Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea (Triennale) in

STATISTICA PER L'ECONOMIA E L'IMPRESA

ANALISI DEI VALORI ESTREMI : PRECIPITAZIONI NEVOSE SU NEW YORK

Relatore: Prof. Matteo Grigoletto
Dipartimento di Scienze Statistiche

Laureando: Erni Deliallisi
Matricola n. 1231407

Anno Accademico 2022/2023

Sommario

Lo scopo di questa relazione è quello di fornire al lettore una serie di metodi adatti all'analisi dei valori estremi di una serie storica; a tale scopo verrà utilizzato come esempio il caso di studio (M. Lee, J. Lee, 2020) relativo alle precipitazioni nevose sull'area urbana di New York, ponendo particolare attenzione alla tempesta di neve che colpì la citta nei giorni dal 22 al 24 Gennaio 2016, evento che causò innumerevoli perdite economiche e disagi alla popolazione. La serie di strumenti e metodi saranno forniti in maniera graduale, quasi "algoritmica", per rendere più chiara la collocazione e lo scopo delle varie tecniche adottate nell'analisi dei valori estremi; inoltre saranno, quando possibile, preferite argomentazioni intuitive a dimostrazioni rigorose. La lettura di questa relazione presuppone una conoscenza di base dell'inferenza statistica; in particolare è richiesta la conoscenza delle procedure di costruzione di modelli statistici basati sulla funzione di verosimiglianza, della formulazione di test d'ipotesi e della costruzione di intervalli di confidenza per variabili casuali; questi ultimi saranno costruiti tramite l'utilizzo di una particolare applicazione della tecnica di "Bootstrapping", della quale verranno richiamati i principi basilari. Dopo la presentazione della metodologia adottata, verranno forniti i risultati dell'analisi, accompagnati da indicazioni generali sulla loro corretta interpretazione.

Indice

Capitolo 1

Introduzione

Nei giorni tra il 22 e il 24 Gennaio 2016 una violenta bufera di neve ha travolto gli stati del nord est americano, provocando accumuli di oltre 90 centimetri giornalieri - 110 cm a Glengary, West Virginia - in alcune zone della regione; le stime meteorologiche prevedevano accumuli fino a 60 cm, ma questo valore fu superato in oltre sette stati diversi in quei giorni. Nello specifico la zona urbana di New York si sono registrati 91 cm di accumulo nevoso giornaliero (registrato presso la stazione meteorologica dell'aeroporto J. F. Kennedy, nel quartiere di Queens). Questi eventi hanno avuto un impatto catastrofico sulla popolazione, facendo registrare più di tre miliardi di dollari di danni e continue interruzioni del traffico civile e commerciale. In generale gli eventi meteorologici hanno una grande influenza sulle attività e su vita delle persone, quindi una loro accurata analisi è doverosa per poter meglio affrontare i cambiamenti che questi comportano. Questa relazione si occupa proprio di analizzare, utilizzando metodi adeguati, gli eventi accaduti nel gennaio 2016 nel nord est americano basandosi sulle rilevazioni nivologiche di quattro stazioni meteorologiche nella zona urbana di NY:

Stazioni Meteo analizzate			
Nome	Nome Completo	Coordinate	Alt.
Central Park	NEW YORK CNTRL PK TWR	40.7789°N 73.9692°W	132 ft
Newark	LIBERTY INTL AP	40.6825°N 74.1694°W	29 ft
La Guardia	LA GUARDIA AP	40.7794°N 73.8803°W	39 ft
JFK	JFK INTL AP	40.6386°N 73.7622°W	11 ft

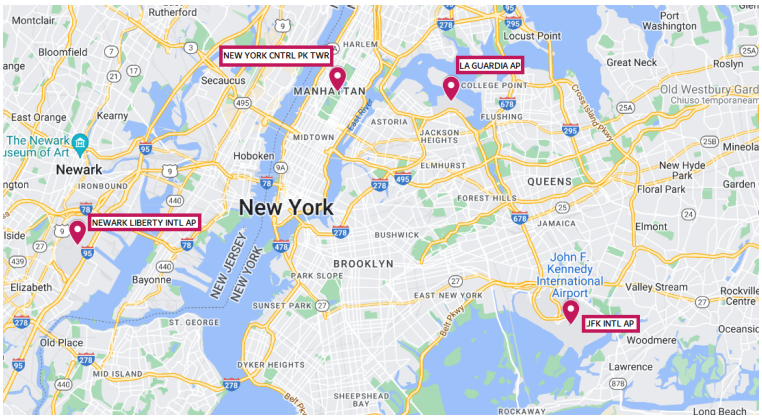


Figura 1.1: Mappa della zona urbana di New York con i segnaposti relativi alle quattro stazioni

Capitolo 2

Dati e analisi

I dati utilizzati nella ricerca fanno riferimento alle serie storiche giornialiere di precipitazione nevosa cumulata (in inch) a partire dal 1 Luglio 1959 fino al 30 Giugno 2015; non sono state incluse le osservazioni relative alla tempesta di neve di Gennaio 2016 per verificare le capacità previsive dei modelli che verranno adattati.
Le informazioni di natura prettamente geografiche riguardanti le stazioni meteorologiche interessate sono presentate tramite la tabella e la mappa presentate precedentemente.

2.1 Rifinitura dei dati

Prima di proseguire con l'applicazione dei modelli è di fondamentale importanza svolgere delle operazioni di rifinitura sui dati.
Sono escluse dall'analisi i giorni nei quali la precipitazione nevosa cumulata registrata sia ≤ 0.1, in quanto potrebbe trattarsi di un errore di misurazione; i nivometri calcolano la quantità di neve previa fusione della massa di ghiaccio accolta, si potrebbe dunque registrare, in giorni particolarmente umidi, il cosidetto "effetto rugiada", che spesso fa registrare valori non nulli nonostante l'effettiva mancanza di precipitazioni nevose.
Altra questione a cui bisogna porre particolare attenzione è il trattamento di rilevazioni su giorni contigui; è ragionevole assumere che la stessa bufera di neve abbia durata superiore al giorno oppure che questa si registri a cavallo di due giornate distinte; nel caso vi siano

due o più rilevazioni > 0.1 consecutive per le giornate $g_1, g_2, ..., g_m$, queste verranno sommate in un unico evento assunto come avvenuto durante g_1.
I valori mancanti (NA) sono esigui e tale problema può essere trascurato.

2.2 Scopi tecnici dell'analisi

Gli scopi tecnici dell'analisi sono sostanzialmente due; Rilevare eventuali trend relativamente ai valori estremi dei dati osservati (vedasi paragrafo 3.4) e prevedere, tramite intervalli di confidenza, livelli di ritorno per periodi di 25, 50, 75 e 100 anni. Verrà valutata anche la robustezza del modello statistico trattato, escludendo prima i dati relativi all'anno 2015 (1 Lug. 2014 - 30 Giu. 2015) per poi reincluderli; un modello robusto non dovrebbe reagire eccessivamente all'aggiunta di nuove osservazioni, fornendo delle stime dei parametri circa uguali.

Capitolo 3

Metodologia

In questo capitolo sarà trattata, passo dopo passo, la serie di strumenti statistici utilizzati nell'analisi proposta , accompagnata , ove necessario , dall' esplicazione di specifici teoremi alla base del funzionamento della metodologia adottata ; questi risultati verranno spiegati tramite argomentazioni intuitive , senza avventurarsi eccessivamente nei principi matematici sottostanti ; queste argomentazioni dovrebbero apparire sufficientemente chiare a chiunque abbia una buona conoscenza di base della teoria statistica e probabilistica. Il capitolo inizierà con l'introduzione dei modelli statistici utilizzati per le analisi, partendo, per ciascuno di essi, dalla presentazione del modello probabilistico caratterizzante, fino ad arrivare all'applicazione dei suddetti modelli tramite la teoria della verosimiglianza. Si noterà fin da subito che l'applicazione di tali modelli porta con sé una serie di problematiche alle quali bisogna porre particolare attenzione; verranno prontamente introdotti particolari metodi utilizzati per eliminare, o quantomeno ridurre la magnitudine, delle suddette criticità.

3.1 Modelli Statistici

Partiamo dalla presentazione di due particolari modelli statistici ampiamente utilizzati nell'analisi di valori estremi; si tratta dei modelli GEV (Generalized extreme values distribution) e GPD (Generalized Pareto distribution); vedasi (S. Coles, 2001, M. R. Leadbetter, 1983 per riferimento).

Questi modelli si occupano di definire in maniera più precisa - di quanto non facciano le più comuni distribuzioni di probabilità - il comportamento dei valori estremi; questi, infatti, nelle usuali analisi che si occupano di caratterizzare il comportamento dei quantili "centrali" di una certa distribuzione, vengono spesso negletti e trattati come eccezioni a cui non porre particolare importanza; Basti pensare alle analisi di validazione del modello basati sui confronti tra quantili empirici e teorici, ove ci si aspetta fin da subito una cattiva rappresentazione delle distribuzioni teoriche dei primi e ultimi quantili.

3.2 Distribuzione GEV

I modelli GEV si applicano a una particolare trasformazione dei dati originali; i Block maxima.
Il concetto di definizione dei Block maxima è alquanto intuitivo: si tratta di "dividere" le nostre osservazioni originali $\mathbf{X_1}$,...,$\mathbf{X_n}$ in m blocchi (non sovrapposti) di egual misura L; il Block maxima del m-esimo blocco, che chiameremo $\mathbf{M_m} = \max(\text{blocco } m - esimo)$, è semplicemente il valore più grande registrato tra le osservazioni incluse nell'm–$esimo$ blocco.
Quello dei Block maxima è uno dei metodi utilizzati nella letteratura statistica per distinguere valori "estremi" dai valori "comuni"; i modelli basati sulla distribuzione GPD si basano invece su un'altra ridefinizione dei dati di partenza, che verrà introdotta a tempo debito.
Una volta ottenuta la serie di Block maxima, ci ritroviamo di fronte al problema di ottenere la loro distribuzione; se si fosse a conoscenza della vera distribuzione cumulata F che genera le nostre osservazioni originali, la distribuzione cumulata G dei Block maxima è facilmente ottenuta:

$$\begin{aligned}\Pr\{M_m \leq z\} &= \Pr\{X_1 \leq z, ..., X_n \leq z\} \\ &= \Pr\{X_1 \leq z \times, ..., \times \Pr\{X_n \leq z\} \\ &= \{F(z)\}^n\end{aligned}$$

Tuttavia per non perdere di generalità, si supponga che la vera distribuzione $F(\cdot)$ che ha generato le osservazioni $X_1, ..., X_n$, sia invece ignota.
A questo punto è necessario introdurre un teorema distributivo di fondamentale importanza; il teorema dei tipi estremi, detto anche teorema di Fisher - Tippett - Gnedenko. Tale teorema afferma che una qualsiasi successione di Block maxima, costruita su una serie di osservazioni $X_1, ..., X_n$, n realizzazioni indipendenti da $F(\cdot)$ ignota , se opportunamente riscalata da due successioni numeriche $\{b_n\}$ e $\{a_n\} > 0$, tale che:

$$M_m* = \frac{M_m - b_n}{a_n}$$

M_m* si distribuisce approssimativamente come tre possibili distribuzioni di probabilità , detti "tipi" (da qui il nome del teorema) o "domini d'attrazione" , in quanto la distribuzione delle M_m* è "attratta" da una di queste tre famiglie:

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

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

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

La successione di Block maxima re scalata M_m* è, appunto, "attratta" da una di queste distribuzioni a seconda della distribuzione $F(\cdot)$ che ha generato i dati di partenza, in particolare è preso in considerazione il comportamento in coda di $F(\cdot)$ ovvero:

$$1 - F(x), \qquad x \to \infty.$$

Le successioni di Block maxima re scalate costruite sulle realizzazioni delle più comuni distribuzioni derivanti dalle famiglie di dispersione esponenziale, ad esempio, tendono a distribuirsi come una

Gumbell. Tuttavia, come detto precedentemente, la distribuzione $F(\cdot)$ è ignota, quindi determinare il dominio d'attrazione diventa tutt'altro che banale; in teoria basterebbe anche solo sapere alcune caratteristiche della distribuzione $F(\cdot)$ per determinarne il dominio d'attrazione, ma senza fare assunzioni distributive ciò diventa molto complicato.

Per ovviare a questo problema è stato proposto di utilizzare invece una particolare famiglia di distribuzione che riprenda le caratteristiche delle tre famiglie sopracitate e ne generalizzi il comportamento; si tratta della distribuzione GEV.
La distribuzione GEV – alla base dei modelli GEV – è caratterizzata da tre parametri;

$$G(z;\mu,\sigma,\xi) = \begin{cases} exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}}\right\}, & \xi \neq 0; \\ exp\left\{-exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, & \xi = 0; \end{cases}$$

μ rappresenta il parametro di posizione, σ parametro di scala e ξ parametro di forma.
Si noti che, nel caso $\xi = 0$, la GEV "decade" alla famiglia di distribuzione Gumbell (I); nel caso pratico si può assumere ciò anche quando ξ tende a zero.
La densità di una GEV con $\xi = 0$ ha dominio in tutto l'asse dei numeri reali (unbounded).
Se $\xi < 0$, la distribuzione GEV presenterà un "limite superiore"; nell'ambito della modellizzazione ciò equivale a dire che c'è una caratteristica nel processo generatore delle nostre osservazioni che rende inverosimile che queste ultime assumano valori maggiori di tale limite. Il limite – che in questo caso è sempre maggiore di μ - man mano che $\xi < 0$ aumenta in valore assoluto, tende a posizionarsi vicino al valore μ (è importante ricordare che la posizione del limite dipende anche dal valore del parametro σ). Se $\xi > 0$, la distribuzione presenterà invece un "limite inferiore"; nell'ambito della modellizzazione questa informazione è trascurabile in quanto ci saranno sempre valori osservati minori dei Block maxima. Il parametro assume quindi un significato leggermente diverso a quello precedente; man mano che ξ diventa più grande, le code della GEV diventano più pesanti.
Generalmente si tenga presente che è altamente improbabile che il

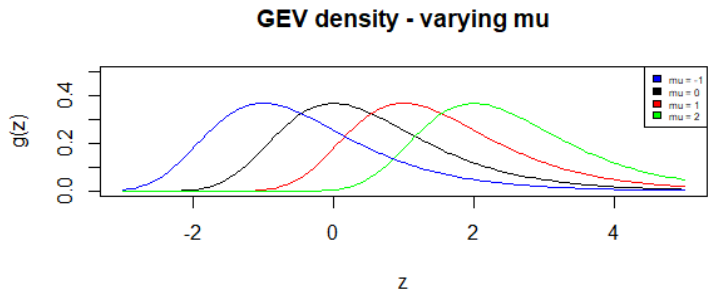


Figura 3.1: Funzione di densità GEV per differenti valori di μ, con $\sigma = 1$, $\xi = 0$.

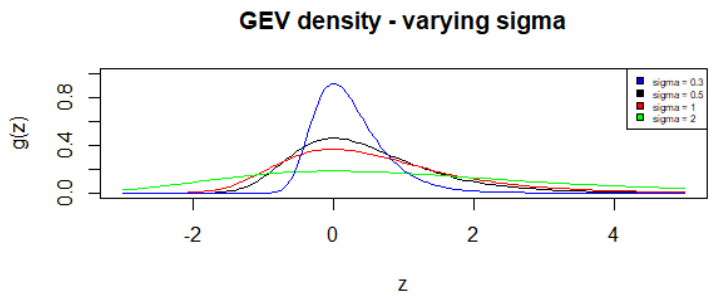


Figura 3.2: Funzione di densità GEV per differenti valori di σ, con $\mu = 0$, $\xi = 0$.

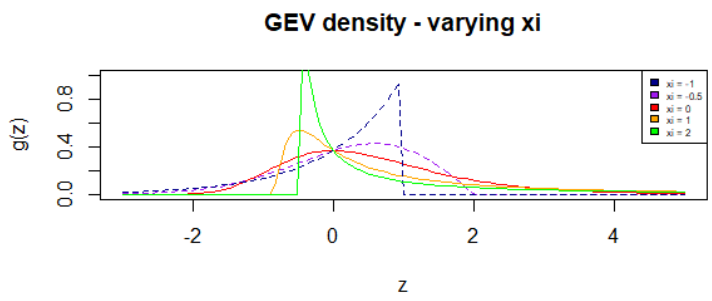


Figura 3.3: Funzione di densità GEV per differenti valori di ξ, con $\mu = 0$, $\sigma = 1$; le linee tratteggiate indicano le distribuzioni con $\xi < 0$; si noti che per $\xi \neq 0$ la funzione assume valore zero in determinati sottoinsiemi di $\mathbb{R}$.

vero valore $\xi < 0.5$.

3.3 Modelli GEV

In questo articolo analizzeremo l'applicazione di modelli statistici GEV basandosi sull'inferenza di verosimiglianza; come in ogni applicazione del genere il nostro obiettivo è quello di ottenere stime puntuali e intervallari dei tre parametri caratterizzanti le distribuzioni GEV, assumendo come fissate le osservazioni a nostra disposizione; utilizzeremo poi le stime ottenute per determinare il livello di ritorno, una particolare funzione dei parametri che discuteremo nella sezione (3.3.4) di questo articolo.
Lo spazio campionario coincide, generalmente, con $\mathbb{R}$, mentre lo spazio parametrico è $\mathbb{R}^2 \times \mathbb{R}^+$. Il modello rispetta le usuali condizioni di regolarità. È possibile ricavare la funzione di log – verosimiglianza a partire dalla densità GEV considerando come osservazioni la serie di Block maxima definita in precedenza, tuttavia non vi sono soluzioni analitiche che permettano di calcolare le stime e le relative misure di variabilità, pertanto è necessario implementare metodi numerici. Di seguito la funzione di log - verosimiglianza per modelli basati sulla GEV:

$$\ell(z;\mu,\sigma,\xi) = -m\log\sigma - (1+1/\xi)\sum_{\imath=1}^{m}\log\left[1+\xi\left(\tfrac{z_\imath-\mu}{\sigma}\right)\right] - \sum_{\imath=1}^{m}\left[1+\xi\left(\frac{z_\imath-\mu}{\sigma}\right)\right]^{-\frac{1}{\xi}},$$

$$Con: \quad 1+\xi\left(\frac{z_\imath-\mu}{\sigma}\right) > 0, \quad \forall \imath = 1,...,m$$

Nel caso sia supposto $\xi = 0$ si avrà invece:

$$\ell(z;\mu,\sigma,\xi) = -m\log\sigma - (1+1/\xi)\sum_{\imath=1}^{m}\left(\frac{z_\imath-\mu}{\sigma}\right) - \sum_{\imath=1}^{m} exp\left\{-\left(\frac{z_\imath-\mu}{\sigma}\right)\right\},$$

La scelta arbitraria di $\{\mathbf{b_n}\}$ e $\{\mathbf{a_n}\} > \mathbf{0}$ (vedasi paragrafo 3.2), nel-

l'ambito della modellizzazione non crea problematiche particolari in quanto la stima dei parametri si "adatterà" facendo si che il modello fornisca comunque una buona rappresentazione delle osservazioni a disposizione.

Nei prossimi paragrafi introdurremo la presenza di eventuali problematiche che potrebbero intaccare l'accuratezza dei risultati del modello; ad esempio, il teorema di Fisher - Tippett - Gnedenko, alla base della distribuzione GEV, assume indipendenza tra le osservazioni, ma questa assunzione non sempre trova riscontro nei casi pratici. I metodi che verranno introdotti aiuteranno a migliorare l'approssimazione dei risultati del modello.

3.4 Non Stazionarietà della Serie Storica

Nell'ambito dell'analisi di una serie storica si intende per stazionarietà (in senso debole) la proprietà della serie di avere media e varianza invarianti nel tempo. L'assunzione di stazionarietà non è sempre soddisfatta, pertanto si è resa necessaria l'introduzione di un metodo che riuscisse a rappresentare efficacemente il cambiamento dei primi due momenti nel tempo; in generale si possono ridefinire i parametri del modello come funzione del tempo:

$$\mu = \mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + ... + \beta_p t^p$$

In questo esempio il parametro μ è stato espresso come funzione del tempo secondo un trend polinomiale di grado p, ma questo tipo di operazione può essere svolto anche sugli altri parametri ($\sigma = \sigma(t)$, $\xi = \xi(t)$) e con diverse tipologie di trend.
A seconda di come viene ridefinito un prarametro vengono aggiunti al modello un certo numero di coefficienti β; nell'esempio precedente, ridefinendo μ si aggiungono p parametri β.

Nel caso di studio trattato si è supposta invarianza nel tempo dei parametri σ e ξ lasciando variare μ per la j-esima stazione secondo un trend lineare;

$$\mu_j = \mu_j(t) = \beta_{0j} + \beta_{1j}\left(\frac{t-1958}{10}\right), \qquad j = 1, .., 4;$$

Dove il termine "1958" è utile per facilità d'interpretazione (la serie inizia nel 1959) e il termine "10" al denominatore cambia l'unità di tempo dall'anno al decennio; operazione ragionevole essendo la precipitazione nevosa un evento poco volatile.
Questo passaggio è di fondamentale importanza in quanto si desidera sapere se, durante il periodo di analisi, vi sono stati dei trend relativi ai soli valori estremi per confrontare tali risultati con quelli di E. Burakowski,2008, secondo la quale vi è un generale trend negativo sulla quantità di precipitazioni nevose su New York.
Per ogni stazione vengono aggiunti al modello un coefficiente di intercetta β_0 e un coefficiente di variazione β_1 in sostituzione al originale parametro μ; verranno eventualmente posti vincoli di uguaglianza tra coefficienti di diverse stazioni per parsimonia del modello.

3.5 Problemi di dipendenza

La modellizzazione proposta dalla letteratura assume che le osservazioni siano indipendenti tra di loro; questa assunzione può essere confermata o meno a seconda dei casi di studio; nel nostro caso specifico si prenda in considerazione la possibile esistenza di strutture di dipendenza temporale e spaziale tra le varie osservazioni.

3.5.1 Dipendenza temporale

Nei modelli GEV la dipendenza temporale è trattata confrontando le distribuzioni cumulate dei Block maxima ottenuti sugli stessi dati con e senza l'assunzione di dipendenza temporale tra i dati (intuitivamente questo specifico problema si pone nell'ambito dello studio su serie storiche, dove è verosimile la presenza di svariate strutture di dipendenza tra le osservazioni).
Risulta che le due distribuzioni sovracitate differiscano di un singolo termine esponenziale theta :

Si assuma che $X_1, ..., X_n$ sia un processo stazionario che ammette correlazione tra le osservazioni, mentre $X_1^*, ..., X_n^*$ sia una sequenza di variabili i.i.d. con la stessa distribuzione marginale di $X_1, ..., X_n$; Si supponga ora di calcolare:

$$M_n = \max\{X_1, ..., X_n\},$$
$$M_n^* = \max\{X_1^*, ..., X_n^*\}.$$

Si può dimostrare che, sotto condizioni di regolarita;

$$\Pr\{(M_n^* - b_n)/a_n \leq z\} \to G_1(z),$$

Per $n \to \infty$, con $\{a_n\} > 0$, $\{b_n\}$ sequenze numeriche e $G_1(z)$ funzione non degenere si otterrà che:

$$\Pr\{(M_n - b_n)/a_n \leq z\} \to G_2(z),$$

Intercorre la relazione:

$$G_2(z) = G_1^{\theta}(z),$$
$$\theta \; costante, \quad 0 < \theta \leq 1.$$

Theta è chiamato extremal index ("indice d'estremità" in italiano) e rappresenta, in maniera inversamente proporzionale, la grandezza dei cluster che tendono a formarsi nei valori estremi; per essere più chiari, un valore di $\theta = 1$ – che rappresenta perfetta indipendenza – sta a indicare che i cluster formati sono di grandezza approsimativamente uguale a $1^{-1} = 1$, vale a dire che i valori tendono a non raggrupparsi e a stare "distanti" , nel tempo, tra di loro. Viceversa, per $\theta \to 0$ – che rappresenta una dipendenza perfetta tra i dati – sarà indice del fatto che la grandezza dei cluster tende a $\frac{1}{0} \to \infty$; nel caso concreto, vi è una tendenza dei valori "similmente" estremi a presentarsi in istanti di tempo contigui, implicando che questi provengano dallo stesso evento generatore.
Tuttavia il problema di dipendenza temporale nelle modellizzazioni basate su distribuzioni GEV viene spesso eliminato tramite la scelta di Block maxima adeguati alla specifica situazione analizzata; nel caso di studio trattato in questo testo - valori estremi di eventi nivologici - ,ad esempio, la scelta di blocchi di durata annua e ricoprenti

il periodo di tempo a partire dal 1 Luglio fino al 30 Giugno dell'anno seguente, rende i Block maxima virtualmente indipendenti, vista l'alta probabilità dell'evenienza che il massimo annuo di precipitazione cumulata nevosa giornaliero sia registrato nei mesi invernali.
Meno banale sarà il caso dei modelli GPD.

3.5.2 Dipendenza spaziale

Un fenomeno a cui va posta particolare attenzione è la dipendenza spaziale tra più stazioni vicine; questo tipo di dipendenza va soprattutto ad intaccare gli standard error associati alle stime di massima verosimiglianza quando vengono posti vincoli sul modello, rendendo le stime più precise di quanto non siano realmente;
Per essere più chiari, ipotizziamo che le osservazioni di stazione A e stazione B (di numerosità rispettivamente n e m) siano incluse, tramite dei vincoli posti sul modello di verosimiglianza, nel processo di stima di un parametro comune (ad esempio può essere posto il vincolo $\mu_A = \mu_B$), per l'usuale principio di parsimonia dei parametri (nel caso di studio è posto il vincolo $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$).
Ciò porta comunque ad ottenere stimatori consistenti per i parametri vincolati; tuttavia, la numerosità campionaria viene artificialmente inflazionata (la nuova numerosità sarà pari a $n+m$), portando a una diminuzione dello standard error associato.
Questo aumento di numerosità non è tuttavia giustificato da un aumento della mole di informazione a disposizione, in quanto le osservazioni allo stesso istante temporale delle due diverse stazioni – ricordiamo, prossime nello spazio – sono, con tutta probabilità, generate dallo stesso evento (in questo caso si tratterà della stessa bufera di neve).
Si può quindi dire che il raggruppamento dei dati delle due stazioni, A e B, porti una informazione minore rispetto a quanta ne porterebbe se fosse vera l'usuale assunzione di indipendenza tra i dati, e minore informazione implica un aumento degli standard error associati allo stimatore comune.
Non essendo a conoscenza delle specifiche strutture di dipendenza spaziale che intercorrono tra i dati, si può utilizzare il metodo di Smith; Il metodo di Smith non fa altro che ricavare la matrice di varianza – covarianza degli stimatori dei parametri "liberi" del

modello (costruito col metodo della verosimiglianza) senza avvalersi dell'approssimazione che vuole:

$$Cov(\hat{\theta}) \approx H^{-1}$$

Dove H rappresenta la matrice di informazione osservata . Tale approssimazione è sufficientemente corretta solo se è assunta indipendenza tra i dati. Il metodo di Smith, invece, ricava la matrice di varianza covarianza del modello analizzando la variazione del gradiente della log – verosimiglianza associata rispetto ai parametri; la seguente riformulazione della matrice di varianza covarianza risulta essere più accurata:

$$Cov(\hat{\theta}) \approx H^{-1} V H^{-1}$$
$$V = Cov(\nabla \ell(\theta_0)) = mCov(\nabla h_1(\theta))$$

Dove $h_\imath(\theta)$ rappresenta l'$\imath$-esimo contributo alla log verosimiglianza da parte delle quattro diverse stazioni.
Le stime $\hat{H}$ e $\hat{V}$ sono ottenute rispettivamente per via numerica e valutando la variazione dei singoli contributi rispetto ai parametri liberi.

Tendendo a mente le metodologie presentate per ovviare ai problemi di non stazionarietà e dipendenza tra le osservazioni, si procederà alla definizione del modello completo e al successivo confronto tra modelli annidati tramite il criterio AICc (Corrected Akaike Information Criterion), così definito;

$$AICc = 2p - 2\ell + \frac{2p(p+1)}{n-p-1}$$

Dove p è il numero di parametri e ℓ è il valore della funzione di log - verosimiglianza in corrispondenza del suo valore massimo. Disponiamo finalmente delle stime di massima verosimiglianza per i parametri della distribuzione GEV che meglio rappresenta i Block maxima campionari.

3.6 Livelli di ritorno

L'ottenimento delle stime di massima verosimiglianza non ha un'utilità divulgativa immediata, in quanto l'interpretazione dei parametri non risulta essere banale, e questi possono sembrare, agli occhi di una persona inesperta, totalmente insignificanti.
Un modo molto più intuitivo di descrivere l'evenienza di diversi valori estremi e la loro magnitudine è tramite l'introduzione del livelli di ritorno x_k associato a un definito periodo di ritorno k;
questo risulta essere un espediente utilizzato anche nella vita quotidiana; sentiamo spesso pronunciare frasi del genere: "un terremoto di scala 7.0 è stato registrato per l'ultima volta 200 anni fa (nella determinata zona)", oppure: "non accadeva da circa 50 anni che si toccassero temperature di 45 gradi celsius (nella data località)".
Nel primo esempio la magnitudine di 7.0 sulla scala Richter rappresenta il livello di ritorno associato al periodo di 200 anni e, analogamente, 50 anni è il periodo di ritorno associato all' evento di livello 45 gradi celsius. Esempi del genere danno un'idea della magnitudine dell'evento registrato e del legame stretto che intercorre tra periodo e livello di ritorno.

Il livello di ritorno è, matematicamente, una funzione dei parametri e di un periodo k; di seguito è espressa la funzione periodo di ritorno per la distribuzione GEV:

$$x_k = \begin{cases} \mu - \frac{\sigma}{\xi}\left[1 - \{-ln(1-k^{-1})\}^{-\xi}\right], & se\ \xi \neq 0; \\ \mu - \{-ln(1-k^{-1})\}, & se\ \xi = 0. \end{cases}$$

x_k è ottenuta a partire dalla inversa della funzione cumulativa di una GEV, ovvero la sua funzione Quantile, che viene poi trasformata in modo tale da restituire i valori più alti in concomitanza delle probabilità prossime a zero; È importante una corretta interpretazione di tale funzione, ricordando che essa restituisce il livello che ci si aspetta sia ecceduto ogni KL, dove L = (lunghezza dei blocchi scelti); dal momento che, come nello studio, è prassi scegliere lunghezza dei blocchi pari a un anno, per facilità interpretativa, il livello sarà ecceduto ogni K anni, il che dovrebbe risultare abbastanza chiaro al lettore di un eventuale ricerca che implementa modelli

basati su distribuzioni di valori estremi.

3.7 Stima intervallare per livelli di ritorno

E' importante fornire intervalli di confidenza per i livelli di ritorno; essendp x_k una funzione dei parametri si potrebbe pensare di implementare il metodo delta per tale scopo.
In questo paragrafo saranno introdotti metodi che possano risultare in intervalli di confidenza più precisi rispetto al metodo delta. Vedremo come stimare la distribuzione dello stimatore $\hat{x_k}$ tramite bootstrap e come sfruttare al meglio tale informazione.

3.7.1 Bootstrap a blocchi

Le statistiche di Wald per x_k, ottenibili tramite metodo Delta sono certamente giustificate dalle proprietà asintotiche degli stimatori ottenuti da un modello di verosimiglianza sotto condizioni di regolarità; tali risultati si possono basare però su assunzioni poco realistiche.
Un metodo che permette di approssimare meglio (di quanto non facciano i metodi analitici) la distribuzione dello stimatore $\hat{x_k}$ è il "Bootstrap a blocchi" (H. R. Kuensch , 1989); rispetto al classico ricampionamento bootstrap, il bootstrap a blocchi riesce a cogliere meglio una possibile struttura di dipendenza (si dice che viene mantenuta la struttura di covarianza) tra le osservazioni; tale tecnica è implementata come descritto di seguito:

- Definiamo n – l + 1 'blocchi' di osservazioni sovrapponibili , di lunghezza l
- Scegliamo n / l blocchi che andranno a comporre un nuovo ricampionamento
- Calcolare, sui dati del nuovo ricampionamento, la stima del livello del ritorno per il periodo k scelto ;

Ripetendo questi tre passi per B volte otteniamo B valori simulati $\hat{x_k}^*$ del livello di ritorno che rappresentano la distribuzione bootstrap dello stimatore $\hat{x_k}$; viene persa comunque la struttura di covarianza nel passaggio da un blocco all'altro; la seguente è una spiegazione intuitiva del perché tale metodo risulti più preciso del metodo bootstrap classico, nel caso sia plausibile l'esistenza di una qualche struttura di dipendenza tra i dati:
Se nel "mondo reale", le osservazioni reali tendono a presentarsi secondo un certo pattern, è di nostro interesse che quest'ultimo sia mantenuto, per quanto possibile, anche nei ricampionamenti del "mondo bootstrap"; per esempio, nel caso di una serie storica trimestrale, la scelta di blocchi di lunghezza $l = 4$ potrebbe aiutare a conservare la componente stagionale. Contrariamente, applicare il bootstrap a blocchi su una serie di osservazioni sulle quali è ragionevole assumere indipendenza tra di esse può portare a conservare pattern frutto del caso anche nel "mondo bootstrap"; ciò conduce a stime meno affidabili.

3.7.2 Intervalli $\mathbf{BC_a}$

Ottenuta la distribuzione bootstrap si potrebbe pensare di ottenere l'intervallo di confidenza per il livello di ritorno prendendo i quantili $[\frac{\alpha}{2}]$ e $[1 - \frac{\alpha}{2}]$ della suddetta distribuzione, come da prassi nel procedimento di stima intervallare bootstrap.
Tuttavia, è stato dimostrato (B. Efron, 1987) che tale modo di procedere si basa su troppe costrizioni che portano in molti casi a intervalli di confidenza imprecisi; infatti , prendere i quantili $[\frac{\alpha}{2}]$ e $[1 - \frac{\alpha}{2}]$ di una distribuzione equivale ad assumere Non Distorsione e Varianza Stabile (ovvero il sunto che questa rimanga costante per ogni valore assunto dal parametro da stimare); ciò equivale a pensare a una distribuzione simmetrica e centrata sul vero valore del parametro.
Tuttavia, non è ovvio che queste due condizioni siano riscontrate in ogni caso di studio, quindi, per avere un metodo di calcolo di intervalli di confidenza su una distribuzione bootstrap più generale si è supposto che la distribuzione bootstrap possa essere potenzialmente "decentrata" rispetto al vero valore del parametro (da qua la Bias Correction) e asimmetrica.

Il decentramento della distribuzione è trattato tramite una costante, z_{BC}, che ha il ruolo di "spostare" in media l'intera distribuzione bootstrap verso il vero valore del parametro; z_{BC} sostanzialmente è una misura della distorsione della stima bootstrap rispetto alla stima puntuale, che per analogia sarà approssimativamente uguale alla distorsione della stima puntuale rispetto al vero valore del parametro. Nel nostro caso specifico la costante di distorsione può essere calcolata come:

$$z_{BC} = \Phi^{-1}\left(\frac{1}{B}\sum_{b=1}^{B} I\left(\hat{x}_k^{(b)} < \hat{x_k}\right)\right),$$

Dove $\sum_{b=1}^{B} I\left(\hat{x}_k^{(b)} < \hat{x_k}\right)$ è una rappresentazione di $\hat{F}^*(\hat{x_k})$, la funzione di ripartizione bootstrap di $\hat{x_k}$.
Si può notare da questa formula come, nel caso in cui l'esatta metà delle B replicazioni bootstrap sia inferiore della stima puntuale del livello di ritorno, la costante $z_{BC} = 0$. Meno banale è l'interpretazione del coefficiente di accelerazione C_a; tale coefficiente rappresenta la variazione della varianza del parametro trattato rispetto al valore del parametro stesso;
nel caso di una distribuzione simmetrica si avrà un valore della costanteC_a pari a 0, in quanto la varianza rimane costante. Analisi su tale coefficiente hanno portato a stabilire che una stima soddisfacente di C_a è ottenuta come come $\frac{1}{6}SKEW(i(\hat{x_k}))$, dove $SKEW(i(\hat{x_k}))$ è l'indice di asimmetria della funzione di score della distribuzione bootstrap d'interesse in questo caso:

$$i(\hat{x_k}) = \frac{\partial log(f^*(\hat{x_k}))}{\partial x}$$

Dove $f^*(\hat{x_k})$ è la funzione di densità bootstrap stimata.
$SKEW(i(\hat{x_k}))$ nel caso di studio può essere stimata tramite (J. A. Hoeting, G. H. Givens) :

$$SKEW(i(\hat{x_k})) = \frac{\sum_{t=1}^{n}(\ddot{x_k}^{(-t)} - \ddot{x_k})^3}{\left[\sum_{t=1}^{n}(\ddot{x_k}^{(-t)} - \ddot{x_k})^2\right]^{3/2}},$$

Dove $\ddot{x}_k{}^{(-t)}$ rappresenta la stima del livello del ritorno omettendo la t-esima osservazione (delete 1- jackknife estimation) e $\ddot{x}_k$ è $\frac{1}{n}\sum_{t=1}^{n}\ddot{x}_k{}^{(-t)}$. Il metodo jackknife si rivela efficace nella stima dei momenti secondo e terzo, che compongono l'indice di asimmetria. Bisogna aver ben presente che le quantità z_{BC} e C_a vanno a modificare i quantili che utilizzeremo nel calcolo dell'intervallo di confidenza, quindi , invece dei classici quantili $[\frac{\alpha}{2}]$ e $[1-\frac{\alpha}{2}]$, considereremo:

$$z\left[\frac{\alpha}{2}\right] = \Phi\left(z_{BC} - \frac{z_{BC} - z_{1-\alpha/2}}{1 - C_a(z_{BC} - z_{1-\alpha/2})}\right),$$
$$z\left[1-\frac{\alpha}{2}\right] = \Phi\left(z_{BC} - \frac{z_{BC} + z_{1-\alpha/2}}{1 - C_a(z_{BC} + z_{1-\alpha/2})}\right),$$

Dove $z_{1-\alpha/2}$ è il quantile 1 - $\alpha/2$ della normale standard.
Di conseguenza l'intervallo di confidenza BCa al livello 1 - α sarà:

$$IC_{1-\alpha} = \left[\hat{F^{-1*}}\left(z\left[\frac{\alpha}{2}\right]\right), \hat{F^{-1*}}\left(z\left[1-\frac{\alpha}{2}\right]\right)\right],$$

Dove $F^{-1*}(\cdot)$ è la funzione quantile bootstrap calcolata.
Questo intervallo dovrebbe essere il più preciso possibile per la descrizione del livello di ritorno per un periodo successivo k; si noti comunque che, per k approssimativamente maggiore di due volte la lunghezza della serie considerata (ad esempio k = 40 (in anni) e lunghezza della serie di 20 anni), il modello GEV tende a fornire stime inaffidabili; pertanto, l'applicazione dei metodi al fine di fornire intervalli corretti potrebbe risultare uno sforzo inutile.

3.8 Distribuzione GPD

Per l'analisi dei valori estremi si può utilizzare un ulteriore modello statistico, basato sulla distribuzione di Pareto generalizzata (GPD, Generalized Pareto Distribution).
Tali modelli utilizzano una diversa sintesi dei dati rispetto ai modelli GEV; questi infatti, invece dei block maxima, vengono applicati sulle osservazioni che risultano essere maggiori di una certa soglia, che

indicheremo come u. La selezione dei soli block maxima può rappresentare una evitabile perdita di informazione se, nei vari blocchi , esistono altri valori estremi, o se, in generale, questi non si distribuiscono uniformemente tra i blocchi stessi. La selezione di valori soglia – eccedenti può risolvere tale problema.

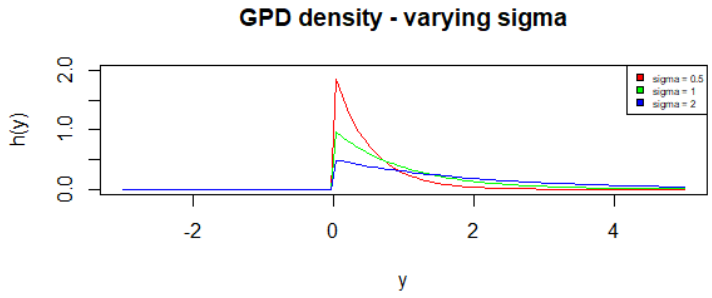


Figura 3.4: Funzione di densità GPD per differenti valori di σ, con $\xi = 0$, u = 0.

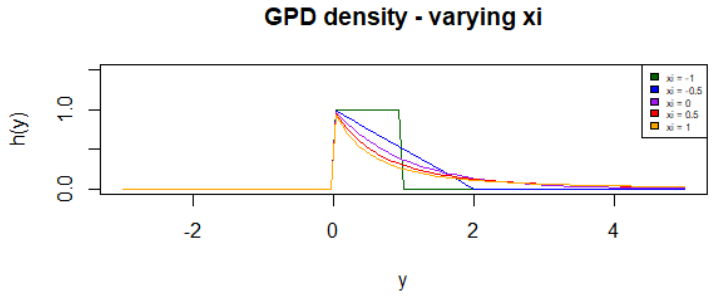


Figura 3.5: Funzione di densità GEV per differenti valori di ξ, $\sigma = 1$, u = 0; il limite inferiore di una GPD è sempre u, mentre il limite superiore esiste solo se $\xi < 0$.

La distribuzione GPD è derivata dalla distribuzione GEV; se infatti, data una sequenza $X_1, X_2, \ldots X_N$ per cui vale il teorema dei tipi estremi, la distribuzione GPD sarà la distribuzione di $(X_i – u)$ condizionata a $X_i > u$:

$$(X_i - u | X_i > u) \dot{\sim} H(y) = 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-\frac{1}{\xi}}$$

Definita in :

$$\{y : y > 0 \wedge (1 + \xi y/\tilde{\sigma}) > 0\},$$
$$con \ : \ \tilde{\sigma} = \sigma + \xi(u - \mu).$$

Con u che rappresenta la soglia scelta. In seguito denoteremo semplicemente come σ il termine $\tilde{\sigma}$; tale parametro, in virtù di questa traformazione, viene chiamato anche "Parametro di scala modificato".
Dalla definizione precedente si può capire come solo le osservazioni $X_i > u$ – valori che chiameremo soglia eccedenti – verranno considerate dalla distribuzione. La famiglia di distribuzione GPD è caratterizzata da due parametri; σ e ξ. Rispetto alla distribuzione GEV manca il parametro μ, ma questo può essere pensato come sostituito dal valore u della soglia; questo infatti, similmente al ruolo del parametro μ nel contesto GEV, fornisce una misura di posizione dei valori estremi nell'asse dei numeri reali; ricordando che la GPD è la distribuzione asintotica dei valori $(X_i - u|X_i > u)$, non vi è realmente bisogno di un ulteriore indicatore di posizione in quanto i valori estremi considerati (i valori soglia eccedenti in questo caso), "partiranno", sempre da zero. Per far capire meglio il concetto si consideri un caso di simulazione da una GPD con una precedentemente fissata soglia u; verranno prima generati una serie di valori dalla densità GPD (che ha forma simil esponenziale), parametrizzata da ξ e σ. A questi valori generati verrà poi sommata la soglia u per "riportare" tali valori a una posizione congrua rispetto alla serie completa di osservazioni.
La soglia u non rappresenta pertanto un parametro vero e proprio, in quanto non influisce sulla funzione di distribuzione ma solo sulla trasformazione dei dati originali; da questo punto di vista, invece, si può pensare ad una analogia con la scelta della lunghezza L dei block maxima nel caso GEV.
Si noti come, nel caso di una $\mathbf{GPD}(\sigma, \xi = 0)$ questa equivalga a una $\mathbf{Esp}(\lambda)$ con $\sigma = \lambda$.

3.9 Modelli GPD

La procedura di applicazione ai dati di un modello basato sulla distribuzione GPD si compone fondamentalmente di due passi:

- Determinazione della soglia u
- Inferenza di verosimiglianza sui parametri

Analizzeremo come approcciarsi a questi due passaggi.

3.8.1 Determinazione della soglia

La scelta della soglia di un modello GPD è un operazione meno banale di quanto non possa sembrare; non esiste infatti una formula analitica per tale scelta, ma bisogna basarsi fondamentalmente su analisi grafiche; a tal proposito vengono considerati il MRL (Mean Residual Life plot) e il Grafico di stabilità ; Il mean residual life plot (esempio in Figura 3.6 a pagina 25) è una rappresentazione grafica del seguente luogo dei punti:

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

Questo luogo dei punti mette a confronto le possibili scelte della soglia u nello spazio campionario con la media dei valori soglia eccedenti associati. Scegliamo il valore u_0 più piccolo oltre il quale la funzione diventa simil lineare, questo poiché, se $(X_i - u|X_i > u) \sim$ **GPD**(σ, ξ), allora:

$$E[(X_i - u|X_i > u)] = \frac{\sigma_{u_0} + \xi u}{1 - \xi},$$
$$Per\ u > u_0$$

Dove u_0 è il valore minimo teorico che fa valere la distribuzione GPD per un certo set di osservazioni e σ_{u_0} è il valore di sigma in corrispondenza della scelta di tale soglia;

Si può notare come il valore atteso dipenda linearmente da u quando u > u_0
Pertanto scartiamo valori di $u < u_0$ come descritto precedentemente poiché un andamento non lineare del MRL è sintomo che per tali valori non sia ancora valida la distribuzione asintotica a una GPD, e seguono ancora la distribuzione ignota $F(\cdot)$ secondo la quale si distribuisce l'intera serie di osservazioni (si ricordino le assunzioni fatte nel paragrafo del GEV); in concomitanza del valore u_0 i valori iniziano ad avere le caratteristiche di una distribuzione GPD e possiamo discriminare tali valori come Estremi; scegliamo proprio u_0 per includere nel modello più osservazioni possibili.

Un altro metodo grafico per determinare u è il grafico di stabilità dei parametri $\hat{\sigma}$ e $\hat{\xi}$; questo metodo consiste nello stimare i valori dei due parametri per diverse scelte di u. Dalle equazioni si può dedurre come, all'aumentare di u, ξ dovrebbe rimanere costante mentre σ dovrebbe variare linearmente; tracciati i grafici (esempio in Figura 3.7 a pagina 25) u contro $\hat{\sigma}$ e $\hat{\xi}$, scegliamo u_0 dove vengono individuate tali caratteristiche.
Dopo aver incluso in un unico campione tutte le osservazioni relative alle quattro stazioni, se ne trova una soglia u_0 comune; questa soglia corrisponderà a un determinato quantile di tale campione, che utilizzeremo per una regressione quantile che ammetta quattro intercette diverse per le serie di ogni stazione; a questo punto si possono porre vincoli di uguaglianza tra i vari coefficienti di intercetta (che corrispondono ai valori soglia) per rendere più parsimonioso il modello.

3.10 Extremal Index per modelli GPD

A questo punto si passa alla stima dei parametri della distribuzione GPD, in maniera simile a quanto è stato fatto con modelli basati sulla distribuzione GEV; durante questo processo di stima la soglia è considerata come valore fissato. Rispetto alla modellizzazione GEV vi è da porre particolare attenzione ai problemi causati dalle strutture di dipendenza temporale; nel caso precedente si è detto che, trattando block maxima annuali questi si possano assumere indi-

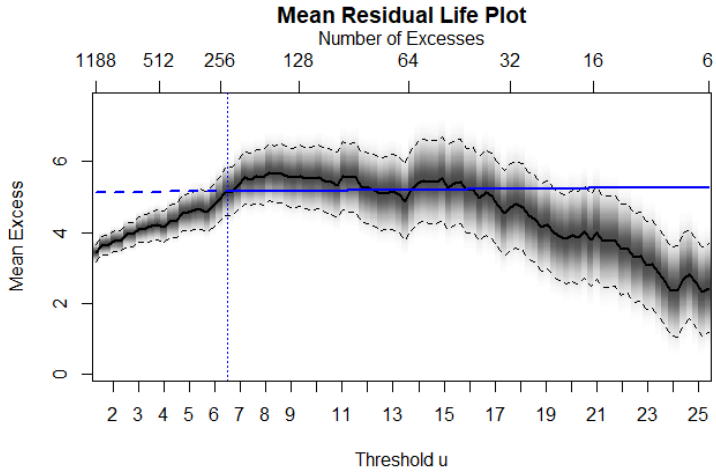


Figura 3.6: Mean Residual Plot per il campione includente le osservazioni di tutte le stazioni, accompagnato da relativi intervalli di confidenza simmetrici basati sulla normalità asintotica della media.

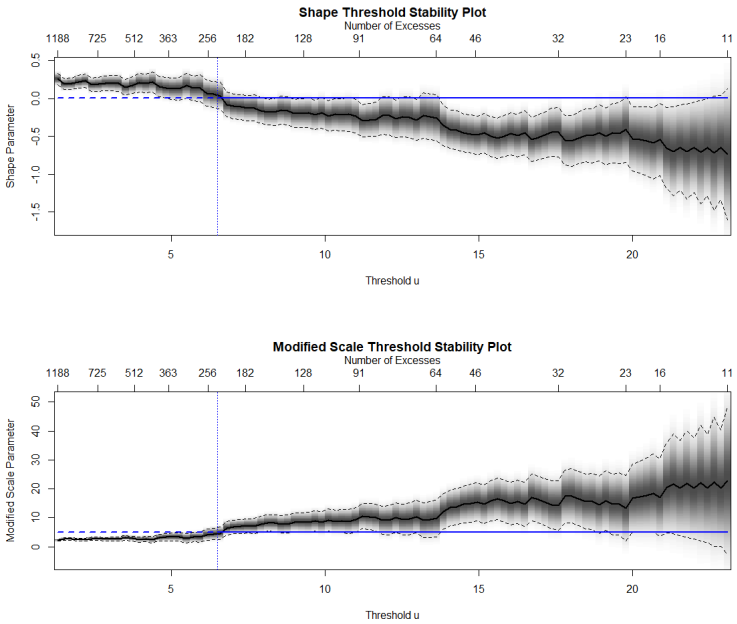


Figura 3.7: Grafici di stabilità dei parametri ξ (primo grafico) e σ (secondo grafico); I grafici sono accompagnati da intervalli di confidenza simmetrici basati sulla normalità asintotica degli SMV. Sia MRL che Grafici di stabilità suggeriscono un valore $u_0 \approx 6$.

pendenti con valori relativi agli anni precedenti o passati; trattando valori sovra eccedenti, è possibile che si presentino due o più osservazioni anomale in un breve periodo di tempo, quindi assumere che tali osservazioni siano indipendenti potrebbe risultare poco accurato.
Non vi è un metodo analitico per determinare l'extremal index(che d'ora in poi indicheremo come θ), ma è stato introdotto da C. A. T. Ferro e J. Segers un metodo di stima puntuale e intervallare per tale parametro. Secondo Ferro e Segers uno stimatore corretto per θ è:

$$\hat{\theta} = min \left\{ \frac{2\left[\sum_{t=1}^{n-1}(\Delta_t - a_1)\right]^2}{(n-1)\sum_{t=1}^{n-1}(\Delta_t - a_1)(\Delta_t - a_2)},\ 1 \right\},$$

Dove:

- N è il numero di valori soglia - eccedenti data la soglia u_0
- $\Delta = \{\Delta_1, ..., \Delta_{n-1}\}$ sono i tempi che intercorrono tra due valori soglia - eccedenti; in particolare il generico Δ_i è una misura del tempo trascorso tra l'i-esimo valore soglia eccedente e il (i+1)-esimo valore soglia eccedente.
- $a_1 = a_2 = 0$ se $max\{\Delta_1, ..., \Delta_n\} \leq 2$, e $a_1 = 1$ e $a_2 = 2$ altrimenti.

Stimato θ puntualmente si può procedere a una sua stima intervallare; Essendo θ approssimativamente la grandezza media dei cluster formatesi si può assumere che nel campione ci siano $L = \theta n$ cluster. Distinguiamo i tempi "Intercluster" come i $L - 1$ tempi Δ_i più grandi che intercorrono tra le osservazioni anomale successive. Implicitamente, i tempi Intercluster definiscono i cluster (di lunghezza variabile) come la serie di osservazioni comprese tra due tempi Intercluster; i tempi Δ_i che intercorrono tra osservazioni anomale successive appartenenti allo stesso cluster sono definiti come tempi "Intracluster". Con queste informazioni si possono ricreare delle sequenze di tempo simulate che seguano la divisione di cluster della serie originale; per fare ciò si supponga che una sequenza di tempo ricampionata sia così costruita:

- Scegliere casualmente uno degli L cluster costituiti e considerare i tempi Intracluster relativi.
- Scegliere casualmente uno degli L-1 tempi Intercluster definiti; accodare ai tempi intracluster selezionati precedentemente tale valore.

Ripetere questi step per L volte (nello step L-esimo non è necessario selezionare un ulteriore valore intercluster in quanto non vi sarà un cluster che lo succederà);

$$
\begin{aligned}
&Tempi\ Intercluster: \\
&\qquad \{\Omega_1, ...\Omega_{L-1}\}, \\
&Tempi\ Intracluster: \\
&\qquad Cluster_1 = \{\omega_{1\,1}, ..., \omega_{1\,n_1}\} \\
&\qquad Cluster_2 = \{\omega_{2\,1}, ..., \omega_{2\,n_2}\} \\
&\qquad \vdots \\
&\qquad Cluster_L = \{\omega_{L\,1}, ..., \omega_{L\,n_L}\}, \\
&Ricampionamento: \\
&\qquad \tilde{\Delta} = \{Cluster_{i_1}, \Omega_{j_1}, ..., Cluster_{i_L}\}, \\
&\qquad Per\ :\ i_{(\cdot)} = 1, ..., L\ e\ j_{(\cdot)} = 1, ..., L-1.
\end{aligned}
$$

Dove n_i è il numero di tempi Intracluster nel Cluster i-esimo.
Calcolando B copie $\hat{\theta}^*$ di $\hat{\theta}$ sostituendo a Δ, B ricampionamenti $\tilde{\Delta}$ si otterà la distribuzione bootstrap stimata $f^*(\cdot)$ di $\hat{\theta}$.

Similmente con quanto accade con gli altri parametri, si possono porre dei vincoli di uguaglianza sui parametri $\theta_j,\ j = 1, ..., 4$ relativi alle quattro serie storiche per rendere più parsimonioso il modello.

La eventuale non stazionarietà della serie in esame viene trattata in modo analogo a quanto discusso riguardo i modelli GEV, ma con la definizione di soglie che seguono un trend lineare come:

$$u = u(t) = u + \beta_1 \left(\frac{t}{365.25 \times 10} \right)$$

La stima di θ risolve il problema della dipendenza temporale, mentre, per ciò che riguarda la dipendenza spaziale, viene sempre applicata la correzione di Smith sugli standard error stimati.
Il tradeoff tra capacità esplicativa del modello e parsimonia è sempre misurato tramite il criterio di Akaike corretto, AICc, col quale si confrontano tra di loro diversi modelli annidati.
Per l'inferenza sui livelli di ritorno si utilizzino i metodi del bootstrap a blocchi e degli intervalli BCa per ottenere stime intervallari più precise possibile; si ricordi di includere nella stimatore l'extremal index θ. La funzione livello di ritorno in un modello GDP è la seguente:

$$x_k = \begin{cases} u + \frac{\sigma}{\xi}[(kd\zeta_u\theta)^\xi - 1], & se\ \xi \neq 0; \\ u + \sigma ln(kd\zeta_u\theta), & se\ \xi = 0. \end{cases}$$

Dove k è il periodo di ritorno considerato, d è il numero medio di osservazioni annue, ζ_u è la probabilità che un valore ecceda la soglia u:

$$\hat{\zeta_u} = \frac{numero\ di\ valori\ soglia\ eccedenti}{numero\ di\ osservazioni\ totali}.$$

La varianza di $\hat{\zeta_u}$ è definita come di seguito:

$$Var(\hat{\zeta_u}) = \frac{\hat{\zeta_u}(1 - \hat{\zeta_u})}{numero\ di\ osservazioni\ totali}.$$

Capitolo 4

Risultati

Per la presentazione dei risultati si decide di utilizzare il seguente modello GPD non stazionario, dove è stato assunto $\xi = 0$; tale modello risulta essere il migliore secondo il criterio AICc:

Modello GPD non stazionario (1959-2014) | ($\ell = -712.8704$)				
Stazione	u	β_1	$\hat{\sigma}$	$\hat{\theta}$
C. Park	4.633(0.638)	0.591(0.091)	4.527(0.501)	0.892(0.082)
Newark	3.985(0.425)	0.591(0.091)	4.527(0.501)	0.901(0.066)
La G.	3.985(0.425)	0.591(0.091)	4.527(0.501)	0.901(0.066)
JFK	3.985(0.425)	0.591(0.091)	4.527(0.501)	0.901(0.066)

Che forniscono i seguenti livelli di ritorno stimati $\hat{x_k}$:

Livelli di ritorno modello GPD (1959-2014)				
Stazione	$k = 25$	$k = 50$	$k = 75$	$k = 100$
C. Park	27.06090	31.72224	35.05050	37.83792
Newark	27.17189	31.83323	35.16149	37.94891
La G.	27.17189	31.83323	35.16149	37.94891
JFK	27.17189	31.83323	35.16149	37.94891

Tra parentesi sono riportati gli standard error associati (per il parametro σ è riportato lo s.e. corretto col metodo di Smith).
Si piò notare come siano stati posti i seguenti vincoli;

- $\sigma_1 = \sigma_2 = \sigma_3 = \sigma_4$.
- $\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14}$.
- $u_2 = u_3 = u_4$.
- $\theta_2 = \theta_3 = \theta_4$.

Il risultato di ciò è che il modello assume uguale distribuzione dei valori soglia eccedenti per le stazioni di Newark ,La Guardia e JFK; dando un occhiata alle caratteristiche geografiche delle stazioni questo potrebbe avere senso, data la diversa altitudine della stazione di Central Park e alla diversa collocazione (Central Park si trova in una area urbana densa di abitazioni mentre le altre tre stazioni sono poste in corrispondenza di aeroporti).

4.1 Robustezza del modello

Si valuta ora la robustezza del modello trattato aggiungendo i dati relativi all'anno 2015; le stime fornite (mantenendo le medesime soglie e l'assunzione $\xi = 0$) sono le seguenti:

Modello GPD non stazionario (1959-2015) | ($\ell = -735.0137$)				
Stazione	u	β_1	$\hat{\sigma}$	$\hat{\theta}$
C. Park	4.633(0.638)	0.563(0.088)	4.763(0.535)	0.901(0.090)
Newark	3.985(0.425)	0.563(0.091)	4.763(0.535)	0.916(0.074)
La G.	3.985(0.425)	0.563(0.088)	4.763(0.535)	0.916(0.074)
JFK	3.985(0.425)	0.563(0.088)	4.763(0.535)	0.916(0.074)

Che forniscono i seguenti livelli di ritorno stimati $\hat{x_k}$:

Livelli di ritorno modello GPD (1959-2015)				
Stazione	$k = 25$	$k = 50$	$k = 75$	$k = 100$
C. Park	27.78221	32.54002	35.89504	38.68096
Newark	27.88201	32.63981	35.99483	38.78075
La G.	27.88201	32.63981	35.99483	38.78075
JFK	27.88201	32.63981	35.99483	38.78075

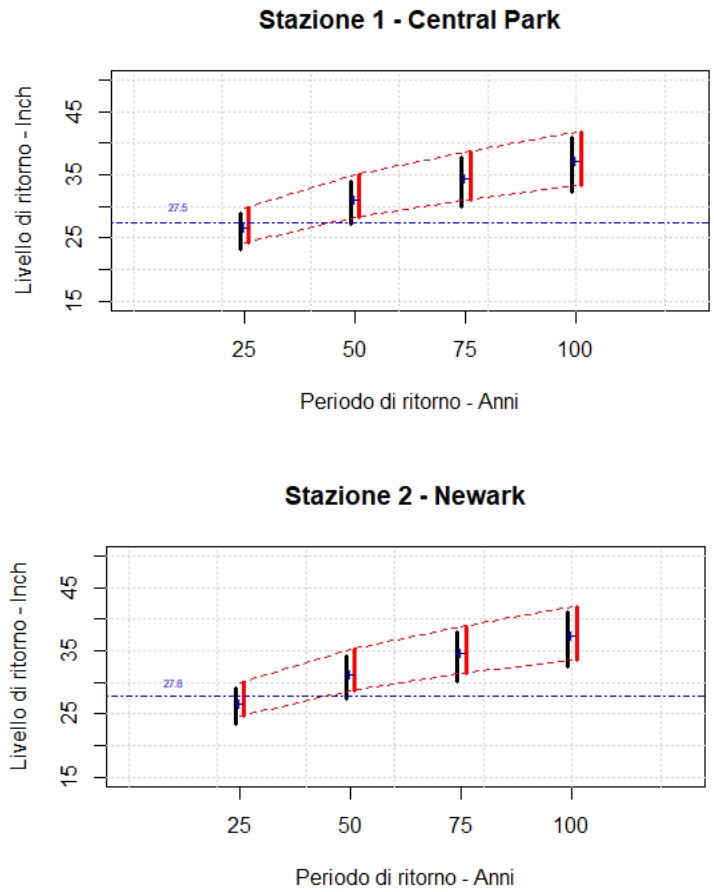
Si può notare che il valore dei parametri stimati dal modello sulla serie (1959-2015) non differisce significativamente dai valori stimati sulla serie (1959-2014); analogamente, rimangono circa uguali ($\approx +1$ *inch*) anche le stime dei livelli di ritorno per i periodi considerati. Possiamo quindi concludere che il modello non subisce particolarmente l'influenza di nuovi dati; riesce quindi a coglierne abbastanza bene il meccanismo generatore senza farsi influenzare troppo dalle sue realizzazioni casuali.

4.2 Livelli previsti

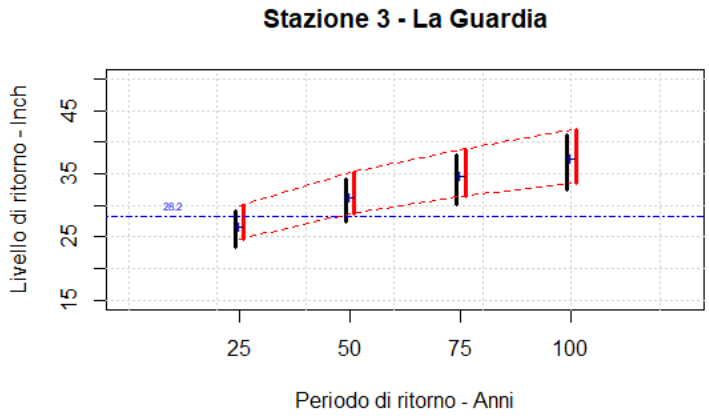
Il modello GPD non stazionario stimato sulla serie (1959-2014) manca dell'osservazione relativa alla bufera di neve del 2016, pertanto, di seguito, si confronterà il valore registrato in quell'occasione con i livelli di ritorno stimati dal modello

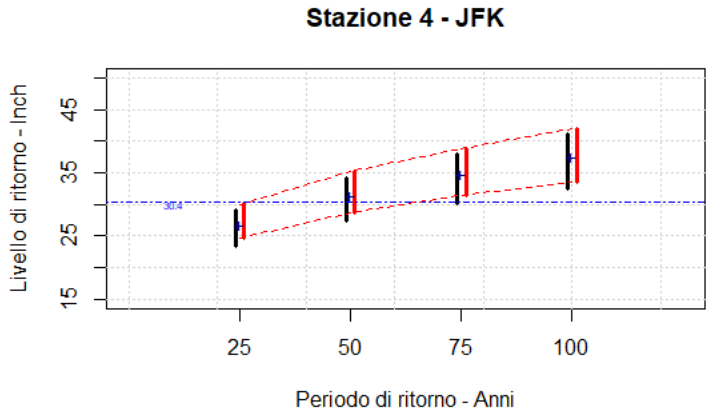
per $k = 25$, 50 75 e 100; verranno inoltre forniti i relativi intervalli di confidenza naive e BC_a.

Figura 4.1: I seguenti grafici confrontano i valori registrati durante la bufera del 2016 (linea blu tratteggiata) con i valori $\hat{x_k}$ previsti dal modello (simbolo +) accompagnati da intervalli di confidenza naive (Linee Nere) e BC_a (Linee Rosse); si può notare come gli intervalli BC_a siano centrati sul valore previsto a differenza degli intervalli naive.



Si può vedere dai grafici (Figura 4.1) come il modello preveda che il livello della bufera di neve del 2016 si ripresenta ciclicamente ogni 40 anni circa;i valori di ≈ 27.5 inch registrati nelle stazioni di Central Park, Newark e La Guardia hanno un periodo di ritorno stimato di circa 25 anni, mentre il picco di 30 inch ($\approx$ 90 cm) registrato presso la stazione JFK ha un periodo di ritorno maggiore di


Stazione 3 - La Guardia
Livello di ritorno - Inch
15
25
35
45
28.2
25
50
75
100
Periodo di ritorno - Anni


Stazione 4 - JFK
Livello di ritorno - Inch
15
25
35
45
30.4
25
50
75
100
Periodo di ritorno - Anni

circa 50 anni.

4.3 Stima del trend dei valori anomali

Abbiamo definito alla fine del paragrafo 3 l'assunzione di non stazionarietà del modello GPD; in particolare è utile verificare la significatività del parametro β_1 secondo il test d'ipotesi:

$$\begin{cases} H_0 & : \beta_1 = 0, \\ H_1 & : \beta_1 \neq 0; \end{cases}$$

Per verificare H_0 si potrebbe usare la statistica di Wald relativa, ma è preferito utilizzare un intervallo di confidenza bootstrap e accettare H_0 nel caso questo contenga il valore 0.

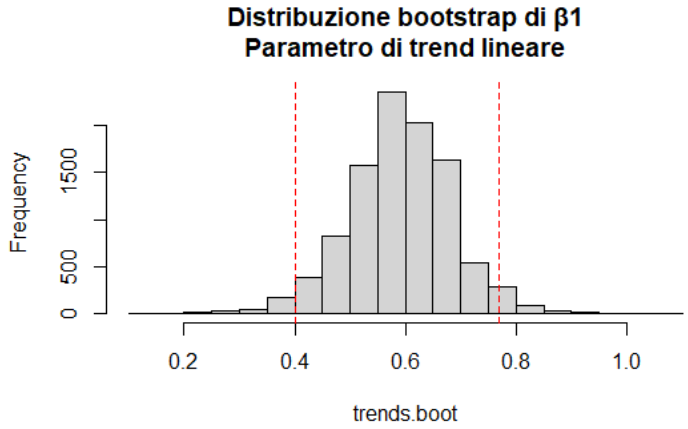


Figura 4.2: Distribuzione bootstrap di $\hat{\theta}$: le linee tratteggiate verticali rosse evidenziano l'intervallo di confidenza di livello $\alpha = 0.05$, $IC_{0.05} = [0.4006, 0.7696]$.

Come evidenziato dalla Figura 4.2, l'intervallo bootstrap non contiene il valore zero, pertanto accettiamo l'ipotesi dell'esistenza del trend lineare stimato $\beta_1 = 0.591$; nella pratica ciò significa che, i valori estremi di precipitazioni nevose cumulate giornaliere tendono ad aumentare di 0.591 Inch ($\approx 1.5\ cm$) a decennio.

Capitolo 5

Conclusioni

Si conclude che modelli GEV e GPD si dimostrano essere adatti per stima e la previsione di livelli di ritorno di un evento anomalo; analizzando la serie dei massimi annuali (Figura 5.1) si evince effettivamente che, nei 57 anni (1959-2016) i livelli della bufera del Gennaio 2016 vengono raggiunti poche volte, confermando i risultati dell'analisi.
Per quanto riguarda il trend degli eventi anomali si percepisce come la magnitudine di questi sia in aumento. Questo va in contrasto con lo studio di Burakowski, E. et al. (2008) che afferma un generale trend negativo sulla quantità di precipitazioni nevose generali; abbiamo quindi un importante informazione sull'andamento delle precipitazioni nevose: sebbene queste stiano generalmente diminuendo (effetto attribuito al riscaldamento globale), stanno aumentando gli eventi estremi, sottolineando un importante risvolto negativo del recente cambiamento climatico.

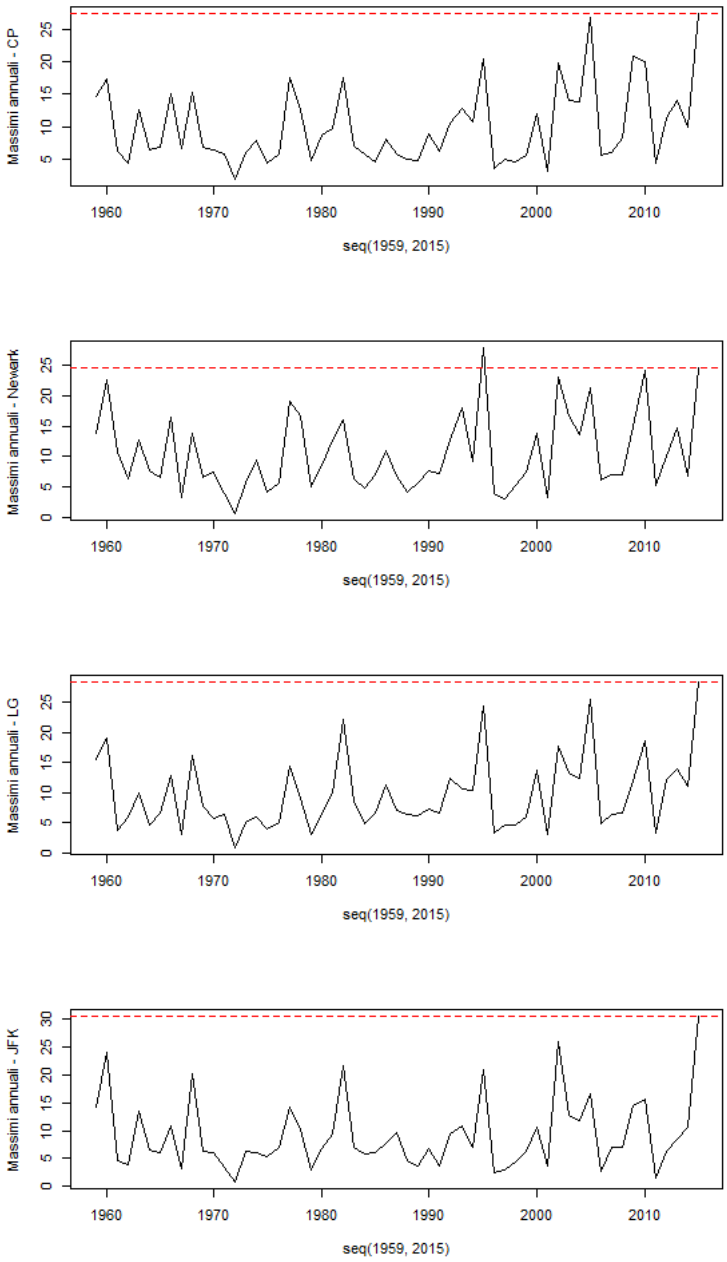


Figura 5.1: Serie storiche dei massimi annui per le quattro stazioni considerate; in rosso tratteggiato il livello della bufera del 2016.

Bibliografia

M. Lee, J. Lee (2020) "Trend and Return Level of Extreme Snow Events in New York City", The American Statistician, 74:3, 282-293, https://doi.org/10.1080/00031305.2019.1592780.

Burakowski, E. A., Wake, C. P., Braswell, B., and Brown, D. P. (2008), "Trends in Wintertime Climate in the Northeastern United States: 1965– 2005," Journal of Geophysical Research.

Coles, S. (2001), "An introduction to Statistical Modelling of Extreme Values", Springer Series in Statistics.

Leadbetter, M. R., Lindgren, G., Rootzén, H.,(1983) "Extremes and Related Proprerties of Random Sequences and Processes", Springer Series in Statistics.

Smith, R. L. (1990), "Regional Estimation From Spatially Dependent Data", Unpublished manuscript.

Efron, B., (1987), "Better Bootstrap Confidence Intervals." Journal of the American Statistical Association, vol. 82, no. 397, pp. 171–85. JSTOR, https://doi.org/10.2307/2289144

Kuensch, H. R.,(1993)"The Jackknife And The Bootstrap For General Stationary Observations", The Annals of Statistics, DOI:10.1214/aos/1176347265.

Givens, G. H. , Hoeting, J. A. ,(2013)"Computational Statistics",Wiley Series in Computational Statistics.

Ferro, C. A. T., Segers, J., (2003), "Inference for Clusters of Extreme Values." Journal of the Royal Statistical Society. Series B (Statistical Methodology), vol. 65, no. 2, pp. 545–56. JSTOR, http://www.jstor.org/stable/3647520.