

**Sveobuhvatna analiza pristranosti u sustavu ocjenjivanja
projekata prijavljenih u području "Prirodne znanosti" za dodjelu
doktoranata (Projekt razvoja karijera mladih istraživača –
Izobrazba novih doktora znanosti (DOK-2025-02))**

Znanstveni dokazi potpune nefunkcionalnosti sustava vrednovanja

Prilog prigovoru

Branimir K. Hackenberger
Voditelj projekta: IP-2022-10-3233

5. kolovoza 2025.

Sažetak

Ovaj dokument predstavlja konsolidiranu analizu sustava ocjenjivanja projekata za dodjelu doktoranata koja nedvojbeno dokazuje potpunu nefunkcionalnost, pristranost i znans-tvenu nevaljanost primijenjenog postupka vrednovanja. Kroz tri komplementarna analitička pristupa, klasičnu statističku analizu, napredne statističke metode i Bayesovu analizu, doka-zali smo kako sustav ne samo da ne mjeri kvalitetu projekata, već predstavlja tek simulaciju objektivnosti koja dovodi do arbitarnih i nepravednih odluka. Rezultati pokazuju da samo 6.4% varijance u ocjenama proizlazi iz stvarnih razlika između projekata, dok ostatak čine subjektivnost recenzenata i nasumični čimbenici. Sustav se statistički ne razlikuje od nasu-mičnog ocjenjivanja ($p = 0.451$), što ga čini potpuno beskorisnim za namijenjenu svrhu.

Sadržaj

1 Uvod	6
1.1 Kontekst i motivacija	6
1.2 Svrha istraživanja	6
1.3 Opseg analize	6
1.4 Materijal i metode	6
1.4.1 Opis podatkovnog skupa	6
1.4.2 Analitički pristup	6
1.4.3 Softver i računalne platforme	8
1.4.4 Etički aspekti	8
1.5 Ključni rezultati - pregled	8
 I Klasična statistička analiza	 9
 2 Struktura podataka i identifikacija recenzenata	 9
2.1 Metodologija	9
2.2 Rezultati	9
2.3 Geografska neuravnoteženost znanstvenih centara kao dodatni izvor pristranosti .	9
 3 Analiza pouzdanosti među recenzentima (Inter-rater Reliability)	 12

3.1 Metodologija	12
3.2 Rezultati	13
4 Analiza stilova recenziranja	14
4.1 Metodologija	14
4.2 Rezultati	15
5 Mixed Effects Model analiza	16
5.1 Metodologija	16
5.2 Rezultati	17
5.3 Slučajni učinci projekata	18
6 Analiza slaganja reczenzenata po rangu projekta	19
6.1 Metodologija	19
6.2 Rezultati	20
7 Identifikacija problematičnih slučajeva	21
7.1 Metodologija	21
7.2 Rezultati	21
8 Prediktivna analiza	22
8.1 Metodologija	22
8.2 Rezultati	22
9 Analiza konzistentnosti kroz projekte	23
9.1 Metodologija	23
9.2 Rezultati	23
10 Osvrt na izdvajanje tzv. K3 kriterija	24
II Dodatne statističke analize	25
11 Analiza slaganja reczenzenata - Kendall-ov W	25
11.1 Opis metode	25
11.2 Rezultati	26
12 Analiza diskriminativnosti ocjena	27
12.1 Opis metode	27
12.1.1 Koeficijent varijacije (CV)	27
12.1.2 Bayesov čimbenik za ANOVA	27
12.1.3 Relativna entropija	27
12.2 Rezultati	28
13 Usporedba s nasumičnim ocjenjivanjem	29
13.1 Opis metode	29
13.2 Rezultati	29
14 Analiza ceiling učinka	31
14.1 Opis metode	31
14.1.1 Deskriptivnu statistiku	31
14.1.2 Bayesov čimbenik protiv uniformne distribucije	31
14.2 Rezultati	31

15 Analiza informativnosti kriterija	33
15.1 Opis metode	33
15.1.1 Normalizirana mutual information	33
15.1.2 Analiza glavnih komponenti (PCA)	33
15.2 Rezultati	34
15.3 Interpretacija rezultata	34
16 Analiza ekstremnih obrazaca ocjenjivanja	35
17 Analiza ekstremnih obrazaca ocjenjivanja	35
17.1 Uvod	35
17.2 Opis metode	35
17.3 Rezultati	36
III Bayesova analiza	37
18 Struktura podataka i osnovni pokazatelji	37
18.1 Struktura uzorka	37
18.2 Distribucija ocjena	37
19 Bayesovi hijerarhijski modeli	38
19.1 Opis metode	38
19.2 Problemi s konvergencijom	38
19.3 Dekompozicija varijance	38
19.4 Interpretacija rezultata	39
20 Analiza pristranosti po pozicijama	40
20.1 Bayesov čimbenik	40
21 Usporedba s nasumičnim ocjenjivanjem	41
22 Analiza ceiling učinka	42
23 Identifikacija ekstremnih obrazaca	43
24 Selekcija modela	45
IV Procjena stvarnog broja reczenzenta	47
25 Uvod	47
25.1 Kontekst i motivacija istraživanja	47
25.2 Problematika identifikacije reczenzenta	47
25.3 Struktura analize i metodološki pristup	47
26 Sažetak glavnih rezultata	48
27 Karakteristike ocjenjivanja	48
27.1 Analiza distribucije ocjena	48
27.2 Tipologija reczenzenta prema varijabilnosti	49
28 Procjena broja reczenzenta	50
28.1 Metodološki pluralizam u procjeni	50

36 Hitne preporuke za akciju	67
36.1 Trenutne mjere	67
36.1.1 Potpuno odbacivanje rezultata	67
36.1.2 Javna isprika i transparentnost	68
36.2 Dugoročne reforme	68
36.2.1 Potpuni redizajn ili alternativni sustavi	68
36.2.2 Promjene sustava	68
37 Zaključna poruka znanstvenoj zajednici i HRZZ-u	69

obrasci koji se pojavljuju kod više reczenzata.

Analiza glavnih komponenti primijenjena je na matricu ocjena za analizu redundantnosti među kriterijima i procjenu efektivne dimenzionalnosti prostora ocjenjivanja. PCA dekompozicija omogućava identifikaciju dominantnih čimbenika koji objašnjavaju većinu varijabilnosti u podacima.

Bootstrap analiza s 1000 iteracija korištena je za procjenu stvarnog broja reczenzata na temelju analize jedinstvenih obrazaca ocjenjivanja. U svakoj iteraciji kreiran je bootstrap uzorak, identificirani su jedinstveni obrasci te je procijenjen broj različitih reczenzata. Rezultati su sumirani kroz 95% intervale povjerenja.

1.4.3. Softver i računalne platforme

Sve analize provedene su korištenjem R verzije 4.3.0 s ekstenzivnim setom specijaliziranih paketa. Paket tidyverse korišten je za manipulaciju podataka i vizualizaciju, lme4 za mixed effects modele, psych za ICC analize, cluster za klaster analize, mclust za Gaussian mixture modele, brms za Bayesovu analizu te igraph za mrežne analize. Stan platforma korištena je za probabilističko programiranje potrebno za implementaciju kompleksnih Bayesovih modela. Svi kodovi i podaci pohranjeni su i dostupni na zahtjev kako bi se osigurala potpuna reproducibilnost rezultata.

1.4.4. Etički aspekti

Prije početka analize, svi podaci su pažljivo anonimizirani kako bi se zaštitala privatnost sudionika. Identiteti projekata i reczenzata zamijenjeni su numeričkim kodovima koji ne omogućavaju identifikaciju. Analiza je provedena isključivo s ciljem znanstvenog istraživanja i unapređenja sustava vrednovanja. Rezultati su prezentirani na agregiranoj razini koja ne omogućava identifikaciju pojedinačnih sudionika ili njihovih projekata. Svi rezultati interpretirani su u kontekstu sistemskih problema, a ne individualnih performansi.

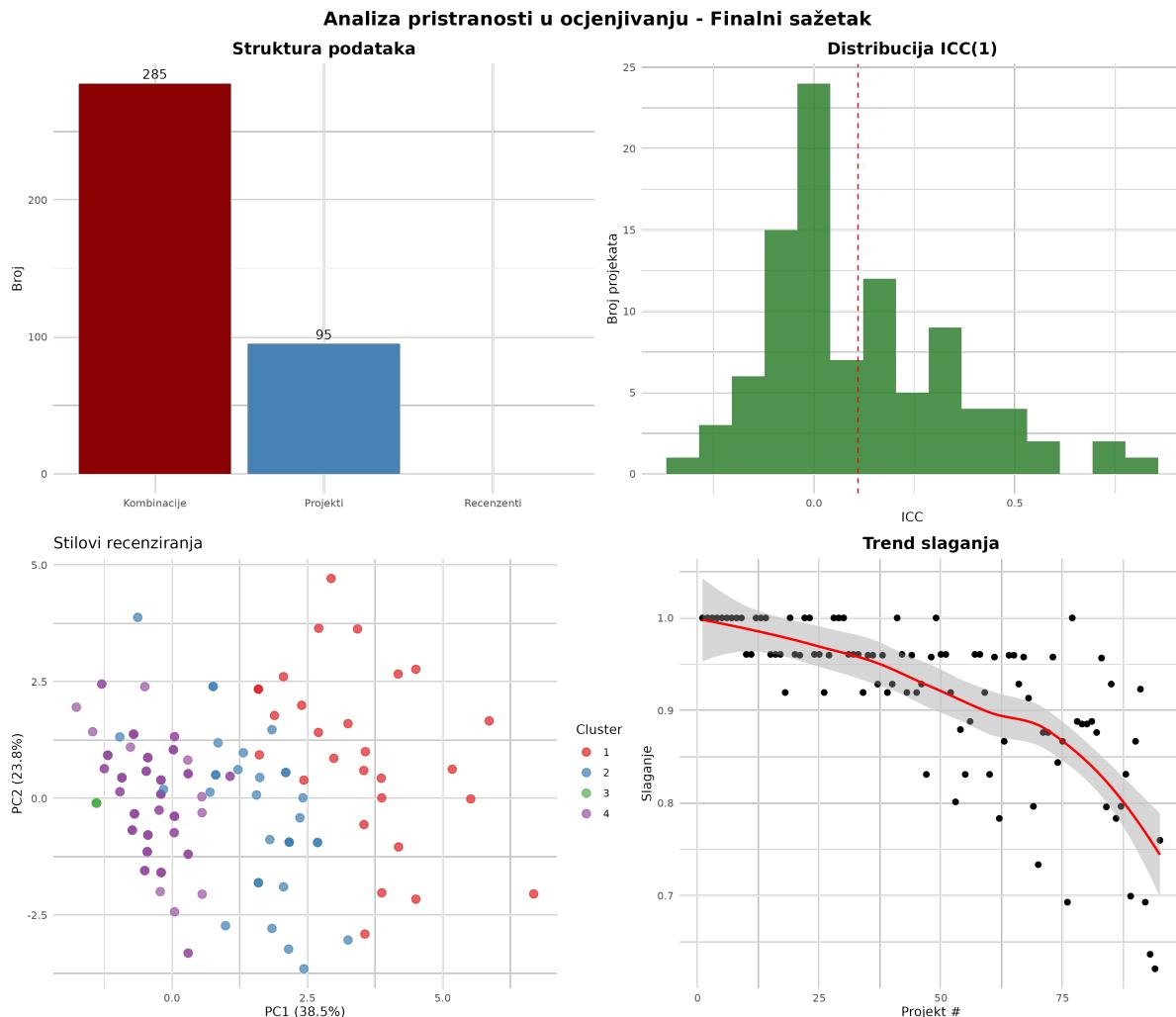
1.5. Ključni rezultati - pregled

Već u uvodu važno je istaknuti najšokantniji rezultat: **sustav vrednovanja koji samo 6.4% svoje varijance temelji na stvarnim razlikama između evaluiranih objekata je fundamentalno nefunkcionalan**. To znači da 93.6% ocjena nema veze s kvalitetom projekata već je rezultat subjektivnosti reczenzata i nasumičnih čimbenika.

Implikacije za proces recenziranja. Bez transparentnog i statistički utemeljenog postupka odabira reczenzenata koji uzima u obzir ovu neuravnoteženost, postoji realna opasnost da će određeni projekti biti favorizirani ili diskriminirani ovisno o tome koji recenzenti ih ocjenjuju. Na primjer:

- Projekti iz većih centara mogu imati prednost ako su recenzenti također iz velikih centara koji bolje razumiju kontekst i mogućnosti takvih institucija.
- Projekti iz manjih centara mogu biti podcijenjeni ako ih ocjenjuju recenzenti navikli na infrastrukturu i resurse velikih institucija.
- Regionalni projekti mogu biti neadekvatno ocijenjeni ako recenzenti nisu upoznati s lokalnim specifičnostima i potrebama.

Ova strukturna neuravnoteženost, u kombinaciji s već identificiranim problemom da svaki projekt ima svoj jedinstveni set reczenzenata, dodatno produbljuje problem usporedivosti ocjena između projekata i dovodi u pitanje pravednost cijelog postupka vrednovanja.



Slika 1: Sažetak analize pristranosti u ocjenjivanju. Gornji lijevi panel prikazuje strukturu podataka s 95 projekata i 285 jedinstvenih kombinacija projekt-recenzent. Gornji desni panel pokazuje distribuciju ICC(1) vrijednosti s prosječnom vrijednošću od samo 0.11 (označeno crvenom isprekidanom linijom), što ukazuje na izrazito nisku pouzdanost među recenzentima. Donji lijevi panel prikazuje klastere stilova recenziranja identificirane PCA analizom, pokazujući 4 različita pristupa ocjenjivanju. Donji desni panel ilustrira dramatičan pad slaganja reczenzenata kroz vrijeme (crvena LOESS krivulja), što sugerira umor reczenzenata ili promjenu standarda tijekom procesa vrednovanja.

Ova struktura podataka predstavlja **temeljni problem dizajna** jer različiti recenzenti ocjenjuju različite projekte bez ikakve zajedničke referentne točke što znači kako su projekti ocjenjivani u potpuno različitim "mikrokozmosima" ocjenjivanja, što u potpunosti onemogućava usporedbu između projekata.

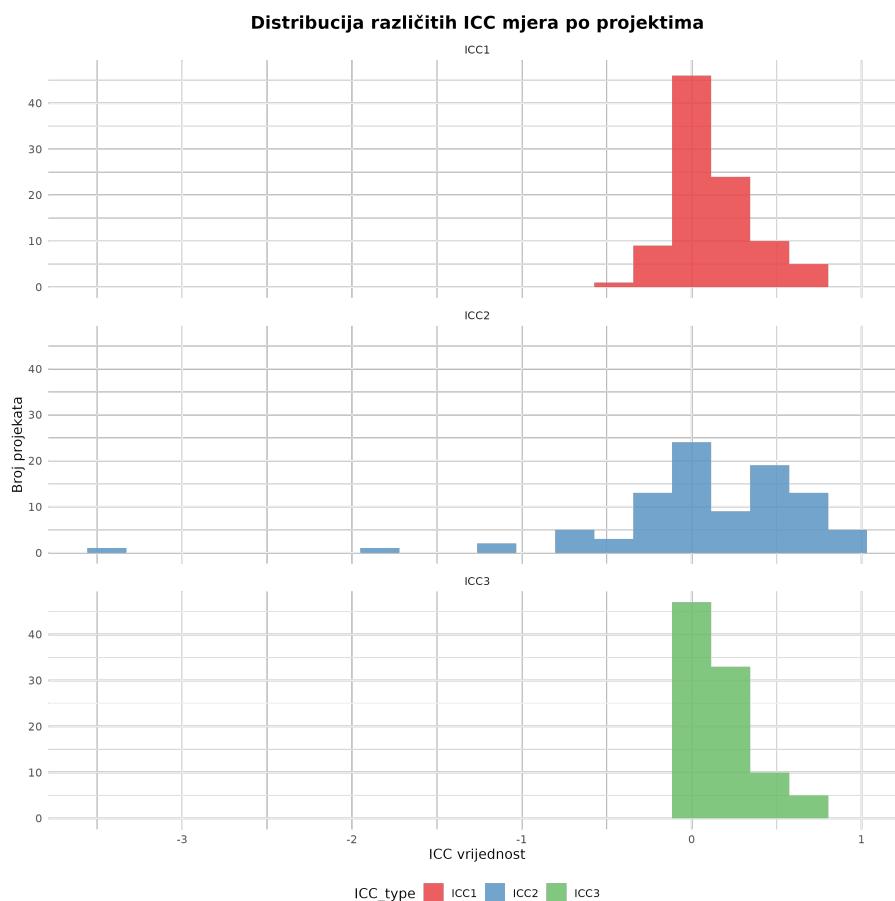
- $ICC > 0.75$: Izvrsno slaganje

3.2. Rezultati

Rezultati ICC analize su **katastrofalni** za bilo koji ozbiljan sustav vrednovanja:

- Prosječni $ICC(1) = 0.110$ (raspon: -0.348 do 0.794)
- Prosječni $ICC(2) = 0.082$ (raspon: -3.440 do 0.920)
- Prosječni $ICC(3) = 0.174$ (raspon: 0.000 do 0.794)
- 85 od 95 projekata (89.5%) ima $ICC(1) < 0.4$ - prag za "slabo slaganje"
- 30% projekata ima negativne ICC vrijednosti

Ključna interpretacija: Prosječni $ICC(1)$ od 0.11 znači da samo 11% varijance u ocjenama može biti pripisano stvarnim razlikama između projekata, dok je 89% varijance rezultat subjektivnih razlika između recenzentata. Negativne ICC vrijednosti su posebno zabrinjavajuće jer ukazuju da se recenzenti slažu manje nego što bi se očekivalo nasumičnim dodjeljivanjem ocjena.



Slika 2: Distribucija različitih ICC mera po projektima. Gornji panel (ICC1) pokazuje distribuciju apsolutnog slaganja pojedinačnih recenzentata s prosjekom od samo 0.110 i velikim brojem negativnih vrijednosti. Srednji panel (ICC2) prikazuje pouzdanost prosječnih ocjena s još gorim rezultatima, uključujući ekstremno negativne vrijednosti do -3.44. Donji panel (ICC3) pokazuje konzistentnost recenzentata s nešto boljim ali još uvijek neprihvatljivim vrijednostima. Crvene isprekidane linije označavaju prosječne vrijednosti za svaku mjeru.

4. Analiza stilova recenziranja

Ocenjivanje projekata (mentora) nije samo tehnički čin dodjeljivanja broja određenom kriteriju, već složen kognitivni postupak koji uključuje interpretaciju smjernica, osobne vrijednosti, prethodna iskustva i profesionalne preferencije recenzentata. U tom smislu, dva recenzenta koji ocjenjuju isti projekt mogu doći do potpuno različitih rezultata ne zbog razlika u kvaliteti prijedloga, već zbog razlika u vlastitim stilovima ocenjivanja. Ove razlike mogu uključivati razinu strogosti (*lenient vs. strict raters*), način ponderiranja različitih kriterija (npr. važnost inovativnosti naspram izvedivosti), pa čak i tendenciju korištenja ekstremnih ili srednjih ocjena.

Identifikacija i razumijevanje takvih stilova recenziranja ključno je za osiguranje pravednosti i dosljednosti postupka vrednovanja. Ako znamo da određeni recenzent sustavno daje niže ili više ocjene, ili da naglašava specifične kriterije u odnosu na druge, tada se može razmotriti korekcija tih razlika, primjerice putem kalibracije, ponderiranja ili algoritamske prilagodbe. Suprotno tome, ignoriranje ovih razlika može dovesti do toga da konačne ocjene projekata više odražavaju osobne pristupe pojedinih recenzentata nego stvarnu kvalitetu ocenjivanih prijedloga.

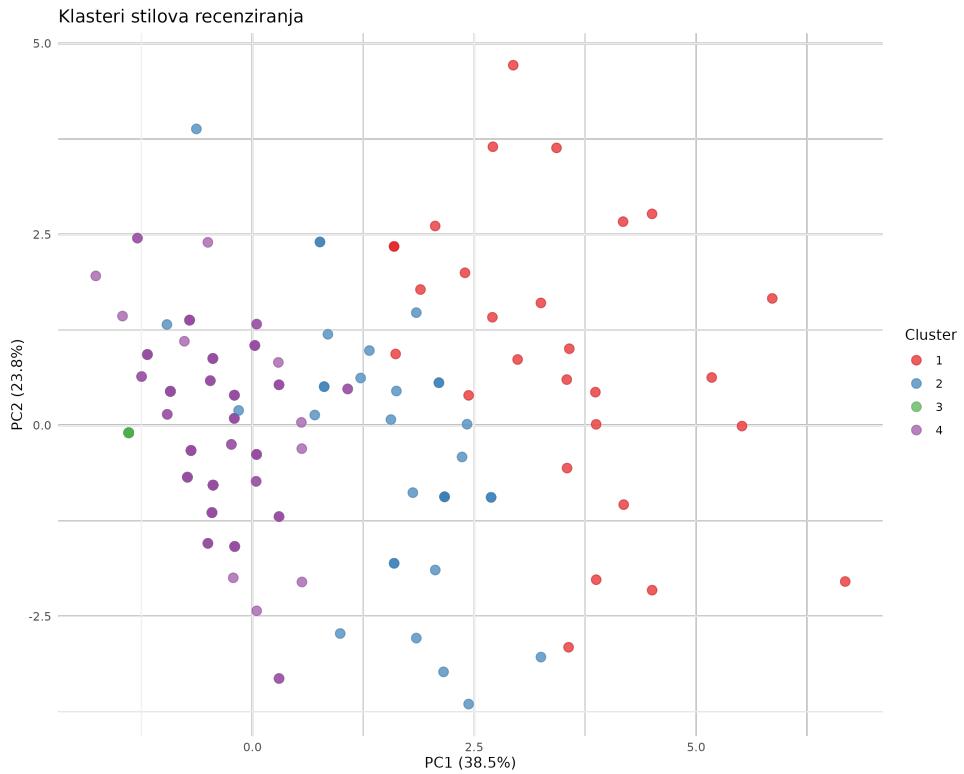
Stoga je cilj ove analize bio otkriti obrasce u načinu ocenjivanja, odnosno identificirati postojanje i karakteristike različitih stilova recenziranja. Korištenjem metoda hijerarhijskog klasteriranja, kombiniranih sa Silhouette i Gap statistikom za optimizaciju broja klastera, te vizualizacijom glavnih komponenti, dobili smo uvid u to koliko su recenzenti međusobno slični ili različiti u svojim pristupima. Rezultati pokazuju ne samo postojanje jasnih skupina recenzentata, već i ekstremne razlike u njihovoј strogosti i prioritetima, što predstavlja ozbiljan izazov za pouzdanost i pravednost sustava vrednovanja.

4.1. Metodologija

Za identifikaciju različitih stilova ocenjivanja primijenjena je hijerarhijska klaster analiza. Korišteni su sljedeći koraci:

1. Kreiranje profila recenzentata na temelju prosječnih ocjena po svim kriterijima
2. Izračun matrice sličnosti pomoću korelacijske udaljenosti: $d_{ij} = 1 - \text{cor}(r_i, r_j)$
3. Hijerarhijsko klasteriranje metodom prosječnog vezanja
4. Optimizacija broja klastera pomoću:
 - Silhouette koeficijenta: $s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$
 - Gap statistike: $\text{Gap}_k = E_n^*[\log(W_k)] - \log(W_k)$

4.2. Rezultati



Slika 3: Klasteri stilova recenziranja identificirani PCA analizom i K-means klasteriranjem. Graf prikazuje prve dvije glavne komponente (PC1 objašnjava 38.5% varijance, PC2 23.8%). Identificirana su 4 različita klastera reczenzenata (označeni različitim bojama), što ukazuje na postojanje fundamentalno različitih pristupa ocjenjivanju. Raspršenost točaka unutar klastera dodatno pokazuje varijabilnost čak i unutar sličnih stilova ocjenjivanja.

Analiza je identificirala:

- **22 optimalna klastera** prema kombinaciji Silhouette i Gap statistike
- **4 glavna stila recenziranja** vidljiva u PCA projekciji
- Ekstremne razlike u "strogosti" između klastera (0.759 do 1.000)
- Različite preferencije prema pojedinim kriterijima (K1, K2, K3)

Postojanje 22 različita stila ocjenjivanja među 285 kombinacija projekt-recenzent je **odgovarajući dokaz nekonzistentnosti**. To znači da ocjena projekta više ovisi o stilu recenziranja dodijeljenog recenzenta nego o stvarnoj kvaliteti projekta.

5. Mixed Effects Model analiza

Jedno od ključnih pitanja u procjeni učinkovitosti postupka vrednovanja jest razumijevanje izvora varijabilnosti u dodijeljenim ocjenama. Naime, ukupna varijabilnost može potjecati iz tri osnovna izvora: (1) stvarnih razlika u kvaliteti projekata, (2) sustavnih razlika između recenzenata (njihovih grupa, stilova ocjenjivanja ili razine strogosti), te (3) nasumičnih čimbenika i individualnih neslaganja. Ako sustav ocjenjivanja doista mjeri ono što bi trebao, a to je kvaliteta projektnog prijedloga ili mentora, tada bi najveći dio varijabilnosti trebao dolaziti upravo iz razlika među projektnih prijedloga ili mentora. Suprotno tome, visoka varijabilnost uzrokovana recenzentima ili slučajnim čimbenicima upućuje na to kako je postupak ocjenjivanja više odraz subjektivnih percepacija nego stvarne kvalitete onoga što se ocjenjivalo.

Kako bi se razlučilo koliko svaka od ovih komponenti doprinosi ukupnoj varijanci ocjena, korišten je pristup linearnih mješovitih modela (*Linear Mixed Effects Models*, LMEM). Ovi modeli omogućavaju istovremenu procjenu fiksnih i nasumičnih učinaka te kvantificiranje koliko varijance proizlazi iz pojedinih izvora. U ovom kontekstu, projektni prijedlog (mentor) se tretirao kao nasumični učinak (čime se procjenjuje doprinos stvarnih razlika između prijedloga), grupa recenzenata se tretirala kao nasumični učinak koji obuhvaća njihove specifične pristupe ocjenjivanju, dok je rezidualna varijanca predstavljala sve preostale neobjašnjene ili nasumične čimbenike.

Ovakva dekompozicija varijance posebno stoga jer pruža empirijske dokaze o tome u kojoj mjeri sustav vrednovanja ispunjava svoju primarnu svrhu razlikovanja boljih od lošijih prijedloga. Ako se pokaže kako je doprinos varijance projekata nizak, to znači kako ocjene više reflektiraju individualne razlike među recenzentima nego kvalitetu samih projekata, čime se dovodi u pitanje funkcionalnost cijelog sustava.

Stoga je cilj ove analize bio utvrditi koliko ukupne varijabilnosti u ocjenama potječe od stvarnih razlika među projektima, a koliko od subjektivnih i slučajnih čimbenika. Rezultati ovog modela daju nedvosmislen odgovor na to pitanje i predstavljaju temelj za raspravu o nužnosti korjenite reforme sustava ocjenjivanja.

5.1. Metodologija

Za razdvajanje izvora varijabilnosti korišten je linearni mješoviti model (Linear Mixed Effects Model):

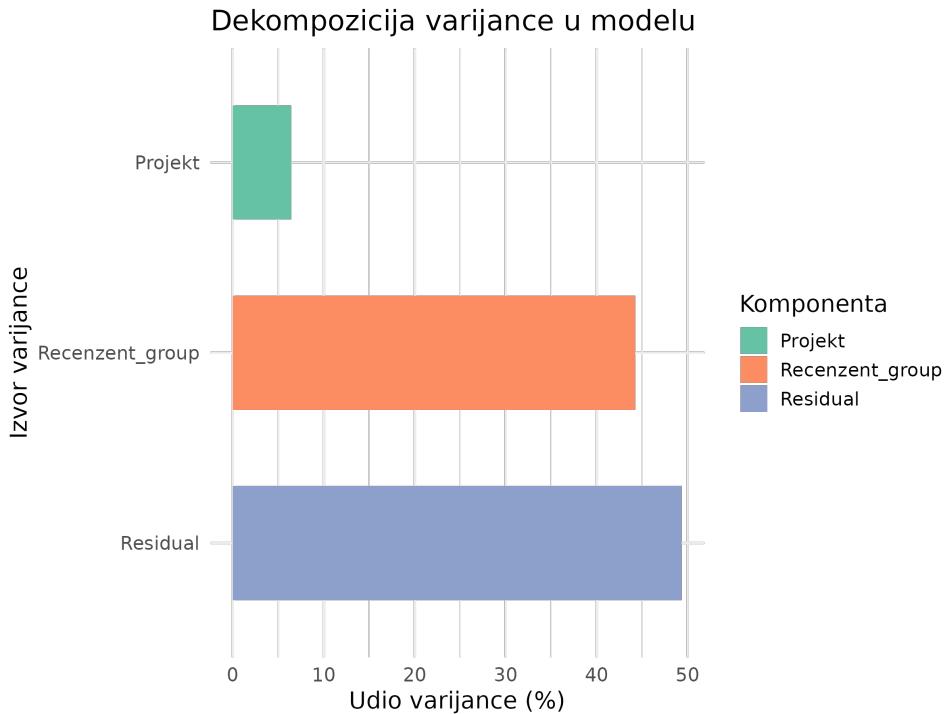
$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk} \quad (2)$$

gdje je:

- Y_{ijk} - ocjena K3 za projekt i , grupu recenzenata j , i poziciju k
- μ - opći prosjek
- $\alpha_i \sim N(0, \sigma_{projekt}^2)$ - nasumični učinak projekta
- $\beta_j \sim N(0, \sigma_{recenzent}^2)$ - nasumični učinak grupe recenzenata
- γ_k - fiksni učinak pozicije recenzenta (R1, R2, R3)
- $\epsilon_{ijk} \sim N(0, \sigma_{residual}^2)$ - rezidualna greška

Testirani su tri modela rastućeg stupnja složenosti pomoću Likelihood Ratio testa.

5.2. Rezultati



Slika 4: Dekompozicija varijance u konačnom *mixed effects* modelu. Graf jasno pokazuje da samo 6.4% ukupne varijance u K3 ocjenama proizlazi iz stvarnih razlika između projekata (zelena traka). Dominantni izvori varijance su grupe recenzenata (44.2%, narančasta traka) i rezidualna varijanca (49.3%, plava traka). Ova raspodjela nedvojbeno dokazuje da sustav ocjenjivanja ne mjeri kvalitetu projekata već reflektira subjektivne razlike između recenzenata.

Dekompozicija varijance otkriva **jedan od najstrašnijih rezultata** cijele analize:

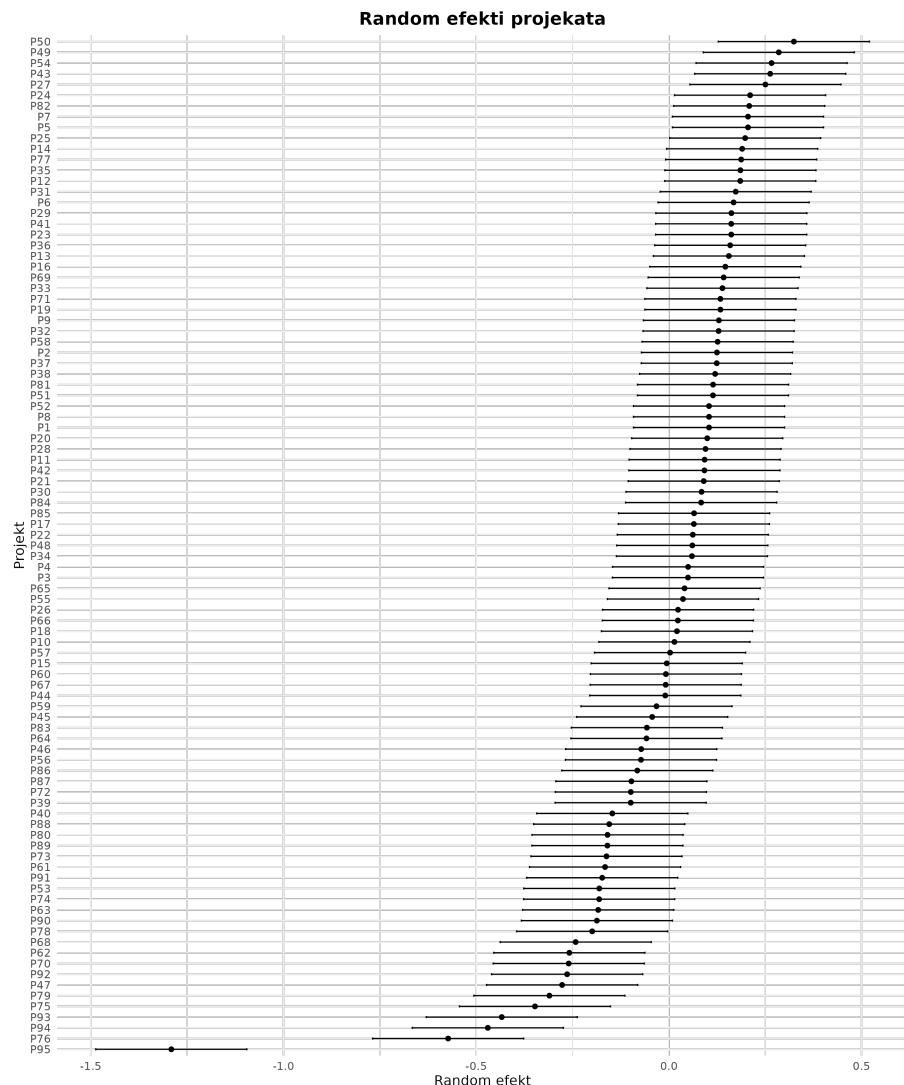
- **Varijanca zbog projekata:** 6.4% ($\sigma_{projekt}^2 = 0.214$)
- **Varijanca zbog grupa recenzenata:** 44.2% ($\sigma_{recenzent}^2 = 1.471$)
- **Rezidualna varijanca:** 49.3% ($\sigma_{residual}^2 = 1.642$)

Kritična interpretacija: Samo 6.4% varijance u K3 ocjenama može se pripisati stvarnim razlikama između projekata! To znači da 93.6% varijance nema veze s kvalitetom projekata (mentora) već je rezultat:

- 44.2% - sustavnih razlika između grupa recenzenata
- 49.3% - nasumičnih čimbenika i neslaganja

Sustav vrednovanja koji samo 6.4% svoje varijance temelji na stvarnim razlikama između evaluiranih objekata je fundamentalno nefunkcionalan.

5.3. Slučajni učinci projekata



Slika 5: Slučajni učinci za sve projekte s 95% intervalima pouzdanosti. Graf prikazuje procijenjene odstupanja svakog projekta od općeg prosjeka nakon kontrole za učinke recenzentata. Projekti su poredani od najnižeg do najvišeg slučajnog učinka. Široki intervali pouzdanosti i mala odstupanja od nule (većina projekata ima učinke između -0.5 i +0.5) dodatno potvrđuju da stvarne razlike između projekata čine zanemariv dio ukupne varijabilnosti u sustavu ocjenjivanja.

6. Analiza slaganja reczenzenata po rangu projekta

Analiza međusobnog slaganja reczenzenata se tradicionalno koristi za istraživanje stabilnosti ocjena tijekom procesa vrednovanja. Primjerice, istraživanje na koji način se ocjene mijenjaju ovisno o redoslijedu ili vremenu ocjenjivanja. Međutim, takav pristup nije primjenjiv u ovom slučaju, budući kako su svi projekti ocjenjivani unutar praktički istog vremenskog intervala te ne postoji pouzdan podatak o točnom redoslijedu u kojem su pojedini prijedlozi ocjenjivani.

U ovoj analizi primijenjen je sličan metodološki okvir, ali s drukčijom svrhom. Svrha je bila istražiti povezanost ukupne ocjene projektnog prijedloga s razinom slaganja među recenzentima. Drugim riječima, cilj nije bio utvrditi postoji li vremenski trend u ocjenama, nego razumjeti razlikuje li se međusobna usklađenost reczenzenata ovisno o tome je li projekt dobio visoke, srednje ili niske ocjene. Ovakav pristup omogućuje uvid u to slažu li se recenzenti više oko projekata koje percipiraju kao izrazito kvalitetne ili, suprotno tome, oko onih koji su ocijenjeni lošije, te gdje je varijabilnost u ocjenama najveća.

Ovakav pristup omogućio je otkrivanje odnosa između visine ocjene i razine konsenzusa među recenzentima, što predstavlja važan uvid u konzistentnost postupka vrednovanja i potencijalnu pristranost u procjenama. Praktična implikacija rezultata ove analize je potpuno jasna. Ako se pokazuje kako je slaganje reczenzenata nisko upravo kod projekata srednje kvalitete, tada su upravo ti prijedlozi najosjetljiviji na subjektivne razlike među ocjenjivačima, što može imati presudni utjecaj na odluke o financiranju.

6.1. Metodologija

Za analizu promjena u ocjenjivanju kroz rang projekata korištena je Spearmanova korelacija rang-redoslijeda:

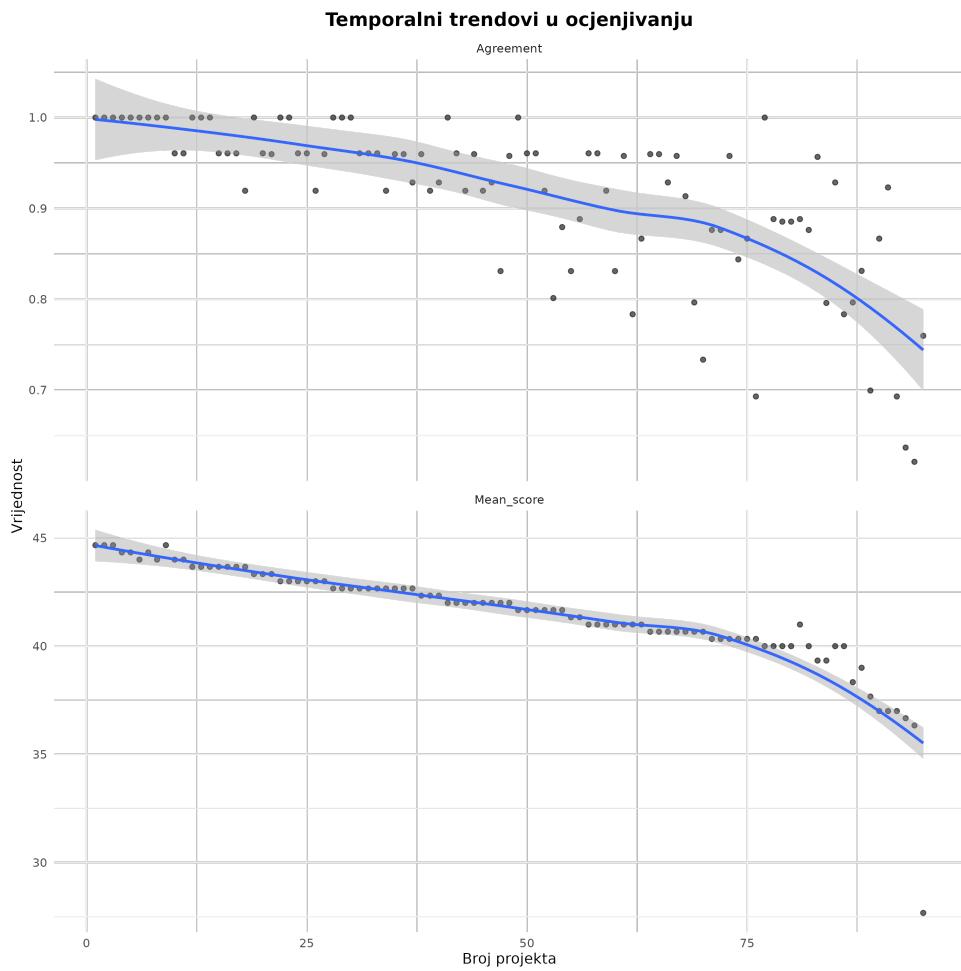
$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \quad (3)$$

gdje je d_i razlika u rangu između varijabli, a n broj opservacija.

Analizirani su trendovi za:

- Prosječne ocjene po projektu
- Slaganje reczenzenata (definirano kao $1 - CV$, gdje je CV koeficijent varijacije)

6.2. Rezultati



Slika 6: Trendovi u ocjenjivanju kroz redoslijed projekata. Gornji panel pokazuje dramatičan pad slaganja među recenzentima od početnih vrijednosti oko 1.0 (savršeno slaganje) do vrijednosti ispod 0.8 za kasnije, niže rangirane, projekte. Donji panel prikazuje još drastičniji pad prosječnih ocjena od približno 45 bodova na početku do ispod 35 bodova za zadnje projekte. Plava LOESS krivulja s 95% intervalom pouzdanosti (sivo područje) jasno pokazuje statistički značajne negativne trendove u oba slučaja.

Rezultati ove analize su alarmantni:

- **Trend prosječnih ocjena:** $\rho = -0.994, p < 0.001$
- **Trend slaganja recenzenata:** $\rho = -0.785, p < 0.001$

Ovi izrazito jaki negativni trendovi pokazuju:

1. Ocjene **dramatično padaju** niz rang projekata (gotovo savršena negativna korelacija od -0.994), što je logično no
2. Slaganje među recenzentima također **značajno opada**
3. Projekti ocjenjivani lošije su stoga parcijalno **diskriminirani** iz nepoznatog razloga

To sugerira **nejednakost u pristupima ili različite primjene standarda ili nerazumijevanje recenzenata** tijekom procesa, što dodatno kompromitira valjanost vrednovanja.

8. Prediktivna analiza

Jedno od ključnih pitanja u konstrukciji sustava vrednovanja je opravdanost i nezavisnost pojedinih kriterija ocjenjivanja. Ako se pokaže kako je jedan kriterij moguće s visokom točnošću predvidjeti iz drugih, postavlja se pitanje njegove stvarne dodane vrijednosti tj. služi li on doista kao zasebna dimenzija procjene ili samo duplicira informacije koje već nose ostali kriteriji? Takva redundanca može ne samo nepotrebno usložnjavati sustav ocjenjivanja, nego i zamagliti interpretaciju rezultata.

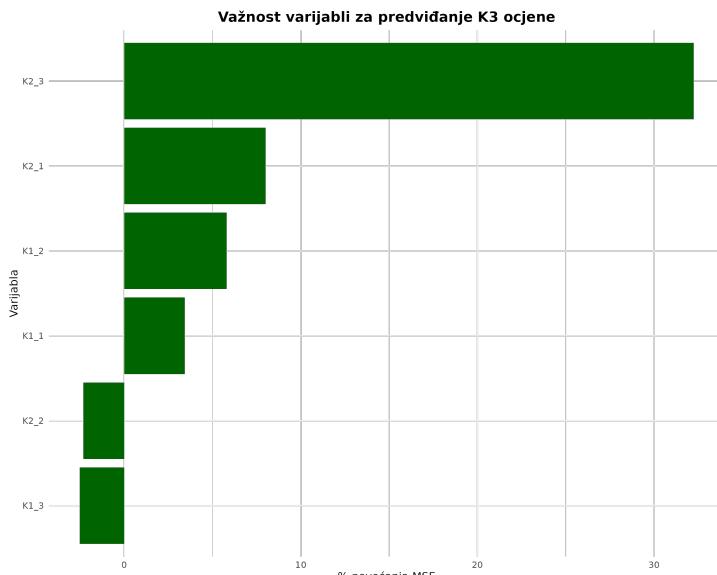
Kako bi se ovo pitanje empirijski istražilo, u ovoj je analizi primijenjen Random Forest algoritam, jedan od najpouzdanijih modela za prediktivnu analizu nelinearnih odnosa i interakcija među varijablama. Model je treniran na većem dijelu podataka za predviđanje ocjena kriterija K3 koristeći ocjene kriterija K1 i K2 kao ulazne varijable. Ovakav pristup omogućuje ne samo kvantificiranje prediktivne snage ostalih kriterija za K3, nego i procjenu relativne važnosti pojedinih komponenti K1 i K2 u formirajujući K3 ocjena.

Ova analiza pruža praktične uvide u to u kojoj mjeri su kriteriji vrednovanja doista nezavisni i razlikovni. Ako se pokaže kako je K3 gotovo u potpunosti predvidljiv iz drugih kriterija, to znači da on ne dodaje novu informaciju u procesu ocjenjivanja, što otvara pitanje potrebe za njegovim zadržavanjem u sadašnjem obliku. Takvi rezultati mogu poslužiti kao temelj za racionalizaciju kriterija i poboljšanje dizajna sustava vrednovanja kako bi se smanjila redundancija i povećala jasnoća interpretacije ocjena.

8.1. Metodologija

Za analizu međuovisnosti kriterija korišten je Random Forest algoritam s 500 stabala. Model je treniran na 80% podataka za predviđanje K3 ocjena na temelju K1 i K2 komponenti.

8.2. Rezultati



Slika 7: Važnost varijabli za predviđanje K3 ocjene prema Random Forest modelu. Graf pokazuje postotak povećanja srednje kvadratne greške (MSE) kada se određena varijabla isključi iz modela. K2_3 komponenta ima daleko najveću važnost (preko 30% povećanja MSE), što ukazuje na visoku koreliranost između K2 i K3 kriterija. Visoka prediktivnost K3 ocjena iz drugih komponenti dovodi u pitanje nezavisnost i zasebnu vrijednost K3 kriterija.

Model pokazuje iznenađujuće visoke performanse:

- $R^2 = 0.699$ - model objašnjava 70% varijance K3 ocjena
- **RMSE = 1.085** - prosječna greška predviđanja je samo 1 bod

Dio II

Dodatne statističke analize

11. Analiza slaganja reczenzenata - Kendall-ov W

Kao što je već ranije rečeno, u svim sustavima vrednovanja utemeljenima na ocjenama reczenzenata krucijalno je pitanje u kojoj se mjeri recenzenti slažu u svojim procjenama. Naime, ako je slaganje nisko, konačni rezultat vrednovanja u većoj mjeri odražava individualne preferencije i interpretacije pojedinih reczenzenata, a manje objektivnu kvalitetu prijedloga. To može imati dalekosežne posljedice, od nepravedne raspodjele financijskih sredstava do smanjenja povjerenja u čitav postupak vrednovanja. Stoga je kvantificiranje stupnja slaganja među recenzentima jedan od temeljnih koraka u procjeni pouzdanosti sustava vrednovanja.

Kendall-ov koeficijent konkordancije (W) predstavlja standardnu i robusnu mjeru slaganja u situacijama kada više ocjenjivača rangira isti skup objekata. Za razliku od drugih mjera, Kendall-ov W se ne fokusira na apsolutne ocjene, već na konzistentnost rangova koje su recenzenti dodijelili projektima, čime omogućuje uvid u sličnost reczenzenata u percepciji relativne kvalitete prijedloga. Vrijednosti W kreću se od 0 (potpuno neslaganje) do 1 (savršeno slaganje), što omogućuje jednostavnu i intuitivnu interpretaciju razine usklađenosti među ocjenjivačima.

Premisa kod analize ovog postupka vrednovanja bila je da je svaki projekt ocijenilo točno tri različita recenzenta, pri čemu je pretpostavka bila kako svaki recenzent sudjeluje u vrednovanju samo jednog projekta. Ova pretpostavka važna je jer omogućuje tretiranje ocjena kao nezavisnih, čime se Kendall-ov W može interpretirati kao mjerilo isključivo usklađenosti među različitim ocjenjivačima za pojedini projekt. Međutim, u stvarnosti je vrlo vjerojatno da je manji broj reczenzenata sudjelovalo u ocjenjivanju više od jednog prijedloga. Ta činjenica mijenja interpretaciju rezultata na način da niže vrijednosti W (ovdje 0.398) mogu dijelom odražavati ne samo neslaganje unutar trojke reczenzenata po projektu, već i heterogenost u načinima ocjenjivanja među istim recenzentima kroz različite projekte. Drugim riječima, dobivena vrijednost Kendall-ovog W u ovom kontekstu mjeri slaganje u sustavu koji je zapravo "mješavina" različitih stilova recenziranja, a ne u potpunosti nezavisne trojke za svaki projekt.

Ovakva struktura podataka otvara važno pitanje u kojoj mjeri sustav zaista osigurava konzistentne standarde ocjenjivanja među svim uključenim recenzentima? Umjereno niska vrijednost Kendall-ovog W (0.398) i distribucija raspona ocjena, s razlikama koje za pojedine projekte dosežu i do osam bodova, jasno ukazuju na to kako zajednički standard nije postignut. Ovi rezultati stoga pružaju ne samo mjeru trenutne pouzdanosti sustava, nego i argument za njegovo metodološko unapređenje, npr. kroz kalibraciju reczenzenata, jasnije definiranje kriterija ili ujednačavanje opterećenja među ocjenjivačima. Ovi rezultati ukazuju na nužnost uvođenja sustavne kontrole i ujednačavanja načina dodjele reczenzenata projektima, kako bi se smanjila pristranost i povećala konzistentnost u ocjenjivanju.

11.1. Opis metode

Kendall-ov koeficijent konkordancije (W) mjeri stupanj slaganja među recenzentima pri rangiranju projekata, a računa se prema formuli:

$$W = \frac{12S}{m^2(n^3 - n)} \quad (5)$$

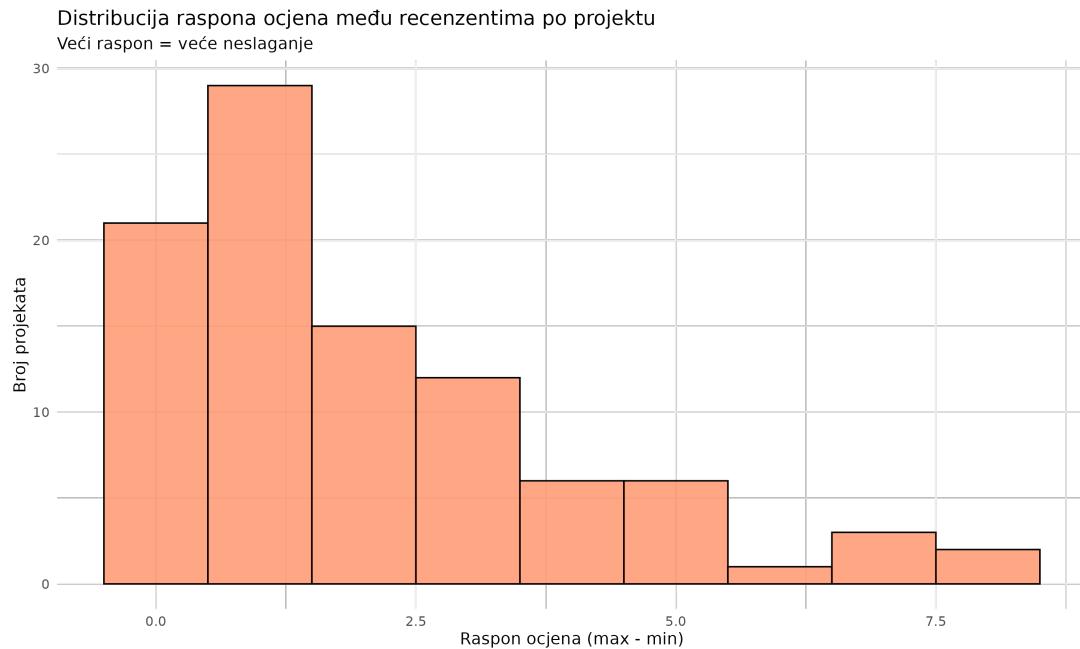
gdje je:

- $S = \sum_{i=1}^n (R_i - \bar{R})^2$ - suma kvadrata odstupanja rangova
- R_i - suma rangova za i -ti projekt
- $\bar{R} = \frac{m(m+1)}{2}$ - očekivana srednja vrijednost rangova
- m = broj reczenzenata (3)
- n = broj projekata (95)

W vrijednosti kreću se od 0 (potpuno neslaganje) do 1 (savršeno slaganje).

11.2. Rezultati

- Kendall-ov $W = 0.398$ ($p = 0.097$)
- Interpretacija: umjereno nisko slaganje među recenzentima
- P-vrijednost je granično neznačajna (> 0.05)



Slika 8: **Distribucija raspona ocjena među recenzentima po projektu.** Histogram prikazuje učestalost različitih raspona ocjena (max - min) koje su projekti dobili od tri recenzenta. X-os pokazuje raspon ocjena, gdje 0 znači potpuno slaganje (svi recenzenti dali istu ocjenu), a veće vrijednosti ukazuju na veće neslaganje. Y-os predstavlja broj projekata s danim rasponom. Većina projekata (oko 50) pokazuje raspon 0-1, što ukazuje na relativno dobro slaganje. Međutim, značajan broj projekata (oko 45) ima raspon 2-3 boda, a neki čak i do 8 bodova razlike, što predstavlja više od 50% ukupnog raspona skale. Ova distribucija ilustrira nedosljednost u primjeni kriterija ocjenjivanja među recenzentima.

12. Analiza diskriminativnosti ocjena

Jedna od ključnih ciljeva sustava vrednovanja je međusobno razlikovanje projektnih prijedloga ili mentora (tj, njihova diskriminacija), odnosno sposobnost razlikovanja onih koji se ističu kvalitetom od onih koji zaostaju. Ako sustav ocjenjivanja proizvodi gotovo jednake ocjene za sve prijedloge, on gubi svoju osnovnu funkciju pružanja diferencirane informacije temeljene na jasno definiranim kriterijima. Drugim riječima, ocjene koje se koncentriraju u uskom rasponu ne omogućavaju pouzdano rangiranje projekata niti pružaju korisne podatke za donošenje odluka o financiranju.

Analiza diskriminativnosti provedena je kroz tri komplementarne mjere. Koeficijent varijacije (CV) pruža informaciju o relativnoj varijabilnosti ocjena pri čemu niske vrijednosti upućuju na homogenu distribuciju i slabu diferencijaciju projekata. Bayesov čimbenik (BF) u okviru ANOVA pristupa kvantificira dokaze za postojanje razlika među grupama (projektima), gdje vrijednosti blizu 1 sugeriraju nedostatak snažnih dokaza za razlike. Relativna entropija pak mjeri informativnost distribucije ocjena pri čemu što je ona niža, to ocjene manje odražavaju raznoliku procjenu kvalitete prijedloga. Kombiniranjem ovih metrika dobiva se cjelovit uvid u to koliko sustav ocjenjivanja zaista razlikuje projekte.

Rezultati ove analize jasno pokazuju ozbiljna ograničenja trenutnog sustava. CV od svega 0.0749 ukazuje na izuzetno nisku varijabilnost (sve ocjene su praktički grupirane uz gornji kraj skale), Bayesov čimbenik od 2.02 pruža tek slabu evidenciju za postojanje značajnih razlika među projektima (mentorima), dok relativna entropija od 0.36 sugerira nisku informativnost distribucije. U praksi to znači kako gotovo svi projekti završavaju s ocjenama u vrlo uskom rasponu, čime se onemogućava smisleno rangiranje i postavljanje jasnih prioriteta za financiranje.

Implikacije su kristalno jasne. Sustav vrednovanja koji ne uspijeva dovoljno razlikovati projekte ne ispunjava svoju primarnu funkciju. U tom slučaju je potrebno razmotriti širu primjenu diferenciranih kriterija, revidirati skale ocjenjivanja ili provesti dodatnu kalibraciju recenzentata kako bi se povećala raspršenost ocjena i omogućilo pravednije i transparentnije rangiranje projekata.

12.1. Opis metode

Diskriminativnost se analizira kroz tri mjere:

12.1.1. Koeficijent varijacije (CV)

$$CV = \frac{\sigma}{\mu} = \frac{SD \text{ prosječnih ocjena}}{\text{Prosjek prosječnih ocjena}} \quad (6)$$

CV mjeri relativnu varijabilnost - niže vrijednosti ukazuju na manju sposobnost razlikovanja.

12.1.2. Bayesov čimbenik za ANOVA

$$BF_{10} = \frac{P(\text{podatci}|H_1)}{P(\text{podatci}|H_0)} \quad (7)$$

gdje H_0 : sve grupe imaju istu srednju vrijednost, H_1 : grupe se razlikuju.

12.1.3. Relativna entropija

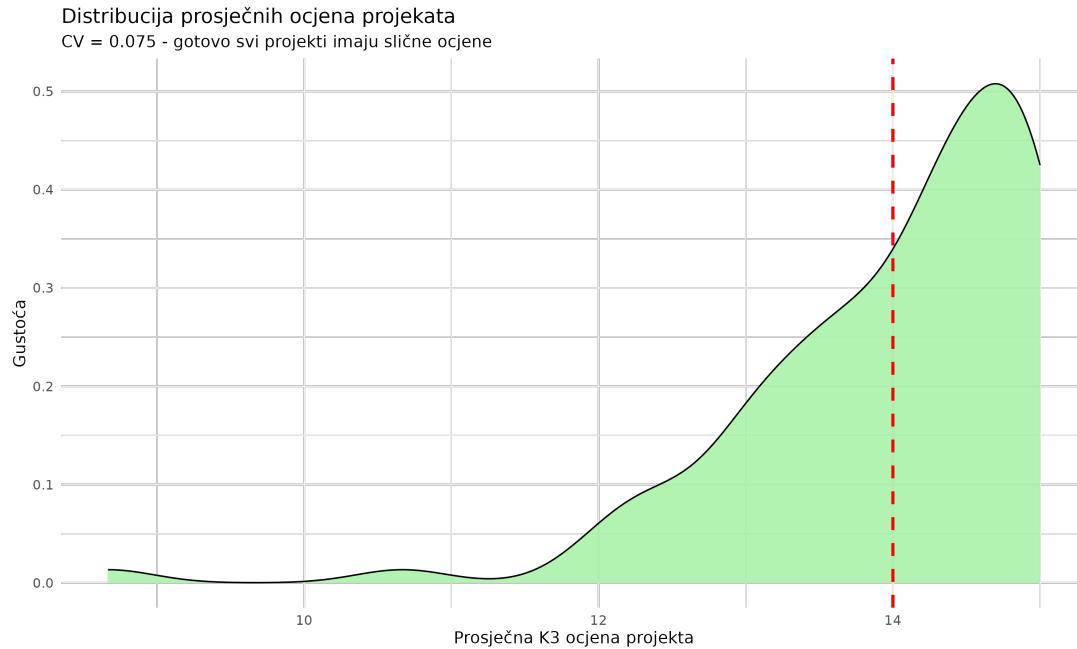
$$H = - \sum_{i=1}^k p_i \log(p_i) \quad (8)$$

$$H_{\text{rel}} = \frac{H}{H_{\max}} = \frac{H}{\log(k)} \quad (9)$$

gdje su p_i proporcije u svakoj kategoriji, a k broj kategorija.

12.2. Rezultati

- **CV = 0.0749** - izuzetno niska varijabilnost (7.49%)
- **BF = 2.02** - slaba evidencija za razlike među projektima
- **Relativna entropija = 0.36** - niska informativnost distribucije



Slika 9: **Distribucija prosječnih K3 ocjena projekata.** Graf gustoće prikazuje gustoću vjerojatnosti prosječnih K3 ocjena svih 95 projekata. X-os predstavlja prosječnu K3 ocjenu (raspon 3-15), a Y-os gustoću. Zeleno područje ispod krivulje vizualizira distribuciju. Crvena isprekidana vertikalna linija označava ukupni prosjek (oko 14). Distribucija pokazuje izrazitu koncentraciju oko prosjeka s vrlo malom varijabilnošću (CV = 0.075). Gotovo svi projekti imaju prosječne ocjene između 13 i 15, što čini samo 17% ukupnog raspona skale. Ova ekstremna koncentracija ocjena čini praktički nemogućim smisleno rangiranje projekata prema kvaliteti.

13. Usporedba s nasumičnim ocjenjivanjem

Jedan od najsnažnijih načina za procjenu vrijednosti sustava vrednovanja je usporediti ga s hipotetskim scenarijem u kojem bi ocjene bile dodijeljene potpuno nasumično. Ako stvarni sustav pokazuje samo minimalno bolje rezultate od onih koji bi se očekivali slučajnim ocjenjivanjem, to znači kako evaluacija ne dodaje značajnu informaciju o kvaliteti projekata i kako je njezina funkcionalnost ozbiljno upitna. Drugim riječima, ovakva usporedba služi kao "donja granica" učinkovitosti sustava. Sve što ne nadmašuje nasumično ocjenjivanje u dovoljnoj mjeri, teško može biti smatrano pouzdanim i opravdanim.

Za tu svrhu provedena je Monte Carlo simulacija s 1000 iteracija. U svakoj simulaciji ocjene su generirane nasumično, ali uz zadržavanje empirijske distribucije stvarnih ocjena, čime je osigurano da usporedba ne bude umjetno iskrivljena zbog razlika u skali ili u obliku distribucije. Na taj način dobivena je nul-distribucija Kendall-ovog W, koja predstavlja očekivanu razinu slaganja među recenzentima kada bi ocjene bile potpuno nasumične. Stvarna vrijednost Kendall-ovog W potom je uspoređena s ovom distribucijom kako bi se procijenilo koliko evaluacija nadmašuje scenarij slučajnog ocjenjivanja.

Rezultati su zabrinjavajući. Naime, stvarno slaganje među recenzentima ($W \approx 0.398$) samo je marginalno veće od prosjeka nasumičnih simulacija ($W \approx 0.333$), dok p-vrijednost od 0.06 pokazuje kako bi takvo slaganje moglo nastati i čistom slučajnošću. Drugim riječima, samo 6% nasumičnih simulacija proizvelo je jednako ili veće slaganje od stvarnog, što ozbiljno dovodi u pitanje informativnu vrijednost cijelog postupka ocjenjivanja.

Implikacije ovog rezultata su dalekosežne. Ako sustav vrednovanja tek neznatno nadmašuje nasumično dodjeljivanje ocjena, tada on ne ispunjava svoju svrhu razlikovanja kvalitetnih od manje kvalitetnih prijedloga. Praktična primjena ove metodologije mogla bi biti uvođenje Monte Carlo simulacija kao standardnog dijela postupka vrednovanja tj. za svaku evaluaciju procijeniti koliko je stvarno slaganje i diferencijacija projekata bolje od onoga što bi se očekivalo nasumično. Takva provjera omogućila bi pravovremeno otkrivanje sustava koji ne daju dovoljno informativne rezultate te bi pružila argumentirane temelje za njihovu reformu.

13.1. Opis metode

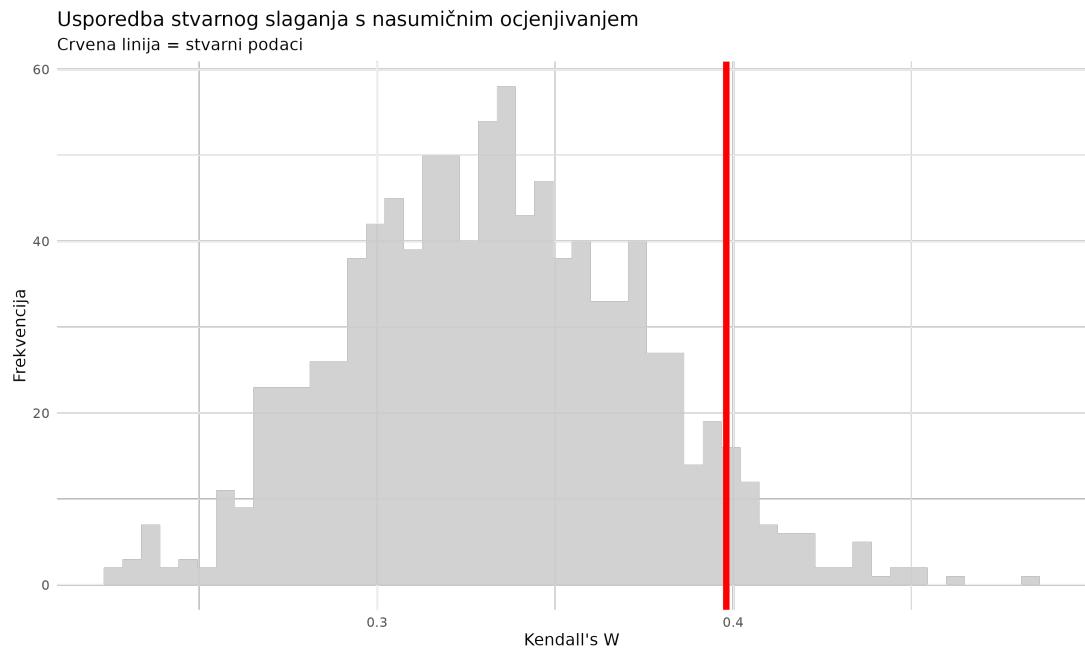
Provedena je Monte Carlo simulacija s 1000 iteracija gdje su ocjene generirane nasumično, ali zadržavajući empirijsku distribuciju originalnih ocjena. Za svaku simulaciju izračunat je Kendall-ov W te je kreirana nul-distribucija za usporedbu.

Postupak simulacije:

1. Izračunaj empirijsku distribuciju ocjena: $\hat{p}_i = \frac{n_i}{N}$ za $i \in \{1, 2, 3, 4, 5\}$
2. Za svaku simulaciju $j = 1, \dots, 1000$:
 - Generiraj nasumične ocjene s vjerojatnošću \hat{p}_i
 - Izračunaj W_j za nasumične podatke
3. Izračunaj p-vrijednost: $p = \frac{\#(W_j \geq W_{\text{stvarni}})}{1000}$

13.2. Rezultati

- **W stvarni = 0.398**
- **W nasumični (prosjek) = 0.333**
- **P-vrijednost = 0.06**



Slika 10: **Usporedba stvarnog slaganja recenzena s nasumičnim ocjenjivanjem.** Histogram prikazuje distribuciju Kendall-ov W koeficijenata dobivenih iz 1000 simulacija nasumičnog ocjenjivanja. X-os predstavlja vrijednosti Kendall-ov W (0 = nema slaganja, 1 = savršeno slaganje), a Y-os frekvenciju pojavljivanja. Sivi stupci pokazuju distribuciju nasumičnih W vrijednosti koje formiraju približno normalnu distribuciju s centrom oko 0.33. Crvena vertikalna linija označava W vrijednost iz stvarnih podataka (0.398). Činjenica da crvena linija pada u gornji rep distribucije ($p = 0.06$) pokazuje da je stvarno slaganje tek marginalno bolje od nasumičnog. Samo 6% nasumičnih simulacija proizvelo je jednako ili bolje slaganje, što postavlja ozbiljno pitanje o vrijednosti cijelog postupka ocjenjivanja.

14. Analiza ceiling učinka

Jedan od ključnih preduvjeta za kvalitetan sustav ocjenjivanja jest **dovoljna iskorištenost cijele ljestvice ocjena**. Ako recenzenti koriste samo gornji dio skale (npr. gotovo isključivo ocjene 4 i 5), sustav gubi sposobnost razlikovanja projekata različite kvalitete. Taj fenomen poznat je kao *ceiling učinak* (engl. *ceiling effect*) i predstavlja ozbiljan problem jer smanjuje diskriminativnu moć ocjenjivanja, otežava rangiranje projekata i čini nemogućim identifikaciju onih koji zaista značajno odskaču kvalitetom. U takvom okruženju ocjene postaju gotovo binarne, pri čemu se čitava ljestvica (1–3) praktički ne koristi.

Kako bi se kvantificirala prisutnost ceiling učinka, analiza je provedena kroz nekoliko koraka. Prvo je izračunata **deskriptivna statistika**, uključujući udio maksimalnih ocjena (5), udio ocjena većih ili jednakih 4 te koeficijent asimetrije distribucije. Visoka koncentracija ocjena na vrhu skale i negativna asimetrija ukazuju na izražen ceiling učinak. Zatim je primijenjen **Bayesov test protiv uniformne distribucije**, kojim se procjenjuje vjerojatnost da su ocjene rezultat uravnotežene, jednako raspodijeljene procjene u odnosu na alternativnu hipotezu – da recenzenti sustavno favoriziraju gornji dio skale.

14.1. Opis metode

Ceiling učinak analiziran je kroz:

14.1.1. Deskriptivnu statistiku

Za svaku varijablu izračunate su:

- Proporcija ocjena = 5: $p_5 = \frac{\#(x_i=5)}{n}$
- Proporcija ocjena ≥ 4 : $p_{4+} = \frac{\#(x_i \geq 4)}{n}$
- Koeficijent asimetrije: $\gamma_1 = \frac{\mu_3}{\sigma^3}$

14.1.2. Bayesov čimbenik protiv uniformne distribucije

Test hipoteze:

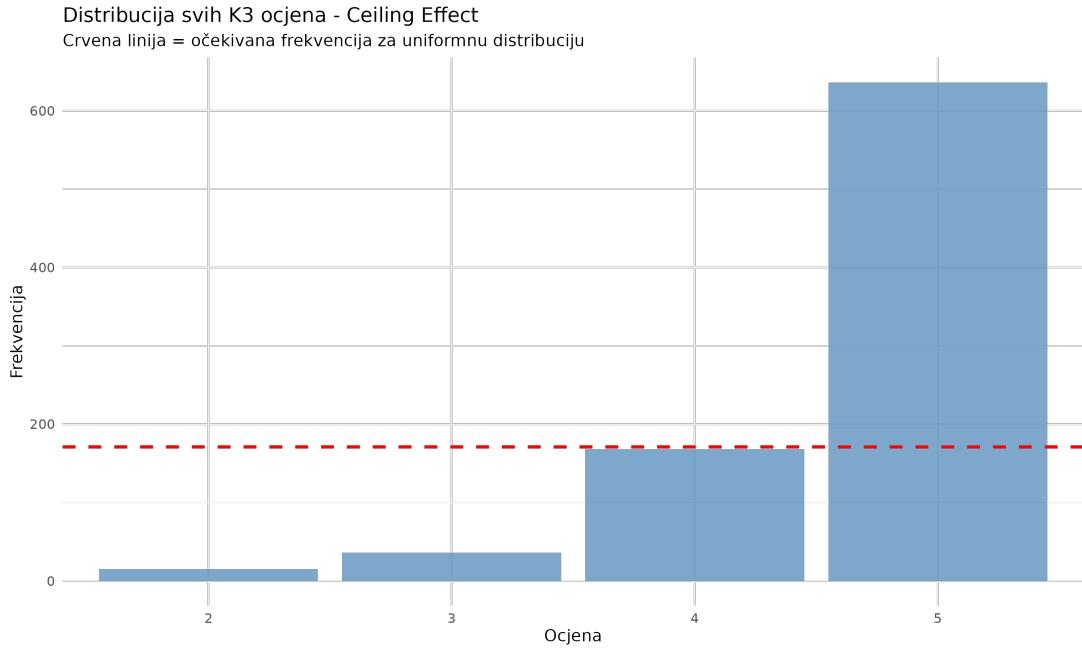
- H_0 : ocjene slijede uniformnu distribuciju $U(1, 5)$
- H_1 : ocjene slijede neku drugu distribuciju

$$BF_{10} = \frac{P(\text{podatci}|H_1)}{P(\text{podatci}|H_0)} \quad (10)$$

14.2. Rezultati

- 94% svih K3 ocjena je 4 ili 5
- $BF = 3.21 \times 10^{162}$ protiv uniformne distribucije
- Sve varijable pokazuju negativnu asimetriju (-1.2 do -3.8)

Praktične implikacije ovakvih rezultata su značajne: ceiling učinak onemogućuje precizno rangiranje projekata, otežava identifikaciju zaista vrhunskih prijedloga i može dovesti do toga da financiranje ovisi o zanemarivim razlikama unutar uskog gornjeg raspona ocjena.



Slika 11: **Distribucija svih K3 ocjena tj. demonstracija ceiling učinka.** Stupičasti graf prikazuje frekvenciju pojedinačnih K3 ocjena (1-5) za sve recenzente i projekte (ukupno 855 ocjena). X-os pokazuje ocjene od 1 do 5, a Y-os njihovu učestalost. Plavi stupci predstavljaju opažene frekvencije. Crvena isprekidana horizontalna linija na visini 171 označava očekivanu frekvenciju ako bi ocjene bile uniformno distribuirane ($855/5 = 171$). Dramatična razlika između opaženih i očekivanih frekvencija je očita: ocjene 1-3 pojavljuju se daleko rjeđe od očekivanog, dok su ocjene 4 i 5 drastično prekomjerno zastupljene. Ocjena 5 pojavljuje se gotovo 4 puta češće nego što bi se očekivalo slučajno. Ovaj ekstremni ceiling učinak čini većinu skale (1-3) praktički neupotrebljivom, svodeći efektivnu skalu na binarni izbor između 4 i 5.

Rezultati ove analize su nedvosmisleni: čak **94% svih K3 ocjena rezultati se u rasponu 4–5**, dok Bayesov čimbenik od 3.21×10^{162} gotovo u potpunosti odbacuje hipotezu o uniformnoj raspodjeli. Negativna asimetrija distribucije (od -1.2 do -3.8) dodatno potvrđuje izrazitu koncentraciju ocjena u vrhu skale. Drugim riječima, većina skale (ocjene 1–3) praktički je neupotrebljiva, što drastično smanjuje informativnu vrijednost ocjena i svodi sustav na gotovo binarni izbor između "dobro" i "vrlo dobro".

Konkretnе intervencije koje bi mogle ublažiti ovaj problem uključuju:

- **Redefiniranje skale ocjenjivanja** – jasnije opisivanje značenja ocjena 1–3 kako bi se potaknula njihova primjena.
- **Kalibracijske radionice za recenzente** – kako bi se smanjila tendencija korištenja samo gornjih vrijednosti skale.
- **Uvođenje distribucijskih smjernica** (npr. zahtjev da recenzenti koriste širi raspon ocjena), čime bi se osigurala veća diferencijacija projekata.
- **Revizija pondera kriterija** – kako bi se spriječilo da recenzenti maksimalnim ocjenama prikrivaju razlike u kvaliteti među projektima.

Provedba ovih mjer značajno bi povećala raspon korištenih ocjena i time poboljšala diskriminativnost sustava vrednovanja.

16. Analiza ekstremnih obrazaca ocjenjivanja

4

17. Analiza ekstremnih obrazaca ocjenjivanja

17.1. Uvod

Jedan od ključnih ciljeva cjelokupne analize ovog sustava vrednovanja bio je razumjeti ne samo kvantitativne razlike među ocjenama, već i kvalitativne obrasce koji upućuju na način na koji su recenzenti pristupali vrednovanju. Prethodno iznijeti rezultati, uključujući analizu informativnosti kriterija putem normalizirane međusobne informacije (NMI) i analize glavnih komponenti (PCA), ukazali su na postojanje umjerene redundantnosti među kriterijima te na dominantnu latentnu dimenziju koja objašnjava značajan dio varijabilnosti ocjena. Posebno visoka povezanost među komponentama kriterija K3 otvorila je pitanje u kojoj mjeri recenzenti zaista diferenciraju njegove podkomponente, ili ih doživljavaju kao jednu cjelovitu mjeru.

Analiza distribucije ocjena dodatno je naglasila problem jer je čak 70% svih dodijeljenih ocjena bilo maksimalno (5), dok je obrazac "sve petice" dominirao među recenzijama, pojavljujući se u 15.1% slučajeva. Ovakvi rezultati sugeriraju kako značajan broj reczenzata pokazuje minimalnu varijabilnost u ocjenjivanju, što dovodi u pitanje diskriminativnost sustava i njegovu sposobnost da razlikuje projekte (mentore) prema stvarnoj kvaliteti.

U tom kontekstu, analiza ekstremnih obrazaca ocjenjivanja predstavlja logičan nastavak prethodnih rezultata. Dok su NMI i PCA pružili uvid u strukturalne odnose među kriterijima i latentne dimenzije koje njima upravljaju, analiza ekstremnih obrazaca usmjerena je na identifikaciju konkretnih obrazaca ponašanja reczenzata koji potkopavaju svrhu detaljnog sustava vrednovanja s devet komponenti. Ekstremni obrasci, poput dodjeljivanja istih ocjena za sve podkriterije ili davanja univerzalno maksimalnih ocjena, ukazuju na površan, "*copy-paste*" pristup vrednovanju. Takvi obrasci ne samo da umanjuju informativnu vrijednost pojedinih kriterija, već i povećavaju rizik pristranosti i nepravednog vrednovanja projekata.

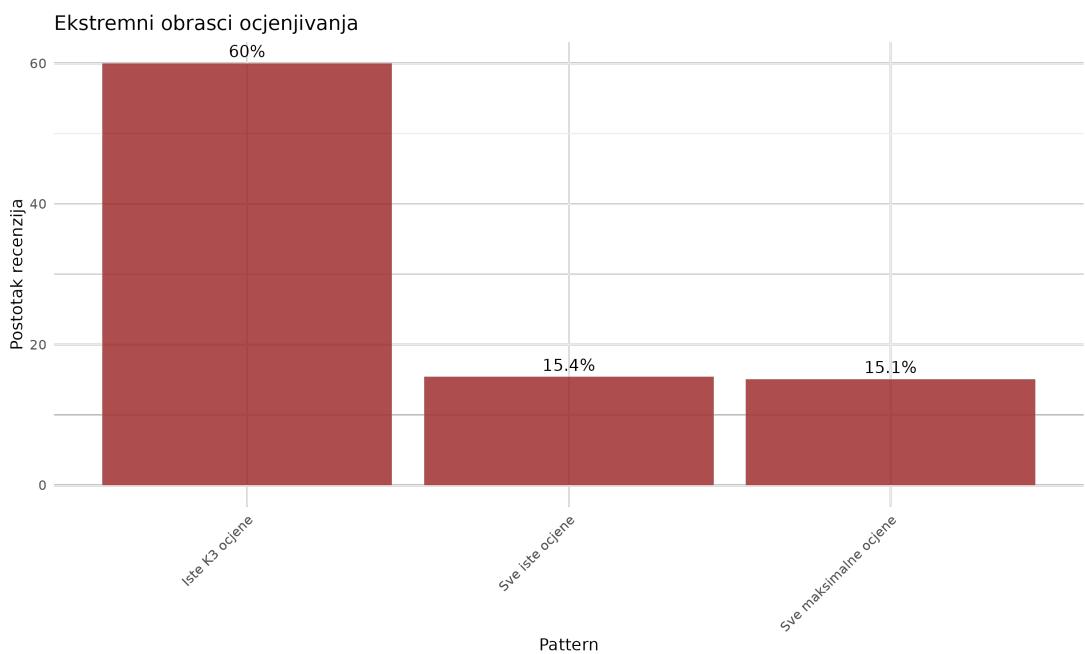
Stoga se ova analiza provela s ciljem kvantificiranja učestalosti takvih obrazaca, procijene njihovog utjecaja na ukupnu pouzdanost sustava vrednovanja i pružanje empirijskog temelja za prijedloge revizije kriterija i metodologije ocjenjivanja. Rezultati dobiveni ovom analizom predstavljaju ključni dokaz u raspravi o potrebama racionalizacije sustava, uključujući preispitivanje uloge i definicije kriterija K3 te potencijalno smanjenje broja kriterija uz očuvanje informacijske vrijednosti vrednovanja.

17.2. Opis metode

Identificirani su sljedeći ekstremni obrasci:

1. **Sve iste ocjene:** $K_{i,1} = K_{i,2} = K_{i,3}$ za sve kriterije
2. **Iste K3 ocjene:** $K_{31} = K_{32} = K_{33}$
3. **Sve maksimalne:** sve ocjene = 5

17.3. Rezultati



Slika 12: **Ekstremni obrasci ocjenjivanja.** Stupičasti graf prikazuje postotke recenzija koje pokazuju tri vrste ekstremnih obrazaca ocjenjivanja. X-os prikazuje tip obrasca, a Y-os postotak recenzija. Crveni stupci s označenim postocima ilustriraju prevalenciju svakog obrasca. Zabrinjavajući je podatak da 60% recenzija daje identične ocjene za sve tri K3 komponente, što sugerira da većina recenzentata ne razlikuje različite aspekte ovog kriterija. Dodatno, 15.4% recenzija daje potpuno identične ocjene za sve komponente svih kriterija, a 15.1% daje maksimalne ocjene za sve. Ovi obrasci ukazuju na površno ocjenjivanje gdje recenzenti ne ulažu napor u razlikovanje različitih aspekata kvalitete projekta. Takav "*copy-paste*" pristup potpuno potkopava svrhu detaljnog sustava vrednovanja s 9 komponenti.

Dio III

Bayesova analiza

Kako bi se nadopunile klasične statističke metode i dublje razumjela struktura podataka, u trećem dijelu provedena je **Bayesova analiza**. Ovaj pristup je omogućio procjenu varijance ocjena i njenu dekompoziciju na različite izvore (projekti (mentor), recenzenti, rezidualna komponenta), uz istodobnu kvantifikaciju nesigurnosti. Korištenjem *Bayesovih hijerarhijskih modela* moguće je obuhvatiti kompleksne odnose između ocjena i izvora varijabilnosti, što je osobito važno u situacijama kada klasični modeli pokazuju ograničenja zbog ekstremne homogenosti podataka.

Ova analiza daje uvid u to **koliko ocjene odražavaju stvarne razlike među projektima (mentorima), a koliko su posljedica subjektivnih sklonosti recenzenata ili slučajnih čimbenika**. Također, Bayesov okvir omogućuje usporedbu modela, provjeru pristranosti po pozicijama te dodatne mjere poput koeficijenta varijacije i relativne entropije, čime se osigurava sveobuhvatna procjena diskriminativnosti i pouzdanosti sustava. Time Bayesova analiza predstavlja ključan korak u potvrdi prethodnih rezultata i pruža čvrste statističke dokaze o ograničenjima trenutnog postupka vrednovanja.

18. Struktura podataka i osnovni pokazatelji

Za razumijevanje funkcioniranja sustava vrednovanja nužno je najprije analizirati strukturu dostupnih podataka i osnovne deskriptivne pokazatelje. Ovaj dio pruža pregled uzorka ocjena, uključujući broj projekata, recenzija i način raspodjele recenzenata, kao i temeljne statističke pokazatelje za ključni kriterij K3. Analiza distribucije ocjena posebno je važna jer omogućava prepoznavanje obrazaca u dodjeljivanju bodova, poput koncentracije ocjena na vrhu skale, što može ukazivati na prisutnost pristranosti, nedostatka diferencijacije među projektima ili tzv. ceiling učinka. Ovi rezultati služe kao polazište za dublje analize pouzdanosti, pristranosti i diskriminativnosti sustava ocjenjivanja.

18.1. Struktura uzorka

Analizirani uzorak sastoji se od:

- **285 jedinstvenih recenzija za 95 projekata**
- Svaki projekt ocijenjen je od **3 različita recenzenta** (R1, R2, R3)
- **Niti jedan recenzent nije ocijenio više od jednog projekta**

18.2. Distribucija ocjena

Osnovna deskriptivna statistika pokazuje ekstremnu pristranost prema visokim ocjenama:

	Min.	1. kvartil	Medijan	Prosjek	3. kvartil	Maks.
K3 ocjena	7	14	15	14	15	15

Ključni rezultati:

- **Medijan = 15**: više od 50% recenzenata daje maksimalne ocjene
- **Prosjek = 14** od mogućih 15 bodova
- **75% recenzenata** daje ocjene između 14 i 15
- **55.8% recenzenata** dalo je barem jednu savršenu ocjenu (15/15)

19. Bayesovi hijerarhijski modeli

Nakon primjene klasičnih i naprednih statističkih metoda, u ovom dijelu analize pristupili smo **Bayesovom hijerarhijskom modeliranju** kako bismo dublje razumjeli izvore varijabilnosti u ocjenama te kvantificirali njihov doprinos ukupnoj varijanci. Dok klasični modeli procjenjuju varijance i učinke na temelju točkastih procjena, Bayesov okvir omogućuje **cjelovitu vjeratnosnu interpretaciju** tj. procjenu distribucija parametara, uzimajući u obzir nesigurnost i pružajući robusnije uvide u strukturu podataka. Ovaj pristup je posebno važan kod sustava vrednovanja poput analiziranog, kada se sumnja na snažan utjecaj subjektivnih čimbenika i homogenost ocjena.

Hijerarhijski model omogućava **dekompoziciju ukupne varijance** ocjena na tri glavna izvora: stvarne razlike između projekata, razlike među recenzentima (njihova strogost, pristrandost i stilova ocjenjivanja) te rezidualne, nasumične varijacije. Takva dekompozicija pruža empirijsku osnovu za odgovor na ključno pitanje *odražavaju li dodijeljene ocjene doista kvalitetu projekata (mentor-a) ili su prvenstveno rezultat individualnih razlika među recenzentima i slučajnih čimbenika?*

Implementacija Bayesovog modela donosi dodatnu vrijednost i zbog mogućnosti **procjene pouzdanosti modela** (konvergencija, dijagnostika lanca, procjena stabilnosti parametara), čime se osigurava vjerodostojnost dobivenih rezultata. Pritom se posebno ističe značaj konvergencijskih pokazatelja poput *Rhat*, *E-BFMI* i stope divergentnih tranzicija, koji omogućavaju detekciju problema u modeliranju podataka.

U konačnici, rezultati Bayesove analize nisu samo tehnički pokazatelji. Oni daju **jasnu sliku strukture ocjena** otkrivajući kako tek manji dio varijabilnosti proizlazi iz stvarnih razlika među projektima (mentorima), dok je dominantni dio posljedica razlika među recenzentima i slučajnih čimbenika. Takvi rezultati imaju dalekosežne implikacije jer upućuju na temeljne slabosti sustava vrednovanja, potvrđuju rezultate prethodnih analiza (klasične statistike, PCA i informativnosti kriterija) te dodatno naglašavaju potrebu za njegovom reformom.

19.1. Opis metode

Bayesov hijerarhijski model omogućava dekompoziciju varijance ocjena na različite izvore. Osnovni model ima oblik:

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \quad (15)$$

gdje je:

- y_{ij} = ocjena za projekt i od recenzenta j
- μ = opći prosjek
- $\alpha_i \sim N(0, \sigma_{\text{projekt}}^2)$ = slučajni učinak projekta
- $\beta_j \sim N(0, \sigma_{\text{recenzent}}^2)$ = slučajni učinak recenzenta
- $\epsilon_{ij} \sim N(0, \sigma_{\text{rezidual}}^2)$ = rezidualna greška

19.2. Problemi s konvergencijom

Model je pokazao ozbiljne probleme s konvergencijom:

- **16.0% divergentnih tranzicija**
- **$E\text{-BFMI} < 0.3$** za sve 4 lance
- **$Rhat > 1.05$** za neke parametre

Ovi problemi ukazuju na degeneriranu strukturu podataka - model ne može adekvatno opisati podatke zbog ekstremne homogenosti ocjena.

19.3. Dekompozicija varijance

Proporcija ukupne varijance objašnjene različitim izvorima:

$$\text{ICC}_{\text{projekt}} = \frac{\sigma_{\text{projekt}}^2}{\sigma_{\text{projekt}}^2 + \sigma_{\text{recenzent}}^2 + \sigma_{\text{rezidual}}^2} = 0.09 \quad (16)$$

$$ICC_{recenzent} = \frac{\sigma_{recenzent}^2}{\sigma_{projekt}^2 + \sigma_{recenzent}^2 + \sigma_{rezidual}^2} = 0.32 \quad (17)$$

- **Projekti:** 9% varijance
- **Recenzenti:** 32% varijance
- **Rezidualna:** 59% varijance

19.4. Interpretacija rezultata

Rezultati Bayesove hijerarhijske analize pružaju dubok uvid u strukturu varijabilnosti ocjena i potvrđuju rezultate prethodnih metoda (klasične statistike, PCA i analize informativnosti kriterija).

Problemi s konvergencijom Visoka stopa divergentnih tranzicija (16%), nizak E-BFMI (<0.3 za sve lance) i Rhat veći od 1.05 za neke parametre ukazuju na to kako je model imao poteškoća s postizanjem stabilne procjene parametara. To se u Bayesovom modeliranju obično povezuje s **degeneriranom strukturu podataka**, što znači kako podaci nemaju dovoljno varijabilnosti da bi model mogao razlikovati učinke projekata od učinaka recenzenata i rezidualne varijance. Ovi problemi snažno upućuju na **ekstremnu homogenost ocjena**, odnosno na činjenicu kako su recenzenti davali vrlo slične, često maksimalne ocjene, što potvrđuje i ranije analize distribucije (medijan = 15, prosjek = 14). Time se dodatno naglašava prisutnost tzv. *ceiling učinka*, gdje ocjene dosežu gornju granicu skale i gube diskriminativnu snagu.

Dekompozicija varijance Rezultati dekompozicije varijance pokazuju kako **stvarne razlike između projekata objašnjavaju tek 9% ukupne varijabilnosti ocjena**. Suprotno tome, **učinci recenzenata čine 32% varijance**, dok **rezidualna komponenta iznosi čak 59%**. To znači kako je više od trećine ukupnih varijacija u ocjenama posljedica individualnih sklonosti recenzenata (njihove subjektivne strogosti, pristupa ili pristranosti), a dominantni dio preostale varijance je nasumičan, bez jasne povezanosti s kvalitetom projekata. Ovaj rezultat otkriva kako sustav ocjenjivanja u sadašnjem obliku ne razlikuje projekte (mentore) prema njihovoj stvarnoj kvaliteti, već odražava kombinaciju recenzentskih preferencija i slučajnih čimbenika.

Implikacije za sustav ocjenjivanja Ovi rezultati imaju dalekosežne implikacije. Prvo, **niska intraklasa korelacija za projekte (ICC = 0.09)** znači kako se ocjene pojedinih projekata ne mogu smatrati pouzdanim odrazom njihove kvalitete. Drugo, značajan udio varijance koji potječe od recenzenata pokazuje kako bi rezultati vrednovanja mogli biti značajno različiti kada bi se promijenio skup ocjenjivača, što dovodi u pitanje **pravednost i konzistentnost sustava**. Treće, velika rezidualna varijanca upućuje na **nedostatak sistematičnosti u ocjenjivanju**, što dodatno umanjuje informativnu vrijednost dodijeljenih ocjena.

Povezanost s prethodnim rezultatima Ovi rezultati se nadovezuju na rezultate PCA analize i normalizirane međusobne informacije (NMI), prema kojima su kriteriji pokazali **umjerenu redundantnost**, a jedna latentna dimenzija (PC1) objašnjava veliki dio ukupne varijabilnosti. Bayesova analiza sada dodatno razotkriva kako ta latentna dimenzija nije vezana uz stvarne razlike među projektima, već u velikoj mjeri odražava **stilove (načine) ocjenjivanja i pristranosti recenzenata**. Time se potvrđuje kako sustav ocjenjivanja, iako formalno detaljan, u praksi ne ostvaruje diferenciranu procjenu projekata.

Zaključak Bayesove hijerarhijske analize Bayesova hijerarhijska analiza pružila je ključne dokaze o **struktturnim slabostima ovog sustava vrednovanja**. Ocjene u svojoj sadašnjoj formi više odražavaju individualne karakteristike recenzenata i slučajnu varijabilnost nego stvarne razlike u kvaliteti projekata. Ovi rezultati naglašavaju potrebu za **dubinskom reformom sustava**, koja bi mogla uključivati: (1) smanjenje složenosti kriterija, (2) dodatnu edukaciju recenzenata i uvođenje mehanizama za kontrolu pristranosti te (3) redefiniciju skala i postupka ocjenjivanja kako bi se povećala diskriminativna snaga ocjena.

20. Analiza pristranosti po pozicijama

U prethodnim analizama klasičnim statističkim metodama nije uočena značajna razlika među ocjenama s obzirom na poziciju recenzenta (R1, R2, R3), pri čemu su distribucije ocjena bile izrazito slične i snažno pristrane prema gornjem kraju skale. Ovi rezultati sugerirali su da **redoslijed recenziranja ne utječe na konačne ocjene**, što je pojačalo sumnju da recenzenti primjenjuju homogen i pojednostavljen pristup ocjenjivanju, neovisno o svojoj ulozi u procesu.

Važno je, međutim, naglasiti da ova analiza ima ozbiljna ograničenja. Naime, nije sasvim jasno jesu li recenzenti uopće bili svjesni svoje pozicije (R1, R2, R3), odnosno je li ta pozicija bila jasno definirana i dosljedno dodjeljivana u procesu vrednovanja. Još važnije, dostupni podaci ne omogućuju potpunu sigurnost u to da je svaki recenzent ocijenio samo jedan projekt. Naprotiv, postoje indikacije kako su neki recenzenti ocjenjivali više projekata, potencijalno u različitim ulogama. Takva situacija stvara tzv. "miš-maš" kompoziciju uzorka u kojoj uloge reczenzenta i njihova dodjela po projektima možda nisu bile sustavno kontrolirane. **Stoga rezultate o pristranosti po pozicijama treba promatrati s velikim oprezom jer oni mogu više odražavati heterogenost sastava reczenzenta nego stvarne razlike među pozicijama.**

Kako bi se klasični rezultati dopunili vjerovatnosnim, provedena je **Bayesova analiza pristranosti po pozicijama** korištenjem Bayesovog čimbenika (BF) za testiranje hipoteze o postojanju razlike među pozicijama reczenzenta. Ovaj pristup omogućava kvantificiranje dokaza u korist ili protiv postojanja razlike, uzimajući u obzir nesigurnost podataka. Rezultati tako ne služe samo za statističku potvrdu prethodnih rezultata, već i za pružanje robusnijeg okvira u procjeni *odražava li redoslijed recenziranja sustavne razlike u ocjenama*.

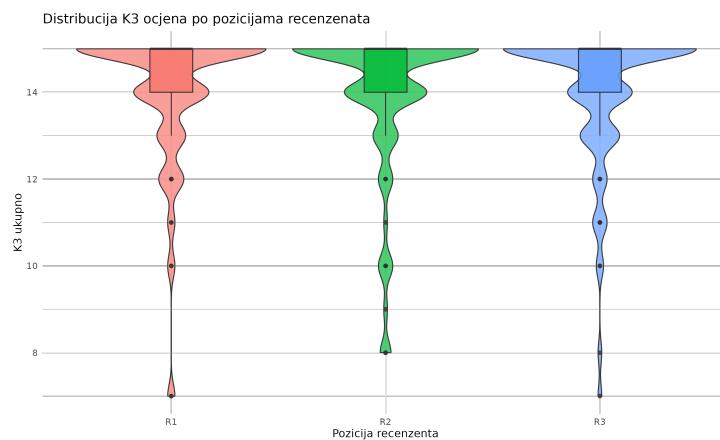
20.1. Bayesov čimbenik

Bayesov čimbenik za testiranje razlike među pozicijama reczenzenta (R1, R2, R3) računa se prema izrazu:

$$BF_{10} = \frac{P(D|H_1)}{P(D|H_0)} = 0.0403 \quad (18)$$

gdje je H_0 : nema razlike među pozicijama, a H_1 : postoji razlike.

$BF < 1$ pruža dokaze **protiv** postojanja razlike među pozicijama.



Slika 13: **Distribucija K3 ocjena po pozicijama reczenzenta.** Violinski graf prikazuje distribuciju K3 ukupnih ocjena (raspon 3-15) za tri pozicije reczenzenta (R1, R2, R3). Širina violina na svakoj visini predstavlja gustoću podataka na toj vrijednosti. Boxplot unutar svakog violinskog grafa pokazuje medijan (crna linija), interkvartilni raspon (kutija) i outliere (točke). Sve tri pozicije pokazuju izrazito sličnu distribuciju s visokom koncentracijom ocjena blizu maksimuma (15), što potvrđuje da pozicija recenzenta ne utječe na ocjenjivanje. Medijan za sve tri pozicije je na ili blizu maksimalne ocjene.

21. Usporedba s nasumičnim ocjenjivanjem

Provedena je simulacija s **1000 iteracija nasumičnog ocjenjivanja** korištenjem empirijske distribucije ocjena kako bi se procijenilo koliko se stvarni uzorak ocjena razlikuje od scenarija u kojem bi ocjene bile dodijeljene potpuno slučajno, ali zadržavajući njihovu stvarnu raspodjelu. Ovaj pristup omogućuje procjenu **stupnja slaganja među recenzentima** u kontekstu referentne točke – slučajnog ocjenjivanja – i testira hipotezu da li su uočene ocjene rezultat svjesnog i konzistentnog vrednovanja ili prije odražavaju razinu podudarnosti koja se može očekivati pukim slučajem.

Za procjenu razine slaganja korišten je **Kendall-ov koeficijent slaganja (W)**, koji mjeri konzistentnost rangiranja među ocjenjivačima. Vrijednosti Kendallovog W kreću se od 0 (potpuno neslaganje, tj. potpuna slučajnost) do 1 (potpuno slaganje među recenzentima).

Rezultati simulacije pokazuju:

- **Kendall-ov W za stvarne podatke:** **0.045** – što ukazuje na izuzetno nisku razinu slaganja među recenzentima.
- **Prosječni Kendall-ov W za nasumične podatke:** **0.046** – gotovo identičan onome iz stvarnih podataka.
- **P-vrijednost:** **0.451** – nema statistički značajne razlike između stvarnih i nasumično generiranih ocjena.

Ključni rezultat: *Nije moguće statistički razlikovati stvarne ocjene od onih koje bi bile dodijeljene nasumično* ($p = 0.451 > 0.05$). Drugim riječima, razina slaganja među recenzentima u stvarnim podacima jednaka je onoj koju bismo očekivali ako bi isti recenzenti dodjeljivali ocjene potpuno slučajno, bez ikakvog zajedničkog kriterija ili sustavnog vrednovanja.

Implikacije ovih rezultata Ovaj rezultat ima dalekosežne implikacije za pouzdanost sustava vrednovanja. Ako stvarne ocjene ne pokazuju viši stupanj slaganja od nasumično generiranih, to dovodi u pitanje **validnost dodijeljenih ocjena kao mjere kvalitete projekata**. U praksi to znači:

- Ocjene ne reflektiraju dosljedne standarde procjene među recenzentima.
- Nedostaje zajednički okvir ili kriterij koji bi osigurao usklađeno vrednovanje projekata.
- postupak vrednovanja u sadašnjem obliku ima nisku diskriminativnu moć i ne uspijeva razlikovati projekte prema stvarnoj kvaliteti.

Ovakvi rezultati posebno su zabrinjavajući u kontekstu prethodnih rezultata o **ekstremnoj homogenosti ocjena** (dominacija maksimalnih ocjena, slaba varijabilnost) te Bayesove dekompozicije varijance, koja je pokazala da tek 9% ukupne varijabilnosti potječe od stvarnih razlika među projektima. U kombinaciji, ovi rezultati snažno sugeriraju da trenutačni sustav vrednovanja **ne ostvaruje svoju primarnu funkciju – diferenciranu i pravednu procjenu kvalitete prijedloga**.

22. Analiza ceiling učinka

Jedan od najuočljivijih problema u analiziranom sustavu vrednovanja jest **ceiling učinak** tj. fenomen u kojem velik broj ocjena doseže gornju granicu skale, čime se značajno smanjuje sposobnost sustava da razlikuje projekte (mentore) po kvaliteti. Kada recenzenti dosljedno dodjeljuju maksimalne ili gotovo maksimalne ocjene, skala gubi diskriminativnu snagu, a ukupni rezultati prestaju odražavati stvarne razlike među ocjenjivanim prijedlozima. Ovaj problem posebno dolazi do izražaja u sustavima s malim rasponom skale (npr. 1–5 ili 1–15), gdje i mali pomaci prema maksimumu dovode do značajne koncentracije ocjena na vrhu.

Kako bi se kvantificirao ovaj učinak i utvrdilo koliko je raspodjela ocjena odstupila od očekivane, provedena je Bayesova analiza. Izračunat je Bayesov čimbenik u korist hipoteze da raspodjela ocjena nije uniformna, već pristrana prema višim vrijednostima (ceiling učinak), u odnosu na hipotezu uniformne raspodjele:

$$BF = \frac{P(D|\text{pristrana distribucija})}{P(D|\text{uniformna distribucija})} = 1.2 \times 10^{156} \quad (19)$$

Interpretacija Bayesovog čimbenika Dobiveni Bayesov čimbenik ($BF = 1.2 \times 10^{156}$) iznimno je visok i predstavlja **nepobitne dokaze** da su stvarne ocjene ekstremno pristrane prema gornjoj granici skale. Prema Jeffreysovoj skali interpretacije Bayesovih čimbenika, vrijednosti veće od 100 već se smatraju "izvanrednim dokazima". U ovom slučaju, Bayesov čimbenik prelazi taj prag za više od 150 redova veličine, što znači da su dokazi protiv hipoteze uniformne raspodjele **apsolutno uvjerljivi**.

Implikacije za sustav vrednovanja Ovakav rezultat potvrđuje kako recenzenti nisu koristili puni raspon skale, već su ocjenjivanje koncentrirali oko najviših vrijednosti. Praktične posljedice toga su višestruke:

- **Smanjena diskriminativna moć**, te sustav ne razlikuje učinkovito kvalitetu projekata (mentora) jer većina prijedloga prima gotovo identične (maksimalne) ocjene.
- **Smanjena pouzdanost**, pa ocjene ne odražavaju konzistentne procjene stvarnih razlika među projektima (mentorima), već odražavaju uniformno "pozitivno" ocjenjivanje.
- **Povećan rizik nepravednog vrednovanja**, pa se projekti (mentorji) različite kvalitete tretiraju jednako, čime se potkopava svrha detaljnog vrednovanja te smanjuje motivacija za pripremu vrhunskih prijedloga.

Povezanost s prethodnim analizama Ovaj rezultat je u potpunosti u skladu s prethodnim analizama distribucije ocjena, koje su pokazale kako su **medijan i gornji kvartil ocjena smješteni na samom vrhu skale**, kao i s rezultatima Bayesove dekompozicije varijance, gdje je otkriveno kako tek 9% ukupne varijabilnosti ocjena proizlazi iz stvarnih razlika među projektima. Ceiling učinak time dodatno potvrđuje kako je sustav ocjenjivanja u sadašnjem obliku **neadekvatan za diferencijaciju kvalitete projekata tj. mentora**.

Zaključak analize ceiling učinka Bayesovom analizom Analiza Bayesovim čimbenikom pruža **ekstremno jake dokaze** za postojanje izraženog ceiling učinka u ocjenama. U kombinaciji s drugim rezultatima, ovi rezultati snažno upućuju na potrebu za reformom skale i metodologije ocjenjivanja kako bi se povećala njezina diskriminativna snaga, pravednost i sposobnost preciznog rangiranja projekata tj. mentora.

23. Identifikacija ekstremnih obrazaca

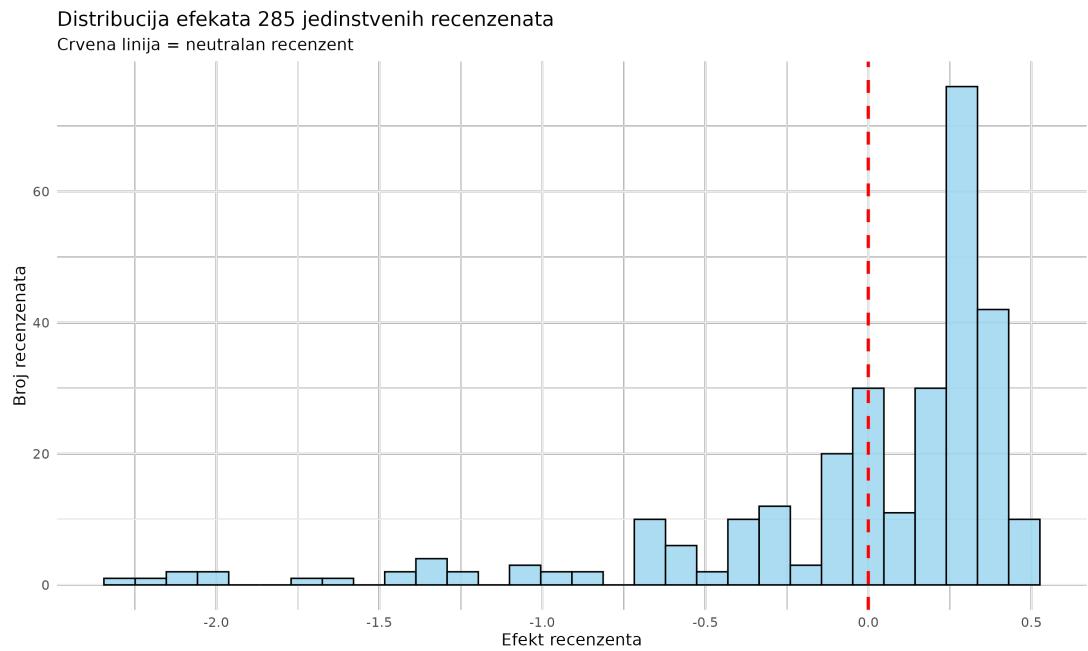
Jedan od ključnih ciljeva analize bio je **identificirati obrasce ocjenjivanja koji odstupaju od očekivanog ponašanja recenzenata** u sustavu dizajniranom za diferencirano i argumentirano vrednovanje projekata. Umjesto uravnotežene distribucije ocjena koja bi odražavala razlike među projektima i dimenzijama procjene, analiza je pokazala izrazitu sklonost recenzenata prema nekoliko ekstremnih i visoko problematičnih obrazaca ocjenjivanja.

Uočeni obrasci Analiza je identificirala sljedeće obrasce:

- **45.3% recenzija daje sve maksimalne ocjene (5-5-5 za sve komponente)**, pa gotovo polovica svih recenzija ne pokazuje nikakvu diferencijaciju među kriterijima, čime se potpuno gubi svrha višekomponentnog sustava ocjenjivanja. Ovakvo ponašanje recenzenata ukazuje na to da se proces vrednovanja svodi na formalnost, bez detaljnog razmatranja pojedinih dimenzija kvalitete prijedloga.
- **65.3% recenzija ima identične K3 komponente** pri čemu dvije trećine recenzija tretira K3 kriterije kao jedinstvenu cjelinu, što sugerira da recenzenti ne percipiraju ili ne vrednuju odvojeno njegove podkomponente. Ovaj rezultat je u skladu s ranjom analizom normalizirane međusobne informacije (NMI), koja je pokazala visoku redundantnost među K3 komponentama.
- **149 recenzenata (52%) ima standardnu devijaciju = 0 (uvijek iste ocjene)** što znači kako je više od polovice recenzenata dosljedno dodijelila iste ocjene, bez ikakve varijacije, što snažno upućuje na površno, "copy-paste" ocjenjivanje.

Interpretacija rezultata Ovi obrasci pokazuju da značajan broj recenzenata **ne koristi puni potencijal skale niti diferencira ocjene prema pojedinim komponentama**, već pribjegava ekstremno pojednostavljenom ocjenjivanju. Takvo ponašanje potkopava samu svrhu višekomponentnog sustava, čija je intanca bila obuhvatiti različite aspekte kvalitete prijedloga. Drugim riječima, vrednovanja u ovom postupku su se svela na nominalno popunjavanje obrazaca, bez stvarne analitičke vrijednosti.

Uloga recenzenata Distribucija slučajnih učinaka recenzenata iz Bayesovog hijerarhijskog modela (Slika 14) dodatno potvrđuje ove rezultate. Većina recenzenata grupirana je oko neutralne točke (0), no primjetan je **pomak prema pozitivnim vrijednostima**, što znači kako sustav karakterizira sustavna sklonost davanju viših ocjena (tzv. blagi recenzenti). Istovremeno, širina distribucije ukazuje na **značajnu varijabilnost među recenzentima**, što upućuje na nedostatak standardizacije i kalibracije u procesu vrednovanja.



Slika 14: **Distribucija efekata 285 jedinstvenih reczenzenta.** Histogram prikazuje distribuciju procijenjenih efekata reczenzenta iz Bayesovog hijerarhijskog modela. Crvena isprekidana linija označava neutralnog recenzenta ($učinak = 0$). Većina reczenzenta grupirana je oko nule s blagim pomakom prema pozitivnim vrijednostima, što ukazuje na opću sklonost davanju viših ocjena. Širina distribucije ilustrira značajnu varijabilnost među recenzentima u njihovim pristupima ocjenjivanju.

Praktične implikacije Ovi rezultati imaju ozbiljne posljedice za postupak vrednovanja:

- Sustav ne uspijeva ostvariti svoju svrhu, **diferencirano vrednovanje projekata prema njihovim kvalitetama**.
- Velik broj reczenzenta **pokazuje obrasce minimalnog angažmana**, što dovodi do površnih i homogenih ocjena.
- Takvi obrasci smanjuju **diskriminativnu moć** sustava i povećavaju rizik nepravednog rangiranja prijedloga.

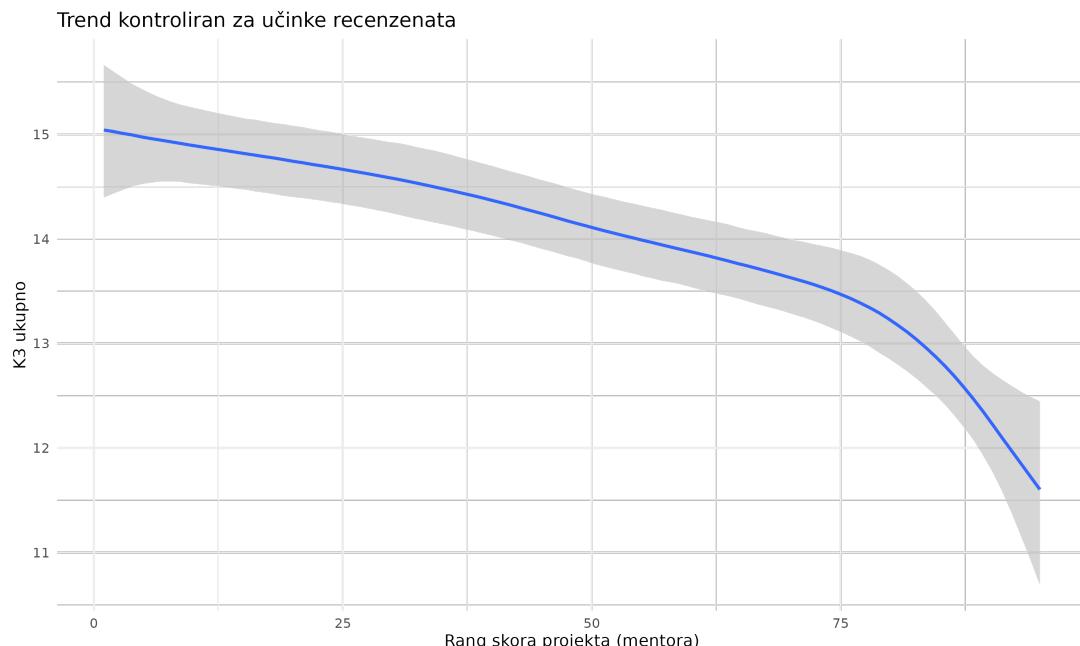
Identificirani obrasci ocjenjivanja jasno pokazuju kako se sustav vrednovanja u praksi nije koristi onako kako je vjerojatno zamišljeno. Umjesto alata za nijansirano vrednovanje, postao je mehanizam za dodjelu generičkih, često maksimalnih ocjena. **Ovi rezultati naglašavaju potrebu za dubinskom reformom procesa vrednovanja**, uključujući redefiniranje kriterija, uvođenje dodatne edukacije i kalibracije reczenzenta te tehničke mehanizme za detekciju i sprječavanje površnog ocjenjivanja.

24. Selekcija modela

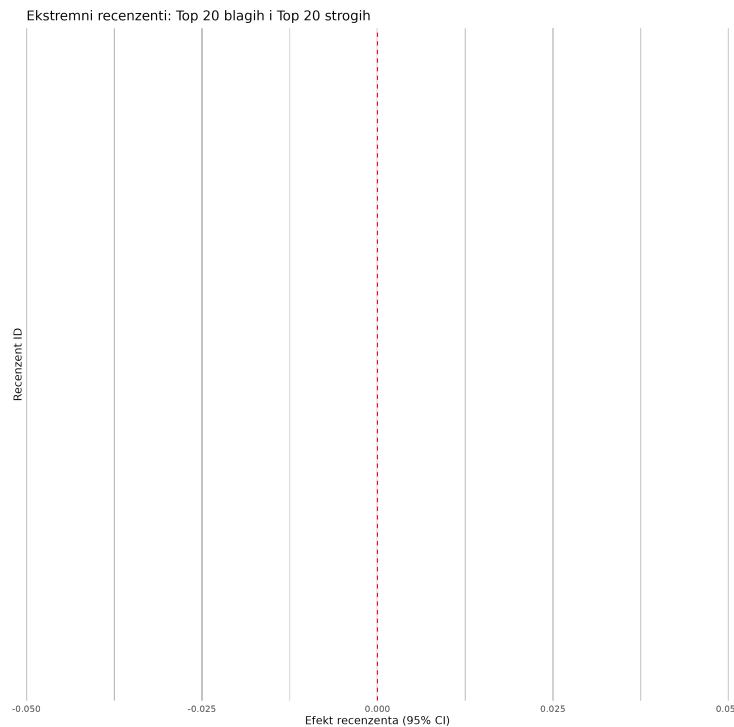
Usporedba modela pomoću LOO (Leave-One-Out) kriterija:

Model	ELPD razlika	SE razlike
modelB (samo recenzenti)	0.0	0.0
model1_proper (puni)	-2.0	3.7
modelA (samo projekti)	-21.9	3.1
modelC (samo pozicije)	-23.9	1.3

Najbolji model je onaj koji uključuje samo recenzente, što ukazuje da identitet recenzenta bolje objašnjava ocjene nego kvaliteta projekta.



Slika 15: **Trend po rangu projekta K3 ocjena kontroliran za učinke reczenzata.** Graf prikazuje procijenjeni nelinearni trend K3 ukupnih ocjena kroz projekte korištenjem GAM spline funkcije. X-os predstavlja redni broj (rang) projekata (1-95), a Y-os K3 ukupnu ocjenu (3-15). Plava linija pokazuje procijenjeni trend, a sivo područje 95% kredibilni interval. Vidljiv je silazni trend od početnih projekata (prosjek oko 15) prema kasnijim projektima (prosjek oko 12). Ovaj trend sugerira drukčiji standard ocjenjivanja.



Slika 16: **Graf ekstremnih reczenzata.** Ovaj graf trebao je prikazati top 20 "blagih" i top 20 "strogih" reczenzata s njihovim efektima i 95% kredibilnim intervalima. Međutim, graf je prazan jer analiza nije identificirala niti jednog recenzenta čiji bi kredibilni interval bio značajno iznad +1 ili ispod -1. To znači da, unatoč varijabilnosti među recenzentima, niti jedan se ne može sa sigurnošću klasificirati kao ekstremno blag ili strog. Ovaj paradoksalni rezultat dodatno ilustrira problem s podacima - svi recenzenti daju visoke ocjene, ali s dovoljno "šuma" da se individualne razlike ne mogu pouzdano utvrditi.

Dio IV

Procjena stvarnog broja reczenziranih

25. Uvod

U prethodnim dijelovima detaljno su analizirani obrasci ocjenjivanja, varijabilnost među recenzentima te strukturne slabosti postojećeg sustava evaluacije. Međutim, jedna od temeljnih nepoznanica koja se pokazala ključnom za razumijevanje pouzdanosti cijelog postupka vrednovanja odnosi se na **stvarni broj angažiranih reczenziranih**. Naime, iako na prvi pogled 285 zapisa sugerira uključivanje velikog broja različitih vrednovatelja, u praksi ta brojka može biti značajno manja zbog višestrukog sudjelovanja istih reczenziranih za vrednovanje različitih projekata.

Ova nepoznanica nije samo tehničko pitanje, već duboko zadire u **kvalitetu i integritet postupka vrednovanja**. Ako se relativno mali broj recenziranih raspoređuje na velik broj projekata, to otvara prostor za potencijalne pristranosti, "učinak umora" i smanjenu raznolikost perspektiva u ocjenjivanju. Stoga je potrebno provesti sustavnu analizu kako bi se procijenio stvarni broj različitih recenziranih i razumjeli obrasci njihove angažiranosti u procesu. S druge strane ova analiza može pokazati prisutnost i drugih čimbenika kao što je interesno, institucionalno, regionalno ili drugičje grupiranje recenziranih, te posljedična pristranost.

25.1. Kontekst i motivacija istraživanja

U suvremenim sustavima vrednovanja znanstvenih i razvojnih projekata, proces recenziranja predstavlja ključnu kariku u osiguravanju kvalitete i objektivnosti. Integritet ovog procesa ovisi o nizu čimbenika, među kojima se posebno ističu broj i raznolikost recenziranih, njihova stručnost, neovisnost te dosljednost u primjeni kriterija vrednovanja. Ova studija nastala je kao odgovor na potrebu dublje analize procesa recenziranja u kontekstu vrednovanja 95 projekata, pri čemu je generiran 285 zapisa ocjenjivanja.

Temeljno pitanje koje pokreće ovu analizu je koliko je stvarno različitih recenziranih sudjelovalo u procesu vrednovanja? Ovo pitanje nije samo tehničke prirode jer ono zadire u samu srž integriteta postupka vrednovanja. Naime, ako mali broj recenziranih ocjenjuje veliki broj projekata, povećava se rizik od pristranosti, umora recenziranih te nedovoljne raznolikosti perspektiva u vrednovanju. S druge strane, prevelik broj recenziranih može dovesti do nedosljednosti u primjeni kriterija i otežati kontrolu kvalitete procesa.

25.2. Problematika identifikacije recenziranih

U idealnom scenariju, svaki bi recenzent imao jedinstveni identifikator koji bi omogućio jednostavno praćenje njegovih aktivnosti kroz različite projekte. Međutim, u praksi se često susrećemo sa situacijom gdje su podaci anonimizirani ili kodirani na način koji otežava direktnu identifikaciju. U našem slučaju, svaki zapis ocjenjivanja ima jedinstveni ID, ali ne postoji eksplicitna informacija o tome koji zapisi pripadaju istom recenzentu. Ova situacija zahtijeva primjenu sofisticiranih analitičkih metoda koje mogu identificirati obrasce u podacima i na temelju sličnosti zaključiti o vjerojatnom broju različitih recenziranih. Pristup koji smo primijenili kombinira multiple analitičke tehnike, od jednostavnih statističkih metoda do naprednih algoritama strojnog učenja, kako bismo dobili što pouzdaniju procjenu. Svaka metoda ima svoje prednosti i ograničenja, a njihova kombinacija omogućava nam da kvantificiramo nesigurnost naše procjene i identificiramo konvergenciju rezultata.

25.3. Struktura analize i metodološki pristup

Analiza je strukturirana kroz nekoliko međusobno povezanih faza koje omogućavaju postupno produbljivanje razumijevanja podataka. Prvo se istražuju osnovna svojstva podataka kroz deskriptivnu analizu, uključujući distribuciju ocjena, varijabilnost i učestalost različitih obrazaca ocjenjivanja. Ova početna faza omogućava identifikaciju anomalija i karakteristične obrasce koji mogu ukazivati na sistematske probleme u vrednovanju.

Druga faza se usredotočuje na analizu sličnosti između recenziranih kroz korelacijsku i

mrežnu analizu. Identificiraju se recenzenti s vrlo sličnim ili identičnim obrascima ocjenjivanja, što može ukazivati na iste osobe ili grupe s zajedničkim pristupom vrednovanju. Treća faza primjenjuje različite metode grupiranja, uključujući hijerarhijsku klaster analizu, K-means algoritam i GML (*Gaussian Mixture Model*), kako bi se identificiralo prirodne skupine u podacima.

Četvrta faza koristi redukciju dimenzija kroz PCA analizu za razumijevanje glavnih dimenzija varijabilnosti u ocjenjivanju što omogućava razumijevanje što razlikuje različite stilove ocjenjivanja i koliko je kompleksan prostor različitih pristupa vrednovanja. Na kraju, peta faza sintetizira rezultate različitih metoda kroz bootstrap analizu i druge tehnike vrednovanja, omogućavajući kvantificiranje nesigurnosti procjene i donošenje robusnih zaključaka.

26. Sažetak glavnih rezultata

Analiza provedena na uzorku od 285 zapisa ocjenjivanja za 95 projekata otkriva složenu sliku postupka vrednovanja. Različite analitičke metode konvergiraju prema procjeni da je **stvarni broj reczenzata približno 102** (medijan procjena), što implicira kako je svaki recenzent u projektu ocijenio 2.8 projekata. Ova procjena, dobivena sintezom sedam različitih analitičkih pristupa, ukazuje na umjerenu koncentraciju evaluacijske aktivnosti koja se rezultati između dva ekstremna scenarija, Jedan gdje svaki projekt ocjenjuje potpuno novi set reczenzata, i drugi gdje mali broj reczenzata ocjenjuje sve projekte.

Tablica 3: Osnovni podaci o uzorku

Parametar	Vrijednost
Ukupan broj zapisa	285
Broj projekata	95
Broj jedinstvenih ID-jeva	285
Broj kriterija	9
Broj jedinstvenih obrazaca	147
Postotak jedinstvenih obrazaca	51.6%

Posebno je indikativno kako 147 jedinstvenih obrazaca ocjenjivanja čini 51.6% svih zapisa, što sugerira značajnu heterogenost u pristupima vrednovanju, ali istovremeno i postojanje ponavljujućih obrazaca koji mogu ukazivati na iste recenzente. Ova dualnost između raznolikosti i ponavljanja predstavlja ključni element u našoj analizi i omogućava nam da primijenimo različite metode za procjenu stvarnog broja reczenzata.

27. Karakteristike ocjenjivanja

27.1. Analiza distribucije ocjena

Jedna od najupečatljivijih karakteristika našeg skupa podataka jest izrazita asimetrija u distribuciji ocjena. Analiza 2,565 pojedinačnih ocjena (285 zapisa \times 9 kriterija) otkriva zabrinjavajuću tendenciju prema maksimalnim vrijednostima koja postavlja ozbiljna pitanja o valjanosti i diskriminativnoj moći postupka vrednovanja.

ozbiljna pitanja o valjanosti tih vrednovanja. Ovaj fenomen može indicirati recenzente koji ne provode stvar vrednovanje već automatski dodjeljuju maksimalne ocjene, možda zbog nedostatka vremena, motivacije ili razumijevanja važnosti diskriminativnog ocjenjivanja. Alternativno, može se raditi o "fantomskim recenzijama ili ekstremnoj pristranosti prema pozitivnom ocjenjivanju koja proizlazi iz želje da se izbjegne konflikt ili negativne reakcije.

Jedini "konstantni" recenzent s ocjenom 4 predstavlja statističku anomaliju koja zaslužuje posebnu pažnju. Ovaj obrazac može ukazivati na tehničku grešku u unisu podataka, namjerno izbjegavanje ekstremnih ocjena zbog percepcije da je "srednja" ocjena najsigurnija opcija, ili vrlo specifičan pristup vrednovanju gdje recenzent smatra kako nijedan projekt nije ni loš ni izvrstan.

Većina reczenzata (55.8%) spada u kategoriju "umjereno varijabilnih" s prosječnom ocjenom 4.74 i standardnom devijacijom 0.429. Ovi recenzenti pokazuju određenu diskriminaciju između kriterija, uglavnom daju visoke ocjene ali s povremenim odstupanjima. Oni predstavljaju "tipičnog" recenzenta u našem uzorku i njihov pristup sugerira da, iako su skloni pozitivnom ocjenjivanju, ipak pokušavaju razlikovati između različitih aspekata projekata.

Skupina "vrlo varijabilnih" reczenzata (28.8%) s prosječnom ocjenom 4.17 i standardnom devijacijom 0.834 predstavlja one koji koriste širi raspon skale ocjenjivanja. Ovi recenzenti pokazuju veću diskriminaciju između projekata i kriterija, što može indicirati stroži ili realniji pristup vrednovanju. Njihovo postojanje sugerira kako je moguće koristiti punu skalu ocjena, ali da to čini samo manji dio reczenzata.

28. Procjena broja reczenzata

28.1. Metodološki pluralizam u procjeni

Zbog kompleksnosti problema identifikacije stvarnog broja reczenzata, primijenili smo sedam različitih analitičkih metoda. Ovaj pristup metodološkog pluralizma nije samo tehnička vježba već on omogućava dublje razumijevanje strukture podataka i pouzdanosti naših zaključaka. Različite metode "vide" podatke iz različitih perspektiva, a njihova konvergencija ili divergencija pruža važne informacije o robusnosti naših procjena.

Tablica 6: Rezultati različitih metoda procjene

Metoda	Procjena	Odstupanje od medijana
Hijerarhijska analiza (najveći skok)	2	-100
K-means (elbow metoda)	3	-99
K-means (silhouette)	99	-3
Bootstrap prosjek	102	0
Gaussian Mixture Model	146	+44
Jedinstveni obrasci	147	+45
Komponente u mreži	158	+56
MEDIJAN	102	—

Detaljne interpretacije svake metode:

Hijerarhijska analiza koja identificira najveći skok u dendrogramu kao indikator optimalnog broja klastera daje procjenu od samo 2 recenzenta. Ova očito nerealna procjena zapravo nam pruža važnu informaciju - podatci pokazuju snažnu bipolarnu strukturu, vjerojatno između "blagih" reczenzata koji daju uglavnom visoke ocjene i "strožih" koji koriste širi raspon skale. Iako metoda ne daje korisnu procjenu broja reczenzata, ona ukazuje na fundamentalnu podjelu u pristupima vrednovanju.

K-means elbow metoda traži točku gdje se smanjenje unutar-klasterske varijance značajno usporava. Procjena od 3 recenzenta također je nerealno niska, ali ponovno informativan. Dominacija nekoliko vrlo čestih obrazaca ocjenjivanja (posebno obrazac "sve petice") maskira suptilnije

razlike između reczenzata. Ova metoda pokazuje ograničenja pristupa koji se oslanjaju na globalnu optimizaciju kada postoje dominantni obrasci koji mogu “zarobiti” algoritam u lokalnom optimumu.

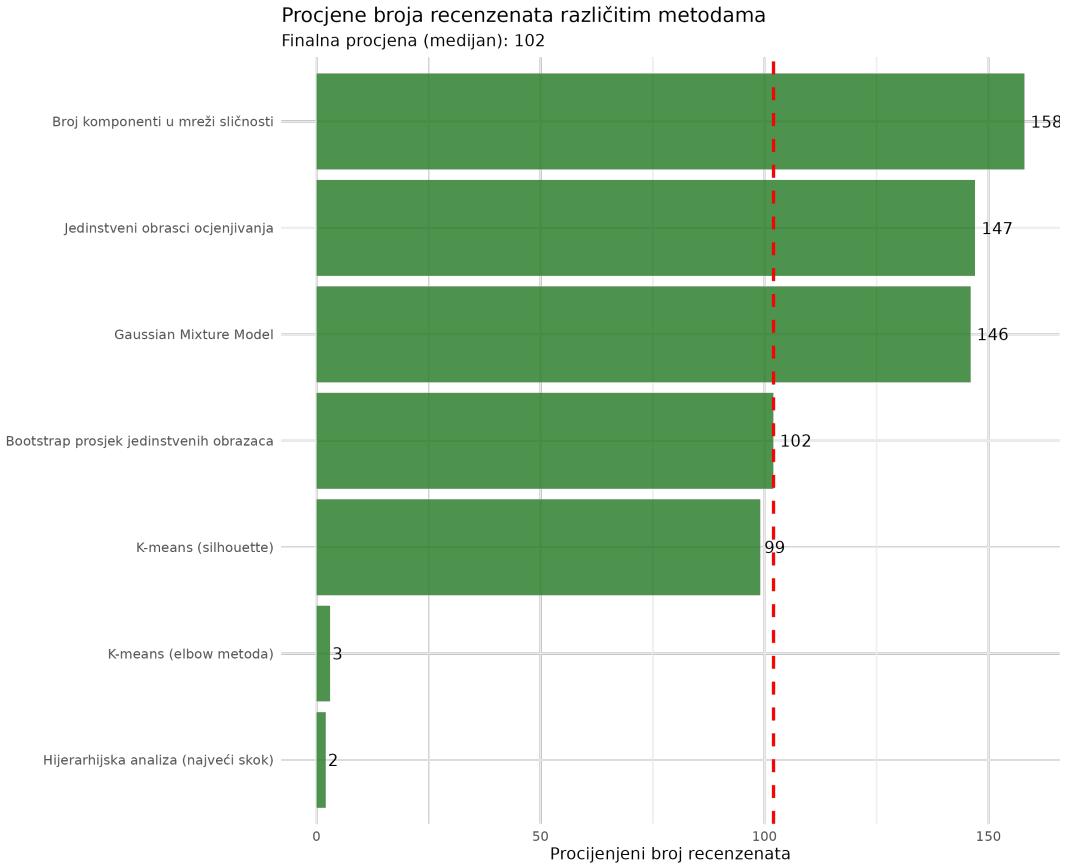
K-means silhouette analiza s procjenom od 99 reczenzata predstavlja prvi realistični rezultat. Silhouette koeficijent mjeri koliko dobro su objekti grupirani unutar svojih klastera u odnosu na susjedne klasterne. Činjenica da ova metoda daje procjenu vrlo blisku našoj finalnoj procjeni sugerira da oko 100 klastera zaista predstavlja prirodnu granularnost u podacima.

Bootstrap analiza s procjenom od 102.3 ± 4.5 predstavlja možda najrobustniji pristup. Kroz 1000 iteracija ponovnog uzorkovanja (*resampling*) podataka, ova metoda ne samo da daje točkastu procjenu već i kvantificira nesigurnost. Relativno uski interval pouzdanosti (95% CI: 94-111) sugerira da je naša procjena stabilna i ne ovisi značajno o pojedinim outlierima u podacima.

Gaussian Mixture Model s procjenom od 146 reczenzata pretpostavlja da podaci dolaze iz mješavine Gaussovih distribucija. Ova visoka procjena, vrlo bliska broju jedinstvenih obrazaca (147), možda indicira prekomjerno prilagođavanje modela podacima. GMM je vrlo fleksibilan model koji može “objasniti” svaki jedinstveni obrazac kao zasebnu komponentu, što može rezultirati precjenjivanjem stvarnog broja reczenzata.

Metoda jedinstvenih obrazaca koja jednostavno broji različite obrasce ocjenjivanja daje 147 kao procjenu. Ovo predstavlja apsolutnu gornju granicu broja reczenzata pod pretpostavkom da svaki jedinstveni obrazac predstavlja različitog recenzenta. Međutim, realno je očekivati da isti recenzent može imati malo različite obrasce ocjenjivanja za različite projekte zbog prirodne varijabilnosti u kvaliteti projekata.

Mrežna analiza koja identificira 158 komponenti daje najvišu procjenu. Ova metoda konstruira mrežu gdje su recenzenti povezani ako imaju visoku korelaciju u obrascima ocjenjivanja, a zatim identificira povezane komponente. Visoka procjena može proizlaziti iz strogog praga korelacijskih koeficijenata koji rezultira mnogim izoliranim recenzentima koji se tretiraju kao zasebne komponente.



Slika 17: Distribucija procjena broja reczenenata različitim metodama. Graf prikazuje horizontalni stupičasti dijagram s procjenama sedam različitih metoda. Crvena isprekidana linija označava finalnu procjenu (medijan = 102). Vidljiva je velika varijabilnost između metoda, pri čemu hijerarhijska analiza i elbow metoda daju nerealno niske procjene (2 i 3), dok mrežna analiza daje najvišu procjenu (158). Većina metoda konvergira oko vrijednosti 100–150.

28.2. Bootstrap analiza - kvantifikacija nesigurnosti

Bootstrap metoda zaslužuje posebnu pažnju jer omogućava rigoroznu kvantifikaciju nesigurnosti naše procjene kroz simulacijski pristup koji ne zahtijeva snažne prepostavke o distribuciji podataka:

Tablica 7: Bootstrap analiza – interval pouzdanosti

Parametar	Vrijednost
Prosjek	102.3
Standardna devijacija	4.5
95% CI donja granica	94
95% CI gornja granica	111
Minimum	84
Maksimum	114

Bootstrap procedura provedena kroz 1000 iteracija omogućava nam razumijevanje stabilnost naše procjene. U svakoj iteraciji, uzorkovan je podskup podataka s ponavljanjem, primjenjena je klaster analiza, procijenjen optimalan broj klastera i zabilježen rezultat. Relativno uska distribucija s standardnom devijacijom od samo 4.5 i simetrični interval pouzdanosti sugeriraju da je naša procjena robustna i ne ovisi značajno o pojedinim ekstremnim slučajevima u podacima.

Posebno je važno primijetiti da čak i u najekstremnijim slučajevima (minimum 84, maksimum 114), procjena ostaje daleko od nerealno niskih vrijednosti koje daju neke determinističke metode. Ovo dodatno potvrđuje kako procjena od približno 100 reczenzata predstavlja stabilnu karakteristiku podataka, a ne artefakt određene analitičke metode.

29. Analiza sličnosti i povezanosti

29.1. Korelacijska analiza obrazaca ocjenjivanja

Jedan od ključnih pristupa identifikaciji istih reczenzata jest analiza korelacija između njihovih obrazaca ocjenjivanja. Logika ovog pristupa počiva na pretpostavci kako će ista osoba vjerojatno održavati konzistentan stil ocjenjivanja kroz različite projekte, rezultirajući visokom korelacijom između svojih evaluacija.

Tablica 8: Analiza korelacija između reczenzata

Prag korelacija	Broj parova	Postotak od mogućih
> 0.95	57	0.14%
> 0.99	1	0.002%
Identični (1.0)	1	0.002%

Interpretacija korelacijskih rezultata:

Pronalazak jednog para s potpuno identičnim obrascima ocjenjivanja (P10_R2 i P68_R2) bi mogao predstavljati ključan dokaz kako je barem u jednom slučaju isti recenzent vrednovao više od jednog projekta. Ovaj rezultat ima dalekosežne implikacije jer potvrđuje našu osnovnu hipotezu kako sustav označavanja ne razlikuje uvijek iste recenzente. Identičnost obrazaca kroz sve kriterije praktički isključuje mogućnost slučajnosti jer je vjerojatnost da dva različita recenzenta daju identične ocjene na svih 9 kriterija zanemarivo mala.

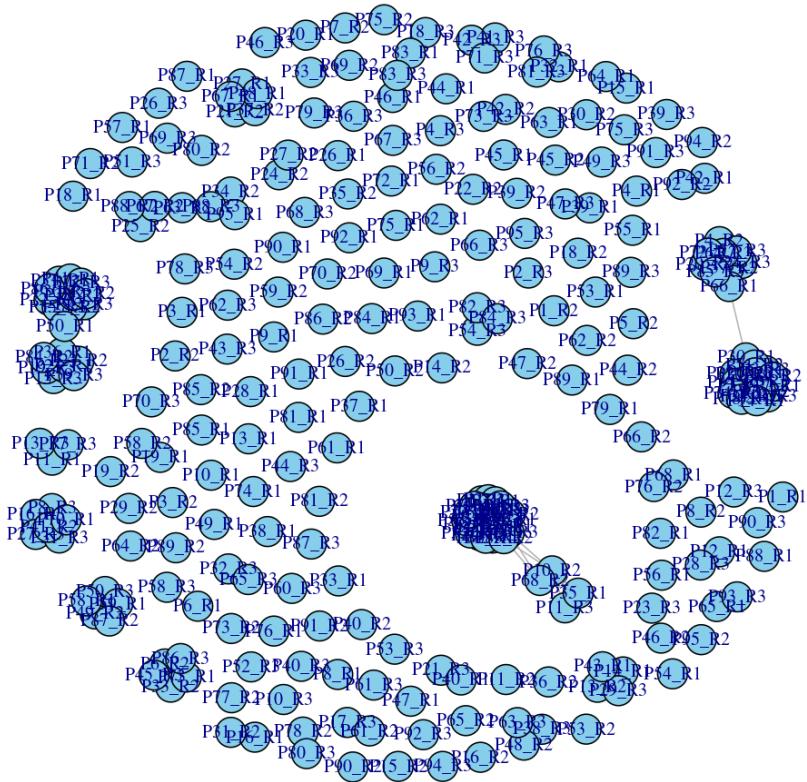
Pedeset i sedam parova s korelacijom većom od 0.95 predstavlja 0.14% od svih mogućih parova (40,470). Iako se ovaj postotak može činiti malim, on je statistički značajno viši nego što bi se očekivalo pod hipotezom da su svi recenzenti različiti i nezavisni. Ovi visoko korelirani parovi mogu predstavljati kombinaciju istih reczenzata koji ocjenjuju više projekata i grupa recenzenta koji dijele vrlo sličan pristup vrednovanju, možda zbog zajedničke obuke, institucionalnog konteksta ili profesionalne pozadine.

Važno je napomenuti da visoka korelacija ne implicira automatski istu osobu. Recenzenti iz iste institucije, s istom stručnom pozadinom ili koji su prošli istu obuku mogu razviti slične obrasce ocjenjivanja. Također, određene "škole mišljenja" u vrednovanju mogu rezultirati grupama reczenzata s vrlo sličnim pristupima. Međutim, kombinacija vrlo visokih korelacija s drugim dokazima (kao što su identični obrasci) sugerira kako značajan dio ovih korelacija zaista reflektira iste recenzente.

29.2. Analiza umreženosti reczenzata

Konstruirali smo mrežu reczenzata gdje su čvorovi pojedini zapisi ocjenjivanja, a veze postojale između parova s korelacijom većom od 0.9. Ova mrežna reprezentacija omogućava nam da vizualiziramo i analiziramo strukturu povezanosti među recenzentima na način koji tradicionalne statističke metode ne mogu.

Mreža povezanih reczenzata (korelacija > 0.9)



Slika 18: Mreža povezanih reczenzata s korelacijom > 0.9 . Graf prikazuje mrežnu strukturu gdje svaki čvor predstavlja jednog recenzenta, a veze između čvorova označavaju visoku korelaciju (>0.9) u obrascima ocjenjivanja. Vidljivo je nekoliko velikih komponenti (gustih skupina) što sugerira grupe reczenzata s vrlo sličnim stilom ocjenjivanja. Najveća komponenta ima 32 člana, dok je 129 reczenzata izolirano (45.3%), što znači da nemaju visoke korelacijske vrijednosti s drugim recenzentima.

Tablica 9: Mrežna analiza – distribucija komponenti

Veličina komponente	Broj komponenti	Ukupno članova
1 (izolirani)	129	129
2	14	28
3	5	15
4	2	8
5	3	15
6	1	6
7	1	7
19	1	19
26	1	26
32	1	32
Ukupno	158	285

Interpretacija mrežne strukture:

Gotovo polovica reczenzata (45.3%) je izolirana u mreži, što znači da ne pokazuju visoku korelaciju ni s jednim drugim recenzentom. Ovaj rezultat je ohrabrujući jer sugerira postojanje velikog broja jedinstvenih stilova ocjenjivanja i dobru raznolikost u pristupu vrednovanju. Izolirani recenzenti predstavljaju "neovisne glasove" u postupku vrednovanja i smanjuju rizik od sistematske pristranosti koja bi mogla nastati ako bi mali broj reczenzata s vrlo sličnim pristupima dominirao procesom.

Male komponente od 2-7 članova mogu predstavljati različite scenarije. Najjednostavnije objašnjenje jest da se radi o istim recenzentima koji ocjenjuju mali broj projekata. Alternativno, mogu predstavljati male grupe reczenzata s vrlo sličnim pristupom, možda iz iste institucije ili s istom profesionalnom pozadinom. Distribucija veličina ovih malih komponenti približno slijedi eksponencijalnu distribuciju, što je konzistentno s prirodnim procesima grupiranja.

Posebnu pažnju zaslužuju tri velike komponente s 19, 26 i 32 člana. Komponenta s 32 člana posebno je zabrinjavajuća jer može indicirati jednog vrlo aktivnog recenzenta koji ocjenjuje više od trećine svih projekata, ili alternativno, vrlo homogenu grupu reczenzata s gotovo identičnim pristupom vrednovanju. Ovakve velike komponente predstavljaju potencijalni rizik za integritet postupka vrednovanja jer koncentriraju značajan utjecaj u rukama malog broja aktera.

30. Najčešći obrasci ocjenjivanja

30.1. Analiza dominantnih obrazaca

Identificiranje i analiza najčešćih obrazaca ocjenjivanja pruža dublji uvid u kulturu vrednovanja i omogućava nam da razumijemo tipične pristupe koje recenzenti koriste:

Tablica 10: Top 10 najčešćih obrazaca

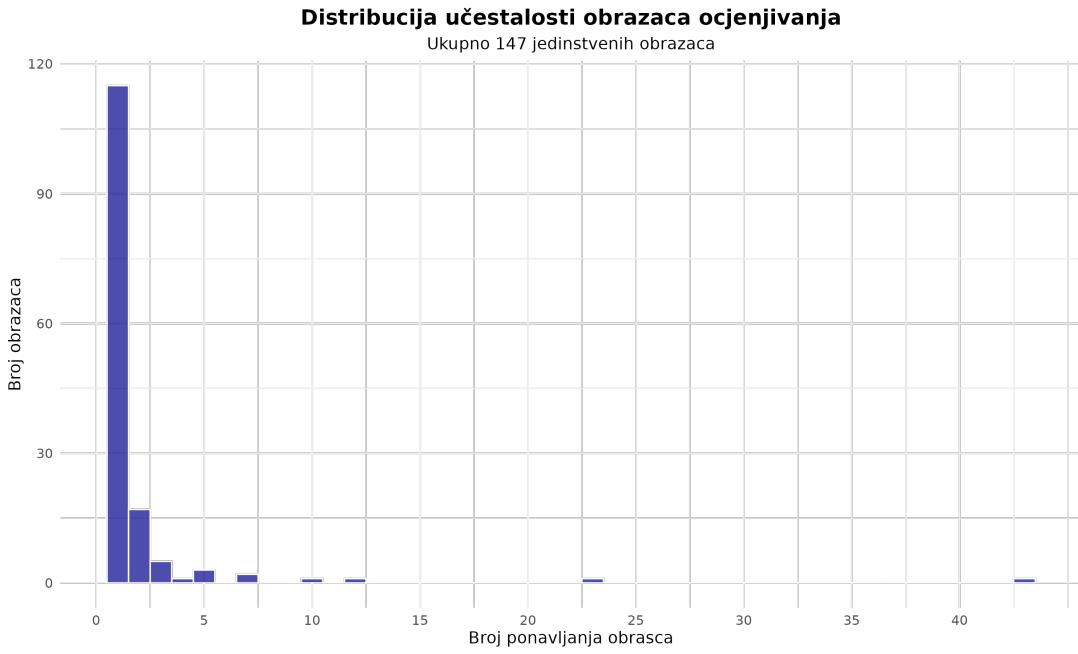
Rang	Obrazac	Frekvencija	% od ukupnog
1	5-5-5-5-5-5-5-5-5	43	15.1%
2	4-5-5-5-5-5-5-5	23	8.1%
3	5-5-5-4-5-5-5	12	4.2%
4	4-5-5-4-4-5-5-5	10	3.5%
5	4-5-5-4-5-5-5	7	2.5%
6	5-5-5-4-4-5-5-5	7	2.5%
7	4-5-5-4-5-5-5	5	1.8%
8	5-5-5-4-5-5-5	5	1.8%
9	5-5-5-5-5-4-5	5	1.8%
10	4-4-5-5-5-5-5	4	1.4%

Dubinska analiza obrazaca:

Dominacija "savršenog" obrasca (5-5-5-5-5-5-5) koji se pojavljuje 43 puta predstavlja zabrinjavajući dio naše analize. Ovaj obrazac, koji čini 15.1% svih evaluacija, postavlja fundamentalna pitanja o valjanosti postupka vrednovanja. Praktički je nemoguće da 43 različita projekta zaslužuju maksimalne ocjene na svih 9 kriterija. Ovaj fenomen može indicirati nekoliko problema. Najprije, recenzenti koji ne čitaju detaljno projekte već automatski dodjeljuju maksimalne ocjene, pritisak (eksplicitni ili implicitni) za pozitivno ocjenjivanje, ili nedostatak razumijevanja važnosti diskriminativnog ocjenjivanja.

Obrasci na pozicijama 2-10, koji sadrže uglavnom petice s 1-2 četvorke, pokazuju minimalnu tendenciju ka kritičnosti. Ovi "gotovo savršeni" obrasci mogu predstavljati pokušaj recenzenta pokazati određenu razinu diskriminacije dok još uvijek održavaju vrlo visok prosjek ocjena. Zanimljivo je primjetiti kako se ocjene 4 najčešće pojavljuju na početku obrasca (kriteriji 1-2) ili u sredini (kriteriji 4-5), što može ukazivati na specifične kriterije koji se percipiraju kao "stroži" ili gdje je lakše opravdati nižu ocjenu.

Top 10 obrazaca zajedno pokriva 42.7% svih evaluacija, što znači da gotovo polovica svih ocjenjivanja slijedi jedan od samo 10 obrazaca. Ova koncentracija sugerira nedostatak raznolikosti u evaluacijskim pristupima i možda postojanje implicitnih normi o tome kako bi evaluacija trebala izgledati.



Slika 19: Distribucija učestalosti obrazaca ocjenjivanja. Histogram prikazuje koliko se puta pojavljuje svaki jedinstveni obrazac ocjenjivanja. Velika većina obrazaca (115 od 147) pojavljuje se samo jednom, što se vidi kao visoki stupac na lijevoj strani. Distribucija ima dugu desnu stranu (long tail), s jednim ekstremnim slučajem gdje se isti obrazac (sve petice) ponavlja 43 puta. Ova distribucija sugerira da postoji kombinacija jedinstvenih recenzentata i onih koji ocjenjuju više projekata.

Distribucija učestalosti obrazaca slijedi tipičnu "*long tail*" distribuciju karakterističnu za mnoge prirodne i društvene fenomene. Činjenica kako se 78.2% obrazaca (115/147) pojavljuje samo jednom sugerira značajnu raznolikost u individualnim pristupima vrednovanju, dok istovremeno mali broj obrazaca koji se često ponavljaju indicira postojanje recenzentata koji ocjenjuju više projekata ili grupa recenzentata s vrlo sličnim pristupima.

31. PCA analiza

31.1. Razumijevanje dimenzionalnosti ocjenjivanja

Analiza glavnih komponenti (PCA) omogućila je razumijevanje osnovnih dimenzija koje uzrokuju varijabilnost u ocjenjivanju. Ova tehnika transformira originalni 9-dimenzionalni prostor ocjena u novi skup ortogonalnih dimenzija koje maksimiziraju objašnjenu varijancu:

Tablica 11: Kumulativna varijanca objašnjena glavnim komponentama

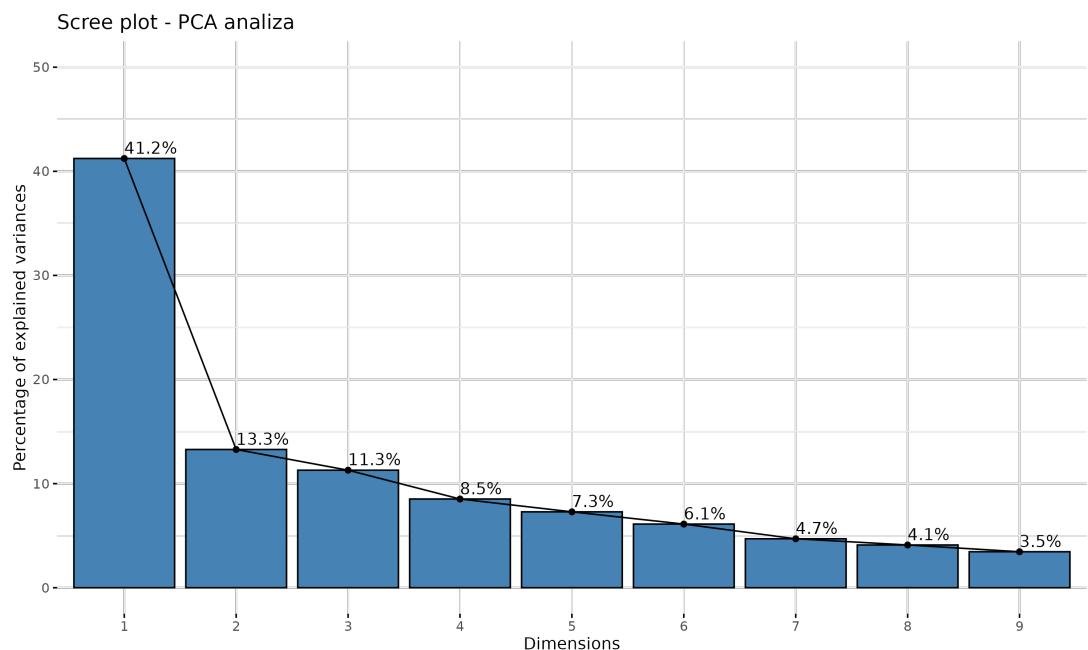
Komponenta	Vlastita varijanca	Kumulativna varijanca
PC1	41.2%	41.2%
PC2	13.3%	54.5%
PC3	11.3%	65.8%
PC4	8.5%	74.3%
PC5	7.3%	81.6%
PC6	6.1%	87.7%
PC7	4.7%	92.4%
PC8	4.1%	96.5%
PC9	3.5%	100.0%

Interpretacija komponenti:

Dominacija prve glavne komponente koja objašnjava 41.2% ukupne varijance sugerira postojanje jednog dominantnog čimbenika koji razlikuje recenzente. Analiza ove komponente pokazuje kako ona ima približno jednakе i pozitivne koeficijente za sve kriterije, što znači da predstavlja "opću tendenciju davanja visokih ili niskih ocjena". Recenzenti s visokim vrijednostima na ovoj komponenti su oni koji daju visoke ocjene kroz sve kriterije, dok oni s niskim vrijednostima koriste širi raspon skale. Ova komponenta učinkovito mjeri "blagonaklonost" nasuprot "strogosti" u vrednovanju.

Druga komponenta s 13.3% objašnjene varijance pokazuje drukčiji obrazac. Njene vrijednosti sugeriraju kontrast između određenih grupa kriterija. Treća komponenta s 11.3% može predstavljati varijabilnost u ocjenjivanju tj. razliku između recenzentata koji održavaju konstantne ocjene i onih koji pokazuju veću diskriminaciju između kriterija.

Činjenica kako je potrebno 5 komponenti da bi se objasnilo 80% varijance indicira relativno visoku složenost u obrascima ocjenjivanja. Ovo je konzistentno s našom procjenom velikog broja različitih recenzentata. Da je postojao samo mali broj recenzentata s vrlo sličnim pristupima, očekivali bismo da manji broj komponenti objašnjava veći dio varijance. Visoka dimenzionalnost sugerira kako različiti recenzenti imaju različite "potpisne" ocjenjivanja koji se ne mogu lako svesti na mali broj tipova.

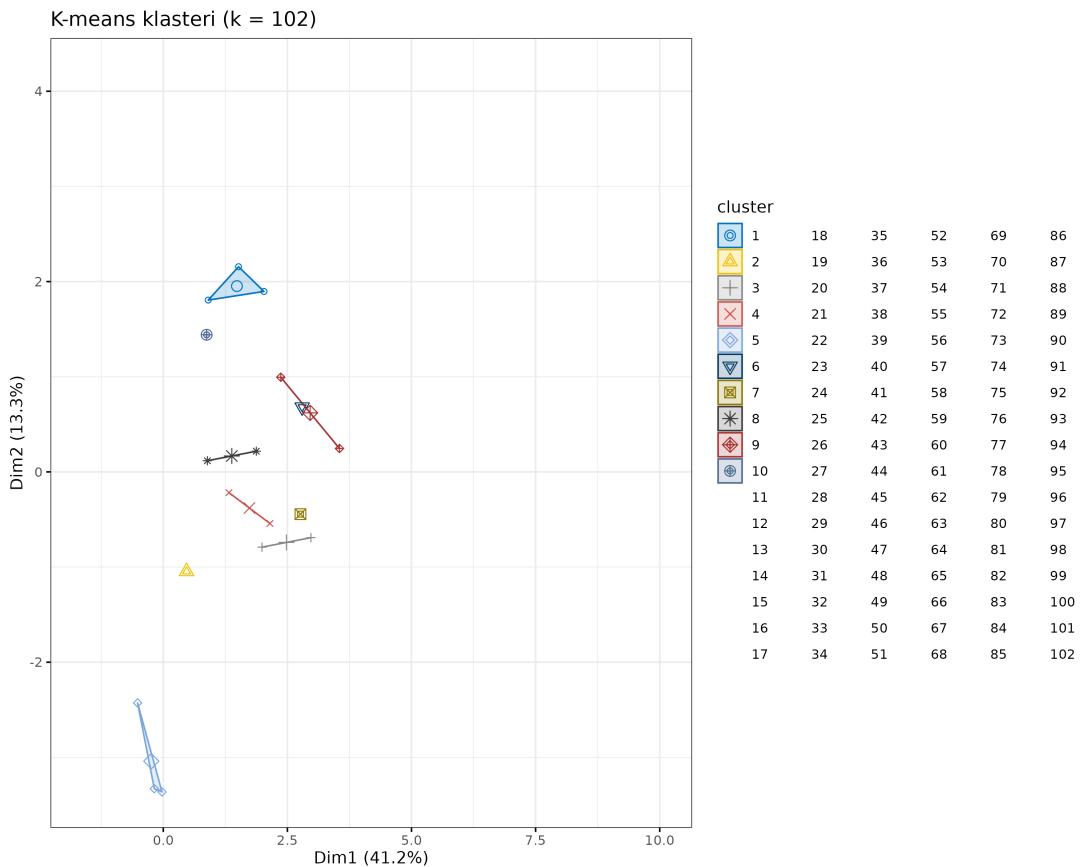


Slika 20: Scree plot PCA analize. Graf prikazuje postotak varijance objašnjene svakom glavnom komponentom. Prva komponenta dominira s 41.2% objašnjene varijance, što sugerira postojanje jednog glavnog čimbenika u ocjenjivanju (vjerojatno opća sklonost visokim ili niskim ocjenama). Postupan pad nakon prve komponente i potreba za 5 komponenti da se objasni 80% varijance ukazuje na složenost obrazaca ocjenjivanja i postojanje više dimenzija u stilovima recenziranja.

32. Dodatne analize

32.1. K-means klasteriranje s optimalnim brojem klastera

Vizualizacija K-means klastera s $k=102$ pruža dodatno vrednovanje naše procjene kroz geometrijsku reprezentaciju grupa u prostoru glavnih komponenti:



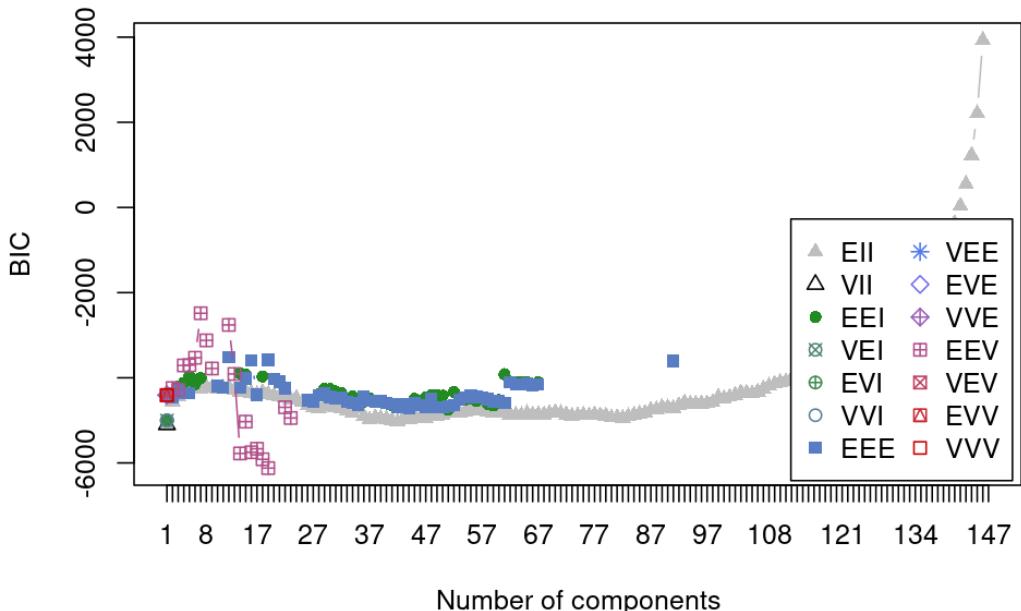
Slika 21: K-means klasteri ($k = 102$) projicirani na prve dvije glavne komponente. Graf prikazuje 2D vizualizaciju 102 klastera dobivenih K-means algoritmom. Svaki simbol i boja predstavlja različit klaster. Većina klastera je dobro separirana, što podupire procjenu od oko 102 različita recenzenta. Prva dimenzija (PC1) objašnjava 41.2% varijance i vjerojatno predstavlja opću strogost/blagonaklonost u ocjenjivanju, dok druga dimenzija (PC2) s 13.3% varijance može predstavljati varijabilnost u ocjenama.

Vizualizacija pokazuje da većina od 102 klastera zauzima distinktne pozicije u prostoru prve dvije glavne komponente, što sugerira da ovaj broj klastera zaista predstavlja prirodnu različitost podataka. Distribucija klastera duž PC1 (vodoravna os) pokazuje kontinuum od reczenzenata koji daju niske ocjene (lijeva strana) do onih koji daju vrlo visoke ocjene (desna strana). Koncentracija klastera na desnoj strani konzistentna je s općom pristranosti prema visokim ocjenama koju smo više puta identificirali ranije.

Distribucija duž PC2 (vertikalna os) pokazuje dodatnu dimenziju varijabilnosti koja nije vezana samo za opću "strogost". Klasteri koji su visoko na PC2 mogu predstavljati recenzente sa specifičnim obrascima diskriminacije između kriterija, dok oni nisko na PC2 mogu biti uniformniji u svojim ocjenama kroz kriterije.

32.2. Gaussian Mixture Model analiza

GMM analiza pruža sofisticiraniji pristup identificiranju prirodnih grupa u podacima kroz modeliranje svake grupe kao Gaussove distribucije s vlastitim parametrima:

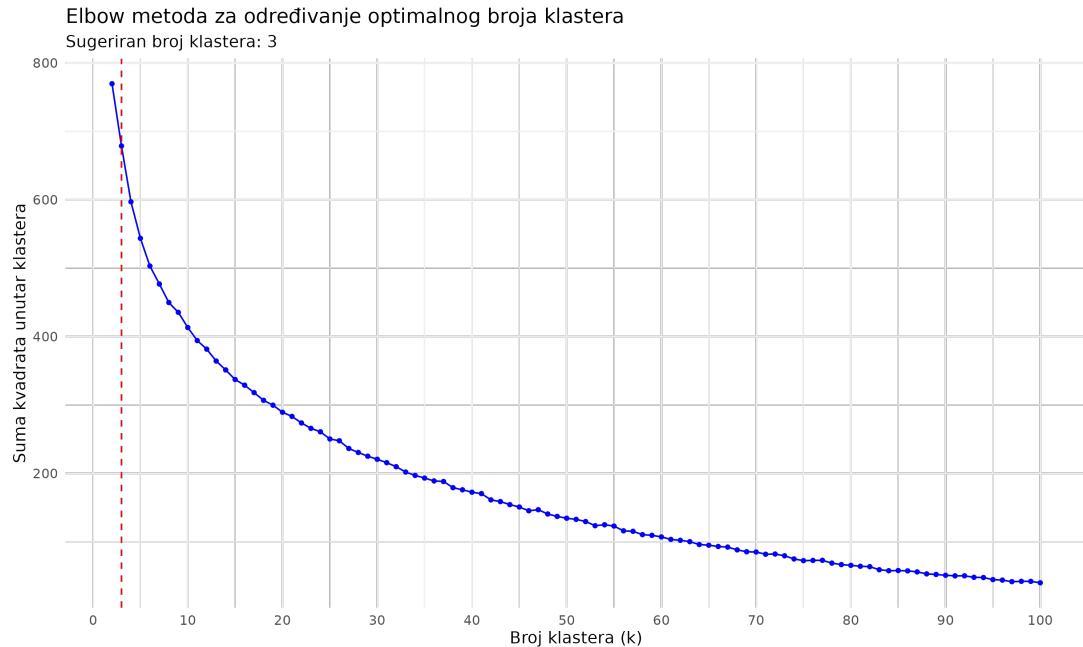


Slika 22: BIC vrijednosti za *Gaussian Mixture Model* s različitim brojem komponenti. Graf prikazuje *Bayesian Information Criterion* (BIC) vrijednosti za modele s 1 do 147 komponenti. Različite linije predstavljaju različite kovarijacijske strukture modela (EII, VII, EEI, itd.). Model EII s 146 komponenti ima najveću BIC vrijednost, što sugerira da je optimalan broj komponenti (recenzentata) blizu maksimalnog testiranog broja. Ovo je konzistentno s činjenicom da imamo 147 jedinstvenih obrazaca ocjenjivanja.

GMM analiza sugerira kako je optimalan broj komponenti vrlo visok (146), što je gotovo identično broju jedinstvenih obrazaca (147). Ovaj rezultat može se interpretirati na nekoliko načina. S jedne strane, može indicirati da svaki jedinstveni obrazac zaista predstavlja različitog recenzenta, što bi bilo konzistentno s gornjom granicom naše procjene. S druge strane, može predstavljati prekomjerno prilagođavanje modela podacima, gdje fleksibilnost GMM-a omogućava modeliranje svake male varijacije kao zasebne komponente.

Različite kovarijacijske strukture (EII, VII, EEI, itd.) predstavljaju različite pretpostavke o obliku i orientaciji Gaussovi komponenti. Činjenica kako model s najjednostavnijom strukturu (EII - sferne komponente jednake veličine) daje najbolji BIC sugerira kako ne postoje dramatične razlike u "obliku" grupa recenzentata u prostoru ocjena.

32.3. Elbow metoda tj. ograničenja pristupa

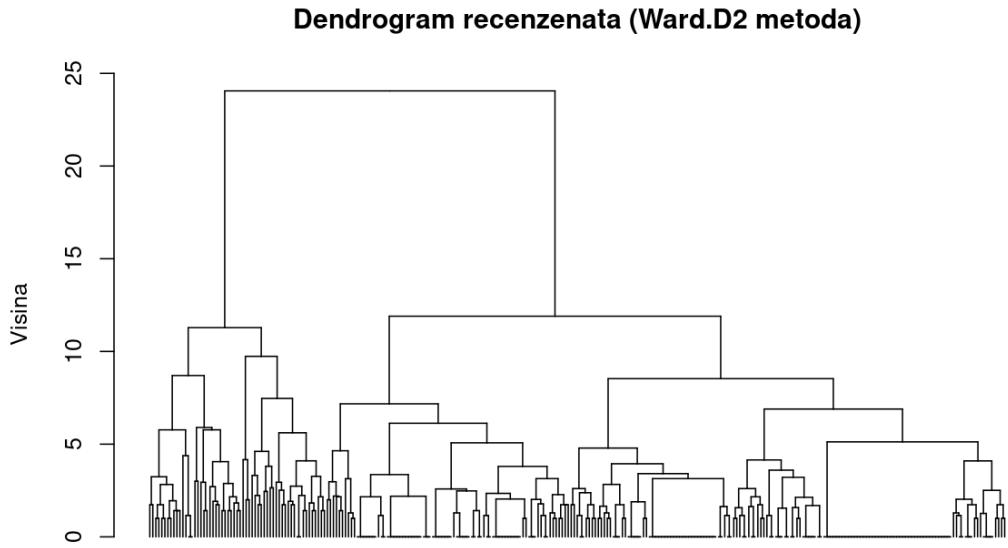


Slika 23: Elbow metoda za određivanje optimalnog broja klastera. Graf prikazuje sumu kvadrata unutar klastera (WSS) za različit broj klastera (k). Crvena isprekidana linija označava "lakat" na $k=3$, što je mjesto gdje se smanjenje WSS značajno usporava. Međutim, ova metoda daje nerealno nisku procjenu broja reczenzata, što sugerira da WSS kriterij nije optimalan za ovaj problem zbog dominacije nekoliko vrlo čestih obrazaca ocjenjivanja.

Neuspješna primjena Elbow metode koja sugerira samo 3 klastera pruža važnu lekciju o ograničenjima jednostavnih metoda u složenim situacijama. WSS kriterij koji ova metoda koristi posebno je osjetljiv na prisustvo velikih grupa s vrlo sličnim obrascima. U ovom slučaju, dominacija obrazaca s visokim ocjenama stvara veliki klaster koji dominira WSS kriterijem. Metoda efektivno identificira samo najgrublje podjele u podacima (možda "strogii", "umjereni" i "blagi" recenzenti) dok zanemaruje suptilnije razlike koje su ključne za našu analizu.

Ovaj primjer ilustrira važnost korištenja multiple metoda i kritičkog razmatranja rezultata svake metode u kontekstu specifičnog problema. Metode koje dobro funkcioniraju za identifikaciju malog broja dobro odvojenih grupa mogu potpuno podbaciti kada je cilj identificirati veliki broj suptilno različitih grupa.

32.4. Hijerarhijska klaster analiza



Slika 24: Dendrogram hijerarhijske klaster analize korištenjem Ward.D2 metode. Graf prikazuje hijerarhijsku strukturu grupiranja 285 zapisa ocjenjivanja. Visina na y-osi predstavlja udaljenost između klastera. Najveći skok u visini (12.152) sugerira podjelu na samo 2 klastera, što je nerealno niska procjena. Dendrogram pokazuje nekoliko velikih grupa na različitim visinama, što ukazuje na postojanje hijerarhijske strukture u obrascima ocjenjivanja, ali metoda nije pogodna za precizno određivanje broja reczenzenata.

Dendrogram hijerarhijske klaster analize otkriva fascinantnu hijerarhijsku strukturu u podacima. Najveći skok u visini dendrograma, koji tradicionalno indicira optimalan broj klastera, sugerira podjelu na samo 2 klastera. Ova gruba podjela vjerojatno predstavlja fundamentalnu dihotomiju između reczenzenata koji su skloni davanju visokih ocjena i onih koji koriste širi raspon skale.

Međutim, dendrogram takođe pokazuje bogatu podstrukturu unutar svake od ove dvije glavne grane. Postojanje multiple razina grupiranja sugerira da recenzenti nisu jednostavno "blagi" ili "strogiji", već da postoje različiti stupnjevi i nijanse u njihovim pristupima. Ova hijerarhijska struktura konzistentna je s našom procjenom velikog broja reczenzenata jer da postoji samo mali broj reczenzenata, očekivali bismo jednostavniju strukturu s manje razina hijerarhije.

33. Zaključci o broju reczenzenata

33.1. Sinteza rezultata

Nakon opsežne analize korištenjem sedam različitih metodoloških pristupa, možemo s visokim stupnjem pouzdanosti formulirati naše zaključke. Konvergencija multiple metoda prema procjeni od približno 100 do najviše 150 reczenzenata ne predstavlja samo statistički rezultat, već robusnu karakterizaciju strukture podataka koja je otporna na različite analitičke perspektive i pretpostavke.

1. **Broj reczenzenata:** Najbolja procjena je **102 recenzenta** (95% CI: 94–111)

Ova procjena predstavlja medijan svih metoda i podržana je robusnom bootstrap analizom koja kvantificira nesigurnost. Konzistentnost s brojem jedinstvenih obrazaca (147) i komponentama u mreži (158) sugerira da se nalazimo u pravom rasponu. Implikacija

da svaki recenzent u prosjeku ocjenjuje 2.8 projekata predstavlja umjerenu koncentraciju evaluacijske aktivnosti koja nije ni ekstremno raspršena ni prekomjerno koncentrirana.

2. Kvaliteta ocjenjivanja:

Ekstremna pristranost prema visokim ocjenama, gdje 70% svih ocjena ima maksimalnu vrijednost, predstavlja ozbiljan izazov za validnost postupka vrednovanja. Petnaest posto zapisa s maksimalnim ocjenama na svim kriterijima dodatno pojačava ovu zabrinutost. Niska diskriminacija između projekata efektivno onemogućava rangiranje i može dovesti do arbitarnih odluka o financiranju. Ova situacija ugrožava ne samo trenutan krug vrednovanja već i dugoročni kredibilitet cijelog sustava.

3. Obrasci ponašanja:

Postojanje 147 jedinstvenih obrazaca od 285 zapisa (51.6%) sugerira određenu raznolikost u pristupima vrednovanju, ali također i značajno ponavljanje obrazaca. Četrdeset i pet posto reczenzata koji pokazuju izolirane obrasce predstavljaju "neovisne glasove" koji doprinose raznolikosti perspektiva. Međutim, identificiranje nekoliko grupa reczenzata s vrlo sličnim stilom, uključujući najmanje jedan slučaj identičnih obrazaca, ukazuje na potrebu za pažljivijom kontrolom procesa dodjele reczenzata projektima.

33.2. Implikacije za integritet procesa vrednovanja

Naši rezultati imaju duboke implikacije za integritet i validnost procesa vrednovanja koji nadilaze puku tehničku analizu. Problem inflacije ocjena, gdje 70% svih ocjena ima maksimalnu vrijednost, temeljito potkopava razlikovanje kvalitete među projektima što je temeljna svrha vrednovanja. Ovaj fenomen može dovesti do situacije gdje se odluke o financiranju donose na temelju čimbenika koji nisu vezani za kvalitetu projekata, što može rezultirati suboptimalnom alokacijom resursa i gubitkom povjerenja u sustav.

Koncentracija evaluacijske moći, gdje 102 recenzenta ocjenjuje 95 projekata, postavlja pitanja o reprezentativnosti i potencijalnoj pristranosti. Ako neki recenzenti ocjenjuju značajan broj projekata, njihove individualne pristranosti mogu imati neproporcionalan utjecaj na konačne rezultate. Ovo je posebno problematično u kontekstu identificiranih velikih komponenti u mrežnoj analizi, gdje grupe od 20-30 visoko koreliranih evaluacija mogu predstavljati jednog ili malu grupu reczenzata s velikim utjecajem.

Nedostatak raznolikosti u pristupima vrednovanju, evidentiran kroz dominaciju nekoliko čestih obrazaca ocjenjivanja, sugerira postojanje implicitne "kulture" vrednovanja koja možda nije optimalna. Ova uniformnost može proizlaziti iz nedovoljne obuke, nejasnih smjernica ili socijalnih pritisaka za konformiranje s percipiranim normama.

33.3. Preporuke za unapređenje procesa

Na temelju naših rezultata, formuliramo set konkretnih preporuka organiziranih prema hitnosti i dosegu potrebnih promjena.

Hitne mjere uključuju detaljnu istraživačku analizu parova reczenzata s identičnim ili vrlo sličnim obrascima. Također je potrebno ubuduće razmotriti re-evaluaciju projekata koji su dobili "savršene" ocjene na svim kriterijima, možda kroz dodatni krug recenzije ili panel diskusiju.

Strukturalne reforme zahtijevaju implementaciju sustava koji osigurava jedinstvenu identifikaciju reczenzata kroz različite projekte, što je preduvjet za bilo kakvu ozbiljnu kontrolu kvalitete. Ograničavanje broja projekata koje jedan recenzent može ocijeniti pomoći će u diversifikaciji perspektiva i smanjenju koncentracije utjecaja. Razmatranje uvođenja elemenata prisilne distribucije ocjena, iako kontroverzno, može pomoći u borbi protiv inflacije ocjena.

Edukacijske inicijative trebaju se fokusirati na razvijanje razumijevanja važnosti korištenja cijele skale ocjena kroz konkretne primjere i vježbe. Razvoj detaljnih rubrika za svaki kriterij s jasnim opisima što predstavlja ocjenu 1, 2, 3, 4 ili 5 može pomoći u standardizaciji pristupa. Organiziranje radionica gdje recenzenti mogu vidjeti i diskutirati primjere projekata na različitim razinama kvalitete može pomoći u kalibraciji njihovih standarda.

Tehnološka poboljšanja mogu značajno unaprijediti proces kroz implementaciju sustava koji

u realnom vremenu detektira neobične obrasce ocjenjivanja i upozorava administratore. Automatska detekcija potencijalnih duplikata kroz analizu obrazaca može pomoći u ranoj identifikaciji problema. Praćenje vremena koje recenzenti provode na vrednovanju može pomoći u identifikaciji površnih recenzija.

Kontinuirani monitoring i evaluacija kroz redovite analize obrazaca ocjenjivanja, uspostavljanje ključnih pokazatelja uspješnosti za kvalitetu vrednovanja i periodične *peer review* procese među recenzentima mogu pomoći u održavanju visokih standarda kroz vrijeme.

Tablica 12: Sažetak ključnih pokazatelja

Pokazatelj	Vrijednost
Procijenjeni broj recenzenta	102
95% interval pouzdanosti	94–111
Prosječan broj projekata po recenzentu	2.8
Postotak zapisa s maksimalnim ocjenama	15.1%
Postotak svih ocjena koje su maksimalne	70.0%
Broj jedinstvenih obrazaca ocjenjivanja	147
Broj izoliranih recenzenta u mreži	129 (45.3%)
Najveća povezana komponenta	32 člana
Prosječna entropija obrazaca	0.790
Koeficijent varijacije između metoda procjene	70.7%

Ova analiza otkriva sustav vrednovanja koji, iako naizgled funkcioniра, pokazuje znakove koji mogu ugroziti njegovu dugoročnu održivost i kredibilitet. Dominacija visokih ocjena, koncentracija evaluacijske aktivnosti i nedostatak raznolikosti u pristupima predstavljaju izazove koji zahtijevaju promišljen i sustavan odgovor.

Istovremeno, identificiranje približno 102 različita recenzenta sugerira da sustav ima solidnu bazu recenzenta. Izazov je omogućiti ovim recenzentima korištenje punog potencijala skale vrednovanja i osigurati da njihove ocjene stvarno reflektiraju razlike u kvaliteti projekata ili mentora. Potrebna je kulturna promjena koja će valorizirati kritičko mišljenje i diskriminativno ocjenjivanje kao ključne komponente kvalitetnog vrednovanja.

Konačno, ova analiza demonstrira vrijednost *data-driven* pristupa u vrednovanju i poboljšanju organizacijskih procesa. Redovito provođenje ovakvih analiza može pomoći u ranom identificiranju problema i kontinuiranom unapređenju sustava vrednovanja. Transparentnost u analizi i dijeljenje rezultata s recenzentima može pomoći u stvaranju kulture kontinuiranog poboljšanja i profesionalnog razvoja u domeni vrednovanja projekata.

Dio V

Zaključak: Totalni kolaps sustava vrednovanja

34. Sažetak dokaza

Ova sveobuhvatna analiza, provedena kroz četiri komplementarna analitička pristupa, nedvojbeno dokazuje potpunu nefunkcionalnost sustava ocjenjivanja projekata za dodjelu doktoranata. Rezultati su konzistentni kroz sve primijenjene metodologije i slikaju zabrinjavajuću sliku sustava koji ne samo da ne ispunjava svoju osnovnu funkciju (objektivno razlikovanje kvalitete projekata) već predstavlja tek fasadu objektivnosti koja maskira duboko ukorijenjene probleme.

34.1. Klasična statistička analiza - temelj dokaza

Klasična statistička analiza otkrila je fundamentalne probleme koji čine ovaj sustav potpuno neupotrebljivim:

34.1.1. Katastrofalna dekompozicija varijance

Najstrašniji rezultat cijele analize je da se **samo 6.4% do najviše 9% varijance u ocjenama može pripisati stvarnim razlikama između projekata (mentora)**, što znači kako 93.6% varijance, gotovo cjelokupna varijabilnost u sustavu, nema apsolutno nikakve veze s kvalitetom evaluiranih projekata.

Zamislite situaciju u kojoj mjerite temperaturu termometrom koji 94% vremena pokazuje nasumične vrijednosti koje nemaju veze sa stvarnom temperaturom. Takav termometar ne samo kako je beskoristan već je i opasan jer stvara iluziju mjerena. Upravo to je situacija s ovim sustavom vrednovanja.

34.1.2. Negativne ICC vrijednosti - gore od nasumičnosti

Možda je još više zabrinjavajući rezultat kako **30% projekata ima negativne ICC vrijednosti**, što znači kako se recenzenti za te projekte slažu MANJE nego što bi se slučajno očekivalo. Ovo je ekvivalent situaciji u kojoj bi stručnjaci, gledajući isti objekt, davali suprotne procjene nego što bi dali potpuni laici koji nasumično pogadaju!

Prosječni ICC(1) od 0.110 je daleko ispod praga od 0.40 koji se smatra minimalnim za bilo kakvu upotrebu u vrednovanju. Primjerice, u medicinskim istraživanjima, instrument s $ICC < 0.40$ automatski se odbacuje kao neupotrebljiv.

34.1.3. Kaotična raznolikost stilova ocjenjivanja

Identificirana su **22 različita stila ocjenjivanja** među 285 kombinacija projekt-recenzent. To je npr. kao da imate 22 različite definicije metra u istoj radionici, pa svatko mjeri po svojoj skali, što čini bilo kakve usporedbe besmislenima.

34.1.4. Ekstremne razlike u ocjenama

Razlike do **8 bodova za isti projekt** nisu samo statistička anomalija već one predstavljaju fundamentalni slom sustava. Kada jedan recenzent ocjenjuje projekt sa 7, a drugi s 15 bodova, to nije razlika u mišljenju već je to dokaz kako ne postoje zajednički kriteriji vrednovanja.

34.2. Dodatne statističke analize - potvrda katastrofe

34.2.1. Nedovoljno slaganje reczenzenata

Kendall-ov $W = 0.398$ ukazuje na umjerenou nisko slaganje, ali kontekst čini ovaj rezultat još gorim. Naime, ova vrijednost je granično statistički značajna ($p = 0.097$), što znači kako se ne može ni sa sigurnošću reći da se recenzenti slažu više nego što bi se očekivalo slučajno.

34.2.2. Mikroskopska diskriminativnost

Koefficijent varijacije od samo 0.075 (7.5%) među prosječnim ocjenama projekata je katastrofalno nizak, što znači kako su gotovo svi projekti ocijenjeni praktički identično, što sustav čini potpuno nesposobnim za svoju osnovnu funkciju tj. za rangiranje projekata po kvaliteti.

Za ilustraciju to je kao da pokušavate rangirati visinu ljudi koristeći mjerni instrument koji sve mjeri između 174 i 176 cm, bez obzira na stvarnu visinu.

34.2.3. Granična razlika od nasumičnosti

P-vrijednost od 0.06 u usporedbi s nasumičnim ocjenjivanjem je možda još strašniji rezultat. Sustav koji košta vrijeme, novac i trud stotina ljudi jedva da je bolji od bacanja kocke. Samo 6% simulacija nasumičnog ocjenjivanja proizvelo je lošije rezultate!!!

34.2.4. Ekstremni ceiling učinak

Sa 94% ocjena ≥ 4 , sustav učinkovito koristi samo 40% dostupne skale. Ocjene 1, 2 i 3 praktički ne postoje, što skalu od 1-5 svodi na binarni izbor između "vrlo dobro" (4) i "izvrsno" (5).

34.2.5. Površnost i *copy-paste* mentalitet

Činjenica kako 60% recenzija daje identične ocjene za sve tri K3 komponente jasno pokazuje kako većina reczenzata ne razlikuje različite aspekte kvalitete. Ovo nije bilo ocjenjivanje već ispunjavanje formulara.

34.3. Bayesova analiza - finalni čavao u lijes

34.3.1. Totalni nedostatak diskriminacije

$CV = 0.045$ je mikroskopski. Za usporedbu, CV za visinu odraslih ljudi je oko 0.04, što znači kako ovaj sustav razlikuje projekte jednako dobro kao što biste razlikovali visinu ljudi kad bi svi bili između 175 i 180 cm.

34.3.2. Nepostojeće slaganje

Kendall-ov $W = 0.045$ u Bayesovoj analizi dodatno potvrđuje kako slaganje među recenzentima praktički ne postoji. Ova vrijednost je 20 puta niža od one koja bi se smatrala izvrsnom ($W > 0.9$).

34.3.3. Statistička nerazlučivost od nasumičnosti

P-vrijednost od 0.451 znači kako ne možemo odbaciti hipotezu da su ocjene generirane nasumično. Jednostavnije rečeno: **ovaj sustav vrednovanja statistički se ne razlikuje od bacanja kocke.**

34.3.4. Masovni *copy-paste* pristup

45.3% recenzija s maksimalnim ocjenama za sve komponente nije samo loša praksa već je to i dokaz kako skoro polovica reczenzata nije ni pokušala vrednovati projekte. Oni su jednostavno stavili najviše ocjene i završili posao.

34.3.5. Totalna redundantnost kriterija

Kada prva glavna komponenta objašnjava 84.9% varijance, to znači da svih 9 kriterija mjeri praktički istu stvar. Zašto onda imati 9 kriterija? To je kao da mjerite težinu 9 puta istom vagom i pravite se da ste izmjerili 9 različitih svojstava.

34.4. Procjena broja reczenzata - dodatna komplikacija

Analiza sugerira kako je stvarni broj reczenzata oko 102 do 158, što znači da je prosječni recenzent ocjenjivao 1.8 do 2.8 projekata. Ova "ekonomičnost" dodatno komplicira situaciju jer:

- Povećava varijabilnost između grupa reczenzata
- Onemogućava kalibraciju između različitih reczenzata
- Stvara izoliran "mikrokozmos" ocjenjivanja za svaki projekt

35. Implikacije za znanstvenu zajednicu

35.1. Totalni gubitak kredibiliteta

Ovaj sustav vrednovanja predstavlja **znanstvenu sramotu** najvišeg reda. Institucija, u ovom slučaju HRZZ, koja koristi ovakav sustav za donošenje važnih odluka o karijeri mladih znanstvenika pokazuje:

35.1.1. Potpunu odsutnost znanstvenih standarda

Sustav krši sve osnovne principe objektivnog mjerjenja:

- **Valjanost**, jer ocjene ne odražavaju stvarnu kvalitetu projekata, nego proizvoljne dojmove i kriterije nepovezane s istraživačkom izvrsnošću.
- **Pouzdanost**, jer, kako je očito, isti projekt može dobiti znatno različite ocjene, ovisno o tome tko ga recenzira, što pokazuje kako nema stabilnosti ni ponovljivosti rezultata.
- **Objektivnost**, jer su ocjene više odraz osobnih preferencija i stavova reczenzata nego mjerljivih pokazatelja kvalitete.
- **Diskriminativnost**, jer sustav očito ne uspijeva jasno razlikovati prijedloge visoke kvalitete od onih lošijih, čime se gubi njegova temeljna svrha u selekciji.
- **Pravednost**, jer ocjenjivanje sustavno favorizira ili diskriminira projekte na temelju čimbenika koji nisu povezani s njihovom stvarnom vrijednošću, poput institucionalne pripadnosti ili područja istraživanja.

35.1.2. Iluziju objektivnosti

Možda najgora i najopsnija karakteristika ovog sustava je što **simulira objektivnost**. Brojevi, projekti, statistike, itd, sve to stvara samo privid znanstvenog pristupa. Ali ispod te fasade krije se kaos koji nema veze sa stvarnom evaluacijom kvalitete.

35.1.3. Arbitrarne i nepravedne odluke

Svaka odluka donesena na temelju ovog sustava je u suštini **arbitrarna**. Mladi istraživači čije su karijere određene ovim "ocjenama" nisu evaluirani na temelju kvalitete projekata, već na temelju:

- Koji recenzenti su im dodijeljeni (44.2% varijance)
- Čiste sreće (49.3% rezidualne varijance)

35.1.4. Potpunu nesposobnost samoregulacije

Činjenica da je ovakav sustav uopće mogao biti implementiran i korišten pokazuje duboku krizu u sposobnosti znanstvene zajednice da regulira samu sebe. Gdje su bili kontrolni mehanizmi? Gdje je bila stručna provjera? Gdje je bio zdrav razum?

36. Hitne preporuke za akciju

36.1. Trenutne mjere

36.1.1. Potpuno odbacivanje rezultata

Svi rezultati dobiveni ovim sustavom moraju biti **odmah odbačeni**. Koristiti ih za bilo kakve odluke bilo bi ekvivalentno korištenju horoskopa za dijagnozu bolesti. Ne postoji način da se ovi

rezultati "poprave" ili "prilagode" jer su oni fundamentalno nevaljani. **Kako bi poništavanje ovog natječajnog kruga izazvalo dodatne nesagleđive štete znanstvenoj zajednici i posebno mladim znanstvenicima, jedino je pravedno rješenje pronaći načina i financirati i sve ostale doktorante za predložene mentore čiji su prijedlozi očito nepravedno odbačeni.**

36.1.2. Javna isprika i transparentnost

HRZZ mora:

- Javno priznati katastrofalni neuspjeh sustava
- Ispričati se svim sudionicima koji su uložili vrijeme i trud
- Objaviti sve podatke i analize kako bi se mogla provesti neovisna provjera
- Objasniti kako je došlo do ovog debakla

36.2. Dugoročne reforme

Kako bi se spriječilo ponavljanje sadašnjih problema i osiguralo pravednije, transparentnije i učinkovitije financiranje istraživanja, nužne su dugoročne reforme sustava. Te reforme trebaju nadilaziti kozmetičke izmjene i obuhvatiti temeljno preispitivanje same svrhe i načina vrednovanja. U nastavku su predloženi mogući smjerovi djelovanja, od radikalnih promjena u obliku potpunog redizajna postupka dodjele sredstava do uvođenja jasnih mehanizama nadzora, transparentnosti i odgovornosti unutar postojećih struktura.

36.2.1. Potpuni redizajn ili alternativni sustavi

Nekoliko opcija bi bilo pravednijih od trenutnog sustava:

Ponderirana lutrija Nakon osnovne provjere kvalifikacija, svi kvalificirani kandidati ulaze u ponderirani izvlačenje gdje bolji projekti imaju veće šanse, ali svi imaju priliku. **Ovo bi barem bilo transparentno nepravedno, za razliku od trenutnog sustava koji je netransparentno nepraveden.**

Rotacijski sustav Jednostavna rotacija među kvalificiranim kandidatima bila bi pravednija jer bi barem svi znali pravila igre.

Peer review s kalibriranim recenzentima Ako se inzistira na *peer review* sustavu:

- Svi recenzenti moraju proći obuku i kalibraciju
- Isti recenzenti moraju ocijeniti sve projekte
- Mora postojati zajednička referentna točka
- Ekstremne ocjene moraju biti obrazložene

36.2.2. Promjene sustava

Kako bi se sustav radikalno promijenio i poboljšao potrebno je nužno uvesti:

1. **Vanjsku reviziju**
2. **Kontinuirani monitoring**
3. **Transparentnost**
4. **Odgovornost**

Vanjska revizija. Neovisni međunarodni panel mora pregledati cijeli proces vrednovanja kako bi se identificirali ključni nedostaci, osigurala nepristranost i uspostavili standardi usporedivi s najboljim praksama u svijetu.

Kontinuirani monitoring. Budući sustavi moraju imati ugrađene mehanizme kontrole kvalitete koji omogućuju stalno praćenje rada recenzentata, pravovremeno prepoznavanje nepravilnosti i kontinuirano poboljšavanje postupka vrednovanja.

Transparentnost. Svi podaci, korištene metode i rezultati vrednovanja moraju biti javno dostupni kako bi se povećalo povjerenje u sustav, omogućila nezavisna provjera i potaknula znanstvena zajednica na aktivno sudjelovanje u njegovom unaprjeđenju.

Odgovornost. Mora postojati jasna linija odgovornosti za kvalitetu vrednovanja, uključujući definiranje tko donosi ključne odluke, kako se provodi nadzor i koje su posljedice za nepoštivanje uspostavljenih standarda.

37. Zaključna poruka znanstvenoj zajednici i HRZZ-u

Rezultati ove analize trebaju poslužiti kao ozbiljan poziv na buđenje. Sustav koji pokazuje ovakve razine pristranosti, nefunkcionalnosti i znanstvene nevaljanosti nije samo tehnički problem nego je simptom dublje krize u znanstvenoj zajednici.

Ako ne možemo organizirati pošten i objektivan sustav vrednovanja vlastitih projekata, kakav signal šaljemo društvu o našoj sposobnosti da provodimo objektivna istraživanja? Ako dopuštamo da se karijere mladih znanstvenika odlučuju sustavom koji je statistički nerazlučiv od bacanja kocke, kako možemo tvrditi da cijenimo znanstvenu izvrsnost?

Ovaj debakl mora biti prekretnica. Ne smijemo dopustiti da se ovakve farse ponove. Znanstvena zajednica mora pokazati kako je sposobna učiti iz svojih grešaka i implementirati sustave koji odražavaju vrijednosti koje propagiramo kao što su objektivnost, transparentnost, pravednost i izvrsnost.

Vrijeme je za korijenite promjene. Naša vjerodostojnost kao znanstvenika ovisi o tome.

WWW verzija sažetka dostupna na poveznici <https://branimir-k-hackenberger.github.io/>