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. Za metodo razvrščanje na podlagi modelov sva pustila, da izbere najboljši model na podlagi BIC vrednosti. Število skupin pri kateri se računa BIC vrednost sva določila glede na trenutne nastavitve (`settings`) v iteraciji. 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. Ker si želiva rezultate, kjer bo ena metoda delovala bolje od druge, bova spreminjala tudi korelacije med spremenljivkami , $cor = c(0, 0.2, 0.9)$.

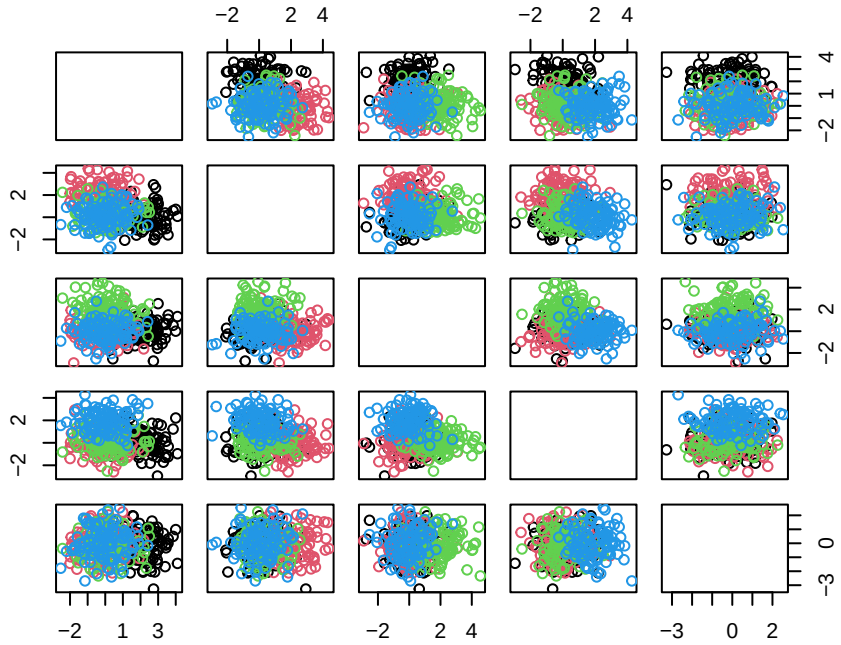
Faktorji, ki jih bova še spreminjala so:

- število skupin, $k = (4, 8, 14)$,
- velikosti skupin, $n = (20, 100, 200)$, pri čemer bodo imele vse skupine vedno enako velikost in
- število neinformativnih spremenljivk, $v = (1, 4, 10)$.

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.

Torej če opišemo postopek generiranja podatkov. Število informativnih spremenljivk sva nastavila na število skupin, število neinformativnih spremenljivk pa bova tekom simulacije spreminjala, saj naju tudi zanima vpliv dodajanja le-teh. Vse spremenljivke generirava torej iz bivariatne multivariatne normalne porazdelitve, kjer imajo informativne spremenljivke povprečje enako parametru `diff`, neinformativne pa enako 0. Standardni odklon pa določiva s pomočjo parametra `cor`, za vse spremenljivke enako(informativne in neinformativne). Število elementov oziroma velikost vzorca, ki ga generirava za spremenljivko pa je določen s pomočjo velikosti skupin.

V naslednjih dveh primerih je število **vseh** spremenljivk enako 5, kjer je 1 spremenljivka **neinformativna**, torej je število skupin enako 4. Na grafu so vidne le 4 spremenljivke, ker ena od vseh ni informativna.



Slika 1: Primer generiranih podatkov za 4 skupine, velikosti n = 100, 5 spremenljivk ter diff = 2 in cor = 0.

3 Simulacija

Izvedla sva simulacijo s 100 ponovitvami (za večje število ponovitev se nisva odločila zaradi časovne zahtevnosti) 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 mero prilagojeni Randov indeks(ARI), ki sva jo uporabila za analizo in primerjavo metod med seboj.

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 (vrednosti blizu 0).

Pri primerjavi metod sva se osredotočala predvsem na prilagojeni Randov indeks(ARI), saj so nekatere druge mere kot naprimer *WSS* pristranske in jo nekatere metode optimizirajo(ravno metoda *kmeans*).

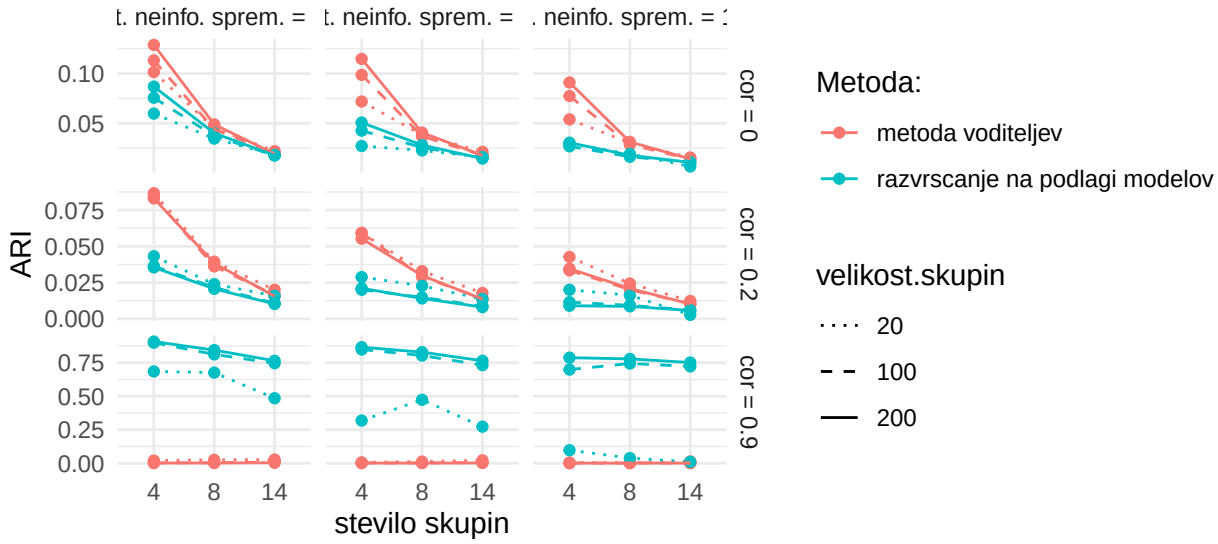
4 Analiza

Analizo rezultatov bova razdelila v več delov, saj pri simulaciji spreminjava veliko število faktorjev (*število skupin*, *število neinformativnih spremenljivk*, *velikost skupin*, razliko med skupinami (*diff*) in korelacijo (*cor*)). Odločila sva se za delitev glede na vrednost povprečja v skupini (razlike povprečja skupin).

Pričakujeva, da bosta metodi pri večjih vrednostih `diff` nekoliko boljše delovali, saj so skupine med seboj dobro ločene oz. je med njimi manj prekrivanja. V nasprotju pa je s spreminjanjem korelacije med spremenljivkami, saj pričakujeva, da bo bo pri niži korelaciji oziroma brez korelacije metoda voditeljev delovala bolje kot razvrščanje na podlagi modelov in ravno nasportno, ko bo korelacija visoka.

4.1 `diff` = 1

Graf prikazuje spreminjanje ARI vrednosti v odvisnosti od števila skupin, pri čemer vrstice predstavljajo korelacijo med spremenljivkami, stolpci pa število neinformativnih spremenljivk.

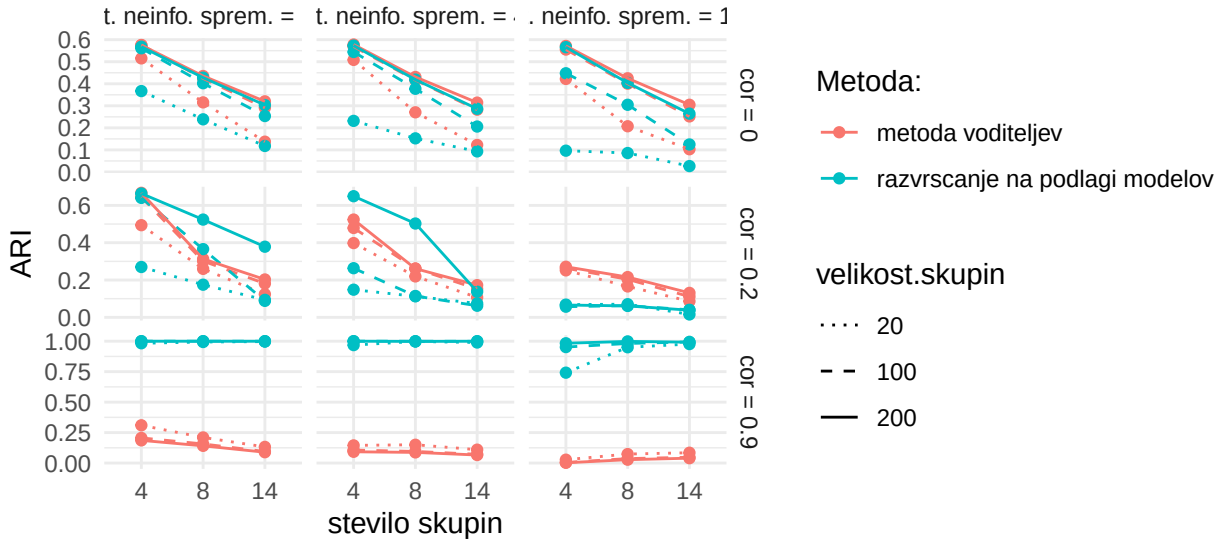


Slika 2: Prikaz ARI vrednosti, ko je razlika povprečja skupin enaka 1.

Najbolj očitna razlika med metodama je vidna s spreminjanjem korelacije med spremenljivkami.
Ko korelacija ni prisotna bi lahko rekli da metoda voditeljev nekoliko bolje razporedi enote v skupine, čeprav je razlika med njima minimalna. Prav tako bi lahko enako trdili v primeru `cor`=2, kjer so razlike majhne in se zmanjšujejo, ko se število neinformativnih spremenljivk povečuje. Ko pa je korelacija med spremenljivkami enaka 0.9, metoda razvrščanja na podlagi modelov veliko bolje razlikuje med skupinami.
V vseh kombinacijah faktorjev je možno videti, da vrednost ARI z naraščanje števila skupin pada. Prav tako vrednost pada, ko število neinformativnih spremenljivk narašča.

4.2 `diff` = 2

Graf prikazuje spreminjanje ARI vrednosti v odvisnosti od števila skupin, pri čemer vrstice predstavljajo korelacijo med spremenljivkami, stolpci pa število neinformativnih spremenljivk.

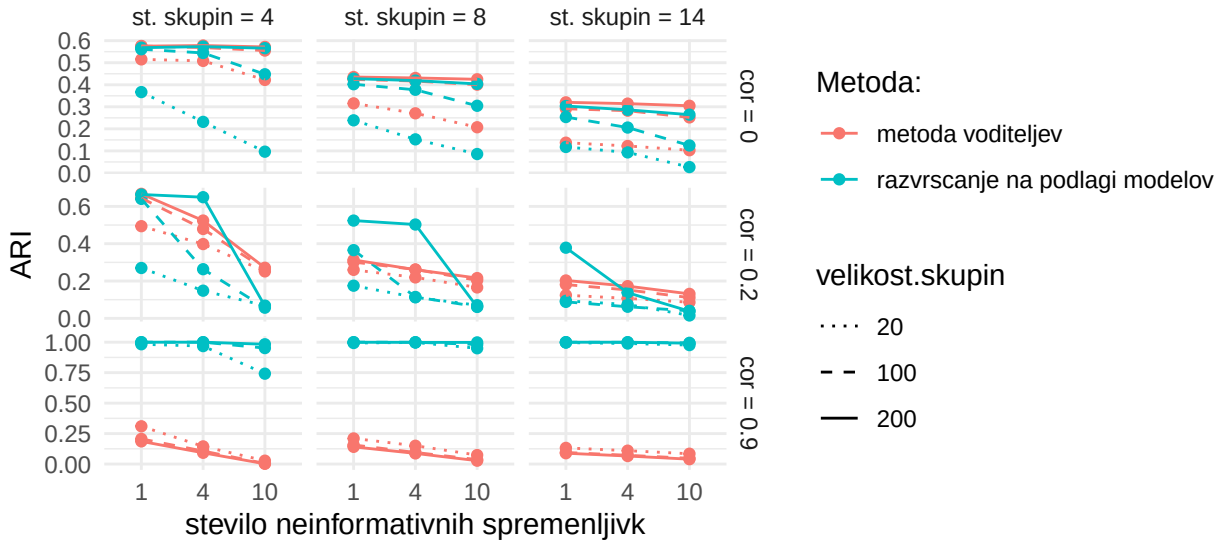


Slika 3: Prikaz ARI vrednosti, ko je razlika povprečja skupin enaka 2.

Tokrat je očitna razlika med metodam vidna le v primeru, ko je korelacija enaka 0.9. V preostalih dveh

primerih pa se vrednosti izračunane po metodi voditeljev in razvrščanja na podlagi modelov med seboj prepletajo. Nekoliko večja razlika je le desnem grafu `cor`=0.2, ko metoda voditeljev v vseh primerih vrne boljše rezultate.

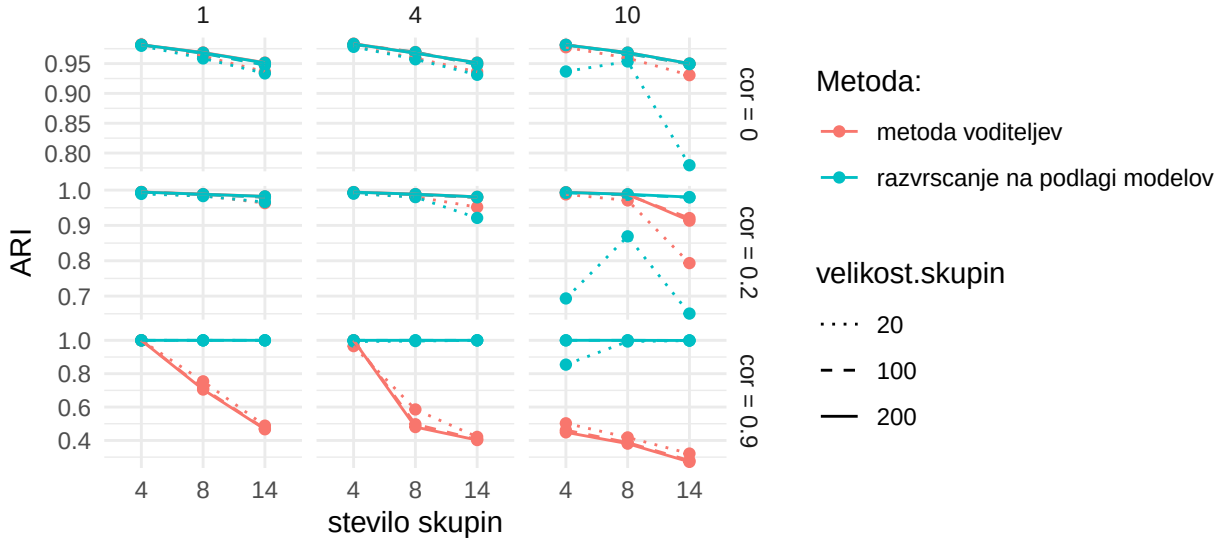
Tako kot v predhodnem razdelku je v vseh kombinacijah faktorjev možno videti, da vrednost ARI z naraščanje števila skupin pada. Prav tako vrednost pada, ko število neinformativnih spremenljivk narašča, kar lahko bolj natančno vidimo tudi na spodnjem grafu



Slika 4: Prikaz ARI vrednosti, ko je razlika povprečja skupin enaka 2 glede na število neinfomrativnih spremenljivk.

4.3 `diff` = 4

Graf prikazuje spreminjanje ARI vrednosti v odvisnosti od števila skupin, pri čemer vrstice predstavljajo korelacijo med spremenljivkami, stolpci pa število neinformativnih spremenljivk.



Slika 5: Prikaz ARI vrednosti, ko je razlika povprečja skupin enaka 4.

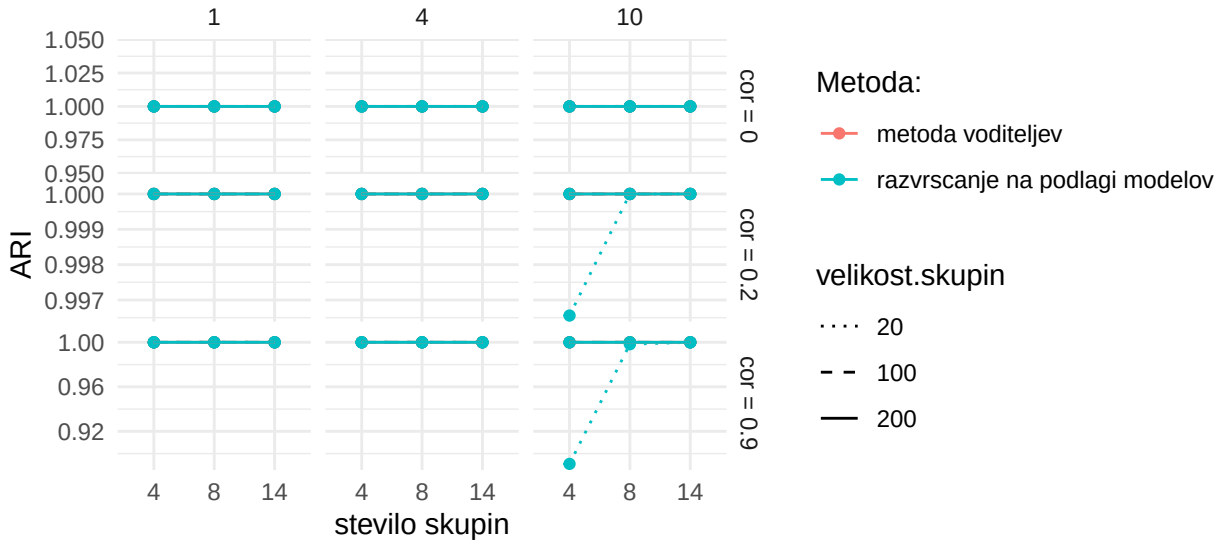
Kot lahko opazimo se z večanjem razlike povprečja skupin, razlike v ARI vrednosti med metodama manjšajo. Izjema je le `cor` = 0.9, kjer z metodo voditeljev ne dobimo dobrih rezultatov.

Če si ogledamo 3. stolpec (*število neinformativnih sprmeneljivk* = 10) opazimo, da metoda *razvrščanja na podlagi modelov* vrne nekoliko slabše rezultate v primeru, ko imamo majhno velikost skupin.

Izpostavimo lahko še zadnjo vrstico (`cor` = 0.9), kjer opazimo, da *metoda voditeljev* vrne dober rezultat, le v primeru, ko je število skupin in število neinformativnih spremenljivk majhno. Ko pa se te dve vrednosti povečata, ARI vrednost dosti pade.

4.4 `diff` = 10

Graf prikazuje spreminjanje ARI vrednosti v odvisnosti od števila skupin, pri čemer vrstice predstavljajo korelacijo med spremenljivkami, stolpci pa število neinformativnih spremenljivk.



Slika 6: Prikaz ARI vrednosti, ko je razlika povprečja skupin enaka 10.

Tokrat dobimo dokaj nezanimive rezultate, saj so vse vrednosti blizu 1, kar naj bi pomenilo smo enote popolnoma pravilno razporedili v skupine. Zaključimo lahko le, da če imamo dovolj ločene skupine (povprečja) je vseeno katero metodo uporabimo, saj obe dobro ločujeta med skupinami.

4.5 Analiza

Iz zgornjih grafov je torej očitno, da spreminjanje faktorja `diff` in `cor` znatno vplivata na to, katera metoda je boljša za razvrščanje v skupine. Pri manjši korelaciji oz. kjer korelacije ni in je ločljivost med skupinami majhna(`diff` = 1), bolje deluje metoda voditeljev, pri večji pa metoda razvršanja na podlagi modelov. Ko pa se ločljivost med skupinami poveča(torej bolje ločimo med skupinami, `diff` = 2, 4, 10), pa je razlika med metodama pri majhnih korelacijah minimalna oz. je skoraj da ni, pokaže se šele pri večji korelaciji, ko je metoda razvrščanja na podlagi modelov vidno boljša.
Faktor `diff` tudi močno vpliva na koeficient ARI, ki je pri večjih ločljivostih med skupinami (skoraj) enak 1, kar kaže na preprleganje modelov oz. predobro razvršanje v skupine i je tud vseeno, ne glede na ostale fakotre, katero metodo vzamemo.

Na mero ARI vpliva tudi večanje števila skupin v katere razvršačmo enote - ta se z večanjem skupin vidno manjša, torej bi lahko rekla, da razvrščanje v (pre)več skupin ne deluje dobro z obe matodama.
Prav tako se kakovost razvrščanja v skupine pri obeh metodah manjša(nobena ni boljša) z dodajanjem neinformativnih spremenljivk.

5 Primerjava metod za razvrščanje v skupine

Metodi bova primerjala s pomočjo ANOVA testa in linearnih mešanih modelov, kjer pričakujeva podobne rezultate. Na podlagi zgornjih grafov predvidevava, da bodo vrednosti fakoraj `diff` in `cor` nekoliko bolj statistično pomembno vplivala na rezultate analize, kot vrednosti drugih faktorjev.

5.1 ANOVA

V spodnji tabeli lahko vidimo količina variabilnosti glede na faktorje, ki jih testiramo in informacijo o statistični značilnosti posameznih spremenljivk in njihovih kombinacij. Razvidno je, da so nekateri faktorji in kombinacije le teh statistično pomembi in razlike med različnimi vrednostmi faktorjev vplivajo na rezultate. Izrazito tako odstopata faktorji `diff` in `cor` ter njuna interakcija, kar bi lahko rekla, da vsa ugotovila že pri zgornji analizi. Prav tako nekoliko bolj na rezultate analize vplivajo interakcije z drugimi faktorji, kjer sta `diff` in `cor` tudi zraven(npr. interakcija `velikost.skupin:diff:cor`). Prav tako so vrednosti faktorja `stevilo.skupin` tud statistično pomembni in vplivajo bolj na rezultate analize, katera metoda za razvrščanje v skupine je boljša.

Tabela 1: Prikaz rezultatov ANOVA testa.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
diff	3	8459.4961960	2819.8320653	82731.011678	0.0000000
diff:cor	6	532.7648857	88.7941476	2605.130197	0.0000000
cor	2	127.6674771	63.8337385	1872.817121	0.0000000
stevilo.skupin	2	67.4521945	33.7260973	989.489474	0.0000000
velikost.skupin:diff:cor	12	66.1068587	5.5089049	161.625680	0.0000000
stevilo.skupin:diff:cor	12	59.9813774	4.9984481	146.649396	0.0000000
stevilo.neinformativnih.sprem	2	51.7574587	25.8787293	759.255661	0.0000000
stevilo.skupin:diff	6	48.9910091	8.1651682	239.557750	0.0000000
velikost.skupin	2	36.1792458	18.0896229	530.731181	0.0000000
stevilo.neinformativnih.sprem:diff:cor	12	25.2243454	2.1020288	61.671392	0.0000000
stevilo.neinformativnih.sprem:diff	6	24.7986902	4.1331150	121.261402	0.0000000
velikost.skupin:diff	6	23.4129226	3.9021538	114.485232	0.0000000
stevilo.neinformativnih.sprem:velikost.skupin:diff:cor	24	17.0985312	0.7124388	20.902231	0.0000000
stevilo.neinformativnih.sprem:stevilo.skupin:diff:cor	24	11.9128094	0.4963671	14.562906	0.0000000
stevilo.neinformativnih.sprem:cor	4	10.8251589	2.7062897	79.399794	0.0000000
stevilo.neinformativnih.sprem:stevilo.skupin	4	7.9764832	1.9941208	58.505481	0.0000000
stevilo.neinformativnih.sprem:stevilo.skupin:diff	12	7.5342930	0.6278577	18.420709	0.0000000
stevilo.neinformativnih.sprem:velikost.skupin:stevilo.skupin:diff:cor	48	4.2976719	0.0895348	2.626861	0.0000000
stevilo.neinformativnih.sprem:velikost.skupin:cor	8	4.2522474	0.5315309	15.594578	0.0000000
stevilo.neinformativnih.sprem:velikost.skupin:diff	12	4.2457012	0.3538084	10.380380	0.0000000
stevilo.neinformativnih.sprem:stevilo.skupin:cor	8	3.6439902	0.4554988	13.363872	0.0000000
stevilo.neinformativnih.sprem:velikost.skupin	4	3.4892031	0.8723008	25.592420	0.0000000
velikost.skupin:stevilo.skupin:diff	12	2.0090603	0.1674217	4.911983	0.0000000
stevilo.skupin:cor	4	2.0021987	0.5005497	14.685620	0.0000000
velikost.skupin:stevilo.skupin:diff:cor	24	1.9302733	0.0804281	2.359678	0.0001888
stevilo.neinformativnih.sprem:velikost.skupin:stevilo.skupin:cor	16	1.8132419	0.1133276	3.324917	0.0000070
stevilo.neinformativnih.sprem:velikost.skupin:stevilo.skupin:diff	24	1.5037998	0.0626583	1.838332	0.0074037
velikost.skupin:cor	4	1.2115417	0.3028854	8.886351	0.0000004
velikost.skupin:stevilo.skupin	4	1.0655540	0.2663885	7.815569	0.0000027
velikost.skupin:stevilo.skupin:cor	8	0.3658987	0.0457373	1.341887	0.2171892
stevilo.neinformativnih.sprem:velikost.skupin:stevilo.skupin	8	0.3643059	0.0455382	1.336046	0.2200132

5.2 Linearni mešani modeli

Primerjajmo metodi še s pomočjo linearnih mešanih modelov, kjer je iz spodnje tabele razvidno, da res dobimo, glede na razvrstitev statistične pomembnosti, enake faktorje kot pri zgornjem testu.

Tabela 2: Prikaz rezultatov linearnih mešanih modelov.

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
diff	7950.2635478	2650.0878493	3	64476	77750.888609	0.0000000
diff:cor	532.7648670	88.7941445	6	64476	2605.130105	0.0000000
cor	127.6673005	63.8336503	2	64476	1872.814531	0.0000000
stevilo.skupin	67.4521888	33.7260944	2	64476	989.489390	0.0000000
velikost.skupin:diff:cor	66.1068541	5.5089045	12	64476	161.625669	0.0000000
stevilo.skupin:diff:cor	59.9813770	4.9984481	12	64476	146.649395	0.0000000
stevilo.neinformativnih.sprem	51.7575440	25.8787720	2	64476	759.256913	0.0000000
stevilo.skupin:diff	48.9910324	8.1651721	6	64476	239.557864	0.0000000
velikost.skupin	36.1791655	18.0895828	2	64476	530.730004	0.0000000
stevilo.neinformativnih.sprem:diff:cor	25.2243444	2.1020287	12	64476	61.671389	0.0000000
stevilo.neinformativnih.sprem:diff	24.7987021	4.1331170	6	64476	121.261460	0.0000000
velikost.skupin:diff	23.4129168	3.9021528	6	64476	114.485204	0.0000000
stevilo.neinformativnih.sprem:velikost.skupin:diff:cor	17.0985320	0.7124388	24	64476	20.902233	0.0000000
stevilo.neinformativnih.sprem:stevilo.skupin:diff:cor	11.9128087	0.4963670	24	64476	14.562905	0.0000000
stevilo.neinformativnih.sprem:cor	10.8251665	2.7062916	4	64476	79.399851	0.0000000
stevilo.neinformativnih.sprem:stevilo.skupin	7.9764790	1.9941197	4	64476	58.505450	0.0000000
stevilo.neinformativnih.sprem:stevilo.skupin:diff	7.5342927	0.6278577	12	64476	18.420708	0.0000000
stevilo.neinformativnih.sprem:velikost.skupin:cor	4.2522497	0.5315312	8	64476	15.594587	0.0000000
stevilo.neinformativnih.sprem:velikost.skupin:diff	4.2456996	0.3538083	12	64476	10.380377	0.0000000
stevilo.neinformativnih.sprem:stevilo.skupin:cor	3.6439913	0.4554989	8	64476	13.363876	0.0000000
stevilo.neinformativnih.sprem:velikost.skupin	3.4892063	0.8723016	4	64476	25.592443	0.0000000
velikost.skupin:stevilo.skupin:diff	2.0090598	0.1674216	12	64476	4.911981	0.0000000
stevilo.skupin:cor	2.0021976	0.5005494	4	64476	14.685611	0.0000000
velikost.skupin:stevilo.skupin:diff:cor	1.9302734	0.0804281	24	64476	2.359678	0.0001888
stevilo.neinformativnih.sprem:velikost.skupin:stevilo.skupin:cor	1.8132419	0.1133276	16	64476	3.324917	0.0000070
stevilo.neinformativnih.sprem:velikost.skupin:stevilo.skupin:diff	1.5037996	0.0626583	24	64476	1.838332	0.0074037
velikost.skupin:cor	1.2115369	0.3028842	4	64476	8.886316	0.0000004
velikost.skupin:stevilo.skupin	1.0655530	0.2663883	4	64476	7.815561	0.0000027
velikost.skupin:stevilo.skupin:cor	0.3658985	0.0457373	8	64476	1.341886	0.2171895
stevilo.neinformativnih.sprem:velikost.skupin:stevilo.skupin	0.3643057	0.0455382	8	64476	1.336045	0.2200136

6 Zaključek

Pri primerjavi dveh metod za razvrščanje v skupine, *metodo voditeljev* ter *razvrščanje na podlagi modelov*, na podlagi mere prilagojeni Randov indeks (ARI) sva ugotovila, da na vrednost ARI pa ne vpliva negativno samo število nepomembnih spremenljivk(tistih, ki imajo enako porazdelitev v vseh skupinah), 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.

Na 'kvaliteto' razporeditve v skupine pa najbolj vpliva ločljivost med skupinami (`diff`) in korelacija(`cor`)(glede na ANOVA test in linearne mešane modele), tudi razlika v metodah je očitna. Če imamo dokaj normalno ločljivost med skupinami(npr. `diff` = 2) in nimamo korelacije je nekoliko boljša metoda voditeljev, ampak že opri majhnih korelacijah, pa je razvrščanje na podlagi modelov boljše. Torej, če imamo med spremenljivkami korelacijo je bolje za razvrščanje v skupine uporabiti metodo razvršanja na podlagi modelov, v nasprotnem primeru pa metodo voditeljev.