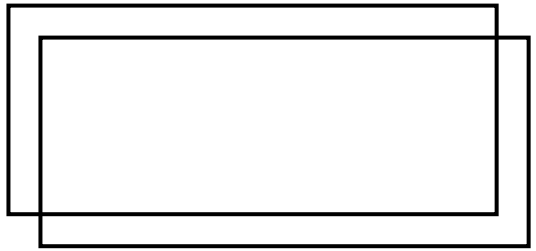


Uspešnost LDA pri različnih mehanizmih manjkajočih vrednosti

Gregor Vavdi, Nace Vreček

Ljubljana, Februar 2020

CILJ NALOGE

Primerjanje uspešnosti razvrščanja v skupine z linearno diskriminantno analizo ob uporabi različnih metod za imputacijo manjkajočih vrednostih ter različnih mehanizmih manjkajočih vrednosti.

SIMULACIJA

FIKSNI DEJAVNIKI

- Porazdelitev spremenljivk: multivariatna normalna
- Število spremenljivk: 4 – X_1, X_2, X_3, X_4
- Število skupin: 3 – vsaka 100 enot
- Povprečja po skupini: {3, 5, 7}
- Korelacija med spremenljivkami: 0.85

SPREMINJAJOČI DEJAVNIKI

- Mehanizem manjkajočih vrednosti: MCAR, NMAR, MAR
- Moč mehanizma manjkajočih vrednosti: {1, 3, 5, 8, 12} *(pri MAR in NMAR)*
- Metode imputacije
- Delež manjkajočih vrednosti: {0.3, 0.4, 0.5, 0.6}

MEHANIZMI MANJKAJOČIH VREDNOSTI

MAR

Mehanizem naključno manjkajočih podatkov. Verjetnost, da vrednost manjka je neodvisna od manjkajoče vrednosti pogojno na vrednosti pri drugih spremenljivkah

Spremenljivka X_1 vpliva na manjkajoče vrednosti X_2, X_3 in X_4.

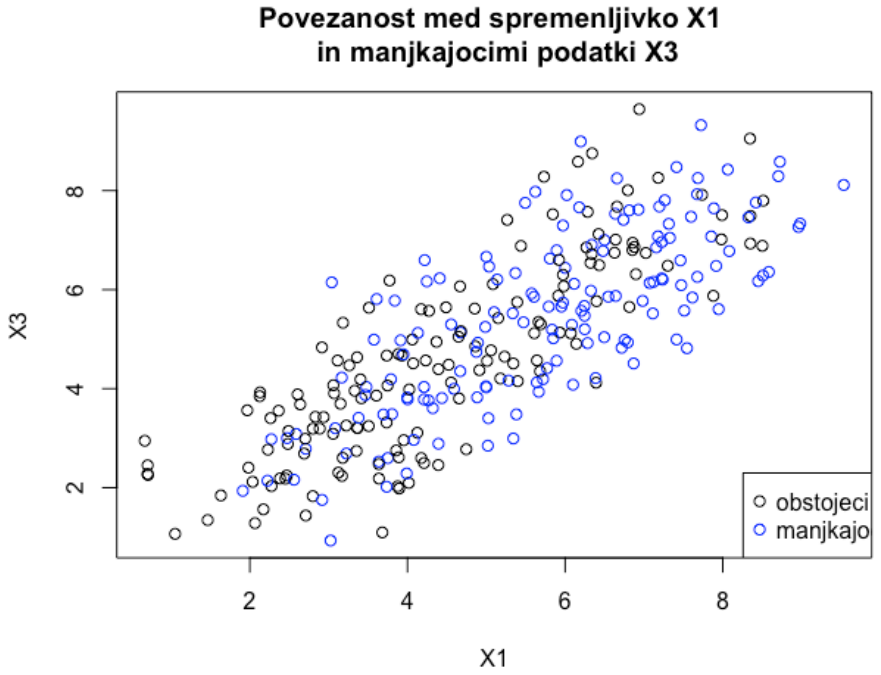
Povezanost med vrednostmi X_1 in manjkajočimi vrednostmi X_2 (X_3 in X_4) sva imenovala **moč** mehanizma.

$$p = \left| \frac{(X_1)^m}{(\sum_{i=1}^{n} X_1)^m} \right|$$

MEHANIZMI MANJKAJOČIH VREDNOSTI

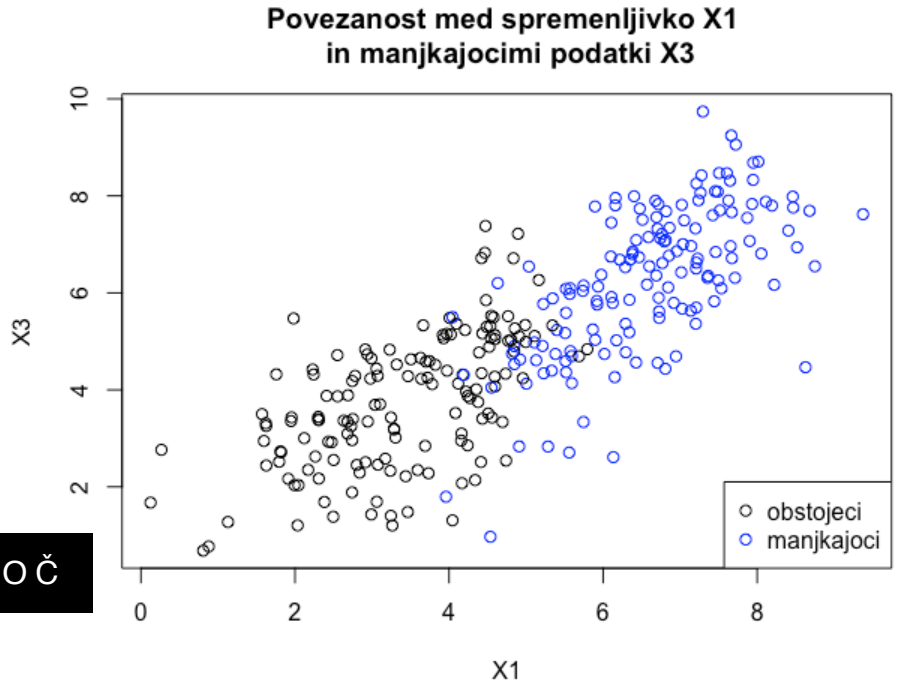
MAR

Mehanizem naključno manjkajočih podatkov. Verjetnost, da vrednost manjka je neodvisna od manjkajoče vrednosti pogojno na vrednosti pri drugih spremenljivkah



MAJHNA MOČ

m = 1



m = 12

VELIKA MOČ

MEHANIZMI MANJKAJOČIH VREDNOSTI

MAR

Mehanizem naključno manjkajočih podatkov. Verjetnost, da vrednost manjka je neodvisna od manjkajoče vrednosti pogojno na vrednosti pri drugih spremenljivkah

MAJHNA MOČ

Delež NA	Skupina I	Skupina II	Skupina III
0.3	19.55	30.53	39.92
0.4	27.07	40.90	52.03
0.5	34.97	51.32	63.71
0.6	43.57	61.90	74.53

Delež NA vrednosti (v %) v posamezni skupini, glede na celotni vzorec pri moči mehanizma m = 1

VELIKA MOČ

Delež NA	Skupina I	Skupina II	Skupina III
0.3	0.46	15.18	74.37
0.4	1.35	29.65	89.00
0.5	3.93	49.60	96.47
0.6	10.13	70.69	99.18

Delež NA vrednosti (v %) v posamezni skupini, glede na celotni vzorec pri moči mehanizma m = 12

MEHANIZMI MANJKAJOČIH VREDNOSTI

NMAR

Mehanizem nenaključno manjkajočih podatkov. Verjetnost, da vrednost manjka je odvisna od manjkajoče vrednosti (torej od spremenljivke, ki ima manjkajočo vrednost).

Spremenljivka X_2 vpliva na manjkajoče vrednosti X_2. Enako za spremenljivko X_3 in X_4.

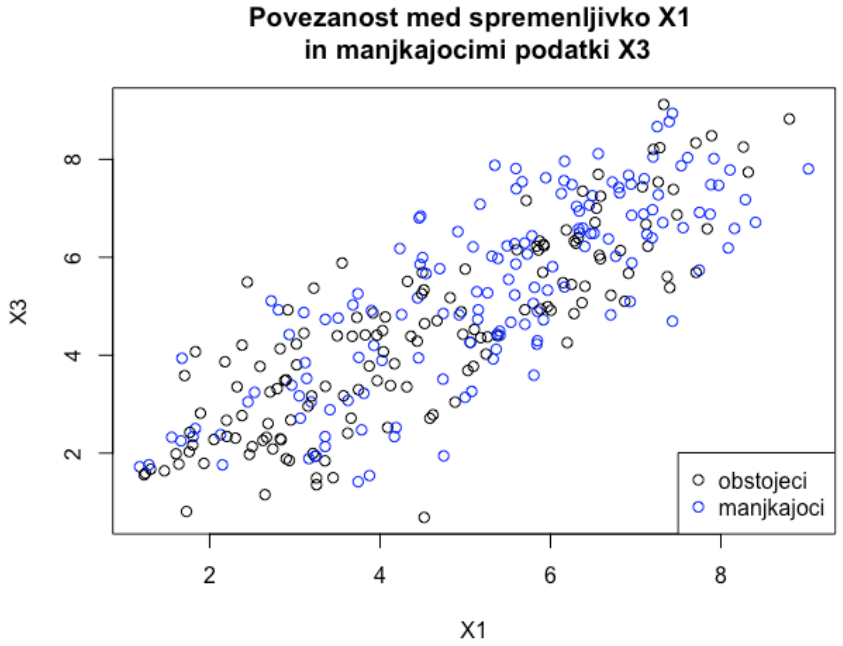
Povezanost med vrednostmi X_i in manjkajočimi vrednostmi X_i sva imenovala **moč** mehanizma, za i = 2, 3, 4.

$$p_i = \left| \frac{(X_i)^m}{(\sum_{i=1}^{n} X_i)^m} \right|$$

MEHANIZMI MANJKAJOČIH VREDNOSTI

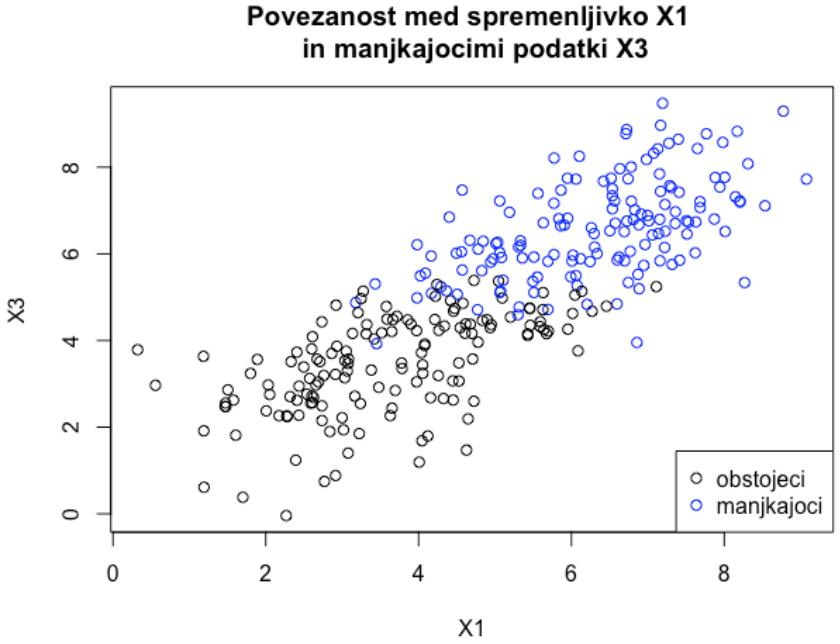
NMAR

Mehanizem nenaključno manjkajočih podatkov. Verjetnost, da vrednost manjka je odvisna od manjkajoče vrednosti (torej od spremenljivke, ki ima manjkajočo vrednost).



MAJHNA MOČ

m = 1



m = 12

VELIKA MOČ

MEHANIZMI MANJKAJOČIH VREDNOSTI

NMAR

Mehanizem nenaključno manjkajočih podatkov. Verjetnost, da vrednost manjka je odvisna od manjkajoče vrednosti (torej od spremenljivke, ki ima manjkajočo vrednost).

MAJHNA MOČ

Delež NA	Skupina I	Skupina II	Skupina III
0.3	19.57	30.47	39.95
0.4	26.76	40.90	52.34
0.5	34.80	51.39	63.81
0.6	43.58	62.06	74.36

Delež NA vrednosti (v %) v posamezni skupini, glede na celotni vzorec pri moči mehanizma m = 1

VELIKA MOČ

Delež NA	Skupina I	Skupina II	Skupina III
0.3	0.48	15.08	74.44
0.4	1.36	29.62	89.02
0.5	3.93	49.59	96.48
0.6	10.07	70.74	99.19

Delež NA vrednosti (v %) v posamezni skupini, glede na celotni vzorec pri moči mehanizma m = 12

MEHANIZMI MANJKAJOČIH VREDNOSTI

MCAR

Mehanizem povsem naključno manjkajočih podatkov. Verjetnost, da določena vrednost manjka je popolnoma neodvisna od manjkajočih vrednosti ter vrednosti pri ostalih spremenljivkah.

- Verjetnost, da določena vrednost manjka je popolnoma neodvisna od vrednosti ki manjka
- Delež manjkajočih vrednosti je bil enak pri vsaki spremenljivki X_2, X_3 in X_4 .
- Pri spremenljivki X_1 manjkajočih vrednosti ni.

METODE IMPUTIRANJA

Analiza na podlagi popolnih enot

- Uporabimo samo enote, ki imajo vse vrednosti na vseh spremenljivkah

Multiple imputacije preko verižnih enačb (MICE)

- manjkajoče vrednosti imputiramo z enostavno metodo
- na podlagi imputiranega podatkovja iz 1. koraka ocenimo model za eno spremenljivko
- z modelom iz 2. koraka imputiramo nove vrednosti za manjkajoče vrednosti
- ponavljamo koraka 2 in 3 dokler se porazdelitve parametrov iz 2. koraka ne stabilizirajo
- m krat ponavljamo korake 1-4

Imputacije preko slučajnih gozdov

- manjkajoče vrednosti napovedujemo preko slučajnih gozdov
- postopek iterativno ponavljamo dokler ne dosežemo konvergence

METODE IMPUTIRANJA

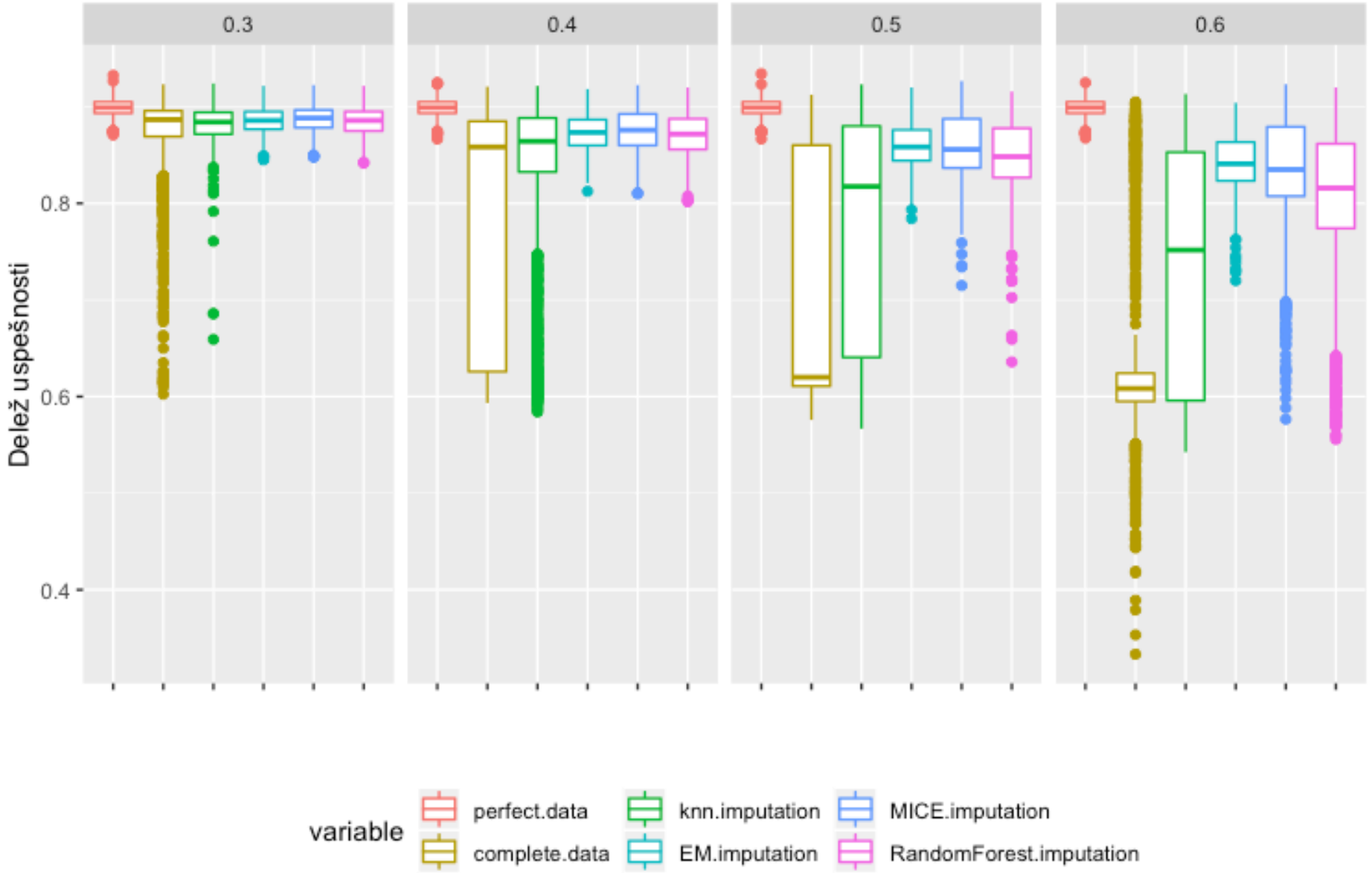
Metoda imputacij na podlagi najbližjih sosedov (kNN)

- metoda, ki temelji na podobnosti med podatki (Beretta & Santaniello, 2016)
- manjkajočo vrednosti ocenimo tako, da pogledamo k najbolj podobnih vrednosti (sosedov) na ostalih spremenljivkah
- kot metodo imputiranja manjkajočih vrednosti sva izbrala obtežena povprečja

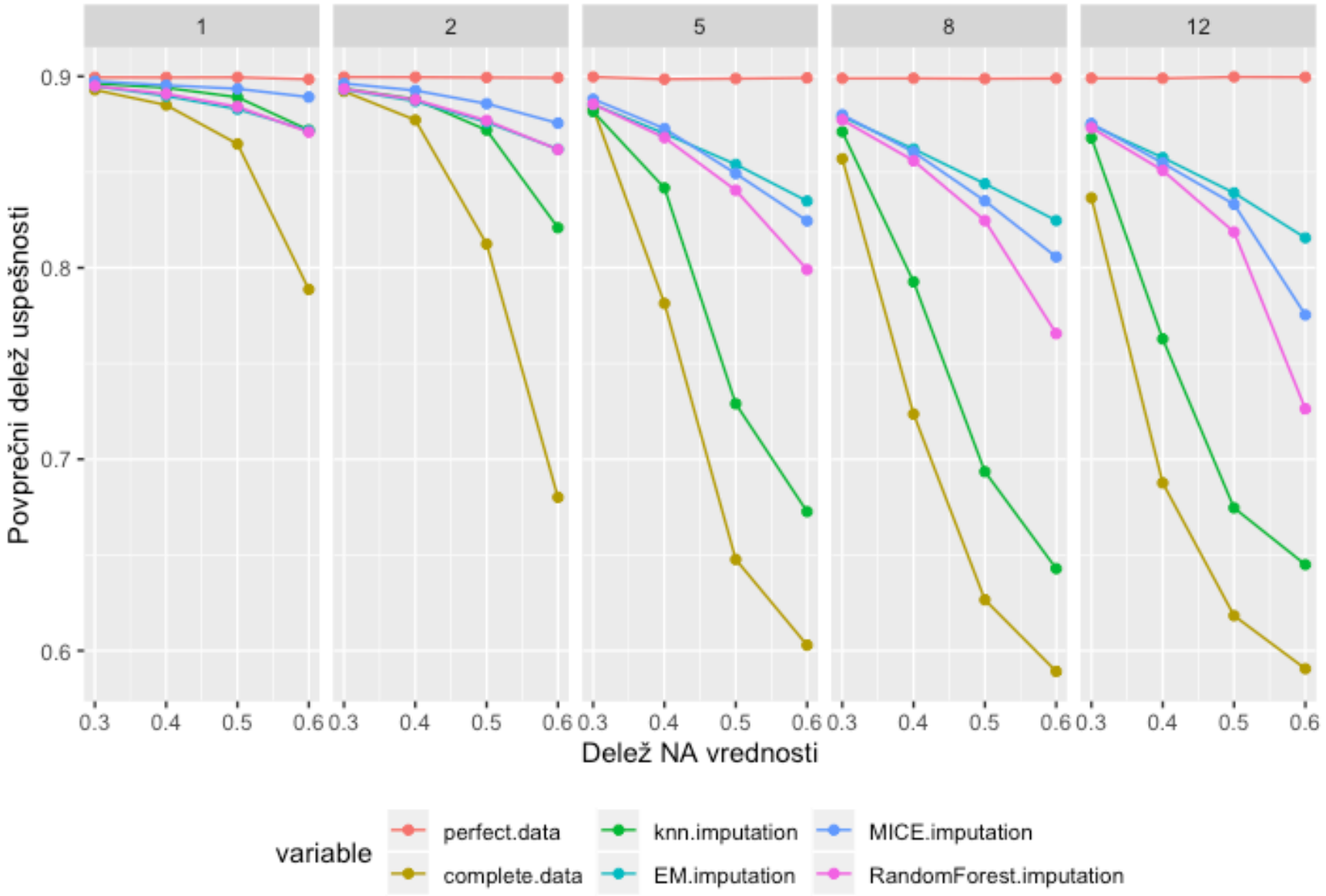
EM algoritem

- predpostavka 1: Opazovane vrednosti so neodvisne v p-razsežnem normalnem prostoru s parametrom mi in sigma.
- predpostavka 2: Podatki manjkajo po mehanizmu MAR, vendar tako, da nikoli ne manjkajo vse komponente.
- uporabila sva knjižnico **norm** in funkcije: *em.norm, in imp.norm*

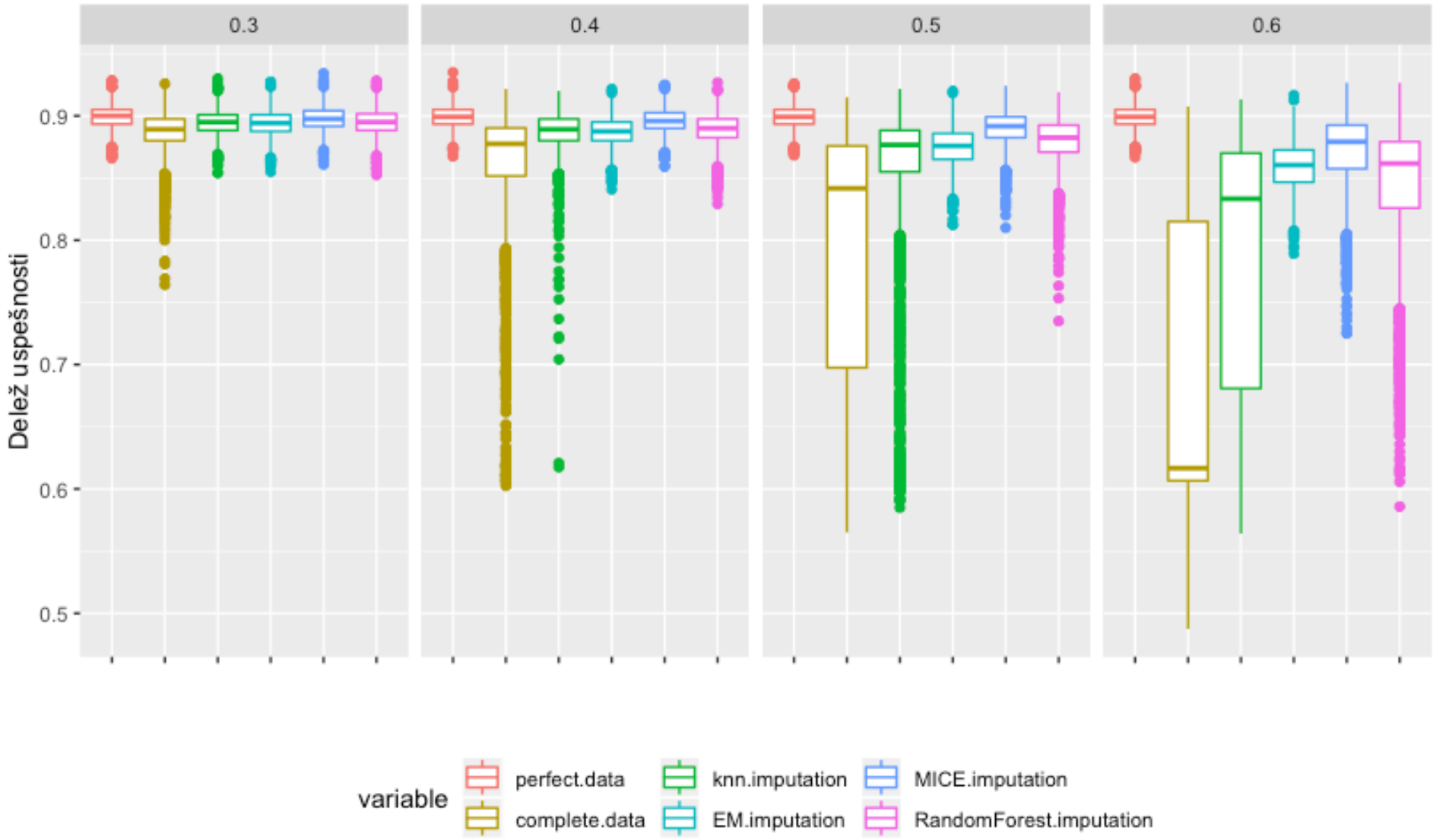
REZULTATI | NMAR



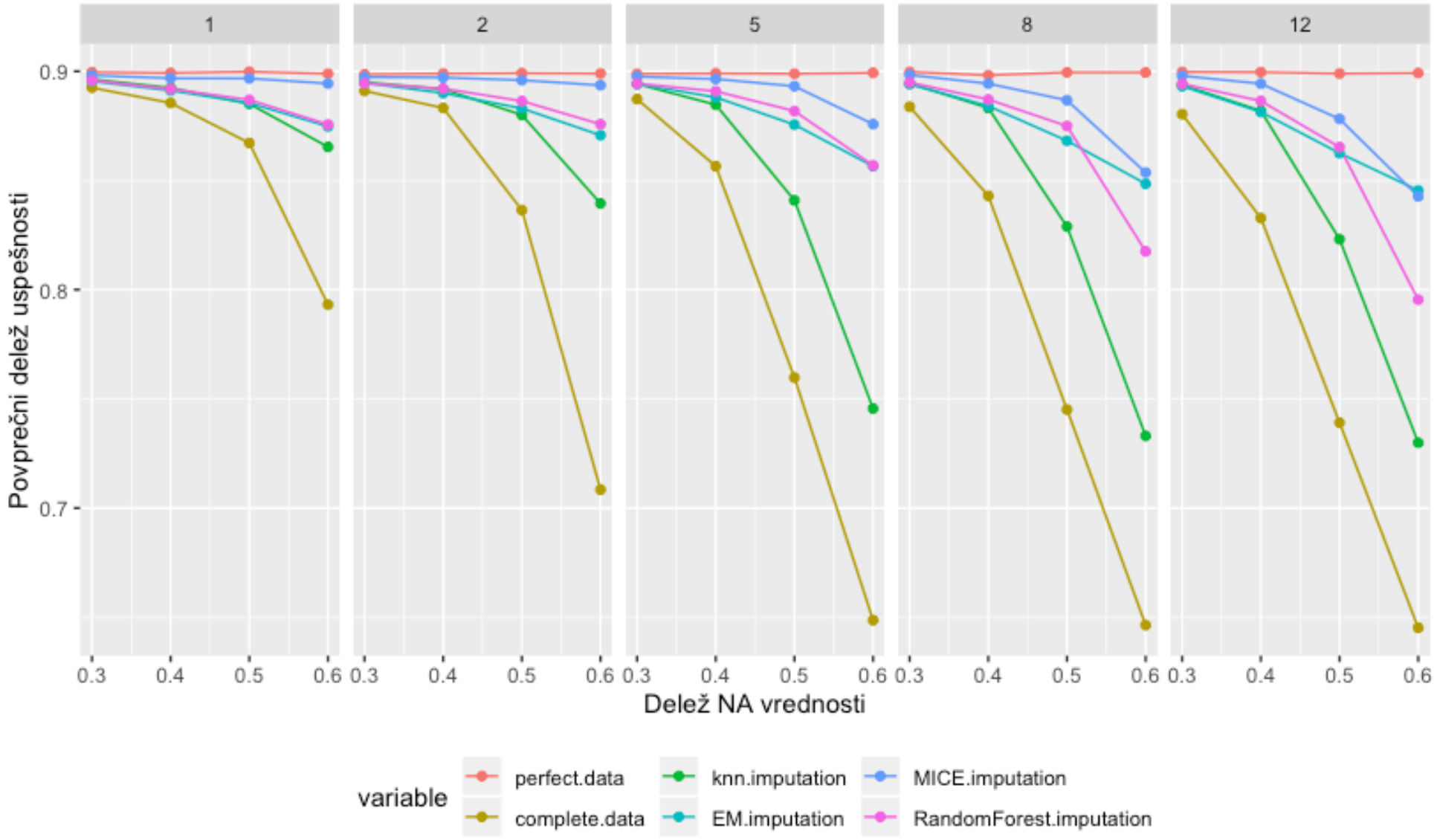
REZULTATI | NMAR



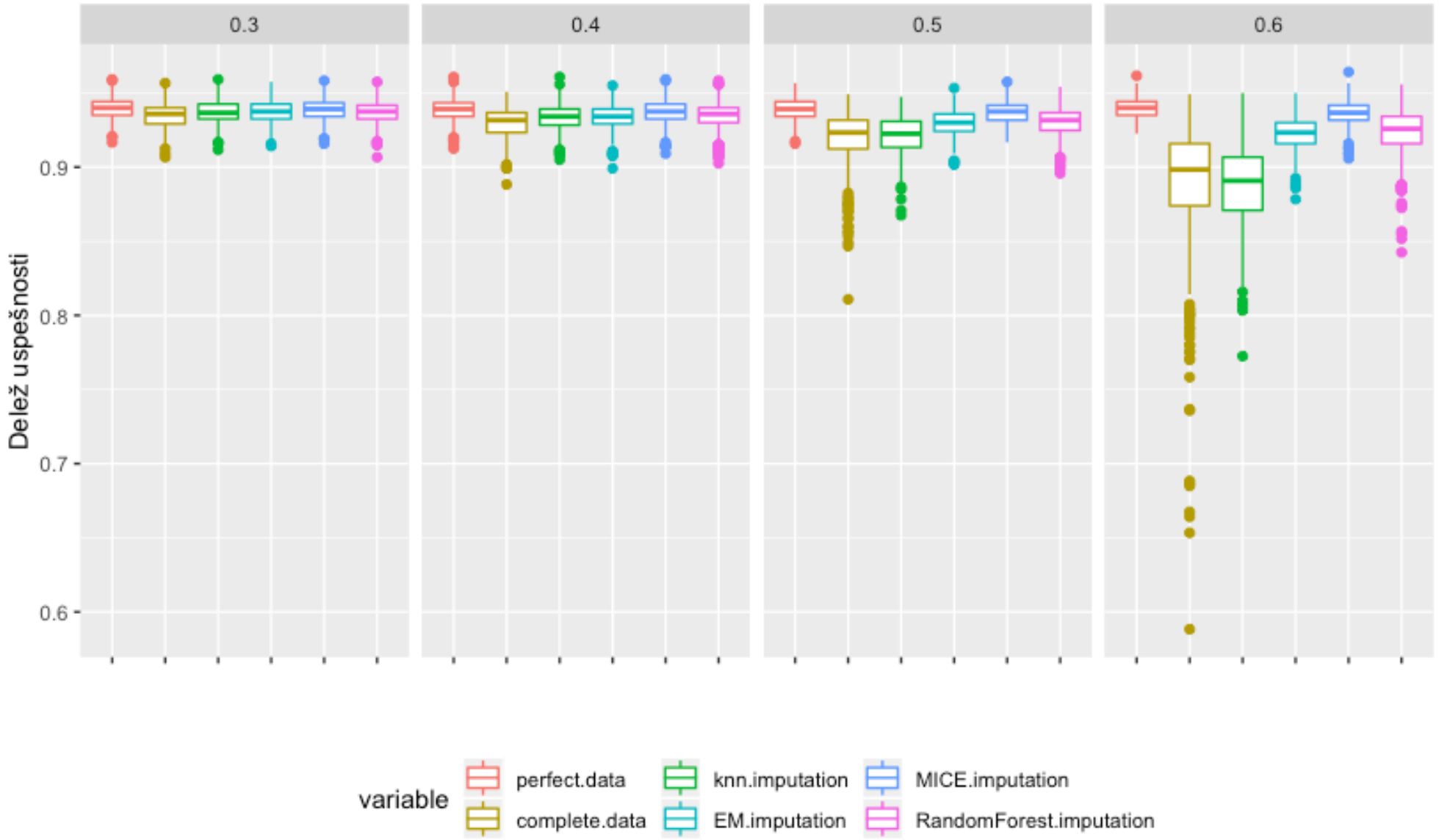
REZULTATI | MAR



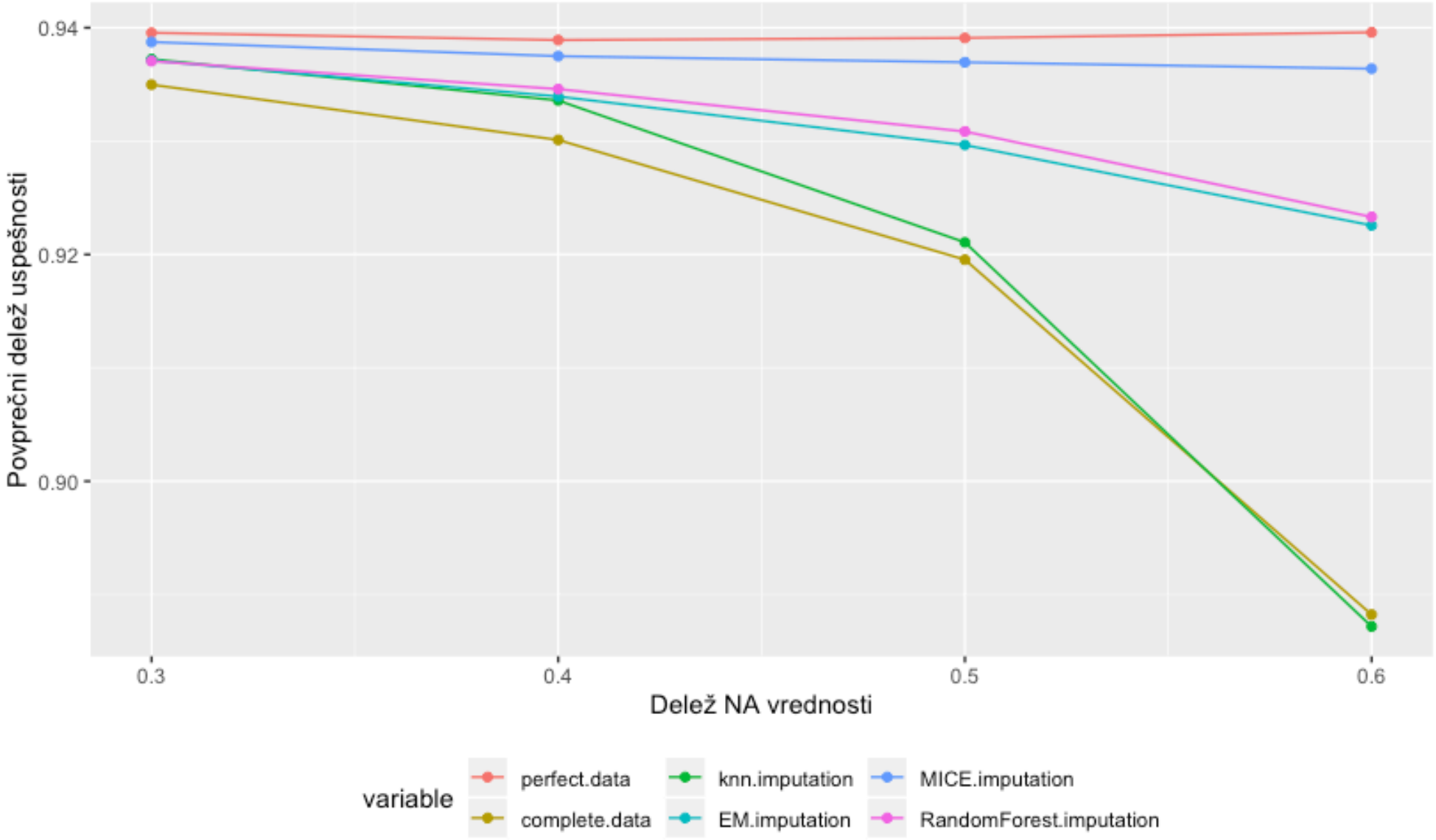
REZULTATI | MAR



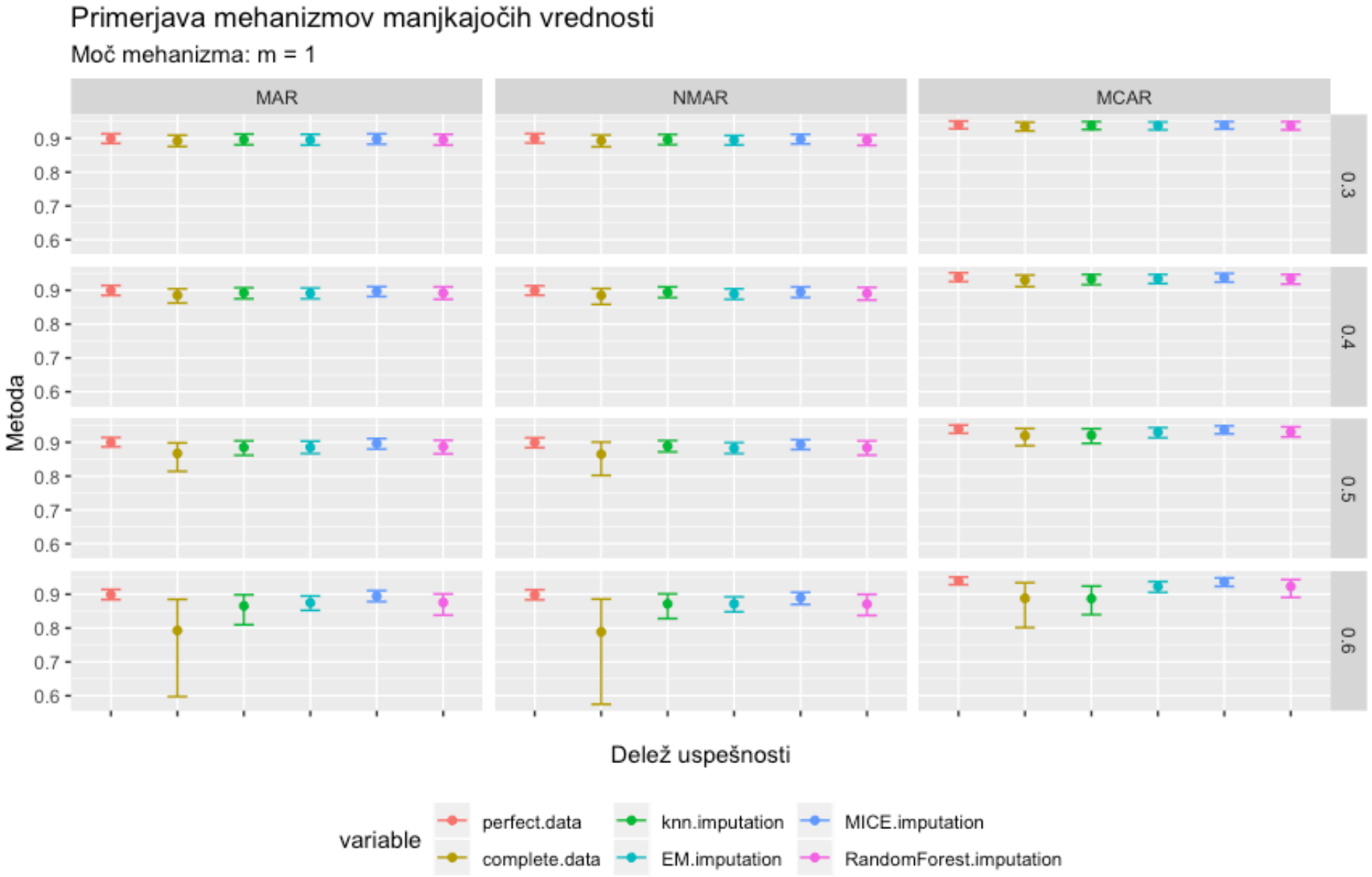
REZULTATI | MCAR



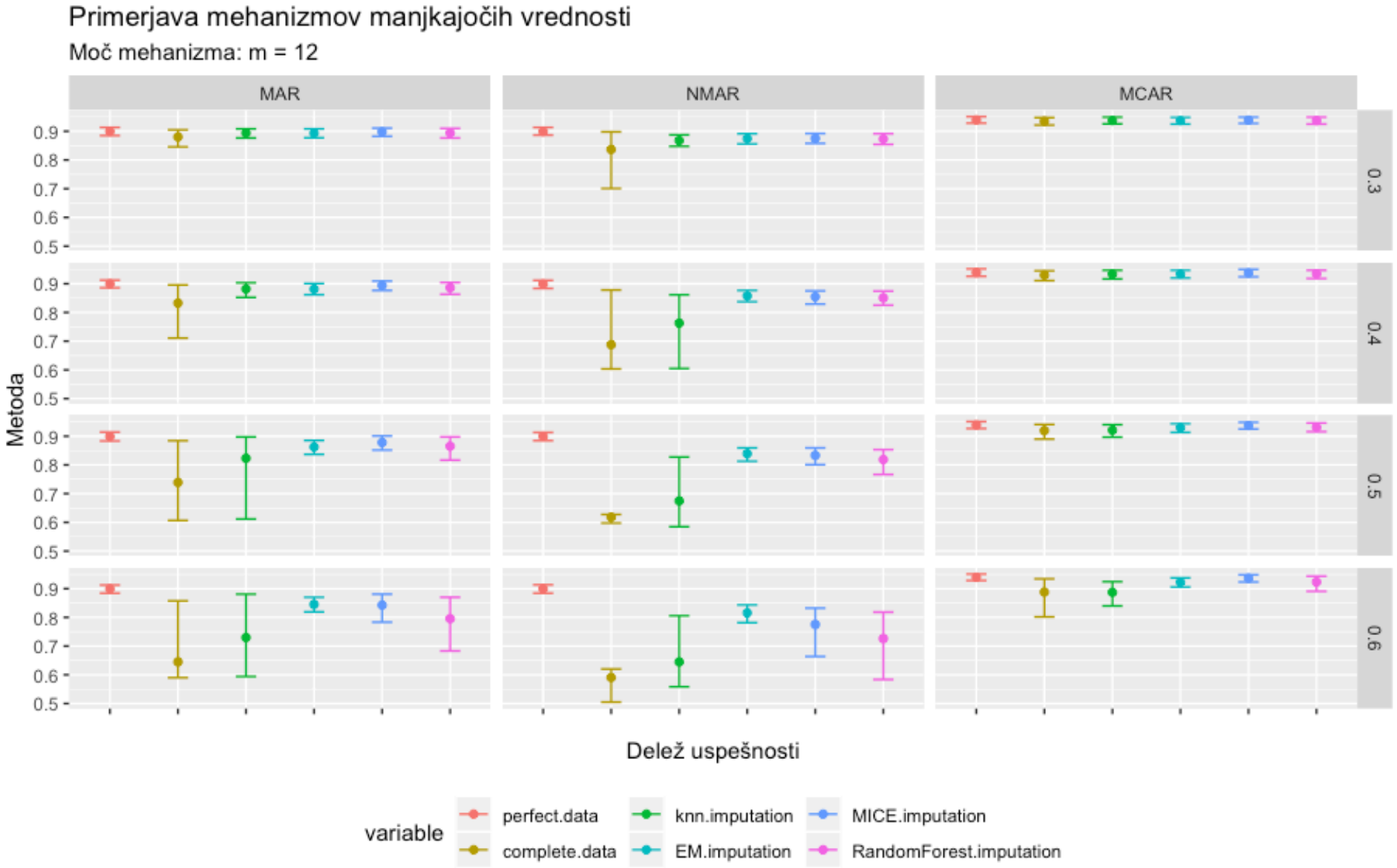
REZULTATI | MCAR



REZULTATI | PRIMERJAVA MEHANIZMOV



REZULTATI | PRIMERJAVA MEHANIZMOV



ZAKLJUČKI in PRIPOROČILA

- Dokler je delež manjkajočih vrednosti manjši ali enak 0.3 je uporaba samo popolnih podatkov še smiselna, pod pogojem, da imamo dovolj velik vzorec.
- Ne priporočava uporabe imputiranja s pomočjo KNN, v kolikor je delež manjkajočih vrednosti večji od polovice vseh podatkov.
- V primeru mehanizma **NMAR** je najboljše uporabiti **EM-algoritem** v primeru velike moči mehanizma, drugače **MICE**.
- V primeru mehanizma **MAR** je najboljše uporabiti **EM-algoritem** v primeru velike moči mehanizma, drugače **MICE**.
- V primeru MCAR je priporočava uporabo **MICE**.

NADALJNA ANALIZA

- Imputirati podatke na vsaki skupini posebej in potem združiti v eno podatkovje.
- Pregledati katera od metod je najhitrejša.
- Ocenjevanje modela na nepopolnih rezultatih