SELEZIONE DEI DATI OMOGENEI E INDIPENDENTI

Uno dei passaggi più delicati dell'inferenza statistica è la selezione dei dati da analizzare dalla serie temporale osservata. In particolare, tale cernita deve assicurare l'omogeneità e l'indipendenza dell'insieme campionario formato. Ne deriva, pertanto, l'esigenza di definire criteri di selezione in base ai quali estrarre dalla serie temporale quei valori che siano, contemporaneamente, descrittivi del medesimo fenomeno fisico e mutuamente indipendenti quindi pensabili come identicamente distribuiti.

Le popolazioni statistiche estratte dai dati osservati possono essere distinte in due diversi tipi, in relazione al procedimento seguito per la loro formazione: il primo prevede il campionamento regolare della serie osservata e genera una popolazione temporalmente aggregata; l'altro prevede un campionamento ad evento e, quindi, genera una popolazione temporalmente disaggregata.

Il primo procedimento è riconducibile al metodo del campione totale, che non assicura a priori né l'indipendenza né l'omogeneità dei dati che formano l'insieme campionario.

Il secondo è riconducibile al metodo delle durate parziali sopra soglia (POT – "*Peak Over Threshold*" - Goda, 1988) al metodo dei selezione a blocchi. Tra questi, qui si è utilizzato il metodo delle durate parziali sopra soglia che, previa la definizione di opportuni criteri d'individuazione degli eventi, permette la cernita di dati sia omogenei sia indipendenti.

In particolare, nel caso delle serie ondametriche, se è relativamente semplice garantire l'indipendenza dei dati selezionati, soprattutto in senso debole, non è altresì agevole garantirne l'omogeneità. Infatti, le osservazioni strumentali raccolte sono relative a stati di mare che hanno avuto diversa genesi e, molto spesso, tale diversità non ha radice totalmente stocastica ma, al contrario, ha origine parzialmente deterministica.

Nel Mediterraneo centrale, per esempio, le limitazioni geografiche delle aree di generazione determinano una risposta del mare alle forzanti meteorologiche funzione della direzione di provenienza del vento. Una medesima perturbazione può, dunque, generare onde molto diverse secondo il settore direzionale forzato, fino addirittura a raggiungere un limite fisico per lo sviluppo del moto ondoso. Questa risposta costante dei bacini si combina tuttavia con le caratteristiche variabili delle perturbazioni meteorologiche che determinano la risposta del mare alle azioni eoliche quali, ad esempio, le loro dimensioni, avvolte comparabili o superiori alla grandezza delle aree di generazione, o alla posizione e al moto del centro di pressione, da cui dipendono la direzione, l'intensità al culmine e la durata della mareggiata.

Pertanto, l'inferenza statistica applicata alle serie ondametriche risulta complicata dalla dipendenza direzionale dei dati e dalla loro relazione con il regime meteorologico e la conformazione del paraggio in cui essi sono misurati.

Inoltre, si deve necessariamente tenere in considerazione la marcata dipendenza statistica tra i dati osservati in condizioni di tempesta, quando le variazioni temporali dei parametri ondosi sono lente.

La definizione dell'evento "mareggiata" permette di semplificare l'identificazione dell'insieme campionario per l'inferenza statistica. La mareggiata si può definire come la successione temporale degli stati di mare caratterizzati da valori d'altezza, periodo e direzione del moto ondoso, variabili in fissati intervalli. In particolare, qui, si considera come "mareggiata" la successione temporale degli stati di mare caratterizzati da:

1. persistenza dell'altezza d'onda sopra la soglia di 1.0 *m* maggiore di 12 ore consecutive,
2. attenuazione dell'altezza d'onda sotto la soglia di 1.0 *m* per meno di 6 ore consecutive,
3. appartenenza della direzione di provenienza al settore angolare di provenienza considerato.

È regola caratterizzare una mareggiata assegnandole i valori d'altezza d'onda, periodo e direzione corrispondenti al culmine d'intensità della successione degli stati di mare.

L'indipendenza campionaria della serie dei colmi è assicurata imponendo un valore di soglia per l'intervallo che intercorre tra due mareggiate successive. L'ampiezza di tale intervallo può essere calcolata in base alla funzione d'autocorrelazione della serie temporale osservata. Imponendo un'indipendenza debole tra gli elementi campionari, si può assumere il ritardo tra due successive mareggiate indipendenti pari all'intervallo temporale che renda il valore della funzione di autocorrelazione sufficientemente piccolo (in letteratura si raccomanda il valore 0.4). Generalmente tale intervallo è pari a 48 ore.

L'indipendenza dell'insieme formato da mareggiate susseguitesi con cadenza non inferiore a due giorni, non ne assicura l'omogeneità. Tale requisito può essere conseguito vagliando ulteriormente il campione indipendente. In particolare, nel metodo delle serie di durata parziale sopra soglia sono analizzati soltanto i valori superiori ad una seconda soglia d'altezza d'onda.

La definizione della soglia di troncamento per la serie disaggregata, quindi, può essere finalizzata a separare, per esempio, le mareggiate generate da venti di tempesta (ciclonici o anticiclonici) da quelle generate da brezze costiere, anche se intense. Nel caso dei nostri mari, in cui le aree di generazione sono limitate e irregolari, la definizione della seconda soglia può essere sfruttata per distinguere le mareggiate

generate da vaste perturbazioni meteorologiche, che si estendono su scale continentali, dalle mareggiate generate da perturbazioni locali, che si estendono su scale regionali.

In ogni caso, la scelta della soglia è un'operazione molto delicata, poiché essa condiziona notevolmente le stime dell'altezza d'onda con assegnato tempo di ritorno.

Una soglia troppo bassa, infatti, conduce ad avere molti elementi nel campione, la maggior parte dei quali non è rappresentativa di condizioni estreme. Tali circostanze determinano previsioni stabili che tendono ad alterare la stima dell'altezza d'onda con tempi di ritorno elevati.

Al contrario, una soglia troppo alta conduce ad avere pochi elementi campionari, anche se tutti rappresentativi di condizioni estreme. Tali circostanze determinano previsioni dell'altezza d'onda variabili con la numerosità campionaria.

Se l'omogeneità climatica può essere ottenuta con il metodo delle serie di durata parziale, l'omogeneità direzionale può essere garantita imponendo l'appartenenza della direzione media di provenienza della mareggiata a determinati settori direzionali. Al fine di individuare tali settori direzionali per ogni boa della RON, si sono prese in considerazione diverse informazioni:

1. la conformazione delle aree di generazione del moto ondoso sottese da ciascuna boa (distribuzione dei fetch geografici),
2. la distribuzione direzionale della frequenza di accadimento delle osservazioni di periodo ed altezza,
3. la distribuzione direzionale dei colmi di mareggiata, in periodo ed altezza.

Si vuole evidenziare che l'esame delle sole curve di frequenza d'accadimento direzionale può portare a sottovalutare alcune preziose informazioni, legate a eventi intensi ma rari che non sono evidenziati dalle curve di frequenza. Per tale ragione, l'individuazione dei settori direzionali, che nelle analisi degli estremi devono comprendere gli eventi intensi ma non frequenti, è stata basata anche sulla distribuzione direzionale dei massimi valori dei colmi di mareggiata.

Il confronto delle informazioni suddette rende l'individuazione dei settori di traversia piuttosto agevole. Infatti, le direttrici di ciascun settore sono immediatamente individuabili grazie ai netti picchi assunti dalle curve di frequenza. I limiti dei settori sono poi definibili esaminando sia la distribuzione direzionale dei colmi di mareggiata sia quella delle lunghezze del fetch geografico.

INDIVIDUAZIONE DEL MODELLO PROBABILISTICO

La rappresentazione dei picchi di mareggiata mediante un'appropriata legge di distribuzione di probabilità introduce nuove problematiche. In particolare, una prima questione è legata alla definizione dell'aggettivo "appropriata". Infatti, la distribuzione di probabilità delle altezze d'onda non è nota a priori e, inoltre, gli studiosi non sono riusciti a identificare univocamente la distribuzione "vera" di un campione d'altezze d'onda. In linea di principio, dunque, qualsiasi distribuzione di probabilità può essere verosimile. Nella prassi operativa, tale presupposto obbliga i progettisti a definire alcuni criteri di scelta, volti ad individuare tra le possibili distribuzioni la legge di probabilità che meglio rappresenta alcune caratteristiche statistiche dei dati (per esempio un migliore adattamento della coda della distribuzione ai valori estremi rilevati).

Indicando con $F(H)$ la probabilità cumulata di non superamento della soglia H, ossia la probabilità che l'argomento H non sia superato da un valore H_i casualmente scelto ($F(H)=F(H_i \le H)$), la $F(H)$ risulta generalmente caratterizzata da due o tre dei seguenti parametri:

- il fattore di scala (A);
- il fattore di posizione (B);
- il fattore di forma (k).

Qui si è adottata la distribuzione di **Weibull** espressa dalla seguente espressione.

$$F(H) = 1 - \exp\left[-\left(\frac{H - B}{A} \right)^k \right]$$

Tra i comuni metodi d'adattamento statistico (massima verosimiglianza, momenti lineari, minimi quadrati) qui si è utilizzato il **metodo dei minimi quadrati**. Una delle sue più diffuse applicazioni trae origine dall'uso della carta di probabilità; in particolare, tale prassi prevede la distorsione degli assi coordinati che individuano il piano di rappresentazione dei valori campionari e della loro probabilità di non superamento, in modo tale che la distribuzione di probabilità cumulata di non superamento si trasformi in una linea retta. Nel caso qui analizzato si ha che la variabile ridotta (x) assume l'espressione

$$x = \left\{ -\ln\left[1 - F(H) \right] \right\}^{1/k}$$

e la relazione $H(F)$ si trasforma nella

$$H = A \cdot x(F) + B$$

Dunque i valori della variabile ridotta sono legati alla probabilità di non superamento della particolare altezza d'onda appartenente all'insieme campionario selezionato. **Tale probabilità non è nota a priori**,

essendo l'incognita del problema. Inoltre, **la variabile ridotta è funzione di uno dei parametri da determinare** (in questo caso il parametro di forma *k*).

Dunque, l'applicazione del metodo della regressione lineare ai minimi quadrati presenta, in generale, la questione dell'attribuzione di una probabilità di non superamento a ciascun elemento dell'insieme campionario selezionato e, in generale, la questione dell'indeterminatezza di uno dei parametri della distribuzione.

Ricordando che la distribuzione di probabilità "vera" non è preventivamente individuabile e, quindi, che qualsiasi distribuzione di probabilità può essere in linea di principio verosimile, l'indeterminatezza di uno dei parametri della distribuzione di Weibull può essere superata nella prassi operativa assegnando al terzo parametro dei valori di tentativo e stimando successivamente i rimanenti parametri incogniti (Goda, 1988). Tale presupposto obbliga il progettista ad utilizzare, in un secondo tempo, dei criteri di scelta tra le diverse distribuzioni identificate.

La questione dell'attribuzione di una probabilità di non superamento a ciascun elemento dell'insieme campionario selezionato è più delicata. Infatti, per aggirare tale indeterminazione si approssima la probabilità di non superamento di un valore d'altezza con la corrispondente frequenza campionaria di non superamento, stimatore non distorno al tendere all'infinito della numerosità campionaria.

Quindi, tale approssimazione è tanto meglio verificata quanto maggiore è la numerosità (*N*) dell'insieme campionario selezionato, che, peraltro, contrasta con la necessità di vagliare solo gli elementi omogenei ed indipendenti. Per bilanciare tali esigenze nel calcolo della frequenza di non superamento, sono state proposte diverse formulazioni stabili e consistenti già a partire da un campione scarsamente numeroso.

Queste formulazioni si basano generalmente sul seguente procedimento:

1. inizialmente si ordina in senso decrescente il campione di altezze d'onda. Il valore dell'altezza di picco alla generica *m*-esima posizione nel campione ordinato è indicata come H_m; quindi, l'insieme ordinato si indica come $(H_1, H_2, ..., H_m, ..., H_N)$, in cui H_1 e H_N sono rispettivamente il massimo e il minimo valore registrato;
2. successivamente si calcola per ciascun elemento della serie ordinata la frequenza di non superamento (chiamata spesso "plotting position"). Indicando con l'apice ($'$) le stime dei rispettivi valori veri, la probabilità di non superamento del *m*-esimo elemento della serie ordinata è data da

$$F'_m = 1 - \frac{m - \mu}{N + \beta}$$

in cui μ e β sono due costanti, il cui valore dipende dalla particolare distribuzione che si sta utilizzando. Per la distribuzione di Weibull, Goda (1988) ottenne che:

$$\mu = 0.20 + 0.27/\sqrt{k}\ ,\ \beta = 0.20 + 0.23/\sqrt{k}$$

Si vuole segnalare che la numerosità da utilizzare nella formula di posizione non è quella del campione censurato (N) ma quella dell'originale insieme campionario omogeneo ed indipendente (N_T).

Determinata la posizione di tutti i valori di altezza d'onda al picco sulla carta di probabilità, come sopra indicato, si è in grado di determinare i coefficienti della distribuzione mediante una semplice regressione lineare.

Come accennato in precedenza, dopo aver ottenuto i parametri della distribuzione, s'incontra l'esigenza di operare una scelta tra tutte le leggi di probabilità considerate e, logicamente, tale scelta deve essere orientata all'individuazione della distribuzione di probabilità più verosimile. In tal senso, in prima approssimazione, si può assumere che la legge di probabilità vera sia quella che si adatta meglio alle osservazioni e, cioè, dal punto di vista dei minimi quadrati, quella che presenti il minimo scarto tra i valori osservati e quelli statisticamente attesi.

A tal riguardo, qui si adotta la metodologia proposta da Goda e Kobune (1990), denominato *minimo rapporto del residuo del coefficiente di correlazione* (*Criterio MIR*), basato sulla normalizzazione del

CALCOLO DELL'INTERVALLO DI CONFIDENZA

Un criterio per la determinazione dell'intervallo di confidenza è stato ricavato da Goda (1988) interpretando i risultati ottenuti mediante numerose simulazioni Monte Carlo. In particolare, l'autore studiò, in funzione del tipo di legge di probabilità, della dimensione campionaria *N* e del parametro di censura ν, il comportamento statistico della variabile adimensionale

$$z = \frac{H'_{T_R} - H_{T_R}}{\sigma_x}$$

in cui H'_{TR} è il valore atteso dell'altezza d'onda con periodo di ritorno T_R, H_{TR} è il valore vero dell'altezza d'onda con medesimo periodo di ritorno (cioè quello calcolato in base alla distribuzione assunta durante la simulazione Monte Carlo) e σ_x è la deviazione standard del campione simulato. Dunque, la variabile *z* è la differenza tra il valore stimato e quello vero, normalizzata rispetto alla deviazione standard del campione analizzato.

L'autore determinò un'espressione empirica della deviazione standard della variabile normalizzata *z*, data dalla

$$\sigma_z = \sqrt{\frac{1.0 + c_1(x_T - c_4 + c_5 \ln\nu)^2 e^{\left[\frac{c_2}{\sqrt[3]{N^4}} + c_3\sqrt{-\ln\nu}\right]}}{N}}$$

in cui i coefficienti risultano funzione del tipo di legge di probabilità adottata. In base alla σ_z è possibile stimare la deviazione standard del valore d'altezza d'onda con assegnato tempo di ritorno $\left(H'_{TR}\right)$ come

$$\sigma\left(H'_{T_R}\right) = \sigma_z \sigma_x$$

Ottenuta la deviazione standard della H'_{TR}, l'intervallo di confidenza della stima corrispondente ad un livello di significatività del 90% si può porre pari a:

$$\pm 1.96\sigma\left(H'_{T_R}\right)$$

RIFERIMENTI

Goda Y., 1988. "On the methodology of selecting design wave height". *Proc. 21st ICCE Conf.*, vol. 1, pp. 899-913.

Goda Y. and Konube K., 1990. "Distribution function fitting for storm wave data". *Proc. 22nd ICCE Conf.*, vol. 1, pp. 1-14.

Piscopia, R., Inghilesi, R., Panizzo, A., Corsini, S. and Franco, L., 2002*a*. "Analysis of 12-year wave measurements by the Italian Wave Network". *Proc. 28th ICCE Conf.*, Cardiff, vol. 1, pp. 121-133.

Piscopia, R., Corsini, S., Inghilesi, R. e Franco, L., 2002*b*. "Misure strumentali di moto ondoso della Rete Ondametrica Nazionale: analisi statistica aggiornata degli eventi estremi". *XVIII Convegno di Idraulica e Costruzioni Idrauliche*, Potenza, Italy, vol. 4, pp. 271-280.

Piscopia, R., Inghilesi, R., Corsini, S. e Franco, L., 2003. "L'atlante delle onde nei mari italiani". *VII ed. Giornate di Ingegneria Costiera*, Trieste, pp. 107-118.

Piscopia, R., Franco, L., Corsini, S., Inghilesi, R. 2004. "Atlante delle onde nei mari Italiani – Italian Wave Atlas", Full Final Report a cura dell'APAT- Università di Roma Tre, Roma.