ELABORAZIONE STATISTICO PROBABILISTICA DELLE GRANDEZZE IDROLOGICHE ESTREME.

Introduzione

Quale che sia la tecnica di indagine, anche la più semplice, adottata per lo studio di un determinato fenomeno naturale, essa non può prescindere dalla disponibilità di osservazioni e misurazioni inerenti il fenomeno oggetto di studio. Non fa eccezione l'idrologia, dove le attività di osservazione e misura in campo rivestono un'importanza del tutto speciale. La semplice interpretazione dei dati raccolti, oltre che limitativa, si presenta alquanto problematica quando i dati sono molto numerosi: la statistica attraverso principi matematici, fornisce gli strumenti per descrivere ed interpretare, attraverso indici sintetici, i dati sperimentali. Questi, a seconda del problema da affrontare, possono essere raccolti sia in relazione al tempo che allo spazio o ad entrambi; in ogni caso abbiamo a che fare con un numero finito di dati che costituiscono un campione di un insieme più vasto che non ci è dato di conoscere che prende il nome di popolazione.

Affinché i dati assumano massima rilevanza ai fini progettuali è necessario che agli stessi sia associata una valutazione di tipo probabilistico, in altre parole bisogna stimare la probabile frequenza di presentazione della grandezza presa a riferimento per problemi dimensionamento di opere di difesa o per l'imposizione di vincoli di varia natura.

Le grandezze più comunemente oggetto di elaborazione in idrologia sono le serie di valori di portata di massima piena, i valori massimi di precipitazione di data durata, le altezze massime del manto nevoso accumulato al suolo, le serie di valori di portata minima.

La presente nota farà principalmente riferimento alle elaborazioni riguardanti le altezze di massima precipitazione di assegnata durata. Nella maggior parte dei casi, infatti, la portata di progetto viene infatti calcolata a partire dalla precipitazione attraverso una descrizione semplificata del funzionamento idrologico del bacino. Tali metodi (*afflussi-deflussi*) consentono di trasformare la pioggia di durata e tempo di ritorno assegnati in valori temibili della portata al colmo, il cui tempo di ritorno coincide, a rigore, con quello della sollecitazione meteorica (*ipotesi di isofrequenza).*

1. Il tempo di ritorno di progetto ed il concetto di rischio idrologico.

La grandezza comunemente presa a riferimento come valore di progetto (per es., per valutare il grado di protezione dagli allagamenti offerto da una certa rete di drenaggio) è il tempo di ritorno T della portata di dimensionamento. Tramite tale espressione si indica il numero di anni in cui il superamento del valore assegnato avviene mediamente una volta; alternativamente, il tempo di ritorno rappresenta il numero di anni che in media separa il verificarsi di due eventi di entità eguale o superiore alla soglia assegnata. Per comprendere l'espressione 'in media', si supponga di suddividere la serie temporale (di durata infinita) in tanti intervalli di durata T pari al tempo di ritorno; in ciascun intervallo la soglia prefissata può essere superata un numero di volte variabile da zero (*la soglia non viene mai superata durante l'intervallo*) a T (*la soglia viene superata tutti gli anni*). La definizione indica che, per un evento caratterizzato da un tempo di ritorno pari a T, il numero medio di tali superamenti sarà pari ad uno.

Quando si deve valutare la probabilità di fallanza di un'opera, il concetto di tempo di ritorno viene spesso sostituito da quello di rischio. Si definisce rischio associato ad una certa portata la probabilità che la portata stessa sia superata almeno una volta in un numero prefissato di anni; pertanto il rischio dipende dall'estensione del periodo considerato e dalla portata in esame, ovvero dal suo tempo di ritorno. Se il dimensionamento dell'opera è stato condotto con riferimento alla portata $x(T)$ di T anni di tempo di ritorno, il rischio $R_N[x(T)]$, ovvero la probabilità che, durante N anni di funzionamento, l'opera risulti insufficiente una o più volte, è esprimibile come:

$$R_N[x(T)] = 1 - \left(1 - \frac{1}{T}\right)^N$$

La Tabella 1 fornisce i valori del rischio di insufficienza di un'opera, espressi dalla (1), corrispondenti ad alcuni valori di N e di T di particolare interesse tecnico. Dalla (1), può essere ottenuto, come caso particolare, il rischio di insufficienza dell'opera in un numero di anni pari al tempo di ritorno, $R_T[x(T)]$,

$$R_T[x(T)] = 1 - \left(1 - \frac{1}{T}\right)^T$$

che, al crescere di T, tende rapidamente al valore asintotico 0.632, come risulta peraltro evidente dall'analisi dei valori del rischio riportati in Tabella 1, relativi a valori identici di N e T. Ciò indica che la probabilità che un'opera diventi insufficiente in un arco di tempo di durata uguale al tempo di ritorno di progetto è pari, per valori non troppo piccoli di quest'ultimo, al 63% circa. Sempre dalla (1), ponendo N pari al numero di anni di vita attesa dell'opera, può ottenersi il rischio di fallanza nell'intero periodo di funzionamento previsto. Quest'ultimo può risultare un indice particolarmente utile soprattutto per aree di grande importanza economica, per le quali la probabilità di allagamento deve risultare ragionevolmente contenuta.

Tabella 1 - Probabilità che, durante *N* anni di funzionamento, la portata marcata da *T* anni di tempo di ritorno venga superata una o più volte.

	N [anni]					
T [anni]	2	5	10	20	50	100
2	0.750	0.969	0.999	0.999	0.999	0.999
5	0.360	0.672	0.893	0.988	0.999	0.999
10	0.190	0.410	0.651	0.878	0.995	0.999
20	0.098	0.226	0.401	0.642	0.923	0.994
50	0.040	0.096	0.183	0.332	0.636	0.867
100	0.020	0.049	0.096	0.182	0.395	0.634

Si consideri, ad esempio, la progettazione di un sistema di drenaggio urbano. I valori di tempo di ritorno comunemente adottati per questi sistemi sono in genere variabili da 2 a 10 anni, e risultano quindi sempre di molto inferiori rispetto al periodo di funzionamento previsto per la rete. Ne consegue che il verificarsi di uno o più crisi di una rete di drenaggio durante il suo periodo di funzionamento sia un evento alquanto probabile, quasi certo. Ciò peraltro corrisponde ad una precisa scelta progettuale, in quanto il contenimento del rischio di fallanza della rete comporta la necessità di incrementare sensibilmente il tempo di ritorno di progetto, con conseguenti (ed in genere inaccettabili) incrementi delle dimensioni ed aggravi dei costi delle canalizzazioni.

2. Analisi statistica delle piogge e identificazione della relazione altezza-durata

Le registrazioni dei pluviografi permettono di ricavare l'andamento dell'intensità di pioggia nel corso di ciascun evento piovoso. Scelto un passo di lettura, si legge l'altezza di pioggia caduta nei successivi intervalli e, riportandone i valori in un grafico altezza-tempo, si costruisce la curva della pioggia cumulata. La curva è tanto più fedele alla realtà del fenomeno quanto più piccolo è l'intervallo scelto come passo di lettura. L'intervallo dovrebbe essere tale da cogliere le variazioni di intensità più significative.
Nel caso dei pluviometri, il minimo passo di lettura utilizzabile è costituito ovviamente da un giorno.

La curva cumulata della precipitazione pone in evidenza le fasi caratterizzate da diverse intensità di pioggia e può essere di determinante importanza nella comprensione della dinamica dei fenomeni di piena e dei dissesti.
Un pluviogramma (detto anche ietogramma) può essere letto in due maniere:

a) ad intervalli di tempo con origine fissa (per esempio, di ora in ora, in modo da avere i valori di pioggia tra le ore 0 ed 1, 1 e 2, 2 e 3, e così via)
b) ad intervallo di tempo costante, ma con origine variabile (per esempio, scelto l'intervallo di un'ora, si esplora il pluviogramma alla ricerca di 60 minuti consecutivi di precipitazione). Il metodo si presta alla ricerca dei valori massimi di pioggia di assegnata durata ed è applicato negli Annali Idrologici del Servizio Idrografico Italiano per ricavare i massimi annuali di 1, 3, 6, 12 e 24 ore.

La relazione fra altezze di precipitazione e loro durata può essere rappresentata tramite una curva che fornisce, per un assegnato valore del tempo di ritorno *T*, la relazione fra la durata della pioggia *t* e la relativa altezza di precipitazione *h*. Tale relazione prende il nome linea segnalatrice di probabilità pluviometrica(LSPP). In pratica non ci si limita mai ad una curva sola, ma si considera un fascio di curve, ciascuna delle quali corrisponde ad un valore diverso del tempo di ritorno.
Diverse formule, piuttosto simili, sono utilizzate per descrivere questa relazione. In Italia viene generalmente utilizzata una legge di potenza del tipo $h_{t,T}=a\ t^n$, dove *a* ed *n* sono coefficienti che dipendono dal tempo di ritorno. La linea segnalatrice viene ricavata in genere tramite una regressione lineare fra i valori di pioggia *T*-ennali di durata assegnata e le durate stesse adottando una trasformata logaritmica.

3. Determinazione dell'altezza di pioggia h relativa ad un assegnato tempo di ritorno per le diverse durate

La determinazione dell'altezza di pioggia relativa ad un assegnato tempo di ritorno viene condotta utilizzando un'impostazione probabilistica. Per favorire la comprensione della terminologia adottata nelle osservazioni ed elaborazioni condotte nei successivi paragrafi, appare utile premettere la definizione di alcune grandezze costantemente richiamate.
Si definisce *funzione di distribuzione di probabilità (cumulata)* di una generica variabile aleatoria *X*, continua e reale, una relazione che per ogni valore fissato, ma generico $x \in X$, fornisce la probabilità di non superamento del valore medesimo:

$$P(x)=Prob[X\leq x]$$

In relazione ad un generico intervallo infinitesimo (x,$x+dx$), la probabilità elementare dP che accorrano osservazioni nello stesso vale:

$dP(x)=p(x)dx$

ove $p(x)$ viene indicata come densità di probabilità e rappresenta la derivata di P nel punto generico x.
La funzione $p(x)$, per come definita, fornisce precise informazioni circa la probabilità di occorrenza dei valori di X. Infatti essi tenderanno a presentarsi più frequentemente nella zona ove $p(x)$ è massima, mentre tenderanno a diradare ove essa assume valori più bassi. La caratterizzazione della forma della distribuzione cumulata di una variabile aleatoria, e della sua densità, può essere effettuata tramite i cosiddetti *momenti*. Questi parametri, definiti come

$$\mu_r'(x)=\int_{-\infty}^{+\infty} x^r\, p(x)dx \qquad r=1,2,3$$

rappresentano i momenti di vario ordine, r, dell'area sottesa alla funzione $p(x)$ rispetto all'origine (x=0). Il loro valore dipende quindi esclusivamente dalla forma di questa funzione.
Indicando con il simbolo μ, o anche $\mu(x)$, il momento di primo ordine, detto media della X, si definiscono come momenti centrali i parametri:

$$\mu_r(x)=\int_{-\infty}^{+\infty} (x-\mu)^r\, p(x)dx \qquad r=1,2,3....$$

Parametri particolarmente importanti per la caratterizzazione delle distribuzioni di probabilità d'uso più comune nelle applicazioni idrologiche sono:
a) il momento centrale del secondo ordine, indicato comunemente con il simbolo σ^2, e chiamato varianza della X; esso rappresenta il momento d'inerzia dell'area sottesa dalla $p(x)$ rispetto alla media; la sua radice quadrata, σ, che è il raggio giratore, viene chiamata, con dizione anglofila, deviazione standard, o, in italiano, scarto quadratico medio, e misura quindi la dispersione delle masse di probabilità intorno alla media; tale dispersione può essere misurata anche con l'indice adimensionale $CV=\sigma/\mu$, detto *coefficiente di variazione*.
b) i momenti centrali di terzo e quarto ordine μ_3 e μ_4, o meglio, i coefficienti adimensionali $\gamma=\mu_3\ /\sigma^3$ e $k=\mu_4\ /\sigma^4$, detti rispettivamente *di asimmetria* (o di *skewness*) e *di appiattimento* (o di *curtosi*); il primo rappresenta il grado ed il tipo di asimmetria di $p(x)$; il secondo ne misura l'appiattimento, cioè la

concentrazione di probabilità intorno al valore di X per cui $p(x)$ è massima (*valore modale*).
Nella pratica statistica uno dei problemi fondamentali è quello dell'inferenza della distribuzione di probabilità da cui possono essere considerati estratti una serie di valori di una generica variabile aleatoria oggetto di studio. In altri termini, dato un insieme di osservazioni x_1, x_2, x_3,....,x_N, (chiamato campione *N-dimensionale*) ottenute tramite misure su una variabile X, ci si pone il problema di determinare la forma di solito analitica, di una funzione di distribuzione $P(x)$ atta a rappresentare, con ragionevole approssimazione, la distribuzione vera, ma incognita, della X.
A tal fine, il procedimento che viene adottato è il seguente:

1) in base ad analisi preliminari ed a ragionamenti sulla naturale variabilità di X, si presuppongono una o più forme analitiche per la distribuzione incognita;
2) si stimano, sulla scorta del campione di osservazioni x_1, x_2, x_3,....,x_N i parametri delle distribuzioni prescelte; in questa fase occorre prestare attenzione alla metodologia adottata in quanto da essa, in larga misura, dipende il risultato dell'analisi;
3) si eseguono quindi dei test di controllo per valutare l'affidabilità delle distribuzioni ipotizzate; in base ai risultati di questi procedimenti è possibile effettuare una ragionevole scelta del modello $P(x)$ più adatto a descrivere la variabilità naturale di X.

Elemento cruciale del procedimento inferenziale è la qualità e la numerosità dei dati disponibili. In base ad essi viene infatti effettuata, come detto, la stima dei parametri delle $P(x)$, ovvero vengono a questi assegnati valori numerici precisi. Una scarsa qualità dei dati o una dimensione eccessivamente modesta del campione di osservazioni rendono infatti inaffidabile l'intera operazione.

L’analisi di frequenza è fondata implicitamente su tre assunzioni fondamentali:

1. I dati analizzati descrivono eventi aleatori.
2. I processi naturali coinvolti nella generazione degli eventi analizzati sono stazionari rispetto al tempo.
3. I parametri della popolazione possono essere stimati sulla base del campione di dati disponibile.

3.1. Il metodo dei momenti per la stima dei parametri di una distribuzione

Un metodo di stima dei parametri delle distribuzioni di probabilità molto semplice e rapido è quello dei momenti.
Per fissare le idee consideriamo il caso di una distribuzione che richiede la specificazione di *s* parametri. Poiché gli *s* parametri della distribuzione si

possono esprimere in funzione dei suoi momenti dei primi *s* ordini, il problema della loro stima si può ricondurre a quello della stima di tali momenti.
Il metodo dei momenti consiste quindi nell'attribuire a ciascun momento della popolazione il valore del corrispondente momento del campione estratto da quella popolazione. Si impone quindi che la media della popolazione della variabile casuale sia eguale alla media delle N osservazioni x_1, x_2, x_3,....,x_N ; che la varianza della popolazione sia eguale alla varianza del campione, etc. Poiché la media e la varianza della popolazione sono esprimibili in funzione dei parametri della distribuzione, le relazioni di eguaglianza che così si scrivono risolvono il problema di stima dei parametri della distribuzione (almeno quando la popolazione è espressa in funzione di due parametri).

Fra le distribuzioni di probabilità più utilizzate per la regolarizzazione dei massimi annuali di pioggia figurano la distribuzione di Gumbel e quella log-normale (o di Galton). Nei successivi paragrafi vengono introdotte tali distribuzioni e viene descritto il processo inferenziale, basato sul metodo dei momenti, che porta all'adattamento di tali distribuzioni al campione in esame.

3.1.1 Distribuzione di Gumbel

La distribuzione di probabilità di Gumbel (detta anche prima forma della distribuzione dei valori estremi; con dizione anglofila: EVI, che sta per Extreme Value Type I) assume la seguente forma (per la funzione di densità di probabilità)

$$p(x) = \frac{1}{\alpha}\exp\left[-\frac{x-u}{\alpha} - \exp\left(-\frac{x-u}{\alpha}\right)\right] \qquad -\infty < x < +\infty$$

dove α ed u sono i due parametri della distribuzione.

La probabilità di non superamento è definita da

$$P(x) = \exp\left[-\exp\left(-\frac{x-u}{\alpha}\right)\right]$$

La stima dei parametri α ed u tramite il metodo dei momenti si articola nei seguenti passaggi:

1) si calcolano la media $\bar{x}$ e la deviazione standard s dei valori del campione;
2) si stimano i valori dei parametri α ed u sulla base delle seguenti relazioni:

$$\alpha = \frac{\sqrt{6}\ s}{\pi}$$

$$u = \bar{x} - 0.5772\,\alpha$$

Il coefficiente di asimmetria γ è per questa distribuzione costante e pari a1.1396.

3.1.2 **Distribuzione log-normale a due parametri**

La distribuzione log-normale è caratterizzata dal fatto che a seguire la legge di distribuzione di probabilità normale non è la variabile originaria *x*, bensì il logaritmo della stessa:

$$y = \ln(x)$$

La funzione di densità di probabilità è espressa dalla:

$$p(x) = \frac{1}{x\sqrt{2\pi}\,\sigma(y)} \exp\left\{-\frac{1}{2}\left[\frac{\ln(x)-\mu}{\sigma}\right]^2\right\}$$

In questo caso, i due parametri da cui dipende la distribuzione di probabilità sono rappresentati da μ e σ.
E' opportuno ricordare che la distribuzione cumulata di probabilità normale non è disponibile in forma chiusa (esplicita). Un'approssimazione, normalmente sufficiente, per la distribuzione di probabilità cumulata di tipo normale, indicata come $\Phi(z)$, dove $z=(\ln x-\mu)/\sigma$, è la seguente

$$P = \Phi(z) = 1 - 0.5\exp\left[-\frac{(83z+351)z+562}{703/z+165}\right]$$

valida per $0<z\leq 5$. Un'approssimazione per la funzione di probabilità normale inversa, tale da consentire la determinazione del valore della variabile *z* a partire dal valore di probabilità cumulata *P*, indicata come $\Phi^{-1}(P)$, è la seguente

$$z = \Phi^{-1}(P) = \frac{P^{0.135} - (1-P)^{0.135}}{0.1975}$$

valida nel range $0.005<P<0.995$.

I parametri μ e σ possono essere stimati utilizzando le seguenti espressioni

$$\mu = \ln\bar{x} - \frac{1}{2}\ln\left(1+\frac{s^2}{\bar{x}^2}\right)$$

$$\sigma = \left[\ln\left(1+\frac{s^2}{\bar{x}^2}\right)\right]^{0.5}$$

dove $\bar{x}$ e s rappresentano rispettivamente la media e lo scarto quadratico medio sperimentale del campione di dati disponibile.

3.2 'Plotting position' e cartogrammi probabilistici

Una valutazione della bontà dell'adattamento della distribuzione al campione disponibile può essere conseguita tramite un esame grafico basato sui diagrammi probabilistici tracciati su carte speciali (cartogrammi probabilistici). Tali carte sono costruite in modo che le curve di probabilità di un certo tipo vi vengano rappresentate come rette.

In tali diagrammi, su di un asse (di solito quello delle ascisse) vengono riportati, in scala lineare o logaritmica, i valori della variabile e sull'altro le probabilità o le frequenze cumulate di superamento (o il tempo di ritorno), adottando una particolare graduazione che dipende dal tipo di distribuzione.

Il problema che immediatamente si pone nel momento in cui si intende porre sul diagramma i punti relativi al campione disponibile riguarda la modalità di calcolo della probabilità di superamento (o di non superamento, o del tempo di ritorno) degli eventi considerati. Come si è detto, infatti, si intende usare il diagramma per valutare l'adattamento di una generica distribuzione al campione: è quindi desiderabile disporre di una definizione della probabilità degli eventi del campione indipendente dalla scelta della distribuzione di probabilità stessa.

A questo punto, si consideri, per fissare le idee, una serie di 10 valori massimi annuali di precipitazione di durata oraria. Si ordini tale serie in senso decrescente (il primo valore è il più elevato). E' evidente che il più elevato dei valori si è verificato una volta in 10 anni. Si può allora affermare che il tempo di ritorno di tale evento è pari a 10 anni? Evidentemente no: il valore massimo annuale di pioggia osservato in 10 anni può avere un periodo di ritorno diverso, minore o maggiore di 10 anni. In teoria, se i dati di partenza sono rappresentativi, un evento caratterizzato da 10 anni di tempo di ritorno a lungo andare sarà uguagliato o superato **in media** una volta ogni 10 anni. Quindi in un campione di numerosità pari a 1000, si dovrebbero rintracciare 1000/10, ovvero 100, eventti così caratterizzati. Tuttavia, se tale periodo di 1000 anni fosse diviso in 100 periodi di 10 anni ciascuno, circa il 37% dei campioni di 10 anni non conterrebbero eventi di 10 anni di tempo di ritorno; circa il 37% ne conterrebbe 1; circa il 18% ne conterrebbe 2; circa il 6% ne conterrebbe 3; circa l'1.5% ne conterrebbe 4; e circa il 0.5% ne conterrebbe 5 o più. Ovviamente, la sommatoria di tutti questi eventi è esattamente 100. E' chiaro che quanto si è detto per il tempo di ritorno vale anche per la probabilità di superamento, ad esso collegata.

Percorrendo una fra le varie alternative disponibili, è possibile attribuire al massimo evento osservato nella serie di numerosità N una probabilità di superamento pari alla media delle probabilità di superamento dei massimi eventi osservati in una serie infinita di campioni di numerosità N. Questa definizione porta alla seguente determinazione della probabilità di superamento *1-P* (dato che con *P* si era definita la probabilità di non superamento)

$$1 - P = 1/(N + 1)$$

Da cui si ricava che, essendo *T=1/(1-P)*, il tempo di ritorno del massimo valore osservato in una serie di numerosità pari a 10 vale 11 anni.
Tale regola di definizione della probabilità di non superamento si può poi applicare per determinare la probabilità per gli altri eventi della serie, tenendo conto della corrispondente posizione *m* nella serie ordinata in senso decrescente, nel modo seguente

$$1 - P = m/(N + 1)$$

Questa definizione della probabilità di superamento prende il nome di '*plotting position*'. Si noti che per la '*plotting position*' sono utilizzabili altre definizioni (tutte molto simili fra loro, peraltro).
Utilizzando la *plotting position* è pertanto possibile rappresentare gli elementi del campione sulla carta probabilistica. Se il tipo di distribuzione corrispondente alla carta probabilistica è adatto ad interpretare le osservazioni, queste devono addensarsi più o meno intorno ad una retta. Questa rappresentazione grafica è quindi utile per verificare l'esattezza dell'assunzione fatta circa la distribuzione di probabilità prescelta.
Una volta selezionato il tipo di distribuzione di probabilità, la rappresentazione del campione sul cartogramma probabilistico prescelto si presta anche per la determinazione dei parametri della distribuzione. E' questa la base del metodo dei minimi quadrati. Generalmente la qualità dei risultati che si ottengono tramite tale procedura è inferiore rispetto a quella conseguibile tramite il metodo dei momenti. Questo perché la plotting position fornisce stime della probabilità di superamento che sono considerevolmente incerte, soprattutto per gli eventi di magnitudo maggiore. Un buon modo per illustrare tale incertezza è quella di considerare la variabilità della 'reale' probabilità di superamento relativa al massimo valore osservato in un campione di 10 anni. Tale probabilità è compresa fra 0.029 e 0.115 il 50% delle volte e fra 0.0052 e 0.25 il 90% delle volte. In termini di tempo di ritorno, si può dire che la metà delle volte il tempo di ritorno sarà compreso fra 8 e 34 anni, ed il 90% delle volte il tempo di ritorno sarà compreso fra 4 e 192 anni.