

Uporaba kopul v hidrologiji

Nejc Bezak*, Mitja Brilly*, Matjaž Mikoš* in Mojca Šraj*

Povzetek

Funkcije kopula se v svetu vse pogosteje uporabljajo na različnih znanstvenih področjih, v hidrologiji pa se njihova uporaba v večji meri pojavlja šele v zadnjem desetletju. V prispevku so prikazani rezultati nekaterih praktičnih možnosti uporabe kopul v hidrologiji. Naredili smo bivariatne ter trivariatne verjetnostne analize elementov visokovodnih valov (konice pretokov, volumni ter časi trajanja), analizirali smo povezanost pretokov, volumnov valov ter vrednosti koncentracij suspendiranih snovi ter definirali model, s katerim lahko na podlagi znanih vrednosti pretokov in padavin ocenimo vrednosti koncentracij suspendiranih snovi. Z uporabo statističnih testov smo pokazali, da je kopula model dal boljše rezultate kot nekateri pogosto uporabljeni regresijski modeli. S tem smo v oceno koncentracij suspendiranih snovi vpeljali dodatne informacije (padavine) in dobili ocenjene vrednosti bližje dejanskim, izmerjenim vrednostim. Pri multivariatnih verjetnostnih analizah pa smo izračunali različne skupne povratne dobe.

Ključne besede: Kopula, bivariatne analize, trivariatne analize, pretok, volumen vala, suspendirane snovi.

Keywords: Kopula, bivariate analysis, trivariate analysis, discharge, wave volume, suspended solids.

Uvod

V hidrologiji so se do sedaj za izvedbo verjetnostnih analiz večinoma uporabljale univariatne porazdelitvene funkcije, kjer upoštevamo le eno spremenljivko (večinoma je to konica pretoka). V začetku 21. stoletja so se v hidrologiji pojavili prvi članki, kjer so raziskovalci za izvedbo različnih analiz uporabili funkcijo kopula (Favre et al., 2004; Salvadori & De Michele, 2004). V drugih vedah, kot sta ekonomija ali biologija, so se kopule uporabljale že nekoliko prej. Teorija kopul temelji na matematičnem teoremu, ki ga je že leta 1959 predstavil Sklar (Sklar, 1959), in bo nekoliko podrobneje predstavljen v nadaljevanju. Temeljno literaturo o funkcijah kopula pa predstavlja naslednje gradivo: Joe (1997), Nelsen (2006), Salvadori et al. (2007). Kopule omogočajo izgradnjo multivariatnega modela, kjer hkrati upoštevamo dve ali več v naravi odvisnih spremenljivk. Tako lahko pri verjetnostnih analizah poleg konic pretokov hkrati upoštevamo tudi volumne in čase trajanja visokovodnih valov ali druge spremenljivke. Ta koncept lahko uporabimo pri vseh multivariatnih problemih, kjer nastopa d medsebojno bolj ali manj odvisnih spremenljivk, kjer odvisnost določa koeficient korelacije (Kendallov ali Spearmanov) oziroma parameter kopule. Tako so bile kopule uporabljene za izvedbo multivariatnih verjetnostnih analiz visokovodnih valov (Favre et al., 2004; Salvadori & De Michele, 2004; Ganguli & Reddy, 2013), verjetnostne analize padavin (Zhang & Singh, 2007), analize suše (Shiau et al., 2007; Wong et al., 2010; Ma et al., 2011), geostatistične interpolacije kot alternativa običajnemu krigingu (Bardossy & Li, 2008), preverjanje ustreznosti prelivnega objekta na jezcu (De Michele et al., 2005) ter še pri mnogih drugih

* Univerza v Ljubljani, Fakulteta za gradbeništvo in geodezijo, Jamova 2, 1000 Ljubljana, Slovenija

hidroloških problemih. Na spletni strani mednarodnega hidrološkega društva STAHY (angl. *International Association of Hydrological Sciences*; www.stahy.org) so zbrani članki, kjer je bila za izvedbo hidroloških analiz uporabljena funkcija kopula.

Da se kopule v hidrologiji vse pogosteje uporabljajo priča tudi dejstvo, da se je število člankov s to metodologijo v bazi Web of knowledge (kategorija *Water resources*), od leta 2004 povzpelo na več kot 180, v še ne končanem letu 2013 pa je bilo do sedaj objavljenih največ prispevkov v zvezi s kopulami (do oktobra je bilo v letu 2013 indeksiranih 37 člankov).

Namen prispevka je prikaz praktične uporabe kopul v hidrologiji. Prikazan in opisan je postopek uporabe bivariatne ter trivariatne verjetnostne analize visokovodnih valov na primeru hidrološke postaje Litija na reki Savi. Funkcijo kopula smo uporabili tudi za modeliranje pretokov, volumnov valov ter koncentracij suspendiranih snovi. Za izvedbo te analize so uporabljeni podatki z vodomerne postaje Gornja Radgona na reki Muri, kjer so se več kot 25 let neprekinjeno izvajale dnevne meritve koncentracij suspendiranih snovi (ARSO, 2013). Poleg tega smo poskušali funkcijo kopula uporabiti za ocenjevanje vrednosti suspendiranih snovi na podlagi izmerjenih vrednosti pretokov in padavin. Za ta namen smo uporabili podatke z vodomerne postaje Ranca na reki Pesnici ter padavinske postaje Polički vrh. Na teh praktičnih primerih so opisani osnovni koraki za izvedbo analiz z uporabo funkcij kopula.

Podatki in metode

Za izvedbo različnih analiz smo uporabili podatke s hidroloških postaj Litija na reki Savi (vzorec letnih maksimalnih pretokov), Gornja Radgona na reki Muri (vzorec letnih maksimalnih pretokov in suspendiranih snovi), Ranca na reki Pesnici (mesečne vsote pretoka in suspendiranih snovi) ter podatke s padavinske postaje Polički vrh (mesečne vsote padavin), ki leži blizu vodomerne postaje Ranca. Preglednica 1 prikazuje nekatere osnovne značilnosti obravnavanih postaj ter lastnosti obravnavanih vzorcev. Zaradi različne narave obravnavanih problemov smo pri različnih postajah obravnavali različna obdobja meritev. Pri analizah koncentracij suspendiranih snovi smo bili omejeni z relativno kratkimi nizi meritev (Bezak et al., 2013b). Več informacij o podatkih in lastnostih vzorcev so podali Bezak et al., 2013a (postaja Litija) ter Bezak et al., 2013b (postaji Gornja Radgona in Ranca).

Preglednica 1: Pregled obravnavanih postaj in nekatere osnovne značilnosti pripadajočih vzorcev

Postaja	Litija	Gornja Radgona	Ranca	Polički vrh
Reka / Območje	Sava	Mura	Pesnica	Porečje Pesnice
Tip postaje	vodomerna	vodomerna	vodomerna	padavinska
Obdobje	1953-2010	1977-2005	1970-1973	1970-1973
Tip vzorca	letni maksimumi	letni maksimumi	mesečne vsote	mesečne vsote
Opazovane spremenljivke	pretok	pretok, suspendirane snovi	pretok, suspendirane snovi	padavine
Velikost vzorca (število podatkov)	58	29	45	45

Za izvedbo analiz smo torej uporabili funkcije kopula. Osnovni teorem je podal Sklar (1959):

$$F(x_1, \dots, x_d) = C(F_1(x_1), \dots, F_d(x_d)), \quad (1)$$

kjer je C d dimenzionalna funkcija kopula. Multivariatno d dimenzionalno porazdelitev lahko zapišemo kot kombinacijo funkcije kopula C in robnih porazdelitvenih funkcij F_1, \dots, F_d (angl. *marginal distributions*) (enačba 1). Gre v bistvu za običajne univariatne porazdelitvene funkcije, kjer lahko uporabimo različne parametrične (npr. normalna, log-normalna, Pearsonova III, generalizirana porazdelitev ekstremnih vrednosti,...) ali neparametrične (kernelova gostota; angl. *kernel density*) porazdelitvene funkcije. Glavna prednost funkcij kopula pred običajnimi multivariatnimi porazdelitvenimi funkcijami (npr. multivariatna normalna porazdelitev) je prav ločenost robnih porazdelitev in funkcij kopula. Ocenjevanje parametrov robnih porazdelitev (običajen postopek) in parametrov funkcije kopula sta tako dva ločena postopka. To dejstvo je pomembno zato, ker lahko kot robne porazdelitve posameznih spremenljivk izberemo različne porazdelitvene funkcije. Tako lahko npr. pri analizah konic, volumnov in časov trajanja visokovodnih valov kot robne porazdelitve izberemo različne porazdelitvene funkcije, npr. Gumbelovo za konice, Pearsonovo III za volumne ter normalno za čase trajanja. Sklarov teorem (Sklar, 1959) pa omogoča, da nato ocenimo parametre kopule povsem ločeno od robnih porazdelitvenih funkcij. Slednje pomeni, da lahko odvisnost med izbranimi spremenljivkami preučujemo ločeno od robnih porazdelitev.

Za modeliranje različnih problemov se uporabljajo različne funkcije kopula iz različnih družin. Tako se v hidrologiji najpogosteje uporabljajo kopule iz Arhimedove družine (angl. *Archimedean family*), kopule iz družine ekstremnih vrednosti (angl. *extreme value family*), nekoliko manj pogosto pa tudi kopule iz eliptične družine (angl. *elliptical family*), ki pa se npr. v ekonomiji uporablja pogosteje (Grimaldi & Serinaldi, 2006; Genest & Favre, 2007; Ganguli & Reddy, 2013). Za modeliranje suše se najpogosteje uporablja kopula Clayton iz Arhimedove družine (Ma et al., 2011), za analize visokovodnih valov se pogosto uporabijo kopule iz družine ekstremnih vrednosti (npr. kopuli Gumbel-Hougaard ali Galambos) (Genest & Favre, 2007), za geostatistične interpolacije se pogosto uporabljajo normalne kopule iz eliptične družine (Bardossy & Li, 2008).

Simetrične kopule iz Arhimedove družine, ki se najpogosteje uporabljajo, imajo večinoma le en parameter. Ta parameter lahko podobno kot pri univariatnih porazdelitvah ocenimo z različnimi metodami. Uporabimo lahko metodo momentov (Kendallov ali Spearmanov koeficient korelacije), metodo največjega verjetja (angl. *maximum likelihood method*), pseudo metodo največjega verjetja (angl. *pseudo likelihood method*), kjer naj bi bila slednja najprimernejša (Joe, 1997; Nelsen, 2006; Salvadori et al., 2007). V primeru, da obravnavamo vsaj trivariatni primer, lahko analize naredimo tudi z asimetričnimi kopulami, kjer imajo te en parameter več kot simetrične različice kopul (Grimaldi & Serinaldi, 2006; Hofert & Mächler, 2011). Slednje se izkažejo za posebej uporabne v primerih, ko je korelacija med enim parom spremenljivk izrazito večja kot korelaciji med drugima dvema paroma (trivariatni primer).

Podobno kot pri univariatnih porazdelitvenih funkcijah so bili tudi za primer funkcij kopula razviti nekateri statistični testi za primerjavo različnih teoretičnih modelov in izbira najustreznejše kopule. V literaturi prevladujeta dva tipa testov, najpogostejši so testi, ki temeljijo na primerjavi teoretične ter empirične funkcije kopula, nekoliko manj pogosto pa se uporabljajo testi, ki temeljijo na Kendallovi porazdelitveni funkciji (Genest et al., 2009; Kojadinovic et al., 2011). Za kontrolo ustreznosti izbrane kopule pa je priporočljivo uporabiti tudi grafične teste, s katerimi lahko potrdimo ali zavrnilo ugotovitve statističnih testov (Genest & Favre, 2007).

Opisane korake (določitev robnih porazdelitvenih funkcij, ocenjevanje parametrov kopul, izbira najustreznejše kopule z uporabo statističnih ter grafičnih testov) je potrebno izvesti pri vseh analizah s kopulami, nato pa se postopki razlikujejo od namena uporabe

modela. Pri verjetnostnih analizah moramo izračunati različne povratne dobe, pri ocenjevanju vrednosti pa moramo izvesti simulacije. Oba koraka bosta malce podrobneje opisana v naslednjih dveh odstavkih.

Pri multivariatnih verjetnostnih analizah z uporabo funkcij kopula lahko izračunamo različne vrednosti pogojnih (angl. *conditional*) ali skupnih (angl. *joint*) povratnih dob (Salvadori et al., 2007; Graler et al., 2013). Skupni povratni dobi imenovani OR ter AND sta definirani z naslednjima izrazoma (bivariatni primer):

$$\begin{aligned} T_{u,v}^{OR} &= \frac{1}{1-C_{\theta}(u,v)}, \\ T_{u,v}^{AND} &= \frac{1}{1-u-v+C_{\theta}(u,v)}, \end{aligned} \quad (2)$$

kjer sta u in v izbrani robni porazdelitveni funkciji, C_{θ} pa predstavlja teoretično funkcijo kopula (v tem primeru bivariatno). Ker pa gre za analize v več dimenzijah, te definicije povratnih dob seveda niso primerljive s tistimi, ki jih večinoma uporabljamo pri univariatnih analizah. To lastnost kopul bi morda lahko ocenili kot največjo težavo, saj je težko (kar naenkrat) v praktično uporabo vpeljati nov koncept povratnih dob. Tako so se v literaturi že pojavili poskusi določitve t.i. kritične plasti (angl. *critical layer*), kjer gre za matematično definicijo prej omenjenega pojma, ki je (bolj) analogna z običajnimi povratnimi dobami (Salvadori et al., 2011). Npr. pri trivariatnem primeru verjetnostnih analiz visokovodnih valov izhajamo iz izbrane vrednosti verjetnosti in na podlagi te vrednosti s pomočjo simulacij določimo ploskev v kopula prostoru, kjer vse točke, ki ležijo nad to ploskvijo, označujejo kritične dogodke (Salvadori et al., 2011). To ploskev lahko nato z uporabo inverznih porazdelitvenih funkcij transformiramo v realni prostor. Povratna doba OR predstavlja primer, ko se zgodi vsaj ena izmed obravnavanih spremenljivk, povratna doba AND pa primer, ko se hkrati zgodijo vse obravnavane spremenljivke. Iz definicije sledi, da je verjetnost pojava povratne dobe AND precej manjša od verjetnosti pojava povratne dobe OR, kar seveda pomeni, da je povratna doba za AND primer precej večja (obravnavamo multivariatni primer).

Funkcije kopula lahko uporabimo tudi za ocenjevanje vrednosti različnih hidroloških spremenljivk na podlagi podatkov, ki jih imamo na razpolago. Prikazana bo praktična uporaba v primeru ocenjevanja vrednosti koncentracij suspendiranih snovi. Pri modelu, kjer bi vrednosti koncentracij suspendiranih snovi ocenili le na podlagi pretokov, lahko izhajamo iz naslednje enačbe:

$$f_q(ssc) = P\{SSC \leq ssc | Q = q\} = \frac{\partial C_{\theta}(q, ssc)}{\partial q}, \quad (3)$$

kjer moramo z uporabo statističnih ter grafičnih testov ponovno izbrati ustrezno funkcijo kopula C_{θ} in kjer funkcija f vedno zavzame vrednost med 0 in 1. To pomeni, da lahko to funkcijo zamenjamo z naključno generiranimi števili, ki pripadajo enakomerni porazdelitveni funkciji. Tako pridemo do enačbe, kjer imamo le eno neznanko, to je vrednost porazdelitvene funkcije za spremenljivko SSC . V večini primerov, zaradi kompleksnosti parcialnih odvodov ne moremo poiskati analitične rešitve enačbe 3, lahko pa z uporabo numeričnih metod (npr. Newton-ova metoda) najdemo ničle, ki kot že rečeno ležijo med 0 ter 1. Te ničle pa so v bistvu vrednosti porazdelitvenih funkcij, ki jih želimo oceniti. Ker smo generirali veliko število podatkov (npr. 10000), dobimo raztros možnih vrednosti. Nato lahko določimo najverjetnejšo vrednost koncentracije (glede na maksimum gostote verjetnosti izbrane neparometrične funkcije) in oblikujemo tudi empirične intervale zaupanja (npr. 10, 20 ali 50 % intervali). Te mejne vrednosti nato še transformiramo v realni prostor z uporabo inverzne oblike porazdelitvene funkcije. Tudi v tem primeru za ocenjevanje vrednosti koncentracij uporabimo le vrednosti pretokov, to pomeni, da ne

moremo pričakovati veliko boljših rezultatov kot pri preprostejših regresijskih modelih (npr. linearna ali eksponentna funkcija), ki se večinoma uporabljajo za ocenjevanje koncentracij suspendiranih snovi. Zato smo se odločili, da bomo poskušali oceno izboljšati tako, da bomo v analizah upoštevali še vrednosti padavin. Model, kjer bi za ocenjevanje vrednosti koncentracij suspendiranih snovi uporabili tudi vrednosti padavin, ima naslednjo obliko:

$$f_{q,p}(ssc) = P\{SSC \leq ssc | Q = q | P = p\} = \frac{\partial^2 C_{\theta}(q,p,ssc)}{\partial q \partial p} / \frac{\partial^2 C_{\theta}(q,p)}{\partial q \partial p}. \quad (4)$$

Koncept določitve najverjetnejše vrednosti koncentracije suspendiranih snovi je enak kot pri uporabi bivariatnih kopul (enačba 3).

Rezultati uporabe kopul na praktičnih primerih

V naslednjih podpoglavjih je prikazana uporaba funkcij kopula na štirih praktičnih primerih. Prikazani so rezultati bivariatnih ter trivariatnih verjetnostnih analiz visokovodnih valov, rezultati trivariatnih analiz konic pretokov, volumnov valov ter koncentracij suspendiranih snovi (SSC) in rezultati ocenjevanja vrednosti suspendiranih snovi na podlagi podatkov o pretokih in padavinah.

Bivariatne verjetnostne analize visokovodnih valov

Za izvedbo bivariatnih verjetnostnih analiz visokovodnih valov smo uporabili podatke z vodomerne postaje Litija na reki Savi. Uporabili smo 58 vrednosti letnih maksimumov. Če želimo določiti vrednosti volumnov in časov trajanja valov, moramo najprej izločiti bazni odtok. Za izločanje baznega odtoka smo uporabili grafično tro-točkovno metodo (Šraj & Bezak, 2013; Šraj et al., 2013b). Naredili smo bivariatne analize parov podatkov: konica pretoka - volumen vala ($Q-V$), konica pretoka - čas trajanja vala ($Q-D$) ter volumen vala - čas trajanja vala ($V-D$). Za modeliranje konic pretokov in časov trajanja visokovodnih valov smo na podlagi statističnih ter grafičnih testov izbrali log-Pearsonovo III porazdelitev, za opis volumnov valov pa Pearsonovo III porazdelitev. Za ocenjevanje parametrov univariatnih porazdelitev smo uporabili metodo momentov L (Hosking & Wallis, 1997), za izbiro najustreznejših porazdelitvenih funkcij pa različne statistične ter grafične teste (Šraj et al., 2013b). Preglednica 2 prikazuje vrednosti Pearsonovega, Kendallovega ter Spearmanovega koeficienta korelacije za obravnavane pare podatkov. Korelacijski koeficienti določajo odvisnost med dvema spremenljivkama, torej če je vrednost koeficienta blizu vrednosti 1 to pomeni, da sta spremenljivki odvisni, če je koeficient enak 0, to pomeni neodvisnost spremenljivk. Korelacijske koeficiente za opis odvisnosti se lahko uporabi tudi za primer, kjer imamo več kot 2 spremenljivki, vendar se izrazi za izračun korelacijskih koeficientov spremijo in zato je boljše za opis odvisnosti uporabiti parameter kopule, ki ga lahko v bivariatnem primeru ocenimo na podlagi Kendallovega koeficienta korelacije ali pa npr. s pseudo metodo največjega verjetja, ki jo je nato lažje uporabiti v multivariatnem primeru. Od vrednosti korelacije oz. dejanske odvisnosti, ki jo opiše korelacijski koeficient, je odvisno katero funkcijo kopula bomo uporabili za analize. Nekaterne kopule so primerne le za pozitivne vrednosti (npr. Gumbel-Hougaard) odvisnosti, spet druge le za korelacije blizu 0 (npr. Ali-Mikhail-Haq kopula). Torej izračunana vrednost korelacijskega koeficienta določa, katere funkcije kopula lahko uporabimo, nadalje pa tudi kakšen bo odziv modeliranih spremenljivk, ki sta medsebojno (ne)odvisni. Iz preglednice 2 lahko vidimo, da je korelacija med konicami pretokov in časi

trajanja visokovodnih valov negativna. To pomeni, da za izvedbo analize (za ta par) ne moremo uporabiti vseh funkcij kopula, ki smo jih uporabili pri drugih dveh parih, saj so nekatere primerne le za pozitivne vrednosti odvisnosti.

Obravnavali smo različne funkcije iz Arhimedove, eliptične družine ter družine kopul ekstremnih vrednosti. Na podlagi Cramér von-Mises ter Kolmogorov-Smirnov statističnih testov (Genest & Favre, 2007) ter različnih grafičnih testov smo kot najprimernejšo funkcijo določili kopulo Gumbel-Hougaard iz Arhimedove družine kopul. Ta kopula spada tudi v družino kopul ekstremnih vrednosti. Te ugotovitve veljajo za para $Q-V$ ter $V-D$, medtem ko smo pri paru $Q-D$ kot najprimernejšo določili Student-t kopulo iz eliptične družine kopul. Za ocenjevanje parametrov kopul smo uporabili metodo momentov (Kendallov koeficient korelacije). Definicije teoretičnih kopul in izraze, ki povezujejo parametre kopul s koeficienti korelacije (Kendall ali Spearman) lahko najdemo v osnovni literaturi (Nelsen, 2006; Salvadori et al., 2007). Preglednica 3 prikazuje izračunane skupne povratne dobe za OR ter AND primera. Izračunali pa smo tudi univariatne (običajne) povratne dobe za pretoke, volumne ter čase trajanja za mediane in maksimalne vrednosti (preglednica 3).

Preglednica 2: Vrednosti Pearsonovega, Kendallovega ter Spearmanovega koeficienta korelacije za obravnavane pare podatkov

	Pearson	Kendall	Spearman
$Q-V$	0,52	0,39	0,54
$V-D$	0,68	0,48	0,63
$Q-D$	-0,15	-0,08	-0,14

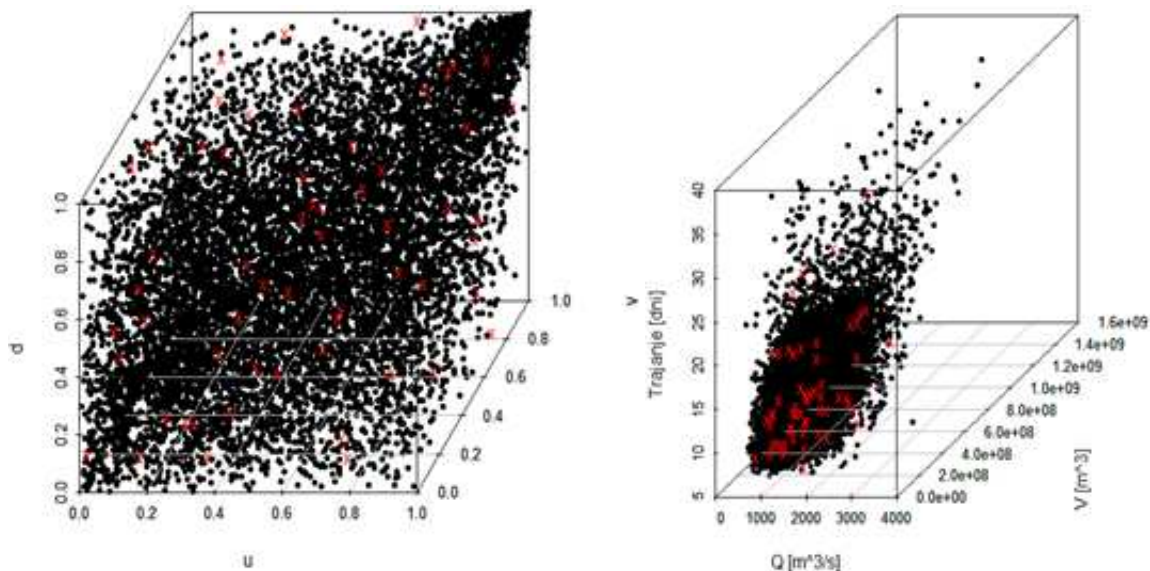
Preglednica 3: Izračunane vrednosti univariatnih in skupnih bivariatnih povratnih dob [leta] za tri funkcije kopula za vsak posamezen par spremenljivk

Univariatni primer Spr.	T^{UNI}		Bivariatni primer					
			Par	Kopula	$T_{u,v}^{OR}$		$T_{u,v}^{AND}$	
	Mediana	Maks.			Mediana	Maks.	Mediana	Maks.
Q	2,0	69,5	$Q-V$	Gumbel	1,2	40,7	2,8	132,8
V	1,9	56,4		Frank	1,5	32,1	2,7	1035,5
D	2,1	70,7		Tawn	1,5	40,7	2,8	133,0
			$V-D$	Gumbel	1,6	43,6	2,8	111,7
				Frank	1,6	32,6	2,7	801,4
				Normal	1,6	36,4	2,8	224,6
$Q-D$				Clayton	1,3	35,2	4,5	5700,3
				AMH	1,3	35,1	4,6	8125,7
				Student-t	1,3	36,8	4,5	708,6

Trivariatna verjetnostna analiza visokovodnih valov

Tudi pri trivariatnih analizah visokovodnih valov smo uporabili podatke o letnih maksimumih z vodomerne postaje Litija na reki Savi. Analizirali smo maksimalne konice pretokov in pripadajoče vrednosti volumnov in časov trajanja visokovodnih valov. To pomeni, da so le konice pretokov maksimalne v vseh letih, volumni ter časi trajanja pa so bili določeni glede na vrednosti pretokov in niso nujno letni maksimumi. Uporabili smo kopule Gumbel-Hougaard, Frank, Joe ter Clayton iz Arhimedove družine in normalno ter

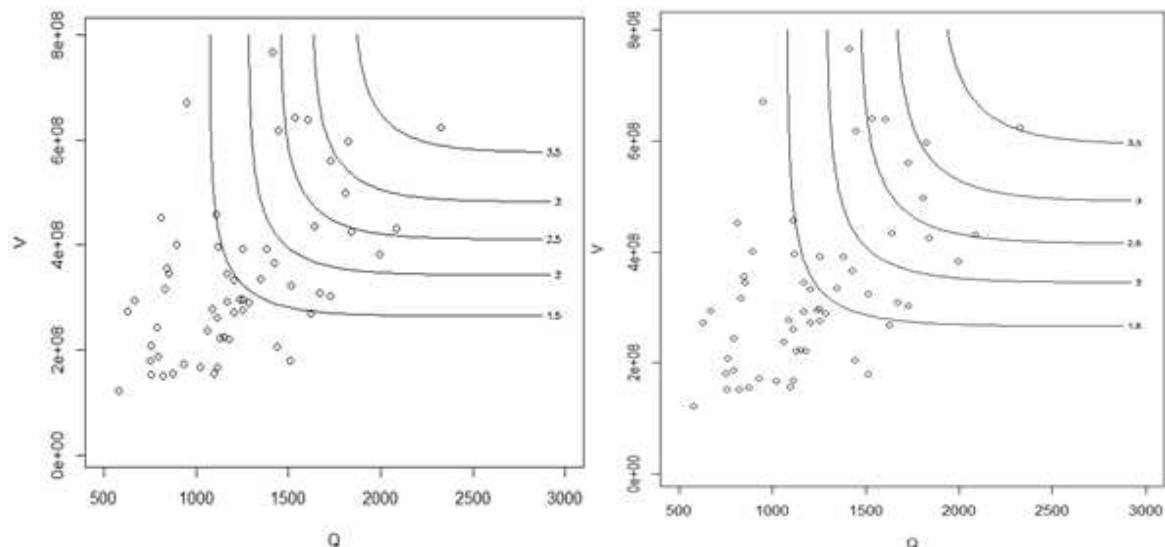
Student-t kopulo iz eliptične družine. Za ocenjevanje parametrov kopul smo uporabili metodo največjega verjetja (Nelsen, 2006; Salvadori et al., 2007), medtem ko so robne porazdelitvene funkcije ostale enake kot pri bivariatnih analizah.



Slika 1: Grafična testa za preverjanje ustreznosti kopule Gumbel-Hougaard pri trivariatni analizi visokovodnih valov Save v Litiji

Slika 1 prikazuje rezultate grafičnega testa za ustreznost kopule Gumbel-Hougaard iz Arhimedove družine kopul, kjer rdeči znaki x označujejo dejanske vrednosti letnih maksimumov, črne oznake pa so generirane vrednosti, ki so bile določene na podlagi popolnoma definiranega modela. Pri izrisu slike 1b smo uporabili tudi robne porazdelitvene funkcije. Na podlagi grafičnih ter statističnih testov (Cramér von-Mises) smo kot najprimernejše določili kopule Gumbel-Hougaard, Frank (Arhimedova družina) ter normalno (eliptična družina). Izračunali smo tudi dve skupni povratni dobi imenovani OR ter AND (enačba 2), ki pa se zaradi tega, ker obravnavamo tri spremenljivke (trivariatni primer) hkrati, malce spremenita (Ganguli & Reddy, 2013)

Slika 2 prikazuje vrednosti pogojne povratne dobe OR, ki označuje primer, ko se zgodi vsaj ena izmed obravnavanih spremenljivk. Prikazani so rezultati za kopuli Frank ter normalno. AND primer pa pomeni, da se hkrati zgodijo vse tri vrednosti spremenljivk. Poleg rezultatov za povratno dobo OR so na sliki 2 prikazane tudi vrednosti letnih maksimumov, ki smo jih uporabili v analizah. Opazimo lahko, da razlika med prikazanima kopulama ni izrazita. Vidimo lahko tudi, da ima večina letnih maksimumov povratno dobo OR manjšo od 1,5 let, le za en dogodek pa je ta večja od 3,5 let, kar si lahko razlagamo tako, da je verjetnost, da se zgodi vsaj ena izmed spremenljivk (konice, volumni ali časi trajanja) relativno velika in posledično je multivariatna vrednost skupne povratne dobe OR dokaj majhna.



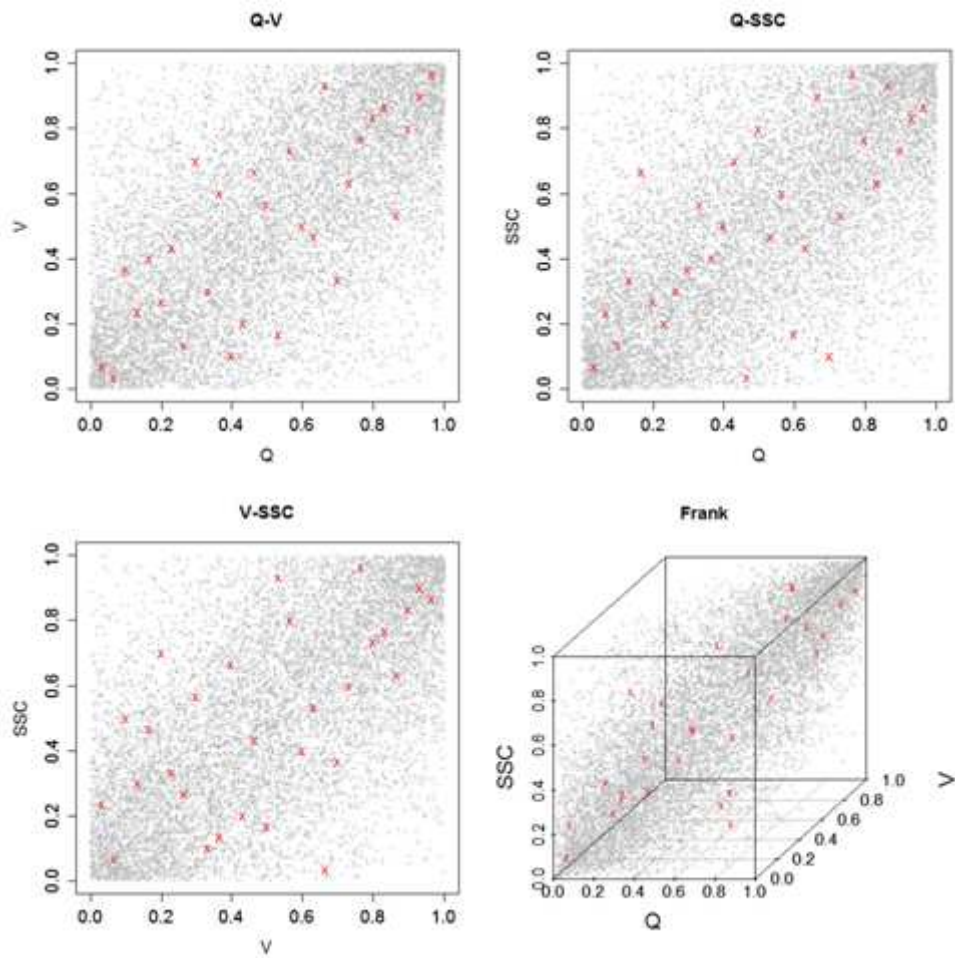
Slika 2: Prikaz izmerjenih podatkov in skupnih povratnih dob (OR primer) za normalno ter Frank kopulo za trajanje visokovodnega vala (D) 15 dni za vodomerno postajo Litija na reki Savi

Trivariatna verjetnostna analiza konic pretokov, volumnov in koncentracij suspendiranih snovi

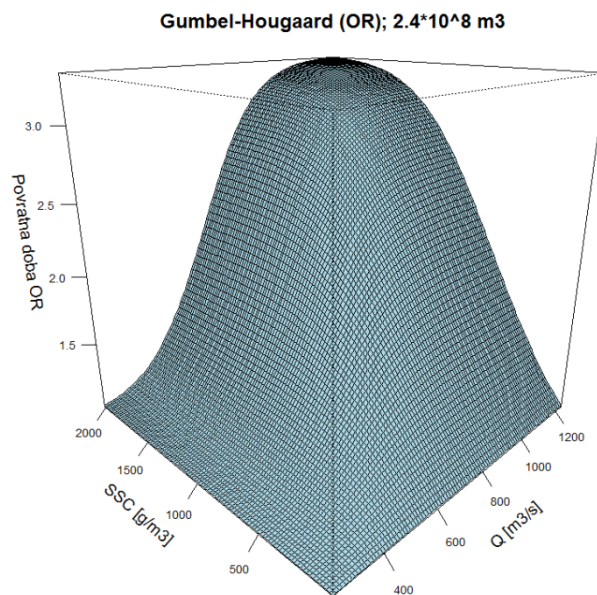
Naslednji problem, ki smo se ga lotili z uporabo funkcij kopula, je analiza konic pretokov (Q), pripadajočih volumnov valov (V) ter pripadajočih vrednosti koncentracij suspendiranih snovi (SSC) (Bezák et al., 2013b). Za izvedbo teh analiz smo uporabili podatke z vodomerne postaje Gornja Radgona na reki Muri (29 vrednosti letnih maksimumov), ki ima najdaljši niz meritev suspendiranih snovi, ki jih je izvajala Agencija RS za okolje (ARSO, 2013). Postopek analize je enak kot pri trivariatnih analizah visokovodnih valov (prejšnje poglavje).

Za izločanje baznega odtoka smo tokrat uporabili dvo-točkovno metodo. Kot robno porazdelitev konic smo izbrali Gumbelovo porazdelitev, za modeliranje volumnov smo določili log-normalno porazdelitveno funkcijo, za opis koncentracij suspendiranih snovi pa generalizirano Pareto porazdelitev. Parametre univariatnih porazdelitev smo ponovno ocenili z uporabo metode momentov L (Hosking & Wallis, 1997). Izračunali smo vrednosti Kendallovega koeficienta korelacije za pare podatkov: Q - V 0,56; Q - SSC 0,50 in V - SSC 0,40. Ker so vse tri vrednosti korelacij statistično značilne (stopnja značilnosti 0,05) in razlike med njimi niso izrazite, smo se odločili za izvedbo verjetnostnih analiz z uporabo simetričnih trivariatnih funkcij kopula (alternativna možnost bi bile asimetrične kopule). Za ocenjevanje parametrov smo tokrat uporabili pseudo metodo največjega verjetja (Salvadori et al., 2007; Ma et al., 2011). Z uporabo grafičnih ter statističnih testov (Cramér-von Mises) smo tudi tokrat poskušali izbrati najustreznejšo kopulo. Sliki 3 in 5 prikazujeta dva grafična testa, kjer so prikazani rezultati za kopulo Frank iz Arhimedove družine kopul. Z rdečimi znaki x so ponovno označeni letni maksimumi, s sivo barvo pa so označeni generirani podatki, ki so bili določeni na podlagi definiranega kopula modela, ki smo ga določili na podlagi podatkov o letnih maksimumih. Z uporabo grafičnih in statističnih testov smo kot najustreznejšo določili kopulo Gumbel-Hougaard iz Arhimedove družine.

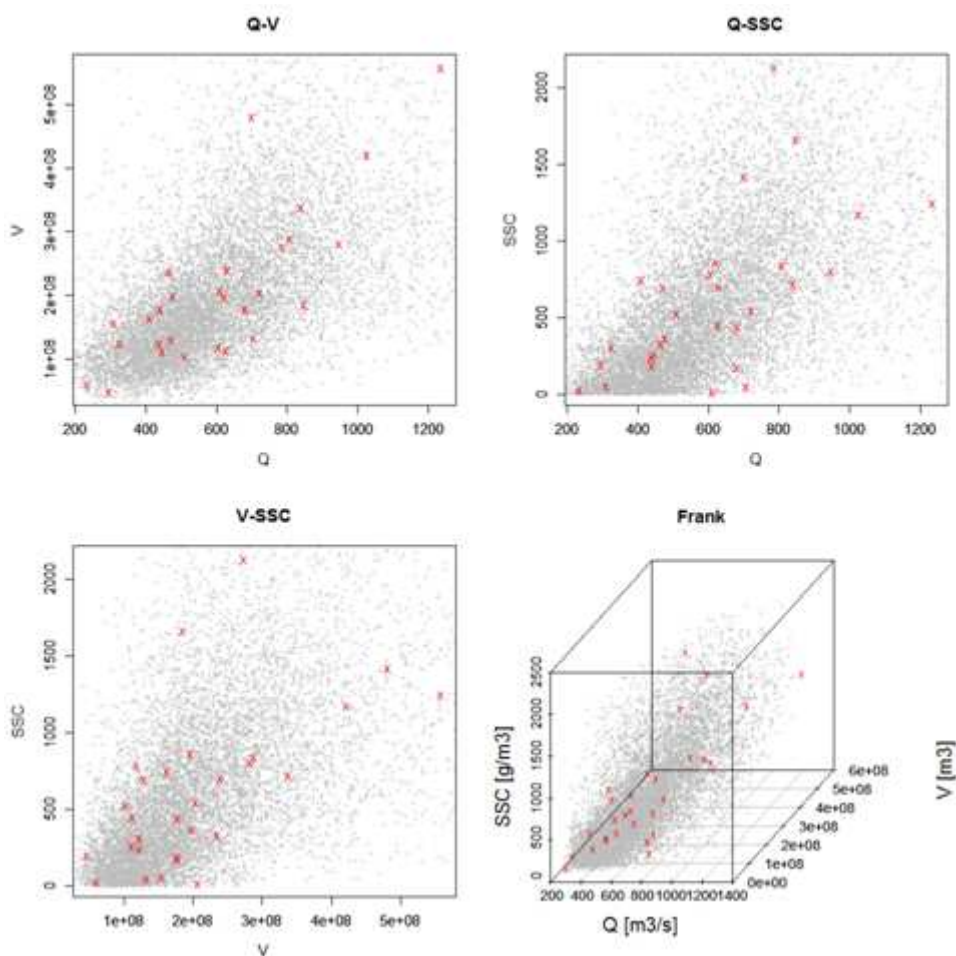
Slika 4 prikazuje rezultate skupne (angl. *joint*) povratne dobe OR za kopulo Gumbel-Hougaard iz Arhimedove družine. Izrisani so rezultati za vrednost volumna visokovodnega vala $2,4 \cdot 10^8 \text{ m}^3$. Pogojna povratna doba OR ponovno označuje primer, ko se zgodi vsaj ena izmed obravnavanih spremenljivk.



Slika 3: Grafični test I za kontrolo ustreznosti kopule Frank za podatke z vodomerne postaje Gornja Radgona na reki Muri



Slika 4: Skupna povratna doba OR za kopulo Gumbel-Hougaard za vrednost pretoka $2,4 \cdot 10^8$ m³ na postaji Gornja Radgona na reki Muri



Slika 5: Grafični test II za kontrolo ustreznosti kopule Frank za podatke z vodomerne postaje Gornja Radgona na reki Muri

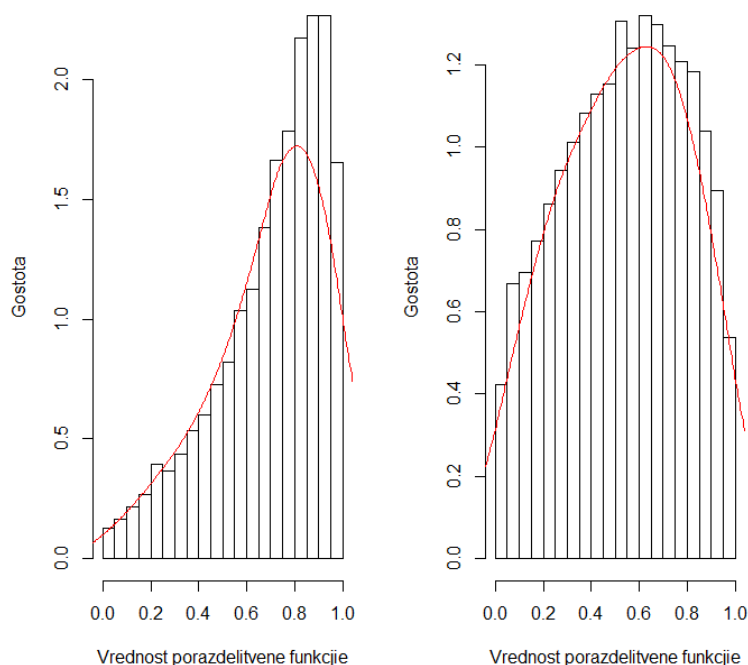
Ocenjevanje vrednosti koncentracij suspendiranih snovi na podlagi vrednosti pretokov in padavin

Kot zadnji praktični primer smo obravnavali model, s katerim lahko ocenimo vrednosti koncentracij suspendiranih snovi (*SSC*) na podlagi vrednosti pretokov in padavin. Ker so meritve koncentracij suspendiranih snovi precej redkejše kot meritve pretokov in padavin, se pogosto pojavi potreba po oceni manjkajočih podatkov. Agencija RS za okolje je v letu 2012 (začasno) prenehala izvajati monitoring te hidrološke spremenljivke (Bezak et al., 2013b). Najpogosteje se za ocenjevanje vrednosti koncentracij suspendiranih snovi uporabljajo različne regresijske funkcije, kot so linearna funkcija, polinomi višjih redov ali eksponentna funkcija. Uporaba teh preprostih modelov je smiselna, ko je korelacija med vrednostmi pretokov in koncentracijami relativno visoka. V primeru, ko je korelacija majhna oz. blizu vrednosti 0, to pomeni, da visoke (nizke) vrednosti pretokov ne pomenijo nujno visokih (nizkih) vrednosti koncentracij suspendiranih snovi. V tem primeru je uporaba regresijskih krivulj vprašljiva. V predstavljenem primeru smo za ocenjevanje manjkajočih *SSC* vrednosti uporabili mesečne vsote treh merjenih spremenljivk (pretoki, koncentracije suspendiranih snovi ter padavine). Možna je uporaba modela tudi z upoštevanjem dnevni serij podatkov, vendar je avtokorelacija v tem primeru izrazitejša, zato je potrebno za opis posameznih robnih porazdelitev uporabiti t.i. časovne modele (angl. *time series model*), kot je npr. ARMA model (angl. *autoregressive moving average model*). Z uporabo teh modelov v bistvu transformiramo izhodiščne dnevne vrednosti

spremenljivk tako, da se znebimo avtokorelacije v podatkih. Pogoji za izvedbo vseh analiz s funkcijo kopula ter tudi za izvedbo verjetnostnih analiz je, da podatki niso avtokorelirani. To pomeni, da vrednost pretoka v nekem dnevu ni odvisna od vrednosti pretokov v prejšnjih dnevih. Zato smo se odločili, da bomo model za ocenjevanje koncentracij suspendiranih snovi najprej poskusili narediti na mesečnih vsotah. Tak model bi npr. lahko uporabili za ocenjevanje masne bilance lebdečih plavin.

Uporaba vrednosti padavin je smiselna, ker smo uporabili podatke z vodomerne postaje Ranca na reki Pesnici, ki ima prispevno površino veliko 84 km^2 , in ker padavinska postaja Polički vrh leži blizu hidrološke postaje, poleg tega pa je korelacija med mesečnimi vsotami padavin in mesečnimi vsotami koncentracij suspendiranih snovi večja kot korelacija med pretoki ter koncentracijami.

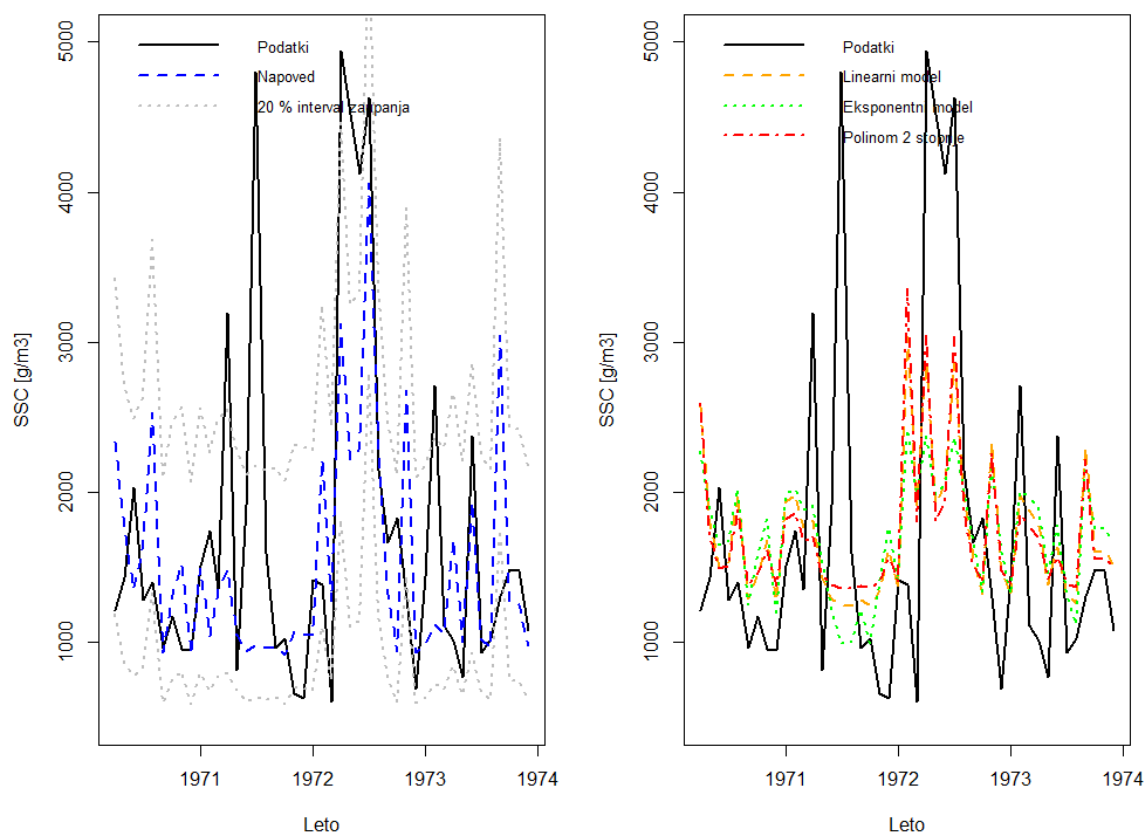
Za izvedbo analize smo uporabili 45 mesečnih vsot pretokov, koncentracij suspendiranih snovi ter mesečnih vsot padavin (preglednica 1). Najprej smo preverili, ali je v vzorcu prisotna avtokorelacija. Z uporabo Box-Pierce testa smo ugotovili, da je v seriji koncentracij suspendiranih snovi prisotna avtokorelacija, smo te podatke najprej transformirali z uporabo Box-Cox transformacije. Za preverjanje stacionarnosti in homogenosti vzorca smo uporabili še Mann-Kendall ter SNHT (angl. *standard normal homogeneity test*) testa. Ugotovili smo, da so vsi vzorci primerni za izvedbo analiz. Z uporabo grafičnih in statističnih testov smo kot robne porazdelitve določili naslednje funkcije: generalizirana porazdelitev ekstremnih vrednosti za pretoke, Gumbelova porazdelitev za padavine ter log-normalno porazdelitev za koncentracije suspendiranih snovi. Za ocenjevanje parametrov smo ponovno uporabili metodo momentov L (Hosking & Wallis, 1997). Z uporabo statističnega testa Cramér-von Mises in grafičnih testov smo kopulo Gumbel-Hougaard določili kot najprimernejšo za izvedbo analize. Nato smo generirali 10000 vzorcev na podlagi enakomerne porazdelitve. Z Newtonovo metodo smo poiskali ničle funkcije, zapisane z enačbo 4.



Slika 6: Primer raztrosa možnih rešitev enačbe 4 v obliki histograma z izrisano neparametrično gostoto verjetnosti

Slika 6 prikazuje raztros možnih rešitev enačbe 4, ki smo jih dobili s simulacijo, na sliki pa je izrisana tudi neparametrična gostota verjetnosti, na podlagi katere smo določili

najverjetnejšo vrednost porazdelitvene funkcije (maksimum neparametrične gostote verjetnosti). Za vsak mesec smo tako dobili tak raztros vrednosti (slika 6), na podlagi teh rezultatov pa smo določili tudi empirične intervale zaupanja.



Slika 7: Primerjava rezultatov kopula modela (levo) za ocenjevanje vrednosti koncentracij suspendiranih snovi z izrisanimi 20 % intervali zaupanja in rezultati običajnih regresijskih modelov (desno)

Primerjava med modelom kopula ter med običajnimi regresijskimi modeli je prikazana na sliki 7. Pri modelu kopula so izrisani tudi empirični intervale zaupanja. Nato smo z uporabo testov primerjali 4 predlagane modele. Odločili smo se za uporabo Nash-Sutcliffe, RMSE (angl. *root mean square error*) ter MAE (angl. *mean absolute error*) testov, poleg tega smo izračunali tudi povprečno razliko med izmerjenimi in napovedanimi (modeliranimi) vrednostmi, preverili pa smo tudi, kakšne so razlike med minimalnimi, maksimalnimi in srednjimi vrednostmi. Rezultati so prikazani v preglednici 4. Opazimo lahko, da pri modelu kopula dobimo boljše rezultate kot pri regresijskih modelih. Rezultati tudi z uporabo modela kopula niso najboljši (slika 7), so pa boljši kot tisti, do katerih bi prišli z uporabo metod, ki se večinoma uporabljajo za ocenjevanje vrednosti koncentracij suspendiranih snovi. Iz preglednice 4 lahko vidimo, da regresijski modeli podcenijo maksimalne vrednosti in precenijo minimalne vrednosti koncentracij. Tudi v tem primeru smo z uporabo modela kopula dosegli izboljšanje rezultatov. Cilj je dobiti čim manjše vrednosti testov RMSE in MAE, ter vrednost testa NSE čim bližje vrednosti 1. Regresijski modeli so dali boljše rezultate le pri oceni povprečne vrednosti, kar pa je verjetno posledica dejstva, da je varianca teh modelov manjša.

Preglednica 4: Primerjava rezultatov različnih modelov za ocenjevanje vrednosti koncentracij suspendiranih snovi

	Podatki	Model kopula	Linearni model	EkspONENTNI model	Polinom 2. stopnje
Nash-Sutcliffe	1	0,22	0,16	0,08	0,16
RMSE	0	1015	1053	1101	1048
MAE	0	682	773	798	762
Povprečna razlika [%]	0	6,7	28,6	29,5	28,6
Minimalna napoved [g/m ³]	604	918	1243	997	1364
Srednja napoved [g/m ³]	1724	1520	1725	1712	1725
Maksimalna napoved [g/m ³]	4939	4058	3037	2421	3360

Analiza rezultatov in razprava

Pri bivariatnih verjetnostnih analizah visokovodnih valov smo ugotovili, da razlike med posameznimi kopulami iz Arhimedove družine kopul, kot so Gumbel-Hougaard ali Frank, niso izrazite (Šraj et al., 2013a). Za modeliranje para konice pretokov in pripadajočih volumnov visokovodnih valov smo izbrali kopulo Gumbel-Hougaard, kar je v skladu z ugotovitvami nekaterih drugih raziskovalcev (Poulin et al., 2007; Karmakar & Simonovic, 2009). Uporaba kopul, ki nimajo poudarka na zgornjem robu porazdelitve (angl. *upper tail dependence*), kar je recimo kopula Frank iz Arhimedove družine kopul, lahko pripelje do tega, da podcenimo vrednosti, ki pripadajo določeni povratni dobi (Poulin et al., 2007). Pri primerjavi univariatnih (običajnih) in skupnih bivariatnih povratnih dob smo ugotovili (Šraj et al., 2013a), da velja zveza $T_{u,v}^{OR} < T_{u,v}^{UNI} < T_{u,v}^{AND}$, kar je v skladu z ugotovitvami drugih raziskovalcev (Salvadori et al., 2007).

Pri trivariatnih verjetnostnih analizah smo uporabili različne v literaturi pogosto uporabljene kopule (Ma et al., 2011; Ganguli & Reddy, 2013). Z uporabo statističnih ter v literaturi pogosto uporabljenih grafičnih testov (Genest & Favre, 2007) smo določili najustreznejšo funkcijo kopula. Ponovno smo izbrali kopulo Gumbel-Hougaard. Izračunali pa smo tudi skupni povratni dobi OR ter AND (Ganguli & Reddy, 2013). Tudi pri trivariatnem primeru smo ugotovili, da velja zveza $T_{u,v,d}^{OR} < T_{u,v,d}^{UNI} < T_{u,v,d}^{AND}$. Vrednosti povratnih dob so bile odvisne od lastnosti funkcij kopul, predvsem od obnašanja na zgornjem robu porazdelitve (Poulin et al., 2007).

Pri trivariatnih verjetnostnih analizah konic pretokov, volumnov valov ter vrednosti koncentracij suspendiranih snovi smo uporabili enak postopek kot pri trivariatnih analizah visokovodnih valov. Z uporabo grafičnih in statističnih testov (Genest & Favre, 2007; Genest et al., 2009; Kojadinovic et al., 2011) smo določili kopulo Gumbel-Hougaard kot najprimernejšo za izvedbo analiz. To kopulo smo izbrali tudi zaradi dejstva, ki smo ga upoštevali že pri bivariatnih analizah, in sicer da lahko nekatere kopule podcenijo ocenjene vrednosti spremenljivk, ki pripadajo izbrani skupni povratni dobi (Poulin et al., 2007). Predstavljena metodologija se je izkazala za uporabno, predvsem pa bi rezultati verjetnostnih analiz morali biti bolj zanesljivi, saj v analizah hkrati upoštevamo tri spremenljivke, medtem ko pri običajnih univariatnih verjetnostnih analizah upoštevamo le eno spremenljivko.

Ocenjene vrednosti koncentracij suspendiranih snovi, ki so rezultat kopula modela, bi lahko označili za bolj zanesljive od tistih, ki jih lahko izračunamo z regresijskimi modeli (slika 7), saj poleg vrednosti pretokov pri ocenjevanju upoštevamo tudi vrednosti padavin, kar predstavlja izboljšavo do sedaj v praksi največkrat uporabljenih metod. Postopek izvedbe analize je enak kot pri prvih treh problemih, ki smo jih obravnavali, s to razliko, da nas tukaj ne zanimajo povratne dobe, ampak želimo izvesti simulacije z uporabo enačbe 4. Teoretično bi lahko v model vpeljali še več spremenljivk in tako poskušali izboljšati končne rezultate modela, vendar bi v tem primeru postopek vseboval dodatne korake (Salvadori et al., 2007). Število uporabljenih parametrov pri kopula modelu je nekoliko večje od števila parametrov pri regresijskih modelih, saj moramo oceniti parameter kopule, ki opisuje odvisnost (vse uporabljene kopule imajo 1 parameter, obstajajo pa tudi kompleksnejše kopule z več parametri), poleg tega pa ocenjene vrednosti določa tudi pasovna širina (angl. *bandwidth*) izbrane neparametrične porazdelitvene funkcije. Določiti pa je potrebno še parametre porazdelitvenih funkcij, ki opisujejo modelirane spremenljivke.

Zaključki

V prispevku smo prikazali uporabo funkcij kopula na štirih praktičnih primerih. Obravnavali smo bivariatne ter trivariatne analize visokovodnih valov reke Save v Litiji, naredili smo trivariatne verjetnostne analize konic pretokov, volumnov valov ter koncentracij suspendiranih snovi, za kar smo uporabili podatke s postaje Gornja Radgona na reki Muri. Poleg tega smo definirali model, s katerim lahko na podlagi vrednosti pretokov ter padavin ocenimo vrednosti lebdečih plavin (*SSC*) in obenem tudi definiramo empirične intervale zaupanja. Za izgradnjo tega modela smo sicer uporabili mesečne vsote hidroloških spremenljivk, vendar bi bila aplikacija možna tudi na dnevni vrednostih, kar pa presega okvir tega prispevka.

Med izvajanjem analiz smo ugotovili, da je za hidrološke probleme v večini primerov najprimernejša kopula Gumbel-Hougaard, ki spada tako v Arhimedovo družino kot v družino kopul ekstremnih vrednosti. Za ocenjevanje parametrov kopul smo uporabili različne metode, kot so metoda momentov, metoda največjega verjetja ter pseudo metoda največjega verjetja. Pri verjetnostnih analizah smo izračunali skupne povratne dobe imenovane OR ter AND, pri ocenjevanju parametrov pa smo s simulacijami ocenili vrednosti koncentracij suspendiranih snovi. Ker smo model za ocenjevanje *SSC* vrednosti definirali sami, rezultatov analiz nismo mogli primerjati z rezultati iz literature, ampak le z rezultati regresijskih modelov, kjer smo ugotovili, da je predstavljena metodologija za ocenjevanje vrednosti koncentracij suspendiranih snovi primerna. Možna pa bi bila tudi razširitev modela, tako da bi upoštevali dnevne vrednosti spremenljivk.

Uporaba funkcij kopula na praktičnih primerih je pokazala, da so kopule uporabno matematično orodje, ki ga lahko uporabimo pri številnih praktičnih problemih v hidrologiji, kjer nastopa več medsebojno bolj ali manj odvisnih spremenljivk.

Zahvala

Zahvaljujemo se Agenciji RS za okolje za posredovane podatke, ki smo jih uporabili za izvedbo analiz. Del rezultatov raziskave je nastal v okviru temeljnega raziskovalnega projekta J2-4096, ki ga financira Javna agencija za raziskovalno dejavnost Republike Slovenije. Del rezultatov pa je prispevek dela UL FGG na mednarodnem raziskovalnem

projektu SedAlp, ki ga financira Evropska unija v okviru programa Alpine Space. Zahvaljujemo se tudi obema recenzentoma za njune koristne komentarje.

Literatura

- ARSO. 2013. Podatki, Hidrološki arhiv, Arhiv površinskih voda. Dostopno na: http://vode.arso.gov.si/hidarhiv/pov_arhiv_tab.php.
- Bardossy, A., Li, J. (2008). Geostatistical interpolation using copulas, *Water resources research* 44(7).
- Bezak, N., Brilly, M., Šraj, M. (2013a). Comparison between the peaks over threshold method and the annual maximum method for flood frequency analyses, *Hydrological Science Journal*, doi: 10.1080/02626667.2013.831174.
- Bezak, N., Šraj, M., Mikoš, M. (2013b). Pregled meritev vsebnosti suspendiranega materiala v Sloveniji in primer analize podatkov, *Gradbeni vestnik* (v tisku).
- De Michele, C., Salvadori, G., Canossi, M., Petaccia, A., Rosso, R. (2005). Bivariate Statistical Approach to Check Adequacy of Dam Spillway, *Journal of Hydrological Engineering* 10(1), 50-57.
- Favre, A. C., Aldouni, E.L., Perreault L., Thiemonge, N., Bobee, B. (2004). Multivariate hydrological frequency analysis using copulas, *Water resources research* 40(1).
- Ganguli, P., Reddy, M. J. (2013). Probabilistic assessment of flood risks using trivariate copulas, *Theoretical and Applied Climatology* 111(1-2), 341-360.
- Genest, C., Favre, A. C. (2007). Everything you always wanted to know about copula modelling but were afraid to ask, *Journal of Hydrological Engineering* 12, 347-368.
- Genest, C., Remillard, B., Beaudoin, D. (2009). Goodness-of-fit tests for copulas: A review and a power study, *Insurance: Mathematics and Economics* 44, 199-213.
- Grimaldi, S., Serinaldi, F. (2006). Asymmetric copula in multivariate flood frequency analysis, *Advances in Water Resources* 29(8), 1155-1167.
- Graler, B., van den Berg, M. J., Vandenberghe, S., Petroselli, A., Grimaldi, S., De Baets, B., Verhoest, N. E. C. (2013). Multivariate return periods in hydrology: a critical and practical review focusing on synthetic design hydrograph estimation, *Hydrology and Earth System Sciences* 17(4), 1281-1296.
- Hofert, M., Mächler, M. (2011). Nested Archimedean Copulas Meet R: The nacopula Package, *Journal of Statistical Software* 39(9).
- Hosking, J. R. M., Wallis, J. R. (1997). *Regional frequency analysis: an approach based on L-moments*. Cambridge, Cambridge University Press, 224 p.
- Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 424 p.
- Karmakar, S., Simonovic, S. P. (2009). Bivariate flood frequency analysis. Part 2: a copula based approach with mixed marginal distributions, *Journal of Flood Risk Management* 2, 32-44.
- Kojadinovic, I., Yan, J., Holmes, M. (2011). Fast large-sample goodness-of-fit for copulas, *Statistica Sinica* 21(2), 841-871.
- Ma, M., Song, S., Ren, L., Jiang, S., Song, J. (2011). Multivariate drought characteristics using trivariate Gaussian and Student t copulas, *Hydrological Processes* 27(8), 1175-1190.
- Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer Series in Statistics, 272 p.
- Poulin, A., Huard, D., Favre, A. C., Pugin, S. (2007). Importance of tail dependence in bivariate frequency analysis, *Journal of Hydrological Engineering* 12, 394-403.
- Salvadori, G., De Michele, C. (2004). Frequency analysis via copulas: Theoretical aspects and applications to hydrological events, *Water resources research* 40(12).
- Salvadori, G., De Michele, C., Durante, F. (2011). On the return period and design in a multivariate framework, *Hydrology and Earth System Sciences* 15, 3293-3305.

- Salvadori, G., De Michele, C., Kottegoda, N., Rosso, R. (2007). *Estremes in nature: An Approach Using Copulas*. Water Science and Technology Library, 292 p.
- Shiau, J. T., Feng, S., Nadaraiyah, S. (2007). Assessment of hydrological droughts for the Yellow River, China, using copulas, *Hydrological Processes* 21(16), 2157-2163.
- Sklar, A. (1959). Fonctions de repartition a n dimensions et leurs marges, *Publ. Inst. Statist. Univ. Paris* 8, 229-231.
- Šraj, M., Bezak, N. (2013). Analiza visokovodnih valov Save v Litiji, *Ujma* 27, 228-235.
- Šraj, M., Bezak, N., Brilly, M. (2013a). Bivariate flood frequency analyses using copula function, *European Geosciences Union* 2013.
- Šraj, M., Bezak, N., Brilly, M. (2013b). Vpliv izbire metode na rezultate verjetnostnih analiz konic, volumnov in trajanj visokovodnih valov Save v Litiji, *Acta Hydrotechnica* (v tisku).
- Wong, G., Lambert, M. F., Leonard, M., Metcalfe, A. V. (2010). Drought Analysis Using Trivariate Copulas Conditional on Climate States, *Journal of Hydrological Engineering* 15, 129-141.
- Zhang, L., Singh, V. P. (2007). Bivariate rainfall frequency distributions using Archimedean copulas, *Journal of Hydrology* 332(1-2), 93-109.