KEITH R. BILLINGSLEY
The University of Georgia, USA
ROBERT MUNZENRIDER
The University of Florida, USA
JAMES E. PRATHER
Georgia State University, USA
MARWIN K. HOFFMAN
Appalachian State University, USA

Slučaj lažne korelacije - ilustracija uobičajene pogrešne upotrebe indeksnih varijabli pomoću kompjuterske simulacije

Ovaj je rad namijenjen istraživačima koji analiziraju agregatne podatke metodom korelacija, a naročito je važan za one koji proučavaju podatke u per capita obliku.

Indeksna varijabla jeste omjer dviju varijabli. Četiri varijable x_1, x_2, x_3 i x_4 omogućuju tako računanje indeksnih varijabli x_1/x_2, x_3/x_4, x_1/x_3, x_4/x_3, itd. Indeksne se varijable često upotrebljavaju na područjima: urbanizacije (omjer gradske populacije prema totalnoj populaciji) troškova po glavi stanovnika (omjer ukupnih troškova i ukupne populacije) i poslovno izbornog rezutata (omjer broja glasova za određenog kandidata i ukupnog broja glasova).

Karl Pearson je još prije više od sedamdeset pet godina pokazao da veliki broj korelacija među indeksnim varijablama (npr. između urbanizacije i troškova po glavi stanovnika) nije odraz neke unutarnje, organske ili teorijski važne povezanosti među osnovnim veličinama ,već posljedica primijenjenog matematičkog postupka, tj. diobe, kojom su indeksi odbiveni.[1] Korelaciju koja je posljedica matematičkog postupka Pearson je nazvao »prividnom«. On je pokazao da korelacija dvije indeksne mjere djelomično odražava korelaciju koja je postojala među varijablama prije nego što su od njih formirani indeksi.

Na osnovi Pearsonovog rada i naše analize koju ćemo prikazati došli smo do zaključka da korelacione tehnike (kao što su faktorska analiza i ana-

[1] Karl Pearson, »Mathematical Contribution to the Theory of Evolution — on a From of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs«, *Proceedings of the Royal Society of London*, LX (1897), pp. 489—98. Ekonometrijska literatura također sadrži reference za ovaj problem. Vidi npr. Edwin Kult and John Meyer, »Correlation and Regression Estimates when the Data are Ratios«, *Econometrica*, 34 (1966), pp. 400—416.

liza multiple regresije) nije opravdano primjenjivati u slučajevima kada baratamo indeksnim varijablama dobivenim iz zajedničke ili koreliranih baza ako istodobno ne posvetimo pažnju uklanjanju problema koji proizlaze iz koreliranja indeksnih varijabli. Iako mnogi istraživači koji analiziraju upravo ovu vrstu podataka poznaju probleme upotrebe korelacija indeksnih varijabli, analiza novijih izdanja važnijih stručnih časopisa pokazuje da ta upozorenja nisu postigla željeni cilj ili se čak potpuno ignoriraju.

Iznijet ćemo kratki prikaz matematičkog dokazivanja hipoteze o lažnosti takvih korelacija. Literatura iz koje crpimo ovo matematičko dokazivanje već je duže vrijeme poznata, ali budući da nije znatnije utjecala na mišljenje istraživača prikazat ćemo i rezultate *kompjutorske simulacije* koju smo izveli da bismo prikazali, nadamo se prilično upečatljivo, opasnosti nekritičnog pristupa. Premda i dalje nastojimo pronaći druga rješenja ovog problema namjera nam je da time otvorimo formalnu diskusiju.

JEDNADŽBE

Slučaj četiriju varijabli. — Uzmimo primjer korelacije između dva indeksa, nazovimo ih *A* i *B*, gdje je $A = x_1/x_3$ i $B = x_2/x_4$. Tako r_{12}, r_{14} i r_{34} možemo zamisliti kao interkorelacije između varijabli — baza, a v_1 v_2, v_3, i v_4 kao koeficijente varijacije (standardna devijacija kroz aritmetičku sredinu) za svaku od četiri varijable — baze. Korelacija r_{AB} u terminima originalnih varijabli može se prikazati na slijedeći način:[2]

$$r_{AB} = \frac{r_{12}v_1v_2 - r_{14}v_1v_4 - r_{23}v_2v_3 + r_{34}v_3v_4}{\sqrt{v_1^2 + v_3^2 - 2\,r_{13}v_1v_3}\ \sqrt{v_2^2 + v_4^2 - 2\,r_{24}v_2v_4}} \quad \text{(I)}$$

Cilj konstrukcije indeksa jest »kontrola« utjecaja x_3 i x_4; međutim, kako pokazuje jednadžba I, utjecaj originalnih varijabli x_3 i x_4 time što su one podijeljene s varijablama x_1 i x_2 uopće nije uklonjen. Na primjer, u brojniku jendadžbe I jasno se vidi da su originalne prvobitne korelacije između varijabli tri i četiri (r_{34}) *aditivne* korelaciji između A i B.

Daljnju ilustraciju problema možemo postići ako damo imena simbolima od x_1 do x_4. Recimo da je x_1 = gradsko stanovništvo određenog izbornog područja, x_2 = ukupni broj glasova za određenog kandidata, x_3 = ukupna populacija i x_4 = ukupni broj glasova. Tako je x_1/x_3 = urbanizacija i x_2/x_3 = podrška kandidatu. Proučava li analitičar koncept urbanizacije u odnosu na podršku određenom kandidatu, njegovi će rezultati biti vrlo nerealni ako se zaključivanje osniva samo na korelaciji između A i B. Iz jednadžbe I možemo uočiti da faktor koji sadrži korelaciju između originalnih bazičnih varijabli ($r_{34}v_3v_4$) jeste sastavni dio indeksne korelacije. Indeksna je varijabla konstruirana zato da bi se kontrolirao utjecaj totalne populacije (x_3) i ukupnog broja glasova (x_4); iz jednadžbe I vidimo da je odnos između urbanizacije i podrške određenom kandidatu posljedica (bar djelomično) činjenice što naseljenije regije imaju više glasača!

Slučaj triju varijabli. — Jednadžba za korelaciju između dva indeksa koji imaju zajednički nazivnik (tj. x_1/x_3, x_2/x_3) može se prikazati na slijedeći način:[3]

[2] Karl Pearson, »Mathematical Contributions. . .«, pp. 493. Ovaj rad sadrži različite pretpostavke primijenjene u razvijanju jednadžbi.

[3] Ibid.

$$r_{CD} = \frac{r_{12}v_1v_2 - r_{13}v_1v_3 - r_{23}v_2v_3 + v_3^2}{\sqrt{v_1^2 + v_3^2 - 2\,r_{13}v_1v_3}\ \sqrt{v_2^2 + v_3^2 - 2\,r_{23}v_2v_3}} \quad \text{(II)}$$

Opet uočavamo da prvobitne bazične varijable nisu uklonjene. Korelacija r_{AB} složen je produkt prvobitnih odnosa između svih triju varijabli, te iz konačne korelacije nije izoliran utjecaj x_3.

Poslužimo se ponovno primjerima kako bismo bolje razumjeli probleme u slučaju triju varijabli. Ako su x_1 = ukupni nacionalni troškovi, x_2 = ukupni osobni prihod i x_3 = ukupna populacija, onda $C = x_1/x_3$ možemo nazvati troškovima po glavi (ulaganje?) a $D = x_2/x_3$ dohodak per capita (bogatstvo?). Na temelju intuicije mogli bismo reći da su mjere »ulaganje« i »bogatstvo« isključile očiti utjecaj koji ima populacija na ukupne prihode i rashode u zemlji. Na žalost, i ovdje intuicija zavarava. Možda je varijabla »populacija« (x_3) prikrivena, ali su njeni utjecaji očiti.

Uobičajena interpretacija produkt-moment korelacije ne može se primijeniti u slučaju korelacije među indeksima koji imaju zajednički nazivnik. Postoji li bilo kakva varijacija unutar varijabli x_1, x_2 i x_3, r_{CD} ne može biti jednak nuli. Razmotrimo pojendostavljenu jednadžbu II, ako su bazične varijable potpuno nezavisne, tj. $r_{12} = r_{13} = r_{23} = 0$: [4]

$$r_{CD} = \frac{v_3^2}{\sqrt{v_1^2 + v_3^2}\ \sqrt{v_2^2 + v_3^2}} \quad \text{(III)}$$

Korelacija među indeksnim varijablama postoji čak i onda kad nema nikakve povezanosti među baznim varijablama; ova je vrsta korelacija među indeksnim varijablama lažna.

Zbog intuitivne privlačnosti indeksnih varijabli, a naročito u per capita obliku, možda je problem prikazan jednadžbom III isuviše apstraktan a da bi bio uvjerljiv. Zato ćemo u slijedećem odjeljku problem ilustrirati u praktičnoj primjeni.

SIMULACIJE

Varijble dobivene pomoću slučjnih brojeva. — Lažna priroda indeksne korelacije omogućuje računanje znatnih korelacija između indeksa dobivenih iz slučajno odabranih brojeva. Očekivali bismo da će prosječna korelacija među parovima slučajnih brojeva biti jednaka nuli. Međutim, kad se ovakve varijable kombiniraju u indekse prema obrascu danom u jednadžbi II, dobivene korelacije gotovo nikad nisu 0. Tabela 1 prikazuje rezultate dobivene kad su tri varijable sastavljene od slučajnih brojeva i kombinirane u dva indeksa: x_1/x_2 i x_2/x_3.[5] U tabeli 1 prikazane su prosječne koleracije dobivene nakon višekratnih ponavljanja odabiranja slučajnih brojeva za različite veličine uzroka i različite osnovne distribucije. Kako pokazuje tabela 1, dobivene korelacije za sve kombinacije velčine uzorka i distribucije koje smo proučavali u našoj su simulaciji u prosjeku iznad 0.5. Na primjer, kad smo sparivali slučajne varijable za četrdeset osam simularnih situacija, dobivena korelacija nakon 200 ponavljanja slučajnog sparivanja bila je u prosjeku 0.53!

---

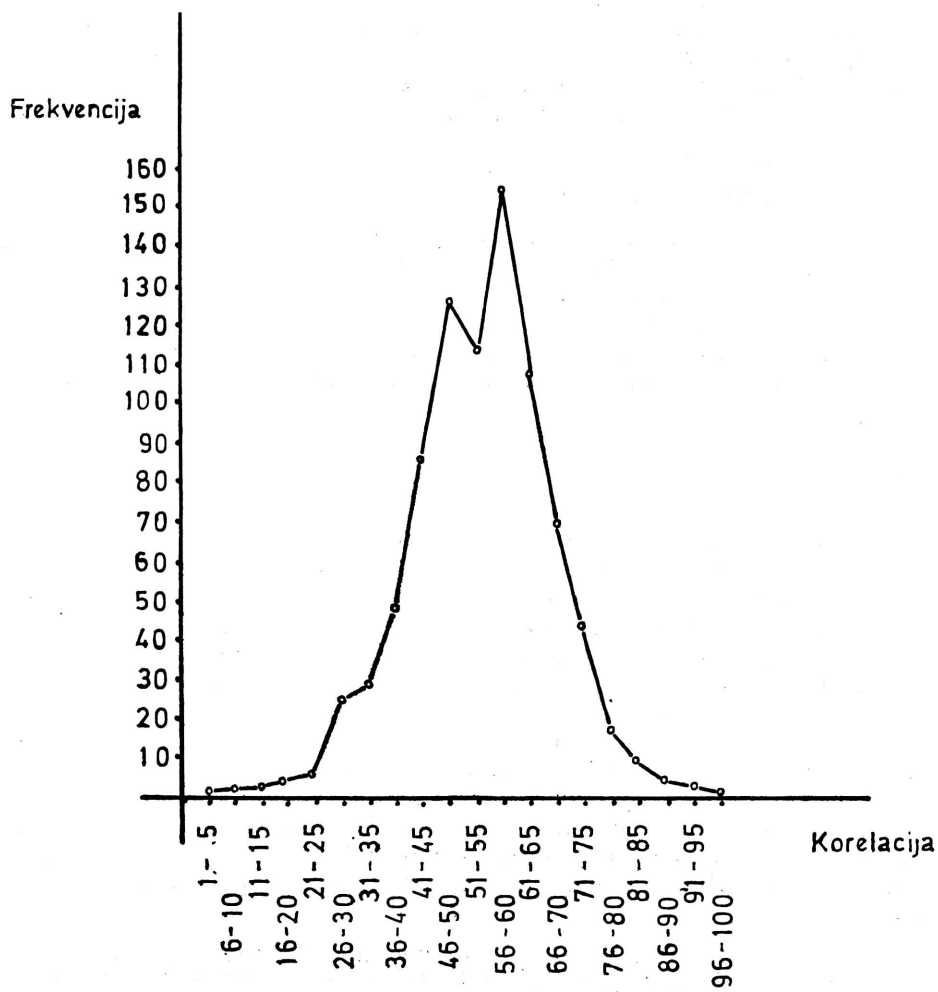
[4] Ibid.

[5] International Business Machines, *System 1360 Scientific Subroutine Package* (360 A-CM-03X) Version III, (Wihte Plains, N. Y., IBM, 1968), p. 77. Vidi navedeni izvor u vezi generatora slučajnih brojeva koji je upotrebljen u ovom radu.

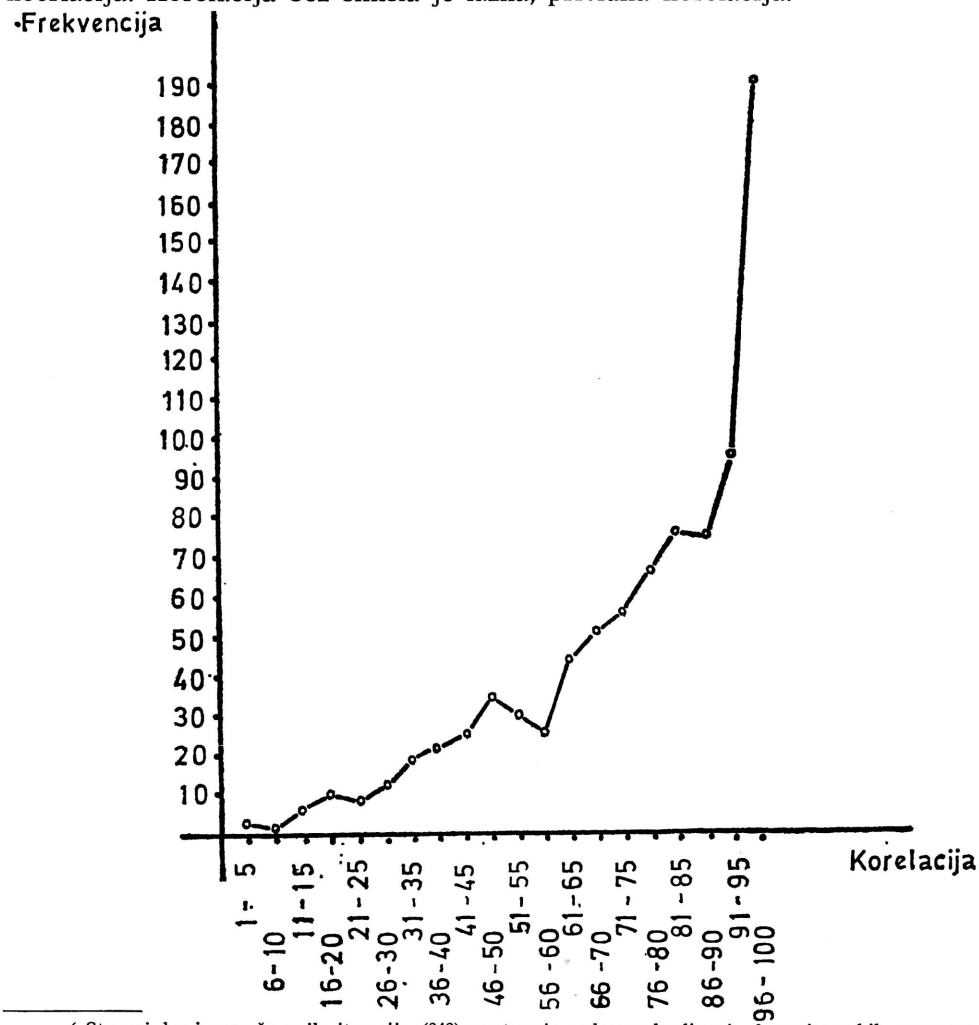
Prosječne indeksne[a)] korelacije za određene veličine uzorka i distribucije prema slučaju raspoređene bazne varijable

Ponavljanja	Osnovna distribucija tri originalne varijable				
	Normalna veličina uzorka				Uniformna veličina uzorka
	25	48	100	1000	48
50	.54	.56	.57	.58	.72
100	.52	.54	.57	b	.71
150	.53	.54	.55	b	.74
200	.53	.53	.55	b	.75

a) Ove su korelacije računate između $C = x_1/x_3$ i $D = x_2/x_3$
Tabela odgovara slučaju triju varijabli prikazanom u jednadžbama II i III.
b) Vrijednost nije izračunata.



Radi toga što se veličina uzorka od 48 često javlja u ekonomskim studijama u SAD, ovu veličinu uzorka podvrgli smo rigoroznijoj analizi. Na grafikonu 1 prikazana je frekvencija distribucija dobivenih korelacija između slučajnih indeksnih vrijednosti, gdje su sumirani rezultati 849 posebnih spariavnja tri varijable kad je veličina uzorka 48.[6] Tri varijable kombinirane su u dva indeksa; koeficijenti korelacije između indeksa predstavljaju jednu os u dijagramu distirbucije frevencija. Iako distribucija pokazuje nepravilnosti, krivulja predstavlja pirbližno normalnu distirbuciju s artimetričkom sredinom 154 i standardnom devijacijom 0.123.[7] Iako slučajno dobivene varijable sigurno nemaju sadržajni smisao ipak među tako dobivenim indeksima postoji znatna koerlacija. Korelacija bez smisla je lažna, prividna korelacija.



[6] Stvarni broj proučavanih iteracija (849) nastao je nakon određivanja kompjuterskih vremena. Namjera je bila da se dobije dovoljno velika frekvencija koja će dozvoliti proučavanje oblika distribucije ali to ne znači da je baš ovaj broj najoptimalniji.

[7] Radi obrade problema identificiranja normalne distribucije iz grafičkog prikaza vidi William C. Hays, *Statistics for Psöchologists* (New York: Holt, 1963), pp. 586—88.

Detaljna analiza tabele 1 upućuje da je prosječni nivo korelacija relativno konstantan za proučavane veličine uzorka, ali da oblik osnovne distribucije baznih varijabli ima znatnu ulogu. Dok normalno distribuirane slučajne varijable daju korelaciju oko 0.5, uniformno distribuirane varijable imaju prosječne indeksne korelacije oko 0.7. Dalje, iz grafikona 2 vidi se da se distribucija takvih indeksnih korelacija u toku višekratnih ponavljanja znatno razlikuje od distribucija karakterističnih za normalno distribuirane varijable. Za indeksne korelacije bazirane na jednolično distribuiranim slučajnim brojevima karakteristična je povezanost veličine korelacija i frekvencija pojavljivanja korelacija. To znači, ima mnogo više koeficijenata u rasponu 0.9 nego u rasponu 0.7, a raspon 0.7 sadrži mnogo više nego raspon 0.3. Koliko nam je poznato, ova osjetljivost na distribuciju dosad nigdje nije bila opisana.

I dok je problem indeksne korelacije vrlo ozbiljan i bez obzira na distribuciju, očito je da je on kod nekih distribucija mnogo značajniji nego kod drugih .

Varijable popisa stanovništva. — Za konačnu ilustarciju problema vezanih uz korelacije indeksnih varijabli u tabeli 2 prikazujemo rezultate analize stvarnih podataka cenzusa, koji se obično najviše koriste u analizama. U prvom stupcu tabele 2 prikazana je korelacija među indeksima dobivenim iz stvarnih podataka cenzusa. Ovi indeksi spadaju u slučaj triju varijabli, koji smo prikazali obrascom jednadžbi II i III. Stupac II predstavlja prosječnu korelaciju među indeksima kad su stvarne vrijednosti za varijablu pridružene slučajnim redoslijedom. Proces pridruživanja analogan je miješanju snopa karata: ne mijenjaju se vrijednosti, već njihov poredak. Nadamo se da je ovaj proces slučajnog raspoređivanja otklonio svaku mogućnost međusobne intuitivne povezanosti varijabli.

Iz stupca 2 je očito da je stupanj antificijelne korelacije dobivene upotrebom ove metode niži nego u ranijem slučaju. Prosječna *slučajna* vrijednost za podatke cenzusa je samo 0.3. Ovaj niži prosjek možemo interpretirati kao znak da se problem indeksne korelacije smanjuje kod podataka koji se obično najčešće koriste. Međutim, to je samo slaba utjeha. Činjenica da je korelacija dobivena za stvarne podatke sparene prema stvarnom stanju veća od *prosječne* korelacije stvarnih podataka, koji su spareni po slučaju ne govori ništa o mogućnosti da je stvarni opis samo artefakt metode.

Vrlo je privlačan zaključak da je odbivena korelacija između urbanizacije i dohotka po glavi stanovnika interesantna (i da odražava stvarnost) ne samo zato što iznosi 0.81, već i zato što slučajni oblik ove povezanosti iznosi samo 0.31. Problem kod ovog zaključivanja je u tome što 0.31 predstavlja prosjek, a izvjesni broj simulirnih vriednosti bio je i veći od njega. Stvarna distribucija korelacija za analizu podataka cenzusa nie dobivena, ali grafikon 1 pokazuje, da je 50% slučajnih korelacija veće od 0.5, a 2% čak veće od 0.8.[8] Bitno je, da korelacija i od 0.81 može biti rezultat znatnog iskrivljenja povezanosti među varijablama »koncentracija populacije« (urbanizacia) i »bogatstvo« (dohodak po glavi stanovnika). Devijacija opažene vrijednosti od prosjeka slučajnih podataka smislena je samo u okviru slučajne distribucije.

[8] U toku je proučavanje distribucije indeksa najčešćih varijabli cenzusa koji su upotrebljeni u cenzusu 1960. i 1970. godine.

Tabela 2:

Korelacije indeksnih varijabli koje su dobivene iz varijabli cenzusa[a] u stvarno dobivenom i slučajnom obliku u 1960. godini

Korelacija među indeksima[b]	Stvarna baza	Slučajna baza
Urbanizacija i prihod po glavi stanovnika (U/P × I/P)	.81	.31
Urbanizacija i troškovi po glavi stanovnika (U/P × E/P)	.32	.33
Prihod po glavi stanovnika i troškovi po glavi stanovnika (U/P × E/P)	.58	.30

a) U = ukupna urbana populacija; I = ukupni prihod; E = ukupni javni rashodi; P = ukupna populacija. P je zajednički u svim indeksima.

b) Za osnovne varijable korelacije su kako slijedi: urbana populacija i ukupni prihod = 0.99; urbana populacija i ukupni troškovi = 0.99; urbana populacija i čitava populacija = 0.99; čitava populacija i ukupni troškovi = 0.97.

ZAKLJUČAK

Iako prikazani podaci predstavljaju samo kratki rezime istraživanja koje je još u toku, iz njih bi jasno trebalo slijediti da, korelacije među indeksima mogu biti vrlo varljive. Pearson je primijetio taj problem u matematičkom postupku za izračunavanje produkt-moment korelacije. Nadamo se da je naša simulacija pokazala da se ova lažna priroda korelacije mora uzeti u obzir u istraživačkoj praksi. Čak ako i nemamo podesne metode za zamjenu, moramo zaključiti da koreliranje indeksnih varijabli može zavesti na pogrešne zaključke, jer su tako dobivene koerlacije lažne.

S engleskog prevela: Mira Čudina-Obradović

K. R. Billingsley, R. Munzenrider, J. E. Prather and M. K. Hoffman

THE CASE OF THE COUNTERFEIT CORRELATION: A COMPUTER SIMULATION ILLUSTRATING A COMMON ERROR WITH THE USE OF INDEX VARIABLES INCLUDING PER CAPITA CENSUS MEASURES

(Summary)

The presented material is particularly directed to scholars who examine aggregate data i.e. index variables through the use of correlational techniques. As K. Pearson already demonstrated a substantial portion of the correlation between the underlying concepts but rather to the operation — division — by whivh the indices were created. He termed this correlation as »spurios« and de-

monstrated that it is due in part to the prior correlation of the variables forming the indices. Here we present a brief overview of the mathematics supporting the assertion of spuriousness and also the results of a computer simulation which we developed to demonstrate the dangers of the approach. Our simulations, hopefully, demonstrate that counterfeit correlations should be of practical concern to the researcher interested in actual applications and that correlating index variables is likely to lead to misleading conclusions because such correlations are counterfeit.

Translated by M. **Čudina-Obradović**