Domača naloga 2

Neža Kržan, Tom Rupnik Medjedovič

1 Cilj naloge

Želiva preučiti 2 različni metodi razvrščanja v skupine. Primerjala bova metodo voditeljev(*k-means*) in razvrščanje na podlagi modelov. Zanima nas katera bo najboljša na podatkih, generiranih iz bivariatne multivariatne normalne porazdelitve.

Zanima naju tudi, kako na metodi vpliva dodajanje nepomembnih spremenljivk, torej tistih, ki imajo enako porazdelitev v vseh skupinah.

Za metodi sva se odločila na podlagi njunih predpostavk, ker so nekatere dokaj podobne, npr. predpostavljajata, da so skupine dovolj ločene oz. ni prekomernega prekrivanja med njimi, homogenosti variance znotraj skupine oz. podatki so v skupinah razmeroma homogeno razporejeni in zahtevata vnaprejšnjo določitev števila skupin, poleg tega pri razvrščanju na podlagi modelov zahtevamo v predpostavkah, da so podatki generirani iz multivariatnih normalnih porazdelitev.

2 Generiranje podatkov

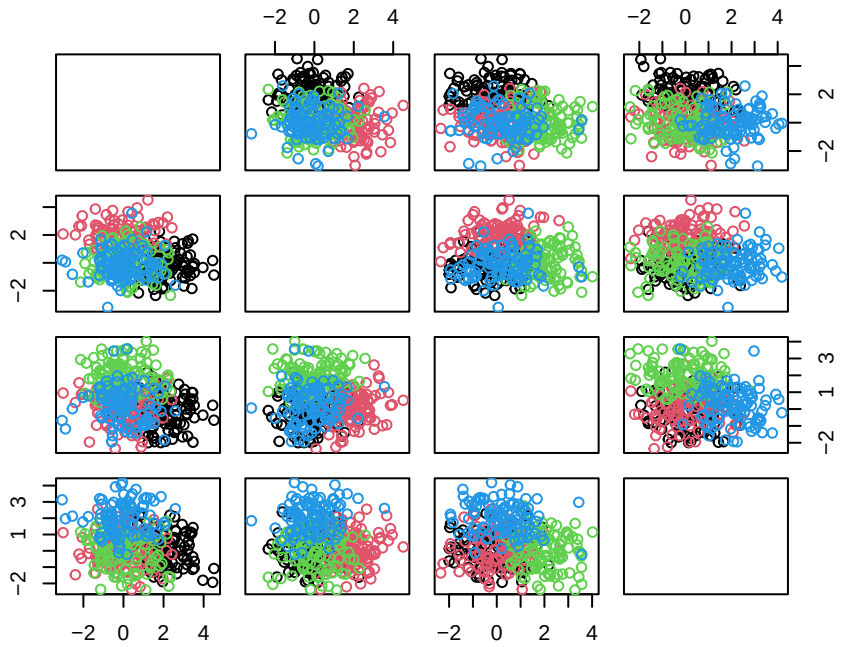
Podatke sva generirala tako, da je njihova porazdelitev bivariatna multivariatna normalna. Zanima naju kako se bodo metode obnesle glede na to kako so si skupine med seboj različne. V ta namen sva si izbrala parameter, ki prilagaja povprečja v skupini, tj. $diff = (1, 2, 4, 10)$. Želiva si, da imava primere, ko so si skupine zelo različne med seboj in ne tako zelo različne. Torej bo pri porazdelitvi povprečja generirana s pomočjo faktorja *diff* in število skupin, kovariančna matrika pa bo po diagonali vsebovala število spremenljivk.

Faktorji, ki jih bova še spreminjala so:

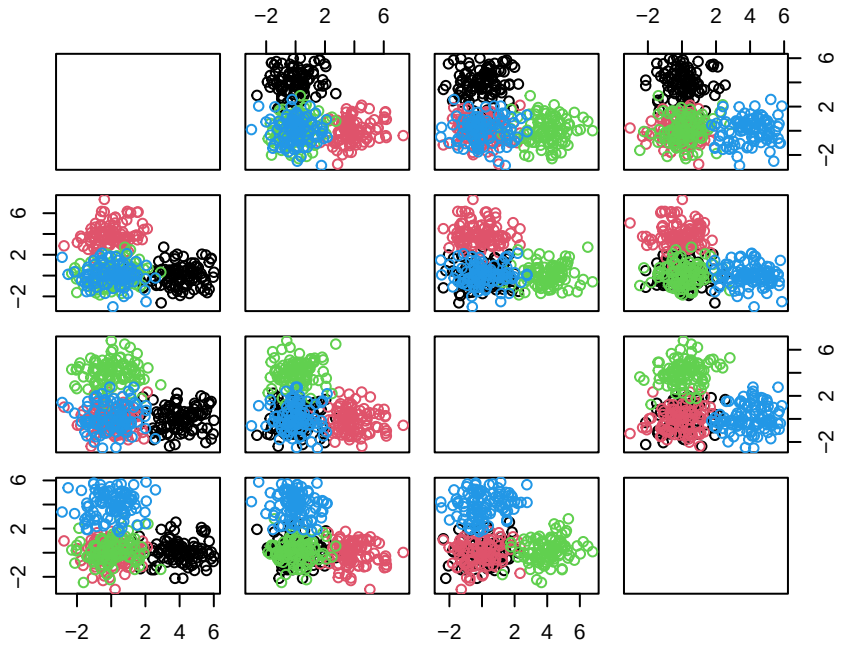
- število skupin, $k = (4, 8, 10)$,
- velikosti skupin, $n = (20, 100, 200)$, pri čemer bodo imele vse skupine vedno enako velikost in
- število spremenljivk, $v = (12, 24, 36)$.

Faktorji so bili izbrani na podlagi tega, da si želiva rezultate, ki bodo dobri in slabi oziroma da bodo za nekatere metode dobri za druge pa slabi.

Pri generiranju podatkov bo število spremenljivk enako številu skupin, vse ostale spremenljivke bodo neinformativne, ker nas zanima tudi kako vpliva dodajanje nepomembnih oz. neinformativnih spremenljivk.



Slika 1: Primer generiranih podatkov za 4 skupine, velikosti n = 100, 12 spremenljivk in diff = 2.



Slika 2: Primer generiranih podatkov za 4 skupine, velikosti n = 100, 12 spremenljivk in diff = 4.

3 Simulacija

Izvedla sva simulacijo s 100 ponovitvami in uporabila t.i. paralelno računanje(angl. *parallel computing*). V simulaciji sva generirala podatke in potem izvedla obe metodi razvrščanja v skupine(*metodo voditeljev* in *razvrščanje na podlagi modelov*).

Za obe metodi sva izračunala 3 različne mere, in sicer prilagojeni Randov indeks(ARI), vsoto kvadratov

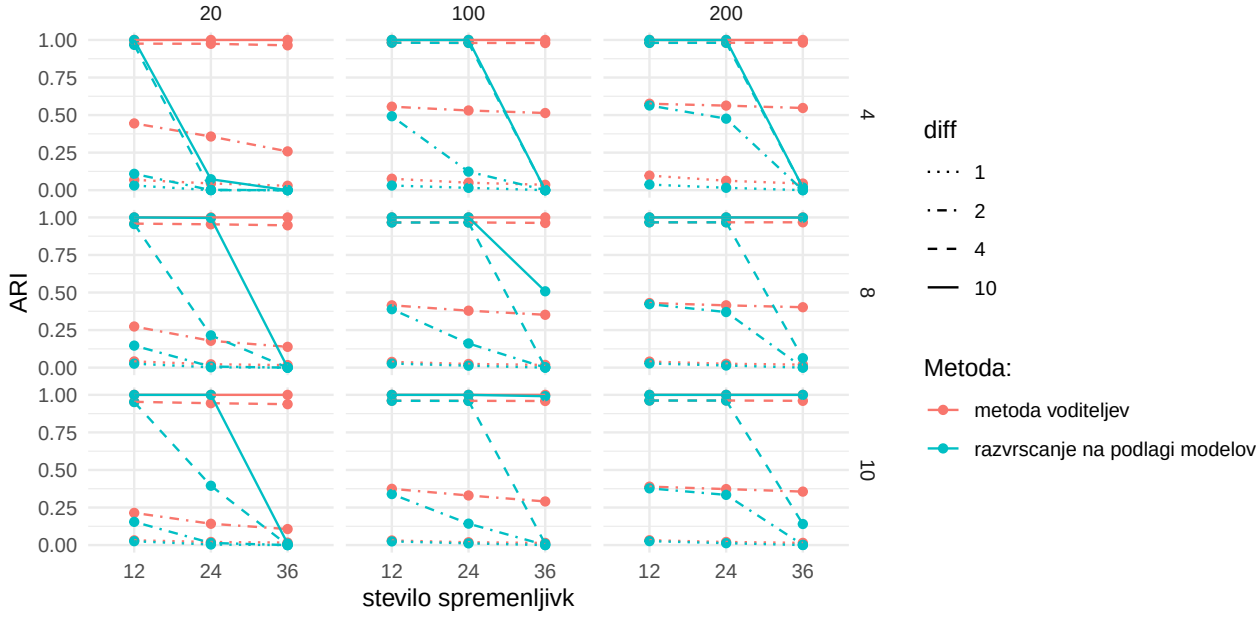
znotraj skupine(WSS) in proporcija vsote kvadratov znotraj skupine($PWSS$). (Zadnji dve meri sva zaradi "neprimernosti" kasneje opustila in ju tudi nisva obravnavala)
Prilagojeni Randovi indeksi zavzemajo vrednosti na intervalu $[-1, 1]$ in želimo si, da so čim bližje 1, torej da gre za dobro ujemanje med razvrstitvami, kar je boljše od naključnega. Pri meri WSS si želimo majhne vrednosti, saj to pomeni, da so skupine bolj kompaktne in s tem so si točke znotraj skupine bolj podobne. Gre sicer za mero, ki je pristranska in jo nekatere metode optimizirajo(ravno metoda *kmeans*). Mera $PWSS$ pa oceni delež variabilnosti v podatkih, ki ga pojasnjujejo skupine, v primerjavi z celotno variabilnostjo podatkov in višji kot je, bolje je, saj to pomeni, da so skupine dobro definirane in točke znotraj skupin tesno sledijo svojim centroidom.

Pri primerjavi metod se bova na začetku osredotočala predvsem na prilagojeni Randov indeks(ARI), saj je mera WSS pristranska in jo nekatere metode optimizirajo(ravno metoda *kmeans*).

4 Popravljen Randov indeks (ARI)

4.1 Spreminjanje faktorja ločljivosti med skupinami

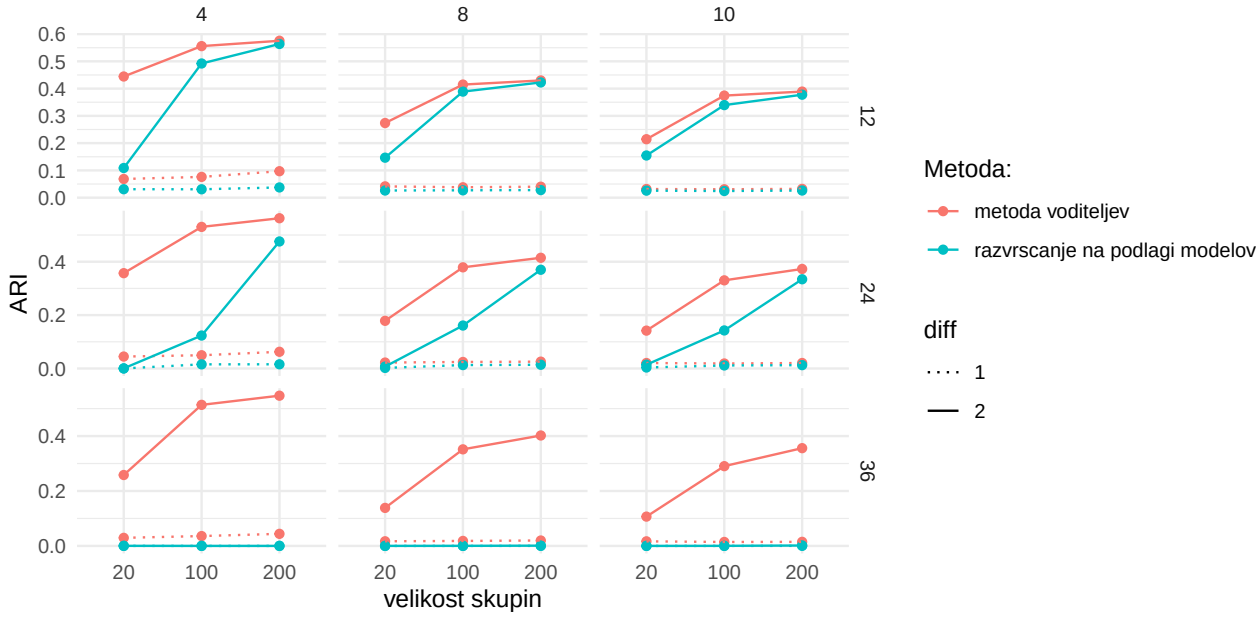
Najprej si poglejmo, kakšne vrednosti ARI imamo, če spremninjamo ločljivost med skupinami(`diff`). Pričakujeva, da bo metoda voditeljev pri večjih vrednostih `diff` nekoliko boljša, saj dobro deluje v primerih, ko so skupine med seboj dobro ločene.



Slika 3: Prikaz ARI vrednosti razdeljen glede na velikos in število skupin.

Takoj opazimo razliko med majhnim in velikim `diff` - pri `diff = 1` so vrednosti ARI dokaj nizke, pri `diff = 10` pa previsoke, kar pomeni, da so skupine preveč ločene in s tem slabi oziroma nerealni rezultati.

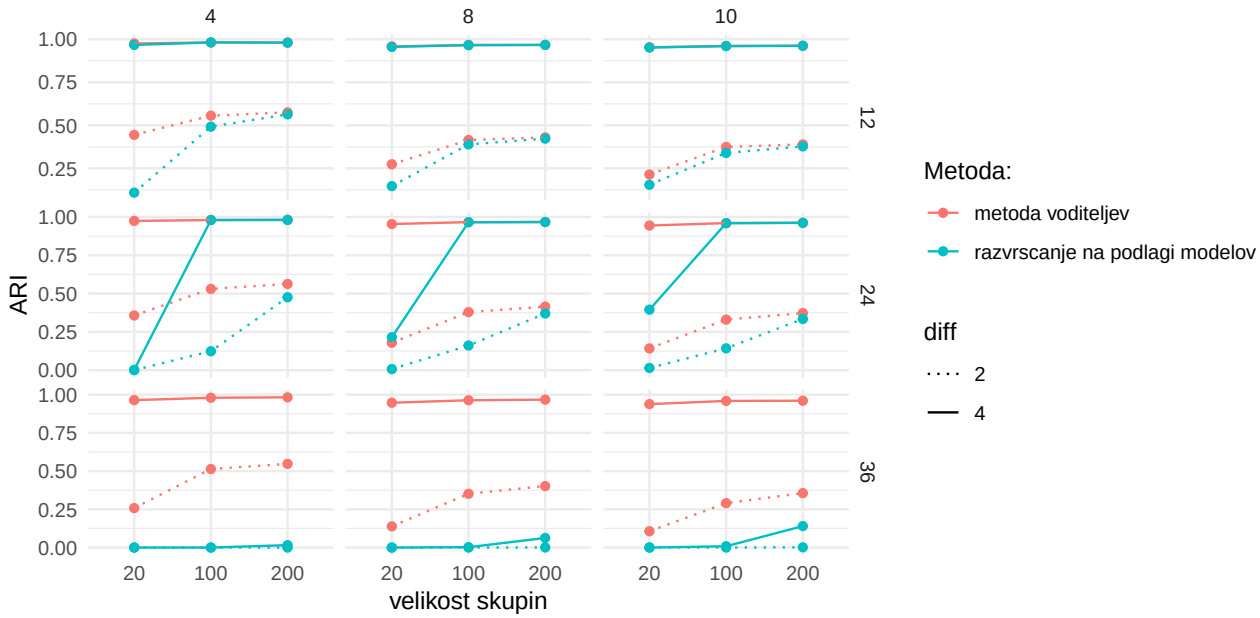
Poglejmo si razliko med `diff = 1` in `diff = 2`.



Slika 4: Prikaz ARI vrednosti razdeljen glede na število spremenljivk in število skupin.

Opazno je *metoda voditeljev* boljša od metode *razvrščanja na podlagi modelov*. Opazimo tudi, da je razlika v metodah večja s povečanjem faktorja `diff` in z manjšimi velikostmi skupin. Z dodajanjem nepomembnih spremenljivk pa se manjša tudi ARI. Lahko bi ocenila, da so rezultati boljši in lažje berljivi za primerjavo metod, če je faktor `diff` nastavljen na `2`, ker pri nastavljeni vrednosti na `1` težko ločimo med metodami, skupine so si med seboj preveč podobne, kar ne ustreza ravno metodi voditeljev in težko ocenimo katera metoda je boljša.

Vseeno pa si poglejmo še razliko med `diff = 2` in `diff = 4`.



Slika 5: Prikaz ARI vrednosti razdeljen glede na število spremenljivk in število skupin.

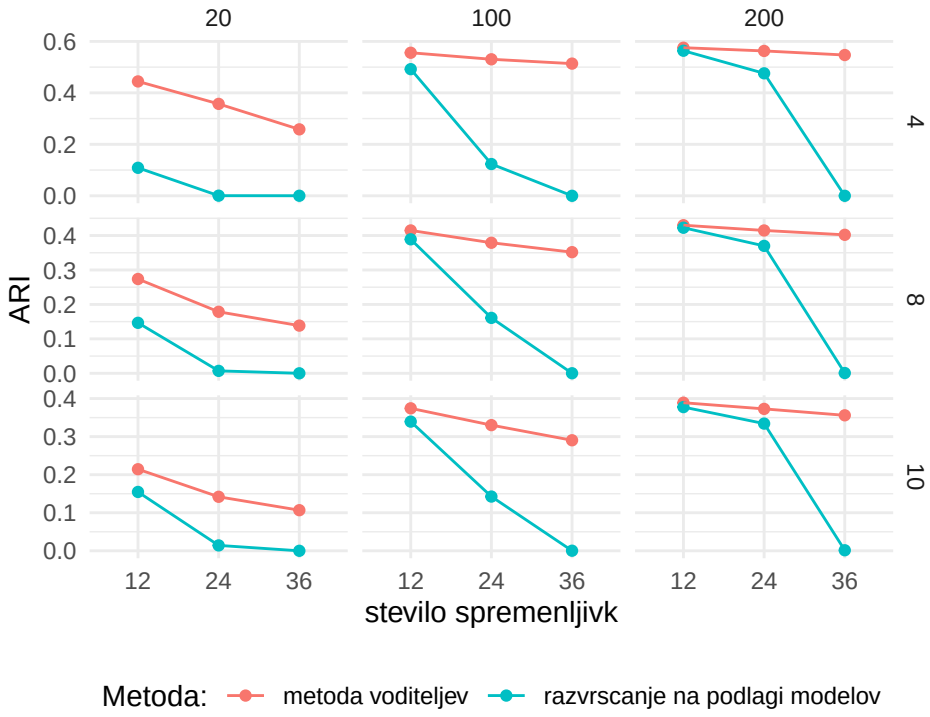
Razlika med metodama pri `diff = 4` je velika, tudi ARI za metodo voditeljev je previsok, da bi lahko rekli, da so rezultati dobri oziroma realni.

Na podlagi zgornje analize se odločiva, da je primerna ločljivost med skupinama nastaljena na `diff = 2` za ustrezno primerjavo med metodama.

4.2 Analiza na podlagi števila in velikosti skupin ter števila spremenljivk

Poglejmo si, kako se spreminja ARI vrednost v primeru, ko spreminjamo število skupin, število spremenljik in velikost skupin, pri tem pa ne upoštevamo, kako so si skupine med seboj različne (`diff`). Vrstice predstavljajo spremembo števila skupin, stolpci pa spremembo velikosti skupin.

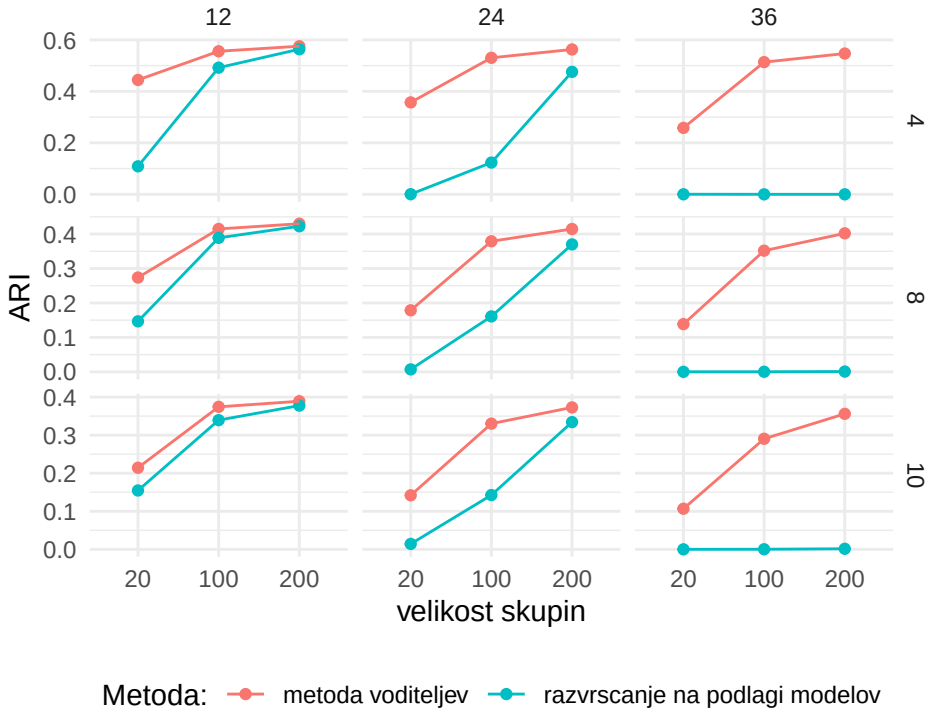
Pričakujeva, da bo pri večjih velikostih skupin indeks večji, saj se ponavadi pri večjih vzorcih moč za razlikovanje skupin običajno poveča. Z večanjem števila skupin pa pričakujeva, da se bo indeks manjšal, ker postane razvrščanje težje, saj pri več skupinah obstaja več kombinacij za razvrščanje, zato je večja verjetnost napačnih ujemanj.



Slika 6: Prikaz ARI vrednosti razdeljen glede na velikost skupin in število skupin.

Vidimo, da je v vseh kombinacijah števila skupin in velikosti skupin trend ARI vrednosti podajoč, ko povečujemo število spremenljivk. Tudi znotraj posamezne velikosti skupin se vrednost manjša z večanjem števila skupin, čeprav je ta razlika dokaj majhna. Vseeno pa lahko opazimo, da ima metoda `k-means` v vseh primerih večjo ali enako ARI vrednost kot metoda `mclust` in je manj občutljiva na povečao števila spremenljivk.

Prikažimo vrednosti še tako, da vrstice predstavljajo spremembo števila skupin, stolpci pa spremembo števila spremenljivk.



Slika 7: Prikaz ARI vrednosti razdeljen glede na število spremenljivk in število skupin.

Jasno se vidi, kako se z večanjem skupin viša tudi ARI, kar sva pričakovala. Na zgornjem grafu pa lahko bolje opazimo očitno razliko med metodama pri večjem številu spremenljivk. Metoda *razvrščanje na podlagi modelov* ima zelo nizke vrednosti indeksa, še posebaj pri 4 skpinah in 36 spremenljivkah(tudi za velike skupine). Pri manjšem številu spremenljivk(npr. 12) pa je razlika med metodama zelo majhna, še posebaj pri velikih skupinah(npr. velikost skupin je 200).

Več spremenljivk naj bi pomagalo izboljšati ARI, vendar samo, če te spremenljivke dejansko prispevajo k razlikovanju skupin. Ker pa sva dodajala t.i. nepomembne spremenljivke(spremenljivke, ki imajo enako porazdelitev v vseh skupinah) pa se indeks ne boljša, pri metodi *razvrščanje na podlagi modelov* se celo manjša.

5 ANOVA

S pomočjo ANOVA testa bova med seboj primerjala metodi za razvrščanja v skupine. Najprej si bova ogledala in med seboj primerjala statistično značilnost spremenljivk (faktorji) in njihove kombinacije, nato pa še modela med seboj.

Tabela 1: Prikaz statistične značilnosti spremenljivk za k-means in mclust

faktorji	**k-means**	**mclust**
stevilo.spremenljivk	0.0000000	0
velikost.skupin	0.0000000	0
stevilo.skupin	0.0000000	0
diff	0.0000000	0
stevilo.spremenljivk:velikost.skupin	0.0000000	0
stevilo.spremenljivk:stevilo.skupin	0.0000001	0
velikost.skupin:stevilo.skupin	0.7387704	0
stevilo.spremenljivk:diff	0.0000000	0
velikost.skupin:diff	0.0000000	0
stevilo.skupin:diff	0.0000000	0
stevilo.spremenljivk:velikost.skupin:stevilo.skupin	0.0000000	0
stevilo.spremenljivk:velikost.skupin:diff	0.0000000	0
stevilo.spremenljivk:stevilo.skupin:diff	0.0000000	0
velikost.skupin:stevilo.skupin:diff	0.0000000	0
stevilo.spremenljivk:velikost.skupin:stevilo.skupin:diff	0.0000000	0

Iz tabele lahko razberemo, da se statistična značilnost posameznih spremenljivk in njihovih kombinacij skoraj ne razlikuje glede na metodo. Pri metodi *razvrščanje na podlagi modelov* dobimo statistično statistično značilnost ($\alpha = 0.05$) v vseh primerih. Prav tako to velja za metodo *metoda voditeljev*, z izjemo interakccije `velikost.skupin:stevilo.skupin`, ki ni statistično značilna.

Sedaj med seboj primerjajmo še modela.

Tabela 2: Primerjava modelov za ANOVA

metoda	**Res.Df**	**RSS**	**Df**	**SS**
mclust	10692	32.025311	NA	NA
kmeans	10692	9.084794	0	22.94052

Če med seboj primerjamo modela, opazimo da se razlikujeta v vrednosti `RSS`. Ta predstavlja variabilnost, ki jo z modelom nismo uspeli pojasniti. Torej želimo si da je vrednost manjša, kar smo v našem primeru dosegli z metodo *metoda voditeljev*.

6 Linearni mešani modeli

Primerjajmo metodi za razvrščanja v skupine še s pomočjo linearnih mešanih modelov.

Tabela 3: Prikaz statistične značilnosti spremenljivk za k-means in mclust

faktorji	**k-means**	**mclust**
stevilo.spremenljivk	0.0000000	0
velikost.skupin	0.0000000	0
stevilo.skupin	0.0000000	0
diff	0.0000000	0
stevilo.spremenljivk:velikost.skupin	0.0000000	0
stevilo.spremenljivk:stevilo.skupin	0.0000001	0
velikost.skupin:stevilo.skupin	0.7376437	0
stevilo.spremenljivk:diff	0.0000000	0
velikost.skupin:diff	0.0000000	0
stevilo.skupin:diff	0.0000000	0
stevilo.spremenljivk:velikost.skupin:stevilo.skupin	0.0000000	0
stevilo.spremenljivk:velikost.skupin:diff	0.0000000	0
stevilo.spremenljivk:stevilo.skupin:diff	0.0000000	0
velikost.skupin:stevilo.skupin:diff	0.0000000	0
stevilo.spremenljivk:velikost.skupin:stevilo.skupin:diff	0.0000000	0

Iz tabele lahko razberemo, da se statistična značilnost posameznih spremenljivk in njihovih kombinacij skoraj ne razlikuje glede na metodo. Rezultati so podobni kot v zgornjem razdelku. Pri metodi *razvrščanje na podlagi modelov* dobimo statistično statistično značilnost ($\alpha = 0.05$) v vseh primerih. Prav tako to velja za metodo *metoda voditeljev*, z izjemo interakccije `velikost.skupin:stevilo.skupin`, ki ni statistično značilna.

Sedaj med seboj primerjajmo še modela.

Tabela 4: Primerjava modelov za ANOVA

metoda	**npar**	**AIC**	**BIC**	**logLik**	**deviance**
kmeans	110	-45607.01	-44805.40	22913.50	-45827.01
mclust	110	-31995.30	-31193.69	16107.65	-32215.30

Tokrat modela med seboj primerjamo glede na `AIC` oz. `BIC` oz. `logLik`. Želimo da je vrednost `logLik` čim večja oziroma vrednosti `AIC` in `BIC` čim manjši. To v našem primeru velja za metodo *metoda voditeljev*.

7 Zaključek

Med seboj sva s pomočjo mere prilagojeni Randov indeks (ARI) primerjala dve metodi za razvrščanje v skupine, in sicer *metoda voditeljev* ter *razvrščanje na podlagi modelov*.

Zanimalo naju je kako na metodi vpliva dodajanje nepomembnih spremenljivk, torej tistih, ki imajo enako porazdelitev v vseh skupinah. Videla sva, da z dodajanjem nepomembnih spremenljivk se rezultat oz. pravila razvrstitev posameznih točk v skupine poslabša.
Na vrednost ARI pa ne vpliva negativno samo število nepomembnih spremenljivk, ampak tudi število skupin v katere želimo razporediti posamezne točke. S tem sva potrdila tudi najno predvidevanje, da se bo z večanjem števila skupin indeks (ARI) manjšal, ker postane razvrščanje težje, saj pri več skupinah obstaja več kombinacij za razvrščanje, zato je večja verjetnost napačnih ujemanj. V primeru, da je število skupin veliko in so te blizu skupaj, je verjetnost za napačno razporeditev velika.
Torej na 'kvaliteto' razporeditve v skupine vpliva tudi ločljivost med skupinami (`diff`). V primeru, da so skupine blizu skupaj metodi slabo razlikujeta posameznimi skupinami. To pa se z večanjem števila vseh spremenljivk še poslabša. Večja kot je ločljivost med skupinami, bolje lahko razlikujemo med skupinami. Vendar pa ko `diff` dosega velike vrednosti, je primerjava med metodami skoraj nemogoča, saj ne glede na

metodo je ARI vrednost visoka oz blizu 1.
Z večanjem velikosti skupin pa se pričakovano povečuje vrednost indeksa.

Vseeno pa se je glede na analizo izkazalo, da metoda *metoda voditeljev* nekoliko bolje razlikuje med skupinamo kot metoda *razvrščanje na podlagi modelov*. Ta je imela v vseh kombinacijah parametrov, ki sva jih spreminjala, višje ali enake vrednosti, kot smo jih dobili po metodi *razvrščanje na podlagi modelov*. Prav tako je primerjava z ANOVO v obeh primerih pokazala, da je model, ki smo ga oblikvali na podlagi metode *metoda voditeljev* primernejši (pojasnimo več variabilnosti).