Razvrščanje v skupine

Domača naloga 2 - Multivariatna analiza

Urh Peček in Alen Kahteran

20. 3. 2021

Kazalo

Slike

Tabele

Cilji naloge

Cilj naloge je razvrstiti enote v skupine tako, da si bodo enote znotraj skupin čim bolj podobne in enote v različnih skupinah čim bolj različne glede na več spremenljivk. Nato bodo ustvarjene skupine glede na izbrani spremenljivki tudi opisane.

Hierarhično razvrščanje

Pri hierarhičnem razvrščanju se enote najpogosteje združujejo. Začnemo s tem, da je vsaka enota v svoji skupini. Sledeče se na vsakem koraku, na podlagi izračunane matrike različnosti v kateri so zapisane razdalje med pari skupin, združita skupini, ki sta si najbližji ter se izračunajo različnosti nove združene skupine do ostalih skupin. Postopek se konča ko so vse enote v eni skupini.

Pred hierarhičnim razvrščanjem v skupine vrednosti spremenljivk vedno standariziramo. Tako bodo imele spremenljivke povprečje 0 in standardni odklon 1. S tem dosežemo enakovreden vpliv spremenljivk na razvrstitev. To je pomembno predvsem v primeru obravnave spremenljivk različnih merskih lestvic. Kot mero različnosti bomo upoštevali kvadrirano evklidsko razdaljo.

Št. skupin na podlagi dendograma

Število skupin lahko določimo na podlagi dendograma, ki grafično prikazuje potek združevanja v skupine in dobra lastnost hierarhičnega razvrščanja je ta, da uporabniku ni potrebno vnaprej določiti števila skupin. Je pa določanje števila skupin na podlagi dendograma je do neke mere subjektivno, saj število skupin določimo tako, da pogledamo, kdaj se začnejo razdalje med skupinami pri združevanju občutneje manjšati.

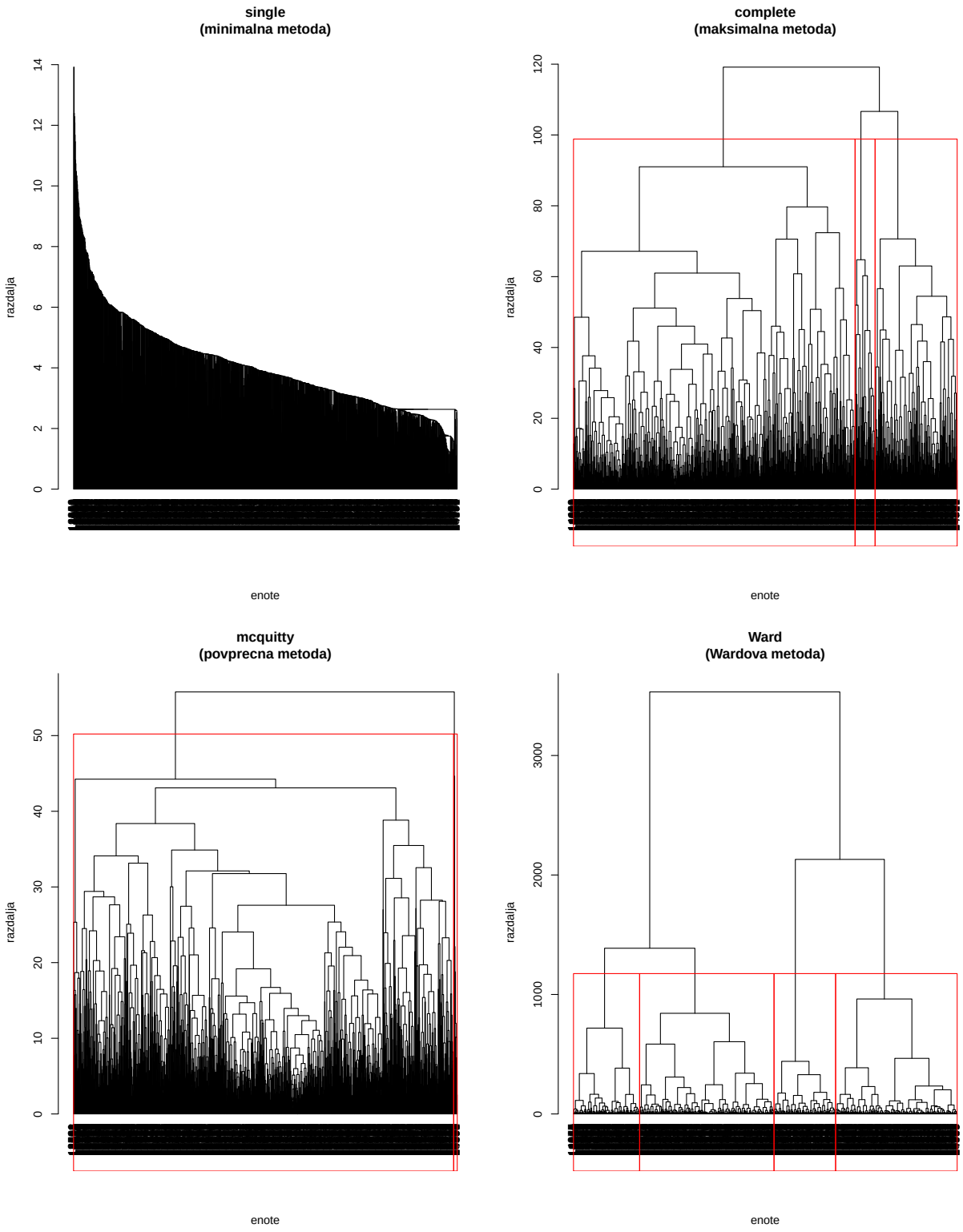
Minimalna metoda (enojna povezanost - single linkage) je primerna za dolge in neeliptične skupine, ki so jasno ločene med seboj. Kadar skupine med seboj niso jasno ločene pri minimalni metodi pride do problema veriženja, kar se pri nas zgodi. Na takem dendogramu ne moremo določiti števila skupin in zato rečemo, da je skupina zgolj ena.

Maksimalna metoda (polna povezanost - complete linkage) je primerna za okrogle skupine. Na podlagi maksimalne metode se odločimo za 3 skupine, lahko pa bi se tudi za 2 ali 4.

Na podlagi McQuittyjeve metode oziroma “povprečne” povezanosti se odločimo za zgolj 2 skupini.

Wardova metoda je primera za eliptične skupine in na podatkih podobnim našim, da Wardova metoda ponavadi najboljše rezultate. Odločimo se za 4 skupine (lahko bi se tudi za 3), saj je vsebinska interpretacija bolj zanimiva kot za 2 skupini, za kar bi se teoretično morali odločiti.

Na koncu se na podlagi Wardovega dendograma ter tudi upoštevanja ostalih metod odločimo za 4 skupine.



Slika 1: Dendogrami različnih metod razvrščanja v skupine

Izbor razvrstitve na podlagi Wardove kriterijske funkcije

Problem razvrščanja v skupine lahko definiramo kot iskanje razbitja, ki minimizira vrednost kriterijske funkcije. Za vsako od zgornjih razvrstitev, kjer smo se za število skupin pri vsaki metodi odločili na podlagi dendograma, izračunamo vrednost Wardove kriterijske funkcije. Za vsako metodo bomo primerjali tudi različno število skupin. Velja upoštevati, da vrednost Wardove kriterijske funkcije z naraščanjem števila skupin pada, zato z različnim številom skupin vrednosti Wardove kriterijske funkcije niso primerljive.

Za vsako metodo, brez minimalne, kjer pride do problema veriženja si oglejmo vrednost Wardove kriterijske funkcije za različno število skupin.

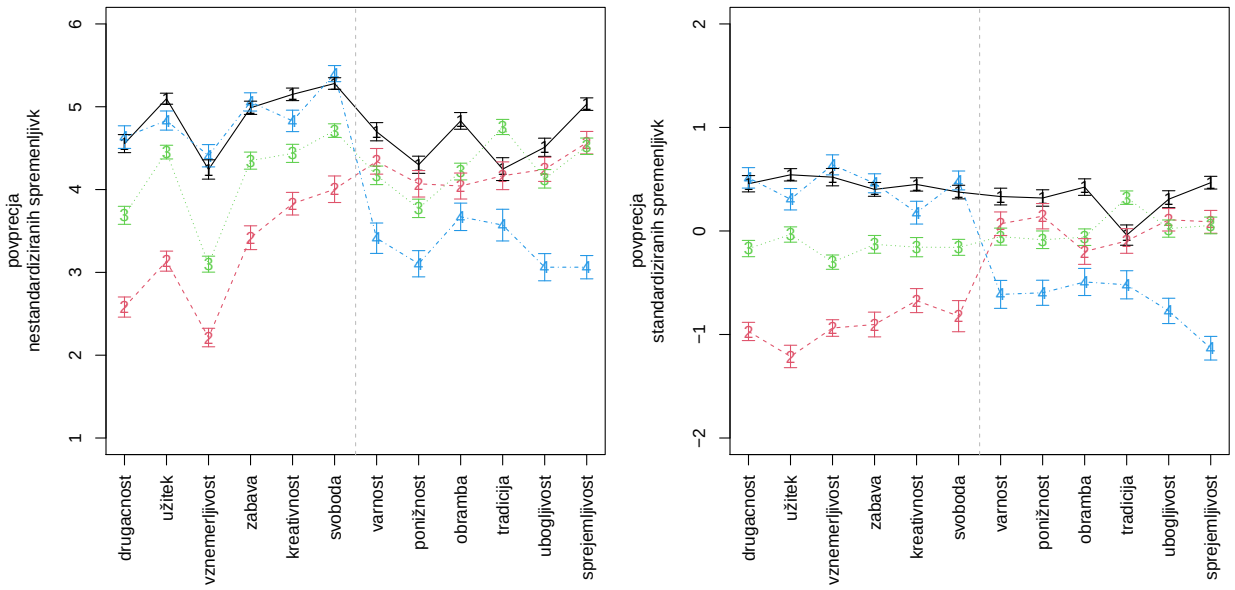
Tabela 1: Vrednost Wardove kriterijske funkcije glede na metodo in število skupin

	število skupin = 2	število skupin = 3	število skupin = 4
maksimalna metoda	16262.09	15961.02	14890.49
povprečna metoda	17811.75	17758.13	17490.97
Wardova metoda	16173.98	15108.69	14415.17

Wardova metoda ima pri dveh skupinah nižjo (boljšo) vrednost kot povprečna metoda pri treh oz. pri štirih skupinah. Pri vseh številih skupin ima Wardova metoda nižjo vrednost tudi od maksimalne metode. Tako se za najboljšo metodo in število skupin, tudi z upoštevanjem dendograma in vsebinske zanimivosti, odločimo za Wardovo metodo s štirimi skupinami.

Povprečja odgovorov po skupinah

Za izbrano razvrstitev tj. Wardovo metodo s štirimi skupinami prikažimo povprečja odgovorov po skupinah na standariziranih in nestandariziranih podatkih.



Slika 2: Povprečja po skupinah za Wardovo metodo s štirimi skupinami. Levo - nestandardizirane vrednosti, desno - standardizirane vrednosti

Če si ogledamo standarizirane vrednosti spremenljivk, vidimo, da skupina 3 odraža povprečje vsebinskih spremenljivk, morda zgolj pri spremenljivki tradicija zavzame nekoliko višjo in pri vznemerljivosti nekoliko nižjo vrednost. Tako jo lahko vzamemo za nekakšno primerjavo. Sledeče poglejmo skupino 4. Ta pri vsebinskih spremenljivkah, ki odražajo odprtost zavzame nadpovprečne vrednosti, medtem ko je pri spremenljivkah konservativnosti podpovprečna. Ravno obratno sliko pa vidimo pri skupini 2, ki ima pri spremenljivki odprtosti podpovprečne, pri konservativnosti pa povprečne oziroma nadpovprečne vrednosti pri varnosti in ponižnosti. Zadnja obravnavana skupina, skupina 1, je pri vseh spremenljivkah, neupoštevajoč tradicije, nadpovprečna, pri tradiciji pa zavzame povprečno vrednost. Pri nestandariziranih vrednostih spremenljivk lahko dodamo zgolj to, da zgoraj omenjena nadpovprečna vrednost niha nekje okoli 5, povprečje okoli vrednosti 4, pri podpovprečni vrednosti pa razumemo okolico vrednosti 3.

Nehierarhično razvrščanje

Metoda voditeljev (k-means)

K-means je posebna metoda metode voditeljev, širše nehierarhičnega razvrščanja. Voditelji so neke vrste predstavniki skupin, vsaka enota pa pripada skupini, kateremu voditelju je najbližje oz. mu je najbolj podobna. Pogoj metode k-means je, da so spremenljivke vsaj intervalne. Kot razdalja se predpostavi evklidska razdalja, voditelje pa predstavljajo povprečja skupin.

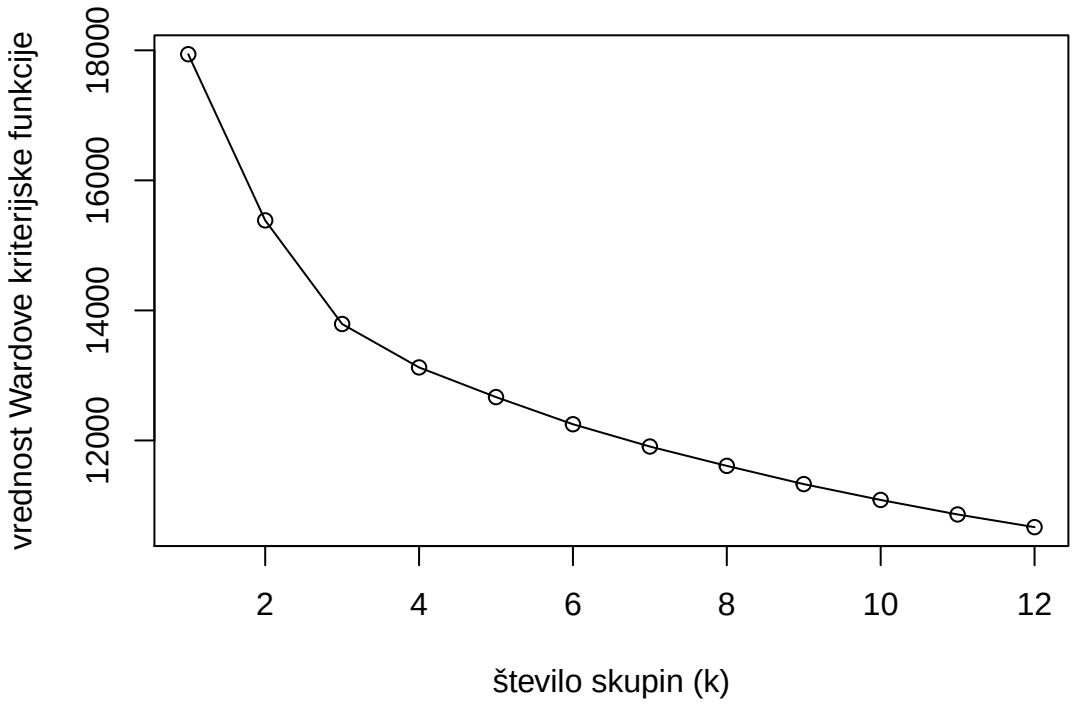
Slaba stran metode voditeljev oziroma k-means je ta, da mora biti število skupin podano vnaprej. Torej imamo na začetku določene voditelje, nato pa na vsakem koraku vsako enoto priredimo voditelju (skupini) kateremu je najbljižja (evklidska razdalja). Na vsakem koraku se tako izračunajo novi voditelji kot povprečja skupin in postopek se zaključi, ko se voditelji ustalijo, torej so stari enaki novim. Za rešitev izberemo tisto razvrstitev, ki ima najmanjšo vrednost Wardove kriterijske funkcije. Postopek običajno večkrat ponovimo, saj za različne začetne voditelje lahko dobimo različne rešitve, torej razvrstitve v skupine.

V našem primeru, bomo upoštevali k-means pristop s standariziranimi podatki.

Scree-diagram

Poglejmo si vrednosti Wardove kriterijske funkcije pri različnem številu skupin. Velja upoštevati, da vrednost Wardove kriterijske funkcije pada z naraščanjem števila skupin, zato za optimalno število skupin velja vrednost k, kjer se "zgodi" koleno. Če koleno ni jasno razvidno je to lahko indikator, da skupine niso jasno ločene.

Na podlagi k-means pristopa in Wardove kriterijske funkcije se odločimo za 3 skupine, kjer je "koleno" najbolj razvidno.



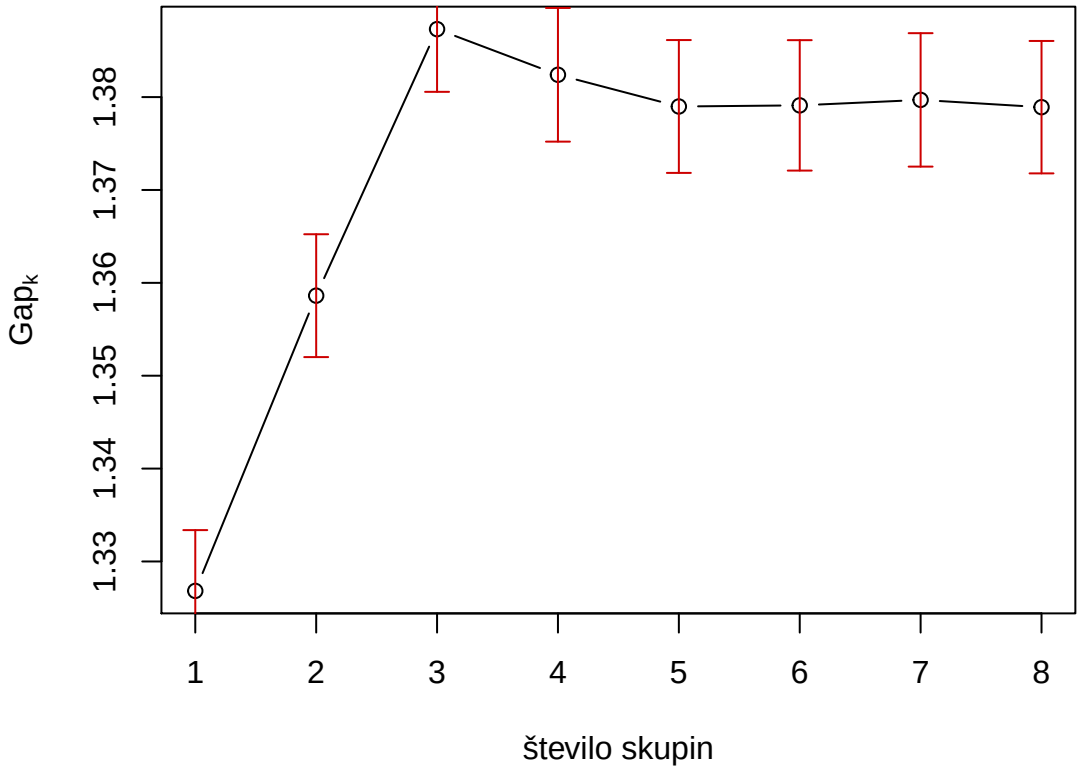
Slika 3: Vrednost Wardove kriterijske funkcije

GAP statistika

Pri določevanju števila skupin si lahko pomagamo tudi z GAP statistiko. V tem primeru iščemo skupine, ki so bolj homogene kot bi jih našli v podatkih brez skupin. Torej primerjamo razdalje znotraj skupin z razdaljami kot bi jih pričakovali glede na referenčne podatke, torej na podatkih brez skupin (iz verzije enakomerne porazdelitve).

Izberemo število skupin, kjer je razlika med opaženimi in referenčnimi podatki največja. Natančneje, izberemo najmanjše število skupin k kjer je vrednost GAP(k) vsaj tako velika kot GAP(k+1) - SE(GAP(k+1)), kjer SE označuje standardno napako GAP statistike.

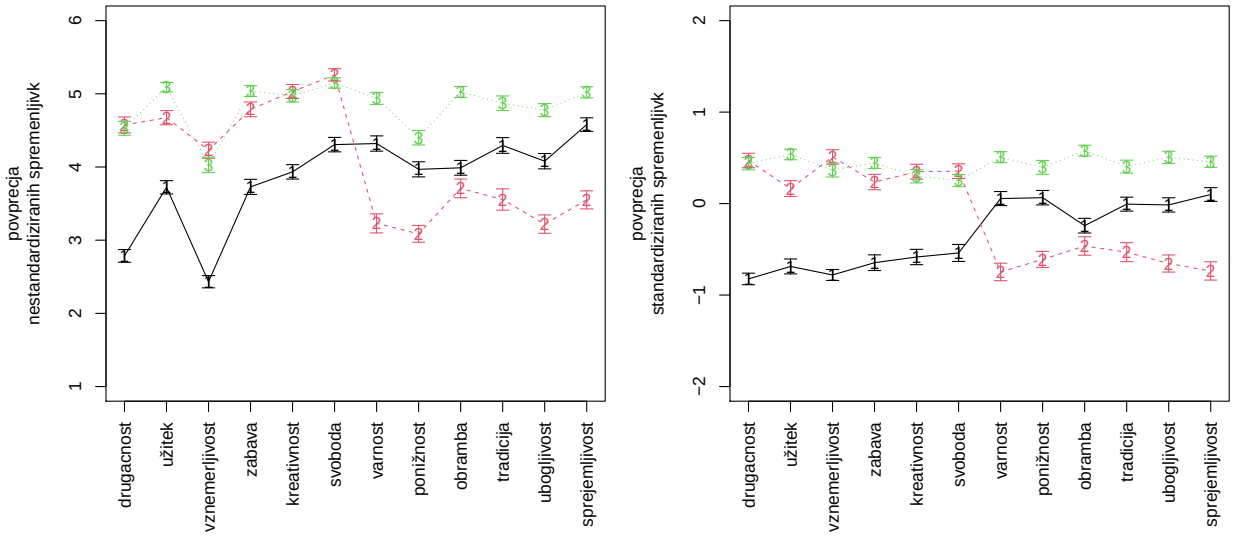
Na podlagi spodnjega grafičnega prikaza vrednosti GAP statistike pri različnem številu skupin se odločimo za 3 skupine.



Slika 4: Vrednost GAP statistike

Povprečja skupin glede na vsebinske spremenljivke

Tako na podlagi scree-diagrama kot GAP statistike smo se odločili za tri skupine. Sedaj pa si poglejmo povprečja skupin glede na nestandarizirane in standarizirane vrednosti vsebinskih spremenljivk.



Slika 5: Povprečja skupin glede na vsebinske spremenljivke

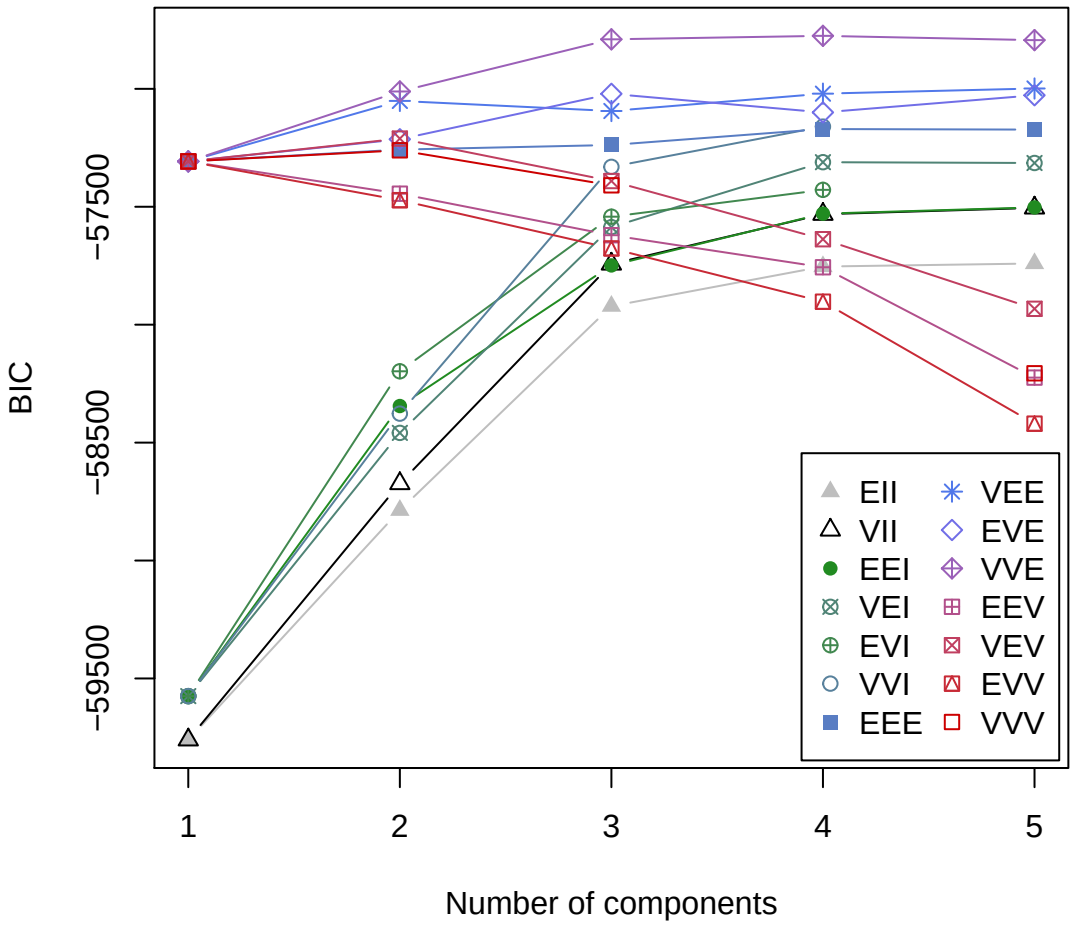
Na grafu povprečij standariziranih spremenljivk vidimo, da se največji skok na prehodu med odprtostjo in konservativnostjo zgodi pri skupini 2. Ta pri spremenljivkah, katere predstavlja odprtost zavzame nadpovprečne vrednosti, pri vseh konservativnih spremenljivka pa je vrednost strogo podpovprečna. Skupina 1 je nekako inverzna skupini 2, pri spremenljivkah odprtosti zavzame močno podpovprečne vrednosti, pri konservativnih spremenljivkah pa odraža relativno povprečje in najnižjo vrednost zavzame pri obrambi. Skupina 3 pri vseh spremenljivkah, tako odprtosti kot konservativnosti zavzame najvišje vrednosti. Za predstavo iz grafa nestandariziranih vrednosti vidimo, da kot nadpovprečno vrednost lahko upoštevamo nekje 4.5, podpovprečne vrednosti pa se nahajajo v okolici vrednosti 3.

Razvrščanje na podlagi modelov

Pri razvrščanju na podlagi modelov predpostavimo, da so naši podatki generirani iz neke mešanice multivariatnih normalnih porazdelitev z različnimi parametri oziroma komponentami. Vsaka skupina namreč prihaja iz svoje multivariatne normalne porazdelitve. Večjo kot ima skupina variabilnost, večja je po volumnu. Omejimo se lahko na posamezne modele, če imamo kakšne domneve o tem, kakšne naj bi skupine bile.

Z metodo, ki temelji na EM algoritmu ocenimo število skupin in parametre za vsako skupino ter kateri skupini posamezna enota pripada. V primeru da predpostavka o multivariatni normalni porazdelitvi drži, se v večini simulacij metoda izkaže kot optimalna. Razvrstitev načeloma naredimo na originalnih (nestandariziranih) podatkih, saj s tem dovoljujemo različno velikost skupin.

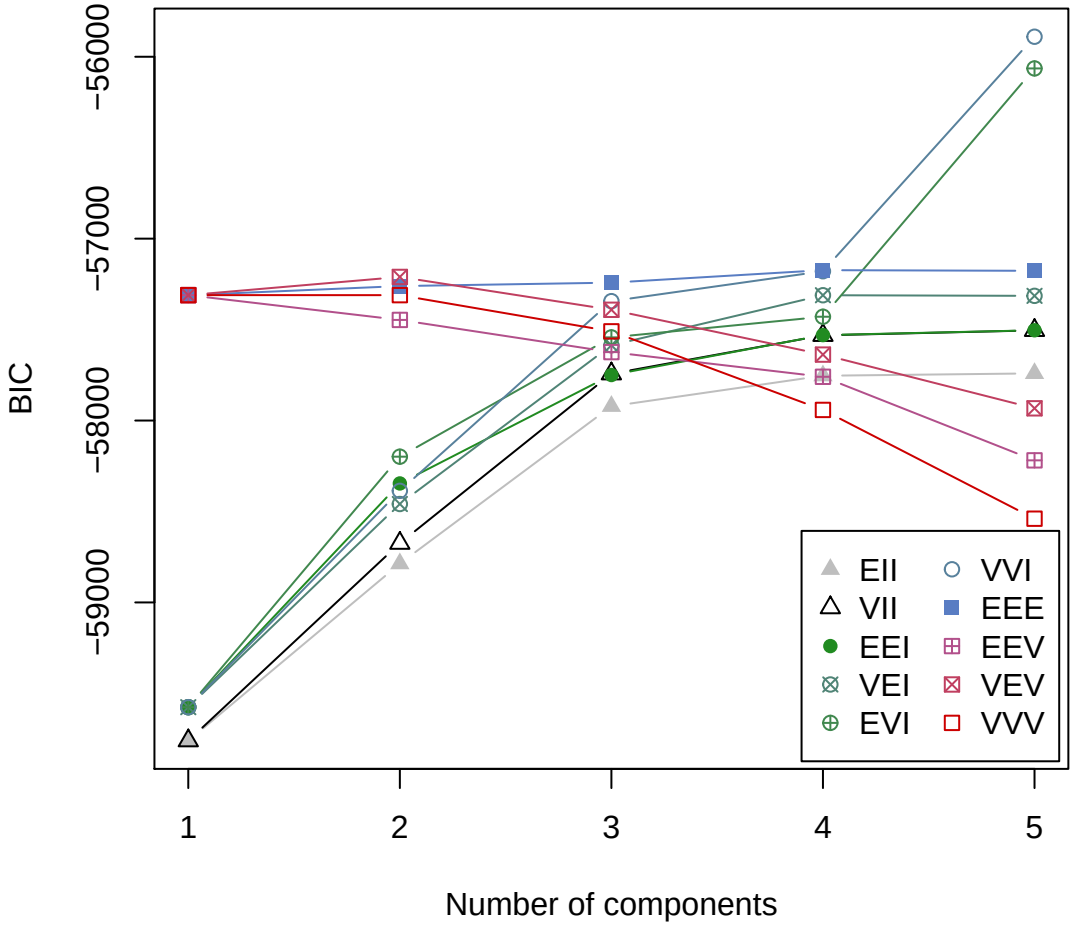
BIC kriterij na originalnih podatkih



Slika 6: BIC kriterij za originalne podatke

Na podlagi BIC kriterija (Bayesian Information Criterion), ki zavzame vrednost -56775.4 izberemo model VVE s štirimi skupinami, kar pomeni, da gre za elipsoidne skupine, ki so različno velike, različnih oblik, hkrati pa so enako usmerjene (od leve zgoraj proti desni spodaj).

Pri oceni modela lahko uporabimo tudi argument priorControl s čimer določimo apriorne verjetnosti, kar rezultira v bolj stabilnih ocenah, a lahko povzroči pristranskost.

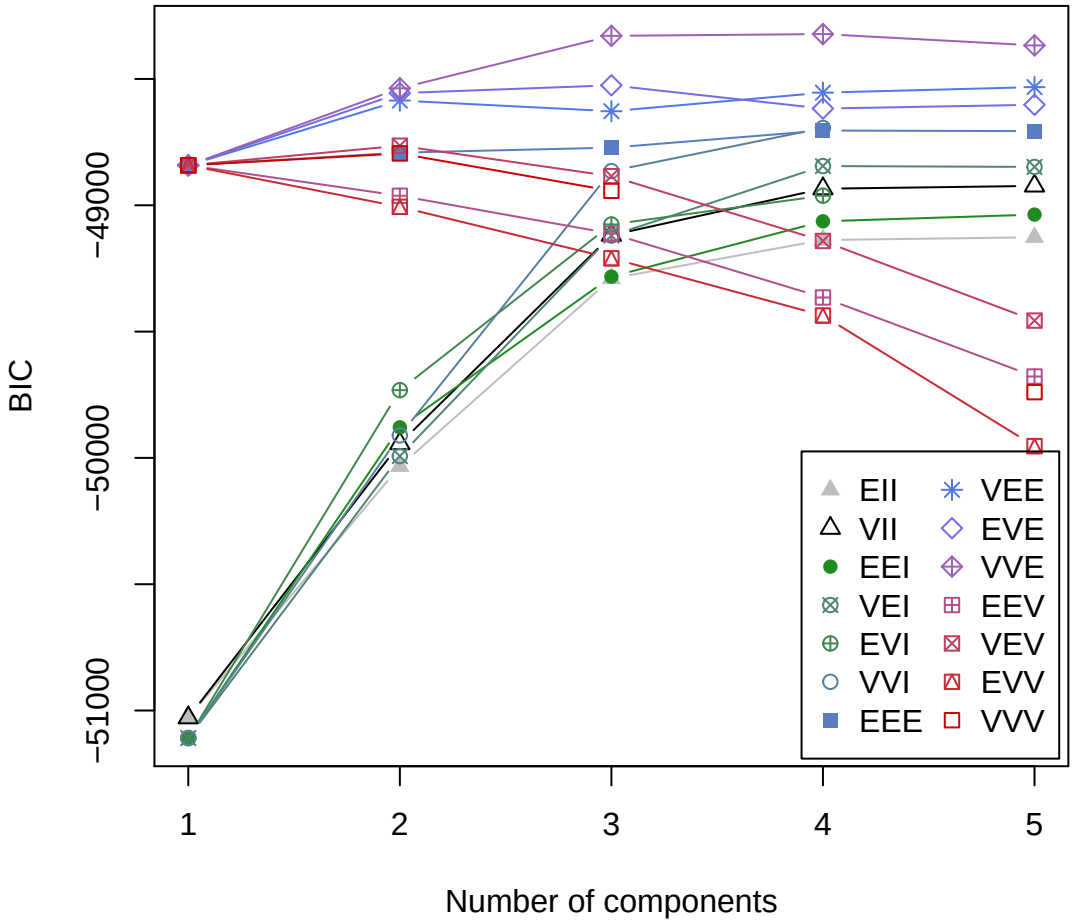


Slika 7: BIC kriterij (priorControl) za originalne podatke

Na podlagi BIC kriterija z uporabljenim argumentom se odločimo za model VVI s petimi skupinami, kar pomeni, da gre za različno velike skupine, različnih oblik in enake usmerjenosti (vzporedne abscisni osi).

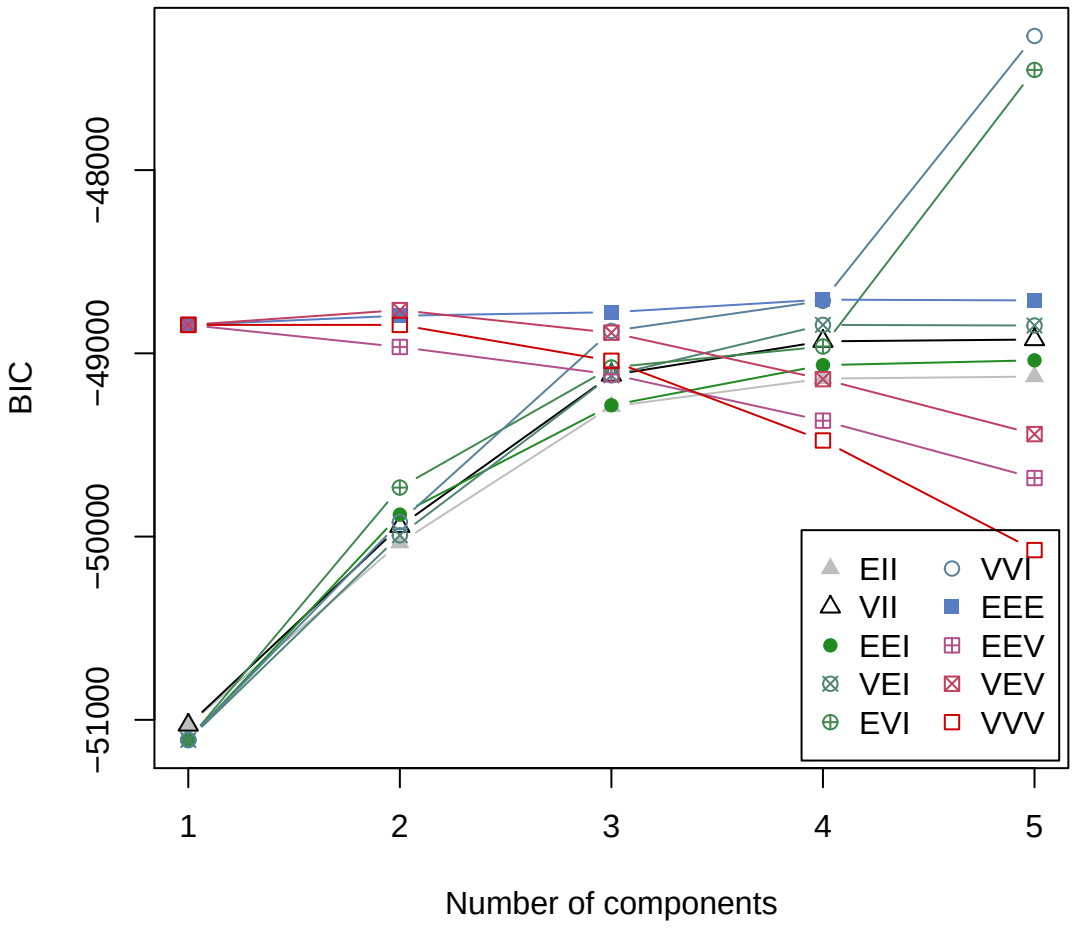
BIC kriterij na standariziranih podatkih

Poglejmo si še kako je z oceno modela na standariziranih podatkih. Po teoriji pri modelih, kjer porazdelitev ni okrogla (spherical), (se pravi pri oznaki modela na sredini ni "I"), ne bi smelo biti razlik. Velja opomniti, da vrednosti BIC kriterija niso primerljive med standariziranimi in nestandariziranimi podatki.



Slika 8: BIC kriterij za standarizirane podatke

Tudi pri oceni modela na standariziranih podatkih na podlagi BIC kriterija z vrednostjo -48322.36 izberemo model VVE s štirimi skupinami (lahko bi tudi 3).



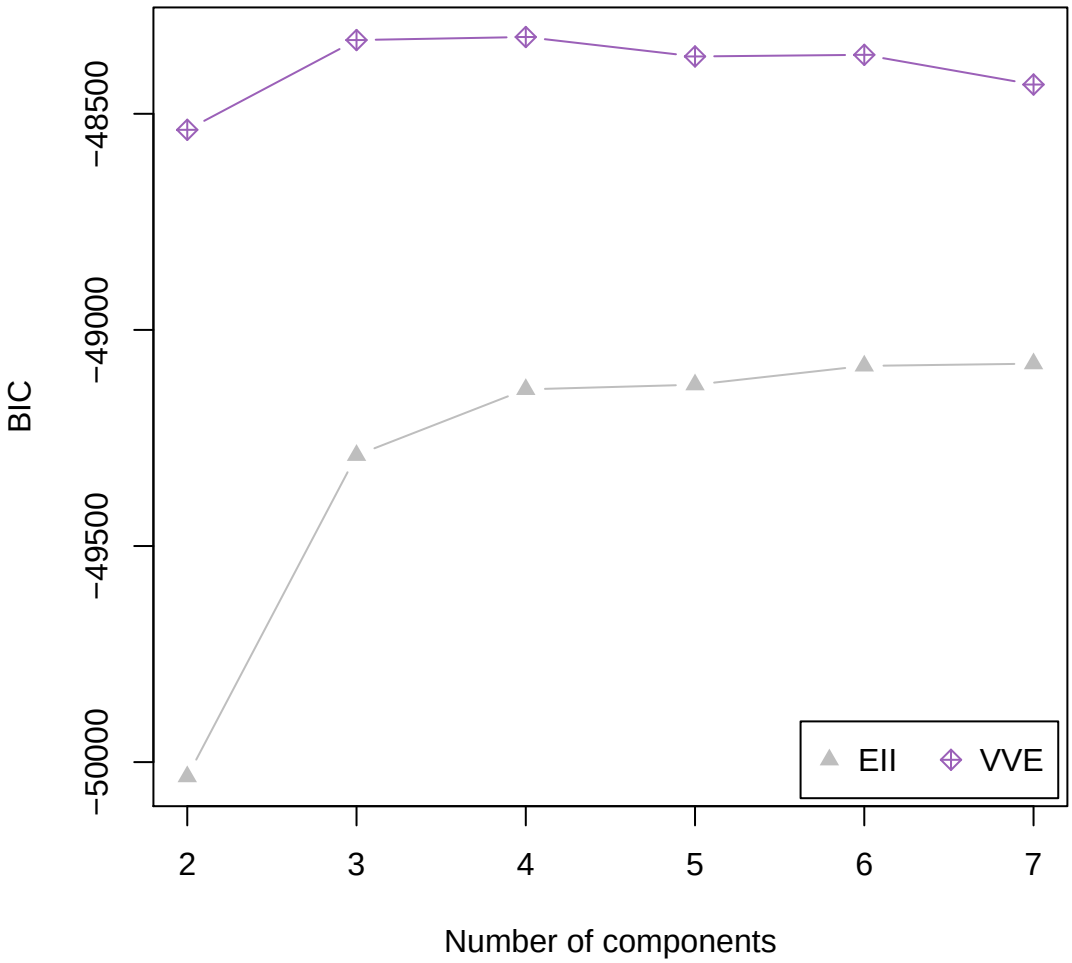
Slika 9: BIC kriterij (priorControl) za standarizirane podatke

Z določitvijo apriornih verjetnosti izberemo model VVI s petimi skupinami.

Na podlagi vseh štirih kriterijev se zaradi enostavnosti odločimo za model VVE s štirimi skupinami (različno velike elipsoidne skupine, ki so različnih oblik in enako usmerjene).

Primerjava modelov

Včasih se lahko omejimo tudi na posamezne modele. Na pogladi BIC kriterja in primerjave modelov EII (okrogle in enako velike skupine), ki je enakovreden metodi k-means in izbranega modela VVE se odločimo za model VVE. Poleg tega bi se pri obeh modelih odločili za tri skupine, saj vrednost BIC kriterija po treh skupinah narašča počasi.



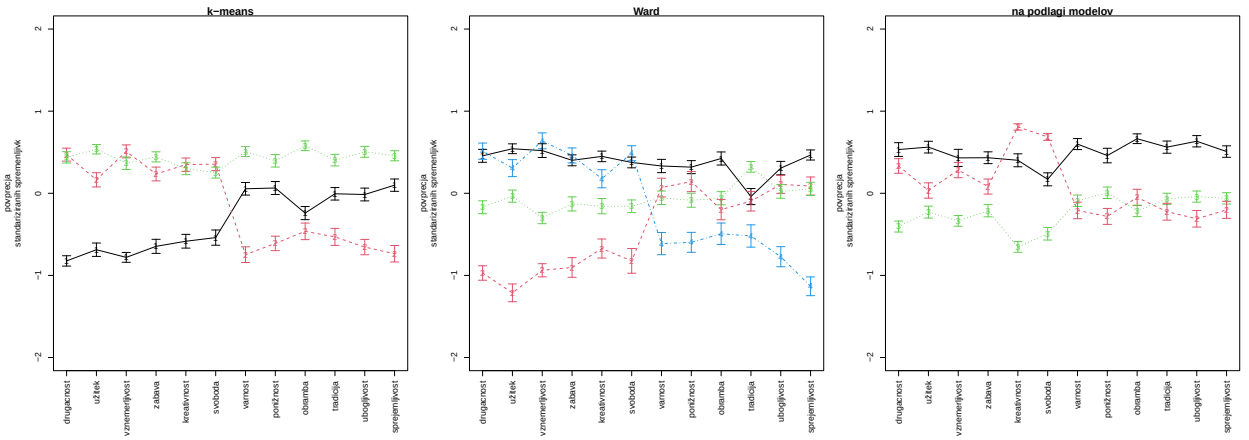
Slika 10: Primerjava VVE in EII modela

Primerjava razvrstitev in izbor

V sklopu primerjave razvrstitev je cilj ugotoviti kako podobne so si naše razvrstitve. V prejšnjih poglavjih smo iz hierarhične in nehierarhične metode ter metode na podlagi modelov izbrali najboljšo razvrstitev, sedaj pa izbrane razvrstitve primerjajmo.

Primerjava povprečij skupin

Ponovno izračunamo povprečja po skupinah in izbrane razvrstitve primerjajmo na standariziranih podatkih.



Slika 11: Primerjava razvrstitev na standariziranih podatkih

Vrstni red skupin v modelih je sicer drugačen, vendar pa sta si model izračunan na podlagi k-means in tisti na podlagi modelov nekoliko podobna, pri Wardovem modelu pa je dodana nekakšna povprečna skupina označena s številko 3. Podrobneje, skupina 3 pri k-means je podobna skupini 1 na podlagi modelov, do odstopanja pride zgolj pri spremenljivkah kreativnosti in svobode, ki pri metodi na podlagi modelov dosežeta nadpovprečni vrednosti. Poleg tega je skupina 1 pri k-means zelo podobna skupini 2 na podlagi modelov. Prav tako pa sta si nekoliko podobni tudi skupini 2 pri k-means in 3 na podlagi modelov, razlikujeta se zgolj v spremenljikah varnosti in ponižnosti, kjer k-means doseže občutno višje vrednosti. Tudi pri Wardovi metodi so skupine 1, 2 in 3 podobne pripadajočim skupinam pri k-means in na podlagi modelov, kot rečeno pa je dodana skupina 3, ki odraža povprečje.

Primerjava razvrstitev na podlagi Wardove kriterijske funkcije

Če si ogledamo katera razvrstitev je najboljša glede na vrednost Wardove kriterijske funkcije je to k-means razvrstitev s tremi skupinami in vrednostjo kriterijske funkcije 13791, medtem ko Wardova metoda s tremi skupinami zavzame vrednost 15109 in metoda na podlagi modelov z modelom VVE zavzame vrednost 15469.

Randov indeks in popravljen Randov indeks

Za ugotavljanje podobnosti dveh razvrstitev lahko uporabimo Randov indeks. Njegova vrednost predstavlja delež parov enot, ki so si v obeh razbitjih usklajeni, torej v obeh razbitjih v isti skupini ali pa v obeh razbitjih v različnih skupinah. Načeloma pa zaradi boljše primerljivosti vrednosti indeksa uporabljamo popravljen Randov indeks (ARI), popravljen za slučajnost. Pri ARI vrednost 1 pomeni identični razbitji, vrednost 0 pa, da sta si razbitji tako podobni, kot bi pričakovali po slučaju, torej večja kot je vrednost ARI bolj sta si razbitji podobni.

Vrednost popravljenega randovega indeksa pri primerjavi Wardove metode s tremi skupinami in k-means s tremi skupinami doseže vrednost 0.31, kar pomeni, da sta si razbitji malo podobni. Nekoliko manjšo

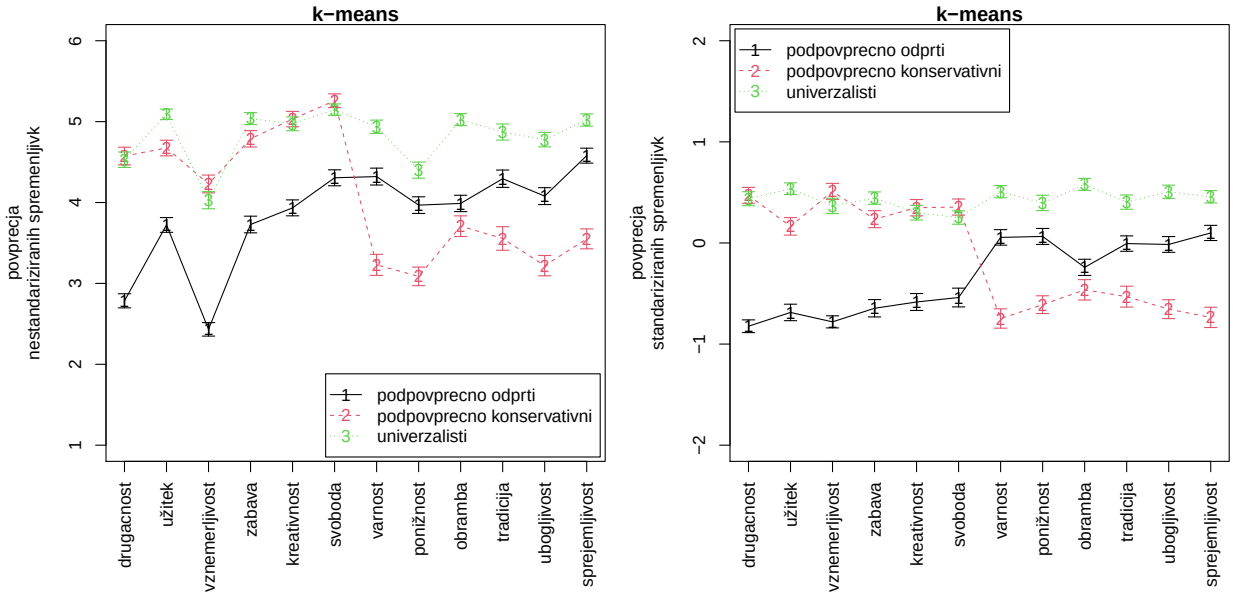
vrednost doseže pri primerjavi k-means in model VVE s tremi skupinami in sicer 0.22. Najmanjša pa je njegova vrednost pri primerjavi Wardove metode z metodo na podlagi modelov, VVE s tremi skupinami, kjer je vrednost enaka 0.16 kar pomeni, da gre za zanemarljivo podobni razbitji. Najbolj podobni sta si torej razvrstitvi na podlagi Wardove metode in k-means.

Interpretacija rezultatov

Na podlagi dosedajšnjih ugotovitev se kot za najboljšo razvrstitev odločimo za razvrstitev na podlagi k-means s tremi skupinami.

Interpretacija izbrane razvrstitve

Da se spomnimo, si še enkrat oglejmo povprečja izračunanih skupin pri nestandariziranih in standariziranih spremenljivkah. Poleg tega se odločimo tudi za poimenovanje skupin in sicer enote, ki pripadajo skupini 1 imenujemo univerzalisti, enote skupine 2 poimenujemo podpovprečno odprti in enote skupine 3 podpovprečno konservativni. Enote so v vse 3 skupine porazdeljene precej enakomerno, in sicer kot univerzaliste imenujemo 552 enot, kot podpovprečno odprte 530 enot in kot podpovprečno konservativne 414 enot.

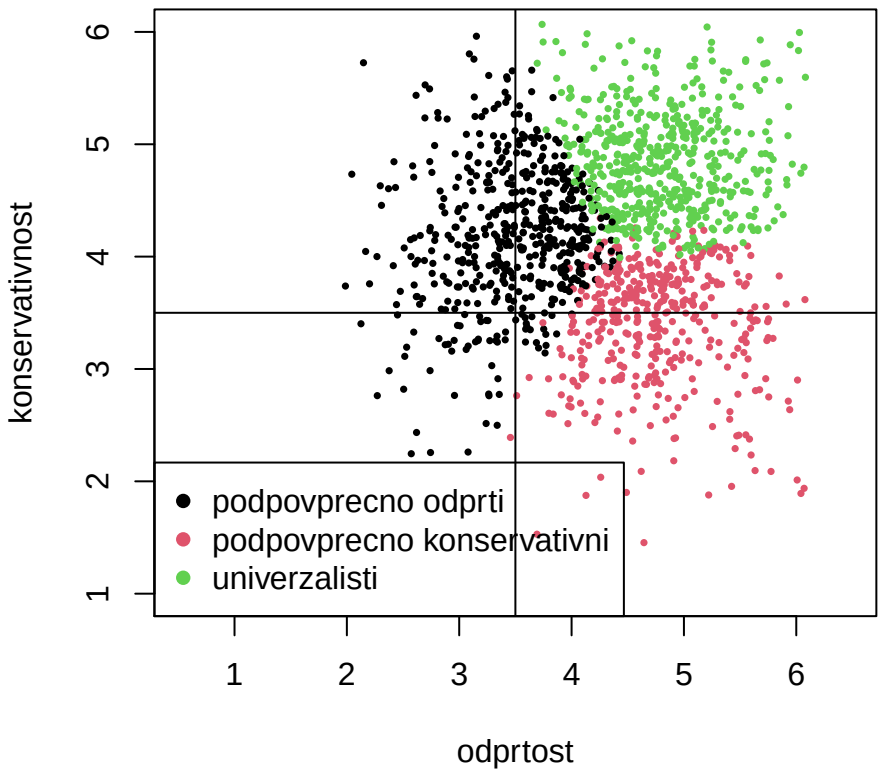


Slika 12: Povprečja skupin za k-means s tremi skupinami

Razsevni grafikon skupin glede na spremenljivki Likartove lestvice

Prikažemo lahko še vrednosti enot glede na spremenljivki Likartove lestvice odprtosti in konservativnosti ter jih pobarvamo glede na pripadnost skupini.

Kot vidimo, skupine niso jasno ločene, saj se držijo precej skupaj. Zgornja slika pa nam lahko služi tudi kot preverba ali smo skupine smiselno poimenovali, seveda relativno glede na ostale spremenljivke, kar v našem primeru drži.



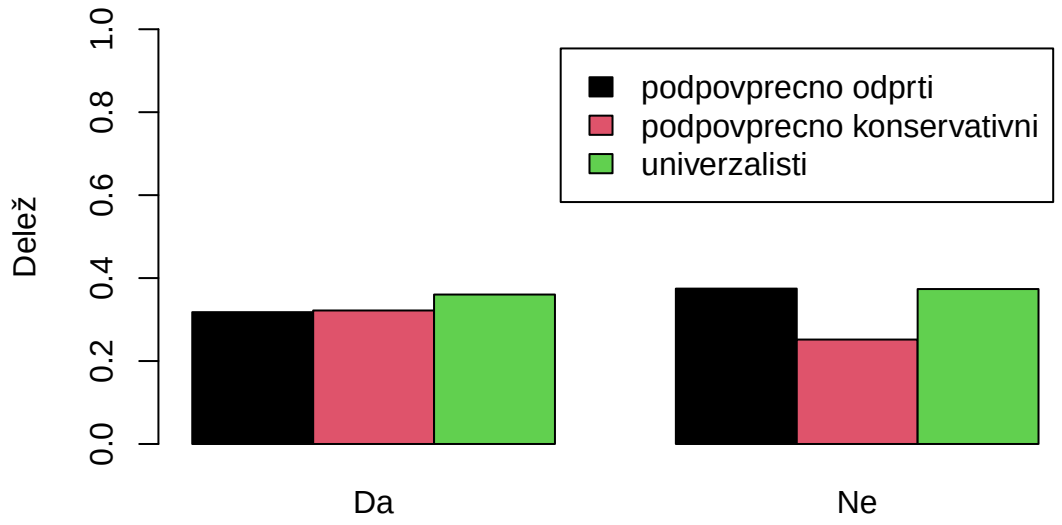
Slika 13: Razsevni grafikon skupin glede na Likartovi spremenljivki

Razlike med skupinami glede na dodatne spremenljivke pri razvrstitvi v skupine

Vpliv dodatnih, v prejšnji nalogi izbranih, spremenljivk na razvrstitev v skupine lahko preverimo grafično, opravimo pa lahko tudi statistični test in obenem preverimo moč povezanosti kjer je to smiselno.

Razlike med skupinami pri spremenljivki Nadzor

Na podlagi spodnjega grafa vidimo, da so zaposleni, ki so odgovorni za nadzor nad sodelavci bolj reprezentativni v skupini podpovprečno konservativni v primerjavi s tistimi ki niso odgovorni za nadzor. Največ odgovornih za nadzor je sicer univerzalistov, med podpovprečno odprte in podpovprečno konservativne pa se porazdeljujejo približno enakomerno. Tudi nadzorovani so v največji meri univerzalisti, sledijo jim podpovprečno odprti in kasneje še podpovprečno konservativni.

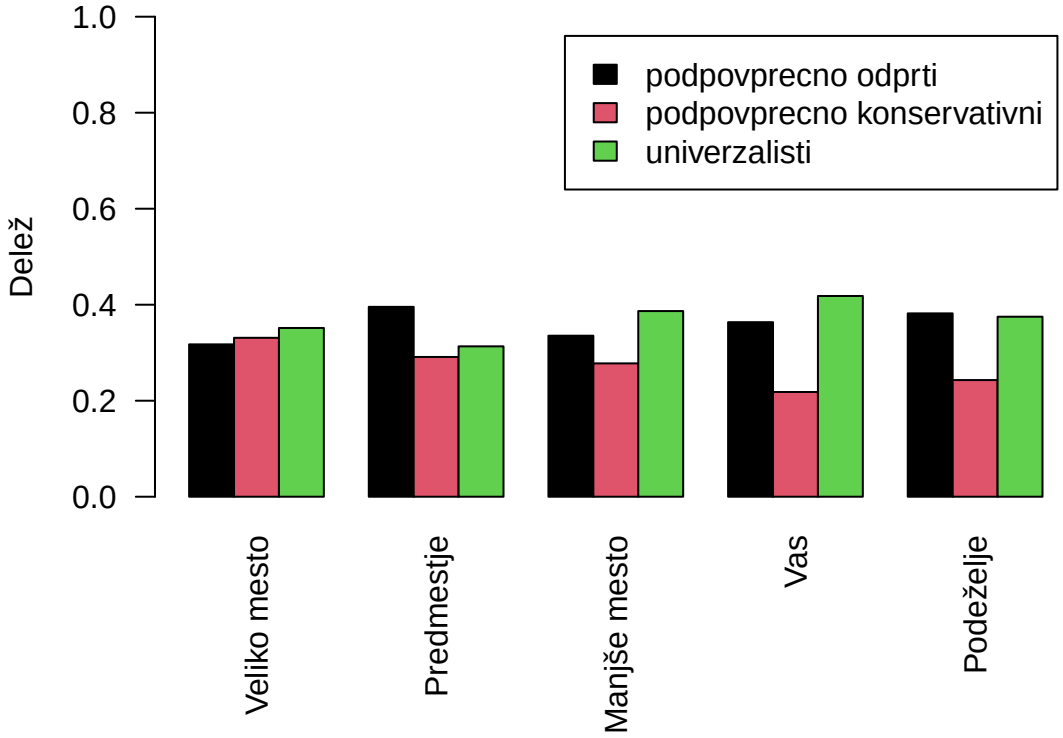


Slika 14: Spremenljivka Nadzor glede na razvrstitev v skupine

Na podlagi testa hi-kvadrat, pri stopnji značilnosti 0.05, na podlagi vrednosti p (p=0.0115), zavrnemo ničelno domnevo, ki pravi, da ni povezanosti med spremenljivko Nadzor in razporeditvijo v skupine. Povemo pa lahko tudi, da na podlagi koeficienta povezanosti Kramerjev V, ki zavzame vrednost 0.0783, rečemo, da je povezanost zelo šibka, pa vendar statistično značilna.

Razlike med skupinami pri spremenljivki Prebivališče

Podobno lahko naredimo tudi glede spremenljivke Prebivališče.



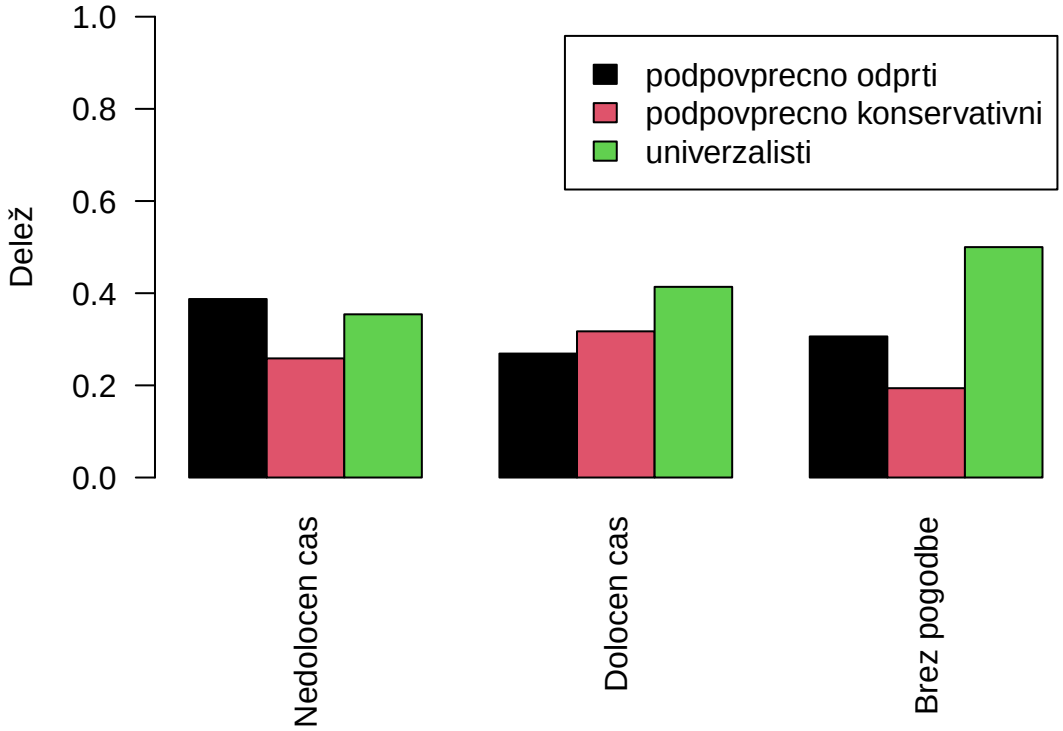
Slika 15: Spremenljivke Prebivališče na razvrstitev v skupine

Na podlagi grafičnega prikaza lahko rečemo, da je v vseh področjih bivanja bolj kot ne največ univerzalistov in najmanj podpovprečno konservativnih, izjema je le predmestje, kjer odstopajo podpovprečno odprti. Najmanjša razlika med deleži je sicer na vasi, kjer je največ univerzalistov in najmanj podpovprečno konservativnih, najbolj pa so si deleži enakovredni v velikem mestu. Opazimo tudi, da delež podpovprečno konservativnih načeloma pada z oddaljevanjem prebivališča od večjih mest.

S Kruskal-Wallisovim testom ugotovimo tudi, da obstaja povezanost med spremenljivko Prebivališče in razvrstitvijo v skupine pri velikosti testa 0.05. Torej obstajajo razlike med vsaj dvema skupinama glede na kraj bivanja. Moč povezanosti ponovno preverimo s Kramerjevim V-jem, ki zavzame vrednost 0.066, kar pomeni, da spremenljivka Prebivališče še nekoliko šibkeje povezana s skupinami kot Nadzor.

Razlike med skupinami pri spremenljivki Razmerje

Preverimo še kako je z vplivom delavnega razmerja na skupine.



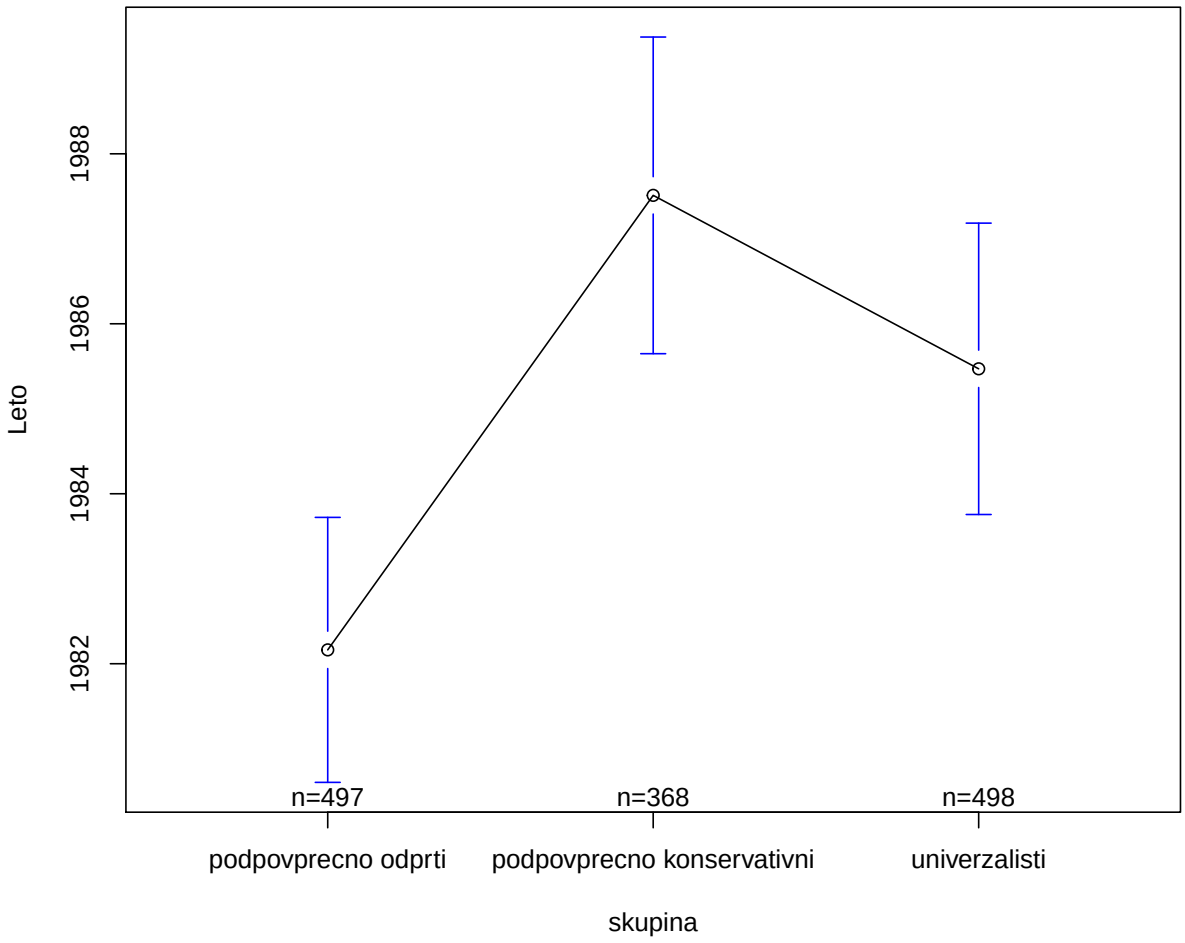
Slika 16: Vpliv spremenljivke Razmerje na razvrstitev v skupine

Največje razlike med deleži v skupinah so v ljudeh brez pogodbe, kjer je največ univerzalistov in najmanj podpovprečno konservativnih. Delež podpovprečno odprtih je največji pri enotah s pogodbo za določen čas, najmanjši pa pri pogodbi za nedoločen čas. Največji delež podpovprečno konservativnih pa je pri tistih s pogodbo za nedoločen čas.

S hi-kvadrat testom, pri štirih stopinjah prostosti lahko pri stopnji značilnosti 0.05 na podlagi vrednosti p (p=0.00395) zavrnemo ničelno domnevo, ki pravi da pri delovnomu razmerju ni povezanosti z razvrstitvijo v skupine. Kramerjev V poda vrednost 0.0761, kar pomeni, da gre za podobno, zelo šibko, povezanost kot pri spremenljivk Nadzor.

Primerjava povprečij glede na spremenljivko Zaposlitev

Kot zadnje pa si oglejmo še primerjavo povprečij po skupinah glede na leto prve zaposlitve.



Slika 17: Primerjava povprečij glede na Zaposlitev

Vidimo, da so se v povprečju podpovprečno odprti zaposlili najprej, okrog leta 1983, kar nekako pomeni, da so starejši ljudje v povprečju najmanj odprti in obenem so tudi najbolj konservativni. Univerzalisti so se v povprečju zaposlili nekje okrog leta 1986, podpovprečno konservativni pa v letu 1985.

Levenov t-test, s katerim preverimo ali so variance po skupinah spremenljivke Zaposlitev enake, pri dveh stopinjah prostosti vrne vrednost p 0.067, kar pomeni da pri stopnji značilnosti 0.05, ne moremo zavrniti ničelne domneve, da so variance po skupinah različne. Tako izvedemo enosmerni anova test s predpostavko enakosti varianc, pri katerih lahko na podlagi vrednosti p ($p<0.001$) pri stopnji značilnosti 0.05 zavrnemo ničelno domnevo, da so povprečja pri vseh skupinah glede na spremenljivko Zaposlitev enaka in z 95% zaupanjem sklenemo, da vsaj pri dveh skupinah obstajajo razlike v povprečjih.

Vsebinski povzetek

V nalogi smo obravnavali različne metode razvrščanja v skupine, ker smo iskali primerno št. skupin da čimbolje razlikujemo skupine po njihovih lastnostih. Med hierarhičnimi metodami smo preverili metode kot so, minimalna, maksimalna, povprečna in Wardova metoda. Pri nehierarhičnih pa metodo voditeljev, scree-diagram, GAP ter še razvrščanje na podlagi modelov.

Pri hierarhičnemu razvrščanju je bila izbrana Wardova metoda s štirimi skupinami. Pri nehierarhičnemu smo se odločili za metodo voditeljev s tremi skupinami. Pri razvrščanju na podlagi modelov pa za model VVE s tremi skupinami. Po primerjanju teh treh metod, na podlagi Wardove kriterijske funkcije, se odločimo za metodo voditeljev. Dobljene skupine smo poimenovali Univerzalisti, podpovprečno konzervativni, podpovprečno odprti.

V nadaljevanju smo pogledali še razvrstive po izbranih spremenljivkah (Nadzor, Prebivališče, Razmerje, Zaposlitev), kjer smo ugotovili, da pri vseh spremenljivkah obstaja šibka povezanost z razvrstitvijo.