POLITECNICO DI TORINO

Corso di Dottorato in Ingegneria Idraulica

Tesi di Dottorato

Metodi statistici non-supervised per la stima di grandezze idrologiche in siti non strumentati

Alberto Viglione

Coordinatore del corso di dottorato
Prof. Luca Ridolfi

Tutore: Prof. Pierluigi Claps

XIX ciclo (2004-2006)

Sono molte le persone che vorrei ringraziare per avermi aiutato a portare a termine il lavoro di questo Dottorato di Ricerca. Prima di tutto la mia famiglia, per l'appoggio ed il costante incoraggiamento di questi anni. Ringrazio poi tutto il Dipartimento di Idraulica del Politecnico di Torino, ed in particolare i miei colleghi di Dottorato, con cui ho avuto la fortuna di condividere questo percorso. In modo particolare Pierluigi Claps e Francesco Laio, per gli insegnamenti, la pazienza e l'amicizia senza di cui questa esperienza non avrebbe mai potuto avere luogo. Inoltre ringrazio la Regione Piemonte per avere finanziato questa ricerca. Infine, dal momento che i ringraziamenti non sarebbero sufficienti, vorrei dedicare questo lavoro alla persona che mi è stata più vicina in questi anni: Paola.

Indice

Introduzione

Alla scala regionale le reti di monitoraggio forniscono misure puntuali delle grandezze idrologiche e climatiche. Tuttavia, per molti problemi pratici dell'idrologia, è importante poter disporre di informazioni che siano non solo accurate, ma anche diffuse sul territorio. Inoltre le serie storiche misurate sono spesso brevi al punto da rendere il campione inadeguato ai fini dell'inferenza statistica su base locale. Uno strumento con il quale si possono affrontare questi problemi è l'*analisi regionale* o *regionalizzazione* delle variabili ambientali. Uno studio che voglia descrivere le variabili statisticamente, ovvero associare loro una distribuzione di frequenza, prende il nome di *analisi di frequenza regionale*.

Lo scopo specifico dell'analisi di frequenza regionale applicata a variabili idrologiche è di utilizzare i dati raccolti in molti punti, attraverso le reti di monitoraggio, per caratterizzare siti di particolare interesse ma non strumentati. Quanto detto vale per qualsiasi variabile idrologica: sia essa di tipo puntuale, come ad esempio l'afflusso o la temperatura, che cumulata. In questo lavoro si farà riferimento principalmente al deflusso fluviale che, per definizione, non caratterizza un singolo punto, ma tutto ciò che sta "a monte" della sezione di interesse. Per variabili cumulate come il deflusso, pertanto, non si può ricorrere alle tecniche di interpolazione spaziale utilizzate per grandezze puntuali, come ad esempio il kriging (v.es. *Cressie*, 1993; *Kitanidis*, 1997). La stima dei deflussi in bacini non strumentati è un argomento sul quale la comunità scientifica internazionale da sempre esprime interesse, e su cui recentemente è stata avviata una iniziativa di rilievo: il progetto PUB (Prediction in Ungauged Basins) promosso dall'International Association of Hydrological Sciences (*Sivapalan et al.*, 2003).

Oltre che per la stima in siti non strumentati, l'analisi regionale può essere utilizzata per migliorare la stima della frequenza degli eventi rari in siti in cui la serie storica di misure è breve. Si può dire, quindi, che l'analisi di frequenza regionale permette di estendere nello spazio e nel tempo l'informazione idrologica

disponibile.

L'analisi di frequenza regionale può essere considerata un argomento di ricerca "classico" in idrologia. A partire dalla pubblicazione di *Dalrymple* (1960), le tecniche proposte in letteratura sono state molte e molte sono le varianti utilizzate in ambito operativo. Tutte si basano, però, sull'ipotesi che la variabilità delle curve di frequenza campionarie, ottenute da misure della variabile idrologica in diversi siti, possa essere attribuita ad una componente casuale dovuta all'aleatorietà del campionamento, e ad una componente deterministica che rispecchia la differenza nelle caratteristiche proprie dei siti (o dei bacini, se la variabile è il deflusso). Le metodologie si differenziano per il modo in cui utilizzano quest'ipotesi per costruire modelli di stima delle distribuzioni di frequenza delle variabili idrologiche a partire da queste caratteristiche. Rispetto a quanto è già stato trattato in letteratura, i contributi innovativi che questo lavoro presenta sono rivolti all'individuazione di procedure statistiche oggettive (non-supervised), e basate sull'analisi dei dati (data-driven), per determinare quali caratteristiche (di sito o di bacino) spiegano meglio la variabilità delle curve di frequenza della grandezza idrologica di interesse.

Il metodo di analisi di frequenza regionale considerato è quello della *grandezza-indice* (Capitolo 1), che è senza dubbio il metodo attualmente più utilizzato. Tale metodo scinde l'analisi della distribuzione di frequenza della variabile in due parti: il parametro di scala, o grandezza-indice, considerato variabile in maniera continua sul territorio, e la distribuzione adimensionalizzata, o *curva di crescita*, che si ritiene essere la stessa per tutti i siti appartenenti alle cosiddette *regioni omogenee*. Per quanto riguarda la prima parte, nel Capitolo 2 viene discussa una procedura completa di regressione multipla per derivare relazioni regionali che legano la grandezza-indice alle caratteristiche dei siti (o dei bacini idrografici). La formazione delle regioni (Capitolo 3), invece, viene effettuata valutando la similitudine tra i siti in termini di alcuni descrittori, detti *variabili di classificazione* che, si presume, permettano di spiegare la variabilità della forma delle distribuzioni di frequenza. La scelta di queste grandezze è spesso condotta in maniera soggettiva, e costituisce per questo motivo una fase delicata della procedura. Un metodo statistico oggettivo di scelta delle variabili di classificazione per la formazione delle regioni omogenee, basato sul concetto di distanza/somiglianza tra entità, è descritto nel Capitolo 4.

La valutazione dell'omogeneità delle regioni, infine, viene eseguita tramite test di omogeneità, altro punto critico nell'analisi di frequenza regionale. Molti

test sono stati proposti in letteratura ma un confronto generale tra di essi non era ancora stato fatto. Il Capitolo 5 è dedicato a questo tipo di confronto: si riportano i risultati ottenuti in *Viglione et al.* (2007a) dove quattro test di omogeneità vengono confrontati con un'impostazione di carattere generale.

La metodologia proposta è stata applicata all'analisi regionale del deflusso annuo in Piemonte e Valle d'Aosta ed i risultati, già presentati in *Viglione et al.* (2006), sono discussi nel Capitolo 6. Le applicazioni dell'analisi regionale a grandezze medie annue, utili alla determinazione della disponibilità idrica di un territorio, sono molto meno numerose di quelle relative alle piene. I risultati dell'analisi svolta sono stati utilizzati per la valutazione della risorsa idrica piemontese in *Viglione* (2007), lavoro di integrazione al Piano di Tutela delle Acque (*Regione Piemonte*, 2004).

Capitolo 1

Analisi di Frequenza Regionale

L'*analisi di frequenza* è la stima di quanto spesso un determinato evento occorre nel tempo. Dal momento che le cause di incertezza legate ai processi fisici che danno luogo agli eventi osservati sono molte, si fa spesso ricorso ad approcci statistici di analisi dei dati, i quali ammettono l'esistenza dell'incertezza legata alla stima e permettono di quantificarne gli effetti.

Le procedure di analisi di frequenza di campioni singoli di dati, che chiameremo *analisi di frequenza locale*, sono oramai ben consolidate. Tuttavia spesso si ha a che fare con l'analisi di molti campioni di dati legati tra loro, quali possono essere, ad esempio, osservazioni meterologiche o ambientali della stessa variabile in diversi punti di misura. Se le frequenze di evento possono essere ritenute simili per i diversi campioni osservati, conclusioni più accurate possono essere raggiunte analizzando tutti i dati insieme piuttosto che i singoli campioni separatamente. Questo approccio è conosciuto come *analisi di frequenza regionale*, perché i campioni di dati analizzati sono tipicamente osservazioni della stessa variabile in siti appartenenti ad una "regione" opportunamente individuata.

1.1 Obiettivi dell'analisi di frequenza

Si supponga di disporre di osservazioni di una qualsivoglia variabile, fatte in alcuni siti di interesse. L'entità dell'evento Q che accade in un certo sito in un certo tempo è trattata come una *variabile aleatoria*, che può potenzialmente

assumere qualsiasi valore tra zero ed infinito. L'analisi di frequenza statistica ha come obiettivo la determinazione della *distribuzione di frequenza*, che specifica quanto spesso i valori di Q si realizzano. Si denoti con $F(x)$ la probabilità che il valore assunto da Q non superi x:

$$F(x) = \Pr[Q \leq x] \ . \tag{1.1}$$

$F(x)$ è la *distribuzione di probabilità cumulata* della variabile Q e costituisce una misura della sua distribuzione di frequenza. La sua funzione inversa, $x(F)$, detta *funzione dei quantili*, esprime l'entità dell'evento in termini della probabilità di non superamento F. Il *quantile* di tempo di ritorno T, Q_T, è l'evento estremo per cui vi è una probabilità $1/T$ che venga superato, nel caso dei massimi, o che non venga raggiunto, nel caso dei minimi, da un qualsiasi singolo evento. Per un evento estremamente alto, nella coda superiore della distribuzione di frequenza, Q_T è dato da:

$$Q_T = x(1 - 1/T), \quad \text{oppure} \quad F(Q_T) = 1 - 1/T \ ; \tag{1.2}$$

per un evento estremamente basso, nella coda inferiore della distribuzione di frequenza, le relazioni corrispondenti sono:

$$Q_T = x(1/T), \quad \text{oppure} \quad F(Q_T) = 1/T \ . \tag{1.3}$$

In Figura 1.1 è rappresentata la curva di frequenza del deflusso annuo (indicato con D) ottenuta per un'ipotetica stazione idrometrica. Oltre alle frequenze di non-superamento F (riportate sull'ascissa), si è indicato anche il tempo di ritorno T per eventi di scarsità (per cui $F = 1/T$).

L'obiettivo dell'analisi di frequenza è di ottenere stime del quantile Q_T per un tempo di ritorno fissato. Quest'ultimo può essere la durata prevista per una struttura (ad esempio $T = 50$ anni), oppure un periodo imposto per legge (ad esempio $T = 10000$ anni in alcune applicazioni di sicurezza delle dighe). Più in generale l'obbiettivo può essere la stima di Q_T per un range di tempi di ritorno, oppure la stima dell'intera distribuzione di frequenza. Per essere considerata utile, una stima dovrebbe non solo essere vicina il più possibile al vero quantile, ma anche essere accompagnata dalla valutazione dell'incertezza ad essa associata.

Se si dispone di misure nel sito di interesse, i dati osservati costituiscono un *campione* di realizzazioni di Q. In molte applicazioni ambientali, e quasi sempre in idrologia, la consistenza del campione, ovvero il numero n delle osservazioni,

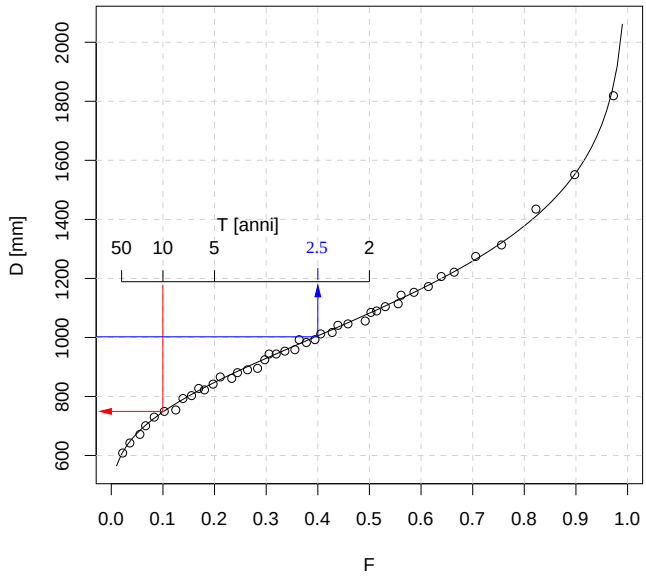

Figura 1.1: Esempio di curva di frequenza del deflusso annuo D (linea continua) adattata ai dati (punti) di una stazione idrometrica; con le frecce si è indicato il valore di deflusso annuo scarso corrispondente al tempo di ritorno di 10 anni ed il tempo di ritorno corrispondente a $D = 1000$ mm.

è insufficiente a rendere efficiente la stima dei quantili. Ovviamente un quantile di tempo di ritorno T non può essere valutato in modo affidabile se $T > n$. Tuttavia, in molte applicazioni ingegneristiche basate su dati a scala annua (ad esempio la precipitazione massima annua, la portata massima annua ecc.), questa condizione si verifica spesso (tipicamente $n < 50$ mentre $T = 100$ o $T = 1000$). Per superare questo problema, sono stati messi a punto molti approcci che fanno uso di fonti di dati alternative o aggiuntive. L'analisi di frequenza regionale è una di queste.

L'analisi di frequenza regionale "aumenta" i dati del sito di interesse usando dati appartenenti ad altri siti ed informazioni aggiuntive, caratteristiche dei siti, facilmente misurabili. Un esempio pratico può chiarire su quale idea si fondi l'analisi di frequenza regionale. Si consideri la variabile piena massime al colmo, ovvero la portata massima istantanea nelle sezioni dei corsi d'acqua. La curva di frequenza empirica delle piene massime al colmo derivata da un campione di dati, andando ad associare ad ogni dato nel campione un valore di frequenza

è una stima della vera curva di frequenza, ma non è certamente una stima esatta perché un campione, benché consistente, non può essere completamente rappresentativo della popolazione. Questo vale anche per le curve di frequenza campionarie ottenute con dati misurati in altri punti. Se tutti i campioni fossero estratti da una medesima popolazione, e se fossero tra loro indipendenti, sarebbe lecito aspettarsi che la stima della curva di frequenza dai dati di tutte le sezioni strumentate sia più efficiente rispetto a quella ottenuta dall'analisi locale nei singoli siti.

Ovviamente nessun gruppo di siti avrà esattamente la stessa distribuzione di frequenza reale dei massimi al colmo. Essa dipende da molti fattori, principalmente dalle caratteristiche di bacino come l'area, la topografia, la geologia, il clima ecc. Se si assume vera quest'ipotesi, la variabilità in un gruppo di curve di frequenza campionarie può essere attribuita a due componenti: una componente casuale dovuta all'aleatorietà del campionamento, e una componente dovuta alla differenza nelle caratteristiche di bacino. Se si potesse distinguere quale parte della variabilità delle curve di frequenza è dovuta alle caratteristiche di bacino, stabilire quali sono queste caratteristiche e relazionarle alle curve stesse, si potrebbe in teoria ottenere la stima esatta della distribuzione di frequenza delle portate massime al colmo non solo nei siti strumentati, ma anche nelle sezioni idrografiche senza dati (per le quali si conoscono le caratteristiche di bacino). In pratica, però, la variabilità totale non può essere separata nettamente nella componente aleatoria ed in quella deterministica. Inoltre, anche se lo fosse, non sarebbe semplice capire quale relazione lega le curve di frequenza alle caratteristiche di bacino. La performance ottenibile con un dato metodo di regionalizzazione dipenderà quindi dalla ripartizione della variabilità dovuta a queste due cause e dalla qualità delle relazioni con le caratteristiche di bacino, ma anche dal soddisfacimento delle ipotesi statistiche su cui il metodo si basa (come ad esempio il grado di indipendenza dei campioni) e, ovviamente, dalla bontà intrinseca del metodo.

1.2 Metodo della grandezza-indice

Uno dei primi contributi nell'ambito dell'analisi di frequenza regionale è quello di *Dalrymple* (1960) che propose la procedura nota come metodo della *piena-indice*, della quale molte varianti e sviluppi sono presenti in letteratura (v.es.

Wiltshire, 1986a,b,c; *Fiorentino et al.*, 1987; *Lettenmaier et al.*, 1987; *Cunnane*, 1988; *Burn*, 1988, 1990; *Hosking & Wallis*, 1993; *Stedinger & Lu*, 1995; *Fill & Stedinger*, 1995, 1998; *Castellarin et al.*, 2001; *Sveinsson et al.*, 2001; *Shu & Burn*, 2004a,b). Ad oggi, il documento probabilmente più completo e ben organizzato sul metodo della grandezza-indice è la monografia di *Hosking & Wallis* (1997) cui si farà spesso riferimento nel seguito. Alla metodologia di Dalrymple si ispirano applicazioni quali il progetto VAPI (VAlutazione delle Portate in Italia) che si ispira al lavoro di *Fiorentino et al.* (1987) e l'FEH ("Flood Estimation Handbook", *Robson & Reed* (1999)) che trattano l'analisi regionale di frequenza delle piene. Nonostante la genesi del metodo, nato per l'analisi statistica delle piene, la procedura, che verrà per questo detta della *grandezza-indice*, può essere usata per qualsiasi tipo di dato.

Si supponga di avere a disposizione i dati di k siti e che per l'i-esimo sito il campione abbia lunghezza n_i e osservazioni Q_{ij} con $j = 1, \ldots, n_i$. Sia $Q_i(F)$, con $0 < F < 1$, la funzione dei quantili nell'i-esimo sito. L'assunzione fondamentale del metodo della grandezza-indice è che i siti formino una *regione omogenea*, ovvero che le distribuzioni di frequenza nei k siti siano identiche a meno di un parametro di scala caratteristico di ogni sito, detto grandezza-indice. Se quest'assunzione è verificata, si può scrivere

$$Q_i(F) = \mu_i \cdot q(F), \quad i = 1, \ldots, k \ , \tag{1.4}$$

dove μ_i è la grandezza-indice ed il fattore $q(F)$ è la *curva di crescita regionale*, una funzione dei quantili adimensionale comune a tutti i siti. La curva di crescita regionale $q(F)$ può essere espressa dalla funzione dei quantili della *distribuzione di frequenza regionale*, la distribuzione comune delle variabili standardizzate Q_{ij}/μ_i.

Nei siti strumentati, la grandezza indice viene stimata come $\hat{\mu}_i = \bar{Q}_i$, la media campionaria, o la mediana campionaria, o un'altra statistica di scala (si veda il Paragrafo 2.1 sull'argomento), dei dati all'i-esimo sito. I dati normalizzati $q_{ij} = Q_{ij}/\hat{\mu}_i$, con $j = 1, \ldots, n_i$ e $i = 1, \ldots, k$, sono la base per la stima della curva di crescita regionale $q(F)$, $0 < F < 1$. Normalmente si assume che la forma di $q(F)$ sia nota a meno di p parametri incogniti $\theta_1, \ldots, \theta_p$, per cui si indica la curva di crescita come $q(F; \theta_1, \ldots, \theta_p)$. Questi parametri possono essere o essere ricondotti al coefficiente di variazione e di asimmetria (skewness) della distribuzione, oppure ai rapporti degli L-momenti τ e τ_3 di Hosking e Wallis (si veda l'Appendice A per la loro definizione formale). Il parametro di scala della distribuzione di frequenza regionale, invece, non è un parametro incognito,

perchè ponendo μ_i nell'Equazione (1.4) pari alla grandezza-indice della distribuzione di frequenza dell'i-esimo sito, si impone che la distribuzione di frequenza regionale abbia parametro di scala uguale ad 1. *Hosking* & *Wallis* (1997) suggeriscono di stimare separatamente gli L-momenti in ogni sito e di valutare i parametri regionali a partire dalla media pesata degli L-momenti locali. Se si indica con $\hat{\theta}_r^{(i)}$ la stima di θ_r nell'i-esimo sito, la stima dell'r-esimo parametro regionale risulterebbe uguale a

$$\hat{\theta}_r^R = \sum_{i=1}^{k} n_i \hat{\theta}_r^{(i)} / \sum_{i=1}^{k} n_i \ . \tag{1.5}$$

Un'altro approccio, utilizzato in questa tesi, consiste nel raggruppare i dati normalizzati q_{ij} in un unico campione di lunghezza $\sum_{i=1}^{k} n_i$, e stimare i parametri regionali $\hat{\theta}_r^R$ di $q(F)$ da quest'unico campione. Il motivo principale di questa scelta è che l'utilizzo di un unico campione, con molti dati, facilita l'individuazione della forma di $q(F)$ con le tecniche di adattamento delle distribuzioni di probabilità ai dati (vedi Paragrafo 3.2). Sostituendo le stime così ottenute in $q(F)$ si ottiene la stima della curva di crescita regionale $\hat{q}(F) = q(F; \hat{\theta}_1^R, \ldots, \hat{\theta}_p^R)$.

La stima dei quantili per l'i-esimo sito è data dalla combinazione delle stime di μ_i e $q(F)$:

$$\hat{Q}_i(F) = \hat{\mu}_i \cdot \hat{q}(F) \ . \tag{1.6}$$

La procedura della grandezza-indice così articolata, si basa su molte assunzioni. Prima di tutto si ipotizza che le osservazioni in ogni sito siano identicamente distribuite e serialmente indipendenti. Queste assunzioni sono plausibili per molti tipi di dati, in particolare per totali annui o eventi estremi, che non sono caratterizzati da variazioni stagionali. L'assunzione che gli eventi osservati nel passato descrivano quello che potrebbe avvenire in futuro è tipica dell'analisi di frequenza. Se ovvie cause di non stazionarietà sono presenti nei dati, ad esempio la costruzione di un invaso, nell'analisi di una serie di dati di portata di piena a valle dell'opera se ne deve tener conto, ad esempio rimuovendo i dati più vecchi, o trasformandoli opportunamente. Per quanto riguarda gli effetti della dipendenza seriale nell'analisi di frequenza di una serie storica di dati, è stato dimostrato che la relativa distorsione della stima dei quantili è ridotta (v.es. *Hosking* & *Wallis*, 1997). Se trend, variazioni periodiche o dipendenza seriale sono presenti in maniera determinante nei dati, occorre trattare questi ultimi con un qualche metodo di analisi statistica delle serie temporali, prima di usarli nella metodologia sopra esposta.

Si ipotizza inoltre che le osservazioni in differenti siti siano tra loro indipendenti, assunzione quasi sempre disattesa se si considerano dati non solo idrologici. Per ogni tipo di dato, sempre si riscontra una qualche correlazione tra i campioni appartenenti a siti geograficamente vicini. Ad esempio eventi meterologici come le tempeste o le siccità tipicamente riguardano aree grandi abbastanza da contenere più siti di misura.

Infine, come già detto, devono valere le ipotesi che le distribuzioni di frequenza nei diversi siti siano identiche a meno di un fattore di scala (ipotesi di omogeneità della regione), e che la forma matematica della curva di crescita regionale sia specificata correttamente, assunzioni che non saranno mai esattamente valide nella pratica. La selezione attenta dei siti da includere in una regione, che è la parte più delicata dell'analisi, può far sì che l'approssimazione fatta sia comunque accettabile. Ad ogni modo è stato dimostrato (v.es *Hosking & Wallis*, 1997) che l'utilizzo dell'analisi regionale permette comunque di dare una stima dei quantili della variabile di interesse più accurata della classica analisi di frequenza locale su singoli campioni.

I capitoli che seguono sono un approfondimento sui due punti costitutivi di questo metodo statistico: la regionalizzazione della grandezza-indice e la regionalizzazione della curva di crescita. Fin qui si è detto che la grandezza-indice μ è un'opportuna statistica di scala della distribuzione di frequenza in ogni sito. Il principale obiettivo dell'analisi regionale, come si è detto, è la valutazione della distribuzione di frequenza di una variabile in siti sprovvisti di dati, in cui non si può stimare direttamente μ. Nel Capitolo 2 si discute di come è possibile fornire una stima della grandezza-indice nei siti dove la variabile non è stata misurata, a partire dalla stessa statistica valutata nei siti strumentati, utilizzando informazioni ovunque disponibili.

Per quanto riguarda la curva di crescita, nel Capitolo 3 si descrivono le metodologie con cui si ottengono le regioni omogenee e si adattano ad esse le distribuzioni di frequenza regionali più opportune. Anche in questo caso vale la considerazione fatta in precedenza: i siti non strumentati devono poter essere associati alle regioni, che devono quindi essere determinabili a partire da informazioni non desumibili dalle osservazioni della variabile analizzata. La scelta di queste "informazioni" è un aspetto alquanto delicato che spesso comporta decisioni soggettive. Nel Capitolo 4 viene discusso un metodo oggettivo di selezione delle informazioni necessarie alla formazione delle regioni per la stima della $q(F)$. Come si è detto, queste regioni devono essere omogenee, ovvero la

distribuzione di frequenza adimensionalizzata della variabile nei diversi siti deve essere pressoché la stessa. La valutazione della correttezza di quest'ipotesi si realizza tramite l'applicazione di test statistici di omogeneità delle distribuzioni di frequenza dei siti della regione. Un confronto tra diversi test di omogeneità utilizzabili nell'analisi di frequenza regionale è descritto nel Capitolo 5.

Capitolo 2

Regionalizzazione della grandezza-indice

L'ipotesi fondamentale del metodo di *Dalrymple* (1960) è che la distribuzione di probabilità della variabile idrologica considerata, in diversi siti appartenenti ad una regione omogenea sia la stessa, a meno del parametro di scala (Paragrafo 1.2). Quest'ultimo viene detto *grandezza-indice* e varia nella regione secondo le caratteristiche (geografiche, fisiche, climatiche, ...) nei siti che la compongono.

Molti approcci metodologici sono disponibili per la stima della grandezza-indice, e le differenze che li contraddistinguono sono legate al grado di informazione disponibile (v.es. *Bocchiola et al.*, 2003). Se si escludono i metodi diretti, che usano le informazioni derivate dalle serie storiche dei dati disponibili nei siti di interesse, i metodi di stima regionale richiedono la conoscenza di informazioni ausiliarie di tipo idrologico e fisico. Questi metodi possono essere suddivisi in due categorie: l'approccio multi-regressivo e la simulazione idrologica. Per entrambi i metodi, lo stimatore migliore è quello che ottimizza un qualche criterio, quali il minimo errore, la minima varianza o la massima efficienza.

Data la sua semplicità, il metodo usato più frequentemente è l'approccio multi-regressivo (v.es. *Kottegoda & Rosso*, 1998) che lega, attraverso equazioni lineari o non-lineari, la grandezza-indice alle caratteristiche dei siti. Se si considera la variabile deflusso, le caratteristiche da utilizzarsi sono quelle di bacino, quali gli indici climatici, i parametri geologici e morfometrici, la copertura del suolo, e così via.

In questo capitolo viene descritto un approccio multi-regressivo per la scelta dei migliori descrittori per la stima della grandezza-indice. Prima, però, è utile

chiarire quale statistica di scala deve essere scelta nei vari casi come grandezza-indice, in base alla robustezza con cui essa può essere stimata.

2.1 Scelta della grandezza-indice

Come grandezza-indice può essere utilizzata una qualsiasi statistica di scala dei campioni, ad esempio la media campionaria (v.es. *Hosking & Wallis*, 1997), la mediana campionaria (v.es. *Robson & Reed*, 1999) o un determinato quantile. Media e mediana sono le statistiche più comunemente adoperate a questo scopo. In questo paragrafo viene esposto il risultato di *Viglione et al.* (2007a) che mostra come, per variabili caratterizzate da un basso coefficiente di asimmetria, la stima della media sia meno distorta di quella della mediana (e viceversa).

Nella formulazione originale del metodo della grandezza-indice di *Dalrymple* (1960), il parametro di scala considerato era la media teorica. Tuttavia il passaggio dalla teoria alla pratica comporta di sostituire alla media teorica la media campionaria. Come è stato chiaramente sottolineato da *Sveinsson et al.* (2001), questa sostituzione non è banale, poiché sostituire la media teorica con la media campionaria può produrre distorsioni evidenti nell'analisi di frequenza regionale. La distorsione introdotta è piuttosto evidente quando la media campionaria non è un "buon stimatore" della media teorica, ovvero quando è distorto o ha una elevata varianza di stima. In questi casi una possibile alternativa potrebbe essere quella di usare la mediana campionaria come grandezza-indice, come proposto ad esempio da *Robson & Reed* (1999). I vantaggi dovuti a questa scelta alternativa sono descritti in quel che segue.

Un'indagine numerica di tipo Monte Carlo è stata condotta per distribuzioni caratterizzate da coefficienti di variazione ed asimmetria diverse (identificate con i rapporti degli L-momenti τ e τ_3 definiti in Appendice A). Per ogni configurazione di τ e τ_3 si sono generati $N = 100000$ campioni lunghi 30 elementi da una distribuzione generalizzata del valore estremo (GEV, vedi Appendice B) con media e mediana note. La distorsione delle stime campionarie di media e mediana è valutata attraverso la radice normalizzata dell'errore quadratico medio,

$$RMSE_{\%} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\bar{x}_i - \mu)^2}}{\mu} \cdot 100 \ , \tag{2.1}$$

dove μ e $\bar{x}_i$ sono, rispettivamente, la media (o mediana) teorica e campionaria. La differenza tra l'$RMSE_{\%}$ per la media e per la mediana è mostrata in Figura

2.1 per tutti i punti presi in considerazione nel piano $\tau - \tau_3$. Dove le differenze

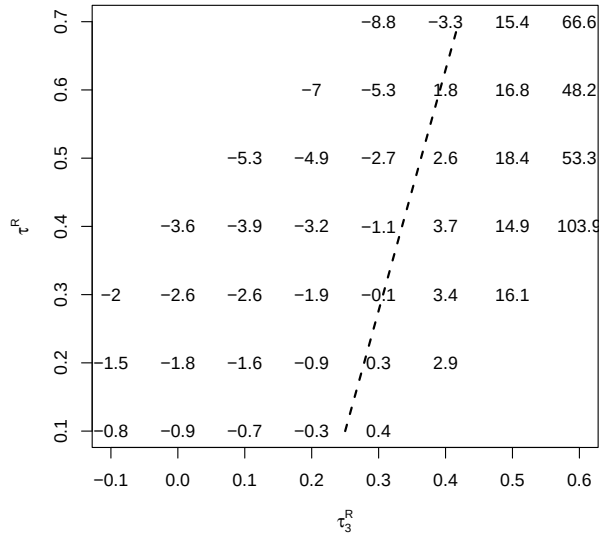

Figura 2.1: Differenza tra l'$RMSE_{\%}$ della media campionaria e l'$RMSE_{\%}$ della mediana campionaria nello spazio $\tau-\tau_3$. La linea tratteggiata indica dove la media campionaria e la mediana campionaria hanno, approssimativamente, lo stesso $RMSE_{\%}$; a destra di questa linea la mediana campionaria è uno stimatore della mediana teorica meno distorto, alla sua sinistra la media campionaria si comporta (leggermente) meglio (da *Viglione et al.*, 2007a).

sono negative, la stima della media è meno distorta della stima della corrispondente mediana, per cui la media può essere considerata una grandezza-indice più affidabile. Risultati analoghi si possono ottenere con distribuzioni differenti dalla GEV. È chiaro dalla Figura 2.1 che le differenze sono quasi trascurabili, eccetto che nella parte più a destra del grafico, corrispondente a campioni fortemente asimmetrici, dove la mediana campionaria si rivela essere uno stimatore considerevolmente migliore della media campionaria. In effetti si sa che la mediana campionaria è molto meno sensibile della media campionaria alla presenza di "outliers", che possono essere facilmente trovati in campioni estratti da distribuzioni molto asimmetriche (*Hampel*, 1974). In generale riteniamo che la Figura 2.1 dimostri i vantaggi di utilizzare la mediana campionaria come grandezza-

indice quando si ritiene che le distribuzioni generatrici siano particolarmente asimmetriche, come nell'analisi regionale delle piene, mentre la media campionaria è migliore, anche se di poco, quando si studiano grandezze non estreme, caratterizzate quindi da distribuzioni poco asimmetriche, come ad esempio il deflusso annuo (vedi Capitolo 6).

2.2 Analisi regressiva lineare

Come si è detto, il metodo usato più frequentemente per la regionalizzazione della grandezza-indice è l'approccio multi-regressivo. L'*analisi regressiva* è una tecnica statistica per investigare e modellare la relazione esistente tra variabili. Nel prosieguo di questo paragrafo verranno analizzati il caso della *regressione lineare semplice*, utile per comprendere il metodo più generale della *regressione lineare multipla*, oltre che alcuni metodi per la verifica di adeguatezza del modello. Si sono riportati solamente i punti fondamentali della tecnica in modo da permettere la comprensione di quanto applicato in questo lavoro; per approfondimenti sulla materia si rimanda ai numerosi testi sull'argomento (ad esempio *Montgomery et al.* (2001)).

2.2.1 Regressione lineare semplice

Supponiamo di voler trovare una relazione $y = f(x)$ tra due grandezze x e y in base alla conoscenza di esse in un numero finito di casi (i punti di Figura 2.2). Solitamente non conosciamo la forma della $f(.)$, quindi, per semplicità, supponiamo che sia di tipo lineare ($y = \beta_0 + \beta_1 x$). Il modello di regressione lineare semplice è quello in cui vi è una sola *variabile esplicativa* x legata alla *variabile dipendente* y da una relazione che, geometricamente, è una linea retta. Il modello viene solitamente espresso come

$$y = \beta_0 + \beta_1 x + \varepsilon \ , \tag{2.2}$$

dove l'intercetta β_0 e la pendenza β_1 sono costanti incognite (*coefficienti* della regressione) ed ε è la componente di errore casuale. Su quest'ultima vengono fatte alcune ipotesi: ovvero che il suo valore atteso sia nullo ($E(\varepsilon) = 0$) e che i valori che assume siano incorrelati tra di loro.

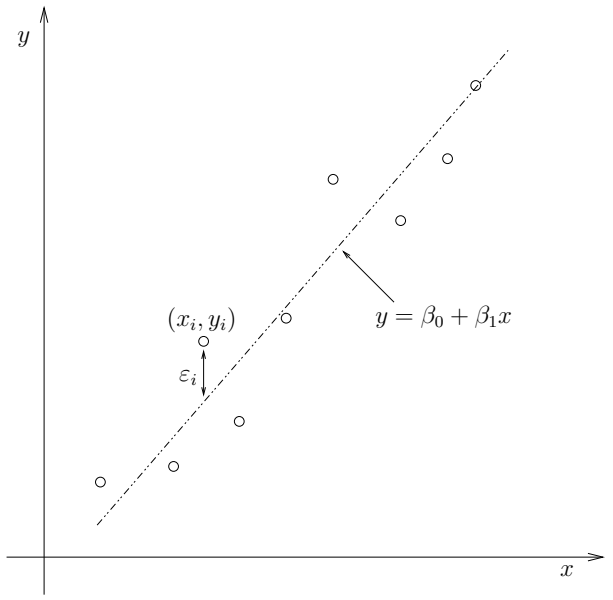

Figura 2.2: Esempio di modello lineare semplice $y = \beta_0 + \beta_1 x$ che lega la variabile indipendente y alla variabile esplicativa x; le coppie di valori (x_i, y_i) misurati (rappresentate dai punti intorno alla retta) sono caratterizzate dagli errori (o residui) ε_i.

Se si suppone di avere n osservazioni indipendenti della coppia (x, y), il metodo dei *minimi quadrati ordinari* (OLS) fornisce una stima dei parametri β_0 e β_1 in modo che la somma dei quadrati delle differenze tra le n osservazioni y_i e la linea retta sia minima. In pratica si vuole minimizzare la funzione

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \ . \tag{2.3}$$

Cercando gli zeri delle derivate della funzione S rispetto ai due coefficienti, si ottiene, dopo alcuni semplici passaggi

$$\begin{cases} \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\ \hat{\beta}_1 &= S_{xy}/S_{xx} \end{cases} \tag{2.4}$$

dove

$$\begin{cases} \bar{x} &= 1/n \sum_{i=1}^{n} x_i \quad ; \quad S_{xx} &= \sum_{i=1}^{n} (x_i - \bar{x})^2 \\ \bar{y} &= 1/n \sum_{i=1}^{n} y_i \quad ; \quad S_{xy} &= \sum_{i=1}^{n} y_i (x_i - \bar{x}) \end{cases} \tag{2.5}$$

sono, rispettivamente, le medie aritmetiche di x_i e di y_i, la somma corretta con la media dei quadrati degli x_i e la somma corretta dei prodotti incrociati tra x_i e y_i.

Le stime della variabile dipendente sono quindi

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \ , \tag{2.6}$$

ed i residui, che rivestono un ruolo importante nel determinare l'adeguatezza del modello regressivo,

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \ , \quad i = 1, 2, ..., n \ . \tag{2.7}$$

Nel caso in cui si utilizzi il Modello (2.4) per la stima dei coefficienti della regressione, oltre alle ipotesi già citate sulla componente ε di errore casuale, occorre che valga l'ulteriore ipotesi di *omoschedasticità*, ovvero che la sua varianza sia costante ($\text{var}(\varepsilon) = \sigma^2 = \text{cost}$). Se valgono tutte queste ipotesi il metodo dei minimi quadrati ordinari fornisce la migliore stima lineare indistorta dei parametri (detti stimatori BLUE, Best Linear Unbiased Estimators).

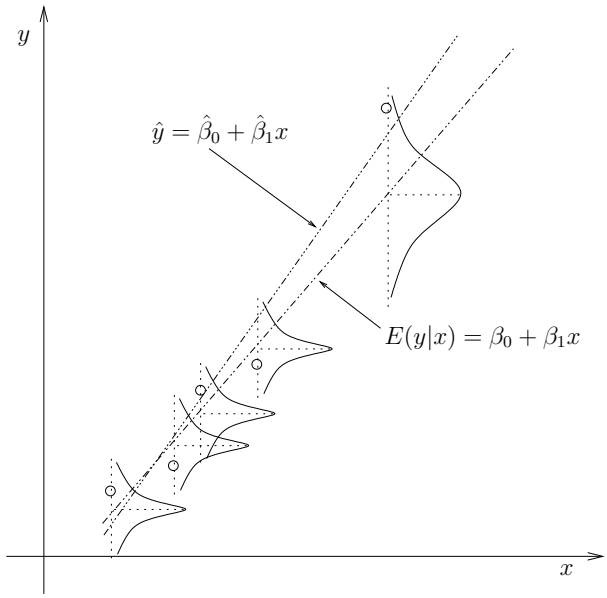

Figura 2.3: Esempio di eteroschedasticità; il modello lineare $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ stimato dai dati con il metodo dei minimi quadrati ordinari è distorto nei confronti del modello "reale" $E(y|x) = \beta_0 + \beta_1 x$ in quanto si è dato lo stesso peso a tutte le osservazioni, quando una di esse, ed in questo caso un'osservazione determinante nella stima della retta poiché staccata da tutte le altre, si conosce con un'incertezza superiore alle altre.

L'importanza dell'ipotesi di omoschedasticità è ben spiegata con l'esempio

di Figura 2.3. Si supponga che esista effettivamente una relazione lineare che lega y ad x, rappresentata dal modello "esatto" $E(y|x) = \beta_0 + \beta_1 x$. I punti (x_i, y_i), ovvero le osservazioni, non ricadono esattamente sulla retta poiché sono affetti da variabilità statistica. Nel caso rappresentato i punti in basso a sinistra si conoscono con un grado di accuratezza superiore al punto in alto a destra (le curve a campana rappresentano la distribuzione di probabilità delle y_i). Se utilizziamo uno stimatore dei coefficienti β_0 e β_1 come il metodo OLS, che non tiene conto dell'eteroschedasticità del campione, rischiamo di stimare un modello $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ distorto (alquanto diverso da quello indistorto $E(y|x)$, come si può vedere in Figura 2.3).

2.2.2 Regressione lineare multipla

Qualora la variabile dipendente y sia messa in relazione con più di una variabile esplicativa, il modello regressivo costruito si dice di regressione lineare multipla ed è del tipo

$$y = \beta_0 + \beta_1 x + \beta_2 x_2 + ... + \beta_{p-1} x_{p-1} + \varepsilon \ , \tag{2.8}$$

dove x_i è una delle $p-1$ variabili esplicative, i β_i sono i p coefficienti della regressione ed ε è il termine di errore che è supposto essere distribuito indipendentemente ed identicamente con media 0 e varianza σ^2.

Nel trattare i modelli di regressione multipla è più conveniente esprimere le equazioni in notazione matriciale, per cui l'Equazione (2.8) diventa

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \ , \tag{2.9}$$

dove

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} , \ \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1,p-1} \\ 1 & x_{21} & x_{22} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n,p-1} \end{bmatrix} , \ \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} , \ \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} .$$

Avendo a disposizione n osservazioni (con n maggiore del numero di parametri p da stimare), ed indicando con y_i la i-esima osservazione della variabile dipendente e con x_{ij} la i-esima osservazione della j-esima variabile esplicativa, come

per l'Equazione (2.3) della regressione lineare semplice il metodo dei minimi quadrati consiste nel minimizzare

$$S(\beta_0, \beta_1, ..., \beta_{p-1}) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p-1} \beta_j x_{ij})^2 \ . \tag{2.10}$$

Se ragioniamo in termini matriciali, si dimostra (v.es. *Montgomery et al.*, 2001) che lo stimatore di $\boldsymbol{\beta}$ col metodo dei minimi quadrati ordinari è

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \ , \tag{2.11}$$

se esiste la matrice inversa $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$ ovvero se le variabili esplicative sono linearmente indipendenti tra di loro. Il vettore delle stime della regressione è

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{y} \ , \tag{2.12}$$

e quello dei residui

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} \ . \tag{2.13}$$

2.3 Stima della grandezza-indice

I metodi multi-regressivi sono i metodi più comunemente utilizzati per la stima della grandezza-indice in siti sprovvisti di dati misurati. Se si considera ad esempio la variabile deflusso (deflusso annuo, portata di piena, ...), l'approccio multi-regressivo lega il deflusso-indice alle caratteristiche di bacino, quali gli indici climatici, i parametri geologici e morfometrici, la copertura del suolo, e così via.

Per la stima della grandezza-indice (che indichiamo qui con y per congruenza alla notazione usata nel Paragrafo 2.2) si possono utilizzare diversi modelli di regressione lineare multipla, ad esempio:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} + \varepsilon \ , \tag{2.14}$$

$$y = \alpha x_1^{\beta_1} x_2^{\beta_2} \ldots x_{p-1}^{\beta_{p-1}} \varepsilon \ , \tag{2.15}$$

oppure

$$y^\lambda = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{p-1} x_{p-1} + \varepsilon \ , \tag{2.16}$$

dove x_i sono le variabili da cui la grandezza-indice è fatta dipendere (nel caso dei deflussi sono i parametri morfoclimatici di bacino) e β_i sono i coefficienti

della regressione. Si noti che anche l'Equazione (2.15) può essere resa lineare nei coefficienti utilizzando una trasformazione logaritmica. Per la stima dei coefficienti delle Equazioni (2.14), (2.15) e (2.16) si utilizza la tecnica dei minimi quadrati ordinari descritta nel Paragrafo 2.2. Il metodo richiede che i descrittori x_i dei siti siano conosciuti con esattezza e che l'unica variabile aleatoria sia la grandezza-indice y. Se si considera ancora il caso dei deflussi, i parametri morfometrici, fisici e geologici dei bacini possono senz'altro essere considerati tali, mentre, a rigore, i parametri climatici sono anch'essi variabili aleatorie. L'approssimazione che si fa nel considerarli fissi è accettabile dal momento che si ritiene siano determinabili con un buon grado di accuratezza, superiore di quello che si presume abbia la stima dei deflussi.

2.3.1 Scelta del migliore modello regressivo

La disponibilità di informazioni legate alla variabile che si vuole stimare è spesso assai consistente. Nel caso del deflusso, ad esempio, le variabili morfoclimatiche che possono essere valutate sui bacini idrografici sono molte. È lecito chiedersi quali variabili esplicative debbano essere utilizzate e quale sia la migliore forma del modello regressivo.

Per ogni tipologia di regressione è utile confrontare i modelli ottenibili da tutte le combinazioni delle variabili morfoclimatiche considerate, per un totale di $k \cdot 2^h$ modelli (dove k è il numero delle tipologie di modello, esemplificate nelle Equazioni (2.14), (2.15) e (2.16), ed h è il numero dei parametri morfoclimatici candidati alla formazione dei modelli).

Significatività dei coefficienti

Innanzitutto si devono escludere quei modelli per i quali anche solo una delle variabili esplicative risulta non significativa in base al test della t di Student. Per spiegare il funzionamento del test si consideri la regressione lineare semplice di Equazione (2.2). Si voglia testare se il coefficiente angolare della retta regressiva β_1 è uguale ad una costante β^*. L'ipotesi nulla e l'ipotesi alternativa siano, rispettivamente, $H_0 : \beta_1 = \beta^*$ e $H_1 : \beta_1 \neq \beta^*$ e gli errori siano indipendenti e distribuiti con distribuzione normale $\varepsilon \sim N(0, \sigma^2)$ (da ciò

consegue che i valori osservati della variabile dipendente sono indipendenti e distribuiti come $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$). Lo stimatore del parametro β_1 dell'Equazione (2.4), calcolato col metodo dei minimi quadrati, è lineare nei confronti dei valori y_i e, come dimostrato in *Montgomery et al.* (2001), distribuito come $\hat{\beta}_1 \sim N(\beta_1, \sigma^2/S_{xx})$, dove S_{xx} è la sommatoria dei quadrati delle differenze rispetto alla media espresso nell'Equazione (2.5). Quindi, in caso di validità dell'ipotesi nulla H_0,

$$Z = \frac{\hat{\beta}_1 - \beta^*}{\sigma/\sqrt{S_{xx}}} \sim N(0,1) \ , \tag{2.17}$$

ovvero Z è distribuita secondo una normale standard.

Se conoscessimo σ^2 potremmo usare Z per testare H_0, ma, tipicamente, la varianza dell'errore non è nota. Si può però dimostrare che la statistica

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2} \ , \tag{2.18}$$

ovvero è distribuita secondo una distribuzione chi-quadro con $n-2$ gradi di libertà, e che le stime di σ^2 e β_1 sono indipendenti. In considerazione delle proprietà della distribuzione t di Student, si può dire che

$$T = \frac{\hat{\beta}_1 - \beta^*}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t_{n-2} \ . \tag{2.19}$$

ovvero che la statistica T è distribuita come una t di Student con $n-2$ gradi di libertà.

L'Equazione (2.19) diventa quindi la statistica test da utilizzarsi nel caso in cui si voglia testare la significatività del parametro β_1: si pone β^* uguale a zero, si calcola T e si va a vedere (ad esempio utilizzando le tabelle riportate su *Kottegoda & Rosso* (1998)) quanto vale la corrispondente t_{limite}, relativa al livello di significatività prescelto. Se $T < t_{limite}$, il parametro non è distinguibile da zero e la variabile esplicativa non deve essere utilizzata nella regressione poiché non è significativamente legata alla variabile dipendente. Perché il test definito dalla statistica T dell'Equazione (2.19) possa essere usato, occorre a rigore che gli errori del modello regressivo siano distribuiti normalmente. Nella pratica si è riscontrato che per deboli "non-normalità" il test risulta essere comunque significativo (v.es. *Montgomery et al.*, 2001).

Nel caso della regressione lineare multipla di Equazione (2.8), il procedimento è analogo a quello presentato per la regressione lineare semplice, e viene utilizzato per valutare la significatività di ognuno dei parametri della regressione. In questo modo si possono eliminare una o più delle variabili esplicative scelte se queste

non danno un contributo significativo alla somma dei quadrati della regressione. Come per l'Equazione (2.19) si dimostra che

$$T = \frac{\hat{\beta}_i - \beta_i^*}{\sqrt{\hat{\sigma}^2 c'_{ii}}} \sim t_{n-p} \ , \tag{2.20}$$

dove c'_{ii} con $i = 1, ..., p-1$ sono gli elementi della diagonale della matrice $(\boldsymbol{X}^T \boldsymbol{X})^{-1}$. Il test definito con l'Equazione (2.20) è da considerarsi solo come test parziale sul parametro in analisi poiché la stima di questo dipende da tutte le variabili esplicative usate nel modello.

Coefficiente di determinazione

Una volta verificata la significatività dei coefficienti delle variabili indipendenti occorre valutare la capacità descrittiva di ogni regressione. Dall'analisi della varianza di una regressione lineare semplice si ottiene l'identità

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2 \ , \tag{2.21}$$

che nel caso della regressione lineare multipla si scrive

$$\left[\boldsymbol{y}^T\boldsymbol{y} - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right] = \left[\hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y} - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right] + \left[\boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}^T \boldsymbol{X}^T \boldsymbol{y}\right], \tag{2.22}$$

dove, in entrambi i casi, il termine a sinistra è la somma corretta con la media dei quadrati delle osservazioni (SS_T) e misura la variabilità totale delle osservazioni, mentre i due termini a destra misurano, rispettivamente, la variabilità delle osservazioni y_i ritrovata nella regressione (SS_R) e la variabilità residua inspiegata (SS_{Res}).

Si dice *coefficiente di determinazione* la quantità

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_{Res}}{SS_T}, \tag{2.23}$$

che rappresenta la proporzione di variazione spiegata dalla variabile esplicativa x. Il valore di R^2 è minore o uguale a 1 e, più la variabilità di y è spiegata dal modello di regressione, più tale valore è elevato.

Se si vogliono confrontare modelli regressivi con un numero diverso di variabili esplicative, il coefficiente di determinazione R^2 non deve essere utilizzato in

quanto il suo valore aumenta sempre quando si aggiunge una variabile esplicativa. Al suo posto si utilizza li *coefficiente di determinazione corretto*

$$R^2_{adj} = 1 - \frac{SS_{Res}/(n-p)}{SS_T/(n-1)}. \tag{2.24}$$

Poiché $SS_{Res}/(n-p)$ è la media quadratica dei residui e $SS_T/(n-1)$ è costante a prescindere dalle variabili considerate nel modello, R^2_{adj} cresce solamente, qualora si aggiunga una variabile esplicativa, se la media quadratica dei residui si riduce. Per questo motivo R^2_{adj} è utile per evitare di aggiungere al modello termini non necessari.

Cross-validazione

Il coefficiente di determinazione R^2_{adj} può essere utilizzato per scegliere il miglior modello tra quelli di una classe (Equazioni (2.14), (2.15) oppure (2.16)) ma non può essere usato per confrontare modelli di differente natura. A questo scopo si consiglia di applicare un metodo di cross-validazione, calcolando la radice dell'errore quadratico medio (RMSE) dei residui $\hat{y}'_i - y_i$, dove $\hat{y}'_i$ è il valore stimato della i-esima grandezza-indice y_i, basato però sulla regressione ottenuta utilizzando tutte le osservazioni eccettuata la i-esima. L'$\mathrm{RMSE}_{\mathrm{cv}}$ è definito come:

$$\mathrm{RMSE}_{\mathrm{cv}} = \sqrt{\frac{1}{n}\sum_{1}^{n}(\hat{y}'_i - \tilde{y}_i)^2}\ . \tag{2.25}$$

Intervalli di confidenza e di predizione

Oltre al coefficiente di determinazione R^2_{adj} e all'$\mathrm{RMSE}_{\mathrm{cv}}$, è utile accompagnare il modello lineare selezionato con la rappresentazione degli intervalli di confidenza delle stime, per associare ad esso l'incertezza che lo contraddistingue. Nel caso in cui i modelli siano stati ricavati con il metodo OLS, questi intervalli si ottengono facilmente. Si consideri il modello di regressione lineare semplice dell'Equazione (2.2) e si voglia stimare la risposta media $E(y)$ per un particolare valore della variabile esplicativa x_0:

$$\hat{E}(y|x_0) = \hat{\mu}_{y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0\ . \tag{2.26}$$

Per ottenere l'*intervallo di confidenza della stima* del $100(1-\alpha)\%$ di $E(y|x_0)$, si dimostra che (v.es. *Montgomery et al.*, 2001) la variabile $\hat{\mu}_{y|x_0}$ è distribuita normalmente e che

$$\mathrm{var}(\hat{\mu}_{y|x_0}) = \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] , \tag{2.27}$$

per cui, per le proprietà della distribuzione t di Student:

$$\begin{cases} E(y|x_0) & \geq & \hat{\mu}_{y|x_0} - t_{\alpha/2,n-2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}} \right)} \\ E(y|x_0) & \leq & \hat{\mu}_{y|x_0} + t_{\alpha/2,n-2} \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}} \right)} \end{cases} \tag{2.28}$$

Un esempio di intervalli di confidenza della media della stima è riportato in Figura 2.4. Si noti che la larghezza dell'intervallo di confidenza di $E(y|x_0)$ è

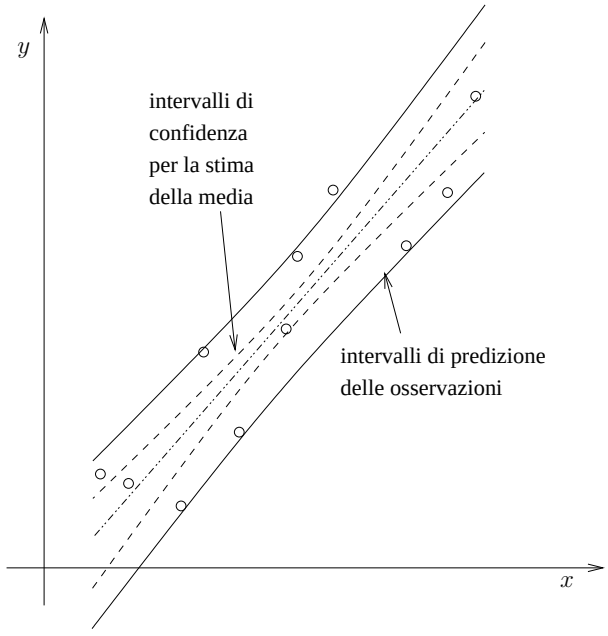

Figura 2.4: Intervalli (al 95% di significatività) di confidenza della stima $E(y|x_0)$ e di predizione per nuove osservazioni y_0.

funzione di x_0 ed è minimo per $x_0 = \bar{x}$. Il fatto che la stima migliore si abbia per un valore di x al centro dei dati utilizzati per il modello, e che vada deteriorandosi verso i bordi dello spazio delle x, è ragionevole anche da un punto di vista intuitivo.

Nel caso invece della regressione lineare multipla di Equazione (2.8) si definenisca il vettore delle variabili esplicative

$$\boldsymbol{x_0} = [1, x_{01}, x_{02}, ..., x_{0,p-1}] \ . \tag{2.29}$$

Poiché, analogamente all'Equazione (2.26), $\hat{E}(y|\boldsymbol{x_0}) = \boldsymbol{x_0}\hat{\boldsymbol{\beta}}$ e, analogamente all'Equazione (2.27), $\text{var}(\hat{E}(y|\boldsymbol{x_0})) = \sigma^2 \boldsymbol{x_0}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_0}^T$, l'intervallo di confidenza del $100(1-\alpha)\%$ di $E(y|\boldsymbol{x_0})$ vale

$$\begin{cases} E(y|\boldsymbol{x_0}) & \geq & \hat{E}(y|\boldsymbol{x_0}) - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \boldsymbol{x_0}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_0}^T} \\ E(y|\boldsymbol{x_0}) & \leq & \hat{E}(y|\boldsymbol{x_0}) + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 \boldsymbol{x_0}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_0}^T} \end{cases} \ . \tag{2.30}$$

Gli intervalli sono diversi se si vuole valutare la variabilità di predizione della variabile dipendente per un determinato valore di quella esplicativa. Si consideri, per semplicità, il caso della regressione lineare semplice, per cui

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \ . \tag{2.31}$$

Poiché la varianza della variabile aleatoria $\psi = y_0 - \hat{y}_0$ vale

$$\text{var}(\psi) = \text{var}(y_0 - \hat{y}_0) = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right] \ , \tag{2.32}$$

se si usa $\hat{y}_0$ per predire y_0, l'errore quadratico di ψ è una statistica appropriata sulla quale basare il cosiddetto *intervallo di predizione*. L'intervallo di predizione del $100(1-\alpha)\%$ su una predizione nella variabile esplicativa x_0 è

$$\begin{cases} y_0 & \geq & \hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \\ y_0 & \leq & \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0-\bar{x})^2}{S_{xx}}\right)} \end{cases} \tag{2.33}$$

Un esempio di intervalli di predizione di nuove osservazioni è riportato in Figura 2.4. Analogamente, nel caso di una regressione lineare multipla, l'intervallo di predizione al $100(1-\alpha)\%$ di una predizione $\hat{y}_0 = \boldsymbol{x_0}\hat{\boldsymbol{\beta}}$ della variabile esplicativa $\boldsymbol{x_0}$ è

$$\begin{cases} y_0 & \geq & \hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + \boldsymbol{x_0}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_0}^T)} \\ y_0 & \leq & \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + \boldsymbol{x_0}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x_0}^T)} \end{cases} \ . \tag{2.34}$$

In pratica, per la selezione del migliore modello regressivo si procede nel modo seguente (un esempio è descritto nel Paragrafo 6.2):

- si scelgono k tipologie di modelli, tra quelli esemplificati dalle Equazioni (2.14)-(2.16). Se h è il numero delle variabili esplicative considerate, il numero dei modelli da confrontare è $k \cdot 2^h$;

- da questi si escludono tutti modelli in cui anche solo uno dei parametri risulta non essere significativo in base al test di Student;

- per ogni tipologia, si scelgono i migliori modelli in base al coefficiente di determinazione R^2_{adj};

- tra questi si sceglie infine il modello o i modelli caratterizzati dal minore errore $\mathrm{RMSE}_{\mathrm{cv}}$;

- per associare a questi una misura di incertezza si identificano gli intervalli di predizione per nuove osservazioni.

2.3.2 Verifiche di adeguatezza del modello

Per i modelli selezionati deve essere condotta la verifica delle assunzioni implicite dell'analisi regressiva lineare: che la relazione tra la variabile dipendente y ed i regressori sia lineare, almeno in prima approssimazione, che non ci sia correlazione lineare tra i regressori (assenza di multicollinearità) e che i residui ε soddisfino alcuni requisiti. In particolare, si richiede che la loro media sia nulla (il che è automaticamente garantito dalla procedura dei minimi quadrati), che la loro varianza sia costante (omoschedasticità), che siano tra loro incorrelati e che siano distribuiti normalmente (quest'ultima assunzione è necessaria per la validità del test di significatività e degli intervalli di confidenza e di predizione). Una decisa violazione di queste assunzioni può dare luogo alla formulazione di un modello instabile, nel senso che un campione differente di osservazioni della stessa variabile potrebbe dar luogo ad un modello completamente differente. La violazione delle assunzioni elencate sopra non può essere valutata da statistiche "globali" quali R^2 o la t di Student. In questo paragrafo vengono presentati alcuni metodi (grafici e statistici) che possono aiutare a rilevare incongruenze con le assunzioni fatte.

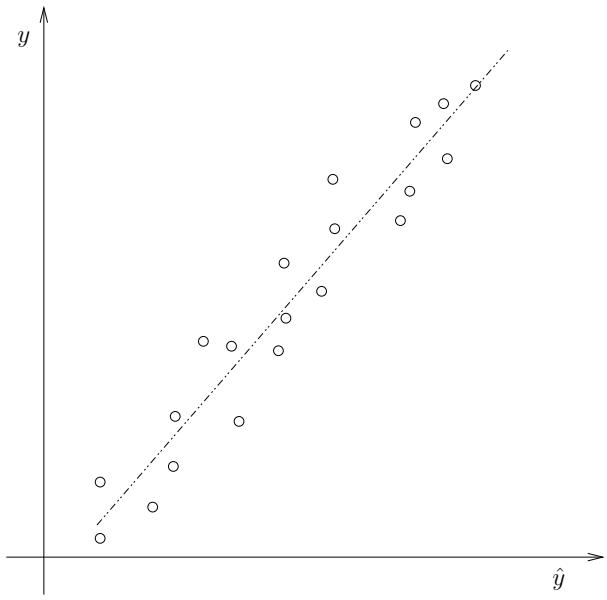

Figura 2.5: Esempio di diagramma diagnostico $\hat{y} - y$, dove y sono i valori osservati della variabile dipendente e $\hat{y}$ i valori stimati con il modello lineare.

Il diagramma diagnostico più immediato è quello che lega variabile dipendente y e variabile esplicativa x (v.es. Figura 2.2) che permette di interpretare immediatamente la bontà del modello. Nel caso delle regressioni multiple, non potendo ricorrere alla rappresentazione nello spazio multidimensionale, si utilizza il diagramma tra y e la stima $\hat{y}$ che dà il modello, come in Figura 2.5. Con questo tipo di rappresentazione, anomalie quali punti che si discostano particolarmente dalla retta, e che corrispondono a casi in cui il modello dà una stima estremamente diversa dalla misura, sono facilmente identificabili.

L'analisi grafica dei residui $\hat{\varepsilon}_i$ della regressione nei confronti dei valori stimati $\hat{y}_i$ può essere molto utile al riconoscimento di alcuni tipi comuni di inadeguatezza del modello. I residui devono essere rappresentati con i valori stimati $\hat{y}_i$ e non con quelli misurati y_i perché normalmente $\hat{\varepsilon}_i$ e y_i sono correlati tra loro. Se la rappresentazione assomiglia al grafico (a) di Figura 2.6, non ci sono difetti evidenti nel modello. I casi (b) e (c), invece, fanno pensare ad una possibile eteroschedasticità dei residui (che, nel caso (b), sembrano avere varianza proporzionale a y). Come già sottolineato nel Paragrafo 2.2.1, Figura 2.3, l'eteroschedasticità (non-costanza della varianza) dei residui implica che la procedura

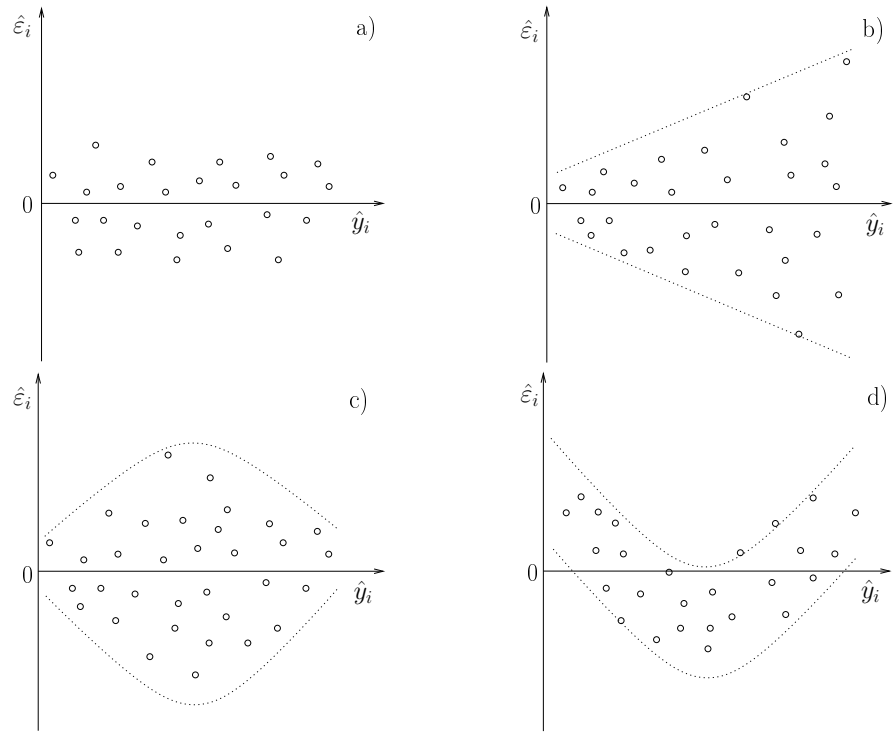

Figura 2.6: Grafici dei residui nei confronti delle stime della regressione. Si sono esemplificate 4 situazioni differenti: a) situazione soddisfacente; b) e c) possibile eteroschedasticità dei residui; d) non-linearità.

OLS non è il migliore stimatore lineare non-distorto (BLUE) dei coefficienti del modello. Per rilevare l'eteroschedasticità si può utilizzare il test di omoschedasticità di *Harrison & McCabe* (1979). La statistica test di Harrison-McCabe è la frazione della somma dei quadrati dei residui corrispondenti ai dati prima di un fissato punto di separazione (ad esempio la frazione della somma dei quadrati dei residui corrispondenti alla prima metà dei dati ordinati). Nell'ipotesi H_0, la statistica test dovrebbe essere pari, all'incirca, a questa frazione, nel nostro caso vicina a 0.5. L'ipotesi nulla è rigettata se la statistica è troppo piccola. Nel caso in cui l'eteroschedasticità venisse confermata, si dovrebbe ricorrere al metodo dei *minimi quadrati pesati* (WLS) oppure ad un'opportuna trasformazione della variabile dipendente, Equazione (2.16), o dei regressori, o di entrambi. Se, oltre al problema dell'eteroschedasticità, le ossevazioni delle coppie (x, y) fossero correlate tra di loro, il metodo che occorrerebbe usare è quello dei *minimi quadrati*

generalizzati (GLS). Una configurazione curva come quella del caso (d) di Figura 2.6 è indice, invece, di non-linearità. Questo può significare che il modello ha bisogno di altre variabili esplicative oppure che si deve ricorrere a un modello non-lineare.

Questi grafici permettono quindi di valutare possibili deviazioni da tutte le assunzioni fatte sui residui, eccetto quella di normalità per la quale si può ricorrere alla rappresentazione in *carta probabilistica normale*. Senza entrare nei particolari (per dettagli si veda *D'Agostino & Stephens*, 1986), questa è un grafico costruito in modo che la funzione dei probabilità cumulata di Gauss (la normale) viene rappresentata su di esso come una linea retta. Se $\hat{\varepsilon}_{[1]}, \hat{\varepsilon}_{[2]}, ..., \hat{\varepsilon}_{[n]}$

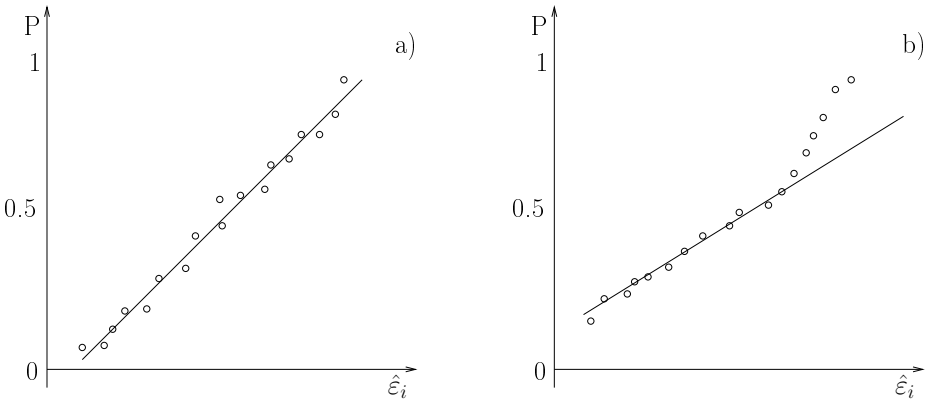

Figura 2.7: Grafici dei residui in carta probabilistica normale. Si sono esemplificate 2 situazioni differenti: a) situazione di normalità; b) non-normalità perché il campione è caratterizzato da asimmetria positiva.

sono i residui ordinati in senso crescente, la loro rappresentazione nei confronti della probabilità cumulata $P_i = (i-1/2)/n$, $i = 1, 2, ..., n$, in carta probabilistica normale dovrebbe giacere approssimativamente su una linea retta (grafico (a) di Figura 2.7). Il grafico (b), al contrario, presenta uno scostamento dalla normale dovuto ad asimmetria positiva. La normalità dei residui è richiesta per la validità del test di significatività (il test t di Student) e per la stima degli intervalli di confidenza/predizione. Per individuare l'assenza di normalità i residui possono essere rappresentati in carta probabilistica normale, e si può utilizzare un test di normalità, ad esempio il test di Anderson-Darling (v.es. *Laio*, 2004). Il test di Anderson-Darling è basato sulla Distribuzione di Frequenza Empirica (EDF)

ed ha statistica test:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}(2i-1)[\ln(p_{(i)}) + \ln(1-p_{(n+1-i)})] \ , \tag{2.35}$$

dove $p_{(i)} = \Psi((x_{(i)} - \bar{x})/s)$. Ψ è la distribuzione di probabilità cumulata della distribuzione normale standard, e $\bar{x}$ e s sono la media e lo scarto quadratico medio del campione di dati (in questo caso i residui della regressione). Il valore della probabilità associata al test può essere valutato per la statistica modificata $Z = A(1.0 + 0.75/n + 2.25/n^2)$ secondo la Tabella 4.9 in *D'Agostino* & *Stephens* (1986).

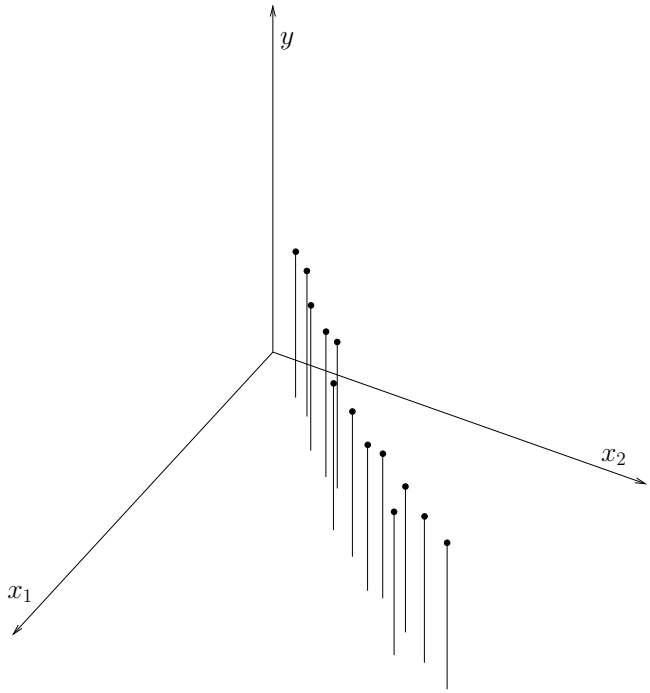

Figura 2.8: Set di dati affetti da multicollinearità.

L'ultimo aspetto che occorre prendere in considerazione è la possibile eccessiva correlazione tra le variabili esplicative (multicollinearità) utilizzate in una regressione lineare multipla. La multicollinearità influenza la procedura dei minimi quadrati determinando problemi di stima dei coefficienti. Il caso rappresentato in Figura 2.8 può essere utile a spiegare gli effetti dovuti alla multicollinearità. Adattare un modello regressivo ai dati (x_1, x_2, y) di Figura 2.8 è analogo a far

passare un piano inclinato tra i punti. Ovviamente l'inclinazione di questo piano sarà molto instabile e sensibile a piccoli cambiamenti dei punti. Inoltre, benché il modello può predire abbastanza bene y in punti (x_1, x_2) vicini a quelli dei dati, qualsiasi estrapolazione al di fuori di questi sarà verosimilmente poco affidabile.

Una semplice statistica adeguata a misurare la presenza di multicollinearità è il *variance inflation factor* (v.es. *Montgomery et al.*, 2001):

$$\text{VIF} = (1 - R_j^2)^{-1} \ , \tag{2.36}$$

dove R_j^2 è il coefficiente di determinazione della regressione lineare tra la variabile indipendente x_j e i $p-1$ regressori rimanenti. Dall'esperienza pratica si desume che se uno dei VIF arriva a valori dell'ordine di 5 o 10, la possibilità che i coefficienti della regressione siano stati stimati male a causa di multicollinearità è elevata.

Il migliore modello lineare deve quindi essere accompagnato dalla verifica di adeguatezza di tutte queste ipotesi. Ad esempio nel Paragrafo 6.2 si sono rappresentati graficamente i residui nei confronti delle stime e in carta probabilistica normale, e si sono riportati i risultati dei test (di Harrison-McCabe per l'omoschedasticità, di Anderson-Darling per la normalità ed il VIF per la non-multicollinearità).

Capitolo 3

Regionalizzazione della curva di crescita

L'identificazione delle regioni omogenee per la stima delle curve di crescita è sicuramente la fase più complessa dell'analisi di frequenza regionale e spesso richiede di fare scelte soggettive. L'obbiettivo è quello di formare gruppi di siti che soddisfino, almeno in prima approssimazione, la condizione di omogeneità, ovvero che le distribuzioni di frequenza dei siti siano identiche a meno della grandezza-indice.

La distribuzione di frequenza della quantità Q di interesse nei siti non è misurata direttamente. I dati di cui si dispone sono statistiche calcolate a partire dalle misure di Q, oppure altri descrittori (o caratteristiche) del sito. Queste caratteristiche possono essere ad esempio la posizione geografica, la quota, le caratteristiche morfometriche e climatiche del bacino sotteso (se la variabile di interesse è il deflusso), ecc. È di gran lunga preferibile utilizzare, nella formazione delle regioni, le caratteristiche del sito, piuttosto che le statistiche legate a Q, per diversi motivi. Tra questi il principale è che se si usano le statistiche di Q, i risultati dell'analisi regionale non possono essere utilizzati per siti senza dati, in quanto risulterebbe impossibile assegnare uno di questi siti ad una regione per il semplice fatto che non si possono stimare le statistiche di Q su di esso. Inoltre se il test dell'omogeneità delle regioni è basato su tali statistiche la sua integrità è compromessa dal fatto di utilizzare le stesse statistiche per la formazione delle regioni.

Per quanto riguarda la suddivisione in regioni, tra i metodi più utilizzati si annovera la ripartizione dei siti in *regioni disgiunte* (v.es. *Hosking & Wallis*,

1997; *Viglione et al.*, 2006). Nel metodo delle regioni disgiunte, le regioni sono ottenute una volta per tutte e separate da confini fissi nello spazio delle caratteristiche usate per la suddivisione, che chiameremo *variabili di classificazione* (vedi Paragrafo 3.1). Un approccio alternativo è quello di definire per ogni sito di interesse, volta per volta, una regione che contiene quei siti i cui dati possono essere vantaggiosamente utilizzati per la stima della distribuzione di frequenza nel sito stesso. Questo approccio è detto della *regione di influenza* (*Burn*, 1990, ROI) ed è stato utilizzato per la redazione del Flood Estimation Handbook inglese (*Robson & Reed*, 1999). Si consideri l'esempio riportato in Figura 3.1 in cui i siti sono rappresentati nello spazio delle variabili di classificazione. Con i

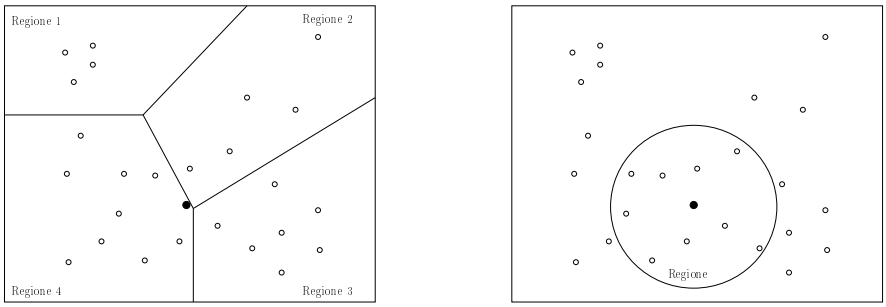

Figura 3.1: Differenze tra regioni disgiunte e regione di influenza: nel grafico a sinistra si sono individuate quattro regioni omogenee disgiunte a partire dai dati dei siti rappresentati con i pallini vuoti; nel grafico a sinistra si è costruita una regione di influenza intorno al sito per cui si vuole stimare la curva di crescita.

pallini vuoti si sono indicati i siti provvisti di dati, mentre con il pallino pieno un ipotetico sito non monitorato per il quale si vuole stimare la curva di crescita relativa alla grandezza misurata negli altri siti. Nel primo caso (grafico a sinistra) si dispone di una suddivisione in regioni omogenee ottenuta con il metodo della grandezza-indice a regioni disgiunte. Il sito di interesse ricade nella Regione 4 per cui si utilizzerà la curva di crescita ottenuta dai dati di quella regione. Nel caso rappresentato, però, il pallino pieno si trova al confine con le Regioni 2 e 3. Sarebbe bastata una piccola differenza nelle variabili di classificazione ad esso associate per determinarne l'appartenenza ad un'altra regione.

Questo problema non si presenta nel caso del metodo ROI (grafico a destra di Figura 3.1), con il quale il sito di interesse è sempre al centro della regione omogenea a cui appartiene. La procedura ROI è quindi particolarmente effica-

ce quando è richiesta una transizione continua tra le regioni. Infatti essa non è soggetta al problema di assegnare siti adiacenti nello spazio delle variabili di classificazione, e quindi presumibilmente aventi distribuzioni di frequenza simili, a due diverse regioni. L'utilizzo dei risultati di un'analisi regionale basata sulla procedura ROI è però molto meno immediato rispetto al metodo delle regioni disgiunte. Questo perchè l'utente deve ogni volta determinare da sè la regione omogenea per il sito di suo interesse ed adattare una distribuzione di probabilità alla curva di crescita regionale. D'altronde un confronto generale tra le due metodologie non è stato ancora fatto. Nel prosieguo del capitolo si farà riferimento al metodo delle regioni disgiunte, che è stato utilizzato nell'applicazione dell'analisi di frequenza regionale presentata nel Capitolo 6.

3.1 Formazione delle regioni

Le procedure proposte in letteratura per la formazione delle regioni di siti simili tra loro sono molte. In passato si sono scelte regioni di siti contigui geograficamente, a volte seguendo addirittura i limiti amministrativi. Il termine "regione" potrebbe infatti far pensare ad un insieme di siti vicini tra loro, ma la prossimità geografica non è necessariamente indice di similarità tra distribuzioni di frequenza. Questo aspetto è molto importante per variabili cumulate come, ad esempio, il deflusso. Si considerino due sezioni di un corso d'acqua poste una subito prima e l'altra subito dopo un importante confluenza. Benché i due punti siano molto vicini tra loro, il deflusso è molto diverso perché molto diversi sono i bacini idrografici che lo determinano. Anche per le variabili puntuali, comunque, l'approccio della contiguità geografica è estremamente arbitrario e soggettivo.

Il metodo più usato oggi è quello di associare ad ogni sito considerato nello studio un set di caratteristiche (e preferibilmente non statistiche di Q), e di dividere o aggregare i siti in base alla similarità nello spazio di queste caratteristiche. Esse possono essere caratteristiche geografiche (latitudine e longitudine) ma anche di tipo fisico, pedologico, morfometrico e climatico. Un indubbio vantaggio che si ha se si determinano gruppi geograficamente dispersi è che le distribuzioni di frequenza nei differenti siti possono essere ritenute con maggiore sicurezza scorrelate tra loro, il che riduce la variabilità dell'eventuale stima dei quantili. La formazione delle regioni può essere eseguita utilizzando metodi standard di *cluster analysis* basati sulle caratteristiche scelte. L'aspetto più delicato, a nostro

avviso, è appunto la scelta di queste caratteristiche, le variabili di classificazione, a cui è dedicato il Capitolo 4. Supponiamo per il momento di conoscere queste variabili e concentriamoci sul come utilizzarle per la formazione delle regioni.

I gruppi (o cluster) devono contenere siti le cui caratteristiche siano simili. Molti algoritmi di cluster analysis misurano la similarità attraverso la distanza Euclidea nello spazio delle variabili di classificazione. La misura di distanza euclidea si esprime con:

$$d_{ij} = \sqrt{\frac{1}{p}\sum_{h=1}^{p}(x_{hi} - x_{hj})^2} \ , \tag{3.1}$$

dove p è il numero delle variabili di classificazione e x_{hi} è il valore della variabile h-esima dell'i-esima entità, standardizzata in modo che il campione degli elementi di tale variabile abbia media 0 e varianza 1. Questa riscalatura, che fa sì che tutte le caratteristiche abbiano la stessa variabilità, dà a tutte le variabili di classificazione lo stesso peso nella formazione delle regioni. Questo potrebbe non essere appropriato, in quanto è probabile che alcune caratteristiche abbiano un'influenza maggiore sulla forma della distribuzione di frequenza, e dovrebbero essere considerate con un maggiore peso nella procedura di formazione dei gruppi. Scegliere i giusti pesi è però un problema tutt'altro che semplice: un possibile approccio è suggerito al Paragrafo 4.2, ma non è stato usato nell'applicazione descritta nel Capitolo 6.

Se si segue la procedura delle regioni disgiunte, la tecnica di cluster analysis che può essere utilizzata a questo scopo è una metodologia mista, costituita da una prima suddivisione dei bacini con l'algoritmo gerarchico di *Ward* (1963) e da una successiva rifinitura dei gruppi con un metodo di minimizzazione della dispersione entro i cluster. L'algoritmo di Ward è di tipo agglomerativo, ovvero parte da una situazione in cui ogni individuo costituisce un cluster. Ad ogni passo dell'analisi si considera l'unione di ogni possibile coppia di cluster e si uniscono quelli la cui fusione determina la minima perdita di informazione, che può essere misurata come la somma delle deviazioni quadratiche di ogni punto dal baricentro del cluster a cui appartiene. L'algoritmo di Ward è utile al nostro scopo perché è costruito in modo da generare gruppi compatti e con un numero di elementi confrontabile (in pratica non dovrebbe dar luogo a gruppi troppo piccoli, poco adatti all'analisi regionale). Come tutti gli algoritmi gerarchici (vedi Figura 3.2), però, anche quello di Ward ha l'inconveniente di non ammettere riallocazioni di elementi tra i gruppi, per cui non è detto che la configurazione

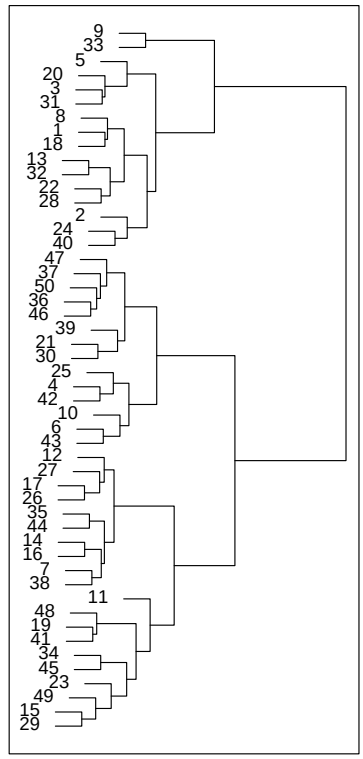

Figura 3.2: Dendogramma di agglomerazione di elementi a mezzo di un algoritmo di cluster analysis di tipo gerarchico. Partendo dalla situazione in cui ogni individuo è considerato singolarmente (parte sinistra del grafico), ad ogni passo dell'analisi si unisce una coppia di individui (o di cluster) fino ad ottenere un unico gruppo (parte destra del grafico).

finale sia effettivamente quella ottimale. Per ovviare a questo inconveniente si può utilizzare, a valle dell'algoritmo di Ward, una tecnica di riallocazione degli elementi.

La procedura viene inizializzata con una suddivisione di partenza, nel nostro caso quella in k cluster ottenuta con Ward. Per questa configurazione può essere calcolata la statistica:

$$W = \sum_{i=1}^{k} \left(\sum_{j=1}^{n_i} \delta_{ij}^2 \right) \ , \tag{3.2}$$

dove δ_{ij} è la distanza euclidea tra l'elemento j-esimo dell'i-esimo gruppo e il baricentro dell'i-esimo cluster ed n_i è il numero di elementi dello stesso. A passi successivi si valuta se lo spostamento di un elemento tra due gruppi comporta una diminuzione di W, nel qual caso viene applicata la correzione. Ciò avviene finché non si giunge ad una configurazione in cui non esistono modifiche in grado di diminuire ulteriormente la dispersione all'interno dei gruppi. Tale metodo comporta la riallocazione dei punti finché tutti gli individui di ogni gruppo sono più vicini (come distanza Euclidea) al centro del proprio cluster che al centro degli altri. Nel caso in cui le variabili di classificazione siano due, ciò comporta che i gruppi possano essere delimitati da "poligoni di Thiessen", rendendo molto semplice la definizione del criterio di assegnazione di una nuova entità ad un cluster (Figura 3.3).

Non è detto che esistano raggruppamenti distinti di siti che soddisfano la condizione di omogeneità. Più realisticamente la forma della distribuzione di frequenza varia in maniera continua nello spazio delle variabili di classificazione, e l'obbiettivo è quello di formare gruppi di siti entro i quali la variabilità delle caratteristiche, e quindi delle distribuzioni di frequenza, è così piccola da far sì che l'analisi di frequenza regionale sia preferibile all'analisi di frequenza locale e all'analisi regionale basata su una regione differente. Non esiste quindi un numero di cluster corretto ed è più importante evitare che i cluster siano troppo piccoli o troppo grandi. Regioni con pochi siti non godono dei maggiori benefici dell'analisi regionale rispetto all'analisi di frequenza locale dei quantili della variabile di interesse; regioni troppo grandi possono risultare meno omogenee e causare una maggiore distorsione nella stima dei quantili per alcuni siti.

Ciò nonostante, data la tecnica di raggruppamento proposta, occorre valutare a che punto della procedura ci si deve arrestare, che in un certo senso corrisponde a valutare il numero dei gruppi. Le tecniche di cluster analysis consentono di raggruppare i dati, ma il numero di gruppi che si formano va scelto in maniera indipendente: nel nostro caso si è cercato di suddividere i bacini nel minor numero possibile di gruppi, verificando che le regioni che si vengono a formare siano statisticamente omogenee. In pratica si parte dalla condizione in cui tutti i bacini sono raggruppati in un unico cluster e si esegue un test per valutare l'omogeneità del macrogruppo; successivamente si suddividono i bacini in due gruppi, in tre gruppi e così via, utilizzando la metodologia descritta; ci si arresta quando tutte le regioni passano il test. I test di omogeneità che possono essere utilizzati a questo scopo sono ampiamente descritti nel Capitolo 5.

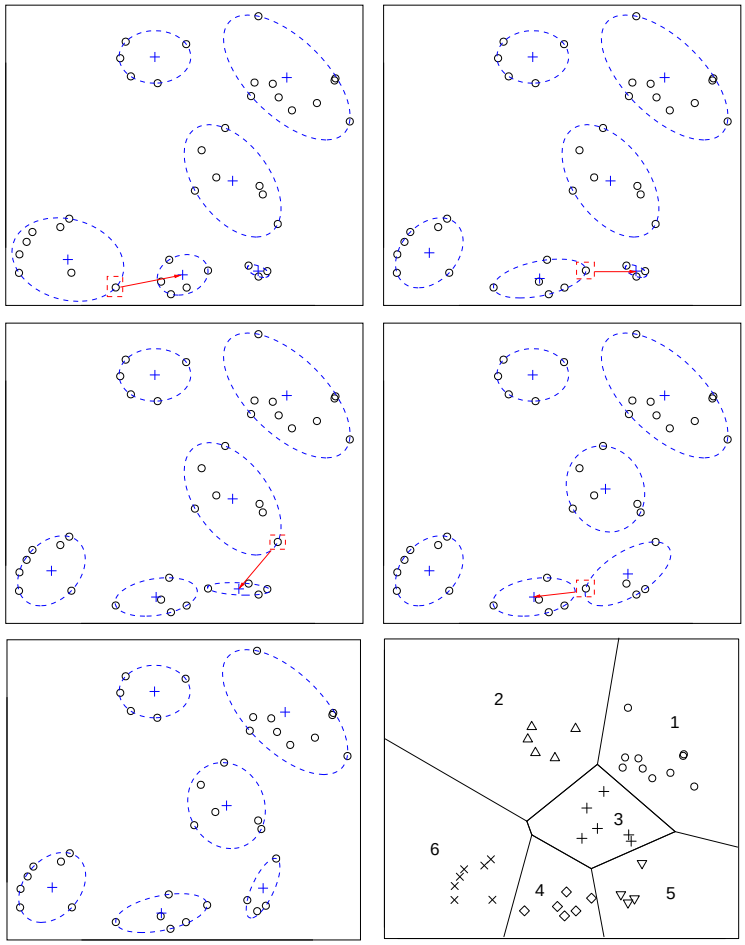

Figura 3.3: Ottimizzazione dei gruppi con un algoritmo di riallocazione degli elementi. Il grafico in alto a sinistra presenta la suddivisione in 6 cluster ottenuta, ad esempio, con l'algoritmo gerarchico di Ward. A passi successivi si valuta se lo spostamento di un elemento tra due gruppi comporta una diminuzione di W dell'Equazione (3.2), ovvero una compattazione dei gruppi intorno ai propri baricentri, nel qual caso viene applicata la correzione. Nella configurazione finale (grafico in basso a sinistra) i gruppi possono essere delimitati da poligoni di Thiessen (grafico in basso a destra), che costituiscono le regioni disgiunte alle quali nuovi elementi possono essere facilmente assegnati.

Una volta determinate le regioni omogenee, i dati di ognuna di esse vengono raggruppati e utilizzati per la stima delle curve di crescita regionali, come già accennato al Paragrafo 1.2. In pratica si costruisce, per ogni raggruppamento,

un campione contenente tutti i valori della variabile nei siti appartenenti alla regione (divisi per i corrispondenti valori indice), ossia si applica il cosiddetto metodo *station-year* (v.es. *Hosking & Wallis*, 1997). Successivamente si valuta quale distribuzione di probabilità descrive meglio il campione.

3.2 Selezione di un modello per la curva di crescita

Nell'analisi di frequenza regionale un'unica curva di crescita (in pratica un'unica distribuzione di frequenza) viene adattata ai dati di più siti. In generale, la regione potrà essere leggermente eterogenea, per cui non esisterà una distribuzione che si adatta esattamente ai campioni di tutti i siti. L'obbiettivo non è, d'altra parte, quello di ottenere una distribuzione "esatta", ma di trovare la distribuzione che permette la migliore stima dei quantili di nostro interesse. Nell'analisi di eventi estremi come ad esempio le piene o le magre dei corsi d'acqua, maggiore attenzione dovrà essere dedicata ad una delle code della distribuzione. In altri casi i quantili delle code della distribuzione potrebbero non essere interessanti agli scopi dell'analisi. Queste considerazioni devono essere fatte al momento della scelta della distribuzione di frequenza.

Le distribuzioni, o meglio le famiglie di distribuzioni, che possono essere candidate all'utilizzo nell'analisi di frequenza regionale, devono essere scelte in base al tipo di variabile analizzata e all'abilità di riprodurre comportamenti di particolare importanza. Una considerazione deve essere fatta a riguardo del numero di parametri incogniti della distribuzione candidata: le stime sono accurate solo quando la distribuzione reale assomiglia alla distribuzione adattata. Soprattutto la stima dei quantili sulle code della distribuzione può essere severamente distorta se questa condizione non è vera. Le distribuzioni con solo due parametri sono spesso troppo rigide e difficilmente adattabili ai dati idrologici. L'utilizzo di una distribuzione con più parametri, quando questi possono essere stimati accuratamente, consente una stima meno distorta dei quantili sulle code della distribuzione. Uno dei maggiori vantaggi dell'analisi di frequenza regionale è che distribuzioni con tre, o anche più, parametri possono essere stimate in modo più affidabile di quanto sarebbe possibile fare con l'analisi di frequenza locale.

Per la stima dei parametri delle distribuzioni esistono molti metodi, i più famosi dei quali sono il metodo dei momenti ed il metodo della massima ve-

rosimiglianza. Per l'analisi di frequenza regionale è stato dimostrato (*Hosking & Wallis*, 1997) che il metodo degli L-momenti, che consta nel sostituire gli L-momenti campionari ai corrispondenti L-momenti della distribuzione, è più efficiente di quello della massima verosimiglianza, quando i campioni sono di lunghezza piccola o moderata (caso che si verifica quasi sempre in idrologia). La definizione degli L-momenti e delle loro proprietà è data in Appendice A, mentre in Appendice B si riporta il legame tra questi e i parametri delle distribuzioni, utile alla stima di questi ultimi con il metodo degli L-momenti.

Le distribuzioni da utilizzarsi dipendono dal tipo di variabile idrologica considerata. Ad esempio se si vuole regionalizzare una variabile estrema, come ad esempio la portata massima annua al colmo, probabilmente è meglio utilizzare la GEV (distribuzione generalizzata del valore estremo), mentre per variabili "medie" quali il deflusso annuo, la distribuzione gamma a tre parametri dovrebbe essere più adatta (Appendice B). Naturalmente la scelta di una distribuzione piuttosto che un'altra deve essere fatta a partire da criteri oggettivi. La valutazione dei meriti delle differenti distribuzioni candidate deve essere basata su quanto bene esse si adattano ai dati disponibili. I criteri con cui questa valutazione può essere fatta sono molti e vengono indicati come *test di bontà di adattamento* delle distribuzioni. Si procede come segue: assegnato un campione di dati x_i $(i = 1, \ldots, m)$ estratto da una distribuzione $F_R(x)$, lo scopo del test è provare l'ipotesi statistica $H_0 : F_R(x) = F(x, \theta)$, dove $F(x, \theta)$ è la distribuzione ipotetica e θ è un vettore di parametri stimati dal campione x_i.

Per determinare se la distribuzione di probabilità scelta si adatta bene agli m dati di cui si dispone in un'assegnata regione, si può utilizzare un test di adattamento basato su una misura dello scostamento medio quadratico tra la distribuzione ipotetica $F(x, \theta)$ e la funzione di frequenza cumulata $F_m(x)$, definita come:

$$\begin{cases} F_m(x) = 0 \,, & x < x_{(1)} \\ F_m(x) = i/m \,, & x_{(i)} \le x < x_{(i+1)} \\ F_m(x) = 1 \,, & x_{(m)} \le x \end{cases} \,, \tag{3.3}$$

dove con $x_{(i)}$ si è indicato l'i-esimo elemento del campione di dati ordinato in senso crescente. In generale questi tipi di test possono essere ricondotti alla formulazione di una statistica test:

$$Q^2 = m \int_x \left[F_m(x) - F(x, \theta)\right]^2 \Psi(x)\, dF(x) \,, \tag{3.4}$$

dove $\Psi(x)$ è una funzione che può valere 1, nel qual caso si ha la statistica di Cramer-von Mises, oppure $\Psi(x) = [F(x, \theta)(1 - F(x, \theta))]^{-1}$, che definisce la stati-

stica di Anderson-Darling (*Laio*, 2004). Nell'applicazione descritta nel Capitolo 6 è stata utilizzata quest'ultima formulazione che, nella pratica, viene calcolata come:

$$A^2 = -m - \frac{1}{m}\sum_{i=1}^{m}\left\{(2i-1)\ln[F(x_{(i)},\theta)] + (2m+1-2i)\ln[1-F(x_{(i)},\theta)]\right\} \quad . \tag{3.5}$$

La statistica A^2 ottenuta dai dati deve essere paragonata alla popolazione delle A^2 che si avrebbero per campioni effettivamente estratti dalla distribuzione ipotetica $F(x,\theta)$, sempre con parametri stimati dal campione. Tale popolazione può essere ricavata con una procedura di tipo Monte Carlo: si genera un grande numero di campioni di lunghezza m dalla $F(x,\theta)$; per ognuno di essi si stimano i parametri $\hat{\theta}$ di F e si calcola la statistica A_0^2 di Anderson-Darling con l'Equazione (3.5) (con $\hat{\theta}$ al posto di θ). L'insieme degli A_0^2 così calcolati permette di determinare la funzione di frequenza $G(A_0^2)$ della statistica test sotto l'ipotesi H_0; ad esempio, se si vuole verificare l'adattamento della distribuzione ai dati originali con significatività del 5%, si rigetta l'ipotesi H_0 se l'A^2 calcolato con i dati originali risulta maggiore del quantile 0.95 di $G(A_0^2)$.

Se si confrontano distribuzioni con lo stesso numero di parametri, le probabilità associate alla statistica di Anderson-Darling possono essere utilizzate per valutare quale distribuzione si adatti meglio al campione. Un'applicazione di questa metodologia è descritta al Paragrafo 6.3.

Capitolo 4

Variabili di classificazione per la formazione delle regioni

La regionalizzaione della curva di crescita di una variabile idrologica con il metodo della grandezza-indice comporta l'identificazione di regioni statisticamente omogenee (vedi Capitolo 3). Una regione si dice omogenea se le distribuzioni di frequenza adimensionalizzate nei siti ad essa appartenenti sono identiche (o almeno simili) tra loro. I metodi proposti in letteratura per la formazione delle regioni sfruttano tutti la somiglianza (o diversità) dei siti, rispetto ad alcune caratteristiche, che chiamiamo *variabili di classificazione* (v.es. *Hall & Minns*, 1999; *Burn & Goel*, 2000; *Castellarin et al.*, 2001). Non si possono tuttavia considerare, ai fini della formazione dei gruppi, caratteristiche direttamente riconducibili alle curve di crescita stesse, dal momento che risulterebbe impossibile assegnare ad una delle regioni un sito dove non si hanno dati (v.es. *Hosking & Wallis*, 1997). Bisogna invece considerare delle caratteristiche (geografiche, fisiche, climatiche, ...), ricostruibili per un sito qualsiasi ma decisamente correlate all'aspetto delle curve di crescita campionarie. La scelta iniziale di queste caratteristiche è di fondamentale importanza, in quanto la successiva suddivisione terrà conto solo di esse.

Il problema è analogo a quello incontrato nella regionalizzazione della grandezza-indice, dove si ricorre all'analisi multi-regressiva tra i valori misurati e descrittori dei siti per poter effettuare una stima anche in siti non strumentati, come mostrato nel Capitolo 2. Si potrebbe pensare ad un approccio analogo per la scelta

delle variabili di calssificazione da utilizzarsi nella formazione delle regioni. Il problema, in questo caso, è che le curve di crescita sono entità più complesse della grandezza-indice, esprimendo un'informazione che, nel suo insieme, non è riconducibile ad un unico valore.

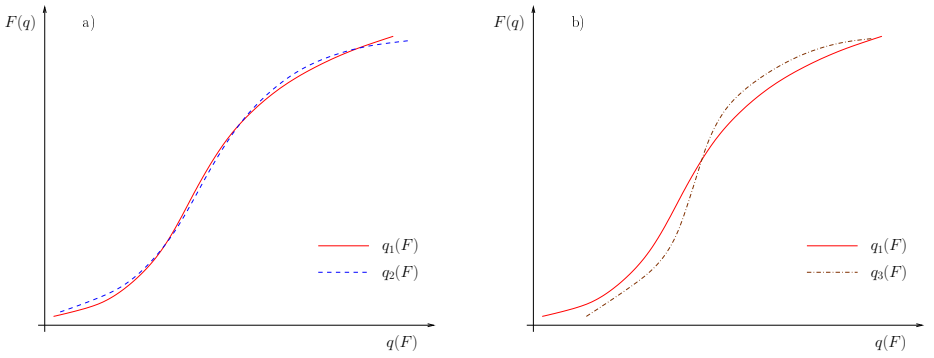

Figura 4.1: Confronto tra curve di crescita: a) le curve $q_1(F)$ e $q_2(F)$ sono simili tra loro; b) le curve $q_1(F)$ e $q_3(F)$ sono tra loro diverse.

Si considerino le curve di crescita $q_1(F)$, $q_2(F)$ e $q_3(F)$ rappresentate in Figura 4.1. Il problema è duplice: da una parte occorre stabire quantitativamente che le curve $q_1(F)$ e $q_2(F)$ del grafico (a) sono simili tra loro mentre quelle del grafico (b), $q_1(F)$, $q_3(F)$, non lo sono; dall'altra bisogna individuare quali variabili (di classificazione) spiegano questa similitudine. Si può pensare di caratterizzare le curve di crescita con un unico parametro, ad esempio una delle statistiche dei campioni, e valutare la correlazione di questo con le grandezze disponibili per ogni sito (ad esempio le grandezze morfoclimatiche dei bacini nel caso in cui la variabile analizzata sia il deflusso). Si tratterebbe essenzialmente di determinare il migliore modello lineare tra questo parametro (ad esempio il coefficiente di variazione o l'L-CV) ed i descrittori dei siti, utilizzando i metodi descritti nel Capitolo 2, e di utilizzare le variabili indipendenti così ottenute come variabili di classificazione per la formazione delle regioni (nel modo indicato nel Capitolo 3). Questo approccio tuttavia presuppone che la statistica considerata sia sufficientemente descrittiva dell'intera forma delle curve di crescita.

In questo capitolo si propone un metodo più generale, legato alla valutazione di diversità, o distanza, tra le curve di crescita considerate nel loro insieme. Se non si può descrivere l'intera curva di crescita a mezzo di un unico valore, si può però associare a due curve di crescita un valore di distanza. Nell'esempio

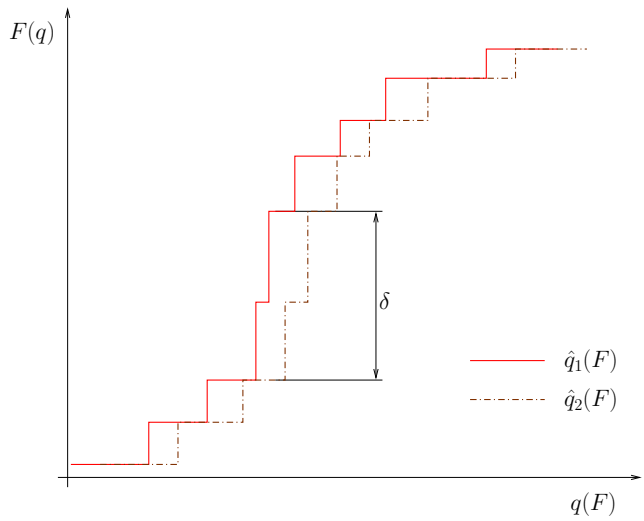

Figura 4.2: Esempio di misura della distanza tra curve di crescita: in questo caso si considera la distanza massima δ tra le funzioni di frequenza cumulate.

rappresentato in Figura 4.2 si è raffigurata la distanza massima δ tra le distribuzioni empiriche di frequenza cumulata $\hat{q}_1$ e $\hat{q}_2$, ma si possono utilizzare anche statistiche più complesse, come le statistiche su cui sono basati i test di omogeneità (vedi Capitolo 5). Questo ci costringe però a ragionare in termini di coppie di siti, invece che di singoli siti. In sostanza, se per due siti si hanno curve di crescita campionarie molto simili (diverse), si vuole valutare in che cosa, in quali grandezze caratteristiche, i due siti sono simili (diversi). Per far ciò si può utilizzare un metodo poco noto in ambito idrologico che si basa sul confronto tra le *matrici delle distanze*. In questo caso, le matrici contengono le distanze tra le distribuzioni e le distanze tra le grandezze che descrivono i diversi siti. Le distanze tra i parametri relativi a coppie di siti vengono raccolte in matrici quadrate $\mathbf{\Delta}$ il cui elemento generico δ_{hk} è una misura della differenza tra i valori

assunti dalla variabile considerata nei siti h e k (v.es. *Fabbris*, 1997):

$$\boldsymbol{\Delta} = \begin{vmatrix} 0 & \cdots & \cdots & \delta_{1h} & \cdots & \cdots & \delta_{1N} \\ \vdots & 0 & & \vdots & & & \vdots \\ \vdots & & 0 & \vdots & & & \vdots \\ \delta_{h1} & \cdots & \cdots & 0 & \cdots & \cdots & \delta_{hN} \\ \vdots & & & \vdots & 0 & & \vdots \\ \vdots & & & \vdots & & 0 & \vdots \\ \delta_{N1} & \cdots & \cdots & \delta_{Nh} & \cdots & \cdots & 0 \end{vmatrix} . \qquad (4.1)$$

Le *distanze* δ_{ij}, sono misure tra entità caratterizzate dalle seguenti proprietà: la distanza tra un'entità e se stessa è nulla ($\delta_{ii} = 0$ per $i = 1, ..., N$); la distanza tra due entità qualsiasi i e j è non negativa ($\delta_{ij} \geq 0$ per $i, j = 1, ..., N$); la distanza tra due entità gode della proprietà di simmetria, ovvero è la stessa se si misura da i a j oppure da j a i ($\delta_{ij} = \delta_{ji}$). Se queste condizioni sono rispettate, le matrici delle distanze risultano quadrate, simmetriche, definite semi-positive, di rango pari al numero di siti considerato, e con valori nulli sulla diagonale. Ad esempio, se si considera la variabile area del bacino S sotteso alle stazioni idrometriche (in un'analisi regionale dei deflussi), δ_{hk} potrebbe essere valutata come differenza, in valore assoluto, tra le aree dei bacini h e k. $\boldsymbol{\Delta}_S$ sarà quindi una matrice simmetrica e con diagonale nulla, contenente la differenza tra le aree di tutte le possibili coppie di bacini.

La determinazione delle variabili di classificazione per la formazione delle regioni può allora essere fatta individuando il migliore modello lineare che lega la matrice delle distanze delle curve di crescita $\boldsymbol{\Delta}_{q(F)}$ con le matrici delle distanze dei descrittori dei siti, e valutandone la significatività. La variabile S verrà quindi utilizzata nella classificazione solo se la relativa matrice delle distanze $\boldsymbol{\Delta}_S$ risulterà correlata alla matrice $\boldsymbol{\Delta}_{q(F)}$. Valutare la significatività della correlazione tra matrici delle distanze non è però immediato dal momento che esse contengono, per loro natura, valori fortemente correlati tra di loro (per cui il test di Student descritto nel Paragrafo 2.3.1 non può essere utilizzato). Se, considerando ancora $\boldsymbol{\Delta}_S$, l'elemento δ_{12} vale 1 e l'elemento δ_{13} vale 2, δ_{23} potrà solo assumere i valori 1 o 3. Per verificare la presenza di correlazione tra due matrici bisogna quindi ricorrere a tecniche statistiche particolari, che non risentano della mutua dipendenza degli elementi in $\boldsymbol{\Delta}$. Il test di Mantel (*Mantel & Valand*, 1970), sviluppato nell'ambito delle scienze biologico-ambientali, è uno di queste.

4.1 Test di Mantel

Molte procedure di analisi dei dati sono basate sulle matrici delle distanze. Il confronto tra due o più matrici delle distanze relative alle stesse entità è spesso condotto per valutare l'esistenza di correlazione tra le matrici. Il metodo più comunemente utilizzato per valutare la relazione esistente tra due matrici delle distanze è il test di *Mantel* (1967). La procedura di raggruppamento spazio-temporale proposta da *Mantel* (1967) è stata originariamente studiata per valutare la relazione tra una matrice di misure di distanza spaziale e una matrice di misure di distanza nel tempo, ma, come formalizzato in *Mantel* & *Valand* (1970), può essere utilizzata in una qualsiasi analisi che coinvolge due matrici delle distanze. Dal lavoro di *Smouse et al.* (1986), che proposero un'estensione del test di Mantel per l'analisi di correlazione parziale, il test applicato a due matrici viene detto *test di Mantel semplice*, mentre quello applicato a tre o più matrici *test di Mantel parziale*. Nei paragrafi seguenti i due test vengono definiti in maniera formale.

4.1.1 Test di Mantel semplice

Si consideri una coppia di matrici delle distanze $\boldsymbol{X}$ e $\boldsymbol{Y}$. Gli elementi di queste matrici, X_{ij} e Y_{ij}, rappresentano le distanze rispetto a due diverse caratteristiche tra due entità i e j $(i, j = 1, \ldots, K)$. Si calcoli la statistica:

$$\tilde{Z}_{YX} = \sum_{ij}(X_{ij}Y_{ij}) \ , \tag{4.2}$$

dove $\sum_{ij}$ indica la somma di tutte le coppie ij. Questa statistica test viene paragonata alla distribuzione attesa di Z_{YX}, ottenuta quando gli elementi corrispondenti delle due matrici non sono associati in nessun modo. Utilizzando una distribuzione empirica ottenuta mediante simulazione Monte Carlo, si calcola la probabilità di ottenere casualmente un valore di Z_{YX} estremo quanto $\tilde{Z}_{YX}$. La procedura Monte Carlo consta nel mantenere rigida una delle due matrici mentre le righe, e le colonne corrispondenti, dell'altra sono permutate casualmente. L'ipotesi alternativa è che ci sia un'associazione tra gli elementi corrispondenti delle due matrici. La probabilità $P = \Pr(Z_{YX} > \tilde{Z}_{YX})$ (associata alla coda superiore della distribuzione nulla) è un'utile misura della significatività statistica della correlazione tra le matrici delle distanze (*Mantel*, 1967).

Mentre P ha un significato ben preciso, $\tilde{Z}_{YX}$ è una misura poco pratica la cui scala è differente al variare delle matrici analizzate. *Smouse et al.* (1986) propongono di utilizzare il coefficiente di correlazione di Pearson r_{XY} al posto di $\tilde{Z}_{YX}$. La procedura da loro suggerita si articola nella maniera seguente. Si considerino le medie dei valori contenuti nelle due matrici

$$\bar{X} = \sum_{ij}(X_{ij}/N) \quad \text{e} \quad \bar{Y} = \sum_{ij}(Y_{ij}/N) \tag{4.3}$$

dove $N = K(K-1)$ è il numero degli elementi delle matrici $\boldsymbol{X}$ e $\boldsymbol{Y}$ eccetto quelli sulla diagonale (che sono tutti nulli). Si calcolino la somma corretta dei prodotti degli elementi delle matrici

$$\mathrm{SP}(X,Y) = (\tilde{Z}_{YX} - N\bar{X}\bar{Y}) \tag{4.4}$$

e le somme corrette dei quadrati degli stessi

$$\mathrm{SS}(X) = \sum_{ij}(X_{ij} - \bar{X})^2 \tag{4.5}$$

e

$$\mathrm{SS}(Y) = \sum_{ij}(Y_{ij} - \bar{Y})^2 \ . \tag{4.6}$$

Mentre $\mathrm{SP}(X,Y)$ varia se si permutano gli elementi di una matrice, sia $\mathrm{SS}(X)$ che $\mathrm{SS}(Y)$ rimangono gli stessi. Se si combinano le Equazioni (4.4), (4.5) e (4.6) si possono calcolare il coefficiente di regressione

$$b_{YX} = \mathrm{SP}(X,Y)/\mathrm{SS}(X) \tag{4.7}$$

ed il corrispondente coefficiente di correlazione di Pearson

$$r_{YX} = \frac{\mathrm{SP}(X,Y)}{\sqrt{\mathrm{SS}(X)\cdot\mathrm{SS}(Y)}} \tag{4.8}$$

del modello regressivo lineare

$$[Y_{ij} - \bar{Y}] = b_{YX}[X_{ij} - \bar{X}] + \epsilon_{ij} \ . \tag{4.9}$$

Questo cambiamento di variabile (r_{YX} è equivalente ad una normalizzazione di $\tilde{Z}_{YX}$) mostra che il test di Mantel può essere visto come test di significatività del coefficiente di una regressione lineare. I classici test di significatività dell'analisi lineare, ad esempio quello della t di Student, non sono validi quando si trattano le matrici delle distanze perché l'ipotesi di indipendenza tra i valori delle variabili

(in questo caso le singole matrici) è in questo caso fortemente violata. Dal momento che il test di Mantel è basato su una distribuzione nulla ottenuta con la simulazione Monte Carlo, la mancanza di indipendenza non costituisce un problema.

4.1.2 Test di Mantel parziale

Spesso si deve valutare se vi è correlazione tra due o più matrici delle distanze ($\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_H$) con una matrice $\boldsymbol{Y}$. In questa situazione, spesso gli elementi corrispondenti delle varie matrici $\boldsymbol{X}$ sono correlati tra loro, per cui l'informazione disponibile può ritenersi ridondante. L'obbiettivo che ci si pone è quello di valutare quanto sono correlate le singole matrici $\boldsymbol{X}$ alla matrice $\boldsymbol{Y}$, e quanta informazione aggiuntiva si fornisce se si considera una particolare matrice $\boldsymbol{X}$, allorquando altre matrici sono state incluse nell'analisi.

Anche in questo caso *Smouse et al.* (1986) fanno riferimento all'analisi regressiva lineare, questa volta multipla. Si consideri per semplicità il caso di due matrici $\boldsymbol{X}$ (l'estensione a più matrici risulterà semplice). L'estensione dell'Equazione (4.9) nel caso delle matrici $\boldsymbol{Y}$, $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ è

$$[Y_{ij} - \bar{Y}] = b_{Y1}[X_{1ij} - \bar{X}_1] + b_{Y2}[X_{2ij} - \bar{X}_2] + \epsilon_{ij} \ . \tag{4.10}$$

Se si indicano le somme dei quadrati corretti con

$$\mathrm{SS}(X_1) = \sum_{ij}(X_{1ij} - \bar{X}_1)^2 \ , \tag{4.11}$$

$$\mathrm{SS}(X_2) = \sum_{ij}(X_{2ij} - \bar{X}_2)^2 \tag{4.12}$$

e

$$\mathrm{SS}(Y) = \sum_{ij}(Y_{ij} - \bar{Y})^2 \ , \tag{4.13}$$

e le somme dei prodotti incrociati con

$$\mathrm{SP}(X_1, Y) = (\tilde{Z}_{Y1} - N\bar{X}_1\bar{Y}) \ , \tag{4.14}$$

$$\mathrm{SP}(X_2, Y) = (\tilde{Z}_{Y2} - N\bar{X}_2\bar{Y}) \tag{4.15}$$

e

$$\mathrm{SP}(X_1, X_2) = (\tilde{Z}_{12} - N\bar{X}_1\bar{X}_2) \ , \tag{4.16}$$

dove

$$\tilde{Z}_{Y1} = \sum_{ij}(X_{1ij}Y_{ij}) \ , \tag{4.17}$$

$$\tilde{Z}_{Y2} = \sum_{ij}(X_{2ij}Y_{ij}) \tag{4.18}$$

e

$$\tilde{Z}_{12} = \sum_{ij}(X_{1ij}X_{2ij}) \ , \tag{4.19}$$

i coefficienti della Regressione (4.10) possono essere calcolati dall'equazione vettoriale analoga all'Equazione (4.7):

$$\boldsymbol{b} = [b_1, b_2]' = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \ , \tag{4.20}$$

dove con la matrice $\boldsymbol{X}'\boldsymbol{X}$ si sono indicate le varianze e le covarianze tra le matrici $\boldsymbol{X}$ delle distanze:

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} \mathrm{SS}(X_1) & \mathrm{SP}(X_1, X_2) \\ \mathrm{SP}(X_1, X_2) & \mathrm{SS}(X_2) \end{bmatrix} \ , \tag{4.21}$$

e con il vettore $\boldsymbol{X}'\boldsymbol{Y}$ la covarianza tra ognuna delle matrici $\boldsymbol{X}$ e la matrice Y:

$$\boldsymbol{X}'\boldsymbol{Y} = \begin{bmatrix} \mathrm{SP}(X_1, Y) \\ \mathrm{SP}(X_2, Y) \end{bmatrix} \ . \tag{4.22}$$

I coefficienti di correlazione di Pearson r_{Y1}, r_{Y2} e r_{12} si ottengono sostituendo gli elementi definiti dall'Equazione (4.11) all'Equazione (4.16) nell'Equazione (4.8). Il coefficiente di correlazione parziale di $\boldsymbol{Y}$ con $\boldsymbol{X}_1$ per valori fissati di $\boldsymbol{X}_2$ vale

$$r_{Y1.2} = \frac{r_{Y1} - r_{Y2}r_{12}}{\sqrt{(1 - r_{Y2}^2)(1 - r_{12}^2)}} \ , \tag{4.23}$$

mentre il coefficiente di correlazione parziale di $\boldsymbol{Y}$ con $\boldsymbol{X}_2$ per valori fissati di $\boldsymbol{X}_1$

$$r_{Y2.1} = \frac{r_{Y2} - r_{Y1}r_{12}}{\sqrt{(1 - r_{Y1}^2)(1 - r_{12}^2)}} \ . \tag{4.24}$$

Infine, il coefficiente di determinazione si può ottenere come

$$R^2 = 1 - (1 - r_{Y1}^2)(1 - r_{Y2.1}^2) = 1 - (1 - r_{Y2}^2)(1 - r_{Y1.2}^2) \ . \tag{4.25}$$

Tutte questi sono risultati standard dell'analisi regressiva lineare tra una variabile dipendente e due variabili indipendenti. Nel caso in cui le variabili indipendenti siano più di due le considerazioni sono analoghe, benché espresse con la simbologia vettoriale. L'unico aspetto complesso della trattazione, che è

anche l'aspetto che ci interessa, è quello legato al test di significatività. Nell'ipotesi classica dell'analisi regressiva lineare, le matrici $\boldsymbol{X}$ sono considerate fisse e misurate senza errore. La differenza che sta nell'analisi di due (o più) matrici $\boldsymbol{X}$, piuttosto che nel considerarne separatamente la correlazione con $\boldsymbol{Y}$, è che queste possono essere tra loro non indipendenti. In questo caso possiamo trattare la dipendenza esistente tra le matrici $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ come fissa ed eseguire il test di Mantel permutando la sola matrice $\boldsymbol{Y}$ e calcolando ogni volta una delle statistiche presentate in precedenza, esattamente come si fa nel caso del test di Mantel semplice.

Se invece $\boldsymbol{X}_1$ e $\boldsymbol{X}_2$ sono esse stesse variabili aleatorie, misurate con una certa incertezza, e associate in qualche modo alla matrice $\boldsymbol{Y}$, il procedimento da seguire è più complesso. In questo caso è più logico indicare le tre matrici con $\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$ e $\boldsymbol{Y}_3$. Il coefficiente di correlazione parziale $r_{ij.k}$ può essere calcolato nel seguente modo: si calcola la regressione tra $\boldsymbol{Y}_i$ e $\boldsymbol{Y}_k$ e si indica con $\boldsymbol{D}_{i.k}$ la matrice i cui elementi sono i residui della regressione $d_{i.k}$. Allo stesso modo si calcolano i residui $\boldsymbol{D}_{j.k}$ della regressione tra $\boldsymbol{Y}_j$ e $\boldsymbol{Y}_k$. La correlazione tra gli elementi corrispondenti delle matrici $\boldsymbol{D}_{i.k}$ e $\boldsymbol{D}_{j.k}$ è la correlazione parziale $r_{ij.k}$ delle matrici $\boldsymbol{Y}_i$ e $\boldsymbol{Y}_j$, data la matrice $\boldsymbol{Y}_k$. La significatività del coefficiente può essere valutata attraverso permutazioni random di una delle matrici dei residui (mantenendo l'altra immutata). Considerazioni aggiuntive sull'aspetto computazionale del test sono contenute in *Legendre* (2000) che mette a confronto diversi metodi di permutazione degli elementi per l'esecuzione del test di Mantel parziale. Per l'applicazione del test è disponibile il pacchetto *vegan* (*Oksanen et al.*, 2005) del software R (*R Development Core Team*, 2006) disponibile all'url `http://cran.r-project.org/`.

Il test di Mantel, così come è stato definito, valuta la significatività della correlazione tra le matrici delle distanze soltanto se questa è di tipo lineare. Come nell'analisi regressiva, le non-linearità possono essere trattate con opportune trasformazioni, oppure sostituendo ai valori della variabile i ranghi. Tuttavia anche in questo caso quello che si può valutare è la significatività della correlazione lineare, ad esempio tra i ranghi. Quando non si conosce affatto la relazione che può intercorrere tra gli elementi delle matrici delle distanze, ovvero nella maggioranza dei casi, il test di Mantel è uno strumento senz'altro utile.

4.2 Identificazione delle variabili di classificazione

Nell'applicazione del test di Mantel all'analisi di frequenza regionale, e più precisamente all'individuazione delle variabili di classificazione per la formazione delle regioni omogenee, l'ipotesi classica dell'analisi regressiva lineare, di considerare le matrici fisse e misurate senza errore, può essere ritenuta valida (Paragrafo 2.3). Se questo è vero, il test di Mantel parziale può essere applicato come test di significatività dei coefficienti di una regressione multipla tra matrici delle distanze.

La procedura di selezione delle variabili di classificazione può quindi seguire una metodologia analoga a quella descritta nel Capitolo 2, dove il test di Mantel parziale viene utilizzato in sostituzione al test della t di Student. Si confrontano tutte le possibili regressioni lineari del tipo:

$$\mathbf{\Delta}_{q(F)} = \beta_0 + \beta_1 \mathbf{\Delta}_{x_1} + \beta_2 \mathbf{\Delta}_{x_2} + \ldots + \beta_{p-1} \mathbf{\Delta}_{x_{p-1}} + \varepsilon \ , \qquad (4.26)$$

dove $\mathbf{\Delta}_{q(F)}$ è la matrice delle distanze tra le curve di crescita campionarie, $\mathbf{\Delta}_{x_i}$ sono le matrici delle distanze tra le variabili candidate ad essere considerate di classificazione per la formazione delle regioni e β_i sono i coefficienti della regressione. Il test di Mantel sui coefficienti può essere eseguito confrontando ogni β_i con i quantili (ad esempio 0.05 se $\beta_i < 0$ o 0.95 se $\beta_i > 0$ volendo eseguire il test al 5% di significatività) della distribuzione del rispettivo coefficiente ottenuta permutando gli elementi della matrice $\mathbf{\Delta}_{q(F)}$, come mostrato nel Paragrafo 4.1. Se significative in base al test di Mantel, le variabili esplicative del modello di Equazione (4.26) possono essere utilizzate come variabili di classificazione nella procedura di formazione delle regioni descritta nel Capitolo 3. Il migliore modello regressivo tra matrici delle distanze può essere scelto in base al coefficiente di determinazione R^2_{adj} analogamente a quanto discusso al Paragrafo 2.3.1.

Nella pratica le variabili di classificazione $x_1, x_2, \ldots, x_{p-1}$ devono essere rese adimensionali in modo da influire con lo stesso peso nella procedura di cluster analysis (Paragrafo 3.1). Nell'applicazione del Capitolo 6 esse sono state adimensionalizzate come $\frac{x_i - m(x_i)}{s(x_i)}$, dove $m(x_i)$ e $s(x_i)$ sono, rispettivamente, la media e la deviazione standard campionarie delle misure del parametro x_i nelle siti monitorati. In alternativa si potrebbe pensare di utilizzare come variabili di classificazione $\beta_1 x_1, \beta_2 x_2, \ldots, \beta_{p-1} x_{p-1}$, ovvero le variabili pesate con il coefficiente ottenuto nell'Equazione (4.26). Ciò costituirebbe di per sè un'adimensionalizza-

zione, e terrebbe conto dell'influenza di ogni variabile nello spiegare la variabilità delle curve di crescita.

Capitolo 5

Test di omogeneità per l'Analisi di Frequenza Regionale

La valutazione dell'omogeneità delle regioni è uno dei punti critici dell'analisi di frequenza regionale. L'ipotesi di omogeneità implica che le distribuzioni di frequenza dei differenti siti appartenenti alla regione siano identiche, a meno di un parametro di scala caratteristico di ogni sito. Parecchi autori hanno proposto test di omogeneità nella letteratura idrologica, tra cui *Dalrymple* (1960), *Wiltshire* (1986a,b,c), *Chowdhury et al.* (1991), *Lu & Stedinger* (1992), *Fill & Stedinger* (1995) e *Hosking & Wallis* (1993, 1997). Tuttavia i confronti tra i test sono stati pochi, con la conseguenza di lasciare l'utente senza un'idea chiara sui meriti e i limiti di ogni tecnica. Le statistiche basate sugli L-momenti (*Hosking & Wallis*, 1993, 1997) sono quelle più usate in questo momento nell'analisi di frequenza regionale, ma non ci sono studi dettagliati che dimostrano la loro superiorità nei confronti degli altri metodi.

In questo capitolo vengono riportati i risultati ottenuti in *Viglione et al.* (2007a): quattro test di omogeneità vengono confrontati con un'impostazione di carattere generale. I primi due test sono quelli proposti da *Hosking & Wallis* (1993), basati sugli L-momenti. Gli altri due test sono nuovi in ambito idrologico: sono il test di Anderson-Darling (*Scholz & Stephens*, 1987), opportunamente modificato per tener conto della normalizzazione dei campioni con la grandezza indice, e il test di *Durbin & Knott* (1971), molto usato come test di bontà di adattamento, ma utilizzato quì per la valutazione di eterogeneità.

Le performance di questi test è stata valutata attraverso la determinazione della loro potenza (e dell'errore di tipo I) con esperimenti di simulazione tipo Monte Carlo. In particolare potenza ed errore di Tipo I sono stati valutati per differenti parametri della distribuzione generatrice, variando il numero di siti appartenenti alla regione, la lunghezza delle serie, il tipo di distribuzione generatrice ed il grado di eterogeneità.

5.1 Test di omogeneità

Si supponga di disporre di k campioni di una stessa variabile in differenti siti di misura, e di voler verificare se possono essere raggruppati per formare una regione statisticamente omogenea: sia Y_{ij} la j-esima osservazione dell'i-esimo campione, ordinato in senso crescente ($Y_{i1} \leq Y_{i2} \leq \ldots \leq Y_{in_i}$, dove $i = 1, \ldots, k$). Se si segue la procedura della grandezza-indice, le osservazioni sono riscalate in un primo momento con $\overline{Y_i}$ (dettagli sulla scelta della grandezza-indice sono riportati nel Paragrafo 2.1) ottenendo $X_{ij} = \frac{Y_{ij}}{Y_i}$. Se le osservazioni sono tra loro indipendenti e l'i-esimo campione riscalato ha distribuzione di frequenza F_i, il test di omogeneità corrisponde a verificare l'ipotesi $H_0 : F_1 = ... = F_k = F$, senza specificare qual è la distribuzione comune F. I vantaggi e gli svantaggi di una statistica test sono quantificabili se si valutano la sua potenza e l'errore di Tipo I. Data l'ipotesi nulla H_0, nel nostro caso l'ipotesi di omogeneità regionale, la potenza del test è definita come la probabilità di rigettare correttamente H_0 quando non è vera. Se invece l'ipotesi è rigettata quando dovrebbe essere accettata, si commette un errore di Tipo I. Il test si dice non-distorto se la probabilità di commettere un errore di Tipo I equivale al livello di significatività scelto, α.

I test di omogeneità comportano di stimare una quantità per ogni sito, θ_i, che misura un qualche aspetto della distribuzione di frequenza empirica (del singolo sito), e di verificare se la dispersione dei θ_i attorno ad un valore regionale, θ^R, è consistente con l'ipotesi di omogeneità. Ciò richiede la definizione della distribuzione di θ nell'ipotesi di validità di H_0, $G_{H_0}(\theta)$, che in molti casi implica che si scelga *a priori* la distribuzione di probabilità comune F. Si tratta di un problema teorico comune a molti test di omogeneità (un'eccezione è il test di *Wiltshire* (1986a) basato sul coefficiente di variazione). La necessità di preselezionare F implica che in realtà il test non permetta di verificare la sola ipotesi

di omogeneità, ma l'ipotesi composita (omogeneità e bontà di adattamento) che la distribuzione di probabilità sia la stessa per ogni sito, ed abbia una forma matematica predefinita, F. Di conseguenza, le possibili ragioni per cui il test non viene superato possono essere sia l'eterogeneità della regione, sia l'inadeguatezza della distribuzione di probabilità adottata F. Torneremo su questo aspetto nel Paragrafo 5.1.2, in cui viene descritto il test di Anderson-Darling.

Un secondo problema è determinato dalla normalizzazione dei campioni con la grandezza-indice che, in alcuni casi, può distorcere la distribuzione $G_{H_0}(\theta)$ della statistica test nel caso di validità dell'ipotesi nulla: è il caso dei test di *Wiltshire* (1986a) e di Anderson-Darling basati sui ranghi. Il problema viene trattato in dettaglio nel Paragrafo 5.1.2. Diamo ora una descrizione dei quattro test di omogeneità selezionati per il confronto. Il pacchetto `homtest` (*Viglione*, 2006) per il software R (*R Development Core Team*, 2006), sviluppato per agevolare l'applicazione dei test, è disponibile sui server CRAN (si veda il sito web `http://www.r-project.org/`).

5.1.1 Le misure di eterogeneità di Hosking e Wallis

L'idea su cui si basano le statistiche di eterogeneità di *Hosking* & *Wallis* (1993) è quella di misurare la variabilità campionaria degli L-momenti e di paragonarla alla variabilità che ci si aspetterebbe abbia una regione omogenea. Quest'ultima è stimata attraverso simulazioni ripetute di regioni omogenee con campioni estratti da una distribuzione kappa a quattro parametri (vedi l'Appendice B). Più in dettaglio i passi sono i seguenti:

1. Per quanto riguarda i k campioni appartenenti alla regione in analisi, si calcolino i rapporti degli L-momenti campionari (si veda l'Appendice A per i dettagli) dell'i-esimo sito: questi sono il coefficiente di L-variazione (L-CV),

$$t^{(i)} = \frac{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{2(j-1)}{(n_i-1)} - 1\right) Y_{i,j}}{\frac{1}{n_i}\sum_{j=1}^{n_i} Y_{i,j}}, \tag{5.1}$$

   l'L-skewness,

$$t_3^{(i)} = \frac{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{6(j-1)(j-2)}{(n_i-1)(n_i-2)} - \frac{6(j-1)}{(n_i-1)} + 1\right) Y_{i,j}}{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{2(j-1)}{(n_i-1)} - 1\right) Y_{i,j}}, \tag{5.2}$$

e l'L-kurtosis

$$t_4^{(i)} = \frac{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{20(j-1)(j-2)(j-3)}{(n_i-1)(n_i-2)(n_i-3)} - \frac{30(j-1)(j-2)}{(n_i-1)(n_i-2)} + \frac{12(j-1)}{(n_i-1)} - 1\right)Y_{i,j}}{\frac{1}{n_i}\sum_{j=1}^{n_i}\left(\frac{2(j-1)}{(n_i-1)} - 1\right)Y_{i,j}}. \tag{5.3}$$

Si noti che i rapporti degli L-momenti non sono influenzati dalla normalizzazione attraverso la grandezza-indice, essendo analogo utilizzare $X_{i,j}$ o $Y_{i,j}$ nelle Equazioni (5.1)-(5.3).

2. Si determinino i coefficienti medi regionali L-CV, L-skewness e L-kurtosis,

$$t^R = \frac{\sum_{i=1}^{k} n_i t^{(i)}}{\sum_{i=1}^{k} n_i} \qquad t_3^R = \frac{\sum_{i=1}^{k} n_i t_3^{(i)}}{\sum_{i=1}^{k} n_i} \qquad t_4^R = \frac{\sum_{i=1}^{k} n_i t_4^{(i)}}{\sum_{i=1}^{k} n_i} \tag{5.4}$$

e si calcoli la statistica

$$V = \left\{\sum_{i=1}^{k} n_i (t^{(i)} - t^R)^2 / \sum_{i=1}^{k} n_i\right\}^{1/2}. \tag{5.5}$$

3. Si ottengano i quattro parametri di una distribuzione kappa a partire dai rapporti degli L-momenti medi regionali t^R, t_3^R e t_4^R, e si generi un elevato numero N_{sim} di realizzazioni di set di k campioni. L'i-esimo sito in ogni campione è quindi estratto dalla distribuzione kappa e ha lunghezza pari a n_i. Per ogni regione omogenea simulata si calcoli la statistica di Equazione (5.5), ottenendo N_{sim} valori. Si determinino da questo vettore di valori di V media μ_V e scarto quadratico medio σ_V, che sono associati all'ipotesi di omogeneità (in realtà all'ipotesi composta di omogeneità e distribuzione generatrice kappa).

4. Una misura di eterogeneità, che verrà qui indicata con HW_1, viene infine calcolata come

$$\theta_{HW_1} = \frac{V - \mu_V}{\sigma_V}. \tag{5.6}$$

θ_{HW_1} può essere approssimata con una distribuzione normale standard (media pari a 0 e varianza pari a 1). Seguendo le indicazioni di *Hosking* & *Wallis* (1997), la regione in analisi può essere considerata "accettabilmente omogenea" se $\theta_{HW_1} < 1$, "possibilmente eterogenea" se $1 \leq \theta_{HW_1} < 2$, e "sicuramente eterogenea" se $\theta_{HW_1} \geq 2$. *Hosking* & *Wallis* (1997) consigliano che questi limiti siano utilizzati come linee guida. Sebbene la statistica θ_{HW_1} sia costruita come un test di significatività, i livelli di significatività

ottenuti sarebbero accurati solo sotto alcune assunzioni: che i dati siano indipendenti sia serialmente che tra i siti, e che la vera distribuzione di probabilità regionale sia la kappa.

La statistica θ_{HW_1} misura l'eterogeneità solo nella dispersione dei campioni, dal momento che è basata solamente sulle differenze tra gli L-CV nella regione. Per questo motivo non è sensibile all'eterogeneità determinata dalla diversità di L-skewness tra i campioni. *Hosking* & *Wallis* (1993) forniscono anche una misura di eterogeneità alternativa (che noi chiamiamo HW_2), nella quale il valore di V dell'Equazione (5.5) è sostituito da:

$$V_2 = \sum_{i=1}^{k} n_i \left\{ (t^{(i)} - t^R)^2 + (t_3^{(i)} - t_3^R)^2 \right\}^{1/2} / \sum_{i=1}^{k} n_i \ , \tag{5.7}$$

La statistica test diventa in questo caso

$$\theta_{HW_2} = \frac{V_2 - \mu_{V_2}}{\sigma_{V_2}} \ , \tag{5.8}$$

con gli stessi limiti di accettabilità di HW_1. *Hosking* & *Wallis* (1997) giudicano θ_{HW_2} molto inferiore a θ_{HW_1} e dicono che raramente essa mostra valori superiori a 2, anche per regioni estremamente eterogenee. Inoltre affermano che nella pratica non è comune trovare siti caratterizzati dallo stesso L-CV e differente L-skewness (siti con elevato L-skewness tendono ad avere anche un elevato L-CV). Ad ogni modo abbiamo deciso di includere anche la statistica HW_2 nell'analisi comparativa di questo lavoro perché essa viene utilizzata nel più sistematico e ben documentato lavoro disponibile sull'analisi regionale delle piene (*Robson* & *Reed*, 1999).

5.1.2 Il test di Anderson-Darling su k campioni

Come accennato, le misure di eterogeneità HW_1 e HW_2 sono caratterizzate dalla limitazione di essere un test di bontà di adattamento + omogeneità, essendo basate sulla distribuzione generatrice kappa. Probabilmente la distribuzione kappa è abbastanza flessibile da limitare le conseguenze di tale assunzione (*Hosking* & *Wallis*, 1997), ma l'inconsistenza teorica rimane tale. Per questo abbiamo deciso di proporre nel confronto anche test che non presentano questo problema. Un possibile candidato potrebbe essere il test di *Wiltshire* (1986a)

basato sul CV, se non fosse stato giudicato inaffidabile dallo stesso Autore. Un altro test che non fa alcuna assunzione sulla distribuzione generatrice è il test di Anderson-Darling (AD) basato sui ranghi (*Scholz & Stephens*, 1987). Il test AD è la generalizzazione del classico test di bontà di adattamento di Anderson-Darling (v.es. *D'Agostino & Stephens*, 1986) descritto al Paragrafo 3.2, e viene usato per testare l'ipotesi che k campioni indipendenti appartengono alla stessa popolazione senza specificare la distribuzione di probabilità comune.

Il test è basato sul confronto tra le distribuzioni di frequenza empiriche locale e regionale. La distribuzione di frequenza empirica, o distribuzione di frequenza campionaria, è definita come $F(x) = \frac{j}{\eta}, x_{(j)} \leq x < x_{(j+1)}$, dove η è la dimensione del campione e $x_{(j)}$ sono le osservazioni riarrangiate in ordine crescente. Si indichino la distribuzione di frequenza empirica dell'i-esimo campione (locale) con $\hat{F}_i(x)$, e quella del raggruppamento di tutti i $N = n_1 + ... + n_k$ dati dei siti (regionale) con $H_N(x)$. La statistica test di Anderson-Darling su k campioni viene quindi definita come

$$\theta_{AD} = \sum_{i=1}^{k} n_i \int_{\text{all } x} \frac{[\hat{F}_i(x) - H_N(x)]^2}{H_N(x)[1 - H_N(x)]} dH_N(x) \ . \tag{5.9}$$

Se si indica il campione costituito da tutti i dati del raggruppamento, ordinato in senso crescente, con $Z_1 < ... < Z_N$, la forma numerica per risolvere l'Equazione (5.9) è:

$$\theta_{AD} = \frac{1}{N} \sum_{i=1}^{k} \frac{1}{n_i} \sum_{j=1}^{N-1} \frac{(NM_{ij} - jn_i)^2}{j(N-j)} \ , \tag{5.10}$$

dove M_{ij} è il numero delle osservazioni nell'i-esimo campione che non superano Z_j. Il test di omogeneità può essere eseguito paragonando i valori di θ_{AD} ottenuti con i valori percentuali riportati in tabella da *Scholz & Stephens* (1987) per differenti livelli di significatività.

La statistica θ_{AD} dipende dai valori dei campioni solo in relazione al loro rango (posizione all'interno del campione ordinato). Ciò garantisce che la statistica test rimanga invariata qualora i campioni subiscano una trasformazione monotona, il che costituisce un'importante proprietà di stabilità che le misure di eterogeneità HW non possiedono. Tuttavia, l'applicazione di questi test alla procedura della grandezza-indice dà luogo a dei problemi. Infatti, in questa procedura i campioni vengono divisi per un valore diverso a seconda del sito corrispondente, modificando così i ranghi nel campione costituito da tutti i dati

del raggruppamento. In particolare, questo fa sì che le distribuzioni di frequenza empiriche locali siano più vicine tra loro, dando l'impressione di omogeneità anche quando i campioni sono molto eterogenei. L'effetto è analogo a quello che si riscontra quando si applica un test di bontà di adattamento a distribuzioni i cui parametri sono stimati dallo stesso campione usato nel test (v.es. *D'Agostino & Stephens*, 1986; *Laio*, 2004). In entrambi i casi i punti percentuali per i test devono essere opportunamente rideterminati. Questo può essere fatto attraverso un approccio di "bootstrap" non parametrico, che si esegue tramite i seguenti passi:

1. Si costruisce il campione $\mathcal{S}$ costituito da tutti i dati osservati adimensionalizzati della regione.

2. Si campiona casualmente (permettendo la riselezione) da $\mathcal{S}$ e si generano k campioni artificiali, di dimensioni $n_1, \ldots, n_k$.

3. Si divide ogni campione per la corrispondente grandezza-indice e si calcola $\theta_{AD}^{(1)}$.

4. Si ripete la procedura per N_{sim} volte e si ottiene un campione di $\theta_{AD}^{(j)}$, $j = 1, \ldots, N_{sim}$ valori, la cui distribuzione di frequenza empirica può essere usata come approssimazione a $G_{H_0}(\theta_{AD})$, la distribuzione di θ_{AD} nel caso di validità dell'ipotesi nulla.

5. I limiti di accettazione per il test, corrispondenti ad un qualsiasi livello di significatività α, sono facilmente determinabili come quantili di $G_{H_0}(\theta_{AD})$ corrispondenti alla probabilità $(1 - \alpha)$.

Chiameremo il test ottenuto in questo modo *test bootstrap di Anderson-Darling*, di qui in poi indicato con *AD*.

5.1.3 Test di Durbin e Knott

L'ultimo test di omogeneità considerato deriva da una statistica di bontà di adattamento originariamente proposta da *Durbin & Knott* (1971). Il test è costruito in modo da evidenziare discrepanze nella dispersione dei campioni, senza tener conto di eventuali discordanze nella media e nello skewness dei dati. Sotto questo punto di vista il test è simile al test HW_1, mentre è analogo al

test AD per il fatto di essere basato sui ranghi. Il test di bontà di adattamento originale è molto semplice: si supponga di avere un campione X_i, $i = 1, ..., n$, con una ipotetica distribuzione di probabilità $F(x)$; nell'ipotesi nulla la variabile casuale $F(X_i)$ è distribuita uniformemente nell'intervallo $(0, 1)$ e la statistica $D = \sum_{i=1}^{n} \cos[2\pi F(X_i)]$ è distribuita approssimativamente secondo una distribuzione normale standard (*Durbin & Knott*, 1971). D serve per individuare discrepanze nella dispersione dei dati: se la varianza di X_i è maggiore di quella dell'ipotetica distribuzione $F(X)$, D è significativamente maggiore di 0, mentre D è significativamente negativo nel caso opposto. Le differenze tra le medie (o le mediane) di X_i e $F(x)$ non sono invece individuabili con D, il che garantisce che la normalizzazione dei campioni con la grandezza-indice non influenzi il test.

L'estensione a test di omogeneità della statistica di *Durbin & Knott* (1971) è chiara: si sostituisce la distribuzione di frequenza empirica ottenuta dai dati raggruppati, $H_N(x)$, a $F(x)$ in D, ottenendo per ogni sito la statistica

$$D_i = \sum_{j=1}^{n_i} \cos[2\pi H_N(X_j)], \tag{5.11}$$

che è distribuita normalmente nell'ipotesi di omogeneità. La statistica $\theta_{DK} = \sum_{i=1}^{k} D_i^2$ deve quindi essere distribuita secondo una distribuzione del chi-quadro con $k - 1$ gradi di libertà, che permette di determinare i limiti di accettabilità per il test, corrispondenti a qualsiasi livello di significatività α. Si noti che l'implementazione del test DK è molto più semplice rispetto alle altre statistiche considerate.

5.2 Principi per il confronto tra i test

Il principale obbiettivo di questa parte del lavoro è l'identificazione, servendosi di simulazioni Monte Carlo, di quale dei test descritti nel Paragrafo 5.1 funziona meglio, ovvero è meno distorto (l'errore di Tipo I è simile al livello di significatività adottato) e più potente. Le simulazioni Monte Carlo richiedono che:

1. si definisca una regione artificiale fornendo il numero di campioni k, la loro lunghezza n (che è stata presa costante per tutti i siti), la distribuzione generatrice (a 3 parametri) $\mathcal{P}$ da utilizzare per la generazione dei campioni, e i rapporti degli L-momenti medi regionali τ^R e τ_3^R;

2. la regione artificiale sia caratterizzata da un'eterogeneità nota, con i rapporti degli L-momenti locali, $\tau^{(i)}$ e/o $\tau_3^{(i)}$, che variano linearmente dal sito 1 al sito k, con un range di variazione complessivo $\Delta\tau$ e $\Delta\tau_3$ (quando $\Delta\tau$ e $\Delta\tau_3$ valgono entrambi 0, la regione è omogenea);

3. per ogni sito nella regione si stimino i tre parametri della distribuzione generatrice $\mathcal{P}$ a partire dagli L-momenti locali, e un campione di lunghezza n è generato da $\mathcal{P}$ e normalizzato con la grandezza-indice;

4. si applichino i quattro test di omogeneità alla regione artificiale ottenuta, dopo aver selezionato un livello di significatività α per i test AD e DK, ed un equivalente limite di accettabilità per le misure di eterogeneità HW_1 e HW_2;

5. si generino 1000 repliche di regioni artificiali come indicato nei punti precedenti e per ogni replica si testi l'omogeneità della regione separatamente con ogni test. La potenza di ogni test (o l'errore di Tipo I) è stimata come percentuale delle 1000 repliche riconosciuta come eterogenea.

Il confronto tra i test deve essere il più generale possibile: valori differenti di k, n, $\mathcal{P}$, τ, τ_3, $\Delta\tau$, $\Delta\tau_3$, e α devono essere considerati, cosa che complica molto la simulazione numerica. In particolare, la dispersione e lo skewness medi dei campioni, τ^R e τ_3^R, influenzano in maniera rilevante la performance dei test. Lo stesso vale per gli altri parametri, ma gli effetti sui test della variazione, ad esempio, di n è molto più semplice da prevedere e, quindi, meno interessante da analizzare. Per questo motivo abbiamo deciso di considerare diversi valori di τ^R e τ_3^R, ovvero di esplorare con le simulazioni gran parte del diagramma $\tau - \tau_3$. *Hosking* & *Wallis* (1997) forniscono dei limiti numerici per τ e τ_3: $0 \leq \tau < 1$, $-1 < \tau_3 < 1$, e $2\tau - 1 < \tau_3$ (limiti validi per variabili che possono assumere solo valori positivi). Lo spazio $\tau - \tau_3$ delimitato da questi limiti rimane però ancora troppo grande in una prospettiva operazionale.

Per scegliere limiti più restrittivi nello spazio $\tau - \tau_3$ si fa riferimento all'esperienza idrologica considerando il lavoro di *Vogel* & *Wilson* (1996), che usano i diagrammi degli L-momenti per selezionare una distribuzione di frequenza regionale per le variabili deflusso minimo, medio e massimo annuo. *Vogel* & *Wilson* (1996) costruiscono questi diagrammi per più di 1400 corsi d'acqua statunitensi. Tutti i valori di τ e τ_3, indipendentemente dal tipo di deflusso, occupano una banda diagonale del grafico esprimibile come $\tau_3 - 0.2 < \tau < \tau_3 + 0.4$ (si veda la Figura 5.1) e pochi punti hanno τ_3 maggiore di 0.5 o minore di -0.1. Si è quin-

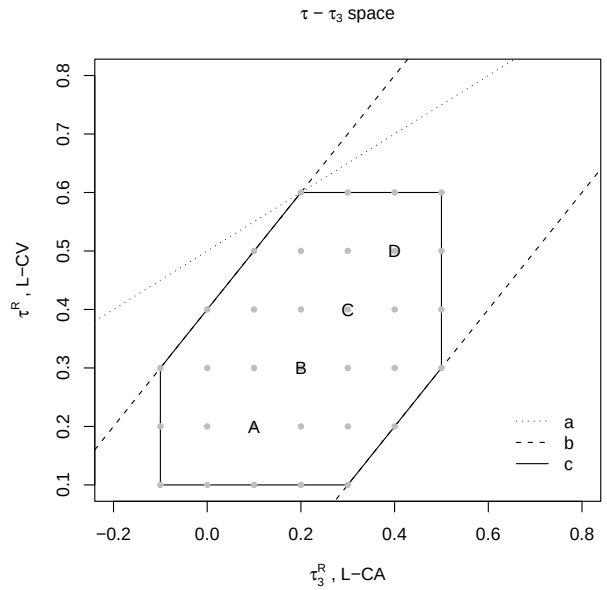

Figura 5.1: Spazio $\tau - \tau_3$ considerato per il confronto dei test (da *Viglione et al.*, 2007a). Le linee sono: (a) limiti numerici forniti da *Hosking* & *Wallis* (1997); (b) banda diagonale individuata utilizzando i dati di *Vogel* & *Wilson* (1996); (c) la regione considerata nello studio. I punti in grigio sono i valori di τ^R e τ_3^R considerati nel caso studio principale (Paragrafo 5.3.1); i punti A, B, C e D sono considerati nell'analisi di sensitività del Paragrafo 5.3.2.

di scelto di limitare l'investigazione alla regione delimitata nel modo seguente (Figura 5.1):

$$\begin{cases} 0.1 & < & \tau & < & 0.6\ , \\ -0.1 & \leq & \tau_3 & < & 0.5\ , \\ \tau_3 - 0.2 & < & \tau & < & \tau_3 + 0.4\ , \end{cases} \tag{5.12}$$

Tutte le coppie τ^R e τ_3^R all'interno della regione, su una griglia di lato 0.1 (punti in grigio della Figura 5.1) sono stati utilizzati per le simulazioni.

Per quanto riguarda le altre variabili coinvolte (k, n, $\mathcal{P}$, $\Delta\tau$, $\Delta\tau_3$, e α), la strategia di simulazione adottata prevede la costruzione di un caso studio principale, con valori dei parametri scelti ragionevolmente, e di effettuare una sorta di analisi di sensitività. I parametri selezionati per il caso studio principale sono: $k = 11$; $n = 30$; $\mathcal{P} \equiv$ distribuzione generalizzata del valore estremo (GEV, Appendice B); $\alpha = 5\%$ (o, equivalentemente, $\theta_{HW} \leq 2$); $\Delta\tau = 0$ e $\Delta\tau_3 = 0$ per

la verifica dell'errore di Tipo I, o $\Delta\tau = 0.5\tau$ e $\Delta\tau_3 = 0$ per la verifica di potenza dei test (vedi Paragrafo 5.3.1). Il tipo ed il grado di eterogeneità, la lunghezza dei campioni, il numero di siti nella regione, il livello di significatività, e le distribuzioni generatrici sono fatte variare una per volta (vedi Paragrafo 5.3.2), ed i risultati sono analizzati in 4 punti nella parte centrale del diagramma $\tau - \tau_3$ (punti A, B, C e D di Figura 5.1).

5.3 Risultati

Un aspetto rilevante nell'analisi di frequenza regionale e legato al soggetto principale di questo capitolo, è la scelta della grandezza-indice, ovvero del parametro utilizzato per normalizzare i campioni. Come già discusso nel Paragrafo 2.1 riteniamo utile sollevare una discussione su questo argomento importante e spesso trascurato, tanto più che la scelta della grandezza-indice può condizionare la performance di alcuni test di omogeneità. In generale riteniamo che la Figura 2.1 mostri i vantaggi di utilizzare la mediana campionaria come grandezza-indice quando si ritiene che le distribuzioni generatrici siano particolarmente asimmetriche, come nell'analisi regionale delle piene. Nel confronto tra i test di omogeneità, la mediana campionaria verrà utilizzata come grandezza-indice.

5.3.1 Caso studio principale

Il caso studio principale corrisponde all'analisi completa della performance dei test per tutti i punti del diagramma $\tau - \tau_3$, con $k = 11$, $n = 30$, $\mathcal{P} \equiv$ distribuzione GEV e $\alpha = 5\%$ (o $\theta_{HW} \leq 2$). Prima di tutto viene considerato l'errore di Tipo I dei test, attraverso la simulazione di regioni omogenee, con $\Delta\tau = 0$ e $\Delta\tau_3 = 0$. La Figura 5.2 riporta sullo sfondo (numeri grigi) la percentuale di regioni considerate eterogenee da ogni test, ed in primo piano (linee nere) una "superficie di trend" che interpola i valori percentuali e permette di comprendere (attraverso isolinee) come varia l'errore di Tipo I nello spazio $\tau - \tau_3$. Si può notare che i valori medi campionari $< t^R >$ e $< t_3^R >$ (ovvero le medie di t^R e t_3^R sulle 1000 repliche) possono essere differenti dai loro corrispettivi teorici τ^R e τ_3^R; i numeri grigi in Figura 5.2 non giacciono precisamente sulla griglia definita

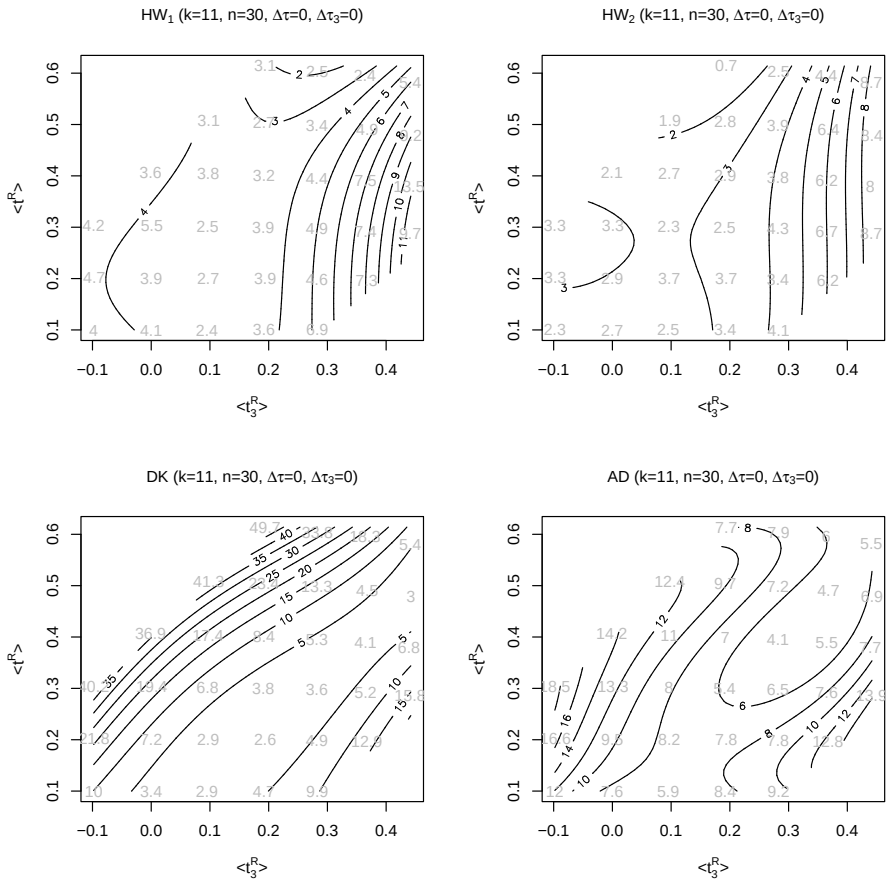

Figura 5.2: Percentuale di regioni erroneamente considerate non-omogenee dai test nello spazio $\tau - \tau_3$ (errore di Tipo I). Le regioni omogenee sono generate utilizzando la distribuzione generalizzata dei valori estremi (GEV) come distribuzione generatrice; i valori degli altri parametri sono riportati nel titolo di ogni sottofigura (da *Viglione et al.*, 2007a).

in Figura 5.1. Questo è dovuto al fatto che per campioni poco numerosi t e t_3 non sono stimatori indistorti di τ e τ_3 (*Hosking* & *Wallis*, 1997).

Nessuno dei test ha l'errore di Tipo I atteso ovunque nello spazio $\tau - \tau_3$. In gran parte dello spazio $\tau - \tau_3$, la percentuale di regioni identificate come non-omogenee dalle misure di eterogeneità di Hosking e Wallis è $2 \div 4\%$; tale percentuale cresce a $8 \div 10\%$ per alti coefficienti di L-skewness ($t_3^R > 0.4$, Figura

5.2). I test basati sui ranghi hanno un errore di Tipo I corretto nella parte diagonale-centrale dello spazio degli L-momenti, mentre la percentuale di regioni erroneamente assunte come eterogenee cresce verso i bordi (soprattutto per il test DK).

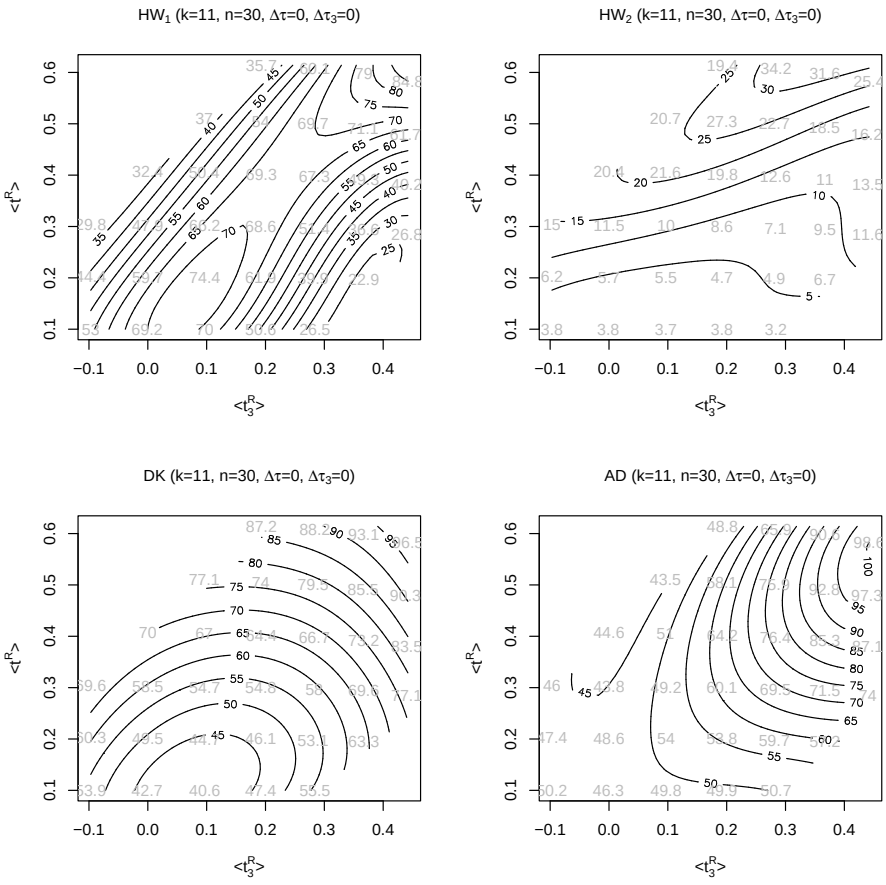

Figura 5.3: Potenza dei test nello spazio $\tau - \tau_3$ quando le regioni eterogenee sono generate utilizzando la distribuzione generalizzata dei valori estremi (GEV) come distribuzione generatrice. L'eterogeneità è dovuta alla variazione della dispersione delle distribuzioni di frequenza nei differenti siti: il range di variazione dell'L-CV ($\Delta\tau$) nella regione è 0.5 volte la media regionale dell'L-CV (τ^R); i valori degli altri parametri sono riportati nel titolo di ogni sottofigura (da *Viglione et al.*, 2007a).

In Figura 5.3 sono riportati i risultati dei test per regioni simulate la cui

eterogeneità è dovuta alla differente dispersione delle distribuzioni di frequenza nei diversi siti. Il range di variazione degli L-CV ($\Delta\tau$) nella regione è pari a 0.5 volte l'L-CV medio regionale (τ^R). Siccome $k = 11$, in una regione con $\tau^R = 0.2$ i campioni sono generati da distribuzioni caratterizzate da valori di τ rispettivamente uguali a $0.15, 0.16, 0.17, ..., 0.25$. I numeri grigi e le linee di tendenza in Figura 5.3 mostrano la potenza dei test, ovvero in percentuale il numero di volte che il test riesce a riscontrare l'eterogeneità. È evidente la mancanza di potenza della misura HW_2, cosa già evidenziata dagli autori stessi (*Hosking* & *Wallis*, 1997). Per tutti gli altri test, la potenza tende ad essere maggiore sulla diagonale centrale del diagramma $\tau - \tau_3$ e a crescere verso l'angolo in alto a destra dello spazio investigato. HW_1, se paragonato ai test DK e AD, è più potente nella parte in basso a sinistra dello spazio degli L-momenti. Invece, per regioni molto asimmetriche ha una potenza considerevolmente inferiore a quella dei test basati sui ranghi, tra i quali il test AD è il migliore.

5.3.2 Analisi di sensitività

Come menzionato nel Paragrafo 5.2, gli effetti di una variazione di k, n, $\mathcal{P}$, $\Delta\tau$, $\Delta\tau_3$, e α è considerata in quattro punti (A, B, C e D) posizionati nella parte centrale del diagramma $\tau - \tau_3$ (Figura 5.1), e non sull'intero diagramma. Come esempio viene riportato in Figura 5.4 il comportamento dei test per regioni la cui eterogeneità è dovuta solamente al parametro di forma ($\Delta\tau = 0$, $\Delta\tau_3 \neq 0$). In questo caso i test non-parametrici, in particolare il test AD, e la misura di eterogeneità di Hosking e Wallis HW_2 sono (ovviamente) più potenti di HW_1. Questo fatto è particolarmente evidente quando il parametro di forma medio è piuttosto grande ($\tau_3^R \geq 0.2$), mentre per bassi valori di τ_3^R (punto A) tutti i test falliscono l'individuazione di eterogeneità. Come previsto, la potenza dei test aumenta all'aumentare del grado di eterogeneità, ovvero all'aumentare di $\Delta\tau_3$.

Come secondo esempio, in Figura 5.5 viene mostrata la potenza dei test per regioni estratte da distribuzioni generatrici differenti, quando l'eterogeneità è dovuta esclusivamente alla diversità nell'L-CV ($\Delta\tau = 0.5\tau^R$). Oltre alla distribuzione GEV, che è considerata nel caso studio principale, le altre distribuzioni a 3 parametri adottate sono la logistica generalizzata (GL), la lognormale a tre parametri (LN), la distribuzione di Pearson tipo III (P3) e la Pareto generalizzata (GP). Si veda l'Appendice B per la descrizione della parametrizzazione di

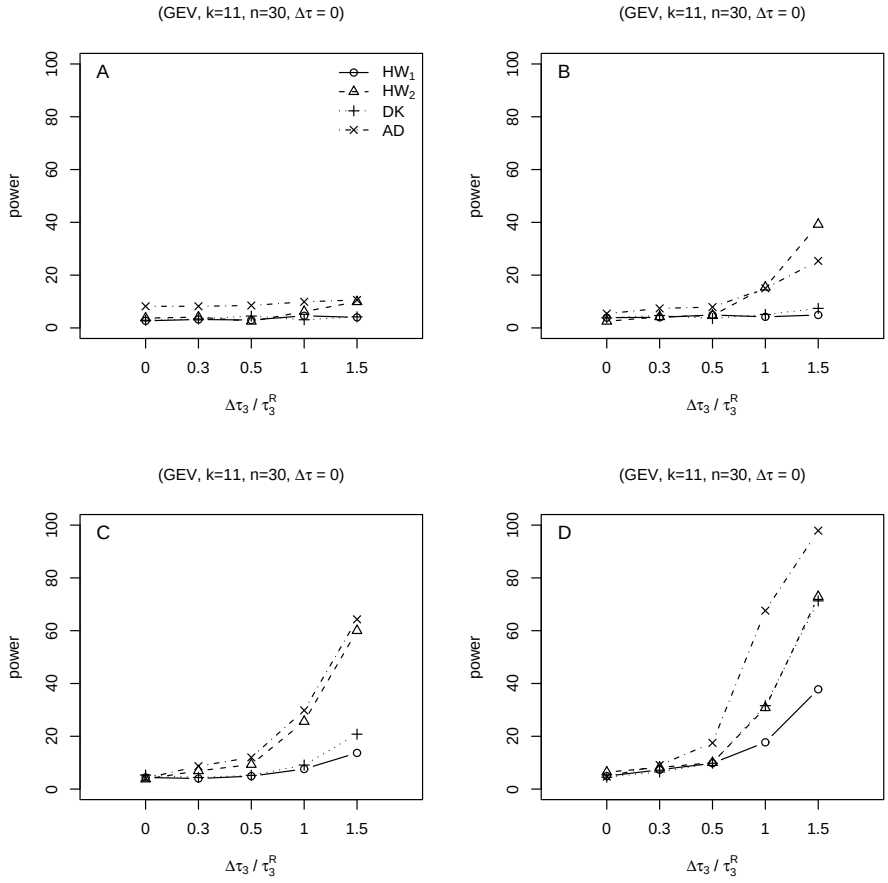

Figura 5.4: Potenza dei test nei punti A, B, C e D (Figura 5.1) quando l'eterogeneità è dovuta al solo parametro di forma τ_3 (vedi Paragrafo 5.3.2); i valori degli altri parametri sono riportati nel titolo di ogni sottofigura (da *Viglione et al.*, 2007a).

queste distribuzioni e le relazioni tre i loro parametri e gli L-momenti. I quattro test si comportano in maniera molto simile al variare della distribuzione generatrice: nel punto A (bassa asimmetria) la misura di eterogeneità di Hosking e Wallis supera i test non-parametrici, mentre nel punto D (alta asimmetria) è vero il contrario. I punti B e C riflettono la transizione tra i due casi estremi, e sono caratterizzati da una sostanziale equivalenza di potenza dei diversi test. In tutti i casi la misura di eterogeneità HW_2 è caratterizzata da una bassa potenza di discriminazione tra regioni omogenee e regioni eterogenee.

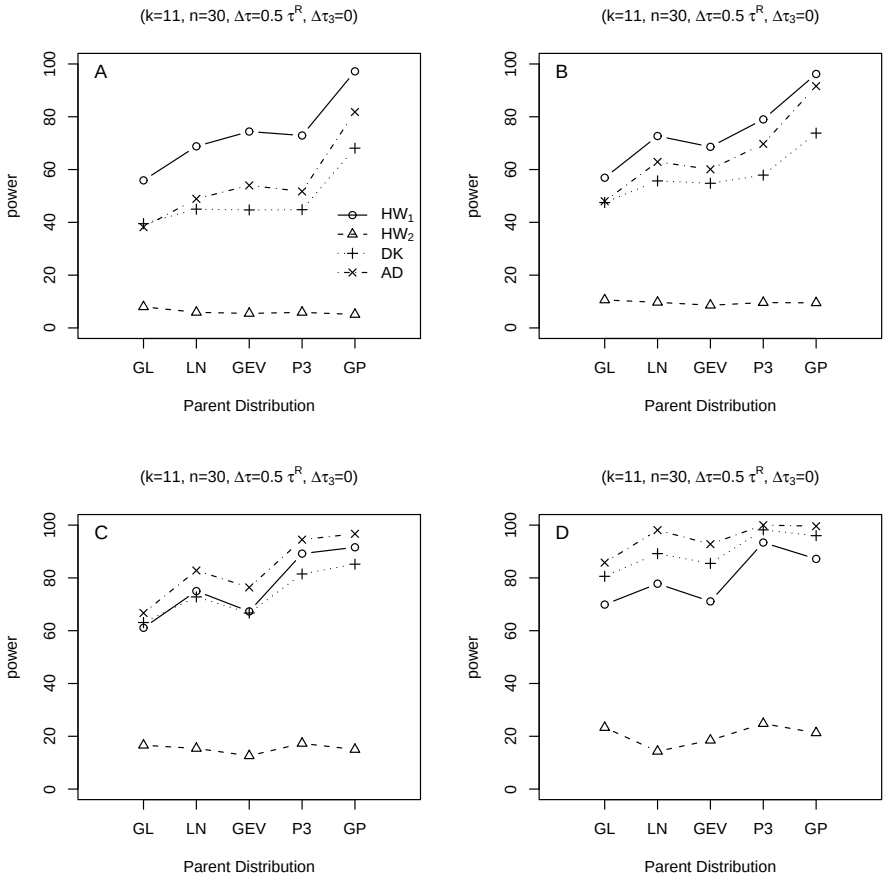

Figura 5.5: Potenza dei test nei punti A, B, C e D (Figura 5.1) al variare della distribuzione generatrice (vedi Paragrafo 5.3.2); i valori degli altri parametri sono riportati nel titolo di ogni sottofigura (da *Viglione et al.*, 2007a).

Gli effetti della variazione degli altri parametri è più banale, e i diagrammi corrispondenti non sono stati riportati: la potenza dei test cresce all'aumentare del numero di siti k nelle regioni, e all'aumentare della lunghezza dei campioni n. I test sono molto più sensibili alla lunghezza delle serie (si è considerata la variazione di n tra 10 e 100) che dal numero di siti k (che si è fatto variare da 3 a 21). Per quanto riguarda l'aumento del grado di eterogeneità nel parametro di dispersione ($\Delta\tau/\tau^R$), il suo effetto è ovviamente di incrementare la potenza dei test. La potenza raggiunge il valore massimo (100%) quando $\Delta\tau/\tau^R = 1$ (tranne

che per HW_2). In tutti i casi considerati il test HW_1 è più potente nei punti A e B, mentre i test DK e AD sono più potenti nei punti C e D. Le differenze in potenza possono essere rilevanti, sotto un punto di vista pratico, specialmente per gradi di eterogeneità intermedi.

5.4 Discussione dei risultati

Un problema pratico dell'analisi di frequenza regionale è la scelta di un test statistico per la valutazione dell'omogeneità delle regioni. In questo capitolo, le misure di eterogeneità di Hosking e Wallis (basate sui rapporti degli L-momenti) sono state confrontate con i test basati sui ranghi di Anderson-Darling e di Durbin e Knott. Il confronto mostra come la misura di eterogeneità di Hosking e Wallis HW_1 (basata solamente sull'L-CV) sia preferibile nel caso di bassa asimmetria, mentre il test bootstrap di Anderson-Darling dovrebbe essere usato per regioni più asimmetriche. Per quanto riguarda HW_2, la misura di eterogeneità di Hosking e Wallis basata su L-CV e L-CA, questo lavoro dimostra ancora una volta quanto essa manchi di potenza.

Come suggerimento per la scelta del test da utilizzarsi, si propone li grafico di Figura 5.6 che è stato ottenuto come compromesso tra potenza ed errore di Tipo I dei test HW_1 e AD. Lo spazio degli L-momenti è diviso in due regioni: se il coefficiente t_3^R della regione analizzata è minore di 0.23, si propone di utilizzare la misura di eterogeneità di Hosking e Wallis HW_1; se $t_3^R > 0.23$, il test bootstrap di Anderson-Darling è preferibile. Ulteriori commenti sorgono dall'osservazione della Figura 5.6 che mostra alcuni punti (t^R, t_3^R) associati a regioni individuate da altri autori. Ognuno di questi punti rappresenta una regione omogenea considerata in tre studi di analisi regionale delle piene: *Hosking & Wallis* (1997) che fornisce i valori di t^R e t_3^R per alcune regioni nell'area dei monti Apalachi; *De Michele & Rosso* (2002) e *Farquharson et al.* (1987) che forniscono i tre parametri (stimati a partire dagli L-momenti) della distribuzione GEV per molte regioni in Italia (*De Michele & Rosso*, 2002) e nel mondo (*Farquharson et al.*, 1987). Si noti come queste regioni costruite empiricamente stiano nella parte dello spazio dei parametri considerata in questo studio, e che la maggior parte di questi punti appartenga alla parte dello spazio $\tau - \tau_3$ dove il test bootstrap di Anderson-Darling è più potente.

La buona performance della misura di eterogeneità di Hosking e Wallis HW_1,

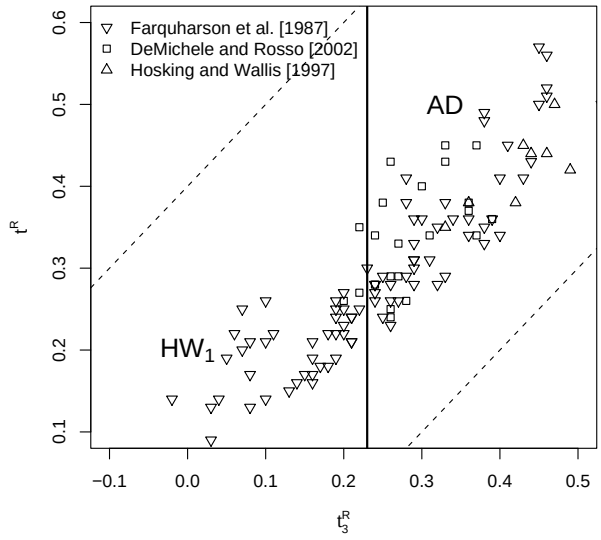

Figura 5.6: Regioni dello spazio $\tau - \tau_3$ dove i test considerati dovrebbero essere utilizzati; a sinistra della linea nera ($t_3^R = 0.23$) la misura di eterogeneità di Hosking e Wallis HW_1 è il test migliore (in base a potenza ed errore di Tipo I), a destra dovrebbe essere usato il test bootstrap di Anderson-Darling AD. Alcuni valori regionali desunti da casi studi reali sono riportati come punti sul grafico: *Farquharson et al.* (1987) ha calcolato questi valori considerando molte stazioni sparse per il mondo, *De Michele* & *Rosso* (2002) considerando l'Italia e *Hosking* & *Wallis* (1997) la regione degli Appalachi (da *Viglione et al.*, 2007a).

molto usata in idrologia, merita un commento ulteriore. Il test HW_1 è esclusivamente basato sul coefficiente L-CV (si vedano le Equazioni (5.5) and (5.6)), ed il fatto che si comporti così bene suggerisce che l'eterogeneità tra i siti sia principalmente dovuta alla variabilità della varianza campionaria. Invece le variazioni in skewness e kurtosis sono in molti casi mascherate dalla variabilità dei momenti, o L-momenti, di ordine superiore. Come conseguenza, altri test di costanza della varianza in differenti campioni potrebbero essere utilizzati come alternative a HW_1. Esempi di questo tipo sono i classici test di Levene e Barlett (*Conover et al.*, 1981) che, comunque, in un primo caso studio considerato, si sono rivelati più deboli di HW_1.

Capitolo 6

Applicazione: Analisi Regionale del deflusso annuo

Per molti problemi pratici dell'idrologia delle acque superficiali, come la gestione delle risorse idriche o lo studio degli eventi di piena, è importante poter far affidamento su informazioni relative ai deflussi medi e di magra che siano nello stesso tempo accurate e diffuse sul territorio.

La variabile presa in considerazione in questo capitolo è il deflusso annuo, ossia il volume d'acqua che transita in una sezione di un corso d'acqua in un anno. Le applicazioni dell'analisi regionale a grandezze medie annue sono molto meno numerose di quelle relative alle piene. A riguardo, *Vogel & Wilson* (1996) danno una breve rassegna dei metodi, soffermandosi sui lavori effettuati negli Stati Uniti. Per quanto riguarda l'Italia ricordiamo i lavori di *Ferraresi et al.* (1988), di *Claps & Mancino* (2002) per la Basilicata e di *Brath et al.* (2004) per la Romagna.

In questo capitolo utilizzeremo il termine deflusso-indice per indicare il valore caratteristico del deflusso annuo. Da quanto detto nel Paragrafo 2.1, si evince che, per variabili aleatorie caratterizzate da asimmetria bassa, la stima della media è meno distorta di quella della mediana. È quindi preferibile usare la media campionaria come valore indice per il deflusso annuo, considerata la sua bassa asimmetria. Indicheremo con D_m la media teorica del deflusso annuo, con $\tilde{D}_m$ quella campionaria in sezioni monitorate e con $\hat{D}_m$ quella stimata in sezioni senza dati.

La curva di frequenza di D, adimensionalizzata rispetto al deflusso indice $q(F) = D(F)/D_m$, è detta curva di crescita (vedi Capitolo 3). Anche in questo caso distingueremo la curva teorica, quella campionaria e quella stimata utilizzando i simboli $q(F)$, $\tilde{q}(F)$ e $\hat{q}(F)$.

L'analisi di frequenza regionale del deflusso annuo viene suddivisa in:

- stima regionale del deflusso indice D_m;
- stima regionale della curva di crescita $q(F)$.

Per quanto riguarda la prima parte dello studio di regionalizzazione si è ricercato un modello che legasse D_m alle caratteristiche morfoclimatiche dei bacini per i quali sono disponibili misure idrometriche. A questo scopo si sono utilizzati i metodi di regressione lineare multipla descritti nel Capitolo 2. Per quanto riguarda la seconda parte i criteri oggettivi di raggruppamento dei bacini idrografici in regioni omogenee proposti nei Capitoli 3 e 4 vengono impiegati, sfruttando come variabili discriminanti alcune grandezze morfologiche e climatiche dei bacini idrografici.

L'analisi di frequenza del deflusso annuo è stata applicata nelle regioni Piemonte e Valle d'Aosta. L'eterogeneità che contraddistingue questo territorio rende particolarmente complicata, e nello stesso tempo interessante, la ricerca di modelli regionali di stima delle variabili idrologiche. In questo ambito spaziale relativamente limitato coesistono infatti conformazioni orografiche e situazioni climatiche estremamente differenti: in poche centinaia di chilometri si va dal clima appenninico-mediterraneo del sud-est collinare piemontese a quello alpino-continentale della montagnosa Valle d'Aosta, passando per tutte le condizioni intermedie.

6.1 Dati utilizzati

Per la stima dei modelli regionali si sono considerati 47 bacini idrografici Piemontesi e Valdostani sottesi da stazioni idrometriche del SIMN (Servizio Idrografico e Mareografico Nazionale) con superfici comprese tra 20 e 8000 km^2 (Figura 6.1).

I dati medi di afflusso e deflusso sono stati reperiti nella Pubblicazione n. 17 del Servizio Idrografico, che riporta i dati caratteristici dei corsi d'acqua italiani aggiornati all'anno 1970. Le serie storiche dei deflussi sono state integrate

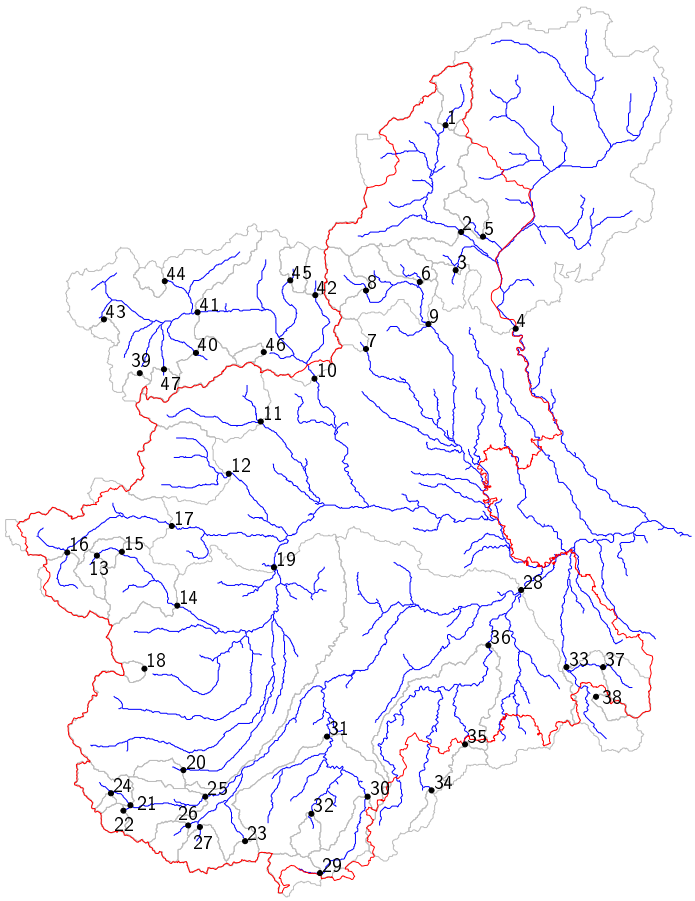

Figura 6.1: Bacini idrografici sottesi da stazioni idrometriche del Servizio Idrografico utilizzati nello studio (la numerazione corrisponde al codice dei bacini come indicato in Tabella 6.I).

fino all'anno 1986, dove possibile, con i dati degli Annali Idrologici del Servizio Idrografico.

Su questi bacini sono state condotte in precedenza un'analisi morfometrica ed uno studio di stima dell'evapotraspirazione media mensile (*Viglione et al.*, 2007b), i risultati dei quali sono stati utilizzati in questo lavoro.

Le variabili morfoclimatiche dei bacini prese in considerazione per l'analisi regionale comprendono (i valori sono riportati in Tabella 6.I per i 47 bacini):

- Afflusso medio annuo A_m [mm];
- Superficie S [km^2];

Tabella 6.I: Parametri morfoclimatici considerati nello studio per i bacini idrografici. (1/2)

cod	nome	$\tilde{D}_m$ [mm]	A_m [mm]	S [km^2]	H_m [m s.l.m.]	P_m [%]
1	Toce a Cadarese	1571	1457	190	2137	22.10
2	Toce a Candoglia	1382	1519	1540	1674	7.70
3	Niguglia a Omegna	1353	1901	122	637	4.80
4	Ticino a Miorina	1395	1695	6692	1286	2.60
5	S.Bernardino a Santino	1730	2113	119	1251	17.90
6	Mastallone a Ponte Folle	1600	1936	147	1319	12.60
7	Cervo a Passobreve	1461	1803	75	1490	20.20
8	Sesia a Campertogno	1275	1427	170	2112	19.80
9	Sesia a Ponte Aranco	1428	1735	703	1491	8.30
10	Dora Baltea a Tavagnasco	918	949	3311	2090	6.60
11	Orco a Pont Canavese	1034	1263	615	1924	12.10
12	Stura di Lanzo a Lanzo	1090	1296	577	1773	10.60
13	Chisone a Soucheres Basses	819	966	92	2222	15.40
14	Chisone a S.Martino	694	1058	581	1730	11.20
15	Chisone a Fenestrelle	654	910	157	2144	15.90
16	Dora Riparia a Oulx	663	851	254	2165	13.20
17	Dora Riparia a S.Antonino	591	841	993	1867	9.90
18	Po a Crissolo	1254	1271	38	2261	28.80
19	Po a Moncalieri	507	952	5032	924	0.80
20	Grana a Monterosso	811	1135	103	1565	15.00
21	Stura di Demonte a Pianche	925	1112	180	2074	17.30
22	Rio Bagni ai Bagni di Vinadio	1241	1398	62	2138	19.10
23	Vermenagna a Limone	1128	1364	57	1677	17.60
24	Rio del Piz a Pietraporzio	1272	1273	21	2194	38.70
25	Stura di Demonte a Gaiola	1011	1219	560	1814	10.20
26	Gesso della Valletta a S.Lorenzo	1384	1392	110	2105	22.30
27	Gesso di Entracque ad Entracque	1404	1468	157	1894	17.00
28	Tanaro a Montecastello	501	997	8024	651	0.80
29	Tanaro a Ponte Nava	1030	1281	148	1576	11.30
30	Tanaro a Nucetto	902	1233	376	1222	7.20
31	Tanaro a Farigliano	776	1120	1516	938	2.70
32	Corsaglia a Molline	1068	1366	89	1513	17.80
33	Scrivia a Serravalle	827	1389	616	688	3.50
34	Bormida di Mallare a Ferrania	965	1228	50	602	5.90
35	Erro a Sassello	882	1200	83	605	5.30
36	Bormida a Cassine	510	971	1542	481	1.80
37	Borbera a Baracche	779	1220	202	867	6.40
38	Vobbia a Vobbietta	835	1461	57	727	9.70
39	Dora di Rhemes a Pelaud	1453	1041	54	2743	24.00
40	Grand'Eyvia a Cretaz	1109	940	179	2593	17.10
41	Dora Baltea a Aosta	898	952	1824	2267	8.40
42	Lys a Gressoney	1357	1191	91	2625	24.90
43	Rutor a Promise	1648	1314	46	2512	31.10
44	Artanavaz a St.Oyen	1023	1283	71	2229	22.30
45	Evancon a Champoluc	977	1048	105	2631	20.20
46	Ayasse a Champorcher	1258	1179	41	2352	30.40
47	Savara a Eau Rousse	1079	987	84	2723	22.90

Tabella 6.I: Parametri morfoclimatici considerati nello studio per i bacini idrografici. (2/2)

cod	L_{LDP} [km]	P_{LDP} [%]	S_{2000} [%]	EST	$NORD$	R_c	X_{bar} [deg]	Y_{bar} [deg]	I_T	I_B
1	31.6	18.20	66.0	−0.29	−0.96	0.52	8.397	46.375	1.52	0.65
2	82.4	10.20	36.4	0.63	−0.78	0.31	8.225	46.149	1.31	0.63
3	16.0	8.30	0.0	0.33	0.94	0.41	8.384	45.821	1.28	0.51
4	168.1	7.80	20.2	0.00	−1.00	0.30	8.652	46.169	1.32	0.56
5	22.6	26.30	2.4	0.51	−0.86	0.53	8.456	46.035	1.92	0.45
6	23.8	22.60	6.3	0.53	−0.85	0.49	8.206	45.888	1.69	0.50
7	14.4	22.90	13.5	0.70	−0.71	0.62	7.978	45.679	1.49	0.54
8	21.8	26.10	57.4	0.90	−0.44	0.49	7.936	45.838	1.10	0.77
9	62.2	16.20	21.9	0.75	−0.66	0.47	8.091	45.833	1.04	0.68
10	110.9	10.80	58.1	0.85	−0.52	0.39	7.395	45.728	0.50	1.04
11	47.9	18.60	46.7	0.93	−0.36	0.43	7.425	45.470	0.98	0.79
12	40.3	21.00	37.4	0.99	−0.17	0.54	7.287	45.290	0.90	0.77
13	17.0	17.60	73.3	0.25	0.97	0.49	6.938	44.974	0.48	1.06
14	56.6	13.70	36.9	0.87	−0.49	0.49	7.084	44.963	0.48	0.96
15	26.6	15.30	64.5	0.88	0.47	0.41	6.965	45.001	0.37	1.12
16	34.9	16.80	63.9	−0.09	1.00	0.46	6.851	44.932	0.24	1.20
17	78.0	11.80	46.3	0.99	0.16	0.24	6.912	45.070	0.16	1.20
18	8.4	29.30	73.7	0.97	0.25	0.74	7.115	44.693	1.03	0.83
19	114.0	5.40	14.5	0.61	0.79	0.39	7.398	44.736	0.18	1.10
20	19.0	19.50	20.5	0.99	0.13	0.54	7.240	44.403	0.67	0.93
21	26.8	16.10	61.8	0.83	−0.55	0.47	7.007	44.356	0.76	0.96
22	9.7	26.00	66.6	0.72	0.70	0.73	7.053	44.267	1.25	0.77
23	10.7	23.10	20.9	0.06	1.00	0.56	7.576	44.178	1.09	0.79
24	8.3	24.70	71.5	0.42	0.91	0.55	7.018	44.311	1.06	0.84
25	55.3	12.10	43.1	1.00	0.09	0.41	7.137	44.316	0.85	0.88
26	17.1	19.50	61.2	0.79	0.61	0.59	7.277	44.212	1.28	0.77
27	16.9	23.80	44.4	−0.10	0.99	0.61	7.407	44.182	1.31	0.73
28	209.8	6.40	6.0	0.75	0.66	0.28	8.064	44.548	0.18	1.07
29	19.5	23.80	17.6	1.00	−0.04	0.50	7.771	44.124	0.90	0.84
30	55.4	16.10	7.8	0.60	0.80	0.28	7.901	44.179	0.65	0.87
31	93.2	12.30	5.2	0.17	0.98	0.54	7.852	44.298	0.47	0.96
32	18.8	20.60	17.0	0.17	0.99	0.58	7.828	44.226	1.19	0.79
33	51.9	8.10	0.0	−0.80	0.60	0.51	9.040	44.628	0.77	0.76
34	18.0	9.60	0.0	0.26	0.97	0.39	8.300	44.297	0.38	0.87
35	17.6	6.40	0.0	0.00	1.00	0.36	8.458	44.447	0.30	0.88
36	131.1	5.80	0.0	0.50	0.87	0.35	8.322	44.500	0.04	1.09
37	25.3	13.40	0.0	−0.80	0.61	0.57	9.112	44.668	0.57	0.86
38	14.9	14.90	0.0	−0.81	0.58	0.55	9.046	44.605	0.88	0.72
39	12.6	19.60	97.8	0.36	0.93	0.57	7.091	45.514	1.09	0.97
40	15.3	25.80	86.4	−0.60	0.80	0.53	7.377	45.584	0.75	1.06
41	55.7	15.40	67.9	0.99	0.11	0.29	7.177	45.718	0.60	1.04
42	16.4	22.50	84.2	−0.03	−1.00	0.59	7.830	45.855	1.23	0.83
43	10.8	27.30	89.7	−0.19	0.98	0.53	6.970	45.672	1.26	0.75
44	11.9	22.10	71.8	0.98	−0.19	0.59	7.151	45.828	1.15	0.76
45	15.1	21.20	88.6	−0.31	−0.95	0.54	7.742	45.872	0.96	0.94
46	12.4	17.90	82.4	0.98	0.19	0.56	7.559	45.613	0.95	0.84
47	11.5	28.30	95.8	0.10	1.00	0.60	7.206	45.523	1.11	1.02

- Altitudine media H_m [m s.m.m.];

- Pendenza media P_m [%]: angolo alla base del triangolo rettangolo che ha per base la radice quadrata dell'area del bacino e per altezza il doppio dell'altitudine mediana (relativa alla sezione di chiusura) del bacino. In sostanza tale pendenza è calcolata rispetto ad un bacino di forma quadrata equivalente a quello reale, e non tiene conto della sua effettiva forma, che può essere più o meno allungata. Si è scelto di calcolare P_m in questo modo, piuttosto che con i classici strumenti di analisi del DEM (Digital Elevation Model), in modo che il suo valore non fosse influenzato dalla risoluzione del DEM stesso;

- Lunghezza del Longest Drainage Path L_{LDP} [km]: lunghezza del percorso tra la sezione di chiusura ed il punto più lontano da essa, sul bordo del bacino, seguendo le direzioni di drenaggio. Esso coincide sostanzialmente con l'asta principale;

- Pendenza media del Longest Drainage Path P_{LDP} [%]: valore medio delle pendenze associate ad ogni pixel del Longest Drainage Path;

- Percentuale d'area a quota superiore ai 2000 m s.m.m. S_{2000} [%];

- Easting EST: seno dell'angolo φ formato dal vettore di orientamento con il nord (Figura 6.2). Il vettore di orientamento è quel segmento che unisce il baricentro del bacino con la sezione di chiusura. Il parametro EST ha valore massimo uguale a 1 se il bacino è orientato verso est e valore minimo pari a -1 se è orientato verso ovest;

- Northing $NORD$: coseno dell'angolo φ formato dal vettore di orientamento con il nord che ha valore massimo uguale a 1 se il bacino è orientato verso nord e valore minimo pari a -1 se è orientato verso sud (Figura 6.2);

- Rapporto di circolarità R_c: rapporto tra l'area del bacino e l'area del cerchio avente lo stesso perimetro del bacino;

- Coordinate del baricentro del bacino X_{bar} e Y_{bar} [deg] espresse come longitudine e latitudine (nel sistema di riferimento di Greenwich) in gradi esadecimali (Figura 6.2);

Figura 6.2: Esempio di calcolo dei parametri geometrici di bacino. S è la superficie del bacino, X_{bar} ed Y_{bar} le coordinate del baricentro del bacino e φ l'angolo del vettore di orientamento (che unisce il baricentro alla sezione di chiusura del bacino) con il nord N. Easting e northing sono, rispettivamente, $EST = \sin\varphi$ e $NORD = \cos\varphi$.

- Indice di Thornthwaite I_T: indice di umidità globale che, nella forma più semplice, si presenta come un indice di bilancio idrico a scala annua:

$$I_T = \frac{A_m - ET_p}{ET_p} \ , \tag{6.1}$$

dove A_m è la precipitazione media annua ed ET_p l'evapotraspirazione potenziale media annua relative al bacino. La seconda grandezza è stata stimata tramite la formulazione di Hargreaves (v.es. *Viglione et al.*, 2007b);

- Indice di Budyko I_B: indice di aridità radiazionale che si esprime come:

$$I_B = \frac{R_n}{\lambda A_m} \tag{6.2}$$

dove R_n è la radiazione netta media annua e λ è il calore latente di vaporizzazione. I valori assunti da I_B sono inferiori all'unità in regioni umide e superiori in regioni aride. La radiazione netta è stata stimata nel lavoro di *Viglione et al.* (2007b).

6.2 Stima della grandezza-indice

Come si è detto, i metodi multi-regressivi sono i metodi più comunemente utilizzati per la stima della grandezza-indice in siti sprovvisti di dati misurati. Se si considera la variabile deflusso (deflusso annuo, portata di piena, ...), l'approccio multi-regressivo lega il deflusso-indice alle caratteristiche di bacino, quali gli indici climatici, i parametri geologici e morfometrici, la copertura del suolo, e così via.

Per la stima di D_m si sono valutati quattro diversi modelli di regressione lineare multipla:

$$D_m = \beta_0 + \beta_1 M_1 + \beta_2 M_2 + \ldots + \beta_{p-1} M_{p-1} + \varepsilon \ , \tag{6.3}$$

$$D_m^{1/2} = \beta_0 + \beta_1 M_1 + \beta_2 M_2 + \ldots + \beta_{p-1} M_{p-1} + \varepsilon \ , \tag{6.4}$$

$$D_m^{1/3} = \beta_0 + \beta_1 M_1 + \beta_2 M_2 + \ldots + \beta_{p-1} M_{p-1} + \varepsilon \ , \tag{6.5}$$

$$\ln(D_m) = \beta_0 + \beta_1 M_1 + \beta_2 M_2 + \ldots + \beta_{p-1} M_{p-1} + \varepsilon \ , \tag{6.6}$$

dove M_i sono parametri morfoclimatici e β_i sono i coefficienti della regressione. Per la stima dei coefficienti delle Equazioni (6.3)-(6.5) si utilizza la tecnica dei minimi quadrati ordinari (si veda il Paragrafo 2.2).

Considerando i quattro tipi di regressione, 14 variabili morfoclimatiche (Paragrafo 6.1) e, per dare maggiore peso alle differenze nelle piccole scale, la trasformazione logaritmica di 4 di esse ($\ln(A_m)$, $\ln(S)$, $\ln(H_m)$ e $\ln(P_m)$), si sono confrontati in tutto più di un milione di modelli (esattamente $4 \cdot 2^{18}$).

Innanzitutto si sono esclusi quei modelli per i quali anche solo una delle variabili esplicative risultasse non significativa in base al test della t di Student all'1% (vedi Paragrafo 2.3.1). Successivamente si è valutata la capacità descrittiva di ogni regressione tramite il coefficiente di determinazione corretto R^2_{adj}. Il coefficiente di determinazione R^2_{adj} è utile per scegliere il miglior modello tra quelli di una delle quattro classi (Equazioni (6.3)-(6.5)) ma non può essere usato per confrontare modelli di differente natura. A questo scopo è stato applicato un metodo di cross-validazione, calcolando la radice dell'errore quadratico medio dei residui $\mathrm{RMSE}_{\mathrm{cv}}$ come spiegato nel Paragrafo 2.3.1.

Per ogni classe (Equazioni (6.3)-(6.5)) si sono scelti i 5 migliori modelli multiregressivi, in base al coefficiente di determinazione corretto R^2_{adj} (Tabella 6.II). Tra questi si sono scelti infine due modelli: quello con minore $\mathrm{RMSE}_{\mathrm{cv}}$ (il migliore), ed il modello che fa uso dei parametri morfoclimatici più facilmente disponibili (il più semplice).

Tabella 6.II: Migliori regressioni tra D_m e le sue trasformate (Dip) e le variabili morfoclimatiche (Ind); sono riportati R^2_{adj} (riferito alla variabile trasformata), RMSE e RMSE_{cv} (riferiti alla variabile originale D_m).

Dip	Ind				R^2_{adj}	RMSE	RMSE_{cv}
D_m	S_{2000}	$\ln(Am)$			0.877	108.7	116.6
	A_m	S_{2000}			0.876	109.3	116.9
	H_m	$\ln(A_m)$			0.865	114.1	122.2
	S_{2000}	$\ln(I_B)$			0.862	115.2	123.0
	NORD	Y_{bar}	$\ln(A_m)$	$\ln(H_m)$	0.862	112.9	127.8
	NORD	$\ln(A_m)$	$\ln(H_m)$	$\ln(Y_{bar})$	0.861	113.0	127.9
$D_m^{1/2}$	S_{2000}	$\ln(A_m)$			0.888	106.0	113.5
	H_m	NORD	$\ln(I_B)$		0.887	104.5	114.6
	H_m	$\ln(A_m)$			0.880	109.2	116.6
	H_m	I_B			0.875	112.8	120.3
	A_m	S_{2000}			0.874	110.5	118.5
	S_{2000}	I_B			0.870	116.0	124.2
$D_m^{1/3}$	S_{2000}	$\ln(A_m)$			0.888	105.7	113.1
	H_m	NORD	$\ln(I_B)$		0.888	104.7	114.9
	H_m	$\ln(A_m)$			0.883	108.5	115.8
	H_m	I_B			0.879	111.9	119.2
	S_{2000}	I_B			0.873	116.0	124.1
	A_m	S_{2000}			0.870	111.8	120.3
$\ln(D_m)$	H_m	NORD	I_B		0.900	101.8	110.5
	A_m	H_m	NORD	$\ln(X_{bar})$	0.888	102.1	116.2
	H_m	NORD	$\ln(I_B)$		0.884	107.3	118.1
	S_{2000}	$\ln(A_m)$			0.884	106.2	113.5
	A_m	S_{2000}	$\ln(I_T)$		0.883	104.6	114.2
	H_m	$\ln(A_m)$			0.883	108.7	116.2

La regressione risultata migliore a tali criteri è la seguente:

$$\ln(\hat{D}_m) = 7.86 + 2.91 \cdot 10^{-4} \cdot H_m + 7.22 \cdot 10^{-2} \cdot NORD - 1.70 \cdot I_B \ , \qquad (6.7)$$

caratterizzata da un coefficiente di determinazione $R^2_{adj} = 0.900$ e da $\text{RMSE}_{\text{cv}} = 110.5$ mm (riferito alla variabile non trasformata D_m). Delle variabili utilizzate per la scrittura dell'Equazione (6.7) l'indice di Budyko I_B non è di facile stima. La determinazione della radiazione solare netta R_n sul bacino (si veda l'Equazione (6.2)) richiede la conoscenza delle distribuzioni spaziali di temperatura, umidità relativa e nuvolosità media (o stime di esse). Per questo motivo si è deciso di considerare anche modelli i cui parametri fossero di più semplice determinazione. Il migliore tra questi risulta essere:

$$\hat{D}_m^{1/3} = -22.7 + 4.37 \cdot \ln(A_m) + 10^{-3} \cdot H_m \ , \qquad (6.8)$$

caratterizzato da un coefficiente di determinazione $R^2_{adj} = 0.883$ e da $\text{RMSE}_{\text{cv}} =$

115.8 mm. Una relazione analoga alla (6.8) è stata ottenuta nella regionalizzazione del deflusso annuo in Basilicata (*Claps et al.*, 1998).

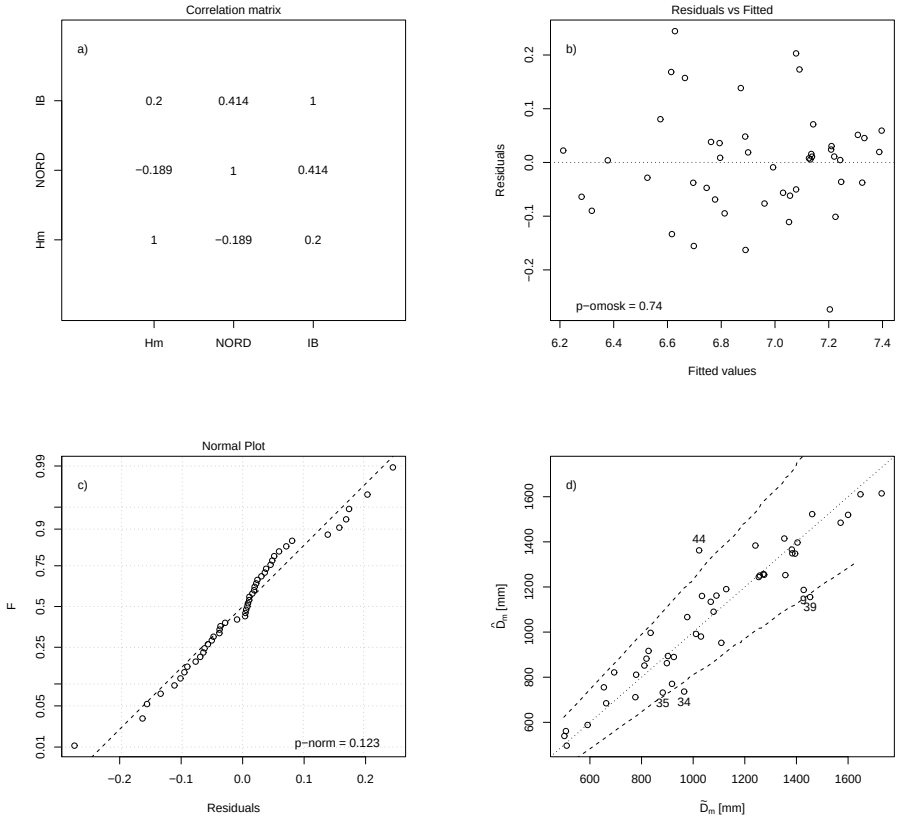

Figura 6.3: Grafici diagnostici della Regressione (6.7) del deflusso medio annuo D_m (da *Viglione et al.*, 2007b): a) coefficienti di correlazione tra le variabili indipendenti; b) rappresentazione stime-residui (variabile trasformata) e probabilità associata al test di omoschedasticità di Harrison-McCabe; c) rappresentazione dei residui in carta probabilistica normale e probabilità associata al test di normalità di Anderson-Darling; d) risultato della cross validazione ed intervalli di predizione di D_m (si sono indicati i siti caratterizzati dagli errori maggiori: Sesia a Ponte Aranco (9), Bormida di Mallare a Ferrania (34), Erro a Sassello (35), Dora di Rhemes a Pelaud (39) e Artavanaz a St.Oyen (44)).

Le Figure 6.3 e 6.4 riproducono alcuni grafici diagnostici delle due regres-

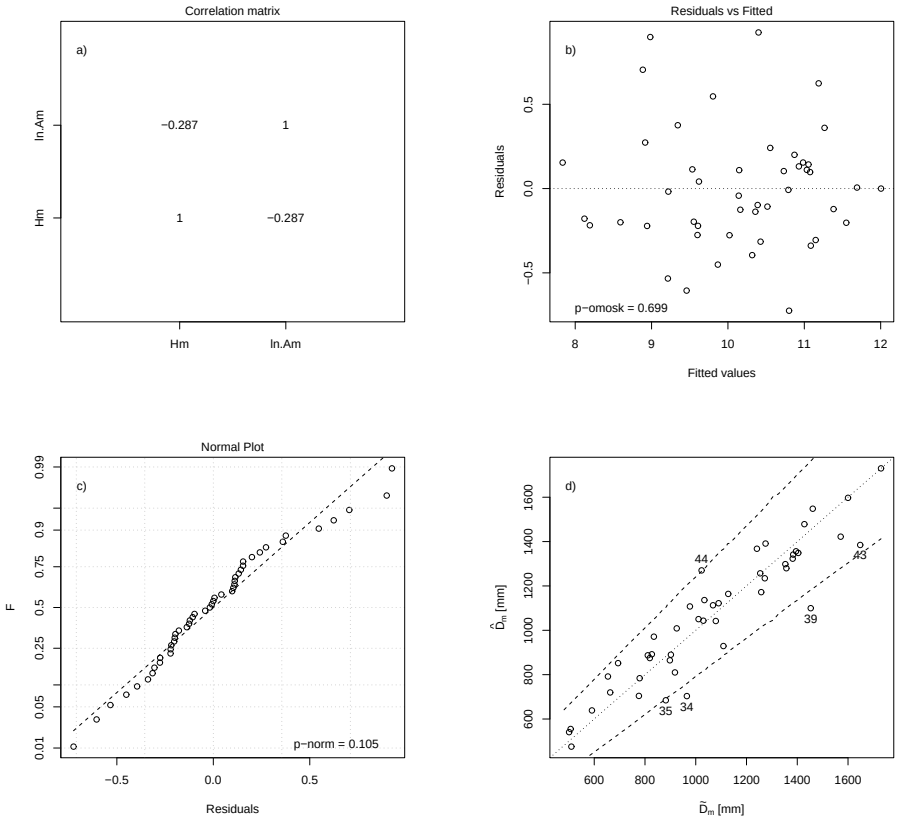

Figura 6.4: Grafici diagnostici della Regressione (6.8) del deflusso medio annuo D_m (da *Viglione et al.*, 2007b): a) coefficienti di correlazione tra le variabili indipendenti; b) rappresentazione stime-residui (variabile trasformata) e probabilità associata al test di omoschedasticità di Harrison-McCabe; c) rappresentazione dei residui in carta probabilistica normale e probabilità associata al test di normalità di Anderson-Darling; d) risultato della cross validazione ed intervalli di predizione di D_m (si sono indicati i siti caratterizzati dagli errori maggiori: Bormida di Mallare a Ferrania (34), Erro a Sassello (35), Dora di Rhemes a Pelaud (39), Rutor a Promise (43) e Artavanaz a St.Oyen (44)).

sioni considerate. Procedendo per righe, il primo grafico rappresenta la matrice dei coefficienti di correlazione tra le variabili indipendenti del modello. La presenza di correlazione tra di essi, quindi valori che si avvicinano a ± 1 (cosa che non avviene nel nostro caso), è da ritenersi problematica perché indice di mul-

ticollinearità (vedi Paragrafo 2.3.2). Oltre ai coefficienti di correlazione è stato calcolato l'indice VIF (vedi Paragrafo 2.3.2): per la Regressione (6.7) i valori dell'indice sono 1.15 per H_m, 1.33 per $NORD$ e 1.34 per I_B; per la regressione (6.8) sono 1.09 sia per H_m che per $\ln(A_m)$. In tutti i casi il valore del VIF è ampiamente inferiore a 5, valore sopra il quale è possibile riscontrare problemi di multicollinearità.

Il secondo grafico rappresenta i residui nei confronti dei valori stimati corrispondenti (valutati per la variabile trasformata $\ln(D_m)$ in Figura 6.3 e $D_m^{1/3}$ in Figura 6.4). Come spiegato nel Paragrafo 2.3.2 la presenza di particolari pattern nella disposizione dei punti può essere indice di eteroschedasticità (diversità nella varianza) dei residui. In questo caso i residui sembrano non dipendere dal valore stimato della variabile dipendente. È stata anche calcolata la probabilità "p-omosk" associata al test di omoschedasticità di *Harrison & McCabe* (1979) (vedi Paragrafo 2.3.2). Abbiamo deciso che si sarebbe dovuta rigettare l'ipotesi di omoschedasticità se p-omosk < 0.05. Dal momento che per i due modelli i valori di p-omosk sono 0.74 e 0.70 rispettivamente, i loro residui possono tranquillamente essere considerati omoschedastici.

Il terzo grafico è la rappresentazione dei residui in carta probabilistica normale. Il complemento ad 1 della probabilità associata al test di normalità di Anderson-Darling (vedi Paragrafo 2.3.2), indicata con "p-norm", è riportato anch'esso sul grafico. Abbiamo deciso che l'ipotesi di normalità dei residui debba essere rigettata se p-norm < 0.05. In questo caso p-norm vale 0.12 per la Regressione (6.7) e 0.10 per la Regressione (6.8).

Il quarto ed ultimo grafico riporta il risultato delle cross-validazioni: il deflusso stimato in ogni bacino è stato ottenuto escludendo il bacino stesso dalla taratura delle Regressioni (6.7) e (6.8); le linee tratteggiate rappresentano l'intervallo di previsione del 95% (vedi Paragrafo 2.3.1), ovvero la fascia entro la quale dovrebbe ricadere il 95% dei valori stimati con il modello se le ipotesi fatte fossero corrette. Le stazioni idrometriche utilizzate nella calibrazione dei modelli le cui stime sono meno buone sono state indicate in figura.

I diagrammi riportati nelle Figure 6.3 e 6.4 consentono quindi di ritenere le Regressioni (6.7) e (6.8) appropriate alla stima del deflusso-indice. Per evidenziare le differenze esistenti tra i 2 modelli, in Figura 6.5 si sono rappresentate le stime ottenute nei siti monitorati con la Regressione (6.7) in funzione di quelle ottenute con la Regressione (6.8). Se si escludono pochi casi, il più critico dei quali è quello dell'Orco a Pont Canavese, i due modelli forniscono essenzialmente

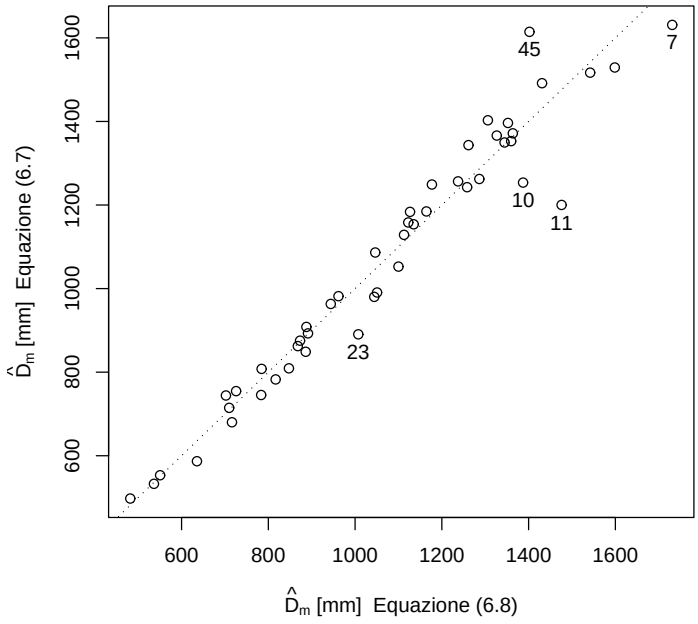

Figura 6.5: Confronto tra le Regressioni (6.7) e (6.8): il risultato delle stime nei 47 siti del SIMN con il Modello (6.7) è rappresentato in funzione dell'analogo risultato ottenuto dal Modello (6.8). Si sono indicate le stazioni in cui la differenza delle stime è superiore a 100 mm: Cervo a Passobreve (7), Dora Baltea a Tavagnasco (10), Orco a Pont Canavese (11), Vermenagna a Limone (23) e Evancon a Champoluc (45).

la stessa stima.

6.3 Regionalizzazione della curva di crescita

La fase più complessa dell'analisi di frequenza regionale è la stima della curva di crescita. Il metodo che si è utilizzato in questo caso è quello della formazione di regioni omogenee disgiunte (vedi Capitolo 3). Si è ricercata una volta per tutte la migliore suddivisione dei siti strumentati in gruppi omogenei, i cui elementi presentassero curve di crescita campionarie simili. Tali curve sono state ottenute dalle stazioni idrometriche la cui serie storica supera i 14 anni di osservazioni.

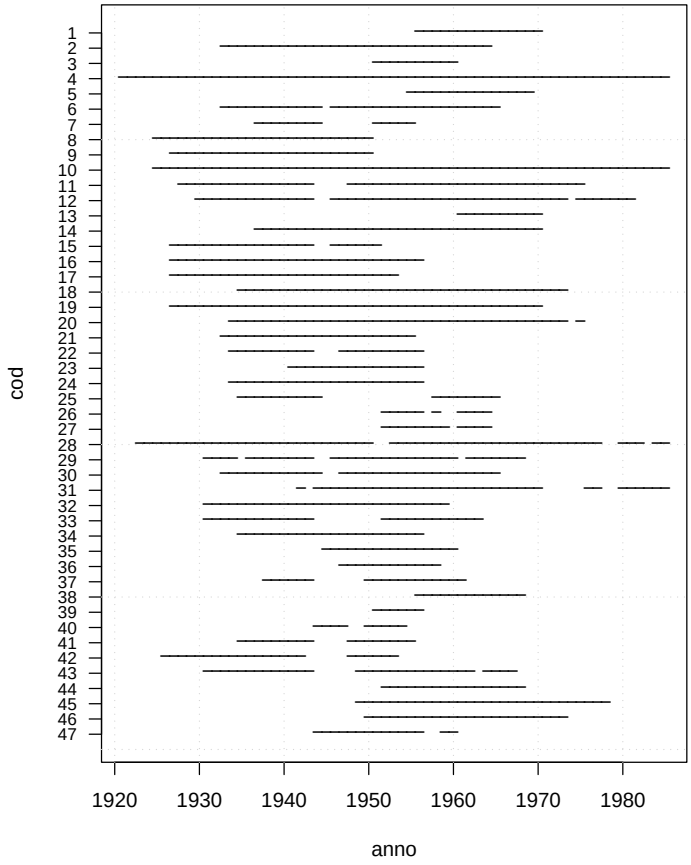

Figura 6.6: Consistenza delle serie storiche delle 47 stazioni idrometriche del SIMN.

La consistenza delle serie storiche delle 47 stazioni idrometriche del SIMN è rappresentata in Figura 6.6. Solo 38 stazioni hanno almeno 15 dati di deflusso annuo (in Tabella 6.III sono riportate alcune statistiche).

Per la scelta delle variabili di classificazione da utilizzarsi nella formazione delle regioni si è proceduto come indicato nel Capitolo 4. Per valutare la distanza tra le curve di crescita campionarie $\tilde{q}(F)$ si è utilizzata la differenza tra gli L-CV campionari delle serie, il che equivale ad utilizzare la statistica test di Hosking e Wallis (Paragrafo 5.1.1). In questo caso l'utilizzo delle matrici delle distanze

Tabella 6.III: Momenti ed L-momenti campionari delle serie storiche dei deflussi annui nelle 38 sezioni idrometriche considerate: numerosità del campione (num), media campionaria ($\hat{\mu}$), scarto quadratico medio ($\hat{\sigma}$), coefficiente di L-variazione (L-CV), L-skewness (L-CA) ed L-kurtosi (L-kur). Le differenze tra $\hat{\mu}$ e $\tilde{D}_m$ di Tabella 6.I è dovuta al fatto che $\tilde{D}_m$ è il deflusso medio annuo riportato nella Pubblicazione 17 (dati fino al 1970), mentre per calcolare $\hat{\mu}$, dove disponibili, si sono utilizzati anche dati di anni successivi (fino al 1986).

cod	num	$\hat{\mu}$ [mm]	$\hat{\sigma}$ [mm]	L-CV	L-CA	L-kur
1	15	1570	197	0.074	0.029	0.018
2	32	1380	361	0.149	0.180	0.061
4	65	1409	361	0.142	0.147	0.146
5	15	1728	475	0.153	0.128	0.292
6	32	1590	571	0.208	0.066	0.082
8	26	1271	259	0.118	0.092	0.049
9	24	1427	328	0.134	−0.091	−0.001
10	61	925	201	0.116	0.257	0.161
11	44	1043	273	0.149	0.124	0.101
12	49	1093	311	0.158	0.116	0.140
14	34	694	231	0.192	0.071	0.071
15	23	654	211	0.179	0.203	0.182
16	30	662	185	0.158	0.152	0.071
17	27	590	163	0.157	0.149	0.079
18	39	1273	339	0.154	0.094	0.049
19	44	506	186	0.207	0.130	0.133
20	41	830	278	0.191	0.135	0.108
21	23	925	261	0.155	0.251	0.183
22	20	1241	334	0.153	0.182	0.123
23	16	1128	350	0.171	0.118	0.298
24	23	1272	259	0.119	0.007	0.005
25	18	1012	324	0.179	0.188	0.172
28	58	516	188	0.201	0.223	0.130
29	34	1032	353	0.190	0.236	0.089
30	31	901	338	0.207	0.265	0.114
31	36	780	235	0.168	0.139	0.175
32	29	1066	343	0.180	0.125	0.211
33	25	828	272	0.183	0.191	0.154
34	22	965	389	0.214	0.325	0.217
35	16	882	296	0.198	−0.011	0.013
37	18	784	411	0.285	0.269	0.215
41	17	897	172	0.107	0.205	0.194
42	23	1351	221	0.094	0.053	0.125
43	31	1649	248	0.084	0.103	0.194
44	17	1023	188	0.107	0.045	0.138
45	30	991	192	0.104	0.170	0.278
46	24	1260	339	0.150	0.233	0.142
47	15	1078	269	0.138	0.246	0.206

non è indispensabile, essendo possibile anche procedere con un'analisi regressiva

tra gli L-CV campionari e le variabili morfoclimatiche (cosa che è stata fatta ottenendo gli stessi risultati). È stata però utilizzata la metodologia descritta nel Paragrafo 4.2 per dimostrarne la validità e per la sua maggiore generalità: nel caso in cui risultasse conveniente utilizzare un altro test di omogeneità (*Viglione et al.*, 2007a, Capitolo 5), ad esempio nell'analisi regionale delle piene, il ricorso all'analisi regressiva classica non sarebbe possibile.

Le grandezze morfoclimatiche utilizzate sono ancora quelle di Tabella 6.I. Si sono quindi ottenute 14 matrici delle distanze $\mathbf{\Delta}_{A_m}, \mathbf{\Delta}_S, \mathbf{\Delta}_{H_m}, \ldots$, da confontarsi, a mezzo di regressioni lineari, con la matrice delle distanze tra gli L-CV, $\mathbf{\Delta}_{L-\mathrm{CV}}$. Analogamente al Paragrafo 6.2 si sono escluse tutte quelle regressioni in cui anche solo una delle matrici non risultasse significativa in base al test di Mantel descritto nel Capitolo 4, e le rimanenti sono state ordinate in funzione dell'R^2_{adj} in senso decrescente. La migliore regressione lineare multipla tra $\mathbf{\Delta}_{L-\mathrm{CV}}$ e le matrici delle distanze delle grandezze morfoclimatiche è risultata essere

$$\mathbf{\Delta}_{L-\mathrm{CV}} = 0.46 + 3 \cdot 10^{-4} \cdot \mathbf{\Delta}_{H_m} + 0.18 \cdot \mathbf{\Delta}_{Y_{bar}} \ .$$

Il test di Mantel dei coefficienti delle variabili indipendenti, applicato con l'1% di significatività, permette di considerare significativa la correlazione tra le distanze delle curve di crescita campionarie e le distanze tra i bacini idrografici in quota media (H_m) e latitudine del baricentro (Y_{bar}). Quindi questi due descrittori dei bacini sono stati scelti come variabili di classificazione per suddividere in gruppi i bacini del SIMN.

In Figura 6.7 si sono rappresentati i "passi" della procedura mista di cluster analysis descritta nel Paragrafo 3.1. Considerare un'unica regione non è corretto, in quanto la misura di eterogeneità di Hosking e Wallis (Paragrafo 5.1.1) vale 8.57, valore molto superiore a 2, limite di accettabilità che abbiamo scelto. Si è quindi proceduto a suddividere i siti in 2, 3 ed infine 4 regioni con l'algoritmo di Ward e la conseguente riallocazione degli elementi (Paragrafo 3.1), arrestando la procedura al superamento del test di omogeneità di *Hosking & Wallis* (1993) in tutti i gruppi. Per due delle regioni individuate i valori della statistica θ_{HW_1} (indicata con H in figura) sono di poco inferiori a 2, per cui tali regioni dovrebbero ritenersi "possibilmente eterogenee", ma si è deciso di accettare questo raggruppamento per evitare di dover utilizzare gruppi troppo piccoli. L'utilizzo della misura di eterogeneità di Hosking e Wallis è da preferirsi rispetto al test bootstrap di Anderson-Darling in quanto la variabile deflusso annuo è, per sua natura, poco asimmetrica. A conferma di quanto fatto, le quattro

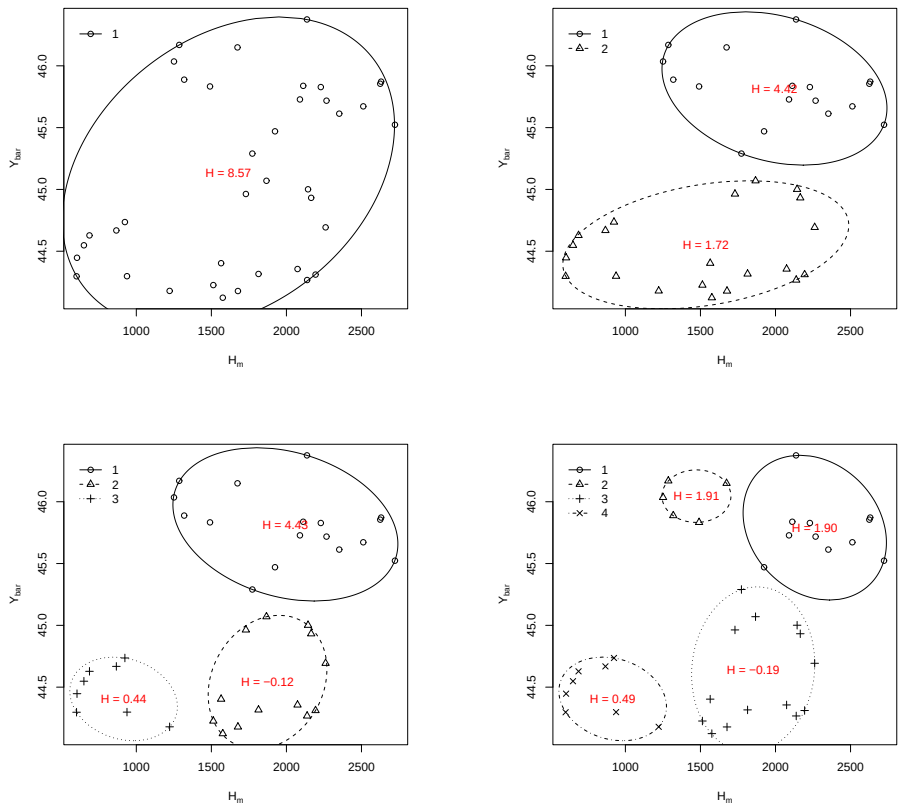

Figura 6.7: Suddivisione dei bacini in 4 regioni omogenee sul piano $H_m - Y_{bar}$. Si sono riportate le quattro fasi della cluster analysis: per ogni gruppo è stato indicato il valore della statistica di omogeneità di Hosking e Wallis (Equazione (5.6)). Si veda Figura 6.9 per la rappresentazione geografica della suddivisione dei bacini idrografici in 4 cluster.

regioni individuate possono essere rappresentate, come si vede in Figura 6.8, sul piano degli L-momenti L-CA ed L-CV, nella parte in cui il test di Hosking e Wallis è più potente (Capitolo 5). In Figura 6.9 le regioni omogenee sono state rappresentate geograficamente. Come si può notare le regioni risultano essere particolarmente compatte, anche in conseguenza del fatto che la latitudine del baricentro del bacino è stata impiegata per formare le regioni: la Regione 1 comprende i bacini valdostani, due bacini limitrofi (Orco a Pont Canavese e Sesia a Campertogno) ed il Toce a Cadarese; la Regione 2 raggruppa le zone del Toce

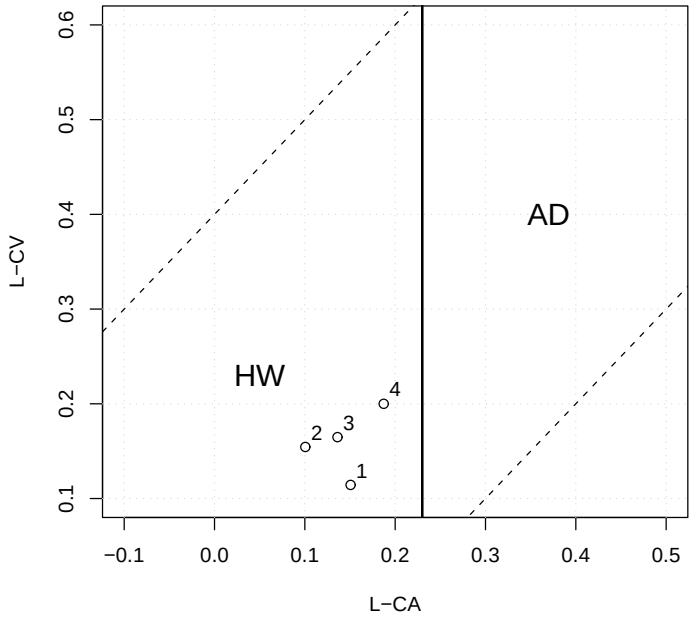

Figura 6.8: Rappresentazione delle quattro regioni sul piano degli L-momenti (L-CA, L-CV) per la decisione del test di omogeneità da utilizzare (vedi Paragrafo 5.4). La misura di eterogeneità di Hosking e Wallis, in questo caso, è da preferirsi rispetto al test bootstrap di Anderson-Darlyng.

e del Sesia (escludendo i due bacini che ricadono nella Regione 1); la Regione 3 comprende i bacini alpini di torinese e cuneese; infine i bacini della pianura e bassa montagna cuneese e dell'appennino alessandrino fanno parte della Regione 4.

La scelta della distribuzione di probabilità caratteristica di ogni gruppo è stata effettuata considerando distribuzioni a 3 parametri (v.es. *Basson et al.*, 1994) per la stima dei quali il numero elevato di osservazioni disponibili per ogni regione determina adeguate condizioni di robustezza. La tecnica di "model selection" descritta nel Paragrafo 3.2 è stata applicata sui quattro raggruppamenti definiti in precedenza, considerando le seguenti distribuzioni di probabilità, descritte in dettaglio in Appendice B:

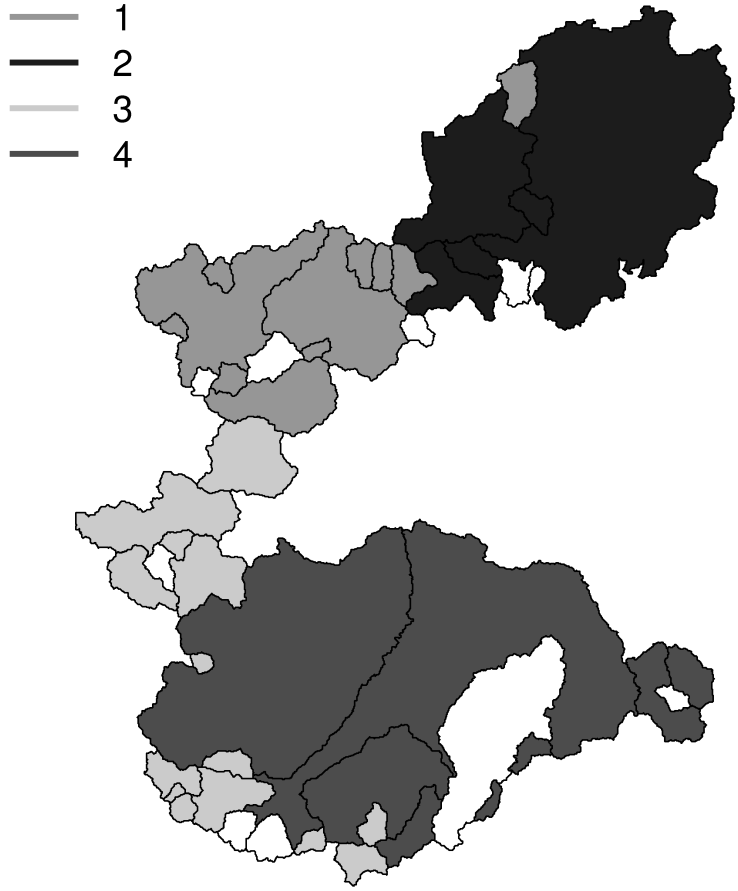

Figura 6.9: Rappresentazione geografica dei bacini appartenenti alle 4 regioni omogenee, che risultano essere particolarmente compatte (se si escludono pochi casi come, ad esempio, quello del Toce a Cadarese). La Regione 1 comprende i bacini Valdostani e limitrofi, la Regione 2 quelli dell'area Toce-Sesia, la Regione 3 i bacini alpini del torinese e del cuneese, mentre la Regione 4 i bacini di pianura e dell'appennino piemontese. I bacini non inseriti nei gruppi (rappresentati in bianco) sono quelli con una serie storica di deflussi annui di numerosità non superiore a 15.

- distribuzione di Pareto generalizzata (GP);
- distribuzione generalizzata dei valori estremi (GEV);
- distribuzione logistica generalizzata (GL);
- distribuzione log-normale a 3 parametri (LN3);

- distribuzione di Pearson tipo III (P3);

delle quali solo le ultime due sono comunemente proposte per la descrizione dei deflussi annui (v.es. *Vogel & Wilson*, 1996). I parametri delle distribuzioni sono stati stimati con il metodo degli L-momenti (*Hosking & Wallis*, 1997) utilizzando il campione ottenuto in ogni regione raggruppando le singole curve di crescita campionarie. In Tabella 6.IV sono riportati i valori della probabilità $\hat{G}(A^2)$, associata alla statistica di Anderson-Darling di Equazione (3.5): tanto più vicino a zero è il valore ottenuto, tanto migliore può ritenersi l'adattamento della distribuzione al campione. Dal momento che tutte le distribuzioni prese in

Tabella 6.IV: Stima della probabilità $G(A^2)$, associata alla statistica di Anderson-Darling di Equazione (3.5); valori superiori a 0.9 comportano il rigetto della distribuzione ipotetica.

	REGIONE			
	1	2	3	4
GP	1.00	1.00	1.00	1.00
GEV	0.172	0.390	0.917	0.849
GL	0.474	0.848	0.999	0.986
LN3	0.300	0.425	0.924	0.867
P3	0.589	0.388	0.863	0.816

considerazione sono a 3 parametri, il valore di $\hat{G}(A^2)$ può essere effettivamente utilizzato per confronto. Poiché si è scelto di utilizzare un solo tipo di distribuzione per tutto il territorio considerato, si è scelta la distribuzione di Pearson (o Gamma a 3 parametri) che, nel caso del terzo gruppo di bacini (Regione 3), è l'unica che non viene rigettata dal test con significatività 10%. La distribuzione di probabilità di Pearson tipo III è definita come:

$$f(q) = \frac{(x-\xi)^{\alpha-1} e^{-(q-\xi)/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \ , \ F(q) = G\left(\alpha, \frac{q-\xi}{\beta}\right)/\Gamma(\alpha) \ , \qquad (6.9)$$

dove $f(q)$ è la funzione densità di probabilità, $F(q)$ è la distribuzione cumulata, ξ è il parametro di posizione, β quello di scala, α quello di forma, Γ è la funzione gamma e G la funzione gamma incompleta (v.es. *Kottegoda & Rosso*, 1998).

Il risultato della model selection è confermato dal diagramma dei rapporti tra gli L-momenti di Figura 6.10 (Appendice A), in cui gli L-momenti regionali L-CA ed L-kur per le regioni formate vengono confrontati con i range di L-momenti che alcune distribuzioni possono assumere. Nel piano L-CA/L-kur le distribuzioni a due parametri sono rappresentate da un punto, mentre quelle a tre parametri

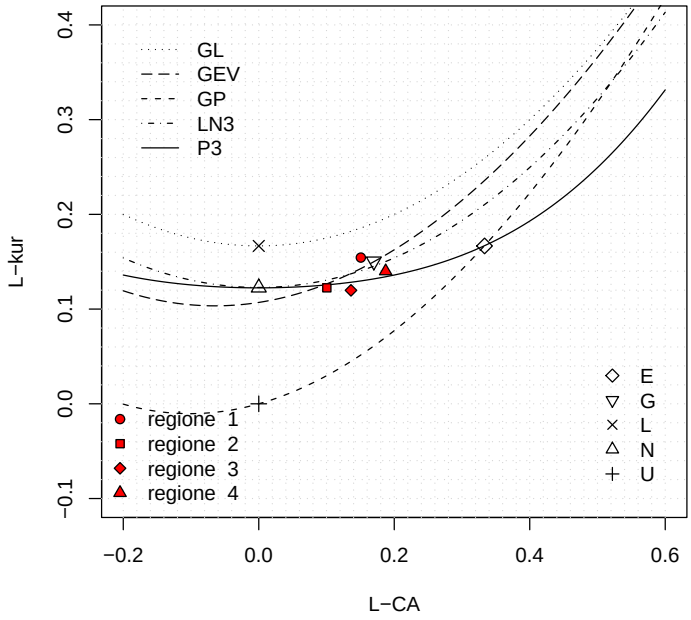

Figura 6.10: Rappresentazione delle quattro regioni e delle distribuzioni di probabilità a due ed a tre parametri sul diagramma degli L-momenti. Le distribuzioni a due parametri sono: esponenziale (E), Gumbel (G), lognormale (L), normale (N) ed uniforme (U); quelle a tre parametri sono: logistica generalizzata (GL), generalizzata del valore estremo (GEV), Pareto generalizzata (GP), lognormale a 3 parametri (LN3) e Pearson tipo III (P3).

da una curva (data la loro maggiore flessibilità). Tre delle regioni identificate si dispongono bene intorno alla linea relativa alla distribuzione Pearson tipo III, mentre la Regione 3 è un po' più distante, cosa che conferma il risultato del test di adattamento (vedi Tabella 6.IV).

La curva di crescita viene quindi definita come:

$$q(F) = \xi + \beta \cdot \Theta(F, \alpha) \ , \tag{6.10}$$

dove $\Theta(F, \alpha)$ è l'inversa della funzione gamma generalizzata. Per le quattro

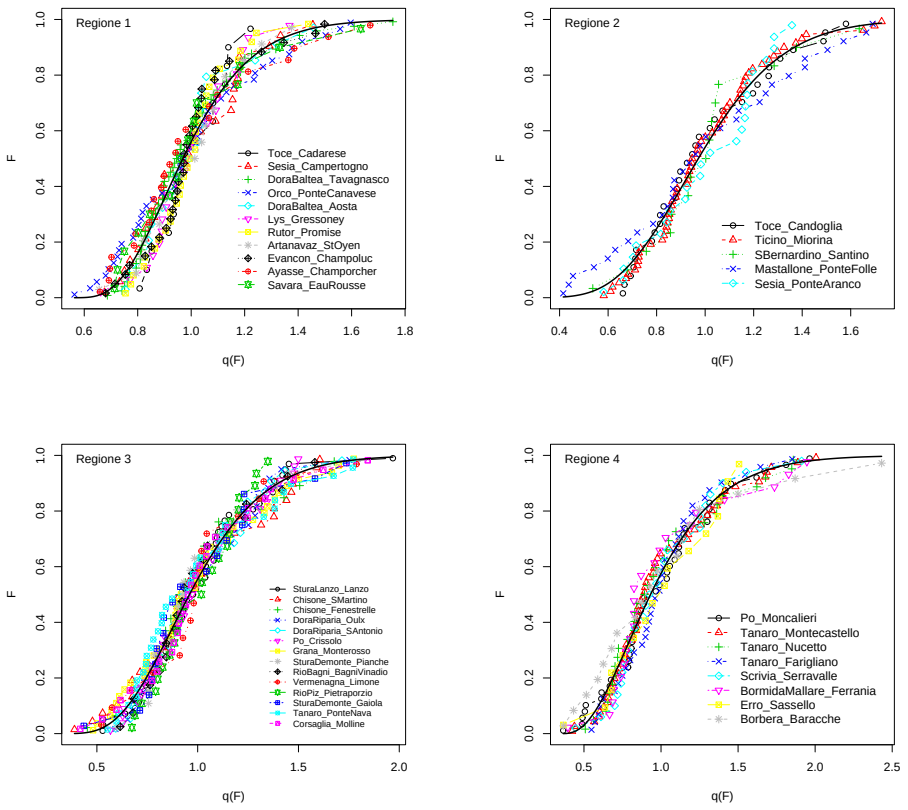

Figura 6.11: Curve di crescita campionarie (punti uniti da spezzate) e curva di crescita regionale (linea nera continua) per le regioni omogenee individuate (da *Viglione et al.*, 2006).

regioni individuate si sono ottenuti i seguenti valori dei parametri:

$$\begin{array}{lll} \alpha_1 = 4.764 & \beta_1 = 9.534 \cdot 10^{-2} & \xi_1 = 0.5458 \\ \alpha_2 = 10.60 & \beta_2 = 8.508 \cdot 10^{-2} & \xi_2 = 9.843 \cdot 10^{-2} \\ \alpha_3 = 5.817 & \beta_3 = 0.1237 & \xi_3 = 0.2801 \\ \alpha_4 = 3.107 & \beta_4 = 0.2093 & \xi_4 = 0.3496 \end{array} \tag{6.11}$$

In Figura 6.11 sono rappresentate le curve di crescita regionali sovrapposte a quelle campionarie relative alle singole stazioni. Osservando i grafici relativi alle Regioni 1 e 2 si comprende perché la misura di eterogeneità di Hosking e Wallis si avvicina al valore di non accettazione.

In Figura 6.12 le quattro curve regionali sono state rappresentate tutte in-

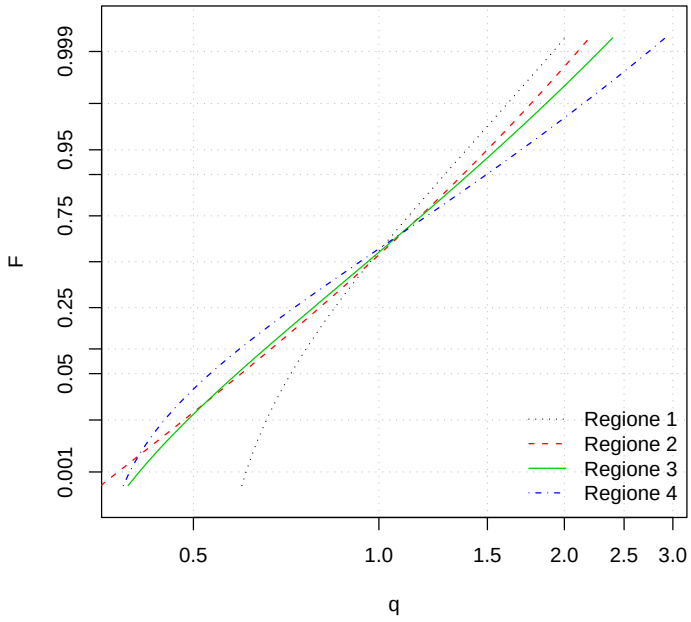

Figura 6.12: Rappresentazione delle curve di crescita regionali in carta probablistica lognormale (da *Viglione et al.*, 2006). Si evidenziano in questo modo le differenze delle distribuzioni, soprattutto per quanto riguarda la coda inferiore, di maggiore interesse nelle analisi di disponibilità idrica (e quindi anche del deflusso annuo).

sieme in carta probabilistica lognormale in modo da farne risaltare le differenze. Si noti come le curve di crescita siano significativamente asimmetriche (non-normali), presentando i seguenti coefficienti di asimmetria:

$$\gamma_1 = 0.974 \ , \quad \gamma_2 = 0.624 \ , \quad \gamma_3 = 0.803 \ , \quad \gamma_4 = 1.106 \ .$$

La densità di probabilità associata alle quattro curve di crescita regionali è riportata in Figura 6.13. Si noti come il coefficiente di variazione dalla Regione 1 sia sensibilmente inferiore a quello delle altre regioni. La ragione di questo comportamento può essere individuata nel fatto che la Regione 1 comprende bacini nivoglaciali (comprende infatti tutti i bacini valdostani), che presentano un andamento più regolare dei deflussi grazie all'apporto di acqua di sciogli-

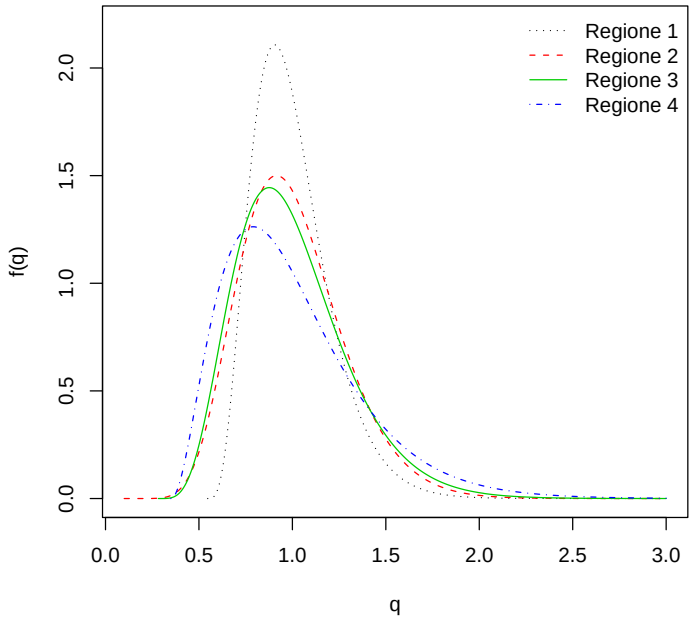

Figura 6.13: Densità di probabilità $f(q)$ associata alle quatto curve di crescita regionali.

mento dei ghiacciai. La Regione 4, caratterizzata da variabilità ed asimmetria più alta, deve queste caratteristiche al fatto di comprendere praticamente tutti i bacini del settore appenninico piemontese, soggetti ad una condizione climatica diametralmente opposta a quella valdostana.

6.4 Utilizzo del modello regionale

Volendo esemplificare la procedura, se si vuole stimare il deflusso annuo $D(T)$ che non viene raggiunto, in media, una volta ogni T anni ($T = 1/F$) si procede come segue:

- se nel sito di interesse è disponibile una serie storica con numerosità suf-

ficiente (almeno $n > 2T$), $D(T)$ può essere stimato con un'analisi di frequenza locale senza ricorrere alla tecnica di regionalizzazione;

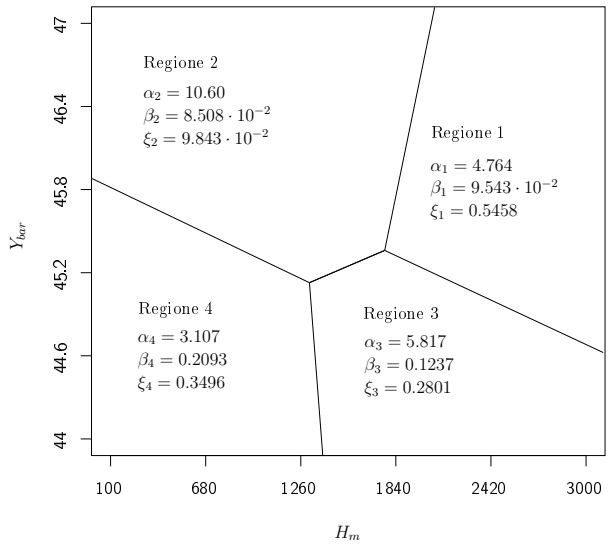

Figura 6.14: Suddivisione del piano $H_m - Y_{bar}$ nelle 4 regioni omogenee (da *Viglione et al.*, 2006); per ognuna di esse sono stati indicati i parametri della curva di crescita regionale (distribuzione Tipo III di Pearson, vedi Appendice B).

- se sono disponibili solo pochi dati, ad esempio 5-10, D_m viene stimato come media aritmetica dei deflussi annui misurati, mentre per $q(T)$ si determina quale tra le quattro curve di crescita individuate nel Paragrafo 6.3 sia da utilizzare. La scelta è fatta in base all'assegnazione del bacino idrografico sotteso dalla sezione considerata ad una delle quattro regioni omogenee, entrando nel diagramma rappresentato in Figura 6.14 con i valori delle variabili quota media H_m e latitudine del baricentro del bacino Y_{bar}. Le linee di separazione nel diagramma di Figura 6.14 sono state ricavate con il criterio della minima distanza tra il baricentro delle regioni omogenee (rappresentate in Figura 6.7) e i punti dello spazio delle variabili morfoclimatiche;

- per una sezione sprovvista di misure, D_m viene stimato tramite uno dei

modelli regressivi proposti (Equazioni (6.7) o (6.8)) mentre per $q(T)$ si procede come nel caso precedente. Anche in questo caso i parametri necessari all'applicazione del modello regionale sono pochi e di facile determinazione: se si utilizza il modello regressivo più semplice (Equazione (6.8)) i parametri richiesti sono solamente l'afflusso medio annuo A_m, la quota media H_m e la latitudine del baricentro del bacino Y_{bar}.

Una volta stimati il deflusso indice e la curva di crescita, il deflusso annuo corrispondente al tempo di ritorno T sarà ottenuto come $\hat{D}(T) = \hat{D}_m \cdot \hat{q}(T)$.

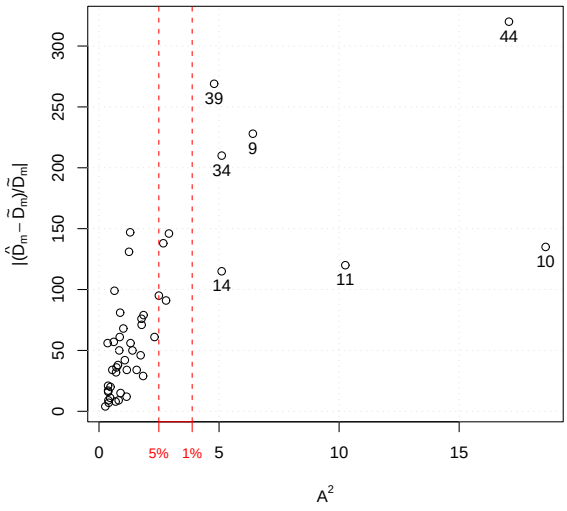

Figura 6.15: Legame tra la statistica di Anderson-Darling A^2 di bontà di adattamento delle distribuzioni $\hat{D}(F)$ alle distribuzioni empiriche $\tilde{D}(F)$ e l'errore relativo di stima della media $|(\hat{D}_m - \tilde{D}_m)/\tilde{D}_m|$. Si sono indicati in figura i limiti percentuali corrispondenti alla significatività 5% ($A^2 = 2.492$) e 1% ($A^2 = 3.880$) del test di Anderson-Darling per distribuzioni completamente specificate. Il test all'1% di significatività rigetta l'ipotesi di adattamento di $\hat{D}(F)$ a $\tilde{D}(F)$ nelle sezioni: Sesia a Ponte Aranco (9), Dora Baltea a Tavagnasco (10), Orco a Pont Canavese (11), Chisone a S.Martino (14), Bormida di Mallare a Ferrania (34), Dora di Rhemes a Pelaud (39) e Artavanaz a St.Oyen (44).

Per avvalorare la procedura, i modelli sono stati applicati alle sezioni idrografiche corrispondenti alle 47 stazioni idrometriche del SIMN. Le distribuzioni di probabilità $\hat{D}(T)$ ottenute (la media è stata stimata con l'Equazione (6.7)) sono state messe a confronto con le funzioni di frequenza empiriche $\tilde{D}(T)$. La bontà dell'adattamento è stata valutata con il test di Andeson-Darling per distribuzioni completamente specificate, dal momento che i parametri della gamma vengono considerati noti a priori, Parametri (6.11), e non vengono stimati dai campioni. In questo caso non occorre derivare la distribuzione della statistica A^2, ma si possono usare i punti percentuali relativi a diversi livelli di significatività di letteratura (v.es. *D'Agostino* & *Stephens*, 1986, Tabella 4.2, pag. 105). Applicando il test al 5% di significatività, le distribuzioni di probabilità stimate risultano adatte (o, più precisamente, non si può rigettare l'ipotesi di buon adattamento) per 36 campioni su 47.

La performance del modello regionale, valutata con il test di adattamento, è particolarmente sensibile agli errori di stima della media. In Figura 6.15 si riportano i valori della statistica A^2 per le 47 stazioni rappresentati con gli errori relativi di stima della media $|(\hat{D}_m - \tilde{D}_m)/\tilde{D}_m|$. Come si può notare, esiste una forte correlazione tra errori di stima della media e risultato del test. Le sezioni che non superano il test al 1% di significatività sono caratterizzate tutte da una stima della media errata di più del 10% del valore osservato. Per questo motivo, come si è detto, se sono disponibili anche solo pochi dati, D_m deve essere stimato come media aritmetica dei deflussi annui misurati, piuttosto che con un modello regionale.

Capitolo 7

Conclusioni

La stima delle variabili idrologiche in siti non monitorati è uno degli aspetti più affascinanti, e nello stesso tempo più utili nella pratica progettuale, dell'idrologia. All'approccio modellistico, sia deterministico che stocastico, adottato per la ricostruzione temporale di eventi idrologici o meteorologici o per la previsione a breve termine, si affiancano metodologie statistiche data-driven di "spazializzazione" della distribuzione di probabilità delle grandezze idrologiche, conosciute come tecniche di analisi di frequenza regionale. Il modello di analisi di frequenza regionale maggiormente utilizzato è quello della grandezza-indice, basato sull'ipotesi che le distribuzioni di probabilità della grandezza in differenti siti appartenenti ad una regione statisticamente omogenea siano identiche, a meno di un parametro di scala.

Le tecniche proposte in letteratura, soprattutto per la determinazione delle regioni omogenee, sono piuttosto complesse e per lo più basate su procedure che comportano scelte soggettive. Questo aspetto fa sì che tali tecniche non portino a risultati univoci, manchino insomma della generalità che si vorrebbe che avessero. Questo lavoro costituisce uno sforzo nella direzione di ridurre al minimo la necessità di ricorrere a scelte soggettive, proponendo criteri semplici ed oggettivi (non-supervised, appunto) a corredo delle tecniche di regionalizzazione in generale, e del metodo della grandezza-indice in particolare.

Nonostante i metodi di regressione multipla per la determinazione del valore indice siano tecniche consolidate, è parso opportuno dedicare la prima parte del lavoro ad un approfondimento su di essi. L'individuazione di un criterio per la scelta della grandezza-indice e la descrizione di una procedura completa e robusta per la sua stima attraverso modelli lineari sono state affrontate nel ten-

tativo di dare ordine alla metodologia, pur non introducendo novità significative. L'approccio proposto ed utilizzato estende l'applicazione della tecnica stepwise regression in modo da escludere il condizionamento dei risultati dovuto all'ordine di inserimento dei parametri. Se da un lato ciò comporta l'esecuzione di un numero elevatissimo di combinazioni, dall'altro rende la procedura completamente automatica ed oggettiva.

I contributi più innovativi del lavoro riguardano la regionalizzazione della curva di crescita, ovvero della distribuzione di frequenza adimensionalizzata con la grandezza-indice. In tal senso si è proposto un criterio di raggruppamento dei siti in regioni omogenee che sfrutta come variabili discriminanti alcune grandezze caratteristiche dei siti stessi. Tale criterio si basa sul legame statistico delle caratteristiche dei siti monitorati con le curve di crescita campionarie in termini di distanza. Si sono introdotti due operatori ad oggi poco utilizzati in idrologia: le matrici delle distanze ed il test di Mantel. Le matrici delle distanze sono matrici quadrate e simmetriche i cui elementi sono ognuno la misura di distanza tra due entità, nel nostro caso distanze delle curve di crescita o delle caratteristiche dei siti. Il test di Mantel permette di valutare la significatività della correlazione tra matrici delle distanze, e quindi di decidere se la similitudine/diversità dei siti, in termini di una o più grandezze caratteristiche, spieghi la similitudine/diversità fra le curve di crescita in essi misurate.

Parallelamente a questo, si è selezionato un criterio oggettivo e di facile applicazione per la scelta, caso per caso, del test di omogeneità che è meglio utilizzare. Da uno studio condotto su potenza e distorsione di vari test con l'ausilio di metodi di simulazione tipo Monte Carlo, è risultato che la misura di eterogeneità di Hosking e Wallis, comunemente utilizzata nell'analisi di frequenza regionale, è preferibile nel caso in cui le curve di crescita siano poco asimmetriche, mentre il test bootstrap di Anderson-Darling, ottenuto dalla modifica del classico test di adattamento di Anderson-Darling, è da preferire in situazioni di elevata asimmetria.

L'analisi regionale dei deflussi annui condotta sui bacini delle regioni Piemonte e Valle d'Aosta fornisce un esempio di applicazione della metodologia della grandezza-indice e dà particolare risalto all'esecuzione operativa delle tecniche proposte. L'eterogeneità che contraddistingue questo territorio, in cui coesistono conformazioni orografiche e situazioni climatiche estremamente differenti, rende particolarmente complicata, e nello stesso tempo interessante, l'applicazione dei metodi regionali. Per il deflusso indice si sono individuati due modelli regressivi

di stima: il "migliore possibile" in base ai dati di calibrazione ed il "più semplice tra i migliori" che richiede solo la conoscenza di quota media e afflusso medio annuo sul bacino. Per la scelta della distribuzione di probabilità del deflusso annuo si propone un grafico utilizzando il quale un bacino sprovvisto di misure può essere assegnato ad una regione, in base alla sua quota media e latitudine del baricentro. Nello spazio di queste due variabili morfometriche si sono individuate quattro regioni per le quali si propongono quattro diverse distribuzioni di probabilità per rappresentare le rispettive curve di crescita.

In conclusione, i risultati ottenuti in questo lavoro contribuiscono a completare e migliorare le tecniche di regionalizzazione attualmente utilizzate. Essi hanno portato alla definizione di alcuni strumenti utili come supporto decisionale per chi utilizza tali tecniche, e, nello stesso tempo, utili per chi le vuole affinare ulteriormente. Tra gli sviluppi futuri della ricerca in questo ambito, sicuramente ci sarà un confronto tra le diverse tecniche di analisi di frequenza regionale: tra il metodo della grandezza-indice a regioni disgiunte, quello delle regioni di influenza, quello delle regioni gerarchiche (v.es. *Fiorentino et al.*, 1987), quelli che ipotizzano una variabilità continua anche dei parametri di forma delle distribuzioni di probabilità (v.es. *Furcolo et al.*, 1998), ecc. I metodi proposti in questa tesi possono essere annoverati tra gli strumenti necessari per la realizzazione di tale confronto, che, per esigenza di ripetibilità, deve essere condotto su basi oggettive.

Appendice A

L-momenti

Nelle procedure di analisi di frequenza, e di analisi di frequenza regionale, si adattano ai dati delle distribuzioni la cui forma si ritiene conosciuta a meno di un numero finito di parametri incogniti. I momenti campionari ordinari, in particolare media, scarto, skewness e kurtosis, sono spesso utilizzati per la stima dei parametri delle distribuzioni di probabilità. *Hosking & Wallis* (1997) suggeriscono invece di utilizzare, al posto dei momenti ordinari, gli L-momenti perché adatti a descrivere più distribuzioni, perché più robusti nella stima da campioni poco consistenti di dati in presenza di outliers e perché meno soggetti a distorsione nella stima. In questa appendice, tratta da *Hosking & Wallis* (1997), si definiscono gli L-momenti in maniera formale. Dopo una breve introduzione sui concetti di distribuzione di probabilità, di stimatori dei parametri e di momenti, vengono definiti gli L-momenti, si discutono alcune loro proprietà, le differenze rispetto ai momenti ordinari ed il loro utilizzo nella stima dei parametri delle distribuzioni (argomento approfondito in Appendice B).

A.1 Distribuzioni di probabilità

Si consideri una variabile casuale X, che può assumere valori appartenenti all'insieme dei numeri reali. La frequenza relativa con cui questi valori si verificano definisce la *distribuzione di frequenza* o *distribuzione di probabilità* di X, che è specificata dalla *distribuzione di frequenza cumulata*

$$F(x) = \Pr[X \leq x] \ , \tag{A.1}$$

dove $\Pr(A)$ indica la probabilità dell'evento A. $F(x)$ è una funzione crescente di x, definita nell'intervallo $[0,1]$. Normalmente in idrologia si ha a che fare con variabili casuali continue, per le quali $\Pr[X = t] = 0$ per ogni t, ovvero a nessun valore è associata una probabilità non-nulla. In questo caso $F(.)$ è una funzione continua ed ha una funzione inversa corrispondente $x(.)$, detta *funzione dei quantili* di X. Data una qualsiasi u, dove $0 < u < 1$, $x(u)$ è l'unico valore che soddisfa

$$F(x(u)) = u \ . \tag{A.2}$$

Per ogni probabilità p, $x(p)$ è il *quantile* di non superamento della probabilità p, ovvero il valore per cui la probabilità che X non superi $x(p)$ è p. L'obiettivo dell'analisi di frequenza è la stima accurata dei quantili della distribuzione di una data variabile casuale. In ingegneria, e nelle applicazioni ambientali in generale, i quantili sono spesso espressi in termini di *tempo di ritorno*, come definito dalle Equazioni (1.2) e (1.3).

Se $F(x)$ è differenziabile, la sua derivata $f(x) = \frac{d}{dx}F(x)$ è la *densità di probabilità* di X.

Il *valore atteso* della variabile casuale X è definito come

$$\mathrm{E}(X) = \int_{-\infty}^{\infty} x dF(x) = \int_{-\infty}^{\infty} x f(x) dx \ , \tag{A.3}$$

ammesso che l'integrale esista. Se si considera la trasformazione $u = F(x)$, si può scrivere

$$\mathrm{E}(X) = \int_0^1 x(u) du \ . \tag{A.4}$$

Una funzione di una variabile casuale $g(X)$ è anch'essa una variabile casuale di valore atteso

$$\mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) dF(x) = \int_{-\infty}^{\infty} g(x) f(x) dx = \int_0^1 g(x(u)) du \ . \tag{A.5}$$

La dispersione dei valori estratti dalla variabile casuale X può essere misurata con la *varianza* di X,

$$\mathrm{var}(X) = \mathrm{E}[\{X - \mathrm{E}(X)\}^2] \ . \tag{A.6}$$

In alcuni casi può essere utile misurare la tendenza di due variabili casuali X e Y ad assumere valori elevati simultaneamente. Questo può essere misurato dalla *covarianza* di X e Y

$$\mathrm{cov}(X,Y) = \mathrm{E}[\{X - \mathrm{E}(X)\}\{Y - \mathrm{E}(Y)\}] \ . \tag{A.7}$$

La *correlazione* tra X e Y

$$\text{corr}(X,Y) = \text{cov}(X,Y)/\{\text{var}(X)\text{var}(Y)\}^{1/2} \ , \tag{A.8}$$

è il corrispettivo adimensionale della covarianza, che può assumere valori compresi tra -1 e $+1$.

A.2 Stimatori

Nella pratica spesso si assume che la forma di una qualche distribuzione di probabilità sia conosciuta a medo di un set di parametri incogniti $\theta_1, \ldots, \theta_p$. Sia $x(u; \theta_1, \ldots, \theta_p)$ la funzione dei quantili di una distribuzione con p parametri incogniti. In molte applicazioni tra i parametri incognti si possono identificare un parametro di posizione ed un parametro di scala. Un parametro ξ di una distribuzione è un *parametro di posizione* se per la funzione dei quantili vale l'eguaglianza

$$x(u; \xi, \theta_2, \ldots, \theta_p) = \xi + x(u; 0, \theta_2, \ldots, \theta_p) \ . \tag{A.9}$$

Si dice, invece, che α è un *parametro di scala* della funzione dei quantili della distribuzione se

$$x(u; \alpha, \theta_2, \ldots, \theta_p) = \alpha \times x(u; 1, \theta_2, \ldots, \theta_p) \ . \tag{A.10}$$

Se per la distribuzione esistono entrambi questi parametri, allora vale l'eguaglianza

$$x(u; \xi, \alpha, \theta_3, \ldots, \theta_p) = \xi + \alpha \times x(u; 0, 1, \theta_3, \ldots, \theta_p) \ . \tag{A.11}$$

I parametri incogniti sono stimati a partire dai dati osservati. Dato un set di osservazioni, una funzione $\hat{\theta}$ di queste deve essere scelta come *stimatore* di θ. Lo stimatore $\hat{\theta}$ è a sua volta una variabile casuale ed ha una distribuzione di probabilità. La bontà di $\hat{\theta}$ come stimatore di θ tipicamente dipende da quanto $\hat{\theta}$ si avvicina a θ. La deviazione di $\hat{\theta}$ da θ può essere scomposta in *distorsione* (tendenza di dare stime sistematicamente più alte, o più basse, del valore vero) e *variabilità* (deviazione casuale dal valore vero, che si verifica anche per gli stimatori che non presentano distorsione).

La performance di uno stimatore $\hat{\theta}$ può essere valutata con due misure, il *bias* ("distorsione") e la *radice dell'errore quadratico medio* (RMSE), definite come

$$\text{bias}(\hat{\theta}) = \text{E}(\hat{\theta} - \theta) \ , \quad \text{RMSE}(\hat{\theta}) = \{\text{E}(\hat{\theta} - \theta)^2\}^{1/2} \ , \tag{A.12}$$

e caratterizzate dall'avere la stessa unità di misura del parametro θ. Si dice che lo stimatore $\hat{\theta}$ è *indistorto* se $\text{bias}(\hat{\theta}) = 0$ ovvero se $\text{E}(\hat{\theta}) = \theta$. Diversi stimatori indistorti dello stesso parametro possono essere paragonati in termini della loro varianza: il rapporto $\text{var}(\hat{\theta}^{(1)})/\text{var}(\hat{\theta}^{(2)})$ si dice *efficienza* dello stimatore $\hat{\theta}^{(2)}$ rispetto allo stimatore $\hat{\theta}^{(1)}$. La radice dell'errore quadratico medio può essere anche scritta come

$$\text{RMSE}(\hat{\theta}) = [\{\text{bias}(\hat{\theta})\}^2 + \text{var}(\hat{\theta})]^{1/2} \ , \tag{A.13}$$

da cui si vede come l'RMSE combina distorsione e variabilità di $\hat{\theta}$ e dà una misura globale dell'accuratezza della stima. Nei classici problemi di statistica in cui la stima dei parametri è basata su un campione di lunghezza n, sia il bias che la varianza di $\hat{\theta}$ sono asintoticamente proporzionali a n^{-1} per n grandi (v.es. *Cox & Hinkley*, 1974), per cui l'RMSE di $\hat{\theta}$ è proporzionale a $n^{-1/2}$.

A.3 Momenti

La forma di una distribuzione di probabilità può essere descritta dai *momenti della distribuzione*, che sono la *media* $\mu = \text{E}(X)$ e i *momenti di ordine superiore* $\mu_r = \text{E}(X - \mu)^r$, con $r = 2, 3, \ldots$. La media definisce il baricentro della distribuzione (si veda l'Equazione (A.3)). La dispersione della distribuzione intorno alla media può essere misurata con la *deviazione standard*

$$\sigma = \mu_2^{1/2} = \{\text{E}(X - \mu)^2\}^{1/2} \ , \tag{A.14}$$

o con la varianza, $\sigma^2 = \text{var}(X)$. Il *coefficiente di variazione* (CV), $C_v = \sigma/\mu$, esprime la dispersione della distribuzione adimensionalizzata con la media. Spesso si utilizzano momenti adimensionalizzati di ordine superiore $\mu_r/\mu_2^{r/2}$, in particolare lo *skewness* (asimmetria)

$$\gamma = \mu_3/\mu_2^{3/2} \ , \tag{A.15}$$

ed il *kurtosis*

$$\kappa = \mu_4/\mu_2^2 \ . \tag{A.16}$$

Quantità analoghe a queste possono essere calcolate da un campione di dati $x_1, x_2, \ldots, x_n$. La *media campionaria*

$$\bar{x} = n^{-1} \sum_{i=1}^{n} x_i \ , \tag{A.17}$$

è lo stimatore naturale di μ. I momenti di ordine superiore

$$m_r = n^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^r \ , \tag{A.18}$$

sono ragionevoli stimatori di μ_r, ma non sono indistorti. Al loro posto vengono spesso usati stimatori indistorti come, nel caso particolare di σ^2, μ_3 e $\kappa_4 = \mu_4 - 3\mu_2^2$, sono, rispettivamente,

$$s^2 = (n-1)^{-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \ , \tag{A.19}$$

$$\tilde{m}_3 = \frac{n^2}{(n-1)(n-2)} m_3 \ , \tag{A.20}$$

$$\tilde{k}_4 = \frac{n^2}{(n-2)(n-3)} \left\{ \left(\frac{n+1}{n-1} \right) m_4 - 3m_2^2 \right\} \ . \tag{A.21}$$

La deviazione standard campionaria, $s = \sqrt{s^2}$, è uno stimatore di σ ma non è indistorto. Gli stimatori campionari del CV, dello skewness e del kurtosis sono, rispettivamente,

$$\hat{C}_v = s/\bar{x} \ , \quad g = \tilde{m}_3/s^3 \quad k = \tilde{k}_4/s^4 + 3 \ . \tag{A.22}$$

Gli stimatori dei momenti hanno alcune proprietà indesiderabili. Ad esempio, gli stimatori g e k possono essere molto distorti in quanto sono caratterizzati da limiti algebrici che dipendono dalla lunghezza del campione; per un campione di n dati si ha che $|g| \leq n^{1/2}$ mentre $k \leq n + 3$ (*Hosking & Wallis*, 1997). Da ciò consegue che, se la distribuzione è particolarmente asimmetrica, potrebbe essere impossibile calcolare questa skewness da un campione di dimensioni fisse. L'inferenza di distribuzioni asimmetriche basata sui momenti campionari può essere estremamente poco affidabile. Per questo motivo in questo lavoro vengono utilizzate altre misure della forma delle distribuzioni, gli L-momenti di Hosking e Wallis.

A.4 L-momenti delle distribuzioni di probabilità

Gli L-momenti sono un sistema alternativo di descrivere la forma delle distribuzioni di probabilità. Storicamente essi nascono come modifica dei *momenti*

pesati in probabilità di *Greenwood et al.* (1979). I momenti pesati in probabilità di una variabile casuale X con distribuzione di frequenza cumulata $F(.)$ sono le quantità

$$M_{p,r,s} = \mathrm{E}[X^p\{F(X)\}^r\{1-F(X)\}^s] \ . \tag{A.23}$$

Particolarmente utili sono i momenti pesati in probabilità $\alpha_r = M_{1,0,r}$ e $\beta_r = M_{1,r,0}$. Per una distribuzione caratterizzata da una funzione dei quantili $x(u)$, dalle Equazioni (A.5) e (A.23) si ottiene

$$\alpha_r = \int_0^1 x(u)(1-u)^r du \ , \quad \beta_r = \int_0^1 x(u)u^r du \ . \tag{A.24}$$

Queste equazioni possono essere paragonate alla definizione dei momenti ordinari, che può essere scritta anche come

$$\mathrm{E}(X^r) = \int_0^1 \{x(u)\}^r du \ . \tag{A.25}$$

Mentre i momenti ordinari considerano successive elevazioni di potenza della funzione dei quantili $x(u)$, i momenti pesati in probabilità considerano successive elevazioni di potenza di u oppure $1-u$ e possono essere visti come integrali di $x(u)$ pesati con i polinomi u^r oppure $(1-u)^r$.

I momenti pesati in probabilità α_r e β_r sono stati usati in letteratura come base di metodi per la stima dei parametri delle distribuzioni di probabilità ma sono difficilmente interpretabili come misure di scala e forma di queste. Queste informazioni sono contenute in certe combinazioni lineari dei momenti pesati in probabilità. Ad esempio, multipli di $\alpha_0 - 2\alpha_1$ o $2\beta_1 - \beta_0$ sono stime dei parametri di scala delle distribuzioni, mentre lo skewness può essere misurato da $6\beta_2 - 6\beta_1 + \beta_0$. Queste combinazioni lineari derivano naturalmente dall'integrazione di $x(u)$ pesata non con i polinomi u^r o $(1-u)^r$, ma con un set di polinomi ortogonali.

Si definiscano *polinomi di Legendre sfasati* (perché definiti nell'intervallo $[0,1]$ invece che nell'intervallo $[-1,+1]$) i polinomi $P_r^*(u)$, con $r = 0,1,2,\ldots$, che godono delle seguenti proprietà:

(i) $P_r^*(u)$ è un polinomio di grado r in u;

(ii) $P_r^*(1) = 1$;

(iii) $\int_0^1 P_r^*(u)P_s^*(u)du = 0$ se $r \neq s$ (condizione di ortogonalità).

I polinomi di Legendre sfasati hanno forma esplicita

$$P_r^*(u) = \sum_{k=0}^{r} p_{r,k}^* u^k \ , \tag{A.26}$$

dove

$$p^*_{r,k} = (-1)^{r-k}\binom{r}{k}\binom{r+k}{k} = \frac{(-1)^{r-k}(r+k)!}{(k!)^2(r-k)!} \ . \tag{A.27}$$

Gli L-momenti di una variabile casuale X con funzione dei quantili $x(u)$ sono definiti come

$$\lambda_r = \int_0^1 x(u)P^*_{r-1}(u)du \ . \tag{A.28}$$

In termini di momenti pesati in probabilità, gli L-momenti sono dati da

$$\begin{array}{ll} \lambda_1 = \alpha_0 & = \beta_0 \ , \\ \lambda_2 = \alpha_0 - 2\alpha_1 & = 2\beta_1 - \beta_0 \ , \\ \lambda_3 = \alpha_0 - 6\alpha_1 + 6\alpha_2 & = 6\beta_2 - 6\beta_1 + \beta_0 \ , \\ \lambda_4 = \alpha_0 - 12\alpha_1 + 30\alpha_2 - 20\alpha_3 & = 20\beta_3 - 30\beta_2 + 12\beta_1 - \beta_0 \ , \end{array} \tag{A.29}$$

e, in generale,

$$\lambda_{r+1} = (-1)^r \sum_{k=0}^{r} p^*_{r,k}\alpha_k = \sum_{k=0}^{r} p^*_{r,k}\beta_k \ . \tag{A.30}$$

È conveniente definire le versioni adimensionali degli L-momenti, cosa che si può ottenere dividendo gli L-momenti di ordine superiore per la misura di scala λ_2. Si ottengono così i *rapporti degli L-momenti*

$$\tau_r = \lambda_r/\lambda_2 \ , \quad r = 3, 4, \ldots \tag{A.31}$$

che misurano la forma di una distribuzione indipendentemente dalla scala. Si definisce inoltre il coefficiente di L-variazione (L-CV) come

$$\tau = \lambda_2/\lambda_1 \ , \tag{A.32}$$

quantità analoga al coefficiente di variazione CV.

A.5 Proprietà degli L-momenti

Gli L-momenti λ_1 e λ_2, l'L-CV τ e i rapporti degli L-momenti τ_3 e τ_4 sono le quantità che *Hosking* & *Wallis* (1997) consigliano di utilizzare per descrivere le distribuzioni di probabilità. Le loro più importanti proprietà sono:

- **Esistenza.** Se esiste la media della distribuzione, allora esistono tutti i suoi L-momenti.

- **Unicità.** Se esiste la media della distribuzione, allora gli L-momenti definiscono tale distribuzione in maniera univoca, ovvero non esistono due distribuzioni diverse con gli stessi L-momenti.

- **Terminologia.** Gli L-momenti (ed i rapporti degli L-momenti) che si utilizzano hanno un determinato significato, paragonabile a quello dei momenti campionari: λ_1 è la L-posizione (o la media) della distribuzione; λ_2 è l'L-scala; τ è l'L-CV; τ_3 è l'L-skewness; τ_4 è l'L-kurtosis.

- **Limiti algebrici.** λ_1 può assumere qualsiasi valore; $\lambda_2 \geq 0$; per una distribuzione che assume solo valori positivi $0 \leq \tau < 1$; i rapporti degli L-momenti soddisfano l'uguaglianza $|\tau_r| < 1$ per ogni $r \geq 3$. Limiti più precisi possono essere trovati per ogni τ_r: ad esempio, dato τ_3, allora $(5\tau_3^2 - 1)/4 \leq \tau_4 < 1$ e, per distribuzioni che assumono solo valori positivi, dato τ si ha che $2\tau - 1 \leq \tau_3 < 1$.

- **Trasformazioni lineari.** Siano X e Y due variabili casuali con L-momenti λ_r e λ_r^* rispettivamente, e si supponga che $Y = aX + b$. Allora $\lambda_1^* = a\lambda_1 + b$; $\lambda_2^* = |a|\lambda_2$; $\tau_r^* = (\text{sign}(a))^r \tau_r$ per $r \geq 3$.

- **Simmetria.** Sia X una variabile casuale simmetrica con media μ, ossia $\Pr[X \geq \mu + x] = \Pr[X \leq \mu - x]$ per ogni x. Allora tutti i rapporti degli L-momenti di ordine dispari valgono 0, ovvero $\tau_r = 0$ se $r = 3, 5, 7, \ldots$.

Gli L-momenti sono stati calcolati per molte distribuzioni (si veda l'Appendice B). La distribuzione che gioca un ruolo centrale nella teoria degli L-momenti, analoga alla distribuzione Normale nella teoria dei momenti ordinari, è la distribuzione uniforme. Si può dimostrare che tutti gli L-momenti λ_r e rapporti degli L-momenti τ_r di ordine superiore (con $r \geq 3$) valgono zero per la distribuzione uniforme. La distribuzione Normale, per il fatto che è simmetrica, presenta gli L-momenti di ordine dispari nulli, ma quelli di ordine pari non sono particolarmente semplici: ad esempio $\tau_4 \approx 0.123$. La distribuzione esponenziale, invece, ha dei rapporti degli L-momenti particolarmente semplici: $\tau_3 = 1/3$, $\tau_4 = 1/6$.

Un modo conveniente per rappresentare gli L-momenti di diverse distribuzioni è il *diagramma dei rapporti degli L-momenti*, esemplificato in Figura A.1. Questo diagramma mostra gli L-momenti in un grafico i cui assi sono l'L-skewness e l'L-kurtosis. Una distribuzione a due parametri, caratterizzata da un parametro di posizione ed uno di scala, viene rappresentata sul diagramma da un punto. Infatti se due distribuzioni differiscono solo nei parametri di posizione

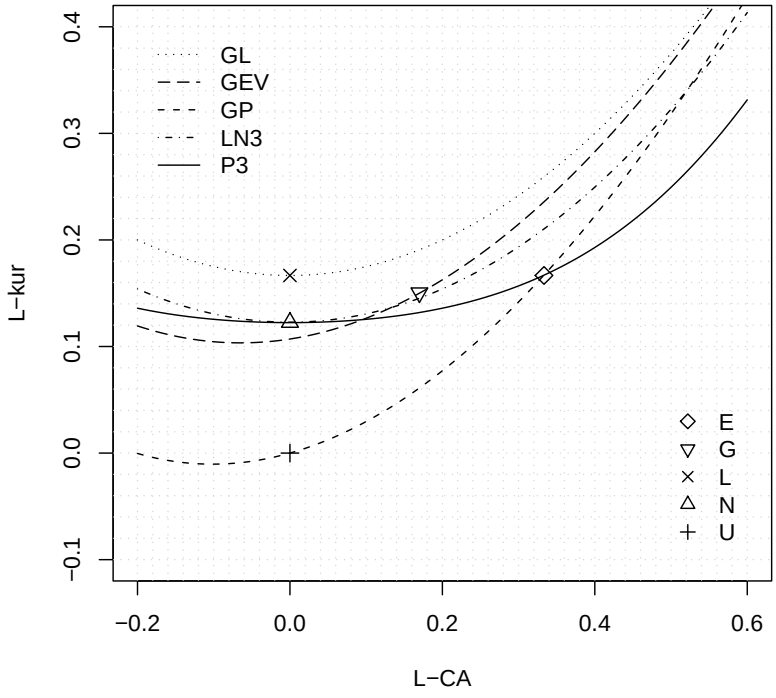

Figura A.1: Diagramma dei rapporti degli L-momenti. Le distribuzioni a due ed a tre parametri sono riportate come punti e come linee rispettivamente. Le distribuzioni a due parametri sono: esponenziale (E), Gumbel (G), lognormale (L), normale (N) ed uniforme (U); quelle a tre parametri sono: logistica generalizzata (GL), generalizzata del valore estremo (GEV), Pareto generalizzata (GP), lognormale a 3 parametri (LN3) e Pearson tipo III (P3).

e di scala, allora sono distribuzioni di due variabili casuali X e $Y = aX + b$ con $a > 0$, per cui, dato la proprietà delle trasformazioni lineari degli L-momenti ($\tau_r^* = (\text{sign}(a))^r \tau_r$), hanno gli stessi L-skewness ed L-kurtosis. Una distribuzione a tre parametri, invece, dal momento che è caratterizzata dai parametri di posizione, scala e forma, viene rappresentata sul diagramma da una linea, i cui punti corrispondono a differenti valori del parametro di forma. Distribuzioni con più di un parametro di forma generalmente ricoprono aree bidimensionali

sul diagramma.

A.6 L-momenti campionari

Gli L-momenti sono stati definiti per una distribuzione di probabilità, ma nella pratica devono essere stimati a partire da campioni finiti. La loro stima è basata su un campione di lunghezza n, ordinato in senso crescente: $x_{1:n} \leq x_{2:n} \leq \cdots \leq x_{n:n}$. È conveniente cominciare dalla definizione dello stimatore del momento pesato in probabilità β_r. Uno stimatore indistorto di β_r è

$$b_r = n^{-1}\binom{n-1}{r}^{-1} \sum_{j=r+1}^{n} \binom{j-1}{r} x_{j:n} \ , \tag{A.33}$$

ovvero

$$b_r = n^{-1} \sum_{j=r+1}^{n} \frac{(j-1)(j-2)\dots(j-r)}{(n-1)(n-2)\dots(n-r)} x_{j:n} \ . \tag{A.34}$$

Analogamente alle Equazioni (A.29) e (A.30), gli *L-momenti campionari* sono definiti come

$$\begin{aligned} l_1 &= b_0 \ , \\ l_2 &= 2b_1 - b_0 \ , \\ l_3 &= 6b_2 - 6b_1 + b_0 \ , \\ l_4 &= 20b_3 - 30b_2 + 12b_1 - b_0 \ , \end{aligned} \tag{A.35}$$

e, in generale,

$$l_{r+1} = \sum_{k=0}^{r} p^*_{r,k} b_k \ , \quad r = 0, 1, \dots, n-1 \ , \tag{A.36}$$

dove i coefficienti $p^*_{r,k}$ sono definiti dall'Equazione (A.27). L'L-momento campionario l_r è uno stimatore indistorto di λ_r.

Analogamente alle Equazioni (A.31) e (A.32), i *rapporti degli L-momenti campionari* sono definiti con

$$t_r = l_r / l_2 \ , \quad r = 3, 4, \dots \ , \tag{A.37}$$

e l'L-CV campionario con

$$t = l_2 / l_1 \ . \tag{A.38}$$

Questi sono gli stimatori naturali di τ_r e di τ rispettivamente, ma non sono indistorti. Ad ogni modo la loro distorsione è molto piccola per campioni di lunghezza moderata o grande. Si può dimostrare che il bias relativo asintotico

di t_3 per una distribuzione Gumbel vale $0.19n^{-1}$, e che vale $0.03n^{-1}$ per t_4 di una distribuzione Normale (n è la lunghezza del campione). *Hosking & Wallis* (1997) hanno calcolato il bias per campioni piccoli estratti da diverse distribuzioni con metodi di simulazione Monte-Carlo. Ad esempio, si è visto come generalmente il bias di t sia trascurabile per campioni con $n \geq 20$. Per quanto riguarda i bias di t_3 e t_4, essi sono certamente piccoli in confronto alle differenze negli L-momenti delle diverse famiglie di distribuzioni. Ciò significa che ci si può aspettare che l'identificazione di un tipo di distribuzione generatrice a partire dagli L-momenti di un campione sia comunque indistorta, a prescindere dalla distorsione degli stimatori. Ad ogni modo sia l'L-skewness che l'L-kurtosis sono molto meno distorti dello skewness e del kurtosis ordinari.

A.7 Momenti e L-momenti

Sia i momenti che gli L-momenti sono misure di posizione, scala e forma delle distribuzioni di probabilità. Il parametro di L-posizione λ_1 è la media e, ovviamente, equivale al momento del primo ordine μ. Per quanto riguarda il parametro di L-scala λ_2, rispetto alla deviazione standard σ vale la relazione $\sigma \geq \sqrt{3}\lambda_2$. Entrambe le misure valutano le differenze tra i valori estratti casualmente da una distribuzione, ma σ dà più peso alle differenze più marcate. La stessa relazione esiste tra CV e L-CV, i cui stimatori soddisfano la disuguaglianza

$$\hat{C}_v \geq \left(\frac{3n}{n+1}\right)^{1/2} t \ . \tag{A.39}$$

Una relazione di questo genere non può essere scritta per le misure di skewness τ_3 e γ, dal momento che la situazione è molto diversa a seconda della distribuzione considerata. Per quanto riguarda il kurtosis, τ_4 è una misura simile a κ e, come quest'ultima, difficilmente interpretabile (spessore delle code, per alcune distribuzioni). Tuttavia τ_4 dà meno peso alle code estreme della distribuzione, rispetto a κ. Come già accennato sia l'L-skewness che l'L-kurtosis sono molto meno distorti dello skewness e del kurtosis ordinari.

Inoltre gli L-momenti godono della proprietà di esistere alla sola condizione di esistenza della media della distribuzione, il che include anche casi in cui i momenti ordinari non esistono. Ad esempio, per una distribuzione GEV (Appendice B) i momenti di ordine 3 e 4 non esistono quando il parametro k di

forma della distribuzione è inferire a $-1/3$ e $-1/4$ rispettivamente. Per questi valori di k i rapporti degli L-momenti assumono valori moderati quali $\tau_3 = 0.403$ e $\tau_4 = 0.241$ rispettivamente (e campioni che presentano L-momenti campionari così elevati sono frequenti nell'analisi di dati come velocità del vento o portate di piena).

Un'altro vantaggio dei rapporti degli L-momenti è quello di essere contenuti nell'intervallo $(-1, 1)$, mentre i rapporti dei momenti possono assumere valori arbitrariamente grandi o negativi. Questa proprietà permette di dare un'interpretazione più semplice ai valori di τ_r.

I limiti algebrici dei momenti campionari sono stati menzionati nel Paragrafo A.3. I rapporti degli L-momenti non sono soggetti a tali restrizioni e i loro stimatori campionari possono assumere qualsiasi valore raggiungibile da quelli teorici.

La differenza principale tra momenti e L-momenti è che i primi danno un peso maggiore alle code estreme delle distribuzioni. Questo può essere visto semplicemente confrontando le Equazioni (A.25) e (A.28). Al crescere di r, il peso assegnato alla coda della distribuzione, $u \approx 1$, cresce come $\{x(u)\}^r$ nell'Equazione (A.25) ma come u^r nell'Equazione (A.28). Per molte distribuzioni $x(u)$ cresce molto più velocemente di u all'avvicinarsi di quest'ultima ad 1; per distribuzioni non limitate superiormente, ovviamente, $x(u) \to \infty$ se $u \to 1$. Anche i momenti campionari, di conseguenza, sono più affetti dalle osservazioni estreme degli L-momenti corrispondenti.

A.8 Stima dei parametri mediante gli L-momenti

Un problema che ci si pone comunemente in statistica è la stima, a partire da un campione casuale di n dati, della distribuzione di probabilità la cui specificazione coinvolge un numero finito, p, di parametri incogniti. Analogamente all'usuale metodo dei momenti, il *metodo degli L-momenti* ottiene una stima dei parametri sostituendo i primi p L-momenti campionari ai corrispondenti L-momenti della distribuzione. Ciò richiede di conoscere l'espressione dei parametri in termini degli L-momenti, il che è riportato nell'Appendice B per molte distribuzioni standard dell'idrologia.

Le distribuzioni esatte degli stimatori dei parametri ottenuti con il metodo degli L-momenti sono generalmente difficili da derivare, ma approssimazioni

per campioni caratterizzati da n elevato possono essere ottenute dalla teoria asintotica. Per molte distribuzioni si è riscontrato che tali stimatori sono asintoticamente distribuiti secondo una distribuzione Normale (*Hosking & Wallis*, 1997), e si sono derivati gli errori standard e gli intervalli di confidenza.

Sempre *Hosking & Wallis* (1997) sostengono che spesso, con campioni di lunghezza piccola o moderata, il metodo degli L-momenti è più efficiente di quello della massima verosimiglianza.

Appendice B

Distribuzioni di Probabilità

Nell'Appendice A si è detto che, analogamente al metodo dei momenti, il metodo degli L-momenti ottiene una stima dei parametri sostituendo i primi p L-momenti campionari ai corrispondenti L-momenti della distribuzione. Ciò richiede di conoscere l'espressione dei parametri in termini degli L-momenti, il che è l'argomento di questo capitolo. Si sono prese in considerazione alcune delle distribuzioni più utilizzate nell'analisi di frequenza regionale, nonché distribuzioni come l'uniforme e l'esponenziale, particolarmente interessanti nell'ambito della teoria degli L-momenti, come già discusso nel Paragrafo A.5.

Le parametrizzazioni riportate sono quelle descritte in *Hosking* & *Wallis* (1997), testo da cui si è attinto per la compilazione di questo capitolo. Per tutte le distribuzioni elencate si forniscono la densità di probabilità $f(x)$, la distribuzione di frequenza cumulata $F(x)$ e la funzione dei quantili $x(F)$ (si veda il Paragrafo A.1 per la loro definizione). Si riportano inoltre le espressioni degli L-momenti in funzione dei parametri e dei parametri in funzione degli L-momenti. Queste ultime possono essere usate per stimare i parametri delle distribuzioni con il metodo degli L-momenti.

B.1 Distribuzione Uniforme

La più semplice tra le distribuzioni continue è la distribuzione uniforme. Essa viene anche detta distribuzione rettangolare a causa della forma della sua distribuzione di frequenza. Una variabile è distribuita uniformemente tra due limiti α e β se la probabilità di ricadere in un intervallo (c, d) interno a (α, β)

è proporzionale all'intervallo stesso e vale $\Pr[c < X < d] = (d - c)/(\beta - \alpha)$. La distribuzione uniforme viene generalmente utilizzata per generare campioni casuali da qualsiasi altra distribuzione, oppure per assegnare una probabilità ad una variabile casuale di cui non si hanno informazioni disponibili, e per cui non si possa ipotizzare un'altra distribuzione a priori. Nell'analisi di frequenza regionale non riveste una particolare importanza. Nella teoria degli L-momenti, però, la distribuzione uniforme è analoga alla distribuzione Normale nella teoria dei momenti ordinari (vedi Paragrafo A.5).

Definizioni

Parametri (2): α (limite inferiore della distribuzione), β (limite superiore).
Campo di esistenza di x: $\alpha \leq x \leq \beta$.

$$f(x) = \frac{1}{\beta - \alpha} \tag{B.1}$$

$$F(x) = \frac{x - \alpha}{\beta - \alpha} \tag{B.2}$$

$$x(F) = \alpha + (\beta - \alpha)F \tag{B.3}$$

L-momenti

$$\lambda_1 = \frac{1}{2}(\alpha + \beta) \tag{B.4}$$

$$\lambda_2 = \frac{1}{6}(\beta - \alpha) \tag{B.5}$$

$$\tau_3 = 0 \tag{B.6}$$

$$\tau_4 = 0 \tag{B.7}$$

Parametri

La stima dei parametri della distribuzione uniforme non è di interesse nell'analisi di frequenza regionale, e pertanto è stata omessa.

B.2 Distribuzione Esponenziale

Se si considera il processo di Poisson (v.es. *Kottegoda & Rosso*, 1998), o degli eventi rari, la probabilità di non-occorrenza della variabile casuale X, che individua un successo, un arrivo, un conteggio o un qualsiasi altro evento, nell'intervallo di tempo t vale $\Pr[X = 0] = e^{-\lambda t}$ dove λ è la frequenza di accadimento dell'evento. Se si considera come variabile casuale il tempo T tra le occorrenze degli eventi di un processo di Poisson, la funzione di frequenza cumulata di T vale $F_T(T \leq t) = 1 - e^{-\lambda t}$. In altre parole, il tempo che occorre aspettare tra eventi successivi di un processo di Poisson ha una distribuzione esponenziale. Naturalmente la distribuzione esponenziale è importante in idrologia perché l'accadimento di molti processi estremi può essere studiato assimilandolo ad un processo di Poisson. Anche quando l'assunzione di Poisson non è pienamente verificata, il modello esponenziale può essere adottato come approssimazione ragionevole (*Kottegoda & Rosso*, 1998).

Definizioni

Parametri (2): ξ (limite inferiore della distribuzione), α (scala).
Campo di esistenza di x: $\xi \leq x < \infty$.

$$f(x) = \frac{1}{\alpha} e^{-\frac{x-\xi}{\alpha}} \tag{B.8}$$

$$F(x) = 1 - e^{-\frac{x-\xi}{\alpha}} \tag{B.9}$$

$$x(F) = \xi - \alpha \log(1 - F) \tag{B.10}$$

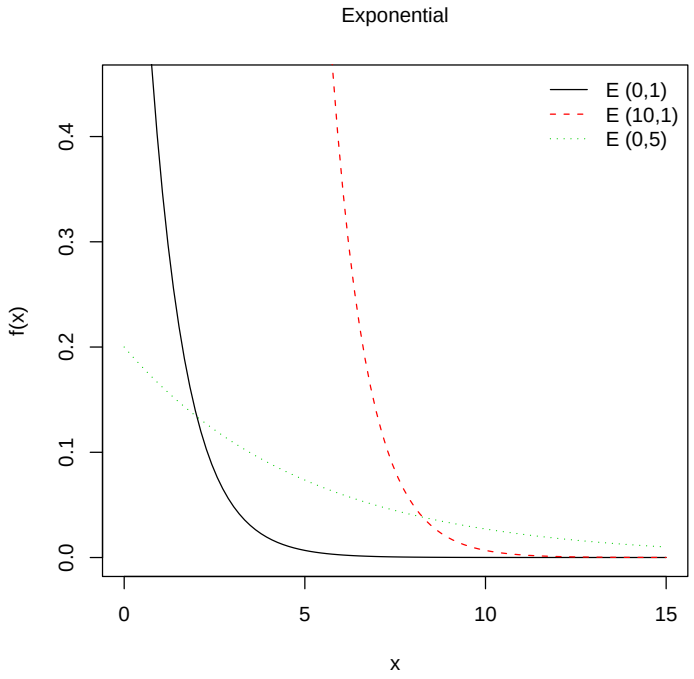

Figura B.1: Esempi di densità di probabilità della distribuzione esponenziale con parametri: (1) $\xi = 0$, $\alpha = 1$; (2) $\xi = 10$, $\alpha = 1$; (3) $\xi = 0$, $\alpha = 5$.

L-momenti

$$\lambda_1 = \xi + \alpha \tag{B.11}$$

$$\lambda_2 = \frac{1}{2}\alpha \tag{B.12}$$

$$\tau_3 = \frac{1}{3} \tag{B.13}$$

$$\tau_4 = \frac{1}{6} \tag{B.14}$$

Parametri

Se si conosce ξ, α è dato da $\alpha = \lambda_1 - \xi$ e gli stimatori con i metodi degli L-momenti, dei momenti e della massima verosimiglianza sono identici. Se ξ è incognito, i parametri sono ottenibili come

$$\alpha = 2\lambda_2 \ , \quad \xi = \lambda_1 - \alpha \ . \tag{B.15}$$

Per stime basate su un singolo campione questi stimatori sono inefficienti, ma nell'analisi di frequenza regionale permettono una stima ragionevole dei quantili della coda superiore della distribuzione.

B.3 Distribuzione di Gumbel

La distribuzione di Gumbel (*Gumbel*, 1958), appartenente alla famiglia esponenziale, è una delle più popolari nella modellazione di distribuzioni di frequenza di eventi naturali estremi. Da numerosi studi effettuati si è riscontrato che la Gumbel fornisce risultati molto consistenti ed è da preferire quando ci si riferisce a tempi di ritorno elevati. Tale distribuzione è caratterizzata da due parametri e rientra in un caso particolare della famiglia GEV (Paragrafo B.6).

Definizioni

Parametri (2): ξ (posizione), α (scala).
Campo di esistenza di x: $-\infty < x < \infty$.

$$f(x) = \frac{1}{\alpha} e^{-\frac{x-\xi}{\alpha}} e^{-e^{-\frac{x-\xi}{\alpha}}} \tag{B.16}$$

$$F(x) = e^{-e^{-\frac{x-\xi}{\alpha}}} \tag{B.17}$$

$$x(F) = \xi - \alpha \log(-\log F) \tag{B.18}$$

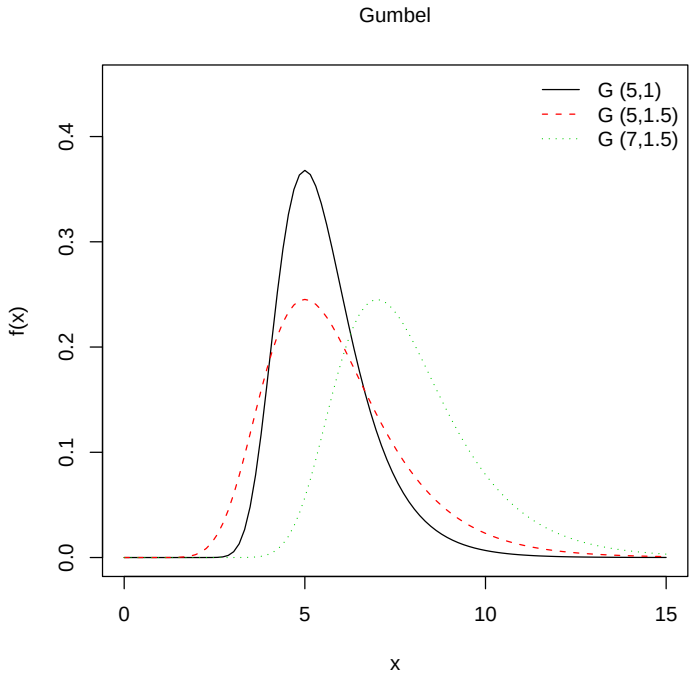

Figura B.2: Esempi di densità di probabilità della distribuzione di Gumbel con parametri: (1) $\xi = 5$, $\alpha = 1$; (2) $\xi = 5$, $\alpha = 1.5$; (3) $\xi = 7$, $\alpha = 1.5$.

L-momenti

$$\lambda_1 = \xi + \alpha\gamma \tag{B.19}$$

dove γ è la costante di Eulero, $0.5772\ldots$.

$$\lambda_2 = \alpha \log 2 \tag{B.20}$$

$$\tau_3 = 0.1699 = \log(9/8)/\log(2) \tag{B.21}$$

$$\tau_4 = 0.1504 = [16\log(2) - 10\log(3)]/\log(2) \tag{B.22}$$

Parametri

$$\alpha = \frac{\lambda_2}{\log 2} \quad \xi = \lambda_1 - \gamma\alpha \ . \tag{B.23}$$

B.4 Distribuzione Normale

La distribuzione Normale, o distribuzione di Gauss, storicamente ha giocato un ruolo di fondamentale importanza in statistica. Questo è principalmente dovuto al teorema del limite centrale. Il teorema sancisce che, sotto certe condizioni, la distribuzione di una somma di variabili casuali tende ad una Gaussiana all'aumentare del numero di termini della somma, qualunque sia la distribuzione originaria delle variabili. La Normale è largamente utilizzata per la modellazione di distribuzioni empiriche con coefficienti di skewness vicini allo zero.

Definizioni

Parametri (2): μ (posizione), σ (scala).
Campo di esistenza di x: $-\infty < x < \infty$.

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \tag{B.24}$$

$$F(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{x} e^{-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2} dt \tag{B.25}$$

Per $x(F)$ non esiste una forma analitica, ma si può ricorrere all'utilizzo di metodi numerici.
Per poter svincolare la distribuzione di probabilità dai due parametri si può trasformare la variabile aleatoria originale X nella nuova variabile aleatoria Z

$$Z = \frac{(X - \mu)}{\sigma} \ . \tag{B.26}$$

Tale nuova variabile prende il nome di *variabile ridotta* o *standardizzata*. Esprimendo la densità di probabilità e la distribuzione di frequenza cumulata in funzione della variabile standardizzata si ottengono le equazioni:

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \ , \tag{B.27}$$

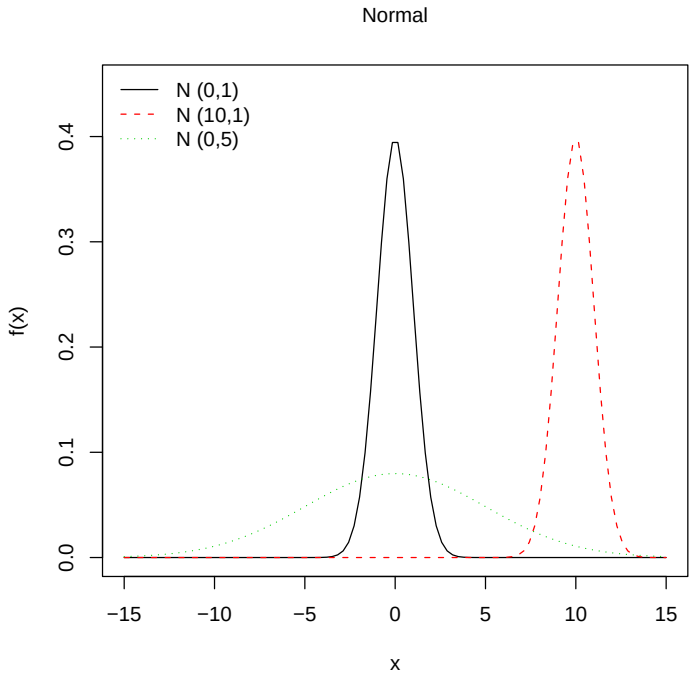

Figura B.3: Esempi di densità di probabilità della distribuzione Normale con parametri: (1) $\mu = 0$, $\sigma = 1$; (2) $\mu = 10$, $\sigma = 1$; (3) $\mu = 0$, $\sigma = 5$.

$$\Phi(z) = \int_{-\infty}^{z} \phi(t)dt \ . \tag{B.28}$$

La distribuzione è ancora una Normale, con media pari a zero e deviazione standard unitaria.

L-momenti

$$\lambda_1 = \mu \tag{B.29}$$

$$\lambda_2 = 0.5642\sigma = \pi^{-1/2}\sigma \tag{B.30}$$

$$\tau_3 = 0 \tag{B.31}$$

$$\tau_4 = 0.1226 = 30\pi^{-1} \arctan\sqrt{2} - 9 \tag{B.32}$$

Parametri

$$\mu = \lambda_1 \ , \quad \sigma = \pi^{1/2}\lambda_2 \ . \tag{B.33}$$

B.5 Distribuzione di Pareto Generalizzata

La distribuzione di Pareto Generalizzata è molto usata nell'analisi degli eventi estremi (*Pickands* (1975) è stato probabilmente il primo ad utilizzarla in questo contesto), specialmente in idrologia e negli studi di affidabilità, quando occorre utilizzare alternative alla distribuzione esponenziale assumendo più spessa o più sottile la coda superiore della distribuzione.

Definizioni

Parametri (3): ξ (posizione), α (scala), k (forma).
Campo di esistenza di $\xi < x \leq \xi + \alpha/k$ se $k > 0$; $\xi \leq x < \infty$ se $k \leq 0$.

$$f(x) = \frac{1}{\alpha} e^{-(1-k)y} \tag{B.34}$$

dove

$$y = \begin{cases} -k^{-1}\log\{1 - k(x-\xi)/\alpha\} \ , & k \neq 0 \\ (x-\xi)/\alpha \ , & k = 0 \end{cases} \tag{B.35}$$

$$F(x) = 1 - e^{-y} \tag{B.36}$$

$$x(F) = \begin{cases} \xi + \alpha[1 - (1-F)^k]/k \ , & k \neq 0 \\ \xi - \alpha\log(1-F) \ , & k = 0 \end{cases} \tag{B.37}$$

Casi speciali: $k = 0$ è la distribuzione esponenziale; $k = 1$ è la distribuzione uniforme nell'intervallo $\xi \leq x \leq \xi + \alpha$.

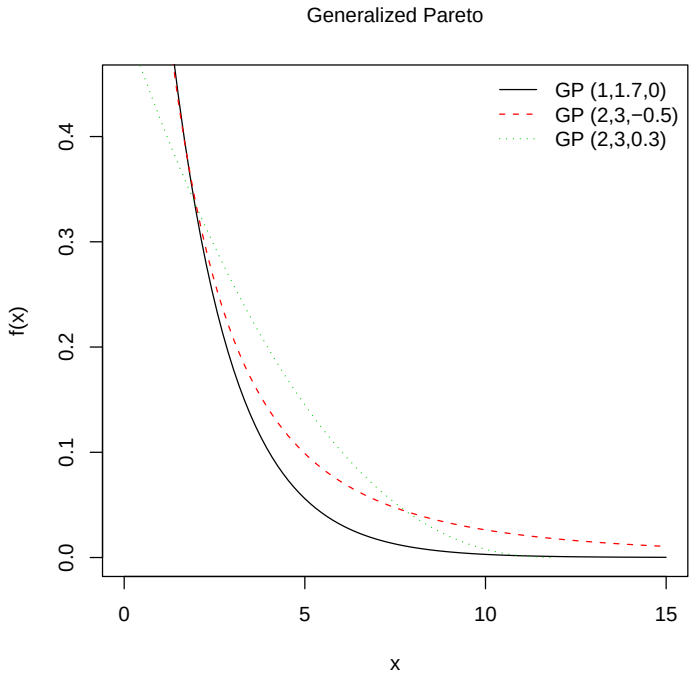

Figura B.4: Esempi di densità di probabilità della distribuzione generalizzata di Pareto con parametri: (1) $\xi = 1$, $\alpha = 1.7$, $k = 0$; (2) $\xi = 2$, $\alpha = 3$, $k = -0.5$; (3) $\xi = 2$, $\alpha = 3$, $k = 0.3$.

L-momenti

Gli L-momenti sono definiti per $k > -1$.

$$\lambda_1 = \xi + \alpha/(1+k) \tag{B.38}$$

$$\lambda_2 = \alpha/[(1+k)(2+k)] \tag{B.39}$$

$$\tau_3 = (1-k)/(3+k) \tag{B.40}$$

$$\tau_4 = (1-k)(2-k)/[(3+k)(4+k)] \tag{B.41}$$

La relazione tra τ_3 e τ_4 è data da

$$\tau_4 = \frac{\tau_3(1+5\tau_3)}{5+\tau_3} \ . \tag{B.42}$$

Parametri

Se si conosce ξ, i due parametri α e k sono dati da

$$k = (\lambda_1 - \xi)/\lambda_2 - 2 \ , \quad \alpha = (1+k)(\lambda_1 - \xi) \ . \tag{B.43}$$

Se invece ξ è incognito, i tre parametri sono dati da

$$k = \frac{1-3\tau_3}{1+\tau_3} \ , \quad \alpha = (1+k)(2+k)\lambda_2 \ , \quad \xi = \lambda_1 - (2+k)\lambda_2 \ . \tag{B.44}$$

B.6 Distribuzione Generalizzata del Valore Estremo

La Generalized Extreme Value, nota anche come GEV, è una distribuzione a tre parametri, derivante dalla teoria dei valori estremi. Fu introdotta da *Jenkinson* (1955) per identificare la distribuzione di frequenza dei valori estremi per dati meterologici. La GEV è largamente utilizzata in ambito idrologico soprattutto per lo studio di piene e piogge intense. Il vantaggio principale della GEV è la sua generalità, ossia il fatto di contemplare tutte le possibili distribuzioni del valore estremo. Infatti, a seconda del valore assunto dal parametro di forma k la GEV è equivalente alle distribuzioni EV-1 (Gumbel), EV-2 ed EV-3.

Definizioni

Parametri (3): ξ (posizione), α (scala), k (forma).
Campo di esistenza di $-\infty < x \leq \xi + \alpha/k$ se $k > 0$; $-\infty < x < \infty$ se $k = 0$; $\xi + \alpha/k \leq x < \infty$ se $k < 0$.

$$f(x) = \frac{1}{\alpha} e^{-(1-k)y - e^{-y}} \tag{B.45}$$

dove

$$y = \begin{cases} -k^{-1}\log\{1-k(x-\xi)/\alpha\}\ , & k \neq 0 \\ (x-\xi)/\alpha\ , & k = 0 \end{cases} \tag{B.46}$$

$$F(x) = e^{-e^{-y}} \tag{B.47}$$

$$x(F) = \begin{cases} \xi + \alpha[1-(-\log F)^k]/k\ , & k \neq 0 \\ \xi - \alpha\log(-\log F)\ , & k = 0 \end{cases} \tag{B.48}$$

Casi speciali: $k = 0$ è la distribuzione Gumbel; $k = 1$ è la distribuzione esponenziale inversa, ovvero $1 - F(-x)$ è la distribuzione di frequenza cumulata di una distribuzione esponenziale.

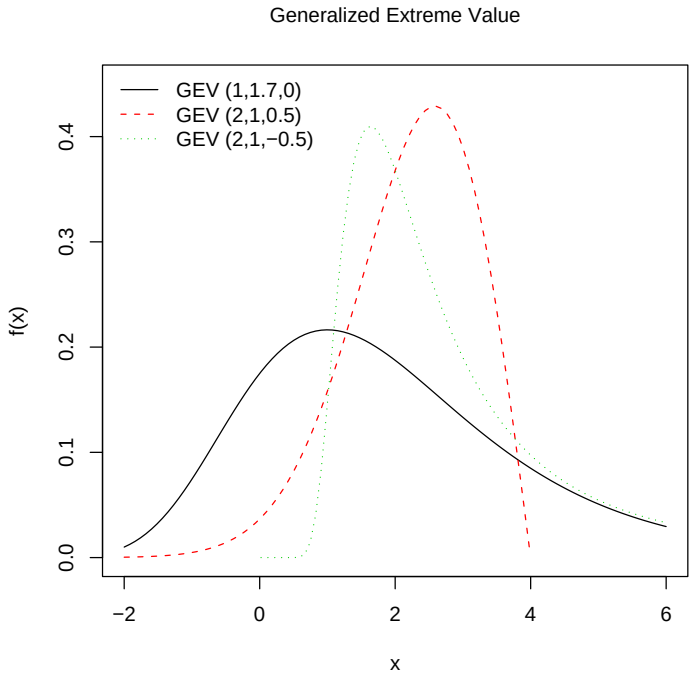

Figura B.5: Esempi di densità di probabilità della distribuzione generalizzata del valore estremo con parametri: (1) $\xi = 1$, $\alpha = 1.7$, $k = 0$; (2) $\xi = 2$, $\alpha = 1$, $k = 0.5$; (3) $\xi = 2$, $\alpha = 1$, $k = -0.5$.

L-momenti

Gli L-momenti sono definiti per $k > -1$.

$$\lambda_1 = \xi + \alpha[1 - \Gamma(1+k)]/k \tag{B.49}$$

dove $\Gamma(.)$ indica la funzione gamma

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \tag{B.50}$$

$$\lambda_2 = \alpha(1 - 2^{-k})\Gamma(1+k)]/k \tag{B.51}$$

$$\tau_3 = 2(1 - 3^{-k})/(1 - 2^{-k}) - 3 \tag{B.52}$$

$$\tau_4 = [5(1 - 4^{-k}) - 10(1 - 3^{-k}) + 6(1 - 2^{-k})]/(1 - 2^{-k}) \tag{B.53}$$

Parametri

Per ricavare k partendo dalla stima degli L-Momenti campionari è necessario invertire l'Equazione (B.52). Purtroppo non ci sono soluzioni esplicite per ricavare k, quindi si deve ricorrere a un'approssimazione numerica proposta da *Hosking & Wallis* (1997), che ha una precisione di 9×10^{-4} per $-0.5 \leq \tau_3 \leq 0.5$:

$$k \approx 7.8590c + 2.9554c^2 \ , \quad c = \frac{2}{3+\tau_3} - \frac{\log 2}{\log 3}. \tag{B.54}$$

Gli altri due parametri sono ricavabili da

$$\alpha = \frac{\lambda_2 k}{(1 - 2^{-k})\Gamma(1+k)} \ , \quad \xi = \lambda_1 - \alpha\left[1 - \Gamma(1+k)\right]/k \ . \tag{B.55}$$

B.7 Distribuzione Logistica Generalizzata

La funzione logistica e la distribuzione logistica sono state utilizzate in moltissimi diversi campi di applicazione (*Johnson et al.*, 1995). Da un punto di vista puramente statistico, la distribuzione logistica nasce come distribuzione limite ($n \to \infty$) delle medie standardizzate dei valori estremi (grandi e piccoli)

di campioni casuali di lunghezza n (*Gumbel*, 1944). Esistono differenti forme di generalizzazione della distribuzione logistica. Quella qui riportata è una versione riparametrizzata della distribuzione log-logistica di *Ahmad et al.* (1988), che permette di mostrare la somiglianza della distribuzione con la Pareto generalizzata (Paragrafo B.5) e la GEV (Paragrafo B.6).

Definizioni

Parametri (3): ξ (posizione), α (scala), k (forma).
Campo di esistenza di $-\infty < x \leq \xi + \alpha/k$ se $k > 0$; $-\infty < x < \infty$ se $k = 0$; $\xi + \alpha/k \leq x < \infty$ se $k < 0$.

$$f(x) = \frac{\alpha^{-1} e^{-(1-k)y}}{(1+e^{-y})^2} \tag{B.56}$$

dove

$$y = \begin{cases} -k^{-1}\log\{1-k(x-\xi)/\alpha\}\ , & k \neq 0 \\ (x-\xi)/\alpha\ , & k = 0 \end{cases} \tag{B.57}$$

$$F(x) = \frac{1}{1+e^{-y}} \tag{B.58}$$

$$x(F) = \begin{cases} \xi + \alpha[1-\{(1-F)/F\}^k]/k\ , & k \neq 0 \\ \xi - \alpha\log\{(1-F)/F\}\ , & k = 0 \end{cases} \tag{B.59}$$

Casi speciali: $k = 0$ è la distribuzione logistica.

***L*-momenti**

Gli L-momenti sono definiti per $-1 < k < 1$.

$$\lambda_1 = \xi + \alpha[1/k - \pi/\sin(k\pi)] \tag{B.60}$$

$$\lambda_2 = \alpha k\pi/\sin(k\pi) \tag{B.61}$$

$$\tau_3 = -k \tag{B.62}$$

$$\tau_4 = (1+5k^2)/6 \tag{B.63}$$

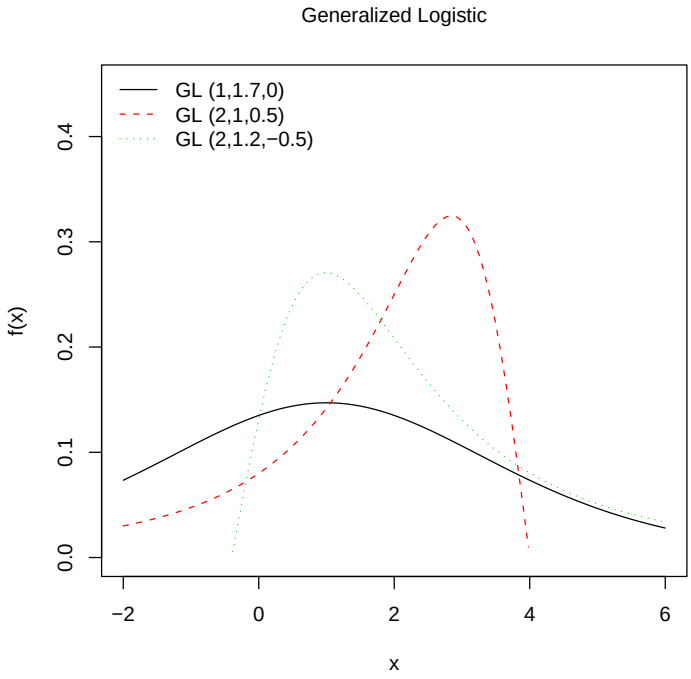

Figura B.6: Esempi di densità di probabilità della distribuzione logistica generalizzata con parametri: (1) $\xi = 1$, $\alpha = 1.7$, $k = 0$; (2) $\xi = 2$, $\alpha = 1$, $k = 0.5$; (3) $\xi = 2$, $\alpha = 1.2$, $k = -0.5$.

Parametri

$$k = -\tau_3 \ , \quad \alpha = \frac{\lambda_2 \sin(k\pi)}{k\pi} \ , \quad \xi = \lambda_1 - \alpha \left(\frac{1}{k} - \frac{\pi}{\sin(k\pi)} \right) \ . \tag{B.64}$$

B.8 Distribuzione Lognormale

È noto come la distribuzione log-Normale sia applicabile ad una gran varietà di fenomeni idrologici, specialmente quando le variabili hanno limite inferiore. Difatti la curva di densità di probabilità si presenta con andamento non simmetrico e limite inferiore. Storicamente si sono susseguiti vari studi che rivelano come

la distribuzione log-Normale sia adattabile all'applicazione in svariati campi, dalla modellazione di picchi di portata (*Chow*, 1954), alla descrizione delle concentrazioni di inquinanti in atmosfera (*Georgopoulos* & *Seinfeld*, 1982). È interessante notare che esiste anche una giustificazione teorica all'uso di tale distribuzione: si consideri il prodotto di una serie di variabili $X = W_1 W_2 \dots W_N$; facendo il logaritmo di ambo i membri si ricava $ln(X) = ln(W_1) + ln(W_2) + \dots + ln(W_N)$, da cui, per il teorema del limite centrale, si ottiene che X deve avere una distribuzione log-Normale, quando il numero di fattori nella moltiplicazione tende ad essere sufficientemente elevato.

Definizioni

Parametri (3): ξ (posizione), α (scala), k (forma).
Campo di esistenza di $-\infty < x \leq \xi + \alpha/k$ se $k > 0$; $-\infty < x < \infty$ se $k = 0$; $\xi + \alpha/k \leq x < \infty$ se $k < 0$.

$$f(x) = \frac{e^{ky - y^2/2}}{\alpha\sqrt{2\pi}} \tag{B.65}$$

dove

$$y = \begin{cases} -k^{-1}\log\{1 - k(x-\xi)/\alpha\} , & k \neq 0 \\ (x-\xi)/\alpha , & k = 0 \end{cases} \tag{B.66}$$

$$F(x) = \Phi(y) \tag{B.67}$$

dove $\Phi(.)$ è la distribuzione di frequenza cumulata della distribuzione Normale standardizzata, definita dall'Equazione (B.28).
Per $x(F)$ non esiste una forma analitica, ma si può ricorrere all'utilizzo di metodi numerici.
Casi speciali: $k = 0$ è la distribuzione Normale con parametri ξ e α.
In questa parametrizzazione della distribuzione lognormale, proposta in *Hosking* & *Wallis* (1997), la variabile casuale X è legata alla variabile casuale Y, che è distribuita secondo una Normale standard, da

$$X = \begin{cases} \xi + \alpha(1 - e^{-kZ})/k , & k \neq 0 \\ \xi + \alpha Z , & k = 0 \end{cases} \tag{B.68}$$

La parametrizzazione standard, esprimibile come

$$F(x) = \Phi\left(\frac{\log(x-\zeta) - \mu}{\sigma}\right) , \quad \zeta \leq x < \infty \tag{B.69}$$

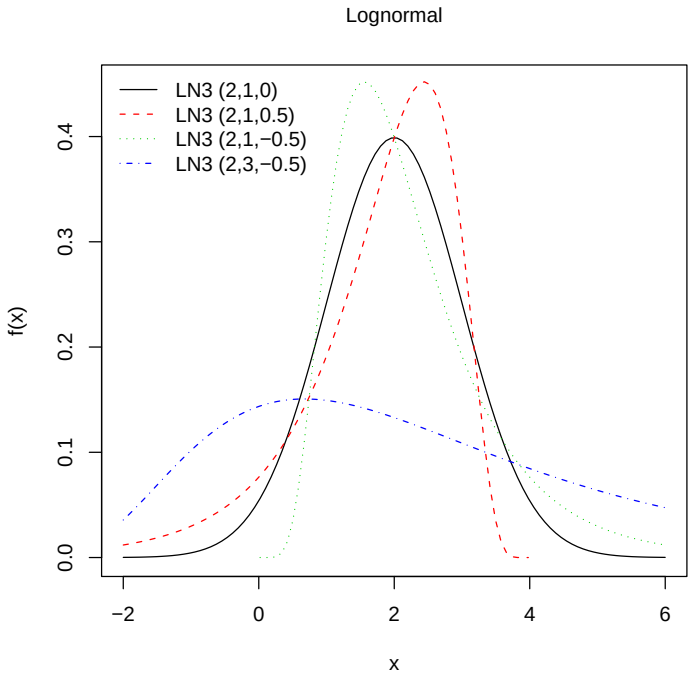

Figura B.7: Esempi di densità di probabilità della distribuzione lognormale con parametri: (1) $\xi = 2$, $\alpha = 1$, $k = 0$; (2) $\xi = 2$, $\alpha = 1$, $k = 0.5$; (3) $\xi = 2$, $\alpha = 1$, $k = -0.5$; (4) $\xi = 2$, $\alpha = 3$, $k = -0.5$.

può essere ottenuta a partire dalla parametrizzazione di Hosking e Wallis considerando che

$$k = -\sigma \ , \quad \alpha = \sigma e^{\mu} \ , \quad \xi = \zeta + e^{\mu} \ . \tag{B.70}$$

L-momenti

Gli L-momenti sono definiti per tutti i valori di k.

$$\lambda_1 = \xi + \alpha(1 - e^{k^2/2})/k \tag{B.71}$$

$$\lambda_2 = \frac{\alpha}{k} e^{k^2/2}[1 - 2\Phi(-k/\sqrt{2})] \tag{B.72}$$

Non esistono espressioni semplici per i rapporti degli L-momenti τ_r con $r \geq 3$. Essi sono funzione del solo parametro k e possono essere calcolati tramite integrazione numerica. *Hosking* & *Wallis* (1997) propongono un approssimazione per τ_3 e τ_4 che ha un'accuratezza rispettivamente migliore di 2×10^{-7} e 5×10^{-7} per $|k| \leq 4$, ovvero per $|\tau_3| \leq 0.99$ e $\tau_4 \leq 0.98$:

$$\tau_3 \approx -k \frac{A_0 + A_1 k^2 + A_2 k^4 + A_3 k^6}{1 + B_1 k^2 + B_2 k^4 + B_3 k^6} \ , \tag{B.73}$$

$$\tau_4 \approx \tau_4^0 + k^2 \frac{C_0 + C_1 k^2 + C_2 k^4 + C_3 k^6}{1 + D_1 k^2 + D_2 k^4 + D_3 k^6} \ , \tag{B.74}$$

dove i coefficienti sono riportati in Tabella B.I.

Tabella B.I: Coefficienti di approssimazione per le Equazioni (B.73), (B.74) e (B.75).

	$\tau_4^0 = 1.2260172 \times 10^{-1}$	
$A_0 = 4.8860251 \times 10^{-1}$	$C_0 = 1.8756590 \times 10^{-1}$	$E_0 = 2.0466534$
$A_1 = 4.4493076 \times 10^{-3}$	$C_1 = -2.5352147 \times 10^{-3}$	$E_1 = -3.6544371$
$A_2 = 8.8027039 \times 10^{-4}$	$C_2 = 2.6995102 \times 10^{-4}$	$E_2 = 1.8396733$
$A_3 = 1.1507084 \times 10^{-6}$	$C_3 = -1.8446680 \times 10^{-6}$	$E_3 = -0.20360244$
$B_1 = 6.4662924 \times 10^{-2}$	$D_1 = 8.2325617 \times 10^{-2}$	$F_1 = -2.0182173$
$B_2 = 3.3090406 \times 10^{-3}$	$D_2 = 4.2681448 \times 10^{-3}$	$F_2 = 1.2420401$
$B_3 = 7.4290680 \times 10^{-5}$	$D_3 = 1.1653690 \times 10^{-4}$	$F_3 = -0.21741801$

Parametri

Il parametro di forma k è funzione del solo τ_3. Non esiste una soluzione esplicita possibile, ma la seguente approssimazione ha un'accuratezza relativa di 2.5×10^{-6} per $|\tau_3| \leq 0.94$, corrispondente a $|k| \leq 3$:

$$k \approx -\tau_3 \frac{E_0 + E_1 \tau_3^2 + E_2 \tau_3^4 + E_3 \tau_3^6}{1 + F_1 \tau_3^2 + F_2 \tau_3^4 + F_3 \tau_3^6} \ . \tag{B.75}$$

I coefficienti usati nell'approssimazione sono quelli indicati in Tabella B.I. Gli altri parametri sono quindi ottenibili come

$$\alpha = \frac{\lambda_2 k e^{-k^2/2}}{1 - 2\Phi(-k/\sqrt{2})} \ , \quad \xi = \lambda_1 - \frac{\alpha}{k}(1 - e^{k^2/2}) \ . \tag{B.76}$$

B.9 Distribuzione di Pearson Tipo III

La distribuzione Gamma, o Pearson tipo III, è stata largamente utilizzata nel campo idrologico, ad esempio per la descrizione di grandezze quali portate massime e minime annue, volumi idrici stagionali e annuali ed anche gli eventi estremi di precipitazione. Nel caso in esame si fa riferimento alla distribuzione Gamma con tre parametri, che produce una PDF asimmetrica che può essere limitata superiormente o inferiormente a seconda che il valore del parametro di scala sia negativo o positivo, rispettivamente.

Definizioni

Parametri (3): ξ (posizione), β (scala), α (forma). Il legame con i momenti ordinari μ, σ e γ è dato da $\alpha = 4/\gamma^2$, $\beta = \frac{1}{2}\sigma|\gamma|$ e $\xi = \mu - 2\sigma/\gamma$.
Se $\beta > 0$, allora il campo di esistenza di x è $\xi \leq x < \infty$, mentre se $\beta < 0$ è $-\infty < x \leq \xi$.

$$f(x) = \frac{1}{|\beta|\Gamma(\alpha)} \left(\frac{x-\xi}{\beta} \right)^{\alpha-1} e^{-(x-\xi)/\beta} \tag{B.77}$$

$$F(x) = \begin{cases} G\left(\alpha, \frac{x-\xi}{\beta}\right)/\Gamma(\alpha) & \text{per } \beta > 0 \\ 1 - G\left(\alpha, \frac{x-\xi}{\beta}\right)/\Gamma(\alpha) & \text{per } \beta < 0 \end{cases} \tag{B.78}$$

dove $\Gamma(.)$ indica la funzione gamma definita nell'Equazione (B.50) e

$$G(\alpha, x) = \int_0^x t^{\alpha-1} e^{-t} dt \tag{B.79}$$

è la funzione gamma incompleta.
La funzione dei quantili $x(F)$ non può essere esplicitata analiticamente, quindi deve essere valutata numericamente.
Casi speciali: se $\alpha > 100$, cioè assume valori elevati, la distribuzione Gamma si comporta come una Normale, con parametri μ e σ.

L-momenti

Gli L-momenti sono definiti per $0 < \alpha < \infty$.

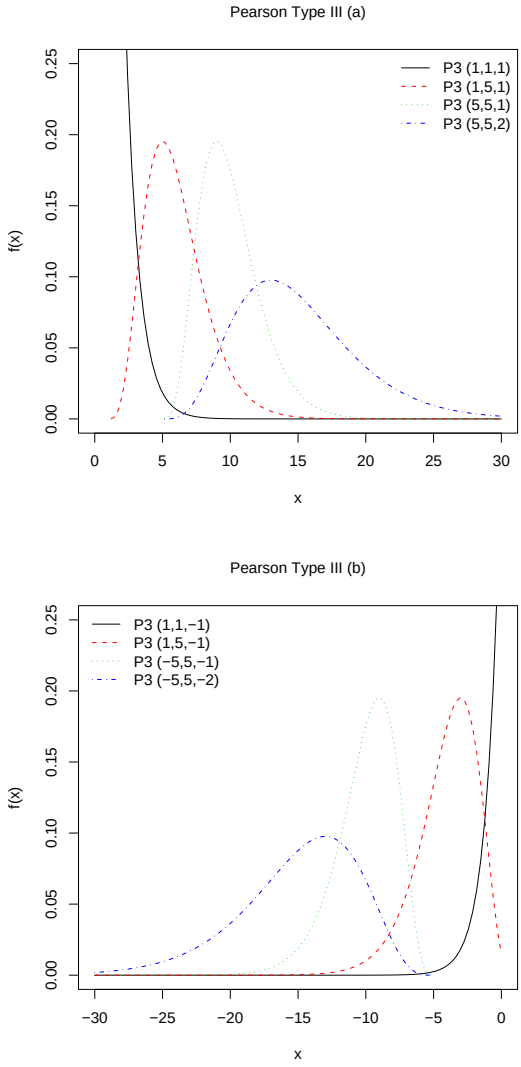

Figura B.8: Esempi di densità di probabilità della distribuzione di Pearson tipo III con parametri: (a.1) $\xi = 1$, $\alpha = 1$, $\beta = 1$; (a.2) $\xi = 1$, $\alpha = 5$, $\beta = 1$; (a.3) $\xi = 5$, $\alpha = 5$, $\beta = 1$; (a.4) $\xi = 5$, $\alpha = 5$, $\beta = 2$; (b.1) $\xi = 1$, $\alpha = 1$, $\beta = -1$; (b.2) $\xi = 1$, $\alpha = 5$, $\beta = -1$; (b.3) $\xi = -5$, $\alpha = 5$, $\beta = -1$; (b.4) $\xi = -5$, $\alpha = 5$, $\beta = -2$.

$$\lambda_1 = \xi + \alpha\beta \tag{B.80}$$

$$\lambda_2 = \pi^{-1/2}\beta\Gamma(\alpha + 1/2)/\Gamma(\alpha) \tag{B.81}$$

$$\tau_3 = 6I_{1/3}(\alpha, 2\alpha) - 3 \tag{B.82}$$

dove $I_x(p, q)$ è la funzione beta incompleta

$$I_x(p, q) = \frac{\Gamma(p + q)}{\Gamma(p)\Gamma(q)} \int_0^x t^{p-1}(1 - t)^{q-1} dt \tag{B.83}$$

Per τ_4 non esiste un'espressione semplice. Per esprimere τ_3 e τ_4 in funzione di α possono essere usate delle approssimazioni con accuratezza 10^{-6}. Se $\alpha \geq 1$

$$\tau_3 \approx \alpha^{-1/2} \frac{A_0 + A_1\alpha^{-1} + A_2\alpha^{-2} + A_3\alpha^{-3}}{1 + B_1\alpha^{-1} + B_2\alpha^{-2}} \ , \tag{B.84}$$

$$\tau_4 \approx \frac{C_0 + C_1\alpha^{-1} + C_2\alpha^{-2} + C_3\alpha^{-3}}{1 + D_1\alpha^{-1} + D_2\alpha^{-2}} \ , \tag{B.85}$$

se $\alpha < 1$

$$\tau_3 \approx \frac{1 + E_1\alpha + E_2\alpha^2 + E_3\alpha^3}{1 + F_1\alpha + F_2\alpha^2 + F_3\alpha^3} \ , \tag{B.86}$$

$$\tau_4 \approx \frac{1 + G_1\alpha + G_2\alpha^2 + G_3\alpha^3}{1 + H_1\alpha + H_2\alpha^2 + H_3\alpha^3} \ . \tag{B.87}$$

I coefficienti delle approssimazioni sono riportati in Tabella B.II.

Tabella B.II: Coefficienti di approssimazione per le Equazioni (B.84)-(B.87).

$A_0 = 3.2573501 \times 10^{-1}$	$C_0 = 1.2260172 \times 10^{-1}$
$A_1 = 1.6869150 \times 10^{-1}$	$C_1 = 5.3730130 \times 10^{-2}$
$A_2 = 7.8327243 \times 10^{-2}$	$C_2 = 4.3384378 \times 10^{-2}$
$A_3 = -2.9120539 \times 10^{-3}$	$C_3 = 1.1101277 \times 10^{-2}$
$B_1 = 4.6697102 \times 10^{-1}$	$D_1 = 1.8324466 \times 10^{-1}$
$B_2 = 2.4255406 \times 10^{-1}$	$D_2 = 2.0166036 \times 10^{-1}$
$E_1 = 2.3807576$	$G_1 = 2.1235833$
$E_2 = 1.5931792$	$G_2 = 4.1670213$
$E_3 = 1.1618371 \times 10^{-1}$	$G_3 = 3.1925299$
$F_1 = 5.1533299$	$H_1 = 9.0551443$
$F_2 = 7.1425260$	$H_2 = 2.6649995 \times 10^{-1}$
$F_3 = 1.9745056$	$H_3 = 2.6193668 \times 10^{-1}$

Parametri

Per stimare α occorre invertire l'Equazione (B.82). È possibile stimare il parametro α con una precisione di 5×10^{-5}. Se $0 < |\tau_3| < \frac{1}{3}$, si assume che $z = 3\pi\tau_3^2$ e si utilizza

$$\alpha \approx \frac{1 + 0.2906z}{z + 0.1882z^2 + 0.0442z^3} \ , \tag{B.88}$$

se $\frac{1}{3} \leq |\tau_3| < 1$, si assume che $z = 1 - |\tau_3|$ e si utilizza

$$\alpha \approx \frac{0.36067z - 0.59567z^2 + 0.25361z^3}{1 - 2.78861z + 2.56096z^2 - 0.77045z^3} \ . \tag{B.89}$$

Noto il parametro α è possibile ricavare gli altri parametri come

$$\gamma = 2\alpha^{-1/2}\text{sign}(\tau_3) \ , \quad \sigma = \lambda_2\pi^{1/2}\alpha^{1/2}\Gamma(\alpha)/\Gamma(\alpha + 1/2) \ , \quad \mu = \lambda_1 \ . \tag{B.90}$$

Se $\alpha > 100$, cioè assume valori elevati, la distribuzione Gamma si comporta come una Normale, con parametri σ e $\mu = \lambda_1$, in quanto τ_3 è circa nullo; in questi casi è possibile ricavare σ utilizzando $\sigma = \lambda_2\pi^{1/2}\frac{1}{1-1/(8\alpha)+1/(128\alpha^2)}$. Una volta noto σ si ricava il parametro β da

$$\beta = \frac{1}{2}\sigma|2\alpha^{-1/2}| \ . \tag{B.91}$$

A questo punto se il parametro β è positivo il parametro di posizione vale $\xi = \lambda_1 - \alpha\beta$, mentre se β è negativo vale $\xi = \lambda_1 + \alpha\beta$.

B.10 Distribuzione Kappa

La distribuzione a quattro parametri kappa ha avuto numerosi utilizzi in campo pratico. La kappa ha come casi speciali la distribuzione di Pareto generalizzata, la GEV, la Gumbel e varie altre distribuzioni. La kappa ha è utilizzata frequentemente quando le distribuzioni a tre parametri non danno un'adeguata rappresentazione dei dati, oppure in studi sulla robustezza.

Definizioni

Parametri (4): ξ (posizione), α (scala), k, h.

Campo di esistenza di x: il limite superiore è $\xi + \alpha/k$ se $k > 0$, ∞ se $k \leq 0$; il limite inferiore è $\xi + \alpha(1 - h^{-k})/k$ se $h > 0$, $\xi + \alpha/k$ se $h \leq 0$ e $k < 0$, e $-\infty$ se $h \leq 0$ e $k \geq 0$.

$$f(x) = \alpha^{-1}[1 - k(x - \xi)/\alpha]^{1/k-1}[F(x)]^{1-h} \tag{B.92}$$

$$F(x) = \{1 - h[1 - k(x - \xi)/\alpha]^{1/k}\}^{1/h} \tag{B.93}$$

$$x(F) = \xi + \frac{\alpha}{k}\left[1 - \left(\frac{1 - F^h}{h}\right)^k\right] \tag{B.94}$$

I casi $h = 0$ e $k = 0$ sono implicitamente inclusi come limiti continui delle Equazioni (B.92)-(B.94).
Casi speciali: $h = -1$ è la distribuzione logistica generalizzata; $h = 0$ è la distribuzione generalizzata del valore estremo; $h = 1$ è la distribuzione generalizzata di Pareto.
Data la sua flessibilità, la distribuzione kappa è utile come distribuzione generatrice per la simulazione di dati artificiali, come ad esempio si fa nel caso delle misure di eterogeneità di Hosking e Wallis (si veda il Capitolo 5).

L-momenti

Gli *L*-momenti sono definiti se $h \geq 0$ e $k > -1$, oppure se $h < 0$ e $-1 < k < -1/h$.

$$\lambda_1 = \xi + \alpha(1 - g_1)/k \tag{B.95}$$

$$\lambda_2 = \alpha(g_1 - g_2)/k \tag{B.96}$$

$$\tau_3 = (-g_1 + 3g_2 - 2g_3)/(g_1 - g_2) \tag{B.97}$$

$$\tau_4 = -(-g_1 + 6g_2 - 10g_3 + 5g_4)/(g_1 - g_2) \tag{B.98}$$

dove

$$g_r = \begin{cases} \frac{r\Gamma(1+k)\Gamma(r/h)}{h^{1+k}\Gamma(1+k+r/h)}\,, & h > 0 \\ \frac{r\Gamma(1+k)\Gamma(-k-r/h)}{(-h)^{1+k}\Gamma(1-r/h)}\,, & h < 0 \end{cases} \tag{B.99}$$

essendo $\Gamma(.)$ la funzione gamma definita nell'Equazione (B.50).

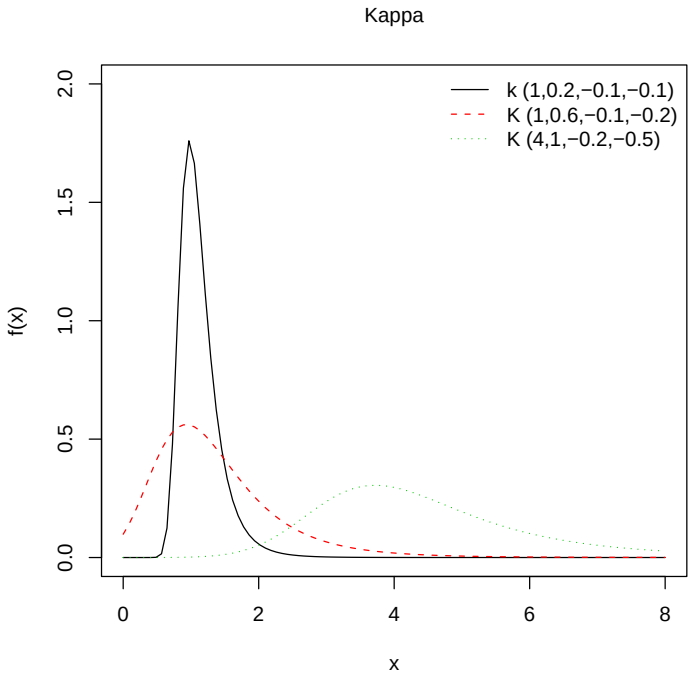

Figura B.9: Esempi di densità di probabilità della distribuzione kappa con parametri: (1) $\xi = 1$, $\alpha = 0.2$, $k = -0.1$, $h = -0.1$; (2) $\xi = 1$, $\alpha = 0.6$, $k = -0.1$, $h = -0.2$; (3) $\xi = 4$, $\alpha = 1$, $k = -0.2$, $h = -0.5$.

Parametri

Per ricavare i parametri a partire dagli L-momenti non ci sono semplici espressioni. Comunque è possibile ricavare k e h in funzione di τ_3 e τ_4 con il metodo numerico di Newton-Raphson a partire dalle Equazioni (B.97) e (B.98).

Elenco delle figure

Elenco delle tabelle

Bibliografia

Ahmad M., Sinclair C., Werritty A., *Log-logistic flood frequency analysis.* Journal of Hydrology (1988), 98, pp. 215–224.

Basson M., Allen R., Pegram G., van Rooyen J., Probabilistic management of water resource and hydropower systems. Water Resources Publications (1994).

Bocchiola D., De Michele C., Rosso R., *Review of recent advances in index flood estimation.* Hydrology and Earth System Sciences (2003), 7 (3), pp. 283–296.

Brath A., Camorani G., Castellarin A., *Una tecnica di stima regionale della curva di durata delle portate in bacini non strumentati.* In *XXIX Convegno di Idraulica e Costruzioni Idrauliche*, volume 2, pp. 391–398, Trento (2004).

Burn D., *Delineation of groups for regional flood frequency analysis.* Journal of Hydrology (1988), 104, pp. 345–361.

Burn D., *Evaluation of regional flood frequency analysis with a region of influence approach.* Water Resources Research (1990), 26, pp. 2257–2265.

Burn D., Goel N., *The formation of groups for regional flood frequency analysis.* Hydrological Sciences - Journal - des Sciences Hydrologiques (2000), 45 (1), pp. 97–112.

Castellarin A., Burn D., Brath A., *Assessing the effectiveness of hydrological similarity measures for flood frequency analysis.* Journal of Hydrology (2001), 241, pp. 270–285.

Chow V., *The log-probability law and its engineering applications.* In *ASCE*, 80, pp. 536–1–536–25 (1954).

Chowdhury J., Stedinger J., Lu L., *Goodness-of-fit tests for regional generalized extreme value flood distributions.* Water Resources Research (1991), 27, pp. 1765–1776.

Claps P., Mancino L., *Impiego di classificazioni climatiche quantitative nell'analisi regionale del deflusso annuo.* In *XXVIII Convegno di Idraulica e Costruzioni Idrauliche*, pp. 169–178, Potenza (2002). 16-19 settembre 2002.

Claps P., Fiorentino M., Silvagni G., *Studio per la valorizzazione e la salvaguardia delle risorse idriche in basilicata. valutazione delle risorse idriche e possibilità di regolazione dei deflussi.* Regione Basilicata (1998).

Conover W., Johnson M., Johnson M., *A comparative study for homogeneity of variances, with applications to the outer cantinental shelf bidding data.* Technometrics (1981), 23 (4), pp. 351–361.

Cox D., Hinkley D., Theoretical statistics. Chapman and Hall, London (1974).

Cressie N., Statistics for Spatial Data. Wiley, NY, revised edizione (1993). 900 pp.

Cunnane C., *Methods and merits of regional flood frequency analysis.* Journal of Hydrology (1988), 100, pp. 269–290.

D'Agostino R., Stephens M., (A cura di), Goodness-of-Fit Techniques, capitolo Tests based on EDF statistics. Marcel Dekker, New York (1986).

Dalrymple T., Flood frequency analyses, volume 1543-A di *Water Supply Paper*. U.S. Geological Survey, Reston, Va. (1960).

De Michele C., Rosso R., *A multi-level approach to flood frequency regionalization.* Hydrology and Earth System Sciences (2002), 6 (2), pp. 185–194.

Durbin J., Knott M., *Components of cramér-von mises statistics.* London School of Economics and Political Science (1971), pp. 290–307.

Fabbris L., Statistica multivariata: analisi esplorativa dei dati. Serie di Matematica. McGraw-Hill Libri Italia srl, Milano (1997).

Farquharson F., Green C., Meigh J., Sutcliffe J., *Comparison of flood frequency curves for many different regions of the world* In *Regional Flood Frequency Analysis.* A cura di Singh V., pp. 223–256, Louisiana State University, Baton Rouge, U.S.A (1987). Proceedings of the International Symposium on Flood Frequency and Risk Analyses, 14-17 May 1986.

Ferraresi M., Todini E., Franchini M., *Un metodo per la regionalizzazione dei deflussi medi.* In *XXI Convegno di Idraulica*, L'Aquila (1988).

Fill H., Stedinger J., *Homogeneity tests based upon gumbel distribution and a critical appraisal of darlymple's test.* Journal of Hydrology (1995), 166, pp. 81–105.

Fill H., Stedinger J., *Using regional regression within index flood procedures and an empirical bayesian estimator.* Journal of Hydrology (1998), 210 (1-4), pp. 128–145.

Fiorentino M., Gabriele S., Rossi F., Versace P., *Hierarchical approach for re-*

gional flood frequency analysis In *Regional Flood Frequency Analysis*. A cura di Singh V., pp. 35–49. Reidel, D., Norwell, Mass. (1987).

Furcolo P., Rossi F., Villani P., *Analisi statistica della variabilità spaziale e temporale degli eventi estremi nelle regioni mediterranee*. In *Tempeste mediterranee, valutazione e previsione degli effetti al suolo*, pp. 95–113. CNR Pubbl. n. 1862 (1998). CIMA, Savona, 1996.

Georgopoulos P. G., Seinfeld J. H., *Statistical distributions of air pollutant concentrations*. Environmental Science & Technology (1982), 16 (7), pp. 401A–416A.

Greenwood J., Landwehr J., Matalas N., Wallis J., *Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form*. Water Resources Research (1979), 15, pp. 1049–1054.

Gumbel E., *Ranges and midranges*. Annals of Mathematical Statistics (1944), 15, pp. 414–422.

Gumbel E., Statistics of Extremes. Columbia University Press, New Jork (1958).

Hall M., Minns W., *The classification of hydrologically homogeneous regions*. Hydrological Sciences - Journal - des Sciences Hydrologiques (1999), 44 (5), pp. 693–704.

Hampel F. R., *The influence curve and its role in robust estimation*. Journal of the American Statistical Association (1974), 69 (346), pp. 383–393.

Harrison M., McCabe B., *A test for heteroscedasticity based on ordinary least squares residuals*. Journal of the American Statistical Association (1979), 74, pp. 494–499.

Hosking J., Wallis J., *Some statistics useful in regional frequency analysis*. Water Resources Research (1993), 29 (2), pp. 271–281.

Hosking J., Wallis J., Regional Frequency Analysis: An Approach Based on L-Moments. Cambridge University Press (1997).

Jenkinson A., *The frequency distribution of the annual maximum (or minimum) value of meterological elements*. Quarterly Journal of the Royal Meteorological Society (1955), 81, pp. 158–171. Generalized Extreme Value distribution.

Johnson N., Kotz S., Balakrishnan N., Continuous Univariate Distributions. John Wiley and Sons, inc., International Edition (1995).

Kitanidis P. K., Introduction to Geostatistics. Cambridge University Press (1997). 249 pp.

Kottegoda N., Rosso R., Statistics, probability, and reliability for civil and environmental engineers. McGraw-Hill, International Edition (1998).

Laio F., *Cramer-von mises and anderson-darling goodness of fit tests for extreme value distributions with unknown parameters*. Water Resources Research (2004), 40, pp. W09308, doi:10.1029/2004WR003204.

Legendre P., *Comparison of permutation methods for the partial correlation and partial mantel tests*. Journal of Statistical Computation and Simulation (2000), 67, pp. 37–73.

Lettenmaier D., Wallis J., Wood E., *Effect of regional heterogeneity on flood frequency estimation*. Water Resources Research (1987), 23 (2), pp. 313–323.

Lu L., Stedinger J. R., *Sampling variance of normalized gev/pwm quantile estimators and a regional homogeneity test*. Journal of Hydrology (1992), 138 (1/2), pp. 223–245.

Mantel N., *The detection of desease clustering and a generalized regression approach*. Cancer Research (1967), 27, pp. 209–220.

Mantel N., Valand R., *A technique of nonparametric multivariate analysis*. Biometrics (1970), 26, pp. 547–558.

Montgomery D., Peck E., Vining G., Introduction to linear regression analysis. Wiley, New York (2001).

Oksanen J., Kindt R., O'Hara R., vegan: Community Ecology Package (2005). R package version 1.6-10.

Pickands J., *Statistical inference using extreme order statistics*. Annuals of Statistics (1975), 3, pp. 491–496.

R Development Core Team, R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2006). ISBN 3-900051-07-0.

Regione Piemonte, *Piano di tutela delle acque*. Direzione Pianificazione Risorse Idriche (2004).

Robson A., Reed D., *Statistical procedures for flood frequency estimation*. In *Flood Estimation Handbook*, volume 3. Institute of Hydrology Crowmarsh Gifford, Wallingford, Oxfordshire (1999).

Scholz F., Stephens M., *K-sample anderson-darling tests*. Journal of American Statistical Association (1987), 82 (399), pp. 918–924.

Shu C., Burn D., *Artificial neural network ensembles and their application in pooled flood frequency analysis*. Water Resources Research (2004a), 40, pp. W09301, doi:10.1029/2003WR002816.

Shu C., Burn D., *Homogeneous pooling group delineation for flood frequency analysis using a fuzzy expert system with genetic enhancement*. Journal of

Hydrology (2004b), 291, pp. 132–149.

Sivapalan M., Takeuchi K., Franks S., Gupta V., Karambiri H., Lakshmi V., Liang X., McDonnell J., Mendiondo E., O'Connell P., Oki T., Pomeroy J., Schertzer D., Uhlenbrook S., Zehe E., *IAHS decade on predictions in ungauged basins (pub), 2003-2012: Shaping an exciting future for the hydrological sciences.* Hydrological Sciences - Journal - des Sciences Hydrologiques (2003), 48 (6), pp. 857–880.

Smouse P. E., Long J. C., Sokal R. R., *Multiple regression and correlation extensions of the mantel test of matrix correspondence.* Systematic Zoology (1986), 35 (4), pp. 627–632.

Stedinger J., Lu L., *Appraisal of regional and index flood quantile estimators.* Stochastic Hydrology and Hydraulics (1995), 9 (1), pp. 49–75.

Sveinsson G., Boes D., Salas J., *Population index flood method for regional frequency analysis.* Water Resources Research (2001), 37 (11), pp. 2733–2748.

Viglione A., homtest: Homogeneity tests for Regional Frequency Analysis (2006). R package version 0.1-4.

Viglione A., *Valutazione delle risorse idriche utilizzabili per obiettivi multipli attraverso la realizzazione di alcuni grandi invasi artificiali in piemonte.* Regione Piemonte (2007). In preparazione.

Viglione A., Claps P., Laio F., *Utilizzo di criteri di prossimità nell'analisi regionale del deflusso annuo.* In *XXX Convegno di Idraulica e Costruzioni Idrauliche* (2006). Roma, 11-16 Settembre 2006.

Viglione A., Laio F., Claps P., *A comparison of homogeneity tests for regional frequency analysis.* Water Resources Research (2007a). In press.

Viglione A., Claps P., Laio F., *Mean annual runoff estimation in north-western italy* In *Water resources assessment and management under water scarcity scenarios.* A cura di La Loggia G. CDSU, Milano (2007b). In press.

Vogel R., Wilson I., *Probability distribution of annual maximum, mean, and minimum streamflows in the united states.* Journal of Hydrologic Engineering (1996), 1 (2), pp. 69–76.

Ward J., *Hierarchical grouping to optimize an objective function.* Journal of the American Statistical Association (1963), 58, pp. 236–244.

Wiltshire S., *Identification of homogeneous regions for flood frequency analysis.* Journal of Hydrology (1986a), 84, pp. 287–302.

Wiltshire S., *Regional flood frequency analysis i: Homogeneity statistics.* Hydrological Sciences - Journal - des Sciences Hydrologiques (1986b), 31,

pp. 321–333.

Wiltshire S., *Regional flood frequency analysis ii: Multivariate classification of drainage basins in britain*. Hydrological Sciences - Journal - des Sciences Hydrologiques (1986c), 31, pp. 335–346.