Quaderni di Statistica
Vol. 5, 2003

Analisi statistiche bivariate di serie idrologiche

Salvatore Grimaldi
Istituto di Ricerca per la Protezione Idrogeologica IRPI-CNR
E-mail:salvatore.grimaldi@irpi.cnr.it

Francesco Serinaldi
Dipartimento di Idraulica Trasporti e Strade,
Università degli Studi di Roma "La Sapienza"

Summary: In this paper main approaches used in hydrology to develop statistical bivariate analyses are described. Among methodologies present in literature, the method based on the Copula, recently introduced in hydrology, is accurately explained. In order to comment and to verify several theoretical and methodological problems some synthetic simulation and a case study are shown.

Keywords: Bivariate distributions, Copula, Linear Parametrical Models.

1. Introduzione

Uno dei principali obiettivi delle analisi idrologiche è determinare la probabilità di accadimento di alcune grandezze caratteristiche, utili per la progettazione delle opere idrauliche (Moisello, 1999).

Nonostante tali informazioni siano presenti nelle registrazioni disponibili, nella pratica si preferisce inferire statisticamente solamente su una di esse e ricavare indirettamente le altre. Ciò per una evidente semplicità delle analisi univariate rispetto a quelle multivariate.

Determinare la probabilità di accadimento di due variabili in maniera congiunta è già di estremo interesse nelle analisi idrologiche.

Nel paragrafo 2 vengono descritti alcuni approcci utilizzati in ambito idrologico evidenziandone i limiti teorici ed applicativi.

Negli ultimi anni, la ricerca si è indirizzata su un nuovo strumento di analisi, definito *copula* o *funzione di dipendenza*, conosciuta nell'ambito della teoria degli spazi metrici probabilistici, ed introdotta solo recentemente in idrologia da De Michele e Salvadori (2003).

Nel presente lavoro, precisamente nel paragrafo 3 e 4, si descrive in dettaglio tale approccio. Si verificano gli effetti pratici dell'impiego di misure di dipendenza diverse da quelle lineari maggiormente utilizzate. Si descrivono quali siano gli effetti di differenti metodi di campionamento di serie idrologiche. Infine, si affronta il problema della numerosità del campione disponibile verificando se sia utile impiegare generatori di serie sintetiche.

2. Metodologie per le analisi statistiche idrologiche bivariate

Le prime applicazioni di analisi bivariate in ambito idrologico risalgono alla metà degli anni ottanta (Hashino, 1985) quando la distribuzione esponenziale bivariata di Freund (1961) viene generalizzata ed utilizzata per rappresentare la distribuzione di probabilità congiunta di due variabili idrologiche.

In seguito, Correia (1987) suggerisce una distribuzione congiunta assumendo che le grandezze analizzate siano distribuite esponenzialmente e che la distribuzione condizionata sia normale. Il limite che emerge è l'assunzione di normalità per le funzioni di densità condizionata. Essa risulta infatti poco generalizzabile a causa della natura delle grandezze idrologiche studiate.

La distribuzione normale bivariata ampiamente studiata in letteratura, è facilmente applicabile, ma il suo impiego presuppone la normalità delle distribuzioni marginali. Tale condizione è ottenuta da Sackl e Bergmann (1987) attraverso una preventiva trasformazione dei dati. Successivamente, Goel et al. (1998) perfezionano la tecnica di normalizzazione sviluppando una procedura a due passi tramite la

trasformazione di Box e Cox (1964). La metodologia è riproposta da Yue (2000a) introducendo, per la stima dei parametri della trasformazione Box-Cox, l'ottimizzazione di una funzione di verosimiglianza.

Tuttavia, le suddette trasformazioni non sempre assicurano che le serie trasformate seguano un distribuzione normale e soprattutto che non vengano alterate alcune proprietà statistiche. Per questo motivo altre distribuzioni bivariate con marginali non normali sono state oggetto di ricerca. Singh e Singh (1991) derivano una funzione di densità di probabilità bivariata con marginali esponenziali applicata alla descrizione della distribuzione congiunta di coppie di variabili idrologiche. Bacchi et al. (1994) propongono un altro schema, introdotto da Gumbel (1960b), con marginali esponenziali,

$$H(x,y) = 1 - e^{-\alpha x} - e^{-\beta y} + e^{-\alpha x - \beta y - \alpha\beta\delta xy} \qquad (1)$$

valida per $x \geq 0$, $y \geq 0$ e $0 \leq \delta \leq 1$.

Osservato che gli eventi idrologici estremi possono essere rappresentati da una distribuzione di Gumbel (1958), Yue et al. (1999) e Yue (2000b, 2001) propongono l'utilizzo di due modelli di distribuzione bivariata con marginali Gumbel (Gumbel, 1958, 1960a; Gumbel e Mustafi, 1967; Tiago de Oliveira, 1975) sintetizzati in *Tabella 1*.

Tabella 1. Distribuzioni mista e logistica di Gumbel.

	Distribuzione bivariata H(x,y)	Parametro $\alpha(\rho)$	Intervallo per ρ
Modello misto	$\exp\left[-\left(e^{-x}+e^{-y}\right)+\alpha\left(e^{x}+e^{y}\right)^{-1}\right]$	$2\left(1-\cos\sqrt{\frac{\pi^2}{6}}\rho\right)$	$[0,\frac{2}{3}]$
Modello logistico	$\exp\left[-\left(e^{-\alpha x}+e^{-\alpha y}\right)^{1/\alpha}\right]$	$(1-\rho)^{-1/2}$	[0,1]

Gli approcci descritti presentano le seguenti limitazioni:

- i modelli sintetizzati in *Tabella 1* non sono in grado di descrivere grandezze con correlazione negativa tipica di alcune coppie di variabili idrologiche;

- tutti i modelli prevedono marginali uguali, mentre nella pratica le coppie di variabili idrologiche analizzate congiuntamente possono presentare distribuzioni marginali differenti;
- infine, nell'adozione della distribuzione bivariata Normale si incorre negli inconvenienti propri della normalizzazione.

3. La Copula

Definizione 1. (Nelsen, 1999) Una *copula bidimensionale* (o *2-copula*, o brevemente, *copula*) è una funzione bivariata $C: [0,1]\times[0,1] \rightarrow [0,1]$ (*Figura 1*) con le seguenti proprietà:

1. C(u,v) è non decrescente in ogni argomento u, v.
2. Per ogni $u, v \in [0,1]$,
   $$C(u,0) = C(0,v) = 0, \quad C(u,1) = u, \quad C(1,v) = v.$$
3. Per ogni $u_1, u_2, v_1, v_2 \in [0,1]$ tali che $u_1 \leq u_2$ e $v_1 \leq v_2$,
   $$C(u_2,v_2) - C(u_1,v_2) - C(u_2,v_1) + C(u_1,v_1) \geq 0.$$

Nel caso in cui u e v siano indipendenti la copula è $C(u,v) = u\times v$. Per ogni copula C ed ogni $(u,v) \in [0,1]^2$ è possibile scrivere:

$$W(u,v) = \max(u + v - 1, 0) \leq C(u,v) \leq \min(u,v) = M(u,v),$$

in cui M e W sono detti rispettivamente, limite superiore e limite inferiore di Frechét-Hoeffding. Ciò significa che ogni copula è compresa tra le due copule limite.

Una terza copula di importanza fondamentale è la copula prodotto $\Pi(u,v) = u\times v$. Dette U e V due variabili casuali con copula C(u,v), U = V se e solo se C(u,v) = M(u,v), U = 1–V se e solo se C(u,v) = W(u,v) e U e V sono statisticamente indipendenti se e solo se $C(u,v) = \Pi(u,v)$. W e M sono, dunque, le funzioni di distribuzioni bivariate dei vettori (U,1–U) e (U,U) rispettivamente: W descrive una dipendenza negativa perfetta (co-monotonia), e M una dipendenza positiva perfetta (contro-monotonia), intendendo, per dipendenza perfetta, la forma più stretta assunta dalla correlazione ovvero da un'altra misura di dipendenza.

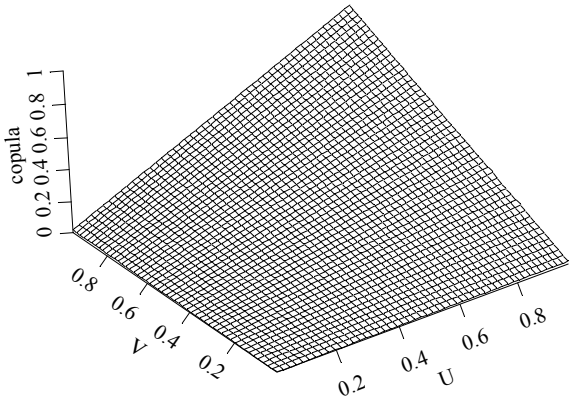

Figura 1. Esempio di copula bidimensionale.

Per l'utilizzo della copula si fa riferimento al seguente teorema e al corollario di Sklar (Nelsen, 1999):

Teorema 1. Sia H una funzione di distribuzione congiunta con marginali F e G. Allora esiste una copula C tale che per ogni x,y ∈ **R**,

$$H(x,y) = C(F(x),G(y)) \tag{2}$$

Se F e G sono continue, allora C è unica; altrimenti C è unicamente determinata in RanF×RanG[1].

Viceversa se C è un copula e F e G sono funzioni di distribuzione, allora la funzione H definita dalla relazione precedente è una funzione di distribuzione congiunta con marginali F e G.

Corollario 1. Sia H(x,y) una distribuzione bivariata di due variabili X, Y con distribuzioni marginali F(x) e G(y) continue. Siano $F^{(-1)}$ e $G^{(-1)}$ le inverse delle distribuzioni marginali. Allora, esiste un'unica copula C : [0,1]×[0,1] → [0,1] tale che ∀ (u,v) ∈ [0,1]×[0,1]

$$C(u,v) = H(F^{(-1)}(u),G^{(-1)}(v)). \tag{3}$$

[1] Ran indica il campo di variazione (range) di una variabile ovvero di una funzione (il suo codominio).

Tali enunciati comportano che la procedura di inferenza bivariata può essere decomposta nella definizione delle singole distribuzioni marginali e della copula. In *Appendice 1* è descritto un semplice esempio per una migliore comprensione dei due enunciati. Si evidenzia la possibilità di estrarre una copula da una data distribuzione multivariata ed usarla per legare insieme due distribuzioni univariate scelte arbitrariamente ottenendo una nuova funzione di ripartizione.

In *Appendice 2* vengono riportate, come esempio, le copule di Gumbel-Hougaard, Clayton-Pareto, Ali-Mikhail-Haq e Frank.

Il parametro di dipendenza α, descritto in *Appendice 2*, si esprime tramite le misure di correlazione di rango τ_k di Kendall e ρ_S di Spearman (Nelsen, 1999; Embrechts et al., 1999, 2002, 2003). In particolare per la τ_k di Kendall vale la seguente relazione (Nelsen, 1999):

$$\tau_K = 4 \iint_{I^2} C(u,v) \cdot dC(u,v) - 1 \tag{4}$$

in cui I^2 è il quadrato unitario $[0,1]\times[0,1] = [0,1]^2$, e l'integrale doppio è il valore atteso della funzione C(u,v).

Per la ρ_S di Spearman vale la relazione (Nelsen, 1999):

$$\rho_S = 12 \iint_{I^2} C(u,v) \cdot dudv - 3 = 12 \iint_{I^2} uv \cdot dC(u,v) - 3 \tag{5}$$

dove I^2 è il quadrato unitario $[0,1]^2$, e l'integrale doppio è il valore atteso della variabile prodotto $U\times V$.

Quindi, note le stime campionarie di τ_K e ρ_S, risolvendo l'integrale doppio presente nelle (4) e (5), si dispone di una relazione la cui unica incognita è il parametro di dipendenza della copula.

Per le applicazioni idrologiche, ci si riferisce alle famiglie di copule appartenenti alla classe *archimedea* definite dalla forma generale

$$C(u,v) = \varphi^{-1}(\varphi(u) + \varphi(v)) \tag{6}$$

in cui φ è una funzione continua, detta generatore della copula, strettamente decrescente e convessa con dominio $\mathbf{I} = [0,1]$ e codominio $[0,\infty)$.

Per una copula *archimedea*, la relazione (4) assume la forma (Nelsen, 1999):

$$\tau_K = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt\,. \tag{7}$$

Le ragioni per le quali tali copule trovano ampio impiego sono la semplicità con cui possono essere costruite, il fatto che molte famiglie parametriche appartengono a questa classe e la grande varietà di strutture di dipendenza rese disponibili.

Individuare la copula che, per una famiglia monoparametrica, meglio si adatta ai dati, vuol dire stimare il parametro di dipendenza.

Un metodo di stima non parametrico del parametro di dipendenza, definita *a un passo*, è suggerita da Genest e Rivest (1993). La procedura consiste nell'utilizzare la (7). Essendo $\varphi(t)$ funzione di α, e la correlazione di rango τ_K stimabile dal campione disponibile si perviene alla stima di α. Per alcune copule l'integrale (7) assume un'espressione semplice (si veda *Appendice 2*) da cui segue un'equazione analiticamente invertibile, per altre copule la soluzione si ottiene con metodi iterativi.

Alternativamente si può applicare un approccio definito *a due passi* (De Matteis, 2001). Il primo passo consiste nella stima delle distribuzioni marginali delle variabili X e Y compiuta usando delle distribuzioni empiriche $F_n(x_i)$ e $G_n(y_i)$ del campione disponibile (x_i,y_i) per $i = 1,\ldots,n$. Il secondo passo prevede la stima di α massimizzando una funzione di massima verosimiglianza. Dalla funzione di densità congiunta, ottenuta derivando la distribuzione congiunta $H(x,y) = C(F(x),G(y))$,

$$h(x,y) = f(x)g(y)C_{12}(F(x),G(y)), \tag{8}$$

in cui

$$C_{12}(u,v) = \frac{\partial^2}{\partial u \partial v}C(u,v), \tag{9}$$

$f(x)$, $g(y)$ sono funzioni di densità e $u = F(x)$, $v = G(y)$ sono funzioni di ripartizione, si ricava (Mood et al., 1988, pag. 285) la funzione di verosimiglianza

$$L(\alpha; u, v) = \prod_{i=1}^{n} C_{12}(u_i, v_i) \tag{10}$$

con $u_i = F_n(x_i)$ e $v_i = G_n(y_i)$.

Eseguita la stima del parametro α per ogni famiglia di copule *archimedee*, occorre analizzare quale di esse sia più appropriata alla descrizione dei dati.

Un test di adattamento è basato sul teorema e sul corollario riportati di seguito (Nelsen, 1999):

Teorema 2. Sia C una copula generata da un generatore φ. K_C denoti una C-misura[2] dell'insieme $\{(u,v) \in \mathbf{I^2} \mid C(u,v) \leq t\}$. Allora per ogni $t \in \mathbf{I}$,

$$K_C(t) = t - \frac{\varphi(t)}{\varphi'(t^+)}. \tag{11}$$

Corollario 2. Siano U e V due variabili casuali uniformi la cui funzione di distribuzione congiunta è una copula *archimedea* C generata da φ. Allora la funzione di distribuzione K_C è la funzione di distribuzione della variabile C(u,v).

Per il *Teorema 2* la copula C è univocamente determinata attraverso la funzione K_C. Per il *Corollario 2*, i valori assunti da K_C hanno distribuzione uniforme standard. Una copula stimata si adatta al campione quando in un QQ-plot ($K_{C(F(x),G(y))}$, quantili uniformi standard) si ha un andamento approssimativamente lineare. La valutazione di tale adattamento si esegue calcolando i p-valori relativi a un test di Kolmogorov-Smirnov e a un test del χ^2, confrontando poi i valori per entrambi i test.

[2] Per una copula *archimedea* si può definire una C-misura della regione in I^2 che giace su, o al di sotto e a sinistra di ogni curva di livello. Essa rappresenta la probabilità

$$P[(u,v) \in I^2 \mid C(u,v) \leq t].$$

4. Verifiche, analisi e applicazioni

In questo paragrafo vengono presentate alcune analisi volte ad evidenziare aspetti e caratteristiche delle procedure per l'inferenza bivariata ed in particolare della copula. In primo luogo, tramite un confronto tra coppie sintetiche generate con distribuzione nota, si vuole verificare quanto influisca nella pratica usare un parametro di dipendenza non lineare. La seconda analisi proposta è relativa al metodo di campionamento delle coppie nel caso di serie di precipitazioni. Vengono presentati tre metodi e si confrontano i risultati ottenuti. L'ultima applicazione riguarda il problema della numerosità. Si è verificato se ampliando il campione disponibile con un generatore di serie sintetiche si migliora la stima della distribuzione congiunta.

4.1 Confronto tra parametri di dipendenza basati sulla correlazione lineare e sulla concordanza

La possibilità di indagare sugli effetti pratici dell'uso di stimatori del parametro di dipendenza della copula basati sulle correlazioni di rango, piuttosto che sulla correlazione lineare, nasce dall'osservazione che la distribuzione logistica bivariata di Gumbel (si veda *Tabella 1*), può essere riguardata come la copula *archimedea* di Gumbel-Hougaard (si veda *Appendice 2*). Per tale distribuzione sono disponibili, quindi, sia uno stimatore del parametro di dipendenza α basato sul coefficiente di correlazione lineare delle variabili sia uno stimatore funzione della correlazione di rango τ di Kendall.

Si è quindi voluto appurare quali siano le differenze che emergono quando si utilizzano i due metodi di stima del parametro α utilizzando campioni non-gaussiani. Ciò è stato possibile impiegando algoritmi in grado di generare campioni con distribuzione congiunta nota a priori.

L'analisi è stata condotta su campioni caratterizzati da distribuzioni marginali Gumbel standard.

Si sono utilizzati due algoritmi. Il primo (Stephenson, 2003) è basato sul modello multivariato dei valori estremi di Tawn (1988, 1990). Il secondo algoritmo è basato sulla copula (Nelsen, 1999).

Con tali algoritmi sono stati generati campioni con marginali Gumbel standard[3] facendo variare la numerosità (40, 80, 160, 320 coppie) ed il valore del parametro di dipendenza imposto $\theta = 1/\alpha$ (0.1, 0.3, 0.5, 0.7, 0.9). Per ogni combinazione della coppia (parametro θ - numerosità del campione), imposti in fase di simulazione, sono state eseguite 100 generazioni indipendenti, per le quali è stata calcolata la media sui 100 valori di θ ottenuti con i due stimatori (correlazione lineare e correlazione di rango di Kendall).

L'obiettivo è evidenziare le possibili diversità di comportamento tra i due stimatori in relazione ad uno stesso campione, e a campioni generati con approcci diversi. Teoricamente si attenderebbe una maggiore precisione da parte del parametro di dipendenza di Kendall, poiché il campione non-gaussiano dovrebbe far insorgere gli inconvenienti propri dell'applicazione della correlazione lineare a dati non normali (Embrechts et al., 1999, 2002). I risultati sono riportati nelle *Tabelle 2* e *3*.

Essi evidenziano al contrario l'assenza di rilevanti differenze numeriche nello stimare θ attraverso il coefficiente di correlazione lineare ovvero tramite la correlazione di rango τ_k di Kendall.

Tabella 2. Analisi su campioni generati con l'algoritmo di Tawn-Stephenson: a) media di θ su 100 generazioni al variare della numerosità e del parametro imposto, calcolato tramite la correlazione lineare; b) media di θ calcolato tramite la correlazione di rango Kendall.

a)

Media	*θ=0.1*	*θ=0.3*	*θ=0.5*	*θ=0.7*	*θ=0.9*
n=40	0.10216	0.31718	0.50398	0.71697	0.92157
n=80	0.10098	0.30606	0.51670	0.70865	0.90247
n=160	0.10162	0.30578	0.50356	0.70474	0.90034
n=320	0.10096	0.30413	0.50471	0.70658	0.90209

[3] Valore medio = 0.577216, varianza = 1.644934, deviazione standard = 1.28255, quantili $\xi_{.001} = -2$, $\xi_{.999} = 7$.

b)

Media	*θ=0.1*	*θ=0.3*	*θ=0.5*	*θ=0.7*	*θ=0.9*
n=40	0.12295	0.32625	0.50743	0.72333	0.92015
n=80	0.10935	0.30850	0.51733	0.70519	0.90384
n=160	0.10588	0.30277	0.50319	0.70362	0.89749
n=320	0.10405	0.30332	0.50213	0.70157	0.89930

Tabella 3. Analisi su campioni generati con l'algoritmo di Nelsen: a) media di θ su 100 generazioni al variare della numerosità e del parametro imposto, calcolato tramite la correlazione lineare; b) media di θ calcolato tramite la correlazione di rango Kendall.

a)

Media	*θ=0.1*	*θ=0.3*	*θ=0.5*	*θ=0.7*	*θ=0.9*
n=40	0.09204	0.33542	0.47937	0.72788	1.02872
n=80	0.10525	0.29100	0.48134	0.67996	0.97128
n=160	0.10585	0.29538	0.49257	0.69654	0.94089
n=320	0.09611	0.29676	0.50902	0.69033	0.96419

b)

Media	*θ=0.1*	*θ=0.3*	*θ=0.5*	*θ=0.7*	*θ=0.9*
n=40	0.09462	0.34103	0.49077	0.71974	1.04293
n=80	0.09500	0.27690	0.48380	0.69323	0.97448
n=160	0.10267	0.29270	0.51170	0.70124	0.96715
n=320	0.09864	0.29716	0.51058	0.70277	0.96952

4.2 Effetti di tre diversi metodi di campionamento per le serie di precipitazione

Altro aspetto approfondito in questo lavoro è la modalità di selezione del campione nel caso di serie di precipitazioni. È stata impiegata una

serie costituita da 40 anni di registrazioni pluviometriche con scala di aggregazione giornaliera. Per poter eseguire un'analisi bivariata è necessario in primo luogo definire un evento. Un evento indipendente è selezionato considerando una successione consecutiva di valori di precipitazione giornaliera diversa da zero preceduta e seguita da un valore nullo. Un evento è dunque caratterizzato da una intensità massima giornaliera (valore massimo dell'evento, mm/giorno), da una durata (numero di valori dell'evento, giorni) e da un volume (somma delle intensità estesa alla durata dell'evento, mm).

Le grandezze scelte per le indagini sono l'intensità ed il volume, le quali sperimentalmente presentano correlazione positiva.

Poiché si vuole operare un'analisi bivariata degli eventi estremi si campiona la serie selezionando un evento per ogni anno osservato. In letteratura (Yue, 2000b, 2001) la scelta delle coppie avviene in genere preferendo l'evento caratterizzato dal massimo picco annuale. Tuttavia, tale scelta non è unica potendo considerare in alternativa l'evento estremo in termini di volume, ovvero assumere gli eventi estremi annuali per ambedue le grandezze. Per evidenziare le eventuali difformità nel comportamento del modello in relazione alla modalità di selezione degli eventi estremi annuali, le coppie sono state scelte secondo tre criteri:

1) si assume l'evento che durante l'anno ha presentato la massima intensità (con il corrispondente volume), ottenendo 40 coppie, relative ai 40 anni di osservazione;
2) si sceglie l'evento che durante l'anno ha presentato il massimo volume (con la corrispondente intensità) ottenendo ancora 40 coppie;
3) si considerano il massimo volume e la massima intensità verificatisi durante l'anno anche se appartenenti a eventi diversi: in questo caso si sono selezionate 56 coppie in quanto solo in 16 occasioni il picco massimo ed il volume massimo non appartengono allo stesso evento.

I confronti sono effettuati sulle curve di livello delle superfici dei tempi di ritorno congiunti, definiti dalla relazione (Yue, 2000b):

$$T_{x,y} = \frac{1}{1 - H(x, y)} \tag{12}$$

con $H(x,y) = Pr\,(X \le x, Y \le y)$.

L'approccio basato sulla copula richiede, per la definizione della distribuzione congiunta, la preventiva scelta delle distribuzioni marginali. Data la natura delle grandezze studiate (massimi annuali di variabili idrologiche) si opta per delle distribuzioni dei valori estremi generalizzate (GEV) a tre parametri, che per $\kappa \to 0$ degenerano in distribuzioni Gumbel:

$$F(x) = \exp^{-\left(1-\kappa\frac{x-\mu}{\sigma}\right)^{\frac{1}{\kappa}}} I_{(\mu+\sigma/\kappa,\infty)}(x) \qquad (13)$$

dove $\mu > -\sigma/\kappa$ è un parametro di posizione, $\sigma > 0$ è un parametro di scala e $\kappa < 0$ è un parametro di forma. In *Tabella 4* si riportano i risultati dei test di adattamento di Kolmogorov-Smirnov (K-S) per le distribuzioni marginali. La stima dei parametri è eseguita tramite il metodo dei momenti (Kottegoda e Rosso, 1997).

Dall'inferenza sulla copula emerge la bontà della copula di Gumbel-Hougaard nell'interpretare la struttura di dipendenza dei dati per i tre metodi di selezione.

A titolo di esempio, nelle *Figure 2* e *3* sono riportate le marginali, la distribuzione congiunta e la superficie dei tempi di ritorno congiunti per il metodo di campionamento 3.

In *Figura 4* vengono quindi riportate le curve di livello di 3 tempi di ritorno ritenuti rappresentativi: 10, 50 e 100. Ogni curva contiene le coppie intensità-volume caratterizzate da un certo tempo di ritorno. Si può osservare la sostanziale equivalenza, ai fini della definizione dei tempi di ritorno congiunti, dei metodi di selezione 1 e 2. Più pronunciata è la differenza tra questi e i tempi relativi al campione selezionato con il terzo metodo.

Se da un lato, nell'introdurre ulteriori dati anche se non estremi annuali, vi è un arricchimento dell'informazione, dall'altro ciò comporta uno spostamento delle curve di livello dei tempi di ritorno congiunti, per cui ad una fissata probabilità di non superamento corrispondono coppie di valori inferiori rispetto a quelli che si ottengono assumendo i campioni selezionati con gli altri due criteri.

Con il primo criterio si ottiene un campione in cui i picchi sono i massimi annuali, ma non necessariamente lo sono anche i volumi; in

modo duale, il secondo criterio individua dei campioni per i quali i volumi sono massimi, ma ad essi possono corrispondere picchi che non siano estremi annuali; infine, con il metodo 3 si opera una "ibridazione" del campione dovuta alla presenza in ambedue le serie di dati (picchi e volumi) di valori che non sono massimi annuali.

Tabella 4. Test di Kolmogorov-Smirnov sulle distribuzioni marginali.

TEST K-S	Criterio 1		Criterio 2		Criterio 3	
	Picchi max	Volumi	Picchi	Volumi max	Picchi max	Volumi max
D_{Gumbel}	0.1258	0.1442	0.1070	0.1400	0.1109	0.1241
D_{GEV}	0.1077	0.0975	0.0826	0.0942	0.0826	0.0765
$D_{95\%}$= 0.2150; $D_{99\%}$= 0.2577						

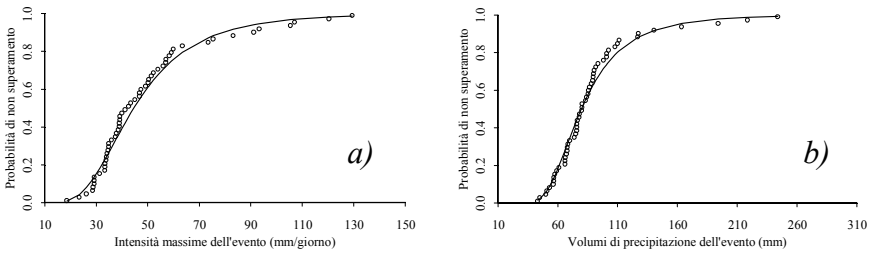

Figura 2. Criterio di selezione 3: confronto tra frequenze empiriche di non superamento e distribuzioni marginali GEV calcolate sul campione delle intensità (a) e sul campione dei volumi (b)

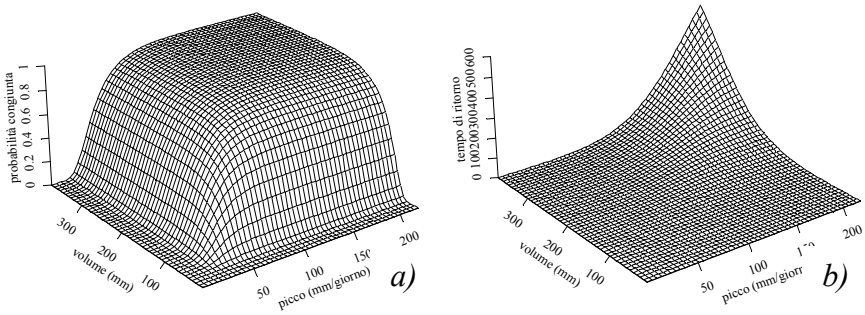

Figura 3. Criterio di selezione 3; a) funzione di distribuzione congiunta ottenuta accoppiando le marginali GEV con una copula di Gumbel-Hougaard; b) superficie dei tempi di ritorno congiunti.

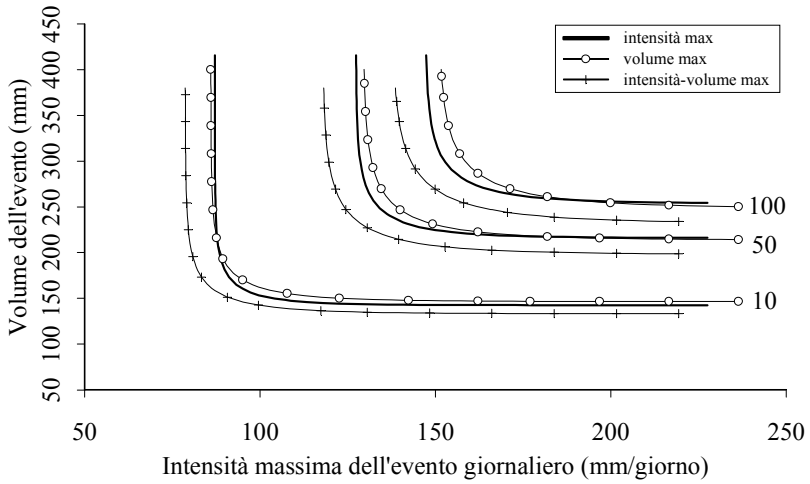

Figura 4. Confronto delle curve di livello dei tempi di ritorno congiunti desunte dai campioni ottenuti con le tre modalità di selezione: picco max, volume max, picco-volume max, e copula di Gumbel con marginali GEV.

4.3 Verifica dell'uso di generatori di serie sintetiche per l'incremento di numerosità del campione disponibile

L'ultimo test presentato in questo lavoro, come già delineato nell'introduzione, consiste nel verificare se generando delle serie sintetiche con Modelli Parametrici Lineari (MPL) si incrementa l'informazione disponibile.

Si considerano 2 serie di precipitazione giornaliere:

a) la serie completa registrata a Perugia dal 1933 al 1972 (40 anni)
b) la serie dei primi 10 anni (dal 1933 al 1942) della stessa registrazione.

Si ipotizza quindi che la serie dei 40 anni abbia una sufficiente informazione per un'analisi idrologica. È evidente che la serie di 10 anni rappresenta un campione insufficiente.

Modellando con un MPL la serie di 10 anni (Grimaldi e Napolitano, 2003; Grimaldi, 2003, in review; Grimaldi et al., 2003, in review) si è generata una terza serie, sintetica, di 500 anni.

Sulle coppie intensità-volume ottenute applicando i primi 2 metodi descritti nel paragrafo 4.2 per le 3 serie (40, 10 e 500 anni) si è applicata la procedura descritta nel precedente paragrafo per identificare e stimare la distribuzione bivariata.

Nelle *Figure 5* e *6* sono confrontate le curve per i tempi di ritorno di 10 e 100 anni. Si può notare una tendenza delle curve relative alle coppie sintetiche ricavate sui 500 anni, ad avvicinarsi alle curve relative al campione di 40 anni. Ciò si ritiene sia un indice di arricchimento dell'informazione in particolare in termini di intensità del campione di 10 anni.

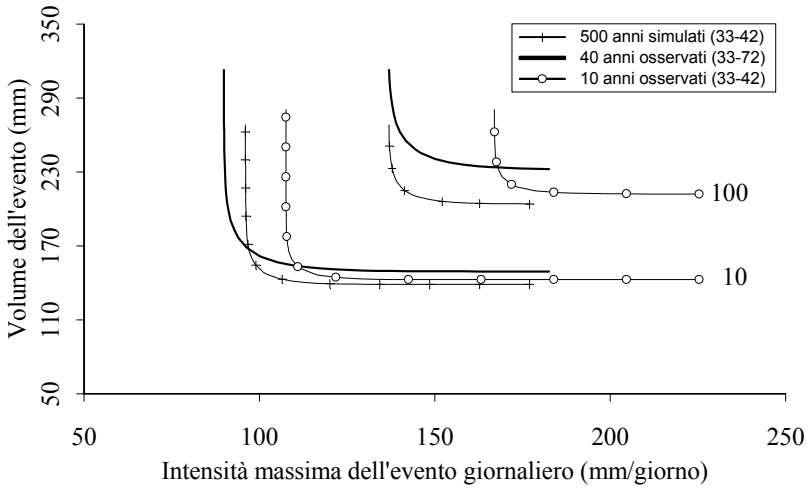

Figura 5. Confronto delle curve dei tempi di ritorno relativi al metodo di selezione della massima intensità annuale.

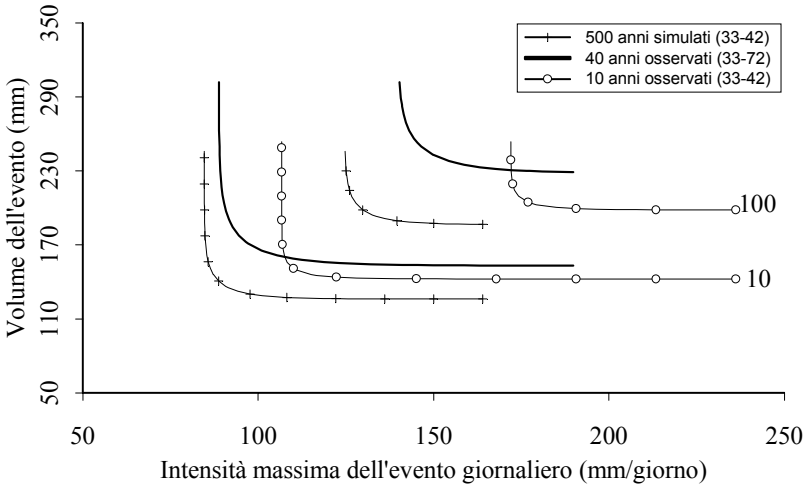

Figura 6. Confronto delle curve dei tempi di ritorno relativi al metodo di selezione del massimo volume annuale.

5. Conclusioni

In questo lavoro è stato descritto ed analizzato il metodo basato sulla copula relativamente alle analisi statistiche idrologiche multivariate. Dopo una breve introduzione sulle metodologie presenti in letteratura utilizzate per analisi statistiche bivariate, viene definita la copula e vengono descritte alcune proprietà e la procedura per la sua applicazione. Inoltre vengono presentati e discussi tre casi di studio.

Il primo di essi ha evidenziato come la differenza tra gli approcci tradizionali e la copula non deve essere attribuita al differente metodo utilizzato per il calcolo del parametro di dipendenza. Infatti i risultati ottenuti nelle simulazioni presentate, lasciano presumere che utilizzando distribuzioni congiunte con parametro desunto tramite il coefficiente di correlazione lineare o il coefficiente di Kendall, non si ottengono, nella pratica, significative differenze.

Si è a riguardo evidenziato che il pratico valore aggiunto offerto da tale metodologia concerne la sua versatilità. Si possono infatti definire

distribuzioni congiunte partendo da distribuzioni marginali differenti accoppiate mediante molteplici funzioni di dipendenza.

La seconda analisi proposta ha evidenziato come influisca il metodo di campionamento nelle analisi di serie giornaliere. È emerso che campionare i valori estremi in modi diversi può produrre risultati difformi. Ulteriori casi di studio sono necessari per generalizzare tale comportamento.

Nella terza ed ultima analisi si è indagato se, generando delle serie sintetiche da quelle osservata, si può avere un incremento di informazione. I risultati in questo caso si ritengono positivi anche se, come nel caso precedente, necessitano di ulteriori verifiche empiriche.

Appendice 1

Si consideri la distribuzione logistica bivariata (Gumbel, 1961):

$$H(x,y) = (1 + e^{-x} + e^{-y})^{-1}. \tag{14}$$

Tale distribuzione ha marginali di tipo logistico standard:

$$F(x) = H(x,\infty) = (1 + e^{-x} + e^{-\infty})^{-1} = (1 + e^{-x})^{-1}$$

$$G(y) = H(\infty,y) = (1 + e^{-\infty} + e^{-y})^{-1} = (1 + e^{-y})^{-1}$$

Si considerino le inverse delle due marginali:

$$x = F^{(-1)}(u) = -\ln(u^{-1} - 1)$$

$$y = G^{(-1)}(v) = -\ln(v^{-1} - 1)$$

con F(x) = u e G(y) = v. Per il *Corollario 1*, dalla (12) si ottiene:

$$\begin{aligned} C(u,v) &= \left(1+e^{-\left(-\ln\left(u^{-1}-1\right)\right)}+e^{-\left(-\ln\left(u^{-1}-1\right)\right)}\right)^{-1} = \left(1+\left(u^{-1}-1\right)+\left(v^{-1}-1\right)\right)^{-1} \\ &= \frac{uv}{u+v-uv} \end{aligned} . \tag{15}$$

Con tale copula è ora possibile costruire una funzione di distribuzione congiunta con marginali qualsiasi; si assumano ad esempio una distribuzione univariata tipo Gumbel ed una tipo Pareto:

$$u = F(x) = \exp(-\exp(-x)) \tag{16}$$

$$v = G(y) = 1 - y^{-c} \quad \text{con} \quad y > 1 \quad \text{e} \quad c > 0 \tag{17}$$

in cui *c* è il parametro di forma. Sostituendo le (16) e (17) nella (15) si ottiene la seguente funzione bivariata:

$$H(x,y) = \left(\frac{1}{e^{-e^{-x}}} + \frac{1}{1-y^{-c}} - 1 \right)^{-1}. \tag{18}$$

Questa, per l'inverso del teorema di Sklar, è una funzione di distribuzione congiunta con marginali F(x) e G(y), infatti:

$$H(x,\infty) = \left(\frac{1}{e^{-e^{-x}}} + \frac{1}{1-\frac{1}{\infty^{c}}} - 1 \right)^{-1} = e^{-e^{-x}} = F(x)$$

$$H(\infty,y) = \left(\frac{1}{e^{-e^{-\infty}}} + \frac{1}{1-y^{-c}} - 1 \right)^{-1} = 1 - y^{-c} = G(y)$$

.

Appendice 2

	Copula C(u,v)	$\tau_K = 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}dt$	Intervallo per τ_K
Gumbel-Hougaard	$\exp\left[-\left((-\ln u)^{\alpha} + (-\ln v)^{\alpha}\right)^{1/\alpha}\right]$	$1-\alpha^{-1}$	[0,1]
Clayton-Pareto	$\left(u^{-\alpha} + v^{-\alpha} - 1\right)^{-1/\alpha}$	$\frac{\alpha}{\alpha+2}$	$[-1,1]\backslash\{0\}$
Ali-Mikhail-Haq	$\frac{uv}{1-\alpha(1-u)(1-v)}$	$1 + 4\frac{\frac{-1}{6\alpha}\left((-1+\alpha)^2\ln(1-\alpha)\right)}{6\alpha^2}$	$[-0.181726, \frac{1}{3}]$
Frank	$\frac{1}{\alpha}\ln\left(1 + \frac{\left(e^{\alpha u}-1\right)\left(e^{\alpha v}-1\right)}{e^{\alpha}-1}\right)$	$1 - \frac{4}{\alpha}\left[D_1(-\alpha) - 1\right]$ *	$[-1,1]\backslash\{0\}$

* con $D_1(\alpha) = \frac{1}{\alpha}\int_0^{\alpha} \frac{t}{e^t - 1}dt$ e $D_1(-\alpha) = D_1(\alpha) + \frac{\alpha}{2}$.

Ringraziamenti: Gli autori ringraziano l'Ing. Francesco Napolitano ed il Dott. Gianfausto Salvadori per i preziosi suggerimenti. Questo lavoro è stato finanziato dal CNR-GNDCI.

Riferimenti Bibliografici

Bacchi B., Becciu G., Kottegoda N.T. (1994), Bivariate exponential model applied to intensities and durations of extreme rainfall, *Journal of Hydrology,* 155, 225-236.

Box G.E.P., Cox D.R. (1964), An analysis of trasformation, *Journal of Royal Statistical Society*, 26, 211-252.

Correia F.N. (1987), Multivariate partial duration series in flood risk analysis, in Singh V.P. (Ed.), *Hydrologic Frequency Modelling*, Reidel, Dordrecht, 541-554.

De Matteis R. (2001), Fitting copulas to data, tesi di laurea, Institute of Mathematics of the University of Zurich.

De Michele C., Salvadori G. (2003), A Generalized Pareto Intensity-Duration Model of Storm Rainfall Exploiting 2-Copulas, *Journal of Geophysical Research,* (D2), 10.1029/2002JD002534, ACL 15, 1-11.

Embrechts P., McNeil A.J., Straumann D. (1999), Correlation: pitfalls and alternatives, *RISK Magazine*, 69-71.

Embrechts P., McNeil A.J., Straumann D. (2002), Correlation and dependence in risk management: properties and pitfalls, in *Risk management: value at risk and beyond*, edito da Dempster M. A. H., pubblicato da Cambridge University Press, Cambridge, 176-223.

Embrechts P., Lindskog F., McNeil A.J. (2003), Modelling dependence with copulas and applications to risk management, in *Handbook of Heavy Tailed Distributions in Finance*, edito da Rachev S.T., pubblicato da Elsevier/North-Holland, Amsterdam, capitolo 8, 329-384.

Freund J.E. (1961), A bivariate extension of the exponential distribution, *Journal of the American Statistical Association*, 56, 961-977.

Genest C., Rivest L. (1993), Statistical Inference Procedures for Bivariate Archimedean Copulas, *Journal of the American Statistical Association*, 88, 1034-1043.

Goel N.K., Seth S.M., Chandra S. (1998), Multivariate modeling of flood flows, *Journal of Hydraulic Engineering ASCE*, (2), 146-155.

Grimaldi S., Napolitano F. (2003), Statistiche multivariate degli eventi estremi derivate dalla generazione di serie giornaliere di precipitazione, Atti della Giornata di Studio *Metodi statistici e matematici per l'analisi di serie idrologiche*, Roma, a cura di Piccolo D. e Ubertini L., Pubblicazione CNR-GNDCI n.2812, 209-219.

Grimaldi S. (2003), Linear Parametric Models Applied on Daily Hydrological Series, inviato per la pubblicazione al *Journal of Hydrologic Engineering ASCE.*

Grimaldi S., Napolitano F. Ubertini L. (2003), A Procedure to use Linear Parametric Models for Daily Rainfall Series Simulation, inviato per la pubblicazione al *Journal of Hydrologic Engineering ASCE.*

Gumbel E.J. (1958), *Statistics of Extremes*, Columbia University Press, New York.

Gumbel E.J. (1960a), Distributions del valeurs extremes en plusieurs dimensions, *Publications de L'Institute de Statistique, Paris*, 9, 171-173.

Gumbel E.J. (1960b), Bivariate exponential distributions, *Journal of the American Statistical Association*, 55, 698-707.

Gumbel E.J. (1961), Bivariate logistic distributions, *Journal of the American Statistical Association*, 56, 335–349.

Gumbel E.J., Mustafi C.K. (1967), Some analytical properties of bivariate extreme distributions, *Journal of the American Statistical Association*, 62, 569-588.

Hashino M. (1985), Formulation of the joint return period of two hydrologic variates associted with a Poisson process, *Journal of Hydroscience and Hydraulic Engineering*, (2), 73-84.

Kottegoda N.T., Rosso R. (1997), *Probability, statistics, and reliability for civil and environmental engineers*, McGraw-Hill, New York.

Moisello U. (1999), *Idrologia tecnica*, La Goliardica Pavese, Pavia.

Mood A.M., Graybill F.A., Boes D.C. (1988), *Introduzione alla statistica*, McGraw-Hill, Milano.

Nelsen R.B. (1999), *An Introduction to Copulas*, Lecture Notes in Statistcs 139, Springer-Verlag, New York.

Sackl B., Bergmann H. (1987), A bivariate flood model and ist application, in Singh V.P. (Ed.), *Hydrologic Frequency Modelling*, Reidel, Dordrecht, 571-582.

Singh K., Singh V.P. (1991), Derivation of bivariate probability density funcions with exponential marginals, *Journal of Stochastic Hydrology and Hydraulics*, 5, 55-68.

Stephenson A.G. (2003), Simulating multivariate extreme value distributions of logistic type, accettato per pubblicazione in *Extremes*.

Tawn J.A. (1988), Bivariate extreme value theory: Models and estimation, *Biometrika* (3), 397-415.

Tawn J.A. (1990), Modelling multivariate extreme value distributions, *Biometrika* (2), 245-253.

Tiago de Oliveira J. (1975), Bivariate extremes: Extensions, *Bullettin of the International Statistical Institute*, (2), 241-251.

Yue S., Ouarda T.B.M.J., Bobée B., Legendre P., Bruneau P. (1999), The Gumbel mixed model for flood frequency analysis, *Journal of Hydrology*, (1-2), 88-100.

Yue S. (2000a), Joint probability distribution of annual maximum storm picks and amounts as represented by daily rainfalls, *Hydrological Sciences Journal*, (2), 315-326.

Yue S. (2000b), The Gumbel mixed model applied to storm frequency analysis, *Water Resources Management*, 14, 377-389.

Yue S. (2001), The Gumbel logistic model for representing a multivariate storm event, *Advances in Water Resources*, 24, 179-185.