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 evaluacijskog sustava

Prilog prigovoru

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

4. kolovoza 2025.

Sažetak

Ovaj dokument predstavlja konsolidiranu analizu sustava ocjenjivanja projekata za dodjelu doktoranata koja nedvojbeno dokazuje potpunu nefunkcionalnost, pristranost i znanstvenu nevaljanost primijenjenog evaluacijskog postupka. Kroz tri komplementarna analitička pristupa, klasičnu statističku analizu, napredne statističke metode i Bayesovu analizu, dokazali smo kako sustav ne samo da ne mjeri kvalitetu projekata, već predstavlja tek simulaciju objektivnosti koja dovodi do arbitrarnih 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 nasumičnog ocjenjivanja ($p = 0.451$), što ga čini potpuno beskorisnim za namijenjenu svrhu.

Sadržaj

1 Uvod **5**
1.1 Kontekst i motivacija . . . 5
1.2 Svrha istraživanja . . . 5
1.3 Opseg analize . . . 5
1.4 Materijal i metode . . . 5
1.4.1 Opis podatkovnog skupa . . . 5
1.4.2 Analitički pristup . . . 5
1.4.3 Softver i računalne platforme . . . 7
1.4.4 Etički aspekti . . . 7
1.5 Ključni nalazi - pregled . . . 7

I Klasična statistička analiza **8**

2 Struktura podataka i identifikacija recenzenata **8**
2.1 Metodologija . . . 8
2.2 Rezultati . . . 8

3 Analiza pouzdanosti među recenzentima (Inter-rater Reliability) **10**

3.1 Metodologija . . . 10
3.2 Rezultati . . . 10

4 Analiza stilova recenziranja **12**
4.1 Metodologija . . . 12
4.2 Rezultati . . . 12

5 Mixed Effects Model analiza **13**
5.1 Metodologija . . . 13
5.2 Rezultati . . . 13
5.3 Slučajni učinci projekata . . . 14

6 Analiza slaganja recenzenata po rangu projekta **15**
6.1 Metodologija . . . 15
6.2 Rezultati . . . 15

7 Identifikacija problematičnih slučajeva **16**
7.1 Metodologija . . . 16
7.2 Rezultati . . . 16

8 Prediktivna analiza **17**
8.1 Metodologija . . . 17
8.2 Rezultati . . . 17

9 Analiza konzistentnosti kroz projekte **17**
9.1 Metodologija . . . 17
9.2 Rezultati . . . 18

10 Fokus na K3 kriterij **18**

II Dodatne statističke analize **19**

11 Analiza slaganja recenzenata - Kendall-ov W **19**
11.1 Opis metode . . . 19
11.2 Rezultati . . . 19

12 Analiza diskriminativnosti ocjena **20**
12.1 Opis metode . . . 20
12.1.1 Koeficijent varijacije (CV) . . . 20
12.1.2 Bayesov čimbenik za ANOVA . . . 20
12.1.3 Relativna entropija . . . 20
12.2 Rezultati . . . 20

13 Usporedba s nasumičnim ocjenjivanjem **21**
13.1 Opis metode . . . 21
13.2 Rezultati . . . 21

14 Analiza ceiling učinka **22**
14.1 Opis metode . . . 22
14.1.1 Deskriptivnu statistiku . . . 22
14.1.2 Bayesov čimbenik protiv uniformne distribucije . . . 22
14.2 Rezultati . . . 22

15 Analiza informativnosti kriterija **23**
15.1 Opis metode . . . 23
15.1.1 Normalizirana mutual information . . . 23
15.1.2 Analiza glavnih komponenti (PCA) . . . 23
15.2 Rezultati . . . 24

16 Analiza ekstremnih obrazaca ocjenjivanja **24**
16.1 Opis metode . . . 24
16.2 Rezultati . . . 24

III Bayesova analiza **25**

17 Struktura podataka i osnovni pokazatelji **25**
17.1 Struktura uzorka . . . 25
17.2 Distribucija ocjena . . . 25

18 Bayesovi hijerarhijski modeli **25**
18.1 Opis metode . . . 25
18.2 Problemi s konvergencijom . . . 25
18.3 Dekompozicija varijance . . . 25

19 Analiza pristranosti po pozicijama **26**
19.1 Bayesov čimbenik . . . 26

20 Analiza slaganja recenzenata **26**
20.1 Kendall-ov W test . . . 26

21 Analiza diskriminativnosti **27**
21.1 Koeficijent varijacije . . . 27
21.2 Relativna entropija . . . 27

22 Usporedba s nasumičnim ocjenjivanjem **27**

23 Analiza ceiling učinka **27**

24 Analiza informativnosti kriterija **27**
24.1 Analiza glavnih komponenti (PCA) . . . 27
24.2 Normalizirana mutual information . . . 27

25 Identifikacija ekstremnih obrazaca **27**

26 Selekcija modela **28**

IV Procjena stvarnog broja recenzenata **31**

27 Sažetak glavnih nalaza **31**

28 Karakteristike ocjenjivanja **31**

29 Procjena broja recenzenata **32**

30 Analiza sličnosti i povezanosti **33**

31 Najčešći obrasci ocjenjivanja **35**

32 PCA analiza **36**

33 Dodatne analize **38**

34 Zaključci o broju recenzenata **41**
34.1 Glavne polaznice za zaključivanje 41

V Zaključak: Totalni kolaps sustava vrednovanja **42**

35 Sažetak dokaza **42**
35.1 Klasična statistička analiza - temelj dokaza 42
35.1.1 Katastrofalna dekompozicija varijance 42
35.1.2 Negativne ICC vrijednosti - gore od nasumičnosti 42
35.1.3 Kaotična raznolikost stilova ocjenjivanja 42
35.1.4 Ekstremne razlike u ocjenama 42
35.2 Dodatne statističke analize - potvrda katastrofe 42
35.2.1 Nedovoljno slaganje recenzenata 42
35.2.2 Mikroskopska diskriminativnost 43
35.2.3 Granična razlika od nasumičnosti 43
35.2.4 Ekstremni ceiling učinak 43
35.2.5 Površnost i *copy-paste* mentalitet 43
35.3 Bayesova analiza - finalni čavao u lijes 43
35.3.1 Totalni nedostatak diskriminacije 43
35.3.2 Nepostojeće slaganje 43
35.3.3 Statistička nerazlučivost od nasumičnosti 43
35.3.4 Masovni *copy-paste* pristup 43
35.3.5 Totalna redundantnost kriterija 43
35.4 Procjena broja recenzenata - dodatna komplikacija 44

36 Implikacije za znanstvenu zajednicu **44**
36.1 Totalni gubitak kredibiliteta 44
36.1.1 Potpunu odsutnost znanstvenih standarda 44
36.1.2 Iluziju objektivnosti 44
36.1.3 Arbitrarne i nepravedne odluke 44
36.1.4 Potpunu nesposobnost samoregulacije 45

37 Hitne preporuke za akciju **45**
37.1 Trenutne mjere 45
37.1.1 Potpuno odbacivanje rezultata 45
37.1.2 Javna isprika i transparentnost 45
37.2 Dugoročne reforme 45
37.2.1 Potpuni redizajn ili alternativni sustavi 45
37.2.2 Promjene sustava 45

38 Zaključna poruka znanstvenoj zajednici i HRZZ-u **46**

1 Uvod

1.1 Kontekst i motivacija

Sustav vrednovanja znanstvenih projekata trebao bi predstavljati temelj pravedne i objektivne raspodjele resursa u akademskoj zajednici. U kontekstu dodjele doktoranata, takav sustav ima dalekosežne posljedice na karijere mladih istraživača i budućnost znanstvenih istraživanja. Stoga je ključno da takav sustav bude utemeljen na čvrstim znanstvenim principima, objektivnosti i transparentnosti.

1.2 Svrha istraživanja

Svrha ovog istraživanja bila je provesti rigoroznu znanstvenu analizu provedenog postupka vrednovanja kako bi se utvrdilo zadovoljava li osnovne standarde objektivnosti i valjanosti. Nažalost, naši rezultati pokazuju **potpunu besmislenost i neobjektivnost provedenog vrednovanja**, te **totalni izostanak bilo kakvog znanstveno utemeljenog sustava vrednovanja**, što je za znanstvenu zajednicu ne samo neprihvatljivo, već i duboko sramotno.

1.3 Opseg analize

Analiza je obuhvatila:

- 95 prijdloga projekata prijavljnih u znanstvenom području "Prirodne znanosti" ocijenjenih u postupku vrednovanja
- 285 jedinstvenih kombinacija projekt-recenzent
- Tri recenzenta po projektu (nepoznate unificiranosti)
- Fokus na tzv. K3 kriterij koji je bio istaknut kao posebno važan

1.4 Materijal i metode

1.4.1 Opis podatkovnog skupa

Analiza je provedena na kompletnom skupu podataka evaluacije projekata prijavljenih za natječaj "Projekt razvoja karijera mladih istraživača – Izobrazba novih doktora znanosti (DOK-2025-02)" **u području prirodnih znanosti**. Podaci su prikupljeni iz službenog evaluacijskog postupka i obuhvaćaju 95 projekata prijavljenih na natječaj s ukupno 285 jedinstvenih kombinacija projekt-recenzent, što predstavlja ukupan broj individualnih recenzija. Svaki projekt ocijenila su tri recenzenta označena kao R1, R2 i R3, pri čemu svaki projekt ima jedinstveni set recenzenata bez preklapanja između projekata.

Evaluacijski sustav temelji se na devet kriterija ocjenjivanja organiziranih u tri glavne kategorije: K1 s komponentama K1_1, K1_2 i K1_3, K2 s komponentama K2_1, K2_2 i K2_3, te K3 s komponentama K3_1, K3_2 i K3_3. Za svaku komponentu korištena je standardna skala ocjenjivanja od 1 do 5. Poseban fokus analize bio je na K3 kriteriju koji je u evaluacijskom postupku identificiran kao posebno važan za procjenu kvalitete projekata.

Za precizno praćenje i analizu kreiran je jedinstveni identifikator za svaku kombinaciju projekt-recenzent u formatu "PX_RY", kao što je "P1_R1". Ovaj pristup je omogućio jednoznačnu identifikaciju svake recenzije, detaljnu analizu preklapanja recenzenata između projekata te praćenje obrazaca ocjenjivanja pojedinih recenzenata kroz različite projekte.

1.4.2 Analitički pristup

Primijenjena je sveobuhvatna metodologija koja kombinira četiri komplementarna analitička pristupa, pri čemu je svaki dizajniran kako bi osvijetlio različite aspekte funkcionalnosti ili nefunkcionalnosti sustava ocjenjivanja.

A. Klasična statistička analiza Analiza pouzdanosti među recenzentima provedena je korištenjem koefivcijenta ICC (*Intraclass Correlation Coefficient*) u tri varijante. ICC(1) koji mjeri apsolutno slaganje pojedinačnih recenzenata i računa se kao omjer varijance između projekata i ukupne varijance, ICC(2) koji procjenjuje pouzdanost prosječne ocjene svih recenzenata, te ICC(3) koji mjeri konzistentnost recenzenata neovisno o apsolutnim razlikama. Za interpreta-

ciju su korišteni standardni pragovi gdje ICC manji od 0.40 označava slabo slaganje, vrijednosti između 0.40 i 0.59 umjereno slaganje, između 0.60 i 0.74 dobro slaganje, a iznad 0.75 izvrsno slaganje.

Za dekompoziciju izvora varijabilnosti primijenjen je LLM (*Linear Mixed Effects Model*) koji omogućava razdvajanje varijance na komponente povezane s projektima, recenzentima i rezidualnu varijancu. Model uključuje opći prosjek, nasumične učinke projekata koji slijede normalnu distribuciju sa srednjom vrijednošću nula, nasumične učinke grupa recenzenata također normalno distribuirane, fiksne efekte pozicija recenzenata te rezidualnu grešku. Model selection proveden je pomoću Likelihood Ratio testova između hijerarhijski ugniježđenih modela.

Hijerarhijska klaster analiza je korištena za identifikaciju različitih stilova recenziranja. Proces je uključivao stvaranje profila recenzenata temeljem njihovih prosječnih ocjena po svim kriterijima, izračun matrice korelacijskih udaljenosti, primjenu Ward.D2 metode vezanja te optimizaciju broja klastera pomoću Silhouette koeficijenta i Gap statistike.

Temporalna analiza (u smislu reda projekata od najviše ocjenjenog prema najniže ocjenjnenom) provedena je korištenjem Spearmanovog koeficijenta korelacije rang-redoslijeda za analizu trendova prosječnih ocjena i slaganja recenzenata kroz vrijeme tj. niz skalu rangova. Slaganje je definirano kao jedan minus koeficijent varijacije, što omogućava praćenje promjena u konzistentnosti ocjenjivanja.

B. Napredne statističke metode Kendall-ov koeficijent konkordancije W korišten je za kvantificiranje stupnja slaganja među recenzentima pri rangiranju projekata. Ovaj koeficijent računa se na temelju sume kvadrata odstupanja rangova i normalizira se prema broju recenzenata i projekata, dajući vrijednosti između 0 (za potpuno neslaganje) i 1 (za savršeno slaganje).

Diskriminativnost sustava analizirana je kroz tri komplementarne mjere. Koeficijent varijacije mjeri relativnu varijabilnost prosječnih ocjena projekata u odnosu na opći prosjek. Bayesov čimbenik za ANOVA testira evidenciju za postojanje stvarnih razlika među projektima naspram nul-hipoteze da su svi projekti jednaki. Relativna entropija kvantificira količinu informacije u distribuciji ocjena u odnosu na maksimalnu moguću entropiju.

Monte Carlo simulacija s 1000 iteracija korištena je za usporedbu opaženih rezultata s očekivanjima pod pretpostavkom nasumičnog ocjenjivanja. U svakoj iteraciji generirane su nasumične ocjene zadržavajući empirijsku distribuciju originalnih ocjena, izračunat je Kendall-ov W za simulirane podatke, te je kreirana nul-distribucija za procjenu statističke značajnosti opaženih rezultata.

Analiza ceiling efekta provedena je kroz izračun proporcije maksimalnih ocjena, koeficijenta asimetrije distribucije te Bayesovog čimbenika koji testira odstupanje od uniformne distribucije. Ove mjere omogućavaju kvantificiranje stupnja u kojem je skala ocjenjivanja efektivno sužena zbog prevelike koncentracije visokih ocjena.

C. Bayesova analiza Bayesov hijerarhijski model implementiran je za procjenu komponenti varijance s potpunom kvantifikacijom nesigurnosti. Model specificira ocjenu kao sumu općeg prosjeka, nasumičnog učinka projekta, nasumičnog učinka recenzenta i rezidualne greške. Prior distribucije specificirane su kao slabo informativne normalne distribucije za fiksne efekte i Half-Cauchy distribucije za standardne devijacije nasumičnih učinaka.

MCMC uzorkovanje provedeno je korištenjem Hamiltonian Monte Carlo algoritma s četiri lance, svaki s 2000 iteracija od kojih je prvih 1000 korišteno za warm-up. Konvergencija je dijagnosticirana pomoću Rhat statistike, Effective Sample Size, E-BFMI te praćenjem divergentnih tranzicija.

Model selection proveden je korištenjem Leave-One-Out cross-validation pristupa koji omogućava usporedbu prediktivnih performansi različitih modela. Testirani su modeli koji uključuju samo projekte, samo recenzente, samo pozicije te puni model sa svim čimbenicima.

D. Analiza obrazaca ocjenjivanja Identifikacija jedinstvenih obrazaca provedena je kreiranjem "potpisa" za svaki set ocjena, što predstavlja jedinstvenu kombinaciju ocjena kroz sve kriterije. Analizirana je frekvencija ponavljanja svakog potpisa te su identificirani dominantni

obrasci koji se pojavljuju kod više recenzenata.

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 recenzenata na temelju analize jedinstvenih obrazaca ocjenjivanja. U svakoj iteraciji kreiran je bootstrap uzorak, identificirani su jedinstveni obrasci te je procijenjen broj različitih recenzenata. Rezultati su sumirani kroz 95% confidence intervale.

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 zaštitila privatnost sudionika. Identiteti projekata i recenzenata zamijenjeni su numeričkim kodovima koji ne omogućavaju identifikaciju. Analiza je provedena isključivo s ciljem znanstvenog istraživanja i unapređenja sustava evaluacije. Rezultati su prezentirani na agregiranoj razini koja ne omogućava identifikaciju pojedinačnih sudionika ili njihovih projekata. Svi nalazi interpretirani su u kontekstu sistemskih problema, a ne individualnih performansi.

1.5 Ključni nalazi - pregled

Već u uvodu važno je istaknuti najšokantniji nalaz: **sustav evaluacije 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 recenzenata i nasumičnih čimbenika.

Dio I
Klasična statistička analiza

2 Struktura podataka i identifikacija recenzenata

2.1 Metodologija

Za identifikaciju jedinstvenih recenzenata kreiran je kompozitni identifikator koji kombinira oznaku projekta i oznaku recenzenta (npr. "P1_R1"). Ovaj pristup omogućava precizno praćenje koji recenzent je ocjenjivao koji projekt te analizu preklapanja recenzenata između projekata.

2.2 Rezultati

Analiza strukture podataka otkriva sljedeće ključne nalaze t epolazi od pretpostavki:

- **95 projekata** je ocjenjivano u postupku
- **285 jedinstvenih kombinacija projekt-recenzent** je identificirano
- Svaki projekt ima točno 3 recenzenta
- **Recenzenti NISU isti kroz sve projekte** - svaki projekt ima svoj jedinstveni set od tri recenzenta

Slika 1: Sažetak analize pristranosti u ocjenjivanju - finalni dashboard. 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 recenzenata kroz vrijeme (crvena LOESS krivulja), što sugerira umor recenzenata ili promjenu standarda tijekom procesa evaluacije.

Ova struktura podataka predstavlja **fundamentalni dizajnerski problem**: različiti recenzenti ocjenjuju različite projekte bez ikakve zajedničke referentne točke. To znači da su projekti ocjenjivani u potpuno različitim "mikrokozmosima" ocjenjivanja, što onemogućava fer usporedbu između projekata.

3 Analiza pouzdanosti među recenzentima (Inter-rater Reliability)

3.1 Metodologija

Za procjenu pouzdanosti ocjenjivanja korišten je Intraclass Correlation Coefficient (ICC), standardna mjera slaganja između recenzenata. ICC se izračunava kao:

$$ICC = \frac{\sigma^2_{between}}{\sigma^2_{between} + \sigma^2_{within}} \tag{1}$$

gdje je $\sigma^2_{between}$ varijanca između objekata ocjenjivanja (projekata), a σ^2_{within} varijanca unutar objekata (među recenzentima).

Interpretirani su sljedeći ICC tipovi:

- **ICC(1)**: Mjeri apsolutno slaganje pojedinačnih recenzenata
- **ICC(2)**: Mjeri pouzdanost prosječne ocjene svih recenzenata
- **ICC(3)**: Mjeri konzistentnost recenzenata

Općeprihvaćene smjernice za interpretaciju ICC vrijednosti:

- ICC < 0.40: Slabo slaganje
- ICC 0.40-0.59: Umjereno slaganje
- ICC 0.60-0.74: Dobro slaganje
- ICC > 0.75: Izvrsno slaganje

3.2 Rezultati

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

- **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 recenzenata. 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 mjera po projektima. Gornji panel (ICC1) pokazuje distribuciju apsolutnog slaganja pojedinačnih recenzenata 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 recenzenata s nešto boljim ali još uvijek neprihvatljivim vrijednostima. Crvene isprekidane linije označavaju prosječne vrijednosti za svaku mjeru.

4 Analiza stilova recenziranja

4.1 Metodologija

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

1. Kreiranje profila recenzenata na temelju prosječnih ocjena po svim kriterijima
2. Izračun matrice sličnosti pomoću korelacijske udaljenosti: $d_{ij} = 1 - cor(r_i, r_j)$
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: $Gap_k = E_n^*[\log(W_k)] - \log(W_k)$

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 recenzenata (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 **dodatni dokaz nekonzistentnosti**. To znači da ocjena projekta više ovisi o stilu recenziranja dodijeljenog recenzenta nego o stvarnoj kvaliteti projekta.

5 Mixed Effects Model analiza

5.1 Metodologija

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

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + \epsilon_{ijk} \tag{2}$$

gdje je:

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

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

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 već je rezultat:

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

Sustav evaluacije 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 recenzenata. Projekti su poredani od najnižeg do najvišeg slučajnog učinka. Široki intervali pouzdanosti i mala odstupanja od nule (većina projekata ima efekte između -0.5 i +0.5) dodatno potvrđuju da stvarne razlike između projekata čine zanemariv dio ukupne varijabilnosti u sustavu ocjenjivanja.

6 Analiza slaganja recenzenata po rangu projekta

6.1 Metodologija

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

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

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

Analizirani su trendovi za:

- Prosječne ocjene po projektu
- Slaganje recenzenata (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 evaluacije.

7 Identifikacija problematičnih slučajeva

7.1 Metodologija

Za identifikaciju projekata s najvećim neslaganjem kreiran je kompozitni indeks problematičnosti:

$$Problem_Score = Range \times (1 - ICC1) \times CV \tag{4}$$

gdje je:

- $Range$ - raspon ocjena K3 među recenzentima
- $ICC1$ - mjera slaganja recenzenata
- CV - koeficijent varijacije

7.2 Rezultati

Najproblematičniji projekti pokazuju ekstremne razlike u ocjenama:

Tablica 1: Top 10 najproblematičnijih projekata prema kompozitnom indeksu

Projekt	**K3 ocjene**	**Raspon**	**ICC1**	**CV**	**Problem Score**
P94	15, 10, 7	8	-0.087	0.379	3.30
P93	7, 15, 14	8	-0.089	0.363	3.16
P89	15, 8, 13	7	-0.220	0.300	2.56
P76	15, 8, 14	7	-0.153	0.307	2.48
P92	14, 15, 8	7	-0.078	0.307	2.32
P70	15, 9, 15	6	-0.009	0.266	1.61
P62	15, 10, 15	5	-0.348	0.217	1.46
P69	14, 15, 10	5	-0.272	0.204	1.29
P84	10, 15, 12	5	-0.173	0.204	1.20
P86	15, 10, 15	5	-0.079	0.217	1.17

Ove razlike su **apsurdno velike** za sustav koji bi trebao objektivno evaluirati projekte:

- Projekt P94: jedan recenzent dao 7 bodova, drugi 15 bodova (više od dvostruke razlike!)
- Projekt P93: raspon od 7 do 15 bodova među tri recenzenta
- Razlika između minimalne i maksimalne ocjene može značiti razliku između financiranja i odbacivanja projekta

8 Prediktivna analiza

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
- Najvažnije varijable: K2_3 (30% važnosti), K2_1 (9%), K1_2 (7%)

Kritična implikacija: Visoka prediktivnost K3 ocjena iz drugih komponenti sugerira da K3 nije nezavisan kriterij već je visoko koreliran s drugim ocjenama. To dovodi u pitanje smisao njegovog izdvajanja kao posebnog kriterija.

9 Analiza konzistentnosti kroz projekte

9.1 Metodologija

Za svaku poziciju recenzenta (R1, R2, R3) izračunata je:

- Prosječna devijacija od prosjeka projekta
- Standardna devijacija tih odstupanja
- Mjera konzistentnosti: $Consistency = 1 - \frac{SD_{deviation}}{SD_{total}}$

9.2 Rezultati

Tablica 2: Konzistentnost pozicija recenzenata

Pozicija	Prosječna devijacija	SD devijacije	Konzistentnost
R1	0.032	1.25	0.233
R2	-0.053	1.29	0.210
R3	0.021	1.22	0.250

Niske vrijednosti konzistentnosti (sve ispod 0.25) pokazuju da čak ni pozicije recenzenata ne pokazuju dosljedne obrasce ocjenjivanja kroz projekte.

10 Fokus na K3 kriterij

K3 ocjene su bile izdvojene kao posebno važan kriterij u evaluaciji. Naša analiza pokazuje da su upravo K3 ocjene:

- Pokazale **najnižu pouzdanost** među recenzentima (ICC < 0.2)
- Imale **najveće raspone ocjena** za iste projekte (do 8 bodova razlike)
- Bile **visoko predvidljive** iz drugih kriterija (nisu nezavisne)
- Pokazale **dramatičan pad kroz vrijeme** evaluacije

Dio II
Dodatne statističke analize

11 Analiza slaganja recenzenata - Kendall-ov W

11.1 Opis metode

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

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

gdje je:

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

12.1 Opis metode

Diskriminativnost se analizira kroz tri mjere:

12.1.1 Koeficijent varijacije (CV)

$$CV = \frac{\sigma}{\mu} = \frac{\text{SD prosječnih ocjena}}{\text{Prosjek prosječnih ocjena}} \tag{6}$$

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

12.1.2 Bayesov čimbenik za ANOVA

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

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

12.1.3 Relativna entropija

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

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

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.** Density plot 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

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 $j = 1, ..., 1000$:
   - 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 recenzenata s nasumičnim ocenjivanjem.** Histogram prikazuje distribuciju Kendall-ov W koeficijenata dobivenih iz 1000 simulacija nasumičnog ocenjivanja. 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 ocenjivanja.

14 Analiza ceiling učinka

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

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

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)

Slika 11: **Distribucija svih K3 ocjena - demonstracija ceiling učinka.** Bar chart 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.

15 Analiza informativnosti kriterija

15.1 Opis metode

15.1.1 Normalizirana mutual information

Mutual information između dvije varijable X i Y:

$$I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{11}$$

Normalizirana mutual information:

$$NMI(X;Y) = \frac{I(X;Y)}{\min(H(X), H(Y))} \tag{12}$$

gdje su $H(X)$ i $H(Y)$ entropije pojedinačnih varijabli.

15.1.2 Analiza glavnih komponenti (PCA)

PCA dekompozira matricu kovarijance:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}\mathbf{V}^T \tag{13}$$

gdje su stupci matrice $\mathbf{V}$ svojstveni vektori, a $\mathbf{D}$ dijagonalna matrica svojstvenih vrijednosti.

Proporcija varijance objašnjene i-tom komponentom:

$$\mathrm{PVE}_i = \frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j} \tag{14}$$

15.2 Rezultati

- **PC1 objašnjava 41.2%** ukupne varijance
- **PC1-3 objašnjavaju 65.8%** varijance
- Normalizirana MI: 0.082-0.211 (umjerena redundantnost)

Rezultati pokazuju umjerenu, ali ne potpunu redundantnost kriterija. K3 komponente pokazuju najvišu međusobnu povezanost (MI = 0.211).

16 Analiza ekstremnih obrazaca ocjenjivanja

16.1 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**: $K3_1 = K3_2 = K3_3$
3. **Sve maksimalne**: sve ocjene = 5

16.2 Rezultati

Slika 12: **Ekstremni obrasci ocjenjivanja.** Bar chart 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. Najzabrinjavajući je podatak da 60% recenzija daje identične ocjene za sve tri K3 komponente, što sugerira da većina recenzenata 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 evaluacijskog sustava s 9 komponenti.

Dio III
Bayesova analiza

17 Struktura podataka i osnovni pokazatelji

17.1 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**

17.2 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 nalazi:

- **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)

18 Bayesovi hijerarhijski modeli

18.1 Opis metode

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

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \tag{15}$$

gdje je:

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

18.2 Problemi s konvergencijom

Model je pokazao ozbiljne probleme s konvergencijom:

- **16.0% divergentnih tranzicija**
- **E-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.

18.3 Dekompozicija varijance

Proporcija ukupne varijance objašnjene različitim izvorima:

$$\text{ICC}_{\text{projekt}} = \frac{\sigma^2_{\text{projekt}}}{\sigma^2_{\text{projekt}} + \sigma^2_{\text{recenzent}} + \sigma^2_{\text{rezidual}}} = 0.09 \tag{16}$$

$$\text{ICC}_{\text{recenzent}} = \frac{\sigma^2_{\text{recenzent}}}{\sigma^2_{\text{projekt}} + \sigma^2_{\text{recenzent}} + \sigma^2_{\text{rezidual}}} = 0.32 \tag{17}$$

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

19 Analiza pristranosti po pozicijama

19.1 Bayesov čimbenik

Bayesov čimbenik za testiranje razlika među pozicijama recenzenata (R1, R2, R3) računa se kao:

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

gdje je H_0: nema razlika među pozicijama, a H_1: postoje razlike.
BF < 1 pruža dokaze **protiv** postojanja razlika među pozicijama.

Slika 13: **Distribucija K3 ocjena po pozicijama recenzenata.** Violin plot prikazuje distribuciju K3 ukupnih ocjena (raspon 3-15) za tri pozicije recenzenata (R1, R2, R3). Širina violina na svakoj visini predstavlja gustoću podataka na toj vrijednosti. Boxplot unutar svakog violina 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.

20 Analiza slaganja recenzenata

20.1 Kendall-ov W test

Kendall-ov W je mjera slaganja među recenzentima, definirana kao:

$$W = \frac{12S}{m^2(n^3 - n)} \tag{19}$$

gdje je:
- S = suma kvadrata odstupanja rangova
- m = broj recenzenata (3)
- n = broj projekata (95)

Rezultat: **W = 0.045** (p < 0.001)

Interpretacija: W blizu 0 ukazuje na praktički nepostojeće slaganje među recenzentima.

21 Analiza diskriminativnosti

21.1 Koeficijent varijacije

Koeficijent varijacije među projektima računa se kao:

$$CV = \frac{\sigma}{\mu} = \frac{0.635}{14.01} = 0.0453 \tag{20}$$

21.2 Relativna entropija

Entropija distribucije projekata:

$$H = -\sum_i p_i \log(p_i) = 1.006 \tag{21}$$

$$H_{\text{rel}} = \frac{H}{H_{\max}} = \frac{1.006}{3.332} = 0.302 \tag{22}$$

Niska relativna entropija (0.302) ukazuje na vrlo malu raznolikost u ocjenama projekata.

22 Usporedba s nasumičnim ocjenjivanjem

Provedena je simulacija s 1000 iteracija nasumičnog ocjenjivanja korištenjem empirijske distribucije ocjena. Rezultati pokazuju:

- Kendall-ov W za stvarne podatke: 0.045
- Prosječni Kendall-ov W za nasumične podatke: 0.046
- P-vrijednost: 0.451

Ključni nalaz: Ne možemo statistički razlikovati stvarne ocjene od nasumično generiranih ($p = 0.451 > 0.05$).

23 Analiza ceiling učinka

Ceiling učinak analiziran je Bayesovim faktorom protiv uniformne distribucije:

$$BF = \frac{P(D|\text{pristrana distribucija})}{P(D|\text{uniformna distribucija})} = 1.2 \times 10^{156} \tag{23}$$

Ovaj ekstremno visok Bayesov čimbenik pruža nepobitne dokaze o odstupanju od očekivane distribucije.

24 Analiza informativnosti kriterija

24.1 Analiza glavnih komponenti (PCA)

PCA analiza pokazala je da prva glavna komponenta objašnjava **84.9%** ukupne varijance, što ukazuje na to da svi kriteriji mjere praktički istu stvar.

24.2 Normalizirana mutual information

Prosječna normalizirana mutual information:

- Između K1 varijabli: 0.65
- Između K2 varijabli: 0.72
- Između K3 varijabli: 0.78
- Između različitih kriterija: 0.68

Visoke vrijednosti (>0.6) ukazuju na veliku redundantnost među kriterijima.

25 Identifikacija ekstremnih obrazaca

Analiza je identificirala sljedeće ekstremne obrasce ocjenjivanja:

- **45.3%** recenzija daje sve maksimalne ocjene (5-5-5 za sve komponente)
- **65.3%** ima identične K3 komponente

- **149 recenzenata** (52%) ima standardnu devijaciju = 0 (uvijek iste ocjene)

Slika 14: **Distribucija učinaka 285 jedinstvenih recenzenata.** Histogram prikazuje distribuciju procijenjenih slučajnih učinaka recenzenata iz Bayesovog hijerarhijskog modela. X-os predstavlja učinak recenzenta (odstupanje od općeg prosjeka), gdje pozitivne vrijednosti označavaju recenzente koji daju više ocjene od prosjeka, a negativne one koji daju niže. Crvena isprekidana linija na 0 označava neutralnog recenzenta. Distribucija pokazuje blagu pristranost prema pozitivnim vrijednostima s većinom recenzenata grupiranima oko neutralne točke, ali s primjetnom skupinom recenzenata s pozitivnim efektima (blagi recenzenti). Širina distribucije ukazuje na značajnu varijabilnost među recenzentima.

26 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 efekte recenzenata.** Graf prikazuje procijenjeni nelinearni trend K3 ukupnih ocjena kroz projekte korištenjem GAM spline funkcije. X-os predstavlja redni broj (rang) projekta (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: **Caterpillar plot ekstremnih recenzenata.** Ovaj graf trebao je prikazati top 20 "blagih" i top 20 "strogih" recenzenata 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 recenzenata

27 Sažetak glavnih nalaza

Analiza je provedena na 285 zapisa ocjenjivanja za 95 projekata. Različite analitičke metode sugeriraju da je **stvarni broj recenzenata približno 102** (medijan procjena), što znači da svaki recenzent u prosjeku ocjenjuje 2.8 projekata.

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%

28 Karakteristike ocjenjivanja

Tablica 4: Distribucija ocjena

Ocjena	**Broj**	**Postotak**
1	10	0.4%
2	23	0.9%
3	140	5.5%
4	605	23.8%
5	1,787	70.0%
Ukupno	**2,565**	**100%**

Interpretacija: Distribucija pokazuje izrazitu pristranost prema visokim ocjenama – 70% svih ocjena je maksimalnih (5), što ukazuje na:

- Moguću pristranost u ocjenjivanju
- Nedostatak diskriminacije između projekata
- Potencijalno zajedničko ocjenjivanje od strane istih recenzenata

Tablica 5: Tipologija recenzenata prema varijabilnosti ocjena

Tip recenzenta	**Broj**	**Postotak**	**Prosječna ocjena**	**Std. dev.**	**Raspon**
Perfekcionisti (sve 5)	43	15.1%	5.00	0.000	0.00
Konstantni	1	0.4%	4.00	0.000	0.00
Umjereno varijabilni	159	55.8%	4.74	0.429	1.01
Vrlo varijabilni	82	28.8%	4.17	0.834	2.33

Interpretacija:

- 15.1% ocjena pokazuje obrazac "sve petice", što je neobično visok postotak

- Većina recenzenata (55.8%) pokazuje umjerenu varijabilnost
- Samo 28.8% pokazuje značajnu varijabilnost u ocjenama

29 Procjena broja recenzenata

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**	–

Interpretacija:
- Velike razlike između metoda (2–158) ukazuju na složenost problema
- Elbow i hijerarhijske metode daju nerealno niske procjene
- GMM, jedinstveni obrasci i mrežna analiza daju više procjene (146–158)
- Bootstrap analiza (102.3 $\pm$ 4.5) podupire medijan od 102

Slika 17: Distribucija procjena broja recenzenata različitim metodama. Graf prikazuje horizontalni stupčasti 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.

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

30 Analiza sličnosti i povezanosti

Tablica 8: Analiza korelacija između recenzenata

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

Interpretacija:

- Pronađen je samo jedan par s identičnim obrascima (P10_R2 i P68_R2)
- 57 parova pokazuje vrlo visoku sličnost (>0.95)
- To sugerira da neki recenzenti koriste vrlo slične obrasce ocjenjivanja

Slika 18: Mreža povezanih recenzenata 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 recenzenata s vrlo sličnim stilom ocjenjivanja. Najveća komponenta ima 32 člana, dok je 129 recenzenata izolirano (45.3%), što znači da nemaju visoke korelacije 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:
- 45.3% (129/285) recenzenata je izolirano – nemaju visoke korelacije s drugima
- Najveća komponenta ima 32 člana – moguće isti recenzent ili grupa s vrlo sličnim stilom
- 158 komponenti sugerira gornju granicu broja različitih recenzenata

31 Najčešći obrasci ocjenjivanja

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	4-5-5-5-5-5-5-5-5	23	8.1%
3	5-5-5-5-4-5-5-5-5	12	4.2%
4	4-5-5-4-4-5-5-5-5	10	3.5%
5	4-5-5-5-4-5-5-5-5	7	2.5%
6	5-5-5-4-4-5-5-5-5	7	2.5%
7	4-5-5-4-5-5-5-5-5	5	1.8%
8	5-5-5-4-5-5-5-5-5	5	1.8%
9	5-5-5-5-5-5-4-5-5	5	1.8%
10	4-4-5-5-5-5-5-5-5	4	1.4%

Interpretacija:
- Top 10 obrazaca pokriva 42.7% svih ocjena
- Dominiraju obrasci s visokim ocjenama (uglavnom 4 i 5)
- Prvi obrazac (sve petice) se pojavljuje na 43 mjesta kroz različite projekte

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 recenzenata i onih koji ocjenjuju više projekata.

32 PCA analiza

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:

- Prva komponenta objašnjava 41.2% varijance – dominantan čimbenik
- Potrebno je 5 komponenti za 80% varijance
- Relativno visoka dimenzionalnost sugerira raznolikost u ocjenjivanju

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.

33 Dodatne analize

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 predstavljaju 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.

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 (recenzenata) blizu maksimalnog testiranog broja. Ovo je konzistentno s činjenicom da imamo 147 jedinstvenih obrazaca ocjenjivanja.

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 recenzenata, što sugerira da WSS kriterij nije optimalan za ovaj problem zbog dominacije nekoliko vrlo čestih obrazaca ocjenjivanja.

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 recenzenata.

34 Zaključci o broju recenzenata

34.1 Glavne polaznice za zaključivanje

1. **Broj recenzenata**: Najbolja procjena je **102 recenzenta** (95% CI: 94–111)
2. **Kvaliteta ocjenjivanja**:
   - Pristranost prema visokim ocjenama (70% petica)
   - 15% zapisa ima maksimalne ocjene na svim kriterijima
   - Niska diskriminacija između projekata
3. **Obrasci ponašanja**:
   - 147 jedinstvenih obrazaca od 285 zapisa
   - 45% recenzenata pokazuje izoliran obrazac
   - Nekoliko grupa recenzenata s vrlo sličnim stilom

Tablica 12: Sažetak

Pokazatelj	**Vrijednost**
Procijenjeni broj recenzenata	102
Prosječan broj projekata po recenzentu	2.8
Postotak maksimalnih ocjena	15.1%
Broj izoliranih recenzenata	129 (45.3%)
Prosječna entropija	0.790
Koeficijent varijacije procjena	70.7%

Dio V
Zaključak: Totalni kolaps sustava vrednovanja

35 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.

35.1 Klasična statistička analiza - temelj dokaza

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

35.1.1 Katastrofalna dekompozicija varijance

Najstrašniji nalaz cijele analize je da se **samo 6.4% varijance u ocjenama može pripisati stvarnim razlikama između projekata**, 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 mjerenja. Upravo to je situacija s ovim sustavom evaluacije.

35.1.2 Negativne ICC vrijednosti - gore od nasumičnosti

Možda je jošp više zabrinjavajuć nalaz 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 suprotnije procjene nego što bi dali potpuni laici koji nasumično pogađaju!

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

35.1.3 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.

35.1.4 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 evaluacije.

35.2 Dodatne statističke analize - potvrda katastrofe

35.2.1 Nedovoljno slaganje recenzenata

Kendall-ov W = 0.398 ukazuje na umjereno nisko slaganje, ali kontekst čini ovaj nalaz 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.

35.2.2 Mikroskopska diskriminativnost

Koeficijent 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.

35.2.3 Granična razlika od nasumičnosti

P-vrijednost od 0.06 u usporedbi s nasumičnim ocjenjivanjem je možda još strašniji nalaz. 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!!!

35.2.4 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).

35.2.5 Površnost i *copy-paste* mentalitet

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

35.3 Bayesova analiza - finalni čavao u lijes

35.3.1 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.

35.3.2 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 ($W > 0.9$).

35.3.3 Statistička nerazlučivost od nasumičnosti

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

35.3.4 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 recenzenata nije ni pokušala evaluirati projekte. Oni su jednostavno stavili najviše ocjene i završili posao.

35.3.5 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.

35.4 Procjena broja recenzenata - dodatna komplikacija

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

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

36 Implikacije za znanstvenu zajednicu

36.1 Totalni gubitak kredibiliteta

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

36.1.1 Potpunu odsutnost znanstvenih standarda

Sustav krši sve osnovne principe objektivnog mjerenja:

- **Valjanost**
- **Pouzdanost**
- **Objektivnost**
- **Diskriminativnost**
- **Pravednost**

Valjanost Valjanost je narušena jer ocjene ne odražavaju stvarnu kvalitetu projekata, nego proizvoljne dojmove i kriterije nepovezane s istraživačkom izvrsnošću.

Pouzdanost Pouzdanost sustava je niska 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 Objektivnost je kompromitirana jer su ocjene više odraz osobnih preferencija i stavova recenzenata nego mjerljivih pokazatelja kvalitete.

Diskriminativnost Diskriminativnost sustava također izostaje jer sustav očito ne uspijeva jasno razlikovati projekte visoke kvalitete od onih lošijih, čime se gubi njegova temeljna svrha u selekciji.

Pravednost Pravednost sustava ozbiljno je dovedena u pitanje jer ocjenjivanje sustavno favorizira ili diskriminira projekte na temelju čimbenika koji nisu povezani s njihovom stvarnom vrijednošću, poput institucionalne pripadnosti ili područja istraživanja.

36.1.2 Iluziju objektivnosti

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

36.1.3 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)
- Kada je njihov projekt ocijenjen (temporalni učinak)
- Čiste sreće (49.3% rezidualne varijance)

36.1.4 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?

37 Hitne preporuke za akciju

37.1 Trenutne mjere

37.1.1 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.

37.1.2 Javna isprika i transparentnost

Institucija 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

37.2 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 evaluacije. 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.

37.2.1 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
- Temporalni učinci moraju biti kontrolirani

37.2.2 Promjene sustava

1. **Externa revizija**
2. **Kontinuirani monitoring**
3. **Transparentnost**
4. **Odgovornost**

Externa revizija. Neovisni međunarodni panel mora pregledati cijeli proces evaluacije 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 recenzenata, pravovremeno prepoznavanje nepravilnosti i kontinuirano poboljšavanje evaluacijskog procesa.

Transparentnost. Svi podaci, korištene metode i rezultati evaluacije 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 evaluacije, uključujući definiranje tko donosi ključne odluke, kako se provodi nadzor i koje su posljedice za nepoštivanje uspostavljenih standarda.

38 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 evaluacije 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 nerazličiv od bacanja kocke, kako možemo tvrditi da cijenimo znanstvenu izvrsnost?

Ovaj debakl mora biti prekretnica. Ne smijemo dopustiti da se ovakve farce ponove. Znanstvena zajednica mora pokazati da 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 radikalne promjene. Naša vjerodostojnost kao znanstvenika ovisi o tome.