Domača naloga 1

Neza Krzan, Tom Rupnik

Kazalo

Slike

Tabele

1 Cilji naloge

V nalogi bova poskušala 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.

2 Podatki

Uporabila bova podatke *Swiss banknotes data*, ki vsebujejo šest meritev, opravljenih na 100 pravih in 100 ponarejenih starih švicarskih bankovcih za 1000 frankov.

Podatki vsebujejo 7 spremenljivk - 6 številskih in eno opisno. Vsebujejo različne izmerjene dolžine in širine bankovca v milimetrih:

- `length`: dolžina bankovca(na sliki x_1),
- `left`: dolžina levega roba(na sliki x_2),
- `right`: dolžina desnega roba(na sliki x_3),
- `bottom`: dolžina spodnjega roba(na sliki x_4) in
- `top`: dolžina zornjega roba(na sliki x_5) ter
- `diag`: dolžina diagonale bankovca(na sliki x_6).

Opisna spremenljivka `status` pa določa ali je bankovec pravi(`genuine`) ali ponarejen(`counterfeit`). V tabeli imamo torej meritve za 200 različnih bankovcev.

2.1 Urejanje podatkov

Imena spremenljivk in vrednosti kategorične spremenljivke sva preimenovala v slovenska imena ter, kot sva že napisala zgoraj, sva podatke skalirala.

Preimenovane spremenljivke:

- `length`: `dolžina`,
- `left`: `levi.rob`,
- `right`: `desni.rob`,
- `bottom`: `spodnji.rob`,
- `top`: `zgornji.rob`,
- `diag`: `diagonala` in
- `status` : `tip`, kjer je potem `counterfeit`:`ponarejen bankovec` in `genuine`:`pravi bankovec`.

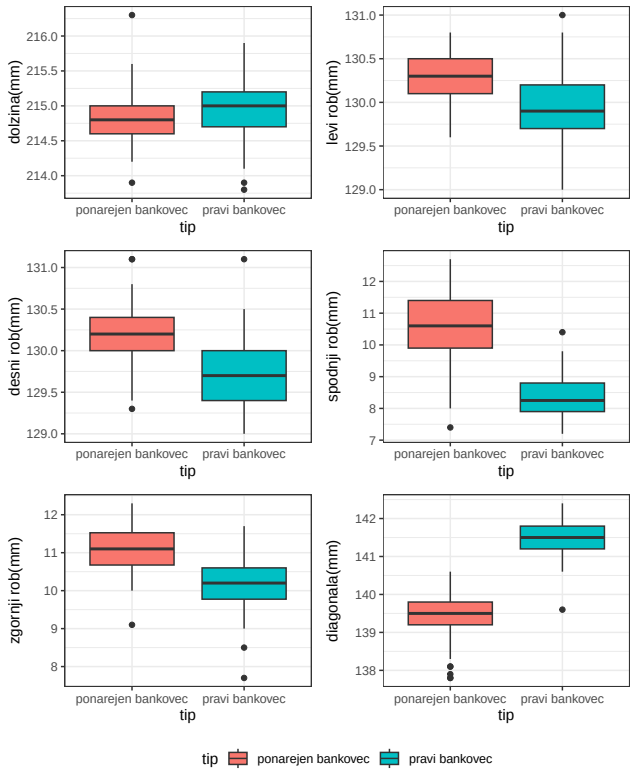
Za lažjo predstavo si poglejmo opisne statistike številskih spremenljivk, da bomo vedeli s kakšnimi podatki imamo opravka.

Tabela 1: Opisne statistike za števillske spremenljivke v podatkovnem okviru Swiss banknotes data.

spremenljivke	N	mean	Std.Dev.	Min	Pctl.25	Pctl.50	Pctl.75	Max
dolzina	200	215	0.4	214	215	215	215	216
levi.rob	200	130	0.4	129	130	130	130	131
desni.rob	200	130	0.4	129	130	130	130	131
spodnji.rob	200	9	1.0	7	8	9	11	13
zgornji.rob	200	11	0.8	8	10	11	11	12
diagonala	200	140	1.0	138	140	140	142	142

Spremenljivke imajo različen razpon vrednosti, zato jih bova, skalirala. Tako bodo imele spremenljivke povprečje 0 in standardni odklon 1. S tem doseževa enakovreden vpliv spremenljivk na razvrstitev. Vidimo pa tudi, da nimamo mankajočih vrednosti v podatkih.

Poglejmo si še porazdelitve spremenljivk.



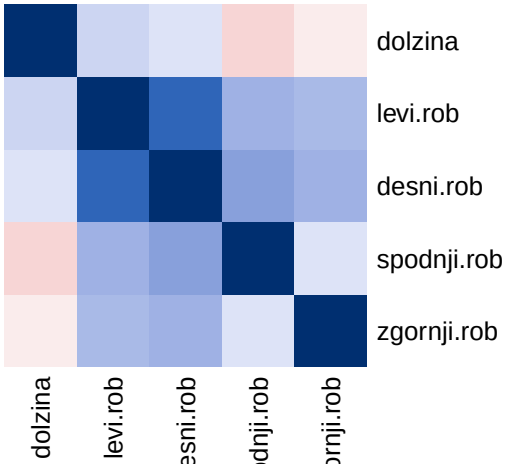
Slika 1: Porazdelitve spremenljivk v podatkovnem okviru `Swiss banknotes data`.

Opazna je razlika med pravimi bankovci in ponarejenimi pri vseh spremenljivkah.

Za razvrščanje bova uporabljala samo številske spremenljivke, in sicer `dolzina`, `levi.rob`, `desni.rob`, `spodnji.rob`, `zgornji.rob`; za analizo pa spremenljivki `tip` in `diagonala`. Ker je `diagonala` edina številska spremenljivka pri analizi, le ta ne bo skalirana.

2.2 Analiza povezanosti med spremenljivkami

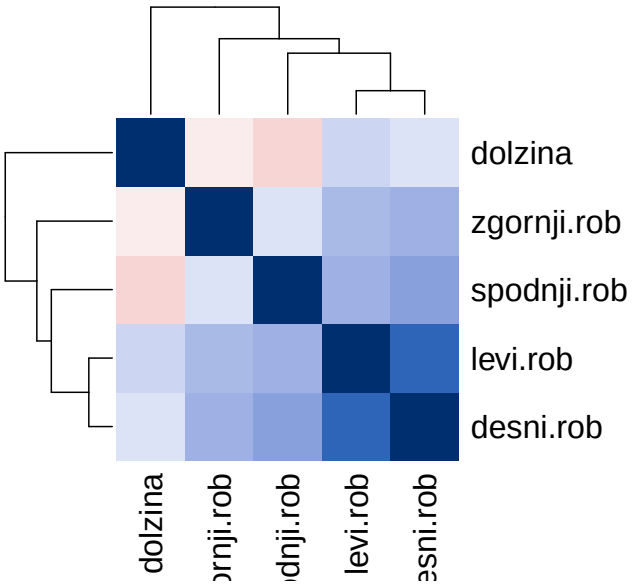
Najprej si poglejmo korelacijo med spremenljivkami.



Slika 2: Korelacija med spremenljivkami.

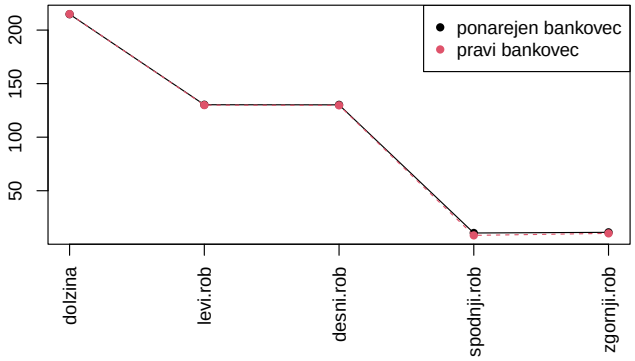
Iz grafa vidimo, da sta levi in desni rob med seboj močno povezana. Prav tako imata zgornji in spodnji rob z njima srednje močno korelacijo. Dolžina ima z vsemi zelo šibko korelacijo iz česar bi lahko zaključili, da tvori svojo skupino.

To je še bolje razvidno iz spodnjega grafa.



Slika 3: Korelacija med spremenljivkami in možna povezava med spremenljivkami.

Poglejmo si povezanost med kategorično spremenljivko `tip` bankovca in številskimi spremenljivkami za razvrščanje.



Slika 4: Povezanost med kategorično spremenljivko(tip bankovca) in številskimi spremenljivkami.

Na tem grafu ne vidimo razlik tako dobro kot na grafu iz slike 1, ker je skala na y-osi drugačna oziroma prilagojena za vse spremenljivke, razlike glede na tip bankovca pa vseeno so, sicer minimalne.

Poglejmo si še korelacijo z številsko spremenljivko za analizo `diagonala` in ostalimi spremenenljivkami.

Tabela 2: Korelacija med spremenljivko diagonala in ostalimi spremenljivkami:

dolzina	0.19
levi.rob	-0.50
desni.rob	-0.52
spodnji.rob	-0.62
zgornji.rob	-0.59

Vidimo, da večjih korelacij ni, so pa skoraj vse negativne.

Torej sklepava lahko, da spremenljivka `dolzina` tvori svoj sklop, ostale spremenljivke pa svojega iz zgornje analize in grafov.

2.3 Konstrukcija in analiza Likertovih spremenljivk

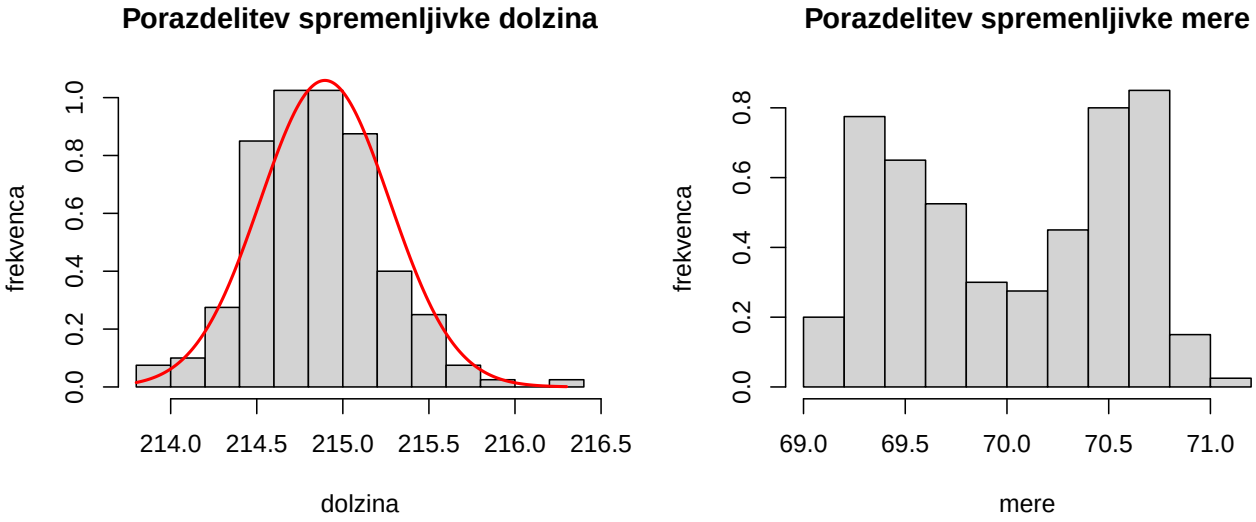
Sedaj bomo za spremenljivke izdelali dve Likertovi lestvici, torej razdelila sva jih v dve skupini, tako da znotraj skupine spremenljivke merijo približno enako stvar. Za vsako skupino sva ustvarili eno Likertovo lestvico, tako da za vsako opazovano enoto izračunamo povprečje izbranih spremenljivk.

- Prvi sklop predstavlja dolžino bankovca in vsebuje `dolzina`.
- Druga sklop predstavlja mere in vsebuje `levi.rob`,`desni.rob`, `spodnji.rob`, `zgornji.rob`.

Poglejmo si porazdelitvi novo ustvarjenih spremenljivk Likertove lestivce tj. dolzina in mere.

Tabela 3: Deskriptivne statistike Likertovih spremenljivk.

	n	mean	sd	median	min	max	range	skew	kurtosis
v1	200	214.90	0.38	214.90	213.80	216.30	2.50	0.19	0.71
v2	200	70.04	0.56	70.09	69.05	71.03	1.97	-0.06	-1.49



Slika 5: Porazdelitvi Likertovih spremenljivk.

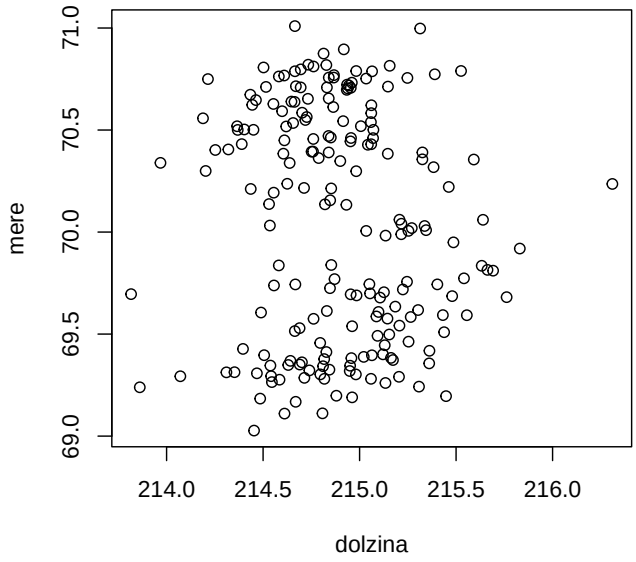
Porazdelitvi `dolzina` in `mere` si nista podobni. Povprečna nagnjenost k `dolzina` je 214.9 s standardnim odklonom 0.38 in je dokaj simetrična. Povprečna nagnjenost k `mere` je 70.04, medtem ko je standardni odklon enak 0.56 in njena porazdelitev je precej nedefinirana.

Da preverimo domnevo o enakosti povprečij `dolzina` in `mere` uporabimo t-test za odvisna vzorca z ničelno domnevo, da je razlika povprečij enaka 0. Ta vrne vrednost $p < 0.001$, torej lahko ničelno domnevo pri stopnji značilnosti 0.05 zavrnemo. 95% interval zaupanja za razliko povprečij med dolzino in mero je [-144.96,-144.76].

Pogledala bova še ali sta ustvarjeni spremenljivki Likertove lestvice povezani med seboj in ali sta povezani z ostalimi spremenljivkami.

2.4 Povezanost Likertovih spremenljivk

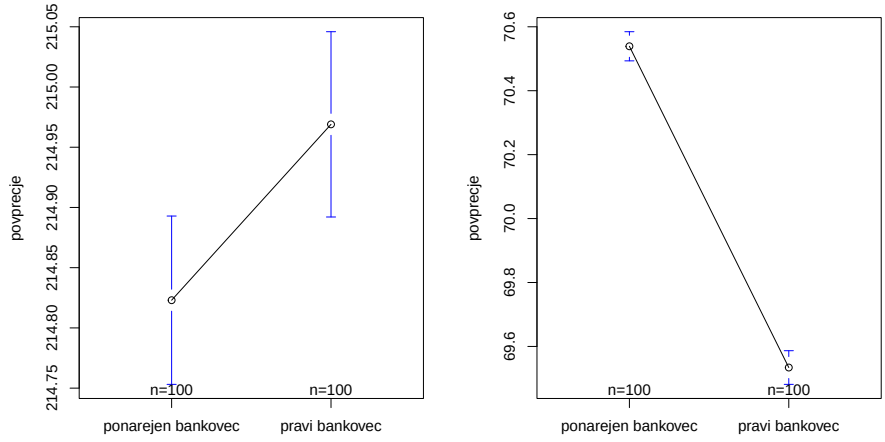
Grafično si oglejmo razsevni diagram dolzine in mer.



Slika 6: Razsevni diagram dolzine in mer.

Da sva preverila ničelno povezanost, sva uporabila korelacijski test za odvisna vzorca na osnovi Pearsonovega koeficienta korelacije, ki vrne kot oceno za Pearsonov korelacijski koeficient vrednost -0.0795, ampak vrednost 0 pa pade v 95% IZ ter vrednost $p = 0.26 > 0.05$, torej ne moreva pri stopnji značilnosti 5% trditi, da povezanost obstaja na populaciji.

2.5 Likertovi spremenljivki in `tip`

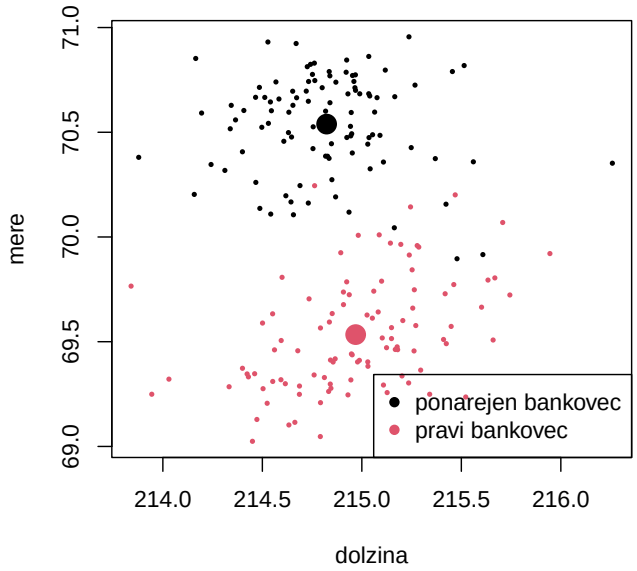


Slika 7: Povprečja dolzina (levo) in mere (desno) po skupinah spremenljivke tip.

Tukaj sva uporabljala Welchev test za primerjavo povprečij, ki pokaže, da se skupini tipa bankovcev pri stopnji značilnosti 5% v povprečju statistično značilno razlikujeta pri obeh Likertovi spremenljivki, ker je $p < 0.05$.

Ocena za povprečje pri dolzini za ponarejen tip bankovca je 214.82 in za pravi bankovec 214.97 pri 95% intervalu zaupanja [-0.25, -0.043]. Ocena za povprečje pri mere pa je za ponarejen tip bankovca 70.54 in za pravi bankovec 69.53 pri 95% intervalu zaupanja [0.94, 1.07].

Nariševa še razsevni diagram spremenljivke `tip`, kjer za osi vzameva Likertovi lestvici, točke pa predstavljajo posamezne enote. Dodamo tudi dve večji točki, ki predstavljata povprečji po skupinah obeh Likertovih spremenljivk.



Slika 8: Razsevni diagram dolzina in mere po spremenljivki tip.

Iz grafa opazimo razlike med tipom bankovca, ampak obstajajo enote, ki se malenkost pomešajo.

Torej spremenljivke sva razdelila v dva sklopa glede na povezanost in tako ustvarila dve Likertovi spremenljivki. Prva spremenljivka predstavlja dolžino, druga pa ostale mere. S 95% zaupanjem lahko trdimo, da dolžina in mere nista statistično značilno povezani. Glede na tip bankovca se skupini statistično razlikujeta pri obeh Likertovi spremenljivki.

3 Hierarhično razvrščanje

Pri hierarhičnem razvrščanju začnemo s tem, da je vsaka enota v svoji skupini. Potem pa se na vsakem koraku, glede na izračunane matrike različnosti, v kateri so razdalje med pari skupin, združujejo skupine, ki so si najbližje. Nato se izračunajo različnosti novih združenih skupin od ostalih, kar se nadaljuje dokler niso vse enote v eni skupini. Dobra lastnost hierarhičnega razvrščanja je, da uporabniku ni potrebno vnaprej določiti števila skupin.

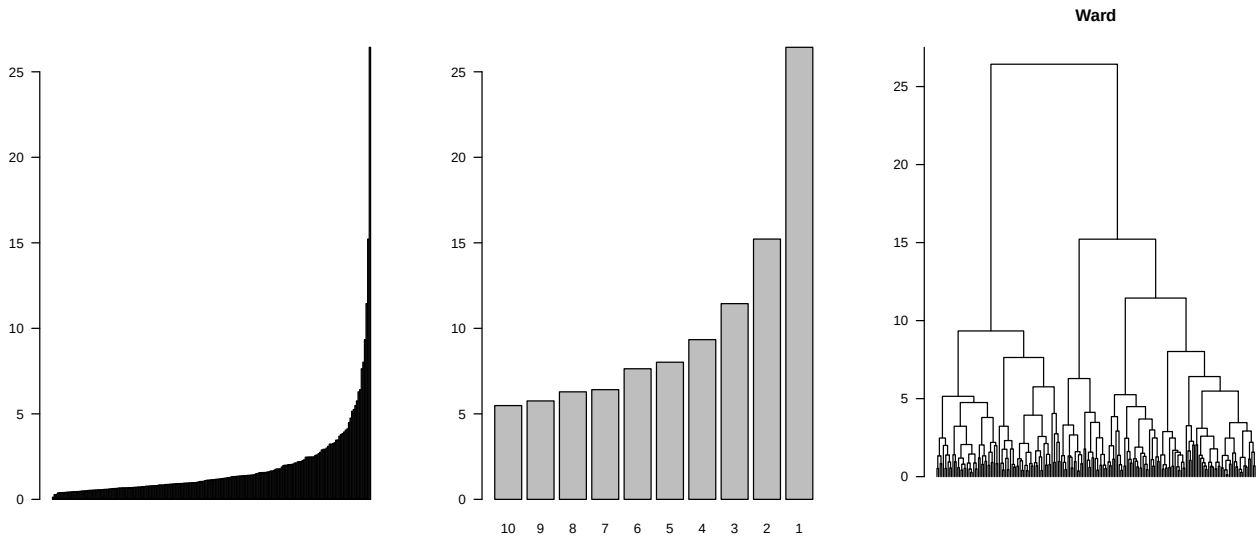
Kot mero različnosti bova uporabila evklidsko razdaljo.

Torej za razvrščanje uporabljava spremenljivke `dolzina`, `levi.rob`, `desni.rob`, `spodnji.rob` in `zgornji.rob` ter primerjala bova tri različne metode in sicer, Wardovo metodo, minimalno metoda (single linkage) in maksimalno metoda (complete linkage).

Število skupin lahko določimo na podlagi dendograma, ki grafično prikazuje potek združevanja v skupine. Število skupin pa določimo tako na podlagi vidnejšega zmanjšanja razdalj skupinami.

3.1 Wardowa metoda

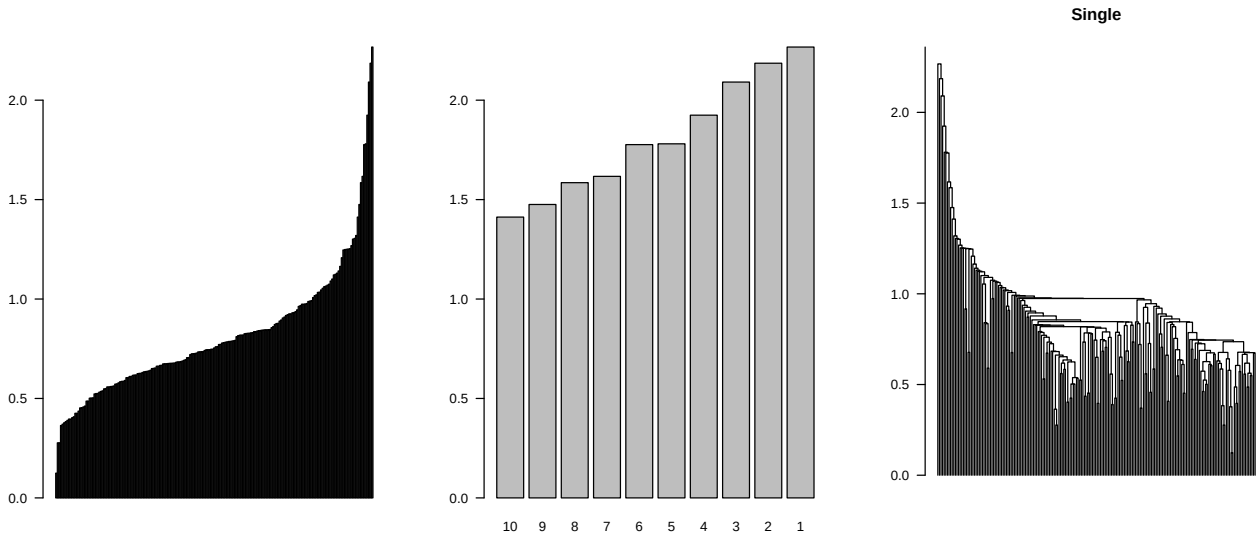
Wardova metoda je primera za eliptične skupine.



Slika 9: Dendogrami Wardowe metode razvrščanja v skupine.

3.2 Minimalna metoda

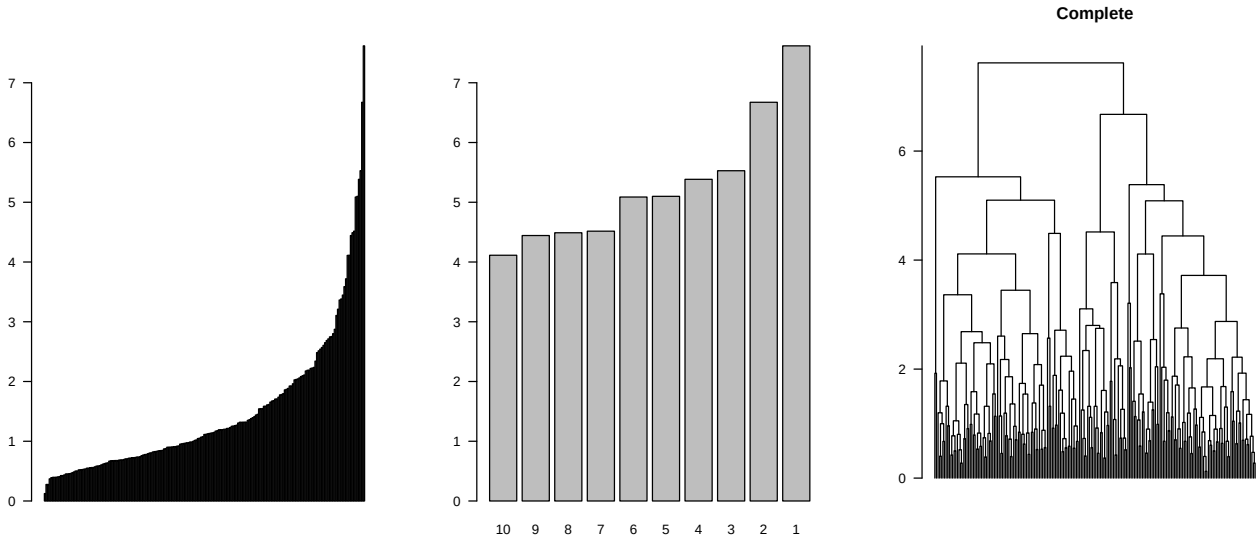
Minimalna metoda (enojna povezanost - single linkage) je primerna za dolge in neeliptične skupine, ki so jasno ločene med seboj. Kadar skupine med seboj niso jasno ločene pri minimalni metodi pride do problema veriženja. Na takem dendogramu ne moremo določiti števila skupin in zato rečemo, da je skupina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



Slika 10: Dendogrami minimalne metode razvrščanja v skupine.

3.3 Maksimalna metoda

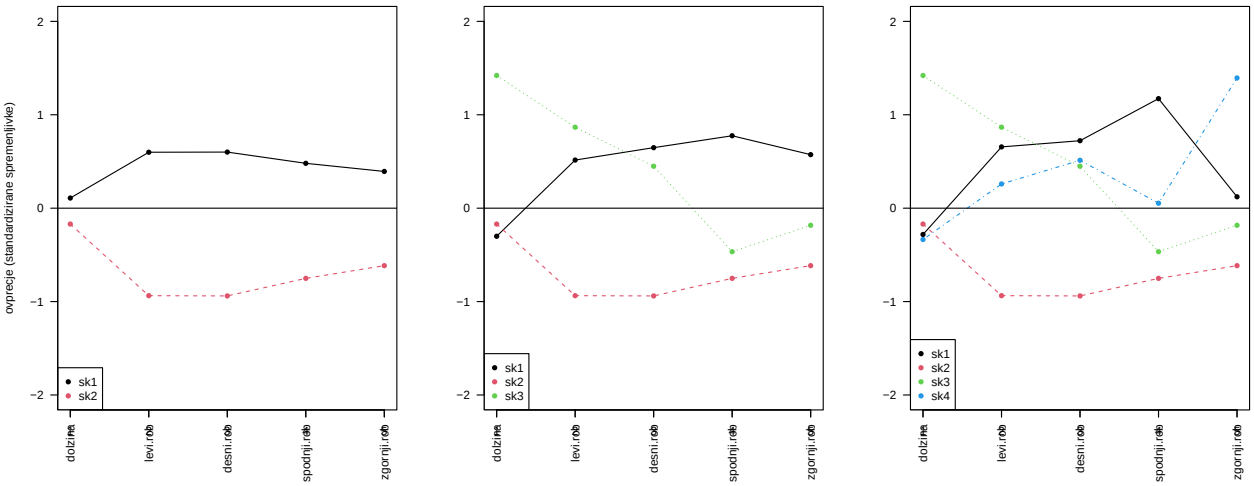
Maksimalna metoda (polna povezanost - complete linkage) pa je primerna za okrogle skupine.



Slika 11: Dendogrami maksimalne metode razvrščanja v skupine.

3.4 Analiza

Glede na izgled grafov (razvrstitve) sva se odločila, da je najbolj primerna razvrstitev po Wardowi metodi. Pri ostalih dveh metodah so različnosti dokaj majhne (ni tako izrazitih skokov v višini). Grafe bomo narisali za *2*, *3* in *4* skupine, saj so tu razlike bolj izrazite.



Slika 12: Povprečja po skupinah za Wardowo metodo.

Če si pogledamo skupino 2 na vseh treh grafih, vidimo, da zavzema podpovprečne vrednosti. Ravno obratno vidimo pri skupini 1, ki na prvem grafu zavzema nadpovprečne vrednosti, na drugih dveh pa zavzema podpovprečne vrednosti samo pri dolžini bankovca. Skupina 3 pa je v nekaterih primerih nadpovprečna v nekaterih pa podpovprečna(`spodnji.rob`, `zgornji.rob`). Pri zadnjem garfu se skupina 4 pri spremenljivki `dolzina` pribliza povprecju zelo dobro, pri vseh ostalih spremenljivkah je nadpovprecna in pri zadnji mocno podpovprecna.

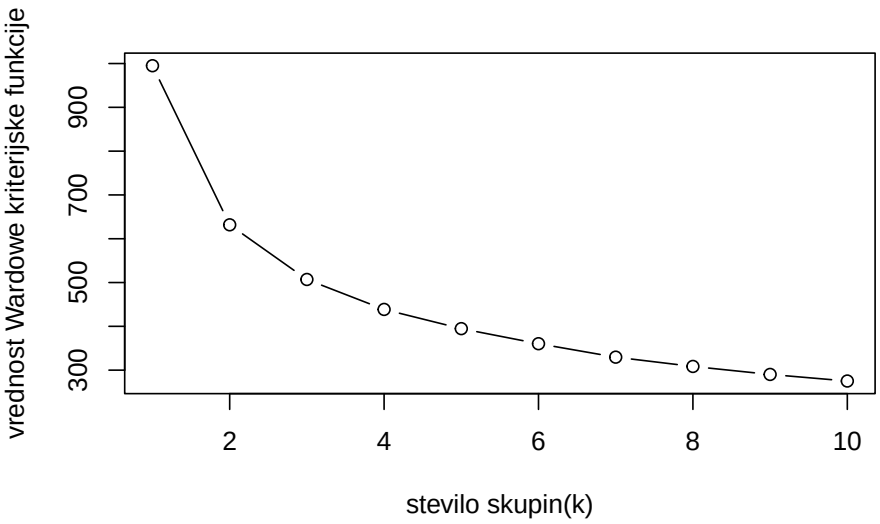
4 Nehierarhično razvrščanje

4.1 Razvrščanje K-means

K-means je metoda voditeljev oz. nehierarhičnega razvrščanja. Voditelji so "predstavniki skupin", vsaka enota pa pripada skupini, kateremu voditelju je najbližje(razdalja je evklidska) oz. mu je najbolj podobna; voditelj predstavlja povprečje skupine. Spremenljivke pri metodi *k-means* morajo biti vsaj intervalne.

Tukaj pri tej metodi mora biti število skupin podamo v naprej, kar je morda slaba lastnost in se glede tega razlikuje od npr. Wardove metode. Na začetku določimo voditelje, potem pa na vsakem koraku vsako enoto priredimo voditelju oz. skupini, kateremu je najbljižja glede na evklidsko razdaljo. Na vsakem koraku se izračunajo novi voditelji kot povprečja skupin. Postopek se zaključi, ko so novi voditelji enaki starim.

Izberemo tisto razvrstitev, ki ima najmanjšo vrednost Wardove kriterijske funkcije, za katero vemo, da pada z naraščanjem števila skupin. Torej za optimalno število skupin ponavadi vzamemo tisto vrednost, kjer se zgodi t.i. "koleno" funkcije. Če to "koleno" ni jasno razvidno, lahko sklepamo, da skupine niso jasno ločene. Postopek običajno večkrat ponovimo, saj za različne začetne voditelje lahko dobimo različne rešitve, torej razvrstitve v skupine.

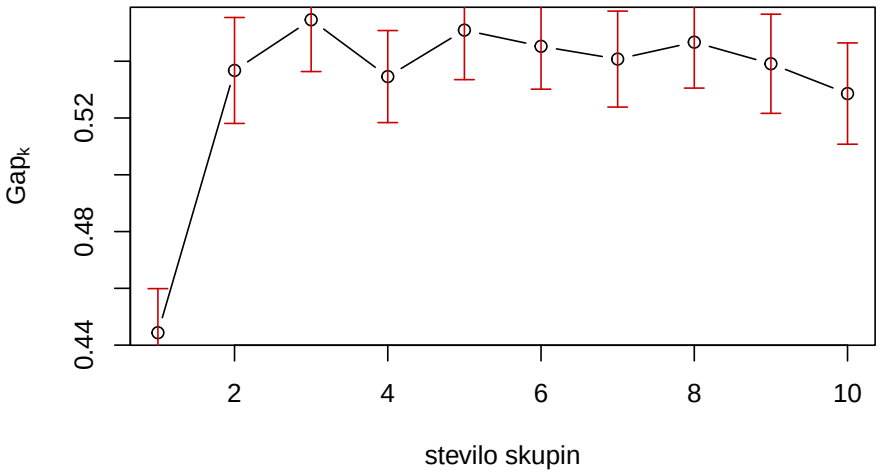


Slika 13: Vrednost Wardove kriterijske funkcije.

Sprememba naklona funkcije izgleda največja pri **2** ali **3** skupinah oziroma je tam "koleno" najbolj razvidno.

4.2 GAP statistika

Pri določevanju števila skupin si lahko pomagamo tudi z GAP statistiko, kjer iščemo skupine, ki so podatki bolj homogeni, kot kjer ni skupin. Gre za primerjavo razdalj znotraj skupin z razdaljami na podatkih brez skupin. Izberemo pa tisto najmanjše število skupin k , kjer je vrednost $GAP(k)$ statistike vsaj tolikšna kot $GAP(k+1) - SE(GAP(k+1))$; SE je standardna napaka GAP statistike.

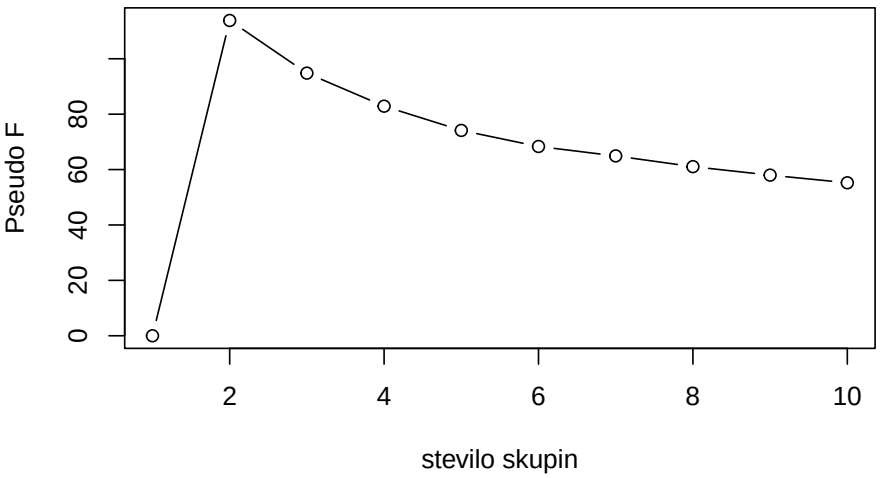


Slika 14: Vrednost GAP statistike.

Na podlagi grafičnega prikaza vrednosti GAP statistike pri različnem številu skupin se odločimo za **3** skupine, saj tam doseže najvišjo točko in začne padati.

4.3 Pseudo F (Calinski - Harabasz indeks)

Uporabimo pa lahko tudi indeks Calinski-Harabasz, ki ocenjuje razmerje med razpršenostjo znotraj skupin in razpršenostjo med skupinami. Uporabljamo ga za oceno primernosti števila skupin v metodi gručenja (angl. *clustering*). Višje vrednosti indeksa Calinski-Harabasz označujejo boljše gručenje, pri čemer optimalno število skupin običajno doseže maksimum tega indeksa.



Slika 15: Vrednost Pseudo F oz. Calinski - Harabasz indeksa.

Tukaj je maksimum dosežen pri **2** skupinah.

Torej, če povzameva celotno analizo, bi, glede na posamezen graf, izbrala

- WSS: sprememba naklona izgleda največja pri **2** skupinah,
- Pseudo F: maksimum doseze pri **2** skupinah,
- gap statistika: najvišjo točko preden začne padati doseže pri **3** skupinah.

Na podlagi zgornjih analiz in ugotovitev pri hierarhičnem razvrščanju, kjer smo se odločali med 2 ali 3 skupinami, bi se tu določili raje za **3** skupine, kot za 2, saj težimo k večjemu številu skupin kot je 2.

4.4 Primerjava vrednosti kriterijske funkcije za Wardowo metodo in K-means

Primerjala sva tudi vrednosti kriterijskih funkcij za Wardowo metodo in metodo K-means, ker sta podobno oziroma delujeta na isti princip. Je pa metoda K-means boljša, ker išče lokalne minimume, za razliko do Wardove, ki deluje hierarhično in vedno poda enak rezultat. Ocenjujeva sva pa po principu, da ima boljša razvrtitev manjšo vrednost karakteristične funkcije. Pomembno pa je tudi to, da so podatki standardizirani, saj drugače med seboj ne bi bilo primerljivo.

Tabela 4: Primerjava vrednosti kriterijske funkcije za Wardowo metodo in K-means.

	k=1	k=2	k=3	k=4
Ward	995	645.6782	529.8618	464.3683
Kmeans	995	631.7882	506.9825	438.5950

Vidimo da ima v vseh primerih (z izjemo prvega kjer sta enaka) K-means manjšo vrednost, kar si tudi želimo. Primerjavo razvrtitev bomo naredili na številu skupin $\mathbf{k = 3}$.

Tabela 5: Kontingenčna tabela.

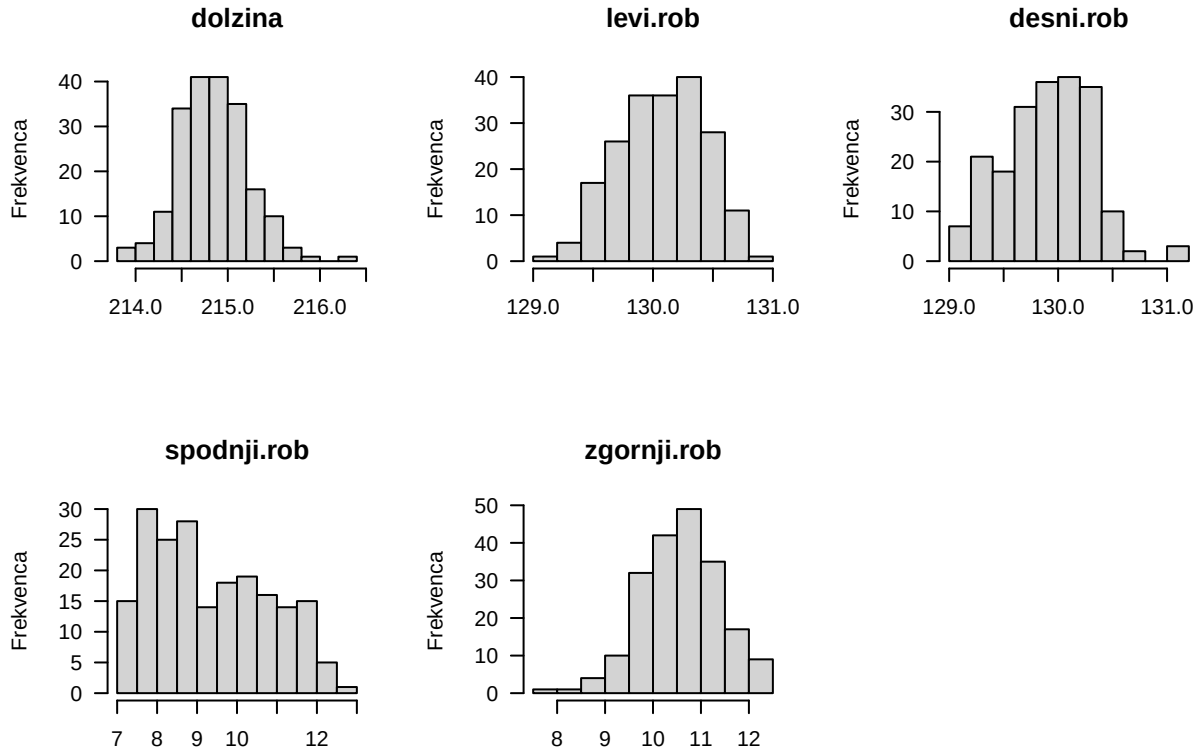
1	2	3
76	0	3
16	27	2
1	2	73

Največje elemente imamo na diagonali kontingenčne tabele, tudi te vrednosti niso ekstremno velike(npr. 400). Za izvendiagonalne elemente si želimo, da bi bili čim manjši oziroma zelo blizu 0, kar pa po večini so, ali pa so celo kar enaki 0(izstopa le ena vrednost - 16).

Poglejmo si še Randov indeks, ki predstavlja delež parov enot, ki so si v obeh razbitjih usklajeni - v obeh razbitjih v isti skupini ali pa v obeh razbitjih v različnih skupinah. Pogledala sva si popravljen Randov indeks, zaradi boljše primerljivosti. Enak je 0.708, kar je blizu 0,5, torej gre skoraj za neko srednjo podobnost, sicer pa večji kot je, boljše je - vrednost 1 pomeni identični razbitji, vrednost 0 pa, da sta si razbitji tako podobni po slučaju.

5 Razvrščanje na podlagi modelov

Tukaj predpostavimo, da so podatki generirani iz multivariatnih normalnih porazdelitev z različnimi parametri oziroma komponentami; vsaka skupina ima svojo multivariatno normalno porazdelitev. Skupina je večja po volumnu, če ima večjo variabilnost, omejimo pa se z domnevami oziroma predhodnim znanjem, kakšne naj bi te skupine bile. Zato si poglejmo porazdelitve spremenljivk ne glede na tip bankovca.



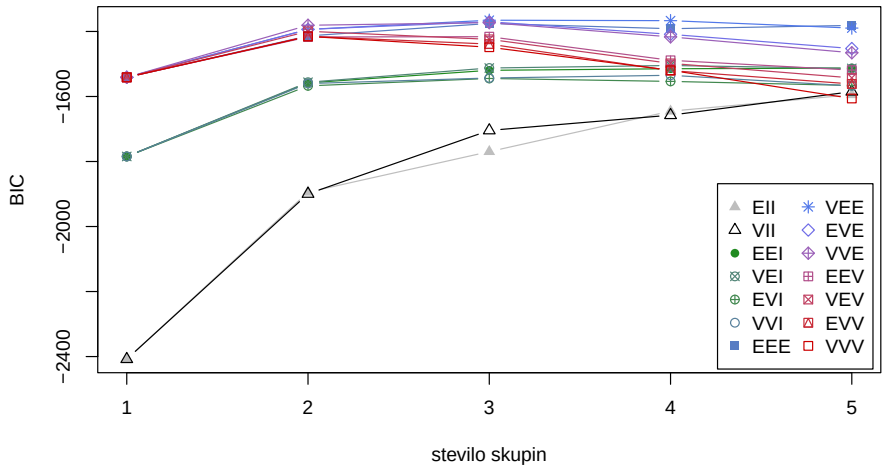
Slika 16: Porazdelitve spremenljivk.

Spremenljivka `spodnji.rob` in tudi `desni.rob` nista porazdeljeni po normalni porazdelitvi, zato ne moremo trditi, da je zadoščen ta pogoj. Ostale so porazdeljene po normalni, nekatere asimetrične v desno(npr. spremenljivka `dolzina`) in nekatere v levo(npr. spremenljivka `levi.rob`).

Tukaj ocenimo število skupin in parametre za vsako skupino ter kateri skupini posamezna enota pripada. V najinem primeru, kjer je predpostavka o multivariatni normalni porazdelitvi kršena, se simulacija ne izkaže za optimalno. Razvrstitev se dela na originalnih podatkih oz. nestandardiziranih podatkih, ker s tem omogočimo različno velikost skupin.

5.1 BIC(Bayes Information Criterion) kriterij

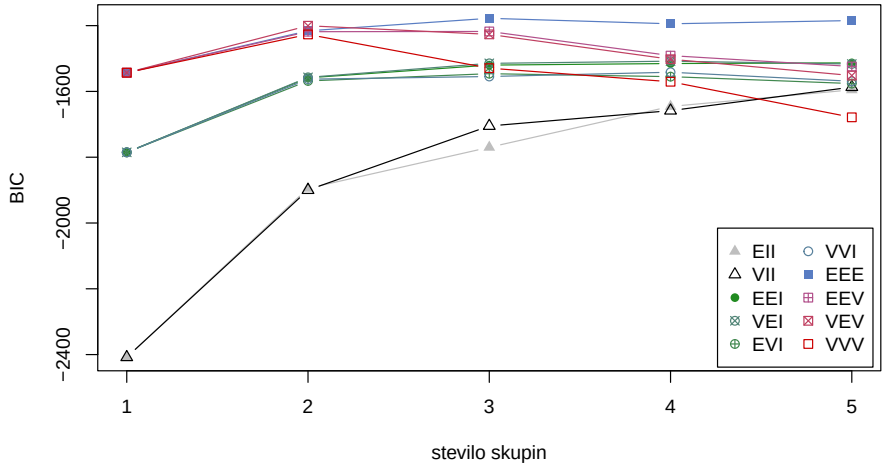
Naredimo torej razvrstitev na originalnih, nestandardiziranih podatkih, kjer funkcija sama izbere naprimernejši model.



Slika 17: BIC kriterij za originalne podatke.

Na podlagi BIC kriterija (Bayesian Information Criterion), ki zavzame vrednost -1365.42 izberemo model VEE s tremi skupinami, kar pomeni, da gre za elipsoidne(angl. *ellipsoidal*) skupine, ki so različno velike, različnih oblik in enako usmerjene.

Kriterij temelji na "Bayesovski" statistiki, zato lahko določimo tudi apriorne verjetnosti(torej neko naše predhodno znanje oziroma prepričanja).

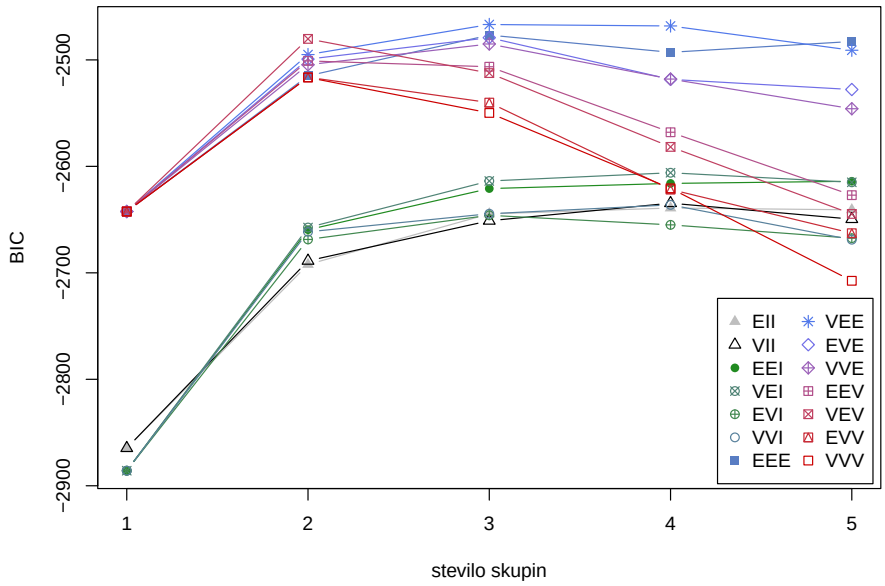


Slika 18: BIC kriterij (priorControl) za originalne podatke.

Na podlagi BIC kriterija z uporabljenim argumentom o apriornih verjetnostih se odločimo za model EEE s tremi skupinami, kar pomeni, da gre za različno velike skupine, različnih oblik in enake usmerjenosti.

5.2 BIC kriterij na standariziranih podatkih

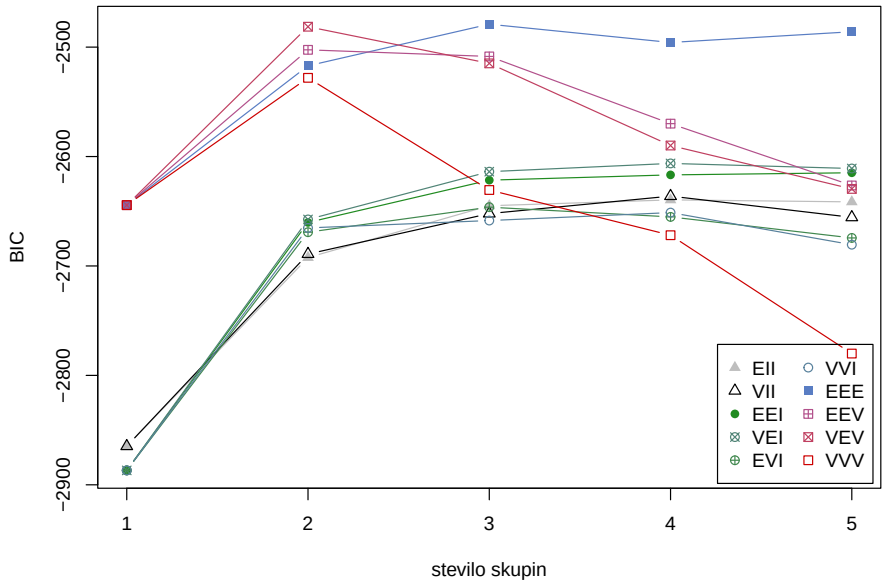
Poglejmo si še, iz radovednosti, kako je z oceno modela na standariziranih podatkih, ampak vrednosti BIC kriterija niso primerljive med standariziranimi in nestandariziranimi podatki.



Slika 19: BIC kriterij za standardizirane podatke.

Na podlagi BIC kriterija (Bayesian Information Criterion), ki zavzame vrednost -2466.76 ponovno izberemo model VVE s tremi skupinami.

Tudi tukaj lahko primerjamo z vključitvijo apriornih verjetnosti.

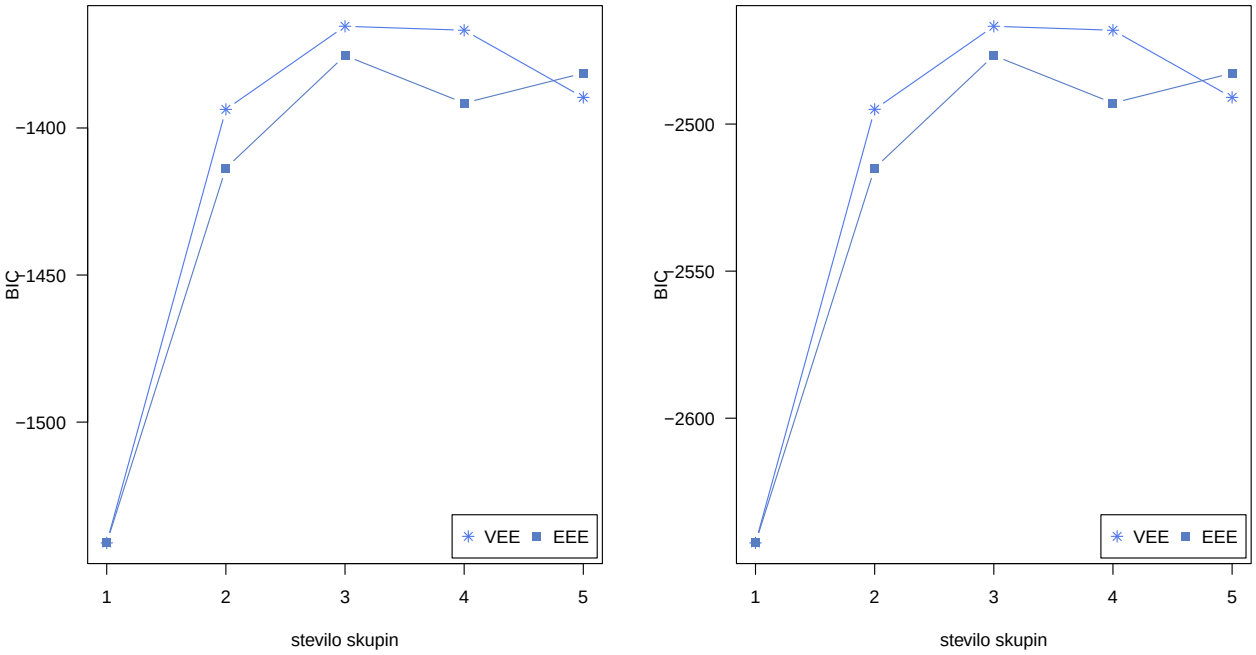


Slika 20: BIC kriterij (priorControl) za standardizirane podatke.

Tudi tukaj se na podlagi BIC kriterija z uporabljenim argumentom o apriornih verjetnostih se odločimo za model EEE s tremi skupinami, kar pomeni, da gre za različno velike skupine, različnih oblik in enake usmerjenosti.

5.3 Primerjava modelov

Na pogladi BIC kriterja, kjer lahko na spodnjem grafu vidimo primerjavo VEE modela in EEE modela za nestandardizirane in standardizirane podatke, se, v obeh primerih, odločimo za model VEE. Bi pa se pri obeh modelih odločila za **3** skupine, saj vrednost BIC kriterija od tam naprej počasi narašča.



Slika 21: Primerjava VEE in EEE modela(levo: nestandardizirani podatki, desno: standardizirani podatki).

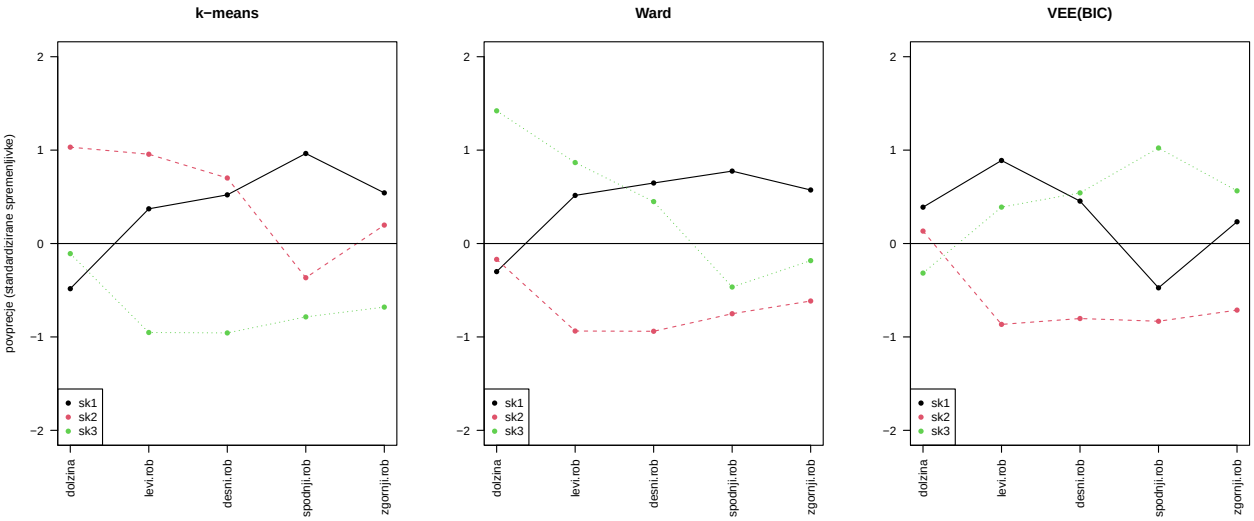
Na podlagi vseh kriterijev se zaradi enostavnosti odločimo za model VEE s tremi skupinami - torej različno velike skupine, različnih oblik in enakih umserjenosti.

6 Najboljša razvrstitev in predstavitev skupin

Tukaj naju pa zanima kako podobne so si naše razvrstitve, ki sva jih v prejšnjih poglavjih izbrala na podlagi različnih modelov. V prejšnjih poglavjih sva izbirala najboljše razvrstitve in sedaj jih bova med seboj primerjala.

6.1 Primerjava povprečij

Na spodnjem grafu si poglejmo povprečja po skupinah in primerjamo razvrstitve na standardiziranih podatkih.



Slika 22: Primerjava razvrstitev na standariziranih podatkih.

Vrstni red skupin se razlikuje v modelih, ampak če pogledamo sta si Wardov model in VEE model nekoliko bolj podobna. Skupina 1 si je pri modelu na podlagi k-means podobna skupini 1 v modelu na podlagi Ward-a. Podobne so si tudi skupina 3 pri modelu na podlagi k-means in skupina 2 pri Wardu in VEE ter skupina 2 pri k-means je podobna skupini 3 pri Wardu. Do odstopanj prihaja le pri posameznih točkah(npr. pri k-means za skupino 2 je nadpovprečno pri spremenljivki `zgornji.rob`, s tem ko je pri modelu Ward za skupino 3 podpovprečno, pri ostalih spremenljivkah pa sta si skupini dokaj podobni). Pri k-means in Wardu- sta si torek skupini 1 podobni, obe sta nadpovprečni pri vseh spremenljivkah, razen pri spremenljivki `dolzina`, kjer sta podpovprečni. Vidimo lahko tudi, da so si skupina 1 pri k-means in Ward-u in skupina 3 pri VEE(BIC) zelo podobne. Skupina 3 pri k-means in skupina 2 pri Ward-u sta edini, ki sta vedno podpovprečni, vedno nadpovprečne skupine pa ni.

6.2 Wardova kriterijska funkcija

Poglejmo si še primerjavo razvrstitev na podlagi Wardove kriterijske funkcije.

Tabela 6: Primerjava vrednosti Wardove kriterijske funkcije za vse metode.

k-means	Ward	VEE(BIC)
506.9825	529.8618	577.464

Glede na vrednost Wardove kriterijske funkcije je najboljša metoda k-means razvrstitev s tremi skupinami, nato sledi Wardova in na koncu VEE. Zato se odločimo za k-means metodo.

6.3 Popravljen Randov indeks

Poglejmo si še kako podobne so razvrstitve glede na popravljen Randov indeks.

Tabela 7: Primerjava popravljenega Randovega indeksa.

Ward in k-means	Ward in VEE(BIC)	k-means in VEE(BIC)
0.708	0.617	0.659

Pri vseh treh primerjavah je vrednost indeksa večja od 0.5, kar pomeni da gre za dokaj podobna razbitja na skupine. Indeks pri Ward in k-means je najvišji, torej gre za najbolj podobno razbitje na skupinah.

6.4 Število enot v skupinah

Na spodnjem izpisu si oglejmo število enot v posamezni skupini in povprečja na nestandardiziranih podatkih za vse modele.

Tabela 8: Velikost skupin pri metodi k -means.

skupina	velikost
sk1	79
sk2	45
sk3	76

Tabela 9: Povprecje Likartovih spremenljivk po skupinah za k-means metodo.

skupina	povprecje Lik. sprem. dolzina	skupina	povprecje Lik. sprem. mere
sk1	214.7139	sk1	70.57975
sk2	215.2844	sk2	70.10111
sk3	214.8553	sk3	69.43355

Tabela 10: Velikost skupin pri Ward metodi.

skupina	velikost
sk1	93
sk2	78
sk3	29

Tabela 11: Povprecje Likartovih spremenljivk po skupinah za Ward metodo.

skupina	povprecje Lik. sprem. dolzina	skupina	povprecje Lik. sprem. mere
sk1	214.7828	sk1	70.54355
sk2	214.8321	sk2	69.46218
sk3	215.4310	sk3	69.95517

Tabela 12: Velikost skupin pri VEE(BIC).

skupina	velikost
sk1	40
sk2	78
sk3	82

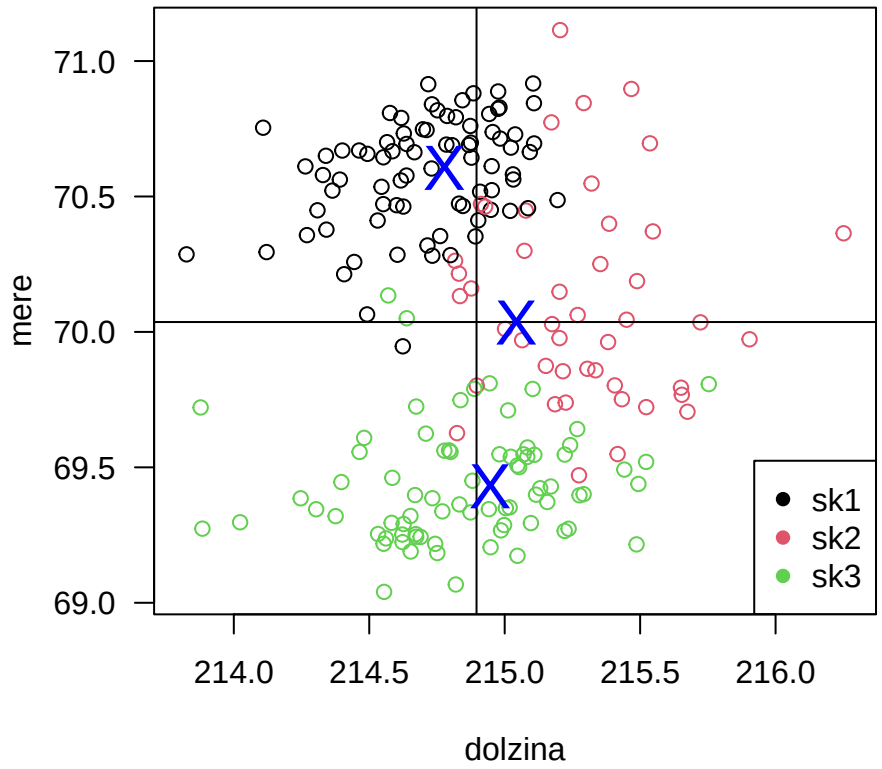
Tabela 13: Povprecje Likartovih spremenljivk po skupinah za VEE(BIC).

skupina	povprecje Lik. sprem. dolzina	skupina	povprecje Lik. sprem. mere
sk1	215.0425	sk1	70.03812
sk2	214.9462	sk2	69.43365
sk3	214.7768	sk3	70.60915

Vidimo, da se število enot glede na skupine pri metodah razlikuje, s tem ko povprečja nestandardiziranih vrednosti Likartovih spremenljivk `dolzina` in `mere` niso tako zelo različne po skupinah.

6.5 Razsevni grafikon skupin glede na Likartovi spremenljivki

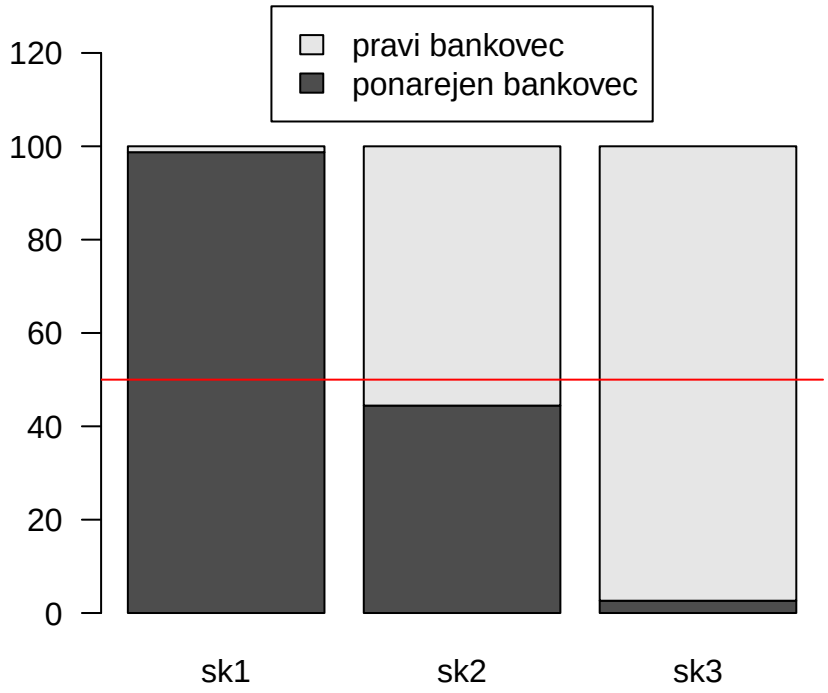
Prikaževa tudi vrednost enot glede na zgoraj izbrani Likartovi spremenljivki `dolzina` in `mere` za modele.



Slika 23: Razsevni grafikon skupin za Likartovi spremenljivki pri k-means.

Kot lahko vidimo iz grafov skupine niso popolnoma jasno ločene, oziroma prihaja do malenkosti mešanja podatkov na robu skupin.

6.6 Povezanost skupin s tipom bankovca



Slika 24: Povezanost skupin pri k-means.

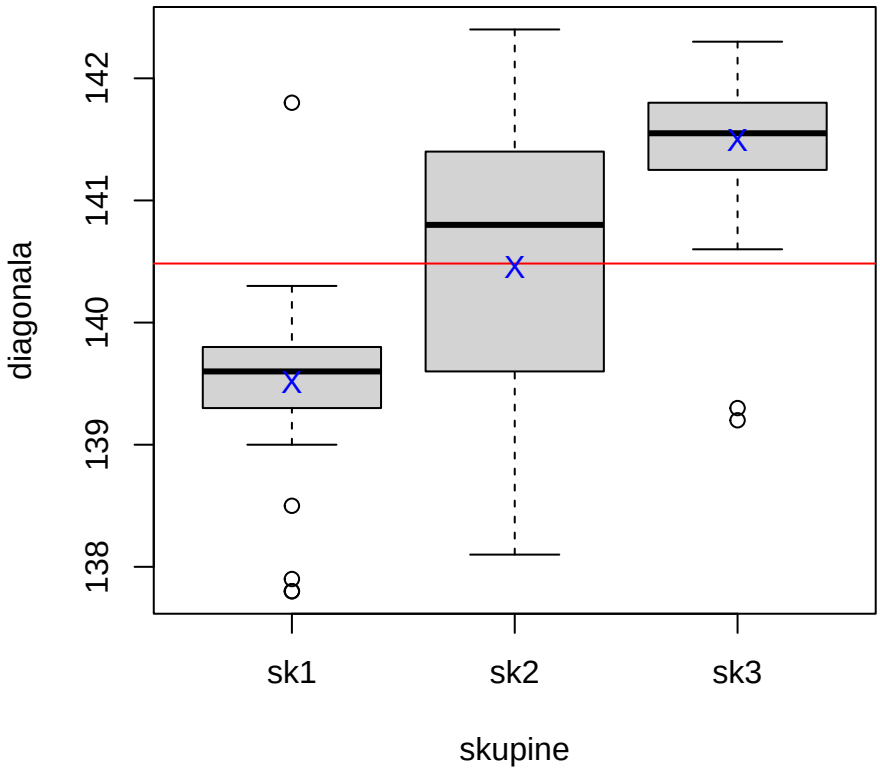
Skoraj vse enote v skupini 1 predstavljajo ponarejene bankovce. Torej skupina 1 je negativno povezana s tipom bankovca (pravimi bankovci).
Skupina 2 ima skoraj enakomerno zastopanost med pravimi in ponarejenimi bankovci.
V skupini 3 pa so skoraj vsi bankovci pravi, torej je pozitivno povezana s tipom bankovca.

Hi-kvadrat test, pri dveh stopinjah prostosti in pri stopnji značilnosti 0.05 je z p-vrednostjo $p < 0.001$ pokazal, da je statistično značilna povezanost med skupino in tipom bankovca.

Preverimo še moč povezanosti s Cramerjevim V koeficientom.

Kramarjev V ima vrednosti 0.85, kar pomeni, da je povezanost med skupinama in tipom bankovcev zelo močna.

6.7 Povezanost skupin z spremenljivko `diagonala`



Slika 25: Povezanost skupin s spremenljivko diagonala pri k-means.

Skupina 1 vsebuje (z izjemo enega) bankovce, ki imajo podpovprečno vrednost diagonale in povezanost skupine 1 in diagonale je negativna. Glede na to, da skupina 1 vsebuje same ponarejene bankovce, ki imajo v povprečju manjšo diagonalo, je to povsem smislno, zato nas tudi nadpovprečne vrednosti diagonale ne presenetijo v tretji skupini.
Povprečje vrednosti diagonal bankovcev v skupine 2 se ujema s povprecjem celotnega vzorca, vendar pa so v njej vsebovani bankovci, ki imajo dosti podpovprečno ali nadpovprečno vrednost diagonale, kar je smiselno, saj vsebuje ponarejene in prave bankovce.
Bankovci v skupini 3 imajo torej vsi nadpovprečno vrednost diagonale saj so vsi pravi. Povezanost te skupine in vrednosti diagonale je pozitivna.

Naredili smo enostranski ANOVA test povezanosti, pri predpostavki različnih varianc. Test je bil statistično značilen ($p < 0.001$), s čimer smo zavrnili ničelno hipotezo, ki pravi, da so povprečja (v našem primeru povprečje vrednosti diagonal) v vseh skupinah enaka.

7 Vsebinski povzetek

V nalogi sva obravnavali različne metode razvrščanja v skupine. Iskala sva najbolj primerno število skupin, da se le-te med seboj čimbolj razlikujejo glede na lastnosti.

Pri hierarhičnemu razvrščanju je bila izbrana Wardova metoda, pri nehierarhičnemu smo se odločili za metodo voditeljev in pri razvrščanju na podlagi modelov pa za model VEE. Povsod smo imeli tri skupine. Po primerjanju teh treh metod, na podlagi Wardove kriterijske funkcije, se odločimo za metodo voditeljev(k-means).

V nadaljevanju pa sva ugotovila, da obstaja povezanost med skupinami in tipom bankovcev, glede na Kramarjev V pa je tudi zelo močna. Ugotovila sva tudi in s testom potrdila, da povprečja diagonal niso enaka v najinih treh skupinah, kar je logično glede na razporeditev bankovcev po skupinah.

8 Viri

Flury, B., Riedwyl, H. (1988). *Multivariate Statistics: A practical approach.* London: Chapman & Hall, Tables 1.1 and 1.2, pp. 5-8.

Pohar P., M.(2024). *Osnove teoretične statistike.*

Polajnar, E.(2024). *Multivariatna analiza.*

Smrekar, J.(2024). *Bayesova statistika.*

Žiberna, A.(2024). *Multivariatna analiza.*