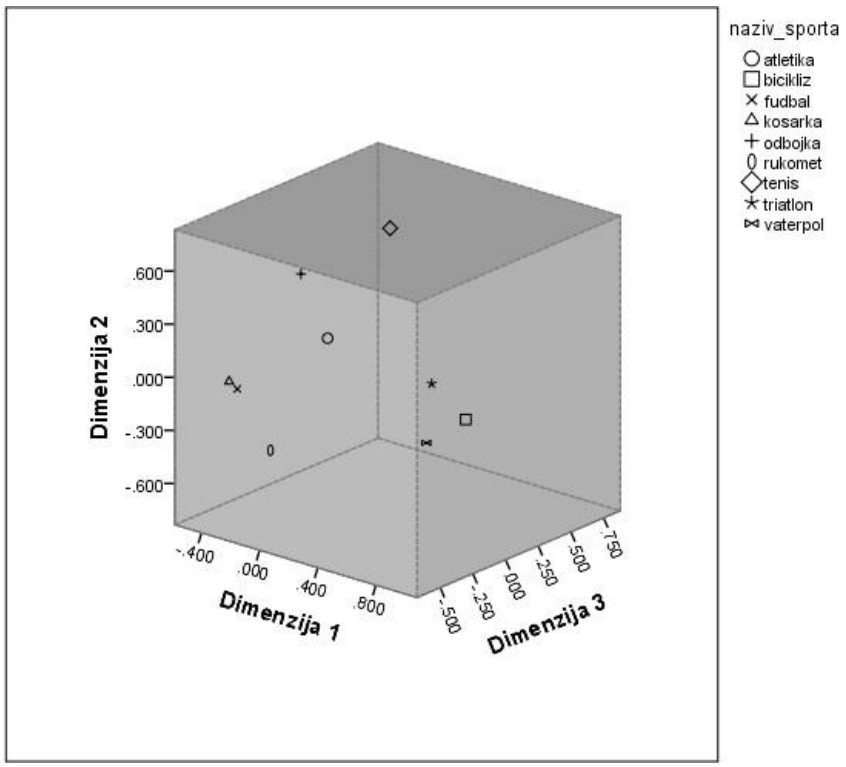
Tabela 4.5: Objektivno ocenjivanje 9 različitih sportova

Prvo ćemo potražiti koliko bi dimenzija bilo potrebno kako bi matricu bila predstavljena što preciznije je moguće.



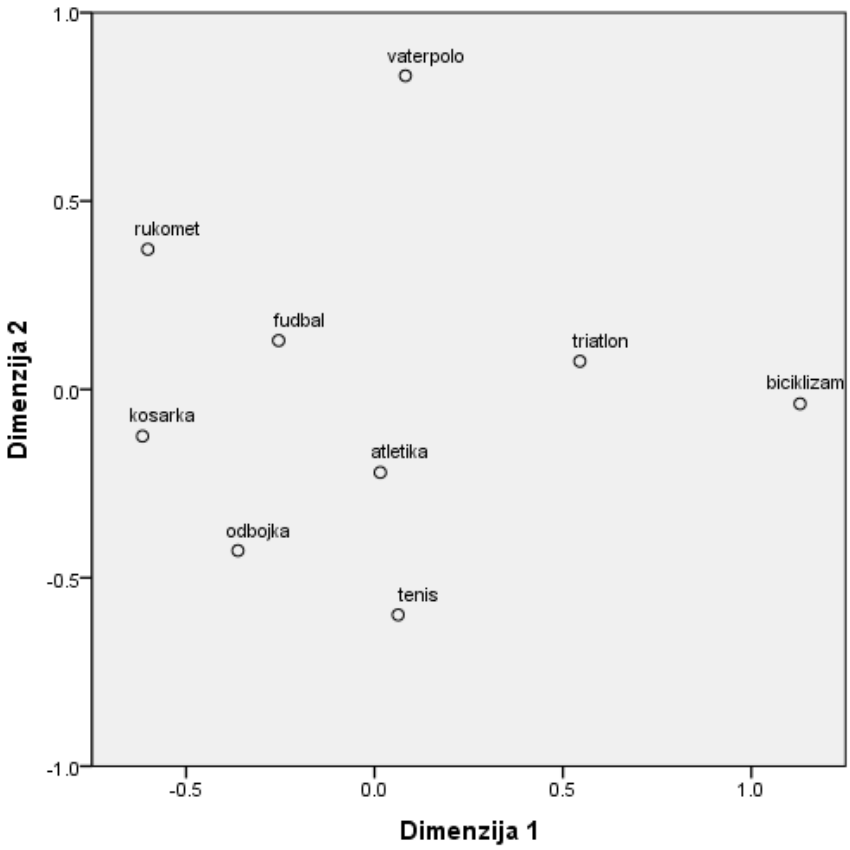
Slika 4.8: Na osnovu grafika traženo rešenje biće u tri dimenzije

Na osnovu Slike 4.8 trebali bi tražiti trodimenzionalno rešenje. Koje izgleda ovako



Slika 4.9: Trodimenzionalni prikaz sportova

S obzirom da je na dvodimenzionalnom grafiku lakše analizirati podatke, a vrednost Stres-a je 0.047 što prema Kruskal-ovom kriterijumu predstavlja veoma dobar fit, bez greške se može uzeti kao konačno rešenje.



Slika 4.10: Dvodimenzionalna konfiguracija

Primer 3.

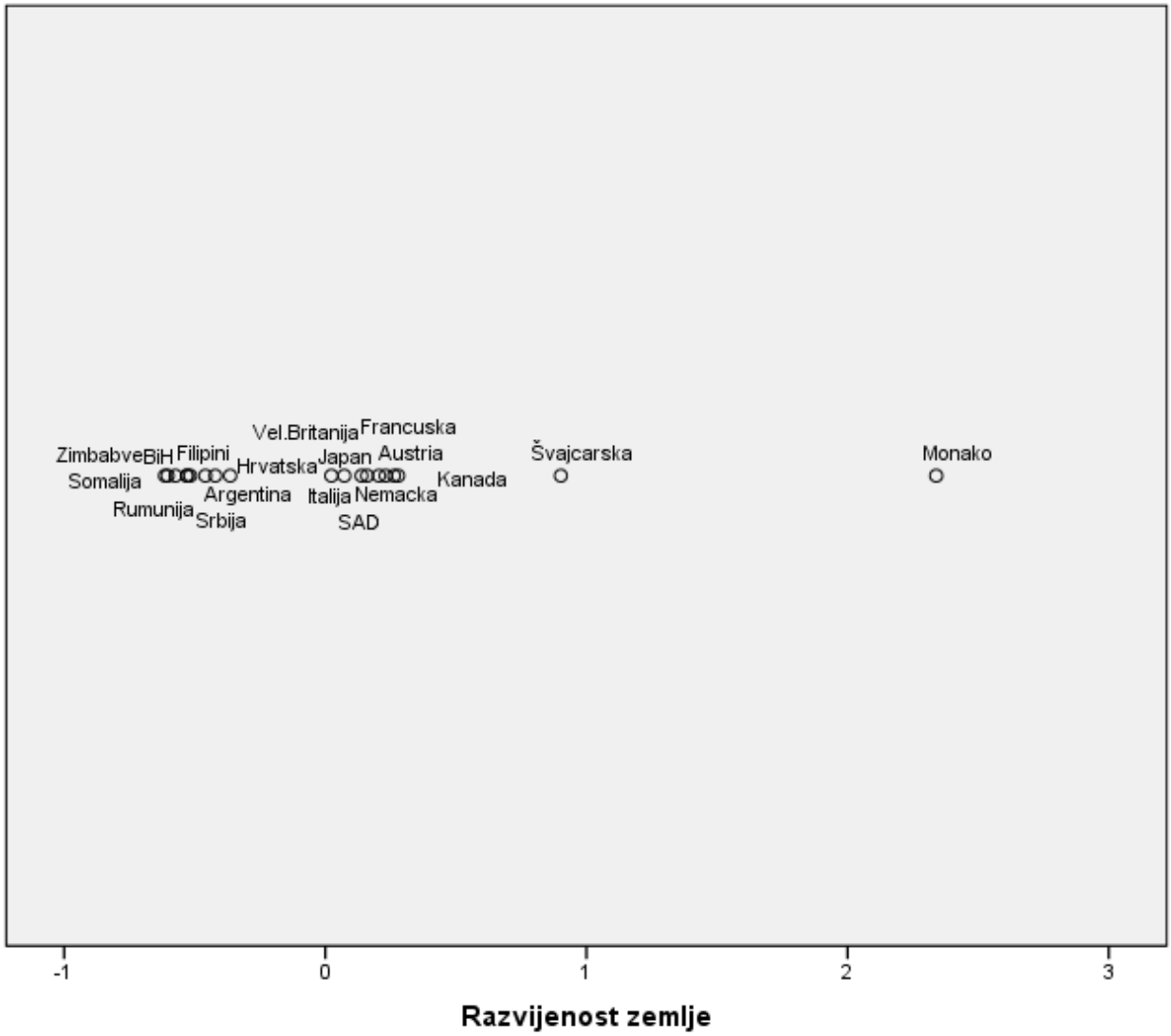
Država	Populacija	Urban	Rast	BDP	Život
Argentina	41 119	92.7	0.9	10 994	80
Austrija	8 429	67.9	0.2	49 686	84
Bosna i Hercegovina	3 744	48.8	-0.2	4 807	78
Kanada	34 675	80.8	0.9	50 565	83
Hrvatska	4 387	58.1	-0.2	14 217	80
Francuska	63 458	86.4	0.5	42 642	85
Nemačka	81 991	74.1	-0.2	43 865	83
Italija	60 964	68.5	0.2	36 124	85
Japan	126 435	91.9	-0.1	46 407	87
Monako	36.1	100	0	167 021	86
Filipini	96 471	49.1	1.7	2 370	73

Rumunija	21 388	52.9	-0.2	8 853	78
Srbija	9 847	56.7	-0.1	5 579	77
Somalija	9 797	38.2	2.6	112	53
Švajcarska	7 734	73.8	0.4	85 794	85
Makednoija	2 067	59.4	0.1	4 925	77
Velika Britanija	62 798	79.7	0.6	38 918	82
SAD	315 791	82.6	0.9	47 882	81
Zimbabve	13 014	39.1	2.2	695	53

Izvor: UN economic and demographic indicators

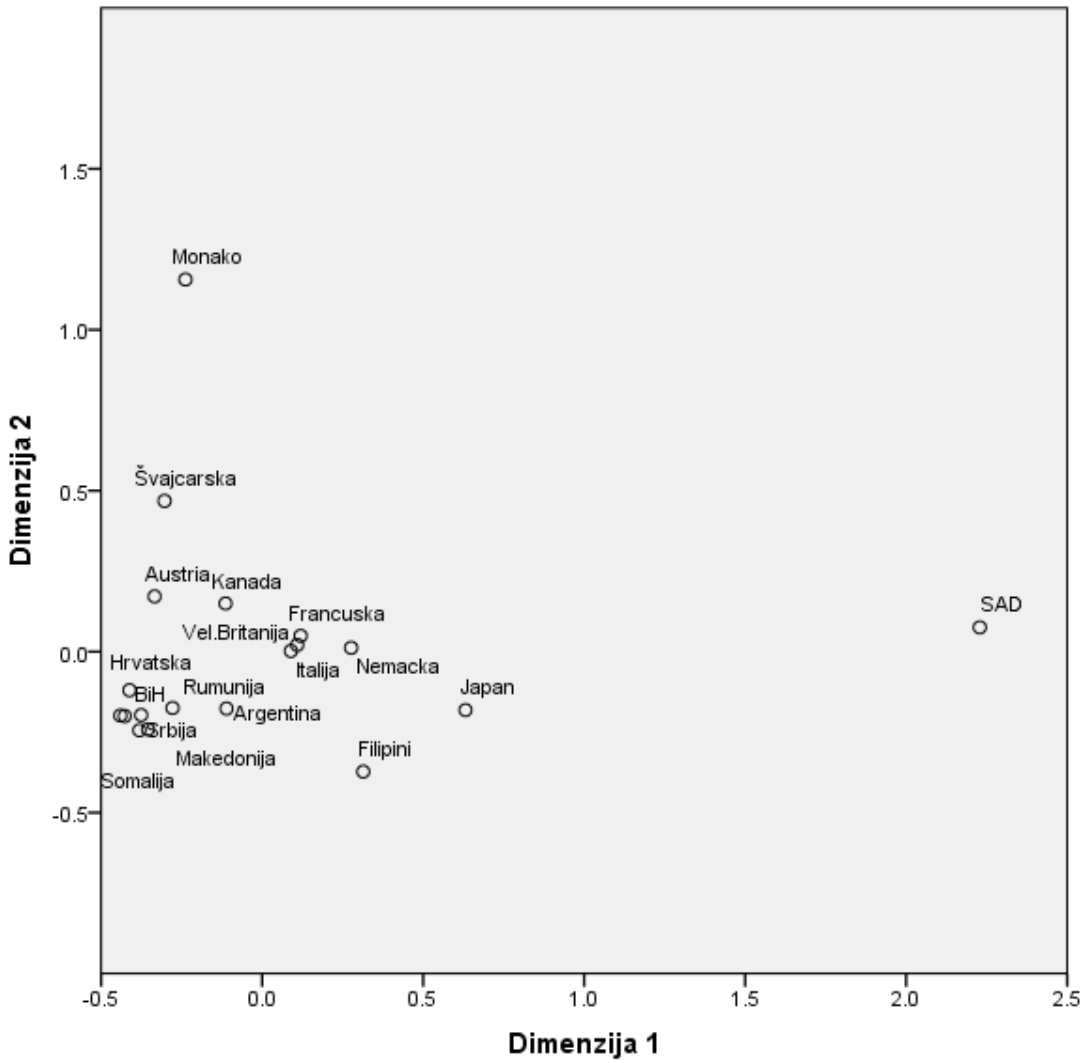
Tabela 4.6: Ekonomski i demografski indikatori za 19 zemalja

Tabela 4.6 pokazuje vrednosti pet ekonomskih i demografskih indikatora za primer od 19 zemalja. Indikatori su broj stanovnika u hiljadama (Populacija), procenat stanovnika koji živi u gradovima (Urban), rast populacije u periodu 2010-2015 (Rast), bruto domaci proizvod po glavi stanovnika u američkim dolarima (BDP), i očekivani životni vek (Život). Kako su podaci u formi pravougane matrice, prvo je potrebno prevesti ih u matricu udaljenosti. Jedan od ciljeva može biti da li je moguće odrediti da li države mogu biti postavljene na skali razvijenosti. Stoga, jednodimenzionalno rešenje je od značajnog interesa. Razvijene zemlje karakteriše nizak rast populacije, visok životni vek, visok BDP. U slučaju jedne dimenzije, razvijene zemlje bi bile na jednoj strani prave, a sve ostale zemlje na drugoj. U ovom slučaju za takvu analizu, problem bi nam bio podatak o veličini populacije, jer veličina ne utiče na razvijenost neke zemlje (Filipini imaju oko 100 miliona stanovnika, a slabo su razvijena zemlja), pa bi to pravilo problem u samoj analizi. Zbog toga, pri crtanju jednodimenzionalnog grafika, podatak o veličini populacije će biti izostavljen.



Slika 4.11: Razvijenost zemalja

Kruskal-ov Stres za MDS u jednoj dimenziji je 0.17 što predstavlja loš Stres, po Kruškalovom kriterijumu, ali ako pogledamo naše rešenje, ono je prilično dobro. Sa desne strane se nalaze Monako i Švajcarska koje predstavljaju napredne zemlje sa dobrim životnim standardom, a sa leve strane Zimbabve i Somalija.
Pokušaćemo sada u dve dimenzije. Kako je sada vrednost Stres-a 0.02, dobijeno rešenje veoma dobro predstavlja podatke u dve dimenzije.

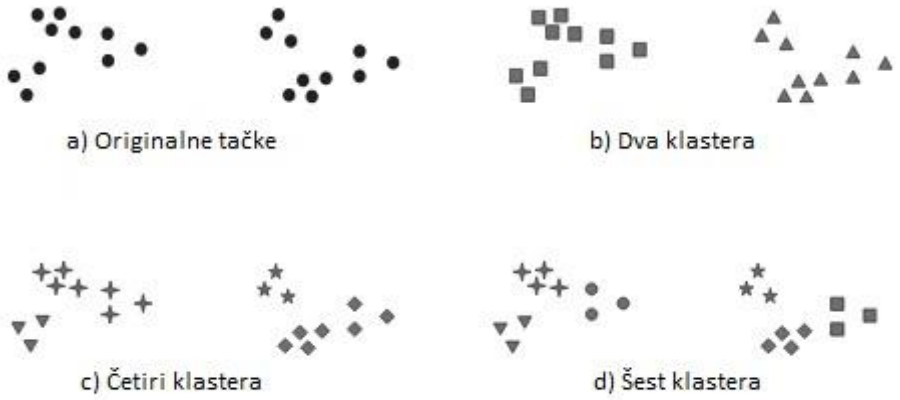


Slika 4.12: Dvodimenzionalni MDS za primer od 19 zemalja

U ovom slučaju Dimenzija 1 se može interpretirati kao veličina zemlje po broju stanovnika, dok Dimenzija 2 predstavlja BDP zemlje.

5 Klaster analiza

Jedna od najosnovnijih osobina ljudskih bića je sposobnost grupisanja sličnih objekata u grupe, tzv. klasifikacija. Samo sortiranje sličnih predmeta u kategorije potiče još od samog nastanka planete Zemlje, kada su prvi ljudi, morali da razlikuju različite objekte koji su delili neke iste ili slične osobine. Na primer, morali su da prepoznaju otrovne biljke, svirepe životinje itd. Ukoliko razmislimo malo bolje, klasifikacija je potrebna i za razvoj jezika, prilikom razgovora, ljudi prepoznaju određene reči, i grupišu ih, da li su to reči njhovog govornog područja ili pak neke strane reči. Klasifikacija je važna stavka svakog naučnog istraživanja. Istraživač je obično zainteresovan za nalaženje „pravila“ u kome su objekti koji se posmtraju sortirani u manji broj *homogenih grupa* ili *klastera,* i to tako, da su objekti unutar grupa slični jedni drugima. Češći zahtev je da se grupe međusobno isključuju (jedan element priprada samo jednoj grupi), nego da se preklapaju (elementi mogu biti u više od jedne grupe). U najmanju ruku, svaka dobijena klasifikaciona šema bi trebala da obezbedi pogodan metod za organizovanje velike i kompleksne grupe podataka. Klaster analiza je generički termin za širok spektar numeričkih metoda za ispitivanje multivarijantnih podataka sa ciljem da se otkriju grupe ili klasteri. U medicini, na primer, otkrivanje da određena grupa pacijenata ima iste simptome i da reaguje isto na određeni lek, može imati veliki uticaj na budući razvoj tretmana te bolesti. U marketing istraživanju, može biti korisno pri grupisanju velikog broja potencijalnih klijenata, prema njihovih potrebama za određenim proizvodima. Na ovaj način marketing kompanije plasiraju svoje reklame, na već targetirane klijente. U poslednje vreme, klaster analiza se koristi i u analizama fotografija. Očigledno, veliki je opseg sfera gde klasificikacija može biti primenjena šta god da klasifikujemo. Ljudska bića mogu biti klasifikovana prema ekonomskom statusu, na nižu klasu, srednju klasu, i višu klasu, ili mogu biti klasifikovani po količini cigareta koje konzumiraju godišnje, na nisku, srednju i visoku količinu. Iz ovoga se može zaključiti da različite klasifikacije mogu postojati i da će svaka od njih prikupljati različite podatke potrebne za samu klasifikaciju. Vrlo je važno napomenuti da neće svaka klasifikacija biti jednako korisna. Tehnike klaster analize nastoje da formalizuju ono što ljudi rade vrlo dobro u dve ili tri dimenzije. Na Slici 5.1 data je prvo originalna udaljenost tačaka, a potom i nekoliko tipova klastera. Klasteri su identifikovani procenom udaljenosti između tačaka, i u ovom slučaju zadatak je bio lak.



Slika 5.1: Jedan od obika klastera

Kao što je i rečeno na samom početku rada, fokus drugog dela će biti na aglomerativnim hijerarhijskim metodoma i metodi k-sredina.

5.1 Hijerarhijske metode

5.1.1 Uvod

Hijerarshijske tehnike grupisanja mogu se podeliti na metode udruživanja (eng.agglomerative methods) i metode deljenja (eng. divisive methods). Aglomerativne metode počinju nizom udruživanja objekata, dok metode deljenja razdvajaju jednu grupu od *n* objekata na više manjih grupa. Hijerarhijske tehnike se mogu grafički predstaviti preko dendograma koji prikazuje spajanje ili deljenje (u zavisnosti od tehnike koja se koristi) napravljeno u svakom koraku analize.

5.1.2 Aglomerativno hijerarhijsko grupisanje

Aglomerativne metode ili metode udruživanja su najrasprostranjenije hijerarhijske metode. U hijerarhijskoj klasifikaciji, podaci se nikada ne dele u određen broj klasa ili grupa u jednom koraku. Klasifikacija se sastoji od serija particija koje mogu krenuti od jednog “klastera” koji sadrži sve elemente, do *n* klastera gde svaki sadrži jedan element. Tehnika aglomerativnog hijerarhijskog grupisanja pravi serije sukcesivnih fuzija (stapanja) od *n* individua u grupe.

Ovakvim postupanjem, stapanja koja su napravljena su ireverzibilna, tj. kada se dva elementa nađu u istoj grupi, nije moguće da se kasnije nađu u dve različite grupe. Kako ova metoda redukuje podakte u jedan klaster koji sadrži sve elemente, cilj je naći broj klastera koji najbolje fituje podatke. O problemu odlučivanja broja klastera će biti reči nešto kasnije.

Procedura aglomerativnog hijerarhijskog grupisanja pravi serije particija podataka, $P_n, P_{n-1}, \dots, P_1$. Prva particija, P_n, sadrži *n* jednočlanih grupa, i poslednja P_1 sadrži grupu u kojoj se nalaze svih *n* pojedinačno. Osnovna operacija za sve metode je veoma slična:

Klasteri $C_1, C_2, \dots, C_n$ svaki sadrži po jedan element.

1. Naći najbliži par različitih klastera, neka to budu C_i i C_j , spojiti ih, i obrisati C_j i posle smanjiti broj klastera za jedan.
2. Ako je broj klastera jednak jedinici, zaustavljamo se, inače se vraćamo na 1.

Pre nego što proces počne, matrica udaljenosti ili matrica sličnosti mora biti izračunata. Postoje mnogi načini za to, ali u ovom radu će se uglavnom raditi sa najčešće korišćenom matricom udaljenosti, Euklidskom udaljnošću. Iako je u prvom delu ovog rada već definisana Euklidska udaljenost, ovde će biti ponovljena jos jednom. Dakle, Euklidsku udaljenost računamo kao:

$$d_{ij} = \sqrt{\sum_{k=1}^{q} \left(x_{ik} - x_{jk}\right)^2}$$

gde je d_{ij} Euklidska udaljenost između i, sa vrednostima promeljvih $x_{i1}, x_{i2}, \dots, x_{iq}$, i j sa vrednostima promeljnivih $x_{j1}, x_{j2}, \dots, x_{jq}$. Euklidska udaljenost može biti zapisana u obliku kvadratne matrice, koja je simetrična, zbog $d_{ij} = d_{ji}$ i koja ima nule na glavnoj dijagonali. Takva matrica je početna tačka svake klaster analize. Kada je takva matrica dobijena, hijerarhijsko grupisanje može da počne, i u svakom koraku procesa, metoda spaja elemente ili grupe formirane ranije, koje su najbliže (ili najsličnije). Kako su grupe formirane, udaljenost između dva elementa, ili dve grupe koje sadrže više elemenata, trebaju biti izračunate. Postoje mnoge tehnike za izračunavanje distance između klastera. Dve jednostvane mere za rastojanje između dve grupe su

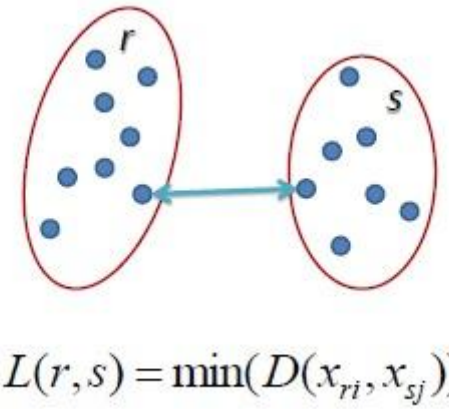
$$d_{AB} = \min_{i \in A,B}(d_{ij})$$

$$d_{AB} = \max_{i \in A,B}(d_{ij})$$

gde je d_{AB} distanca između dva klastera A i B, i d_{ij} je udaljenost između dva elementa *i* i *j* . Ovo može biti Euklidska distanca ili neka od distanci navedenih u *Glavi 3*. Prvi metod se naziva jednostruko povezivanje (eng. single linkage), a drugi metod je kompletno povezivanje (eng. complete linkage). Druga mogućnost za merenje međugrupne udaljenosti ili razlike je

$$d_{AB} = \frac{1}{n_A n_B} \sum_{i \in A} \sum_{i \in B} d_{ij}$$

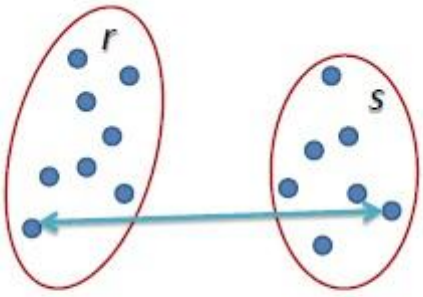
gde su n_A i n_B broj elemenata u klasterima A i B. Ovaj metod se još naziva i prosečno povezivanje (eng. group average).



Izvor: http://www.saedsayad.com/clustering_hierarchical.htm

Slika 5.2: Jednostruko povezivanje

Kod jednostrukog povezivanja, udaljenost između klastera je definisana kao najkraća udaljenost između dva objekta svakog klastera.

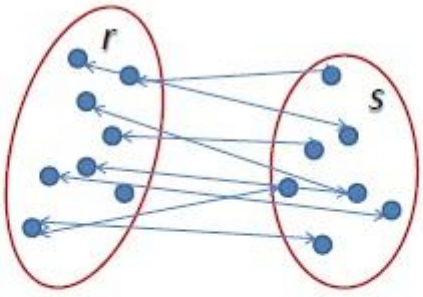


$$L(r,s) = \max(D(x_{ri}, x_{sj}))$$

http://www.saedsayad.com/clustering_hierarchical.htm

Slika 5.3: Kompletno povezivanje

Kod kompletnog udruživanja, distanca između dva klastera je definisana kao najduža udaljenost između dva objekta svakog klastera.



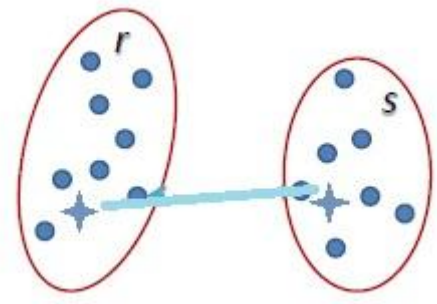
$$L(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} D(x_{ri}, x_{sj})$$

http://www.saedsayad.com/clustering_hierarchical.htm

Slika 5.4: Prosečno povezivanje

Kod prosečnog povezivanja, distanca je definisana kao prosečna udaljenost između svaka dva objekta iz oba klastera.

Sledeća metoda udruživanja je metoda centroida. U ovom postupku dve grupe se spajaju ako su njihovi centroidi najmanje udaljeni u odnosu međusobnu udaljenost svih mogućih parova grupa koje postoje.



http://www.saedsayad.com/clustering_hierarchical.htm

Slika 5.5: Centroidi

Još jedna metoda grupisanja koja će biti spomenuta u ovom radu je Ward-ova metoda. Ona predstavlja alternativni pristup klaster analizi. On posmatra klaster analizu kao problem analize varijanse, umesto korišćenja udaljenosti kao mere povezanosti. Ovaj metod (videti u [13]) će početi sa tzv. "listovima" i na kraju završiti sa "stablom". Ward-ov metod počinje sa *n* klastera veličine 1, i zaustavlja se kad se sve observacije nalaze u jednom klasteru. U ovoj metodi koristi se greška sume kvadrata.

Definicija 1. Greška sume kvadrata je data sledećom formulom

$$ESS = \sum_i \sum_j \sum_k \left| x_{ijk} - \overline{x_{\imath \cdot k}} \right|^2$$

gde je x_{ijk} vrednost promenljive *k* u observaciji *j* koja pripada klasteru *i,* $\overline{x_{\imath \cdot k}}$ je srednja vrednost *i*-tog klastera za *k*-tu promenljivu. Ovde sumiramo po svim promenljiva. Porede se pojedinačne observacije za svaku promenljivu sa srednjom vrednošću klastera za tu promenljivu. Ukoliko je ESS mala vrednost, to govori da su naši podaci blizu srednje vrednosti klastera, što dalje znači da imamo klaster sličnih jedinica.

Definicija 2. Totalna suma kvadrata definisana je kao

$$TSS = \sum_{i}\sum_{j}\sum_{k}\left|x_{ijk} - \overline{x_{..k}}\right|^2$$

Definicija 3. R kvadrat (R^2) predstavlja proporciju varijacije koja je objašnjena određenim grupisanjem observacija i data je formulom

$$R^2 = \frac{TSS - ESS}{TSS}$$

Koristeći Ward-ov metod počinje se sa n klastera veličine 1, tj. u svakom klasteru se nalazi po jedan element. U prvom koraku algoritma, formira se $n-1$ klastera, jedan veličine 2, a ostali veličine 1. Potom se računa ESS i R^2 . Par elemenata koji daje najmanju ESS ili najveću R^2 vrednost, će formirati prvi klaster. Potom, u sledećem koraku, $n-2$ klastera su formirana od pomenutih $n-1$ klastera definisana u drugom koraku. Ovde mogu nastati klasteri veličine 2, ili jedan klaster veličine 3 koji sadrži dva elementa grupisana u prvom koraku. Ponovo, vrednost R^2 je maksimizarana. Dalje se postupak ponavlja, tako da se u svakom koraku algoritmi klasteri kombinuju tako da se minimizira greška sume kvadrata ili da se maksimizira vrednost R^2. Proces se zaustavlja kada su svi elementi sadržani u jednom velikom klasteru veličine n.

5.1.3 Lance i Williams formula

Lance i Williams-ova rekurentna formula daje rastojanje između grupe k i grupe $(i+j)$ koja je nastala spajanjem grupa i i j. Kako su grupe i i j bile najbliže jedna drugoj, nastao je novi klaster, a pojedinčne grupe su obrisane. Sada umesto da se traži odstojanje između k i i, i grupe k i j, kao što je rečeno, sada se traži odstojanje između grupe k i grupe $(i+j)$ kao

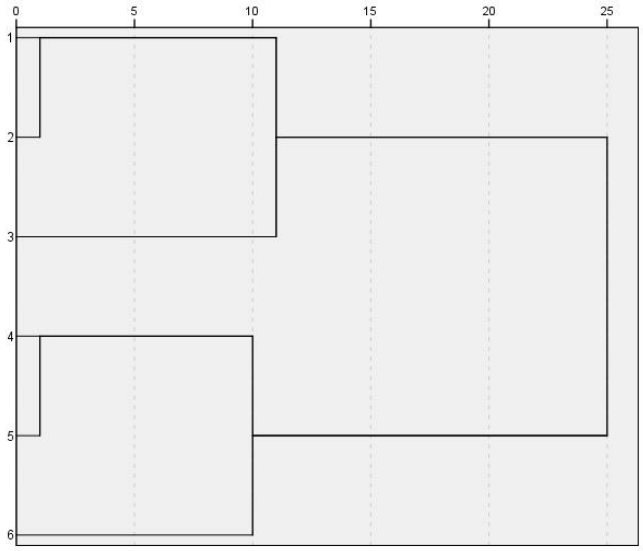
$$d_{k,i+j} = \alpha_i d_{k,i} + \alpha_j d_{k,j} + \beta d_{i,j} + \gamma\left|d_{k,i} - d_{k,j}\right|$$

gde je $d_{i,j}$ rastojanje između grupa i i j. U zavisnosti od toga kakve vrednosti parametri α_i, α_j, β i γ uzimaju, dobijamo različite formule. U Tabeli 5.1 date su vrednosti parametara kao i kojoj vrsti povezivanja pripadaju.

Metoda	$\alpha_i = \alpha_j$	β	γ
Jednostruko povezivanje	$\frac{1}{2}$	0	$-\frac{1}{2}$
Kompletno povezivanje	$\frac{1}{2}$	0	$\frac{1}{2}$
Prosečno povezivanje	$\frac{n_i}{n_i + n_j}$	0	0
Metoda centroida	$\frac{n_i}{n_i + n_j}$	$\frac{-n_i n_j}{(n_i + n_j)^2}$	0
Ward-ov metod	$\frac{(n_k + n_i)}{(n_k + n_i + n_j)}$	$\frac{-n_k}{(n_k + i + n_j)}$	0

Tabela 5.1: Lance-Williams parametri

Hijerarhijska klasifikacija može biti predstavljena kao dvodimenzionalni diagram, poznatiji kao *dendogram*. Dendogram prikazuje spajanja koja se dogode u svakom koraku analize. Čvorovi dendograma predstavljaju klastere, a dužina grana udaljenost na kojima su klasteri udruženi.



Slika 5.4: Dendogram za pet objekata

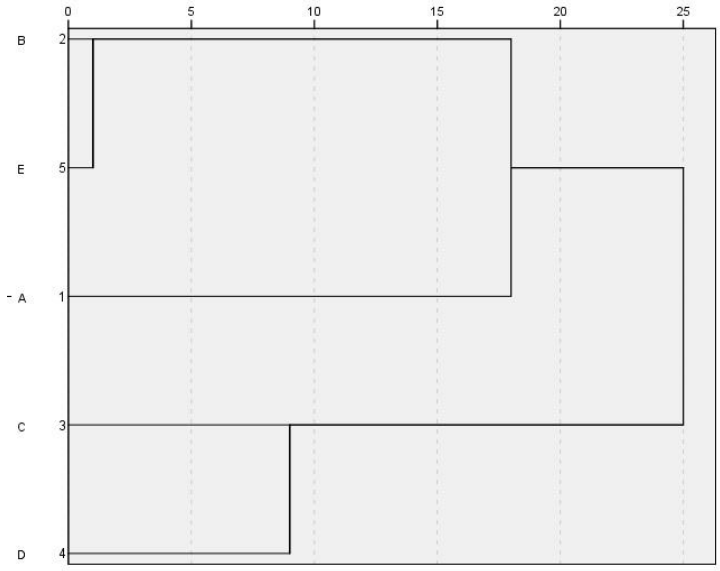
Dalje će biti prikazan jedan prost primer, i kako hijerarhijska klaster analiza može biti primenjena.

Primer 1. Imamo pet studenata, A, B, C, D, E i date njihove godine, grad iz kog dolaze, prosek na fakultetu kao i mesečna primanja. Izabrali smo centroid metodu spajanja.

Case	4 Clusters	3 Clusters	2 Clusters
1:A	1	1	1
2:B	2	2	1
3:C	3	3	2
4:D	4	3	2
5:E	2	2	1

Tabela 5.2: Tabela pokazuje kom klasteru pripadaju studenti u zavisnosti od samog broj klastera

U Tabeli 5.2 mogu se videti, u zavisnosti od broja klastera (u ovom slučaju 2, 3 i 4), kom klasteru pripada koji student. Još jedan od načina prikazivanja, kao što je rečeno je dendogram.



Slika 5.5: Dendogram za 5 studenata

U slučaju dendograma, treba odabrati određenu particiju podataka (rešenje sa određenim brojem grupa). Odgovor je da mi "sečemo" dendogram na nekoj visini i to daje određen broj grupa. U ovom slučaju, na ovako prostom primeru, to nije naročito teško, ali kada je broj elemenata (ispitanika) veliki, (recimo 300, 500 ili više) to postaje malo teži zadatak. Pitanje je gde je optimalno preseći dendogram? Jedan od pristupa je izučavanje koliko su velike promene u visini dendograma, i tamo gde je promena "velika" označava odgovarajući broj klastera.

5.1.4 Metode deljenja

Metode deljenja počinju iz suprotnog pravca u odnosu na metode udruživanja (aglomerativne metode). Naime, one počinju jednom velikom grupom koja se potom deli na više manjih. U ovom radu neće se puno govoriti o ovom tipu klaster analize, ali vredi pomenuti da se metode deljenja sastoje iz jednodimenzionalnih i višedimenzionalnih metoda deljenja.

Kod jednodimenzionalnih metoda deljenja, izbor promenljive po kojoj će se praviti podela, zavisi od homogenosti klastera ili od povezanosti sa ostalim promenljivim. Primer kriterijuma homogenosti je informacija sadržaja, *C*, koji je definisan sa *p* promenljivih i *n* objekata

$$C = pn \log n - \sum_{i=1}^{p} [f_i \log f_i - (n - f_i) \log(n - f_i)]$$

gde je f_i broj elementa koji imaju *i*-ti atribut. Umesto homogenosti klastera, može se koristiti povezanost sa ostalim promenljivima (vidi u [1]).

Višedimenzionalne metode udruživanja su sličnije aglomerativnim metodama, jer za razliku od jednodimenzionalnih metoda, koje koriste jednu promenljivu, koriste sve promeljive zajedno i takođe rade sa matricom sličnosti. U prvom koraku se počinje sa nalaženjem elementa koji je najudaljeniji od svih ostalih unutar grupe, i upravo taj element je početna stanica (osnova) za odvojenu grupu. U drugom koraku se svaki objekat posmatra kao ulaz u odvojenu grupu, i svaki element koji je bliži odvojenoj grupi se pripaja toj grupi. Ovi koraci se ponavljaju dok se ne dođe do konačne podele, tj. dok se svaki element na nalazi pojedinačno u jednom klasteru. Već je spomenut problem izbora grupa, tj. koliko je grupa optimalno tražiti kao rešenje. Vrlo često je slučaj da sam istraživač nije zainteresovan za totalnu podelu, već za recimo jednu ili dve particije. Kod standardnih aglomerativnih metoda kao i metoda deljenja, particiju postižemo odabirom jednog od rešenja, što je ekvivalentno presecanju dendograma na određenoj visini, kao što je već rečeno.

5.2 Nehijerarhijske metode

Nehijerarhijske metode se razlikuju od hijerarhijskih, i aglomerativnih i deljivih, po tome što postoji mogućnost da objekti promene ranije formirane grupe, za razliku od hijerarhijskih gde je to nije moguće. Ukoliko neki od kriterijuma sugeriše da element treba da promeni grupu u kojoj se nalazi, to je moguće. Jedna od pretpostavki je da je broj grupa unapred određen.

U prvoj fazi nehijerarhijskog grupisanja počinje se podelom skupa objekata na unapred izabran broj manjih grupa. U drugoj fazi se određuje distanca između svakog objekta i svake grupe. Ukoliko je objekat najbliži određenoj grupi,biće smešten u tu grupu. U trećoj fazi, nakon pridruživanja objekta grupi, izračunava se centroid grupe iz koje je objekat otišao kao i one u koju je došao. Zatim se ponovo računa rastojanje od centroida grupe i vrši se preraspodela dokle god to izabrani kriterijum sugeriše. Najrasprostranjenija nehijararhijska metoda, je metoda *k-sredina* (eng. k-means method) o kojoj se govori u narednom poglavlju.

5.2.1 Metoda k-sredina

Ideja podela skupa podataka je da obezbedi particiju od *n* objekata u *k* disjunktnih klastera. Po definiciji, objekat može da pripada jednom klasteru i svaki klaster mora da ima barem jedan objekat (drugačije bi imali manje od *k* klastera). Klasifikacioni algoritam je obično iterativan; inicijalni korak je obično poboljšan u svakom sledećem koraku sve dok ima poboljšanja. Definisanje inicijalne particije zahteva *a priori* specifikaciju broja klastera. Pretpostavimo da je „mera“ koliko je dobra particija, predstavljena nekom funkcijom J, čija se vrednost smanjuje koliko je to moguće kako bi postigli dalju optimizaciju rezultata. Opšti algoritam za sve metode ovakvog tipa bi bio:

1. Odredi inicijalnu particiju u *k* klastera i izvedi vrednost za funkciju J
2. Promeni particiju kako bi se smanjila vrednost J što je više moguće, ostavljajući *k* nepromenjeno
3. Ako nova smanjenja J nisu moguća, proces će stati, i broj klastera koji postoje u tom momentu, će biti i konačan broj klastera. U suprotnom se vraćamo na *Korak 2*.

Postoje različite procedure u definisanju koliko dobro fituje funkcija J i kakve operacije su dozvoljene u drugom koraku. Često se može desiti da opšti algoritam završi u lokalnom minimumu, tj. da dobijemo rešenje koje nije najbolje klasifikovalo objekte u *k* grupa prema postojećem kriterijumu J. Takođe, inicijalna particija može biti problem, koji dovodi do

veoma lošeg rešenja. Ovaj problem se prevazilazi izvođenjem desetina različitih početnih particija koje dovode do „najboljeg“ rešenja. Nikada ne možemo biti 100% sigurni da smo dobili globalni optimum; da bi to uradili, morali bi da proverimo sve moguće particije, što je za velik broj *n* nemoguće. Particija može biti modifikovana u *Koraku 2.* na dva načina:

- Proučiti za svaki objekat pojedinačno, kako njegova relokacija, iz grupe u kojoj se nalazi u neku drugu grupu, utiče na vrednost J. Objekti koji smanjuju J se premeštaju u grupu u kojoj je ovo smanjivanje maksimalno. Moguće je da mnogo objekata ili čak i svi objekti moraju biti premešteni u jednom koraku, i tada se samo nadamo da će nova vrednost J biti manja nego prethodna.
- Objekat za koji je primećeno maksimalno opadanje vrednosti J je izabran i premešten u novu grupu. Ova strategija zahteva monotono opadanje J, i definitivno je sporija nego prethodni metod.

Kod metoda k-sredina traži se da se particija od *n* objekata razdeli u *k* grupa ili klastera $(G_1, G_2, \dots G_k)$, gde G_i predstavlja skup od n_i pojedinaca u *i*-toj grupi, i k je dato, minimizirajući neki numerčki kriterijum (recimo funkcija J), gde se niske vrednosti smatraju „dobrim“ rezultatom. Jedna od najčešće korišćenih implementacija metoda k-sredina nastoji da nađe particiju *n* individua sa *k* grupa koje minimiziraju sumu kvardata u grupi (eng. within-group sum squares) ili skraćeno WGSS po svim promenljivima.

$$WGSS = \sum_{j=1}^{q} \sum_{l=1}^{k} \sum_{i \epsilon G_l} (x_{ij} - \bar{x}_j^{(l)})^2$$

gde je $\bar{x}_j^{(l)} = \frac{1}{n_i} \sum_{i \epsilon G_1} x_{ij}$ srednja vrednost za pojedinačne elemente u grupi G_l po promenljivoj *j.*

Problem deluje lagan; razmotriti svaku moguću particiju od *n* elementa u *k* grupa, i izabrati jednu sa najmanjim WGSS. Nažalost, u praksi problem nije tako jednostavan. Brojevi koji su uključeni u deljenje su toliko veliki da kompletno nabrajanje nije moguće ni sa najboljim kompjuterima. Kako proučavanje svake moguće particije nije moguće, razvijeni su neki algoritmi koji su ubrzali i olakšali sam proces. Ovi algoritmi traže minimalne vrednosti za kreiranje klastera, prekomponovanje postojećih particija i zadržavajući nove, samo ukoliko je postignuto neko poboljšanje. Kao i do sada, ni ovi algoritmi ne garantuju pronalazak globalnog minimuma. Koraci koje se koriste su sledeći:

1. Pronalazak inicijalne particije objekata, u potreban broj klastera.
2. Računanje promene u zadatom kriterijum, pri premeštanju jednog objekta iz klastera u drugi klaster.
3. Pravljenje promena koje vode do najvećeg poboljšanja u vrednosti postvaljenog kriterijuma.
4. Ponavljanje koraka (2) i (3) dok god individualne promene objekata utiču na poboljšanje kriterijuma.

Pristup metode k-sredina koji koristi minimalizaciju u grupama po sumi kvadrata po svakoj promenljivoj je široko rasprostranjen ali naučnici se susreću sa dva problema. Prvi je da metoda k-sredina je skalarano invarijantna, tj. da različita rešenja mogu biti dobijena ukoliko se radi sa "sirovim" podacima ili ukoliko se radi sa standardizovanim podacima. Drugi problem je što se nameće "sferan" oblik grupa podataka, tj. podaci će uvek biti oblikovani kao hiper-lopte iako su "stvarne" grupe u podacima nekog drugog oblika. Pored toga, metoda k-sredina je i dalje veoma popularna. Pomoću ove metode, istraživač sam bira u koliko klastera bi želeo da rasporedi podatke. Postoje mnogi načini za određivanje broja klastera, ali nijedan nije potpun. Metoda koju ćemo koristiti je crtanje sume kvadrata naspram rešenja metode k-sredina. Kako broj grupa raste, suma kvadrata će opadati, ali opet (kao i u *Glavi* 4) ćemo tražiti "lakat" koji može biti najbolji indikator koliko grupa nam je dovoljno.

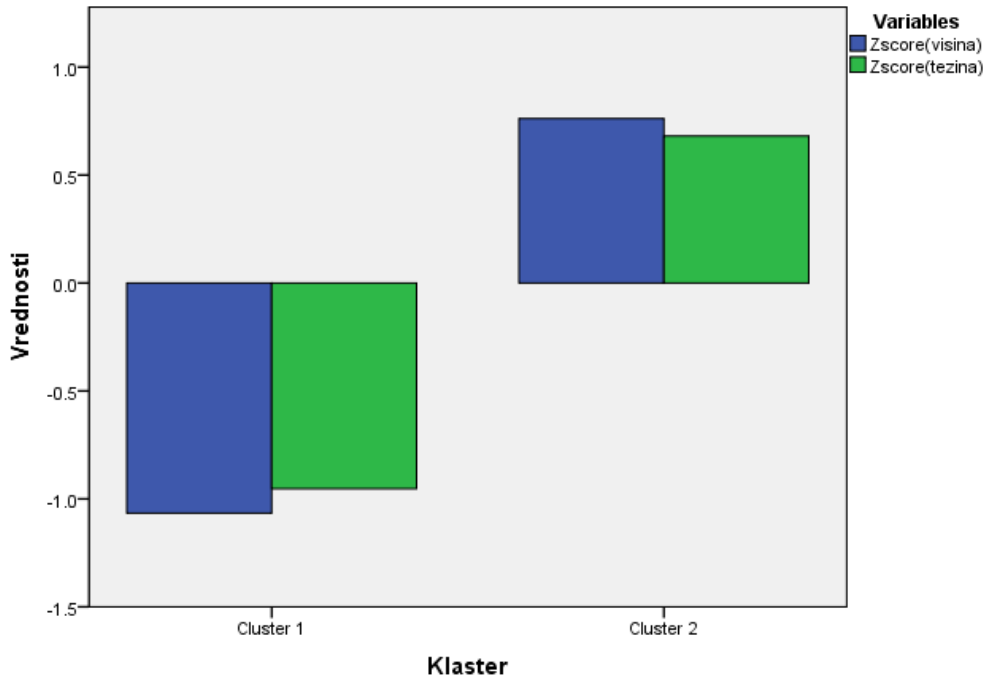
Primer1. U ovom primeru uradićemo analizu 12 različitih rasa pasa. U Tabeli 5.3 su dati podaci na osnovu kojih će biti urađena analiza.

Rasa	Visina	Težina
Labrador	25.00	80.00
Nemački ovčar	26.00	95.00
Jorkšir	9.00	6.00
Zlatni retriver	24.00	75.00
Bigl	15.00	30.00
Bokser	25.00	80.00
Buldog	15.00	55.00
Dahšun	9.00	32.00
Pudla	21.00	65.00

Doberman	28.00	90.00
Rotfajler	27.00	135.00
Ši Cu	11.00	16.00

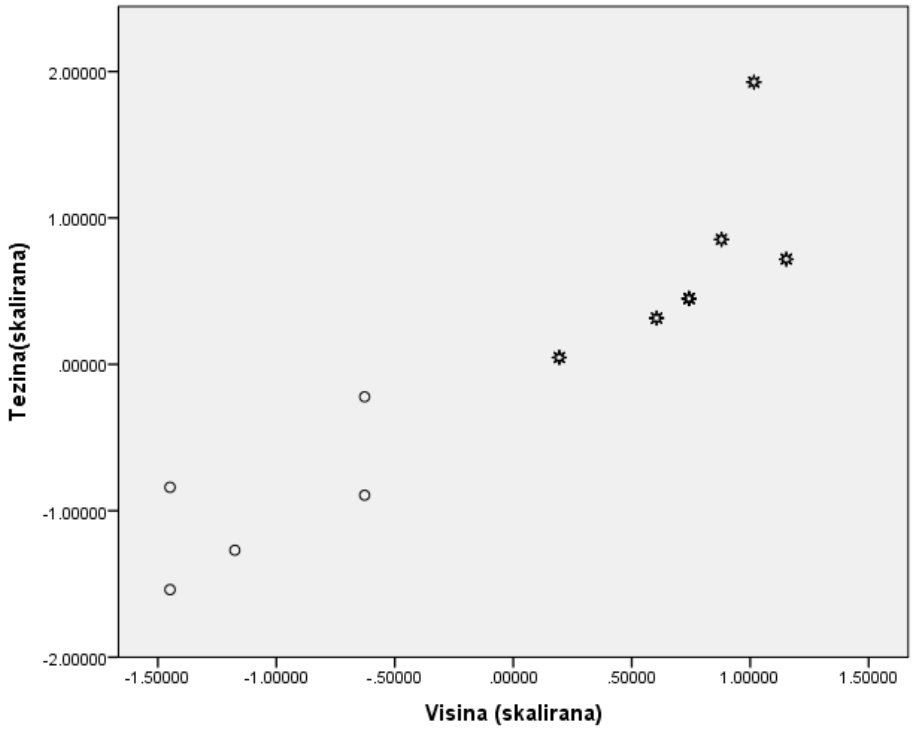
Tabela 5.3: Prikaz dvanaest različitih rasa pasa

S obzirom na mali broj podataka, odabir je rešenje sa dva klastera.



Slika 5.6: "Osobine" klastera

Iz Slike 5.6 se vidi da postoje dva klastera, i takođe se vidi da u prvom klasteru nalaze elementi koji imaju niske standardizovane vrednosti (z vrednosti) visine i težine, dok je u drugom klasteru situacija obrnuta. Ukoliko bi to nacrtali, izgledalo bi ovako



Slika 5.7: Dvodimenzionalan prikaz dva klastera

5.3 Primeri

Primer 1. Veoma često, pogotovo u poslednje vreme, ljudi čitaju sastojke hrane koju jedu. U ovom primeru razmotrićemo koliko je koji sendvič sličan. U Tabeli 5.4 dati su podaci za svaki sendvič.

Sendviči	Kalorije	Masti total	Zasićene masti(g)	Trans masti (g)	Holesterol (mg)	Vlakna(g)	Šećeri (g)	Proteini(g)
Cheesesteak	503	21	5	0	84	1	6	38
Cheesesteak Sandwich(with BBQ)	573	21	5	0	84	1	22	38
Cheesesteak Sandwich with Mayo	683	41	8	0	84	1	6	38
Chicken Melt	884	42	19	0	174	8	17	69
Fried Chicken Finger Sandwich	630	28	5	0	93	5	9	29
Fried Chicken Finger Sandwich Mushrooms	638	28	5	0	93	5	9	30
Fried Chicken Finger Wrap	264	14	2	0	50	3	4	20
Fried Chicken Finger Wrap B.Olives	302	17	2	0	50	4	4	20
Grilled Cheese Sandwich	410	15	3	0	0	2	18	8
Grilled Cheese Sandwich A1 steak sauce	440	15	3	0	0	2	22	8
Grilled Chicken Breast Sandwich	584	20	5	0	141	2	5	48
Grilled Chicken Breast Sandwich Bacon	643	25	6	0	151	2	5	52
Grilled Chicken Breast honey mustard	744	36	7	0	151	2	9	48
Grilled Chicken Breast Wrap	218	6	1	0	99	0	0	39
Grilled Chicken Breast Wrap buffalo sauce	225	6	1	0	99	0	0	40
Grilled Chicken Breast Wrap cheddar cheese	380	20	9	0	145	0	0	49
Grilled Portobello Mushroom Melt	1010	60	22	0	118	7	15	39
Grilled Portobello Mushroom Melt with Creamy Jalapeno S	1422	103	30	1	121	7	17	40
Grilled Portobello Mushroom Sandwich	626	34	7	0	43	5	11	15
Grilled Portobello Mushroom Sandwich with 1000 Island D	766	47	9	0	58	5	14	15
Grilled Portobello Mushroom Sandwich with Onion Rings	641	34	7	0	43	6	12	15
Grilled Portobello Mushroom Sandwich with Pickles	627	34	7	0	43	6	11	15
Grilled Portobello Mushroom Sandwich with Ranch Dressin	756	47	9	0	53	5	12	16
Grilled Portobello Mushroom Sandwich with Teriyaki Sauc	671	34	7	0	43	5	20	15
Grilled Portobello Mushroom Wrap	260	20	3	0	0	4	6	6
Grilled Portobello Mushroom Wrap with Bleu Cheese Crumb	360	28	8	0	25	4	6	12
Grilled Portobello Mushroom Wrap with Chipotle Mayo	450	40	6	0	10	4	6	6
Grilled Portobello Mushroom Wrap with Roasted Red Peppe	266	20	3	0	0	4	7	6
Grilled Portobello Mushroom Wrap with Swiss Cheese	425	32	11	0	38	4	6	18
Hot Dog	607	41	14	0	80	1	5	20
Hot Dog with Guacamole	702	49	15	0	80	3	7	22
My Bleu Chicken	1058	53	25	0	199	2	18	77
Our Famous Pounder	1415	83	30	0	395	2	5	109
Our Famous Pounder Wrap with White Wrap	1350	77	29	0	352	1	2	108
Patty Melt	1192	71	31	0	251	8	17	80
Patty Melt with Zesty Horseradish Sauce	1762	131	40	0	281	8	23	80
Ribeye Steak Sandwich	698	37	11	0	185	2	5	39
Ribeye Steak Sandwich with Peanut Butter	875	51	14	0	185	4	8	45
Roast Beef Sandwich	662	26	7	0	163	2	9	49
Salmon Burger	646	32	6	0	148	2	6	38
Sausage Burger Sandwich	983	61	24	0	183	2	5	53
The Classic with Dijon Mustard	698	36	11	0	139	3	6	39
The Delirious with American Cheese	1309	81	34	0	334	2	5	91
The Delirious Wrap with Banana Peppers	753	49	18	0	247	2	1	71
The Semi Serious with Substitute Wheat Bun	953	41	13	0	166	2	11	50
Turkey Burger Platter with Bacon	329	22	5	0	100	1	1	31
Turkey Burger Platter with Pineapple	283	17	4	0	90	1	4	28
Turkey Burger Sandwich with BBQ Sauce	696	31	7	0	133	2	21	36
Veggie Burger	621	23	6	0	83	7	7	22
Veggie Burger with Cole Slaw	666	26	7	0	86	7	10	22

Tabela 5.4: Sendviči i nutricione vrednosti

Recimo da su najbitniji elementi, tj. sastojci koje određena hrana sadrži, kalorije, masti, holesterol i proteini, pa ćemo te elemente i uzeti u obzir u našoj analizi. Prvo se moraju standardizovati vrednosti pomenutih elemenata. Postoje razne vrste standardizacije, u ovom radu će biti korišćene Z vrednosti. Primenjena je metoda k-sredina, i u ovom slučaju odabrano je postojanje tri klastera. U Tabeli 5.5 se vidi koliko se elemenata nalazi u kom klasteru.

Klaster	1	35.00
	2	10.00
	3	5.00
Ispravni		50.00
Nedostajući		0.00

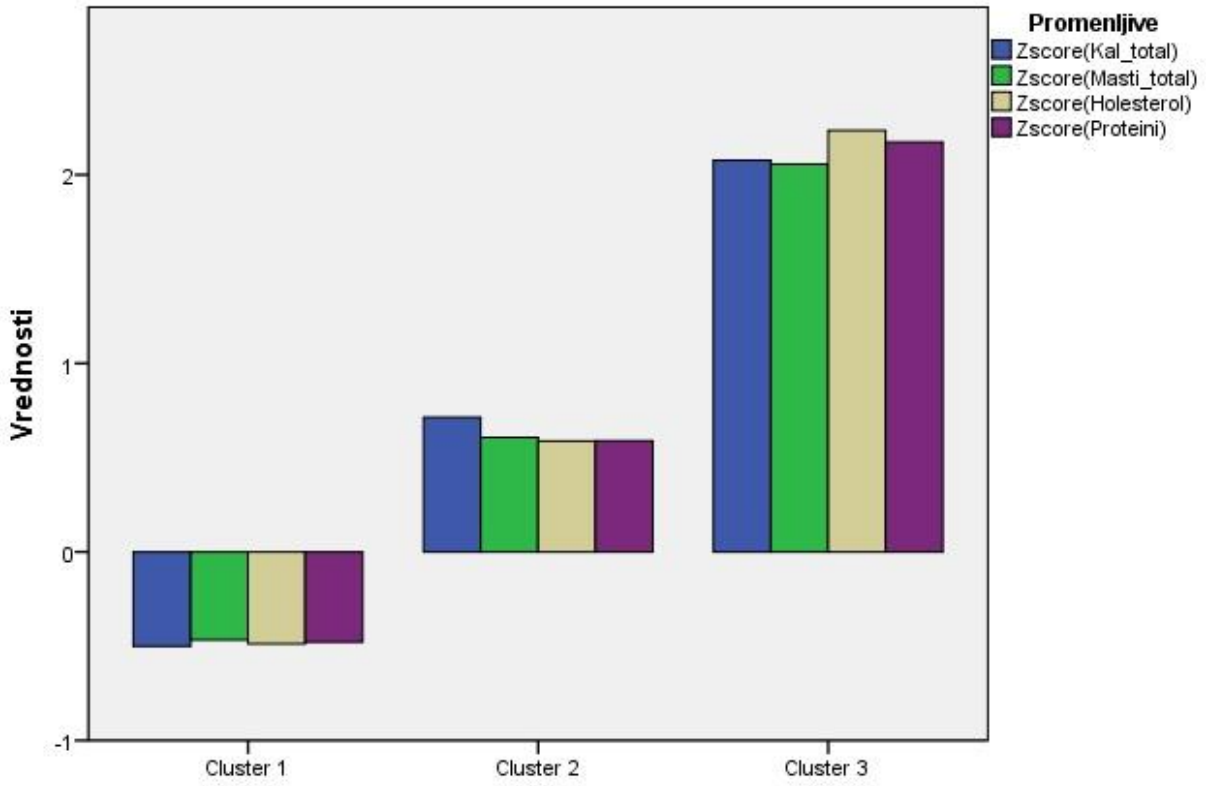
Tabela 5.5: Broj elemenata u svakom klasteru

Potom, skrećemo pažnju na to, koliko su daleko centri klastera jedan od drugog, što se može videti u sledećoj tabeli. Ona pokazuje prosečne vrednosti za standardizovane vrednosti.

	Cluster		
	1	2	3
Zscore(Kal_total)	-.50075	.71390	2.07743
Zscore(Masti_total)	-.46708	.60678	2.05599
Zscore(Holesterol)	-.48725	.58793	2.23487
Zscore(Proteini)	-.47873	.58902	2.17305

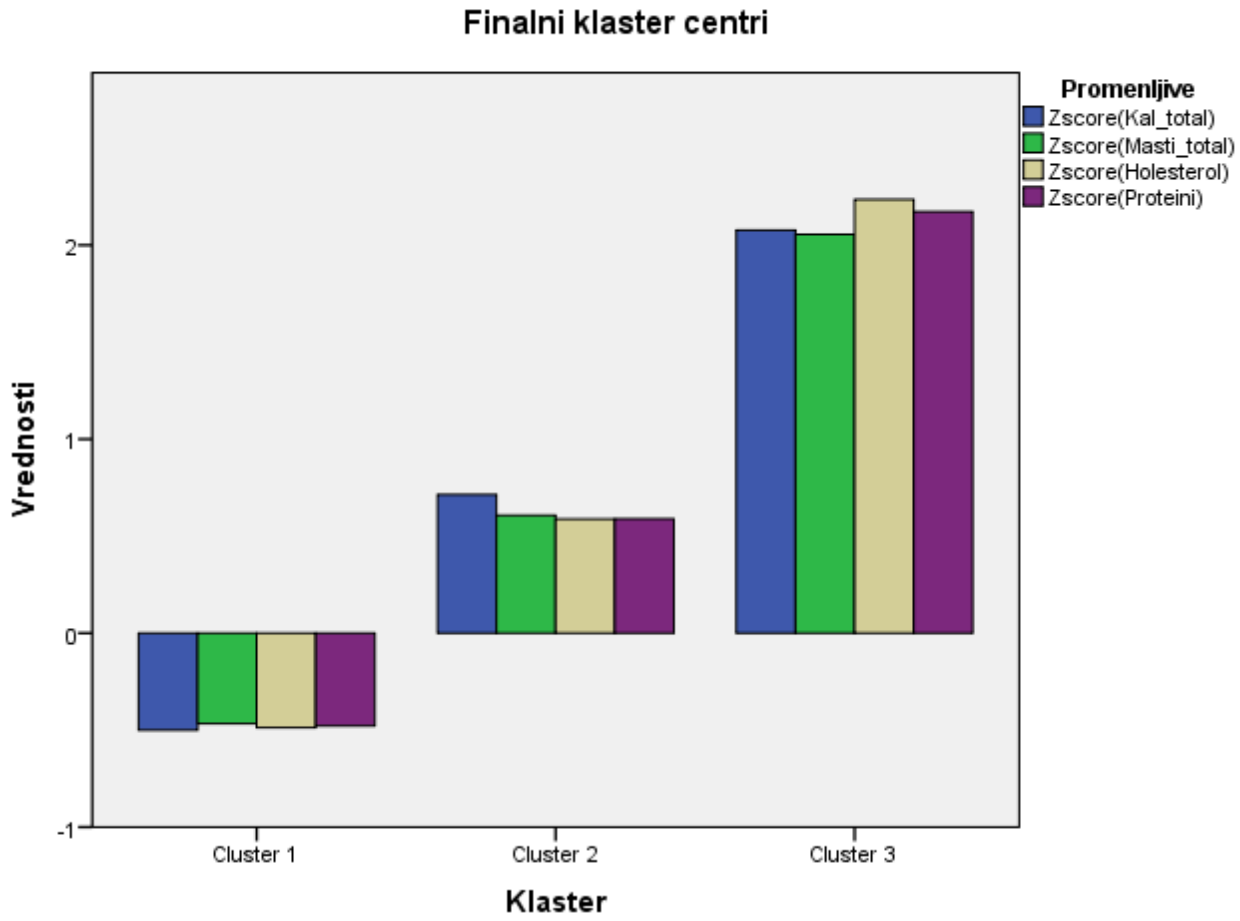
Tabela 5.6: Vrednosti centara klastera

Dalje, sledeća slika jasno pokazuje kakvi su nam klasteri, tj. njihove osobine .



Slika 5.8: "Osobine" klastera

U ovom grafičkom prikazu, se vidi da je prva grupa sastavljena od sendviča sa veoma niskim vrednostima kalorija, masti, holesterola i proteina. U drugoj grupi se nalaze sendviči sa nešto višim vrednostima u odnosu na prvu grupu, dok se u trećoj nalaze sendviči sa veoma visokim vrednostima koje su uzete kao parametri. Pomoću ovog grafičkog prikaza, možemo odrediti kojoj grupi pripada određen sendvič.



Slika 5.9: Osobine finalnih klastera

U prvom klasteru, se nalaze elementi sa veoma niskim vrednostima, masti, kalorija, holesterola i proteina. U drugom klasteru, imamo elemente sa nešto većim vrednostima, dok se u trećem klasteru nalaze veoma visoke vrednosti.

Primer 2. U drugom primeru, primenjuje se hijerarhijska metoda na grupu automobila datu u Tabeli 5.6. U ovom primeru biće razmatrano više različitih parametara, i videće se kako to utiče na formiranje samih klastera.

Model	Kubikaža	Snaga motora	Gorivo	0-100kmh	Max brzina	Potrošnja	Rezervoar
Alfa Romeo 156	1598	120	benzin	10.2	124	34	63
Citroen C5 Saloon	1560	110	dizel	11.2	118	56	71
Honda Accord Saloon	1997	153	benzin	10.5	132	38	65
Audi A3 Sportback	1197	103	benzin	10.2	120	56	50
Renault Grand Espace	2188	147	dizel	11.1	117	35	83
Fiat 500L	1248	93	dizel	15	101	70	50
Seat Toledo	1968	138	dizel	9.7	125	47	55
Peugeot 607	2230	160	benzin	9.3	136	31	80
MINI Coooper S	1598	181	benzin	6.8	142	48	50
Aston Martin DB9	5935	470	benzin	4.6	190	17	85
Ferrari 599 GTB	5999	612	benzin	3.6	205	13	105
Maserati GranTurismo	4691	433	benzin	4.8	183	18	86
Mercedes ML350	2987	254	dizel	7.2	139	39	93
BMW X6	4395	566	benzin	4.1	155	25	85

Tabela 5.7: Modeli automobila sa njihovim specifikacijama

Prvo ćemo prikaziati matricu sličnosti koja će biti ista za oba slučaja.

Case	1:Alfa Romeo 156	2:Citroen C5 Saloon	3:Honda Accord Saloon	4:Audi A3 Sportback	5:Renaul Grand Espace	6:Fiat 500L	7:Seat Toledo
1:Alfa Romeo 156	.000	2.193	.178	2.316	.216	7.451	.798
2:Citroen C5 Saloon	2.193	.000	1.410	.150	2.266	2.085	.615
3:Honda Accord Saloon	.178	1.410	.000	1.629	.220	6.044	.363
4:Audi A3 Sportback	2.316	.150	1.629	.000	2.733	2.789	.659
5:Renaul Grand Espace	.216	2.266	.220	2.733	.000	6.992	.943
6:Fiat 500L	7.451	2.085	6.044	2.789	6.992	.000	4.830
7:Seat Toledo	.798	.615	.363	.659	.943	4.830	.000
8:Peugeot 607	.288	3.385	.527	3.557	.354	9.784	1.269
9:MINI Coooper S	2.021	1.974	1.727	1.360	2.829	7.908	.839
10:Aston Martin DB9	13.266	21.284	13.855	21.909	12.264	33.719	15.205
11:Ferrari 599 GTB	18.919	28.790	20.082	29.470	17.833	42.687	21.724
12:Maserati GranTurismo	8.804	16.236	9.549	16.509	8.355	28.039	10.778
13:Mercedes ML350	1.567	3.471	1.353	3.391	1.680	10.539	1.232
14:BMW X6	6.640	12.088	6.901	11.954	6.611	23.249	7.400

8:Peugeot 607	9:MINI Coooper S	10:Aston Martin DB9	11:Ferrari 599 GTB	12:Maserati GranTurismo	13:Mercedes ML350	14:BMW X6
.288	2.021	13.266	18.919	8.804	1.567	6.640
3.385	1.974	21.284	28.790	16.236	3.471	12.088
.527	1.727	13.855	20.082	9.549	1.353	6.901
3.557	1.360	21.909	29.470	16.509	3.391	11.954
.354	2.829	12.264	17.833	8.355	1.680	6.611
9.784	7.908	33.719	42.687	28.039	10.539	23.249
1.269	.839	15.205	21.724	10.778	1.232	7.400
.000	2.285	9.666	14.669	5.930	.896	4.292
2.285	.000	15.376	21.748	10.615	1.193	6.435
9.666	15.376	.000	.937	.636	8.461	2.399
14.669	21.748	.937	.000	2.238	13.754	5.590
5.930	10.615	.636	2.238	.000	5.230	.964
.896	1.193	8.461	13.754	5.230	.000	2.629
4.292	6.435	2.399	5.590	.964	2.629	.000

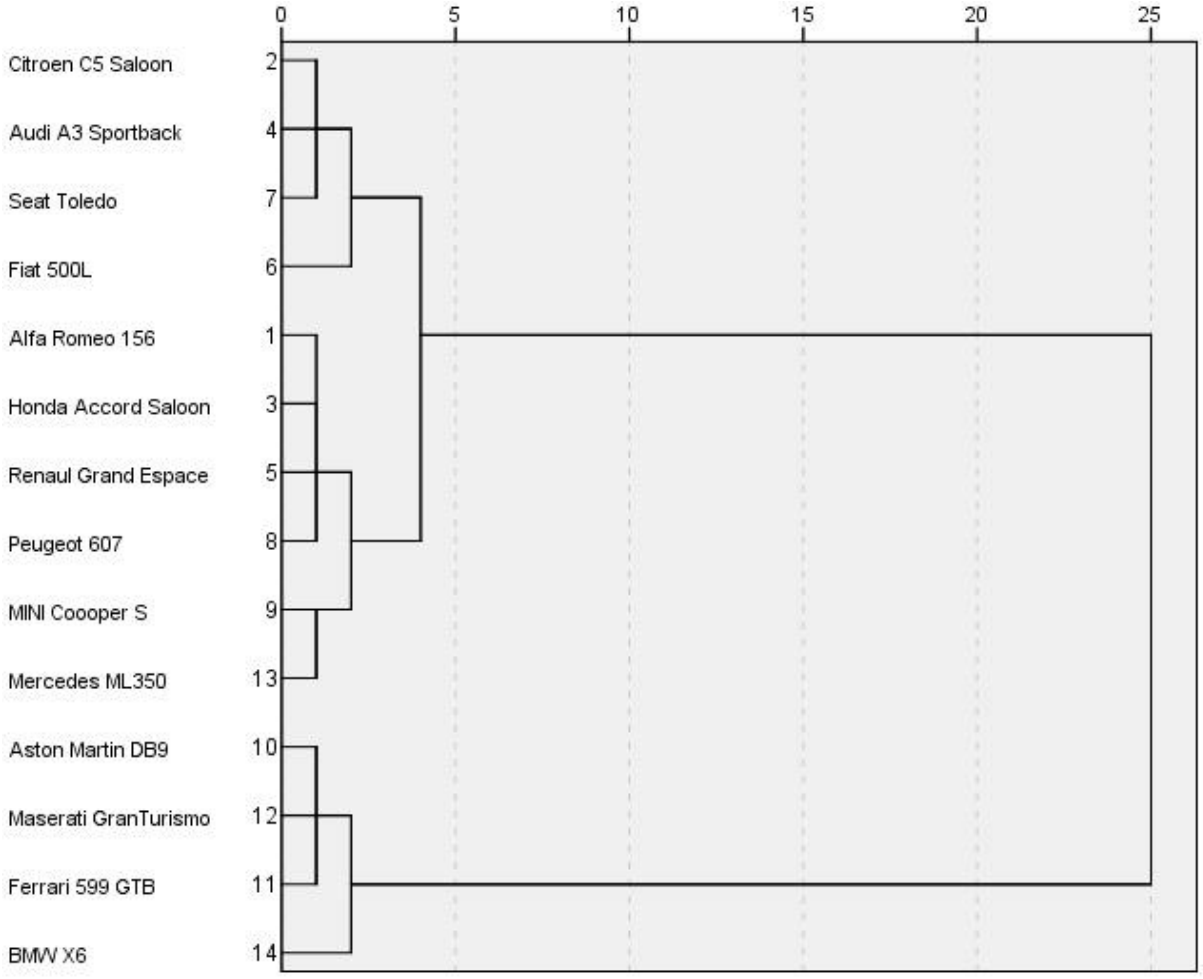
Tabela 5.8: Matrica sličnosti za 14 modela automobila

U prvom slučaju modeli automobila će biti grupisani po kubikaži, ubrazanju od 0-100km/h, potrošnji (milja/galon) i emisiji ugljen-dioksida. Takođe, korisno je videti kojoj grupi će neki model pripasti, u zavisnosti od broja grupa(u našem slučaju 2, 3, 4). Prvo će biti korišćen Ward-ov metod. Takođe, kao i do sada u radu, vrednosti su standardizovane kao Z vrednosti.

Model	4 Clusters	3 Klastera	2 Clusters
1:Alfa Romeo 156	1	1	1
2:Citroen C5 Saloon	2	2	1
3:Honda Accord Saloon	1	1	1
4:Audi A3 Sportback	2	2	1
5:Renaul Grand Espace	1	1	1
6:Fiat 500L	3	2	1
7:Seat Toledo	2	2	1
8:Peugeot 607	1	1	1
9:MINI Coooper S	1	1	1
10:Aston Martin DB9	4	3	2
11:Ferrari 599 GTB	4	3	2
12:Maserati GranTurismo	4	3	2
13:Mercedes ML350	1	1	1
14:BMW X6	4	3	2

Tabela 5.9: Podela modela automobila po klasterima

U gornjoj tabeli jasno se vidi kojoj grupi pripada koji model, za unesene parametre u zavisnosti od broja grupa.



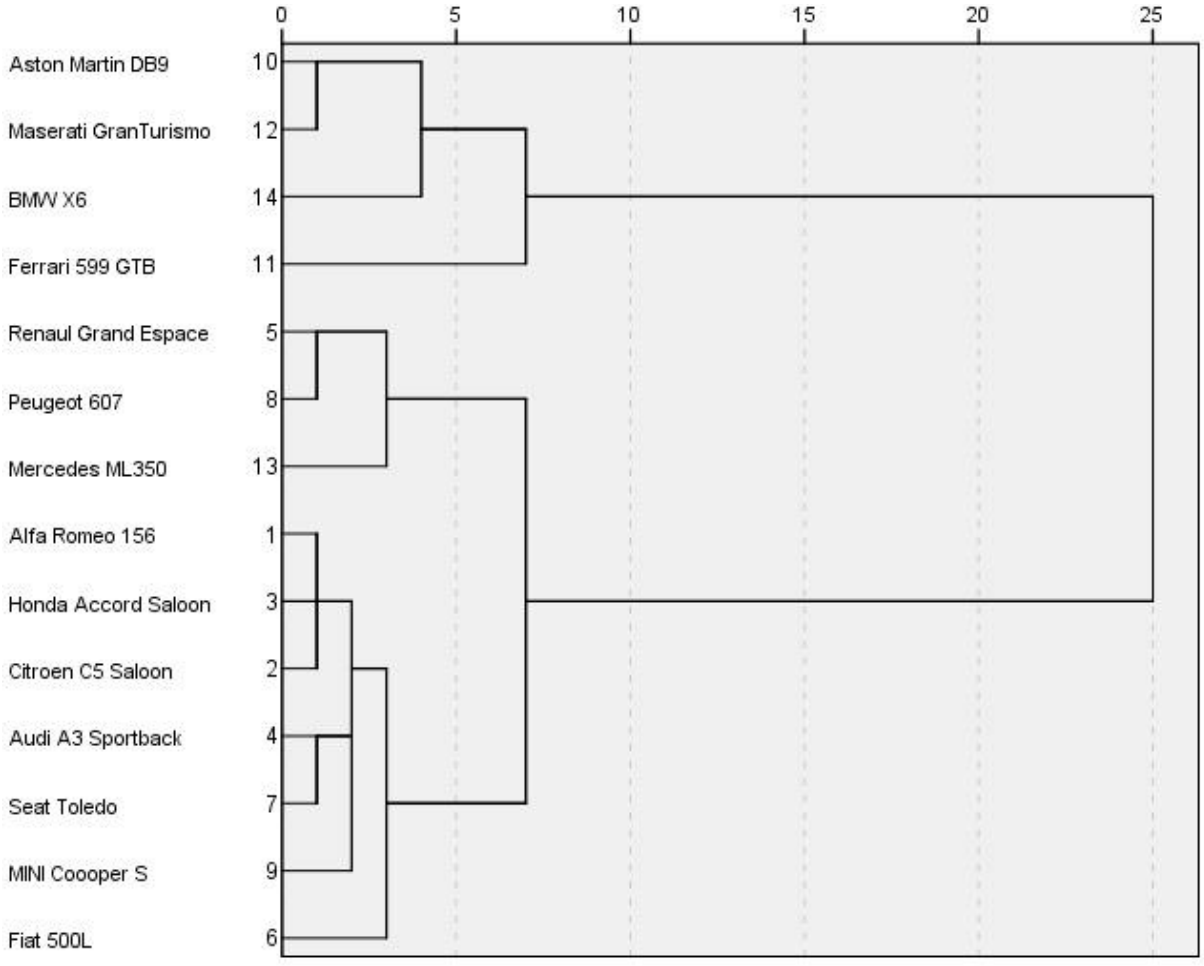
Slika 5.10: Dendogram za 14 modela automobila

Iz dobijenog dendograma, vidimo da bismo mogli da imamo i 6 klastera. Takođe vidi se, za svaki klaster u kojoj iteraciji je nastao. Dodatno, vidi se i koji klaster je sličniji kojem. Na primer, grupa u kojoj su Citroen C5, Audi A3 i Seat Toledo, sličnija je grupi u kojoj se nalazi Fiat 500L, nego recimo grupi u kojoj su Mini Cooper S i Mercedes ML, i svim ostalim grupama takođe.

Sada će biti urađen isti primer, samo što će biti uneseni drugi parametri za poređenje da vidimo da li će doći do promene samih klastera. U ovom slučaju biće korišćeni snaga motora(konjske snage), maksimalna brzina (mp/h) i zapremina rezervoara (l). Osim drugačijih parametara, biće korišćena i druga metoda povezivanja, metoda centroida.

Model	4 Klastera	3 Klastera	2 Klastera
1:Alfa Romeo 156	1	1	1
2:Citroen C5 Saloon	1	1	1
3:Honda Accord Saloon	1	1	1
4:Audi A3 Sportback	1	1	1
5:Renaul Grand Espace	2	2	1
6:Fiat 500L	1	1	1
7:Seat Toledo	1	1	1
8:Peugeot 607	2	2	1
9:MINI Coooper S	1	1	1
10:Aston Martin DB9	3	3	2
11:Ferrari 599 GTB	4	3	2
12:Maserati GranTurismo	3	3	2
13:Mercedes ML350	2	2	1
14:BMW X6	3	3	2

Tabela 5.10: Podela modela automobila po klasterima



Slika 5.11: Dendogram sa centroid metodom povezivanja

Kao što je bilo očekivano, rezulati se u ovom primeru razlikuju od prethodnog primera. Aston Marin i Maserati su u prvom koraku svrstani u istu grupu, dok su recimo Alfa Romeo, Honda i Citroen u zajedničkoj grupi.

Zaključak

Na osnovu do sada pokazanog, jasno je da obe tehnike, i multidimenzionalno skaliranje i klaster analiza, imaju veoma široku primenu u mnogim naučnim oblastima. Multidimenzionalno skaliranje (MDS) je istraživačka tehnika koje se koristi u testiranju hipoteze o postojanju broja dimenzija za određen skup podataka. Korišćenje MDS u analizi podataka nudi nekoliko prednosti. MDS slika strukturu skupa podataka koji aproksimiraju distance između parova objekata. U radu je pokazano kako funkcionišu klasično multidimenzionalno skaliranje i ne-metričko multidimenzionalno skaliranje, kao i razlike u dvema tehnikama.

Treba napomenuti dve važne stavke kod MDS. Prva je, da ose (x-osa, i y-osa) kao takve, nemaju značaja, a druga je, da je orijentacija slike proizvoljna. Ovo je pokazano u primeru sa gradovima Srbije, gde mapa ne mora nužno da bude okrenuta tako da je sever gore i istok desno. Jedino što je bitno, je koja tačka se nalazi u blizini neke druge tačke.

Prilikom samog rada na MDS-u i interpretaciji slike, bitne su dimenzija i klasteri. Veoma je važno primetiti, da stvaran broj dimenzija ili atributa, ne mora da odgovara matematičkom broj dimenzija (osa) koje definišu vektorski prostor (MDS mapu). U radu su prikazane najčešće greške koje se javljaju prilikom sprovođenja analize i načini za prevazilaženje istih.

Interpretacija multidimenzionalnog skaliranja može biti različita, što implicira, da dva istraživača ne moraju da dobije iste rezultate.

Tehnike klaster analize se koriste kako bi se pronašli klasteri tj. grupe, u a priori ne klasifikovanom skupu multivarijantnih podataka. Iako su tehnike klaster analize veoma korisne za samu analizu podataka, zahtevaju pažnju u primeni ukoliko želimo da izbegnemo pogrešna rešenja. Razvijene su mnoge metode klaster analize i brojna istraživanja su pokazala da ne postoji najbolja metoda, već to isključivo zavisi od toga šta želimo da dobijemo, tj. pokažemo.

Da bi se uspešno primenjivale obe tehnike, MDS i klaster analiza, potrebno je dobro proučiti podatake iz kojih će biti izvučen zaključak i smer u kojem će sama analiza ići, kako bi odgovarajući model bio primenjen.

Literatura

[1] Everitt B., Hothorn T. (2012), An Introduction to Applied Multivariate Analysis with R, Springer
[2] Everitt B., Rabe-Hesketh S. (1997), The Analysis of Proximity Data. London: Chapman and Hall/CRC
[3] Kruskal J., Wish M. (1977), Multidimensional Scaling
[4]Everit B. (2005), An R and S-Plus Companion to Multivariate Analysis, Springer
[5]Lozanov-Crvenković Z. (2011), Statistika, Novi Sad
[6]Patric L. Odel and Benjamin S. Duran, (1974), Cluster Analysis: A Survey, Springer
[7] Borg I.,Groenen P.J.F., Mair P. (2013), Applied Multidimensional Scaling, Springer
[8] Chambers J. M., Cleveland W. S., Kleiner B. and Tukey P. A. (1983), Graphical Methods for Data Analysis, London, UK: Chapman & Hall/CRC
[9] Izenman A. J. (2008), Modern Multivariate Techniques, New York: Springer-Verlag
[10] Kaufman, L. and Rousseeuw, P. J. (1990), Finding Groups in Data: An Introduction to Cluster Analysis, New York: John Wiley & Sons
[11] Krzanowski, W. J. (1988), Principles of Multivariate Analysis, Oxford, UK: Oxford University Press
[12] Borg I., Groenen P. J. F. (2005), Modern multidimensional scaling, New York: Springer.
[13] Everitt B., Landau S., Leese M. (2001), Cluster Analysis, Fourth edition, Arnold.
[14] Rencher A.C. (2002), Methods of Multivariate Analysis, Second edition, Wiley.
[15] https://en.wikipedia.org/wiki/Cluster_analysis
[16] https://www-users.cs.umn.edu/~kumar/dmbook/ch8.pdf
[17] http://www.stat.columbia.edu/~madigan/W2025/notes/clustering.pdf

Biografija

Miloš Banjanin rođen je u 27.06.1990. u Novom Sadu. Osnovnu školu „Žarko Zrenjanin“ završio je 2005. godine., kao nosilac Vukove diplome. Iste godine upisuje Gimnaziju „Svetozar Marković“ u Novom Sadu, koju završava 2009. godine. Po završetku gimnazije, iste godine upisao je osnovne studije na Prirodno – matematičkom fakultetu, u Novom Sadu, smer matematika finansija. Osnovne studije završava 2013. godine i upisuje master studije na istom usmerenju. Zaključno sa septembarskim rokom 2015. godine, položio je sve ispite predviđene nastavnim planom i programom i stekao uslov za odbranu master rada.

Novi Sad, 2016. godine Miloš Banjanin

UNIVERZITET U NOVOM SADU
PRIRODNO – MATEMATIČKI FAKULTET
KLJUČNA DOKUMENTACIJSKA INFORMACIJA

Redni broj:
RBR

Identifikacioni broj:
IBR

Tip dokumentacije: Monografska dokumentacija
TD

Tip zapisa: Tekstualni štampani materijal
TZ

Vrsta rada: Master rad
VR

Autor: Miloš Banjanin
AU

Mentor: dr Zagorka Lozanov-Crvenković
MN

Naslov rada: Analiza sličnosti podataka
NR

Jezik publikacije: srpski (latinica)
JP

Jezik izvoda: srpski/engleski
JI

Zemlja publikovanja: Republika Srbija
ZP

Uže geografsko poreklo: Vojvodina
UGP

Godina: 2016
GO

Izdavač: Autorski reprint
IZ

Mesto i adresa: Novi Sad, Departman za matematiku i informatiku, Prirodno-matematički fakultet, Univerzitet u Novom Sadu, Trg Dositeja Obradovića
MA

Fizički opis rada: (5/70/8/16/27/0)
(broj poglavlja/broj strana/broj citata/broj tabela/broj grafika/broj priloga)
FO

Naučna oblast: Matematika
NO

Naučna disciplina: Statistika
ND

Predmetna odrednica/ Ključne reči: Multidimenzionalno skaliranje, Klaster analiza, Euklidska distanca, matrica sličnosti, bliskost podataka, Stres, fit, razlike, sličnosti
PO

UDK

Čuva se: Biblioteka departamana za matematiku i informatiku, Prirodno-matematički fakultet, Univerzitet u Novom Sadu
ČU

Važna napomena:
VN

Izvod:
Ovaj master rad se bavi analizom sličnosti podataka. U radu su izložene dve tehnike, multidimenzionalno skaliranje i klaster analiza, pomoću kojih se može analizirati bliskost podataka. U prvom delu rada obrađeno je multidimenzionalno skaliranje, jedna od multivarijantnih statističkih tehnika pomoću koje se pronalazi skup tačaka malih dimenzija koji najbolje aproksimira visoko dimenzionalnu konfiguraciju podataka, predstavljenu početnom matricom bliskosti. Multidimenzionalno skaliranje otkriva strukturu skupa podataka, crtajući tačke u jednoj, dve ili tri dimenzije, ukoliko je to moguće. Na početku poglavlja je izloženo kako se dobija matrica bliskosti iz skupa podataka, kao i matematički pristup multidimenzionalnom skaliranju. Nastavak poglavlja je posvećen dvema metodama, klasičnom multidimenzionalnom skaliranju i ne-metričkom multidimenzionalnom skaliranju. Kroz nekoliko primera pojašnjeno je kako to funkcioniše u praksi. U drugom delu rada obrađena je klaster analiza. U ovom radu objašnjeni su pojmovi odstojanja, kao što su Euklidsko, Menhetn, odstojanje Minkovskog. Navedene su neke tehinke klaster analize: hijerarhijske metode i nehijerarskijske metode (metoda k-sredina). Na kraju rada, obrađeni su neki primeri, i sumirani rezultati celog rada.
IZ

Datum prihvatanja teme od strane NN veća: 19.Februar 2016.
DP

Datum odbrane:
DO

Članovi komisije:
KO

Predsednik: dr Ljiljana Gajić, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

Mentor: dr Zagorka Lozanov-Crvenković, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

Član: dr Ivana Štajner-Papuga, redovni profesor, Prirodno-matematički fakultet, Univerzitet u Novom Sadu

UNIVERSITY OF NOVI SAD
FACULTY OF SCIENCES
KEY WORDS DOCUMENTATION

Accession number:
ANO

Identification number:
INO

Document type: Monograph type
DT

Type of record: Printed text
TR

Contents code: Master's thesis
CC

Author: Miloš Banjanin
AU

Mentor: Zagorka Lozanov-Crvenković, PhD
MN

Title: Analysis of Proximity Data
TI

Language of text: Serbian
LT

Language of abstract: Serbian/English
LA

Country of publication: Serbia
CP

Locality of publication: Vojvodina
LP

Publication year:2015
PY

Publisher: Author's reprint
PU

Publication place: Novi Sad, Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad, Dositeja Obradovića Sq.4
PP

Physical description: (5/70/8/16/27/0)
(number of sections/ pages/ tables/ pictures/ graphs/appendices)
PD

Scientific field: Mathematics
SF

Scientific discipline: Statistics
SD

Subject/Key words: Multidimensional Scaling, Cluster Analysis, Euclidian distance, proximity matrix, closeness of data, Stress, fit, dissimilarities, similarities
SKW

UC

Holding data: The Library of the Department of Mathematics and Informatics, Faculty of Sciences, University of Novi Sad
HD

Note:
N

Abstract:
This thesis deals with analysis of proximity data. Two techniques are presented, multidimensional scaling and cluster analysis, by which one can analize proximity of data. In the beginning of the paper it was spoken about multidimensional scaling, one of the statistical techniques which is used to find a set of low dimensional points that best approximates high dimensional data configuration, represented by proximity matrix. Multidimensional scaling reveals structure of a data set, plotting points in one, two or three dimensions, if possible. In the beginnig of the chapter it was showed how to get proximity matrix, and mathematical approach as well. Later on, one can read about classical multidimensional scaling and ordinal multidimensional scaling. Through few examples it is shown how all that work in practice. Second part of the paper is based on cluster analysis. It is explained what is distance, Euclidian, Manhattan, and Minkowski distance too. Furthermore, special attention of thesis was on hierarchical and non-hierarchical clustering methods. At the end, few examples were showed, and the results of paper were summarized.
AB

Accepted by the Scientific Board on: 19th February, 2016.
ASB

Defended:

DE

Thesis defend board:
DB

President: Ljiljana Gajić, Phd, full professor, Faculty of Sciences, University of Novi Sad

Mentor: Zagorka Lozanov-Crvenković, Phd, full professor, Faculty of Sciences, University of Novi Sad

Member: Ivana Štajner-Papuga, Phd, full professor, Faculty of Sciences, University of Novi Sad