SULLA STIMA PARAMETRICA DELLE DISTRIBUZIONI DEI VALORI ESTREMI: LE PIOGGE INTENSE

Dr. Andrea Evangelista

Tutor: Dr. Attilio Colagrossi

Abstract

Per lunghi anni gli studi riguardanti le distribuzioni di probabilità di variabili aleatorie sono stati rivolti all'analisi dei valori centrali e alle loro tendenze generali.

Le variabili ambientali,quali le temperature,le altezze di precipitazione e le portate di piena ,si prestano invece ad essere analizzate in ragione dei fenomeni più estremi.

La teoria dei valori estremi quindi ha attratto solo in tempi recenti i ricercatori,che nel corso degli anni si sono adoperati alla ricerca di metodologie adeguate a modellare le leggi di distribuzione dei valori più elevati.

Nel presente lavoro sono presentate, in forma necessariamente sintetica, due delle metodologie più utilizzate negli studi di modellazione dei valori estremi:il metodo classico,dei massimi per blocchi e il metodo soglia (o delle eccedenze).

Queste due pratiche sono state applicate alla serie pluviometrica 1951-1998 della stazione di Nardò, che è stata scelta in funzione di una maggiore completezza dei dati a disposizione.

I risultati sono stati valutati in funzione dei livelli di ritorno associati a tempi variabili da 10 a 100 anni. I due metodi, risultati comunque adeguati a descrivere la distribuzione delle grandezze in esame, hanno mostrato una differente capacità di cogliere i valori estremi.

Per semplicità e flessibilità, per la stima dei parametri delle distribuzioni sono stati utilizzati degli stimatori di massima verosimiglianza.

Probabilmente per il metodo soglia poteva essere più adeguato un approccio bayesiano per la stima dei parametri,ma la difficoltà di disporre di pacchetti statistici che potessero implementare una simile analisi ci permette in questo momento di evidenziarne solo i limiti.

INDICE

Introduzione

Per lunghi anni gli studi riguardanti le distribuzioni di probabilità di variabili aleatorie sono stati rivolti all'analisi dei valori centrali e alle loro tendenze generali.

La teoria dei valori estremi quindi ha attratto solo in tempi recenti i ricercatori,che nel corso degli anni si sono adoperati alla ricerca di metodologie adeguate a modellare le leggi di distribuzione dei valori più elevati.

Le variabili ambientali,quali le temperature,le altezze di precipitazione e le portate di piena ,si prestano invece ad essere analizzate in ragione dei fenomeni più estremi.

In particolare in questo lavoro abbiamo voluto focalizzare la nostra attenzione sugli eventi piovosi di natura più estrema.

Avere delle informazioni circa la loro intensità, periodicità e,soprattutto , sulle loro possibilità di ritorno è una questione che riveste un grande interesse in idrologia perché hanno molteplici applicazioni.

Possiamo pensare al dimensionamento delle reti di drenaggio in ambito urbano e rurale,che richiede una stima attendibile delle piogge da smaltire, oppure alla progettazione di tutti gli interventi di difesa e di conservazione del suolo e del territorio.

Nel titolo abbiamo parlato di stima non a caso. Infatti utilizzeremo per la realizzazione dei modelli di distribuzione dei valori estremi un approccio stocastico e non deterministico. L'evento piovoso infatti verrà valutato in funzione probabilità di accadimento a partire dallo studio delle serie storiche pluviometriche disponibili.

Nello studio di certi eventi estremi, assume notevole interesse conoscere la legge di probabilità con cui si distribuiscono tali valori in campione assegnato.

Volendo indicare con P(x) la distribuzione di probabilità di un valore x (non estremo), e con $P_N(x)$ la distribuzione del massimo in un campione, per l'assioma delle probabilità composte si avrà:

$$P_N(x)=P(x)^N$$

In pratica se N sono gli anni di osservazione, la $P_N(x)$ mi fornisce la probabilità di non superamento del valore x in N anni.

E' infatti intuitivo che tale probabilità decresce con l'aumentare degli anni,come mostra bene la relazione indicata.

Se la distribuzione di probabilità di x fosse nota, il problema sarebbe già risolto,ma questo nella realtà non è possibile e quindi nel nostro lavoro verranno analizzate delle soluzioni asintotiche per questo problema.

Per la soluzione del problema verranno utilizzati due diversi metodi:il metodo classico,detto dei massimi per blocchi e quello soglia o delle eccedenze .

Una volta determinata la $P_N(x)$,potremo esplicitare i risultati in termini di tempi di ritorno ,concetto legato alla distribuzione dei massimi dalla seguente relazione:

$$\mathbf{T(x)=1/1-}\ P_N(x)$$

che ci fornisce il numero di anni in cui l'evento di intensità x viene eguagliato o superato in media una volta.

Metodologia

Lo scopo del lavoro era quello di mettere a confronto due diverse pratiche statistiche per la determinazione delle distribuzione di probabilità degli eventi estremi in pluviometria.Le due differenti tecniche avevano bisogno di due tipologie di dati differenti.

Per il metodo classico avevamo bisogno di una serie di osservazioni pluviometriche che riportasse i massimi valori di precipitazione in 24h consecutive.

Per il metodo soglia avevamo bisogno di una serie completa delle piogge cumulate giornaliere,comprensiva quindi dei giorni non piovosi.

Per ottenere un campione numeroso di osservazioni,si è cercato di individuare una stazione pluviometrica italiana caratterizzata da un grande numero di osservazioni, e con il minor numero possibile di dati mancanti.

Questo procedimento è stato facilitato da un precedente lavoro della dr.ssa Fabrizio, teso a valutare l'affidabilità delle serie storiche dei compartimenti di Bari. Sulla base di quelle risultanze,la nostra scelta è caduta sulla stazione di Nardò. I dati, provenienti dall'archivio idrologico dell'APAT, riportano le suddette serie per un periodo che va dal 1951 al 1998. Altre osservazioni,precedenti il periodo considerato non sono state considerate a causa della presenza di numerosi dati mancanti.

Per le implementazioni statistiche si è fatto riferimento alle prassi più diffuse.

La carenza di testi specifici per questo tipo di analisi ha spostato la nostra attenzione sui lavori di ricercatori pubblicati regolarmente in rete, e di cui riportiamo gli estremi nella bibliografia.

Per la manualistica ,si è visionato il testo di Maione-Moisello "Elementi di statistica per l'idrologia".

Per quanto riguarda le metodologie statistiche,rimandiamo alla visione delle pagine seguenti, in cui le varie tecniche vengono analizzate in maniera più rigorosa.

In ultimo, va sottolineata la difficoltà a reperire risorse computazionali per il trattamento dei dati.

Difatti la maggior parte dei pacchetti statistici disponibili sul mercato non implementano un tipo di analisi così particolare. Si è fatto quindi ricorso a al software open-source “R”, che,grazie all’apporto di numerosi ricercatori, è riuscito a dare un supporto alle nostre stime.

Cap 1-Il metodo classico

1.1 La distribuzione GEV (Generalized Extreme Value)

Il modello è basato sul comportamento di $M_n = \max(X_1,...,X_n)$, dove $X_1,...,X_n$ è una successione di variabili casuali indipendenti caratterizzate dal fatto di avere una comune funzione di distribuzione F.

Di norma, nelle applicazioni X_i ,rappresenta il valore di un processo misurato su una scala temporale regolare, cosicché M_n è il massimo del processo in n unità di osservazioni di tempo. Considerando n come il numero di osservazioni in un anno, di conseguenza M_n corrisponde al valore massimo annuale.

Da un punto di vista prettamente teorico,la distribuzione di M_n si può ottenere esattamente per tutti i valori di n,qualora fosse nota la funzione F:

$$\Pr\{M_n \le z\} = \Pr\{X_1 \le z,...,X_n \le z\} = \Pr\{X_1 \le z\}...\Pr\{X_n \le z\} = \{F(z)\}^n$$

Nel nostro caso tale formula non ci può essere d'aiuto. Dato che nella maggior parte dei casi fissiamo lo studio al massimo di un grande numero di variabili, nella teoria statistica si è seguito un approccio asintotico per modellizzare M_n ; questa metodologia è simile a quella di approssimare la distribuzione delle medie campionarie con la distribuzione normale.

Per le grandezze ideologiche,quali ad esempio le stesse precipitazioni giornaliere,vengono rese disponibili le serie storiche dei valori massimi annuali.

Per l'interpretazione di questo tipo di osservazioni si prestano in maniera particolare diverse leggi probabilistiche.

Il tipo di forma asintotica dipende dalla distribuzione originaria e da questo punto di vista si possono suddividere in tre grandi famiglie ,a ciascuna delle quali corrisponde un diverso tipo di forma asintotica.
Nel dettaglio:

$$I: G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\} \qquad -\infty < z < \infty$$

$$II: G(z) = \begin{cases} 0 & z \le b \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\} & z > b \end{cases}$$

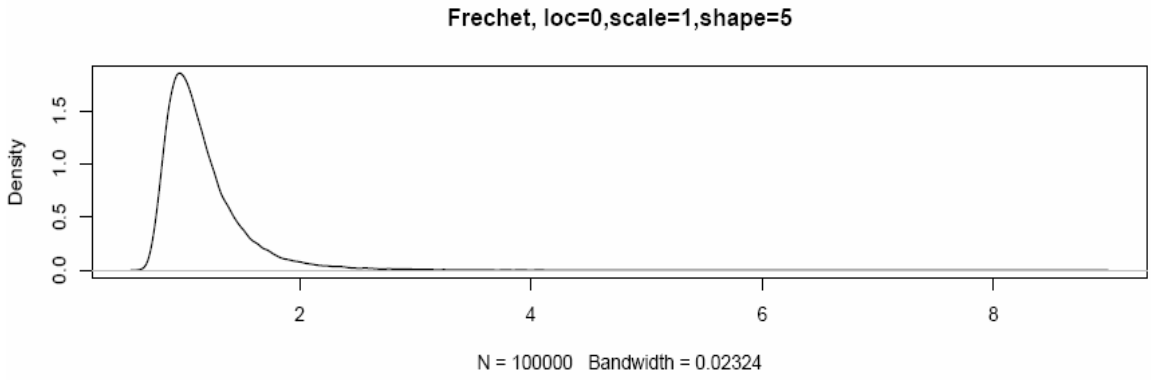
$$III: G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\} & z < b \\ 1 & z \ge b \end{cases}$$

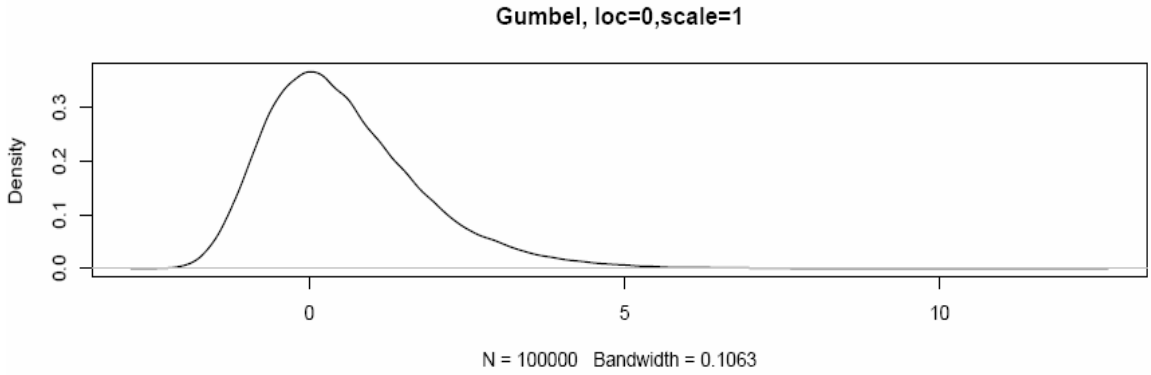
per parametri a>0, b e, nel caso delle famiglie II e III, α >0 .

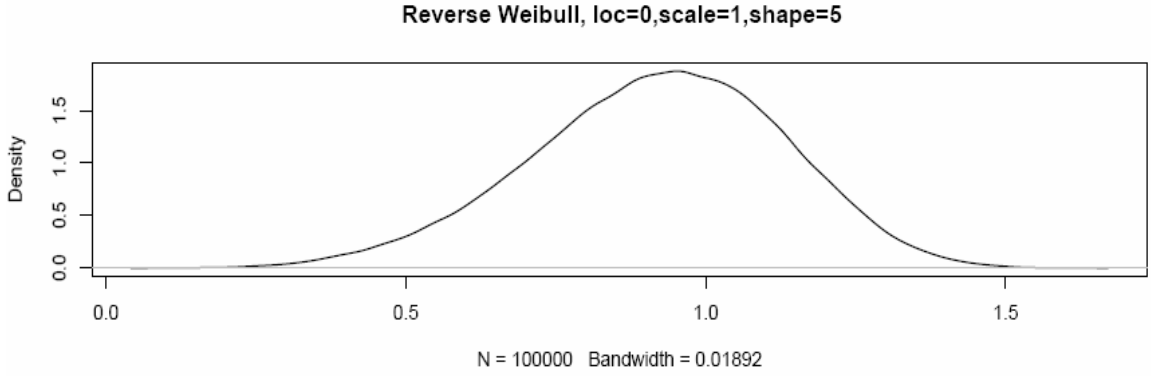
In particolare, la distribuzione dei massimi normalizzati $(M_n - b_n)/a_n$ converge ad una variabile che ha una distribuzione tra una delle tre appena scritte.

Collettivamente queste tre classi di distribuzioni sono chiamate le distribuzioni dei valori estremi, con i tipi I, II e III universalmente noti come le famiglie di Gumbel, Frechet e Weibull rispettivamente.

Queste tre distribuzioni presentano un comportamento diverso a seconda del comportamento delle code della distribuzione F.







Come il grafico evidenzia,la distribuzione di Weibull è limitata superiormente, a differenza di quelle di Gumbel e Frechet.

Successivamente, le famiglie di Gumbel, Frechet e Weibull sono state ricombinate in una singola famiglia di modelli aventi funzione di distribuzione della forma

$$G(z) = \exp\left\{-\left[1+\varepsilon\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\varepsilon}\right\}$$

definita sull'insieme $\{z: 1+\varepsilon(z-\mu)/\sigma>0\}$, dove i parametri soddisfano $-\infty<\mu<\infty$, $\sigma>0$ e $-\infty<\varepsilon<\infty$.

Questa è la famiglia di distribuzioni Generalizzata di Valore Estremo (GEV). Il modello ha tre parametri: un parametro di locazione, μ; un parametro di scala, σ, e un parametro di forma, ε.

Le classi di valori estremi precedenti si possono riottenere in questo modo:

- La Fréchet corrisponde al caso $\varepsilon>0$; GEV$(1,\alpha^{-1},\alpha^{-1})$ =Fréchet (α)
- La Weibull corrisponde al caso $\varepsilon<0$; GEV$(-1,\alpha^{-1},-\alpha^{-1})$ =Weibull(α)
- La Gumbel corrisponde al caso $\varepsilon=0$

Avere ricondotto le tre famiglie di valori estremi in una singola famiglia semplifica enormemente le implementazioni statistiche.

Attraverso l'inferenza su ε sono i dati stessi a determinare il tipo di coda più appropriato, e di conseguenza non c'è necessità di dare giudizi soggettivi a priori riguardo quale famiglia individuale di valori estremi adottare.

L' applicazione della GEV consiste nel bloccare i dati in successioni di uguale lunghezza e di adattare la GEV all'insieme dei massimi delle suddette successioni.

Nell'adozione di questo modello per un qualunque insieme, la scelta dell'ampiezza n dei blocchi può essere critica.

Si tratta in pratica di un trade-off tra distorsione e varianza: blocchi troppo piccoli rendono l'approssimazione del teorema GEV poco significativa, portando a distorsione nella stima; d'altra parte blocchi grandi generano pochi massimi e di conseguenza alta varianza.

Nel nostro lavoro cercheremo di valutare il modello GEV sui nostri dati. In particolare utilizzeremo i valori dei massimi annuali di precipitazione in 24h.Quindi per tornare alla terminologia fin qui adottata,per M_n si adotterà n=365.

Per la stima dei parametri della distribuzione GEV sono adottate diverse metodologie.

In questo lavoro utilizzeremo la tecnica della massima verosimiglianza, facendo particolarmente attenzione alle condizioni di regolarità richieste al fine di verificare le proprietà asintotiche dello stimatore di massima verosimiglianza.

Tali condizioni sono state valutate da Smith R.L.,che riportò i seguenti risultati:

- Quando $\varepsilon > -0.5$, gli stimatori di massima verosimiglianza sono completamente regolari.
- Quando $-1 < \varepsilon < -0.5$, gli stimatori di max verosimiglianza esistono ma sono non regolari.
- Quando $\varepsilon < -1$, gli stimatori di max verosimiglianza non esistono.

La massimizzazione della seguente funzione,detta di log-verosimiglianza,porta alla stima di massima verosimiglianza:

$$l(\mu,\sigma,\varepsilon) = \sum_{i=1}^{k} \left\{ -\log\sigma - (1+1/\varepsilon)\log\left[1+\varepsilon\left(\frac{x_i-\mu}{\sigma}\right)\right]^{-1/\varepsilon} \right\}$$

dato:

$$1+\varepsilon\left(\frac{x_i-\mu}{\sigma}\right) > 0, \text{ per i=1,..,k.}$$

L'utilizzo di una tipologia di famiglia parametrica come la GEV per descrivere le distribuzioni dei massimi annuali di precipitazione ha molti vantaggi.

Primo tra tutti la sua facilità di implementazione, dato che sono frequentemente rese disponibili le serie storiche delle precipitazioni di massima intensità per diverse durate,e, nonostante i pochi parametri sta stimare risulta essere statisticamente robusta.

Numerosi lavori hanno affrontato anche i limiti di una simile trattazione dei dati estremi.

Utilizzare infatti i soli massimi annuali comporta una notevole perdita di informazione dovuta al fatto che un picco elevato possa nascondere altri eventi estremi verificatisi nello stesso anno.

Altro limite è quello di assumere costanti i parametri nell'arco di periodo considerato, che se da un lato può essere plausibile nel breve periodo,certamente risulta essere falso per tempi di ritorno più lunghi.

I metodi soglia,che vedremo successivamente permettono di superare questi limiti.

1.2-Analisi delle precipitazioni estreme secondo il metodo classico

Le figure 3.1 e 3.2 mostrano, le precipitazioni massime annuali registrate dalla stazione di Nardò nel periodo 1951-1998

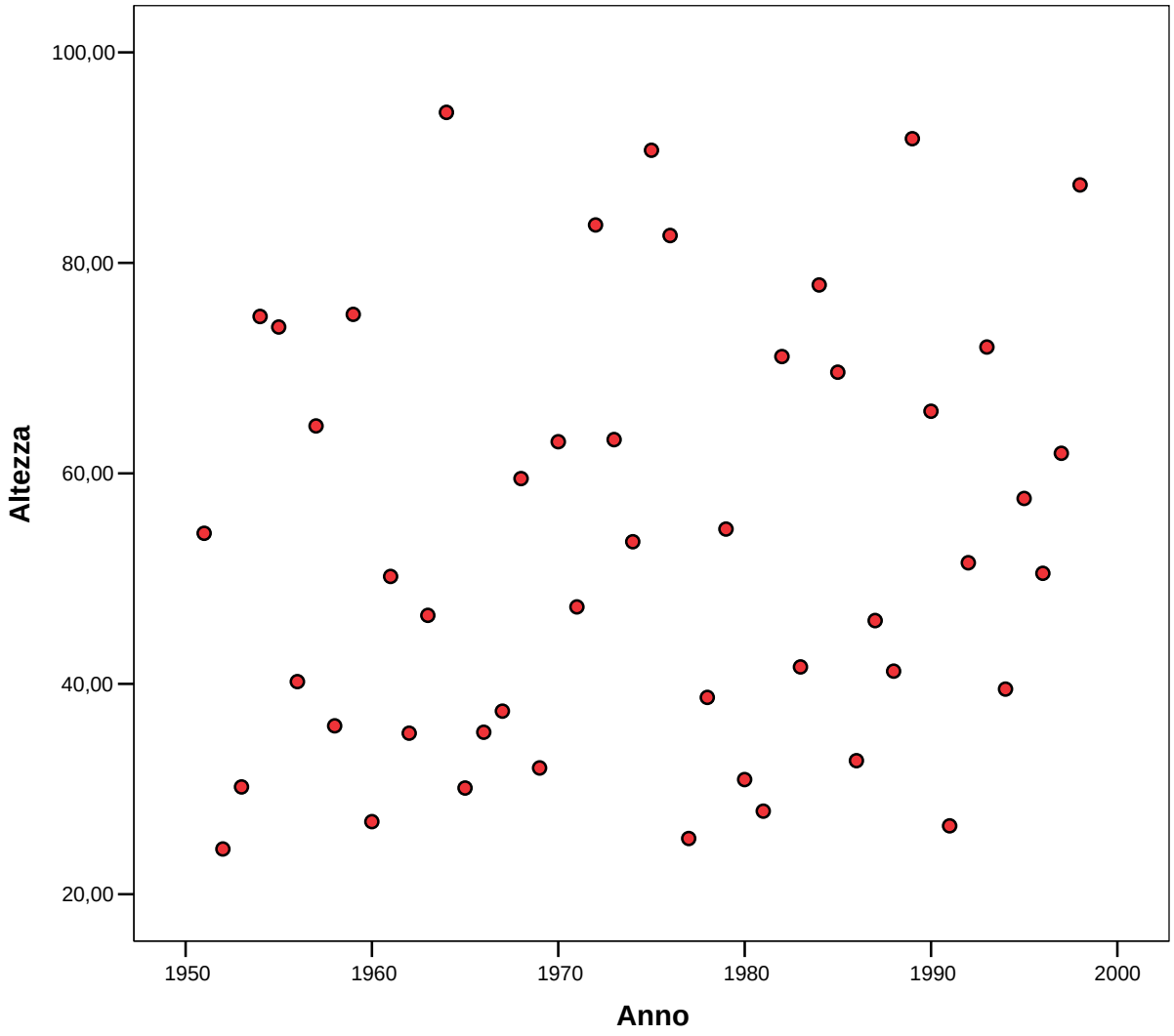


Fig. Grafico a dispersione delle precipitazioni massime annuali a Nardò

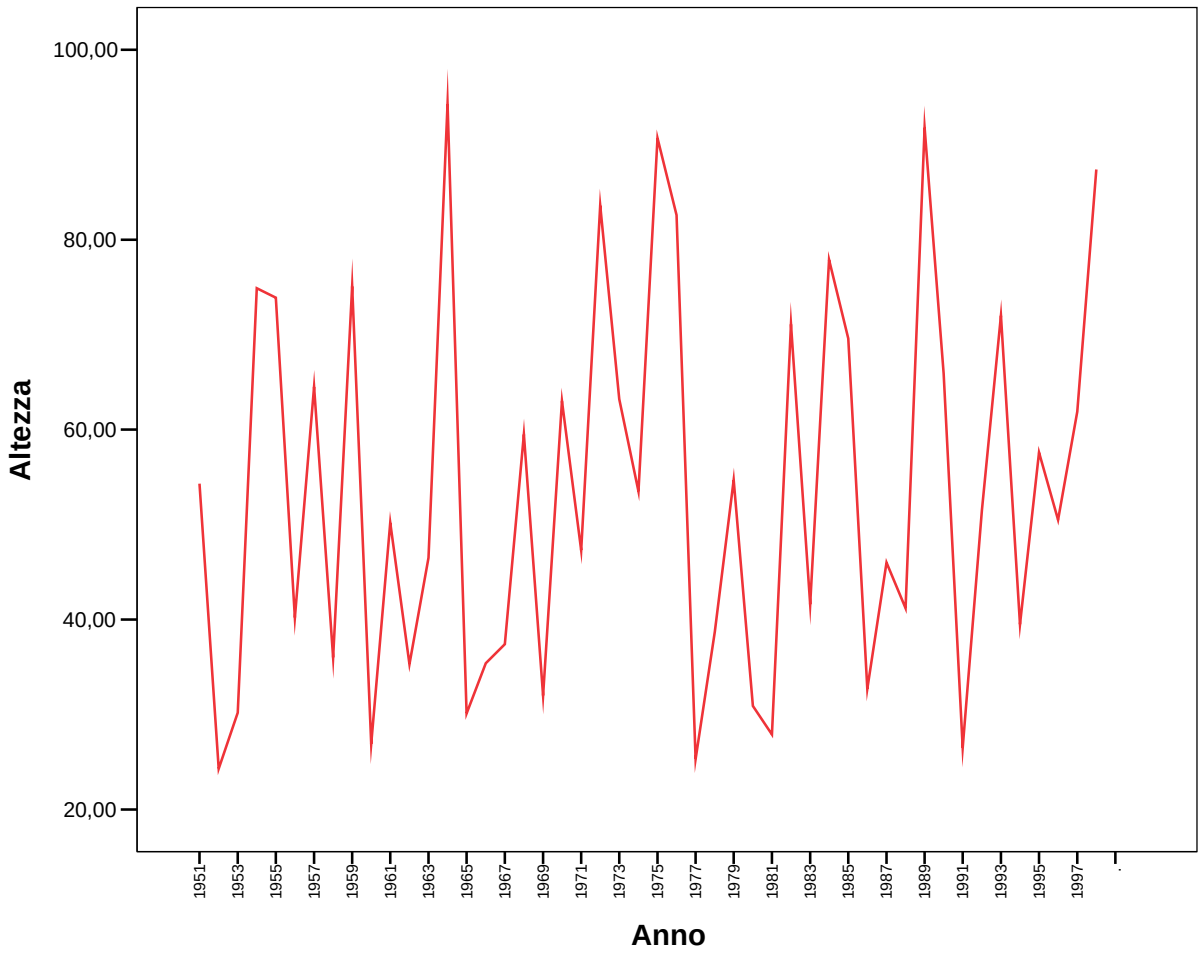


Fig. Serie temporale delle precipitazioni massime annuali a Nardò

I grafici si mostrano abbastanza irregolari, e risulta evidente nessun tipo di trend.

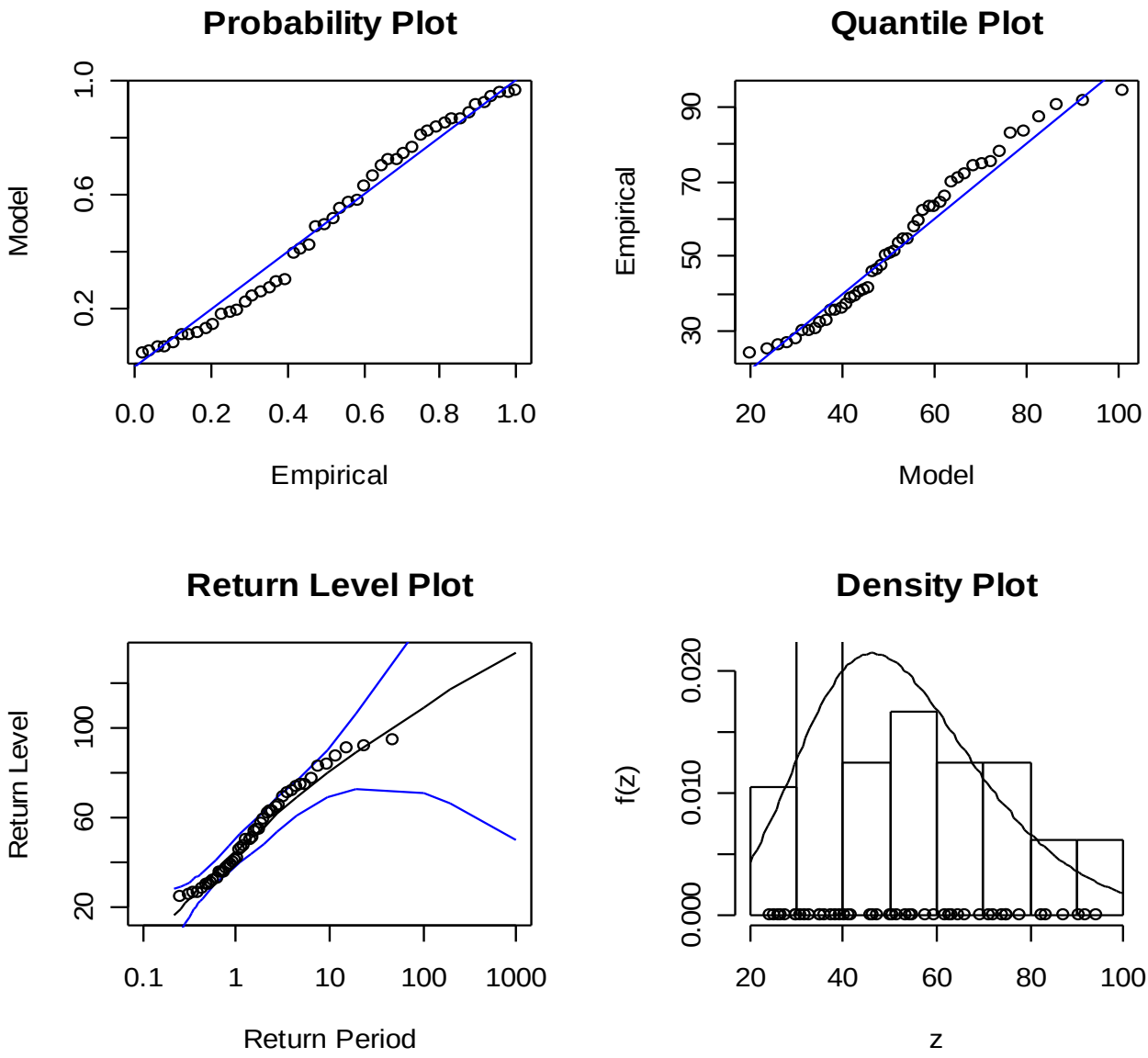
Procediamo ora ad applicare il modello GEV sui nostri dati,assumendo che le osservazioni siano indipendenti per tutto il periodo studiato.

Procedendo con l'utilizzo degli stimatori di massima verosimiglianza otteniamo le seguenti stime per i parametri della GEV:

$$(\hat{\mu}, \hat{\sigma}, \hat{\varepsilon}) = (44.672, 17.280, -0.090)$$

Il valore di ε risulta maggiore di -1 per cui le condizioni poste sono state rispettate.

Di seguito vengono riportati i grafici diagnostici per verificare il buon adattamento del modello GEV ai nostri dati;



In pratica i punti empirici dei dati vengono confrontati con quelli del modello.

Se la GEV è un modello ragionevole per la distribuzione della popolazione, i punti del probability plot e del quantile plot dovrebbero essere situati in prossimità della diagonale.

Notevoli scostamenti del grafico dalla linearità segnalano una non buona precisione del modello per i dati.Il Return Level Plot invece riproduce le stime dei livelli di ritorno per T=0 fino a 1000 anni.Tutti i punti empirici si trovano allineati sulla curva e soprattutto rientrano all'interno delle bande di confidenza del modello stimato.

Meno informativo è l’istogramma di frequenza, su cui è sovrapposta la funzione di densità di probabilità del modello GEV,in quando le ampiezze degli intervalli possono essere variabili.

I grafici nel complesso fanno propendere per l’ipotesi che la GEV rappresenti un modello adeguato per descrivere i nostri dati.

Cap 2-Il metodo soglia

2.1 La Distribuzione Generalizzata di Pareto (GPD)

Come si è visto in precedenza,lo studio dei massimi valori delle grandezze idrologiche viene basato usualmente sulla distribuzione dei massimi annuali.
Una simile trattazione statistica, che concettualmente risulta essere corretta, è di frequente l'unica possibile,in quanto non sono resi noti i dati originali ma si dispone solamente dei valori riferiti alle serie storiche dei massimi annuali.
Nei casi in cui fossero disponibili anche le registrazioni di tutta la serie dei dati, sarebbe opportuno utilizzare altre implementazioni statistiche che permettono di utilizzare maggiormente le informazioni contenute nei dati.
Nel nostro caso infatti siamo riusciti a risalire alla serie storica delle precipitazioni giornaliere della stazione di Nardò,per il periodo che va dal 1951 al 1998. I dati riportano le altezze di pioggia cadute nelle 24 ore che vanno dalle ore 09:00 alla stessa ora del giorno successivo.
Per questo motivo,in questa serie, i massimi annuali,differentemente da quelli utilizzati nel modello classico,possono essere leggermente inferiori in virtù del diverso intervallo orario di registrazione.
Una metodologia che può rispondere al nostro caso è quello detto del "metodo soglia" o delle "eccedenze".
In assenza però di conoscenze riguardo al modello statistico appropriato da usare per i nostri dati, facciamo ancora ricorso ad un modello asintotico come approssimazione. Ma piuttosto che raggruppare i dati in blocchi di un anno ed estrarre il massimo da ogni blocco, definiamo un evento come estremo se cade oltre un determinato livello.
L'intuizione su cui si basa una tale pratica,è quello di individuare una soglia u adeguatamente alta ed analizzare tutti i superamenti di questa soglia in periodo di tempo determinato.
Siano $X_1, X_2, \ldots$ una successione di variabili casuali.

Per un numero di osservazioni sufficentemente grande e per una soglia u elevata la funzione di distribuzione di (X-u) , condizionata a X>u, è approssimativamente

$$H(y) = 1 - \left(1 + \frac{\varepsilon\, y}{\tilde{\sigma}}\right)^{-1/\varepsilon}$$

definita su $\left\{ y : y > 0 \;\; e \;\; \left(1 + \varepsilon\, y / \tilde{\sigma}\right) > 0 \right\}$, dove

$$\tilde{\sigma} = \sigma + \varepsilon\, (u - \mu)$$

La famiglia di distribuzioni definita è chiamata famiglia di Pareto generalizzata (GPD=Generalized Pareto Distribution).

In sostanza, quando u si tende al limite superiore, $F_u(y) \approx GPD(y; \sigma_u, \varepsilon)$. Praticamente, per una soglia adeguatamente alta u, c'è qualche σ_u (che dipende da u) e qualche ε (indipendente da u) per cui la GPD è una buona approssimazione della funzione di distribuzione delle eccedenze F_u.

Se $\varepsilon<0$ la distribuzione degli eccessi ha un estremo superiore in $\mu - \tilde{\sigma}/\varepsilon$; se $\varepsilon>0$ la distribuzione non ha un estremo superiore.

Dopo aver scelto una soglia adeguata, i parametri della GPD possono essere stimati attraverso il metodo della massima verosimiglianza.

Supponiamo che i valori $y_1, ..., y_k$ siano i k valori oltre la soglia u. Quando $\varepsilon \neq 0$ la log-verosimiglianza si ottiene come

$$l(\sigma, \varepsilon) = -k \log \sigma - (1 + 1/\varepsilon) \sum_{i=1}^{k} \log\left(1 + \varepsilon\, y_i / \sigma\right)$$

dato $\left(1+\sigma^{-1}\varepsilon y_i\right)>0$ per i=1,…,k; altrimenti, $l(\sigma,\varepsilon)=-\infty$.

Nel caso $\varepsilon=0$ la log-verosimiglianza è ottenuta come

$$l(\sigma)=-k\log\sigma-\sigma^{-1}\sum_{i=1}^{k}y_i$$

2.2 La scelta della soglia

La scelta della soglia è un punto molto importante nei modelli basati sulle eccedenze.
Su questo punto molti ricercatori hanno aperto un dibattito volto ad individuare le soluzioni più efficienti per l'applicazione della GPD.
Comunque sembra superato il concetto che individuava la soglia in valori stabiliti a priori ed indipendenti dalle grandezze dei dati elaborati. Infatti ogni realizzazione di osservazioni ha in se un concetto relativo di "evento estremo".
Nello specifico, scegliere una soglia troppo bassa può portare delle distorsioni nel modello, o comunque non renderlo "capace" di intercettare quei valori più elevati ai quali siamo interessati,facendo cadere tra l'altro le proprietà asintotiche della GPD.
Viceversa,una soglia troppo alta porta ad individuare poche eccedenze, determinando una alta varianza e un modello difficilmente utilizzabile a causa di intervalli di confidenza troppo ampi in fase di stima dei parametri.
Nella pratica statistica più in voga, le metodologie di individuazione della soglia si basano principalmente su due metodi grafici.
Il primo viene effettuato prima della stima dei parametri e un secondo simulando ripetutamente la GPD sui dati utilizzando soglie diverse per valutarne la stabilità dei parametri nelle soglie successive.
Più dettagliatamente, se Y ha segue una GPD con parametri σ e ε, allora

$$E(Y) = \frac{\sigma}{1-\varepsilon}$$

Ipotizziamo la GPD valida come modello per le eccedenze di una generica soglia u_o, generata da una serie $X_1,...,X_n$, il valore atteso di questa funzione sarà:

$$E(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1-\varepsilon}$$

dove σ_{u_0} è utilizzato per indicare il parametro di scala relativamente alla soglia u_o.

Ma se la distribuzione di Pareto è valida per la soglia u_o, dovrebbe essere adeguata per tutte le soglie u> u_o, avendo preventivamente apportato una appropriata trasformazione del parametro di scala in σ_u. Quindi, per u> u_o,

$$E(X - u | X > u) = \frac{\sigma_u}{1-\varepsilon} = \frac{\sigma_{u_0} + \varepsilon\mu}{1-\varepsilon}$$

In questo modo, per u> u_o, $E(X - u | X > u)$ è una funzione lineare di u.

Il luogo dei punti

$$\left\{ \left(u, \frac{1}{n_u} \sum_{i=1}^{n_u} \left(x_{(i)} - u \right) \right) : u < x_{max} \right\},$$

dove $x_{(1)},...,x_{(n_u)}$ sono le n_u osservazioni che superano u e x_{max} è la più grande delle X_i, è chiamato il diagramma dell'eccesso medio residuo (mean residual life plot).

Aldilà di una soglia u_o, per i cui gli eccessi di una GPD fornisce una valida approssimazione, il diagramma della vita media residua dovrebbe essere approssimativamente lineare in u.

Il mean residual life plot tuttavia si presta come un metodo indicativo, e viene utilizzato per individuare inizialmente un range di soglie papabili per il nostro modello.

Infatti,nella prassi, il diagramma della vita media residua viene integrato con un'altra procedura grafica.

I presupposti sono gli stessi del precedente,solo che in questa fase interviene una simulazione sui dati empirici di diversi modelli GPD.

Se la GPD è un modello ragionevole per superamenti di una soglia u_o, allora le eccedenze di una soglia $u > u_o$ seguono anch'essi una GPD. I parametri di forma delle due distribuzioni sono identici, mentre chiamando con σ_u il valore del parametro di scala per la soglia u diventa

$$\sigma_u = \sigma_{u_0} + \varepsilon\,(u - u_0)$$

Al di sopra u_o, le stime di σ^* e ε dovrebbero essere costanti qualora u_o sia una soglia adeguata. I parametri stimati $\hat{\sigma}^*$ e $\hat{\varepsilon}$ contro u, vengono raffigurati su quello che viene indicato come il grafico della stabilità del parametro (parameter stability plot), e scegliamo la soglia corrispondente al valore più basso di u per cui le stime rimangono quasi costanti.

2.3-Analisi delle precipitazioni estreme secondo il metodo soglia

L'utilizzo del metodo soglia viene utilizzato a questo punto per la modellizzazione secondo una GPD dei valori empirici della stazione di Nardò.Per rendere possibile un confronto dei risultati ottenuti con la distribuzione GEV, viene considerato lo stesso periodo di riferimento,che va 1951 al 1998.I dati riportano le altezze di pioggia cadute nelle 24 ore

che vanno dalle ore 09:00 alla stessa ora del giorno successivo. Per questo motivo,in questa serie, i massimi annuali,differentemente da quelli utilizzati nel modello classico,possono essere leggermente inferiori in virtù del diverso intervallo orario di registrazione.
La lettura dei dati può essere facilitato dall'utilizzo del seguente grafico

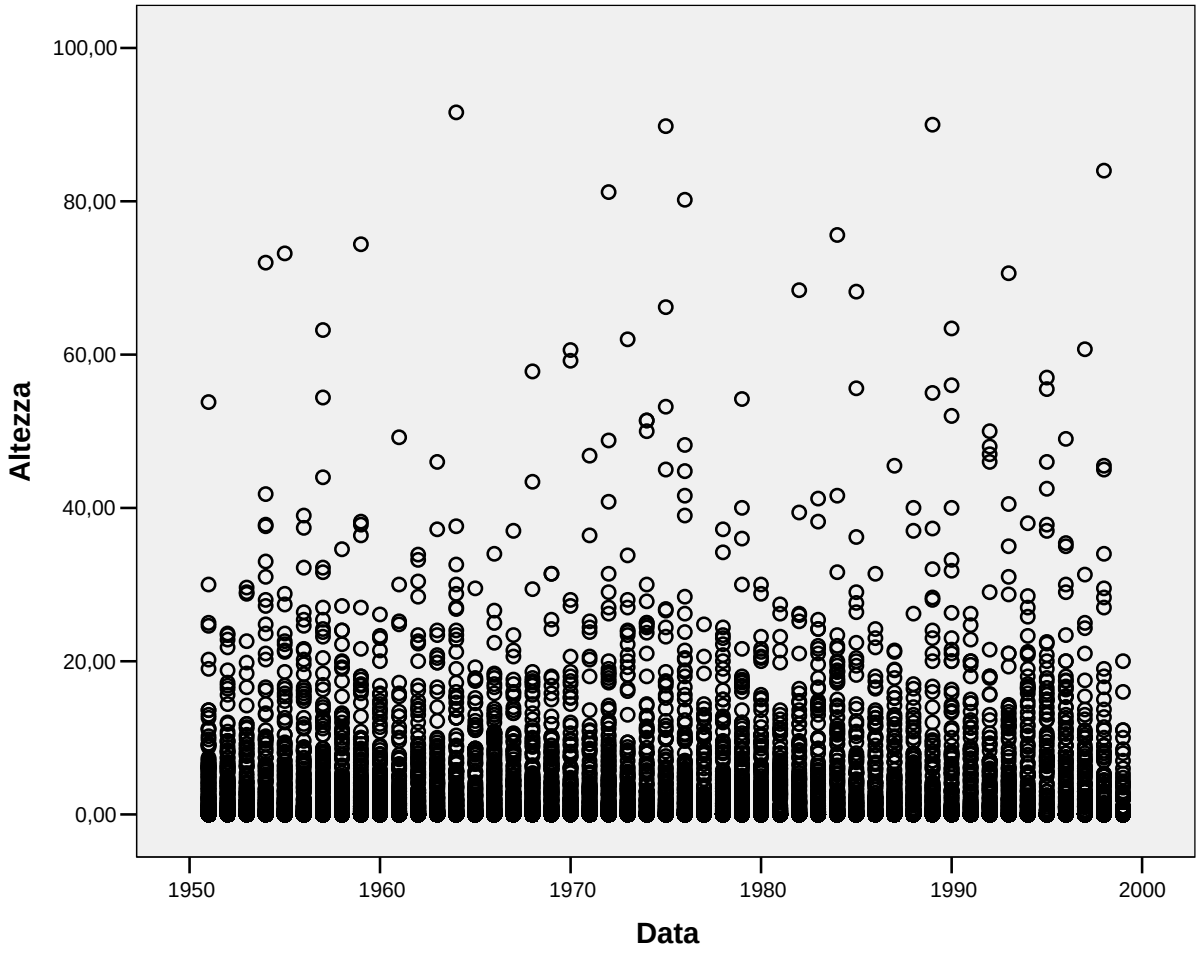


Fig. Grafico a dispersione delle precipitazioni giornaliere

Come per il caso della distribuzione dei massimi, non risulta apprezzabile un trend per le precipitazioni più intense. Il grafico a dispersione comunque ci comincia già a fornire un'idea su quale possa essere la soglia prescelta per la GPD, poiché i la nube dei punti si fa via via più rada a partire dall'altezza 40.
Per avere ulteriori conferme alla nostra intuizione,procediamo anche ad integrare il precedente diagramma con i procedimenti grafici descritti nel precedente paragrafo.
Quello che segue è il mean residual life plot

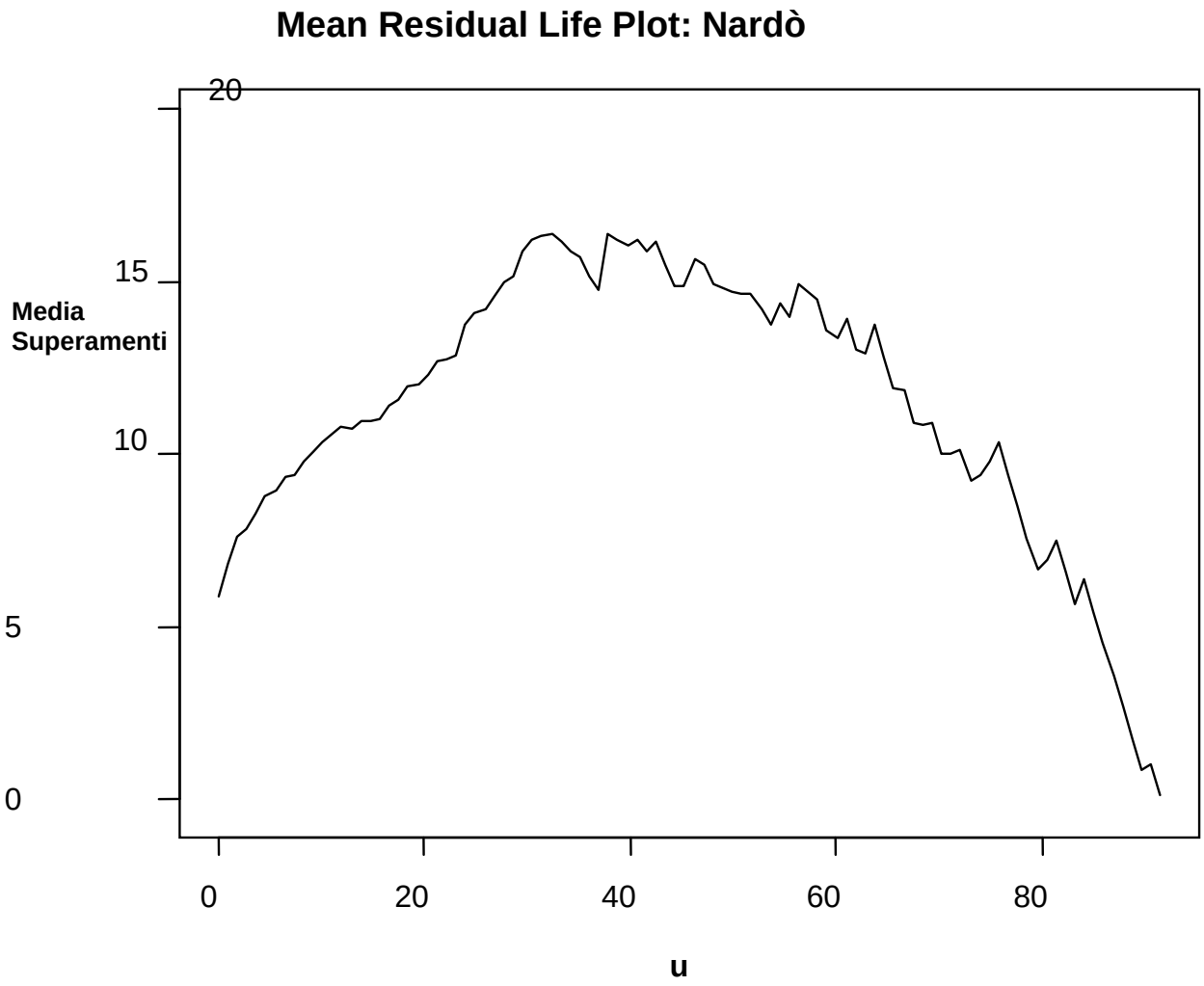


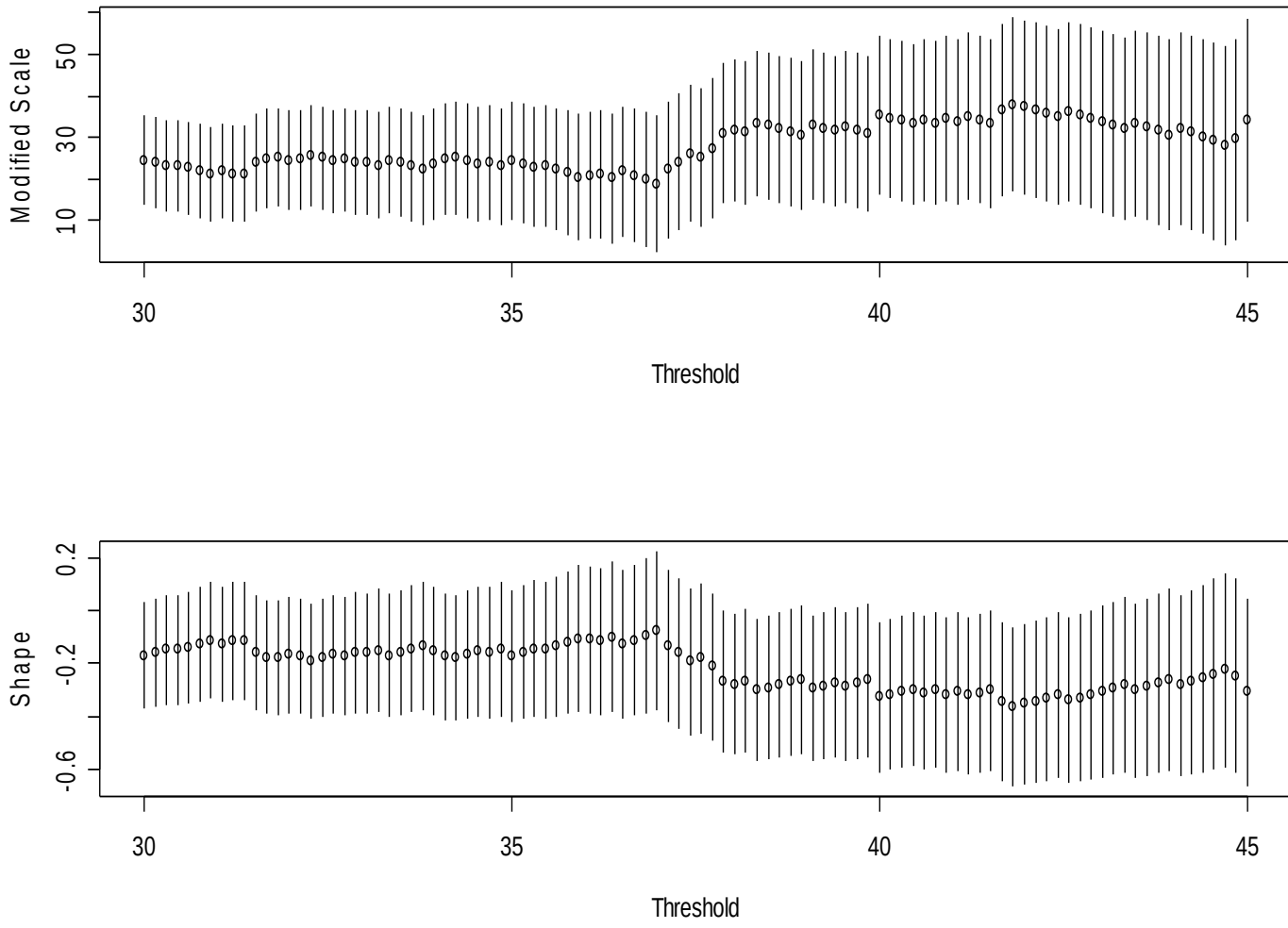
Fig. Mean residual life plot per le precipitazioni giornaliere

Il grafico mostra un rapido incremento fino approssimativante i valori 35-40.

Dopo tali soglie il livello prima tende a stabilizzarsi e poi a diventare irregolare a causa delle poche eccedenze determinate dalle soglie più elevate. Nessuna indicazione netta può essere tratta da questo grafico.

Per determinare una soglia più adeguata può essere d'aiuto ricorrere al secondo metodo grafico illustrato, cioè il diagramma di stabilità dei parametri.

Come detto, sono qui tracciati i parametri $\hat{\sigma}^*$ e $\hat{\varepsilon}$ contro u, e il range di soglie considerato è 30-45.



. Fig. Parameter stability plot, i parametri di scala e di forma contro u

A questo punto la situazione è più chiara e appare netto il cambio di trend a cavallo del valore 38. Scegliamo quindi tale valore come soglia per il nostro modello.

La soglia di 38 mm risulta superata 67 volte, e quindi rispetto alla GEV verranno utilizzati un maggior numero di valori.

Procedendo alle stime di massima verosimiglianza per la GPD si ottengono i seguenti risultati:

$$(\hat{\sigma},\hat{\varepsilon}) = (21.24, 0.28)$$

Il valore di ε risulta maggiore di -1 per cui le condizioni poste sono state rispettate.

Di seguito vengono riportati i grafici diagnostici per verificare il buon adattamento del modello GEV ai nostri dati;

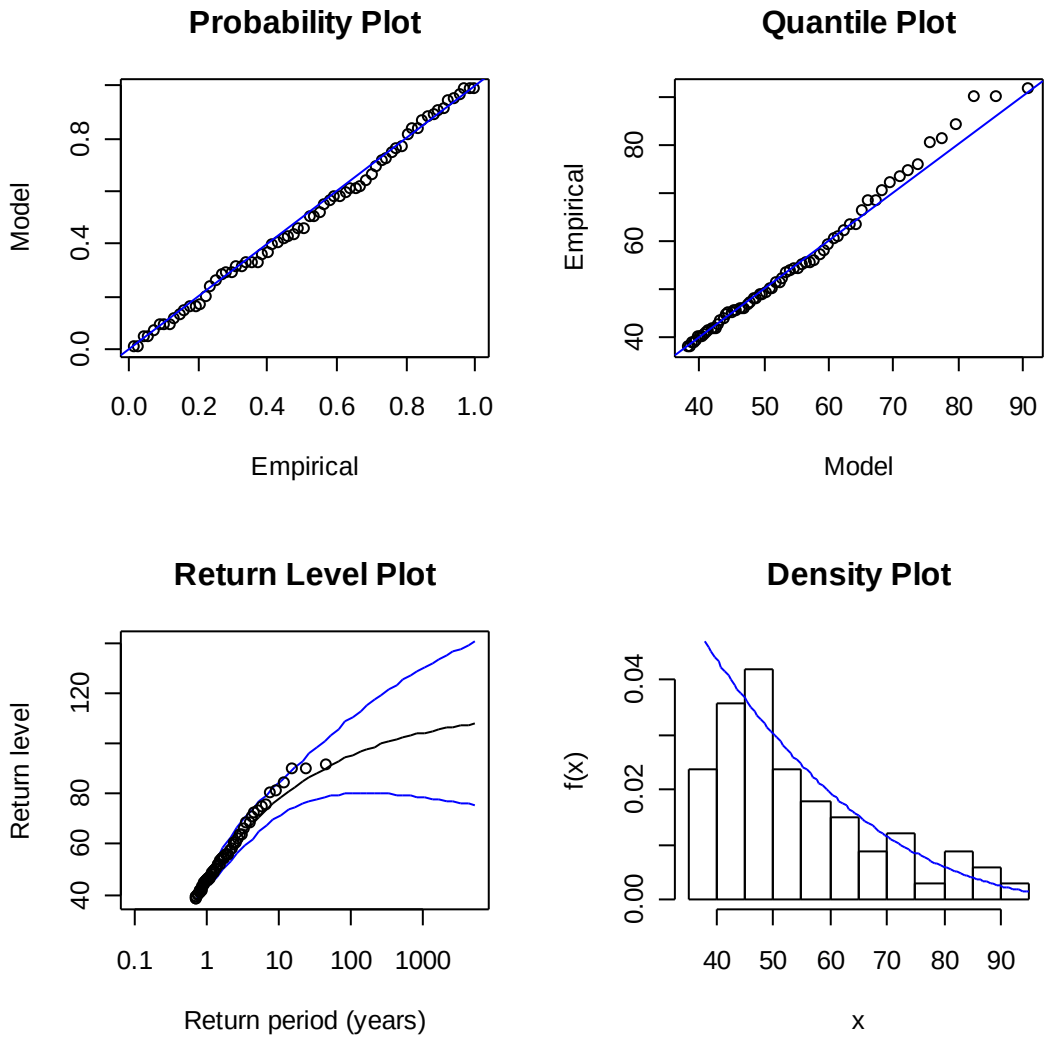


Fig. Grafici diagnostici della GPD

I punti empirici dei dati vengono confrontati con quelli del modello.Se la GPD è un modello ragionevole per la distribuzione delle eccedenze della soglia 38, i punti del probability plot e del quantile plot dovrebbero essere situati in prossimità della questa diagonale.Osservando i grafici situazione sembra verificata, anche se rispetto alla GEV il

modello”fatica” maggiormente a stimare le probabilità dei punti più estremi,che rispetto agli altri sono più distanti rispetto alla diagonale. Probabilmente scegliendo una soglia più elevata avremmo dato un maggior peso a queste osservazioni a scapito però di un maggiore precisione nella stima degli intervalli di confidenza dei parametri.

Sulla bontà del modello prescelto ,Return Level Plot,che riproduce le stime dei livelli di ritorno per T=0 fino a 1000 anni, mostra come i punti empirici si trovano allineati sulla curva e soprattutto rientrano all’interno delle bande di confidenza del modello stimato.

I grafici nel complesso fanno propendere per l’ipotesi che la GPD rappresenti un modello adeguato per descrivere i nostri dati.

Cap 3-Un confronto tra il metodo classico e il metodo soglia

3.1 I livelli di ritorno

I metodi utilizzati per parametrizzare le leggi di distribuzione dei valori estremi contengono in se delle informazioni che possono essere esplicitate in altre forme per rendere più immediata la lettura dei risultati.

Come si era già parlato nell'introduzione del lavoro, il campione di dati utilizzato contiene in se , oltre che una informazione riguardo i livelli di pioggia caduta, anche una informazione riguardo ai tempi in cui questi fenomeni si sono verificati.

Non dimentichiamo infatti che la numerosità del campione è determinata dal numero di anni di osservazione.

Conoscere quindi le leggi distributive dei valori estremi risulta utile per rispondere ad alcune domande che negli studi ambientali rivestono un ruolo molto importante.

Quale è la probabilità che il massimo valore di una precipitazione superi un livello dato in un periodo di tempo futuro? Per quali livelli di altezza di pioggia la probabilità di superamento è sufficientemente piccola?

Se la prima domanda contiene già una risposta nella funzione di ripartizione dei valori estremi, determinati con i metodi illustrati, la seconda trova una soluzione focalizzando l'attenzione sul problema inverso.

Precisiamo con M la variabile aleatoria del massimo in un determinato anno e con G la sua funzione di ripartizione :

$$\Pr(M > x) = 1 - G(x)$$

Se indichiamo un livello di probabilità p, si avrà che:

$$\Pr\left(M > z_p\right) = p$$

che rappresenta la probabilità di superamento del quantile p.
Il concetto di livello di ritorno,che è strettamente correlato a quello del quantile, risulta molto più incisivo.
Nella terminologia usuale il quantile z_p di G è il livello di ritorno associato al periodo di ritorno 1/p.
Infatti per un intervallo di tempo futuro di T anni, , il numero di volte in cui un determinato livello u viene superato è una variabile aleatoria N(u), che assume i valori interi tra 0 e T

$$N(u)=\sum_{i=1}^{T} I_{[M_i>u]}.$$

e quindi si distribuisce come il numero di successi in T prove Bernoulliane.,con una probabilità di successo nella singola prova Pr(M>u).
Il valore atteso di N(u) è dato allora da

$$E\left[N(u)\right] = T \cdot \Pr(M > u) = T(1 - G(u))$$

Il valore u_t tale che

$$E[N(u)]=1$$

è dunque il valore tale che

$$1\text{-}G(u)=1/T$$

In pratica il quantile 1/T di G.
La relazione appena descritta associa al (1/T)-quantile di G il significato di livello che mediamente viene superato una volta in T anni, ovvero, come si usa dire, di livello che viene superato in media una volta ogni T anni.

Esportando questo ragionamento alle distribuzione GEV e GPD, illustrate precedentemente, i livelli di ritorno sono determinati nel seguente modo.
Per la GEV, le stime dei quantili z_p per la distribuzione del massimo annuale ,si ottengono nel seguente modo:

$$z_p = \begin{cases} \mu - \dfrac{\sigma}{\varepsilon}\left\{ 1 - [-\log(1-p)]^{-\varepsilon} \right\} & \text{per } \varepsilon \neq 0 \\ \mu - \sigma \log[-\log(1-p)] & \text{per } \varepsilon = 0 \end{cases}$$

dove i parametri hanno lo stesso significato illustrato nella impostazione della GEV illustrata precedentemente
Per la GPD, invece, stabilito

$$\varsigma_u = \Pr\{X > u\}$$

il livello di ritorno di N anni è definito da

$$z_N = u + \frac{\sigma}{\varepsilon}\left[(N n_y \varsigma_u)^{\varepsilon} - 1\right]$$

Qualora ε=0, la precedente espressione diventa

$$z_N = u + \sigma \log\left(N n_y \varsigma_u\right)$$

Anche in questo caso i parametri utilizzati hanno il significato illustrato nella trattazione della GPD.

3.2 I livelli di ritorno con le due metodologie

In questo paragrafo verranno illustrate le stime dei livelli di ritorno, per permettere un confronto, seppur approssimativo,tra le due distinte metodologie fin qui implementate nella trattazione dei nostri dati pluviometrici.

La tabella che segue illustra i livelli di ritorno per periodi di T che vanno da 10 a 100 anni.

Bisogna però premettere che questi risultati vanno trattati con una certa cautela, in quanto se le stime possono avere una certa credibilità per tempi di ritorno relativamente piccoli, certamente sono distorte per periodi più lunghi,in quanto è inverosimile che le condizioni iniziali possano mantenersi identiche per periodi di tempo così lunghi.

P	**Metodo Classico**	**Metodo Soglia**
1/10	79,86	77,67
1/20	89,69	84,14
1/30	92,67	87,38
1/40	98,73	89,47
1/50	101,5	90,98
1/60	103,72	92,14
1/70	105,56	93,08
1/80	107,14	93,86
1/90	108,51	94,53
1/100	109,73	95,11

Tab.Livelli di ritorno secondo le due metologie

Le stime dei livelli di ritorno presentano un comportamento simile, associando a tempi di ritorno più lunghi livelli di di precipitazione più elevati.

Tuttavia,se per p più piccoli le stime sono più vicine,per tempi più lunghi le due serie tendono a distaccarsi più marcatamente.

Nel dettaglio, il metodo dei massimi per blocchi, quello classico , tende a fornire stime più elevate.

Per rendere meglio l'idea su questo problema può essere utile visionare il grafico seguente , nel quale sono riportate in asse delle ascisse i tempi di ritorno variabili da 10 a 100 anni, con intervalli di 10 anni, e rispettivi livelli di ritorno che sono stati stimati con la GEV e la GPD.

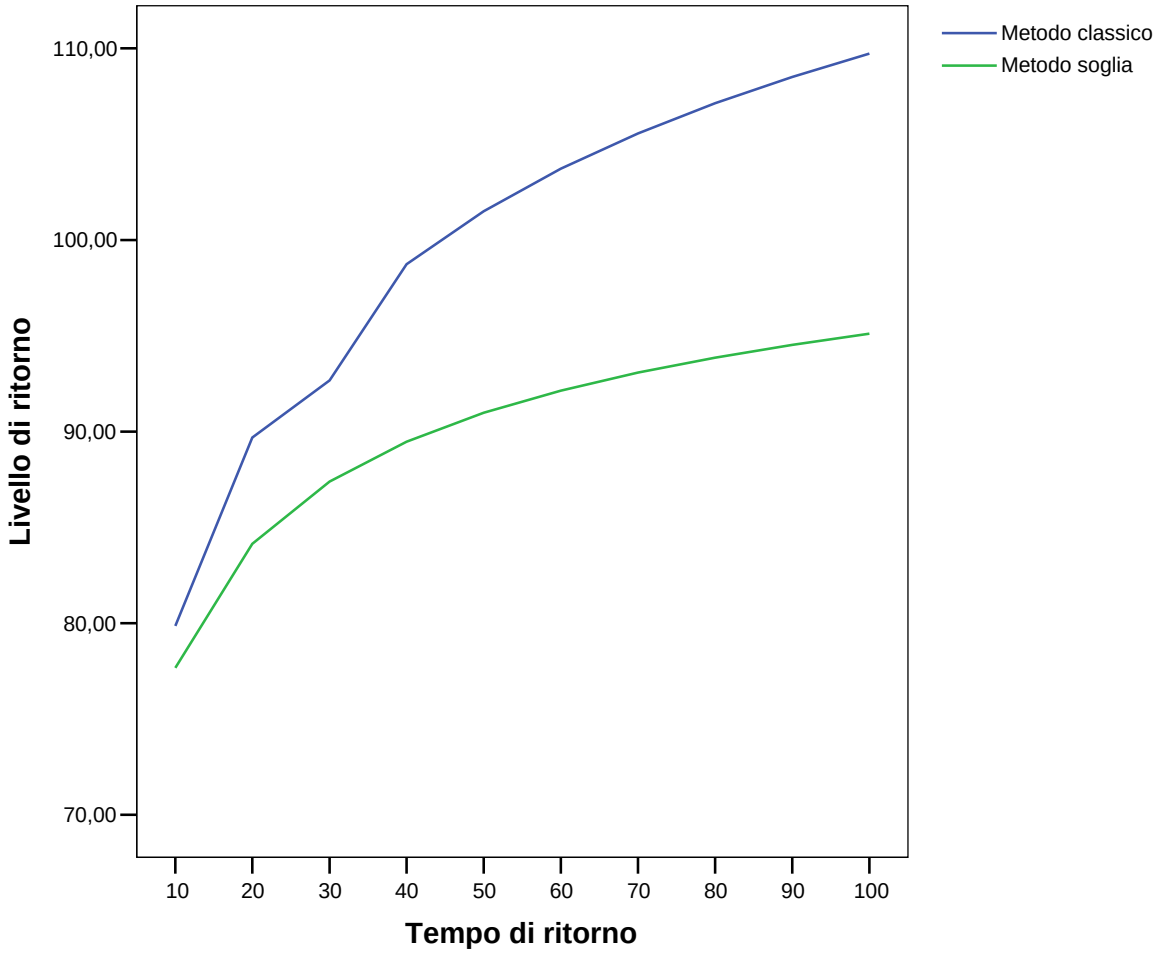


Fig.Livelli di ritorno dei valori massimi secondo il tempo di ritorno

Il grafico mostra in maniera più netta la differenza di comportamento segnalata precedentemente.

Infatti la curva del metodo classico cresce più rapidamente rispetto a quella stimata per il metodo soglia, e con l'aumentare degli anni di osservazione le curve tendono a distanziarsi.

Per rendere tuttavia più intuitivo il ragionamento fin qui sviluppato, può essere più utile rappresentare i valori riportati nella tabella precedente rapportandoli ai valori massimi che si sono registrati nel periodo di osservazione delle serie pluviometriche.

Un risultato negativo nella tabella indica che le stime non superano in valore assoluto gli estremi delle serie storiche di cui disponiamo.

E' logico attendersi per tempi più lunghi la precipitazione più intensa venga eguagliata o superata.

P	**Metodo Classico**	**Metodo Soglia**
1/10	-14,44	-13,93
1/20	-4,61	-7,46
1/30	-1,63	3,51
1/40	4,43	-2,13
1/50	7,2	-0,62
1/60	9,42	0,54
1/70	11,26	1,48
1/80	12,84	2,26
1/90	14,21	2,93
1/100	15,43	3,51

Tab.Differenza tra i livelli di ritorno stimati e il massimo valore registrato

Dalla tabella ci accorgiamo che entrambi i modelli prevedono che ci saranno degli aumenti negli estremi di precipitazione di massima intensità.

Precisamente, il valori massimi saranno superati in media una volta ogni 40 anni secondo il metodo classico o 50 anni secondo l'approccio basato sul metodo soglia ,periodi non particolarmente differenti da quello d'osservazione.

In analogia con il grafico presentato precedentemente, illustriamo nella rappresentazione successiva le differenze tra i livelli di ritorno stimati e il massimo valore osservato nel periodo di esame,sempre rapportati ai rispettivi tempi di ritorno.

Allo scopo di rendere più agevole la lettura del diagramma, è stata tracciata una linea al livello 0, corrispondente all'uguagliamento del massimo valore osservato nella serie storica.

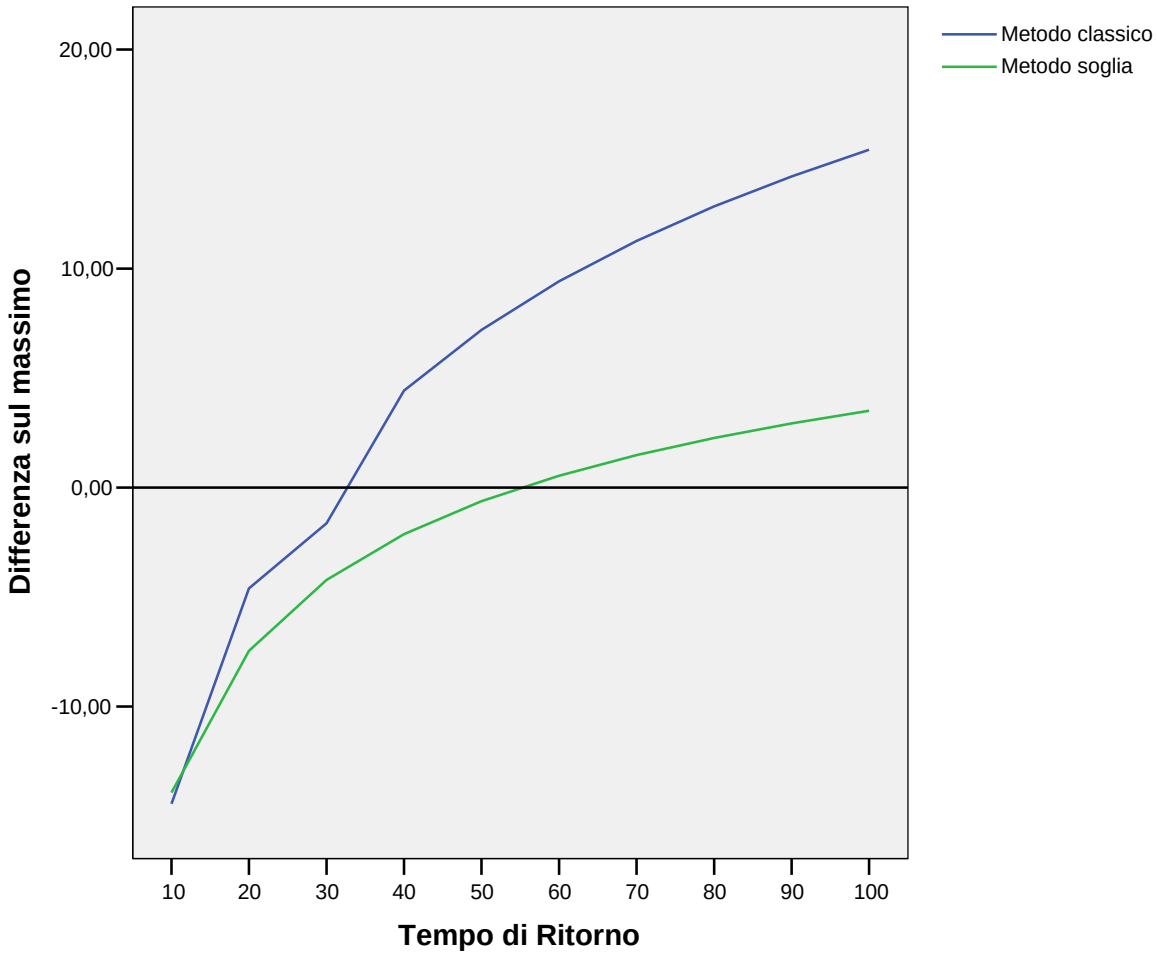


Fig. Differenze tra i livelli di ritorno stimati e il massimo osservato

Seppur su una scala diversa, il grafico porta alle stesse considerazioni effettuate per il diagramma dei valori assoluti.

Conclusioni

Nei passaggi precedenti di questo lavoro si sono analizzate due diverse metodologie maggiormente diffuse per la parametrizzazione delle distribuzioni di probabilità dei valori estremi.
In particolare si sono valutati i punti di forza e i limiti di due approcci differenti, quello classico, basato sui massimi per blocchi, e quello alternativo del metodo delle eccedenze.
In letteratura l'approccio classico gode di maggiore successo a causa della sua facilità di implementazione e sulla sua relativa affidabilità.
E' risaputo infatti ,che spesso,per le serie pluviometriche,come per altre variabili ambientali, sono largamente diffuse le serie storiche dei massimi annuali piuttosto che la serie completa delle osservazioni.
Quindi ,se da una parte il metodo classico risulta più agevole da implementare , dall'altra rinuncia per costruzione ad utilizzare una grande quantità di informazione qualora questa si fosse resa disponibile.
Nello studio della nostra serie pluviometrica comunque la tecnica classica si è rivelata adeguata a parametrizzare le nostre osservazioni, anche in presenza di eventi piovosi più estremi, che,da un punto di vista ambientale , rivestono un interesse particolare.

In sede di elaborazione del lavoro abbiamo utilizzato,sulla stessa serie pluviometrica,considerata nella sua completezza e non solo sui massimi annuali, un approccio differente, basato sul metodo soglia.
Nella pratica,quest'ultima metodologia,più che parametrizzare i valori massimi, si occupava di determinare una distribuzione di probabilità relativa ai valori che eccedevano un dato livello fissato a priori.
Sulla difficoltà di selezione della soglia si è già dibattuto precedentemente, ma è bene qui ricordare che questo aspetto anche in letteratura è stato affrontato in maniera controversa. Infatti non si è ancora giunti ad una visione univoca del problema, e le diverse metodologie adottate nella pratica statistica portano spesso a soluzioni estremamente differenti.

Abbiamo potuto constatare che il modello da noi stimato secondo la metodologia della soglia aveva una maggiore difficoltà ad “intercettare” i valori più estremi,sebbene questi ricadessero all’interno delle bande di confidenza stimate.
Quindi ,se statisticamente è stato ritenuto un modello appropriato per i nostri dati, nella pratica ambientale,e per i dati da noi utilizzati, il modello tende a sottostimare gli eventi più estremi .
Il fatto che alcuni massimi della serie delle osservazioni dei massimi per blocchi fossero leggermente superiori rispetto alle osservazione giornaliere,ha potuto comportare questa sostanziale differenza tra i modelli stimati. Sarebbe quindi interessante riconfrontare le due metodologie sulla base degli stessi valori estremi, dove il massimo annuale viene valutato non sulle 24h consecutive ma sul massimo giornaliero,onde verificare se le differenze tra i due modelli continuino a persistere.

Un altro aspetto da considerare è senz’altro quello relativo alle metodologie di stima, che per i modelli considerati si sono basate sugli stimatori di massima verosimiglianza, che comunque in letteratura vengono ritenuti adeguati allo scopo.
Tuttavia , questa pratica, dovrebbe essere valutata più attentamente in particolare per la stima dei parametri del modello soglia.
Una condizione per l’affidabilità delle stime di massima verosimiglianza risiede nel fatto che le osservazioni debbano essere incorrelate. Se questo può essere plausibile per la metodologia dei massimi per blocchi, la situazione è diversa per il metodo soglia,in quanto si rischia di inserire nel modello osservazioni tra loro correlate, coincidenti con periodi particolarmente piovosi. Pertanto sarebbe interessante valutare sulla stessa serie pluviometrica l’affidabilità di altri approcci più “raffinati”,come quello bayesiano, che non risentono delle limitazioni espresse dagli stimatori di massima verosimiglianza.

BIBLIOGRAFIA

COLDWELL, R. (2002). *Estreme value theory and applications to flood probabilty calculations.*

COLES, S. (1999). *Estreme value theory and applications. Notes in support of the course on extreme value theory and applications at Bucato, Sao Paulo, Brasil.*

COLES, S. (2001). *An introduction to statistical modeling of extreme values*. *Springer*.

COLES, S. & PERICCHI, L. (2001). *Anticipating catastrophes through estreme values modeling. Applied statistics.*

DALLA VALLE, G. & GAETAN, C. (2001). *Materiale didattico per il corso di inferenza statistica I.*

DAVISON, A. C. & SMITH, R. L. (1990). *Models for exceedances over high thresholds (with discussion). J.R.Statistical Society, B, 52*.

FABRIZIO, R.(2003) S*tudio statistico sulle serie storiche di precipitazione*

GILLELAND, E., KATZ, R., YOUNG, G. (2003). Extremes toolkit. Weather and climate applications of extreme value statistics.

GUMBEL, E. (1954). *Statistical theory of extreme values and some practical applications. A series of lectures*. Columbia university press.

GUMBEL, E.J. (1958). *Statistics of extremes*. Columbia university press.

MAIONE ,U.,MOISELLO U.,(2003)*Elementi di statistica per l'idrologia,*

KATZ, R. (2003). Statistics of weather and climate extremes.

REISS, R.D. & THOMAS, M. (1997). *Statistical analysis of extreme values*. Birkhauser Verlag

SMITH, R. L. (1991). Extreme value theory. In *Handbook of Applicable Mathematics*, volume 7, chapter 14. Wiley.

SMITH, R.L. (2001). Lecture notes on environmental statistics. University of Washington.

SMITH, R.L. (2003). The statistics of extremes.