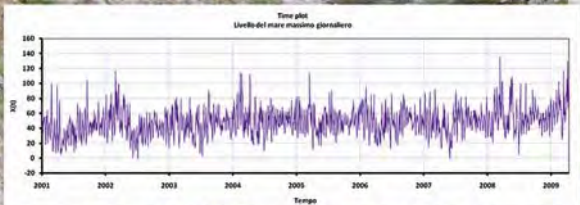
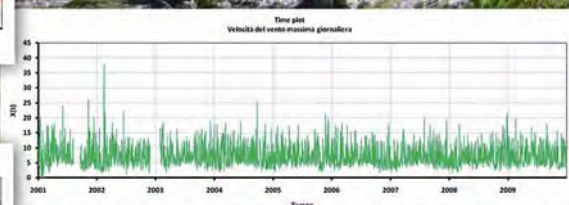
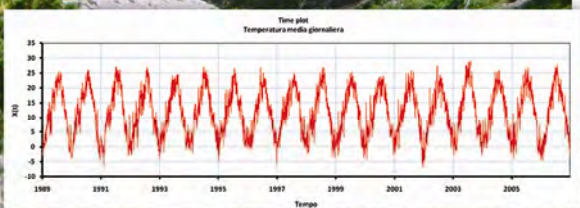
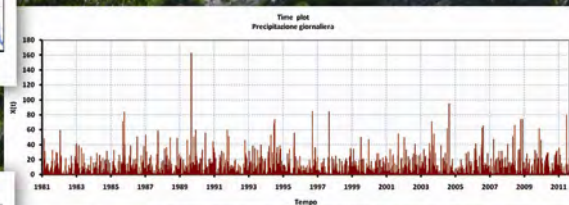
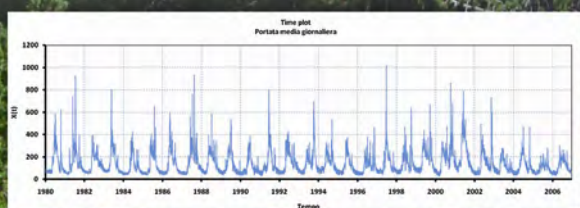




ISPRA

Istituto Superiore per la Protezione
e la Ricerca Ambientale

Linee guida per l'analisi e l'elaborazione statistica di base delle serie storiche di dati idrologici



MANUALI E LINEE GUIDA



ISPRA

Istituto Superiore per la Protezione
e la Ricerca Ambientale

Linee guida per l'analisi e l'elaborazione statistica di base delle serie storiche di dati idrologici

L'Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA) e le persone che agiscono per conto dell'Istituto non sono responsabili per l'uso che può essere fatto delle informazioni contenute in questo manuale.

ISPRA - Istituto Superiore per la Protezione e la Ricerca Ambientale
Via Vitaliano Brancati, 48 – 00144 Roma
www.isprambiente.gov.it

ISPRA, Manuali e Linee Guida 84/13
ISBN 978-88-448-0584-5

Riproduzione autorizzata citando la fonte

Elaborazione grafica
ISPRA

Grafica di copertina: Franco Iozzoli
Foto di copertina: Martina Bussetini

Coordinamento editoriale:
Daria Mazzella
ISPRA – Settore Editoria

Gennaio 2013

Autori

Giovanni Braca, Martina Bussetini, Barbara Lastoria, Stefano Mariani
ISPRA – Dipartimento Tutela delle Acque Interne e Marine – Servizio Monitoraggio e Idrologia delle
Acque Interne – Settore Idrologia

Corresponding author

Giovanni Braca email: giovanni.braca@isprambiente.it

Ringraziamenti

Si ringraziano il prof. Salvatore Grimaldi dell'Università della Tuscia (Viterbo) e l'ing. Francesco Serinaldi dell'Università "La Sapienza" di Roma per lo studio a base delle presenti Linee Guida.

Si ringrazia il prof. Francesco Napolitano dell'Università "La Sapienza" di Roma per gli utili commenti allo studio a base delle presenti Linee Guida.

Indice

1. INTRODUZIONE	1
1.1 Oggetto delle Linee Guida	1
1.2 Obiettivo delle Linee Guida	1
1.3 Struttura delle Linee Guida	2
1.4 Target delle Linee Guida	3
2. I DATI IDROLOGICI	4
2.1 Principali enti che raccolgono e diffondono serie di dati idrologici	4
2.1.1 ISPRA (ex SIMN)	5
2.1.1.1 Banche dati Annali ex SIMN	5
2.1.1.2 Banca dati mareografica e ondametria	7
2.1.1.3 Banca dati della laguna di Venezia	8
2.1.1.4 Banca dati climatica SCIA	8
2.1.2 Centri Funzionali	9
2.1.3 CRA - CMA (ex UCEA)	9
2.1.4 CNMCA e Servizio Meteorologico dell'Aeronautica	10
2.1.5 Servizi Meteorologici e AgroMeteorologici Regionali	10
2.1.6 Elenco di siti web per l'accesso ai dati meteo-idrologici e agrometeorologici	10
2.2 Grandezze idrologiche d'interesse	12
2.2.1 Temperatura dell'aria	12
2.2.2 Precipitazione	13
2.2.3 Altezza idrometrica, livello freatico, livello mareografico	15
2.2.4 Portata liquida	15
2.2.5 Trasporto solido al fondo e in sospensione	16
2.2.6 Pressione atmosferica	17
2.2.7 Umidità relativa dell'aria	17
2.2.8 Direzione e velocità del vento	17
2.2.9 Evaporazione ed evapotraspirazione	18
2.2.10 Radiazione solare a suolo	18
2.2.11 Eliofania	18
3. ANAGRAFICA E METADATI	19
4. FOGLIO ANÁBASI	21
5. CARATTERIZZAZIONE STATISTICA DI UNA SERIE STORICA	22
5.1 Descrizione statistica	22
5.1.1 Lunghezza, frequenza e numero di dati della serie	22
5.1.2 Continuità e completezza	23
5.1.3 Indici di posizione	23
5.1.3.1 Media, moda e mediana	23
5.1.3.2 Quantili, percentili, quartili, minimo e massimo	24
5.1.4 Indici di dispersione	24
5.1.4.1 Varianza, scarto quadratico, range, coefficiente di variazione	24
5.1.5 Indici di forma	25
5.1.5.1 Asimmetria e curtosi	25

5.2	Analisi esplorativa	25
5.2.1	Time plot	26
5.2.2	Distribuzione frequenza campionaria	26
5.2.3	Distribuzione frequenza cumulata campionaria	27
5.2.4	Box plot	28
5.3	Outlier e robustezza delle statistiche di una serie	29
5.4	Trattamento dei dati mancanti	29
6.	QUALITÀ DELLA SERIE	31
7.	ANALISI STATISTICA DI BASE	33
7.1	Autocorrelazione	33
7.1.1	Test Ljung-Box per la presenza di autocorrelazione	35
7.2	Verifica della “normalità” dei dati	35
7.3	Analisi della “lunga memoria”	36
7.4	Stazionarietà	37
7.4.1	Componenti di una serie storica	38
7.4.2	Serie destagionalizzata	38
7.4.2.1	Considerazioni sui dati mancanti	40
7.4.3	Change point	40
7.4.3.1	Test di Pettitt	40
7.4.3.2	Test CUSUM	41
7.4.3.3	Esempio di applicazione dei test per il <i>change point</i>	41
7.4.3.4	Considerazioni sui dati mancanti	43
7.4.4	Trend	43
7.4.4.1	Test di Mann-Kendall	44
7.4.4.2	Test di Pearson	44
7.4.4.3	Test di Spearman	44
7.4.4.4	Esempio di applicazione dei test per il <i>trend</i>	44
7.4.4.5	Considerazioni sui dati mancanti	45
7.5	Analisi dei valori estremi	47
7.5.1	Serie AM	49
7.5.2	Serie POT/PDS	51
7.5.3	Scelta del valore di soglia della GPD	54
7.5.4	Incertezza sulla stima dei parametri della GEV e della GPD	57
7.5.5	Scelta dell’approccio AM o POT	58
7.5.6	Relazione tra la GEV e la GPD	59
7.5.7	Considerazioni sui dati mancanti	59
7.5.8	Ulteriori approfondimenti	60
8.	SCHEMA PROCEDURA ED ESEMPI DI APPLICAZIONE	61
8.1	Schema procedura	62
8.2	Analisi di una serie di portate	63
8.2.1	Scala giornaliera	63
8.2.1.1	Metadato e caratteristiche	63
8.2.1.2	Descrizione statistica e grafici standard	64
8.2.1.3	Analisi di stazionarietà	65
8.2.1.4	Analisi degli estremi	68

8.2.2	Scala mensile	75
8.2.2.1	Metadato e caratteristiche	75
8.2.2.2	Descrizione statistica e grafici standard	76
8.2.2.3	Analisi di stazionarietà	77
8.2.3	Scala annuale	80
8.2.3.1	Metadato e caratteristiche	80
8.2.3.2	Descrizione statistica e grafici standard	81
8.2.3.3	Analisi di normalità	82
8.2.3.4	Analisi di stazionarietà	83
8.2.3.5	Analisi degli estremi	85
9.	BIBLIOGRAFIA	88
10.	APPENDICE A	89
10.1	Scheda A: metadati per le serie idrologiche	89
10.2	Scheda B: descrizione statistica	90
10.3	Scheda C: statistiche di base	90
10.4	Scheda D: analisi di stazionarietà	90
10.5	Scheda E: analisi degli estremi	91
11.	APPENDICE B. APPROFONDIMENTI DI STATISTICA	92
11.1	Definizioni e concetti di base di probabilità e statistica	92
11.1.1	Elementi di probabilità	92
11.1.1.1	Definizione di evento e principali relazioni tra eventi	92
11.1.1.2	Definizione di probabilità	92
11.1.1.3	I tre assiomi della teoria della probabilità	92
11.1.2	Variabili casuali, leggi di probabilità, momenti	94
11.1.3	Elementi di statistica descrittiva e analisi esplorativa dei dati	98
11.1.3.1	Misure di posizione	98
11.1.3.2	Quantili	99
11.1.3.3	Misure di variabilità	100
11.1.3.4	Misure di forma	101
11.1.3.5	Coefficiente di correlazione lineare di Pearson e coefficiente di correlazione di rango di Kendall	103
11.1.3.6	Alcuni metodi grafici di indagine	104
11.1.4	Definizioni di stazionarietà e di funzione di autocorrelazione	107
11.1.5	Test delle ipotesi	110
11.1.6	Tecniche di ricampionamento di tipo "bootstrap"	111
11.2	Analisi di non stazionarietà	112
11.2.1	Aspetti generali	113
11.2.2	Analisi preliminare dei dati	113
11.2.3	Applicazione dei test statistici	114
11.2.4	Interpretazione dei risultati	115
11.2.5	Stagionalità	115
11.3	Analisi di probabilità di dati idrologici	116
11.3.1	Distribuzioni di probabilità	117
11.3.2	Metodi di stima di parametri	117
11.3.3	Test di adattamento della distribuzione empirica ad una distribuzione teorica	118
11.3.4	Analisi dei valori estremi	119

11.3.4.1	Tempo di ritorno -----	119
11.3.4.2	Definizione del campione per l'analisi dei valori estremi-----	120
11.3.5	Precipitazioni e analisi dei valori estremi: curve Intensità-Durata-Frequenza (IDF) -----	122
11.4	Procedure di analisi -----	125
11.4.1	Calcolo del parametro di Hurst -----	125
11.4.1.1	Metodo della varianza aggregata-----	127
11.4.1.2	Metodo R/S -----	128
11.4.1.3	Metodo di Higuchi -----	129
11.4.2	Componenti stagionali -----	131
11.4.3	Analisi di non-stazionarietà -----	134
11.4.3.1	Test Ljung-Box per la presenza di autocorrelazione-----	134
11.4.3.2	Test per i change point -----	134
11.4.3.2.1	Test di Pettitt -----	134
11.4.3.2.2	Test CUSUM con procedura bootstrap -----	136
11.4.3.3	Test per i trend-----	138
11.4.3.3.1	Test di Mann-Kendall -----	138
11.4.3.3.2	Test di Pearson -----	140
11.4.3.3.3	Test di Spearman -----	140
11.4.4	Analisi degli eventi estremi -----	141
11.4.4.1	Introduzione-----	141
11.4.4.2	Modello per i massimi: distribuzione dei valori estremi generalizzata (GEV)-----	141
11.4.4.3	Modello sopra soglia: distribuzione generalizzata di Pareto (GPD) -----	142
11.4.5	Curve Intensità-Durata-Frequenza (IDF) basate sulle proprietà di scala della precipitazione-----	144
11.5	Bibliografia dell'approfondimento-----	146
11.5.1	Definizioni e concetti di base -----	146
11.5.2	Serie temporali e stazionarietà -----	146
11.5.3	Analisi dei trend e change points -----	147
11.5.4	Analisi degli eventi estremi Curve IDF-----	148

Elenco delle figure

Figura 1.1 - Esempio di alcune diverse tipologie di serie di dati idrologici oggetto delle Linee Guida. 2	
Figura 2.1 - Compartimenti idrografici all'epoca del SIMN.....	5
Figura 2.2 - Pagina di accesso alle serie storiche di precipitazione, temperatura e idrometria http://www.sintai.sinanet.apat.it/	6
Figura 2.3 - Pagina di accesso agli annali idrologici Parte I e Parte II in formato pdf http://www.acq.isprambiente.it/annalipdf/	6
Figura 2.4 - Pagina di accesso a dati storici digitalizzati http://193.206.192.243/storico/	7
Figura 2.5 - Pagina di accesso alla banca dati PLUTER http://www.acq.isprambiente.it/pluter/index.html	7
Figura 2.6 - Pagina di accesso alla banca dati mareografica http://www.mareografico.it/	7
Figura 2.7 - Pagina di accesso alla banca dati della Rete Ondametrica Nazionale (RON) http://www.telemisura.it/	8
Figura 2.8 - Pagina di accesso alla banca dati della laguna di Venezia http://www.venezia.isprambiente.it/ispra/index.php?folder_id=20	8
Figura 2.9 - Pagina di accesso alla banca dati di SCIA(http://www.scia.sinanet.apat.it/#).....	9
Figura 2.10 - Pagina di accesso alla banca dati agrometeorologica del CRA-CMA (http://www.cra-cma.it/homePage.htm).....	10
Figura 2.11 - Pagina di accesso alla banca dati meteorologica del Servizio Meteorologico dell'Aeronautica (http://clima.meteoam.it/RichiestaDatiGenerica.php).....	10
Figura 2.12 - Contenuto delle tabelle termometriche: estratto da Annali Idrologici parte I anno 1951 Compartimento di Napoli.....	13
Figura 2.13 - Contenuto delle tabelle pluviometriche: estratto da Annali Idrologici parte I anno 1951 Compartimento di Napoli.....	14
Figura 2.14 - Contenuto della tabella idrometrica: estratto da Annali Idrologici parte II anno 1951 Compartimento di Napoli.....	15
Figura 2.15 - Contenuto delle tabelle delle misure di portata estratto da Annali Idrologici parte II anno 1951 Compartimento di Napoli.....	16
Figura 4.1 - Pagina iniziale (<Home>) della macro ANÁBASI.....	21
Figura 4.2 - Pagina principale (<Input>) della macro ANÁBASI.....	21
Figura 5.1 - Esempio di time plot per la rappresentazione di una serie di dati.....	26
Figura 5.2 - Esempio di istogramma per la rappresentazione di una serie di dati.....	26
Figura 5.3 - Esempi di istogramma della distribuzione di frequenza percentuale campionaria.....	27
Figura 5.4 - Esempio di distribuzione di frequenza cumulata campionaria.....	28
Figura 5.5 - Esempi di box-plot per la rappresentazione statistica di un set di dati. A destra il box plot con l'indicazione in verde di valori estremi e in rosso di valori anomali.....	28
Figura 7.1 - Esempio di una funzione di autocorrelazione per le portate medie mensili di un singolo mese per ciascun anno (frequenza 1). I valori assoluti del coefficiente di correlazione per lag superiori a 0 sono inferiori a 0.4.	34
Figura 7.2 - Confronto delle funzioni di autocorrelazione di una serie di portate (a destra) e una serie di precipitazioni (a sinistra) a scala giornaliera.....	34
Figura 7.3 - Grafico per la stima del parametro λ della trasformata di Box e Cox.....	36
Figura 7.4 - Serie storica delle portate giornaliere.....	39
Figura 7.5 - Componente stagionale della media (a) e della deviazione standard (b).....	39
Figura 7.6 - Serie storica destagionalizzata delle portate medie giornaliere.....	39
Figura 7.7 - Serie storica della grandezza idrologica X simulata con change point.....	41
Figura 7.8 - Curva della statistica di Pettitt della grandezza idrologica X simulata con change point.	42
Figura 7.9 - Curva CUSUM della grandezza idrologica X simulata con change point.....	42
Figura 7.10 - Serie storica della grandezza idrologica Y simulata senza change point.....	42
Figura 7.11 - Curva della statistica di Pettitt della grandezza idrologica Y simulata senza change point.....	42
Figura 7.12 - Curva CUSUM della grandezza idrologica Y simulata senza change point.....	43
Figura 7.13 - Serie storica della grandezza idrologica X simulata con trend lineare.....	44
Figura 7.14 - Serie storica della grandezza idrologica Y simulata senza trend.....	45

Figura 7.15 - Rappresentazione dei valori estremi.	47
Figura 7.16 - Rappresentazione degli approcci utilizzati nella EVA applicata alla stessa serie di dati. (a) Approccio Block Maximum in cui si considerano i valori massimi (in rosso) per ciascun blocco. (b) Approccio POT in cui si considerano i valori (in rosso) sopra una soglia prefissata.	48
Figura 7.17 - Rappresentazione dei tipi di distribuzioni GEV	50
Figura 7.18 - Schema del livello di ritorno.....	51
Figura 7.19 - Esempio della PDF della GP per diversi valori del parametro di forma.	53
Figura 7.20 - Bilanciamento della scelta della soglia per l'analisi dei valori estremi con l'approccio POT	55
Figura 7.21 - Schema di declustering generale proposto nelle LG per l'approccio POT.....	55
Figura 7.22 - Grafico "mean residual life". In rosso sono riportate le fasce di confidenza al 95%	56
Figura 7.23 - Stabilità del parametro di scala della GPD. In rosso sono riportate le fasce di confidenza al 95%.....	57
Figura 7.24 - Stabilità del parametro di forma della GPD. In rosso sono riportate le fasce di confidenza al 95%.....	57
Figura 8.1 - Diagramma di flusso della procedura di analisi di una serie storica di dati idrologici. ..	62
Figura 8.2 - Diagramma cronologico della serie (time plot). Portate medie giornaliere - Adige a Bronzolo.	64
Figura 8.3 - Grafici della frequenza percentuale campionaria e della frequenza cumulata. Portate medie giornaliere - Adige a Bronzolo.	64
Figura 8.4 - Box plot con e senza l'indicazione del VAI e VAS . Portate medie giornaliere - Adige a Bronzolo.	65
Figura 8.5 - Funzione di autocorrelazione per un periodo di 2 anni (a) e per i primi 20 giorni (b). Portate medie giornaliere - Adige a Bronzolo.	66
Figura 8.6 - Componente stagionale della media (a) e della deviazione standard (b). Portate medie giornaliere - Adige a Bronzolo.....	66
Figura 8.7 - Diagramma cronologico della serie destagionalizzata. Portate medie giornaliere - Adige a Bronzolo.	67
Figura 8.8 - Funzione di autocorrelazione della serie destagionalizzata su un periodo di 20 giorni (a), 2 anni (b) e 5 anni (c). Portate medie giornaliere - Adige a Bronzolo.	67
Figura 8.9 - Stima del parametro di Hurst relativo alla serie destagionalizzata $H = 0.85$. Portate medie giornaliere - Adige a Bronzolo.	68
Figura 8.10 - Mean residual life. Portate medie giornaliere - Adige a Bronzolo.	69
Figura 8.11 - Numero di superamenti in funzione del livello di soglia. Portate medie giornaliere - Adige a Bronzolo.	70
Figura 8.12 - Stabilità dei parametri della distribuzione GPD in funzione del livello di soglia. Portate medie giornaliere - Adige a Bronzolo.	70
Figura 8.13 - Serie storica dei valori sopra la soglia di $400 \text{ m}^3/\text{s}$ senza declustering. 185 superamenti e un crossing rate di 6.85. Portate medie giornaliere - Adige a Bronzolo.	70
Figura 8.14 - Serie storica dei valori sopra la soglia di $400 \text{ m}^3/\text{s}$ con declustering: 63 superamenti e un crossing rate di 2.62. Portate medie giornaliere - Adige a Bronzolo.	70
Figura 8.15 - Grafici diagnostici per la GPD. Serie POT senza declustering. Portate medie giornaliere - Adige a Bronzolo.....	71
Figura 8.16 - Grafici diagnostici per la GPD. Serie POT con declustering. Portate medie giornaliere - Adige a Bronzolo.	72
Figura 8.17 - Diagramma cronologico della serie (time plot). Portate medie mensili - Adige a Bronzolo.	76
Figura 8.18 - Grafici della frequenza percentuale campionaria e della frequenza cumulata. Portate medie mensili - Adige a Bronzolo.....	76
Figura 8.19 - Grafici box plot con e senza l'indicazione del VAI e VAS . Portate medie mensili - Adige a Bronzolo.	77
Figura 8.20 - Funzione di autocorrelazione per un periodo di 5 anni (a) e per i primi 20 mesi (b). Portate medie mensili - Adige a Bronzolo.....	77
Figura 8.21 - Componente stagionale della media (a) e della deviazione standard (b). Portate medie mensili - Adige a Bronzolo.	78
Figura 8.22 - Diagramma cronologico della serie destagionalizzata. Portate medie mensili - Adige a Bronzolo.	78
Figura 8.23 - Funzione di autocorrelazione della serie destagionalizzata su un periodo di 60 mesi (a), 20 mesi (b). Portate medie mensili - Adige a Bronzolo.	79

Figura 8.24 - <i>Diagramma cronologico della serie (time plot). Portate giornaliere massime annuali - Adige a Bronzolo.</i>	81
Figura 8.25 - <i>Grafici della frequenza percentuale campionaria e della frequenza cumulata. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	81
Figura 8.26 - <i>Box plot in cui il VAI e VAS coincidono rispettivamente con il minimo e massimo. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	82
Figura 8.27 - <i>QQ plot Normali per la verifica della normalità dei dati e della loro trasformata Box e Cox. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	82
Figura 8.28 - <i>Diagramma del coefficiente della trasformata di Box e Cox in funzione dell'asimmetria dei dati. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	83
Figura 8.29 - <i>Funzione di autocorrelazione della serie 15 lag. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	83
Figura 8.30 - <i>Time plot con indicazione della media. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	84
Figura 8.31 - <i>Curva CUSUM con bootstrap (1000 campioni). Portate giornaliere massime annuali - Adige a Bronzolo.</i>	84
Figura 8.32 - <i>Distribuzione della statistica S_{diff} nel test CUSUM con bootstrap (1000 campioni). Portate giornaliere massime annuali - Adige a Bronzolo.</i>	84
Figura 8.33 - <i>Curva della statistica di Pettitt. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	84
Figura 8.34 - <i>Time plot con indicazione della tendenza lineare. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	85
Figura 8.35 - <i>Grafici diagnostici per la GEV. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	86
Figura 11.1 - <i>Esempio di due campioni distribuiti secondo una legge Gaussiana con due diverse medie e uguale varianza.</i>	99
Figura 11.2 - <i>Esempio di due campioni distribuiti secondo una legge gaussiana con uguale media e diversa varianza.</i>	100
Figura 11.3 - <i>Esempi di distribuzioni con le tre configurazioni possibili in termini di simmetria.</i>	102
Figura 11.4 - <i>Esempi di distribuzioni leptokurtica e platikurtica rispetto alla distribuzione normale.</i>	102
Figura 11.5 - <i>Esempi di campioni con diversi valori del coefficiente di correlazione lineare di Pearson.</i>	104
Figura 11.6 - <i>Esempi di alcuni grafici diagnostici.</i>	106
Figura 11.7 - <i>Illustrazione dei passi dell'algoritmo LOESS applicato allo scatter-plot mostrato in Figura 11.6.</i>	106
Figura 11.8 - <i>La figura riporta tre possibili risultati ottenibili nell'applicazione di un qq-plot.</i>	107
Figura 11.9 - <i>Esempi di serie temporali con diverse forme di dipendenza temporale e relative ACF. Il simbolo ϵ indica un termine casuale con distribuzione normale standard. Le linee tratteggiate blu nei grafici delle ACF indicano gli intervalli di confidenza al 95% dei coefficienti di autocorrelazione nell'ipotesi di autocorrelazione nulla.</i>	109
Figura 11.10 - <i>Esempi di tipologie di variabilità sistematica che possono essere presenti in una serie temporale.</i>	112
Figura 11.11 - <i>Esempio di decomposizione additiva (eq. 11.2.3).</i>	116
Figura 11.12 - <i>Esempio di curve di caso critico riferite a dati riportati in Tabella 11.2B. Il grafico in scala bi-logaritmica evidenzia un andamento approssimativamente lineare.</i>	123
Figura 11.13 - <i>Esempio di curve IDF. Ogni curva interpola i valori di intensità di pioggia corrispondenti a una fissata probabilità di non superamento p, calcolati tramite distribuzioni parametriche stimate sui dati relativi ad ogni durata disponibile.</i>	125
Figura 11.14 - <i>Serie delle portate annuali del Nilo ad Assuan (linea tratteggiata) e serie degli scarti cumulati (linea continua). Il range è indicato con R.</i>	126
Figura 11.15 - <i>Metodo della varianza aggregata: applicazione ad una serie simulata di numerosità 10000 con $H = 0.5$ (pannello a sinistra) e $H = 0.75$ (pannello a destra). In entrambi i pannelli, la stima della pendenza della retta di regressione (linea continua) è stata condotta con i punti compresi nell'intervallo temporale logaritmico (0.5-2.5). Le linee tratteggiate indicano le rette con pendenze corrispondenti ad $H = 0.5$ (assenza di lunga memoria).</i>	128
Figura 11.16 - <i>Metodo R/S: applicazione ad una serie simulata di numerosità 10000 con $H = 0.5$ (pannello a sinistra) e $H = 0.75$ (pannello a destra). In entrambi i pannelli, la stima della pendenza della retta di regressione (linea continua) è stata condotta con i punti compresi nell'intervallo</i>	

<i>temporale logaritmico (1.0-3.0). Le linee tratteggiate indicano le rette con pendenze corrispondenti ad $H = 0.5$ (assenza di lunga memoria).</i>	129
Figura 11.17 - Metodo di Higuchi: applicazione ad una serie simulata di numerosità 10000 con $H = 0.5$ (pannello a sinistra) e $H = 0.75$ (pannello a destra). In entrambi i pannelli, la stima della pendenza della retta di regressione (linea continua) è stata condotta con i punti compresi nell'intervallo temporale logaritmico (1.0-3.0). Le linee tratteggiate indicano le rette con pendenze corrispondenti ad $H = 0.5$ (assenza di lunga memoria).	130
Figura 11.18 - Box-plot dei valori del parametro di Hurst calcolato su 200 serie simulate con $H = 0.75$ (linea grigia) con i tre metodi della varianza aggregata (VA), del rescaled range (R/S) e di Higuchi. Ogni metodo è applicato alla serie completa ("serie"), alla serie privata del 10% dei dati selezionati in modo casuale ("random") ed alla serie in cui il 10% dei dati è stato sottratto come un intervallo continuo in una posizione casuale all'interno della serie ("blocchi").	131
Figura 11.19 - Il grafico riporta le osservazioni corrispondenti ad ogni giorno dell'anno estratte da una serie giornaliera di portate di 27 anni (per ogni giorno sono riportati 27 valori). Le curve rappresentano le componenti stagionali della media e della deviazione standard ottenute con il metodo classico e con l'algoritmo di Grimaldi (2004).	132
Figura 11.20 - Esempio di serie con e senza change point e relative serie della statistica $U_{t,T}$ usata per il calcolo della statistica test di Pettitt.	136
Figura 11.21 - Esempio di serie con e senza change point e relative curve CUSUM.	137
Figura 11.22 - Distribuzioni bootstrap della statistica test S_{diff} ottenute tramite ricampionamento delle due serie con e senza change point riportate in Figura 11.21 a-b. Le linee rosse verticali indicano il valore della statistica test calcolata sulle due serie originali.	138
Figura 11.23 - Esempi di serie simulate con e senza trend monotono.	139
Figura 11.24 - Grafico "mean residual life" e grafici dei parametri di scala e forma in funzione della soglia per la serie simulata descritta nel testo.	144

Elenco delle tabelle

Tabella 2.1 - <i>Enti e relativi web-link per l'accesso ai dati meteo-idrologici e agrometeorologici</i>	11
Tabella 5.1 - <i>Frequenza di una serie storica</i>	22
Tabella 6.1 - <i>Coefficienti per la determinazione dell'indice di qualità della serie</i>	31
Tabella 6.2 - <i>Indice di qualità della serie iQuaSI</i>	32
Tabella 7.1 - <i>Esito dei test per il "change point detection". Grandezza idrologica X simulata con change point</i>	41
Tabella 7.2 - <i>Esito dei test per il "change point detection". Grandezza idrologica Y simulata senza change point</i>	42
Tabella 7.3 - <i>Esito dei test per il "trend detection". Grandezza X simulata con trend lineare</i>	44
Tabella 7.4 - <i>Esito dei test per il "trend detection". Grandezza Y simulata senza trend</i>	45
Tabella 7.5 - <i>PRO e CONTROLLO degli approcci BM e POT</i>	58
Tabella 8.1 - <i>Metadati della serie storica. Portate medie giornaliere - Adige a Bronzolo</i>	63
Tabella 8.2 - <i>Caratteristiche quantitative della serie. Portate medie giornaliere - Adige a Bronzolo</i>	64
Tabella 8.3 - <i>Statistiche univariate campionarie per la descrizione sintetica della serie. Portate medie giornaliere - Adige a Bronzolo</i>	65
Tabella 8.4 - <i>Esito dei test di autocorrelazione (test sui primi 10 lag). Portate medie giornaliere - Adige a Bronzolo</i>	66
Tabella 8.5 - <i>Esito dei test di autocorrelazione della serie destagionalizzata (test sui primi 5 lag). Portate medie giornaliere - Adige a Bronzolo</i>	68
Tabella 8.6 - <i>Tabella di sintesi dell'analisi di stazionarietà. Portate medie giornaliere - Adige a Bronzolo</i>	68
Tabella 8.7 - <i>Stima dei parametri della distribuzione generalizzata di Pareto con diversi metodi e relativi standard error. Serie POT senza declustering. Portate medie giornaliere - Adige a Bronzolo</i>	73
Tabella 8.8 - <i>Stima dei parametri della distribuzione generalizzata di Pareto con diversi metodi e relativi standard error. Serie POT con declustering. Portate medie giornaliere - Adige a Bronzolo</i>	73
Tabella 8.9 - <i>Quantili corrispondenti a tempi di ritorno notevoli. Serie POT senza declustering. Portate medie giornaliere - Adige a Bronzolo</i>	73
Tabella 8.10 - <i>Quantili corrispondenti a tempi di ritorno notevoli. Serie POT con declustering. Portate medie giornaliere - Adige a Bronzolo</i>	73
Tabella 8.11 - <i>Sintesi dell'analisi degli estremi. Portate medie giornaliere - Adige a Bronzolo</i>	74
Tabella 8.12 - <i>Metadato della serie storica. Portate medie mensili - Adige a Bronzolo</i>	75
Tabella 8.13 - <i>Caratteristiche della serie. Portate medie mensili - Adige a Bronzolo</i>	76
Tabella 8.14 - <i>Statistiche univariate campionarie per la descrizione sintetica della serie. Portate medie mensili - Adige a Bronzolo</i>	77
Tabella 8.15 - <i>Esito dei test di autocorrelazione (test sui primi 10 lag). Portate medie mensili - Adige a Bronzolo</i>	78
Tabella 8.16 - <i>Esito dei test di autocorrelazione (test sui primi 10 lag) della serie destagionalizzata. Portate medie mensili - Adige a Bronzolo</i>	79
Tabella 8.17 - <i>Tabella di sintesi dell'analisi di stazionarietà. Portate medie mensili - Adige a Bronzolo</i>	79
Tabella 8.18 - <i>Metadati della serie storica. Portate giornaliere massime annuali - Adige a Bronzolo</i>	80
Tabella 8.19 - <i>Caratteristiche della serie. Portate giornaliere massime annuali - Adige a Bronzolo</i>	81
Tabella 8.20 - <i>Statistiche univariate campionarie per la descrizione sintetica della serie. Portate giornaliere massime annuali - Adige a Bronzolo</i>	82
Tabella 8.21 - <i>Esito del test di normalità dei dati. Portate giornaliere massime annuali - Adige a Bronzolo</i>	83
Tabella 8.22 - <i>Esito del test di normalità dei dati trasformati con la trasformata di Box e Cox con $\lambda = 0.53$. Portate giornaliere massime annuali - Adige a Bronzolo</i>	83
Tabella 8.23 - <i>Esito dei test di autocorrelazione (test sui primi 5 lag). Portate giornaliere massime annuali - Adige a Bronzolo</i>	84
Tabella 8.24 - <i>Esito dei test per il "change point detection". Portate giornaliere massime annuali - Adige a Bronzolo</i>	84
Tabella 8.25 - <i>Esito dei test per il "trend detection". Portate giornaliere massime annuali - Adige a Bronzolo</i>	85

Tabella 8.26 - <i>Tabella di sintesi dell'analisi di stazionarietà. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	85
Tabella 8.27 - <i>Stima dei parametri della distribuzione GEV con il MoM e il PWM. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	87
Tabella 8.28 - <i>Quantili corrispondenti a tempi di ritorno notevoli elaborati con la distribuzione GEV con parametri PWM. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	87
Tabella 8.29 - <i>Tabella di sintesi dell'analisi degli estremi. Portate giornaliere massime annuali - Adige a Bronzolo.</i>	87
Tabella 11.1 - <i>Tempi di ritorno T_p relativi alle serie PDS e corrispondenti valori T relativi alle serie AM ottenuti dalla relazione eq. 11.3.11</i>	122
Tabella 11.2 - <i>Valori massimi annuali di intensità di pioggia riferiti a diverse durate d.</i>	123

Elenco degli acronimi e abbreviazioni

<i>AM:</i>	<i>Annual Maximum</i>
<i>ANPA:</i>	<i>Agenzia Nazionale per la Protezione dell'Ambiente</i>
<i>ARPA:</i>	<i>Agenzia Regionale per la Protezione dell'Ambiente</i>
<i>APAT:</i>	<i>Agenzia per la Protezione dell'Ambiente e per i Servizi Tecnici</i>
<i>BM:</i>	<i>Block Maximum</i>
<i>CFC:</i>	<i>Centro Funzionale Centrale</i>
<i>CFD:</i>	<i>Centri Funzionali Decentrati</i>
<i>CDF:</i>	<i>Cumulative Distribution Function</i>
<i>CNMCA:</i>	<i>Centro Nazionale di Meteorologia e Climatologia Aeronautica</i>
<i>CRA:</i>	<i>Consiglio per la Ricerca e la Sperimentazione in Agricoltura</i>
<i>CMA:</i>	<i>Unità di Ricerca per la Climatologia e la Meteorologia Applicate all'Agricoltura</i>
<i>DSTN:</i>	<i>Dipartimento per i Servizi Tecnici Nazionali</i>
<i>DPC:</i>	<i>Dipartimento della Protezione Civile</i>
<i>EVA:</i>	<i>Extreme Value Analysis</i>
<i>EVT:</i>	<i>Extreme Value Theory</i>
<i>iid:</i>	<i>indipendenti e identicamente distribuite / independent and identically distributed</i>
<i>ISPRA:</i>	<i>Istituto Superiore per la Protezione e la Ricerca Ambientale</i>
<i>LG:</i>	<i>Linee Guida</i>
<i>MATM:</i>	<i>Ministero per l'Ambiente e per la Tutela del Territorio e del Mare</i>
<i>MiPAF:</i>	<i>Ministero delle Politiche Agricole e Forestali</i>
<i>ML:</i>	<i>Maximum Likelihood</i>
<i>MoM:</i>	<i>Method of Moments</i>
<i>OMM:</i>	<i>Organizzazione Meteorologica Mondiale (vedi WMO)</i>
<i>PCM:</i>	<i>Presidenza del Consiglio dei Ministri</i>
<i>PDF:</i>	<i>Probability Density Function</i>
<i>PDS:</i>	<i>Partial Duration Series</i>
<i>POT:</i>	<i>Peak Over Treshold</i>
<i>PWM:</i>	<i>Probability Weighted Moment</i>
<i>SII:</i>	<i>Servizio Idrografico Italiano</i>
<i>SIMN:</i>	<i>Servizio Idrografico e Mareografico Nazionale</i>
<i>SINA:</i>	<i>Sistema Informativo Nazionale Ambientale</i>
<i>SINTAI:</i>	<i>Sistema Informativo Nazionale per la Tutela delle Acque Italiane</i>
<i>RMN:</i>	<i>Rete Mareografica Nazionale</i>
<i>RNDT:</i>	<i>Repertorio Nazionale dei Dati Territoriali</i>
<i>RON:</i>	<i>Rete Ondametrica Nazionale</i>
<i>UCEA:</i>	<i>Ufficio Centrale di Ecologia Agraria</i>
<i>WMO:</i>	<i>World Meteorological Organization (vedi OMM)</i>

Premessa

L'Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA) ha istituzionalmente il compito di definire uno standard metodologico per l'elaborazione dei dati idrologici, avendo ricevuto, per effetto dell'art. 28 della L.133/2008, tutte le attribuzioni dell'Agenzia per la Protezione dell'Ambiente e i Servizi Tecnici (APAT), che, a sua volta, era stata costituita in base all'art.38 del DLgs 300/1999 dalla fusione del Dipartimento per i Servizi Tecnici della Presidenza del Consiglio dei Ministri (DSTN), e dell'Agenzia Nazionale per la Protezione dell'Ambiente (ANPA).

Tale compito deriva da un dettato normativo che ha origine dall'art.22 del DPR 85/91 dove al Servizio Idrografico e Mareografico Nazionale (SIMN) veniva assegnato, tra gli altri, il compito "di provvedere al rilievo sistematico ed alla elaborazione delle grandezze relative al clima marittimo, allo stato dei litorali ed ai livelli marini, di provvedere alla pubblicazione sistematica degli elementi osservati ed elaborati e di cartografia e di predisporre criteri, metodi e standard di raccolta, analisi e consultazione dei dati relativi all'attività conoscitiva svolta".

Tale compito veniva successivamente ribadito nell'ambito della Conferenza Stato – Regioni nella seduta del 24 maggio 2001 che ha avuto per oggetto l'accordo tra il Governo e le Regioni ai fini dell'attuazione dell'art. 92, comma 4 del DLgs 31 marzo 1998, n. 112, concernente il trasferimento alle Regioni degli uffici periferici del DSTN – SIMN, nel cui verbale veniva, tra le altre cose, concordato che "... 9) *Per l'esercizio dei compiti di rilievo nazionale di cui agli artt. 2 e 9, comma 4, della legge 18 maggio 1989, n. 183 e dell'art. 88 del decreto legislativo 31 marzo 1998, n. 112, le regioni debbono assicurare la trasmissione al Dipartimento per i Servizi Tecnici Nazionali dei dati rilevati sia dalle stazioni di rilevamento locale che in telemisura; inoltre sono stipulati accordi tra le regioni e il Dipartimento per i Servizi Tecnici Nazionali, aventi per oggetto: a) la standardizzazione dei criteri, metodi e standard di raccolta, elaborazione e consultazione dei dati relativi all'attività conoscitiva e di gestione e manutenzione delle reti di monitoraggio;...*"

Il presente rapporto metodologico costituisce pertanto il primo di una serie di Linee Guida che ISPRA produrrà e che potranno costituire una base di discussione per pervenire a procedure di analisi ed elaborazione statistica dei dati idrologici concordate e condivise con le strutture regionali competenti.

Le presenti Linee Guida sono in continua evoluzione e aggiornamento con l'aggiunta, ove opportuno, di ulteriori procedure e metodi di analisi.

Le presenti Linee Guida fanno riferimento allo studio, commissionato da ISPRA al Dipartimento GEMINI dell'Università della Tuscia e coordinato dal prof. Salvatore Grimaldi, avente per oggetto: "*Studio di metodologie di analisi statistica di base di serie storiche di dati idrologici a diverse scale di aggregazione, finalizzato alla definizione di 'Linee Guida' sull'elaborazione di dati idrologici*"

Il documento è strutturato, per quanto possibile, secondo le specifiche GLISC 1.1 2007 "*Guidelines for the production of scientific and technical reports: how to write and distribute grey literature*" Version 1.1 July 2007 (<http://www.glisc.info/>).

Sommario

Il presente documento propone uno standard metodologico per la caratterizzazione delle serie storiche di dati idrologici. Vengono individuati e descritti un set di parametri, di test e procedure statistiche al fine di uniformare, a livello nazionale, le informazioni minime necessarie per un'efficace elaborazione, una corretta interpretazione e una uniforme diffusione dei dati idrologici e dei risultati delle loro elaborazioni.

La necessità di definire uno standard nazionale sull'elaborazione dei dati idrologici deriva, oltre che da un dettato normativo, da una reale esigenza di uniformità delle stesse, sempre più avvertita a seguito del grande sviluppo delle possibilità di accesso e di scambio di informazioni idrologiche.

La standardizzazione delle procedure di analisi ed elaborazione statistica ha il principale obiettivo di rendere i dati e i risultati facilmente confrontabili.

Parole chiave: *serie storica, dati idrologici, metadato idrologico, qualità della serie, analisi univariata, valori estremi, non-stazionarietà, tempo di ritorno.*

Abstract

The present document proposes a methodological standard for the characterization of hydrological time series. A set of statistical parameters, tests and procedures are identified and described in order to standardize, at national level, the minimum set of essential information for an effective elaboration, a correct interpretation and uniform dissemination of hydrological data and results.

The need of a national standard for hydrological data elaboration and presentation comes not only from normative requirements, but also from a real demand of a standardized hydrological analysis, because of huge amount of hydrological information widely available nowadays.

The standardization of the statistical analysis procedures allows data and the elaboration results to be easily comparable, as well.

Keywords: *time series, hydrological data, hydrologic metadata, time series quality, extreme value statistics, univariate analysis, non stationarity, return period.*

1. Introduzione

1.1 Oggetto delle Linee Guida

Oggetto delle presenti Linee Guida (d'ora in avanti abbreviate con **LG**) è la “serie storica” o “serie temporale” di dati idrologici (*hydrological time serie*) intesa come un insieme di dati costituiti da una sequenza ordinata di osservazioni relative a un determinato fenomeno idrologico d'interesse effettuate in istanti di tempo consecutivi (come ad esempio la temperatura dell'aria o dell'acqua) e di norma, anche se non necessariamente, equispaziati, ovvero su intervalli (come ad esempio la precipitazione) e di norma, anche se non necessariamente, della stessa ampiezza. Nel seguito con il simbolo $t = 1, 2, \dots, N$, si indica il tempo, che costituisce il criterio ordinatore delle osservazioni, e con $\{X_1, X_2, \dots, X_N\}$ o con $\{X_t; t = 1, 2, \dots, N\}$ la sequenza dei dati.

Nella Figura 1.1 sono riportati alcuni esempi di serie storiche che costituiranno oggetto delle LG. Come risulta evidente, pur essendo riferite al medesimo intervallo di campionamento (giornaliero), presentano caratteristiche (*pattern*) profondamente diverse che, come sarà mostrato in seguito, si rifletteranno sulle analisi statistiche.

Non sempre, tuttavia, l'analisi statistica dei dati di una serie è legata alla sua struttura temporale.

In molte analisi, come ad esempio quella di valori estremi, la struttura temporale non è importante; tutte le osservazioni, cioè, si riferiscono al medesimo periodo/istante di tempo. Si parla in questo caso di *cross-sectional data*.

Nel seguito indicheremo, quindi, con “serie di dati idrologici” o “serie storica” l'insieme dei dati per i quali è importante la sequenza temporale. Indicheremo semplicemente con “dati idrologici” quando l'analisi non richiede di specificare la sequenza temporale.

Non appare superfluo ricordare che le serie di dati idrologici costituiscono l'osservazione di fenomeni intrinsecamente aleatori e non deterministici.

Di maggiore interesse, per il loro contenuto d'informazione idrologica, sono ovviamente le serie lunghe di dati (*long time series*), che abbracciano un arco temporale maggiore di 30-50 di anni. Queste, tuttavia, possono presentare, rispetto a serie di dati brevi, problematiche peculiari, quali ad esempio l'omogeneità, che devono essere affrontate in maniera per quanto possibile standard per garantire la confrontabilità e la riproducibilità dei risultati e la correttezza delle interpretazioni. Risultati forniti da metodi differenti possono essere poco confrontabili: questo rende difficoltosa l'interpretazione dei risultati, soprattutto nell'ambito della caratterizzazione della variabilità climatica.

Anche il miglioramento della qualità delle misure in ambito idrologico con l'adozione di strumentazione sempre più tecnologicamente avanzata e accurata, che costituisce tra l'altro un obiettivo strategico del *World Meteorological Organization (WMO)*, comporta una sensibile variazione e/o una discontinuità nelle serie lunghe di dati, tale da richiedere procedure per la sua identificazione ed eventualmente anche procedure per la correzione e l'omogeneizzazione dei dati.

A fronte tuttavia della necessità di standardizzare i metodi e i risultati delle determinazioni, ciò non vuole comunque significare l'adozione del “metodo unico” e che le analisi statistiche dei dati non possano essere eseguite con altri metodi.

1.2 Obiettivo delle Linee Guida

Nel presente documento ci si propone di individuare, nell'ambito della vasta letteratura tecnico-scientifica, alcune procedure da definire come standard per l'elaborazione statistica di base e finalizzata a valutazioni di tipo idrologico (bilancio idrologico, progettazione opere idrauliche, prevenzione rischio idraulico, ecc.), i cui risultati possano essere sintetizzati in una scheda sulle caratteristiche della serie dei dati e sull'individuazione di particolari comportamenti.

In particolare, si individuano metodi e procedure relative a:

- caratteristiche intrinseche della serie;
- struttura di autocorrelazione;
- stagionalità;
- stazionarietà, trend, cambiamenti repentini, memoria a lungo termine;
- valori estremi (massimi)

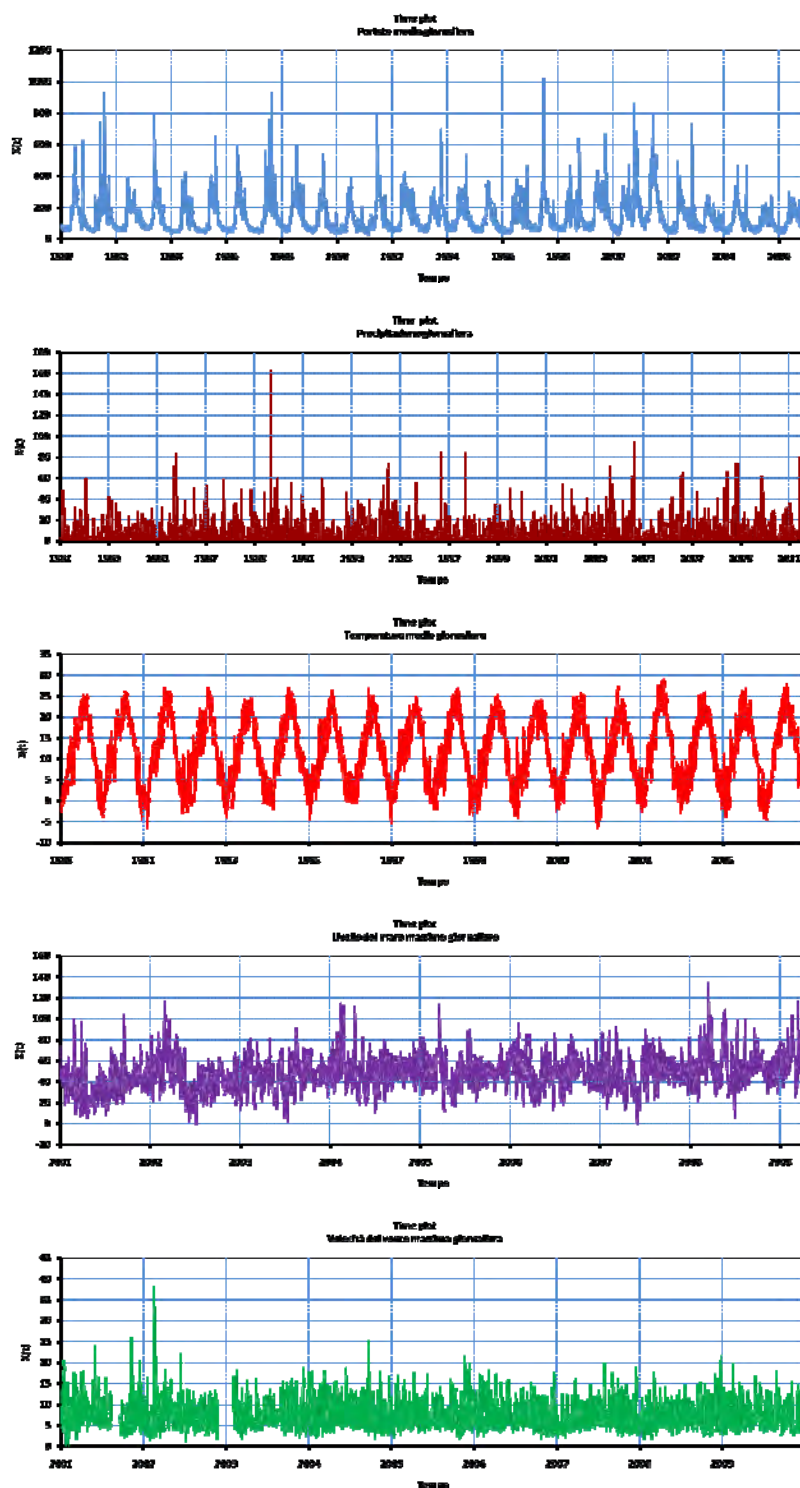


Figura 1.1 - Esempio di alcune diverse tipologie di serie di dati idrologici oggetto delle Linee Guida

1.3 Struttura delle Linee Guida

Dopo una breve disamina delle tipologie di dati idrologici, cui le elaborazioni statistiche si riferiscono, delle principali criticità che essi possono presentare e dopo un breve riferimento agli enti che raccolgono e diffondono dati idrologici e alla loro organizzazione, in relazione all'oggetto "serie storica" il documento è strutturato sostanzialmente in tre parti:

- la prima relativa ai metadati che dovrebbero corredare la serie di dati;

-
- la seconda relativa alla sua descrizione statistica
 - la terza relativa alle analisi statistiche di base da effettuare.

Ciascuna di queste parti è sintetizzata in una scheda riportata nell'”Appendice A”.

Si farà riferimento principalmente all'analisi statistica univariata delle serie temporali cioè le serie costituite da una sola variabile scalare, né si procederà alla identificazione di un modello statistico dei dati per effettuare previsioni (*forecast*).

Non saranno fatti riferimenti all'analisi bi-variata o multivariata (derivante dall'osservazione congiunta di due o più fenomeni) che si propone in generale di interpretare il meccanismo dinamico che ha generato la serie sulla base delle informazioni fornite da una altra serie di dati (e.g. analisi di regressione).

Né viene affrontato il problema della variabilità spaziale e del suo trattamento che sarà oggetto di un altro rapporto tematico

Le serie storiche sulle quali applicare le procedure proposte possono essere costituite da dati direttamente rilevati, opportunamente validati, che indicheremo come “serie primitive” (*primitive data*) ovvero da dati dedotti da questi mediante l'applicazione di procedure di aggregazione, di selezione ecc., che indicheremo come “serie derivate” (*derived data*).

Le analisi e le elaborazioni proposte vengono effettuata su serie di dati che si considerano validate, nelle quali ciascun dato ha superato i controlli di qualità standard.

Nel presente documento si affronta, invece, il problema della qualità della serie storica nella sua interezza, considerando che anche serie storiche con tutti dati validati possono presentare caratteristiche che ne riducono la complessiva qualità (e.g. elevato numero di dati mancanti, serie non omogenee prodotte con strumentazione diversa, non stazionarie ecc.).

Viene pertanto effettuata una proposta di un indice di qualità della serie di dati, denominato **iQuaSI** (indice di **Qualità** della **Serie** di dati **Idrologici**), definito in funzione della numerosità e della qualità dei singoli dati che la costituiscono, da mettere in relazione alle tipologie di elaborazioni da effettuare.

Ove possibile è stato inserito tra parentesi anche il corrispondente termine inglese per facilitare la ricerca dell'argomento anche sulla letteratura tecnico-scientifica in lingua anglosassone.

Nelle versioni successive potrà essere ampliato il numero delle caratteristiche delle serie di dati idrologici e la gamma di procedure di analisi dei dati.

Le presenti linee guida fanno in gran parte riferimento agli standard definiti a livello internazionale del WMO (WMO, 1988, 1994, 2000, 2003, 2008, 2009)

Nelle presenti LG sarà evidenziato attraverso un riquadro colorato particolari suggerimenti (in verde) o attenzioni (in arancio) da porre in relazione all'argomento trattato.

Per rendere più agevole l'applicazione di quanto riportato nel documento, le procedure proposte, i test statistici e il calcolo dei parametri sono implementati in un foglio di calcolo in MS Excel 2007 denominato **ANÁBASI** dall'acronimo “**AN**alisi **S**tatistica di **BA**se delle **S**erie di dati **Idrologiche**” le cui caratteristiche generali di funzionamento sono riportate nel capitolo 4.

Non è infine superfluo rilevare che le presenti LG non costituiscono un manuale di teoria della probabilità e statistica ma hanno il precipuo scopo di identificare e riportare un set di procedure che sono state individuate come standard per le analisi statistiche di base dei dati idrologici. Un primo livello di approfondimento delle procedure utilizzate è riportato nell'”Appendice B. Approfondimenti di statistica”, nella quale si fa anche riferimento a un'ampia letteratura specialistica. Tali approfondimenti sono tratti dallo studio, commissionato da ISPRA al Dipartimento GEMINI dell'Università della Tuscia, elaborato dal prof. Salvatore Grimaldi e l'ing. Francesco Serinaldi, avente per oggetto: “*Studio di metodologie di analisi statistica di base di serie storiche di dati idrologici a diverse scale di aggregazione, finalizzato alla definizione di ‘Linee Guida’ sull'elaborazione di dati idrologici*”.

1.4 Target delle Linee Guida

Il documento si rivolge principalmente agli operatori dei Centri Funzionali regionali e degli enti competenti a supporto dell'attività di elaborazione e di diffusione dei dati idrologici.

Si rivolge anche a coloro che utilizzano le serie idrologiche per scopi di studio e di progettazione in maniera tale che possano essere dotati delle informazioni minime necessarie per un appropriato e maggiormente consapevole utilizzo del dato idrologico.

2. I dati idrologici

I dati idrologici cui si farà riferimento nelle presenti LG sono quelli “tradizionali”, principalmente raccolti e pubblicati sistematicamente dal 1917 fino al 2002 dal Servizio Idrografico e Mareografico Nazionale (SIMN, fino al 1989 Servizio Idrografico Italiano, SII) nella sua articolazione territoriale in Compartimenti Idrografici, negli Annali Idrologici suddivisi in Parte I e Parte II, e in seguito raccolti e pubblicati dagli enti regionali competenti nei quali sono confluiti in parte gli uffici compartimentali del SIMN. Nello specifico, quelli “tradizionali” d’interesse prettamente idrologico sono:

- Temperatura dell’aria
- Precipitazione (liquida e solida)
- Altezza del manto nevoso
- Altezza idrometrica (fiumi e laghi)
- Livello freatico
- Portata liquida
- Livello di marea
- Trasporto solido al fondo e in sospensione

mentre quelli “non tradizionali”, di maggiore interesse in ambito meteorologico o agro-meteorologico, raccolti e pubblicati da altri enti con competenza principalmente meteorologica e agro-meteorologica, sono:

- Pressione atmosferica
- Umidità relativa
- Direzione e velocità del vento
- Evaporazione ed evapotraspirazione
- Radiazione solare
- Eliofoania

Negli ultimi anni, tuttavia, non essendo più così marcata la distinzione tra idrologia e meteorologia, gli enti competenti raccolgono l’intero set di dati meteo-idrologici per cui nel seguito faremo riferimento in generale a tali dati. Tuttavia per il fatto che solo recentemente gli enti competenti hanno iniziato a raccogliere l’intero set di dati le serie storiche di alcuni di essi non sono particolarmente lunghe.

Il riferimento principale ai dati idrologici prodotti dal SIMN è legato al fatto che essi sono i più numerosi e uniformemente distribuiti sul territorio italiano, nonché dal fatto che l’ISPRA ha ereditato, per effetto di successivi dettati normativi, le competenze e l’esperienza del Servizio stesso.

Non si escludono, tuttavia, serie temporali di dati meteo-idrologici prodotte da altri enti e istituzioni.

Si ritiene opportuno riportare una breve descrizione dell’organizzazione dei principali enti, a carattere nazionale e regionale/locale, che hanno raccolto e raccolgono dati di interesse meteo-idrologico.

Si ritiene, altresì, opportuno riportare, per ciascuna tipologia di dato meteo-idrologico, una breve descrizione delle modalità di raccolta e dei principali aspetti critici che si possono riscontrare nelle serie attualmente reperibili, di cui è necessario tenere conto nell’elaborazione per una corretta interpretazione dei risultati.

Ad esempio, una serie storica molto lunga può essere costituita da dati rilevati con modalità e strumentazioni diverse nel corso del tempo per cui è essenziale effettuare, come si propone nelle LG, una preliminare analisi di stazionarietà della serie per individuarne trend e/o discontinuità.

2.1 Principali enti che raccolgono e diffondono serie di dati idrologici

I principali enti pubblici che attualmente raccolgono e/o archiviano e/o diffondono dati di interesse meteo-idrologico sono:

- ISPRA (ex SIMN)
- Centri funzionali decentrati e centrali
- CRA - CMA (ex UCEA)
- CNMCA - Servizio Meteorologico Aeronautica
- Servizi Meteorologici e Agro-Meteorologici Regionali

2.1.1 ISPRA (ex SIMN)

2.1.1.1 Banche dati Annali ex SIMN

L'Istituto Superiore per la Protezione e la Ricerca Ambientale (ISPRA), come già detto in premessa, ha raccolto l'eredità tecnica del SIMN e attualmente non raccoglie direttamente dati idrologici ma costituisce il Centro presso il quale confluiscono i dati raccolti dagli enti regionali competenti per elaborare e diffondere lo stato idrologico nazionale.

D'altra parte l'ISPRA ha predisposto la Banca Dati contenente tutti i dati idrologici pubblicati dal SIMN dalla sua istituzione nel 1917 al 2002. Il Servizio Idrografico è stato, infatti, istituito dal 1917 e da quella data sono stati pubblicati in maniera sistematica dati idrologici per tutto il territorio nazionale. Quella del Servizio Idrografico costituisce la più lunga nel tempo, densa e uniformemente distribuita nello spazio e sistematica raccolta di dati idrologici (principalmente temperatura, precipitazione e portata dei corsi d'acqua) disponibile per il territorio italiano.

Il SIMN era articolato sul territorio in compartimenti idrografici (Figura 2.1) ciascuno dei quali pubblicava autonomamente gli Annali Idrologici Parte I e Parte II.



Figura 2.1 - Compartimenti idrografici all'epoca del SIMN

Gli uffici dei compartimenti idrografici erano articolati nella maniera seguente:

- Ufficio Compartimentale con sede in Venezia e con sezioni staccate ad Udine e Padova e con officina a Strà, competente sui bacini sfocianti sul litorale Alto Adriatico a nord del Fiume Po e sul tratto costiero compreso tra il confine italo-sloveno e Porto Levante compreso, incluse le superfici lagunari venete;
- Ufficio Compartimentale con sede a Parma e con sezioni staccate a Milano, Torino Sondrio, competente sul bacino del Fiume PO e sul tratto costiero compreso tra la foce di Porto Levante e la foce di Porto Garibaldi compreso;
- Ufficio Compartimentale con sede a Bologna competente sui bacini con foce sul litorale adriatico dal fiume Reno al fiume Tronto e sul tratto costiero compreso tra Porto Garibaldi e la foce del Fiume Tronto compresa;
- Ufficio Compartimentale con sede a Pescara competente sui bacini con foce sul litorale adriatico dal fiume Salinello al fiume Fortore e nel tratto costiero compreso tra la foce del Fortore e la foce del Lato compresa;
- Ufficio Compartimentale con sede a Bari competente sui bacini con foce sul litorale adriatico e ionico dal fiume Candelaro al fiume Lato e nel tratto costiero compreso tra la foce del Fortore e la foce del Lato compresa;
- Ufficio Compartimentale con sede a Catanzaro con sezione staccata a Potenza competente sui bacini con foce sul litorale ionico e tirrenico dal fiume Bradano al Fiume Noce e nel tratto costiero compreso tra la foce del Lato e la foce del Noce compresa;
- Ufficio Compartimentale con sede a Napoli competente sui bacini con foce sul litorale tirrenico dal fiume Garigliano al fiume Bussento e nel tratto costiero compreso tra la foce del Noce e la foce del Garigliano compresa;

- Ufficio Compartimentale con sede a Roma competente sui bacini con foce sul litorale tirrenico dal fiume Fiora al Lago di Fondi e nel tratto costiero compreso tra la foce del Garigliano e la foce del Fiora compresa;
- Ufficio Compartimentale con sede a Pisa, con sezione staccata a Firenze, competente sui bacini con foce sul litorale tirrenico dal fiume Serchio al fiume Albenga e nel tratto costiero compreso tra la foce del Fiora e la foce del Magra compresa;
- Ufficio Compartimentale con sede a Genova competente sui bacini del litorale ligure dal confine italo-francese al fiume Magra e nel tratto costiero compreso tra la foce del Magra ed il confine italo-francese.

Nel periodo di massima sviluppo il Servizio Idrografico disponeva di oltre 3000 stazioni di rilevamento termo-pluviometrico funzionanti con una densità uniforme sul territorio italiano superiore ad una stazione ogni 100 km²

Attualmente presso l'ISPRA sono archiviati e, tramite il sito WEB, gratuitamente e liberamente accessibili e scaricabili una grande quantità di dati idrologici in diverso formato.

Nelle Figura 2.2, Figura 2.3, Figura 2.4, Figura 2.5 sono riportati le pagine del portale WEB ISPRA nell'ambito del Sistema Informativo Nazionale Ambientale (SINAnet) e del Sistema Informativo Nazionale per la Tutela delle Acque Italiane (SINTAI), attraverso le quali si può accedere ai dati idrologici provenienti dagli annali parte I e parte II, nonché da altre pubblicazione del Servizio Idrografico



Figura 2.2 - Pagina di accesso alle serie storiche di precipitazione, temperatura e idrometria <http://www.sintai.sinanet.apat.it/>



Figura 2.3 - Pagina di accesso agli annali idrologici Parte I e Parte II in formato pdf <http://www.acq.isprambiente.it/annalipdf/>



Figura 2.4 - Pagina di accesso a dati storici digitalizzati <http://193.206.192.243/storico/>

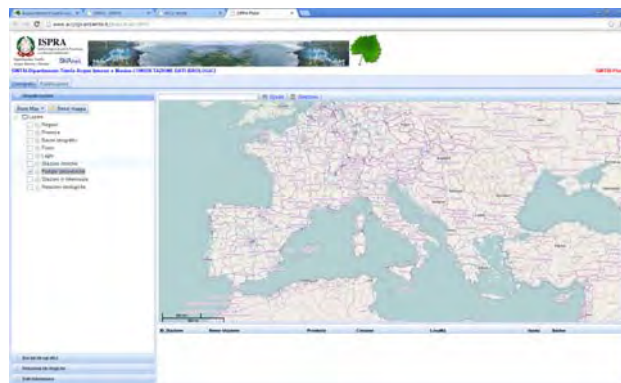


Figura 2.5 - Pagina di accesso alla banca dati PLUTER <http://www.acq.isprambiente.it/pluter/index.html>

2.1.1.2 Banca dati mareografica e ondametrica

L'ISPRa, mediante reti di sua proprietà, raccoglie, archivia e pubblica dati rilevati dalla Rete Mareografica Nazionale (RMN) e alla Rete Ondametrica Nazionale (RON).

La RMN è costituita da 33 stazioni di misura uniformemente distribuite sul territorio nazionale ed ubicate prevalentemente all'interno delle strutture portuali e che misurano, ad intervalli orari e sub-orari:

- Livello del mare
- Temperatura dell'aria
- Temperatura dell'acqua
- Pressione atmosferica
- Umidità relativa
- Velocità e direzione del vento

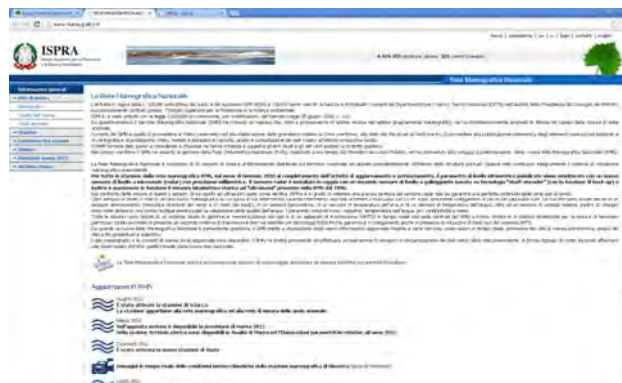


Figura 2.6 - Pagina di accesso alla banca dati mareografica <http://www.mareografico.it/>

La RON è attualmente costituita da 15 boe oceanografiche dislocate lungo le coste italiane. Ciascuna boa è equipaggiata con un ondometro direzionale accelerometrico a stato solido di altissima precisione e una stazione meteorologica completa.

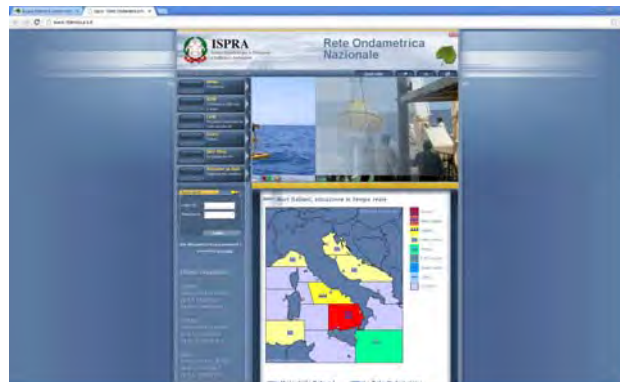


Figura 2.7 - Pagina di accesso alla banca dati della Rete Ondametrica Nazionale (RON) <http://www.telemisura.it/>

I dati della RMN e della RON sono gratuiti e l'accesso è consentito previa registrazione dell'utente.

2.1.1.3 Banca dati della laguna di Venezia

L'ISPRA gestisce anche il Servizio Laguna di Venezia, che ha ereditato le competenze dell'Ufficio Idrografico sulla laguna di Venezia, che dispone oggi di una rete di 52 stazioni meteo-mareografiche distribuite all'interno del bacino lagunare e lungo il litorale Nord – Adriatico .

Alla banca dati si accede tramite interfaccia WEB (Figura 2.8)



Figura 2.8 - Pagina di accesso alla banca dati della laguna di Venezia http://www.venezia.isprambiente.it/ispra/index.php?folder_id=20

I dati della banca dati della laguna di Venezia sono gratuiti e l'accesso è consentito previa registrazione dell'utente.

2.1.1.4 Banca dati climatica SCIA

L'ISPRA, nell'ambito dei propri compiti di gestione e sviluppo del sistema informativo nazionale ambientale, in collaborazione con il Servizio Meteorologico dell'Aeronautica Militare, l'Unità di Ricerca per la Climatologia e la Meteorologia applicate all'Agricoltura (CMA-CRA), numerose Agenzie Regionali per la Protezione dell'Ambiente e i Servizi Agrometeorologici Regionali della Sicilia e delle Marche, ha predisposto il "Sistema nazionale per la raccolta, l'elaborazione e la diffusione di dati Climatologici di Interesse Ambientale" (SCIA), che raccoglie alcuni dati provenienti da diverse reti di osservazione di grandezze climatiche e attraverso la loro elaborazione, rende disponibili i valori decadali, mensili e annuali denominati convenzionalmente indicatori climatologici.

Le informazioni prodotte da SCIA, di ambito più strettamente climatologico, sono gratuitamente e liberamente accessibili e scaricabili attraverso il sito web dedicato all'interno del sito dell'ISPRA (Figura 2.9)



Figura 2.9 - Pagina di accesso alla banca dati di SCIA(<http://www.scia.sinanet.apat.it/#>)

2.1.2 Centri Funzionali

La Rete Nazionale dei Centri Funzionali per le finalità di protezione civile, introdotta dalla Dir.P.C.M. del 27 febbraio 2004, è costituita da 21 Centri Funzionali Decentrati (CFD), uno per ciascuna regione e provincia autonoma, e da un Centro Funzionale Centrale (CFC) presso il Dipartimento della Protezione Civile (DPC). Questa rete è supportata da diversi Centri di Competenza che forniscono servizi, informazioni, dati, elaborazioni e contributi tecnico-scientifici in specifici ambiti, di cui uno dei primi a essere istituito è quello presso l'ISPRA.

Presso alcuni dei centri funzionali decentrati o regionali sono stati trasferiti, per effetto del DPCM del 24 luglio 2002, gli Uffici Compartimentali del SIMN comprensivi anche dei relativi beni strumentali, delle stazioni di misura delle portate e le reti di rilevamento manuale, automatico e in telemisura dei parametri idro-meteo-pluviometrici con i sistemi di collegamento in ponte radio. I centri funzionali decentrati rappresentano, quindi, il proseguimento dei rilevamenti della rete del SIMN. Parte dei CFD sono collocati presso le ARPA.

2.1.3 CRA - CMA (ex UCEA)

L'Unità di Ricerca per la Climatologia e la Meteorologia Applicate all'Agricoltura (CMA) del Consiglio per la Ricerca e la Sperimentazione in Agricoltura (CRA) ex Ufficio Centrale di Ecologia Agraria (UCEA) del Ministero delle Politiche Agricole e Forestali (MiPAF) è un ente pubblico che, tra le altre cose, pubblica la Banca Dati Agrometeorologica Nazionale. Questa banca dati raccoglie le misure di tre reti:

- 1) la Rete Agrometeorologica Nazionale (RAN), costituita da una rete di sensori propri del CRA - CMA;
- 2) la rete dell'Ente Nazionale Assistenza al Volo (ENAV), costituita da stazioni situate presso gli aeroporti civili;
- 3) la rete dell'Aeronautica Militare (AM), costituita da stazioni situate presso gli aeroporti o le installazioni militari.

I dati sono pubblicati e scaricabili gratuitamente e liberamente dal sito web <http://www.cra-cma.it/homePage.htm> (Figura 2.10).



Figura 2.10 - Pagina di accesso alla banca dati agrometeorologica del CRA-CMA (<http://www.cra-cma.it/homePage.htm>)

2.1.4 CNMCA e Servizio Meteorologico dell'Aeronautica

Il Centro Nazionale di Meteorologia e Climatologia Aeronautica (CNMCA), è l'organo operativo centrale del Servizio Meteorologico dell'Aeronautica.

Il CNMCA è il custode nazionale centrale di tutti i dati prodotti dall'intero Servizio Meteorologico dell'AM. Ha così ereditato, conserva e rende disponibili agli utenti pubblici e privati osservazioni, misure ed elaborazioni meteo effettuate sul territorio nazionale dai diversi enti militari a partire dalla costituzione dell'AM. Inoltre custodisce la documentazione meteorologica relativa a osservazioni effettuate sul territorio delle ex-colonie italiane (Eritrea, Somalia, Etiopia, Libia, Istria, Dalmazia, Dodecaneso) a partire dal XIX secolo.

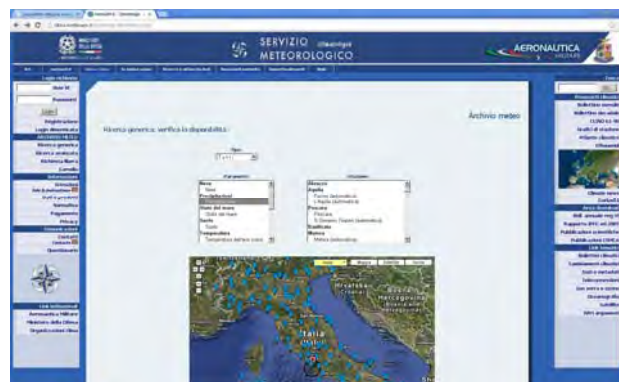


Figura 2.11 - Pagina di accesso alla banca dati meteorologica del Servizio Meteorologico dell'Aeronautica (<http://clima.meteoam.it/RichiestaDatiGenerica.php>)

I dati non sono liberamente accessibili né gratuiti. Tramite il sito WEB si può verificare solo la disponibilità dei dati nel database (Figura 2.11).

2.1.5 Servizi Meteorologici e Agrometeorologici Regionali

Altri enti pubblici a carattere locale e regionale con competenze di natura essenzialmente agrometeorologica producono dati d'interesse idrologico.

2.1.6 Elenco di siti web per l'accesso ai dati meteo-idrologici e agrometeorologici

Si riporta l'elenco dei siti web (Tabella 2.1) da cui è possibile accedere, anche se con modalità e funzionalità differenti, ai dati meteo-idrologici e agrometeorologici.

Tabella 2.1 - Enti e relativi web-link per l'accesso ai dati meteo-idrologici e agrometeorologici

Regione/ Provincia	Ente	Tipo dato	URL
ABR	Regione	Meteo-Idro	http://www.regione.abruzzo.it/xIdrografico/index.asp
BAS	ARPA	Meteo-Idro	http://www.arpab.it/idrometeorologico/index.asp
BZ	Provincia	Meteo-Idro	http://www.provincia.bz.it/hydro/index_i.asp
CAM	ARPA	Meteo	http://www.meteoarpac.it/
	Regione	Agro	http://www.sito.regione.campania.it/agricoltura/meteo/agrometeo.htm
CAL	ARPA		http://www.cfc Calabria.it/index.php/dati-storici.html
EMR	ARPA	Meteo-Idro	http://www.arpa.emr.it/sim/?idrologia/dati_e_grafici
FVG		Meteo-Idro	http://www.protezionecivile.fvg.it/ProtCiv/default.aspx/Cae/MappaCae.aspx
			http://www.osmer.fvg.it/
LAZ	Regione	Meteo-idro	http://www.idrografico.roma.it/default.aspx
		Agro	http://www.arsial.it/portalearsial/agrometeo/
LIG	ARPA		http://www.arpal.gov.it/index.php?option=com_flexicontent&view=items&cid=56&id=206&Itemid=244
			http://www.arpal.gov.it/index.php?option=com_flexicontent&view=items&cid=56&id=166&Itemid=199
LOM	ARPA	Meteo-Idro	http://ita.arpalombardia.it/meteo/meteo.asp http://cmg.arpalombardia.it/webcmgfrontend/ http://idro.arpalombardia.it/pmapper-4.0/map.phtml
	Regione	Meteo	http://sinergie.protezionecivile.regione.lombardia.it/sinergie_wsp5/html/public/report/mapHPMNetwork.jsf
	Consorzio Grandi Laghi	Meteo-Idro	http://www.laghi.net/homepage.aspx
MAR	Regione	Meteo-idro	http://www.protezionecivile.marche.it/moduli.asp?modulo=statsol
		Agro	http://meteo.regione.marche.it/
MOL	Regione	Meteo-Idro	http://www.protezionecivile.molise.it/index.php/centro-funzionale/settore-geo.html
		Agro	http://www.arsiam.it/meteo/index.html
PIE	ARPA	Meteo-Idro	http://www.arpa.piemonte.it/approfondimenti/temi-ambientali/idrologia-e-neve/accesso-ai-dati-meteo-idro-nivologici/banca-dati-meteorologica-e-banca-dati-idrologica
PUG	Regione	Meteo-Idro	http://www.regione.puglia.it/idrografico/ http://www.protezionecivile.puglia.it/public/page.php?73
		Agro	http://www.agrometeopuglia.it/opencms/opencms/Agrometeo/home_agro
SAR	Regione	Meteo-Idro	http://www.regione.sardegna.it/j/v/25?s=131338&v=2&c=5650&t=1
	ARPA	Agro	http://www.sar.sardegna.it/
SIC	Regione	Meteo-Idro	http://www.osservatorioacque.it/?cmd=datitlm%20%20
		Agro	http://www.sias.regione.sicilia.it/
Trento	Provincia	Meteo	http://www.meteotrentino.it/Default.aspx
		Idro	http://www.floods.it
TOS	Regione	Meteo-Idro	http://www.sir.toscana.it/
UMB	Regione		http://www.cfumbria.it/
		Meteo-Idro	http://www.idrografico.regione.umbria.it/annali/default.aspx
VEN	ARPA	Meteo-Idro	http://www.arpa.veneto.it/temi-ambientali/idrologia/dati
		Agro	http://www.arpa.veneto.it/temi-ambientali/agrometeo/dati/dati-disponibili
VDA	Regione	Meteo-Idro	http://www.regione.vda.it/Territorio/centrofunzionale/settoreidrografico/default_i.asp

2.2 Grandezze idrologiche d'interesse

2.2.1 Temperatura dell'aria

La temperatura dell'aria (*surface air temperature*) è, insieme alla precipitazione, la grandezza idrologica di cui si trovano le serie storiche, alle diverse scale di aggregazione, più lunghe. Ciò testimonia il fatto che, sin dall'istituzione dei primi servizi per la raccolta di dati idrologici, essa è stata riconosciuta, al pari della precipitazione, di primaria importanza per la valutazione di alcune fondamentali componenti del ciclo idrologico. Essa infatti interviene in maniera determinata nella valutazione dell'evapotraspirazione.

La temperatura dell'aria, attualmente, è misurata secondo standard e procedure normate dal *WMO* e, per quanto riguarda il SIMN, recepite nel quaderno "Norme tecniche per la raccolta e l'elaborazione dei dati idrometeorologici – parte I" (SIMN, 1998a).

Secondo tali norme, gli strumenti di misura dovrebbero essere posti ad una altezza dal suolo compresa tra 1.25 e 2 metri, all'interno di capannine meteorologiche di colore bianco per ridurre l'influenza della radiazione solare e della temperatura del suolo. La capannina dovrebbe essere posta su di un prato, lontano da edifici che potrebbero creare dei micro-climi locali. Queste brevi notizie sull'attuale standard di misura della temperatura per evidenziare che per le serie disponibili che coprono un arco temporale di alcuni decenni non è possibile garantire che i dati siano stati rilevati tutti con i medesimi standard di qualità degli strumenti, delle procedure e delle installazioni, che potrebbero con elevata probabilità essere cambiati nel tempo.

Le serie storiche di temperatura dell'aria, generalmente reperibili, sono costituite dai valori di:

- Temperatura minima giornaliera - $T_{g,min}$;
- Temperatura massima giornaliera - $T_{g,max}$;
- Temperatura media giornaliera - $T_{g,med}$;
- Temperatura minima mensile - $T_{m,min}$;
- Temperatura massima mensile - $T_{m,max}$;
- Temperatura media mensile - $T_{m,med}$.

È tuttavia opportuno anche fare alcune precisazioni sulle modalità di rilevamento della temperatura che, ovviamente, si è profondamente modificata nel tempo. Ad esempio il SIMN ha pubblicato dati di temperatura per una medesima stazione rilevati sia con termometro a minima e massima a lettura visiva (simbolo T), ovvero con termometro a lamina bimetallica registratore su tamburo rotante (simbolo T_r), ovvero ancora con termometro elettronico (termocoppie, termoresistenze ecc., simbolo T_e).

È del tutto evidente come l'accuratezza degli strumenti utilizzati sia molto diversa e potrebbe condizionare significativamente le elaborazioni statistiche.

Altro elemento al quale prestare attenzione è la definizione della grandezza che costituisce la serie storica.

Infatti, negli Annali Idrologici del SIMN, la temperatura media è definita, contrariamente alla definizione matematica, come semplice media aritmetica del valore minimo e del valore massimo. Tale convenzionale definizione è da attribuire ai tempi in cui la temperatura era rilevata solo con termometri a minima e a massima a lettura visiva e quindi solo due dati giornalieri in cui il massimo veniva attribuito al giorno precedente e il minimo al giorno dell'osservazione che avveniva alle ore 9:00, ipotizzando che la temperatura minima fosse sempre quella notturna (Figura 2.12). Ciò poteva comportare di attribuire la temperatura minima al giorno dopo, qualora questa avveniva durante le ore diurne.

Tale definizione, tuttavia, è generalmente stata mantenuta anche per la temperatura rilevata con termometri registratori, anche se ormai la tendenza è utilizzare il valore medio giornaliero come media aritmetica dei valori campionati ad intervalli orari o sub-orari.

I dati sono trasmessi da Osservatori o da stazioni termopluviometriche controllati o dipendenti direttamente dalla Sezione.

Ogni stazione è fornita di un termometro a massima e di un termometro a minima, oppure di termometro a massima e minima uniti, che vengono osservati ogni giorno alle ore 9 antimeridiane.

Il valore massimo rilevato viene assegnato al giorno precedente; quello minimo al giorno stesso dell'osservazione.

Le stazioni sono ordinate, nelle tabelle, secondo la rispettiva posizione idrografica.

Le tabelle sono precedute dall'elenco e caratteristiche delle stazioni termometriche che hanno funzionato nell'anno.

TABELLA I. — Sono riportati, per le stazioni che hanno regolarmente funzionato nell'anno, i valori massimi e minimi rilevati giornalmente, e le rispettive medie mensili,

unitamente alla temperatura media del mese e dell'anno cui si riferiscono le osservazioni e le corrispondenti medie del periodo.

TABELLA II. — Per tutte le stazioni della tabella I sono riportate:

a) le medie mensili ad annue delle massime e delle minime temperature osservate giornalmente e le medie mensili ed annue delle temperature diurne. Come "temperatura diurna" è assunto il valore della semisomma delle temperature massima e minima osservate in uno stesso giorno.

b) le temperature estreme (massima e minima) osservate in ogni mese e nell'anno, ed il giorno nel quale sono state osservate.

Tutte le temperature riportate sono espresse in gradi centigradi e corrispondono alle letture effettivamente eseguite, non essendo effettuata la riduzione al livello del mare.

Figura 2.12 - *Contenuto delle tabelle termometriche: estratto da Annali Idrologici parte I anno 1951 Compartimento di Napoli*

Anche per quanto riguarda l'orario di acquisizione dei dati, fino all'introduzione di sistemi di acquisizione elettronici, non è stata mai effettuata una distinzione tra ora solare e ora legale. I moderni sistemi di acquisizione attualmente fanno riferimento al tempo assoluto UTC (*Coordinated Universal Time*) o GMT (*Greenwich Mean Time*).

Per una serie lunga di dati termometrici è necessario prestare attenzione alla definizione e alle modalità di rilevamento che, con elevata probabilità, potrebbero essere mutate nel tempo



2.2.2 Precipitazione

La precipitazione al suolo (*precipitation*), che si manifesta in forma liquida e solida, è la grandezza di maggiore interesse idrologico per le sue fondamentali implicazioni sulla vita dell'uomo sia quando essa si manifesta con scarsità sia quando si manifesta in maniera abbondante ed estrema.

Anch'essa è attualmente misurata secondo standard e procedure normate dal *WMO* e, per quanto riguarda il SIMN, recepite nel quaderno "Norme tecniche per la raccolta e l'elaborazione dei dati idrometeorologici – parte I" (SIMN, 1998a).

Secondo tali norme, gli strumenti di misura dovrebbero essere posti ad una altezza dal suolo compresa tra 0,5 e 1,5 metri dal suolo e posti lontano almeno 10 metri da ostacoli verticali, quali edifici o alberi che ne impediscano l'accumulo della pioggia o neve soprattutto in caso di precipitazioni trasversali.

Le serie storiche di precipitazione, generalmente reperibili, sono costituite dai valori:

- Precipitazione cumulata annuale - P_a ;
- Precipitazione cumulata mensile - P_m ;
- Precipitazione cumulata giornaliera - P_g ;
- Precipitazione cumulata oraria - P_h ;
- Precipitazioni cumulate su intervalli di X minuti - $P_{X \text{ min}}$.

Le serie di precipitazione oraria e sub-oraria di X minuti sono prodotte solo negli ultimi decenni con l'utilizzo di pluviometri a registrazione elettronica.

In realtà anche i pluviometri registratori meccanici, registrando con tratto continuo su supporto cartaceo, avrebbero registrato serie storiche orarie e sub-orarie ma che risulta difficile e oneroso rendere in forma digitale. Sono stati effettuati diversi esperimenti di sviluppo di software per la lettura automatica o semi-automatica per la lettura dei diagrammi che, tuttavia, sono rimasti solo semplici tentativi sperimentali non utilizzati su vasta scala per il recupero dell'informazione idrologica contenuta nei diagrammi.

Già a partire dall'introduzione di pluviometri registratori su tamburo rotante (pluviografi) si è invece proceduto alla definizione di serie storiche dei massimi annuali di precipitazione orarie per alcune durate prefissate:

- Precipitazione durata 1 ora massima annuale - $P_{1h,max}$;
- Precipitazione durata 3 ore massima annuale - $P_{3h,max}$;
- Precipitazione durata 6 ore massima annuale - $P_{6h,max}$;
- Precipitazione durata 12 ore massima annuale - $P_{12h,max}$;
- Precipitazione durata 24 ore massima annuale - $P_{24h,max}$.

Generalmente la precipitazione giornaliera è cumulata tra le ore 9:00 del giorno precedente e le ore 9:00 del giorno in cui è effettuata l'osservazione. Tale definizione, adottata per le osservazioni rilevate e pubblicate dal SIMN (Figura 2.13), convenzionale era legata alle particolari condizioni di rilevamento effettuato da un osservatore umano che misurava manualmente la quantità di acqua precipitata e raccolta dal pluviometro semplice.

CONTENUTO DELLE TABELLE

Le tabelle sono precedute dall'elenco e caratteristiche delle stazioni di osservazione che hanno funzionato nell'anno. I valori delle precipitazioni riportati sono espressi in millimetri di acqua e comprendono pioggia e neve fusa.

TABELLA I. — Per ogni stazione riporta la quantità di pioggia caduta giornalmente ed i totali mensili ed annuo della precipitazione e del numero dei giorni piovosi.

Per le stazioni dotate di apparecchiatura a lettura diretta (pluviometri comuni e pluviometri) le osservazioni vengono eseguite ogni giorno generalmente alle ore 9 ed il risultato viene attribuito al giorno stesso della misura; il valore segnato rappresenta quindi la quantità di precipitazione caduta nelle 24 ore che hanno preceduto la misura.

Per le stazioni dotate di pluviografo, si riporta, per ogni giorno, la quantità di pioggia che dal diagramma risulta caduta nelle 24 ore comprese fra le ore 9 del giorno precedente e le ore 9 del giorno di cui si tratta.

Con carattere **grassetto** è stampato il quantitativo giornaliero misurato per ogni mese.

TABELLA II. — Per le stesse stazioni di cui alla tabella I, riportati i totali mensili ed annui delle quantità di precipitazione.

Per ciascuna stazione è riportato in **grassetto** il più elevato dei valori mensili ed in **corsivo** il più basso.

TABELLA III. — Per le stazioni dotate di pluviografo, riporta i dati relativi ai valori più elevati delle precipitazioni registrate, nell'anno, per 1, 3, 6, 12, e 24 ore consecutive appartenenti o no allo stesso giorno.

Sono considerate le precipitazioni iniziate dopo le ore 0 del primo gennaio e quelle, eventualmente terminate dopo le ore 24 del 31 dicembre.

TABELLA IV. — Per alcune stazioni, opportunamente scelte, riporta i massimi valori delle precipitazioni verificatesi per 1, 2, 3, 4 e 5 giorni consecutivi, appartenenti o no allo stesso mese. Sono considerati solamente i periodi il cui inizio cade entro l'anno anche se eventualmente sono terminati nell'anno successivo.

TABELLA V. — Riporta il valore, la durata e la data delle precipitazioni di maggiore intensità e di breve durata registrate dai pluviografi.

TABELLA VI. — Riporta, per alcune determinate stazioni, per i mesi da gennaio a maggio e da ottobre a dicembre nei quali possono verificarsi precipitazioni nevose:

a) le altezze, in centimetri, degli strati nevosi sul suolo presenti nell'ultimo giorno delle tre decadi mensili.

b) il numero dei giorni nei quali si sono avute precipitazioni nevose,

c) il numero complessivo dei giorni di permanenza della neve sul suolo.

Figura 2.13 - *Contenuto delle tabelle pluviometriche: estratto da Annali Idrologici parte I anno 1951 Compartimento di Napoli*

Anche con l'avvento dei pluviometri registratori e successivamente da quelli a memorizzazione elettronica con campionamento sub-orario per uniformità di definizione, si è continuato a definire la precipitazione giornaliera come quella cumulata tra le ore 9:00 dei due giorni consecutivi. Tuttavia si possono trovare serie storiche di precipitazione cumulata giornaliera in cui viene esplicitamente indicato che il dato è riferito all'intervallo tra le ore 0:00 e le 24:00 del medesimo giorno.

Sensibili all'intervallo di campionamento del dato originario e alle modalità con cui viene estratto il valore massimo, sono le serie storiche dei massimi annuali delle precipitazioni orarie di diversa durata. Con l'avvento degli strumenti a registrazione elettronica, infatti, la finestra mobile di ampiezza pari alla durata della precipitazione di cui si vuole trovare il massimo, viene fatta scorrere di una quantità uguale all'intervallo minimo di campionamento che può essere anche inferiore a 10 minuti. Tale modalità individua valori massimi sensibilmente maggiori (specie per le brevi durate) di quelli che potevano individuarsi con una lettura manuale sul diagramma cartaceo con una risoluzione non inferiore all'ora. Tale effetto, come è evidente, comporta una discontinuità nella serie dei massimi orari di cui dovrebbe tenersene adeguatamente conto.

Dalle serie temporali di precipitazione giornaliera si derivano generalmente serie temporali di:

- Numero mensile **giorni piovosi**;
- Numero annuo **giorni piovosi**;
- Numero di **giorni non piovosi** consecutivi;
- Numero massimo di **giorni piovosi** consecutivi nell'anno;
- Numero massimo di **giorni non piovosi** consecutivi nell'anno;

nonché di precipitazione al di sopra di un valore soglia prefissata (serie POT, *Peak Over Threshold*)

- Precipitazione giornaliera sopra la soglia $X - P_g, x$;
- Precipitazione mensile sopra la soglia $X - P_m, x$.

Le modalità di definizione della precipitazione giornaliera potrebbero essere diverse: dalle 9:00 alle 9:00 o dalle 0:00 alle 24:00.



Il valore del massimo annuale di precipitazione oraria di assegnata durata potrebbe essere influenzato significativamente dall'intervallo di campionamento del dato grezzo di base e dall'intervallo di scorrimento della finestra mobile.



2.2.3 Altezza idrometrica, livello freatico, livello mareografico

L'altezza idrometrica (*hydraulic depth*) è una grandezza relativa nel senso che è misurata rispetto ad un riferimento arbitrario detto "zero idrometrico" che non necessariamente coincide con il fondo dell'alveo, che, come è noto, non è fisso ma variabile. È una grandezza che, pertanto, può assumere anche valori negativi. Se all'altezza idrometrica rilevata si somma la quota topografica assoluta, rispetto al livello del mare, dello zero idrometrico si ottiene il livello idrometrico assoluto rispetto al livello del mare.

Il livello freatico (*phreatic depth*) è generalmente riferito al piano campagna. Data la sua lenta variabilità veniva rilevato dal SIMN ogni tre giorni

Il livello mareografico (*sea level*) è invece riferito al livello medio mare e per la sua alta variabilità è generalmente campionato a livello orario o sub-orario.

Le serie storiche relative a livelli idrici di fiumi, laghi, acquiferi e del mare, generalmente reperibili sono:

- altezza idrometrica media giornaliera - h_g (per laghi e fiumi);
- livello freatico tri-giornaliero f_{3g} ;
- livello freatico medio mensile f_m ;
- livello mareografico orario e sub-orario.

L'altezza idrometrica media giornaliera rilevata e pubblicata dal SIMN (Figura 2.14) è convenzionalmente definita sulla base delle modalità di lettura manuale del livello letto alla stadia da un operatore umano prima che venissero introdotti gli idrometri registratori. La convenzione è comunque rimasta anche per le stazioni dotate di idrometro registratore: l'altezza media giornaliera è quella rilevata alle ore 12:00 dall'operatore, ovvero dedotta in corrispondenza del mezzogiorno nello spoglio dei diagrammi per le stazioni registratrici.

CONTENUTO DELLA TABELLA

La tabella è preceduta dall'elenco e caratteristiche delle stazioni idrometriche che hanno funzionato nell'anno.

TABELLA 1. — Riporta per alcune stazioni, le altezze idrometriche meridia-

ne rilevate direttamente all'idrometro da parte dell'osservatore oppure dedotte in corrispondenza del mezzogiorno, dallo spoglio dei diagrammi per le stazioni fornite di apparecchio registratore.

Figura 2.14 - *Contenuto della tabella idrometrica: estratto da Annali Idrologici parte II anno 1951 Compartimento di Napoli*

L'altezza idrometrica è una grandezza che non presenta di per sé un grande interesse idrologico poiché, a causa dell'estrema variabilità dell'alveo legata alla dinamica morfologica, è fortemente non stazionaria. La sua utilità risiede invece nella sua relativa facilità ed economicità di misura che consente, in virtù di una relazione matematica biunivoca, sotto opportune condizioni, di determinare il valore della portata idrica che transita nella sezione. In tal modo dalle serie delle altezze idrometriche si determinano le serie delle portate.

La serie di altezza idrometrica presenta una sensibile non stazionarietà legata alla dinamica morfologica dei corsi d'acqua e quindi non risulta particolarmente interessante nelle analisi statistiche delle serie storiche.



La definizione del valore medio dell'altezza idrometrica non corrisponde alla definizione matematica. Esso corrisponde convenzionalmente al valore letto alla stadia alle ore 12:00.



2.2.4 Portata liquida

La portata liquida (*discharge*) che transita in un corso d'acqua è la grandezza idrologica di maggiore interesse per la sua diretta implicazione su gran parte delle attività umane.

La portata liquida in un corso d'acqua è definita come il volume di acqua che attraversa una sezione nell'unità di tempo. In generale si trova misurata in m³/s. Meno frequentemente si trova anche espressa in l/s.

Le serie storiche di dati di portata liquida in un corso d'acqua non derivano generalmente da misure dirette di portata ma sono derivate dalle serie dei dati di altezza idrometrica attraverso un modello matematico di trasformazione costituita dalla cosiddetta "scala di deflusso" o "scala delle portate". Tale circostanza, ovviamente, influenza significativamente l'accuratezza del dato di portata.

Le serie temporali di portata generalmente reperibili sono:

- Portata media giornaliera - Q_g ;
- Portata media mensile - Q_m ;
- Portata media annua - Q_a ;
- Portata istantanea massima annuale (colmo di piena) - $Q_{c,max}$.

Dalle serie temporali di portata media giornaliera si deducono generalmente serie temporali di:

- Portata media giornaliera massima annuale - $Q_{g,max}$;
- Portata media giornaliera minima annuale - $Q_{g,min}$;
- Portata di assegnata durata "d" - Q_d ;
- Portata media giornaliera al di sopra di una soglia prefissata X (serie POT, *Peak Over Threshold*) - $Q_{g,X}$.

I dati di portata media giornaliera del SIMN erano determinati convenzionalmente dalla trasformazione, mediante la scala di deflusso, dell'altezza idrometrica media giornaliera rilevata alle ore 12:00 come viene generalmente indicato negli annali (Figura 2.15). I dati di portata al colmo di piena invece erano dedotti dalla trasformazione mediante la scala di deflusso dell'altezza idrometrica al colmo dedotta dal diagramma dell'idrometro registratore ovvero, dove questo non era installato, dalle letture alla stadia ad intervalli ravvicinati (solo durante la piena) da parte dell'osservatore idrografico.

<p>La tabella è preceduta dall'elenco e caratteristiche delle stazioni idrometriche che hanno funzionato nell'anno.</p> <p>TABELLA 1. — Riporta per alcune stazioni, le altezze idrometriche meridia-</p>	<p>ne rilevate direttamente all'idrometro da parte dell'osservatore oppure dedotte in corrispondenza del mezzogiorno, dallo spoglio dei diagrammi per le stazioni fornite di apparecchio registratore.</p>
---	--

Figura 2.15 - Contenuto delle tabelle delle misure di portata estratto da *Annali Idrologici parte II anno 1951 Compartimento di Napoli*

La portata media giornaliera Q_g e la portata massima annuale istantanea al colmo di piena Q_c sono dati pubblicati rispettivamente negli annali Parte II e nella Pubblicazione n. 17 del Servizio Idrografico "Dati caratteristici dei corsi d'acqua italiani". Le altre serie dell'elenco sono da queste derivate.

Anche per le serie di portate al colmo o per le portate medie giornaliere si possono verificare disomogeneità per effetto non solo dell'utilizzo di nuova strumentazione per la misura, ma anche per effetto di una diversa modalità di calcolo della portata media e della portata al colmo.

La portata liquida in un corso d'acqua è una grandezza non rilevata direttamente ma derivata mediante una trasformazione matematica con la scala di deflusso ed è pertanto affetta da una maggior incertezza rispetto alle grandezze direttamente rilevate



2.2.5 Trasporto solido al fondo e in sospensione

Il trasporto solido al fondo (*bed load*) e in sospensione (*wash load or suspended load*) costituisce un'importante grandezza idrologico-idraulica per la valutazione, ad esempio, della dinamica morfologica dei corsi d'acqua e dei litorali. Tuttavia non esistono attualmente in Italia rilievi sistematici di tali grandezza da costituire serie storiche di lunghezza statisticamente significativa.

Il Servizio Idrografico ha effettuato in passato la pubblicazione sistematica nell'annale Parte II Sezione E dati di trasporto torbido cioè trasporto in sospensione

Le serie temporali di trasporto solido generalmente reperibili sono quelle relative ai sedimenti in sospensione e derivano dai rilievi torbiometrici effettuati nelle stazioni di misura del SIMN:

- Portata torbida massima mensile - $QT_{max,m}$ (kg/s);
- Portata torbida minima mensile - $QT_{min,m}$ (kg/s);
- Portata torbida media mensile - $QT_{med,m}$ (kg/s);
- Torbidità specifica massima mensile - $TS_{max,m}$ (kg/m³);
- Torbidità specifica minima mensile - $TS_{min,m}$ (kg/m³);
- Torbidità specifica media mensile - $TS_{med,m}$ (kg/m³);
- Deflusso torbido mensile - VT_m (ton);
- Deflusso torbido unitario mensile - VTU_m (ton/km²);
- Deflusso torbido annuo - VT_a (ton);
- Deflusso torbido unitario annuo - VTU_a (ton/km²).

I dati riportati sugli annali provengono da analisi su campioni prelevati generalmente una volta al giorno alle ore 12 (in caso di piena venivano prelevati due e anche tre campioni al giorno), nella sezione di misura, nella zona centrale dell'alveo, mediante bottiglia torbiometrica.

2.2.6 Pressione atmosferica

La pressione atmosferica (*atmospheric pressure*), generalmente misurata in *mbar* o in *hPa*, costituisce una grandezza d'interesse prevalentemente meteorologico, e per questo fino a qualche decennio fa era rilevata solo dai servizi meteorologici. Ciononostante, la sua utilità in ambito idrologico è indubbia poiché entra nella stima dell'evaporazione e dell'evapotraspirazione, grandezze essenziali per la conoscenza del bilancio idrico e idrologico.

Attualmente viene rilevata anche dai servizi idrografici (e.g. ISPRA – RMN). Le serie storiche che sono reperibili presso i servizi meteorologici (principalmente quello dell'AM, UGM) sono:

- Pressione atmosferica oraria - B_h ;
- Pressione atmosferica tri-oraria - B_{3h} ;
- Pressione atmosferica media giornaliera - B_g .

Le serie sono in genere prodotte dall'aggregazione di dati con campionamento orario.

2.2.7 Umidità relativa dell'aria

L'umidità relativa dell'aria (*relative humidity*), definita come il rapporto espresso in percentuale tra la quantità di vapore contenuto nell'aria e la quantità massima che la stessa aria può contenere nelle medesime condizioni di temperatura e pressione (cioè a saturazione), è anch'essa una grandezza generalmente di maggiore interesse meteorologico.

Tuttavia attualmente anch'essa viene rilevata dai servizi idrografici (e.g. ISPRA – RMN). Non mancano infatti applicazioni in ambito idrologico come, ad esempio, la valutazione del bilancio idrologico ed in particolare nella valutazione dell'evapotraspirazione.

Le serie temporali di umidità relativa dell'aria generalmente reperibili sono:

- Umidità relativa media oraria - U_h ;
- Umidità relativa media tri-oraria - U_{3h} ;
- Umidità relativa media giornaliera - U_g .

Le serie sono in genere prodotte dall'aggregazione di dati con campionamento orario o sub-orario.

2.2.8 Direzione e velocità del vento

La velocità del vento si rileva generalmente come “vento filato” (*wind run*) espresso in km/giorno e come “raffica” (*wind speed*) espresso in km/h. Il primo dato è equivalente alla velocità media giornaliera e il secondo alla velocità istantanea massima.

Secondo le specifiche del WMO il vento deve essere misurato a 10 m dal suolo e a una distanza pari a 10 volte l'ostacolo più alto nei dintorni per evitare effetti di turbolenza locale.

Le serie temporali di direzione e velocità del vento generalmente reperibili sono:

- direzione e velocità del vento media oraria - DV_h e VV_h ;
- direzione e velocità del vento media giornaliera - DV_g e VV_g ;
- direzione e velocità del vento massima giornaliera - $DV_{g,max}$ e $VV_{g,max}$.

2.2.9 Evaporazione ed evapotraspirazione

Il fenomeno dell'evaporazione è in generale associato a quello della traspirazione delle piante dando origine al fenomeno dell'evapotraspirazione (*evapotranspiration*). Generalmente è espressa in mm (al pari della precipitazione) come volume per unità di superficie. Oltre ad essere una grandezza d'interesse agro-meteorologico costituisce anche una grandezza importante in ambito idrologico poiché costituisce un termine essenziale nella valutazione del bilancio idrico e idrologico.

Rare sono le serie storiche di dati di evapotraspirazione per l'intrinseca difficoltà di misurare direttamente tale grandezza. Più frequentemente si trovano serie di dati di evaporazione rilevate, più agevolmente, mediante evaporimetri.

Le serie temporali di evaporazione e di evapotraspirazione generalmente reperibili sono:

- Evaporazione oraria - E_h ;
- Evaporazione giornaliera - E_g ;
- Evapotraspirazione mensile - ET_m .

In genere le serie storiche di evapotraspirazione sono ricostruite mediante formule empiriche o modelli di bilancio del suolo a partire dalle serie storiche di precipitazione, temperatura e altre grandezze.

2.2.10 Radiazione solare a suolo

La radiazione solare al suolo globale (*global solar radiation*), somma dei contributi di radiazione diretta, diffusa e riflessa, è il flusso di energia, sottoforma di onde elettromagnetiche, che raggiunge la superficie terrestre ed è misurata generalmente in W/m^2 . Più frequentemente la radiazione solare al suolo è espressa come energia totale che raggiunge la superficie terrestre in un intervallo di tempo ed è misurata in kJ/m^2 .

In ambito idrologico è un grandezza utilizzata per la valutazione dell'evapotraspirazione mediante ad esempio la formula di Penman-Monteith (Allen et. al. 1998) o Priestley-Taylor (Priestley e Taylor, 1972).

Le serie temporali di radiazione solare al suolo generalmente reperibili sono:

- Radiazione totale giornaliera - R_{g_g} ;
- Radiazione totale mensile - R_{g_m} ;

Serie temporali di radiazione solare al suolo stimata si possono trovare sul Atlante Italiano della Radiazione Solare dell'ENEA (<http://www.solaritaly.enea.it/index.php>).

2.2.11 Eliofania

L'eliofania (*effective heliophany* o anche *hours of bright sunshine*) misura la durata media del soleggiamento (somma degli intervalli di tempo in cui il sole è visibile e non oscurato dalle nubi) in una località o zona specifica ed è misurata in ore. Si può trovare anche espressa come "eliofania relativa" (*relative heliophany*) come rapporto percentuale tra l'eliofania e la durata astronomica della permanenza del sole sull'orizzonte (*absolute heliophany*) in un dato giorno dell'anno e in una data località.

Le serie di dati di eliofania possono essere utilizzate in ambito idrologico nella valutazione delle serie storiche di evapotraspirazione (ad esempio mediante la formula di Penman-Monteith) o per la valutazione della radiazione solare globale al suolo per la stima del bilancio idrologico o idrico.

Le serie temporali di eliofania e di eliofania relativa generalmente reperibili sono:

- Eliofania giornaliera - EF_g ;
- Eliofania relativa giornaliera - EF_{r_g} ;
- Eliofania giornaliera media mensile - EF_m ;
- Eliofania relativa giornaliera media mensile - EF_{r_m} .

Per ulteriori dettagli sulle modalità di misura delle grandezze idrologiche si può fare riferimento al rapporto "Manuale di riferimento per la misura al suolo delle grandezze idrometeorologiche" del CNR-GNDICI (1993)

3. Anagrafica e metadati

La definizione di uno standard per il trattamento statistico delle serie di dati idrologici, oggetto delle LG, richiede necessariamente anche la definizione di un set minimo d'informazioni di tipo descrittivo, tecnico e amministrativo, per l'identificazione univoca e la caratterizzazione del punto di rilevamento della grandezza idrologica.

Il WMO nelle *“Guidelines on climate metadata and homogenization”* (WMO, 2003a) afferma che: “Good metadata are needed to ensure that the final data user has no doubt about the conditions in which data have been recorded, gathered and transmitted, in order to extract accurate conclusions from their analysis”.

In altri termini è necessario che la serie dei dati idrologici sia accompagnata da un insieme d'informazioni sui dati (*“information about information”*), indicate con il termine “metadato”, nel quale siano indicati ad esempio il nome o un codice univoco, la posizione georiferita del punto di rilevamento, l'ente proprietario dei dati, i diritti di utilizzo, il tipo di strumento di rilevamento, la classe di precisione dello strumento, lo standard di rilevamento, ecc..

La disponibilità di tale insieme d'informazioni consentirebbe un'elaborazione e un'analisi più idonea e soprattutto una più corretta interpretazione dei risultati.

È del tutto evidente che maggiore è la quantità d'informazioni che accompagna la serie di dati, migliore e più corretta sarebbe l'analisi e l'interpretazione dei risultati. Ad esempio, disporre della precisa localizzazione del punto di rilevamento di una grandezza idrologica può fornire utili informazioni circa il contesto ambientale nel quale la grandezza è stata rilevata e quindi poter fornire una corretta interpretazione di eventuali anomalie della serie (trend, salti, ecc.). Anche disporre dell'informazione circa la tipologia dello strumento di rilevamento e della sua classe di precisione, che potrebbe per serie storiche molto lunghe (quasi sicuramente) essere variata nel corso degli anni, consentirebbe di interpretare correttamente andamenti anomali dei dati che potrebbero essere legati proprio al cambio di tipologia di strumento di rilevamento.

Attualmente i dati mete-idrologici (in particolare gli Annali Idrologici), in quanto “dati territoriali di interesse generale” ai sensi del DPCM 10 novembre 2011 (recante *Regole tecniche per la definizione del contenuto del Repertorio nazionale dei dati territoriali, nonché delle modalità di prima costituzione e di aggiornamento dello stesso*), dovrebbero essere corredati di un metadato secondo le specifiche dello stesso DPCM (e derivate dallo standard UNI EN ISO 19115:2005) per essere inseriti nel Repertorio Nazionale dei Dati Territoriali (RNDT), istituito dall'articolo 59¹, comma 3, del DLgs 7 marzo 2005, n. 82 recante “Codice dell'amministrazione digitale”.

In particolare l'art. 1, c. 2, lettera e), del DPCM 10 novembre 2011, definisce “metadato” come le “informazioni che descrivono i dati territoriali e i servizi ad essi relativi e che consentono di registrare, ricercare e utilizzare tali dati e servizi”, mentre l'articolo 59, comma 1, del DLgs 82/2005 definisce i dati territoriali come “qualunque informazione geograficamente localizzata”. La tabella dell'Allegato 1 del DPCM 10 novembre 2011, che riporta l'elenco di tutti i dati territoriali di interesse generale che dovrebbero essere inseriti nel RNDT, al punto 29 indica come dati di interesse generale, come già detto, gli annali idrologici e quindi, per estensione, le serie storiche dei dati idrologici di cui si tratta nelle presenti LG.

L'istituzione del RNDT ha l'obiettivo della diffusione e della condivisione delle conoscenze e delle informazioni di carattere ambientale e territoriale, in una logica d'integrazione tra sistemi informativi ambientali e territoriali esistenti sia all'interno della Pubblica Amministrazione che presso gli altri soggetti che operano sul territorio.

In questa LG tuttavia si propone di corredare la serie storica dei dati idrologici di un metadato costituito da un sott'insieme di quello previsto dallo standard ISO al solo fine di supportare l'elaborazione statistica e l'interpretazione dei risultati.

Si propone quindi un metadato che faccia riferimento essenzialmente a:

¹ Art. 59.

Dati territoriali

1. Per dato territoriale si intende qualunque informazione geograficamente localizzata.

2. E' istituito il Comitato per le regole tecniche sui dati territoriali delle pubbliche amministrazioni, con il compito di definire le regole tecniche per la realizzazione delle basi dei dati territoriali, la documentazione, la fruibilità e lo scambio dei dati stessi tra le pubbliche amministrazioni centrali e locali in coerenza con le disposizioni del sistema pubblico di connettività di cui al decreto legislativo 28 febbraio 2005, n. 42.

3. Per agevolare la pubblicità dei dati di interesse generale, disponibili presso le pubbliche amministrazioni a livello nazionale, regionale e locale, presso il CNIPA e' istituito il Repertorio nazionale dei dati territoriali.

... omissis

- *metadati descrittivi* (anagrafica) che forniscono le informazioni per l'univoca identificazione e geolocalizzazione del sito di rilevamento e della grandezza idrologica rilevata;
- *metadati amministrativi* che forniscono informazioni sul soggetto titolare dei diritti di proprietà intellettuale, sulle modalità di cessione e passaggi per la riproduzione, nonché informazioni circa la disponibilità e il reperimento dei dati delle serie storica (e.g. URL);
- *metadati tecnici* che forniscono informazioni sulla struttura della serie (formato, aggregazione, struttura del record, intervallo campionamento, ecc.), la tipologia degli strumenti di rilevamento e caratteristiche di precisione (classe A, B, ..), modalità di campionamento, validazione dei dati, e quant'altro necessario per una esaustiva caratterizzazione fisica dei dati.

Considerata quindi l'importanza della definizione di opportuni standard per la rappresentazione delle informazioni in Appendice A al paragrafo "Scheda A:" è riportata una proposta di scheda anagrafica e metadati costituita da un sottoinsieme derivata dallo standard UNI EN ISO 19115:2005 e dal DPCM 10 novembre 2011 (relativamente ai soli metadati tecnici) adottato per le informazioni geografiche, che contiene le informazioni minime per un corretta elaborazione e interpretazione della serie storica che dovrebbe costituire parte integrante dell'oggetto "serie storica".

Particolarmente importante per il trattamento statistico dei dati è la conoscenza dei periodi in cui il campionamento dei dati della serie sia avvenuto in maniera omogenea, potendo la serie essere costituita dall'unione di due o più serie di dati derivati da grandezze rilevate nel medesimo punto ma con modalità e strumentazione diverse.

Ad esempio, una serie di dati di portata giornaliera di un corso d'acqua potrebbe, per un primo periodo, essere costituita da dati determinati dal rilevamento giornaliero (alle ore 12:00) dell'altezza idrometrica tramite la curva di deflusso; per un secondo periodo, con strumentazione più recente, potrebbe essere costituita da dati determinati attraverso la media aritmetica dei 24 valori delle portate orarie determinate a partire dal rilevamento orario delle altezze idrometriche; infine con strumentazione avanzata potrebbe essere costituita da dati determinati attraverso la media degli N valori (144 se l'intervallo di campionamento è 10 minuti) delle portate sub-orarie dedotte dal rilevamento sub-orario delle altezze idrometriche. È del tutto evidente che, benché si tratti della medesima grandezza idrologica, potrebbero evidenziarsi delle anomalie statistiche nella serie legate al diverso tipo di campionamento.

Quando tali anomalie risultano particolarmente evidenti sarebbe opportuno trattare le serie separatamente.

È opportuno, per una migliore comprensione del contesto ambientale in cui il dato è rilevato, che le stazioni di rilevamento, oltre alle coordinate spaziali, siano indirizzate anche attraverso i sistemi di geolocalizzazione su WEB tipo Google Maps®, ecc..



È opportuno che le serie storiche di dati idrologici contengano come metadato l'URL da dove sia possibile accedere ai dati sia liberamente sia con accesso limitato.



È opportuno trattare separatamente serie di dati idrologici costituite da dati rilevati in maniera sostanzialmente diversa da determinare anomalie statistiche



La definizione di grandezze derivate può dipendere dall'intervallo di campionamento del dato grezzo che dovrebbe quindi essere specificato nel metadato.



4. Foglio ANÁBASI

A supporto delle Linee Guida è stata sviluppata nel linguaggio VBA (*Visual Basic for Application*, implementazione di Visual Basic inserita all'interno di alcuni software) una *macro* nel software MS Excel 2007, denominata *ANÁBASI* (*ANAlisi statistica di Base delle Serie storiche di dati Idrologici*), nella quale sono implementati il calcolo dei parametri, le procedure e i test statistici proposti nel presente documento.

È opportuno sottolineare che la macro *ANÁBASI* non costituisce un software di statistica, di cui è ricco il panorama commerciale e open source, ma uno strumento semplice e rapido che supporti l'operatore nell'applicazione delle procedure proposte nelle LG.

La scelta di utilizzare un foglio elettronico per implementare le procedure è stata dettata principalmente dalle seguenti ragioni di ordine pratico:

- 1) generale facilità di utilizzo e automazione;
- 2) familiarità degli operatori con questa tipologia di software;
- 3) estrema diffusione del software;
- 4) software generalmente in dotazione agli operatori;
- 5) possibilità di sfruttare le capacità grafiche e di calcolo già implementate.



Figura 4.1 - Pagina iniziale (<Home>) della macro ANÁBASI

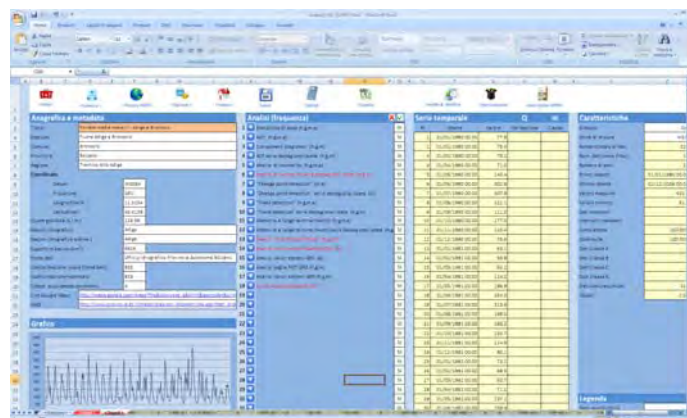


Figura 4.2 - Pagina principale (<Input>) della macro ANÁBASI

La macro è organizzata secondo la nota struttura a “fogli” di Excel.

Nel foglio principale <Input> attraverso una semplice operazione di “copia e incolla” viene inserita la serie storica dei dati idrologici e, contemporaneamente all’operazione di “incollaggio”, ne viene effettuata la verifica della completezza e della validità secondo gli intervalli di convalida definiti dall’utente. I risultati numerici e grafici delle elaborazioni sono memorizzati e visualizzati ciascuno in un foglio separato.

Il manuale di istruzioni del foglio *ANÁBASI*, con tutte le informazioni sulle caratteristiche grafiche e numeriche, non fa parte delle presenti LG (si suppone difatti che le macro di tale software siano aggiornate con più frequenza delle LG) ed è prodotto e pubblicato separatamente.

5. Caratterizzazione statistica di una serie storica

I dati idrologici che costituiscono la serie storica si considerano come valori di una variabile aleatoria (o anche casuale) e che la stessa serie costituisca l'esito di un'estrazione casuale (benché ordinata secondo la variabile temporale) di un numero finito di valori, detto "il campione", da un insieme infinito di possibili valori che generalmente viene indicato come "la popolazione".

Il primo passo da compiere prima di procedere all'elaborazione di una serie storica è quello di effettuare un'analisi visuale, la cosiddetta analisi esplorativa dei dati (*exploratory data analysis*, *EDA*), mediante grafici diagnostici, seguita o preceduta da una caratterizzazione quantitativa, mediante un insieme di parametri che riassumono le principali caratteristiche statistiche dei dati, denominata comunemente statistica descrittiva (*descriptive statistics*).

Approfondimenti sono riportati nell'Appendice B. Approfondimenti di statistica.

5.1 Descrizione statistica

È fondamentale per la descrizione statistica della serie fornirne alcune caratteristiche intrinseche come la lunghezza, la continuità, la completezza e la frequenza.

I parametri che descrivono statisticamente la serie sono riconducibili a tre categorie:

- *posizione*, ossia la tendenza dei dati ad assumere certi valori.
- la *variabilità*, ossia la "mutevolezza" o dispersione dei dati
- la *forma*, vale a dire l'aspetto complessivo della distribuzione rispetto a configurazioni di riferimento (in generale rispetto ad un comportamento della variabile "normale").

Quest'analisi preliminare può sintetizzarsi in una scheda e in alcuni grafici standard.

In Appendice A, nella "Scheda B: descrizione statistica" è riportata una proposta di scheda per un insieme minimo di parametri da utilizzare per la descrizione statistica della serie storica.

Di seguito si riportano le definizioni adottate nella Scheda B: descrizione statistica.

5.1.1 Lunghezza, frequenza e numero di dati della serie

La **lunghezza della serie storica** misurata in anni, indipendentemente dalla frequenza di campionamento, viene definita come la differenza tra l'anno dell'ultimo rilevamento e l'anno del primo rilevamento aumentata di un'unità.

$$\text{lunghezza} = \text{anno primo rilevamento} - \text{anno ultimo rilevamento} + 1 \quad \text{eq. 1}$$

Anche se la serie ha inizio negli ultimi giorni di un anno, quest'ultimo viene comunque considerato come anno d'inizio.

La lunghezza della serie storica è una delle fondamentali caratteristiche della serie poiché definisce la capacità dei dati di fornire informazioni idrologiche affidabili.

La **frequenza della serie storica** viene definita come il numero di dati rilevati in un anno (considerato sempre di 365 giorni).

Tabella 5.1 - Frequenza di una serie storica

Campionamento	Frequenza
Orario	8760
Triorario	2920
Giornaliero	365
Settimanale	52
Decadale	36
Mensile	12
Annuale	1

In genere nelle analisi di serie con aggregazione giornaliera si omette il 29 febbraio. Tuttavia alcuni autori, per tener conto dell'anno bisestile, pongono la frequenza delle serie giornaliera pari a 365,25.

Il **numero massimo di dati** N_m è il numero di dati che dovrebbero essere contenuti nell'intervallo tra il primo rilevamento e l'ultimo con l'assegnata frequenza di campionamento (sono compresi, quindi, anche i dati mancanti). Indichiamo, invece, con N il **numero totale di dati** presente nella serie, con la ovvia condizione che $N \leq N_m$.

5.1.2 Continuità e completezza

Per caratterizzare e quantificare la capacità di una serie idrologica di poter fornire informazioni affidabili, oltre alla lunghezza della serie stessa, è necessario definire alcuni altri semplici parametri. Definiamo **continuità** della serie il valore:


$$\text{continuità} = 1 - 2 \times \frac{\text{numero di intervalli di dati mancanti}}{\text{numero massimo di dati}} \quad \text{eq. 2}$$


La continuità è definita in maniera tale che una serie che presenti tutti i dati validi (uguale al numero massimo di dati) abbia un indice di continuità pari a 1 mentre, dal lato opposto, una serie che presenti un dato valido alternato ad un dato mancante e che quindi presenta il massimo valore di intervalli di dati mancanti (uguale a circa la metà del numero di dati quando $N_m \rightarrow \infty$) abbia valore 0.

La **completezza** è, invece, così definita:

$$\text{completezza} = \frac{\text{numero di dati validi}}{\text{numero massimo di dati}} \quad \text{eq. 3}$$

e fornisce un'indicazione di quanto la serie sia completa, ovvero sia quanti dati validi contenga rispetto alla totalità massima dei dati compresi tra il primo valore rilevato e l'ultimo.

Una serie di dati, ancorché lunga ma interrotta da numerosi intervalli di dati mancanti ovvero con una percentuale elevata di dati mancanti non è in grado di fornire informazioni affidabili. 

Una serie può presentare un elevato indice di completezza ma una bassa continuità qualora presenti molte piccole interruzioni. Viceversa una serie può presentare un'elevata continuità ma una bassa completezza se presenta un'unica ampia interruzione. 

5.1.3 Indici di posizione

5.1.3.1 Media, moda e mediana

Sono i cosiddetti indici di posizione e forniscono, per l'appunto, la posizione sulla scala dei numeri, ovvero l'ordine di grandezza dei valori esistenti nel campione e sono così definiti. Posto $\{X_1, X_2, \dots, X_N\}$ la serie di dati e con N il numero di dati la media campionaria (*sample mean*) è notoriamente definita come:

$$m = \frac{1}{N} \sum_{i=1}^N X_i \quad \text{eq. 4}$$

La moda (*mode*):

$$\text{moda} = \text{valore dell'osservazioni che compare con maggiore frequenza} \quad \text{eq. 5}$$

La mediana (*median*):

$$\text{mediana} = \text{valore in posizione centrale tra i valori osservati ordinati} \quad \text{eq. 6}$$

La media troncata (*truncated mean*) all'X% può essere utilizzata per eliminare l'influenza degli outlier. Si ottiene calcolando la media di un set di dati in cui sono stati eliminati simmetricamente i valori estremi in una percentuale del numero totale di dati. Tale percentuale caratterizza la media troncata. Ad esempio la media troncata al $\alpha\%$ si ottiene da un set di dati in cui sono stati eliminati $\alpha/2\%$ dei dati all'estremo superiore e $\alpha/2\%$ a quello inferiore. È un parametro di posizione robusto che, come la mediana, non risente degli *outlier*.

$$m_\alpha = \frac{1}{(N_{(1-\alpha/2)} - N_{(\alpha/2)} + 1)} \sum_{i=N_{\alpha/2}}^{N_{(1-\alpha/2)}} X_i \quad \text{eq. 7}$$

5.1.3.2 Quantili, percentili, quartili, minimo e massimo

I quantili e i percentili sono anch'essi indici di posizione del campione. Il quantile di ordine α (o in alternativa il percentile che è espresso in percentuale) rappresenta il valore che divide il campione ordinato in due parti di ampiezza pari a α e $(1 - \alpha)$ e caratterizzati da valori minori e maggiori del quantile o del percentile. Ad esempio il percentile 25% è quel valore che divide il campione in parti tali che il 25% dei valori sono inferiori e il 75% sono superiori. È evidente che la mediana, per com'è definita, costituisce il percentile 50% ovvero il quantile $1/2$. Analogamente il valore minimo e il valore massimo di una serie possono essere considerati rispettivamente come percentile 0% e 100%. Particolarmente indicativi sono i percentili 25% e 75% (ovvero quantili $1/4$ e $3/4$) poiché definiscono un significativo indice di variabilità della serie dei dati. I percentili 25%, 50% e 75% sono anche detti quartili e indicati con Q_1, Q_2, Q_3 (rispettivamente 1°, 2° e 3° quartile).

5.1.4 *Indici di dispersione*

5.1.4.1 Varianza, scarto quadratico, range, coefficiente di variazione

I cosiddetti indici di dispersione della distribuzione misurano la variabilità del campione.

La varianza (*variance*) è definita come la media degli scarti quadratici rispetto alla media. Da notare che la somma degli scarti quadratici è divisa per $N - 1$ e non per N . Ciò al fine di ottenere un cosiddetto "stimatore non distorto":

$$s^2 = \frac{1}{N-1} \sum_{i=1}^N (X_i - m)^2 \quad \text{eq. 8}$$

Lo scarto quadratico medio o deviazione standard (*standard deviation*) è la radice quadrata della varianza:

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^N (X_i - m)^2} \quad \text{eq. 9}$$

Il *range* misura l'ampiezza massima del campione dei dati. Non è un indice robusto poiché è molto sensibile agli *outlier*:

$$R = X_{\max} - X_{\min} \quad \text{eq. 10}$$

L'IQR (*inter quartile range*) è invece un indice molto robusto per misurare l'ampiezza della distribuzione dei dati della serie:

$$IQR = Q_3 - Q_1 = P_{75\%} - P_{25\%} \quad \text{eq. 11}$$

La MAD (*median absolute deviation*) costituisce anch'esso un indice molto robusto per misurare la dispersione dei dati.

$$\text{MAD} = \text{mediana}|X_i - \text{mediana}(X_i)| \quad \text{eq. 12}$$

Infine il coefficiente di variazione (*coefficient of variation*) campionario rappresenta la variabilità normalizzata rispetto alla media.

$$CV = \frac{s}{|m|} \quad \text{eq. 13}$$

5.1.5 Indici di forma

5.1.5.1 Asimmetria e curtosi

Altri indici per caratterizzare i dati di una serie sono i cosiddetti indici di forma che forniscono indicazioni sulla forma della distribuzione dei dati.

In particolare si utilizza l'indice di asimmetria (*skewness*) definito come:

$$g = \frac{\frac{1}{N} \sum_{i=1}^N (X_i - m)^3}{\left[\frac{1}{N} \sum_{i=1}^N (X_i - m)^2 \right]^{3/2}} \quad \text{eq. 14}$$

e la curtosi (*kurtosis*) definita come:

$$k = \frac{\frac{1}{N} \sum_{i=1}^N (X_i - m)^4}{\left[\frac{1}{N} \sum_{i=1}^N (X_i - m)^2 \right]^2} - 3 \quad \text{eq. 15}$$

Il primo indice misura la non simmetria, nel senso che se è pari a zero, i dati si distribuiscono in maniera simmetrica. Viceversa quando più è distante da zero tanto più i dati della serie si distribuiscono in maniera asimmetrica. Se l'asimmetria è positiva, la distribuzione dei dati presenterà maggiore frequenza dei valori inferiori alla mediana (coda a destra). Viceversa se l'asimmetria è negativa presenterà valori inferiori alla mediana con minore frequenza (coda a sinistra). Se la distribuzione è simmetrica (e unimodale) media moda e mediana coincidono. Se invece è asimmetricamente positiva, la media è inferiore alla moda e alla mediana. Se è asimmetrica negativamente, la media è superiore alla moda e alla mediana.

In generale i dati idrologici presentano asimmetria positiva. È tuttavia possibile rendere simmetrici i dati delle serie mediante opportune trasformazioni.

Il secondo indice misura l'allontanamento della distribuzione dei dati dalla distribuzione normale di Gauss che presenta un valore di curtosi costante uguale a 3 (qualunque sia la media e la varianza). In particolare misura l'appiattimento della distribuzione ovvero lo spessore delle code rispetto alla distribuzione gaussiana.

5.2 Analisi esplorativa

L'analisi esplorativa dei dati tramite appropriati grafici diagnostici (Tukey 1977; Cleveland 1993, 1994) è una fase fondamentale di ogni studio quantitativo. Spesso la sua importanza è sottovalutata, omettendo l'analisi visiva a favore del calcolo di statistiche sintetiche o di analisi numeriche incapaci di evidenziare importanti aspetti delle serie deducibili solo tramite un controllo visivo diretto. Ciò può condurre peraltro a risultati errati, che potrebbero essere evitati da una semplice analisi esplorativa.

Il modo più semplice di eseguire un'analisi grafica consiste nell'esaminare i dati grezzi per evidenziare problemi (dati mancanti, *outlier*), andamenti temporali (stagionalità, trend, cambiamenti repentini) presenti nei dati.

Quindi oltre al set di parametri per la descrizione statistica della serie vengono definiti alcuni grafici diagnostici standard che sintetizzano graficamente le caratteristiche e il comportamento della serie.

5.2.1 Time plot

Il grafico cronologico o *time plot* costituisce il primo grafico da presentare in una analisi statistica di una serie poiché mediante esso si individua immediatamente il suo comportamento (*pattern*) e le modalità con cui si manifesta nel tempo la grandezza idrologica. La serie storica può essere rappresentata come grafico a linee (Figura 5.1) o come istogramma (Figura 5.2). In particolare quest'ultimo tipo è preferibile per quelle grandezze idrologiche rilevate su intervalli, come ad esempio la precipitazione.

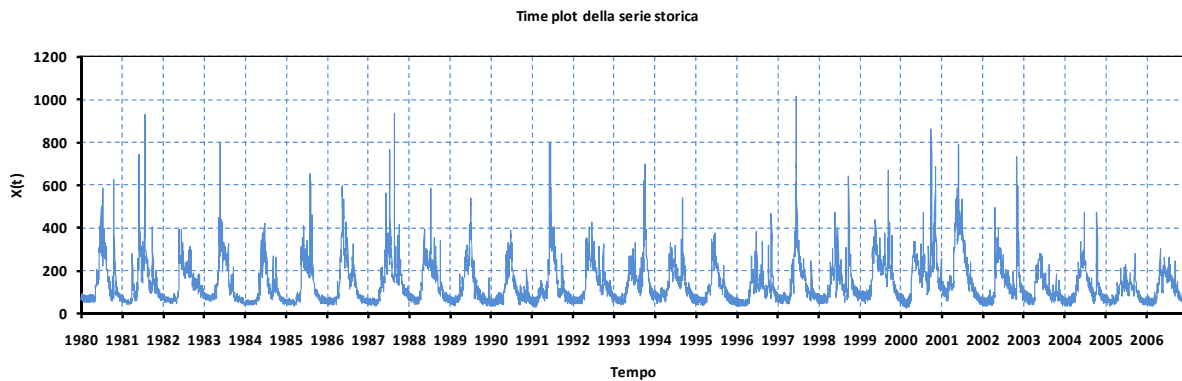


Figura 5.1 - Esempio di time plot per la rappresentazione di una serie di dati

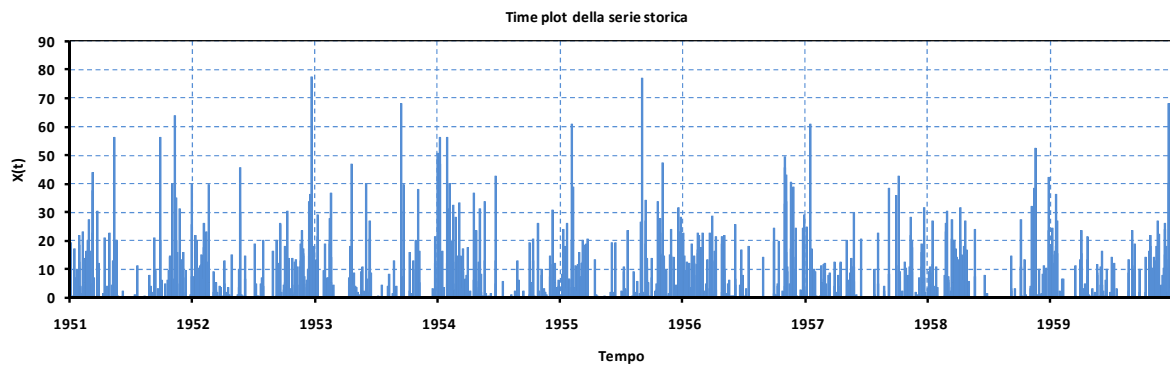


Figura 5.2 - Esempio di istogramma per la rappresentazione di una serie di dati

5.2.2 Distribuzione frequenza campionaria

In questo tipo di grafico si rappresenta la distribuzione della frequenza percentuale del campione suddiviso in classi di valore. La barra dell'istogramma esprime in percentuale il numero di valori della serie, rispetto al numero totale N , compresi in un determinato intervallo. Mediante tale grafico si evidenzia immediatamente l'asimmetria della distribuzione della grandezza. La Figura 5.3a rappresenta per esempio una distribuzione di dati campionari fortemente e positivamente asimmetrica, mentre la Figura 5.3b mostra invece una frequenza campionaria simmetrica.

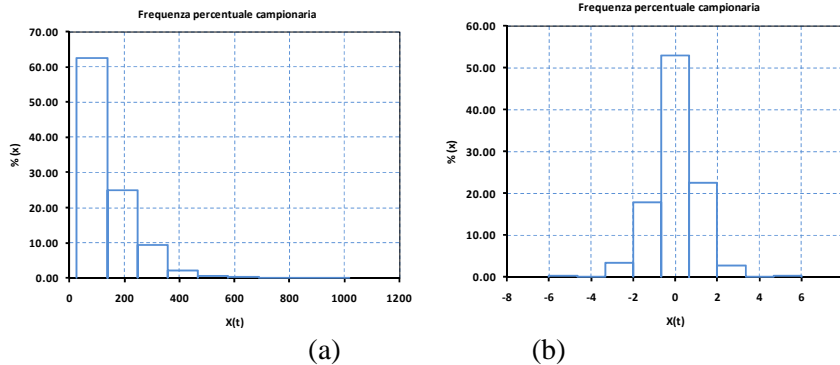


Figura 5.3 - Esempi di istogramma della distribuzione di frequenza percentuale campionaria

5.2.3 Distribuzione frequenza cumulata campionaria

Se con i è indicata la i -esima posizione che occupa il dato della serie ordinata in senso crescente ed N è il numero totale dei dati, allora la frequenza cumulata campionaria è definita dalla relazione:

$$\frac{i}{N} \quad \text{eq. 16}$$

esprimibile anche come percentuale, che indica che lo i -esimo dato ordinato (in senso crescente) supera una percentuale di dati pari $i/N \times 100$

Tuttavia per confrontare la frequenza cumulata campionaria con la probabilità di un determinato modello statistico dei dati, si possono usare anche delle stime della frequenza più raffinate (*Plotting Position*), generate da una formulazione del tipo: $F = (i - \alpha)/(N + 1 - 2\alpha)$ con α compreso tra 0 e 1. Le più comuni stime di questo tipo sono la formula di Weibull ($\alpha = 0$):

$$\frac{i}{N + 1} \quad \text{eq. 17}$$

la formula di Hazen ($\alpha = 0.5$):

$$\frac{i - 0.5}{N} \quad \text{eq. 18}$$

e la formula di Gringorten ($\alpha = 0.44$), che è ottimizzata per l'analisi dei valori estremi:

$$\frac{i - 0.44}{N + 0.12} \quad \text{eq. 19}$$

Il diagramma che riporta la frequenza cumulata campionaria in funzione del valore della grandezza idrologica fornisce una rappresentazione immediata della distribuzione dei dati.

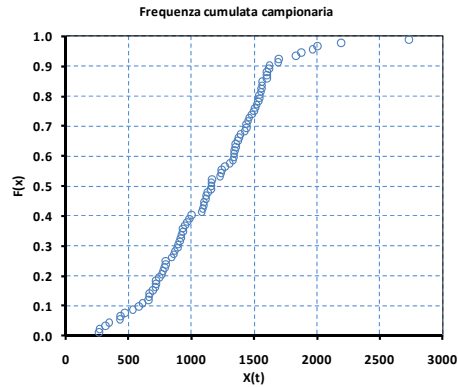


Figura 5.4 - Esempio di distribuzione di frequenza cumulata campionaria.

5.2.4 Box plot

Questo tipo di grafico costituisce una rappresentazione grafica dei dati che descrivono la distribuzione di un campione mediante alcuni indici di posizione.

È costituito da un rettangolo diviso in due parti. I lati opposti del rettangolo sono posti in corrispondenza dei percentili 25% e 75% (1° e 3° quartile) mentre la divisione è posta in corrispondenza della mediana o, come detto in precedenza, del percentile 50% (2° quartile).

Dal rettangolo escono due segmenti, cosiddetti baffi, che terminano al valore minimo e al valore massimo.

In definitiva per costruire un box plot occorrono cinque valori: Min, Q_1 , Q_2 , Q_3 e Max (Figura 5.5a).

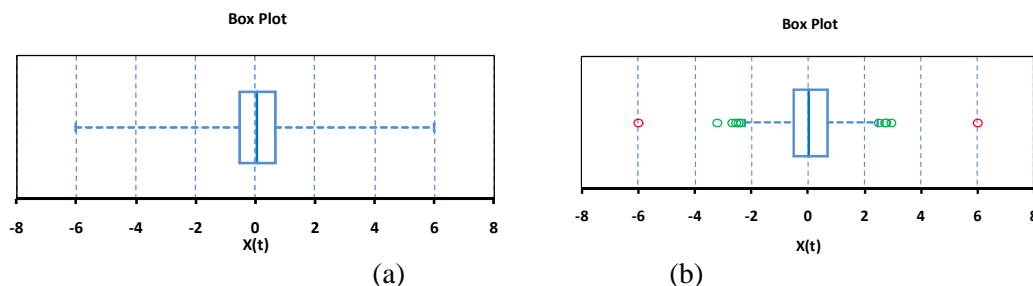


Figura 5.5 - Esempi di box-plot per la rappresentazione statistica di un set di dati. A destra il box plot con l'indicazione in verde di valori estremi e in rosso di valori anomali.

La distanza interquartile è una misura della dispersione della distribuzione. Il 50% delle osservazioni infatti si trovano comprese tra questi due valori. Se l'intervallo interquartile è piccolo, tale metà delle osservazioni si trova fortemente concentrata intorno alla mediana; all'aumentare della distanza interquartile aumenta la dispersione del 50% delle osservazioni centrali intorno alla mediana.

Le distanze tra ciascun quartile e la mediana forniscono informazioni riguardo alla forma della distribuzione. Se una distanza è diversa dall'altra, allora la distribuzione è asimmetrica.

Se si indica con $r = (Q_3 - Q_1)$ il range interquartile, il valore adiacente inferiore (VAI) è il valore più piccolo tra le osservazioni che risulta maggiore o uguale a $Q_1 - 1,5r$.

Il valore adiacente superiore (VAS), invece, è il valore più grande tra le osservazioni che risulta minore o uguale a $Q_3 + 1,5r$. Pertanto se gli estremi della distribuzione sono contenuti tra $Q_1 - 1,5r$ e $Q_3 + 1,5r$ essi coincideranno con gli estremi dei baffi, altrimenti come estremi verranno usati i valori $Q_1 - 1,5r$ e $Q_3 + 1,5r$.

La definizione del VAS e del VAI potrebbe fornire un criterio per stabilire quando un dato può considerarsi anomalo.

Potrebbe, ad esempio, essere considerato valore estremo un valore con scostamento positivo dal 3° quartile superiore a 1,5 volte il range interquartile o, simmetricamente, un valore con scostamento negativo dal primo quartile superiore (in valore assoluto) a 1,5 volte il range interquartile (valori riportati in verde in Figura 5.5b).

Potrebbe essere considerato valore anomalo un valore con scostamento (positivo) dal terzo quartile o (negativo) dal primo quartile superiore a 3 volte il range interquartile (valori riportati in rosso in Figura 5.5b).

Un'alternativa per definire quando un dato può considerarsi anomalo è quello di considerare l'intervallo con ampiezza pari a 2 o 3 volte la deviazioni standard. Ciò tuttavia presuppone una assunzione di modello normale dei dati.

5.3 Outlier e robustezza delle statistiche di una serie

Gli *outlier* sono dei valori nella serie talmente distanti dalla maggior parte dei dati della serie stessa che, quando non siano errati, sono da ritenersi anomali ovvero appartenenti a una diversa popolazione e trattati in maniera opportuna.

I valori anomali possono influenzare molte statistiche, come la media o la deviazione standard. Essi possono anche influenzare gli indici di associazione tra le variabili come il coefficiente di correlazione di Pearson.

In presenza di casi anomali che influenzano i risultati delle analisi è possibile utilizzare delle statistiche sintetiche che per il fatto di essere meno influenzate dalla presenza di tali valori si dicono robuste. Ad esempio, la mediana spesso può risultare più affidabile della media. Sono inoltre disponibili alcune misure di sintesi che risultano robuste alla presenza di tali valori, come ad esempio la media troncata.

Né risulta corretto, a meno di evidenti ragioni e fondate motivazioni, rimuovere di valori anomali che risultano influenti, ovvero che hanno un impatto eccessivo sulle misure di sintesi che si vogliono considerare.

L'utilizzo della media come indice di posizione e della varianza come indice di dispersione sono adatti se i dati fossero distribuiti secondo una curva di Gauss. I dati idrologici in generale sono lontani dall'essere distribuiti come una normale e quindi sarebbe più corretto descriverli mediante i quantili.

I dati idrologici per la loro non-normalità e per la presenza di *outlier* sarebbero più correttamente descritti mediante la mediana e i quantili.



5.4 Trattamento dei dati mancanti

I dati mancanti (*missing values*), dovuti ad esempio al malfunzionamento dello strumento di misura o all'interruzione del servizio di rilevazione per manutenzione o per altre ragioni, rappresentano un problema molto frequente nelle serie storiche di dati idrologici che può portare ad analisi poco significative, fortemente distorte e/o addirittura errate.

Il problema dei dati mancanti ha implicazioni sulle serie derivate mediante aggregazione. Ad esempio, uno dei problemi più frequenti è stabilire quale percentuale di dati mancanti di precipitazione giornaliera è accettabile per poter considerare affidabile il dato della precipitazione aggregata annua; oppure quanti dati di precipitazione oraria sono necessari per considerare accettabile il dato giornaliero; o, ancora, quale percentuale di dati mancanti è ammissibile per poter estrarre il valore massimo (o minimo) da un campione non completo. E così via.

Gli intervalli di valori mancanti possono essere di vario tipo: ad esempio, intervalli continui di lunghezza rilevante rispetto a quella della serie completa dovuti presumibilmente ad azioni sistematiche, quali l'interruzione della misura o la perdita degli archivi, oppure valori isolati distribuiti in modo pressoché casuale all'interno della serie e legati prevalentemente a disfunzioni temporanee degli strumenti di misura. Naturalmente, le varie tipologie di valori mancanti possono essere presenti contemporaneamente in una serie e combinarsi in modo più o meno complesso.

Una classificazione delle tipologie dei *missing value* è chiaramente di difficile realizzazione, così come la definizione di un insieme di metodologie univoche per il loro trattamento. Oltre alla causa che può avere generato un valore mancante occorre, infatti, considerare anche la natura della grandezza registrata (pioggia, temperatura, portata, ecc.), la sua scala di risoluzione temporale (oraria, giornaliera, mensile, ecc.), nonché l'effetto dei *missing value* sulle analisi che si desidera condurre sulla serie studiata.

Il problema della possibilità di ricostruzione dei dati idrologici procede comunque parallelamente al problema dell'interpolazione spaziale dei dati che sarà oggetto di un altro rapporto tematico; sia per la possibilità di poter ricostruire il dato mancante da quelli rilevati allo stesso istante in punti circostanti, sia per la verifica e la cross-validazione dei dati ricostruiti mediante le tecniche di ricostruzione sopra citate.

Data la particolare complessità del problema, esso verrà affrontato estesamente in una successiva versione del presente LG, parallelamente alla definizione di linee guida per l'interpolazione spaziale dei dati con l'obiettivo di definire un approccio standard e più possibile semplice per ricostruire i dati mancanti nelle analisi delle serie dei dati idrologici.

Nei capitoli successivi al termine della descrizione delle analisi statistiche proposte si riportano alcune considerazioni relative alla presenza dei valori mancanti e alla loro possibile influenza sulle analisi stesse, con l'obiettivo di fornire delle indicazioni, se pur di massima, sulla percentuale di *missing value* ammissibile per l'applicazione delle analisi statistiche senza il ricorso a tecniche di ricostruzione dell'informazione mancante.

6. Qualità della serie

Una proposta di caratterizzare la qualità di una serie di dati si ritrova già nel progetto “*Analyzing Long Time Series of Hydrological Data with Respect to Climate Variability*” del WMO nell’ambito del World Climate Programme (WMO, 1988), nel quale si proponeva di definire:

- 1) molto buona: una serie di dati rilevati in una stazione della quale era ben nota la storia, i cui dati erano stati verificati in relazione all’omogeneità e per i quali non era stata ritenuta necessaria alcuna correzione;
- 2) buona: una serie di dati rilevati in una stazione della quale era ben nota la storia ma i cui dati erano stati omogeneizzati (corretti);
- 3) accettabile: una serie di dati rilevati in una stazione della quale non era ben nota la storia e i cui dati non erano stati omogeneizzati (corretti) ma comunque potevano essere ritenuti accettabili.

Si evidenzia, quindi, l’importanza di considerare non solo la qualità del singolo dato attraverso procedure di validazione, ma anche la qualità dell’insieme dei dati: una serie con tutti i dati validati, infatti, può presentare caratteristiche che ne riducono la qualità complessiva (e.g. elevato numero di dati mancanti, non omogeneità prodotta con strumentazione diversa nel corso del tempo, ecc.).

Viene qui di seguito riportata una proposta di definizione di un “**indice di Qualità della Serie di dati Idrologici**” *iQuaSI* basato sulla lunghezza della serie espressa in anni (dal primo all’ultimo rilevamento) e sulla classe di qualità del singolo dato.

A tale scopo, possiamo supporre di classificare i dati in base a quattro classi di qualità opportunamente definite:

- ✓ classe A: dati rilevati direttamente con strumentazione di elevata accuratezza (< 3%) (e.g. pluviometro registratore elettronico in perfetta efficienza);
- ✓ classe B: dati rilevati con strumentazione con media accuratezza (3–5%) (e.g. pluviometro registratore meccanico);
- ✓ classe C: dati rilevati con strumentazione con bassa accuratezza (> 5%) o stimati mediante grandezze indirette (e.g. pluviometro semplice o totalizzatore, radar meteo per la precipitazione, portata stimata mediante scale di deflusso, ecc.);
- ✓ classe D: dato mancante o ricostruito mediante modellistica matematica.

L’indice che qui si propone, compreso tra 0 e 1, è definito come una combinazione lineare dei rapporti tra la lunghezza della parte della serie costituita da dati di una determinata classe di qualità e la lunghezza totale, con coefficienti dipendenti dalla lunghezza della serie.

$$iQuaSI = a_L \times \left(\frac{L_A}{L} \right) + b_L \times \left(\frac{L_B}{L} \right) + c_L \times \left(\frac{L_C}{L} \right) + d_L \times \left(\frac{L_D}{L} \right) \quad \text{eq. 20}$$

Nella Tabella 6.1 sono riportati i coefficienti utilizzati nell’eq. 20 per la determinazione dell’indice di qualità della serie, definiti per ciascuna classe di dati e in funzione della lunghezza della serie; mentre nella Tabella 6.2 sono riportate le classi di qualità della serie corrispondenti ai possibili valori assunti dall’*iQuaSI*.

Tabella 6.1 - Coefficienti per la determinazione dell’indice di qualità della serie

		Lunghezza della serie (anni)			
		$L \geq 30$	$15 \leq L < 30$	$5 \leq L < 15$	$L < 5$
Classe dei dati	A	1	3/4	1/2	0
	B	3/4	1/2	1/4	0
	C	1/2	1/4	0	0
	D	0	0	0	0

Tabella 6.2 - Indice di qualità della serie *iQuaSI*

Qualità della serie	<i>iQuaSI</i>
<i>Elevata</i>	$0.90 < iQuaSI \leq 1$
<i>Buona</i>	$0.70 < iQuaSI \leq 0.90$
<i>Sufficiente</i>	$0.30 < iQuaSI \leq 0.70$
<i>Scadente</i>	$0.10 < iQuaSI \leq 0.30$
<i>Pessima/Inutilizzabile</i>	$0 \leq iQuaSI \leq 0.10$

Di seguito si riportano alcuni esempi per meglio comprendere l'indice di qualità della serie:

- 1) una serie risulta di qualità *elevata* quando sia costituita da dati di classe A rilevati per non meno di 30 anni e ammettendo eventualmente dati di classe inferiore (classe B, C, e D) per non più di 3 anni (10%);
- 2) una serie risulta di qualità *buona* se costituita da dati di classe A rilevati per un numero di anni compreso tra 15 e 30, ovvero costituita da dati di classe B rilevati per un numero di anni maggiore di 30 con pochissimi dati mancanti (meno del 2-3%);
- 3) una serie risulta di qualità *sufficiente* se costituita da dati di classe A rilevati per un numero di anni compreso tra 5 e 15, ovvero costituita da dati di classe B rilevati per un numero di anni compreso tra 15 e 30 con un numero di dati mancanti superiore al 5%;
- 4) una serie risulta di qualità *scadente* se costituita da dati di classe B rilevati per un numero di anni compreso tra 5 e 15, ovvero costituita da dati di classe C rilevati per un numero di anni maggiore tra 15 e 30 con un numero di dati mancanti inferiore al 5%;
- 5) una serie costituita da un numero di dati rilevati per un periodo inferiore a 5 anni è sempre inutilizzabile indipendentemente dalla classe del dato rilevato.

L'importanza di associare a ogni serie di dati un valore di *iQuaSI* deriva dalla necessità di valutarne l'utilizzo per le diverse applicazioni.

Le serie con *iQuaSI* = *elevato*, per esempio, potrebbero essere quelle che rispondono alle esigenze di applicazioni che richiedono un'accuratezza elevata del singolo dato ma anche una adeguata lunghezza e una completezza e continuità elevata della serie (e.g. studi sui dei cambiamenti climatici, ecc.).

Le serie con *iQuaSI* = *buono*, invece, potrebbero rispondere alle esigenze di applicazioni che richiedono un'accuratezza media e una media lunghezza (e.g. idrologia, stime valori estremi, ecc.).

Le serie con *iQuaSI* = *sufficiente* rispondono alle esigenze di applicazioni che richiedono un'accuratezza bassa, ovvero una numerosità della serie non elevata, come nel caso di valutazioni e stime tecniche con grossi margini di incertezza intrinseca, applicazioni di agronomia, valutazioni e stime del bilancio idrico.

Non sempre è possibile conoscere la classe di qualità dello strumento con cui è stata rilevata una grandezza, soprattutto per serie molto estese e per i periodi iniziali molto lontani nel tempo. In questo caso si dovrebbe assumere una classe non superiore alla B.

Una serie con tutti i dati validati può presentare caratteristiche che ne riducono la qualità complessiva, ad esempio, se è elevato il numero di dati mancanti o se le serie non sono omogenee poiché prodotte con strumentazione diversa.



7. Analisi statistica di base

L'analisi statistica di base è volta a valutare:

1. la dipendenza temporale tra i dati tramite la funzione di autocorrelazione;
2. la "normalità dei dati";
3. la "memoria lunga" della serie;
4. la stazionarietà della serie con l'individuazione:
 - delle componenti stagionali della serie;
 - dei cambiamenti repentini (*change point*);
 - dei cambiamenti gradualmente (*trend*);
5. il comportamento dei valori estremi.

Nell'Appendice B. "Approfondimenti di statistica" è riportato un approfondimento delle procedure indicate nel presente capitolo.

È opportuno che i dati idrologici siano analizzati, per quanto possibile, usando metodi non parametrici (*distribution free*) cioè metodi che non si basano sull'assunzione che i dati seguano una particolare distribuzione di probabilità. Questo perché i dati idrologici sono spesso non-normali, mentre molti metodi sono basati sull'ipotesi di normalità. L'uso di metodi parametrici è comunque possibile ma ciò richiede trattamenti statistici avanzati (WMO, 2000, pag.60)



7.1 Autocorrelazione

L'autocorrelazione definisce il grado di dipendenza lineare tra i dati di una serie. Si esprime attraverso una funzione, detta appunto di autocorrelazione (**ACF** – *Auto Correlation Function*).

La funzione di autocorrelazione di una serie stazionaria si può stimare usando la seguente formula:

$$\rho[k] = \frac{\text{Cov}[k]}{\text{Cov}[0]} \quad \text{eq. 7.1.1}$$

in cui k è il *lag* temporale (numero di intervalli temporali tra gli elementi della serie).

Essa risulta particolarmente importante per le analisi preliminari delle serie storiche. La forma di tale funzione infatti fornisce una chiara rappresentazione della struttura di dipendenza permettendo di cogliere visivamente alcune proprietà fisiche generali dei dati idrologici analizzati (stagionalità, legame temporale nel breve e lungo periodo, non stazionarietà).

Solo per dare una pratica interpretazione della funzione, se $\rho[1] = 0.7$ significa che in media tutti i dati delle serie distanti (o traslati di) 1 passo temporale presentano, mediamente, una forte dipendenza. Al variare del *lag* può variare il valore della dipendenza dei dati (solitamente più debole al crescere della distanza, ma con le dovute eccezioni nel caso di dati affetti da stagionalità o non stazionari).

Dato un *lag* temporale k , il valore di autocorrelazione $\rho(k)$ della serie $\{x_1, x_2, \dots, x_N\}$ è ottenuto calcolando la correlazione tra la serie $\{x_1, x_2, \dots, x_{N-k}\}$ e la serie $\{x_{1+k}, \dots, x_N\}$, ossia tra la serie in cui vengono eliminati gli ultimi k termini e quella ottenuta eliminando i primi k termini.

La funzione $\rho(k)$ assume valori sempre compresi tra -1 e 1 , e per $k = 0$ si ha che $\rho(0) = 1$, ossia una serie è perfettamente correlata linearmente con se stessa. Il grafico dell'ACF ha sempre quindi il valore nell'origine dei tempi (o dei *lag*) pari a 1 (Figura 7.1).

Si noti che per valori di *lag* temporali superiori alla metà della lunghezza N della serie, la significatività statistica del coefficiente di autocorrelazione ρ viene meno in quanto ottenuto utilizzando un numero di dati molto inferiore a quello della serie analizzata. Pertanto, è consigliabile effettuare il calcolo della funzione di autocorrelazione al massimo fino al $\text{lag} = N/4$ per avere serie sufficientemente lunghe che forniscono valori più affidabili del coefficiente di correlazione.

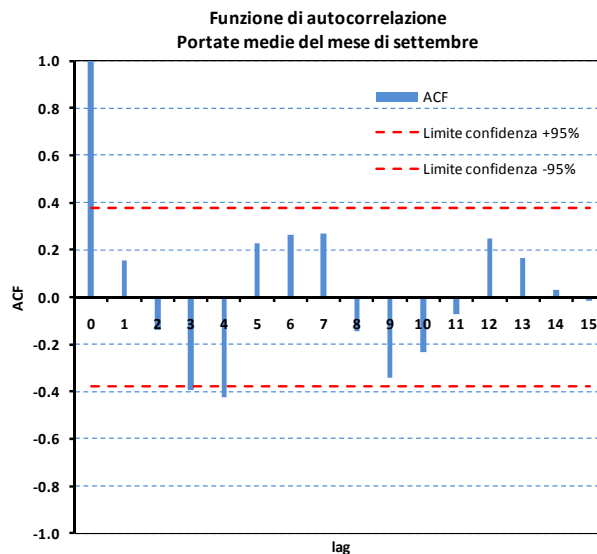


Figura 7.1 - Esempio di una funzione di autocorrelazione per le portate medie mensili di un singolo mese per ciascun anno (frequenza 1). I valori assoluti del coefficiente di correlazione per lag superiori a 0 sono inferiori a 0.4.

La forma di tale funzione fornisce una chiara rappresentazione della struttura di dipendenza temporale dei dati della serie, permettendo di cogliere visivamente alcune proprietà fisiche generali dei dati idrologici analizzati:

- stagionalità;
- legame temporale nel breve e lungo periodo;
- non stazionarietà, ecc.

Ad esempio, se si considera la serie storica delle portate giornaliere di un corso d'acqua che sottende un ampio bacino idrografico, il valore del coefficiente $\rho(1)$ è in genere molto alto (Figura 7.2a). Ciò è giustificabile in considerazione del fatto che se un giorno dell'anno si manifesta un valore di portata di $500 \text{ m}^3/\text{s}$, il giorno dopo è presumibile che il valore sia analogo. L'insieme e la complessità dei processi fisici agenti sul bacino idrografico (deflussi superficiali, deflussi sub-superficiali, deflusso profondo) limitano la variabilità dei deflussi creando una forte dipendenza dei dati a scala giornaliera. Se invece si considera una serie giornaliera delle precipitazioni, la funzione di autocorrelazione (Figura 7.2b) mostra in genere i primi valori significativi con valori modesti di ρ e un veloce decadimento a zero dopo pochi passi.

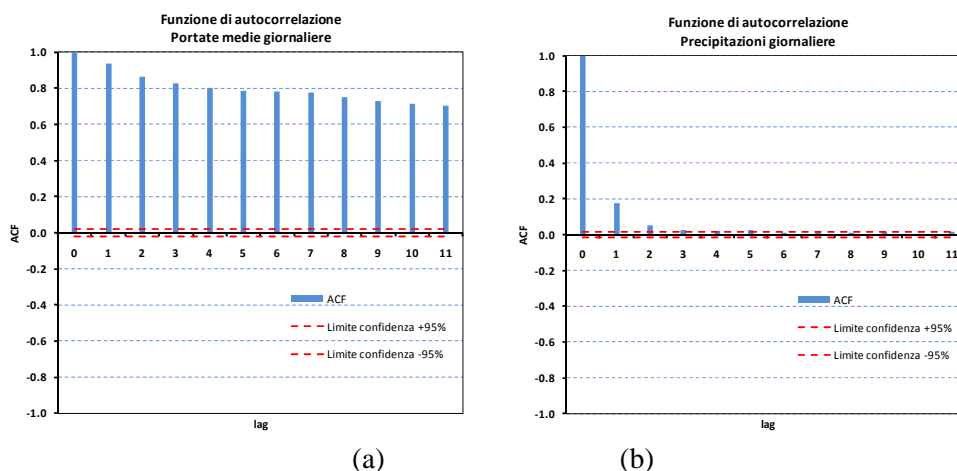


Figura 7.2 - Confronto delle funzioni di autocorrelazione di una serie di portate (a destra) e una serie di precipitazioni (a sinistra) a scala giornaliera

Tale andamento rispecchia il fenomeno fisico degli eventi di pioggia che in media si manifestano per intervalli brevi (da 1 a 5 giorni). La presenza in una serie osservata di pochi eventi superiori a 4-5 giorni di durata fa sì che statisticamente la correlazione tra i dati per un intervallo di durata maggiore

sia pressoché nulla. In sostanza la funzione ACF suggerisce che se in un determinato giorno piove, è presumibile che anche il giorno successivo possa piovere, ma nulla si può dire sull'eventualità di una precipitazione dopo 6 giorni.

Ovviamente la correlazione temporale espressa dalla funzione ACF può essere di varia natura. Oltre al comportamento nel breve periodo dovuto principalmente alle dinamiche di base del processo (genesi e propagazione di un evento meteorico, formazione di una piena) si possono riscontrare anche evidenze di processi fisici più generali (cambiamenti climatici, stagionalità).

Le linee tratteggiate in rosso nelle figure relative alla funzione ACF rappresentano i cosiddetti limiti di confidenza a livello α (nel caso specifico al 95%) che indicano i limiti entro i quali con probabilità α ricadono i valori che possono essere considerati nulli (e quindi non correlati) poiché il valore diverso da zero è solo un effetto del campionamento limitato.

Una serie è correlata quando un dato dipende da un certo numero di dati precedenti. Il numero di dati precedenti che influenzano il valore al tempo t è rilevabile dall'ACF.

Per approfondimenti circa l'ACF si veda Appendice B, paragrafo 11.1.4

7.1.1 Test Ljung-Box per la presenza di autocorrelazione

Come sarà chiarito in seguito, quando si affronteranno l'analisi dei *change point* e dei *trend*, è importante verificare se, in maniera statisticamente significativa, una serie è non autocorrelata per i primi s lag.

A tal fine si utilizza il test di **Ljung Box** o, equivalentemente, il test di **Box-Pierce** (si veda Appendice B paragrafo 11.4.3.1). In entrambi i test, l'ipotesi nulla H_0 è che l'ACF per i primi s lag non sia significativamente diversa da zero, ossia che la serie storica non presenta un'autocorrelazione significativa.

Quando l'esito del test è quello per cui l'ipotesi nulla H_0 non è rigettabile rispetto al livello di significatività fissato α , la serie non è significativamente autocorrelata per i lag testati e l'errore che si commetterebbe nel non rigettare H_0 quando questa è falsa (errore di II tipo) ha probabilità β . La quantità $1-\beta$ si chiama *potenza del test* ed esprime la capacità di un test statistico di riconoscere la falsità di H_0 quando questa è effettivamente falsa.

Quando l'esito del test è invece quello per cui l'ipotesi nulla H_0 è rigettabile, la serie è significativamente autocorrelata per i lag testati e l'errore che si commetterebbe nel rigettare H_0 quando questa è vera (errore di I tipo) ha probabilità α .

7.2 Verifica della “normalità” dei dati

Poiché molte delle procedure statistiche sono derivate dall'ipotesi di normalità dei dati (e.g., test sul trend di Pearson), è utile verificare se i dati siano effettivamente distribuiti significativamente secondo la legge normale.

In realtà, come è noto, i dati idrologici non sono quasi mai distribuiti secondo la legge normale ma anzi presentano caratteristiche come l'asimmetria, talvolta anche molto accentuata, che li allontanano dalla “normalità”. Possono, tuttavia, mediante un'opportuna trasformazione (e.g., Box e Cox, lognormale, quantile normale) distribuirsi secondo una normale.

Nel foglio ANÁBASI i dati della serie possono essere trasformati con la relazione di Box e Cox che ha la seguente espressione:

$$x_i^{(\lambda)} = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \log x_i & \text{se } \lambda = 0 \end{cases} \quad \text{per } i = 1, \dots, N \quad \text{eq. 7.2.1}$$

in cui il parametro λ viene stimato imponendo che l'asimmetria dei dati trasformati sia nulla.

Mediante il grafico, come quello riportato in Figura 7.3 - **Grafico per la stima del parametro λ della trasformata di Box e Cox**. Figura 7.3, in cui il coefficiente di asimmetria dei dati è posto in funzione del coefficiente della trasformazione di Box e Cox, si individua il valore di λ che annulla l'asimmetria dei dati.

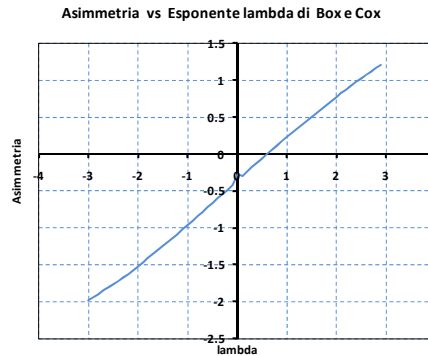


Figura 7.3 - Grafico per la stima del parametro λ della trasformata di Box e Cox.

Per verificare la “normalità” dei dati si utilizza il test di **Jarque-Bera** basato sulle misure delle due particolari caratteristiche della distribuzione normale e cioè l’asimmetria, che assume sempre valore nullo ($\gamma = 0$), e la curtosi, che assume anch’essa sempre valore nullo ($k = 0$).

L’ipotesi nulla H_0 del test è che l’asimmetria e la curtosi siano significativamente nulle. Pertanto quando il livello di significatività osservato detto p -valore (p -value) è maggiore del livello di significatività α fissato, la statistica ricade nella zona di non rigetto e quindi i dati sono significativamente normali. Quando il p -value è invece minore del livello di significatività α la statistica ricade nella zona di rigetto, per cui i dati si possono significativamente considerare non normali e l’errore che si commetterebbe, qualora l’ipotesi nulla H_0 fosse vera, è pari al livello di significatività.

7.3 Analisi della “lunga memoria”

Oltre alla funzione di autocorrelazione, un’importante analisi dei dati idrologici che si propone di effettuare nelle presenti LG è quella che consiste nel verificare la cosiddetta “lunga memoria” della serie dei dati, per stabilire se essi risentano dei valori assunti in tempi anche oltre i *lag* analizzati nella funzione di autocorrelazione.

La presenza di fluttuazioni di periodo molto ampio, per le quali le serie si manifestano localmente non-stazionarie pur essendo il segnale globalmente stazionario, introduce il concetto che solitamente è denominato *lunga memoria*. I principali elementi distintivi di una serie di tale natura si riscontrano nella funzione di autocorrelazione che presenta un decadimento lento con una forma estesa. Il concetto di lunga memoria è strettamente correlato con la definizione del parametro di Hurst.

A tal fine si calcola, con i metodi descritti in Appendice B l’esponente di Hurst, indicato con H e compreso tra 0 e 1, i cui valori vanno così interpretati:

- $H = 0.5$: la serie ha un comportamento stazionario;
- $H > 0.5$: la serie ha un comportamento definito persistente. In questo caso, un trend positivo (negativo) nel passato implica un trend positivo (negativo) in futuro;
- $H < 0.5$: la serie ha un comportamento definito antipersistente. In questo caso, un trend positivo (negativo) nel passato implica un trend negativo (positivo) in futuro.

I comportamenti persistenti (antipersistenti) sono tanto maggiori quanto più H è prossimo rispettivamente all’unità (allo zero). In questi casi, la correlazione delle osservazioni non si annulla mai, per cui le osservazioni passate influiscono su quelle future (in senso positivo o negativo), indipendentemente dal *lag* che le separa. Da questa proprietà deriva la definizione di *lunga memoria*. Ad esempio, nelle serie storiche delle portate dei corsi d’acqua il parametro di Hurst è generalmente maggiore di 0.5 evidenziando un comportamento persistente. In particolare, Hurst ha stimato per il fiume Nilo, per il quale scoprì l’effetto memoria, un esponente $H = 0.9$ (molto persistente).

Il calcolo di H , come detto nel paragrafo 11.4.1, dovrebbe essere eseguito su serie di lunghezza almeno di 100 elementi. Fissando l’attenzione sulle scale di risoluzione temporale più comuni (giornaliera, mensile e annuale), in generale, poiché H misura la lunga memoria del sistema, la presenza di valori mancanti in percentuale limitata (e.g., 5-10% della lunghezza della serie) non dovrebbe influenzare in modo determinante il calcolo. Ciò in ragione del fatto che i metodi di stima si basano su procedure di aggregazione progressiva dei dati su intervalli di tempo crescenti, e il calcolo è

effettuato escludendo gli intervalli più brevi (in cui si risente dell'effetto della memoria di breve periodo) e quelli più lunghi (in cui si ha un'elevata incertezza della stima dovuta alla riduzione della lunghezza del campione). Per dati annuali, la cui disponibilità supera raramente alcune decine di anni, il calcolo di H è sconsigliato in quanto decisamente poco affidabile a causa della scarsa numerosità del campione; in ogni caso i dati non dovrebbero contenere più di 1-3 valori mancanti isolati.

7.4 Stazionarietà

Gran parte delle usuali analisi statistiche in ambito idrologico sono basate sull'assunzione della stazionarietà della serie dei dati. Ciò significa che tutti i dati della serie, considerati come variabili aleatorie, dovrebbero provenire da un'unica popolazione caratterizzata da un'unica distribuzione di probabilità.

La verifica della stazionarietà di una serie di dati idrologici dovrebbe costituire, quindi, la fase propedeutica per ogni valutazione idrologica. In realtà, già l'analisi esplorativa e visuale dei dati, con la semplice visualizzazione del *time plot* della serie, potrebbero fornire informazioni sulle caratteristiche di non-stazionarietà, almeno per quelle più evidenti.

Per gli scopi delle presenti LG, una serie si può assumere stazionaria se la media e la varianza risultano costanti nel tempo (stazionarietà in senso debole). Tuttavia non è né semplice né immediato dal punto di vista pratico stabilire la stazionarietà di una serie a partire dal numero finito di dati della serie stessa. A tale scopo si ricorre a procedure di test che forniscono valutazioni oggettive, ma sempre di natura probabilistica, che possono supportare l'interpretazione del comportamento dei dati.

Molteplici possono essere per le serie di dati idrologici le cause di non stazionarietà e le modalità con cui esse si manifestano. Ad esempio, la non stazionarietà può interessare il valor medio e/o la varianza e/o i valori estremi con modalità (*pattern*) che possono essere essenzialmente di tipo:

- 1) graduale (*trend*);
- 2) repentino (*change point*);
- 3) stagionale (*seasonal*);
- 4) ciclico non stagionale (*cyclic not seasonal*);
- 5) combinazione dei precedenti.

Nel primo caso le cause della non stazionarietà potrebbero essere, ad esempio, ascrivibili a una variabilità legata a fenomeni di cambiamenti climatici sul lungo periodo, ovvero a fenomeni evolutivi come urbanizzazione o deforestazione. Ma fenomeni di trend nella serie di dati idrologici potrebbero essere determinati anche (e non raramente) da derive dello strumento di misura ovvero da modifiche locali e graduali delle condizioni di misura. Variazioni repentine potrebbero invece essere associate ad interventi antropici di grande impatto come, nel caso della serie di portate, la costruzione di opere idrauliche, invasi, argini, scolmatori, ecc.. Ma, come nel caso di variazioni graduali, un cambiamento repentino potrebbe essere associato alla modifica della tipologia di strumento di misura, ovvero allo spostamento, mantenendo la stessa denominazione, del punto di rilevamento. Questi ultimi casi, non infrequenti, si riscontrano soprattutto nelle serie storiche molto lunghe che abbracciano molti decenni. Risulta, quindi, particolarmente importante poter disporre di strumenti statistici che possano verificare la possibilità che una lunga serie di dati abbia le caratteristiche di stazionarietà per poter essere utilizzata nella sua interezza in analisi statistiche di base.

Qualora non sia verificata tale ipotesi sarebbe necessario operare correzioni, talvolta anche molto complesse. Nelle presenti LG si sono trattate le sole procedure per rimuovere la non-stazionarietà legata alla stagionalità che è quella più evidente e facilmente individuabile e interpretabile. Per quanto riguarda le altre tipologie di non stazionarietà ci si limiterà a metterle in evidenza e a verificare il soddisfacimento delle ipotesi di stazionarietà alla base di molte procedure statistiche.

Le analisi statistiche che tengono conto della non-stazionarietà, come ad esempio quelle nell'analisi degli estremi, costituiscono tuttora argomento di ricerca.

Una linea guida per un approccio sistematico a questo tipo di analisi è presente nel report n° 45 redatto nell'ambito del "World Climate Data and Monitoring Program (WCDMP)" (WMO, 2000)². ed è sintetizzato nel lavoro di Kundzewicz e Robson (2004).

² Il report costituisce l'esito del workshop WMO/UNESCO/CEH tenutosi nel Centro per l'Ecologia e Idrologia di Wallingford (UK) nel dicembre del 1998, con lo scopo di operare una revisione delle tecniche di valutazione dei trend nelle serie idrologiche.

7.4.1 Componenti di una serie storica

Le serie storiche di dati, e in particolare di quelli idrologici, possono ritenersi costituite da diverse componenti, ciascuna delle quali esprime un particolare comportamento sia di natura deterministica (o sistematica) sia di natura aleatoria. Tali componenti possono essere additive o moltiplicative.

Nel caso additivo, ad esempio, una serie può essere considerata come la somma di:

- 1) un *trend* o una tendenza deterministica (o sistematica) che esprime l'andamento nel lungo periodo (di fondo) del fenomeno di cui la serie è la rappresentazione, che può essere crescente, decrescente o costante, ed è in generale rappresentabile mediante una funzione analitica del tempo (e.g., semplicemente lineare);
- 2) una componente ciclica deterministica con periodicità diversa da quella annuale;
- 3) una componente stagionale deterministica che, per i dati ambientali e idrologici, esprime generalmente la variabilità legata all'evoluzione delle stagioni e quindi una componente che presenta una frequenza annuale ma che in generale manifesta valori diversi da un anno all'altro (come vedremo in seguito, si può anche individuare una componente stagionale uguale per tutti gli anni);
- 4) una componente irregolare, che può essere o non essere totalmente aleatoria e contenere ancora componenti deterministiche. In genere, può essere facilmente trattata come una serie di dati rappresentabile come un processo stocastico cosiddetto "white noise" (rumore bianco) gaussiano caratterizzato dal fatto di avere distribuzione normale standard con media 0 e varianza unitaria.

7.4.2 Serie destagionalizzata

Nelle serie storiche di dati idrologici con campionamento sub-annuale (giornalieri, settimanali e mensili), l'effetto legato alla stagionalità può rendere difficile l'analisi delle altre componenti della serie, che in genere risultano più interessanti da analizzare. Quella stagionale costituisce, infatti, una componente nota e quindi relativamente meno interessante, ma al contempo tale da schermare altri andamenti legati a fenomeni ben più interessanti (e.g., cambiamenti climatici).

È necessario quindi individuare metodi che permettano di isolare la componente stagionale. Il procedimento di eliminazione dell'effetto stagionale è detto destagionalizzazione (*deseasonalizing*) e la serie che si ottiene si denomina serie destagionalizzata (*deseasonalized*).

Nell'ambito idrologico, in particolare, si è interessati alla stima di una componente stagionale periodica che sia rappresentativa della variabilità stagionale della serie completa, e dunque uguale per tutti gli anni di osservazione.

Inoltre, assume importanza la definizione della variabilità stagionale della varianza della serie, che misura l'entità delle oscillazioni intorno all'andamento medio della serie (definito dalla componente stagionale della media). Come mostrato da Grimaldi (2004) il calcolo delle componenti stagionali può essere eseguito indipendentemente dalla definizione delle componenti di tendenza, la cui stima in molti casi può non essere necessaria.

Molteplici sono i metodi che consentono di decomporre e isolare le componenti di una serie storica. Nelle presenti LG si propone di utilizzare il metodo elaborato da Grimaldi e derivato dal metodo *Seasonal Trend decomposition based on LOESS (STL)* sviluppato da Cleveland. La procedura di destagionalizzazione elaborata da Grimaldi (detta anche STL Modificato Robusto) è descritta in dettaglio in Appendice B paragrafo 11.4.2.

Con il metodo STL Modificato Robusto si individua la componente stagionale della media e della varianza, uguale per tutti gli anni.

Nel caso di una serie con risoluzione giornaliera (Figura 7.4), essa viene costruita effettuando per ciascun giorno dell'anno la media degli N (numeri di anni) valori della grandezza idrologica corrispondenti al giorno dell'anno e successivamente applicando ai 365 valori di ciascuna media, corretta con particolari pesi dipendenti dalle stesse fluttuazioni della media, l'algoritmo di regressione denominato LOESS o LOWESS (*LOcally WEighted Scatterplot Smoothing*) ottenendo una curva, molto regolare (ma non parametrica) che rappresenta la componente deterministica della variabilità della grandezza idrologica dovuta alla stagionalità ((Figura 7.5a). Analogamente calcolando la varianza o lo scarto quadratico medio per ciascun giorno dell'anno (ciascuno con gli N dati) e quindi applicando l'algoritmo LOESS a valori dello scarto corretto con opportuni pesi si ottiene la componente stagionale della varianza (Figura 7.5b). L'adozione dei particolari pesi per la stima della

media e della varianza costituisce la differenza del metodo Grimaldi rispetto al cosiddetto metodo Classico

La grandezza destagionalizzata (e standardizzata) è fornita quindi da:

$$Z_{t+365(i-1)} = \frac{X_{t+365(i-1)} - \hat{m}_t^*}{\hat{\sigma}_t^*} \quad \text{eq. 21}$$

La serie così ottenuta (Figura 7.6) è priva di stagionalità nella media e nella varianza.

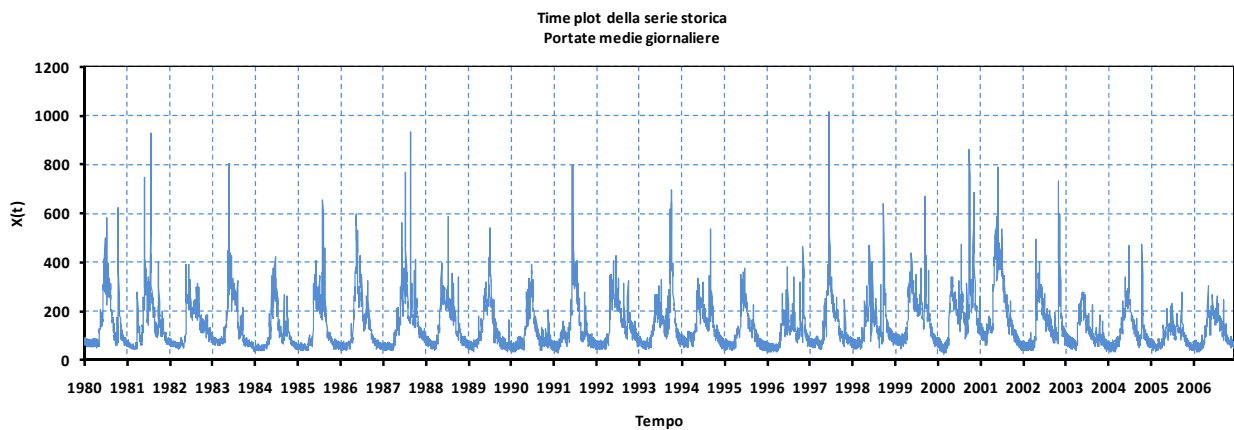


Figura 7.4 - Serie storica delle portate giornaliere

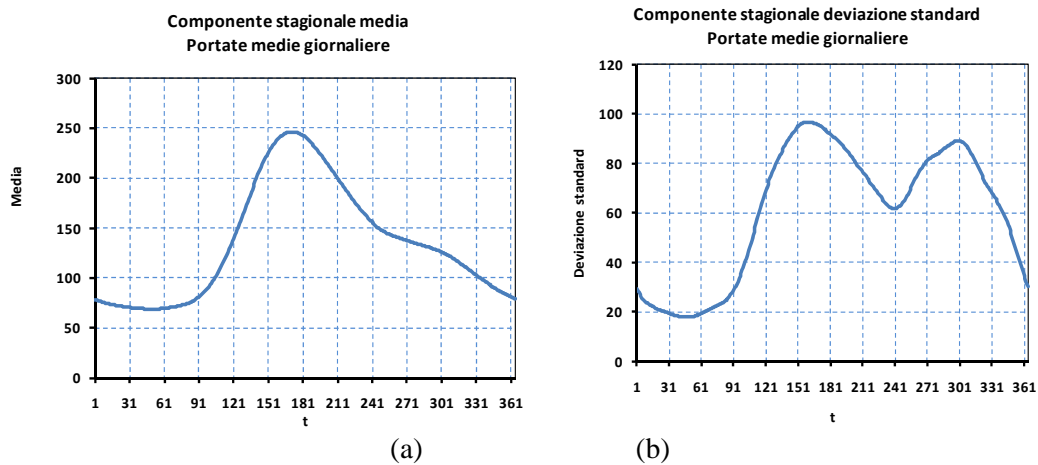


Figura 7.5 - Componente stagionale della media (a) e della deviazione standard (b)

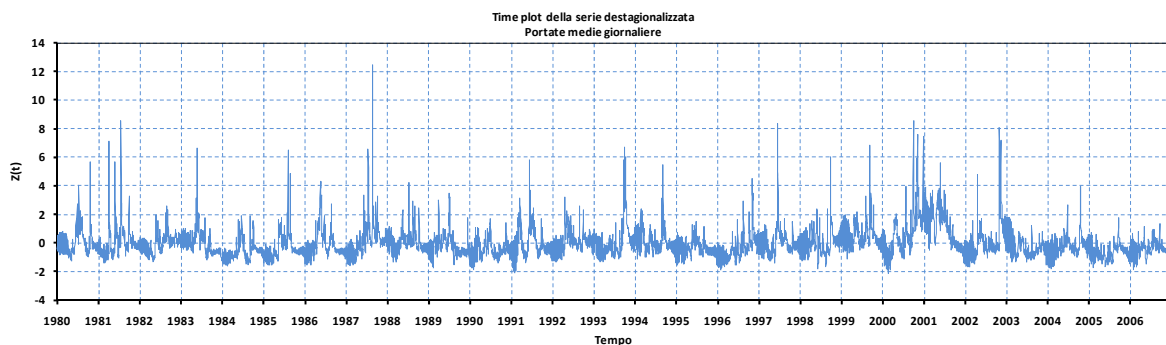


Figura 7.6 - Serie storica destagionalizzata delle portate medie giornaliere

A questo proposito si osserva l'importanza della stagionalità della varianza poiché anche la variabilità dei dati intorno alla media ha un'evidente stagionalità.

Analogamente si può procedere per serie mensili o settimanali in cui la componente stagionale è costituita rispettivamente da 12 o 52 valori.

Lo studio delle periodicità stagionali che caratterizzano serie storiche di grandezze ambientali è di importanza rilevante per la definizione dei regimi climatici sottesi al fenomeno osservato. La rimozione delle componenti stagionali in media e varianza consente, inoltre, di evidenziare proprietà della serie (in genere altre forme di periodicità, o lunga memoria) che altrimenti non sarebbero rilevabili, perché adombrate dalla periodicità dominante.

Per la descrizione approfondita del metodo si veda Appendice B paragrafo 11.2.5, 11.4.2 e 7.4.2.1

Nell'applicazione degli algoritmi di destagionalizzazione a serie giornaliere, è opportuno rimuovere i 29 febbraio presenti negli anni bisestili per avere serie annuali di uguale numerosità (365 giorni).



La scelta delle finestre d'interpolazione da utilizzare nell'applicazione del LOESS deve essere eseguita con un approccio prova-errore e dipende dalla risoluzione temporale della serie. Per serie giornaliere si suggerisce di applicare valori tra 30 e 90 giorni associati alla durata del mese e di un stagione.



7.4.2.1 Considerazioni sui dati mancanti

Il calcolo delle componenti stagionali è effettuato in genere sui dati a scala temporale giornaliera o mensile. La procedura proposta si basa sul calcolo di una media ponderata locale dei valori relativi a ogni giorno o mese dell'anno e sul successivo *smoothing* di ognuna di queste medie basate sui valori delle medie dei giorni o mesi contigui. La struttura del calcolo non richiede la continuità temporale delle registrazioni e consente dunque, implicitamente, la presenza di valori mancanti, che influiscono quasi esclusivamente sulla robustezza della media locale (media ponderata dei valori di un fissato giorno o mese) in ragione del numero di giorni o mesi disponibili. Ad esempio, se una serie di 15 anni di dati giornalieri è priva di quattro registrazioni del primo gennaio, il solo effetto sul calcolo della componente stagionale è che la media dei "primo gennaio" è eseguita su 11 valori piuttosto che su 15. La successiva procedura di *smoothing* riduce ulteriormente l'effetto del numero ridotto di osservazioni utilizzato per il calcolo della media dei "primo gennaio". Ne segue che la presenza di valori mancanti, anche in numero consistente, non pregiudica una definizione sufficientemente affidabile delle componenti stagionali.

7.4.3 *Change point*

L'analisi statistica per l'individuazione dei cambiamenti repentini (*change point*) cerca di rispondere alle seguenti domande:

- 1) se sono presenti uno o più cambiamenti repentini;
- 2) quando tali cambiamenti sono accaduti;
- 3) quanto essi siano statisticamente significativi.

L'analisi è effettuata sui dati ordinati secondo la struttura temporale ed è effettuata mediante due test non parametrici, quello di Pettitt e quello del CUSUM (*CUMulative SUM*) che hanno la particolarità di non richiedere ipotesi particolari sulla distribuzione dei dati.

Per entrambi i test l'ipotesi nulla H_0 è che non ci siano *change point*:

- 1) quando l'ipotesi nulla H_0 non è rigettabile i dati non presentano significativi *change point*;
- 2) quando l'ipotesi nulla H_0 è rigettabile i dati presentano significativi *change point*.

I test per i *change point* vengono effettuati solo su serie mensili e annuali.

7.4.3.1 Test di Pettitt

Il test è basato sui valori assunti dalla seguente statistica:

$$U_{t,T} = \sum_{i=1}^t \sum_{j=t+1}^T \text{sgn}(X_i - X_j) \quad t = 1, \dots, T, \quad \text{eq. 7.4.1}$$

in cui T è il numero dei dati della serie e $\text{sgn}(\cdot)$ indica la funzione segno. Per ogni valore di t (istante della serie) si ottiene quindi un valore di $U(t,T)$ che viene diagrammato. L'analisi del grafico della funzione $U(t,T)$ consente l'individuazione dell'istante in cui può essere avvenuto un *change point*.

Infatti, il punto di massimo o di minimo della funzione $U(t,T)$ rappresenta l'istante temporale in cui si collocherebbe il *change point* se l'esito del test fosse tale che l'ipotesi nulla H_0 fosse rigettabile al livello di significatività assegnato. Per la descrizione dettagliata del metodo si veda l'Appendice B paragrafo 11.4.3.2.1.

7.4.3.2 Test CUSUM

Per la descrizione dettagliata del metodo si veda l'Appendice B paragrafo 11.4.3.2.2

Quest'approccio è basato sull'analisi del grafico della serie cronologica delle somme cumulate degli scarti tra le osservazioni e la media del campione $(X_t - \hat{\mu}_X)$.

L'interpretazione del grafico della curva CUSUM è relativamente semplice. Poiché tale metodo è basato sulla somma cumulata delle differenze rispetto alla media $(X_t - \hat{\mu}_X)$, quando in un periodo i valori della serie sono, per la maggior parte, maggiori della media allora il grafico CUSUM è evidentemente crescente poiché si aggiungono valori per la maggior parte positivi.

Quindi, in un periodo in cui la curva CUSUM è crescente significa che i valori della serie sono in massima parte maggiori della media. Analogamente in un periodo in cui la curva CUSUM è decrescente i valori della serie sono in massima parte minori della media. Quando si verifica un netto cambiamento dell'andamento della curva significa, quindi, che si è verificato un netto cambiamento nei valori dei dati. Il punto in cui avviene l'inversione è il punto in cui è avvenuto il cambiamento repentino. Quando invece l'andamento della curva CUSUM è relativamente costante con andamento irregolare allora la media è costante.

È tuttavia necessario stabilire se tali variazioni possano considerarsi statisticamente significative. A tal fine si effettua il test i cui dettagli sono riportati in Appendice B paragrafo 11.4.3.2.2

7.4.3.3 Esempio di applicazione dei test per il *change point*

Per verificare la capacità dei test di evidenziare un cambiamento repentino si riporta l'applicazione dei test di Pettitt e del CUSUM su una serie temporale simulata della grandezza X in cui è stato inserito a metà dell'anno 1992 un *change point* (Figura 7.7).

Com'è evidente, entrambi i test (che sono sostanzialmente equivalenti) forniscono l'esito (Tabella 7.1) che l'ipotesi nulla è rigettabile al livello di significatività del 5%. I diagrammi nella Figura 7.9 e nella Figura 7.8 mettono in evidenza la presenza di un *change point* esattamente nel periodo in cui è stato artificialmente inserito.

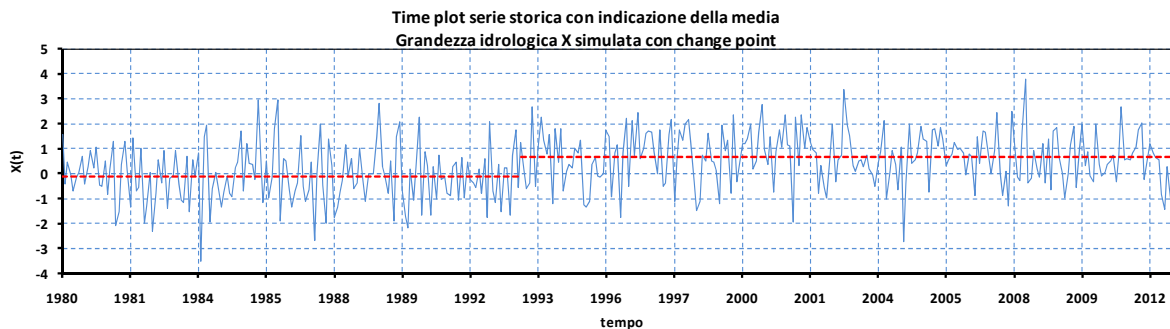


Figura 7.7 - Serie storica della grandezza idrologica X simulata con *change point*.

Tabella 7.1 - Esito dei test per il "change point detection". Grandezza idrologica X simulata con *change point*.

Test	Statistica	P-Value	Livello di significatività	Esito
CUSUM	79.5	0.001	5%	IPOTESI H_0 RIGETTABILE
Pettitt	15344.0	0.000	5%	IPOTESI H_0 RIGETTABILE

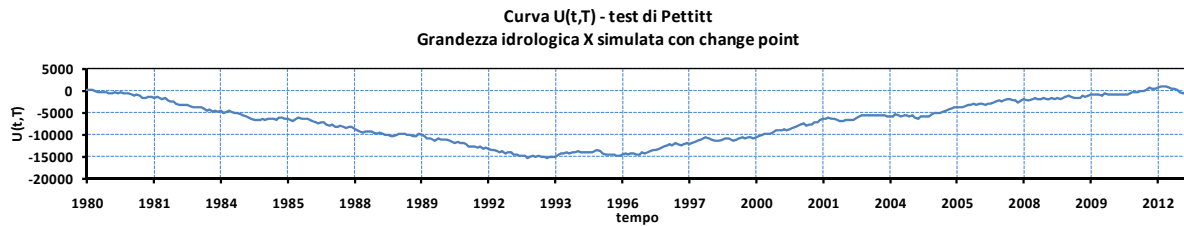


Figura 7.8 - Curva della statistica di Pettitt della grandezza idrologica X simulata con change point.

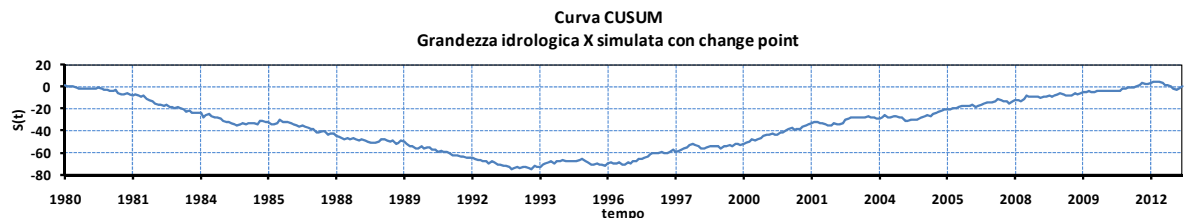


Figura 7.9 - Curva CUSUM della grandezza idrologica X simulata con change point.

Per controprova sono stati applicati i test di Pettitt e del CUSUM su una serie temporale simulata della grandezza Y senza *change point* (Figura 7.10).

Com'è evidente, anche in questo caso, entrambi i test forniscono il medesimo esito (Tabella 7.2) che l'ipotesi nulla non è rigettabile al livello di significatività del 5%, che significa che non esistono significativi cambi repentini. I diagrammi nella Figura 7.11 e nella Figura 7.12 non evidenziano, infatti, alcun punto di cambiamento nelle caratteristiche della serie.

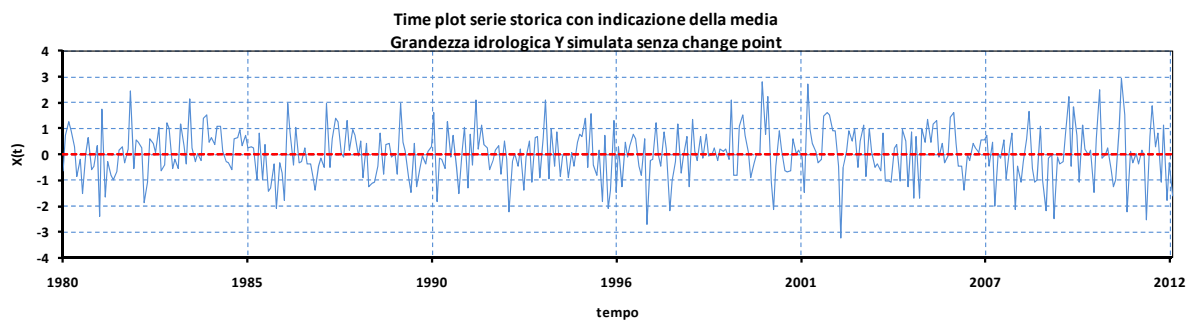


Figura 7.10 - Serie storica della grandezza idrologica Y simulata senza change point.

Tabella 7.2 - Esito dei test per il "change point detection". Grandezza idrologica Y simulata senza change point.

Test	Statistica	P-Value	Livello di significatività	Esito
CUSUM	17.5	0.891	5%	IPOTESI H ₀ NON RIGETTABILE
Pettitt	2129.0	0.646	5%	IPOTESI H ₀ NON RIGETTABILE

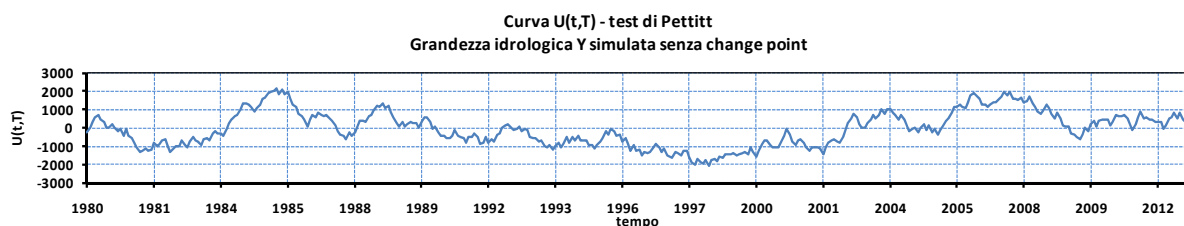


Figura 7.11 - Curva della statistica di Pettitt della grandezza idrologica Y simulata senza change point.

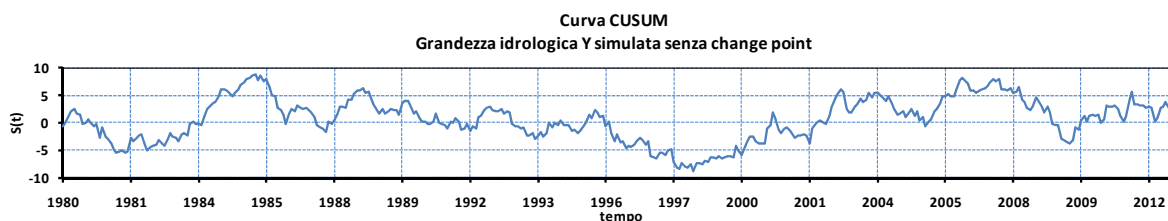


Figura 7.12 - Curva CUSUM della grandezza idrologica Y simulata senza change point.

7.4.3.4 Considerazioni sui dati mancanti

Nei paragrafi precedenti si suggerisce di non applicare i test per i *change point* su dati con aggregazione giornaliera o più fine, poiché questi dati possono presentare delle oscillazioni casuali di ampiezza tale da indurre la rilevazione di *change point* anche se non realmente presenti. In più, la sensibilità di questi test cresce al crescere del campione, perciò in presenza di una numerosità superiore ad alcune centinaia anche piccoli salti fisicamente non significativi sono rilevati come statisticamente rilevanti.

Focalizzando l'attenzione alla scala temporale mensile, una serie destagionalizzata (e con autocorrelazione trascurabile) di 20-30 anni presenta 240-360 elementi. In questi casi, il primo passo è un'analisi visiva della serie e la valutazione delle potenziali cause di eventuali *change point* presenti nella serie. Quest'operazione richiede inevitabilmente un intervento soggettivo da parte dell'analista, sia nelle rilevare visivamente potenziali *change point* sia nella ricerca di meta-informazioni che aiutino a caratterizzare la serie.

Se, ad esempio, si sospetta la presenza di un salto in una serie di portata dovuta al cambiamento della collocazione di uno strumento o alla costruzione di un invaso, l'assenza di 12 mesi di dati in coincidenza del cambio può non influire sul suo rilevamento, in quanto il periodo di transitorio legato al cambiamento ha una durata ragionevolmente inferiore all'anno. Il solo effetto è la mancanza di una collocazione esatta del cambio all'interno dell'anno, che peraltro è soggetta ad incertezza a causa del transitorio stesso. Se il cambiamento è di natura climatica e verosimilmente la fase di transitorio del fenomeno è superiore all'anno, sono ammissibili anche intervalli mancanti superiori. In altri termini, i test possono essere applicati alle serie senza considerare i dati mancanti, ma preservando l'informazione relativa alla posizione cronologica di ogni osservazione, in modo tale da poter discriminare tra *change point* repentini (collocati all'interno di un intervallo di osservazioni disponibili con continuità) e *change point* associati ad istanti in corrispondenza dei quali è presente un intervallo di dati mancanti. Chiaramente anche per questo tipo di test (come per il calcolo di H), è opportuno che il numero di dati mancanti non sia eccessivo (e.g., 5%) e non sia concentrato in un intervallo ristretto della serie. Se, ad esempio, i dati mancanti sono concentrati nella parte iniziale di una serie, è opportuno non considerare questa parte della serie. In altre parole, i dati disponibili non dovrebbero rappresentare delle informazioni isolate tra ampi intervalli di *missing value*.

Applicare i test per i *change point* a serie giornaliere non è consigliabile, poiché le oscillazioni casuali dovute alla variabilità intrinseca del fenomeno a questa scala di risoluzione temporale può condurre al rilevamento di *change point* statisticamente significativi, ma fisicamente inconsistenti. In genere è opportuno, se possibile, approfondire l'indagine per valutare la natura e la causa del *change point* (ad esempio, influenza antropica, variazioni climatiche, errori di misura, ecc.).



7.4.4 Trend

Come già detto, nelle presenti LG non si individuano metodi per determinare la forma del *trend* e quindi “*detrendizzare*” (cioè eliminare il *trend*) ma solo stabilire se i dati presentino o meno un *trend* monotono statisticamente significativo. A titolo di completezza, tuttavia, si citano alcuni metodi per la stima dei *trend*.

La stima del *trend* può essere effettuata sulla serie, eventualmente destagionalizzata, attraverso diverse procedure, che possono essere ricondotte a due filoni principali.

- a) il metodo analitico;
- b) il metodo delle medie mobili ponderate (WMA, *Weighted Moving Average*)

Nel primo caso s'individua una funzione analitica (e.g. lineare) e quindi si stimano, con i minimi quadrati, i parametri che caratterizzano tale funzione sulla base dei valori della serie destagionalizzata. I secondi metodi consistono in curve non parametriche derivanti da medie mobili o più in generale da regressioni locali tipo LOESS di cui le medie mobili ponderate costituiscono un caso particolare. Per la verifica delle variazioni graduali è possibile applicare il test della regressione lineare basato sul coefficiente di correlazione ρ di Pearson, il test di Mann-Kendall e il test per il coefficiente di correlazione ρ_s di Spearman. Il test di Pearson è un test parametrico per la verifica di trend lineari che richiede la normalità dei dati, mentre i test di Mann-Kendall e Spearman sono test non parametrici per la verifica di trend monotoni (lineari e non lineari). Per la descrizione completa dei test si veda Appendice B paragrafo 11.4.3.3.

7.4.4.1 Test di Mann-Kendall

Uno dei test non parametrici più usati per il rilevamento di trend monotoni (lineari e non) è il test di Mann-Kendall e si basa sul confronto delle coppie di osservazioni x_i, x_j ($i > j$) per accertare se $x_i > x_j$ ovvero $x_i < x_j$. L'ipotesi nulla H_0 è che la serie sia priva di trend significativo.

7.4.4.2 Test di Pearson

Il coefficiente di correlazione ρ di Pearson può essere usato per misurare l'associazione lineare tra valori adiacenti in una sequenza di valori ordinati (autocorrelazione), ovvero tra due vettori di osservazioni. L'esistenza di un trend significativo per i valori assunti da una grandezza fisica (Y) nel tempo (X) è spesso valutata definendo la significatività della pendenza b_1 della retta di regressione $Y = b_0 + b_1X$. Il test è di tipo parametrico e si basa sull'ipotesi che il campione abbia distribuzione normale. Qualora questa ipotesi non sia soddisfatta, la variabile è, nelle procedure implementate nel foglio ANABASI, preventivamente trasformata tramite la trasformata di Box e Cox. L'ipotesi nulla H_0 è che la serie non presenti correlazione significativa, ovvero trend lineare monotono.

7.4.4.3 Test di Spearman

Il coefficiente di correlazione di Spearman ρ_s è l'equivalente non parametrico del coefficiente di correlazione di Pearson. Analogamente al test di Mann-Kendall, il test di Spearman non richiede preventive trasformazioni dei dati, poiché è basato sui ranghi, ossia sugli indici che denotano le posizioni occupate dalle osservazioni nel campione ordinato (in modo crescente o decrescente). Anche il test di Spearman ha come ipotesi nulla H_0 l'assenza di trend.

7.4.4.4 Esempio di applicazione dei test per il trend

Come per i test per i change point, si riporta un esempio di una serie storica di una grandezza simulata X in cui è stato artificialmente inserito un trend lineare e di una serie storica di una grandezza simulata Y senza trend.

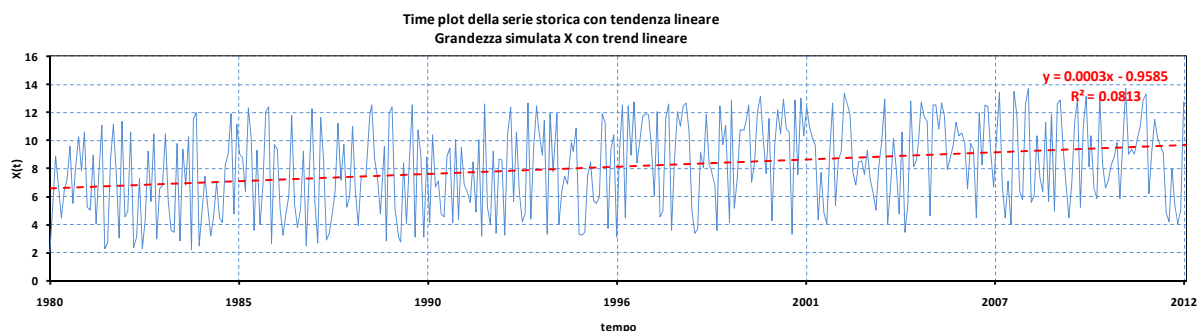


Figura 7.13 - Serie storica della grandezza idrologica X simulata con trend lineare.

Tabella 7.3 - Esito dei test per il "trend detection". Grandezza X simulata con trend lineare.

Test	Statistica	Tau	P-Value	Livello di significatività	Esito
Mann-Kendall	5.6545	0.190	0.000	5%	IPOTESI H_0 RIGETTABILE
Test	Statistica	Rho	P-Value	Livello di significatività	Esito
Pearson	6.6127	0.316	0.000	5%	IPOTESI H_0 RIGETTABILE
Spearman	5.6335	0.283	0.000	5%	IPOTESI H_0 RIGETTABILE

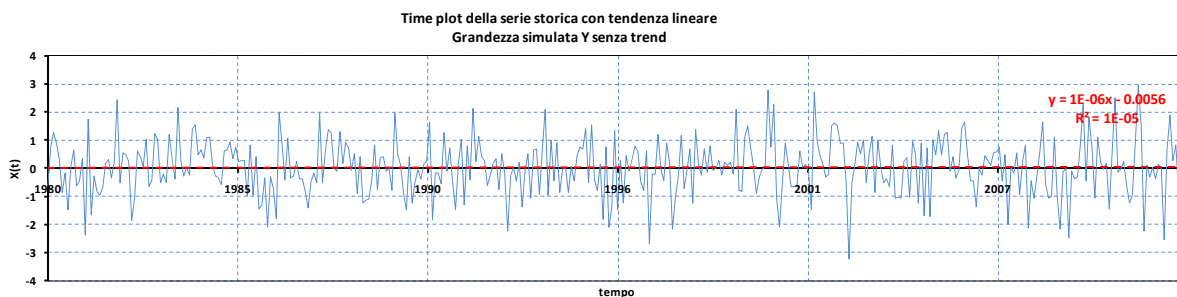


Figura 7.14 - Serie storica della grandezza idrologica Y simulata senza trend.

Tabella 7.4 - Esito dei test per il “trend detection”. Grandezza Y simulata senza trend.

Test	Statistica	Tau	P-Value	Livello di significatività	Esito
Mann-Kendall	0.2389	0.008	0.811	5%	IPOTESI H ₀ NON RIGETTABILE
Test	Statistica	Rho	P-Value	Livello di significatività	Esito
Pearson	-0.0737	-0.004	0.941	5%	IPOTESI H ₀ NON RIGETTABILE
Spearman	0.2301	0.012	0.818	5%	IPOTESI H ₀ NON RIGETTABILE

È evidente dalla Figura 7.13 e dalla Figura 7.14 e dalle corrispondenti Tabella 7.3 e Tabella 7.4, come nel primo caso i test forniscano una conferma della presenza del trend mentre nel secondo caso escludono un *trend* significativo.

7.4.4.5 Considerazioni sui dati mancanti

Analogamente ai test per i *change point*, i test per i *trend* non richiedono esplicitamente la continuità della serie. Evidentemente però i dati mancanti influenzano l’interpretazione dei risultati. Ad esempio, la presenza di un numero elevato n di *missing value*, isolati in una serie a scala di risoluzione annuale di lunghezza N , può dare luogo ad una serie più breve di quella osservata (qualora gli anni con dati mancanti siano trascurati), per cui un eventuale *trend* non significativo in N anni appare significativo sulla serie contratta di $N-n$ valori. La situazione è ancora più evidente in presenza di intervalli di valori mancanti. Un approccio generale per affrontare il problema dei dati mancanti è difficilmente proponibile. In genere occorre giudicare di volta in volta, tuttavia nell’applicazione dei test per i trend si suggerisce di utilizzare serie o parti di serie continue e di numerosità adeguata (possibilmente > 50). Ogni risultato su serie più brevi o con un numero rilevante di dati mancanti dovrebbe essere considerato con cautela.

Una *change point* può essere interpretato come trend dai test di Mann-Kendall, Pearson e Spearman. Al contrario, un trend non è rilevato come *change point* dai test di Pettitt e CUSUM. Ne segue che i test per trend monotoni dovrebbero essere eseguiti dopo l’applicazione dei test per i *change point* sulle parti di serie delimitate dai *change point*, qualora presenti



I test per i *change point* sono eseguiti in genere su sottoserie mensili o stagionali per eliminare l’effetto della stagionalità e l’eventuale autocorrelazione dei dati. Alternativamente il problema della stagionalità può essere aggirato usando le serie destagionalizzate. In questo caso però occorre verificare che l’autocorrelazione residua sia trascurabile.



Per una serie a scala giornaliera non si procede all’analisi dei *trend* e dei *change point* in quanto tale analisi non fornisce in genere buoni risultati a causa della estrema variabilità delle osservazioni (fluttuazioni casuali).



Utilizzare le serie destagionalizzate (nelle quali è assente la componente di autocorrelazione dovuta alla ciclicità annuale) non è corretto quando, a causa della presenza di persistenza, non è trascurabile l’autocorrelazione non stagionale



Il non tener conto della non-stazionarietà delle serie idrologiche potrebbe comportare pericolose sottostime o costose sovrastime dei parametri di progetto.



7.5 Analisi dei valori estremi

In ambito idrologico, una delle analisi statistiche di maggiore interesse da eseguire su una serie di dati e che costituisce uno dei principali obiettivi delle presenti LG, è quella di descrivere il comportamento statistico dei valori estremi, cioè di quei valori che rappresentano i fenomeni idrologici più estremi, come ad esempio fenomeni di inondazione (*floods*), di magra (*extreme low flows*), siccità estreme (*extreme droughts*), precipitazioni intense (*heavy rain*) (quando si analizzano dati di portata e/o di livello idrometrico e/o di precipitazione), ovvero ondate di calore (*heat waves*) (quando si analizzano dati di temperatura), ovvero maree estreme (*extreme sea levels*) (e.g. acqua alta a Venezia quando si analizzano dati mareografici), altezze d'onda estreme (*extreme sea waves*) (relativamente ai dati ondametrici), venti estremi (*extreme wind gust*), e così via.

L'interesse per l'analisi dei valori estremi è generalmente legato alla necessità di disporre della conoscenza del fenomeno estremo per valutarne il rischio e per dimensionare interventi per la mitigazione degli effetti e la predisposizione di azioni di natura diversa.

Dal punto di vista operativo, lo scopo dell'analisi dei valori estremi (*Extreme Value Theory, EVT o Extreme Value Analysis, EVA*) è essenzialmente quello di:

- stimare il valore della grandezza idrologica che viene superato mediamente una volta ogni T anni, che indichiamo con x_T , anche con T molto superiore all'intervallo di osservazione disponibile. L'intervallo di tempo T è il cosiddetto "periodo di ritorno" e x_T il quantile corrispondente a tale periodo di ritorno;
- valutare l'incertezza associata alla stima del valore x_T .

Ciò si persegue essenzialmente:

- individuando un'opportuna distribuzione di probabilità;
- calibrando i parametri sulla base delle osservazioni disponibili;
- estrapolando la distribuzione ai valori di interesse, in genere superiori a quelli osservati.

La teoria dei valori estremi fornisce, in definitiva, gli strumenti per quantificare la rarità e la severità degli estremi di un fenomeno idrologico osservato e per estrapolare valori corrispondenti a eventi più intensi di quelli rilevati.

L'interesse per i valori estremi in ambito idrologico non esclude, tuttavia, l'importanza dei valori medi che sono d'interesse per altre categorie di problemi, altrettanto importanti, come, ad esempio, quelli relativi alla gestione e l'utilizzo delle risorse idriche.

Nella presente versione delle LG l'attenzione viene focalizzata solo sui valori estremi massimi (coda superiore o destra della distribuzione, *upper tail*) (Figura 7.15).

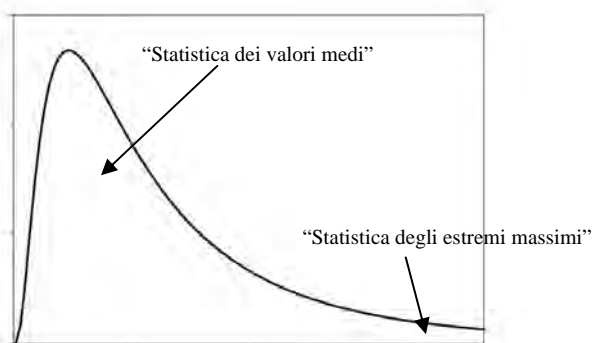


Figura 7.15 - Rappresentazione dei valori estremi.

Nelle successive versioni delle LG verrà affrontato il problema degli estremi minimi (coda inferiore o sinistra della distribuzione, *lower tail*).

L'analisi dei valori estremi prevede, come primo passo da compiere, quello di selezionare dalla serie completa dei dati, i valori estremi sui quali effettuare l'analisi. Le modalità di selezione dei valori estremi determinano anche il tipo di distribuzione di probabilità utilizzabile.

Riguardo a tali modalità, gli approcci classici nella EVA sono principalmente due (Figura 7.16):

- 1) serie dei massimi estratti da intervalli temporali regolari o blocchi (*Block Maximum, BM*) che, nel caso in cui, come generalmente accade, sono costituiti dall'anno solare, costituisce la serie dei massimi annuali (*Annual Maximum, AM*);
- 2) serie dei picchi sopra una fissata soglia o serie dei massimi di durata parziale (*Peak Over Threshold, POT*, ovvero *Partial Duration Series, PDS*)

Nel primo caso, fissato un intervallo temporale (come detto, in genere l'anno), si considera l'insieme di dati costituito dal valor massimo di ciascun intervallo. Nel secondo caso, invece, l'approccio consiste nell'individuare un'opportuna soglia, sufficientemente elevata, e considerare l'insieme dei valori al di sopra di tale soglia. Meno intuitivo è il fatto che tale approccio sia identificato anche come serie dei massimi di durata parziale, poiché ciascun massimo si riferisce ad un evento che ha una durata diversa dagli altri.

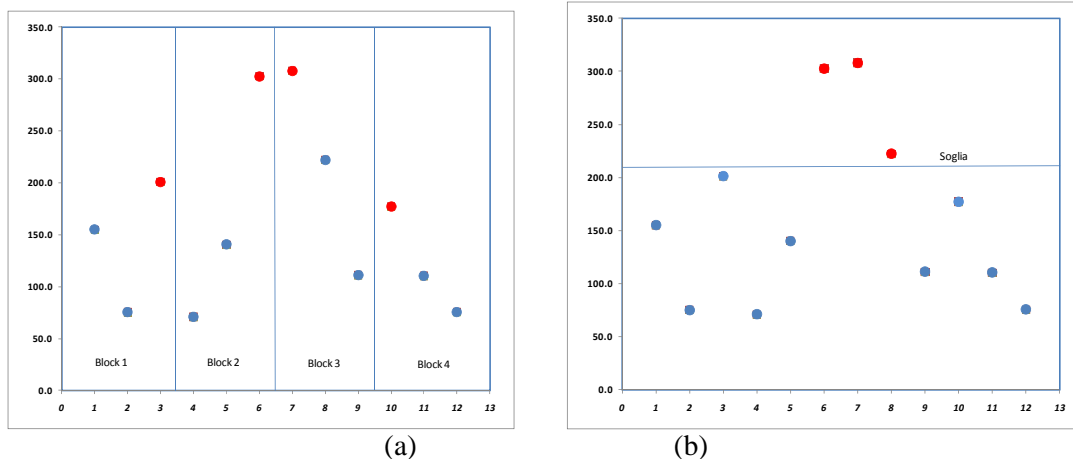


Figura 7.16 - Rappresentazione degli approcci utilizzati nella EVA applicata alla stessa serie di dati. (a) Approccio *Block Maximum* in cui si considerano i valori massimi (in rosso) per ciascun blocco. (b) Approccio *POT* in cui si considerano i valori (in rosso) sopra una soglia prefissata.

Gli approcci sopraelencati, ai fini dell'applicazione dei teoremi della EVT, dovrebbero garantire essenzialmente le seguenti due ipotesi:

- 1) indipendenza statistica del campione di variabili estratte: dei massimi di ciascun blocco o dei picchi soprastoglia;
- 2) medesima distribuzione di probabilità per le variabili estratte

Le due ipotesi vengono indicate come “variabili indipendenti e identicamente distribuite” e abbreviate con *i.i.d.* anche nella letteratura di origine anglosassone (*independent and identically distributed variables*).

Ricordiamo che due eventi sono statisticamente indipendenti tra loro se la valutazione della probabilità di un evento non dipende dal verificarsi dell'altro. In questo caso la probabilità del verificarsi dell'evento congiunto è pari al prodotto delle probabilità del verificarsi di ciascun evento separatamente.

Come sarà chiarito in seguito, nessuna delle due ipotesi è, in generale, rigorosamente rispettata per le variabili idrologiche. Queste infatti sono, più o meno, caratterizzate da legami autocorrelativi sul breve periodo e/o da lunga memoria; inoltre, anche solo tenendo conto della stagionalità, e non anche di un eventuale non stazionarietà o *trend* climatico, verrebbe meno l'ipotesi che le variabili aleatorie selezionate abbiano la medesima distribuzione di probabilità.

Le analisi descritte nei paragrafi precedenti sono effettuate in particolar modo per verificare la possibilità, anche se non proprio rigorosa, di sussistenza delle due ipotesi che consentono di applicare i risultati della *EVT*.

Mentre nell'approccio *Block Maximum*, e in particolare l'approccio *AM*, l'indipendenza statistica è praticamente assicurata dalle modalità di estrazione del dato e dal fatto che l'intervallo è ampio (se l'intervallo del *Block Maximum* è piccolo, ad esempio mensile, l'indipendenza non è così scontata), non lo è invece nel secondo caso in cui i valori sopra soglia potrebbero essere vicini ed essere statisticamente dipendenti (appartenere, per esempio, allo stesso evento idrologico).

Il principale problema nell'analisi dei valori estremi risiede nell'individuare la stima più corretta e robusta dei valori della grandezza idrologica per periodi di ritorno elevati, anche molto maggiori del

periodo di osservazione. Tale problema dipende fortemente dalla distribuzione adottata, dalla forma della sua coda nonché dal metodo di stima dei parametri.

In realtà non è possibile stabilire quale degli approcci sopraelencati e quale metodo di stima dei parametri fornisca le stime migliori per i periodi di ritorno. Com'è facilmente intuibile, ciascun approccio e ciascun metodo di stima presenta dei vantaggi e degli svantaggi e l'analisi consiste proprio nel verificare, caso per caso, quale sia il metodo più affidabile.

7.5.1 Serie AM

Le serie AM sono analizzate mediante la distribuzione di probabilità generalizzata del valore estremo (*Generalized Extreme Value distribution, GEV*) in virtù del teorema di Fisher–Tippett–Gnedenko o detto anche “primo teorema del valore estremo” per cui la variabile aleatoria definita dal massimo di n variabili *i.i.d* (con legge di distribuzione anche non nota):

$$M_n = \max(X_1, \dots, X_n) \quad \text{eq. 22}$$

sotto opportune condizioni, si distribuisce asintoticamente (per $n \rightarrow \infty$) secondo una GEV.

Ad esempio, il vettore (X_1, \dots, X_n) potrebbe essere la serie dei valori massimi annuali di portata media giornaliera ovvero i valori di precipitazione giornaliera massimi annuali, cosicché la variabile M_n rappresenta il valore massimo della variabile X su n anni di osservazione.

Diverse sono le formulazioni che si trovano nella letteratura tecnico-scientifica dell'espressione della GEV. Nelle LG e nel foglio elettronico ANÁBASI è utilizzata la seguente espressione della funzione di ripartizione (*Cumulative Distribution Function, CDF*), che segue i principali riferimenti internazionali sull'argomento (e.g. Coles, 2001):

$$\Pr(X \leq x) = F(x) = \begin{cases} \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\} & \xi \neq 0 \\ \exp\left(-\left[e^{-\left(\frac{x-\mu}{\sigma}\right)}\right]\right) & \xi = 0 \end{cases} \quad \text{eq. 23}$$

dove σ , μ e ξ denotano rispettivamente i cosiddetti parametri di **scala**, di **posizione** e di **forma**.

L'espressione utilizzata è anche quella riportata su *Wikipedia* in lingua inglese (http://en.wikipedia.org/wiki/Generalized_extreme_value_distribution).

La distribuzione per $\xi \neq 0$ è definita, evidentemente, se:

$$\left[1 - \xi\left(\frac{x-\mu}{\sigma}\right)\right] > 0 \quad \text{eq. 24}$$

Se il parametro di forma è $\xi = 0$ la distribuzione ricade, come caso particolare, nella distribuzione del valore estremo di tipo I (*Extreme Value type 1, EV1*), più comunemente nota come distribuzione di Gumbel.

Una nota particolare merita il parametro di forma ξ , il cui valore determina il tipo di coda destra della distribuzione, cioè le modalità con cui essa decade (tende a zero). Infatti, per i problemi di estremi che si analizzano è di fondamentale importanza definire il corretto comportamento della coda della distribuzione che governa la frequenza e l'intensità (la magnitudine) degli estremi.

In sintesi, si ha (Figura 7.17) che per:

- $\xi=0$: la distribuzione (di Gumbel o EV1) e i dati di cui la distribuzione ne è la modellazione statistica presentano una coda che decresce come un'esponenziale ed è detta *normal tailed o light tailed* (dalla parola inglese *tail=coda*);
- $\xi>0$: la distribuzione (di Fréchet o EV2) e i dati di cui la distribuzione ne è la modellazione statistica presentano una coda (destra) che decresce meno rapidamente di un legge

esponenziale per cui è detta *heavy tailed* o *fat tailed* (coda pesante o spessa). In questo caso la distribuzione è limitata inferiormente ed esiste solo se $x > (\mu - \sigma / \xi)$;

- $\xi < 0$: la distribuzione (di Weibull o EV3) in questo caso la distribuzione è limitata superiormente e decade molto più rapidamente dell'esponenziale (*thin tailed* o *light tailed*). Esiste solo se $x < (\mu - \sigma / \xi)$ e generalmente non presenta grande utilità in quei fenomeni idrologici che non sono limitati superiormente. E' invece utilizzata per i valori estremi minimi. È anche detta *short tailed*.

L'analisi sulla forma della coda della distribuzione fornisce importanti informazioni sulla tipologia dei valori estremi della serie e di come essi si manifestano.

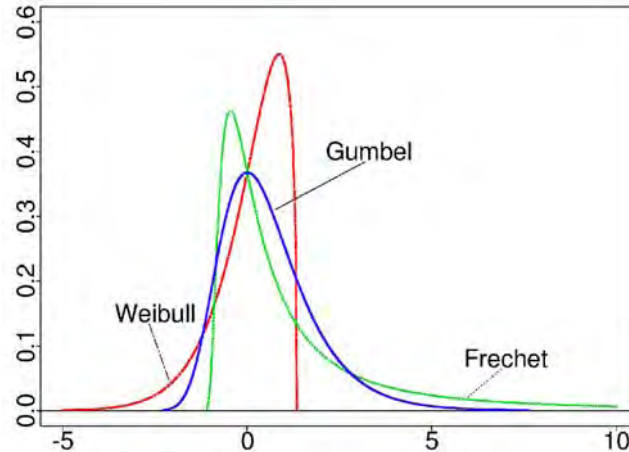


Figura 7.17 - Rappresentazione dei tipi di distribuzioni GEV

Dati che presentano una distribuzione *heavy tailed* sono caratterizzati dal fatto che eventi eccezionali notevolmente superiori a quelli osservati hanno una probabilità di superamento relativamente più alta (ovvero un periodo di ritorno più basso) rispetto ad eventi che presentano una distribuzione *normal tailed*. Ciò significa, d'altro canto che, a parità di periodo di ritorno T (e quindi di frequenza) la distribuzione *heavy tailed* fornisce valori di ritorno più elevati della *normal tailed* (Gumbel).

Bisogna prestare particolare attenzione alla formulazione della CDF della GEV e in particolare al parametro di forma ξ che viene spesso assunto di segno opposto, come ad esempio nel lavoro di Hosking et al., (1985), nel *Handbook of Hydrology* di Maidment (1993) e nel testo di Kottegodda e Rosso (1997 pag. 450), e quindi le considerazioni sul segno di ξ si ribaltano.

Un metodo grafico per visualizzare il decadimento della coda dei dati è quello di rappresentarli sul *QQ Plot* Esponenziale (relativo cioè alla distribuzione di Gumbel, con il dato osservato sull'asse delle ascisse). Se i dati presentano un comportamento *normal tailed* i valori elevati si allineano lungo la linea retta; se il comportamento è *heavy tailed* i punti presentano una concavità verso il basso. Se invece presentano un comportamento *light tailed* la concavità è rivolta verso l'alto.

Detto $T > 1$ il periodo di ritorno espresso in anni, il valore della grandezza idrologica associata al periodo di ritorno T , detto quantile x_T , con probabilità di superamento pari a $1/T$ (o di non superamento pari a $[1-1/T]$) (Figura 7.18) è fornito dalla formulazione inversa della eq. 23:

$$x_T = \begin{cases} \mu - \frac{\sigma}{\xi} \left\{ 1 - \left[-\ln \left(1 - \frac{1}{T} \right) \right]^{-\xi} \right\} & \xi \neq 0 \\ \mu - \sigma \ln \left[-\ln \left(1 - \frac{1}{T} \right) \right] & \xi = 0 \end{cases} \quad \text{eq. 25}$$

Nella terminologia della *EVT/EVA* il quantile x_T è detto livello di ritorno (*return level*) associato al periodo di ritorno T (*return period*).

Più in generale, se l'intervallo del *Block Maximum* non è l'anno, si considera il quantile che viene superato mediamente una volta ogni m osservazioni con probabilità, quindi, pari a $1/m$. Se λ è il numero di valori estratti in un anno (12 se il *Block Maximum* è il mese) il livello di ritorno corrispondente a T anni è:

$$x_T = \begin{cases} \mu - \frac{\sigma}{\xi} \left\{ 1 - \left[-\ln \left(1 - \frac{1}{\lambda T} \right) \right]^{-\xi} \right\} & \xi \neq 0 \\ \mu - \sigma \ln \left[-\ln \left(1 - \frac{1}{\lambda T} \right) \right] & \xi = 0 \end{cases} \quad \text{eq. 26}$$

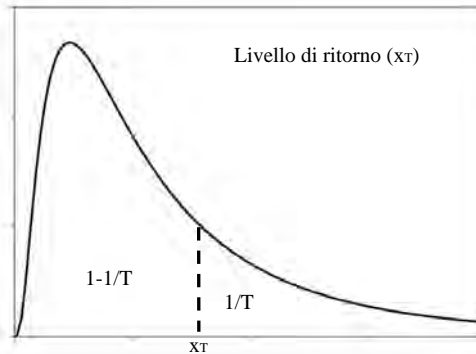


Figura 7.18 - Schema del livello di ritorno

La stima dei parametri viene effettuata nel foglio ANÁBASI con i tre metodi classici:

- metodo dei momenti (*Method of Moments, MoM*)
- metodo dei momenti pesati in probabilità o L-momenti (*Probability Weighted Moment, PWM* o *L-Moments, LM*)
- metodo della massima verosimiglianza (*Maximum Likelihood, ML*)

e vengono utilizzati per il calcolo dei livelli di ritorno i parametri forniti dal metodo di stima *PWM*.

Non è prevista in questa fase delle LG, anche per la notevole complessità, esporre la teoria e le procedure della stima parametrica che si rimanda, per il momento, ai testi e agli articoli specialistici.

In generale, come anche riportato nel “*Guidelines on Analysis of extremes in a changing climate in support of informed decisions for adaptation*” (WMO, 2009), il metodo della massima verosimiglianza è preferibile quando il campione dei valori estremi è sufficientemente numeroso (> 500) e quando la serie può presentare una non-stazionarietà. In questo caso, il metodo *ML*, infatti, può tenere conto della non-stazionarietà. Il metodo degli *PWM* o *L-Moment* è invece preferibile quando il campione è limitato (< 500 come accade nella gran parte dei casi). Il classico metodo dei momenti invece non è raccomandabile poiché tende sottostimare i valori per elevati periodi di ritorno.

7.5.2 Serie POT/PDS

Le serie POT/PDS sono invece analizzate mediante la distribuzione di probabilità generalizzata di Pareto (*Generalized Pareto distribution, GP*) introdotta da Pickands nel 1975 per l’analisi degli eventi estremi e basata sul teorema di Pickands-Balkema-De Haan (noto anche come il “secondo teorema del valore estremo”). Il teorema dimostra che, considerata una serie (X_1, \dots, X_n) di variabili i.i.d, la funzione di distribuzione delle eccedenze $Y_i = (X_i - \mu)$ rispetto ad una soglia μ , sufficientemente alta, condizionata al valore della soglia stessa, è ben approssimata dalla GP (per una ampia categoria di funzioni di distribuzione di X). In altri termini la variabile:

$$\{X_i - \mu | X_i > \mu\} \text{ per } \mu \text{ molto alta} \quad \text{eq. 27}$$

segue approssimativamente una distribuzione GP qualunque sia la distribuzione comune delle X_i .

In questo caso il vettore (X_1, \dots, X_n) potrebbe essere, ad esempio, la sequenza dei valori massimi della portata giornaliera registrati per ciascun evento di piena indipendente.

Anche per l’espressione della GP, diverse sono le formulazioni che si trovano nella letteratura tecnico-scientifica. Nelle LG e nel foglio elettronico ANÁBASI è utilizzata la seguente espressione della funzione di ripartizione (*Cumulative Distribution Function, CDF*) che, mantenendo la medesima simbologia utilizzata per la GEV (poiché, come mostrato nei paragrafi successivi, è in stretta relazione

con la GPD), segue anch'essa i principali riferimenti internazionali (e.g. Coles, 2001). Anche per la GPD, l'espressione è quella riportata sul sito WEB di *Wikipedia* in lingua inglese (http://en.wikipedia.org/wiki/Generalized_Pareto_distribution).

Per il teorema della probabilità condizionata la distribuzione cumulata (CDF) delle eccedenze può essere espressa in funzione della CDF della variabile X:

$$F_{\mu}(y) = \Pr\{X - \mu \leq y | X > \mu\} = \frac{F(y + \mu) - F(\mu)}{1 - F(\mu)} \quad \text{eq. 28}$$

che, per il secondo teorema dei valori estremi, è approssimativamente pari a:

$$F_{\mu}(y) = \frac{F(x) - F(\mu)}{1 - F(\mu)} \approx G(x - \mu) = \begin{cases} 1 - \left[1 + \xi \left(\frac{x - \mu}{\sigma} \right) \right]^{-\frac{1}{\xi}} & \xi \neq 0 \\ 1 - \left[e^{-\left(\frac{x - \mu}{\sigma} \right)} \right] & \xi = 0 \end{cases} \quad \text{eq. 29}$$

Quindi, la distribuzione GPD è la distribuzione delle eccedenze condizionata al valore della soglia. Per il calcolo del quantile corrispondente al periodo di ritorno T dobbiamo fare riferimento alla distribuzione incondizionata F , stimando empiricamente la probabilità di superamento della soglia:

$$\Pr(X > \mu) = 1 - F(\mu) = \frac{N_{\mu}}{n} \quad \text{eq. 30}$$

Ottenendo, per la distribuzione incondizionata, l'espressione:

$$F(x) = \begin{cases} 1 - \frac{N_{\mu}}{n} \left[1 + \xi \frac{x - \mu}{\sigma} \right]^{-\frac{1}{\xi}} & \xi \neq 0 \\ 1 - \frac{N_{\mu}}{n} \left[e^{-\frac{x - \mu}{\sigma}} \right] & \xi = 0 \end{cases} \quad \text{eq. 31}$$

Con $x > \mu$ e N_{μ} il numero di eccedenze e n il numero del campione da cui sono estratte le eccedenze. Anche in questo caso, σ , μ , e ξ denotano rispettivamente i cosiddetti parametri di scala, di posizione e di forma.

Il parametro di scala σ deve essere sempre positivo, mentre il parametro di posizione μ coincide con la soglia assegnata per cui, di fatto, la GPD è una distribuzione a 2 parametri.

Particolare attenzione va posta all'inferenza sul parametro di forma ξ poiché è quello che determina sia le principali caratteristiche della distribuzione e in particolare la coda, sia l'efficienza della stima dei parametri.

La funzione CDF e la densità di probabilità (*Probability Density Function, PDF*) (Figura 7.19) sono definite quando:

- $\xi \geq 0$ per $\mu \leq x < \infty$ (distribuzione di Pareto tipo II) e presenta un comportamento *heavy tailed*;
- $\xi = 0$ la distribuzione diventa un'esponenziale (tipo I) con media $(\mu + \sigma)$ e presenta un comportamento *normal tailed*;
- $\xi < 0$ per $\mu \leq x \leq (\mu - \sigma/\xi)$ (tipo III) e risulta limitata superiormente (*short tailed* che generalmente non presenta utilità per quei fenomeni idrologici che non sono limitati superiormente).

Inoltre nel caso particolare in cui:

- $\xi = -1$ la distribuzione diventa una distribuzione uniforme nell'intervallo $[\mu, \mu + \sigma]$ anch'essa di scarso interesse idrologico

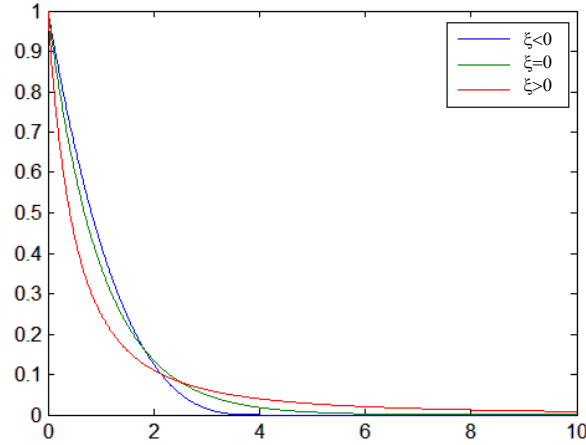


Figura 7.19 - Esempio della PDF della GP per diversi valori del parametro di forma.

Anche in questo caso bisogna prestare particolare attenzione alla formulazione della CDF (o della PDF) della GP e in particolare al parametro di forma ξ che viene spesso assunto di segno opposto, come, ad esempio, nel lavoro di Hosking e Wallis (1987), e quindi le considerazioni sul segno di ξ si ribaltano.

Il parametro di forma ξ è quindi definito in maniera tale che valori positivi implicano il comportamento *heavy tailed*, mentre valori negativi implicano una distribuzione limitata superiormente.

Particolare attenzione deve essere posta nella valutazione del livello di ritorno x_T corrispondente al periodo di ritorno T , considerato che con l'approccio *POT*, a differenza dell'approccio *AM*, possono essere selezionati più eventi indipendenti nell'arco di un anno.

Se il numero medio di eventi indipendenti (quelli da cui si estrae il campione sopra soglia) per anno è dato dal valore λ (detto *crossing rate*), il numero di eventi indipendenti in T anni è uguale a λT . Pertanto, l'evento che in T anni è superato una sola volta avrà una probabilità di superamento pari a 1 sul numero di eventi in T anni che è proprio λT :

$$\Pr(X > x_T) = 1 - F(x_T) = \frac{1}{\lambda T} \quad \lambda = \frac{n}{N_{anni}} \quad \text{eq. 32}$$

con la condizione che

$$1 - \frac{1}{\lambda T} > \frac{N_{\mu}}{n} \quad \text{eq. 33}$$

da cui:

$$T = \frac{1}{\lambda[1 - F(x_T)]} \Rightarrow x_T = F^{-1}\left(1 - \frac{1}{\lambda T}\right) \quad \text{eq. 34}$$

$$x_T = \begin{cases} \mu - \frac{\sigma}{\xi} \left[1 - \left(\frac{n}{N_{\mu}} \frac{1}{\lambda T} \right)^{-\xi} \right] & \xi \neq 0 \\ \mu - \sigma \ln \left(\frac{n}{N_{\mu}} \frac{1}{\lambda T} \right) & \xi = 0 \end{cases} \quad \text{eq. 35}$$

Quindi considerando che:

$$\frac{n}{N_\mu} \frac{1}{\lambda} = \frac{n}{N_\mu} \frac{N_{anni}}{n} = \frac{N_{anni}}{N_\mu} = \frac{1}{\lambda_\mu} \quad \text{eq. 36}$$

e avendo indicato con λ_μ il numero di superamenti di soglia per anno. Il quantile, pertanto, diventa:

$$x_T = \begin{cases} \mu - \frac{\sigma}{\xi} \left[1 - \left(\frac{1}{\lambda_\mu T} \right)^{-\xi} \right] & \xi \neq 0 \\ \mu - \sigma \ln \left(\frac{1}{\lambda_\mu T} \right) & \xi = 0 \end{cases} \quad \text{eq. 37}$$

Che è il medesimo risultato cui si sarebbe pervenuti considerando direttamente il quantile dei superamenti di soglia.

La stima dei parametri viene effettuata nel foglio ANÁBASI con i tre metodi classici:

- metodo dei momenti (*Method of Moments, MoM*)
- metodo dei momenti pesati in probabilità o L-momenti (*Probability Weighted Moment PWM or L-Moments LM*)
- metodo della massima verosimiglianza (*Maximum Likelihood ML*)

e vengono utilizzati per il calcolo dei livelli di ritorno i parametri forniti dal metodo di stima *PWM*.

Non è prevista in questa fase delle LG, anche per la notevole complessità, esporre la teoria della stima parametrica che si rimanda, per il momento, ai testi specializzati.

È tuttavia utile conoscere, anche per una migliore interpretazione del comportamento dei dati, alcuni risultati relativi ai metodi utilizzabili per stima dei parametri della *GPD*, che, a seconda dei metodi utilizzati, possono essere anche molto diversi e quindi influenzare in maniera determinante i valori di ritorno.

In particolare, come riportato in Deidda e Puliga (2008), il metodo dei momenti per la *GPD*, a fronte della estrema semplicità di applicazione, è teoricamente applicabile solo per valori del parametro di forma $\xi < 1/2$, poiché per tale valore la varianza diventa infinita. Tuttavia alcuni autori (Hosking e Wallis, 1987) suggeriscono di utilizzare questo metodo solo per $\xi < 1/4$. Per $\xi \cong 0$ l'accuratezza del metodo dei momenti è paragonabile a quella del metodo della massima verosimiglianza *ML*. È tuttavia utile notare che il metodo dei momenti è molto sensibile agli *outlier* poiché la media e la varianza campionaria sono poco robusti e quindi un singolo valore anomalo può modificare sensibilmente la parametrizzazione della distribuzione.

Il metodo *PWM* fornisce le migliori prestazioni per $\xi \cong 0.2$ mentre per $\xi < 0$ le stime risultano molto distorte. Diversi autori hanno dimostrato che il metodo *ML* costituisce il metodo migliore (più efficiente non distorto) per un elevato numero di dati. Per piccoli set di dati ($n < 100$) invece il metodo *PWM* è più efficiente.

7.5.3 Scelta del valore di soglia della *GPD*

Il problema principale che si presenta nell'utilizzo della *GPD* per analizzare i valori estremi è sicuramente quello della selezione della soglia. Non esiste, infatti, una formula che indichi in maniera univoca il valore ottimale da scegliere.

La scelta della soglia è il frutto di un compromesso tra la distorsione e la varianza della stima. Se, infatti, si pone una soglia troppo alta per assicurare l'indipendenza dei dati, potranno essere selezionati pochi eventi e la varianza di stima dei parametri sarebbe troppo elevata. Al contrario, la scelta di una soglia eccessivamente bassa, andrebbe contro le ipotesi d'indipendenza su cui si fondano i risultati illustrati facendo aumentare la distorsione delle stime. Inoltre un elevato numero di eventi statisticamente dipendenti porterebbe a una forte sottostima della varianza di stima, fornendo un'illusoria accuratezza del livello di ritorno (Figura 7.20).

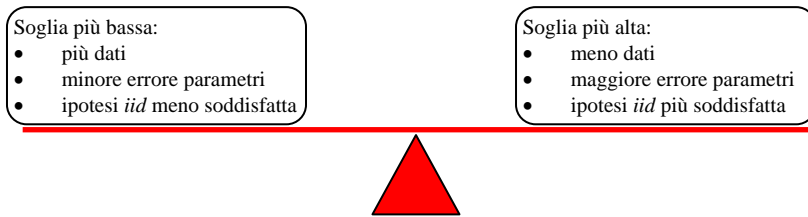


Figura 7.20 - Bilanciamento della scelta della soglia per l'analisi dei valori estremi con l'approccio POT

Non esiste una procedura per estrarre un insieme di variabili statisticamente indipendenti al di sopra di una soglia (*declustering*), ma è possibile definire un criterio che possa individuare variabili che si possono ritenere ragionevolmente indipendenti.

In genere il criterio dipende anche dalla particolare natura della variabile idrologica. Ad esempio, per una serie di portate di un corso d'acqua, alcuni autori (e.g., Willems, 2003) selezionano i massimi indipendenti considerando:

- che il tempo tra due picchi sia superiore alla costante di esaurimento della fase di recessione;
- che la portata minima tra due picchi sia inferiore al 37% del massimo colmo.

Nelle LG e nel foglio ANÁBASI, poiché le variabili idrologiche di cui si tratta sono di natura anche molto diversa (precipitazione, portata, temperatura, vento, ecc.) si propone un criterio di *declustering* secondo lo schema riportato in Figura 7.21:

- 1) si individuano inizialmente i valori di massimo relativo (escludendo quindi anche quei valori comunque elevati della grandezza che si trovano nella fase crescente o decrescente dell'evento);
- 2) tra i valori di massimo relativo si considerano indipendenti solo quei massimi separati da un intervallo temporale Δt maggiore di un tempo caratteristico t^* definito dall'operatore esperto in funzione delle caratteristiche della grandezza idrologica (e.g. portata) e del contesto ambientale dell'evento (e.g. superficie del bacino, tempo di corrvazione, ecc.);
- 3) tra i valori di massimo relativo indipendenti, si considerano solo quelli al di sopra della soglia prefissata.

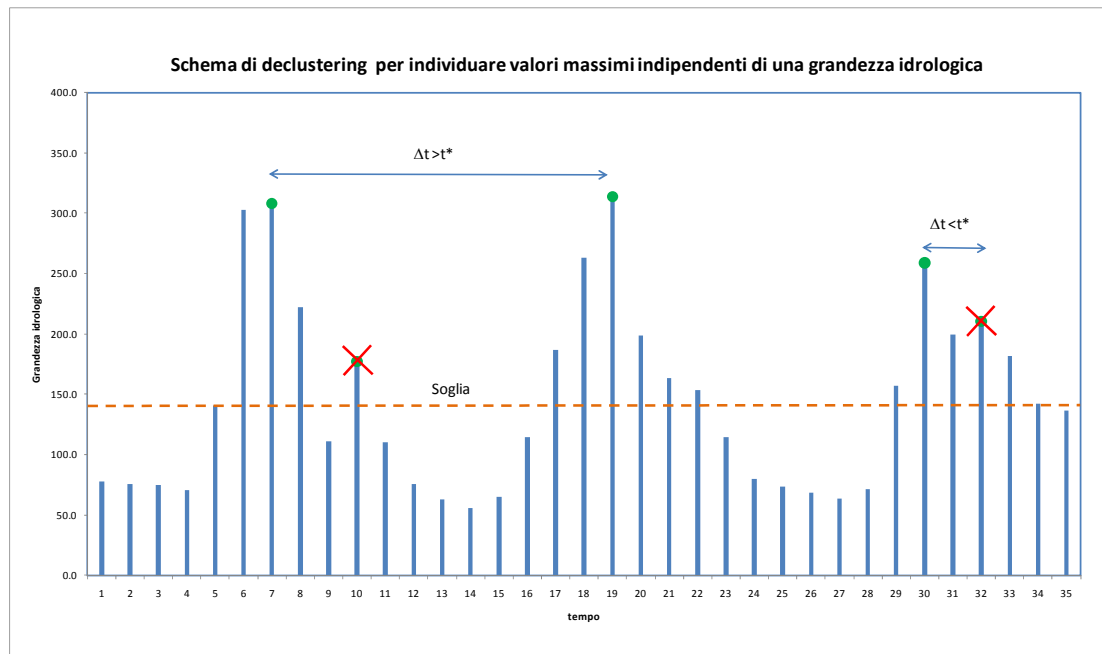


Figura 7.21 - Schema di declustering generale proposto nelle LG per l'approccio POT.

Un'alternativa al semplice *declustering* per la scelta del valore di soglia ottimale è quello di utilizzare una procedura diagnostica basata sulle caratteristiche della *GPD* e in particolare sulla media. In particolare si dimostra (vedi, Appendice B. Approfondimenti di statistica paragrafo 11.4.4.3) che teoricamente la media delle eccedenze sopra la soglia μ è una funzione lineare della soglia stessa. Se, per il campione di dati della serie, si riporta su un diagramma, la media delle eccedenze in funzione della soglia, si ottiene un diagramma detto "*mean residual life*" che per i valori di soglia μ , per il quale

la distribuzione generalizzata di Pareto risulta adatta, dovrebbe essere pressoché lineare. Pertanto la soglia ottimale si può scegliere come il valore più basso del tratto lineare nel diagramma *mean residual life*.

Il grafico viene corredato di intervalli di confidenza (di norma al 95%) basandosi sull'approssimazione di normalità della media. Nella Figura 7.22 il grafico *mean residual life* mostra un andamento approssimativamente lineare intorno al valore 400, cui segue un andamento irregolare dovuto alla riduzione della numerosità del campione al crescere della soglia.

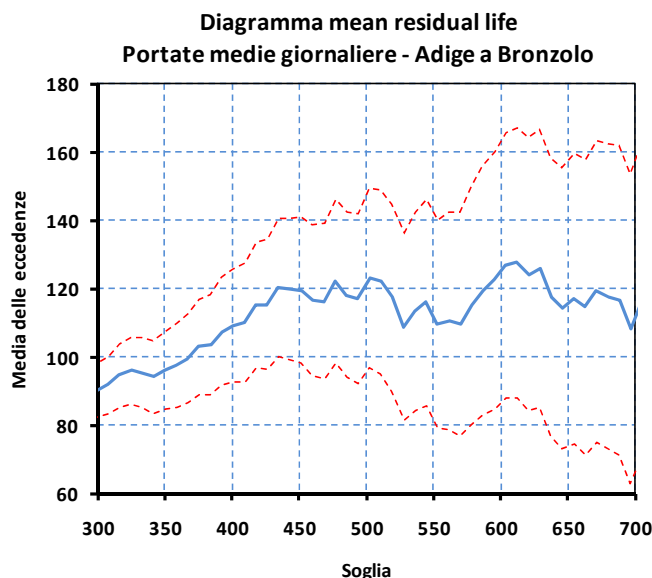


Figura 7.22 - Grafico “*mean residual life*”. In rosso sono riportate le fasce di confidenza al 95%

Poiché l'interpretazione del grafico *mean residual life* non è sempre facile e immediata, per procedere alla scelta della soglia si tiene conto anche di un altro metodo, basato sulla proprietà di stabilità dei parametri che caratterizzano la *GPD*.

In particolare se la distribuzione *GPD* è valida per una soglia μ_0 , allora lo è anche per una soglia $\mu > \mu_0$, con parametro di forma invariato e parametro di scala che varia linearmente con μ secondo la relazione:

$$\sigma_\mu = \sigma_{\mu_0} + \xi(\mu - \mu_0) \tag{eq. 38}$$

Il parametro σ_μ può essere quindi riparametrizzato come ($\sigma^* = \sigma_\mu - \xi\mu$) che è costante al variare della soglia μ . Quindi, σ^* e ξ dovrebbero essere costanti sopra μ_0 , se questa è una soglia valida affinché le eccedenze seguano una distribuzione *GPD*. Nella Figura 7.23 e nella Figura 7.24 i valori dei parametri della *GPD* sono approssimativamente costanti tra i valori 350 e 400. Pertanto la soglia ottimale dovrebbe essere compresa tra i valori 350 e 400.

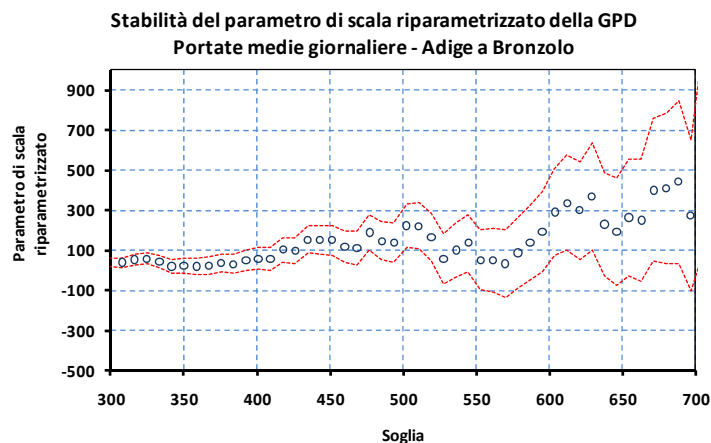


Figura 7.23 - Stabilità del parametro di scala della GPD. In rosso sono riportate le fasce di confidenza al 95%

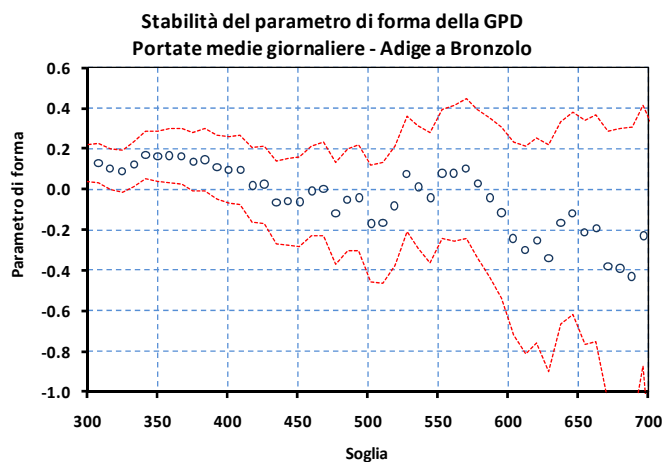


Figura 7.24 - Stabilità del parametro di forma della GPD. In rosso sono riportate le fasce di confidenza al 95%

7.5.4 Incertezza sulla stima dei parametri della GEV e della GPD

Nell'analisi degli estremi di una variabile idrologica, una volta individuata la più adatta distribuzione di probabilità e dopo averne effettuato la stima dei relativi parametri, le domande da porsi sono:

- 1) quanto accurata è la stima dei parametri;
- 2) quanto accurata è la stima dei livelli di ritorno;
- 3) quanto è accettabile l'estrapolazione a elevati periodi di ritorno.

La stima dei parametri, qualunque sia il metodo adottato, presenta un'incertezza legata alla finita numerosità del campione, che è di fondamentale importanza tenerne conto nella valutazione di valori di progetto di variabili idrologiche, generalmente espressi come valori relativi ad assegnati periodi di ritorno.

L'incertezza sul parametro è espressa attraverso il cosiddetto *standard error* (*s.e.*) che altro non è che la deviazione standard (radice quadrata della varianza) dello stimatore del parametro stesso.

Lo stimatore di un parametro θ , (e.g. la media campionaria) è esso stesso una variabile aleatoria (in quanto funzione di variabili aleatorie) e come tale ha una media e una varianza. La sua deviazione standard è lo *standard error*.

Ipotizzando che asintoticamente (quando cioè la dimensione del campione N tende all'infinito) lo stimatore si distribuisce come una variabile normale, l'intervallo di confidenza di assegnata probabilità α dello stimatore diventa:

$$\hat{\theta} - z_{1-\alpha/2} \cdot (s.e.) \leq \theta \leq \hat{\theta} + z_{1-\alpha/2} \cdot (s.e.) \quad \text{eq. 39}$$

Dove $z_{1-\alpha/2}$ è il quantile della distribuzione normale standardizzata $N(0,1)$.
 In genere poiché $\alpha = 0.95$, l'intervallo di confidenza degli stimatori è semplicemente:

$$\hat{\theta} - 1.96 \cdot \sqrt{\text{var}(\hat{\theta})} \leq \theta \leq \hat{\theta} + 1.96 \cdot \sqrt{\text{var}(\hat{\theta})} \quad \text{eq. 40}$$

L'incertezza della stima dei parametri della distribuzione si riflette, quindi, sulla stima dei quantili e dei livelli di ritorno. È pertanto necessario stimare lo *s.e.* del quantile x_q . Non sempre è possibile ottenere espressioni semplici per lo *s.e.* del quantile x_q in funzione dello *s.e.* dei parametri della distribuzione che, in generale, diventa (sempre considerando l'intervallo di confidenza al 95%)

$$\hat{x}_q - 1.96 \cdot \sqrt{\text{var}(\hat{x}_q)} \leq x_q \leq \hat{x}_q + 1.96 \cdot \sqrt{\text{var}(\hat{x}_q)} \quad \text{eq. 41}$$

È sempre necessario indicare nella stima del quantile x_q o del livello di ritorno x_T il corrispondente *standard error*. Ciò consentirebbe di valutare l'incertezza (l'intervallo di confidenza) insita nel valore che generalmente viene posto a base di una progettazione.

Tale incertezza, com'è intuitivo, dipende in maniera determinante dalla dimensione del campione e anche dal metodo di stima dei parametri. Ad esempio, il metodo di stima *ML* fornisce, per alcuni valori del parametro di forma, stime con uno *s.e.* inferiore rispetto al metodo *PWM*. In questo caso viene detto che uno stimatore è più efficiente dell'altro.

7.5.5 Scelta dell'approccio AM o POT

Nell'analisi dei valori estremi emerge il problema di individuare quale sia l'approccio migliore per un determinato problema. Ovviamente la risposta è che non esiste un metodo migliore dell'altro ma la scelta è conseguenza della valutazione dei pro e dei contro di ciascun approccio, in relazione al problema e al dato idrologico che si vuole analizzare (Tabella 7.5).

Uno dei maggiori limiti delle serie *AM* è quello di considerare un solo valore per anno, tralasciando altri eventi significativi che possono verificarsi nello stesso anno, trascurando, di fatto, un'importante informazione. Il secondo evento più intenso in un anno che viene tralasciato, potrebbe essere di gran lunga maggiore di eventi massimi occorsi in altri anni. Un altro limite è quello per cui, considerando un solo evento per anno, generalmente le serie sono molto limitate.

Tabella 7.5 - PRO e CONTRO degli approcci BM e POT

Block Maximum (BM)	Peak-Over-Threshold (POT)
PRO	
Indipendenza dei dati assicurata dalla selezione di blocchi di grande ampiezza (e.g. anno)	Possibilità di selezionare un campione più numeroso e quindi eseguire stime più efficienti (anche se rimane il dubbio che un numero maggiore di valori estremi bassi possa aggiungere informazioni utili sul comportamento dei valori estremi elevati)
Facilità di applicazione	
Ipotesi teoriche più realizzabili in pratica	
CONTRO	
Elevata incertezza della stima (bassa efficienza) a causa della dimensione limitata del campione	Arbitrarietà della scelta della soglia, benché basata su procedure diagnostiche
	Minore facilità di applicazione
	Indipendenza tra le variabili difficile da assicurare, anche utilizzando tecniche di <i>declustering</i> .

Il maggior limite invece delle serie *POT* è quello di poter considerare eventi che non siano indipendenti, venendo quindi a mancare una delle ipotesi base della *EVA*. La possibilità di scegliere un campione più numeroso, rispetto all'approccio *AM*, ma con la possibilità di eventi non indipendenti, potrebbe comportare una forte sottovalutazione dell'incertezza connessa alla stima dei livelli di ritorno e un'illusoria accuratezza, se non si tiene in debito conto la dipendenza statistica tra le variabili selezionate.

Anche la scelta della soglia è un problema che, benché basato su strumenti diagnostici, è connotato da forti elementi di arbitrarietà. Inoltre mediante l'approccio *POT*, in generale, si possono selezionare un numero maggiore di dati rispetto all'approccio *AM* ma che, essendo per lo più relativi a valori estremi bassi, potrebbero non aumentare e migliorare, o peggio alterare, la conoscenza sul comportamento dei valori estremi elevati.

7.5.6 Relazione tra la *GEV* e la *GPD*

Tra le distribuzioni derivanti dai due diversi approcci sul medesimo insieme di dati esiste uno stretto legame. La distribuzione asintotica delle eccedenze rispetto a una soglia μ è strettamente legata alla distribuzione asintotica del massimo annuale (in generale del massimo a blocchi) di una variabile aleatoria. Si dimostra che se la distribuzione del massimo annuale è una *GEV* con i parametri di forma, di scale e di posizione indicati con il pedice "*GEV*":

$$F(x) = \exp \left\{ - \left[1 + \xi_{GEV} \frac{x - \mu_{GEV}}{\sigma_{GEV}} \right]^{-\frac{1}{\xi}} \right\} \quad \text{eq. 42}$$

allora la distribuzione delle eccedenze sopra soglia $Y = (X - \mu)$, condizionata ad $X > \mu$, è approssimativamente:

$$F(y) = 1 - \left[1 + \xi_{GPD} \frac{y}{\sigma_{GPD}} \right]^{-\frac{1}{\xi_{GPD}}} \quad \text{eq. 43}$$

E tra i parametri delle due distribuzioni sussistono le seguenti relazioni:

$$\begin{aligned} \xi_{GPD} &= \xi_{GEV} = \xi \\ \sigma_{GPD} &= \sigma_{GEV} + \xi(\mu_{GPD} - \mu_{GEV}) \end{aligned} \quad \text{eq. 44}$$

Tale circostanza suggerisce un'efficace utilizzo del binomio *GEV-GPD* e in particolare per l'inferenza statistica sui parametri. Infatti, essendo le due distribuzioni legate fra loro, l'inferenza statistica sulla *GPD*, che può essere condotta su una base dati molto ampia, può essere usata per irrobustire l'inferenza statistica sulla *GEV*, i cui parametri vengono usualmente stimati su un numero molto ridotto di osservazioni (soltanto sui massimi annuali).

7.5.7 Considerazioni sui dati mancanti

Nell'analisi degli eventi estremi, il problema dei dati mancanti emerge prevalentemente nello studio dei massimi annuali, in cui il calcolo delle probabilità di superamento e dei tempi di ritorno è basato spesso su campioni di lunghezza limitata (comunemente < 100). Poiché il calcolo dei quantili con un determinato tempo di ritorno definisce i valori superati un certo numero di volte in N anni consecutivi di osservazione, qualora vi siano n osservazioni mancanti, la stima risulta distorta in quanto l'informazione è disponibile solo per $N-n$ anni. Ne segue che l'assenza di dati può essere trascurata se inferiore a 2-5 osservazioni (in cui il limite inferiore è riferito alle serie più brevi).

Il problema è meno evidente per le analisi eseguite con il metodo dei picchi sopra soglia, in cui il tempo di ritorno in anni è calcolato in base al numero medio annuo dei superamenti della soglia selezionata. In questo caso, la numerosità del campione è superiore a quella dei massimi annuali e il tempo di ritorno è definito in modo indiretto a partire dal numero dei superamenti dalla soglia, il quale non è vincolato alla continuità degli anni di osservazione. È tuttavia opportuno che per ogni anno, la serie utilizzata per estrarre i valori sopra soglia (in genere dati giornalieri o a scala temporale più fine) sia disponibile con sufficiente continuità, affinché la selezione sopra soglia abbia un carattere di omogeneità su tutta la serie. La continuità durante l'anno è in ogni modo meno stringente di quella richiesta per i massimi annuali, in ragione del fatto che vengono selezionati più valori per ogni anno e

i dati mancanti possono influire in modo decisivo solo se la percentuale è elevata e/o essi sono concentrati, ad esempio, nelle stagioni in cui la grandezza studiata presenta i valori più elevati.

7.5.8 Ulteriori approfondimenti

Per approfondimenti teorici sull'analisi degli estremi si veda Appendice B, paragrafo 11.4.4. Nell'ambito della teoria degli estremi rientrano anche le curve IDF (*Intensity Duration Frequency*) che nella presente versione delle LG vengono esposte solo nell'Appendice B, paragrafo 11.4.5.

8. Schema procedura ed esempi di applicazione

In questo capitolo si mostra l'applicazione delle metodologie descritte in precedenza ad alcune serie idrologiche. Gli esempi sono svolti con l'intento di indicare un percorso di analisi coerente che a partire da una caratterizzazione qualitativa evolva progressivamente verso una sintesi quantitativa dell'informazione di interesse.

In generale, gran parte dei dati idrologici sono disponibili a una scala di risoluzione temporale giornaliera. Per alcune analisi è tuttavia più opportuno lo studio delle serie a scale temporali più elevate (mensile e annuale). Nel seguito, le analisi saranno eseguite alle tre scale temporali giornaliera, mensile e annuale, scegliendo di volta in volta la scala (o le scale) di risoluzione più adeguate per l'applicazione di ciascun analisi.



8.1 Schema procedura

Il percorso di analisi di una serie storica si può sintetizzare nel diagramma seguente.

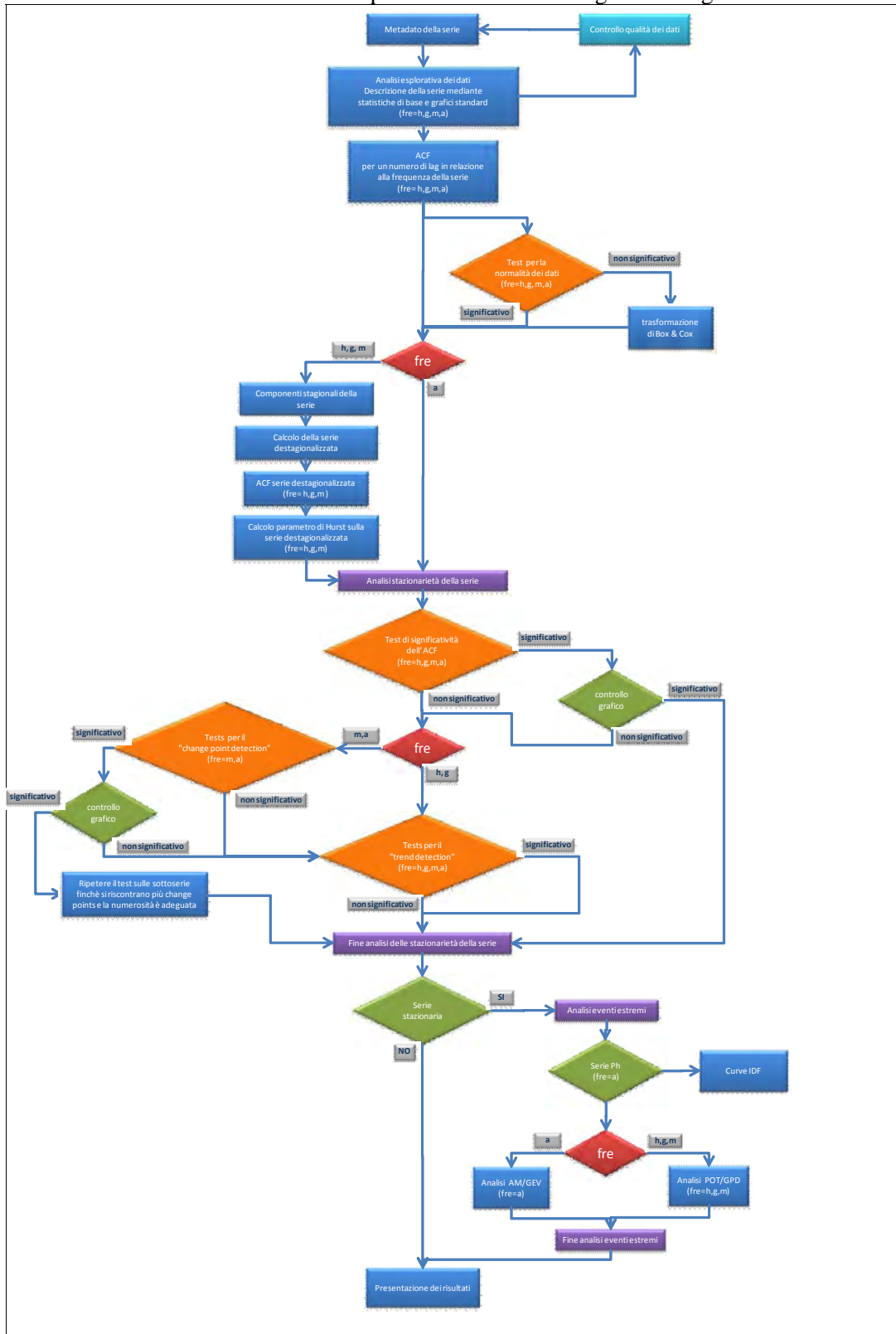


Figura 8.1 - Diagramma di flusso della procedura di analisi di una serie storica di dati idrologici.

8.2 Analisi di una serie di portate

Vengono di seguito analizzate, alle diverse scale di aggregazione, le portate del fiume Adige registrate nella stazione di Bronzolo.

8.2.1 Scala giornaliera

8.2.1.1 Metadato e caratteristiche

Il primo passo nell'analisi di una serie di dati idrologici è quello di procedere alla raccolta dei metadati e alla definizione delle principali caratteristiche quantitative. Il set dei metadati è costituito da 3 schede (scheda A1, scheda A2, scheda A3) rispettivamente relative ai metadati tecnico-amministrativi, ai metadati geografici e ai metadati relativi alle modalità di rilevamento (Tabella 8.1).

Tabella 8.1 - Metadati della serie storica. Portate medie giornaliere - Adige a Bronzolo.

SCHEDA A1: metadati tecnico-amministrativi	
metadato	valore
Titolo della serie	Portate medie giornaliere - Adige a Bronzolo
Grandezza idrologica	Portata media giornaliera
Unità misura	m ³ /s
Simbolo	Q _g
Tipo grandezza (primitiva/derivata)	derivata
Identificatore (codice) locale	855
Identificatore (codice) nazionale	855
Ente responsabile e fonte del dato	Ufficio Idrografico della Provincia Autonoma di Bolzano
Disponibilità (URL)	http://www.provincia.bz.it/meteo/stazioni-idrometriche.asp?stat_stid=133
Ultimo aggiornamento	

SCHEDA A2: metadati geografici	
metadato	valore
Nome stazione	Adige a Bronzolo
Comune	Bronzolo
Provincia	BZ
Regione	Trentino Alto Adige
Coordinate:	
Datum (ellissoide)	WGS84
Proiezione	GEO
Long/X (°/m)	11.31536
Lat/Y(°/m)	46.41377
Quota geoidica (m s.l.m.)	226.98
Link geolocalizzazione su web	http://maps.google.com/maps?f=q&source=s_q&hl=it&geocode=&sll=46.41377,11.31536&z=16&q=46.41377+11.31536
Bacino idrografico	Adige
Superficie bacino chiusura (km ²)	6926

SCHEDA A3: metadati modalità rilevamento	
metadato	valore
Numero periodi di campionamento omogeneo	1
Intervallo di campionamento omogeneo. Da	01/01/1980
a	31/12/2006
Grandezza idrologica primitiva	h _g
Intervallo campionamento grandezza idrologica primitiva	1g
Funzione applicata alla grandezza idrologica primitiva	Trasformazione non lineare
Grandezza idrologica derivata	Q _g
Classe di accuratezza della grandezza derivata	C
Funzione Aggregazione/Selezione	nessuna
Percentuale massima dati mancanti nell'aggregazione/selezione	0
Standard di rilevamento	SIMN

Per completare l'informazione sulla serie di dati viene definita una scheda standard (scheda B) con le principali caratteristiche quantitative della serie (Tabella 8.2)

Tabella 8.2 - Caratteristiche quantitative della serie. Portate medie giornaliere - Adige a Bronzolo.

SCHEDA B: descrizione statistica	
caratteristica	valore
Numero massimo di dati	9862
Numero totale di dati	9862
Frequenza (numero massimo dati/anno)	365
Numero di anni	27
Istante primo dato	01/01/1980 00.00
Istante ultimo dato	31/12/2006 00.00
Valore massimo	1018.0
Valore minimo	26.9
Dati mancanti (e/o ricostruiti)	0
Intervalli di dati mancanti	0
Completezza	100.00%
Continuità	100.00%
iQuaSI	0.25

8.2.1.2 Descrizione statistica e grafici standard

Il secondo passo è quello di procedere all'analisi esplorativa dei dati mediante grafici standard e la tabella riassuntiva dei principali indici statistici (Tabella 8.3).

Il primo grafico da realizzare è il diagramma cronologico (*time plot*) con il quale si evidenziano gli andamenti sistematici rispetto al tempo.

La Figura 8.2 mostra la serie temporale completa delle portate medie giornaliere.

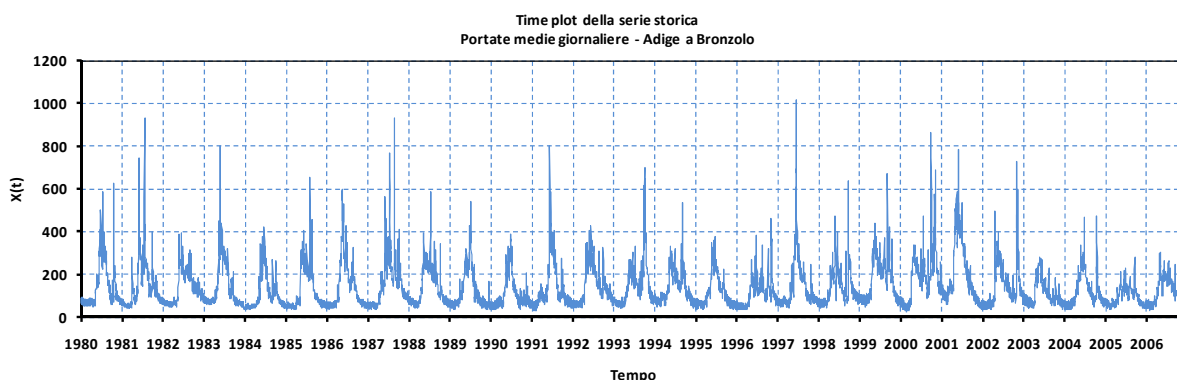


Figura 8.2 - Diagramma cronologico della serie (*time plot*). Portate medie giornaliere - Adige a Bronzolo.

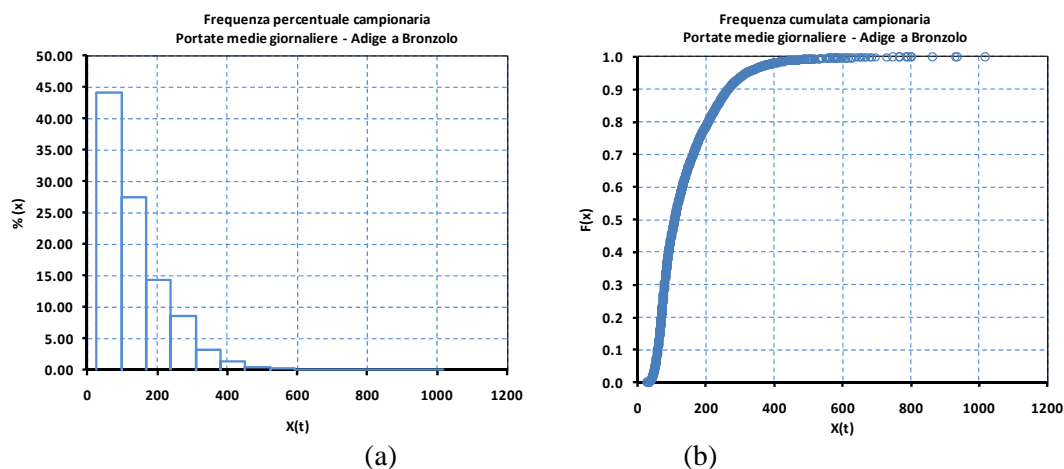


Figura 8.3 - Grafici della frequenza percentuale campionaria e della frequenza cumulata. Portate medie giornaliere - Adige a Bronzolo.

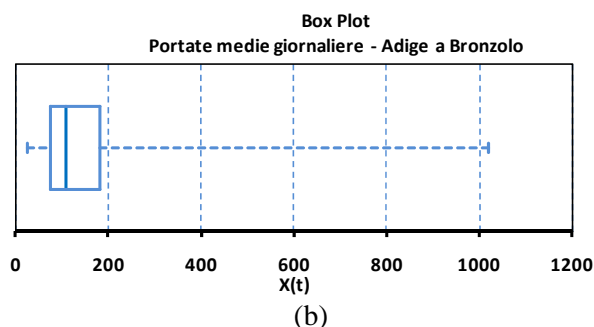
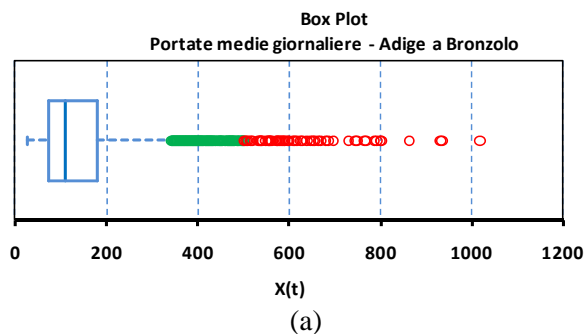


Figura 8.4 - Box plot con e senza l'indicazione del VAI e VAS. Portate medie giornaliere - Adige a Bronzolo.

Tabella 8.3 - Statistiche univariate campionarie per la descrizione sintetica della serie. Portate medie giornaliere - Adige a Bronzolo.

SCHEDA C: statistiche base			
statistica	simbolo	valore	s.e.
Indici di posizione			
Media	m	140.48	
Moda		-	
Minimo (percentile 0%)	Min	26.90	
Percentile 25% (1° quartile Q_1)	Q_1	74.30	
Mediana (percentile 50%)	Q_2	108.50	
Percentile 75% (3° quartile Q_3)	Q_3	181.10	
Massimo (percentile 100%)	Max	1018.00	
Indici di dispersione			
Range Inter Quartile (Q_3-Q_1)	IQR	106.80	
Range (Max-Min)	R	991.10	
Valore adiacente inferiore	VAI	26.90	
Valore adiacente superiore	VAS	341.30	
Scarto quadratico medio	s	93.9	
Varianza	s^2	8809.5	
Median Absolute Deviation	MAD	42.0	
Coefficiente di variazione	CV	0.7	
Indici di forma			
Asimmetria	g	2.03	
Curtosi	k	6.8	
Percentuale di dati su cui sono calcolate le statistiche		100.0%	

I dati di portata giornaliera dell'Adige a Bronzolo, come tutti i dati di portata giornaliera, si presentano positivamente asimmetrici. Tale comportamento è legato ai fenomeni che presiedono alla formazione dei deflussi e si manifesta maggiormente per i piccoli corsi d'acqua a carattere torrentizio, alimentati prevalentemente dalle precipitazioni. L'istogramma della frequenza percentuale campionaria (Figura 8.3a) e i box plot (Figura 8.4) forniscono un'immediata visualizzazione di questa caratteristica. Sempre per effetto della dinamica della formazione dei deflussi, si trova un valore della curtosi positiva e molto lontana dal valore della distribuzione normale.

8.2.1.3 Analisi di stazionarietà

La mutua dipendenza temporale delle osservazioni è evidenziata dalla funzione di autocorrelazione mostrata nella Figura 8.5.

La serie presenta un'evidente periodicità annuale (stagionalità) legata al regime climatico che interessa l'arco alpino. La periodicità stagionale è chiaramente evidenziata dall'andamento della funzione di autocorrelazione della serie giornaliera che presenta dei massimi a *lag* (giorni) multipli di 365. Ancorché assolutamente evidente, si procede comunque ai test per la verifica dell'autocorrelazione e riportati nella Tabella 8.4. In questo caso, ovviamente, l'ipotesi nulla di assenza di correlazione è rigettata da entrambi i test.

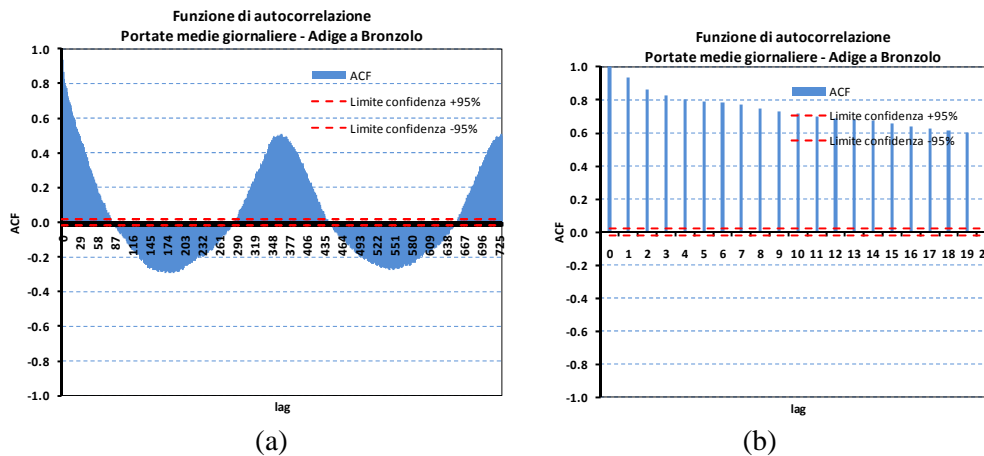


Figura 8.5 - Funzione di autocorrelazione per un periodo di 2 anni (a) e per i primi 20 giorni (b). Portate medie giornaliere - Adige a Bronzolo.

Tabella 8.4 - Esito dei test di autocorrelazione (test sui primi 10 lag). Portate medie giornaliere - Adige a Bronzolo.

Test	Statistica	P-Value	Livello di significatività	Esito
Ljung-Box	63091.320	0.000	5%	IPOTESI H ₀ RIGETTABILE
Box-Pierce	63046.129	0.000	5%	IPOTESI H ₀ RIGETTABILE

Eseguita la verifica della correlazione lineare tra i dati, nella quale emerge con evidenza l'andamento stagionale, si procede alla individuazione delle componenti stagionali della media e della varianza che caratterizzano in modo univoco la periodicità stagionale.

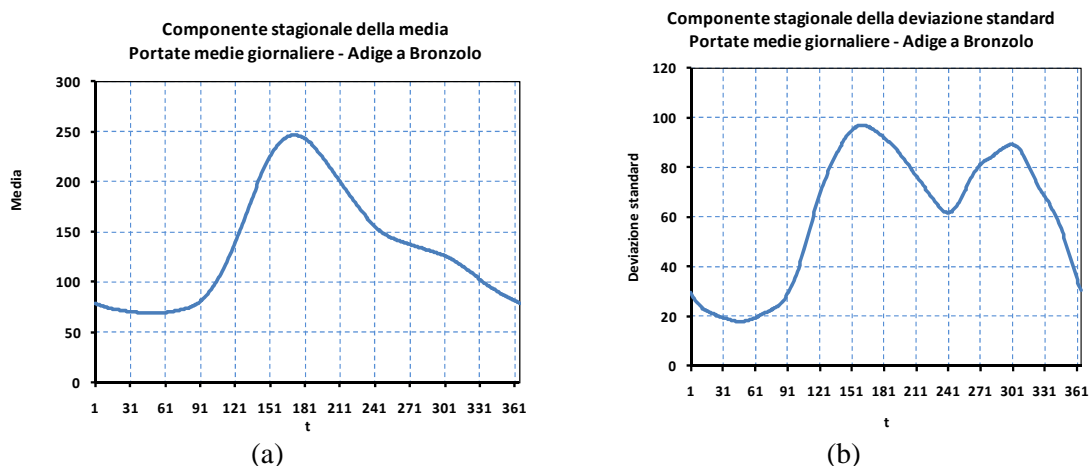


Figura 8.6 - Componente stagionale della media (a) e della deviazione standard (b). Portate medie giornaliere - Adige a Bronzolo.

La componente stagionale della media presenta un unico massimo in estate con un andamento analogo a quello del regime pluviale continentale, che interessa l'arco alpino tra il lago di Lugano e le sorgenti Tagliamento. Di contro, la componente della deviazione standard ha un andamento del tutto simile a quello del regime pluviale sublitoreo alpino. In altri termini, è probabile che il regime pluviale che

interessa le sorgenti e gli affluenti influenzi la componente media, mentre la variabilità sia più influenzata dal regime pluviometrico che interessa la stazione di misura.

Note le componenti stagionali della media e della varianza riportate rispettivamente in Figura 8.6a e in Figura 8.6b, si procede alla destagionalizzazione della serie riportata in Figura 8.7. Come si vede la serie destagionalizzata non presenta, con evidenza, un andamento stagionale, ma dall'analisi dell'ACF (Figura 8.8a, Figura 8.8b, Figura 8.8c) presenta un andamento ciclico con periodo di 7 giorni e una forte autocorrelazione che si attenua, ma che non si estingue sul lungo periodo. Tale comportamento è sintomo di una persistenza e di una memoria su lungo periodo.

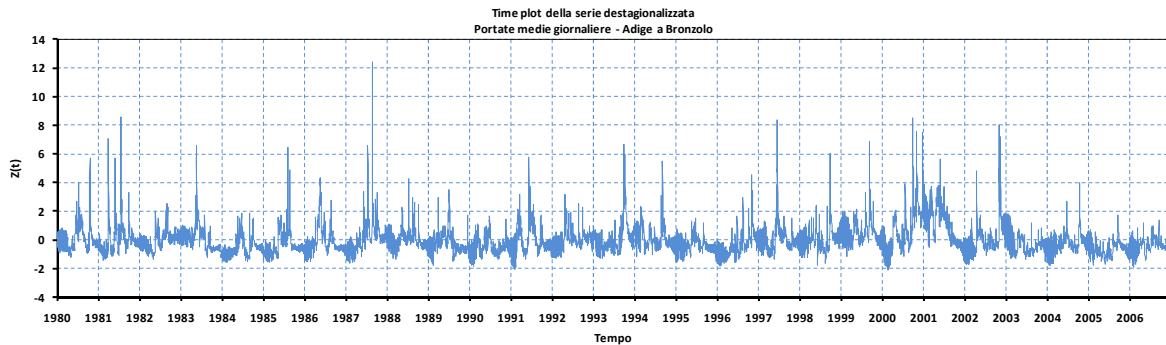
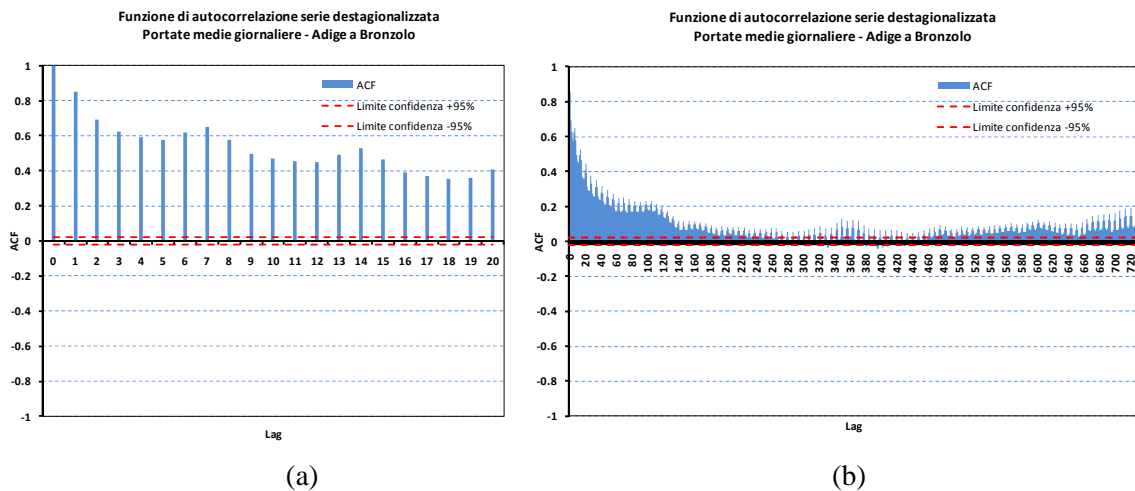
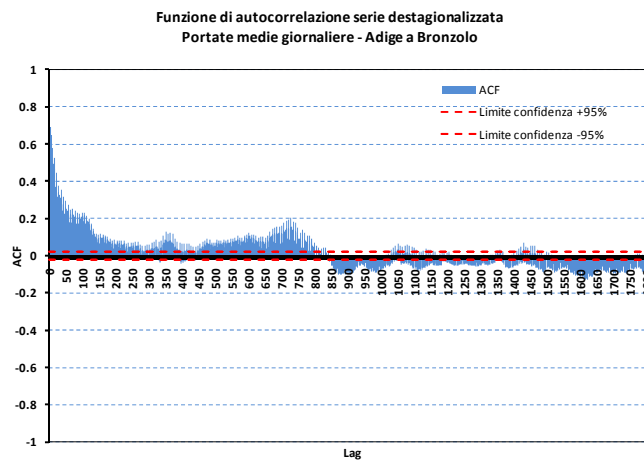


Figura 8.7 - Diagramma cronologico della serie destagionalizzata. Portate medie giornaliere - Adige a Bronzolo.



(a)

(b)



(c)

Figura 8.8 - Funzione di autocorrelazione della serie destagionalizzata su un periodo di 20 giorni (a), 2 anni (b) e 5 anni (c). Portate medie giornaliere - Adige a Bronzolo.

Tabella 8.5 - Esito dei test di autocorrelazione della serie destagionalizzata (test sui primi 5 lag). Portate medie giornaliere - Adige a Bronzolo.

Test	Statistica	P-Value	Livello di significatività	Esito
Ljung-Box	38269.2	0.00	5%	IPOTESI H ₀ RIGETTABILE
Box-Pierce	38243.2	0.00	5%	IPOTESI H ₀ RIGETTABILE

Anche in questo caso, benché assolutamente evidente, si procede comunque alla valutazione dei test di autocorrelazione per la serie destagionalizzata che forniscono esito negativo (rigettano, cioè, l'ipotesi nulla H₀).

Per rilevare la persistenza nella serie si esegue il calcolo del parametro di Hurst sulla serie destagionalizzata. Il metodo della varianza aggregata fornisce un valore di circa $H = 0.85$ (Figura 8.9a e Figura 8.9b) confermando così la presenza di memoria lunga con un comportamento persistente, ossia intervalli temporali in cui andamenti crescenti (decrementi) nella serie tendono a preservarsi. Questo comportamento è evidenziato anche dall'andamento dell'ACF della serie destagionalizzata (Figura 8.8c), il cui lento decadimento indica la presenza di persistenza nel segnale.

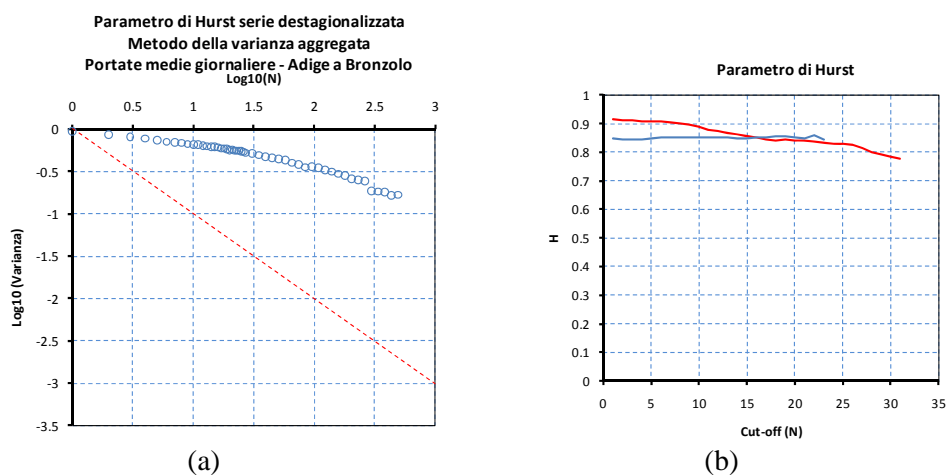


Figura 8.9 - Stima del parametro di Hurst relativo alla serie destagionalizzata $H = 0.85$. Portate medie giornaliere - Adige a Bronzolo.

Nella Tabella 8.6 è riportata la sintesi dell'analisi di stazionarietà della serie. Non sono stati eseguiti i test per evidenziare i *trend* e i *change point* poiché a scala giornaliera non forniscono risultati affidabili a causa della variabilità delle osservazioni, né sono stati eseguiti sulla corrispondente serie destagionalizzata per la presenza di persistenza che induce un'autocorrelazione non stagionale e non trascurabile.

Tabella 8.6 - Tabella di sintesi dell'analisi di stazionarietà. Portate medie giornaliere - Adige a Bronzolo.

SCHEDA D: analisi di stazionarietà	
analisi	livello
Autocorrelazione	significativa
Stagionalità	presente
Ciclicità	presente
Memoria a lungo termine	persistente
Trend	non effettuato
Change point	non effettuato
Normalità	non effettuata
Non-stazionarietà in media	significativa
Non-stazionarietà in varianza	significativa

8.2.1.4 Analisi degli estremi

Nella serie a scala giornaliera si procede all'analisi dei valori estremi mediante l'approccio POT. L'approccio AM viene invece utilizzato analizzando la serie dei massimi annuali delle portate

giornaliere. Il confronto tra i due approcci e la verifica delle importanti relazioni tra le due distribuzioni viene effettuata contestualmente all'analisi AM.

La scelta della soglia è effettuata con il supporto dei grafici diagnostici *mean residual life* (Figura 8.10) e di stabilità dei parametri (Figura 8.12a e Figura 8.12b) nei quali si riconosce che il valore di soglia pari a 400 m³/s può costituire un valore accettabile sufficientemente alto per garantire l'indipendenza dei valori e non troppo elevato per selezionare un numero adeguato di eventi che superano la soglia.

Nell'esempio in esame, un comportamento approssimativamente lineare nel grafico "*mean residual life*" è visibile tra 350 e 450 m³/s. La figura mostra che le stime dei parametri di scala e forma si stabilizzano e sono pressoché costanti nell'intervallo di soglia compreso tra 350 e 400 m³/s, in cui i valori dei parametri definiscono approssimativamente un plateau e l'incertezza rimane contenuta. La soglia di 400 m³/s appare una scelta ragionevole che consente di ottenere un campione di 185 eventi (Figura 8.11), con un *crossing rate* $\lambda = 6.85$ (ben oltre 6 volte il numero di anni della serie che costituisce la numerosità della serie dei massimi annuali) (Figura 8.13). A un'analisi più approfondita, tuttavia, molti di tali valori sono risultati appartenere ai medesimi eventi. Per ovviare si procede alla selezione di eventi (o *cluster*) indipendenti secondo il criterio per cui si considerano ragionevolmente indipendenti quei massimi relativi di portata giornaliera (picchi locali) che sono separati da un intervallo temporale minimo di 3 giorni (intervallo di tempo presumibilmente superiore al tempo di corrivazione del bacino in esame).

Il *declustering* ha invece consentito, considerando lo stesso valore di soglia uguale a 400 m³/s, di selezionare 63 eventi con un *crossing rate* $\lambda = 2.62$ (Figura 8.14).

Si procede, quindi, alla stima dei parametri della GPD per entrambi i casi, con e senza *declustering*, e alla stima dei livelli di ritorno.

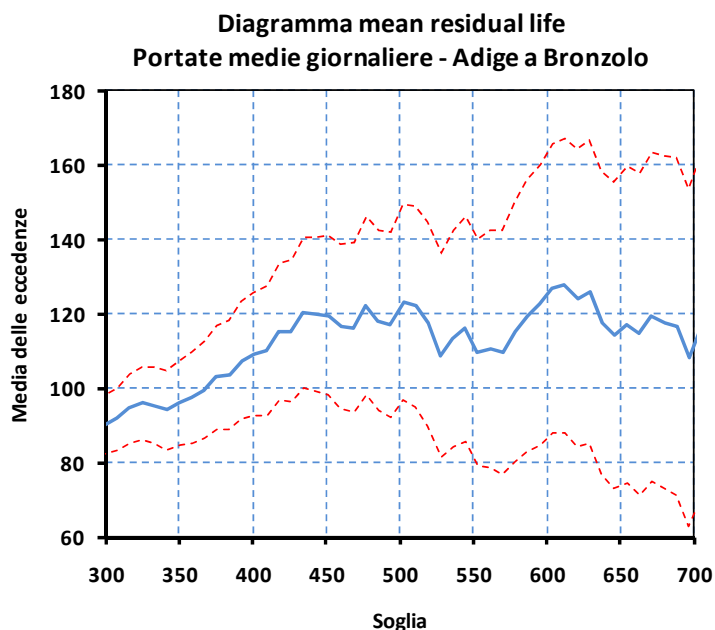


Figura 8.10 - Mean residual life. Portate medie giornaliere - Adige a Bronzolo.

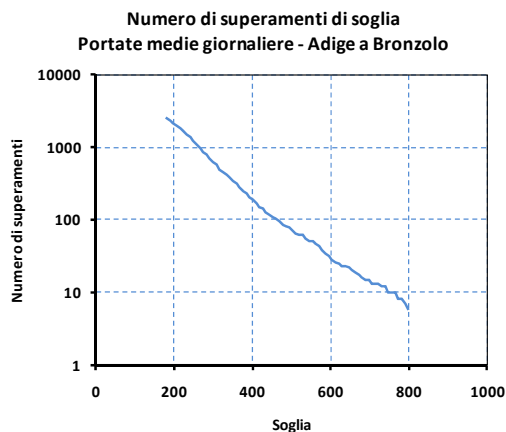


Figura 8.11 - Numero di superamenti in funzione del livello di soglia. Portate medie giornaliere - Adige a Bronzolo.

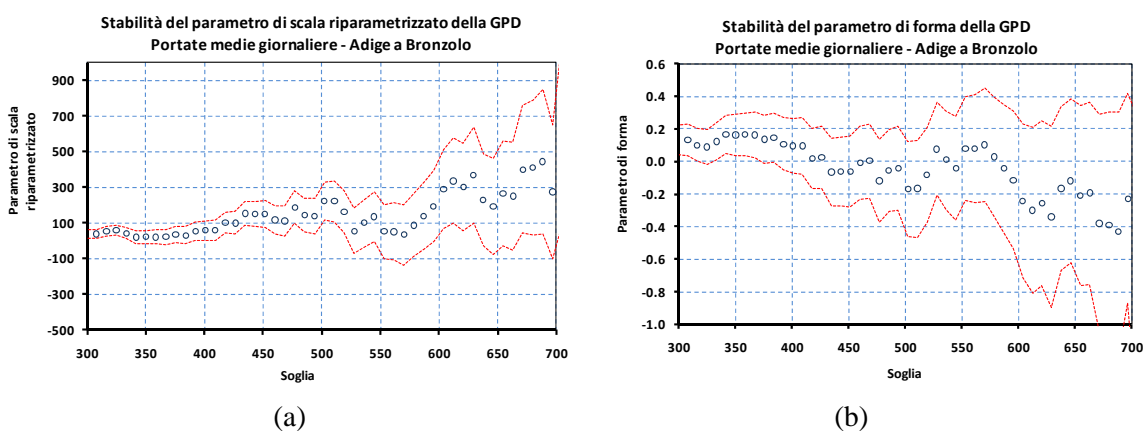


Figura 8.12 - Stabilità dei parametri della distribuzione GPD in funzione del livello di soglia. Portate medie giornaliere - Adige a Bronzolo.

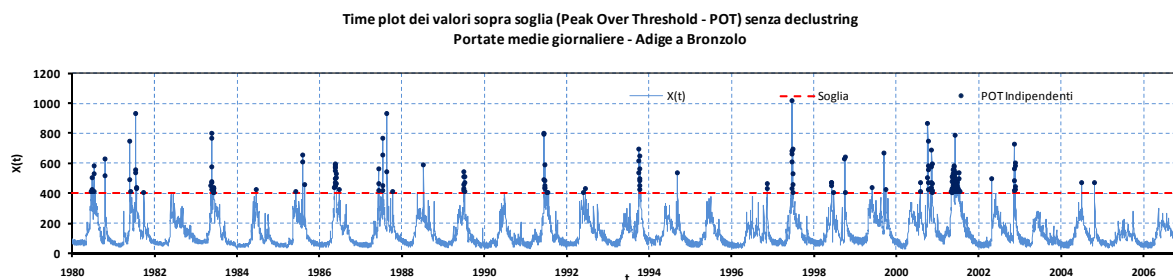


Figura 8.13 - Serie storica dei valori sopra la soglia di $400 \text{ m}^3/\text{s}$ senza declustering. 185 superamenti e un crossing rate di 6.85. Portate medie giornaliere - Adige a Bronzolo.

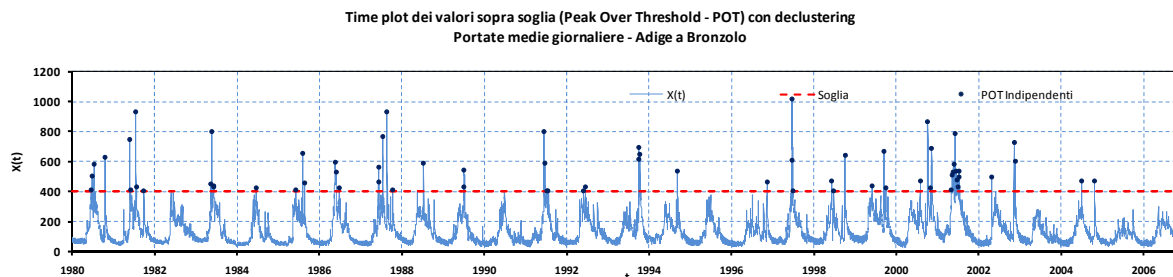


Figura 8.14 - Serie storica dei valori sopra la soglia di $400 \text{ m}^3/\text{s}$ con declustering: 63 superamenti e un crossing rate di 2.62. Portate medie giornaliere - Adige a Bronzolo.

Come si evince dalla Tabella 8.7e dalla Tabella 8.8 la stima dei parametri con e senza *declustering* differisce sensibilmente. Ma ciò che è ancora più significativo, è la notevole differenza tra gli *s.e.* della stima dei parametri. Tale differenza dipende dal numero degli eventi selezionati, maggiore nel caso senza *declustering*. Il fatto di aver selezionato molti eventi dipendenti comporta una non corretta applicazione dell'approccio POT con un'illusoria minore incertezza, ma soprattutto una pericolosa sottostima dei livelli di ritorno alti, come si evince facilmente dalla Tabella 8.9 e dalla Tabella 8.10.

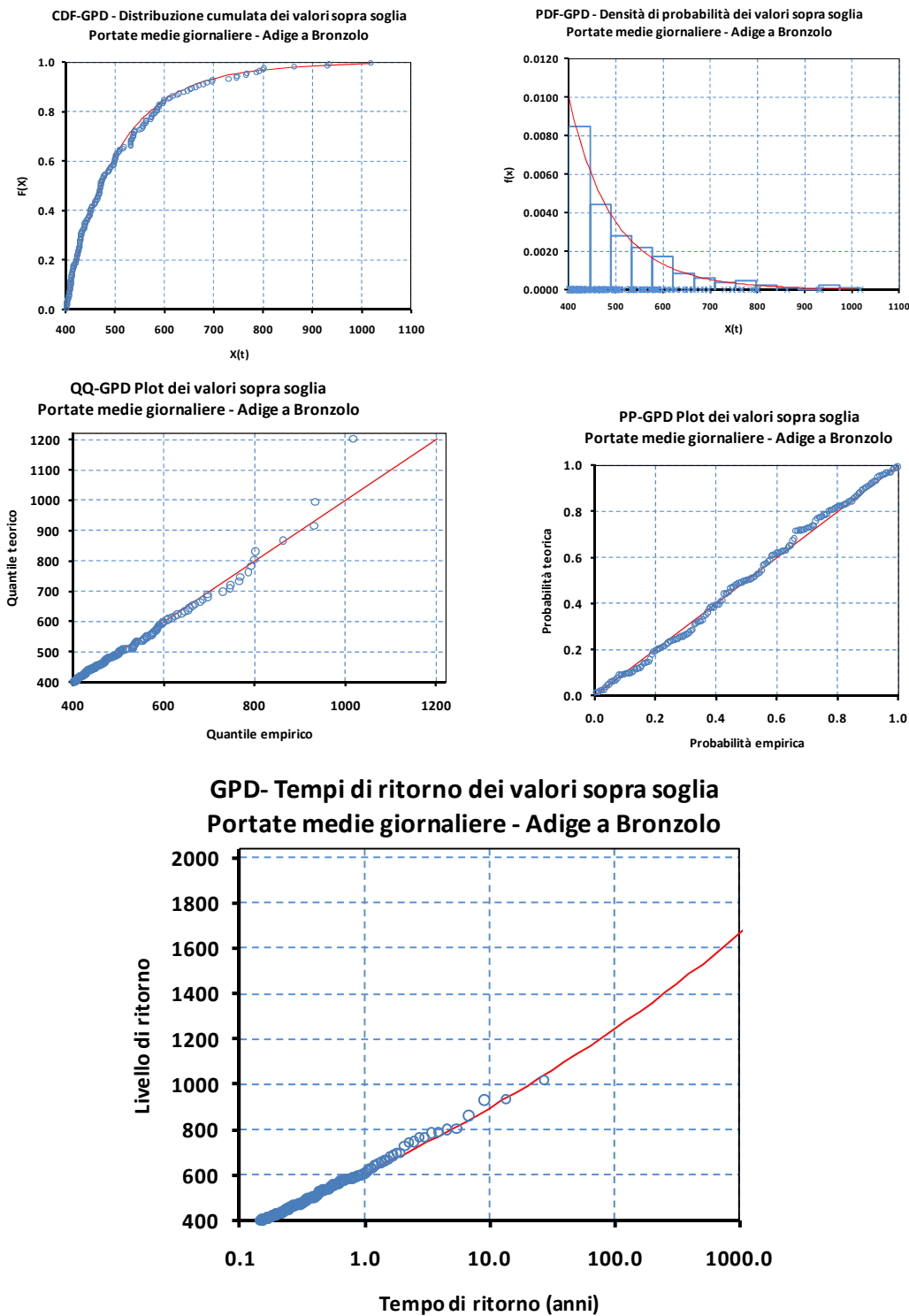


Figura 8.15 - Grafici diagnostici per la GPD. Serie POT senza *declustering*. Portate medie giornaliere - Adige a Bronzolo.

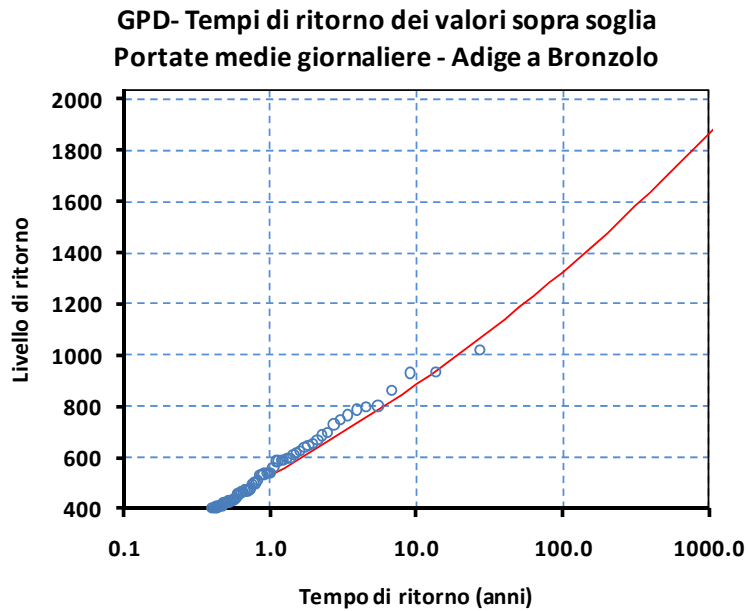
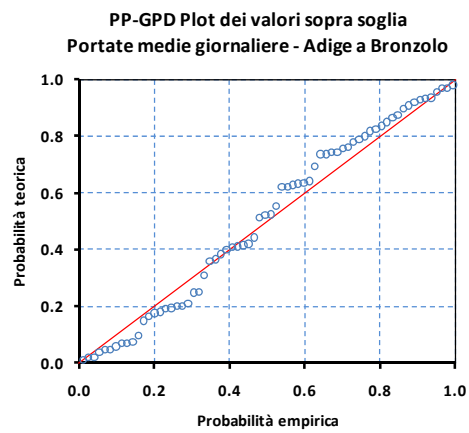
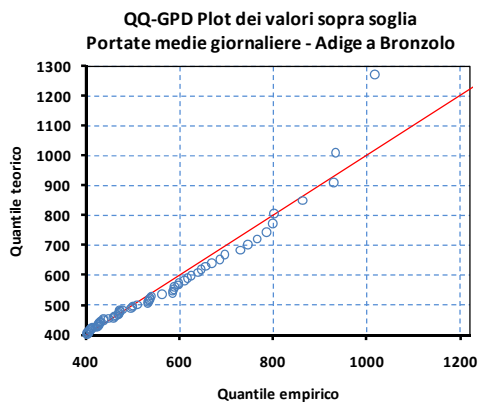
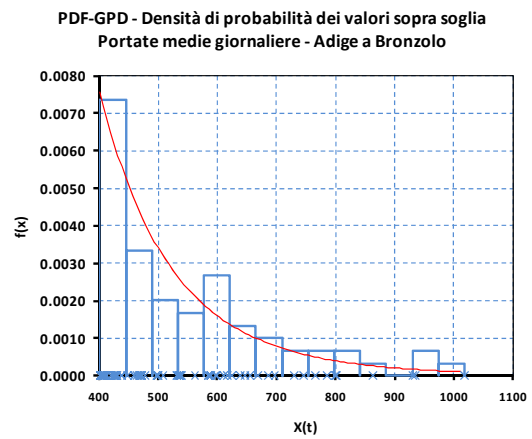
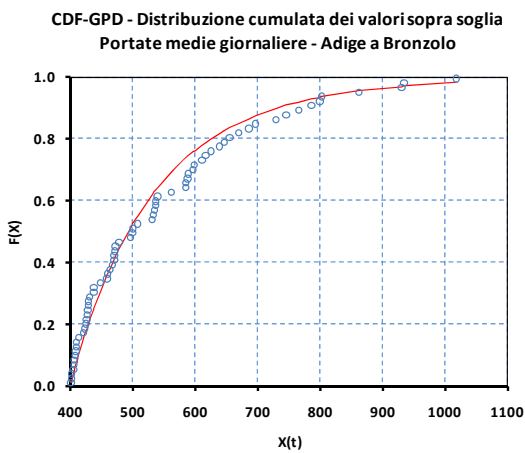


Figura 8.16 - Grafici diagnostici per la GPD. Serie POT con declustering. Portate medie giornaliere - Adige a Bronzolo.

Tabella 8.7 - Stima dei parametri della distribuzione generalizzata di Pareto con diversi metodi e relativi standard error. Serie POT senza declustering. Portate medie giornaliere - Adige a Bronzolo.

Parametri della distribuzione GPD		
	MoM	PWM
Posizione (μ)	400.0	400.0
(s.e.)	-	-
Scala (σ)	105.74	98.99
(s.e.)	(11.30)	(11.18)
Forma (ξ)	0.0403	0.0786
(s.e.)	(0.069)	(0.085)

Tabella 8.8 - Stima dei parametri della distribuzione generalizzata di Pareto con diversi metodi e relativi standard error. Serie POT con declustering. Portate medie giornaliere - Adige a Bronzolo.

Parametri della distribuzione GPD		
	MoM	PWM
Posizione (μ)	400.0	400.0
(s.e.)	-	-
Scala (σ)	151.37	130.70
(s.e.)	25.86	24.21
Forma (ξ)	-0.0086	0.0866
(s.e.)	0.123	0.140

Tabella 8.9 - Quantili corrispondenti a tempi di ritorno notevoli. Serie POT senza declustering Portate medie giornaliere - Adige a Bronzolo.

Quantili corrispondenti a tempi di ritorno notevoli									
T (anni)	10	20	30	50	100	200	300	500	1000
X_T	896.3	994.6	1054.6	1133.1	1244.6	1362.4	1434.4	1528.3	1662.0

Tabella 8.10 - Quantili corrispondenti a tempi di ritorno notevoli. Serie POT con declustering. Portate medie giornaliere - Adige a Bronzolo.

Quantili corrispondenti a tempi di ritorno notevoli*									
T (anni)	10	20	30	50	100	200	300	500	1000
X_T	886.5	1010.0	1085.7	1185.0	1326.9	1477.7	1570.1	1691.3	1864.6

*La stima dello s.e. del livello di ritorno X_T non è ancora implementata nel foglio ANABASI.

Tabella 8.11 – Sintesi dell'analisi degli estremi. Portate medie giornaliere - Adige a Bronzolo.

SCHEDA E: analisi degli estremi		
Approccio	POT con declustering	
Distribuzione	GPD	
Metodo stima parametri	PWM	
Parametri		
	valore	(s.e)
Posizione	400.0	-
Scala	130.70	24.21
Forma	0.0866	0.140
Livelli ritorno notevoli		
	valore	(s.e.)*
X _{T10}	886.50	-
X _{T20}	1010.0	-
X _{T30}	1085.7	-
X _{T50}	1185.0	-
X _{T100}	1326.9	-
X _{T200}	1477.7	-
X _{T300}	1570.1	-
X _{T500}	1691.3	-
X _{T1000}	1864.6	-

*La stima dello s.e. del livello di ritorno X_T non è ancora implementata nel foglio ANABASI.

8.2.2 Scala mensile

La serie delle portate medie mensili è ottenuta dalla media in ciascun mese dei valori delle portate medie giornaliere.

8.2.2.1 Metadato e caratteristiche

Anche per la serie delle portate mensili si procede, innanzitutto, con l'esame dei metadati (Tabella 8.12) e delle caratteristiche quantitative della serie (Tabella 8.13).

Tabella 8.12 - Metadato della serie storica. Portate medie mensili - Adige a Bronzolo.

SCHEDA A1: metadati tecnico-amministrativi	
metadato	valore
Titolo della serie	Portate medie mensili - Adige a Bronzolo
Grandezza idrologica	Portata media mensile
Unità misura	m ³ /s
Simbolo	Q _m
Tipo grandezza (primitiva/derivata)	derivata
Identificatore (codice) locale	855
Identificatore (codice) nazionale	855
Ente responsabile e fonte del dato	Ufficio Idrografico della Provincia Autonoma di Bolzano
Disponibilità (URL)	http://www.provincia.bz.it/meteo/stazioni-idrometriche.asp?stat_stid=133
Ultimo aggiornamento	

SCHEDA A2: metadati geografici	
metadato	valore
Nome stazione	Adige a Bronzolo
Comune	Bronzolo
Provincia	BZ
Regione	Trentino Alto Adige
Coordinate:	
Datum (ellisoide)	WGS84
Proiezione	GEO
Long/X (°/m)	11.31536
Lat/Y (°/m)	46.41377
Quota geoidica (m s.l.m.)	226.98
Link geolocalizzazione su web	http://maps.google.com/maps?f=q&source=s_q&hl=it&geocode=&sll=46.41377,11.31536&z=16&q=46.41377+11.31536
Bacino idrografico	Adige
Superficie bacino chiusura (km ²)	6926

SCHEDA A3: metadati modalità rilevamento	
metadato	valore
Numero periodi di campionamento omogeneo	1
Intervalli di campionamento omogeneo. Da a	01/01/1980 31/12/2006
Grandezza idrologica primitiva	h _g
Intervallo campionamento grandezza idrologica primitiva	1g
Funzione applicata alla grandezza idrologica primitiva	Trasformazione non lineare
Grandezza idrologica derivata	Q _g
Classe di accuratezza della grandezza derivata	C
Funzione Aggregazione/Selezione	media mensile
Percentuale massima dati mancanti nell'aggregazione/selezione	0
Standard di rilevamento	SIMN

Le serie derivate dall'aggregazione e/o dalla selezione di dati, come quella delle portate medie mensili derivata dalla serie delle portate giornaliere, sono soggette a un ulteriore fonte di peggioramento della

qualità del dato, dovute alla possibilità di non disporre dell'insieme completo dei dati su cui viene effettuata l'operazione.

Tabella 8.13 - Caratteristiche della serie. Portate medie mensili - Adige a Bronzolo.

SCHEDA B: descrizione statistica	
caratteristica	valore
Numero massimo di dati	324
Numero totale di dati	324
Frequenza (numero massimo dati/anno)	12
Numero di anni	27
Istante primo dato	01/01/1980 00.00
Istante ultimo dato	01/12/2006 00.00
Valore massimo	432.7
Valore minimo	51.8
Dati mancanti (e/o ricostruiti)	0
Intervalli di dati mancanti	0
Completezza	100.00%
Continuità	100.00%
iQuaSI	0.25

8.2.2.2 Descrizione statistica e grafici standard

Dal *time plot* (Figura 8.17) è evidente che anche la serie mensile presenta un pattern stagionale, che dovrebbe ritrovarsi nell'ACF. Analogamente si trova un'asimmetria positiva ma molto meno accentuata che nelle portate giornaliere, così come la curtosi. Pochi valori sono superiori al VAS e nessuno inferiore al VAI che coincide con il minimo. Non presenta valori che potrebbero essere considerati *outlier*.

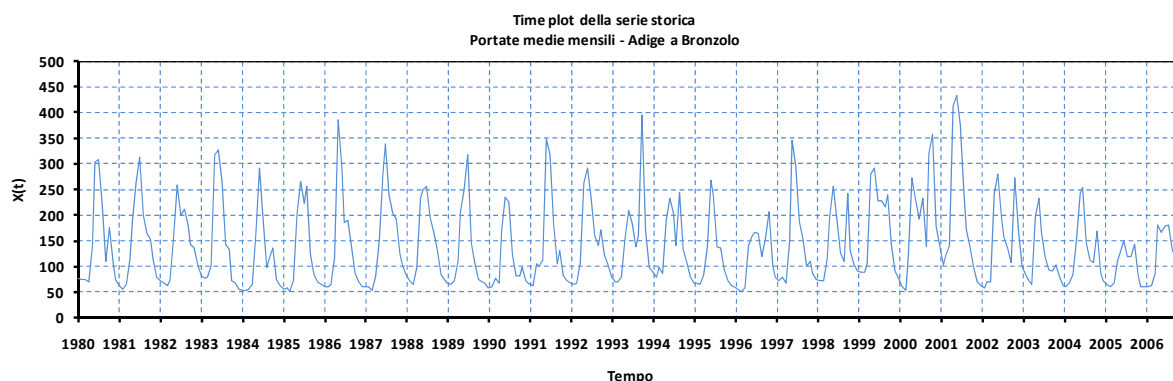


Figura 8.17 - Diagramma cronologico della serie (*time plot*). Portate medie mensili - Adige a Bronzolo.

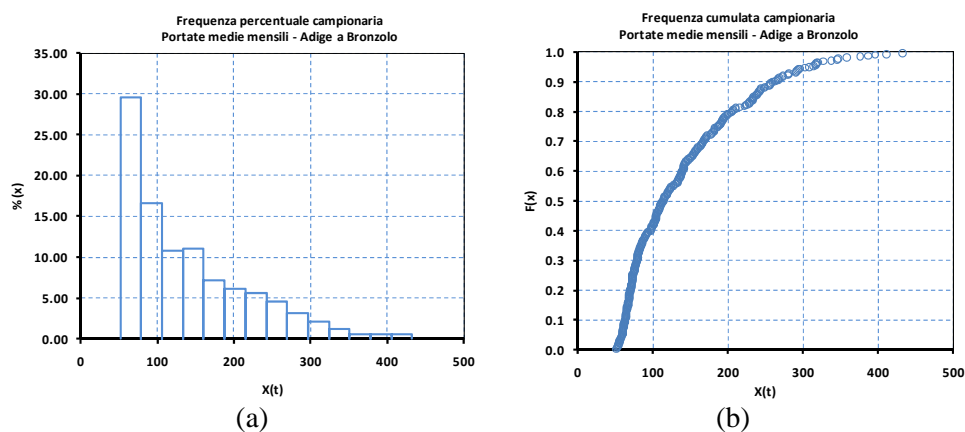


Figura 8.18 - Grafici della frequenza percentuale campionaria e della frequenza cumulata. Portate medie mensili - Adige a Bronzolo.

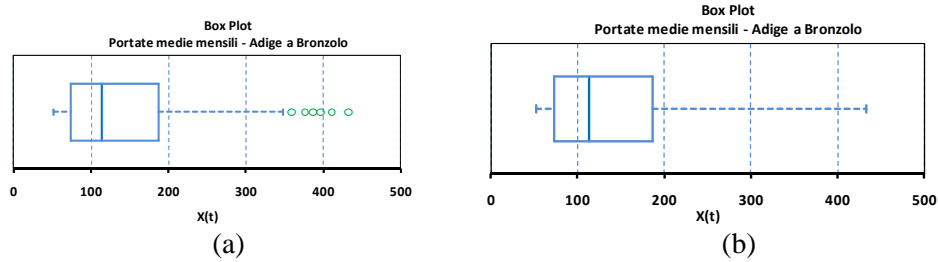


Figura 8.19 - Grafici box plot con e senza l'indicazione del VAI e VAS . Portate medie mensili - Adige a Bronzolo.

Tabella 8.14 - Statistiche univariate campionarie per la descrizione sintetica della serie. Portate medie mensili - Adige a Bronzolo.

SCHEDA C: statistiche base			
statistica	simbolo	valore	s.e.
Indici di posizione			
Media	m	140.10	
Moda	-	-	
Minimo (percentile 0%)	Min	51.79	
Percentile 25% (1° quartile Q ₁)	Q ₁	73.15	
Mediana (percentile 50%)	Q ₂	113.85	
Percentile 75% (3° quartile Q ₃)	Q ₃	186.54	
Massimo (percentile 100%)	Max	432.67	
Indici di dispersione			
Range Inter Quartile (Q ₃ -Q ₁)	IQR	113.39	
Range (Max-Min)	R	380.88	
Valore adiacente inferiore	VAI	51.79	
Valore adiacente superiore	VAS	347.22	
Scarto quadratico medio	s	80.5	
Varianza	s ²	6487.8	
Median Absolute Deviation	MAD	45.3	
Coefficiente di variazione	CV	0.6	
Indici di forma			
Asimmetria	g	1.15	
Curtosi	k	0.7	
Percentuale di dati su cui sono calcolate le statistiche		100.0%	

8.2.2.3 Analisi di stazionarietà

L'analisi dell'ACF conferma la stagionalità della serie con periodo di 12 mesi (Figura 8.20), come era previsto.

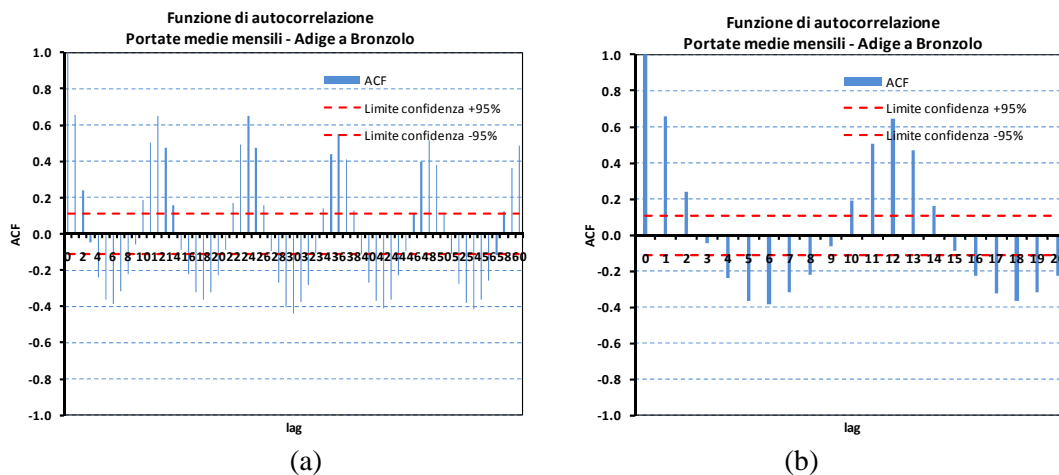


Figura 8.20 - Funzione di autocorrelazione per un periodo di 5 anni (a) e per i primi 20 mesi (b). Portate medie mensili - Adige a Bronzolo.

I test sulla significatività dell'autocorrelazione rigettano l'ipotesi nulla H_0 di assenza di autocorrelazione nei primi 10 lag (Tabella 8.15).

Tabella 8.15 - Esito dei test di autocorrelazione (test sui primi 10 lag). Portate medie mensili - Adige a Bronzolo.

Test	Statistica	P-Value	Livello di significatività	Esito
Ljung-Box	336.479	0.000	5%	IPOTESI H_0 RIGETTABILE
Box-Pierce	330.520	0.000	5%	IPOTESI H_0 RIGETTABILE

L'analisi delle componenti stagionali, effettuata con una finestra di *smoothing* dell'algoritmo del LOESS di 2 mesi, fornisce il medesimo risultato ottenuto per le portate giornaliere. La componente in media, infatti, presenta un massimo nel periodo giugno-luglio caratteristico del regime pluviale continentale, mentre la componente della deviazione standard presenta due massimi: uno nel mese di giugno e un altro nel mese di ottobre, praticamente uguali, caratteristico del regime sublitoraneo alpino.

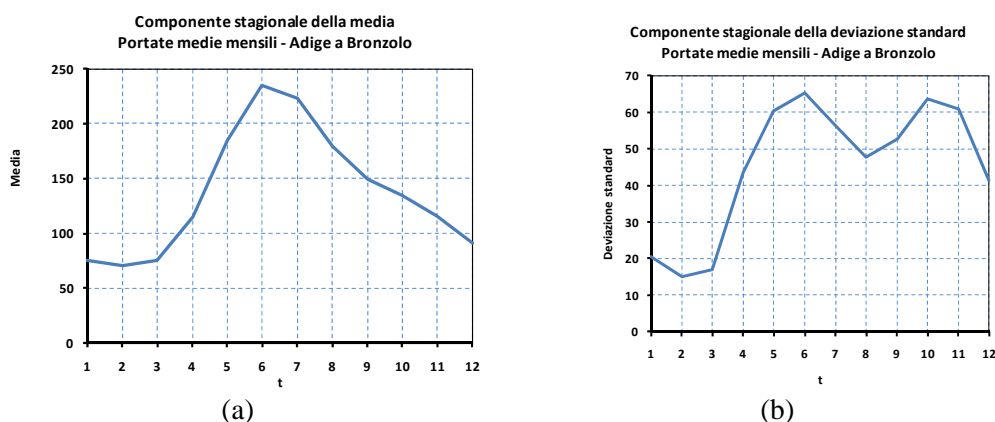


Figura 8.21 - Componente stagionale della media (a) e della deviazione standard (b). Portate medie mensili - Adige a Bronzolo.

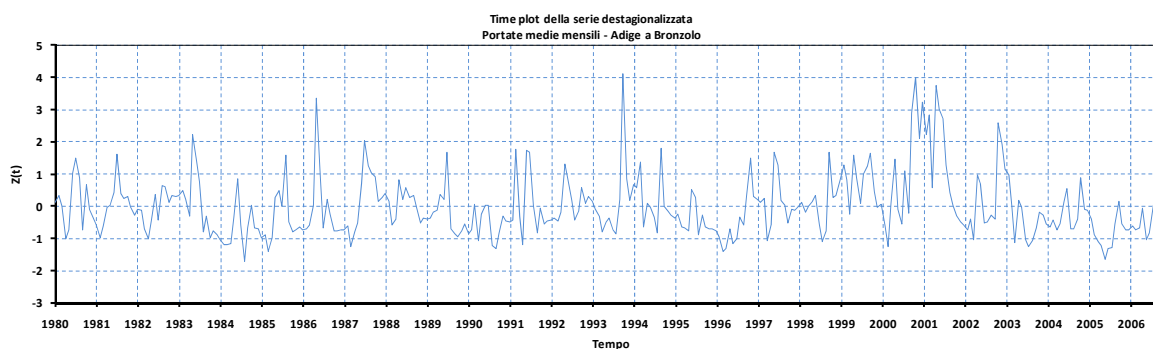


Figura 8.22 - Diagramma cronologico della serie destagionalizzata. Portate medie mensili - Adige a Bronzolo.

La serie destagionalizzata (Figura 8.22) non sembra, apparentemente, presentare una correlazione. In realtà l'ACF mette in evidenza una correlazione assolutamente non trascurabile nei primi lag e che, all'aumentare dei lag, non si attenua ma mostra un pattern ciclico. Ovviamente i test danno entrambi esito negativo (rigettano l'ipotesi nulla H_0 di non correlazione).

Per la serie delle portate medie mensili non si procede ulteriormente.

L'analisi di lunga memoria non viene effettuata poiché i dati mensili della serie destagionalizzata sono in numero non sufficientemente elevato per ottenere stime affidabili del parametro di Hurst con il metodo della varianza aggregata.

Per le medesime ragioni esposte per la serie giornaliera, non sono stati eseguiti i test per evidenziare i *trend* e i *change point* poiché, anche a questa scala di aggregazione, non forniscono risultati affidabili a causa della variabilità delle osservazioni, né sono stati eseguiti sulla corrispondente serie

destagionalizzata per la presenza, anche in questo caso, di un'autocorrelazione non stagionale e non trascurabile.

Più significativo per evidenziare *trend* o *change point* su una serie di portate mensili è effettuare le analisi sulle serie di ciascun mese dell'anno.

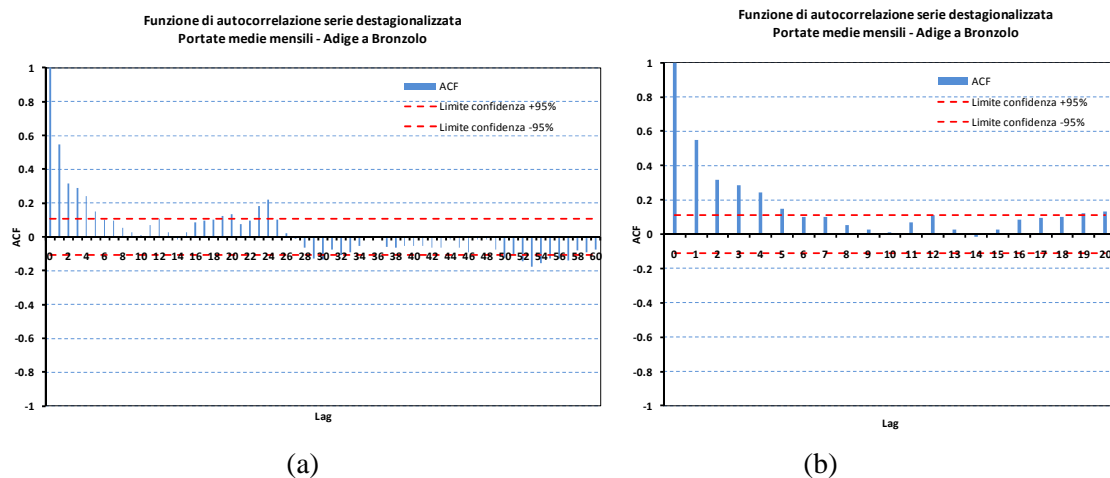


Figura 8.23 - Funzione di autocorrelazione della serie destagionalizzata su un periodo di 60 mesi (a), 20 mesi (b). Portate medie mensili - Adige a Bronzolo.

Tabella 8.16 - Esito dei test di autocorrelazione (test sui primi 10 lag) della serie destagionalizzata. Portate medie mensili - Adige a Bronzolo.

Test	Statistica	P-Value	Livello di significatività	Esito
Ljung-Box	193.2	0.000	5%	IPOTESI H ₀ RIGETTABILE
Box-Pierce	190.7	0.000	5%	IPOTESI H ₀ RIGETTABILE

Tabella 8.17 - Tabella di sintesi dell'analisi di stazionarietà. Portate medie mensili - Adige a Bronzolo.

SCHEDA D: analisi di stazionarietà	
analisi	livello
Autocorrelazione	significativa
Stagionalità	presente
Ciclicità	presente
Memoria a lungo termine	non effettuata
Trend	non effettuato
Change point	non effettuato
Normalità	non effettuata
Non-stazionarietà in media	significativa
Non-stazionarietà in varianza	significativa

Non viene, infine, eseguita l'analisi degli estremi mediante l'approccio POT, poiché dagli strumenti diagnostici del *mean residual life* e dalla stabilità dei parametri della GPD, non sembra riconoscersi valori di soglia coerenti con il modello della distribuzione di probabilità GPD.

8.2.3 Scala annuale

Alla scala annuale si procede all'analisi dei massimi delle portate medie giornaliere.

La serie dei massimi annuali delle portate medie giornaliere viene estratta direttamente dalla serie completa dei dati giornalieri.

8.2.3.1 Metadato e caratteristiche

Tabella 8.18 - Metadati della serie storica. Portate giornaliere massime annuali - Adige a Bronzolo

SCHEDA A1: metadati tecnico-amministrativi	
metadato	valore
Titolo della serie	Portate giornaliere massime annuali - Adige a Bronzolo
Grandezza idrologica	Portata giornaliera massima annuale
Unità misura	m ³ /s
Simbolo	Q _{g,max}
Tipo grandezza (primitiva/derivata)	derivata
Identificatore (codice) locale	855
Identificatore (codice) nazionale	855
Ente responsabile e fonte del dato	Ufficio Idrografico della Provincia Autonoma di Bolzano
Disponibilità (URL)	http://www.provincia.bz.it/meteo/stazioni-idrometriche.asp?stat_stid=133
Ultimo aggiornamento	

SCHEDA A2: metadati geografici	
metadato	valore
Nome stazione	Adige a Bronzolo
Comune	Bronzolo
Provincia	BZ
Regione	Trentino Alto Adige
Coordinate:	
Datum (ellissoide)	WGS84
Proiezione	GEO
Long/X (°/m)	11.31536
Lat/Y (°/m)	46.41377
Quota geoidica (m s.l.m.)	226.98
Link geolocalizzazione su web	http://maps.google.com/maps?f=q&source=s_q&hl=it&geocode=&sll=46.41377,11.31536&z=16&q=46.41377+11.31536
Bacino idrografico	Adige
Superficie bacino chiusura (km ²)	6926

SCHEDA A3: metadati modalità rilevamento	
metadato	valore
Numero periodi di campionamento omogeneo	1
Intervallo di campionamento omogeneo. Da a	01/01/1980 31/12/2006
Grandezza idrologica primitiva	h _g
Intervallo campionamento idrologica primitiva	grandezza 1g
Funzione applicata alla grandezza idrologica primitiva	Trasformazione non lineare
Grandezza idrologica derivata	Q _g
Classe di accuratezza della grandezza derivata	C
Funzione Aggregazione/Selezione	massimo annuale
Percentuale massima dati mancanti nell'aggregazione/selezione	0
Standard di rilevamento	SIMN

Tabella 8.19 - Caratteristiche della serie. Portate giornaliere massime annuali - Adige a Bronzolo

SCHEDA B: descrizione statistica	
caratteristica	valore
Numero massimo di dati	27
Numero totale di dati	27
Frequenza (numero massimo dati/anno)	1
Numero di anni	27
Istante primo dato	01/01/1980 00.00
Istante ultimo dato	01/01/2006 00.00
Valore massimo	1018.0
Valore minimo	278.4
Dati mancanti (e/o ricostruiti)	0
Intervalli di dati mancanti	0
Completezza	100.00%
Continuità	100.00%
iQuaSI	0.25

8.2.3.2 Descrizione statistica e grafici standard

Dall'analisi degli elementi descrittivi, grafici e numerici (Figura 8.24, Figura 8.25, Tabella 8.20), si rileva che la serie dei dati dei massimi annuali delle portate giornaliere presenta una distribuzione pressoché simmetrica senza particolari valori estremi o anomali che possono considerarsi *outlier*.

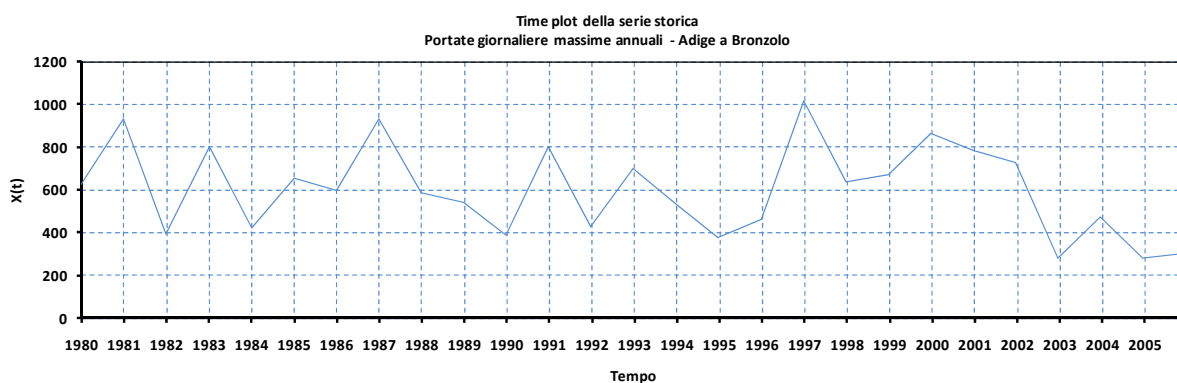


Figura 8.24 - Diagramma cronologico della serie (time plot). Portate giornaliere massime annuali - Adige a Bronzolo.

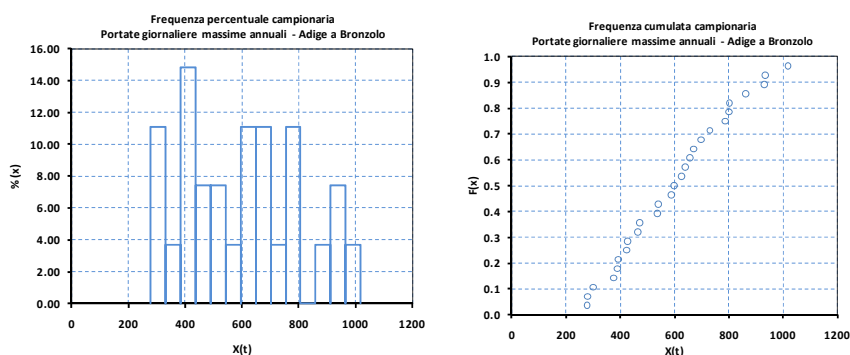


Figura 8.25 - Grafici della frequenza percentuale campionaria e della frequenza cumulata. Portate giornaliere massime annuali - Adige a Bronzolo.

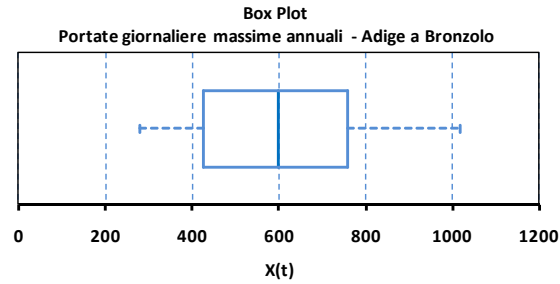


Figura 8.26 - Box plot in cui il VAI e VAS coincidono rispettivamente con il minimo e massimo. Portate giornaliere massime annuali - Adige a Bronzolo.

Tabella 8.20 - Statistiche univariate campionarie per la descrizione sintetica della serie. Portate giornaliere massime annuali - Adige a Bronzolo

SCHEDA C: statistiche base			
statistica	simbolo	valore	s.e.
Indici di posizione			
Media	m	600.66	
Moda		0.0	
Minimo (percentile 0%)	Min	278.40	
Percentile 25% (1° quartile Q_1)	Q_1	425.65	
Mediana (percentile 50%)		597.30	
Percentile 75% (3° quartile Q_3)	Q_3	758.15	
Massimo (percentile 100%)	Max	1018.00	
Indici di dispersione			
Range Inter Quartile (Q_3-Q_1)	IQR	332.50	
Range (Max-Min)	R	739.60	
Valore adiacente inferiore	VAI	278.40	
Valore adiacente superiore	VAS	1018.00	
Scarto quadratico medio	s	206.6	
Varianza	s^2	42693.1	
Median Absolute Deviation	MAD	173.8	
Coefficiente di variazione	CV	0.3	
Indici di forma			
Asimmetria	g	0.21	
Curtosi	k	-0.9	
Percentuale di dati su cui sono calcolate le statistiche		100.0%	

8.2.3.3 Analisi di normalità

La simmetria, già messa in evidenza da un semplice esame visivo del box plot e della distribuzione percentuale campionaria, è associata al fatto che i dati possono essere significativamente distribuiti con legge normale, come emerge dal test di Jarque-Bera (Tabella 8.21e Tabella 8.22) che fornisce un esito positivo (ipotesi nulla H_0 non rigettabile) anche per i dati non trasformati secondo la trasformata di Box e Cox.

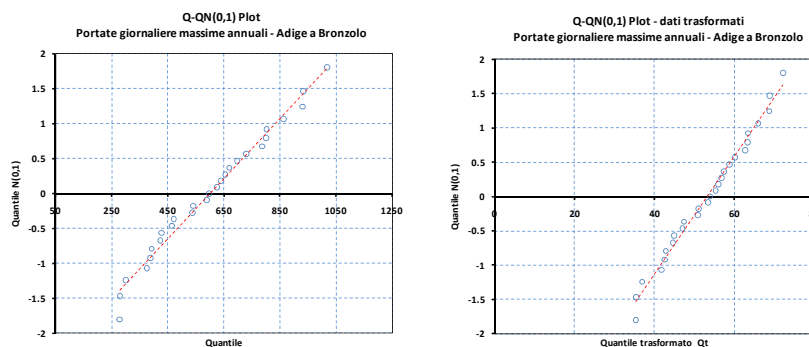


Figura 8.27 - QQ plot Normali per la verifica della normalità dei dati e della loro trasformata Box e Cox. Portate giornaliere massime annuali - Adige a Bronzolo.

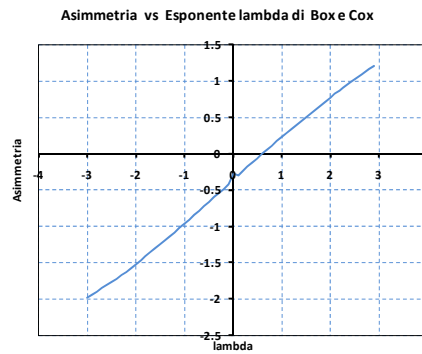


Figura 8.28 - Diagramma del coefficiente della trasformata di Box e Cox in funzione dell'asimmetria dei dati. Portate giornaliere massime annuali - Adige a Bronzolo.

Tabella 8.21 - Esito del test di normalità dei dati. Portate giornaliere massime annuali - Adige a Bronzolo.

Test	Statistica	P-Value	Livello di significatività	Esito
Jarque-Bera	1.1537	0.56167	5%	IPOTESI H ₀ NON RIGETTABILE

Tabella 8.22 - Esito del test di normalità dei dati trasformati con la trasformata di Box e Cox con $\lambda = 0.53$. Portate giornaliere massime annuali - Adige a Bronzolo.

Test	Statistica	P-Value	Livello di significatività	Esito
Jarque-Bera	1.0717	0.58517	5%	IPOTESI H ₀ NON RIGETTABILE

8.2.3.4 Analisi di stazionarietà

L'analisi, attraverso l'ACF (Figura 8.29e

Tabella 8.23), i test per il *change point detection* (Figura 8.30 e Tabella 8.24) e i test per il *trend detection* (Figura 8.34e Tabella 8.25) evidenziano che la serie dei massimi annuali è significativamente stazionaria.

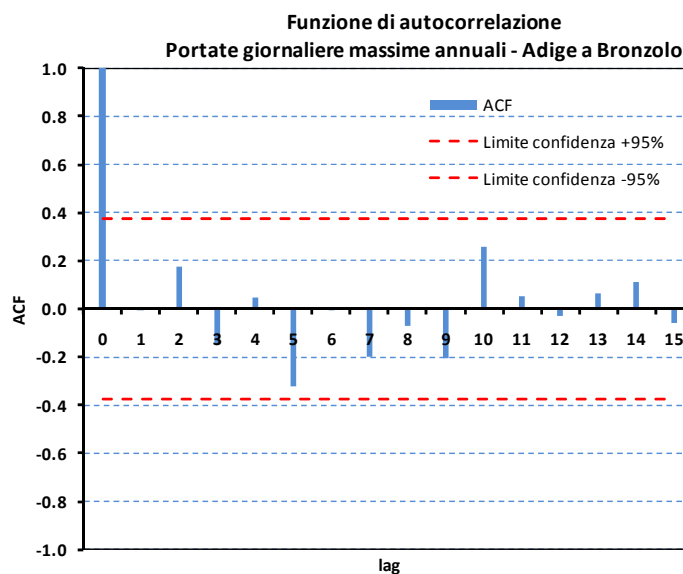


Figura 8.29 - Funzione di autocorrelazione della serie 15 lag. Portate giornaliere massime annuali - Adige a Bronzolo.

Tabella 8.23 - Esito dei test di autocorrelazione (test sui primi 5 lag). Portate giornaliere massime annuali - Adige a Bronzolo.

Test	Statistica	P-Value	Livello di significatività	Esito
Ljung-Box	12.127	0.277	5%	IPOTESI H_0 NON RIGETTABILE
Box-Pierce	8.438	0.586	5%	IPOTESI H_0 NON RIGETTABILE

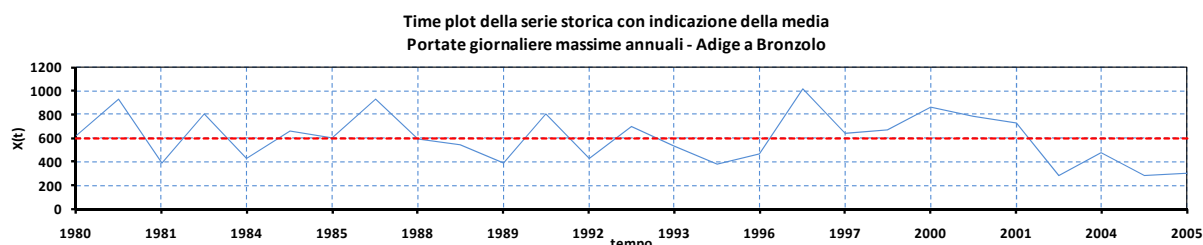


Figura 8.30 - Time plot con indicazione della media. Portate giornaliere massime annuali - Adige a Bronzolo.

Tabella 8.24 - Esito dei test per il “change point detection”. Portate giornaliere massime annuali - Adige a Bronzolo.

Test	Statistica	P-Value	Livello di significatività	Esito
CUSUM	1101.6	0.449	5%	IPOTESI H_0 NON RIGETTABILE
Pettitt	80.0	0.152	5%	IPOTESI H_0 NON RIGETTABILE

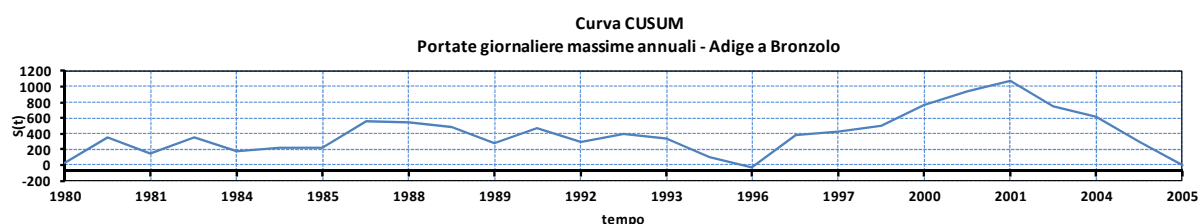


Figura 8.31 - Curva CUSUM con bootstrap (1000 campioni). Portate giornaliere massime annuali - Adige a Bronzolo.

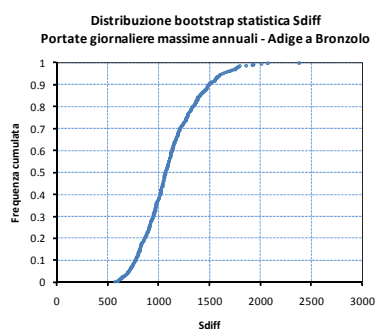


Figura 8.32 - Distribuzione della statistica S_{diff} nel test CUSUM con bootstrap (1000 campioni). Portate giornaliere massime annuali - Adige a Bronzolo.

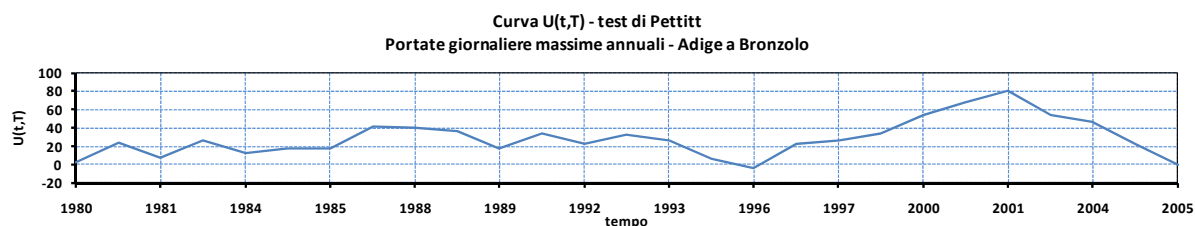


Figura 8.33 - Curva della statistica di Pettitt. Portate giornaliere massime annuali - Adige a Bronzolo.

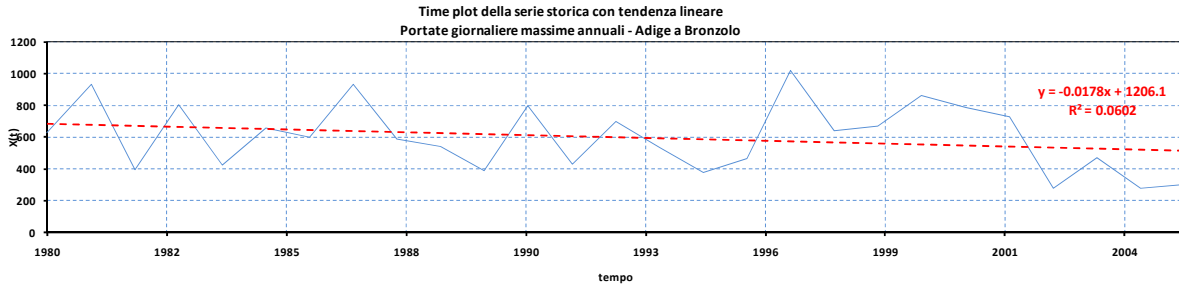


Figura 8.34 - Time plot con indicazione della tendenza lineare. Portate giornaliere massime annuali - Adige a Bronzolo.

Tabella 8.25 - Esito dei test per il “trend detection”. Portate giornaliere massime annuali - Adige a Bronzolo.

Test	Statistica	Tau	P-Value	Livello di significatività	Esito
Mann-Kendall	-1.1674	-0.162	0.243	5%	IPOTESI H ₀ NON RIGETTABILE
Test	Statistica	Rho	P-Value	Livello di significatività	Esito
Pearson	-1.7569	-0.332	0.091	5%	IPOTESI H ₀ NON RIGETTABILE
Spearman	-1.1923	-0.234	0.233	5%	IPOTESI H ₀ NON RIGETTABILE

Tabella 8.26 - Tabella di sintesi dell’analisi di stazionarietà. Portate giornaliere massime annuali - Adige a Bronzolo.

SCHEDA D: analisi di stazionarietà	
analisi	livello
Autocorrelazione	non significativa
Stagionalità	non presente
Ciclicità	non presente
Memoria a lungo termine	non effettuata
Trend	non significativo
Change point	non significativo
Normalità	significativa
Non-stazionarietà in media	non presente
Non-stazionarietà in varianza	non presente

8.2.3.5 Analisi degli estremi

A fronte del fatto che il rispetto dell’ipotesi d’indipendenza è generalmente garantito quando si considerino i massimi annuali, è evidente che la numerosità della serie è relativamente bassa per ottenere estrapolazioni verso probabilità di non superamento corrispondenti a tempi di ritorno superiori al numero di anni disponibile.

Nel caso in esame si nota che il valore del parametro di forma è negativo, indicando che la distribuzione ha un limite superiore. La presenza di un limite superiore può condurre alla sottostima dei valori con elevato tempo di ritorno.

Quando il parametro di forma può essere assunto uguale a zero in ragione dell’incertezza della stima, la distribuzione GEV converge alla distribuzione dei valori estremi di Gumbel. È dunque opportuno ripetere la stima per il distribuzione a due parametri di Gumbel per valutare se un modello più semplice (a due parametri) è sufficiente a descrivere i dati.

La maggiore incertezza nella distribuzione GEV è data principalmente dalla presenza del parametro di forma, la cui stima è molto sensibile alla bassa numerosità del campione.

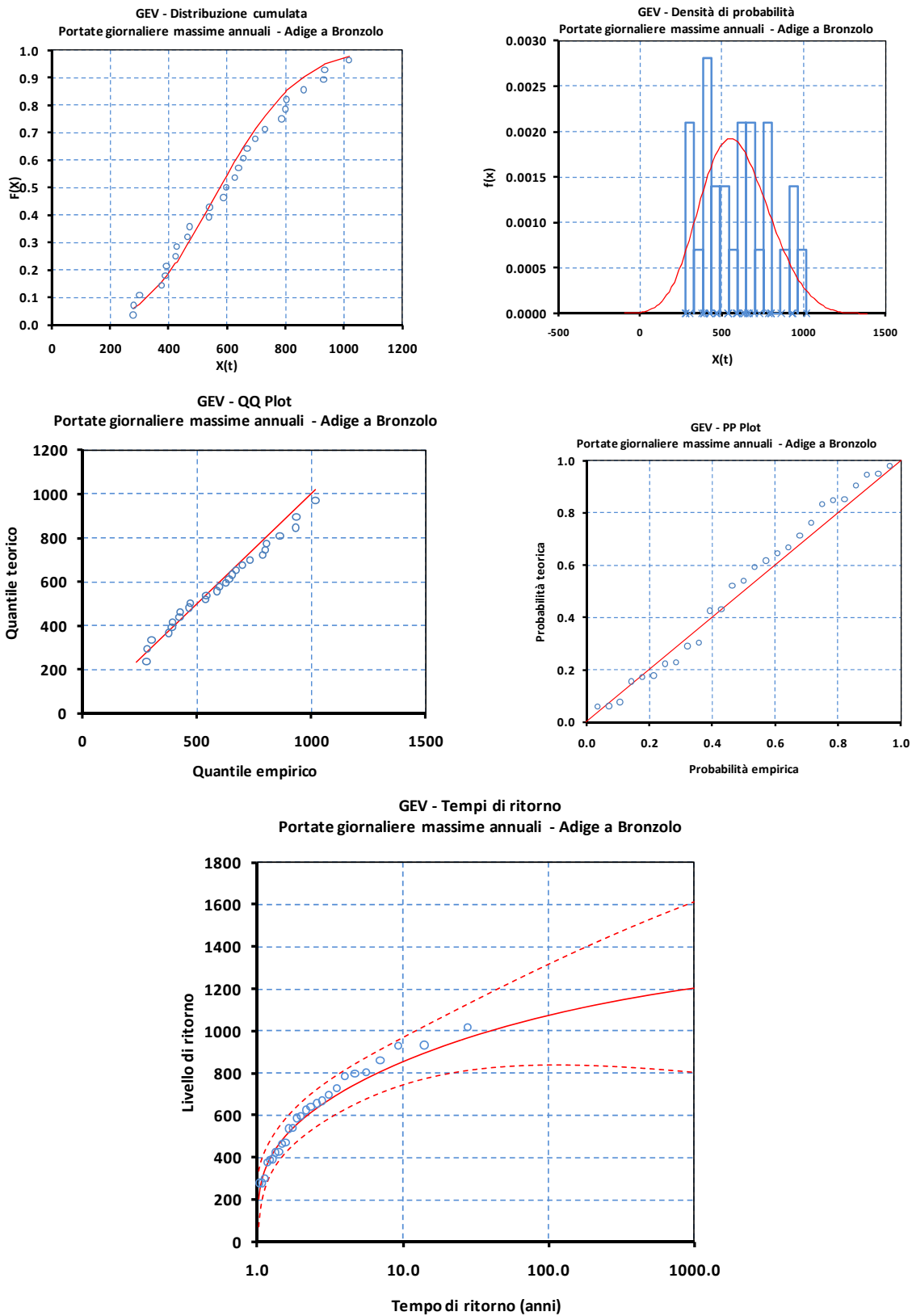


Figura 8.35 - Grafici diagnostici per la GEV. Portate giornaliere massime annuali - Adige a Bronzolo.

Tabella 8.27 - Stima dei parametri della distribuzione GEV con il MoM e il PWM. Portate giornaliere massime annuali - Adige a Bronzolo.

Parametri della distribuzione GEV*		
	MoM	PWM
Posizione (μ)	507.7	506.9
(s.e.)		
Scala (σ)	161.17	196.38
(s.e.)		
Forma (ξ)	0.000	-0.219
(s.e.)		

*La stima dello s.e. dei parametri GEV non è ancora implementata nel foglio ANABASI.

Tabella 8.28 - Quantili corrispondenti a tempi di ritorno notevoli elaborati con la distribuzione GEV con parametri PWM. Portate giornaliere massime annuali - Adige a Bronzolo.

Quantili corrispondenti a tempi di ritorno notevoli									
T (anni)	10	20	30	50	100	200	300	500	1000
X_T	855.9	935.8	976.4	1022.3	1076.4	1122.8	1146.7	1174.0	1206.5
s.e.	57.60	71.64	82.43	98.06	121.68	146.80	161.76	180.62	205.81

Tabella 8.29 - Tabella di sintesi dell'analisi degli estremi. Portate giornaliere massime annuali - Adige a Bronzolo.

SCHEDA E: analisi degli estremi		
Approccio	AM	
Distribuzione	GEV	
Metodo stima parametri	PWM	
Parametri		
	valore	(s.e.)*
Posizione	506.9	
Scala	196.38	
Forma	-0.219	
Livelli ritorno notevoli		
	valore	(s.e.)
XT10	855.9	57.60
XT20	935.8	71.64
XT30	976.4	82.43
XT50	1022.3	98.06
XT100	1076.4	121.68
XT200	1122.8	146.80
XT300	1146.7	161.76
XT500	1174	180.62
XT1000	1206.5	205.81

*La stima dello s.e. dei parametri GEV non è ancora implementata nel foglio ANABASI.

9. Bibliografia

- Allen R.G., Pereira L.S., Raes D., Smith M., 1998, *Crop evapotranspiration: Guidelines for computing crop water requirements*. Irr. & Drain. Paper 56. UN-FAO, Rome, Italy.
- CNR-GNDCL, 1993, *Manuale di riferimento per la misura al suolo delle grandezze idrometeorologiche*. a cura di Virgilio Anselmo, UO 1.39 Università di Torino.
- Cleveland W.S., 1993, *Visualizing Data*. Hobart Press, New Jersey.
- Coles S., 2001, *An Introduction to Statistical Modeling of Extreme Values*. Springer, London.
- Deidda R., Puliga M., 2009, *Performances of some parameter estimators of the Generalized Pareto Distribution over rounded-off samples*. Physics and Chemistry of the Earth, 34, 626–634.
http://unica2.unica.it/rdeidda/papers/2009_PCE_Deidda_Puliga.pdf
- Grimaldi S., 2004, *Linear parametric models applied to daily hydrological series*. Journal of Hydrologic Engineering, 9(5), 383–391.
- Hosking J.R.M., Wallis J.R., Wood E., 1985, *Estimation of the Generalized Extreme Value distribution by the method of the Probability-Weighted moments*. Technometrics, 27, 251-261.
- Hosking J.R.M., Wallis J.R., 1987, *Parameter and Quantile Estimation for the Generalized Pareto Distribution*. Technometrics, 29, 339-349.
- Kottegoda N.T., Rosso R., 1997, *Statistics, probability, and reliability for civil and environmental engineers*. McGraw Hill, New York.
- Kundzewicz Z.W., Robson A., 2004, *Change detection in hydrological records a review of the methodology*. Hydrological Sciences Journal, 49(1).
- Little R.J.A., Rubin D.B., 2002, *Statistical Analysis with Missing Data*. Second Edition, John Wiley & Sons, Hoboken, New Jersey.
- Maidment D.R., (Editor in Chief), 1993, *Handbook of Hydrology*. McGraw-Hill, New York.
- Priestley C.H.B., Taylor R.J., 1972, *On the assessment of surface heat flux and evaporation using large-scale parameters*. Mon. Weather Rev., 100,81-82.
- SIMN, 1998 a, “Norme tecniche per la raccolta e l’elaborazione dei dati idrometeorologici – parte I”.
- SIMN, 1998 b, “Norme tecniche per la raccolta e l’elaborazione dei dati idrometeorologici – parte II”.
- Tukey J.W., 1997, *Exploratory data analysis*. Addison-Wesley, Reading, MA.
- Willems P., 2003, *WETSPRO: Water Engineering Time Series PROCESSING tool tool, Reference Manual and User’s Manual*. Hydraulics Laboratory K.U.Leuven, Leuven, Belgium.
- WMO, 1988, *Analyzing long time series of hydrological data with respect to climate variability. Project description*. WCAP-3, WMOTFD-No. 224, Geneva, Switzerland.
<http://water.usgs.gov/osw/wcp-water/WCAP-3.pdf>
- WMO, 1994, *Guide to hydrological practices: Data acquisition and processing, analysis, forecasting and other applications*. Fifth edition, 1994, WMO No.168, Geneva Switzerland.
http://hydrologie.org/BIB/OMM/WMOENG_v5.pdf
- WMO, 2000, *Detecting Trend and Other Changes in Hydrological Data*. World Climate Programme—Water, World Climate Programme Data and Monitoring, WCDMP-45, WMO/TD no. 1013 Kundzewicz, Z. W. & Robson, A. (eds), WMO, Geneva, Switzerland.
<http://water.usgs.gov/osw/wcp-water/detecting-trend.pdf>
- WMO, 2003a, *Guidelines on climate metadata and homogenization*. Llanso, P., Ed., WMO Tech. Doc. 1186, 51 pp, WMO, Geneva, Switzerland.
http://www.wmo.int/pages/prog/wcp/wcdmp/wcdmp_series/documents/WCDMP-53.pdf
- WMO, 2003b, *Hydrological Data Management: Present State and Trends*. Operational Hydrology Report No. 48, 2003, WMO-No. 964, WMO, Geneva, Switzerland.
http://www.whycos.org/IMG/pdf/964_E.pdf
- WMO, 2008, *Guide to Meteorological Instruments and Methods of Observation*. Seventh edition 2008, WMO, Geneva, Switzerland.
http://www.wmo.int/pages/prog/gcos/documents/gruanmanuals/CIMO/CIMO_Guide-7th_Edition-2008.pdf
- WMO, 2009, *Guidelines on Analysis of extremes in a changing climate in support of informed decisions for adaptation*. WCDMP-No. 72, WMO, Geneva, Switzerland.
http://www.wmo.int/pages/prog/wcp/wcdmp/wcdmp_series/documents/WCDMP_72_TD_1_500_en_1.pdf

10. Appendice A

10.1 Scheda A: metadati per le serie idrologiche

SCHEDA A1: metadati tecnico-amministrativi	
metadato	valore
Titolo della serie	
Grandezza idrologica	
Unità misura	
Simbolo	
Tipo grandezza (primitiva/derivata)	
Identificatore (codice) locale	
Identificatore (codice) nazionale	
Ente responsabile e fonte del dato	
Disponibilità (URL)	
Ultimo aggiornamento	

SCHEDA A2: metadati geografici	
metadato	valore
Nome stazione	
Comune	
Provincia	
Regione	
Coordinate:	
	Datum (ellissoide)
	Proiezione
	Long/X (°/m)
	Lat/Y (°/m)
Quota geoidica (m s.l.m.)	
Link geolocalizzazione su web	
Bacino idrografico	
Superficie bacino chiusura (km ²)	

SCHEDA A3: metadati modalità rilevamento	
metadato	valore
Numero periodi di campionamento omogeneo	
Intervalli di campionamento omogeneo. Da a	
Grandezza idrologica primitiva	
Intervallo campionamento grandezza idrologica primitiva	
Funzione applicata alla grandezza idrologica primitiva	
Grandezza idrologica derivata	
Classe di accuratezza della grandezza derivata	
Funzione Aggregazione/Selezione	
Percentuale massima dati mancanti nell'aggregazione/selezione	
Standard di rilevamento	

10.2 Scheda B: descrizione statistica

SCHEDA B: descrizione statistica	
caratteristica	valore
Numero massimo di dati	
Numero totale di dati	
Frequenza (numero massimo dati/anno)	
Numero di anni	
Istante primo dato	
Istante ultimo dato	
Valore massimo	
Valore minimo	
Dati mancanti (e/o ricostruiti)	
Intervalli di dati mancanti	
Completezza	
Continuità	
iQuaSI	

10.3 Scheda C: statistiche di base

SCHEDA C: statistiche base			
statistica	simbolo	valore	s.e.
Indici di posizione			
Media	m		
Moda			
Minimo (percentile 0%)	Min		
Percentile 25% (1° quartile Q ₁)	Q ₁		
Mediana (percentile 50%)			
Percentile 75% (3° quartile Q ₃)	Q ₃		
Massimo (percentile 100%)	Max		
Indici di dispersione			
Range Inter Quartile (Q ₃ -Q ₁)	IQR		
Range (Max-Min)	R		
Valore adiacente inferiore	VAI		
Valore adiacente superiore	VAS		
Scarto quadratico medio	s		
Varianza	s ²		
Median Absolute Deviation	MAD		
Coefficiente di variazione	CV		
Indici di forma			
Asimmetria	g		
Curtosi	k		

10.4 Scheda D: analisi di stazionarietà

SCHEDA D: analisi di stazionarietà	
analisi	livello
Autocorrelazione	
Stagionalità	
Ciclicità	
Memoria a lungo termine	
Trend	
Change point	
Normalità	
Non-stazionarietà in media	
Non-stazionarietà in varianza	

10.5 Scheda E: analisi degli estremi

SCHEDA E: analisi degli estremi	
Approccio	
Distribuzione	
Metodo stima parametri	
Parametri	
	valore (s.e)
Posizione	
Scala	
Forma	
Livelli ritorno notevoli	
	valore (s.e)
X_{T10}	
X_{T20}	
X_{T30}	
X_{T50}	
X_{T100}	
X_{T200}	
X_{T300}	
X_{T500}	
X_{T1000}	

11. Appendice B. Approfondimenti di statistica

In questa sezione sono riportate alcune definizioni e concetti di probabilità e statistica utili per l'approfondimento di quanto precedentemente esposto.

Benchè costituisca un approfondimento, questa appendice non ha lo scopo di fornire una descrizione esaustiva per la quale si rimanda ai riferimenti bibliografici suggeriti di volta in volta.

11.1 Definizioni e concetti di base di probabilità e statistica

11.1.1 Elementi di probabilità

11.1.1.1 Definizione di evento e principali relazioni tra eventi

Si definisce evento l'esito di un esperimento concettuale o di una fenomeno che presenta un connotato di casualità. L'insieme di tutti i possibili eventi A , ossia di tutti i possibili esiti, costituisce lo spazio campionario Ω . Il complemento A^c di un evento A consiste in tutti i possibili esiti di Ω che non sono inclusi in A .

Due eventi A_1 e A_2 sono mutuamente esclusivi se il verificarsi dell'uno esclude l'altro, in simboli $A_1 \cap A_2 = A_1 A_2 = \emptyset$, in cui " \cap " è l'operatore intersezione e " \emptyset " indica l'insieme vuoto. L'unione di due eventi, indicata con $A_1 \cup A_2$ ovvero $A_1 + A_2$, rappresenta il loro accadimento congiunto.

Dato lo spazio campionario Ω , una famiglia di eventi $\mathbf{A} = \{A_i \subseteq \Omega, i = 1, 2, \dots\}$ è detta σ -algebra o spazio degli eventi se contiene Ω ed è chiusa rispetto alle operazioni di unione numerabile e complementazione, ossia se valgono le seguenti tre proprietà:

- 1) $\Omega \in \mathbf{A}$ (ossia, \mathbf{A} ricomprende tra i suoi elementi lo spazio campionario Ω);
- 2) se $A \in \mathbf{A}$, allora $A^c \in \mathbf{A}$ (lo spazio degli eventi contiene ogni insieme A e il suo complemento; dalla 1) segue che $\emptyset \in \mathbf{A}$);
- 3) se gli eventi $A_i \in \mathbf{A}, i = 1, 2, \dots$, allora la loro unione $\cup A_i \in \mathbf{A}$.

11.1.1.2 Definizione di probabilità

Il concetto di probabilità può essere introdotto in modi diversi. Un approccio classico è legato allo studio di esperimenti, quali il lancio di monete o dadi e l'estrazione di una carta da un mazzo. Se il risultato dell'esperimento (e.g., il lancio di un dado) prevede n esiti mutuamente esclusivi e ugualmente possibili (i 6 numeri corrispondenti alle 6 faccine), e n_A è il numero di esiti che possiede l'attributo A (e.g., il numero di esiti associati all'uscita di un particolare numero tra i sei possibili è uguale a 1), allora la probabilità che si verifichi l'evento A (e.g., esca il 6) è pari alla frazione n_A/n (e.g., 1/6). Questa definizione di probabilità è detta *a priori* poiché può essere calcolata tramite ragionamenti deduttivi senza bisogno di eseguire un esperimento reale.

Una seconda definizione di probabilità nasce dalla ripetizione di un esperimento, in condizioni costanti, e dalla misurazione di una particolare proprietà durante ogni prova (e.g., la misura della tensione di rottura a trazione di un certo numero di provini). In molti casi i valori misurati cadono all'interno di alcune classi di valori, in cui le frequenze relative sono stabili. La probabilità che il valore osservato cada all'interno di un particolare intervallo è quindi approssimata dalla frequenza osservata con cui i valori risultanti da un numero elevato di esperimenti cadono all'interno dell'intervallo considerato. Questa definizione è detta *frequenziale* o *a posteriori*, in quanto legata alla possibilità di eseguire più volte uno stesso esperimento in condizioni controllate.

In molti casi, come per molti fenomeni naturali, non è possibile ripetere l'osservazione in condizioni analoghe (e.g., la temperatura media di un determinato giorno può essere misurata una sola volta).

La probabilità legata a tali eventi è definita *probabilità soggettiva*, in quanto non legata a degli esperimenti o ad una teoria, ma al grado di fiducia che l'esperienza e il giudizio personale assegnano al verificarsi di un determinato evento.

11.1.1.3 I tre assiomi della teoria della probabilità

Sebbene le precedenti definizioni di probabilità abbiano natura diversa, per tutte valgono alcuni assiomi che permettono di sviluppare strumenti analitici utili agli scopi applicativi. Si definisce funzione di probabilità $\text{Pr}[\cdot]$ una funzione il cui dominio è lo spazio degli eventi \mathbf{A} e il codominio l'intervallo $[0,1]$, $\text{Pr}: \mathbf{A} \rightarrow [0,1]$, tale che:

- 1) $\text{Pr}[A] \geq 0$ per ogni $A \in \mathbf{A}$;

$$2) \Pr[\Omega] = 1;$$

$$3) \text{ Se } A_1 \in \mathbf{A}, A_2 \in \mathbf{A} \text{ e } A_1 A_2 = \emptyset, \text{ allora } \Pr[A_1 + A_2] = \Pr[A_1] + \Pr[A_2]$$

Dai tre assiomi precedenti possono essere dedotte alcune relazioni utili ai fini del calcolo.

Il terzo assioma può essere esteso a qualunque sequenza di eventi mutuamente esclusivi. Se $A_1, A_2, \dots, A_k \in \mathbf{A}$, e $A_i A_j = \emptyset$ per ogni $i \neq j$, con $i, j = 1, \dots, k$, allora la probabilità dell'unione degli eventi è uguale alla somma delle probabilità dei singoli eventi:

$$\Pr[A_1 + A_2 + \dots + A_k] = \Pr[A_1] + \Pr[A_2] + \dots + \Pr[A_k] \quad \text{eq. 11.1.1}$$

La probabilità del complemento di un evento è uguale alla differenza tra l'unità e la probabilità di quell'evento:

$$\Pr[A^c] = 1 - \Pr[A] \quad \text{eq. 11.1.2}$$

La probabilità dell'unione di due eventi A e B non necessariamente mutuamente esclusivi è uguale alla differenza tra la somma delle probabilità di questi eventi e la probabilità della loro intersezione:

$$\Pr[A \cup B] = \Pr[A + B] = \Pr[A] + \Pr[B] - \Pr[AB] \quad \text{eq. 11.1.3}$$

La probabilità dell'evento nullo è zero:

$$\Pr[\emptyset] = 0 \quad \text{eq. 11.1.4}$$

La probabilità dell'unione di n eventi non supera la somma delle loro probabilità:

$$\Pr[A_1 + A_2 + \dots + A_k] \leq \Pr[A_1] + \Pr[A_2] + \dots + \Pr[A_k] \quad \text{eq. 11.1.5}$$

in cui l'uguaglianza vale per eventi mutuamente esclusivi.

Dati due eventi A e B nello spazio Ω , e $\Pr[B] \neq 0$, la probabilità condizionata dell'evento A dato il verificarsi dell'evento B , indicata con $\Pr[A|B]$, è definita come:

$$\Pr[A|B] = \Pr[AB] / \Pr[B] \quad \text{eq. 11.1.6}$$

Il concetto di probabilità condizionata può essere esteso a più eventi e da esso si può desumere la seguente legge moltiplicativa, ad esempio per tre eventi:

$$\Pr[ABC] = \Pr[A|BC] \Pr[B|C] \Pr[C] \quad \text{eq. 11.1.7}$$

e per k eventi:

$$\Pr[A_1 A_2 \dots A_k] \leq \Pr[A_1|A_2 \dots A_k] \Pr[A_2|A_3 \dots A_k] \dots \Pr[A_k] \quad \text{eq. 11.1.8}$$

Quando la probabilità di accadimento di un evento è indipendente dal verificarsi di un altro evento, i due eventi sono detti statisticamente indipendenti e valgono le seguenti relazioni:

$$\Pr[A|B] = \Pr[A] \text{ se } \Pr[B] \neq 0 \quad \text{eq. 11.1.9}$$

ovvero

$$\Pr[B|A] = \Pr[B] \text{ se } \Pr[A] \neq 0 \quad \text{eq. 11.1.10}$$

L'indipendenza implica:

$$\Pr[A \cap B] = \Pr[AB] = \Pr[A] \Pr[B] \quad \text{eq. 11.1.11}$$

e

$$\Pr[A \cup B] = \Pr[A + B] = \Pr[A] + \Pr[B] - \Pr[A] \Pr[B] \quad \text{eq. 11.1.12}$$

Per una trattazione estesa, ulteriori dettagli ed esempi, si rimanda al testo Kottegota e Rosso (2008; Cap. 2).

11.1.2 Variabili casuali, leggi di probabilità, momenti

Una variabile casuale reale può essere definita come una funzione il cui dominio è lo spazio campionario e il cui codominio è l'insieme (o un sottoinsieme) dei numeri reali,

$$X: \Omega \rightarrow \mathfrak{R},$$

ossia una funzione di un esperimento che associa un valore numerico (della variabile casuale) ad ogni possibile esito dell'esperimento. Esempi di variabili casuali sono il numero di volte in cui si ottiene "testa" lanciando una moneta ad esempio cinque volte, l'altezza di una persona scelta per un esperimento, il numero di incidenti stradali che si verificano in una città nel corso di un determinato anno, il valore di portata in osservata in fissate sezioni di un corso d'acqua.

Una variabile casuale può essere discreta o continua. Una variabile discreta può assumere solo un numero finito di valori (e.g., numeri interi positivi), mentre una variabile continua può assumere ogni valore in un intervallo delimitato da valori legati a limiti fisici o teorici della grandezza descritta dalla variabile (e.g., le portate di un fiume possono assumere solo valori reali ≥ 0).

Le proprietà statistiche di una variabile casuale sono descritte dalla sua distribuzione o legge di probabilità, ossia una funzione che assegna un valore di probabilità ad ogni valore della variabile stessa.

Nel seguito, una variabile casuale sarà indicata convenzionalmente con una lettera maiuscola (e.g., X), mentre una sua realizzazione, ossia un suo valore osservato, con la corrispondente lettera minuscola (e.g., x).

Per una variabile discreta X , si definisce funzione densità di probabilità discreta, la funzione:

$$p_X(x) = \Pr[X = x] \quad \text{eq. 11.1.13}$$

per la quale valgono le seguenti relazioni derivanti dai tre assiomi della probabilità:

- 1) $0 \leq p_X(x) \leq 1$, per tutti i possibili valori x
- 2) $p_X(x) = 0$, per tutti i valori x non ammissibili
- 3) $\sum p_X(x) = 1$, allorché la somma è estesa a tutti i possibili valori x .

Si definisce funzione di ripartizione (o funzione di ripartizione cumulata, o distribuzione cumulata), la probabilità che la variabile casuale X non superi un valore x , ossia:

$$F_X(x) = \Pr[X \leq x] \quad \text{eq. 11.1.14}$$

La funzione F_X è una funzione monotona crescente per valori crescenti di X e tale che $0 \leq F_X(x) \leq 1$, per tutti i possibili valori x . Nel caso di una variabile discreta, F_X è la somma delle probabilità di tutti i possibili valori di X minori o uguali all'argomento x :

$$F_X(x) = \sum_{x_k \leq x} p_X(x_k) \quad \text{eq. 11.1.15}$$

La legge di probabilità di una variabile continua è completamente specificata dalla cosiddetta funzione di densità di probabilità f_X , una funzione continua e positiva nell'insieme dei possibili valori di X , e tale che:

$$\Pr[x_1 \leq X \leq x_2] = \int_{x_1}^{x_2} f_X(x) dx \quad \text{eq. 11.1.16}$$

Per una variabile casuale continua può essere applicata la definizione di funzione di ripartizione sostituendo al simbolo di sommatoria quello di integrazione, per cui:

$$F_X(x) = \int_{-\infty}^x f_X(z) dz \quad \text{eq. 11.1.17}$$

per $-\infty < x < \infty$, da cui segue che la densità di probabilità è la derivata prima della funzione di ripartizione:

$$f_X(x) = \frac{dF_X(x)}{dx} \quad \text{eq. 11.1.18}$$

Si definisce media teorica o valore atteso di una variabile casuale la media della variabile ponderata in base alla distribuzione di probabilità. Da un punto di vista meccanico la media rappresenta il baricentro dell'area sottesa dalla funzione di densità di probabilità. In particolare, per variabili casuali discrete e continue la media è espressa rispettivamente dalle relazioni:

$$\mu_X = E[X] = \sum_{\forall x_i} x_i p_X(x_i) \quad \text{eq. 11.1.19}$$

$$\mu_X = E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \quad \text{eq. 11.1.20}$$

in cui $E[\cdot]$ è detto operatore valore atteso.

Si definisce momento di ordine r di una variabile casuale intorno ad un valore a , la funzione di una variabile casuale discreta (continua) X :

$$\mu_r^* = E[(X - a)^r] = \sum_{\forall x_i} (x_i - a)^r p_X(x_i) \quad \text{eq. 11.1.21}$$

$$\mu_r^* = E[(X - a)^r] = \int_{-\infty}^{\infty} (x - a)^r f_X(x) dx \quad \text{eq. 11.1.22}$$

Se $a = 0$, il generico momento r -esimo è detto momento assoluto. Il primo momento assoluto coincide con la media μ_X . Se $a = \mu_X$, il momento è detto centrale:

$$\mu_r = E[(X - \mu_X)^r] = \sum_{\forall x_i} (x_i - \mu_X)^r p_X(x_i) \quad \text{eq. 11.1.23}$$

$$\mu_r = E[(X - \mu_X)^r] = \int_{-\infty}^{\infty} (x - \mu_X)^r f_X(x) dx \quad (1.24) \quad \text{eq. 11.1.24}$$

I momenti di una variabile casuale riassumono importanti proprietà della sua distribuzione. In particolare, la media è un cosiddetto indice di posizione. Essa indica il valore intorno al quale tende a concentrarsi la distribuzione della variabile.

Un altro importante indicatore è la varianza, ossia il momento centrale di ordine $r = 2$, che riassume la variabilità di X intorno alla media:

$$\text{Var}[X] = E[(X - E[X])^2] = E[X^2] - (E[X])^2 \quad \text{eq. 11.1.25}$$

³ Il termine di destra dell'equazione si ottiene tramite semplici passaggi:

$$\text{Var}[X] = E[(X - E[X])^2] = E[X^2 - 2X \cdot E[X] + (E[X])^2] = E[X^2] - E[2X \cdot E[X]] + E[(E[X])^2]$$

ricordando che $E[E[X]] = E[X]$ segue che:

$$\text{Var}[X] = E[X^2] - 2(E[X])^2 + (E[X])^2 = E[X^2] - (E[X])^2$$

Dall'equazione precedente emerge che la varianza ha le dimensioni della variabile casuale al quadrato (ad esempio, se la variabile ha le dimensioni di una lunghezza, allora la varianza ha le dimensioni di una lunghezza al quadrato). Per esprimere la variabilità intorno alla media tramite valori con dimensioni congruenti a quelle della variabile casuale, si ricorre alla deviazione standard $\sigma_x = \sqrt{\text{Var}[X]}$. Due ulteriori indicatori legati ai momenti sono il coefficiente di asimmetria (o skewness) e il coefficiente di kurtosi:

$$\gamma_1[X] = \frac{\mu_3}{\sqrt{\mu_2^3}} = \frac{E[(X - E[X])^3]}{\sqrt{\{E[(X - E[X])^2]\}^3}} \quad \text{eq. 11.1.26}$$

$$\gamma_2[X] = \frac{\mu_4}{\mu_2^2} = \frac{E[(X - E[X])^4]}{\{E[(X - E[X])^2]\}^2} \quad \text{eq. 11.1.27}$$

Il primo indica l'eventuale asimmetria delle code della distribuzione: se $\gamma_1 = 0$ allora la distribuzione è simmetrica, se γ_1 ha valore positivo (negativo) allora la distribuzione è detta asimmetrica positiva (negativa) e la coda destra (sinistra) è più lunga della sinistra (destra). Il coefficiente di kurtosi dà un'indicazione del peso delle code della distribuzione (ossia delle parti estreme della distribuzione) rispetto alla parte centrale. Questi concetti saranno approfonditi nella sezione successiva in riferimento a campioni di numerosità finita.

Una delle finalità più comuni della raccolta di dati è la ricerca di relazioni causa effetto tra fenomeni. Per questo tipo di studi è necessario analizzare il contemporaneo presentarsi delle osservazioni di più variabili, anziché limitarsi allo studio delle singole distribuzioni. Quando un fenomeno è descritto da più variabili (e.g., la portata massima, il volume e la durata di un'onda di piena) si parla di distribuzioni congiunte o multivariate. Una distribuzione è detta bivariata quando il numero delle variabili considerate è uguale a due (e.g., la portata massima e il volume di un'onda di piena). I concetti esposti in precedenza, relativi ad una variabile casuale, possono essere estesi a due o più variabili considerate simultaneamente.

Nel caso bivariato per due variabili casuali X e Y discrete, la probabilità che le due variabili assumano simultaneamente valori x e y , ossia la probabilità dell'evento $(X = x) \cap (Y = y)$, è descritta dalla densità di probabilità discreta:

$$p_{XY}(x,y) = \Pr[(X = x) \cap (Y = y)] \quad \text{eq. 11.1.28}$$

e dalla corrispondente funzione di ripartizione congiunta:

$$F_{XY}(x, y) = \Pr[(X \leq x) \cap (Y \leq y)] = \sum_{x_i \leq x} \sum_{y_j \leq y} p_{XY}(x_i, y_j) \quad \text{eq. 11.1.29}$$

Le precedenti possono essere estese al caso di k variabili. In generale, la distribuzione congiunta di k variabili X_1, \dots, X_k , ossia di una variabile k -dimensionale, è definita come la probabilità dell'intersezione dei k eventi $(X_1 = x_1), \dots, (X_k = x_k)$, in cui (x_1, \dots, x_k) sono dei punti nello spazio campionario k -dimensionale.

Analogamente, per una variabile bidimensionale continua è possibile definire la densità di probabilità congiunta f_{XY} , tale che :

$$\Pr[(x_1 \leq X \leq x_2) \cap (y_1 \leq Y \leq y_2)] = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{XY}(x, y) dx dy \quad \text{eq. 11.1.30}$$

e la corrispondente distribuzione cumulata congiunta:

$$F_{XY}(x, y) = \Pr[(-\infty \leq X \leq x) \cap (-\infty \leq Y \leq y)] = \int_{-\infty}^x \int_{-\infty}^y f_{XY}(u, v) du dv \quad \text{eq. 11.1.31}$$

dalle quali segue che:

$$f_{XY}(x, y) = \frac{d^2 F_{XY}(x, y)}{dx dy} \quad \text{eq. 11.1.32}$$

Analogamente al caso di variabili discrete, anche per variabili continue le precedenti relazioni possono essere estese a qualunque dimensione k . L'operatore valore atteso introdotto per una variabile casuale può essere esteso al caso in cui si considerino più variabili. In particolare, si definisce covarianza di due variabili casuali X e Y il valore atteso del prodotto delle rispettive deviazioni dalle medie. In formule:

$$Cov[X, Y] = E[(X - E[X])(Y - E[Y])] = E[XY] - E[X]E[Y] \quad \text{eq. 11.1.33}$$

in cui

$$E[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{XY}(x, y) dx dy \quad \text{eq. 11.1.34}$$

Quando due variabili sono indipendenti ($f_{XY} = f_X f_Y$), allora $E[XY] = E[X]E[Y]$ da cui segue che $Cov[X, Y] = 0$.

Se $X = Y$, allora

$$Cov[X, Y] = Cov[X, X] = E[(X - E[X])(X - E[X])] = E[(X - E[X])^2] = Var[X]$$

La covarianza assume valori elevati e positivi se entrambe le variabili tendono ad assumere contemporaneamente valori elevati o valori piccoli, mentre assume valori elevati e negativi se a valori elevati di una variabile tendono a corrispondere valori piccoli dell'altra. La covarianza è una misura della relazione lineare tra le variabili X e Y .

Si definisce coefficiente di correlazione ρ la covarianza normalizzata tramite le deviazioni standard σ_X e σ_Y delle variabili X e Y :

$$\rho = \frac{Cov[X, Y]}{\sigma_X \sigma_Y} \quad \text{eq. 11.1.35}$$

per il quale si può dimostrare che $-1 \leq \rho \leq 1$. L'uso di ρ permette di confrontare in modo omogeneo il grado di associazione lineare tra coppie di variabili $(X_1, Y_1), \dots, (X_k, Y_k)$.

Nello studio delle serie temporali, è di interesse valutare il grado di associazione lineare tra i valori di una variabile X osservati ad un generico istante s e quelli osservati ad un generico istante $t = s + h$, in cui h è un intervallo temporale che separa gli istanti s e t , detto *lag*. Supponendo di disporre di un campione di numerosità n , i due vettori di osservazioni $\{x_1, \dots, x_{n-h}\}$ e $\{x_{h+1}, \dots, x_n\}$, di numerosità $(n-h)$, possono essere pensati come realizzazioni di due variabili denominate convenzionalmente X_s e X_t . Ponendo $X = X_s$ e $Y = X_t$, è possibile applicare i concetti di covarianza e correlazione, che in questo caso assumono le denominazioni di *autocovarianza* e *autocorrelazione*, poiché il grado di associazione lineare non è misurato tra i valori simultanei di due variabili eterogenee (e.g., valori simultanei di volume e durata di un'onda di piena), ma tra i valori di una stessa variabile osservati ad istanti separati da un intervallo temporale h . Precisamente, l'autocovarianza è definita dalla relazione:

$$\begin{aligned} Cov[X_s, X_t] &= E[(X_s - E[X_s])(X_t - E[X_t])] = \\ &= E[X_s X_t] - E[X_s]E[X_t] \end{aligned} \quad \text{eq. 11.1.36}$$

mentre l'autocorrelazione è data da:

$$\rho = \frac{\text{Cov}[X_s, X_t]}{\sigma_{X_s} \sigma_{X_t}} \quad \text{eq. 11.1.37}$$

Al variare di h , le precedenti relazioni definiscono le cosiddette funzioni di autocovarianza e autocorrelazione. Per ulteriori dettagli si rimanda al testo di Kottegotte e Rosso (2008; Cap. 3).

11.1.3 Elementi di statistica descrittiva e analisi esplorativa dei dati

Se un fenomeno è descritto da una variabile casuale X , le sue proprietà statistiche sono definite in modo completo dalla distribuzione di probabilità F_X (o dalla densità f_X) ad essa associata. Dunque, il confronto tra due o più fenomeni o l'esame di uno stesso fenomeno in circostanze diverse possono essere ricondotti al confronto delle rispettive distribuzioni. L'analisi statistica diviene più semplice se si ricorre a degli indici di sintesi che riassumano le proprietà delle variabili in esame. Tali indici devono essere scelti con riferimento esplicito allo scopo dell'indagine statistica, avendo presente il principio generale che ogni sintesi comporta una perdita di informazione rispetto ai dati elementari, per cui essa va ricercata in modo da minimizzare tale perdita.

È quindi necessario esplicitare quali aspetti della distribuzione si intendono esaminare. I tre aspetti fondamentali, a cui si è già parzialmente accennato al paragrafo precedente nella discussione riguardante l'interpretazione dei momenti centrali, sono:

- la *posizione*, ossia la misura della centralità complessiva della distribuzione in relazione ai possibili valori assumibili da X e alle rispettive probabilità di accadimento. La sintesi è dunque un valore rappresentativo della variabile nella sua globalità.
- la *variabilità*, ossia la "mutevolezza" o dispersione dei dati.
- la *forma*, vale a dire l'aspetto complessivo della distribuzione rispetto a configurazioni di riferimento. In particolare, la sintesi deve misurare la simmetria della distribuzione (rispetto ad un punto notevole, come una misura di posizione), il peso degli estremi in rapporto ai valori centrali della distribuzione, ecc..

11.1.3.1 Misure di posizione

Nel paragrafo 11.1.2 si è visto che media, varianza (deviazione standard) e coefficienti di asimmetria e kurtosi sono degli indici di sintesi che descrivono alcuni aspetti di una variabile X con distribuzione di probabilità F_X . Le espressioni di questi indici riportate nel 11.1.2 sono riferite all'intera popolazione, ossia a tutti i valori che possono essere assunti da X . Poiché operativamente si analizzano campioni di dimensioni finite, è utile introdurre le corrispondenti espressioni campionarie e alcune misure alternative di posizione, variabilità e forma che forniscono una stima delle corrispondenti grandezze teoriche.

In generale, questi metodi sono detti stimatori e il valore da loro restituito per un particolare campione è detto stima; ad esempio la media aritmetica nell'eq. 11.1.38 rappresenta uno stimatore delle medie teoriche descritte nelle eq. 11.1.19 e eq. 11.1.20. Il valore restituito da uno stimatore per un particolare campione è detto stima.

Dato un campione $\{x_1, \dots, x_n\}$, allora la misura di posizione più nota è senz'altro la media aritmetica:

$$\hat{\mu}_X = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{eq. 11.1.38}$$

La Figura 11.1 mostra due campioni ottenuti da due distribuzioni Gaussiane⁴ con medie diverse (uguali rispettivamente a 3 e 9). I trattini lungo l'asse delle ascisse rappresentano i valori campionati (il grafico è detto strip-plot o strip-chart) ed evidenziano la tendenza del campione a addensarsi intorno alle due medie (indicate con le linee tratteggiate), assunte come indici di posizione. Le corrispondenti densità di probabilità indicano l'elevata frequenza di accadimento dei valori intorno alle medie.

⁴ Si definisce distribuzione Gaussiana, o di Gauss, o normale la legge di probabilità di una variabile casuale X con densità di probabilità:

$$f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(x-\mu)^2}{\sigma^2}\right)$$

per $-\infty < x < \infty$, e in cui $-\infty < \mu < \infty$ è il parametro di posizione e $\sigma^2 > 0$ è il parametro di scala.

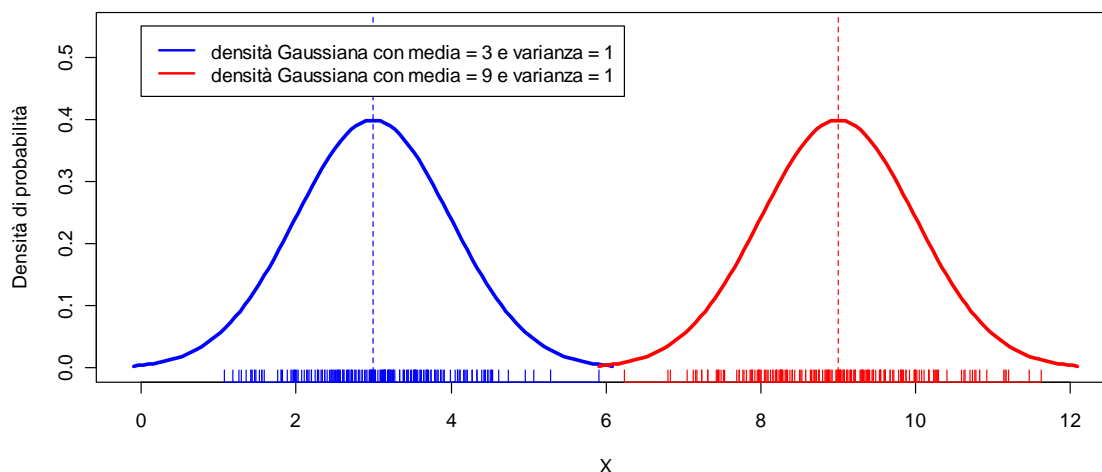


Figura 11.1 - Esempio di due campioni distribuiti secondo una legge Gaussiana con due diverse medie e uguale varianza.

La media ha il difetto di essere sensibile al verificarsi di valori notevolmente diversi dal resto del campione (*outlier*), dovuti ad esempio ad una errata rilevazione. Ad esempio, si ipotizzi che una variabile X possa assumere solo valori nell'insieme $\{1, 2, 3, 4, 5\}$. La media di X è uguale a 3. Se fosse presente un valore atipico $\{1, 2, 3, 4, 50\}$, dovuto ad esempio ad un errore di trascrizione, la media sarebbe 12.

Questo difetto, evidente soprattutto in campioni di numerosità limitata, dipende dal fatto che la media rappresenta il baricentro della distribuzione: un valore molto diverso da tutte le altre osservazioni attrae il baricentro nella sua direzione, cosicché la media diviene un semplice punto di equilibrio e non più un valore rappresentativo dell'intero campione.

Due misure alternative o complementari alla media sono la *moda* e la *mediana*.

La moda è il valore della variabile X a cui corrisponde la massima frequenza (picco massimo della densità di probabilità). Usare la moda come indice di sintesi vuol dire assumere come valore più rappresentativo quello che si è verificato più spesso. In campioni di numerosità limitata o in particolari fenomeni, la funzione di densità di probabilità può mostrare più picchi. In tal caso, si parla di distribuzioni bi-modali o multi-modali. L'uso della moda è dunque opportuno solo se la distribuzione del campione presenta un solo massimo, ossia è uni-modale.

La mediana è il valore di una variabile X cui corrisponde una probabilità di superamento (o non superamento) del 50%, ossia tale che $F_X(x_{\text{mediana}}) = 0.5$, quindi, il valore che occupa la posizione centrale nella serie ordinata (in ordine crescente o decrescente) delle osservazioni. La mediana è dunque determinata in modo che metà delle osservazioni abbia valori inferiori alla mediana e metà valori superiori. Una proprietà della mediana particolarmente interessante è quella di essere rappresentativa della *posizione* della distribuzione anche in presenza di valori estremi molto diversi da tutti gli altri. Questa proprietà è detta *robustezza* ed è dovuta al fatto che il calcolo della mediana tiene conto solo dell'ordinamento delle osservazioni, considerando solo il valore dell'osservazione collocata al centro della graduatoria ordinata. Questa caratteristica rappresenta un vantaggio rispetto alla media. Di contro, la mediana è sensibile alle variazioni nella parte centrale della distribuzione, proprio perché determinata dall'osservazione posta al centro del campione ordinato.

In conclusione la scelta dell'indice di sintesi deve essere guidata dagli obiettivi da perseguire: se i dati non presentano valori anomali ed è necessario tener conto di tutte le osservazioni, allora la media è l'indicatore di posizione più opportuno; se si desidera eliminare gli effetti di misure anomale, allora la mediana è la misura di posizione più appropriata.

11.1.3.2 Quantili

Il concetto di mediana può essere generalizzato in quello di *quantile*. Come la mediana è il valore cui corrisponde una probabilità di non superamento $F_X(x_{\text{mediana}}) = 0.5$, il generico quantile p -esimo, x_p , è il valore con probabilità cumulata p , ossia tale che $F_X(x_p) = p$. I quantili con $p = 0.25, 0.50, 0.75$ sono detti rispettivamente primo, secondo e terzo *quartile* (il secondo quartile coincide con la mediana). I quantili con $p = 0.10, 0.20, \dots, 0.90$ sono detti anche *decili*. Il quantile p -esimo (e.g., 0.30-esimo) è detto anche *percentile* $100p$ -esimo (30-esimo).

Per funzioni di ripartizione strettamente crescenti (la maggior parte di quelle usate nelle applicazioni) il quantile è calcolato tramite la funzione inversa della distribuzione: $x_p = F_X^{-1}(p)$. Nel caso in cui la funzione di ripartizione non è strettamente monotona, allora x_p è definito come il più piccolo valore tale che $F_X(x_p) = p$, ossia $F_X^{-1}(p) = \inf \{x : F_X(x) \geq p\}$.

11.1.3.3 Misure di variabilità

La variabilità di un fenomeno è la sua attitudine ad assumere valori differenti (Piccolo 2004; p. 98). È opportuno che una misura di variabilità rispetti alcuni requisiti, quali

- 1) essere sempre positiva o uguale a zero,
- 2) essere zero per variabili che assumono un solo valore,
- 3) essere invariabile allorché si aggiunge una costante alla variabile X .

Come accennato nel 11.1.2, una misura di variabilità è la varianza. Per un campione finito di numerosità n della variabile X , l'espressione della varianza campionaria è:

$$\hat{\sigma}_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \mu_X)^2 \quad \text{eq. 11.1.39}$$

La varianza misura la dispersione delle osservazioni rispetto alla media μ_X , assunta come valore di riferimento. Se la media della popolazione μ_X non è nota, essa può essere sostituita dalla media campionaria $\hat{\mu}_X$. La radice quadrata della varianza campionaria restituisce la deviazione standard del campione. La Figura 11.2 illustra due campioni con medie e numerosità uguali, ma diversa varianza. Gli strip-plot in essa riportati evidenziano come uno dei campioni sia maggiormente disperso intorno al valore centrale rispetto all'altro campione.

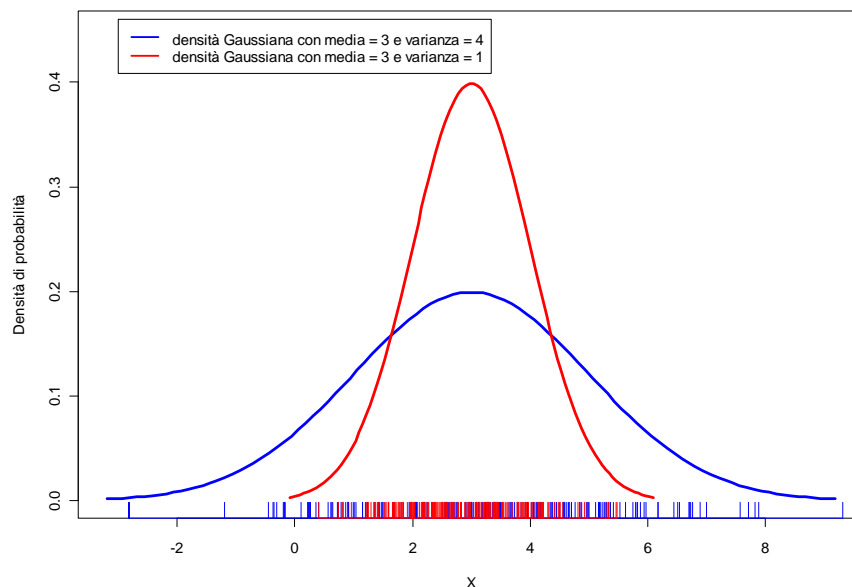


Figura 11.2 - Esempio di due campioni distribuiti secondo una legge gaussiana con uguale media e diversa varianza.

Misure di variabilità alternative possono essere desunte dal campione (di numerosità n) ordinato in ordine crescente o decrescente. Tra queste è utile ricordare:

- il campo di variazione (*range*) $R(X)$, ossia la differenza tra i valori di massimo e minimo osservato, $R(X) = \max\{x_1, \dots, x_n\} - \min\{x_1, \dots, x_n\}$. Questa misura è influenzata dai due valori estremi, quindi va usata con cautela se si sospetta la presenza di valori estremi anomali;
- la distanza inter-quartile (*inter-quartile range*) $IQR(X)$, ossia la differenza tra il terzo e il primo quartile, $IQR(X) = x_{0.75} - x_{0.25}$. Questo indice è basato sulle informazioni contenute nel

⁵ La formula eq. 11.1.39 rappresenta lo stimatore puntuale non distorto della varianza, come si evince dal denominatore $(1-n)$. Per approfondimenti riguardo i concetti di stimatore distorto e non distorto si rimanda a (Piccolo 2004; pp. 326-328).

50% del campione collocato al centro della distribuzione. Può dunque essere opportuno considerare indici fondati su una maggiore porzione del campione;

- la distanza inter-decile (*inter-decile range*) $IDR(X)$, ossia la differenza tra il nono e il primo decile, $IDR(X) = x_{0.90} - x_{0.10}$. Questo indice è robusto rispetto a possibili valori anomali presenti agli estremi del campione. Inoltre, esso tiene conto della variabilità espressa dall'80% del campione.

11.1.3.4 Misure di forma

La posizione e la variabilità di una distribuzione non esauriscono l'informazione contenuta in un campione, poiché due variabili possono avere indici di posizione e variabilità uguali, ma differire per il peso dei valori estremi rispetto al valore centrale, a causa del diverso comportamento delle code della distribuzione. Questi aspetti possono essere quantificati tramite degli indici.

Una distribuzione che presenta una frazione di osservazioni con valori elevati molto più numerosa rispetto alla frazione di osservazioni con valori bassi è detta asimmetrica positiva (o asimmetrica a destra). Viceversa, se la distribuzione presenta una frazione di osservazioni con valori bassi molto più numerosa rispetto alla frazione di osservazioni con valori alti, essa è detta asimmetrica negativa (o asimmetrica a sinistra). Se la distribuzione ha una sola moda, allora valgono le seguenti relazioni:

- asimmetria a destra (positiva) \Rightarrow moda $<$ mediana $<$ media,
- asimmetria a sinistra (negativa) \Rightarrow media $<$ mediana $<$ moda.

Una misura di asimmetria è il coefficiente di asimmetria di Fisher. Se si considera la variabile standardizzata $Z = (X - \mu_X) / \sigma_X$, questo indice è definito come la media aritmetica delle terze potenze di Z :

$$\hat{\gamma}_1 = \frac{1}{n} \sum_{i=1}^n (z_i)^3 = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \mu_X}{\sigma_X} \right)^3 \quad \text{eq. 11.1.40}$$

ed è positivo, negativo o nullo per una distribuzione, rispettivamente, asimmetrica positiva, negativa o simmetrica, analogamente a quanto già visto al paragrafo 11.1.2 relativamente al corrispondente indice γ_1 (valido per l'intera popolazione di una variabile X con distribuzione F_X).

La Figura 11.3 mostra le tre possibili configurazioni di una distribuzione in termini di simmetria, la mutua posizione di media, mediana e moda, e i corrispondenti valori del coefficiente di asimmetria.

Un altro aspetto importante della forma di una distribuzione riguarda il peso più o meno accentuato delle code rispetto alla parte centrale. La presenza di un numero elevato di osservazioni nella zona centrale ovvero in prossimità delle code causa una concentrazione delle densità di probabilità nella parte centrale della distribuzione ovvero un appiattimento dovuto al maggiore peso delle code. Questo comportamento è definito con il termine kurtosi.

Un indice che sintetizza questo aspetto è il coefficiente di kurtosi di Pearson, definito come la media aritmetica delle potenze quarte della variabile standardizzata Z :

$$\hat{\gamma}_2 = \frac{1}{n} \sum_{i=1}^n (z_i)^4 = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \mu_X}{\sigma_X} \right)^4 \quad \text{eq. 11.1.41}$$

Per la distribuzione Gaussiana, il coefficiente di kurtosi di Pearson è uguale a 3; per distribuzioni con una maggiore concentrazione della densità intorno ai valori centrali rispetto alla distribuzione Gaussiana, il coefficiente di kurtosi è maggiore di 3, e la distribuzione è detta leptokurtica (Piccolo 2004; Cap. 6, p. 126); per distribuzioni con minore concentrazione della densità intorno ai valori centrali rispetto alla distribuzioni Gaussiana, il coefficiente di kurtosi è minore di 3 e la distribuzione è detta platikurtica (Piccolo 2004; Cap. 6, p. 126).

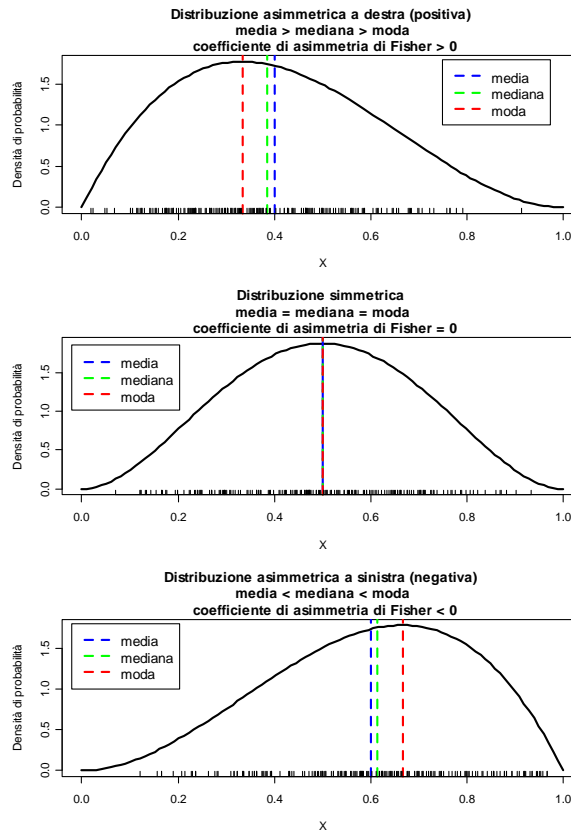


Figura 11.3 - Esempi di distribuzioni con le tre configurazioni possibili in termini di simmetria.

La Figura 11.4 mostra il confronto di una distribuzione leptokurtica (distribuzione logistica⁶) ed una platikurtica (distribuzione uniforme⁷) con una distribuzione normale. I parametri delle tre distribuzioni sono stati scelti in modo tale che le tre densità sono simmetriche intorno allo zero ed hanno la stessa varianza. L'aspetto differenziante è la modalità in cui la massa di probabilità è concentrata intorno al picco e sulle code.

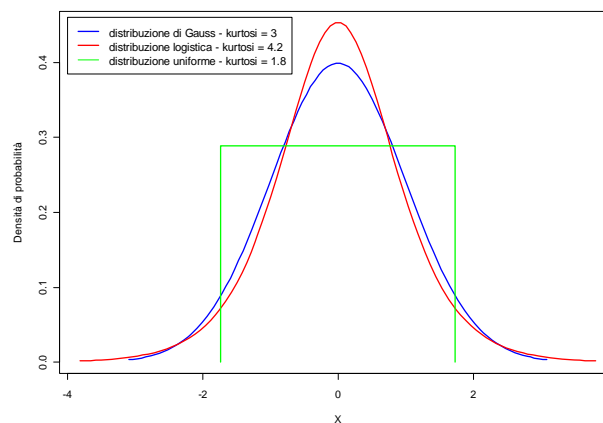


Figura 11.4 - Esempi di distribuzioni leptokurtica e platikurtica rispetto alla distribuzione normale.

⁶ Si definisce distribuzione logistica la legge di probabilità di una variabile casuale X con funzione di ripartizione:

$$F_X(x; \mu, \sigma) = \frac{1}{1 + \exp\left(-\frac{x - \mu}{\sigma}\right)}$$

per $-\infty < x < \infty$ e in cui $-\infty < \mu < \infty$ è il parametro di posizione e $\sigma > 0$ è il parametro di scala.

⁷ Si definisce distribuzione uniforme la legge di probabilità di una variabile casuale X con densità di probabilità:

$$f_X(x; a, b) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & x < a, x > b \end{cases}$$

in cui $a \leq x \leq b$, $-\infty < a, b < \infty$ sono due parametri di posizione.

11.1.3.5 Coefficiente di correlazione lineare di Pearson e coefficiente di correlazione di rango di Kendall

Nel paragrafo 11.1.2 è stato introdotto il coefficiente di correlazione lineare come misura dell'associazione lineare tra due variabili X e Y . La versione campionaria per il calcolo su serie di osservazioni con numerosità finita n è data dalla relazione:

$$\hat{\rho} = \frac{1}{n} \sum_{i=1}^n \left(\frac{x_i - \mu_X}{\sigma_X} \right) \left(\frac{y_i - \mu_Y}{\sigma_Y} \right) = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \mu_X)(y_i - \mu_Y)}{\sigma_X \sigma_Y} \quad \text{eq. 11.1.42}$$

La Figura 11.5 illustra l'interpretazione del valore del coefficiente di correlazione lineare in alcuni campioni simulati imponendo un prefissato valore di ρ . Essi mostrano come nel passare da valori negativi a valori positivi la pendenza della retta che attraversa idealmente le nuvole si inverte, e l'addensamento dei punti è tanto più vicino alla retta che li attraversa quanto maggiore è il valore di ρ . Quando il valore di ρ tende a zero la nuvola non individua una direzione definita, come è atteso per il fatto che la variabile X non segue le variazioni di Y .

È utile precisare che il coefficiente di correlazione di Pearson evidenzia esclusivamente delle relazioni lineari tra due variabili, ma non altri tipo di correlazioni. Ad esempio, si consideri una variabile X con distribuzione normale standard ed una variabile $Y = X^2$. Le due variabili sono legate da una legge funzionale, per cui ci si attende che un indice di associazione rilevi tale legame. Tuttavia, ricordando le relazioni che esprimono la covarianza e la correlazione eq. 11.1.33 e eq. 11.1.35, e che il momento di ordine tre della distribuzione normale standard è nullo (la distribuzione è simmetrica), si ha che la covarianza (e dunque la correlazione lineare) è nulla:

$$\text{Cov}[X,Y] = E[XY] - E[X] E[Y] = E[X^3] - E[X] E[X^2] = 0 - 0 \cdot E[X^2] = 0.$$

In generale, ogni insieme di proprietà auspicabili per una misura di associazione $\kappa(X,Y)$ dovrebbe includere le seguenti:

1. $\kappa(X,Y) = \kappa(Y,X)$ (simmetria).
2. $-1 \leq \kappa(X,Y) \leq 1$ (normalizzazione).
3. $\kappa(X,Y) = -1 \Leftrightarrow X,Y$ sono contro-monotone, ossia legate in modo funzionale monotono tale che a valori elevati (piccoli) di X corrispondono valori piccoli (elevati) di Y ;
4. $\kappa(X,Y) = 1 \Leftrightarrow X,Y$ sono co-monotone ossia legate in modo funzionale monotono tale che a valori elevati (piccoli) di X corrispondono valori elevati (piccoli) di Y .
5. Per $T: \mathfrak{R} \rightarrow \mathfrak{R}$ strettamente monotona nell'intervallo di variazione di X :

$$\kappa(T(X), Y) = \begin{cases} \kappa(X, Y) & T : \text{funzione crescente} \\ -\kappa(X, Y) & T : \text{funzione decrescente} \end{cases} .$$

6. $\kappa(X,Y) = 0 \Leftrightarrow X,Y$ sono indipendenti.

La correlazione lineare soddisfa soltanto le prime due proprietà. Misure di associazione alternative come i coefficienti di correlazione di rango di Kendall e Spearman, l'indice di cograduazione di Gini e il coefficiente di correlazione mediale di Blomqvist soddisfano anche le proprietà 3 e 4 se X e Y sono variabili continue (Embrechts et al. 1999).

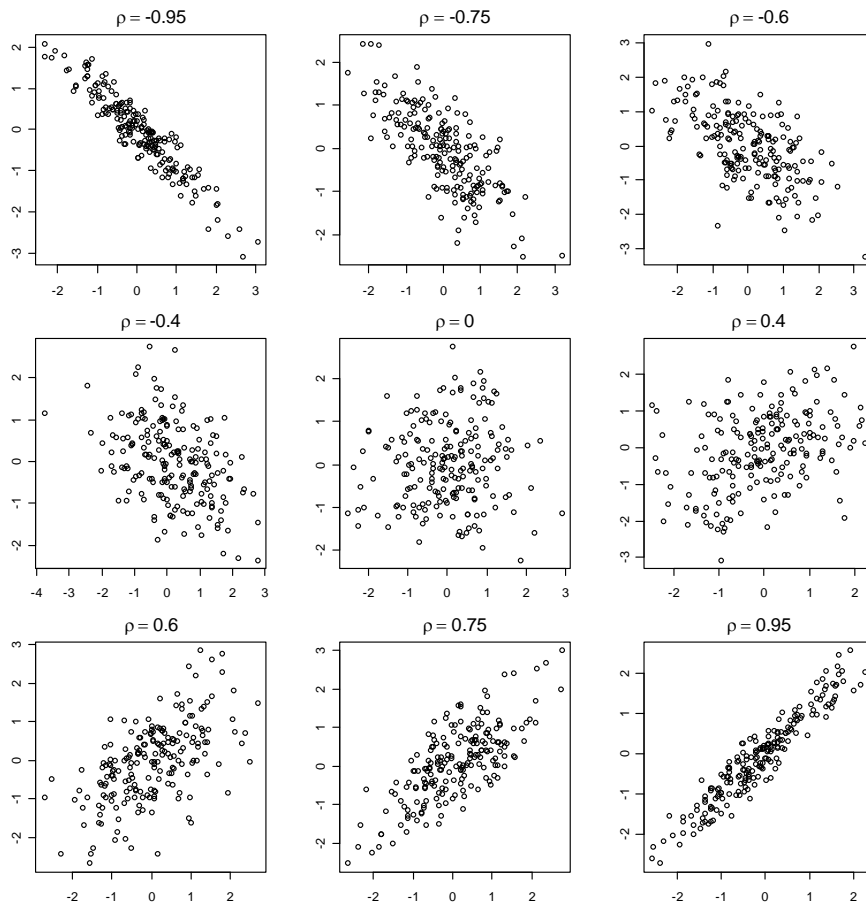


Figura 11.5 - Esempi di campioni con diversi valori del coefficiente di correlazione lineare di Pearson.

Tra queste misure, i coefficienti di Kendall e Spearman sono probabilmente quelli più noti. Allo scopo di illustrare le idee che sono alla base di queste misure di associazione si riporta la descrizione del coefficiente di Kendall, rimandando al lavoro di Nelsen (2006; pp.157-185) per un'esauriva trattazione degli altri coefficienti.

I coefficienti di Kendall e Spearman sono delle misure di dipendenza che valutano un legame tra le variabili noto con il nome di concordanza. In modo informale, una coppia di variabili casuali è detta concordante se “grandi” valori dell’una tendono ad essere associati a “grandi” valori dell’altra, e “piccoli” valori dell’una a “piccoli” valori dell’altra. Più precisamente, siano (x_i, y_i) e (x_j, y_j) due osservazioni estratte da un vettore di variabili casuali continue (X, Y) . Si dice che (x_i, y_i) e (x_j, y_j) sono *concordanti* se $x_i < x_j$ e $y_i < y_j$, ovvero se $x_i > x_j$ e $y_i > y_j$. Analogamente si afferma che (x_i, y_i) e (x_j, y_j) sono *discordanti* se $x_i < x_j$ e $y_i > y_j$, o se $x_i > x_j$ e $y_i < y_j$. In modo alternativo, (x_i, y_i) e (x_j, y_j) sono *concordanti* se $(x_i - x_j)(y_i - y_j) > 0$, e *discordanti* se $(x_i - x_j)(y_i - y_j) < 0$.

Il coefficiente di correlazione di rango τ_K di Kendall è definito in termini di concordanza come segue. Si indichi con $\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ un campione di n osservazioni ricavate da un vettore (X, Y) di variabili casuali continue. Nel campione sono presenti $n(n-1)/2$ coppie distinte di osservazioni (x_i, y_i) e (x_j, y_j) , ed ogni coppia di esse è concordante o discordante. Indicando con c il numero di coppie concordanti e con d quello delle coppie discordanti, la stima della τ_K di Kendall per il campione è data da:

$$\hat{\tau}_K = \frac{c - d}{c + d} = \frac{2(c - d)}{n(n - 1)} \quad \text{eq. 11.1.43}$$

11.1.3.6 Alcuni metodi grafici di indagine

L’analisi esplorativa dei dati tramite appropriati grafici diagnostici (Tukey 1977; Cleveland 1993, 1994) è una fase fondamentale di ogni studio quantitativo. Spesso la sua importanza è sottovalutata, omettendo l’analisi visiva a favore del calcolo di statistiche sintetiche o di analisi numeriche incapaci

di evidenziare importanti aspetti delle serie deducibili solo tramite un controllo visivo diretto. Ciò può condurre peraltro a risultati errati, che potrebbero essere evitati da una semplice, ma informativa, analisi esplorativa.

Il modo più semplice di eseguire un'analisi grafica consiste nell'esaminare i dati "grezzi" per evidenziare problemi (dati mancanti, outliers), andamenti temporali (stagionalità, trend, cambiamenti repentini), strutture spaziali e regionali presenti nei dati. Una buona analisi esplorativa comprende il disegno, lo studio e il raffinamento dei grafici, con l'intento di evidenziare particolari aspetti e identificare ulteriori grafici ritenuti utili.

Tra i grafici più comunemente impiegati nell'analisi visiva è utile ricordare gli istogrammi, i grafici quantile-quantile (qq-plot), gli scatter plot, le curve smoothing (e.g., Cleveland 1993, 1994; Grubb e Robson 2000), i grafici di serie temporali (time-series plot) e i grafici delle funzioni di autocorrelazione. I primi quattro tipi sono riportati nelle Figura 11.6 e Figura 11.8, mentre gli ultimi due tipi saranno descritti ed applicati nella sezione relativa allo studio di serie autocorrelate.

Il pannello centrale della Figura 11.6 mostra il grafico a punti (scatter-plot) di due grandezze dedotte da una serie di valori di *Standardized Precipitation Index (SPI)* calcolati su serie pluviometriche a scala temporale semestrale registrate in Sicilia tra il 1921–2003. La disposizione dei punti indica l'esistenza di una relazione per la quale valori piccoli (grandi) di una grandezza corrispondono a valori piccoli (grandi) dell'altra.

Questo aspetto è confermato dall'andamento della linea blu, che rappresenta una curva di smoothing di tipo *LOESS* (o *LOWESS*, Locally Weighted Regression - regressione locale ponderata - Cleveland 1994). La curva di smoothing permette di definire l'andamento della relazione esistente tra i dati (ossia tra un variabile di risposta *Area* a una variabile esplicativa *SPI*) seguendo le variazioni locali senza imporre una legge di variazione analitica scelta a priori e potenzialmente non corretta.

Il principio dell'algoritmo alla base del *LOESS* utilizzato per la costruzione della curva riportata in Figura 11.6 è illustrato nella Figura 11.7. Per un determinato valore della variabile in ascissa, ad esempio $SPI = 1.4$, si definisce un intorno (delimitato dalle linee tratteggiate) contenente una prefissata frazione di punti α , il cui valore va scelto in base alle proprietà del campione e in genere varia da 0.25 a 1 (Cleveland 1994; p.172). Ad ogni osservazione della variabile *Area* contenuta nell'intorno è assegnato un peso tramite una funzione (pannello in alto a destra) che ha un massimo in corrispondenza del valore $SPI = 1.4$ e decresce fino ad annullarsi in corrispondenza degli estremi dell'intorno. La funzione, che in genere ha la forma $(1-u^3)^3$, $0 \leq u < 1$, assegna quindi un peso maggiore alle osservazioni di *Area* cui corrisponde un valore di *SPI* prossimo ad 1.4, e un peso minore a quelle con un valore di *SPI* più distante da 1.4. Si applica poi una regressione lineare ponderata (pannello in basso a sinistra). Il valore della retta di regressione in corrispondenza di $SPI = 1.4$ è il valore della curva *LOESS* per $SPI = 1.4$, ossia il valore atteso di *Area* per $SPI = 1.4$ restituito dalla regressione locale (pannello in basso a destra). Ripetendo la procedura per diversi valori di *SPI* si ottiene la curva riportata in Figura 11.6, che esprime l'interpolazione dei risultati delle regressioni locali.

Tornando alla Figura 11.6, i grafici a sinistra e in basso rispetto al pannello centrale rappresentano dei grafici "box-whiskers" (o box-plot) in cui gli estremi del box indicano il primo e terzo quartile, la linea intermedia rappresenta la mediana, e i segmenti tratteggiate forniscono un'indicazione della lunghezza delle code della distribuzione. I box-plot descrivono in modo sintetico la densità di probabilità, rappresentata in modo alternativo tramite gli istogrammi in alto e a destra.

Un istogramma è costruito dividendo il range dei valori osservati in un numero di classi tale che esse contengano un numero minimo di osservazioni per avere un grafico informativo (in letteratura sono disponibili diverse formule empiriche per la scelta del numero di classi più opportuno; si veda, ad esempio, Kottegota e Rosso (2008; Cap. 1)). Per ogni classe si calcola il numero di osservazioni che ricadono nella classe stessa. Infine, per ogni classe si disegna una barra la cui area è proporzionale alla frequenza delle osservazioni che in essa ricadono.

I trattini lungo gli assi del pannello centrale di Figura 11.6 sono gli strip-plots già introdotti in precedenza. Per ulteriori dettagli si rimanda a Kottegota e Rosso (2008; Cap. 1).

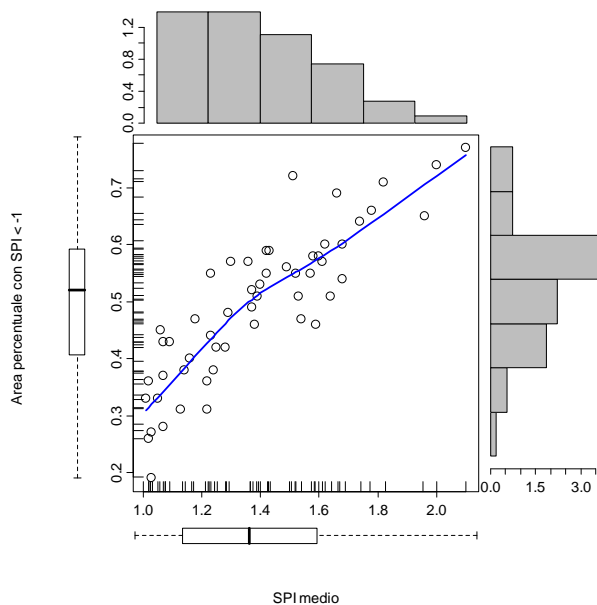


Figura 11.6 - Esempi di alcuni grafici diagnostici.

Il grafico in Figura 11.8 è detto grafico quantile-quantile (qq-plot) e riporta le osservazioni (ordinate) vs. i quantili che si otterrebbero allorché il campione fosse estratto da una distribuzione nota. In altri termini, dato un campione $\{y_1, \dots, y_n\}$, le osservazioni sono ordinate in ordine crescente e ad ognuna è associata una frequenza di non superamento empirica $F_n(x_i) = i / (n + 1)$, che individua la cosiddetta *funzione di ripartizione empirica*. La funzione F_n rappresenta una possibile controparte non parametrica (e dunque non vincolata a nessuna forma predefinita) di una funzione parametrica F_x . Se F_n e F_x sono simili, allora i quantili $x_i = F_x^{-1}(F_n(y_i))$ associati a F_x , corrispondenti alle frequenze empiriche associate alle osservazioni y_i dovrebbero essere uguali o molto simili (punti blu).

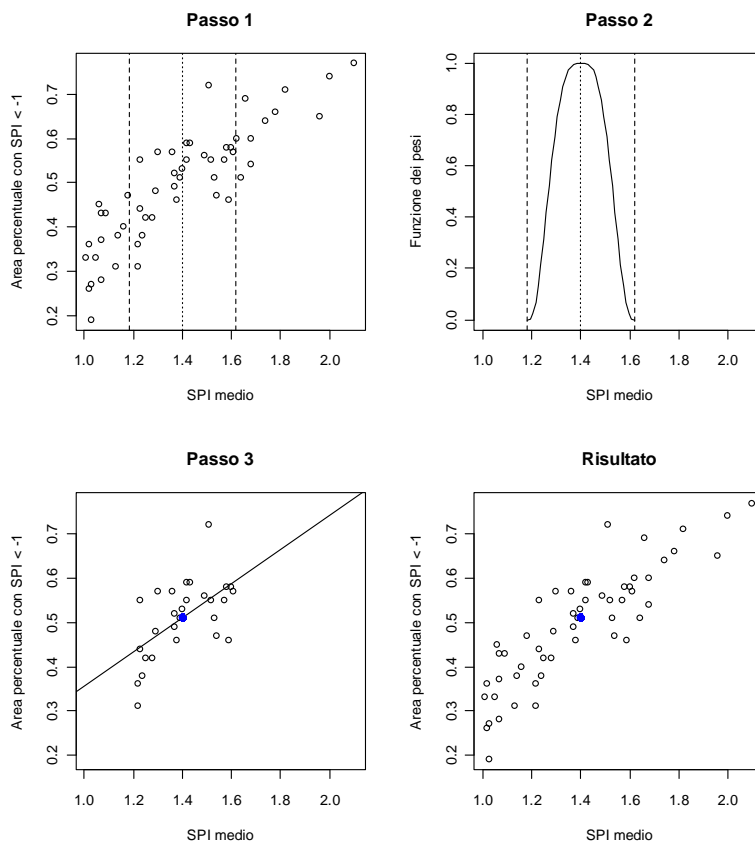


Figura 11.7 - Illustrazione dei passi dell'algoritmo LOESS applicato allo scatter-plot mostrato in Figura 11.6

In tal caso, x_i e y_i si dispongono lungo una retta a 45° , indicando che la distribuzione parametrica ben rappresenta le frequenze empiriche, e dunque è un buon modello probabilistico per le osservazioni. I punti rossi e verdi descrivono invece due possibili situazioni in cui F_x si discosta da F_n : come mostrato qualitativamente dai box-plot, nel caso specifico, una forma a “U” indica che la distribuzione empirica F_n è asimmetrica a destra rispetto alla teorica ipotizzata (in questo caso una Gaussiana standard), mentre una forma ad “U” rovesciata indica che la distribuzione empirica è asimmetrica a sinistra. Il tipo di scostamento fornisce un’utile indicazione per la scelta di distribuzioni alternative. Ad esempio, nel primo (secondo) caso il grafico suggerisce l’adozione di una distribuzione asimmetrica a destra (sinistra) in luogo della Gaussiana.

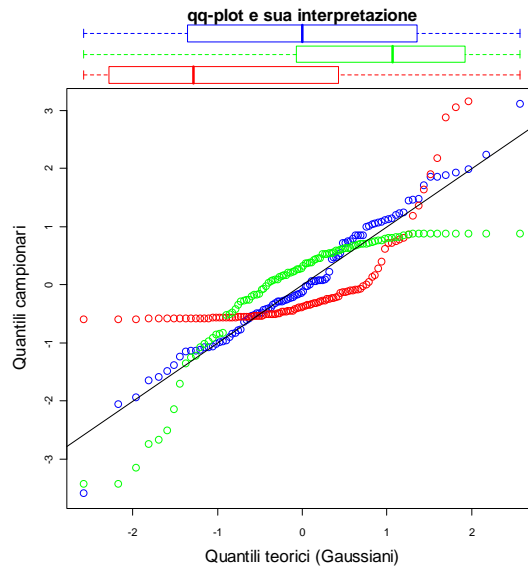


Figura 11.8 - La figura riporta tre possibili risultati ottenibili nell’applicazione di un qq-plot.

11.1.4 Definizioni di stazionarietà e di funzione di autocorrelazione

Una serie temporale è detta *stazionaria in senso forte* se la distribuzione di ogni insieme di valori $\{x_{t_1}, \dots, x_{t_k}\}$ è invariante nel tempo, ovvero tutte le funzioni di ripartizione multivariate per sottoinsiemi di variabili devono essere in accordo con le corrispondenti distribuzioni per gli insiemi traslati nel tempo, per tutti i valori del lag h . Questo si traduce analiticamente nella relazione:

$$\Pr[x_{t_1} \leq c_1, \dots, x_{t_k} \leq c_k] = \Pr[x_{t_1+h} \leq c_1, \dots, x_{t_k+h} \leq c_k] \quad \text{eq. 11.1.44}$$

per tutti i $k = 1, 2, \dots$, tutti gli istanti temporali t_1, t_2, \dots, t_k , tutti i valori c_1, c_2, \dots, c_k , e tutti i lag $h = 0, \pm 1, \pm 2, \dots$.

Per $k = 1$, la relazione precedente implica che le variabili che compongono la serie seguano la stessa distribuzione univariata, ossia $\Pr[x_t \leq c] = \Pr[x_{t+h} \leq c]$ per ogni istante temporale t e $t+h$. Inoltre, se la funzione media, $\mu_t = \mu(t)$, della serie x_t esiste, allora $\mu_{t+h} = \mu_t$ per ogni t e $t+h$, e quindi $\mu_t = \mu$ deve essere costante.

Quando $k = 2$, si ha allora che $\Pr[x_{t_1} \leq c_1, x_{t_2} \leq c_2] = \Pr[x_{t_1+h} \leq c_1, x_{t_2+h} \leq c_2]$, per ogni istante temporale t_1 e t_2 , e lag h . Se il processo ha varianza finita, la stazionarietà in senso forte (stazionarietà della distribuzione) implica che la funzione di autocovarianza della serie $Cov[t_1, t_2]$ soddisfa la relazione:

$$\begin{aligned} Cov[t_1, t_2] &= E[(x_{t_1} - \mu_{t_1})(x_{t_2} - \mu_{t_2})] = E[(x_{t_1+h} - \mu)(x_{t_2+h} - \mu)] = \\ &= Cov[t_1 + h, t_2 + h] \end{aligned} \quad \text{eq. 11.1.45}$$

per ogni valore di t_1 , t_2 e h . Questo risultato può essere interpretato dicendo che la funzione di autocovarianza del processo dipende solo dalla distanza sull'asse dei tempi tra t_1 e t_2 , e non dallo specifico istante temporale (Shumway e Stoffer, pp. 23-24).

In altre parole, la stazionarietà della distribuzione implica la stazionarietà dei momenti (media, varianza e covarianza), i quali sono degli indicatori di proprietà della distribuzione stessa.

Il concetto di stazionarietà in senso forte è spesso troppo restrittivo per le applicazioni, ed è difficile da verificare su una singola serie osservata. Pertanto, invece di imporre delle condizioni su tutte le possibili distribuzioni, ci si limita a condizioni sui primi due momenti della serie, pervenendo ad una definizione di stazionarietà detta "in senso debole". Una serie *stazionaria in senso debole* è un processo a varianza finita tale che: (1) la funzione valore medio, $\mu_t = E[x_t]$, è costante e non dipende dal tempo t ; (2) la funzione di covarianza $Cov[t_1, t_2]$ dipende dagli istanti temporali t_1 e t_2 solo tramite la differenza $|t_1 - t_2|$. Per semplicità, il termine *stazionario* può essere usato per indicare la stazionarietà in senso debole (a cui ci si riferisce con maggiore frequenza nelle applicazioni).

Chiaramente una serie stazionaria in senso forte è necessariamente stazionaria, mentre non vale in generale il viceversa, tranne per processi Gaussiani, caratterizzati da distribuzioni Gaussiane multivariate completamente definite dal vettore delle medie e dalla matrice di covarianza: in tale caso la stazionarietà dei primi due momenti implica la stazionarietà delle distribuzioni di ogni possibile sotto-campione.

Poiché la funzione media, $\mu_t = E[x_t]$, di un processo stazionario è indipendente dal tempo t , è possibile scrivere $\mu_t = \mu$, e semplificare la notazione per la covarianza come segue:

$$Cov[t+h, t] = E[(x_{t+h} - \mu)(x_t - \mu)] = E[(x_h - \mu)(x_0 - \mu)] = Cov[h, 0] \quad \text{eq. 11.1.46}$$

La precedente non dipende da t , richiede l'ipotesi di varianza finita $Var[x_t] = Cov[0, 0] < \infty$, e, per semplicità, si scrive con $Cov[h, 0] = Cov[h]$. Da quanto precede derivano le definizioni di seguito riportate.

La *funzione di autocovarianza di una serie temporale stazionaria* è definita come:

$$Cov[h] = E[(x_{t+h} - \mu)(x_t - \mu)] \quad \text{eq. 11.1.47}$$

la quale è simmetrica rispetto all'origine, $Cov[h] = Cov[-h]$.

Dalla relazione eq. 11.1.47 si definisce la *funzione di autocorrelazione (ACF) di una serie temporale stazionaria* come il rapporto dell'autocovarianza e della varianza del processo:

$$\rho[h] = \frac{Cov[t+h, t]}{\sqrt{Cov[t+h, t+h]Cov[t, t]}} = \frac{Cov[t+h, t]}{\sqrt{Var[t+h]Var[t]}} = \frac{Cov[h]}{Cov[0]} \quad \text{eq. 11.1.48}$$

e per la disuguaglianza di Cauchy-Schwarz, $-1 < \rho[h] < 1$ per ogni h (Shumway e Stoffer; pp. 24-25).

La funzione di autocorrelazione è utile per dedurre informazioni sulla dipendenza temporale presente nelle serie. La Figura 11.9 illustra tre serie con diverse strutture di dipendenza temporale, la cui funzione generatrice è riportata in forma analitica in testa a ciascun grafico, e le relative funzioni di autocorrelazione.

La prima serie rappresenta un segnale ad andamento periodico e l'ACF mostra in modo chiaro che il periodo del segnale è 50, come riportato nell'espressione analitica. I valori positivi dell'ACF a lag 50 indicano che osservazioni separate da 50 unità di tempo tendono ad assumere valori analoghi. Al contrario, osservazioni separate da 25 unità di tempo tendono ad assumere valori discordanti a causa dell'andamento ciclico. L'ACF rappresenta dunque uno strumento efficace per evidenziare componenti periodiche nella serie.

La seconda serie si riferisce ad un segnale detto autoregressivo, in cui l'osservazione al generico tempo t è legata al valore assunto dalla serie all'istante $t-1$ tramite un coefficiente pari a 0.9. In questo caso l'ACF decade in modo approssimativamente esponenziale, indicando che i valori assunti dalla serie sono legati tra loro tramite la relazione $x_t = 0.9x_{t-1}$ fino ad un certo lag, ossia i valori al tempo t

preservano memoria dei valori ai tempi precedenti sino ad un certo intervallo, che nel caso specifico è di circa 20-30 istanti temporali.

La terza serie è descritta da una funzione simile alla precedente a meno del segno del coefficiente che lega due osservazioni successive x_{t-1} e x_t . La relazione e il grafico della serie indicano che a valori x_{t-1} tendono a corrispondere all'istante successivo valori x_t discordanti, da cui segue la tendenza ad un continuo cambio di segno tra osservazioni successive. L'ACF evidenzia in modo chiaro questa tendenza tramite il continuo cambiamento di segno, nonché il progressivo decadimento dell'influenza delle osservazioni precedenti su quelle successive al crescere del lag che le separa.

Infine, la quarta serie rappresenta un segnale le cui osservazioni hanno autocorrelazione nulla, media uguale a zero e varianza costante, ossia un cosiddetto "white noise". In questo caso, l'auto correlazione è prossima a zero.

La funzione di autocorrelazione può essere calcolata per ogni valore del lag h , tuttavia, data un serie di numerosità N , il calcolo dell'ACF a lag h corrisponde al calcolo della correlazione tra le sottoserie $\{x_1, \dots, x_{t+k}\}$ e $\{x_{1+k}, \dots, x_N\}$, la cui numerosità decresce al crescere di h . Ne segue che per ottenere valori affidabili dell'ACF è opportuno limitare il calcolo ad $h = N/4$ (Hipel e McLeod 1994; p. 72). Per processi privi di correlazione, la stima dell'ACF ai vari lag oscilla intorno allo zero con distribuzione approssimativamente Gaussiana con media zero e varianza approssimata $1/N$ (Hipel e McLeod 1994; p. 72). Questa proprietà permette di definire degli intervalli di confidenza approssimati intorno al valore zero dell'ACF stimata su un campione. Ad esempio, l'intervallo di confidenza approssimato al 95% nell'ipotesi che il processo sia decorrelato è definito dai due limiti $\pm 1.96/\sqrt{N}$, in cui 1.96 è il 0.975-esimo quantile della distribuzione Gaussiana standard. Le linee blu riportate in Figura 11.9 rappresentano questi limiti: se il campione proviene da un processo decorrelato, su 100 valori di lag (ad esempio, da 1 a 100), il valore dell'ACF stimato dovrebbe ricadere all'esterno della fascia individuata dai due limiti in media 5 volte. L'ACF della serie white noise (la cui correlazione teorica è nulla) illustra questo comportamento.

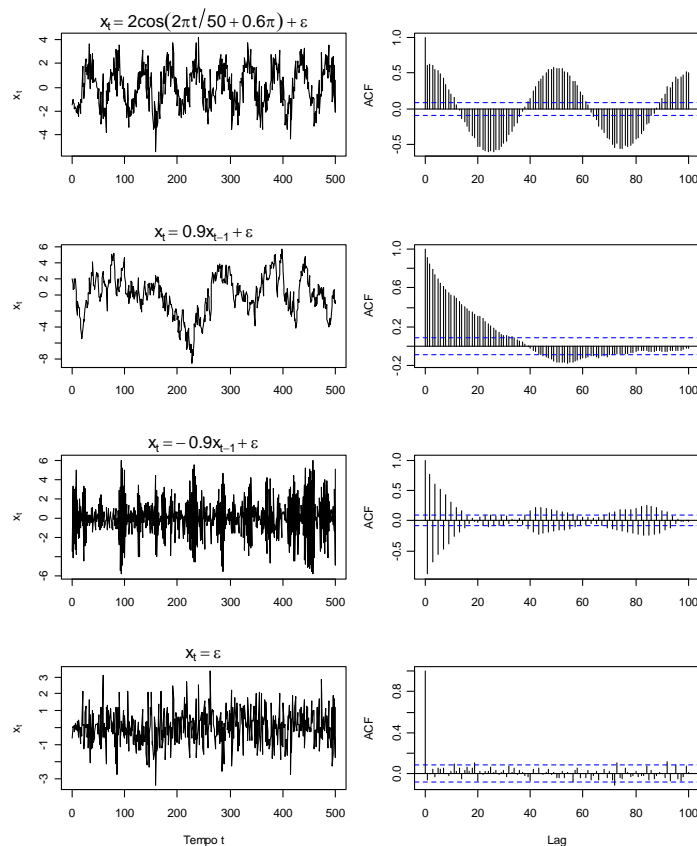


Figura 11.9 - Esempi di serie temporali con diverse forme di dipendenza temporale e relative ACF. Il simbolo ε indica un termine casuale con distribuzione normale standard. Le linee tratteggiate blu nei grafici delle ACF indicano gli intervalli di confidenza al 95% dei coefficienti di autocorrelazione nell'ipotesi di autocorrelazione nulla.

11.1.5 Test delle ipotesi

Un test delle ipotesi è una procedura il cui obiettivo è definire se un'ipotesi (ad esempio, la presenza di una proprietà in un campione osservato) può essere ritenuta valida con un certo grado di fiducia, quantificato tramite un valore di probabilità.

Prima di introdurre la struttura di un generico test è opportuno evidenziare alcuni aspetti. In primo luogo, come sempre avviene in ambito sperimentale, la natura delle conoscenze che si presuppongono note determina anche le metodologie statistiche che si adottano per l'analisi. In particolare, dato un campione proveniente da una variabile casuale X , si può derivare un test delle ipotesi supponendo nota la distribuzione di X , per cui l'inferenza si riferisce ai soli parametri che la specificano. In questo caso il test è detto parametrico. D'altra parte, si può derivare un test senza introdurre assunzioni stringenti sulla forma analitica della distribuzione di X . In questo secondo caso il test è detto non-parametrico (distribution-free). La scelta fra i due approcci dipende dal grado di conoscenza della distribuzione da cui è estratto il campione a disposizione.

In secondo luogo, un test non è concepito per provare o confutare un'ipotesi, ma semplicemente per mostrare se un assunto deve essere scartato in quanto associato ad un valore di probabilità tale da non poter essere accettato, ossia ad un livello di credibilità o fiducia più piccolo di un valore minimo ritenuto accettabile.

Infine, occorre osservare che l'ipotesi che si sottopone al test è quella che non implica cambiamento. Ad esempio, si supponga che un corso d'acqua sia stato oggetto di interventi di sistemazione, e si voglia verificare se tali interventi abbiano generato un cambiamento statisticamente significativo della portata media misurata in una stazione a valle delle opere. Assumendo che le serie delle osservazioni antecedenti e successive agli interventi siano disponibili, l'ipotesi sottoposta a test è che la portata media non sia cambiata, ossia che la portata media ante operam sia uguale a quella post operam. Analogamente, nel caso in cui si voglia testare se due serie sono correlate, si sottopone a test l'ipotesi che non ci sia correlazione.

Questo approccio al problema è legato al fatto che l'ipotesi che non implica cambiamento è generalmente semplice e ciò agevola la definizione degli elementi necessari per l'applicazione della procedura descritta nel seguito. Di contro, l'ipotesi che implica un cambiamento è associata a molteplici possibilità: ad esempio testare che due medie μ_1 e μ_2 siano diverse può voler dire verificare se $\mu_1 - \mu_2 > 0$ oppure $\mu_1 - \mu_2 < 0$ oppure $|\mu_1 - \mu_2| > 0$. Ciò comporta una maggiore difficoltà nel definire le grandezze richieste dalla procedura di verifica.

Dalle osservazioni precedenti segue che l'assunto sottoposto a test è convenzionalmente definito come ipotesi nulla H_0 . La procedura generale per un test delle ipotesi può essere riassunta in cinque passi:

Passo 1

Formulare il problema pratico in termini di ipotesi. Questo passaggio, in alcuni casi, può non essere semplice. È opportuno focalizzare quella che è comunemente definita ipotesi alternativa H_1 , poiché essa è la più importante da un punto di vista pratico, in quanto esprime lo scenario di situazioni che si desidera che il test sia in grado di diagnosticare. L'ipotesi nulla deve essere molto semplice e rappresentare lo stato di fatto o più precisamente che non c'è differenza tra i processi che si stanno testando. In altre parole l'ipotesi nulla rappresenta lo standard o la situazione di controllo con cui confrontare le evidenze a sostegno dell'ipotesi alternativa. Ad esempio, dati due campioni di una stessa variabile, si voglia testare se la differenza delle rispettive medie μ_1 e μ_2 è dovuta alle variazioni di campionamento (casualità) oppure è tale da lasciare presupporre una causa. In questo caso, l'ipotesi nulla è che le medie sono uguali (ossia le differenze sono casuali) $H_0: \mu_1 - \mu_2 = 0$, mentre $H_1: \mu_1 - \mu_2 \neq 0$ (potendo essere μ_1 maggiore o minore di μ_2).

Passo 2

Definire e calcolare una statistica test T , ossia una funzione che dipende esclusivamente dal campione e non contiene nessun parametro incognito. Una statistica test dovrebbe presentare due proprietà:

- (1) dovrebbe comportarsi in modo diverso quando H_0 è vera rispetto a quando è vera H_1 ;
- (2) la sua distribuzione dovrebbe essere nota nella circostanza che H_0 è vera.

Passo 3

Scegliere una regione critica, ossia il tipo di valori di T che meglio evidenziano che H_1 è vera contro l'ipotesi che H_0 è vera. Le regioni critiche possono essere di tre tipi:

- (1) di destra (right-sided), tale che H_0 è rigettata se la statistica test è maggiore o uguale a un fissato valore critico;

(2) di sinistra (left-sided), tale che H_0 è rigettata se la statistica test è minore o uguale a un fissato valore critico; bilaterale (two-sided) tale che H_0 è rigettata se la statistica test è minore o uguale a un valore critico sinistro o maggiore o uguale a un valore critico destro.

Un valore di T che ricade all'interno di un'opportuna regione critica conduce a rigettare H_0 in favore di H_1 . Se il valore di T ricade al di fuori della regione critica, H_0 non è rigettata. La conclusione del test non dovrebbe mai essere che H_0 è accettata, poiché un test può mostrare soltanto che non ci sono evidenze per il rigetto, ma non fornisce informazioni riguardo all'esclusiva validità di H_0 .

Passo 4

Scegliere l'ampiezza della regione critica, ossia specificare il valore del rischio che si è disposti a correre nel caso in cui il test porti ad una conclusione errata. Si definisce livello di significatività del test, indicandolo con α , il valore del rischio che si assume nel rigettare l'ipotesi H_0 quando questa è vera. Questo errore è comunemente definito del I tipo (type I error). In relazione alla natura della conseguenze che si hanno nel commettere questo tipo di errore, il valore di α è fissato tra 0.01 e 0.10.

Passo 5

Sebbene sia sufficiente definire se il valore di T ricade all'interno o all'esterno della regione critica, è utile tuttavia considerare in quale zona della regione critica ricade T . Se T si trova in prossimità dei limiti (valori critici) della regione critica, sussiste una moderata evidenza che H_0 debba essere rigettata. Se invece T è distante dai limiti della regione critica, allora l'evidenza è più marcata. In altre parole, il reale livello di significatività associato a T sotto l'ipotesi nulla, detto p -valore (p -value), può fornire un'utile informazione oltre al fatto che T ricada nella zona critica.

Il concetto di robustezza introdotto in relazione alle proprietà di media e mediana, è applicabile anche ai test delle ipotesi o a una qualunque procedura e indica che la procedura è insensibile o poco sensibile al mancato rispetto delle assunzioni che sono alla sua base.

Come evidenziato all'inizio del paragrafo, i test parametrici si basano sull'ipotesi che il campione sia estratto da una distribuzione di cui si assume nota la forma analitica. Nei test parametrici, la conoscenza della distribuzione della statistica test T nell'ipotesi nulla è dunque legata alla conoscenza della distribuzione di X . Allorché la distribuzione di X non fosse nota, tali test non potrebbero essere applicati. Tuttavia, in alcuni casi, deviazioni dalle assunzioni di base hanno un'influenza moderata sui test, indicando che tali test sono robusti.

Ad esempio, se un test si basa sull'ipotesi che X abbia una distribuzione Gaussiana ed è robusto, allora esso può essere applicato anche se la distribuzione del campione osservato mostra moderati scostamenti dalla distribuzione Gaussiana (e.g., presenza di moderata asimmetria).

In genere, nei test non parametrici si impongono invece delle assunzioni minimali sulla distribuzione di X : ad esempio se è continua o discreta, se è simmetrica, se assume valori in un intervallo di ampiezza finita o infinita. Poiché queste ipotesi sono meno stringenti rispetto alla definizione di una specifica distribuzione, i test non parametrici sono generalmente più robusti di quelli parametrici. Tuttavia, quando l'ipotesi nulla è falsa, i test non parametrici tendono a rigettarla correttamente con una frequenza inferiore rispetto alle procedure parametriche, le quali, proprio perché associate a particolari modelli probabilistici, presentano una maggiore capacità di discriminazione quando H_0 è falsa. Per ulteriori approfondimenti si rimanda a Piccolo (2004; Cap. 14, pp. 355-405).

Un elenco esteso di test utili in una varietà di applicazioni è fornito in Kanji (2006).

11.1.6 Tecniche di ricampionamento di tipo "bootstrap"

Una notevole flessibilità nella scelta della statistica test è fornita dalle tecniche di ricampionamento, di cui la più nota è probabilmente quella definita *bootstrap* (e.g., Efron e Tibshirani 1993; Davison e Hinkley 1997). La logica del metodo *bootstrap* è quella di costruire dei campioni non osservati, ma statisticamente simili a quelli osservati, ricampionando la serie osservata tramite un procedimento di estrazione con reinserimento delle osservazioni. La procedura è analoga all'estrazione di un numero da un'urna, con successivo reinserimento del numero prima della successiva estrazione. Una volta scelta una statistica test, la si calcola sia sul campione osservato che su un numero grande N (e.g., $N = 1000$) di campioni della stessa numerosità di quello osservato, ottenuti tramite il ricampionamento (detti *campioni bootstrap*). Gli N valori della statistica test così ottenuti permettono di definire la distribuzione campionaria (ossia la distribuzione empirica) della statistica scelta. Poiché i campioni *bootstrap* derivano da un processo di estrazione casuale con reinserimento dalla serie originale,

l'eventuale struttura di correlazione temporale della serie osservata non viene conservata. Ne segue che i campioni *bootstrap* hanno proprietà analoghe al campione osservato (e.g., i momenti), ma rispettano, almeno approssimativamente, l'ipotesi di indipendenza. Questo li rende adatti al calcolo delle distribuzioni delle statistiche test assumendo come ipotesi nulla l'assenza di *trend*, *change point*, o di un generico andamento temporale di tipo sistematico. Una volta nota la distribuzione campionaria della generica statistica test sotto l'ipotesi nulla, è possibile confrontare il valore della statistica stessa calcolato sul campione osservato con i quantili, ad esempio uguali a 0.95 e 0.90, dedotti dalla distribuzione campionaria, e verificare se il valore cade nelle regioni critiche con livello di significatività pari al 5 e 10%. In alternativa si può definire la percentuale di volte che il valore della statistica calcolato sul campione osservato è superato dai valori provenienti dagli N campioni, ossia il p -value relativo alla statistica per il campione osservato, e controllare quanto questa percentuale è distante dai livelli di significatività comunemente adottati del 5 e 10%.

11.2 Analisi di non stazionarietà

Come descritto nel paragrafo 11.1 una serie si può assumere stazionaria se la media e la varianza risultano essere invarianti nella serie stessa (stazionaria in senso debole). In pratica, considerando due finestre temporali all'interno della serie e stimate al loro interno la media e la varianza i valori ottenuti dovrebbero essere simili. Si intuisce come sia complesso identificare un comportamento non stazionario di una serie. Infatti è immediato il problema di quanto dovrebbero essere ampie le finestre temporali e di quanto dovrebbero essere simili tali stime per asserire che una serie sia caratterizzata da una variabilità sistematica (e quindi non stazionaria). I test proposti in letteratura aiutano a capire il comportamento, ma è intuitivo che per alcuni casi non riescono a fornire risultati assoluti o non privi di una successiva interpretazione soggettiva.

La variabilità sistematica può essere ciclica, monotona, repentina (*change points*), o più complessa. Nella Figura 11.10 sono riportati alcuni esempi di tipologie di variabilità riscontrabili in una serie temporale.

In letteratura emerge una sostanziale coerenza nelle tecniche usate per lo studio della variabilità sistematica di serie storiche di natura ambientale, quali precipitazioni, temperature, deflussi, indici climatici, ecc. (Peterson 2005). Gran parte dei lavori si collocano nell'ambito dello studio dei cambiamenti climatici (e.g., Brunetti et al. 2000, 2001a-b, 2002, 2006; Xu et al. 2003; Xiong e Guo 2004; Yue e Pilon 2004; Aksoy et al. 2008; Rusticucci e Renom 2008).

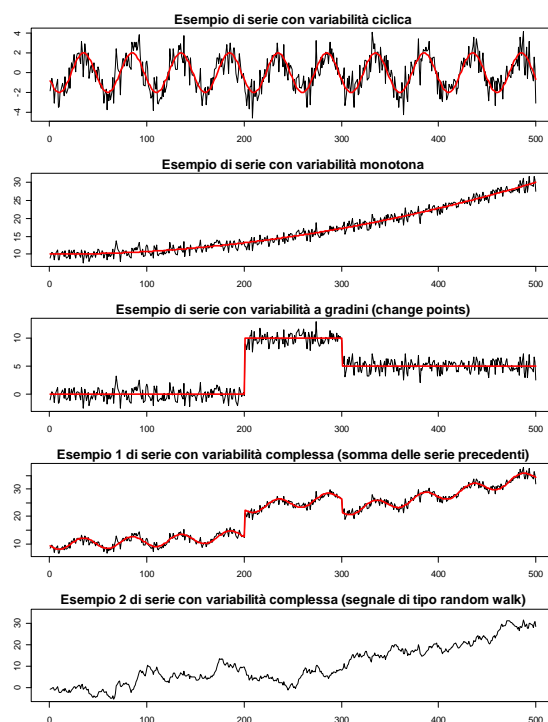


Figura 11.10 - Esempi di tipologie di variabilità sistematica che possono essere presenti in una serie temporale.

Nel seguito si riporta una sintesi dei principali metodi utilizzati per l'analisi della non-stazionarietà in serie temporali.

11.2.1 Aspetti generali

L'individuazione di trend in serie storiche di grandezze climatiche, ambientali e idrologiche è di fondamentale importanza, poiché le procedure di progettazione, pianificazione e gestione sono comunemente basate sull'ipotesi di stazionarietà dei processi di interesse.

Un trend climatologico può essere dovuto ad una variabilità intrinseca riconducibile a cambiamenti climatici, ossia a variazioni di lungo termine nel regime climatico. Esso può interessare i valori medi, la variabilità (varianza, valori estremi, persistenza) e la distribuzione intra-annuale (stagionalità). Cambiamenti graduali sono, in genere, legati a fenomeni evolutivi come urbanizzazione, deforestazione, ecc.. Di contro, cambiamenti repentini sono associati ad interventi antropici di grande impatto (e.g., costruzione di invasi), ovvero a cambiamenti e variazioni climatiche con impatto spiccatamente non lineare sulle quantità osservate.

Le fasi fondamentali attraverso le quali si conduce un processo di analisi di una serie sono:

- analisi preliminare dei dati;
- applicazione di test statistici formali;
- interpretazione dei risultati.

11.2.2 Analisi preliminare dei dati

Le serie possono essere di natura molto diversa, essere registrate a diverse scale di risoluzione temporale ed essere aggregate a diverse scale spaziali. Lo studio dei *trend* monotoni e dei *change point* è influenzato da questi fattori nonché dalla presenza di intervalli con dati mancanti, da stagionalità e trend di breve periodo, dalla mancanza di omogeneità (cambiamento dello strumento e/o delle unità di misura). Questi aspetti devono essere considerati preventivamente tramite un accurato controllo della qualità del dato, supportato dai corrispondenti metadati, e dall'eventuale trattamento delle serie allo scopo di ottenere dati adatti alle analisi (ad esempio, riempimento degli intervalli di dati mancanti).

Un fattore di notevole impatto è la presenza di correlazione temporale nelle serie. La correlazione indica, infatti, che le osservazioni ad un tempo t influiscono su quelle relative ad un istante temporale successivo. Questo legame implica che un'osservazione ad un generico istante contiene una parte di informazione relativa alle osservazioni registrate ai tempi precedenti. In altre parole, ogni osservazione contiene una parte di informazione già contenuta in altre osservazioni. Questa ridondanza di informazione non si verifica per serie prive di correlazione, in cui ogni osservazione contiene un'informazione che non è presente in nessuna delle altre. Da un punto di vista operativo, queste considerazioni si traducono nel fatto che serie correlate contengono meno informazione statistica rispetto a serie non autocorrelate di uguale numerosità.

I test usualmente impiegati nello studio dei trend richiedono l'ipotesi di indipendenza, per cui occorre adottare particolari accorgimenti per tener conto dell'eventuale autocorrelazione della serie in esame, al fine di evitare conclusioni errate dovute ad un'errata valutazione dell'informazione realmente disponibile. L'influenza della correlazione può essere limitata riducendo la frequenza delle osservazioni (aggregazione), usando misure sintetiche (medie, massimi annuali, ecc.), correggendo i valori critici teorici dei test tramite pre-whitening (Yue et al 2002)⁸, o deducendo i valori critici tramite tecniche di ricampionamento (e.g., *bootstrap*; Yue e Pilon, 2004). Quest'ultima, di cui vedremo un'applicazione nel paragrafo successivo, consente di ottenere, a partire dalla serie osservata, campioni tramite estrazione con reinserimento. Tali campioni non preservano per definizione le eventuali variabilità sistematiche presenti nella serie originaria. Ciò permette di calcolare le statistiche test su un numero di campioni che in linea di principio soddisfano l'ipotesi nulla (assenza di andamenti sistematici).

⁸ Una procedura di pre-whitening è così definita in quanto il suo obiettivo è restituire un segnale privo di autocorrelazione con media nulla e varianza finita, detto "white noise" (rumore bianco) (Shumway e Stoffer 2006, p. 12).

Inoltre, alcune proprietà dei dati che non rispettano le ipotesi richieste dai test possono essere rimosse tramite opportune trasformazioni. Infatti, per l'applicazione di alcuni test statistici è necessario che i dati seguano una distribuzione Gaussiana. Ci si può ricondurre a tale distribuzione tramite l'applicazione di alcune trasformazioni, quali:

la trasformazione di Box-Cox (Box e Cox 1964):

$$x_i^{(\lambda)} = \begin{cases} \frac{x_i^\lambda - 1}{\lambda} & \text{se } \lambda \neq 0 \\ \log x_i & \text{se } \lambda = 0 \end{cases} \quad \text{per } i = 1, \dots, N \quad \text{eq. 11.2.1}$$

in cui λ è un parametro che deve essere stimato per ottenere una soddisfacente approssimazione della distribuzione di $x_i^{(\lambda)}$ alla distribuzione Gaussiana;

la trasformazione del quantile normale canonico, detto anche "normal score" (NS; Robson et al (2000)):

$$x_i^{(NS)} = \Phi^{-1}(F_X(x_i)) \quad \text{per } i = 1, \dots, N \quad \text{eq. 11.2.2}$$

in cui F_X è la distribuzione delle osservazioni (teorica o empirica) e Φ^{-1} è la funzione inversa della distribuzione cumulata Gaussiana standard.

11.2.3 Applicazione dei test statistici

L'applicazione dei test si svolge attraverso le seguenti fasi:

- scelta della variabile in base agli obiettivi di interesse;
- scelta del tipo di cambiamento che si intende studiare (ciclico, graduale, ecc.);
- controllo delle ipotesi richieste per l'applicabilità dei test di stazionarietà;
- selezione dei test di stazionarietà (è opportuno scegliere più test, ognuno basato su un principio diverso, in modo da avere una cross-validazione dei risultati);
- valutazione dei livelli di significatività;
- interpretazione dei risultati.

Alcuni tra i test usati frequentemente in letteratura per l'analisi dei *trend* graduali e dei *change point* (e.g., Robson et al 2000) rappresentano un buon compromesso tra semplicità e *performance*. Questi aspetti uniti al criterio di adottare metodi basati su principi di diversa natura (parametrici, non parametrici, *bootstrap*, ecc.) conducono a suggerire l'adozione dei seguenti test, la cui formulazione sarà descritta in dettaglio nel paragrafo 11.4.3:

Test per i cambiamenti repentini (*change point*)

Per la verifica della presenza di *change point* è possibile applicare il test di *Pettitt* e il test CUSUM. Entrambi i test sono basati sui ranghi e sono concepiti per individuare dei salti repentini nelle misure di tendenza centrale di una serie, ossia variazioni della media o mediana. Inoltre non richiedono la conoscenza a priori della collocazione cronologica del *change point*. Il test CUSUM richiede l'applicazione del *bootstrap* per la definizione della distribuzione della statistica test sotto l'ipotesi nulla, mentre il test di *Pettitt* si avvale di risultati analitici per la definizione della regione critica.

Test per il trend graduale

Per la verifica delle variazioni graduali è possibile applicare il test della regressione lineare basato sul coefficiente di correlazione ρ di Pearson, il test di Mann-Kendall ed il test per il coefficiente di correlazione ρ_s di Spearman. Il test di Pearson è un test parametrico per la verifica di trend lineari che richiede la normalità dei dati, mentre i test di Mann-Kendall e Spearman sono test non parametrici per la verifica di trend monotoni (lineari e non lineari).

11.2.4 Interpretazione dei risultati

Nell'interpretazione di un test è necessario ricordare che un livello di confidenza del 5% denota che, 1 volta su 20 il test fallirà rigettando l'ipotesi nulla qualora questa sia vera. I risultati vanno confrontati con le evidenze visive e il più possibile con metadati relativi alla stazione di misura (cambi di strumento, urbanizzazione, interventi antropici di varia natura). È fondamentale distinguere la variabilità meteorologica, la cui estensione è limitata nel tempo, da quella imputabile ai cambiamenti climatici, la cui persistenza si protrae per intervalli temporali molto ampi. Un'analisi dei cambiamenti climatici richiede serie di notevole lunghezza, possibilmente superiore ai 50 anni. Occorre inoltre tener conto del fatto che una variabilità molto accentuata può coprire delle tendenze di lungo periodo, e che trend apparenti possono rappresentare le fasi ascendenti o discendenti di forzanti climatiche cicliche il cui periodo è superiore alla lunghezza della serie osservata. Infine, non appare superfluo ricordare che trend statisticamente significativi possono non esserlo da un punto di vista fisico. È sempre necessario, pertanto, avere una conoscenza adeguata del fenomeno fisico studiato, e in questo senso è auspicabile adottare un approccio interdisciplinare che coinvolga soggetti esperti nelle diverse aree, qualora l'analista non abbia in sé tutte le competenze necessarie ad un adeguato grado di approfondimento.

11.2.5 Stagionalità

Per un approfondimento delle procedure disponibili in letteratura per il calcolo delle componenti stagionali relative ai primi due momenti di una serie storica si rimanda al paragrafo 11.4.2. Tra i metodi disponibili in letteratura, una procedura non-parametrica particolarmente adatta a serie con scala di aggregazione temporale uguale o inferiore al giorno, per le quali la presenza di outliers può influire sulla definizione delle componenti stagionali, è la *Seasonal Trend decomposition based on LOESS (STL)*; Cleveland et al., 1990). La procedura *STL* permette la decomposizione di una serie storica secondo lo schema classico di tipo additivo (e.g., Chatfield 2000). Più precisamente, al variare del tempo t una variabile ambientale x_t è rappresentata tramite la somma di quattro componenti Figura 11.11, ognuna associata ad un diverso tipo di variabilità, ossia:

$$x_t = S_t + T_t + C_t + R_t \quad \text{eq. 11.2.3}$$

in cui:

- S_t rappresenta la variabilità stagionale, generalmente annuale, che emerge in serie misurate a scala giornaliera, mensile, stagionale o comunque sub-annuale (si pensi, e.g., alla ciclicità annuale che emerge dalle serie di portata di un fiume o dalle temperature);
- T_t è il trend, ossia un tipo di variazione che è presente quando la serie mostra un andamento crescente o decrescente per periodi lunghi e non riconducibili agli intervalli crescenti o decrescenti delle fluttuazioni stagionali;
- C_t rappresenta le variazioni cicliche a scala diversa da quella annuale, quali, ad esempio, i cicli solari con periodo 11 anni, o fenomeni, come *El Niño Southern Oscillation (ENSO)*;
- R_t è il termine che descrive le fluttuazioni irregolari residue dopo la rimozione dei termini sistematici S_t , T_t e C_t . Se questi ultimi descrivono accuratamente le variazioni sistematiche, allora R_t rappresenta un residuo puramente casuale. Il termine R_t può tuttavia contenere ancora un'informazione sistematica residua nel caso in cui gli altri termini non descrivono compiutamente il comportamento della serie legato, ad esempio, a cause fisiche o antropiche.

La struttura dell'algoritmo di decomposizione adottato dalla procedura *STL* permette l'estrazione della componente stagionale della media che è variabile di anno in anno durante il periodo di osservazione. In idrologia è invece prevalente l'interesse per la stima di una componente stagionale periodica che sia rappresentativa della variabilità stagionale della serie completa, e dunque uguale per tutti gli anni di osservazione. Inoltre, assume importanza la definizione della variabilità stagionale della varianza della serie, che misura l'entità delle oscillazioni intorno all'andamento medio della serie (definito dalla componente stagionale della media). Come mostrato da Grimaldi (2004) il calcolo delle componenti stagionali può essere eseguito indipendentemente dalla definizione delle componenti di tendenza, la cui stima in molti casi può non essere necessaria. Il metodo per la stima delle componenti stagionali sarà descritto in dettaglio nel paragrafo 11.4.

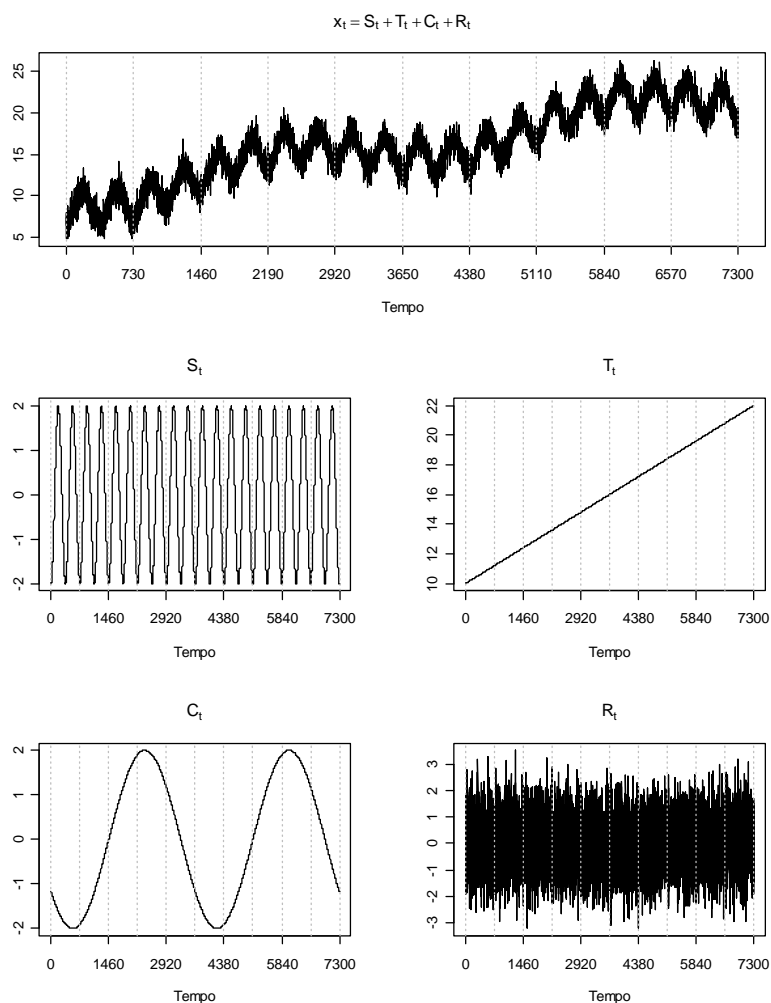


Figura 11.11 - Esempio di decomposizione additiva (eq. 11.2.3).

11.3 Analisi di probabilità di dati idrologici

Il verificarsi di molti fenomeni idrologici estremi non può essere previsto sulla base di assunzioni deterministiche con un anticipo e un'affidabilità compatibili con le decisioni e gli interventi (piani di allerta, allarme, intervento) inerenti al loro accadimento (WMO, 1994). In questi casi, si ricorre ad un approccio probabilistico che consente di tener conto nel processo decisionale dell'incertezza del fenomeno. Qualora gli eventi possano essere assunti indipendenti, allora la probabilità associata ad un evento o ad una combinazione di eventi può essere dedotta tramite un'analisi delle frequenze di accadimento. Le grandezze idrologiche comunemente descritte tramite questo tipo di analisi sono le precipitazioni (intense), le portate (massime e minime relative a definiti intervalli temporali) e le temperature.

L'analisi delle frequenze di accadimento si avvale di strumenti sia grafici che analitici. I primi ricorrono a grafici diagnostici per visualizzare, secondo molteplici criteri, la relazione esistente tra le osservazioni della variabile di interesse e le corrispondenti frequenze di superamento. I punti corrispondenti alle coppie osservazione-frequenza sono quindi interpolate tramite delle curve parametriche per ottenere le frequenze di accadimento di eventi non osservati. L'approccio analitico si basa, invece, sull'assunzione di uno specifico modello matematico, detto distribuzione di probabilità, allo scopo di definire, in modo analitico, l'equivalente della curva interpolante usata nel metodo grafico. In genere, il modello matematico è caratterizzato tramite dei parametri la cui stima (basata sul campione osservato) consente di adattare il modello ai dati osservati secondo opportuni criteri che verranno discussi nel seguito.

11.3.1 Distribuzioni di probabilità

Le distribuzioni di probabilità sono usate in un'ampia varietà di studi idrologici, quali quelli sulle risorse idriche, sui valori estremi di portata (massimi e minimi), sulla siccità, sui volumi dei serbatoi, sulle grandezze di pioggia (intensità, precipitazione cumulata, ecc.) e, in generale, in modelli per l'analisi di serie temporali.

I valori totali annui di grandezze, come il volume di deflusso o l'altezza di pioggia, tendono ad essere distribuiti approssimativamente secondo una distribuzione normale. I valori totali mensili e settimanali mostrano in genere un'asimmetria definita (principalmente positiva; si veda paragrafo 11.1.3.4), per cui la loro distribuzione può essere descritta da modelli con una coda (in genere la destra) più lunga dell'altra. Le serie degli estremi annuali (massimi e minimi), e dei valori estremi sopra o sotto una soglia hanno in genere una distribuzione con evidente asimmetria positiva. La parte del campione concentrata in prossimità del valore medio può essere spesso descritta da diverse distribuzioni. D'altra parte, le distribuzioni possono differire significativamente l'una dall'altra per le alte probabilità di non superamento (eventi rari). Poiché i valori usati per la gestione, la pianificazione e la progettazione sono basati sulla stima degli eventi rari e dunque più gravosi, è importante essere in grado di determinare tali eventi nel modo più accurato possibile, associandovi anche un'informazione relativa all'incertezza della valutazione.

Quando l'interesse è focalizzato sui quantili estremi, la scelta della distribuzione deve essere particolarmente appropriata. Uno strumento per guidare la scelta è quello fornito dalla classificazione delle distribuzioni proposta da Al Eldouni et al. (2008) e basata sulle proprietà delle code. Gli stessi autori introducono dei metodi grafici per selezionare la classe più appropriata per la descrizione del campione in termini di comportamento di coda. In generale, alla luce della varietà di distribuzioni disponibili, è opportuno che la scelta sia guidata, in un primo momento, dalla natura della grandezza che si vuole descrivere e dalla conformità delle basi teoriche del modello a tale natura (e.g., per descrivere una serie di massimi annuali, la scelta privilegerà, in una fase iniziale, distribuzioni degli eventi estremi). Accanto a considerazioni concettuali, che conducono a restringere il numero dei modelli potenzialmente adatti a descrivere il campione, si deve necessariamente valutare la bontà dell'adattamento della distribuzione teorica (modello parametrico) a quella empirica tramite grafici diagnostici e test numerici. Occorre sottolineare che tali procedure valutano l'adattamento assumendo un particolare criterio, quale, ad esempio, l'adattamento nella parte centrale della distribuzione, l'adattamento nelle code, la minimizzazione di una distanza, ecc.. L'uso di più test o la scelta di un particolare criterio deve essere sempre guidata dallo scopo dell'analisi e principalmente dalla proprietà che si vuole riprodurre nel modo più accurato.

11.3.2 Metodi di stima di parametri

La stima dei parametri di una distribuzione può essere condotta tramite diversi metodi. I metodi più usati sono:

- il metodo dei momenti
- il metodo della massima verosimiglianza
- il metodo degli L-momenti.

Il metodo della massima verosimiglianza

è considerato efficiente in quanto restituisce la minore varianza dei parametri stimati rispetto ad altri metodi, ma ha lo svantaggio di restituire, a volte, stime distorte, ossia il valore atteso dello stimatore è diverso da quello del parametro che si desidera stimare (e.g., Martins e Stedinger 2000). Tale metodo è basato sull'espressione della densità di probabilità ed è disponibile una vasta letteratura a supporto di risultati asintotici e approssimati per la definizione di intervalli di confidenza e test di significatività dei parametri (e.g., Coles (2001), e riferimenti ivi contenuti).

Il metodo dei momenti

è probabilmente il più noto e più usato metodo di stima dei parametri di una distribuzione grazie alla semplicità di applicazione e alla possibilità di interpretare i parametri in termini di proprietà geometriche della distribuzione. Tuttavia, le stime ottenute con il metodo dei momenti sono meno efficienti di quelle dedotte con il metodo della massima verosimiglianza, soprattutto per distribuzioni con 3 o più parametri, poiché i valori dei momenti di ordine elevato sono spesso distorti nel caso di campioni limitati. Inoltre, per alcune distribuzioni, i momenti di ordine superiore (≥ 2) possono non essere definiti, rendendo inapplicabile il metodo.

Il metodo degli L-momenti

formalizzato da Hosking (1990), fornisce stime confrontabili con quelle ottenute con il metodo della massima verosimiglianza, ed in alcuni casi le procedure di stima sono più semplici e i calcoli meno complessi. Per campioni limitati, l'efficienza del metodo degli L-momenti può essere superiore a quella del metodo della massima verosimiglianza. A differenza del metodo della massima verosimiglianza, per il metodo degli L-momenti, i risultati asintotici utili per la valutazione dell'incertezza degli stimatori hanno, in generale, una forma analitica complessa, per cui è preferibile, se non necessario, valutare l'incertezza della stima dei parametri e dei quantili tramite procedure di simulazione di tipo Monte Carlo (Kottegoda e Rosso 2008, pp. 517-519).

11.3.3 Test di adattamento della distribuzione empirica ad una distribuzione teorica

La scelta delle distribuzioni da usare nell'analisi delle frequenze di accadimento è da lungo tempo oggetto di ricerca. Come accennato in precedenza, per tale scelta sono necessari, oltre a considerazioni teoriche e grafici diagnostici, test delle ipotesi di tipo formale. Nel caso specifico, l'ipotesi nulla è che il campione sia una realizzazione di una definita distribuzione teorica, mentre l'ipotesi alternativa è che il campione non derivi da tale distribuzione. I vari test proposti in letteratura si differenziano per le grandezze assunte come indicative della bontà di adattamento. Accanto ai test formali, esistono una serie di indici o criteri di selezione, che possono essere usati per operare la scelta tra due o più modelli generici (non necessariamente delle distribuzioni di probabilità).

Tra gli indici di selezione dei modelli proposti in letteratura, è possibile ricordare, tra gli altri, l'*Akaike Information Criterion* (AIC - Akaike 1974) e lo *Schwarz Information Criterion* (SIC - Schwarz 1978), noto anche come *Bayesian Information Criterion* (BIC) o *Schwarz Bayesian Criterion* (SBC). Le espressioni analitiche di tali indici sono:

$$AIC = -2 \ln(L) + 2(k) \quad \text{eq. 11.3.1}$$

$$SBC = -2 \ln(L) + \ln(n) \cdot (k) \quad \text{eq. 11.3.2}$$

dove L indica il valore massimizzato della funzione di verosimiglianza del modello stimato, k il numero di parametri del modello e n la numerosità del campione.

In entrambe le espressioni, il primo termine rappresenta l'adattamento di un modello ai dati. L'adattamento risulta tanto migliore quanto maggiore è la complessità del modello e dunque il numero dei parametri. Il primo termine decresce al crescere del numero dei parametri. D'altro lato, un eccessivo numero di parametri produce un incremento dell'incertezza delle stime e conduce alla definizione di un modello che pur seguendo fedelmente i dati è poco utile operativamente, poiché riproduce non solo il comportamento sistematico di interesse, ma anche le fluttuazioni casuali dovute all'incertezza intrinseca del fenomeno (problema di "sovra-parametrizzazione"). Il secondo termine introduce dunque una penalizzazione (un termine positivo crescente al crescere del numero dei parametri) che limita il numero dei parametri in accordo con il principio di "parsimonia". I due indici conducono alla selezione di modelli che meglio descrivono i dati con un numero limitato di parametri. Come detto in precedenza, tali indici non costituiscono dei test, ma sono indicatori di performance che ordinano i modelli considerati in base al bilanciamento tra l'errore di adattamento (o analogamente il valore della massima verosimiglianza) e il numero dei parametri. Il modello migliore è quello che restituisce il migliore adattamento preservando un numero limitato di parametri.

Tra i test formali, è opportuno ricordare il test del χ^2 che si basa su una misura della differenza tra i valori di densità di probabilità empirica E_i e teorica O_i in ogni intervallo i di un insieme di l intervalli in cui è preventivamente suddiviso il campo di esistenza della variabile studiata:

$$\chi^2 = \sum_{i=1}^l (O_i - E_i)^2 / E_i \quad \text{eq. 11.3.3}$$

Il test richiede la scelta preventiva del numero (o dell'ampiezza) degli intervalli, per compiere la quale si possono adottare opportune formule empiriche (e.g., Moore 1986).

Test alternativi sono quelli basati sulle funzioni di ripartizione empiriche. A differenza del test del χ^2 , non è richiesta nessuna scelta preventiva ed i test sono condotti direttamente sulle distribuzioni di probabilità cumulate. A questa famiglia di test appartengono il test di Kolmogorov-Smirnov, di Cramer-von Mises e di Anderson-Darling (Stephens 1986).

Il test di Kolmogorov-Smirnov assume come statistica la massima distanza tra la funzione di ripartizione empirica F_n e quella teorica F sottoposta test:

$$D = \max_x |F_n(x) - F(x)| \quad \text{eq. 11.3.4}$$

Esso pertanto concentra l'attenzione in un punto della distribuzione, che potrebbe essere localizzato nella zona centrale della distribuzione.

Per i test di Cramer-von Mises e di Anderson-Darling le statistiche sono definite a partire dalla seguente espressione:

$$Q = n \int_{-\infty}^{\infty} (F_n(x) - F(x))^2 \psi(x) dF(x) \quad \text{eq. 11.3.5}$$

in cui n è la numerosità del campione e ψ è una funzione tale che per $\psi(x) = 1$, Q restituisce la statistica di Cramer-von Mises, mentre per $\psi(x) = [F(x)(1 - F(x))] - 1$, Q restituisce la statistica di Anderson-Darling.

Il test di Cramer-von Mises si basa sulla somma dei quadrati delle differenze tra funzione di ripartizione empirica F_n e quella teorica F per tutti i valori osservati e dà in qualche modo conto dell'adattamento dell'intera distribuzione, mentre il test di Anderson-Darling dà maggiore peso alle differenze che si manifestano sulle code della distribuzione. Sebbene il test di Kolmogorov-Smirnov sia il più noto, gli altri due sono generalmente più potenti.

Infine, occorre sottolineare che tutti i test richiedono che la distribuzione teorica sia completamente specificata, ovvero sia nota la forma analitica e il valore dei parametri. Se i parametri sono calcolati a partire dal campione testato, i valori critici delle statistiche test con assegnato livello di significatività, riportati in letteratura, sono generalmente troppo elevati e quindi poco utili alla discriminazione tra diverse distribuzioni, che di fatto non vengono rigettate anche se l'adattamento al campione non è soddisfacente. Nel caso del test del χ^2 , questo fenomeno non si verifica solo se i parametri sono stimati con il metodo della minimizzazione della statistica χ^2 . Per gli altri test, sono disponibili in letteratura delle formule correttive associate alle distribuzioni più comunemente usate in idrologia (Laio, 2004).

11.3.4 Analisi dei valori estremi

11.3.4.1 Tempo di ritorno

Lo scopo di un'analisi di frequenza effettuata sulle serie di dati idrologici è quello di associare il valore assunto da eventi estremi ad una corrispondente probabilità di accadimento tramite l'uso di distribuzioni di probabilità (Chow et al. 1988). Nelle analisi idrologiche, la frequenza di accadimento è spesso descritta tramite il concetto di tempo di ritorno T definito come *l'intervallo di tempo medio per cui un evento è superato o eguagliato*.

Per introdurre questa grandezza, si supponga che un evento estremo si verifichi allorché una variabile casuale X assuma un valore maggiore o uguale ad un fissato valore x_T . Indicando con t l'intervallo di inter-arrivo tra due eventi, ossia tra due istanti successivi tali che $X \geq x_T$, il tempo di ritorno è definito come il valore atteso di t , $E[t]$.

La probabilità di accadimento dell'evento $X \geq x_T$, esprimibile come $q = \Pr[X \geq x_T]$, è legata al tempo di ritorno sulla base delle seguenti considerazioni. Ad ogni osservazione corrispondono due possibili esiti: o l'esito è *positivo* e quindi $X \geq x_T$ con probabilità q oppure l'esito è *negativo* e quindi $X < x_T$ con probabilità $(1-q)$.

Nell'ipotesi di osservazioni indipendenti, la probabilità di osservare un intervallo di interarrivo t tra due superamenti è pari al prodotto delle probabilità di osservare $(t-1)$ esiti negativi (in cui $X < x_T$) seguiti da un esito positivo (in cui $X \geq x_T$), ossia $(1-q)^{t-1}q$.

Pertanto, il valore atteso di t è:

$$\begin{aligned}
E[t] &= \sum_{t=1}^{\infty} t(1-q)^{t-1} q \\
&= q + 2(1-q)q + 3(1-q)^2 q + 4(1-q)^3 q + \dots \\
&= q[1 + 2(1-q) + 3(1-q)^2 + 4(1-q)^3 + \dots]
\end{aligned}
\tag{eq. 11.3.6}$$

in cui l'espressione in parentesi quadra rappresenta lo sviluppo in serie di potenze:

$$(1+z)^n = 1 + nz + [n(n-1)/2]z^2 + [n(n-1)(n-2)/2]z^3 + \dots \tag{eq. 11.3.7}$$

assumendo $z = -(1-q)$ e $n = -2$. Ne segue che il valore atteso di t può essere riscritto come:

$$T = E[t] = \frac{q}{[1-(1-q)]^2} = \frac{1}{q} \tag{eq. 11.3.8}$$

Il tempo di ritorno T di un evento è quindi pari all'inverso della sua probabilità di accadimento. A volte, T è erroneamente interpretato come se l'evento con T anni di tempo di ritorno dovesse verificarsi una sola volta ogni T anni.

Dato che la probabilità che un evento con T anni di tempo di ritorno non venga superato in un periodo di N anni è $p^N = (1 - 1/T)^N$, allora la probabilità che l'evento si verifichi almeno una volta in N anni è il complemento della probabilità di non superamento, vale a dire $1 - p^N = 1 - (1 - 1/T)^N$. Per $T \geq 25$ anni e $N = T$, la probabilità che un evento con T anni di tempo di ritorno si verifichi almeno una volta in T anni è circa del 63.3%, ovvero, in un esperimento ideale, in circa due casi su tre.

11.3.4.2 Definizione del campione per l'analisi dei valori estremi

L'analisi dei valori estremi di una grandezza idrologica si concretizza nella definizione della relazione che lega i valori della grandezza stessa al tempo di ritorno (Maione 1999). Tale relazione può essere ottenuta analizzando due diverse serie idrologiche estratte, con opportuni criteri, dalla serie temporale dei valori assunti dalla grandezza in esame.

La prima di queste serie è quella dei massimi annuali (AM), ossia dei valori massimi assunti dalla grandezza in esame per ognuno degli anni di osservazione disponibili; la seconda è la cosiddetta serie di durata parziale (Partial Duration Series, PDS) o serie dei massimi che eccedono una soglia assegnata (Peaks Over Threshold, POT). Per un prefissato valore di soglia x_0 della variabile X , la serie PDS è ottenuta esaminando la serie temporale ed estraendo il valore massimo di X in ognuno degli intervalli di tempo in cui la variabile assume continuamente valori maggiori di x_0 (Maione 1999).

Entrambi i metodi di selezione dei valori estremi presentano aspetti a favore della scelta dell'uno o dell'altro. Poiché i dati relativi ai massimi annuali sono registrati e riportati nei rapporti pubblicati dagli enti preposti al monitoraggio, un importante vantaggio nell'uso delle serie AM è la loro disponibilità. Un limite è invece l'uso di un unico evento massimo per ogni anno, indipendentemente dal fatto che, ad esempio, il secondo evento più intenso per un determinato anno possa superare i massimi osservati negli altri anni. In queste circostanze, l'uso delle serie AM può portare ad escludere dall'analisi alcuni dati importanti per la valutazione dei quantili con elevati tempi di ritorno. Un altro esempio in cui l'uso delle serie AM può condurre a conclusioni poco affidabili, è quello relativo alle serie di portata in alcune zone aride o semi-aride. In queste circostanze, alcuni massimi annuali potrebbero essere nulli, implicando l'introduzione nella serie AM di valori che non possono essere definiti correttamente eventi di piena (Stedinger et al. 1993).

Questi inconvenienti non si presentano utilizzando le serie PDS, che considerano tutti i picchi indipendenti che eccedono una prefissata soglia. Le analisi sulle PDS possono restituire stime dei quantili estremi più accurate delle corrispondenti analisi dei massimi annuali, poiché in genere le serie PDS hanno una numerosità maggiore. I due aspetti principali che limitano l'impiego del metodo PDS sono:

1. la disponibilità dei dati e
2. la necessità di rispettare la condizione di mutua indipendenza delle osservazioni.

Mentre i massimi annuali sono generalmente disponibili, la selezione dei picchi sopra una soglia necessita la conoscenza della serie completa delle osservazioni, la cui acquisizione richiede un notevole lavoro di raccolta. Mentre nelle serie AM la condizione di mutua indipendenza tra due osservazioni successive è, in genere, garantita dal notevole intervallo temporale che le separa, nelle

serie PDS ciò potrebbe non essere verificato e, pertanto, occorre introdurre un criterio per garantire il rispetto dell'ipotesi di indipendenza.

Nell'analisi PDS emergono due aspetti di cui occorre tener conto:

1. occorre modellare il numero di eventi che superano la soglia x_0 in un anno generico;
2. occorre definire un modello per i valori assunti da questi eventi.

La distribuzione di Poisson⁹ è spesso usata per descrivere il numero dei superamenti, mentre si utilizza una distribuzione di tipo esponenziale¹⁰ per descrivere i valori dei picchi sopra soglia.

Considerate le diverse modalità di campionamento, è utile definire la relazione esistente tra i tempi di ritorno desumibili utilizzando i due diversi metodi. Le distribuzioni relative alle procedure AM e PDS sono, infatti, strettamente legate da relazioni, anche se non è sempre agevole determinarle. In alcuni casi, tuttavia, il legame assume una forma particolarmente semplice. Per una serie PDS, supponiamo che sia λ il numero medio di eventi per anno eccedenti una fissata soglia x_0 , e sia $G(x)$ la probabilità che il valore di tali eventi, quando si verificano, sia minore di x , ossia che il valore ricada nell'intervallo (x_0, x) . Allora il numero di eventi per ogni livello x , con $x \geq x_0$, è $\lambda^* = \lambda[1 - G(x)]$.

La distribuzione $F(x)$ per la corrispondente serie AM descrive qual è la probabilità che il massimo annuale non ecceda x in un determinato anno. Per eventi indipendenti, assumendo che la probabilità di non superamento di x su un periodo di un anno sia descritta dalla distribuzione di Poisson, ne segue che:

$$\begin{aligned} F(x) &= \Pr[X \leq x] \\ &= \Pr[\text{nessun evento tale che } X > x] \\ &= \exp(-\lambda^*) \\ &= \exp\{-\lambda[1 - G(x)]\} \end{aligned} \quad \text{eq. 11.3.9}$$

Questa equazione rappresenta la relazione tra la distribuzione dei massimi annuali $F(x)$, il numero medio di eventi nell'unità di tempo e la distribuzione dei picchi sopra soglia $G(x)$. Se la probabilità di superamento annuale $(1 - F(x))$ è indicata con $1/T$, e la corrispondente probabilità di superamento $(1 - G(x))$ per un livello x nella PDS è indicata con q , allora è possibile scrivere:

$$\frac{1}{T} = 1 - \exp(-\lambda q) = 1 - \exp\left(-\frac{1}{T_p}\right) \quad \text{eq. 11.3.10}$$

in cui $T_p = 1/\lambda q$ è il tempo di ritorno per il livello x nella serie PDS. Risolvendo in T_p , si ottiene:

$$T_p = -\frac{1}{\ln(1 - 1/T)} \quad \text{eq. 11.3.11}$$

⁹ La distribuzione di Poisson descrive il numero di eventi che si verificano in un intervallo temporale. La sua densità di probabilità è definita come:

$$p_x(x; \nu) = \Pr[X = x] = \frac{\nu^x e^{-\nu}}{x!}$$

per $x = 0, 1, 2, \dots$ e in cui $\nu > 0$ è il parametro che rappresenta il numero medio di eventi in un intervallo temporale di riferimento (e.g., l'anno).

¹⁰ La distribuzione esponenziale descrive il tempo che intercorre tra due eventi in un processo in cui il numero di accadimenti in un fissato intervallo temporale segue una distribuzione di Poisson. Se infatti l'accadimento di un evento segue una distribuzione di Poisson, allora la probabilità che l'evento (e.g. il superamento di una soglia) non si verifichi in intervallo di tempo t è $\Pr[X = 0] = e^{-\lambda t}$ in cui $\lambda = \nu/t$ è il numero medio di eventi nell'unità di tempo. Da questo risultato e considerando il tempo tra due eventi (superamenti) come una variabile casuale, segue che

$$F_T(t) = \Pr[T \leq t] = 1 - \Pr[X = 0] = 1 - e^{-\lambda t},$$

ossia, il tempo di attesa tra due eventi successivi di un processo di Poisson segue una distribuzione esponenziale.

Sostituendo T con una generica variabile X , la distribuzione esponenziale è:

$$F_x(x) = 1 - e^{-\lambda x}$$

per $x \geq 0$ e in cui $\lambda > 0$.

T_p è minore di T , poiché più di un evento per anno può verificarsi nella serie PDS. Le relazioni precedenti trasformano il numero medio di eccedenze nell'unità di tempo λq per eventi maggiori di x nella probabilità di superamento annuale $1/T$ della serie AM. Per valori x con $T > 10$ anni, corrispondenti a eventi poco frequenti, la probabilità di superamento annuale $1/T$ eguaglia sostanzialmente $\lambda q = \lambda[1 - G(x)]$ della corrispondente PDS, per cui $T = T_p$. Dalla relazione eq. 11.3.11 è possibile ottenere la seguente Tabella 11.1 di conversione per valori di T minori o uguali a 10 anni:

Tabella 11.1 - Tempi di ritorno T_p relativi alle serie PDS e corrispondenti valori T relativi alle serie AM ottenuti dalla relazione eq. 11.3.11

T_p	0.50	1.00	1.45	2.00	5.00	10.00
T	1.16	1.58	2.00	2.54	5.52	10.50

11.3.5 Precipitazioni e analisi dei valori estremi: curve Intensità-Durata-Frequenza (IDF)

Lo studio statistico delle precipitazioni intense e di breve durata di tipo puntuale (ossia registrate in una determinata località) è comunemente svolto ricorrendo alle cosiddette curve di *Intensità-Durata-Frequenza (IDF)*, o *Altezza-Durata-Frequenza (Depth-Duration-Frequency, DDF)*. Queste curve, che in Italia sono note col nome di *curve di probabilità pluviometrica*, rappresentano la relazione tra l'intensità i_d o l'altezza h_d di pioggia che si registra nella località considerata e la sua durata d , per un assegnato valore di tempo di ritorno. L'intensità (media) di precipitazione nella durata d è il rapporto tra l'altezza di pioggia h_d caduta nell'intervallo di durata d e la durata stessa. Nell'ambito delle piogge intense, l'intensità si misura usualmente in mm h^{-1} e, a volte in mm giorno^{-1} . Di seguito viene riportata una breve sintesi relativa alla determinazione di questo tipo di curve. Per una trattazione più approfondita si consiglia di consultare i testi di Calenda e Margaritora (1993) e di Maione (1999).

I dati necessari per la definizione delle curve *IDF* sono i valori delle intensità massime annuali di pioggia per un insieme di durate di interesse. Ad esempio, gli Annali Idrologici, pubblicati dagli Uffici Idrografici Regionali dell'ex SIMN, riportano le altezze di pioggia (da cui è possibile dedurre le intensità) per le durate di 1, 3, 6, 12, 24 ore e di 1, 2, 3, 4, 5 giorni consecutivi. Al riguardo, è utile ricordare che la procedura di acquisizione di questi dati consiste nell'esame delle serie cronologiche registrate dagli strumenti di misura (ietogrammi) e nell'individuazione degli intervalli di tempo di durata $d = \Delta t, 2\Delta t, 3\Delta t, \dots$, in cui si è verificata la massima intensità (o altezza) di precipitazione. Ripetendo l'analisi per ogni anno, si possono ricavare delle tabelle simili alla Tabella 11.2-A, in cui ogni colonna riporta la serie dei valori di intensità massima annuale corrispondenti ad un intervallo temporale di prefissata durata d registrati in un periodo di 36 anni. I valori evidenziati nella Tabella 11.2-A in blu rappresentano i massimi dell'intensità per ogni colonna, dunque per ogni durata considerata. Ordinando le serie in ordine decrescente, si ottiene una tabella Tabella 11.2-B in cui la prima riga contiene i cinque valori di intensità massima per le durate d considerate (evidenziati in Tabella 11.2-A) che costituiscono quindi il cosiddetto primo caso critico, la seconda riga contiene i cinque secondi valori più elevati che rappresentano il secondo caso critico, e così a seguire.

Le curve che uniscono i punti rappresentativi dello stesso caso critico per ogni durata in un grafico (d, i_d) sono dette *curve di caso critico* (Figura 11.12). Ai valori appartenenti ad ogni caso critico è possibile associare una frequenza di superamento corrispondente all'ordine del caso critico. Ad esempio, le osservazioni relative al quinto caso critico dei dati riportati in Tabella 11.2-B hanno una probabilità di essere uguagliate o superate pari a $5/(36 + 1)$ per ogni durata. Poiché, in genere, possono essere di interesse i valori di precipitazione relativi a durate e frequenze di accadimento diverse da quelle osservate, è opportuno regolarizzare le curve di caso critico tramite relazioni analitiche. La regolarizzazione è eseguita in due fasi.

In primo luogo, si adatta una distribuzione parametrica alla funzione di ripartizione empirica della serie degli n massimi annuali $i_{j,d}, j = 1, 2, \dots, n$, riferiti ad una fissata durata d . In particolare, trattandosi di valori massimi, si trova spesso che la distribuzione di Gumbel¹¹ interpreta in modo soddisfacente le

¹¹ La distribuzione di Gumbel, detta anche dei valori estremi del primo tipo è espressa dalla relazione:

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

in cui $\infty \leq x \leq \infty, \infty \leq u \leq \infty$ è un parametro di posizione, $\alpha > 0$ è un parametro di scala.

osservazioni campionarie, anche se possono essere usate anche altre distribuzioni, qualora mostrino un adattamento migliore ai dati.

Tabella 11.2 - Valori massimi annuali di intensità di pioggia riferiti a diverse durate d.

(A)						(B)					
Intensità di pioggia						Intensità di pioggia					
Anno	Durata					Caso critico	Durata				
	1 h	3 h	6 h	12 h	24 h		1 h	3 h	6 h	12 h	24 h
1950	22.0	9.3	5.7	3.4	2.3	1	42.2	23.8	16.5	11.0	6.2
1951	22.4	12.1	7.3	4.0	2.6	2	41.4	20.0	13.3	9.8	5.3
1952	31.8	11.2	5.6	3.3	2.0	3	38.8	18.3	13.2	6.7	4.7
1953	23.7	9.1	7.8	5.4	3.3	4	38.6	16.9	11.7	5.9	4.6
1954	21.0	13.5	7.2	3.9	2.0	5	37.6	15.9	8.9	5.8	4.1
1955	25.0	23.8	13.2	6.7	3.3	6	35.6	15.6	8.7	5.6	3.5
1956	16.0	6.0	4.9	2.9	2.0	7	34.0	14.8	8.3	5.5	3.3
1957	20.0	14.8	8.3	5.6	3.5	8	32.0	14.5	7.8	5.4	3.3
1958	20.8	10.3	7.1	5.1	2.9	9	31.8	13.5	7.8	5.4	3.3
1959	31.4	11.7	6.2	3.8	2.5	10	31.4	13.3	7.8	5.1	3.2
1960	37.6	18.3	11.7	5.8	3.0	11	29.6	12.4	7.7	4.8	3.0
1961	11.8	9.1	7.7	5.5	4.7	12	27.0	12.1	7.5	4.7	3.0
1962	41.4	15.6	7.8	4.8	3.3	13	27.0	12.0	7.5	4.7	2.9
1963	38.6	13.3	6.7	3.3	1.9	14	25.2	11.7	7.3	4.6	2.9
1964	23.8	10.0	5.1	2.5	1.7	15	25.0	11.7	7.2	4.4	2.7
1965	29.6	9.9	4.9	3.3	2.3	16	25.0	11.2	7.2	4.4	2.6
1966	38.8	16.9	16.5	11.0	6.2	17	24.4	10.6	7.1	4.3	2.5
1967	20.0	9.1	5.0	3.0	2.1	18	23.8	10.3	6.9	4.2	2.5
1968	27.0	10.3	7.5	4.4	2.0	19	23.7	10.3	6.7	4.2	2.5
1969	21.4	10.6	5.3	4.7	2.5	20	23.0	10.3	6.2	4.0	2.3
1970	27.0	9.6	4.8	2.9	1.5	21	22.4	10.0	6.0	3.9	2.3
1971	23.0	8.5	4.8	4.2	2.1	22	22.2	9.9	6.0	3.9	2.3
1972	25.2	9.3	5.3	2.9	1.9	23	22.0	9.7	5.7	3.8	2.3
1973	35.6	12.0	7.8	5.9	4.1	24	21.4	9.6	5.6	3.6	2.1
1974	24.4	12.4	8.9	5.4	2.7	25	21.0	9.3	5.5	3.5	2.1
1975	20.0	9.7	6.0	4.3	2.9	26	20.8	9.3	5.3	3.4	2.0
1976	34.0	11.7	7.2	4.4	3.0	27	20.2	9.1	5.3	3.3	2.0
1977	25.0	14.5	7.5	4.2	3.2	28	20.0	9.1	5.1	3.3	2.0
1978	18.0	6.7	5.5	4.7	4.6	29	20.0	9.1	5.0	3.3	2.0
1979	20.2	8.9	6.9	3.9	2.3	30	20.0	8.9	4.9	3.0	2.0
1980	11.4	6.7	4.6	3.5	2.3	31	18.8	8.5	4.9	2.9	1.9
1981	42.2	15.9	8.7	4.6	2.5	32	18.0	6.8	4.8	2.9	1.9
1982	32.0	20.0	13.3	9.8	5.3	33	17.0	6.7	4.8	2.9	1.9
1983	22.2	10.3	6.0	3.6	1.8	34	16.0	6.7	4.6	2.9	1.8
1984	18.8	6.8	3.8	2.9	1.9	35	11.8	6.2	4.3	2.7	1.7
1985	17.0	6.2	4.3	2.7	2.0	36	11.4	6.0	3.8	2.5	1.5

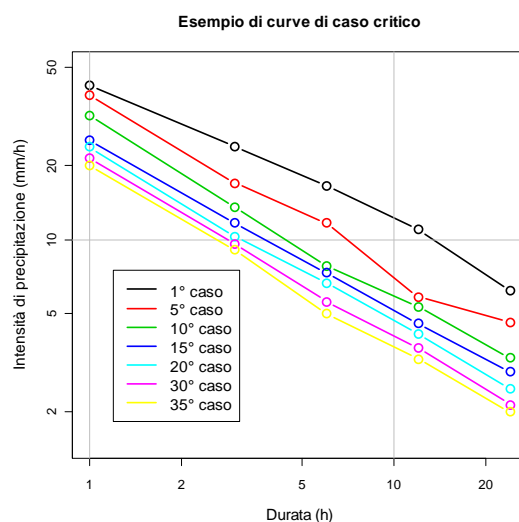


Figura 11.12 - Esempio di curve di caso critico riferite a dati riportati in Tabella 11.2B. Il grafico in scala bi-logaritmica evidenzia un andamento approssimativamente lineare.

Una volta stimati i parametri della distribuzione scelta, è possibile calcolare l'intensità di pioggia relativa alle durate disponibili (e.g. 1, 3, 6, 12, 24 h) con qualsiasi tempo di ritorno T (o probabilità di superamento), $i_{d,T}$. Se i valori $i_{d,T}$ si dispongono su un piano bi-logaritmico in cui in ascissa si riportano le durate d e in ordinata la corrispondente $i_{d,T}$, si osserva che i valori che corrispondono ad uno stesso tempo di ritorno tendono a disporsi su delle curve approssimativamente lineari e parallele simili alle curve di caso critico.

La seconda fase della procedura di regolarizzazione è l'interpolazione delle intensità di pioggia con assegnato tempo di ritorno T tramite una curva analitica che permetta il calcolo delle intensità con fissata probabilità di superamento per durate diverse da quelle per cui sono disponibili i dati. In genere l'interpolazione può essere eseguita tramite una legge del tipo¹²:

$$i_{d,T} = ad^{n-1} \quad \text{eq. 11.3.12}$$

in cui a e n sono parametri dipendenti dal tempo di ritorno T . Questa relazione è detta *curva di probabilità pluviometrica* di tempo di ritorno T o *curva Intensità-Durata-Frequenza (IDF)*, ed esprime, per ogni durata, la massima intensità di pioggia puntuale con tempo di ritorno T . La corrispondente curva per l'altezza di pioggia $h_{d,T}$ è data da:

$$h_{d,T} = ad^n \quad \text{eq. 11.3.13}$$

I parametri a e n possono essere stimati individuando la retta che meglio interpola le intensità di pioggia con uguale tempo di ritorno nel piano bi-logaritmico, ricorrendo, ad esempio, al metodo dei minimi quadrati. Sebbene entrambi i parametri varino con il tempo di ritorno, tuttavia, mentre il parametro a , che rappresenta la pioggia di durata unitaria (e.g., la pioggia di un'ora se le durate sono misurate in ore), è crescente al crescere del tempo di ritorno, il parametro n presenta spesso variazioni modeste e può essere assunto costante, così come mostrato dal parallelismo delle curve riportate in Figura 11.13.

Un metodo alternativo per la definizione delle curve IDF, basato sulle proprietà di scala delle serie di precipitazione, è descritto ad esempio da Burlando e Rosso (1996). Invece di definire i parametri delle distribuzioni delle serie dei massimi per ogni durata e adattare poi una curva IDF ai valori corrispondenti ad uguali tempi di ritorno, il metodo, basato sull'invarianza di scala, consiste nel definire una distribuzione i cui parametri dipendono dalla durata. In questo modo, per ogni fissato valore della durata d , è disponibile l'espressione della corrispondente distribuzione di probabilità, tramite la quale calcolare ogni possibile quantile per quella durata. Il metodo sarà descritto in dettaglio nel paragrafo 11.4.

¹² La letteratura fornisce anche altre espressioni più complesse e con un maggiore numero di parametri per tener conto, ad esempio, della curvatura che si osserva in Figura 11.13 per durate inferiori ad 1 ora. Ad esempio, il WMO (1994) suggerisce le tre espressioni:

$$i_{d,T} = \frac{a}{b+d}, \quad i_{d,T} = \frac{a}{(d-b)^n}, \quad i_{d,T} = \frac{a+b \log T}{(1+d)^n}$$

mentre Chow et al. (1988) suggeriscono

$$i_{d,T} = \frac{aT^m}{d+b}, \quad i_{d,T} = \frac{aT^m}{d^n+b}$$

in cui a , b , m e n sono parametri che dipendono dal tempo di ritorno T e variano da stazione a stazione.

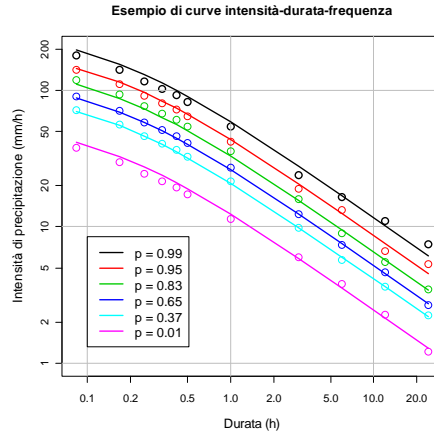


Figura 11.13 - Esempio di curve IDF. Ogni curva interpola i valori di intensità di pioggia corrispondenti a una fissata probabilità di non superamento p , calcolati tramite distribuzioni parametriche stimate sui dati relativi ad ogni durata disponibile.

11.4 Procedure di analisi

In questa sezione si descrivono in dettaglio le procedure introdotte.

11.4.1 Calcolo del parametro di Hurst

In questo paragrafo sono descritti in dettaglio i metodi per il calcolo del parametro di Hurst. Saranno analizzati i metodi della varianza aggregata, del rescaled range R/S , e di Higuchi.

Il calcolo del parametro di Hurst è riconducibile ad un classico studio di tipo idrologico condotto da Hurst (1951) per il progetto di un invaso artificiale ottimale in relazione alle portate di efflusso osservate (Feder 1988). Considerando l'anno come unità temporale, la gestione dell'invaso comporta che per ogni anno t , il serbatoio riceverà una portata in ingresso $\xi(t)$, e rilascerà una portata in uscita $\bar{\xi}_\tau$. Lo studio di Hurst era volto alla definizione della capacità di invaso necessaria per avere una portata in uscita uguale alla media degli afflussi per un fissato periodo in esame. La portata media in entrata in un periodo di τ anni è data dalla relazione:

$$\bar{\xi}_\tau = \frac{1}{\tau} \sum_{t=1}^{\tau} \xi(t) \quad \text{eq. 11.4.1}$$

Si consideri ora la serie degli scarti cumulati, $X(t, \tau)$, degli afflussi $\xi(t)$ dalla media $\bar{\xi}_\tau$:

$$X(t, \tau) = \sum_{u=1}^t (\xi(u) - \bar{\xi}_\tau) \quad \text{per } t = 1, \dots, \tau, \quad \text{eq. 11.4.2}$$

che rappresentano i volumi accumulati nell'invaso al netto della portata costante rilasciata. La differenza tra il valore massimo e quello minimo della serie nell'intervallo considerato, $R(\tau) = \max_{1 \leq t \leq \tau} X(t, \tau) - \min_{1 \leq t \leq \tau} X(t, \tau)$, rappresenta il range R . Il range è la capacità richiesta per garantire la portata media in uscita per tutto il periodo τ . Per un invaso con sufficiente capacità, R è la differenza tra i volumi massimo e minimo in esso contenuti durante il periodo τ .

La Figura 11.14 illustra un esempio della curva descritta dalla serie $X(t, \tau)$ per la serie delle portate del Nilo ad Assuan tra il 1871 e il 1930 ($\tau = 60$ anni) e tra il 1871 e il 1970 ($\tau = 100$ anni).

Chiaramente il range dipende dal periodo di tempo τ e ci si attende che R cresca all'aumentare di τ , come è possibile desumere dai due grafici in Figura 11.14 ($R \cong 4000 \cdot 10^8 \text{ m}^3$ per il periodo 1871-1930 e

$R \cong 5000 \cdot 10^8 \text{ m}^3$ per il periodo 1871-1970). Hurst ha analizzato un elevato numero di serie temporali relative a diversi fenomeni ambientali (si veda, e.g., Feder (1988)), per indagare le proprietà del parametro R e dedurre delle generalizzazioni riguardo al suo comportamento. Poiché i valori di R possono essere molto eterogenei al variare delle grandezze esaminate (temperature, portate, ecc.), Hurst ha adimensionalizzato il parametro R dividendolo per la deviazione standard S di $\xi(t)$:

$$S = \left[\frac{1}{\tau} \sum_{t=1}^{\tau} (\xi(t) - \bar{\xi}_{\tau})^2 \right]^{1/2} \quad \text{eq. 11.4.3}$$

ottenendo un parametro detto *rescaled adjusted range*, R/S . Hurst ha osservato che R/S varia in funzione del tempo τ secondo una legge di potenza del tipo:

$$R/S = (\tau/2)^H \quad \text{eq. 11.4.4}$$

in cui H è detto esponente o parametro di Hurst.

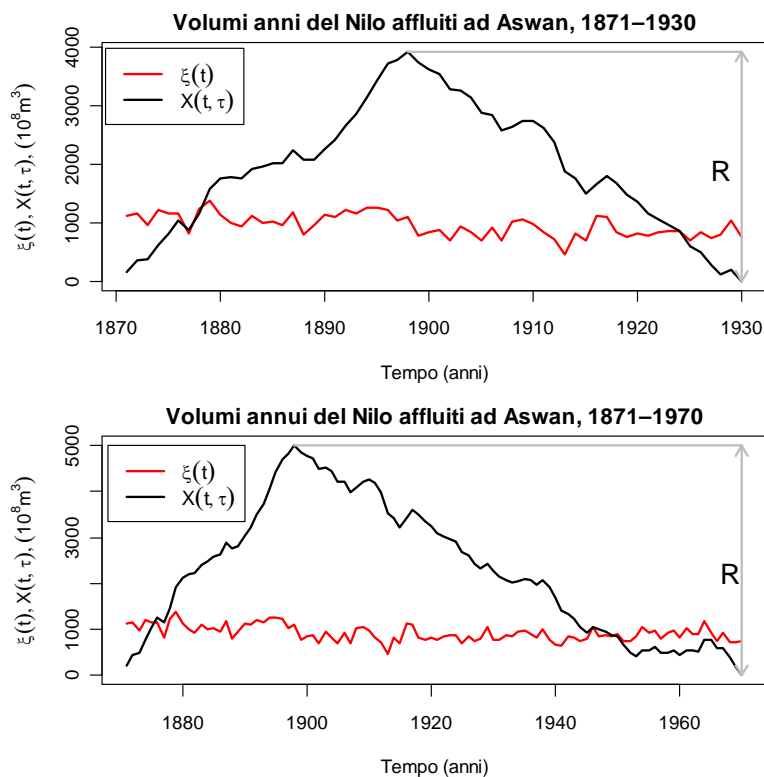


Figura 11.14 - Serie delle portate annuali del Nilo ad Assuan (linea tratteggiata) e serie degli scarti cumulati (linea continua). Il range è indicato con R .

Dallo studio empirico di serie di diversa natura (portate, intensità di precipitazione, temperature, pressioni, ecc.) e numerosità, Hurst ha mostrato che per molti fenomeni naturali il valore di H è distribuito in modo pressoché simmetrico con media di circa 0.726 e deviazione standard pari a circa 0.09. Il valore del parametro di Hurst indica l'entità della correlazione tra osservazioni passate e future e in particolare:

- per $H = 0.5$, la serie è stazionaria e non presenta persistenza;
- per $H > 0.5$, la serie presenta persistenza o memoria di lungo periodo. In questo caso se i dati osservati in un intervallo di tempo nel passato mostrano un incremento positivo (negativo), allora si avrà in media un incremento (decremento) in futuro. In altre parole, un trend positivo (negativo) nel passato implica verosimilmente un trend positivo (negativo) in futuro. Inoltre, per $H > 0.5$, la correlazione delle osservazioni non si annulla mai, per cui le osservazioni passate influiscono su quelle future, indipendentemente dal lag che le separa. Da questa proprietà deriva la definizione di *lunga memoria*.

- per $H < 0.5$, la serie presenta antipersistenza. In questo caso, un trend positivo (negativo) nel passato implica un trend negativo (positivo) in futuro.

Oltre al metodo R/S , basato sulla relazione $R/S = (\tau/2)^H$ altri metodi utilizzati per l'identificazione e la stima iniziale del parametro di Hurst sono:

- il metodo della varianza aggregata;
- il metodo di Higuci.

Occorre, tuttavia, rilevare che l'efficacia di questi metodi di stima è influenzata da altre proprietà presenti nelle serie reali, quali la non stazionarietà (come trend, stagionalità ecc.), la presenza di correlazione di breve memoria (la più comune struttura di correlazione, caratterizzata da un andamento con veloce decadenza al variare del tempo) e l'insufficiente numerosità del campione.

11.4.1.1 Metodo della varianza aggregata

Un metodo semplice per il calcolo del parametro di Hurst è quello che si basa sulle proprietà della media campionaria. Data una sequenza $\{X_1, \dots, X_n\}$ di variabili indipendenti e identicamente distribuite, per il teorema del limite centrale¹³, al crescere della numerosità n , la varianza $Var[\hat{\mu}_X]$ della media campionaria $\hat{\mu}_X$, tende ad essere inversamente proporzionale alla numerosità. In presenza di memoria di lungo periodo, la relazione tra la deviazione standard della sequenza di variabili e quella della corrispondente media campionaria assume la forma generale (e.g., Koutsoyiannis 2006b):

$$Var[\hat{\mu}_X] = \frac{\sigma_X^2}{n^{2-2H}} \quad \text{eq. 11.4.5}$$

in cui H è il parametro di Hurst. Questa relazione restituisce il risultato fornito dal teorema del limite centrale per $H = 0.5$ (assenza di lunga memoria). L'eq. 11.4.5 implica una relazione lineare tra i logaritmi di n e $Var[\hat{\mu}_X]$:

$$\log Var[\hat{\mu}_X] = \log \sigma_X^2 + 2(H - 1) \log n \quad \text{eq. 11.4.6}$$

la quale suggerisce un metodo per la stima di H . Il valore della pendenza della retta di regressione tra n e $Var[\hat{\mu}_X]$ nel piano bi-logaritmico determina il parametro di Hurst allorché si calcoli $Var[\hat{\mu}_X]$ per vari valori di n . In pratica, da un campione di numerosità m si estraggono dei sottocampioni disgiunti di numerosità ridotta n , per ognuno dei quali si calcola la media campionaria. Si ottiene così una serie di (m/n) medie campionarie $\hat{\mu}_X$ relative a sottocampioni di numerosità n , dalla quale è possibile stimare $Var[\hat{\mu}_X]$. Il calcolo può essere ripetuto per diversi valori di n , in modo da ottenere un numero sufficiente di coppie $(n, Var[\hat{\mu}_X])$ da disporre nel piano bi-logaritmico per valutare l'allineamento, la pendenza della retta di regressione e dunque H .

In particolare, per un campione di lunghezza m , se si assumono sottocampioni di numerosità unitaria, $n = 1$, si ha che:

$$\hat{\mu}_X^{(1)} = X_1, \hat{\mu}_X^{(2)} = X_2, \dots, \hat{\mu}_X^{(m)} = X_m,$$

e quindi $Var[\hat{\mu}_X] \equiv \sigma^2$ (per qualunque valore di H).

Nel caso in cui la numerosità n dei sottocampioni sia uguale a 2, si definisce un numero di campioni pari a $(m/2)$, per i quali si possono calcolare le $(m/2)$ medie:

$$\hat{\mu}_X^{(1)} = (X_1 + X_2)/2, \hat{\mu}_X^{(2)} = (X_3 + X_4)/2, \dots, \hat{\mu}_X^{(m/2)} = (X_{m-1} + X_m)/2,$$

¹³ Teorema del limite centrale: se $\{X_1, \dots, X_n\}$ è una successione di variabili casuali indipendenti e identicamente distribuite con valore medio μ_X e varianza $0 < \sigma_X^2 < +\infty$, allora la media campionaria $\hat{\mu}_X$ tende ad avere una distribuzione Gaussiana con media uguale a μ_X e varianza $Var[\hat{\mu}_X] = \sigma_X^2 / n$.

e quindi la varianza $Var[\hat{\mu}_X]$ relativa alla numerosità $n = 2$.

Il calcolo è ripetuto per valori di n crescenti, ad esempio, fino a $n = m/10$ (per avere almeno 10 valori $\hat{\mu}_X$ su cui calcolare la varianza $Var[\hat{\mu}_X]$). La Figura 11.15 illustra il risultato dell'applicazione della procedura a due serie simulate ottenute imponendo $H = 0.5$ e 0.75 .

La retta di regressione è stimata scartando i punti corrispondenti ai valori di n più bassi (che corrispondono a intervalli temporali brevi in cui domina la memoria breve) e i valori più elevati in cui è presente una maggiore variabilità nell'allineamento dovuto alla riduzione della numerosità del campione di $\hat{\mu}_X$ su cui si calcola $Var[\hat{\mu}_X]$. La scelta dei punti da scartare (cut-off) introduce chiaramente un elemento di soggettività nella procedura. Un semplice metodo di scelta è quello di considerare diversi valori limite di n e controllare la variabilità della stima per le diverse scelte.

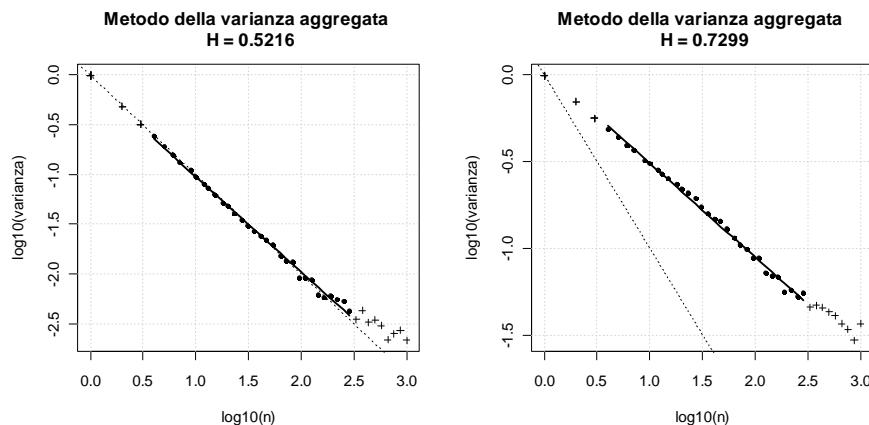


Figura 11.15 - Metodo della varianza aggregata: applicazione ad una serie simulata di numerosità 10000 con $H = 0.5$ (pannello a sinistra) e $H = 0.75$ (pannello a destra). In entrambi i pannelli, la stima della pendenza della retta di regressione (linea continua) è stata condotta con i punti compresi nell'intervallo temporale logaritmico (0.5-2.5). Le linee tratteggiate indicano le rette con pendenze corrispondenti ad $H = 0.5$ (assenza di lunga memoria).

11.4.1.2 Metodo R/S

Il metodo R/S si basa sulla relazione $R/S = (\tau/2)^H$, che implica una relazione lineare tra il range riscalato R/S e il tempo τ nel piano bi-logaritmico. La pendenza della retta di regressione stimata sulle coppie di punti $(\log R/S, \log \tau)$ coincide con il valore del parametro H . Operativamente, si calcolano i valori di R/S per alcuni valori di τ equispaziati nel piano logaritmico. Analogamente al metodo della varianza aggregata, per stabilire il valore di H è opportuno stimare la retta di regressione escludendo i valori di R/S relativi a intervalli di tempo τ brevi ed elevati; la scelta dei punti da escludere può essere condotta analogamente a quanto esposto per il metodo della varianza aggregata. Dalle analisi di robustezza di diversi metodi per il calcolo di H , tra i quali il metodo della varianza aggregata, del rescaled range R/S , e di Higuchi (Taquq et al. 1995; Montanari 1996), si è potuto constatare che il metodo R/S tende a sovrastimare i valori di H minori di 0.7, e a sottostimare quelli maggiori di 0.7. Si è inoltre osservato che il metodo è molto sensibile alla presenza di periodicità e di non stazionarietà. In Figura 11.16 è mostrato un esempio dell'applicazione del metodo alle stesse serie usate per illustrare il metodo della varianza aggregata.

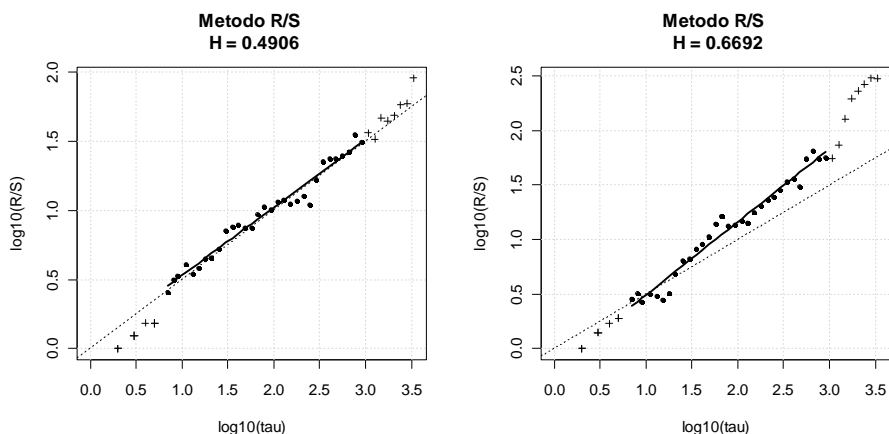


Figura 11.16 - Metodo R/S: applicazione ad una serie simulata di numerosità 10000 con $H = 0.5$ (pannello a sinistra) e $H = 0.75$ (pannello a destra). In entrambi i pannelli, la stima della pendenza della retta di regressione (linea continua) è stata condotta con i punti compresi nell'intervallo temporale logaritmico (1.0-3.0). Le linee tratteggiate indicano le rette con pendenze corrispondenti ad $H = 0.5$ (assenza di lunga memoria).

11.4.1.3 Metodo di Higuchi

L'approccio suggerito da Higuchi (1988) è basato sulla similitudine tra una serie temporale con lunga memoria e una curva frattale. Supponendo di approssimare una generica curva con una successione di $N(k)$ segmenti di lunghezza k , ci si attende che la lunghezza $L(k) = N(k) \times k$ approssimi sempre meglio la lunghezza reale L_R della curva per valori sempre minori di k . In realtà ciò non sempre accade, ed in alcuni casi la lunghezza $L(k)$ è legata a k tramite una relazione di proporzionalità del tipo $L(k) \propto k^D$, in cui D , detta dimensione frattale, assume valori non interi (frazionari) compresi tra 1 e 2. In queste circostanze $L(k)$ cresce al decrescere di k invece di convergere ad un valore L_R . Un tipico esempio di questo fenomeno è la misura della lunghezza della costa di un'isola: immaginando di approssimare la lunghezza con una successione di segmenti di lunghezza sempre minore (e.g. da 1 km a 1 cm), si otterranno in genere lunghezze sempre maggiori, poiché l'uso di segmenti progressivamente più corti permette di considerare irregolarità della costa che sono trascurate utilizzando segmenti più lunghi.

Il parametro di Hurst è legato alla dimensione frattale D tramite la relazione $D = 2 - H$ (Feder 1988, pp. 184-187). Nel metodo proposto da Higuchi, l'attenzione è dunque rivolta alla stima della dimensione frattale di una serie tramite la relazione di proporzionalità $L(k) \propto k^D$ in cui L è una lunghezza del frattale associato alla serie calcolata alla scala di aggregazione temporale k .

Data una serie temporale $X(1), \dots, X(N)$, si costruisce una nuova serie, X_k^m , definita come segue (Higuchi 1988):

$$X_k^m = \left\{ X(m), X(m+k), X(m+2k), \dots, X\left(m + \left\lfloor \frac{N-m}{k} \right\rfloor k\right) \right\} \quad \text{eq. 11.4.7}$$

$$m = 1, 2, \dots, k$$

in cui $\lfloor \cdot \rfloor$ indica la funzione parte intera inferiore, e m e k sono interi ed indicano rispettivamente il valore iniziale della serie e l'intervallo temporale. Per un intervallo temporale uguale a k si ottengono k nuove serie. Ad esempio, per $k = 3$ e $N = 100$, si ricavano le tre serie:

$$\begin{aligned} X_3^1 &= \{X(1), X(4), X(7), \dots, X(97), X(100)\} \\ X_3^2 &= \{X(2), X(5), X(8), \dots, X(98)\} \\ X_3^3 &= \{X(3), X(6), X(9), \dots, X(99)\} \end{aligned}$$

La lunghezza della curva X_k^m è definita come:

$$L_m(k) = \frac{1}{k} \left\{ \frac{N-1}{\lfloor \frac{N-m}{k} \rfloor k} \left(\sum_{i=1}^{\lfloor \frac{N-m}{k} \rfloor} |X(m+ik) - X(m+(i-1)k)| \right) \right\} \quad \text{eq. 11.4.8}$$

in cui $(N-1)/\lfloor (N-m)/k \rfloor k$ rappresenta un fattore di normalizzazione per la lunghezza della curva k -esima del generico intervallo k . La lunghezza della curva relativa a ogni intervallo k è definita infine come la media $\langle L_m(k) \rangle$ dei k valori $L_m(k)$. Se $\langle L_m(k) \rangle$ è proporzionale a k^{-D} allora la curva è un frattale con dimensione $D = 2-H$, e le coppie di punti $\langle L_m(k) \rangle$ e k dovrebbero allinearsi su una retta con pendenza $-D$ in un piano bi-logaritmico. La Figura 11.17 illustra i risultati dell'applicazione del metodo alle stesse serie usate nei due paragrafi precedenti. Anche in questo caso, analogamente ai metodi esposti precedentemente, è opportuno stimare la pendenza della retta di regressione scartando i valori corrispondenti alle scale di aggregazione più basse e più alte.

Nota 11.1: Per l'applicazione pratica del metodo di Higuchi è necessario sottrarre alla serie analizzata la media campionaria, ottenendo così una serie "trasformata" avente media nulla.

Nota 11.2: Per il calcolo del parametro di Hurst è opportuno che la serie abbia una lunghezza maggiore 100 (e.g., Hosking 1984).

Nota 11.3: I metodi per il calcolo del parametro di Hurst descritti in precedenza possono essere influenzati in varia misura dalla presenza della periodicità stagionale che caratterizza le serie idrologiche (e.g., Montanari et al 1999). È quindi opportuno calcolare H su serie preventivamente destagionalizzate tramite le procedure di seguito esposte nel paragrafo 11.4.2.

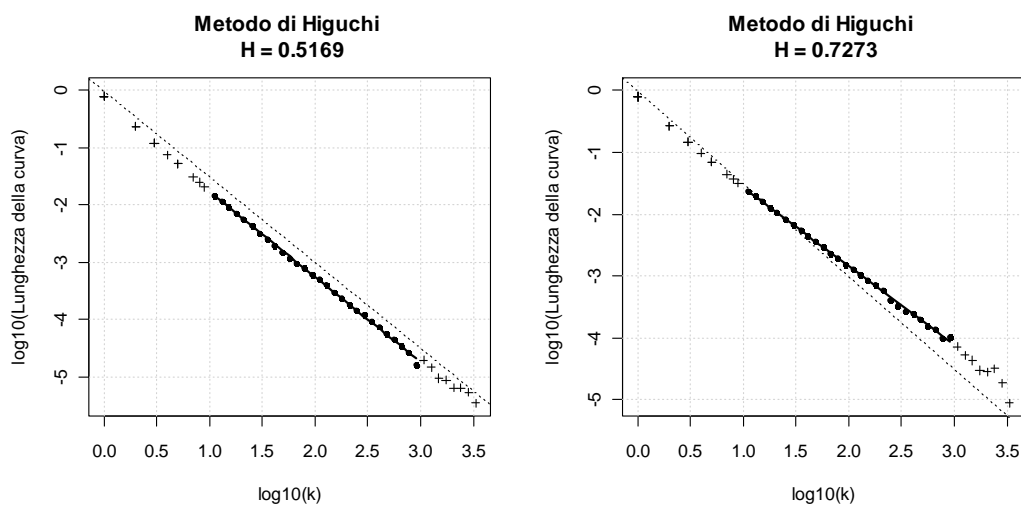


Figura 11.17 - Metodo di Higuchi: applicazione ad una serie simulata di numerosità 10000 con $H = 0.5$ (pannello a sinistra) e $H = 0.75$ (pannello a destra). In entrambi i pannelli, la stima della pendenza della retta di regressione (linea continua) è stata condotta con i punti compresi nell'intervallo temporale logaritmico (1.0-3.0). Le linee tratteggiate indicano le rette con pendenze corrispondenti ad $H = 0.5$ (assenza di lunga memoria).

La Figura 11.18 mostra i box-plot dei valori del parametro di Hurst calcolato su 200 serie simulate con $H = 0.75$ per i tre metodi di stima della varianza aggregata (VA), del rescaled range (R/S) e di Higuchi. Ogni metodo è applicato alle serie complete ("serie"), alle serie private del 10% dei dati, selezionati in modo casuale ("random"), ed alle serie in cui il 10% dei dati è stato sottratto come un intervallo continuo a partire da un punto casuale all'interno della serie ("blocchi"). La distorsione mostrata dai metodi della varianza aggregata e di Higuchi è dovuta alla limitata numerosità del campione, che è stata scelta uguale a 240, ossia la lunghezza di una serie destagionalizzata di 20 anni a scala mensile. Prescindendo dalla distorsione delle stime, i box-plot mostrano che la rimozione del 10% dei valori (sia come valori isolati che come blocchi continui) non influisce in modo rilevante sull'esito della stima. Chiaramente, il risultato delle simulazioni è lontano dall'essere esaustivo, tuttavia supporta le indicazioni dedotte in precedenza da considerazioni concettuali.

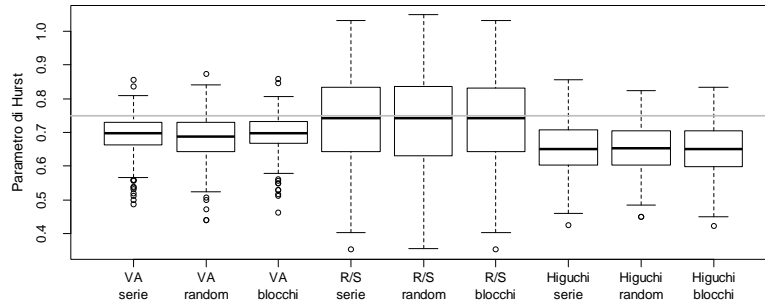


Figura 11.18 - Box-plot dei valori del parametro di Hurst calcolato su 200 serie simulate con $H = 0.75$ (linea grigia) con i tre metodi della varianza aggregata (VA), del rescaled range (R/S) e di Higuchi. Ogni metodo è applicato alla serie completa (“serie”), alla serie privata del 10% dei dati selezionati in modo casuale (“random”) ed alla serie in cui il 10% dei dati è stato sottratto come un intervallo continuo in una posizione casuale all’interno della serie (“blocchi”).

11.4.2 Componenti stagionali

Per la stima delle componenti stagionali della media e della varianza è possibile applicare il metodo descritto da Grimaldi (2004). Questo approccio ha una natura non parametrica; si garantisce così una flessibilità maggiore rispetto ai metodi basati su funzioni analitiche per la definizione delle curve che descrivono gli andamenti stagionali. Per semplicità di esposizione, la notazione usata per la descrizione dell’algoritmo è riferita a serie giornaliere le cui componenti stagionali hanno quindi un periodo di 365 giorni. La notazione può essere agevolmente adattata allorché si considerino delle serie a scala di risoluzione temporale diversa, ad esempio, settimanale (periodo di 52 settimane) o mensile (periodo di 12 mesi).

Il calcolo si articola nelle seguenti sei fasi:

1ª fase: il primo passo consiste in una stima iniziale delle componenti stagionali tramite il cosiddetto “metodo classico”. Data una serie giornaliera di n anni, il metodo classico prevede la costruzione di 365 sotto-serie di numerosità n , ognuna contenente le n osservazioni relative ad un fissato giorno dell’anno. Si calcolano quindi media e varianza (o deviazione standard) di ognuna delle 365 sotto-serie. In formule si ha:

$$\hat{\mu}_t = \frac{1}{n} \sum_{i=1}^n X_{t+365(i-1)} \quad t = 1, \dots, 365 \quad \text{eq. 11.4.9}$$

$$\hat{\sigma}_t^2 = \frac{1}{n-1} \sum_{i=1}^n (X_{t+365(i-1)} - \hat{\mu}_t)^2 \quad t = 1, \dots, 365 \quad \text{eq. 11.4.10}$$

La Figura 11.19 illustra l’applicazione del metodo classico ad una serie di 27 anni di portate giornaliere del fiume Adige registrate a Bronzolo: per ogni giorno dell’anno sono riportate le 27 osservazioni corrispondenti (punti), la media e la deviazione standard. Le sequenze delle 365 medie (linea blu) e deviazioni standard (linea rossa) individuano l’andamento delle due componenti stagionali. La figura evidenzia che le due componenti presentano delle oscillazioni irregolari attribuibili ad una aleatorietà residua dovuta al campionamento (medie e deviazioni standard sono stimate su sotto-serie di numerosità pari a 27) piuttosto che alla stagionalità. I passi successivi dell’algoritmo permettono di eliminare le oscillazioni casuali ed ottenere un andamento regolare congruente con la natura fisica della stagionalità.

2ª fase: indicando con $X_{t,i}$ l’osservazione del giorno t -esimo, $t = 1, \dots, 365$, del generico anno i -esimo, $i = 1, \dots, n$, per ogni giorno dell’anno t si definisce una serie di residui $Y_{t,i}$ ottenuti sottraendo alle osservazioni $X_{t,i}$ la quantità $(\hat{\mu}_t - E[\hat{\mu}_t])$, ossia:

$$Y_{t,i} = X_{t,i} - \hat{\mu}_t + E[\hat{\mu}_t] \quad t = 1, \dots, 365; i = 1, \dots, n \quad \text{eq. 11.4.11}$$

3^a fase: per ogni giorno dell'anno t , gli n valori $Y_{t,i}$ forniscono un'indicazione dello scostamento delle osservazioni $X_{t,i}$ dalla stima della componente stagionale. I valori di $Y_{t,i}$ più elevati sono associati alle osservazioni $X_{t,i}$ che si discostano maggiormente dalla componente stagionale e che influiscono sulla stima causando le oscillazioni che si desidera rimuovere.

Queste considerazioni inducono ad eseguire un aggiornamento della stima delle componenti stagionali ricorrendo ad una media ponderata (in luogo di quella aritmetica) in cui i pesi $k_{t,i}$ tengano conto dei valori degli scostamenti $Y_{t,i}$. Tali pesi sono calcolati tramite una funzione bi-quadratica:

$$\begin{cases} k_{t,i} = (1 - u_{t,i}^2)^2 & u_{t,i} < 1 \\ k_{t,i} = 0 & u_{t,i} \geq 1 \end{cases} \quad t = 1, \dots, 365; i = 1, \dots, n \quad \text{eq. 11.4.12}$$

in cui:

$$u_{t,i} = \frac{|Y_{t,i}|}{f \cdot \text{mediana}|Y_{t,i}|} \quad \text{eq. 11.4.13}$$

il simbolo $|\cdot|$ indica l'operatore valore assoluto, la mediana è calcolata sugli n valori $|Y_{t,i}|$ per ogni fissato valore di t , e f è un coefficiente positivo il cui valore è scelto in modo da ridurre l'influenza dei valori elevati degli scostamenti $|Y_{t,i}|$.

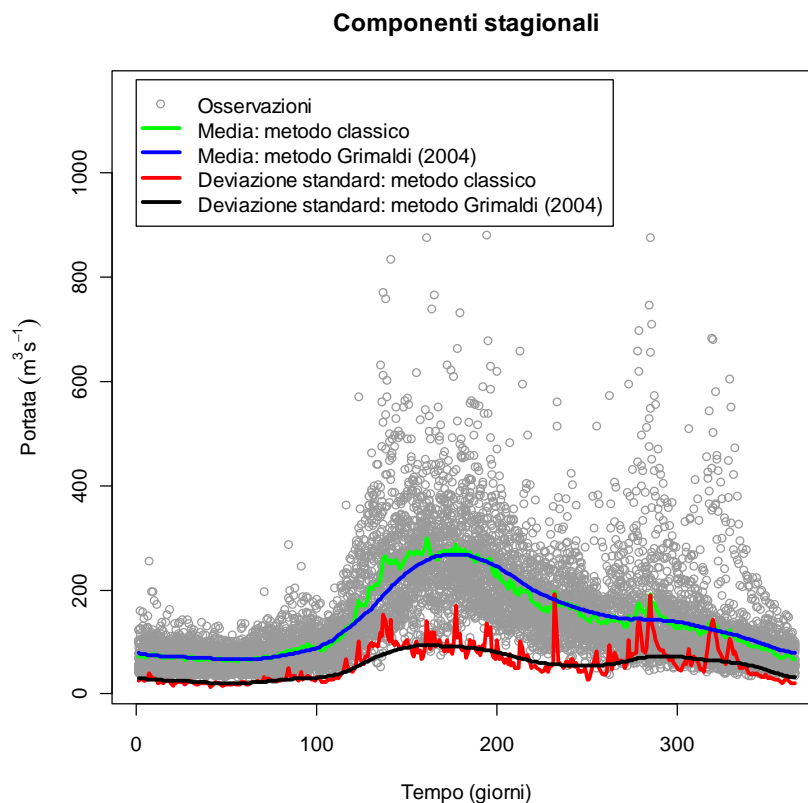


Figura 11.19 - Il grafico riporta le osservazioni corrispondenti ad ogni giorno dell'anno estratte da una serie giornaliera di portate di 27 anni (per ogni giorno sono riportati 27 valori). Le curve rappresentano le componenti stagionali delle media e della deviazione standard ottenute con il metodo classico e con l'algoritmo di Grimaldi (2004).

Il coefficiente f è legato all'efficienza della regressione ponderata basata sui pesi definiti dall'eq. 11.4.12. Gross (1977) suggerisce di adottare un valore uguale a 6 per scostamenti $|Y_{t,i}|$ che mostrino una coda pesante. Lo stesso valore è adottato da Cleveland et al. (1990), mentre Grimaldi (2004) suggerisce di assumere un valore pari a 36 per gli scarti relativi alla componente stagionale della varianza per tenere conto delle maggiore variabilità associata ai momenti del secondo ordine (si veda la differenza, in termini di ampiezza delle oscillazioni, tra la curva verde e quella rossa in Figura 11.19). I pesi da usare rispettivamente per il calcolo delle componenti stagionali della media e della varianza sono indicati nel seguito con la notazione $k_{t,i}^\mu$ e $k_{t,i}^\sigma$, assumendo quindi $f = 6$ per i primi e $f = 36$ per i secondi.

4^a fase: applicando i pesi calcolati al passo precedente, si determinano le componenti stagionali aggiornate tramite le medie ponderate:

$$\hat{\mu}_t^* = \frac{\sum_{i=1}^n X_{t+365(i-1)} k_{t+365(i-1)}^\mu}{\sum_{i=1}^n k_{t+365(i-1)}^\mu} \quad t = 1, \dots, 365, \quad \text{eq. 11.4.14}$$

$$\hat{\sigma}_t^{*2} = \frac{n}{n-1} \frac{\sum_{i=1}^n \left(X_{t+365(i-1)} k_{t+365(i-1)}^\sigma - \frac{1}{n} \sum_{i=1}^n X_{t+365(i-1)} k_{t+365(i-1)}^\sigma \right)^2}{\sum_{i=1}^n k_{t+365(i-1)}^\sigma} \quad t = 1, \dots, 365 \quad \text{eq. 11.4.15}$$

in cui $\frac{1}{n} \sum_{i=1}^n X_{t+365(i-1)} k_{t+365(i-1)}^\sigma$ è la media aritmetica della serie moltiplicata per i pesi $k_{t,i}^\sigma$ e $n/(n-1)$ è un fattore di correzione che si introduce affinché il valore atteso della varianza campionaria $E[\hat{\sigma}_t^{*2}]$ sia uguale al valore σ_t^{*2} relativo alla popolazione.

5^a fase: le sequenze $\hat{\mu}_t^*$ e $\hat{\sigma}_t^{*2}$ sono ulteriormente regolarizzate applicando la tecnica *LOESS*, che consiste nel sostituire ai valori $\hat{\mu}_{t_0}^*$ e $\hat{\sigma}_{t_0}^{*2}$ di ogni fissato giorno t_0 i valori che si ottengono tramite una regressione locale. Le serie così ottenute, \hat{m}_t^* e \hat{s}_t^{*2} , rappresentano le componenti stagionali definitive. I risultati ottenuti sono illustrati nella Figura 11.19. L'effetto di regolarizzazione dell'algoritmo è evidente soprattutto nella componente della deviazione standard, che è più sensibile della media alla presenza di valori anomali, in quanto funzione dei quadrati delle osservazioni.

6^a fase: infine, la serie destagionalizzata è calcolata tramite la relazione:

$$Z_{t+365(i-1)} = \frac{(X_{t+365(i-1)} - \hat{\mu}) - \hat{m}_t^*}{\hat{\sigma}_t^*} \quad t = 1, \dots, 365; i = 1, \dots, n. \quad \text{eq. 11.4.16}$$

in cui $\hat{\mu}$ è la media campionaria della serie osservata.

Nota 11.4. Le finestre di smoothing per la tecnica *LOESS* alla base del metodo devono essere scelte con una procedura di tipo try-and-error, ovvero occorre impiegare diversi valori sino ad ottenere delle componenti regolari prive di oscillazioni non attribuibili alla variabilità stagionale. Per le serie giornaliere, finestre di 60 giorni consentono in genere di filtrare le oscillazioni non stagionali. È tuttavia opportuno controllare l'esito della procedura per finestre di interpolazione comprese indicativamente tra i 30 i 90 giorni. Per le serie mensili è sufficiente usare una finestra di smoothing uguale ad un mese, in ragione della maggiore regolarità della serie rispetto alle serie giornaliere

Nota 11.5. Per serie mensili, il metodo di destagionalizzazione descritto fornisce risultati pressoché uguali al metodo classico in regione della regolarità delle serie (che non mostrano le oscillazioni

casuali presenti nelle serie a scala di risoluzione temporale più fine). Per l'applicazione del metodo descritto a serie mensili, il valore più adeguato per finestre di interpolazione LOESS è pari a 1 mese.

11.4.3 Analisi di non-stazionarietà

11.4.3.1 Test Ljung-Box per la presenza di autocorrelazione

L'applicazione dei test per trend e *change point* richiede che le osservazioni in esame non siano correlate nel tempo. È dunque opportuno valutare preventivamente la significatività che l'ACF della serie non sia significativamente diverso da zero (ipotesi nulla). Anziché verificare l'assenza di correlazione a ogni singolo lag, il test di Ljung-Box (Ljung e Box 1978), che fa parte della famiglia dei *test portmanteau*, controlla l'assenza di autocorrelazione nella serie per un gruppo di lag compresi tra 1 e un prescelto passo S (l'ipotesi nulla è che l'ACF non sia significativamente diversa da zero). La statistica test è data dalla relazione:

$$T_{LB} = N(N+2) \sum_{s=1}^S \frac{\rho^2(s)}{N-s} \quad \text{eq. 11.4.17}$$

in cui N è la numerosità della serie e $\rho(s)$ è l'autocorrelazione campionaria al passo s -esimo. Sotto l'ipotesi nulla T_{LB} si distribuisce asintoticamente come una variabile casuale χ^2 con S gradi di libertà. Valori elevati della statistica, che ricadono agli estremi della coda superiore di questa distribuzione, rappresentano un'evidenza che l'autocorrelazione fino al lag S può essere (globalmente) significativamente diversa da zero.

11.4.3.2 Test per i *change point*

11.4.3.2.1 Test di Pettitt

Dato un campione $\{x_1, \dots, x_T\}$ delle variabili $X_t, t = 1, \dots, T$, il test di Pettitt (1979) permette di rilevare i cambiamenti repentini nelle caratteristiche delle variabili X_t ad un istante t non noto a priori. Il test è basato sui valori assunti dalla seguente statistica:

$$U_{t,T} = \sum_{i=1}^t \sum_{j=t+1}^T \text{sgn}(X_i - X_j) \quad t = 1, \dots, T, \quad \text{eq. 11.4.18}$$

in cui $\text{sgn}(\cdot)$ indica la funzione segno, la quale restituisce il valore 1 se $x > 0$, 0 se $x = 0$ e -1 se $x < 0$. La serie dei valori $\{u_{1,T}, \dots, u_{T,T}\}$ assunti da $U_{t,T}$ permette di definire la statistica per il test dell'ipotesi nulla " H_0 : assenza di *change point*" contro l'ipotesi alternativa " H_1 : presenza di *change point*". Tale statistica è:

$$K_T = \max_{1 \leq t < T} |U_{t,T}| \quad \text{eq. 11.4.19}$$

se la direzione del cambiamento (incremento o decremento repentino) non è nota a priori, ovvero:

$$K_T^+ = \max_{1 \leq t < T} U_{t,T} \quad \text{eq. 11.4.20}$$

$$K_T^- = -\min_{1 \leq t < T} U_{t,T} \quad \text{eq. 11.4.21}$$

allorché sia noto a priori che un cambiamento atteso sia un incremento o un decremento.

Per variabili continue, la statistica $U_{t,T}$ può essere calcolata in modo semplice tramite la relazione ricorsiva (Pettitt 1979):

$$U_{t,T} = U_{t-1,T} + V_{t,T} \quad \text{eq. 11.4.22}$$

per $t = 2, \dots, T$, in cui $V_{i,T} = \sum_{j=1}^T \text{sgn}(X_i - X_j)$ e $U_{i,T} = V_{i,T}$.

Si dimostra che le distribuzioni delle variabili K_T^+ , K_T^- e K_T sotto l'ipotesi nulla sono date dalle relazioni approssimate (Pettitt, 1979):

$$\Pr[K_T^+ > k^+] = \exp\left\{-6(k^+)^2/(T^3 + T^2)\right\} \quad \text{eq. 11.4.23}$$

per K_T^+ (e K_T^- , poiché le due statistiche hanno comportamento simmetrico), e:

$$\Pr[K_T > k] = 2 \sum_{r=1}^{\infty} (-1)^{r+1} \exp\left\{-6kr^2/(T^3 + T^2)\right\} \cong 2 \exp\left\{-6k^2/(T^3 + T^2)\right\} \quad \text{eq. 11.4.24}$$

per K_T , in cui k^+ , k^- e k sono i valori assunti dalle variabili K_T^+ , K_T^- e K_T . Fissato un livello di significatività α , ad esempio 0.05, e calcolato il valore k (oppure k^+ o k^-) per la serie di interesse, se

$\Pr[K_T > k] < \alpha$ (oppure $\Pr[K_T^+ > k^+] < \alpha$), allora H_0 può essere rigettata, altrimenti H_0 non può essere rigettata (a meno di un rischio pari a α di commettere un errore del primo tipo). È utile sottolineare che la distribuzione approssimata (eq. 11.4.24) è accurata alla seconda cifra decimale per valori di probabilità minori di 0.5 (Pettitt 1979), ossia quelli di interesse per rilevare un *change point* significativo ai comuni livelli di significatività usati nelle analisi (0.10, 0.05, 0.001). Qualora si rilevi un *change point* significativo, l'istante t associato al valore k (oppure k^+ o k^-), ossia il punto in cui la serie $\{u_{1,t}, \dots, u_{T,t}\}$ presenta il massimo scostamento rispetto allo zero, rappresenta l'istante temporale in cui si colloca il *change point*.

Le Figura 11.20 mostra due serie simulate con e senza *change point*, pannelli (a) e (b), e le corrispondenti serie di valori $\{u_{1,t}, \dots, u_{T,t}\}$ della statistica $U_{i,t}$, pannelli (c) e (d). I pannelli (a) e (c) si riferiscono ad un campione di 200 elementi simulato da una distribuzione normale standard (media = 0 e varianza = 1). Il pannello (b) mostra un campione i cui primi 50 elementi seguono una distribuzione normale con media = 1 e varianza = 1, mentre i restanti 150 elementi hanno distribuzione normale standard. Le linee blu nei pannelli (a) e (b) indicano i valori medi teorici delle distribuzioni da cui sono campionati i dati. La presenza nella seconda serie del *change point* in corrispondenza della 50-esima osservazione induce un andamento regolare nella corrispondente serie delle $U_{i,t}$, la quale mostra un punto di massimo in prossimità del *change point* e un evidente cambiamento del trend da crescente a decrescente. È utile rilevare che, in generale, i segmenti con pendenze costanti nella serie della statistica $U_{i,t}$, sono indicativi di assenza di cambiamento, mentre cambi repentini di pendenza sono indicativi di un possibile *change point*.

La serie dei valori di $U_{i,t}$, corrispondente invece alla prima serie che è senza *change point*, presenta un andamento irregolare con molteplici picchi. I valori della statistica K_T associati alle due serie (senza e con *change point*) sono rispettivamente uguali a 845 e 4862, a cui corrispondono valori di probabilità, secondo l'eq. 11.4.24, maggiori di 0.5 e uguale a $4.36 \cdot 10^{-8}$. Per la serie senza *change point* non è possibile fornire un valore esatto (e per questo è usata la notazione $\Pr \geq 0.5$) poiché il valore di K_T ricade nel centro della distribuzione della statistica test dove l'approssimazione fornita dall'eq. 11.4.24 non è accurata. Questo limite non costituisce tuttavia un problema, poiché valori elevati di probabilità sono associati a valori piccoli di K_T che indicano la non significatività del *change point*. Per la serie con *change point*, il valore di probabilità uguale a $4.36 \cdot 10^{-8}$ è molto minore di 0.05 (ma anche di 0.001), da cui segue una forte evidenza della presenza di un *change point*.

Una volta individuato un *change point* significativo all'istante t è opportuno ripetere l'analisi di Pettitt sui sottocampioni $\{x_1, \dots, x_t\}$ e $\{x_{t+1}, \dots, x_T\}$ per verificare la presenza di ulteriori *change point*. L'analisi è pertanto ripetuta in modo iterativo finché si rilevano *change point* significativi.

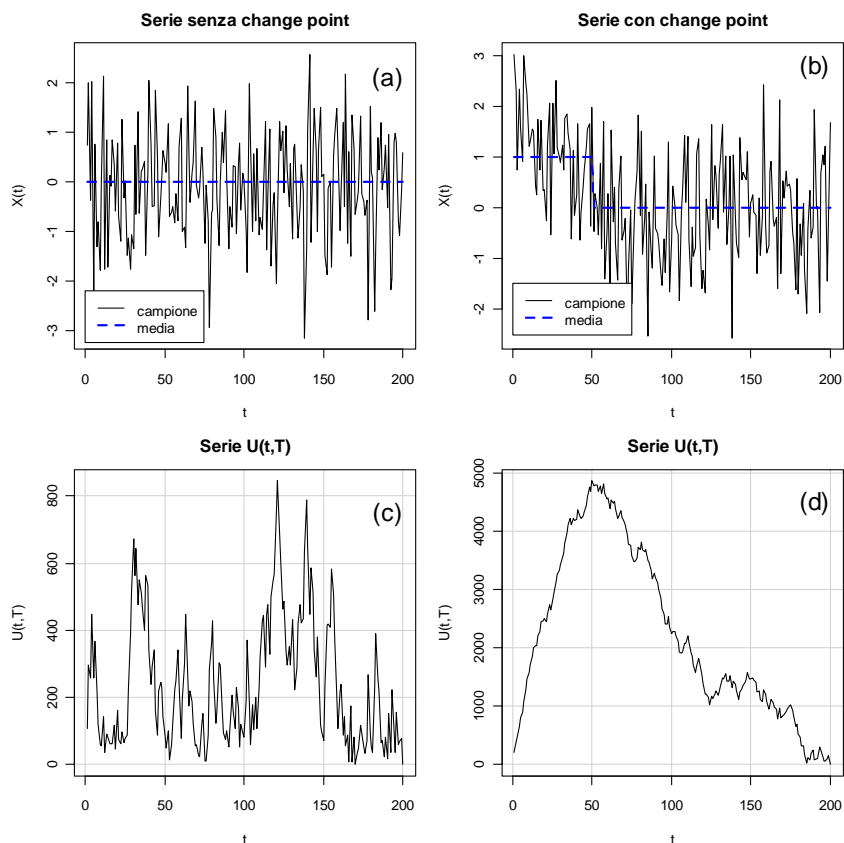


Figura 11.20 - Esempio di serie con e senza *change point* e relative serie della statistica $U_{t,T}$ usata per il calcolo della statistica test di Pettitt.

11.4.3.2.2 Test CUSUM con procedura bootstrap

Questo approccio, impiegato per eseguire l'analisi dei *change point* (e.g., Smadi e Zghoul, 2006), è basato sull'analisi del grafico della serie cronologica delle somme cumulate (CUSUM) degli scarti tra le osservazioni e la media del campione ($X_t - \hat{\mu}_X$). Analogamente alla serie dei termini $U_{t,T}$ su cui si basa il test di Pettitt, la somma cumulata dei termini ($X_t - \hat{\mu}_X$) descrive un andamento che permette di valutare la presenza di *change points*: tratti lineari nel grafico indicano che i termini ($X_t - \hat{\mu}_X$) sono approssimativamente costanti e quindi non ci sono variazioni nella media; cambi di pendenza repentini, ossia tratti lineari crescenti (decescenti) seguiti da tratti lineari decrescenti (crescenti) indicano delle variazioni repentine nel valore medio in prossimità dell'istante in cui si verifica il cambio. Un andamento irregolare, ma stabile intorno allo zero, indica l'assenza di cambiamenti repentini. La Figura 11.21, in cui sono riportate le serie con e senza *change point* già introdotte nel paragrafo precedente e le relative curve CUSUM, mostra come la curva CUSUM relativa alla serie senza *change point* oscilli intorno allo zero, mentre la curva CUSUM relativa alla serie con *change point* sia caratterizzata da un punto di massimo in prossimità del *change point* (50-esima osservazione) e da un evidente cambiamento del trend approssimativamente lineare da crescente a decrescente. La significatività del *change point* è valutata tramite un test delle ipotesi, descritto nel seguito, la cui statistica test è basata sulla serie dei termini ($X_t - \hat{\mu}_X$).

Dal confronto delle curve CUSUM e di quelle associate alle serie $U_{t,T}$ relative al test di Pettitt, emerge un'analogia tra l'approccio CUSUM e quello di Pettitt. La differenza tra i due test consiste nel fatto che il test di Pettitt è basato sui ranghi e dispone di una distribuzione asintotica della statistica test sotto l'ipotesi nulla per il calcolo dei p -value, mentre il test CUSUM è basato sulle osservazioni e non dispone di una distribuzione unica della statistica test, poiché quest'ultima dipende dai valori della grandezza analizzata. Nel test CUSUM, la distribuzione della statistica test sotto l'ipotesi nulla è dunque approssimata dalla distribuzione empirica ottenuta tramite una procedura di ricampionamento bootstrap.

Si illustrano nel seguito i passaggi necessari per l'applicazione del test. Si consideri un campione indipendente e identicamente distribuito delle variabili X_t , con $t = 1, \dots, T$, allora si definiscono, attraverso un processo iterativo, le seguenti somme cumulate S_0, \dots, S_T :

- sia $\hat{\mu}_X = \frac{1}{T} \sum_{i=1}^T X_i$;
- si pone $S_0 = 0$;
- si calcolano in modo ricorsivo le quantità $S_t = S_{t-1} + (X_t - \hat{\mu}_X)$, per $t = 1, \dots, T$.

Riportando in un grafico i valori S_t al variare di t si ottiene la curva CUSUM. Come detto in precedenza, in presenza di un *change point* in media, tale curva presenta un andamento crescente (decescente) seguito da un andamento decrescente (crescente). L'intervallo in cui la curva CUSUM è relativamente lineare indica invece un periodo in cui la media non mostra variazioni. In assenza di *change point* invece la curva oscilla intorno allo zero senza mostrare un andamento definito Figura 11.21.

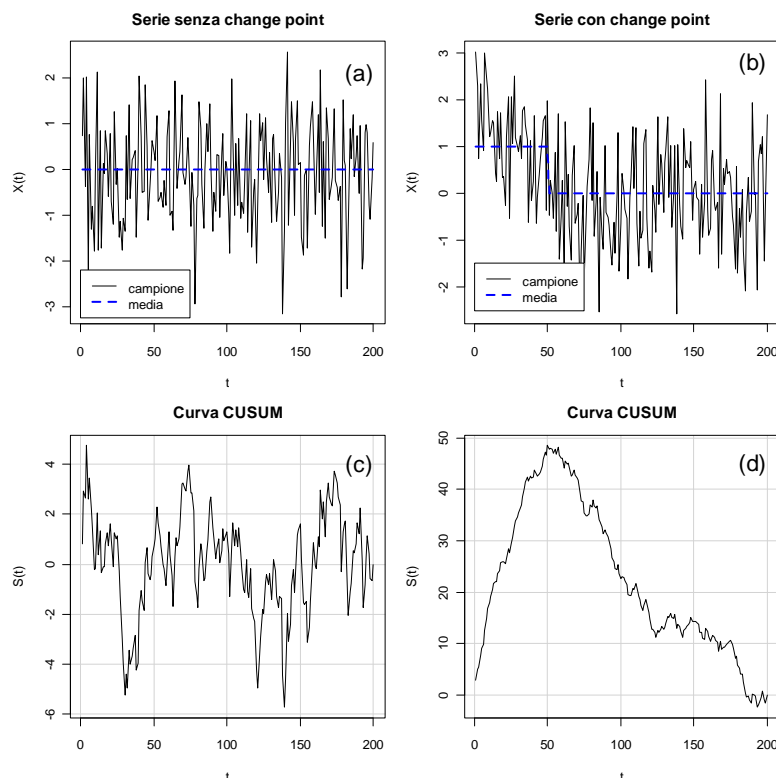


Figura 11.21 - Esempio di serie con e senza *change point* e relative curve CUSUM.

La statistica su cui è eseguito il test è la differenza $S_{\text{diff}} = S_{\text{max}} - S_{\text{min}}$, in cui $S_{\text{max}} = \max_{1 \leq t < T} S_t$ e $S_{\text{min}} = \min_{1 \leq t < T} S_t$. La scelta di utilizzare la quantità S_{diff} come statistica test è dovuta al fatto che in assenza di *change point* la serie delle S oscilla intorno allo zero e dunque S_{diff} assume valori relativamente bassi, mentre in presenza di *change point* S_{diff} assume valori elevati dovuti all'andamento della curva CUSUM (si vedano, ad esempio, le ordinate dei pannelli (c) e (d) in Figura 11.21). Poiché la statistica test dipende dai valori assunti dalla grandezza studiata, i valori critici (e i p -value) della statistica sotto l'ipotesi nulla devono essere definiti tramite una procedura di tipo bootstrap. Si ricorda che il ricampionamento (bootstrap) produce un rimescolamento dei dati che elimina possibili *trend* e *change point* presenti nella serie originale, restituendo delle serie per le quali è valida l'ipotesi nulla. L'approccio bootstrap consiste nei seguenti passi:

- si genera un campione bootstrap X_1^0, \dots, X_T^0 tramite ricampionamento con o senza ripetizione;
- a partire da questo campione si calcola la serie CUSUM S_1^0, \dots, S_T^0 ;

- dalla serie S_1^0, \dots, S_T^0 si calcolano massimo, minimo e statistica test per il campione bootstrap considerato: S_{\max}^0 , S_{\min}^0 e S_{diff}^0 ;
- si ripetono i passi (1)-(3) N volte (ad es., $N = 1000$).

La serie delle N differenze S_{diff}^0 definisce la distribuzione empirica della statistica test utilizzata per accertare che sia valida l'ipotesi H_0 (assenza di *change point*). Posto n il numero di casi in cui $S_{\text{diff}}^0 > S_{\text{diff}}$, allora la quantità $n/N = P$ rappresenta il p -value approssimato per S_{diff} , ossia la probabilità di superamento associata al valore osservato S_{diff} nell'ipotesi che esso segua la distribuzione empirica della statistica test valida sotto l'ipotesi nulla. La Figura 11.22 riporta le distribuzioni bootstrap della statistica test S_{diff} relative alle due serie riportate in Figura 11.21 a-b.

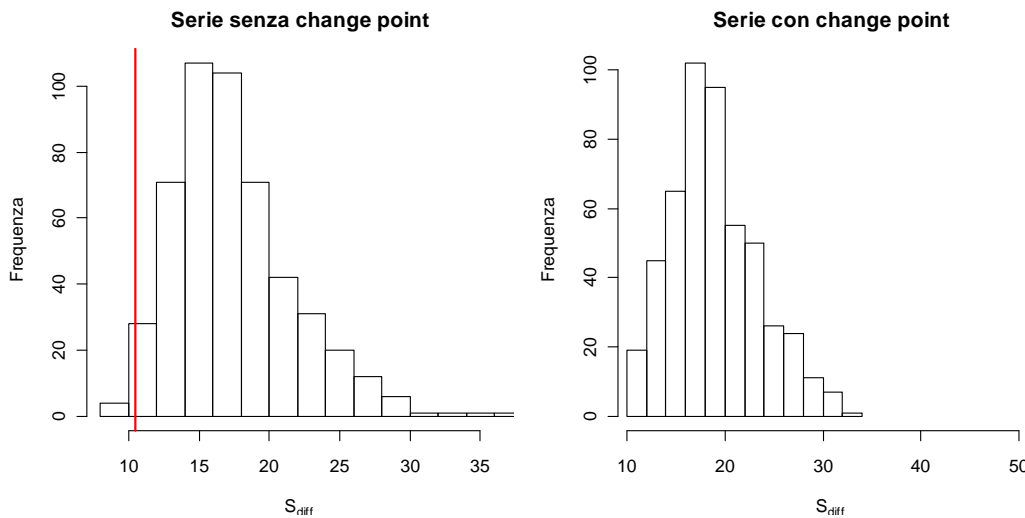


Figura 11.22 - Distribuzioni bootstrap della statistica test S_{diff} ottenute tramite ricampionamento delle due serie con e senza *change point* riportate in Figura 11.21 a-b. Le linee rosse verticali indicano il valore della statistica test calcolata sulle due serie originali.

Per la serie senza *change point*, il valore di S_{diff} (linea rossa) si colloca sulla coda sinistra della distribuzione ed è superato dal 98.6% dei valori di S_{diff} calcolati sulle serie ricampionate. Ciò indica che l'ampiezza massima delle oscillazioni della curva CUSUM relativa alla serie originale in Figura 11.21a è compatibile con le ampiezze che si potrebbero osservare in assenza di *change point*. Per la serie con *change point* in Figura 11.21 b, il valore di S_{diff} ricade a destra della distribuzione bootstrap ed ha un valore molto superiore a quelli ottenuti con ricampionamento ed associati a serie senza *change point*. Ciò indica la presenza di oscillazioni nella curva CUSUM con un andamento simile a quello riportato in Figura 11.21d, legate alla presenza di *change point*. La collocazione del *change point* corrisponde all'indice m della serie per il quale $|S_m| = \max_{1 \leq t < T} |S_t|$, cioè al punto più lontano da zero nel grafico della curva CUSUM.

11.4.3.3 Test per i trend

11.4.3.3.1 Test di Mann-Kendall

Uno dei test non parametrici più usati per il rilevamento di trend monotoni (lineari e non) è il test di Mann-Kendall. Data una serie temporale di osservazioni x_t , con $t = 1, \dots, T$, il test si basa sul confronto delle coppie di osservazioni (x_i, x_j) , con $i > j$ e $i, j \in [1, T]$, per accertare se $x_i > x_j$ ovvero $x_i < x_j$. Indicando con C le coppie del primo tipo e con D quelle del secondo tipo, la statistica test è definita come $S = C - D$, che rappresenta la differenza tra il numero delle volte in cui un'osservazione x_j è superata dalle successive osservazioni e il numero di volte in cui una osservazione x_j è maggiore delle osservazioni successive. La logica del test si basa sull'ipotesi che in assenza di trend una osservazione x_j sia seguita da un numero approssimativamente uguale di osservazioni con valore maggiore e minore di x_j , cosicché $S \cong 0$. La statistica test si esprime in modo formale tramite la relazione:

$$S = \sum_{k=1}^{T-1} \sum_{j=k+1}^T \text{sgn}(X_j - X_k) \quad \text{eq. 11.4.25}$$

in cui $\text{sgn}(x) = 1$ se $X > 0$, 0 se $X = 0$, -1 se $X < 0$, e T è la numerosità. Operativamente, il test è condotto sulla statistica:

$$Z = \begin{cases} (S-1)/V^{0.5} & \text{per } S > 0 \\ 0 & \text{per } S = 0 \\ (S+1)/V^{0.5} & \text{per } S < 0 \end{cases} \quad \text{eq. 11.4.26}$$

in cui V è la varianza di S , data da:

$$V = \frac{T(T-1)(2T+5)}{18} - \frac{\sum_{i=1}^g n_i(n_i-1)(2n_i+5)}{18} + \frac{\sum_{i=1}^g (n_i^2 - n_i)(n_i-2)}{9T(T-1)(T-2)} + \frac{\sum_{i=1}^g (n_i^2 - n_i)}{2T(T-1)} \quad \text{eq. 11.4.27}$$

Nell'eq. 11.4.27, i termini successivi al primo sono introdotti per operare la correzione necessaria in presenza di gruppi di osservazioni uguali che generano i cosiddetti nodi. In particolare, g è il numero dei gruppi di osservazioni con valore uguale (nodi), e n_i è il numero di nodi in un generico gruppo i . La statistica Z è usata in luogo di S poiché si può dimostrare che la sua distribuzione è normale standard. Ciò permette di calcolare in modo semplice i p -value corrispondenti al valore di Z calcolato, o in alternativa calcolare i valori della statistica test sotto l'ipotesi nulla (assenza di trend) per un fissato livello di significatività (e.g., 5%). L'ipotesi nulla è rigettata se il valore di Z calcolato per la serie in esame ha un p -value minore del livello di significatività prescelto, o analogamente se il valore di Z è superiore al valore della variabile normale standard con probabilità di superamento uguale al livello di significatività fissato.

La Figura 11.23 mostra due serie simulate con e senza trend monotono. La statistica test calcolata per la serie stazionaria ha un valore uguale a 0.36 cui corrisponde un p -value uguale a 0.72, il quale indica che il valore della statistica ricade nella parte centrale della distribuzione normale standard valida nell'ipotesi nulla di assenza di trend. La statistica test ed il corrispondente p -value assumono invece valori uguali a -1.72 e 0.08 per la serie con trend. Sebbene il trend imposto sia modesto (come è possibile rilevare visivamente dalla Figura 11.23), il valore della statistica test si colloca sulla coda della distribuzione normale standard.

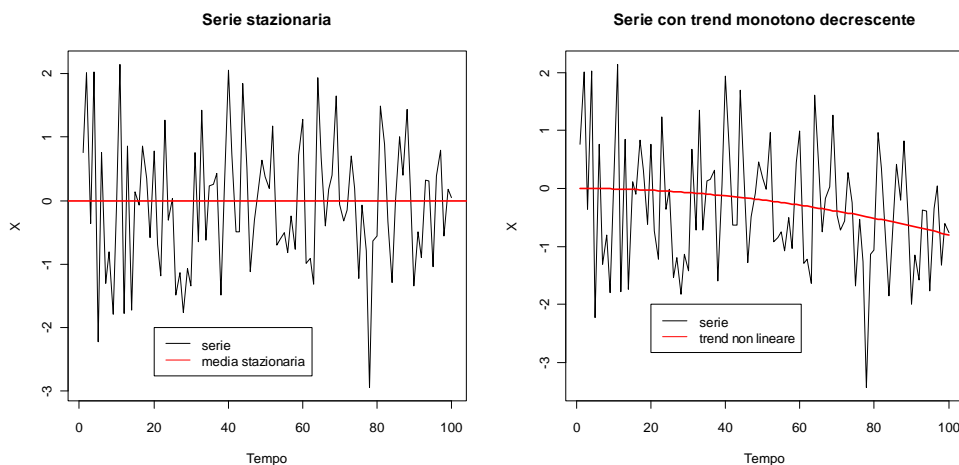


Figura 11.23 - Esempi di serie simulate con e senza trend monotono.

Il valore del p -value indica che il trend è significativo se il livello di significatività prescelto è del 10%, ma non lo è per un livello di significatività del 5%. In questo caso, ed in casi analoghi in cui il p -value si colloca tra due valori assunti comunemente come livelli di significatività (5% e 10%), la presenza del trend dovrebbe essere considerata sospetta, in quanto la statistica test ricade su una coda

della distribuzione, ma in una posizione tale che il rifiuto o il non rifiuto dell'ipotesi nulla dipende dal rischio si è disposti ad accettare di commettere un errore del primo tipo (rigetto dell'ipotesi nulla quando è vera).

11.4.3.3.2 Test di Pearson

Il coefficiente di correlazione ρ_P di Pearson può essere usato per misurare l'associazione lineare tra valori adiacenti in una sequenza di valori ordinati (autocorrelazione), ovvero tra due vettori di osservazioni. L'esistenza di un trend significativo per i valori assunti da una grandezza fisica (Y) nel tempo (X) è spesso valutata definendo la significatività della pendenza $b_1 \equiv \rho_P$ della retta di regressione $Y = b_0 + b_1X$. Il test è di tipo parametrico e si basa sull'ipotesi che il campione abbia distribuzione normale. Qualora questa ipotesi non sia soddisfatta, la variabile è preventivamente trasformata in una variabile distribuita secondo una gaussiana tramite, ad esempio, leggi logaritmiche o di potenza, in relazione alle proprietà di skewness della serie originale, o tramite la trasformazione del quantile normale. Ricordando che il coefficiente di correlazione tra le due variabili X e Y ha espressione:

$$\rho_P = \frac{T \sum_{t=1}^T x_t y_t - \sum_{t=1}^T x_t \sum_{t=1}^T y_t}{\left(T \sum_{t=1}^T x_t^2 - \left(\sum_{t=1}^T x_t \right)^2 \right)^{0.5} \left(T \sum_{t=1}^T y_t^2 - \left(\sum_{t=1}^T y_t \right)^2 \right)^{0.5}} \quad \text{eq. 11.4.28}$$

in cui T è la numerosità del campione, segue che la statistica usata per testare la significatività del coefficiente di Pearson è:

$$r = \frac{\rho_P}{\left[(1 - \rho_P^2) / (T - 2) \right]^{0.5}} \quad \text{eq. 11.4.29}$$

Sotto l'ipotesi nulla H_0 (correlazione non significativa) segue una distribuzione t di Student¹⁴ con $(T-2)$ gradi di libertà. Una volta calcolata la statistica test sul campione analizzato, il test è svolto con la procedura descritta nel paragrafo 11.1.5 ed applicata nel paragrafo precedente per il test di Mann-Kendall.

Considerando le due serie riportate in Figura 11.23, i valori delle statistiche test (e i rispettivi p -value) assumono i valori $r = 0.13$ (p -value = 0.90) per la serie stazionaria, e $r = -2.13$ (p -value = 0.04) per la serie con trend. Questi risultati sono analoghi a quelli ottenuti con il test di Mann-Kendall, ma il test di Pearson rigetta l'ipotesi nulla sia al 10% sia al 5% di livello di significatività seppure con un p -value prossimo al 5%.

11.4.3.3.3 Test di Spearman

Il coefficiente di correlazione di Spearman ρ_S è l'equivalente non parametrico del coefficiente di correlazione di Pearson. Analogamente al test di Mann-Kendall, il test di Spearman non richiede preventive trasformazioni dei dati, poiché è basato sui ranghi, ossia sugli indici che denotano le posizioni occupate dalla osservazioni nel campione ordinato (in modo crescente o decrescente). Date due serie di osservazioni x_t e y_t , con $t = 1, \dots, T$, definendo i rispettivi ranghi con r_{xt} e r_{yt} , il coefficiente di correlazione di Spearman è dato dall'espressione:

¹⁴ Si definisce distribuzione t di Student la legge di probabilità di una variabile casuale X con densità di probabilità:

$$f_X(x; \nu) = \frac{\Gamma[(\nu+1)/2]}{\sqrt{\pi\nu}\Gamma[\nu/2]} \frac{1}{\left[(x^2/\nu) + 1 \right]^{(\nu+1)/2}}$$

con $-\infty < x < \infty$ e in cui ν è il parametro che indica i cosiddetti gradi di libertà della distribuzione e $\Gamma(r)$ è la funzione gamma completa:

$$\Gamma(r) = \int_0^{\infty} t^r \exp(-t) dt \quad \text{per } r > 0 \cdot \\ = 0 \quad \text{per } r \leq 0$$

Per campioni con numerosità maggiore di 30 la distribuzione t di Student converge alla distribuzione normale.

$$\rho_s = 1 - \frac{6 \sum_{t=1}^T (r_{xt} - r_{yt})^2}{T^3 - T} \quad \text{eq. 11.4.30}$$

La statistica per testare la significatività di ρ_s è:

$$s = \rho_s \sqrt{T-1} \quad \text{eq. 11.4.31}$$

la quale sotto l'ipotesi nulla H_0 (correlazione non significativa) e $T > 30$ segue una distribuzione normale standard. L'applicazione alle due serie mostrate in Figura 11.23 restituisce valori (e p -value) della statistica test pari a $s = 0.33$ (p -value = 0.74) per la serie stazionaria, e $r = -1.71$ (p -value = 0.09) per la serie con trend. Questo risultato è analogo a quello ottenuto con il test di Mann-Kendall.

11.4.4 Analisi degli eventi estremi

11.4.4.1 Introduzione

La definizione di un modello per i valori estremi può essere ricondotto operativamente alla scelta di una opportuna funzione di ripartizione che possa essere usata per la stima dei quantili con prefissati tempi di ritorno. Sebbene la scelta tra diverse distribuzioni sia ampia, come osservato da Stedinger e Griffis (2008), le differenze tra gli stimatori dei quantili corrispondenti a diverse leggi di probabilità potenzialmente adatte a descrivere un campione sono in genere sostanzialmente inferiori all'incertezza di campionamento dovuta alla numerosità del campione (e.g., Hosking e Wallis 1997, pp. 134–138, 142). Inoltre, ogni famiglia parametrica rappresenta soltanto un'approssimazione della reale distribuzione dai cui proviene il campione, che rimane comunque incognita. Ne segue che il punto fondamentale è usare una distribuzione che sia consistente con le proprietà dei dati analizzati (dominio di esistenza, comportamento delle code, ecc.) e stimare i suoi parametri tenendo conto possibilmente delle proprietà del processo analizzato (precipitazione, deflusso, ecc.) e dei dati disponibili.

Seguendo questo approccio, l'adozione della distribuzione dei valori estremi generalizzata (GEV) per i massimi annuali e della distribuzione di Pareto generalizzata (GPD) per i valori sopra soglia rappresenta una scelta che permette di coniugare una consolidata base teorica (teoria dei valori estremi; e.g., Coles (2001)) e la flessibilità di distribuzioni a tre parametri in grado di rappresentare diverse tipologie di dati. Nelle sezioni seguenti sono sinteticamente descritte le due distribuzioni GEV e GPD, evidenziandone le principali proprietà utili per la stima dei parametri, la selezione, e la diagnostica. Il metodo adottato per la stima dei parametri è quello della massima verosimiglianza, il quale presenta proprietà teoriche utili nella procedura di inferenza (Coles 2001; pp. 27-43).

11.4.4.2 Modello per i massimi: distribuzione dei valori estremi generalizzata (GEV)

Si consideri una sequenza di variabili indipendenti e identicamente distribuite X_1, \dots, X_n con una distribuzione comune F , in cui le X_i sono i valori assunti da un processo misurato ad una scala temporale regolare (e.g., temperature medie giornaliere, portate medie giornaliere, ecc.) e n è il numero di osservazioni in un fissato periodo (e.g., un anno). Assumendo $M_n = \max\{X_1, \dots, X_n\}$, per la teoria degli eventi estremi, al tendere di n ad infinito, la quantità riscalata $M_n^* = [(M_n - b_n) / a_n]$ (in cui $a_n > 0$ è una costante di scala e $b_n > 0$ una costante di posizione) converge ad una delle tre distribuzioni degli eventi estremi, note come famiglie di Gumbel, Fréchet e Weibull (e.g., Kottegoda e Rosso 2008, pp. 415-422). Queste possono essere combinate in un'unica famiglia detta dei valori estremi generalizzata (GEV):

$$G(Z \leq z) = \exp\left(-\left(1 + \xi\left(\frac{z - \mu}{\sigma}\right)\right)^{-1/\xi}\right) \quad \text{eq. 11.4.32}$$

definita nell'insieme $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$, dove i parametri soddisfano $-\infty < \mu < \infty$, $\sigma > 0$ e $-\infty < \xi < \infty$. Il modello GEV ha tre parametri: un parametro di posizione μ , uno di scala σ , e uno di forma ξ . In questa parametrizzazione, le distribuzioni di Fréchet e Weibull corrispondono rispettivamente ai casi $\xi > 0$ e $\xi < 0$. Il valore $\mu - \sigma/\xi$ rappresenta il limite inferiore per distribuzione di Fréchet e il limite

superiore per la distribuzione di Weibull. Il sottoinsieme della famiglia GEV con $\xi = 0$ è interpretato come limite della eq. 11.4.32 per $\xi \rightarrow 0$ e conduce alla famiglia di Gumbel, infatti:

$$\begin{aligned} \lim_{\xi \rightarrow 0} G(z) &= \lim_{\xi \rightarrow 0} \exp\left(-\left(1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right)^{-1/\xi}\right) = \\ &= \lim_{\xi \rightarrow 0} \exp\left(-\exp\left(-\frac{1}{\xi} \log\left(1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right)\right)\right) = \\ &= \lim_{\xi \rightarrow 0} \exp\left(-\exp\left(-\frac{\log\left(1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right)}{\xi\left(\frac{z-\mu}{\sigma}\right)}\right)\right) = \\ &= \exp\left(-\exp\left(-\left(\frac{z-\mu}{\sigma}\right)\right)\right) \end{aligned}$$

per cui

$$G(Z \leq z) = \exp\left(-\exp\left(-\left(\frac{z-\mu}{\sigma}\right)\right)\right) \quad -\infty < z < \infty \quad \text{eq. 11.4.33}$$

L'unificazione delle tre famiglie semplifica l'implementazione delle procedure di stima e verifica e riconduce l'analisi allo studio di un'unica distribuzione. Attraverso l'inferenza su ξ , il tipo di comportamento di coda è determinato dai dati stessi, senza necessità di un giudizio soggettivo a priori su quale famiglia dei valori estremi adottare. Inoltre, l'incertezza inerente la stima di ξ è una misura del livello di incertezza legato a quale delle tre famiglie originali (Gumbel, Fréchet, Weibull) sia la più appropriata a descrivere la distribuzione dei dati analizzati.

Dal punto di vista dell'inferenza, occorre ricordare alcuni risultati relativi all'applicabilità delle proprietà asintotiche degli stimatori di massima verosimiglianza:

- quando $\xi > -0.5$, gli stimatori di massima verosimiglianza sono regolari, nel senso che hanno le proprietà asintotiche standard, e.g. distribuzione Gaussiana degli stimatori;
- quando $-1 < \xi < -0.5$, gli stimatori di massima verosimiglianza sono generalmente ottenibili, ma non hanno le proprietà asintotiche standard;
- quando $\xi < -1$, gli stimatori di massima verosimiglianza sono difficilmente ottenibili in quanto la funzione di massima verosimiglianza ha un andamento non regolare caratterizzato dalla presenza di più massimi locali.

Sebbene la regolarità degli stimatori non sia verificata per $\xi < -0.5$, questo limite non rappresenta un ostacolo all'applicazione del metodo della massima verosimiglianza, poiché valori negativi del parametro di forma ξ restituiscono distribuzioni con coda destra superiormente limitata, che raramente trovano applicazione nella modellazione dei valori estremi di variabili idrologiche (Coles 2001).

11.4.4.3 Modello sopra soglia: distribuzione generalizzata di Pareto (GPD)

Si consideri una sequenza di variabili indipendenti e identicamente distribuite X_1, \dots, X_n con una distribuzione comune F , e sia $M_n = \max\{X_1, \dots, X_n\}$ tale che, per n tendente ad infinito, valga la eq. 11.4.32. Allora, per un valore di soglia u abbastanza elevato (da definire in base al campione), la distribuzione di $Y = (X - u)$, condizionata a $X > u$, è approssimativamente:

$$H(Y \leq y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \quad \text{eq. 11.4.34}$$

definita nell'insieme $\{y : y > 0 \text{ e } (1 + \xi y / \tilde{\sigma}) > 0\}$, in cui:

$$\tilde{\sigma} = \sigma + \xi(u - \mu) \quad \text{eq. 11.4.35}$$

La famiglia rappresentata dall'eq. 11.4.34 è detta distribuzione di Pareto generalizzata (GPD). Se i massimi (e.g., annuali) hanno distribuzione GEV, allora le eccedenze sopra soglia hanno una corrispondente distribuzione GPD. Inoltre, i parametri della GPD sono univocamente determinati da quelli della GEV associata. In particolare, il parametro ξ è identico per le due distribuzioni, mentre $\tilde{\sigma}$ e σ sono legate tramite la eq. 11.4.35. La dualità tra GEV e GPD implica che il parametro di forma ξ è dominante nel determinare il comportamento qualitativo della distribuzione. Se $\xi < 0$ la distribuzione ha limite superiore $u - \tilde{\sigma}/\xi$. Infatti, se $\xi > 0$ la distribuzione è superiormente illimitata. La distribuzione è illimitata anche nel caso $\xi = 0$, che deve essere interpretato come limite per $\xi \rightarrow 0$ della eq. 11.4.34. Tale limite restituisce:

$$\begin{aligned} \lim_{\xi \rightarrow 0} H(Y \leq y) &= \lim_{\xi \rightarrow 0} 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right) = \\ \lim_{\xi \rightarrow 0} 1 - \exp\left[-\frac{1}{\xi} \frac{y}{\tilde{\sigma}} \frac{\tilde{\sigma}}{y} \log\left(\frac{\xi y}{\tilde{\sigma}}\right)\right] &= \\ \lim_{\xi \rightarrow 0} 1 - \exp\left[-\frac{\log\left(\frac{\xi y}{\tilde{\sigma}}\right)}{\frac{\xi y}{\tilde{\sigma}}} \frac{y}{\tilde{\sigma}}\right] &= \\ 1 - \exp\left[-\frac{y}{\tilde{\sigma}}\right] \end{aligned}$$

per cui:

$$H(Y \leq y) = 1 - \exp\left(-\frac{y}{\tilde{\sigma}}\right), \quad y > 0, \quad \text{eq. 11.4.36}$$

corrispondente alla distribuzione esponenziale con parametro $1/\tilde{\sigma}$.

Appare utile richiamare due proprietà teoriche della distribuzione GPD che possono essere usate nell'inferenza per la scelta della soglia. Questa deve essere sufficientemente elevata per garantire la validità delle ipotesi di base (le osservazioni siano indipendenti e descrivano un processo estemale) e non introdurre distorsione nelle stime, ma al tempo stesso dovrebbe essere tale da non generare campioni di numerosità ridotta e dunque stime troppo incerte. Due sono le tecniche usate abitualmente: una basata su un'analisi esplorativa che precede la stima dei parametri, ed una seconda che valuta la stabilità dei valori dei parametri, stimando il modello per un intervallo di soglie.

La prima tecnica è basata sulla media della GPD:

$$E[Y] = \frac{\tilde{\sigma}}{1-\xi} < \infty \quad \text{per } \xi < 1 \quad \text{eq. 11.4.37}$$

Se il modello GPD è valido per una soglia u_0 , allora lo è anche per una soglia $u > u_0$, a meno di un cambiamento di scala del parametro $\tilde{\sigma}_u$ tramite la relazione:

$$E[X - u | X > u] = \frac{\tilde{\sigma}_u}{1-\xi} = \frac{\tilde{\sigma}_{u_0} + \xi u}{1-\xi} \quad \text{eq. 11.4.38}$$

ottenuta dall'eq. 11.4.35. Allora, per $u > u_0$, $E[X - u | X > u]$ è una funzione lineare della soglia. Inoltre, $E[X - u | X > u]$ è semplicemente la media delle eccedenze sopra la soglia u , per la quale la media campionaria rappresenta una stima empirica. Quindi, il modello GPD è appropriato per i valori di u tali che la media delle eccedenze varia linearmente. Il luogo dei punti

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u) \right) : u < x_{\max} \right\}$$

in cui $x_{(1)}, \dots, x_{(n_u)}$ sono le n_u osservazioni che eccedono u , e x_{\max} è il valore massimo delle osservazioni, è detto grafico “*mean residual life*”. Sopra la soglia u_0 per la quale la GPD è valida, il grafico dovrebbe essere approssimativamente lineare in u .

Il secondo metodo di selezione di u è basato sui seguenti argomenti. Se la distribuzione GPD è valida per una soglia u_0 , allora lo è anche per una soglia $u > u_0$, con parametro di forma invariato e parametro di scala che varia linearmente con u secondo la relazione $\tilde{\sigma}_u = \tilde{\sigma}_{u_0} + \zeta(u - u_0)$. Il parametro $\tilde{\sigma}_u$ può essere riparametrizzato come $\tilde{\sigma}^* = \tilde{\sigma}_u - \zeta u$, costante in u . Quindi, $\tilde{\sigma}^*$ e ζ dovrebbero essere costanti sopra u_0 se questa è una soglia valida affinché le eccedenze seguano una distribuzione GPD.

La Figura 11.24 riporta i grafici “*mean residual life*” e i grafici dei parametri di scala $\tilde{\sigma}^*$ e forma ζ al variare della soglia u per un campione simulato di numerosità 120, in cui 20 osservazioni sotto il valore di soglia $u = 3$ seguono una distribuzione beta¹⁵ mentre i valori sopra la soglia seguono una distribuzione GPD. Il grafico “*mean residual life*” mostra un andamento approssimativamente lineare crescente a partire da $u \cong 3$ fino a $u \cong 7$, cui segue un andamento irregolare dovuto alla riduzione della numerosità del campione al crescere della soglia. I grafici dei parametri di scala $\tilde{\sigma}^*$ e forma ζ al variare della soglia u mostrano chiaramente che i valori di $\tilde{\sigma}^*$ e ζ si stabilizzano per soglie superiori a 3. Il valore minimo della soglia u a partire dal quale si manifesta un comportamento lineare crescente nel grafico “*mean residual life*” e un comportamento stazionario nel grafico di $\tilde{\sigma}^*$ e ζ in funzione di u , rappresenta il valore di soglia al di sopra dal quale è possibile assumere che la distribuzione GPD descriva adeguatamente il campione.

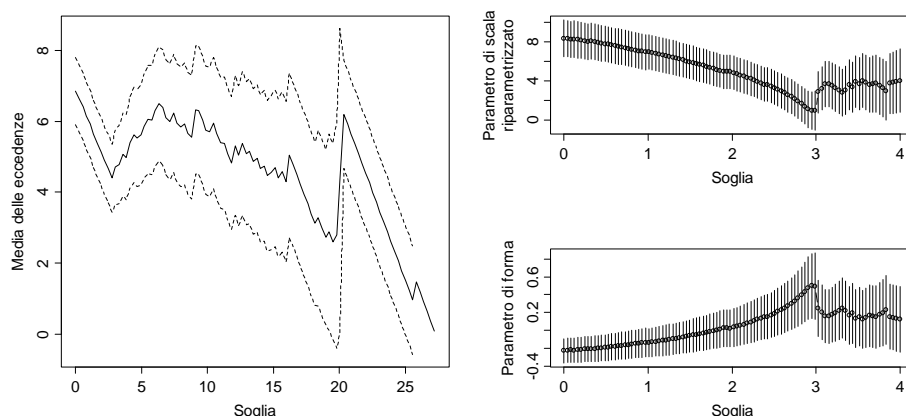


Figura 11.24 - Grafico “*mean residual life*” e grafici dei parametri di scala e forma in funzione della soglia per la serie simulata descritta nel testo.

11.4.5 Curve Intensità-Durata-Frequenza (IDF) basate sulle proprietà di scala della precipitazione

Lo studio volto a definire le altezze di pioggia per una determinata durata ed un fissato tempo di ritorno T_r (o probabilità di superamento) può essere eseguito impiegando la metodologia descritta in Salvadori e De Michele (2001) e Salvadori et al. (2007, pp. 39-50), basata sulle proprietà di scala

¹⁵ Si definisce distribuzione beta la legge di probabilità di una variabile casuale X con densità di probabilità:

$$f_X(x; b, p, q) = \frac{x^{p-1}(b-x)^{q-1}}{b^{p+q-1}B(p, q)}$$

con $0 \leq x \leq b < \infty$ e in cui $p > 0, q > 0$ e $B(p, q)$ è la funzione beta:

$$B(p, q) = \int_0^1 x^{p-1}(1-x)^{q-1} dx.$$

delle distribuzioni GEV e GPD introdotte nel paragrafo precedente. Come illustrato nel paragrafo 11.3.5, il metodo tradizionale per la definizione delle curve IDF richiede la stima dei parametri di una distribuzione per i dati corrispondenti ad ogni durata (generalmente 1, 3, 6, 12, 24 ore) e la successiva stima dei parametri di una relazione intensità-durata (o altezza-durata). L'approccio introdotto da Burlando e Rosso (1996), ed aggiornato in seguito da Salvadori e De Michele (2001), associa all'evidenza empirica dei risultati una corretta base teorica fondata sulla teoria degli eventi estremi, e fornisce una rappresentazione delle curve IDF basata su un numero di parametri ridotto. In particolare, il metodo consiste nell'esprimere il quantile dell'altezza di precipitazione H relativo ad una durata e ad un tempo di ritorno prefissati tramite la funzione inversa della distribuzione GEV, i cui parametri variano con la durata secondo una semplice legge di potenza. Ricordando l'eq. 11.4.32, la distribuzione GEV fornisce la seguente espressione per la probabilità di non superamento di H :

$$\Pr[H \leq h] = \exp\left(-\left(1 + \xi\left(\frac{h - \mu}{\sigma}\right)\right)^{-1/\xi}\right) \quad \text{eq. 11.4.39}$$

Il valore dell'altezza di precipitazione h per un fissato tempo di ritorno $T_r = 1/(1 - \Pr[H \leq h])$ può essere desunto direttamente dalla eq. 11.4.39 invertendo la relazione, ed ottenendo:

$$h(T_r) = \mu + \frac{\sigma}{\xi} \left(\left(-\log\left(1 - \frac{1}{T_r}\right) \right)^{-\xi} - 1 \right) \quad \text{eq. 11.4.40}$$

Indicando con $D' = rD$ una generica scala temporale, in cui r è il rapporto di scala, e con $H_{D'}$ il valore del processo osservato alla scala D' , la dipendenza di h dalla durata D può essere introdotta tramite le seguenti leggi di potenza per i parametri:

$$\begin{cases} \mu_{D'} = \mu_D r^\delta \\ \sigma_{D'} = \sigma_D r^\delta \\ \xi_{D'} = \xi_D \end{cases} \quad \text{eq. 11.4.41}$$

Allorché le precedenti relazioni sono supportate dall'evidenza empirica, la distribuzione di $H_{D'}$ è uguale alla distribuzione di $r^\delta H_D$ (in cui H_D è il processo osservato alla scala D) e i suoi parametri possono essere calcolati tramite le eq. 11.4.41 per una qualsiasi durata. Questo approccio permette di calcolare $h(T_r)$ per una qualunque durata, a partire dai valori stimati dell'esponente di scala δ e dei parametri della distribuzione GEV per le durate osservate. Occorre osservare che le proprietà di invarianza di scala della precipitazione valgono in genere per intervalli di durate tipiche, ad esempio tra 5 minuti e 1 ora o tra 1 e 12 ore. Ogni intervallo i può essere caratterizzato da un valore δ_i che specifica il comportamento di invarianza di scala nell'intervallo stesso. Le variazioni del valore di δ nei diversi intervalli indica dunque un cambiamento delle proprietà di scala del fenomeno, le quali possono essere associate a variazioni delle proprietà fisiche della pioggia alle diverse scale. Ad esempio, variazioni del valore di δ tra le scale sub-orarie e quelle al di sopra dell'ora potrebbero essere associate al fatto che a scala sub-oraria le altezze massime annuali sono legate prevalentemente a precipitazioni di natura convettiva, mentre alle scale superiori il fenomeno dominante è di natura frontale. Chiaramente, eventuali conclusioni sul legame tra le proprietà fisiche della pioggia e le variazioni delle proprietà di scala richiedono ulteriori analisi e approfondimenti. Tuttavia, le possibili variazioni di δ al variare delle scale temporali possono indicare la presenza di cambiamenti nella struttura del fenomeno e suggerire la necessità di ulteriori analisi. L'approccio tradizionale descritto nel paragrafo 11.3.5 non permette questo tipo di interpretazione poiché è basato su curve di derivazione puramente empirica.

La definizione delle curve IDF si basa sulla stima dei parametri di distribuzioni per i massimi annuali di altezza (intensità) di precipitazione a diverse scale temporali. Le difficoltà causate dalla presenza di dati mancanti possono essere ricondotte alle stesse esposte per l'analisi dei valori estremi.

11.5 Bibliografia dell'approfondimento

11.5.1 Definizioni e concetti di base

- Embrechts P, McNeil AJ, Straumann D (1999) Correlation: Pitfalls and alternatives. *Risk Magazine*, 12(5) 69–71
- Kanji GK (2006) 100 statistical tests (third edition). SAGE Publications
- Kottegoda NT, Rosso R (2008) *Applied Statistics for Civil and Environmental Engineers* (second edition). Wiley-Blackwell.
- Nelsen RB (2006) *An Introduction to Copulas* (second edition). Springer
- Piccolo D (2004) *Statistica per le decisioni*. Il Mulino

11.5.2 Serie temporali e stazionarietà

- Akaike H (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723
- Brockwell PJ, Davis RA (2002) *Introduction to time series and forecasting* (second edition). Springer
- Chatfield C (2000) *Time series forecasting*. Chapman & Hall / CRC
- Cleveland WS (1993) *Visualizing Data*. Hobart Press, New Jersey
- Cleveland BR, Cleveland WS, McRae JE, Terpenning I (1990) STL: A seasonal trend decomposition procedure based on loess. *Journal of Official Statistics*, 6, 3–73
- Feder J (1988) *Fractals*, Plenum Press
- Grimaldi S., (2004) Linear parametric models applied to daily hydrological series. *Journal of Hydrologic Engineering*, 9(5), 383–391
- Gross AM (1977) Confidence intervals for bisquare regression estimates. *Journal of the American Statistical Association*, 72(358), 341–354 Published
- Hosking JRM (1984) Modeling persistence in hydrological time series using fractional differencing. *Water Resources Research*, 20(12), 1898–1908
- Higuchi T (1988) Approach to an irregular time series on the basis of the fractal theory, *Physica D*, 31, 277–283
- Hipel KW, McLeod AI (1994) *Time series modelling of water resources and environmental systems*, Elsevier, Amsterdam, The Netherlands
- Hurst HA, Long Term Storage Capacity of Reservoirs. *Transactions of the American Society of Civil Engineers*, 116, 770–799
- Ljung GM, Box GEP (1978) On a Measure of a Lack of Fit in Time Series Models. *Biometrika*, 65, 297–303
- Mandelbrot BB, Taqu MS (1979) Robust R/S analysis of long serial correlation. Proc. 42nd Session of the ISI, Manila, Book 2, 69–99
- Montanari A (1996) *Modellistica stocastica di variabili idrologiche affette da persistenza a lungo termine*. Tesi di Dottorato, VIII ciclo, Milano
- Montanari A, Rosso R, Taqu MS (2000) A seasonal fractional ARIMA model applied to the Nile River monthly flows at Aswan. *Water Resources Research*, 36(5), 1249–1265
- Montanari A, Taqu MS, Teverovski V (2000) Estimating long-range dependence in the presence of periodicity: An empirical study, *Mathematical and Computer Modelling*, 29, 217–228
- R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org/>.
- Schwarz G (1978) Estimating the dimension of a model. *Ann. Statist.*, 6, 461–464
- Shumway RH, Stoffer DS (2006) *Time series analysis and its applications*. Springer, New York
- Taqu MS, Teverovsky V, Wilinger W (1995) Estimators for long-range dependence: an empirical study. *Fractals*, 3(4), 785–788
- Tukey JW (1997) *Exploratory data analysis*. Addison-Wesley, Reading, MA, 688 pp

Villarini G, Serinaldi F, Smith JA, Krajewski WF (2009) On the stationarity of annual flood peaks in the continental United States during the 20th century, *Water Resources Research*, doi:10.1029/2008WR007645, in press.

11.5.3 *Analisi dei trend e change points*

- Aksoy H, Erdem Unal N, Alexandrov V, Dakova S, Yoon J (2008a), *Hydrometeorological analysis of north western Turkey with links to climate change*. *International Journal of Climatology*, 28(8), 1047-1060
- Aksoy H, Gedikli A, Unal NE, Kehagias A (2008b) Fast segmentation algorithms for long hydrometeorological time series. *Hydrological Processes*, H22(23H), 4600-4608
- HBox GEPH, HCoX DRH (1964) H An analysis of transformations H. *Journal of the Royal Statistical Society, Series B* 26, 211–246
- Brunetti M, Buffoni L, Maugeri M, Nanni T (2000) Precipitation intensity trends in northern Italy. *International Journal of Climatology*, 20(9), 1017-1031
- Brunetti M, Colacino M, Maugeri M, Nanni T (2001a) Trends in the daily intensity of precipitation in Italy from 1951 to 1996. *International Journal of Climatology*, 21(3), 299-316
- Brunetti M, Maugeri M, Nanni T (2001b) Changes in total precipitation, rainy days and extreme events in northeastern Italy. *International Journal of Climatology*, 21(7), 861-871
- Brunetti M, Maugeri M, Nanni T, Navarra A (2002) Droughts and extreme events in regional daily Italian precipitation serie. *International Journal of Climatology*, 22(5) 543-558
- Brunetti M, Maugeri M, Monti F, Nanni T (2006) Temperature and precipitation variability in Italy in the last two centuries from homogenised instrumental time series. *International Journal of Climatology*, 26(3), 345-381
- Peterson TC (2005) Climate Change Indices. *WMO Bulletin*, 54 (2), 83-86
- Chiew F, Siriwardena L (2005) *Trend: Trend/change detection software*. CRC (Catchment hydrology), F. Chiew ed., Australia
- Cleveland WS (1993) *Visualizing Data*. Hobart Press, New Jersey
- Cleveland WS (1994) *The Elements of Graphing Data*. Hobart Press, New Jersey
- Davidson AC, Hinkley DV (1997) *Bootstrap Methods and Their Application*. Cambridge University Press
- Efron B, Tibshirani RJ (1998) *An Introduction to the Bootstrap*. Chapman and Hall
- Grubb H, Robson A (2000) Exploratory/visual analysis. In: *Detecting Trend and Other Changes in Hydrological Data* (ed. by Z. W. Kundzewicz & A. Robson), 19–47. World Climate Programme—Water, World Climate Data and Monitoring Programme, WCDMP-45, WMO/TD no. 1013. World Meteorological Organization, Geneva, Switzerland.
- Kundzewicz ZW, Robson A (eds) (2000) *Detecting Trend and Other Changes in Hydrological Data; World Climate Programme—Water, World Climate Programme Data and Monitoring, WCDMP-45, WMO/TD no. 1013, World Meteorological Organization, Geneva, Switzerland.*
- Kundzewicz ZW, Robson A (2004) Change detection in river flow records—review of methodology. *Hydrological Science Journal* 49(1), 7–19
- Lee PM (1997) *Bayesian Statistics. An Introduction* (second edition), Arnold, 344pp.
- Pettitt AN (1979) A non-parametric approach to the change point problem. *Applied Statistics*, 28(2), 126-135
- Robson A, Bárdossy A, Jones D, Kundzewicz ZW (2000) Statistical methods for testing for change. In *Kundzewicz ZW, Robson A (eds) (2000) Detecting Trend and Other Changes in Hydrological Data; World Climate Programme—Water, World Climate Programme Data and Monitoring, WCDMP-45, WMO/TD no. 1013, World Meteorological Organization, Geneva, Switzerland*

-
- Rusticucci M, Renom M (2008) Variability and trends in indices of quality-controlled daily temperature extremes in Uruguay. *International Journal of Climatology*, 28(8), 1083-1095
- Smadi MM, Zghoul A (2006) A sudden change in rainfall characteristics in Amman, Jordan during the mid 1950s. *American Journal of Environmental Sciences*, 2(3), 84-91
- Xu ZX, Takeuchia K, Ishidairaa H (2003) Monotonic trend and step changes in Japanese precipitation. *Journal of Hydrology*, 279, 144–150
- Xiong LH, Guo SL (2004) Trend test and change-point detection for the annual discharge series of the Yangtze River at the Yichang hydrological station. *Hydrological Science Journal* 49(1), 99–112
- Yue S, Pilon P (2004) A comparison of the power of the t test, Mann-Kendall and bootstrap tests for trend detection. *Hydrological Science Journal* 49(1), 21–37
- Yue S, Pilon P, Phinney B, Cavadias G (2002) The influence of autocorrelation on the ability to detect trend in hydrological series. *Hydrological Processes* 16, 1807–1829

11.5.4 Analisi degli eventi estremi Curve IDF

- Burlando P, Rosso R (1996) Scaling and multiscaling models of depth-duration-frequency curves for storm precipitation. *Journal of Hydrology*, 187(1-2): 45–64
- Chow VT, Maidment DR, Mays LW (1988) *Applied Hydrology*. New York, USA, McGraw-Hill
- Salvadori G, De Michele C (2001) From Generalized Pareto to Extreme Values laws: scaling properties and derived features, *J. Geophys. Res.*, 106(D20), 24063–24070
- Salvadori G, De Michele C, Kottegoda NT, Rosso R (2007) *Extremes in nature: an approach using copulas*. Springer

