UNIVERSITÀ DEGLI STUDI DI PADOVA

FACOLTÀ DI SCIENZE STATISTICHE

CORSO DI LAUREA
IN STATISTICA E TECNOLOGIE INFORMATICHE

RELAZIONE FINALE

APPLICAZIONE DELLA TEORIA DEI VALORI ESTREMI PER LA STIMA DELLA DIMENSIONE MASSIMA DELLE CELLULE DEGLI ALBERI

RELATORE: CH.MO PROF. STUART COLES

LAUREANDA: ELISA RIZZO

ANNO ACCADEMICO 2003-04

A Nicola e ai miei genitori

Indice

Introduzione

Il tronco sostiene meccanicamente la pianta, attraverso esso fluisce l'acqua che dalle radici viene trasportata alle foglie. Nel tronco ci sono tante cellule di dimensione diversa entro le quali scorre l'acqua. Ovviamente più la cellula è grande e più acqua riesce a trasportare. La portata d'acqua dipende, infatti, dalla quarta potenza del raggio al quadrato. Naturalmente la dimensione cellulare non è infinita ma fortemente sotto controllo da parte della pianta stessa. La nostra analisi ha come scopo quello di stimare il valore della dimensione massima possibile. Poiché questo valore dipende molto dall'albero in questione, infatti è un carattere utilizzato per la classificazione delle diverse specie di piante, andremmo a vedere come cambia questo limite in tre diverse specie di alberi: Larice, Pioppo nero e Robinia.

Per fare questa analisi utilizzeremo la Teoria dei Valori Estremi, un insieme di procedure scientificamente e statisticamente razionali utili a stimare il comportamento estremo di variabili o processi casuali. La Teoria dei Valori Estremi è intrinsecamente collegata all'estrapolazione. Data una serie di dati indipendenti $X_1 ... X_n$ da un'ignota distribuzione F, il problema consiste nello stimare accuratamente la coda di F. La difficoltà è data dal fatto che i dati sono normalmente concentrati vicino al centro della distribuzione, quindi i dati estremi sono scarsi e la stima diviene difficile.

Si potrbbe pensare che studiare l'andamento di probabilità rare non abbia molto senso, ma ci sono molte aree di applicazione dove questi calcoli sono indispensabili, diventa così ragionevole sviluppare tecniche scientifiche adeguate. La Teoria dei Valori Estremi è applicata in molti campi, il più comune è l'area

dell'ambiente e dei processi ambientali, come: il livello del mare, la velocità del vento, il livello dei fiumi o delle piogge, la concentrazione di inquinamento. Livelli estremi di questi processi possono portare a problemi gravi, l'intento è quindi quello di studiare la probabilità di questi eventi in modo da prevedere e ridurre i problemi ad essi collegati.

Nella tesina verranno approffonditi solo alcuni modelli della Teoria dei Valori Estremi. In particolare, dopo il primo capitolo, in cui presentiamo i dati, nel Capitolo 2 verrà approfondito il Modello Classico della stima dei valori estremi, basato sulla distribuzione Generalizzata dei Valori Estremi (GEV). Nel Capitolo 3 vedremmo alcune interressanti considerazioni su questa distribuzione. Nel Capitolo 4, invece presentiamo il Modello della Soglia che è basato sulla distribuzione Generalizzata di Pareto (DGP). Nel Capitolo 5, infine, confronteremo i risultati ottenuti dai due metodi e vedremmo i vantaggi che darebbe un'analisi Bayesiana applicata a questo contesto.

Capitolo 1

Dati sulla dimensione cellulare

I dati considerati riguardano tre tipi di albero, il Larice, il Pioppo nero e la Robinia. Per ognuno si sono considerate delle sezioni trasversali alla base del tronco e si sono considerati alcuni campioni di 1 cm^2 in diverse parti della circonferenza. Per ogni campione si sono misurati i diametri cellulari.

Lo scopo dello studio è stimare, con tecniche opportune, il valore massimo assoluto di lume cellulare che la pianta ammette. Partendo dunque, solo da alcuni campioni della superficie, vogliamo conoscere il massimo dell'intera superficie. I dati sono stati raccolti in due diversi modi:

1. Per ogni campione si sono considerati solo i valori massimi. L'unità di misura sono i micrometri quadrati ($10e^{-9}$ millimetri quadrati). Il dataset consta di 25 osservazioni, ogni colonna rappresenta un albero. Per motivi tecnici, di misurazione, il numero di osservazioni è diverso per i tre alberi. I valori mancanti sono indicati con NA. Abbiamo dunque 25 osservazioni per il Larice, 21 per il Pioppo e solo 14 per la Robinia.

	Larice	Piopponero	Robinia
1	1031.38	11244.70	26486.04
2	1151.46	11458.30	27658.45
3	1152.12	11486.78	30316.55
4	1649.90	11510.51	30947.85

```
5  1837.67   11833.28   34298.95
...
22 2836.30         NA         NA
23 3336.27         NA         NA
24 3353.95         NA         NA
25 3632.65         NA         NA
```

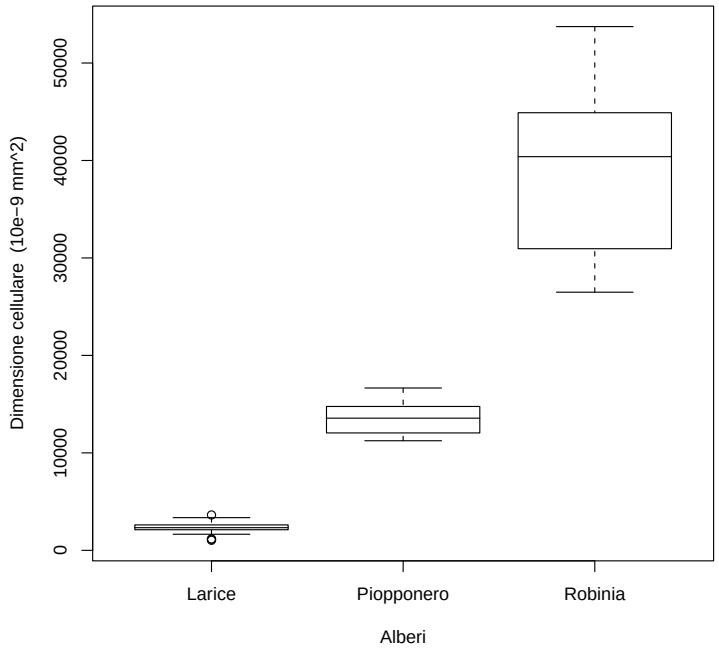

Figura 1.1: Box-plot per i dati sui massimi.

I diagrammi a scatola della fig. 1.1 consentono di confrontare la distribuzione dei 3 alberi. Essi indicano che la dimensione cellulare è molto differente tra i 3 alberi. Il Larice è sicuramete quello con dimensioni più piccole, la sua mediana è uguale a 2325, seguito dal Pioppo e dalla Robina con mediane rispettivamente pari a 13561 e 40394. C'e una grande differenza anche per quanto riguarda la varianza. Rispetto alla grande

variabilità della Robinia, mediana e quantili del Larice quasi coincidono, tanto che il grafico risulta molto schiacciato. Questa differenza è probabilmnte dovuta al diverso numero di osservazioni per i tre gruppi. Il 50% dei dati del Larice, quelli compresi tra il primo e il terzo quantile, stanno circa nell'intervallo [2100 , 2600], quelli del Pioppo nel [12000 , 15000], mentre quelli della Robinia nel [32000,53000]. La dimensione cellulare della Robinia, dunque, è in media, il doppio di quella del Pioppo che a sua volta è 5 volte quella del Larice. Dalla fig. 1.2 notiamo che tutte e tre le distribuzioni sono abbastanza simmetriche. Il Larice è l'unico che presenta valori anomali. Come indicato nella tabella 1.1, il valore massimo riscontrato per i tre alberi è rispettivamente 3633, 16656 e 53736.

	Larice	Pioppo nero	Robinia
Min.	1031	11245	26486
1^0 Qu.	2116	12047	31786
Mediana	2325	13561	40394
Media	2315	13622	39425
3^0 Qu.	2612	14762	44743
Max.	3633	16656	53736
NA		4	11

Tabella 1.1: Principali indici di distribuzione (1^0 Qu. = primo quantile, 3^0 Qu. = secondo quantile).

2. Si sono considerati, oltre ai massimi, anche tutti quei valori vicini al massimo e maggiori di una certa soglia. Il dataset consta di 236 osservazioni, però ci sono dei valori mancanti per il Larice e la Robinia. I dati sono i seguenti:

```
      Larice       Piopponero   Robinia
1   1003.653   10001.09  5159.557
2   1031.381   10024.82  5178.543
```

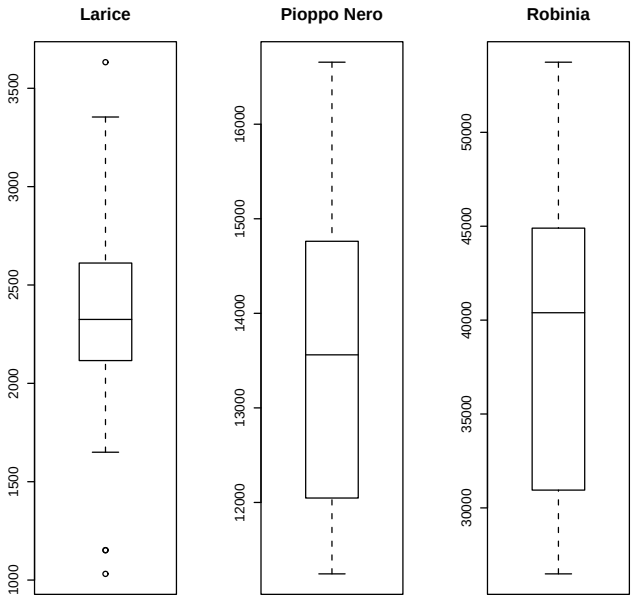

Figura 1.2: Box-plot per ogni albero.

```
3   1095.024   10029.57  5330.434
4   1151.461   10029.57  5344.674
5   1152.116   10029.57  5719.656
...
233       NA    15896.37        NA
234       NA    15943.84        NA
235       NA    16304.58        NA
236       NA    16655.83        NA
```

I diagrammi a scatola della fig. 1.3 confermano le osservazioni fatte precedentemente. Però, come conseguenza del fatto che stiamo considerando valori minori rispetto a prima, le mediane diminuiscono. Di poco per Larice e Pioppo, notevolmente per la Robinia che passa da 40394 a 19978. Aumenta notevolmente la variabilità della Robinia, diminuisce invece quel-

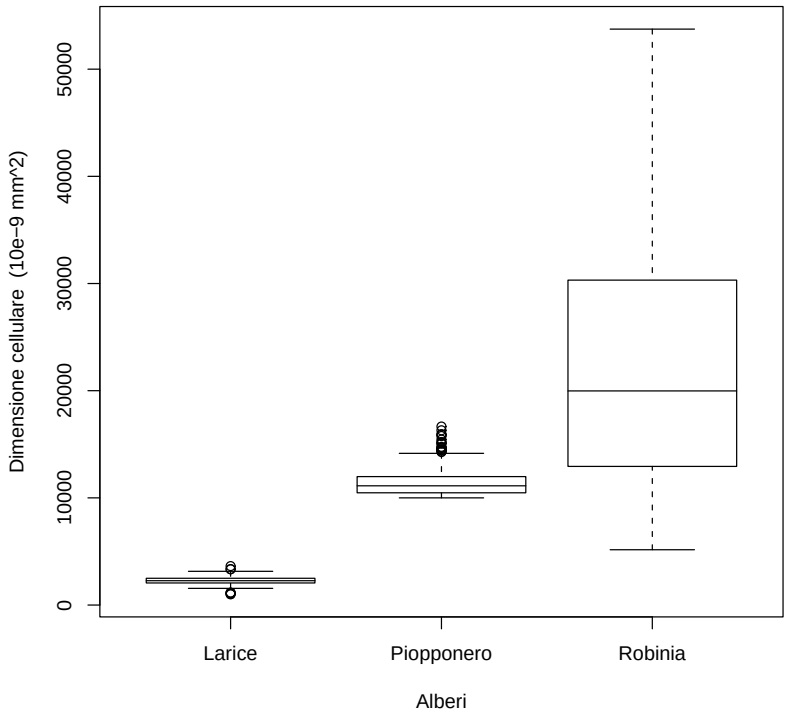

Figura 1.3: Box-plot per i dati della soglia.

la del Pioppo, che è l'unico basato su tutte le 236 osservazioni.
Dalla fig. 1.4, possiamo notare che il Pioppo nero presenta molti valori anomali, significa che abbiamo casi isolati di valori molto elevati.

Come si vede dalla tabella 1.2, i valori massimi per i tre alberi sono uguali a quelli trovati prima.

	Larice	Pioppo nero	Robinia
Min.	1004	10001	5160
1^0Qu.	2060	10475	12934
Mediana	2273	11126	19978
Media	2231	11504	21677
3^0 Qu.	2498	11980	30317
Max.	3633	16656	53736
NA	183		99

Tabella 1.2: Principali indici di distribuzione (1^0 Qu. = primo quantile, 3^0 Qu. = secondo quantile).

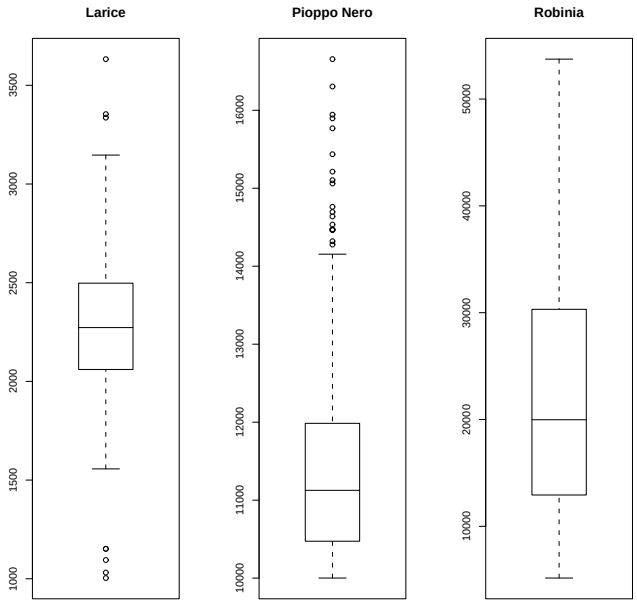

Figura 1.4: Box-plot per i singoli alberi.

Capitolo 2

Teoria dei valori estremi

2.1 Modello Classico dei valori estremi

Date n variabili casuali i.i.d. $X_1 ... X_n$ da una distribuzione sconosciuta F, il punto di partenza per l'analisi dei valori estremi è lo studio del comportamento di:

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

chiamato Ordine statistico massimo. La sua funzione di ripartizione è:

$$\begin{aligned} P\{M_n \leq x\} &= P\{X_1 \leq x, ..., X_n \leq x\} \\ &= P\{X_1 \leq x\} ... P\{X_n \leq x\} \\ &= F(x)^n. \end{aligned} \tag{2.1}$$

però non conoscendo F, tale probabilità non risulta semplice da stimare.

Una possibiltà sarebbe quella stimare F con tecniche statistiche e sostituire la stima in (2.1). Però piccole discrepanze nella stima di F ci porterebbero a sostanziali discrepanze per F^n.

L'alternariva è quella di adottare un approccio asintotico, cioè studiare i limiti per M_n, con n tendente a infinito e usare questa famiglia come un'approssimazione di M_n con n finito. Sappiamo che necessariamente, con $P = 1$, la

distribuzione di M_n per $n \to \infty$ converge con l'estremo superiore di F, quindi $F^n(x) \to 0$. A questo punto adottiamo lo stesso approccio usato per la stima della distribuzione della media campionaria, giustificato dal Teorema del limite Centrale. In quel caso, con $P = 1$ la media di X_n convergeva con μ, quindi si adottava una normalizzazione dei dati e si trovava che si distribuiva come una Normale standardizzata:

$$\frac{X_n - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

Nel caso di M_n usiamo invece quest'altra modifica, che come vedremo, si distribuirà come una $GEV(\mu, \sigma, \xi)$

$$\frac{M_n - b_n}{a_n} \sim G(\mu, \sigma, \xi),$$

Con a_n e b_n sequenze di coefficienti normalizzanti.

2.1.1 Distribuzione Generalizzata dei Valori estremi

L'intero rango di possibili distribuzioni per $M_n^* = (M_n - b_n)/a_n$ è dato dal Teorema 1.

Teorema 1, Extremal types theorem: Se esistono sequenze di costanti $a_n > 0$ e b_n, con n tendente ad infinito, tale che, la probabilità

$$P\Big\{\frac{Mn - b_n}{a_n} \leq x\Big\} \to G(x),$$

dove G è una distribuzione non degenerata, allora G segue una delle seguenti funzioni di ripartizione:

$$\text{I} : \text{G(x)} = \exp\{-\exp(-\tfrac{x-b}{a})\} \quad -\infty < x > \infty; \text{ distrib. Gumbel}$$

$$\text{II} : \text{G(x)} = \begin{cases} 0 & \text{se } x \leq b \\ \exp\{-(\tfrac{x-b}{a})^{-\alpha}\} & \text{se } x > b, \alpha > 0; \end{cases} \text{ distrb. Fréchet}$$

$$\text{III} : \text{G(x)} = \begin{cases} \exp\{-[-(\tfrac{x-b}{a})]^{\alpha}\} & \text{se } x < b, \alpha > 0 \\ 1 & \text{se } x \geq b; \end{cases} \text{ distrib. Weibull.}$$

Ogni famiglia ha un parametro di posizione b e un parametro di scala a, in più la Frechet e la Weibull hanno un ulteriore parametro α .
Le tre distribuzioni hanno una forma di comportamento distinto, in corrispondenza al differnte comportamento delle code della distribuzione F.
Se consideriamdo il comportamento di G nel punto estremo superiore x_+ notiamo che: per la dist Weibull x_+ è finito, invece per le dist Frechet e Gumble x_+ tende a infinito. Anche se, la densità di Gumbel scende esponenzialmente, quella di Fréchet invece in modo polinomiale.

In analogia con il Teorema del Limite Centale troviamo che la distribuzione limite di un campione di massimi segue una delle distribuzione elencate sopra, qualsiasi sia la distribuzione F di partenza. Nell'applicazione statistica infatti non daremo più alcuna considerazione alla distribuzione della popolazione F, ma andremmo ad applicare una G della famiglia GEV direttamente alla serie dei massimi M_n.

Distribuzione GEV Per fini statistici non è conveniente lavorare con 3 classi di distribuzioni diverse, si usa perciò utilizzare una parametrizzazione che li contenga tutti e tre. Questa distribuzione è la Distribuzione Generalizzata dei Valori estremi $G(\mu.\sigma, \xi)$, dall'inglese GEV (Generalized Extreme Value distribution), la cui funzione di ripartizione è:

$$G(x) = \exp\Big\{ -\Big[1 + \xi(\frac{x-\mu}{\sigma})\Big]^{-\frac{1}{\xi}}\Big\}, \tag{2.2}$$

definita per $1 + \xi(\frac{x-\mu}{\sigma}) > 0; -\infty < \mu > +\infty; \sigma > 0; -\infty < \xi < +\infty$.

La distribuzione ha tre parametri:

Il parametro μ è il parametro di posizione, al suo variare si sposta la distribuzione. Se μ aumenta, la distribuzione si sposta a destra, se diminuisce, si sposta a sinistra.

Il parametro σ è il parametro di scala , se aumenta la distribuzione si allarga, altrimenti si restringe.

Il parametro ξ è il parametro di forma.

- Quando $\xi < 0$ i quantili della distribuzione sono limitati, esiste quindi e si può calcolare il punto massimo della distribuzione dei massimi. Analiticamente: La G(x) é valida per $1 + \xi(\frac{x-\mu}{\sigma}) > 0$ quindi $\xi(\frac{x-\mu}{\sigma}) > -1$ Dividiamo per ξ e otteniamo $x > \mu - \sigma/\xi$, quindi

$$\tau = \mu - \frac{\sigma}{\xi},$$

è il limite superiore della distribuzione e dipende da tutti i parametri di G.

- Quando $\xi > 0$ invece la distribuzione non è limitata superiormente Dividendo l'spressione di prima per ξ otteniamo $x > \mu - \frac{\sigma}{\xi}$, x_+ quindi tende a infinito.

- Quando $\xi = 0$ la distribuzione è illimitata da entrambi i lati.

Il I tipo della classe delle distribuzioni dei valori estremi si ricava con il limite di $\xi \to 0$. Il II e III tipo invece corrispondono rispettivamente alle classi con $\xi > 0$ e $\xi < 0$. Quindi solo facendo inferenza su ξ sono gli stessi dati che determinano il tipo di comportamento delle code più appropriato e non è necessario fare ipotesi a priori su famiglia di distribuzione da adottare.

In sostanza abbiamo trovato che, per valori grandi di n,

$$P\Big(\frac{M_n - b_n}{a_n} < x\Big) \approx G(x) \qquad \text{con } n \to \infty.$$

L'unica cosa che ancora non conosciamo sono le costanti a_n e b_n, ma questo problema viene presto risolto,

$$P(M_n < x) \approx G\Big(\frac{x - b_n}{a_n}\Big) = G^*(x),$$

dove $G^*(x)$ è un'altro membro della famiglia GEV e poichè in ogni caso i parametri di GEV vengono stimati di volta in volta, è irrilevante che i parametri di G siano diversi da quelli di G^*.

Ricapitolando, abbiamo scoperto che il massimo di una serie di valori indipendenti segue una distribuzione GEV. Per stimare i parametri di GEV però abbiamo bisogno di una serie, $M_{n,1}, ...M_{n,m}$, di massimi di m gruppi diversi, ognuno di n osservazioni.

Spesso, poichè la teoria viene applicata a serie di dati misurati nel tempo, per gruppo si intende un periodo di tempo di lunghezza un anno, per cui$n$ diventa il numero di osservazioni in un anno, e la serie dei massimi diventa una serie di massimi annuali. In questo caso, però stiamo trattando con dati di alberi misurati in una superficie. Qui non si può parlare di anni, ma di campioni. Lo scopo dello studio è quello di trovare la dimensione cellulare massima possibile rispetto all'intera superficie del tronco. Poichè era improponibile poter misurare l'intera superficie, si sono selezionati aleatoriamente solo alcuni campioni per ciascuna specie di albero e sulla base di essi si vuole calcolare il massimo possibile per l'intera superfice; n diventa quindi il numero di osservazioni per campione e la serie dei massimi diventa una serie di massimi campionari.

2.2 Valori di ritorno

Dalla G(x) otteniamo l'espressione dei quantili della distribuzione dei massimi campionari:

$$x_p = \begin{cases} \mu - \frac{\sigma}{\xi}[1 - \{-\log(1-p)\}^{-\xi}] & \text{con } \xi \neq 0 \\ \mu - \sigma[\log\{-\log(1-p)\}] & \text{con } \xi = 0, \end{cases} \tag{2.3}$$

dove $G(x_p) = 1 - p$.

Sempre nel contesto di dati temporali, il valore x_p indica quel quantile che ha una probabilità p di essere superato nel corso degli anni. In altre parole, il livello x_p è quel valore che ci si aspetta venga superato in media una volta ogni 1/p

anni. Con la terminologia specifica si dice che x_p è il valore di ritorno associato al periodo di ritorno $1/p$.

Per capire meglio questa relazione ci serviamo di un grafico, il Return Level Plot, che mette in ralazione 1/p con x_p. Il grafico indica in media con quale probabilità ogni livello x_p puo essere superato. La curva risulta lineare se $\xi = 0$, convesso con limite asintotico che tende a τ, se $\xi < 0$ e concavo, senza limite superiore, se $\xi > 0$.

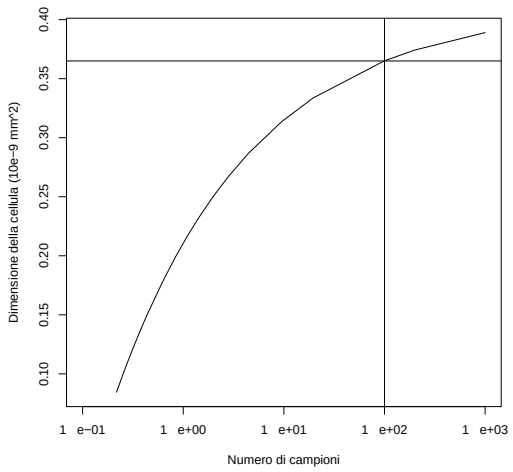

Figura 2.1: Grafico del livello di ritorno per il Larice.

Nel nostro caso il grafico può essere così interpretato: il livello x_p è quel valore che verrà superato, in media, una volta ogni 1/p campioni della superficie dell'albero. L'asse delle ascisse rappresenta il numero di campioni e l'asse delle ordinate, x_p, la dimensione cellulare. Ad esempio, osservando il grafico in fig. 2.1 possiamo affermare che:

- in media, 1 cellula in ogni 100 campioni avrà diametro maggiore di 0.365 micrometri quadrati.

- ciascuna cellula ha una probabilità $\frac{1}{100}$ di avere un diametro maggiore di 0.365.

Dal grafico possiamo ricavare anche il valore massimo possibile dell'intera

superficie della pianta. Questo valore lo troviamo nel punto in cui la curva comincia a stabilizzarsi, cioè dove il periodo di ritorno $1/p$ è molto alto e la probabilità p è molto bassa.

2.3 Inferenza sulla distrubuzione GEV

Assumendo che n sia sufficientemente grande da applicare il modello GEV ai dati, adesso il problema è fare inferenza sui parametri (μ, σ, ξ). Tra le varie tecniche possibili, la tecnica della Verosimiglianza risulta preferibile a tutte, anche se presenta i suoi svantaggi. Una dificoltà concerne le proprietà asintotiche a cui questo metodo deve sottostare. Queste condizioni non sono soddisfatte dal modello GEV perchè i punti finali delle distribuzione sono funzioni dei parametri ($\mu - \sigma/\xi$ per $\xi<0$). Questa violazione delle condizioni regolari sta a significare che i risultati standard di questo metodo non sono applicabili automaticamente. Smith (1985) studiò questo problema in dettaglio e ottenne i seguenti risultati:

- quando $\xi > -0.5$ gli stimatori di massima verosimiglianza esistono, sono regolari e godono delle usuali proprietà;
- quando -1< $\xi < -0.5$ gli stimatori esistono ma non sono regolari;
- quando $\xi < -1$ gli stimatori non esistono.

Il caso $\xi < -0.5$ corrisponde alla distribuzione con una piccola limitata coda superiore. Questa situazione si incontra raramente in applicazioni pratiche, quindi le limitazioni teoriche dell'approccio della massima verosimiglianza non ostacolano la pratica.

Calcoliamo la log-verosimiglianza, $\ell(\mu, \sigma, \xi) = log \prod g(x)$, dove $g(x)$ è la densità di distribuziobe di GEV,

$$g(x) = \frac{dG(x)}{dx} = \frac{1}{\sigma}\Big[1 + \xi(\frac{x-\mu}{\sigma})\Big]^{-\frac{1}{\xi}-1} \exp\Big[1 + \xi(\frac{x-\mu}{\sigma})\Big]^{-\frac{1}{\xi}}.$$

La log-verosimiglianza è:

$$\ell(\mu,\sigma,\xi)=\sum_{i=1}^{k}\Big\{-\log\sigma-(1+\xi)\log\Big[1+\xi\Big(\frac{x-\mu}{\sigma}\Big)\Big]-\Big[1+\xi\Big(\frac{x-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}\Big\}, \quad (2.4)$$

con $1+\xi(\frac{x-\mu}{\sigma})>0$ per $i=1,...,n$. La massimizzazione di questa funzione porta alle stime di massima verosimiglianza. Non c'è una soluzione analitica, ma per ogni dataset, la massimizzazione è facile usando algoritmi numerici di ottimizzazione.

Il vettore delle stime di μ,σ e ξ si distribuisce approssimativamente come una normale multivariata con media $(\hat{\mu},\hat{\sigma},\hat{\xi})$ e matrice di varianza pari all' inversa dell'informazione osservata. Gli intervalli di confidenza sono dunque facili da trovare, ad esempio per μ sarebbe $\hat{\mu}\pm z_{\alpha/2}*\sqrt{var}$.

Maggiore accuratezza dell'intervallo di confidenza può essere ottenuto dal profilo di verosimiglianza. Per ottenere il profilo di ξ, fissiamo un valore $\xi=\xi_0$ e massimizziamo la log-verosimiglianza $\ell(\mu,\sigma,\xi_0)$ rispetto agli altri due parametri rimanenti. Questo viene ripetuto per un range di valori ξ_0. L' insieme di massimi di log-verosimiglianza così ottenuti costituiranno il profilo. Questo metodo può essere applicato anche quando si vuole fare inferenza su combinazioni di parametri. Possiamo, ad esempio ottenere l'intervallo di confidenza per x_p. Dobbiamo riparametrizzare il modello GEV in modo che x_p sia un parametro, $GEV(x_p,\sigma,\xi)$ e poi, come prima, possiamo massimizzare la log-verosimiglianaza rispetto ai parametri rimanenti.

La stima di massima verosimiglianza per x_p, con $0<p<1$ a livello di ritorno $1/p$ è:

$$x_p=\begin{cases}\hat{\mu}-\frac{\hat{\sigma}}{\hat{\xi}}[1-y_p^{-\hat{\xi}}] & \text{con } \xi\neq 0\\ \hat{\mu}-\hat{\sigma}[\log y_p] & \text{con } \xi=0,\end{cases} \quad (2.5)$$

dove $y_p=-\log(1-p)$.

I periodi di ritorno più interessanti sono caratterizzati da probabilità p piccola. Per $p=0$ e cioè per periodo di ritorno infinito, possiamo fare inferenza sul punto massimo della distribuzione chiamata appunto "l'infinitesima osservazione del periodo di ritorno". Per $\xi<0$ la stima di questo punto è $x_0=\mu-\sigma/\xi$.

2.4 Model check

Definita la distribuzione e stimatone i suoi parametri non ci resta che controllare l'adattabilità dei nostri dati al modello. Per questo ci si serve di seguenti grafici.

- **Probability plot:**
  Compara la funzione di distribuzione empirica con quella ricavata dai dati. Partendo dai dati $x_1 < x_2 < ... < x_m$ ordinati in modo crescente
  - la funzione di distribuzione empirica è : $\widetilde{G}(x_i) = \frac{i}{m+1}$,
  - la funzione basata sui dati è: $\widehat{G}(x_i) = \exp\Big\{ - \Big[1 + \hat{\xi}(\frac{x_i - \hat{\mu}}{\hat{\sigma}})\Big]^{-1/\hat{\xi}}\Big\}$.

  L'insieme dei punti del grafico è:

  $$\{\widetilde{G}(x_i), \widehat{G}(x_i), i = 1, ..., m\}.$$

  Sostanziali scostamenti del grafico dalla linearità indicano una mala approssimazione del modello ai dati. $\widetilde{G}(x_i)$ e $\widehat{G}(x_i)$ sono limitati a 1 al crescere di x_i.

- **Quantile plot:**
  Compara i quantili empirici: $\widetilde{G}^{-1}(\frac{i}{m+1}) = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}}\Big[1 - \Big\{ - log(\frac{i}{m+1})\Big\}^{-\hat{\xi}}\Big]$ con quelli stimati dai dati x_i.

  L'insieme dei punti del grafico è:

  $$\Big\{\widetilde{G}^{-1}\Big(\frac{i}{m+1}\Big), x_i, i = 1...m\Big\}.$$

- **Return level plot:**
  Un grafico dei quantili x_i contro y_p su scala logaritmica $y_p = -\log(1-p)$

  $$\{\log y_p, x_p\} \qquad \text{con } 0 < p < 1.$$

  Il grafico è lineare nel caso di $\xi = 0$. É convesso con limite asintotico, per $p \to 0$, uguale a $\mu - \sigma/\xi$, se $\xi < 0$, è concavo e non ha limiti finiti se $\xi > 0$.

Grazie alla sua semplicità di interpretazione, e grazie alla scelta della scala, che comprime la coda della distribuzione cosi da evidenziare l'effetto della estrapolazione, return level plot è particolarmente conveniente sia per la presentazione del modello che per la verifica di esso.

- **Istogramma:**
  Compara la funzione di densita dei valori massimi con l'istogramma dei dati. Questo grafico è molto meno informativo degli altri, anche perchè l'istogramma può cambiare molto con la scelta degli intervalli.

2.5 Applicazione di GEV ai dati

I dati dei massimi, per le tre specie di alberi, sono contenuti in questi tre insiemi:

- $M_L = \{x_1, ..., x_{25}\}$: valori massimi misurati su **25** campioni di **Larice**.
- $M_P = \{x_1, ..., x_{21}\}$, massimi misurati su **21** campioni di **Pioppo Nero**.
- $M_R = \{x_1, ..., x_{14}\}$; massimi misurati su **14** campioni di **Robinia**.

Per ogni insieme andiamo a stimare il modello GEV e i suoi parametri e individuiamo il massimo valore possibile. Per una questione computazionale, poichè i dati sono di ordine elevato, dividiamo i dati originali per 10000.

Larice: Per applicare il modello ai dati ci serviamo di un software staristico. Il programma preso in considerazione è R, scaricabile gratuitamente da http://$www.r-project.org$. Le funzioni utilizzate sono state realizzate da Alec Stephenson, statistico dell'Università di Lancaster, e sono raccolte nella libreria *ismev*. Applicando la funzione $gev.fit$, otteniamo le seguenti stime, (approssimiamo alla terza cifra decimale):

$$(\mu, \sigma, \xi) = (0.211, 0.064, -0.320),$$

con valore massimo della log-verosimiglianza pari a -33.929. La corrispondente matrice di varianza-covarianza stimata è:

$$V = \begin{pmatrix} 1.969e-04 & -8.404e-06 & -0.0006 \\ -8.400e-06 & 9.375e-05 & -0.0006 \\ -6.250e-04 & -6.568e-04 & 0.0141 \end{pmatrix}$$

La radice quadrata dei valori nella diagonale di questa matrice dà lo standard error di μ ,σ e ξ:

$$S.E. = (0.014, 0.009, 0.119).$$

Sappiamo che il vettore dei parametri si distribuisce approssimativamente come una normale multivariata, possiamo quindi calcolare l'intervallo di confidenza al 95%, con $z_{\alpha/2} = 1.96$,

$$\mu \to 0.211 \pm 1.96 \cdot 0.014 = (0.183, 2.955)$$

$$\sigma \rightarrow 0.064 \pm 1.96 \cdot 0.009 = (0.046, 0.081)$$

$$\xi \rightarrow -0.320 \pm 1.96 \cdot 0.119 = (-0.543, -0.086).$$

Ricordiamo che il metodo di massima verosimiglianza è valido solo quando la stima del parametro di forma è maggiore di -0.5; in questo caso non solo$\xi = -0.32$ ma pure l'intervallo di confidenza è spostato tutto verso valori negativi e maggiori di -0.5, per cui siamo certi che le stime sono corrette.

L'accuratezza dell'intervallo di confidenza può essere testata anche grazie al profilo della log-verosimiglianza: la fig. 2.2, che mostra il profilo perξ, evidenzia un intervallo compreso tra (-0.565, -0.05), intervallo molto simile a quello stimato prima.

Certi che ξ sia negativo, siamo certi anche che la distribuzione è limitata superiormente. Il punto superiore della distribuzione è:

$$x_0 = \mu - \sigma/\xi = 0.411,$$

che è il quantile x_p calcolato ponendo $p = 0$. La sua varianza, calcolata con il metodo delta, risulta uguale a 0.0032, quindi il suo intervallo di confidenza è [0.299 , 0.522].

Tornando alla scala originale, la dimensione massima di una cellula di Larice misura al massimo 4110 micrometri quadrati. Il suo intervallo di confidenza al 95% oscilla tra [2990 , 5220].

Per stimare il 100-campione return level, poniamo p=1/100 e sostituendo in (2.6), troviamo che $\hat{x}_{0.01} = 0.365$ e la $Var(\hat{x}_{0.01}) = 0.0005$. Quindi il suo intervallo di confidenza al 95% è $0.365 \pm 1.96 \cdot \sqrt{0.0005} = [0.319, 0.4108]$.

Maggiore esatezza è data dal profilo di log-verosimiglianza. Le figure 2.3 e 2.4 mostrano il profilo di x_p con p = 0.01 e p = 0, rispettivamente. Da esse ricaviamo i relativi intervalli di confidenza, [0.337 , 0.475] e [0.35 , 1.5]. Mentre il primo intervallo è abbastanza simile a quello ottenuto con il metodo delta, il secondo no. Questa discrepanza è dovuta all'assimetria del profilo, che si amplifica con l'aumentare del return period. Tale discrepanza è abbastan-

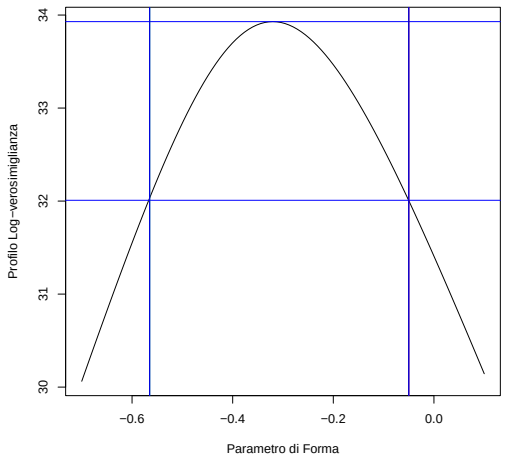

Figura 2.2: Profilo di verosimiglianza per ξ per i dati sul Larice

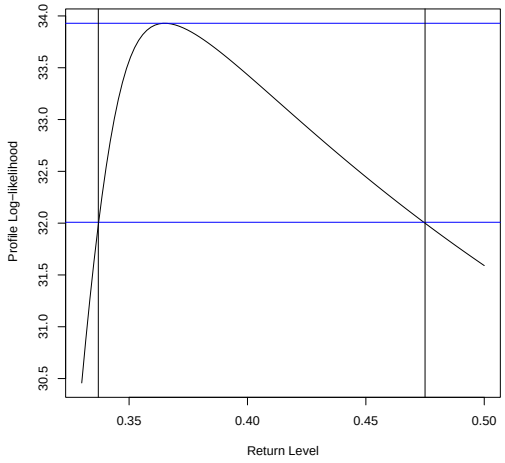

Figura 2.3: Profilo di verosimiglianza per 100-campione return level per i dati sul Larice

za concepibile, visto che i dati danno deboli informazioni verso valori alti del processo.

I grafici diagnostici del modello confermano la bontà del modello. Anche

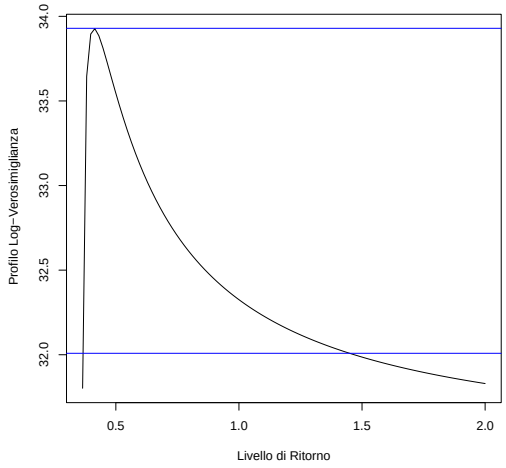

Figura 2.4: Profilo di verosimiglianza per x_0 per i dati sul Larice

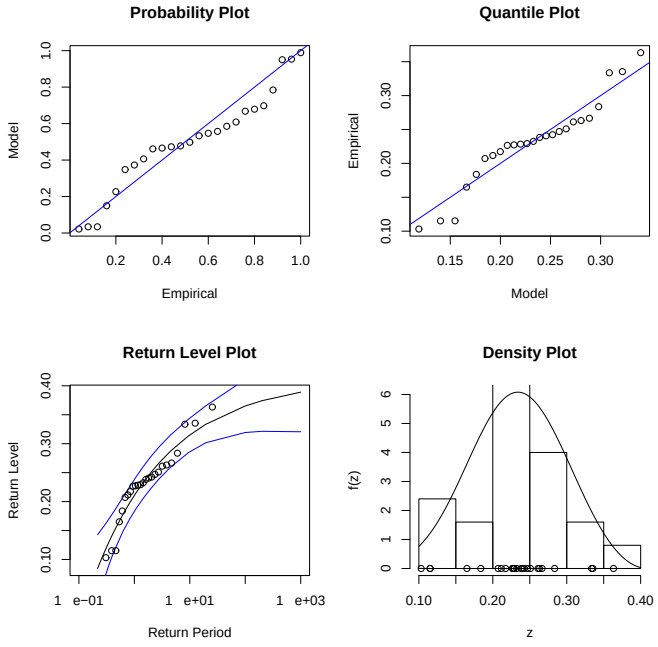

Figura 2.5: Grafici diagnostici per GEV stimato sui dati di Larice.

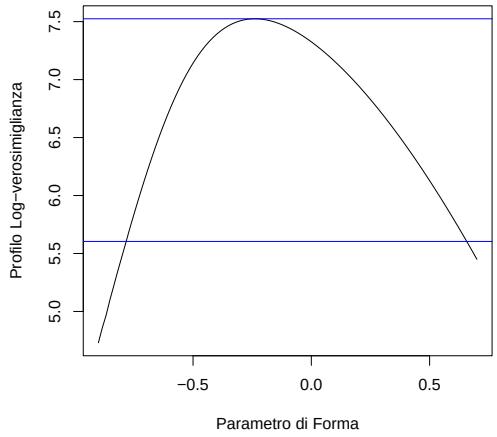

Figura 2.6: Profilo di verosimiglianza per ξ per i dati sul Pioppo nero

se i grafici di probabilità e dei quantili non sono molto lineari, il terzo grafico dei valori di ritorno è abbastanza buono, perché tutti i punti stanno dentro l'intervallo di confidenza. Dall'istogramma notiamo che la curva segue bene l'andamento dei dati.

Pioppo nero: Anche in questo caso, procediamo massimizziando la log-verosimiglianza per ottenere le stime dei parametrei del modello GEV. Per il valore massimo della log-verosimiglianza pari a -7.524 otteniamo le seguenti stime:

$$(\mu, \sigma, \xi) = (1.297, 0.162, -0.239).$$

La matrice di varianza-covarianza è:

$$V = \begin{pmatrix} 0.0022 & 0.0007 & -0.0100 \\ 0.0007 & 0.0015 & -0.0104 \\ -0.0100 & -0.0104 & 0.1147 \end{pmatrix}$$

dalla quale otteniamo lo standard error rispettivamente di μ, σ e ξ,

$$S.E. = (0.046, 0.039, 0.338).$$

La stima del parametro di forma è negativa e maggiore di -0.5, però, diversamente dal caso precedente, come è evidente dalla fig. 2.6, l'intervallo di confidenza per ξ comprende valori positivi, per cui potremmo dubitare sulla sua negatività. Se andiamo ad analizzare il profilo di verosimiglinza (fig. 2.7) di x_0:

$$x_0 = \mu - \sigma/\xi = 1.789,$$

notiamo che il suo intervallo di confidenza non è superiormente limitato, per cui, secodo il modello, il valore x_0 tenderebbe a infinito. Nella realtà questo non è possibile.

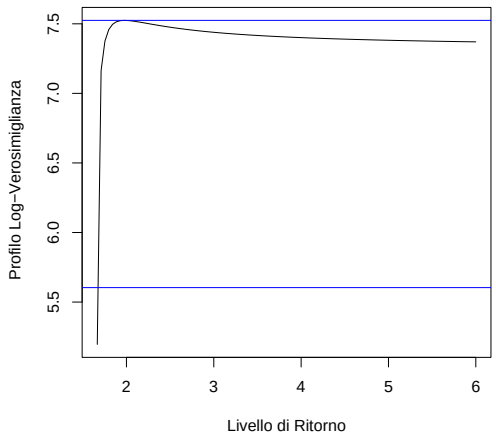

Figura 2.7: Profilo di verosimiglianza per x_0 per i dati sul Pioppo nero

Il modello comunque si adatta bene ai dati: dalla fig. 2.8 notiamo che i punti del grafico dei quantili seguono un andamento abbastanza lineare e quelli del return level plot non escono dall'intervallo di confidenza. Come consequenza del fatto che l'intervallo di ξ comprendeva anche valori positivi, la curva risulta abbastanza lineare, però pur sempre tendente a un valore finito, e ciò ci conferma che ξ può essere considerato negativo.

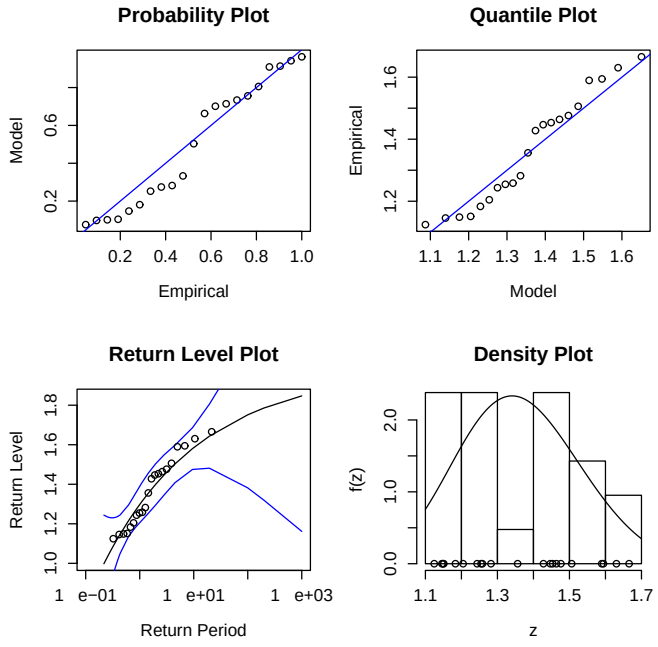

Figura 2.8: Grafici diagnostici per GEV stimato sui dati di Pioppo Nero.

Robinia: Per il valore massimo della log-verosimiglianza pari a 16.644 otteniamo le seguenti stime:

$$(\mu, \sigma, \xi) = (3.699, 0.839, -0.392).$$

La matrice di varianza-covarianza è:

$$V = \begin{pmatrix} 0.0644 & -0.0009 & -0.025 \\ -0.0009 & 0.0379, & -0.0304, \\ -0.0258 & -0.0303 & 0.0519 \end{pmatrix}$$

da questa otteniamo lo standard error rispettivamente di μ, σ e ξ,

$$S.E. = (0.254, 0.195, 0.228).$$

La stima del parametro di forma è negativa e maggiore di -0.5, e dalla fig. 2.9 notiamo che l'intervallo di confidenza è [-0.88 , 0.26] e comprende alcuni valori positivi. Il valore limite superiore della distribuzione risulta uguale a:

$$x_0 = \mu - \sigma/\xi = 5.841.$$

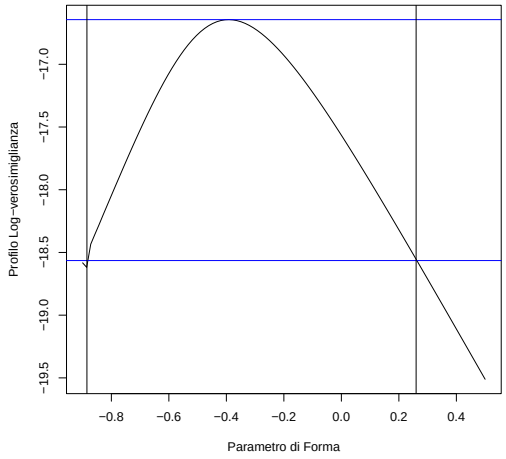

Figura 2.9: Profilo di verosimiglianza per ξ per i dati sulla Robinia

Dalla fig. 2.10 vediamo che anche in questo caso, come conseguenza del fatto che l'intervallo di ξ è spostato su valori positivi, il profilo non ha un limite superiore definito. I grafici diagnostici (fig. 2.11) però, confermano la buona adattabilità dei dati al grafico.

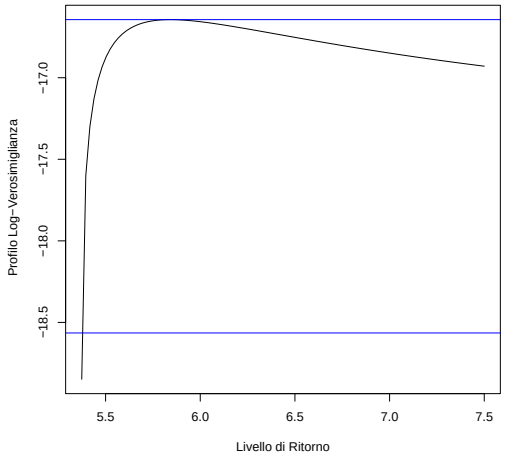

Figura 2.10: Profilo di verosimiglianza per x_0 per i dati sulla Robinia

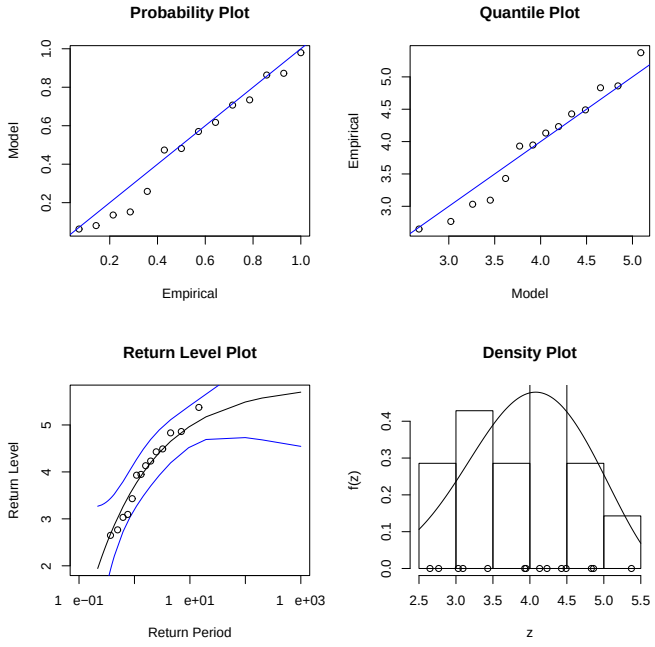

Figura 2.11: Grafici diagnostici per GEV su dati sulla Robinia

2.6 Confronto tra i 3 alberi

	Larice	Pioppo nero	Robinia
μ	0.211	1.297	3.699
σ	0.064	0.162	0.839
ξ	-0.320	-0.239	-0.392
x_0	0.411	1.789	5.841

Tabella 2.1: Confronto tra le stime dei parametri dei 3 alberi. (μ è la media, σ è il parametro di scala, ξ il parametro di forma e x_0 è il massimo possibile).

In conclusione, il modello GEV si adatta abbastanza bene a tutti e tre i tipi di alberi. Non è molto realistico il risultato sull'intervallox_0, è impossibile infatti che la dimensione cellulare possa tendere a infinito. Le stime dei massimi delle tre popolazioni risultno sostanzialmente diverse. La Robinia, come era emerso dall'analisi grafica, è quella con media e varianza maggiori. I parametriξ sono invece molto simili, in particolare quello del Larice e della Robinia.

Capitolo 3

Considerazioni

3.1 Come si comporta la dimensione dei massimi dell'intera superficie?

Finora abbiamo svolto l'analisi su un numero limitato di campioni della superficie dell'albero. Ora vogliamo estendere i risultati ottenuti a tutta la superficie del tronco. Non conosciamo però il numero totale di campioni presenti nella sezione di albero.

Vedremo, attraverso alcune semplici considerazioni, che, anche se non abbiamo nessuna informazione sui restanti campioni, è possibile ottenere una stima del comportamento dei massimi di tutta la superficie del tronco. Ad esempio, se dalla sezione dei nostri alberi consideriamo 4 campioni, e per ogni campione troviamo il massimo, otteniamo 4 osservazioni massime.

$$m = 4 \qquad M(4) = \{M_1, M_2, M_3, M_4\},$$

dove $M_i = \max\{x_{1i}, ..., x_{ni}\}$ per $i = 1, ..., m$ si distribuisce come una $GEV(\mu, \sigma, \xi)$. Consideriamo ora 100 campioni,

$$m = 100 \qquad M(100) = \{M_1, ..., M_{100}\},$$

la $M(4)$ e $M(100)$ si distribuiranno entrambe come una GEV, ma con parametri leggermente distinti. Ciò significa che al variare di m, numero di campioni,

cambiano anche i rispettivi parametri, e di conseguenza cambia di volta in volta il loro comportamento. Quello che non cambia però è il punto massimo in cui termina la coda superirore della distribuzione (x_0).

Vediamo come si comporta la distribuzione $M(m)$:

$$M(m) = \{M_1, ..., M_m\}, \qquad \text{con } M_i = max\{x_{1i}, ..., x_{ni}\},$$

con $i = 1...m$, $m =$ numero campioni e $n =$numerosità campione. Partiamo dal fatto che il massimo di un'insieme di n valori si distribuisce come una GEV,

$$M(1) \sim GEV(\mu, \sigma, \xi).$$

Andiamo a studiare come varia il comportamento di M(m) rispetto a M(1). La funzione di ripartizione di $M(m)$ è

$$\begin{aligned} P(M(m) < x) &= P(M_1 < x, ..., M_m < x) \\ &= P(M_1 < x)...P(M_m < x) \\ &= G(x)^m = \exp\Big\{ -\Big[1 + \xi(\frac{x-\mu}{\sigma})\Big]^{-\frac{1}{\xi}}\Big\}^m. \end{aligned}$$

Con alcuni passaggi logici ottengo:

$$G(x)^m = \exp\Big\{ -\Big[1 + \xi\Big(\frac{x-\mu^*}{\sigma^*}\Big)\Big]^{-\frac{1}{\xi}}\Big\}, \tag{3.1}$$

$$M(m) \sim GEV(\mu^*, \sigma^*, \xi),$$

$$\begin{aligned} \text{con} \qquad \mu^* &= \mu - \frac{\sigma}{\xi}(1 + m^\xi) \\ \sigma^* &= \sigma m^\xi. \end{aligned}$$

I parametri di $M(m)$ dipendono dai parametri di $G(\mu, \sigma, \xi)$. La stima di x_0^* è

$$x_0^* = \mu^* - \frac{\sigma^*}{\xi^*}.$$

Possiamo velocemente verificare che: $\mu^* - \frac{\sigma^*}{\xi^*} = \mu - \frac{\sigma}{\xi}(1 + m^\xi) - \frac{\sigma m^\xi}{\xi} = \mu - \frac{\sigma}{\xi}$ quindi,

$$x_0 = x_0^*.$$

Il punto di massimo non varia al variare di m, ciò significa che μ^* e σ^* sono due parametri che si compensano. Dalla $M(m)$ possiamo ora ricavare i quantili. Sostituiamo μ^* e σ^* in (2.6) e tracciamo il grafico dei valori di ritorno.

Analizziamo il caso del Larice: partendo dalle stime di $\hat{\mu}$, $\hat{\sigma}$ e $\hat{\xi}$, sulla base di soli 25 campioni, possiamo, stimando m (il totale di campioni del fusto), ricavare il comportamento dei massimi di tutta la superficie del tronco. Per stimare m adottiamo due approcci:

Primo approccio: m, numero fisso discreto Ipotizziamo m=10,50,100,1000. Ci fermiamo a 1000 perche come si vede dal grafico la curva resta abbastanza costante.

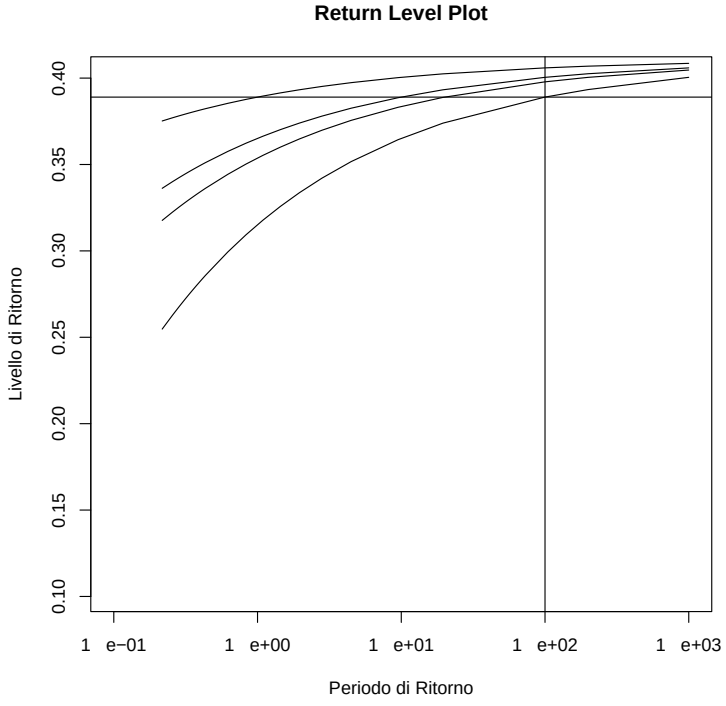

Figura 3.1: Return level plot al variare di m =10,50,100,1000

Dal grafico in fig. 3.1 appare chiaro che al variare di m il comportamento della serie dei massimi rimane simile. A destra del grafico, per probabilità piccole, tutte le rette tendono allo stesso valore massimo. Più aumenta m e più la curva si sposta verso l'alto, cioè cresce la dimensione massima ammessa dalla cellula. Questo si adatta alla realtà, infatti più sono i campioni studiati, più di conseguenza saranno i massimi e maggiore dunque sarà la probabilità di incontrare valori elevati.

Secondo approccio: m, variabile Poisson Pensiamo m come una Poisson di media k

$$m \sim Poisson(k) \qquad \text{con funzione di densità } p(m) = e^{-k}\frac{k^m}{m!},$$

dove k è la media dei campioni, che vengono considerati, per superfice di tronco. Poichè in questo caso m non è un numero noto, indicheremo la serie dei massimi non più con $M(m)$, ma semplicemente con M. La funzione di ripartizione di M quindi diventa:

$$\begin{aligned} P(M<x) &= \sum_{m=0}^{\infty} P(M<x|m)\cdot P(m) \\ &= \sum_{m=0}^{\infty} \exp\Big\{-[1+\xi(\frac{x-\mu^*}{\sigma^*})]^{-\frac{1}{\xi}}\Big\}^m \cdot e^{-k}\frac{k^m}{m!} \\ &= e^{-k}\sum_{m=0}^{\infty}\frac{\Big[k\cdot exp\{-[1+\xi(\frac{x-\mu^*}{\sigma^*})]^{-\frac{1}{\xi}}\}\Big]^m}{m!}, \end{aligned} \tag{3.2}$$

poichè $\sum_{n=0}^{\infty}\frac{x^n}{n!} = e^x$, la funzione di ritartizione è,

$$F(M) = \exp\Big\{\kappa\cdot\exp\Big\{-[1+\xi(\frac{x-\mu}{\sigma})]^{-\frac{1}{\xi}}\Big\}-1\Big\}. \tag{3.3}$$

Ponendo $F(M) = 1-p$, troviamo i quantili e costruiamo il grafico del livello di ritoro,

$$x_p = \frac{\sigma}{\mu}(-1+log(y_p)^{-\xi})+\mu,$$

dove $y_p = \frac{1}{k}\log(1-p)+1$. Vediamo in effetti che pensando k come Poisson otteniamo un grafico simile a quello di prima. La fig. 3.2 può essere interpre-

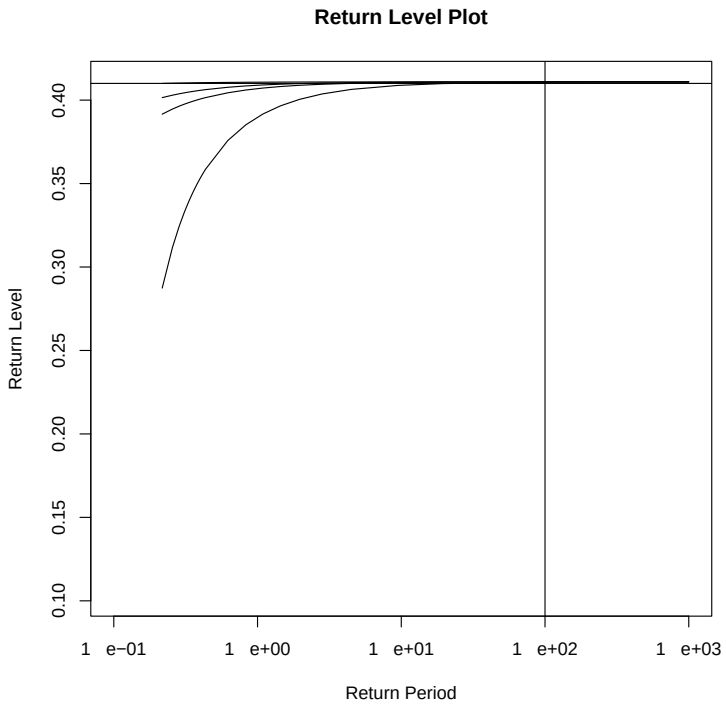

Figura 3.2: Return level plot al variare di k, media Poisson =10,50,100,1000

tata esattamente allo stesso modo del grafico precedente. All'aumentare di k, cioè all'aumentare della media di campioni analizzati per sezione, le curve si avvicinano quasi a coincidere. Si hanno differenze solo per probabilità basse, ma in fondo queste non sono oggetto del nostro studio. In fig 3.3 confrontando le curve ricavate dai due approcci, notiamo che la curva con m Poisson dà sempre stime più alte, però per probabilità piccole le due curve vanno sempre a coincidere.

Otteniamo gli stessi risultati anche facendo le stesse considerazioni ai dati sugli altri due alberi, Pioppo e Robinia. Quindi a maggior ragione possiamo affermare che al variare di m, il comportamento di $M(m)$ rimane costante e le stime trovate nel capitolo precedente possono essere generalizzate a tutta la superficie dell'albero.

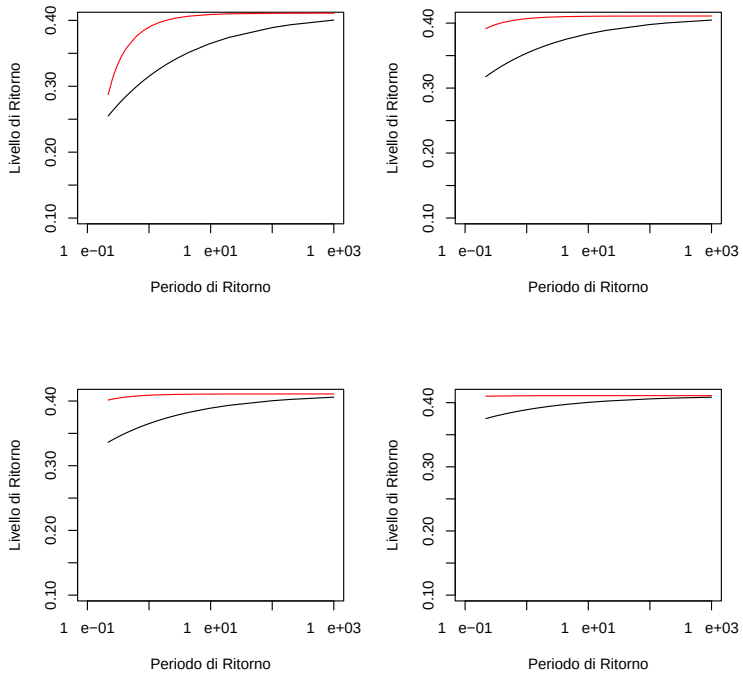

Figura 3.3: Confronto tra m come numero fisso (linea continua) e m come Poisson (linea tratteggiata)

3.2 C'e differenza tra i tre alberi?

Vogliamo vedere se il comportamento della serie dei massimi dei tre alberi è simile o meno. Per poter fare questo confronto utilizziamo due modelli annidati. Vediamo in dettaglio questi due metodi di analisi.

Modello A: Nel primo caso, come abbiamo fatto nel capitolo 2, andiamo ad anlizzarre separatamente le tre popolazioni : Larice, Pioppo nero e Robinia. Tutte e tre si distribuiscono come una GEV, ognuna con i suoi parametri, abbiamo quindi 9 stime differenti:

- Larice $\sim GEV(\mu_1, \sigma_1, \xi_1)$: log-ver. = -33.929,
- Pioppo Nero $\sim GEV(\mu_2, \sigma_2, \xi_2)$: log-ver. = -7.524,

- Robinia $\sim GEV(\mu_3, \sigma_3, \xi_3)$: log-ver. $= 16.643$.

Ogni popolazione avrà la sua log-verosimiglianza e le sue stime. I valori delle stime sono riportate in tabella 2.1. La somme delle 3 log-verosimiglianze darà il valore della log-verosimiglianza di A.

$$\text{Log-ver. del Modello A} = -33.929 - 7.524 + 16.643 = -24.81.$$

Modello B: Supponiamo che il comportamento delle tre popolazioni non sia così differente, in fondo si tratta sempre di alberi simili. Ipotizziamo che il parametro di forma, cioè quello che stabilisce la forma del fenomeno in studio, sia uguale tra le tre. In questo caso avremmo 7 parametri. Il modello B è dunque contenuto in A. Andiamo a verificare l'ipotesi,

$$H_0 : \xi_1 = \xi_2 = \xi_3.$$

- Larice $\sim GEV(\mu_1, \sigma_1, \xi)$
- Pioppo Nero $\sim GEV(\mu_2, \sigma_2, \xi)$
- Robinia $\sim GEV(\mu_3, \sigma_3, \xi)$

$$\text{Log-ver. del Modello B} = -24.73.$$

Possiamo applicare il test della log-verosimiglianza per verificare la validità di B:

$$Test = 2(\text{Log-ver. B} - \text{Log-ver. A}) \sim \chi^2_2.$$

Il test sotto l'ipotesi nulla si distribuisce come una Chi-quadrato con 2 gradi di libertà.

$$T = 2[-24.73 - (-24.81)] = 0.153 \sim \chi^2_2,$$

l'α-osservato è 0.926, quindi, per $\alpha = 0.05$, accettiamo ampiamente l'ipotesi nulla. Il comportamento delle dimensini delle cellule può dunque essere considerato simile tra specie di alberi diverse. I grafici in fig. 3.4 confermano tale conclusione perchè mostrano una buonissima affidabilità dei dati al modello.

Questo risultato è di grandissima importanza nell'ambito dello studio delle piante. Il parametro ξ è il parametro di forma, e come dice il nome, è quello

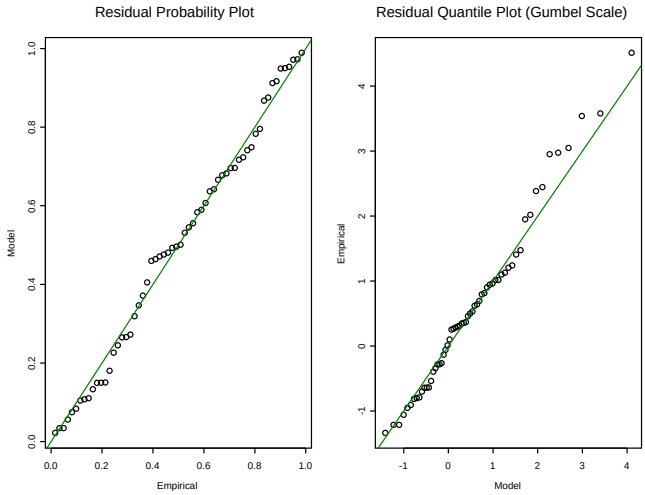

Figura 3.4: Grafici diagnostici per GEV sui dati dei tre tipi di alberi.

che determina la forma delle code della distribuzione. Quindi, affermare cheξ è uguale per tutte e tre le specie di alberi significa che le code delle distribuzioni sono molto simili, per cui si può concludere che il comportamento dei valori estremi della dimensione cellulare è simile in tutti gli alberi, indipendentemente dalla specie.

Capitolo 4

Modello della soglia

Il Modello Classico dei Valori Estremi è un approccio dispendioso dell'analisi dei valori estremi in quanto considera solo i dati nelle estremità, anche se altri dati sono accessibili. Nel nostro caso risulta più naturale considerare come eventi estremi non solo i massimi ma tutti quei valori X che eccedono una determinata soglia u.

Siano $x_1 ... x_n$ realizzazioni i.i.d. di una ignota distribuzione F. Consideriamo quindi quegli x_i per cui

$$x_i > u \qquad \text{con } u\text{= soglia,}$$

che vengono rinominati $x_{(1)}, ... x_{(k)}$. L'analisi si basa su,

$$y_j = x_{(j)} - u, \qquad \text{per } j = 1, ..., k.$$

Indicando con x, un termine arbitrario della sequenza degli $x_{(i)}$, segue che una descrizione del comportamento stocastico di questi punti estremi è data dalla probabilità condizionata,

$$P(x > u + y | x > u) = \frac{1 - F(u + y)}{1 - F(u)} \qquad \text{con } y > 0.$$

4.1 Distribuzione generalizzata di Pareto

Non conoscendo F, a partire dal Teorema 2, troviamo un'approssimazione utile ripercorrendo la logica utilizzata per la distribuzione GEV.

Teorema 2: Data una sequenza di variabili indipendenti $X_1 ... X_n$ con distribuzione comune F, se vale il Teorema 1, secondo cui:

$$P(M_n \leq z) \approx G(z) \qquad \text{dove } G(x) = \exp\Big\{ -\Big[1 + \xi\Big(\frac{x-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}\Big\},$$

allora la distribuzione di Y, definita come $X - u$ con $x > u$ e u molto grande, è approssimativamente

$$H(y) = 1 - \Big(1 + \frac{\xi y}{\tilde{\sigma}}\Big)^{-\frac{1}{\xi}} \tag{4.1}$$

$$\text{con } \tilde{\sigma} = \sigma + \xi(u - \mu), \tag{4.2}$$

dove $y > 0, \qquad 1 + \frac{\xi y}{\tilde{\sigma}} > 0.$

Questa famiglia di distribuzioni si chiama Distribuzione Generalizzata di Pareto (DGP). Il teorema implica che le due distribuzioni, la GEV e la DGP, sono in correlazione: i parametri della famiglia di Pareto sono infatti determinati unicamente da quelli della distribuzione GEV e, in particolare ξ è uguale tra le due. Questo significa che il parametro di forma ξ è dominante nel determinare il comportamento qualitativo della distribuzione DGP, come lo era per GEV.

- Se $\xi < 0$ la distribuzione degli eccessi ha un limite superiore,

- se $\xi > 0$ la distribuzione non ha limiti,

- se $\xi = 0$ si prende in considerazione il limite di $\xi \to 0$ che porta a

$$H(y) = 1 - \exp\Big(-\frac{y}{\tilde{\sigma}}\Big) \qquad y > 0,$$

corrispondente a un'esponenziale con parametro $1/\tilde{\sigma}$.

Possiamo quindi concludere che gli y_j sono visti come realizzazioni indipendenti di una variabile aleatoria, la cui distribuzione può essere approssimata da un membro della famiglia generalizzata di Pareto. Vedremo di seguito come fare inferenza con i dati osservati, verificare il modello e fare previsioni.

Giustificazioni del teorema Partiamo dal risultato del Teorema 1

$$F^n(z) \approx \exp\Big\{ -\Big[1+\xi\Big(\frac{x-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}\Big\},$$

per parametri $\mu, \sigma > 0$ e ξ,

$$n\log(F) \approx -\Big[1+\xi\Big(\frac{x-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}. \tag{4.3}$$

A questo punto ci avvaliamo dell'espressione di Taylor: $\log F(z) \approx -1 - F(z)$, valida per valori grandi di z, che sostituita in (4.3) ci dà, proprio come stavamo cercando, un' approssimazione si F

$$1 - F(u) \approx \frac{1}{n}\Big[1+\xi\Big(\frac{x-\mu}{\sigma}\Big)\Big]^{-\frac{1}{\xi}}.$$

Possiamo dunque stimare la probabilità:

$$\begin{aligned} P(X > u+y|X>u) &\approx \frac{n^{-1}[1+\xi(u+y-\mu)/\sigma]^{-\frac{1}{\xi}}}{n^{-1}[1+\xi(u-\mu)/\sigma]^{-\frac{1}{\xi}}}, \\ &= \Big[1+\frac{\xi y}{\tilde{\sigma}}\Big]^{-\frac{1}{\xi}}, \end{aligned}$$

dove $\tilde{\sigma} = \sigma + \xi(u-\mu)$.

4.2 Selezione della soglia

Dobbiamo scegliere la soglia u in modo da assicurare un equilibrio tra devianza e varianza: una soglia troppo bassa va contro le ipotesi asintotiche del modello e fa aumentare la devianza, una soglia troppo alta genera pochi valori estremi su cui stimare il modello, e quindi la varianza sarebbe elevata. Esistono due metodi per individuare u:

1. Tecnica esplorativa, basata sulla media della distribuzione generalizzata di Pareto.

2. Ricerca della stabilità dei parametri, basata sulla stima di vari modelli in un rango di differenti soglie.

1. Quando $\xi < 1$, la media della DGP è $E(Y) = \sigma/(1-\xi)$, quando $\xi \geq 1$ è, invece, infinita. Se u_0 è la soglia prescelta, la media è

$$E(X - u_0 | X > u_0) = \frac{\sigma_{u_0}}{1-\xi} \qquad \text{con } \xi < 1,$$

dove indichiamo con σ_{u_0} il parametro di scala riferito a u_0. Ma se la distribuzione DGP è valida per u_0, allora è valida ugualmente anche per tutte le soglie $u > u_0$. Quindi

$$\begin{aligned} E(X - u | X > u) &= \frac{\sigma_u}{1-\xi} \\ &= \frac{\sigma_{u_0} + \xi u}{1-\xi}, \end{aligned} \tag{4.4}$$

in virtù della (4.2). Per $u > u_0$ l'espressione della media è una funzione lineare in u, dove $u < x_{max}$. Rappresentando, in un grafico, la stima della media al variare della soglia, otteniamo un grafico di punti:

$$\Big\{u, \frac{1}{n_u} \sum_{i=1}^{n_u} (x_{(i)} - u)\Big\},$$

dove n_u rappresenta il numero di valori che eccedono la soglia u.

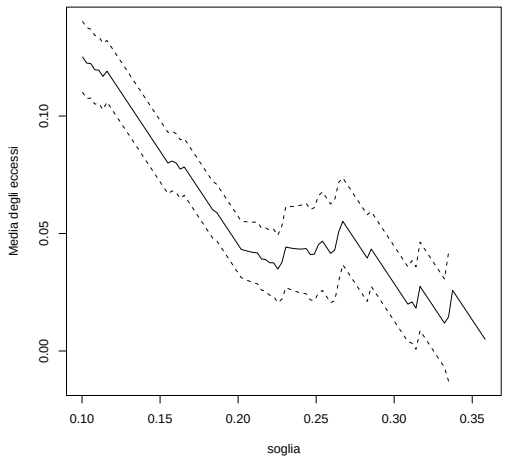

Figura 4.1: Mean Residual life plot del Larice

Questo grafico si chiama Mean residual life plot e permette di individuareu_0 come quel punto dell'asse delle ascisse per cui, al si sopra di esso, la curva delle medie appare approssimativamente lineare. Il grafico in fig 4.1 ci suggerisce che una possibile u_0 potrebbe essere la soglia 0.255. Non sempre, però il grafico è facile da interpretare. Per maggiori conferme utilizziamo la seconda procedura.

2. La seconda metodo per la scelta di u_0 è quello di stimare il modello sotto un rango di soglie, fino a raggiungere la stabilità dei parametri.
Al variare degli u, maggiori di u_0, per i quali l'ipotesi asintotica della distribuzione generalizzata di Pareto rimane valida, si verifica che:

- la stima di $\boldsymbol{\xi}$ deve rimanere *costante* perchè i parametri di forma sono identici al variare di u, $\xi_{u_0} = \xi_u$,

- la stima di $\boldsymbol{\sigma}$ deve essere *lineare in* u perchè, per la (4.2)

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

il parametro di scala cambia con u. Per renderlo costante lo riparametrizziamo: $\sigma^* = \sigma_u - \xi u$.

Verifichiamo queste due ipotesi rappresentando in un grafico ξ e σ^* contro u, e selezioniamo la soglia u_0 come il valore più basso di u per il quale le stime rimangono quasi-costanti.

4.3 Inferenza sulla distribuzione DGP

Determinando la soglia, abbiamo di conseguenza determinato i dati su cui si baserà tutta la nostra analisi. Su di essi, attraverso la massima verosimiglianza, possiamo ora stimare i parametri della distribuzione DGP. La log-verosimiglianza per i k eccessi $y_1 ... y_k$ è

$$\ell(\sigma, \xi) = -k \log \sigma - (1 + 1/\xi) \sum_{i=1}^{k} log(1 + \xi y_i/\sigma), \tag{4.5}$$

$$\ell(\sigma) = -k \log \sigma - \sigma^{-1} \sum_{i=1}^{k} y_i, \qquad \text{con } \xi \neq 0.$$

Non è possibile massimizzare analiticamente la log-verosimiglianza, sono necessarie tecniche matematiche. Si deve prestare attenzione a evitare le instabilità numeriche quando $\xi \approx 0$.

4.4 Valori di ritorno

Stimando i parametri determiniamo la distribuzione e quindi il comportamento dei nostri valori estremi, questo ci permette di fare previsioni e di conoscere il comportamento dell'intera popolazione da cui deriva il campione. Nel nostro caso ci permette di stimare il comportamento della dimensione delle cellule dell'intera superfice del tronco dell'albero analizzato.

Ritornando al campione $X_1 ... X_n$, sappiamo che per $x > u$ la probabilità è

$$P(X > x | X > u) = \Big[1 + \xi\Big(\frac{x-u}{\sigma}\Big)\Big]^{-1/\xi},$$

da cui ricaviamo che la probabilità di X è

$$P(X > x) = \zeta_u \Big[1 + \xi\Big(\frac{x-u}{\sigma}\Big)\Big]^{-1/\xi},$$

dove $\zeta_u = P(X > u)$. Pongo $P(X > x) = 1/m$ e trovo i quantili,

$$x_m = \begin{cases} u + \frac{\sigma}{\xi}\big[(m\zeta_u)^{\xi} - 1\big] & \text{con } \xi \neq 0 \\ x_m = u + \sigma \log(m\zeta_u) & \text{con } \xi = 0. \end{cases} \tag{4.6}$$

dove m, chiamata m-osservazione del livello di ritorno, è un numero sufficentemente grande da assicurare che $x_m > u$ e x_m è il quantile che viene superato, in media, una volta ogni m osservazioni.
Rappresentando in un grafico, x_m contro m, in scala logaritmica, otteniamo un grafico con le stesse caratteristiche qualitative del return level plot basato sui modelli GEV: lineare con $\xi = 0$, concavo con $\xi > 0$, convesso con $\xi < 0$.

In genere, per facilitare l'interpretazione, è conveniente usare i valori di ritorno in scala annuale, cioè al posto di m usare N=numero di anni. Otteniamo quindi un grafico di z_N contro N che ci dirà, in media, una volta ogni quanti anni un determinato livello verrà superato. In questo contesto, sapendo che ogni anno si hanno n_y osservazioni, $m = N \cdot n_y$.

Nel nostro caso, dove non si ha a che fare con il tempo ma con lo spazio, poichè non possiamo misurare tutta la superfice dell'albero, ma solo alcuni campioni, vogliamo sapere qual è, in media, la probabilità di incontare, in ogni C campioni, valori superiori a una determinata media. Poniamo quindi$m = C \cdot n_c$ dove, n_C= numero di osservazioni in ogni campione e C= numero di campioni. In questo caso z_C indica il livello che ci si aspetta possa essere superato, in media, una volta ogni C campioni analizzati,

$$z_C = u + \frac{\sigma}{\xi}\Big[(Cn_c\zeta_u)^{\xi} - 1\Big]. \tag{4.7}$$

La stima dei valori di ritorno richiede la sostituzione delle stime dei parametri. Ci manca la stima di ζ, che è la probabilità di un'osservazione di superare la soglia u. Una stima naturale è la proporzione dei punti che superano$u$,

$$\hat{\zeta}_u = \frac{k}{n}.$$

4.5 Model check

Una volta scelta la soglia u e quindi determinati gli eccessi $y_{(1)}...y_{(k)}$, per testare la qualità del modello generalizzato di Pareto stimato, ci serviamo dei seguenti grafici.

- **Probabilità plot:**
  Sapendo che $\widehat{H}(y) = 1 - \left(1 + \frac{\hat{\xi}y}{\hat{\sigma}}\right)^{-1/\hat{\xi}}$ con $\hat{\xi} \neq 0$ l'insieme dei punti del grafico è cosi definito:

$$\{(i/(k+1), \widehat{H}(y_{(i)})), i = 1, ..., k\}.$$

- **Quantile plot:**
  Sapendo che $\widehat{H}^{-1}(y) = u + \frac{\hat{\sigma}}{\hat{\xi}}[y^{-\hat{\xi}} - 1]$ sempre assumando $\hat{\xi} \neq 0$
  l'insieme dei punti del grafico è cosi definito:

$$\{(\widehat{H}^{-1}(i/(k+1)), y_{(i)}), i = 1, ..., k\}.$$

  Se il modello generalizzato di Pareto è ragionevole per modellare gli eccessi di u, entrambi i grafici dovrebbero essere aprossimativamente lineari.

4.6 Applicazione di DGP ai dati

Applichiamo il metodo della soglia ai dati sui tre alberi, per convenzione i dati sono stati riscalati ($x/10000$).

Larice Per prima cosa dobbiamo affrontare il problema della scelta della soglia u. Scelta che servirà a definire quante, tra tutte le osservazioni del Larice, saranno i valori estremi. Il primo metodo, basato sull'analisi del Mean residual plot, rappresentato in fig. 4.1, come commentato prima, ci porta a scegliere $u = 0.25$. Anche il secondo metodo, quello basato sull'analisi dei grafici di figura 4.2, ci conferma questa scelta. Il cambio di comportamento per soglie molto alte, che era stato osservato nel grafico precedente, è apparente anche qui: per valori maggiori di 0.25 le curve perdono linearità. Le stime di massima verosimiglianza sono dunque,

$$(\sigma, \xi) = (0.0728, -0.585),$$

con corrispondente valore di verosimiglianza -28.662. La $\hat{\xi}$ è leggermente superiore a -0.5, per cui la stima potrebbe non essere regolare. Non riusciamo infatti a costruire un regolare profilo di verosimiglianza per ξ.

Gli errori standar per le due stime sono,

$$S.E. = (0.033, 0.395),$$

quindi l'intervallo al 95% per ξ è [-1.359,0.1891]. Anche se comprende alcuni valori positivi, consideriamo ξ negativo e stimiamo il punto massimo della

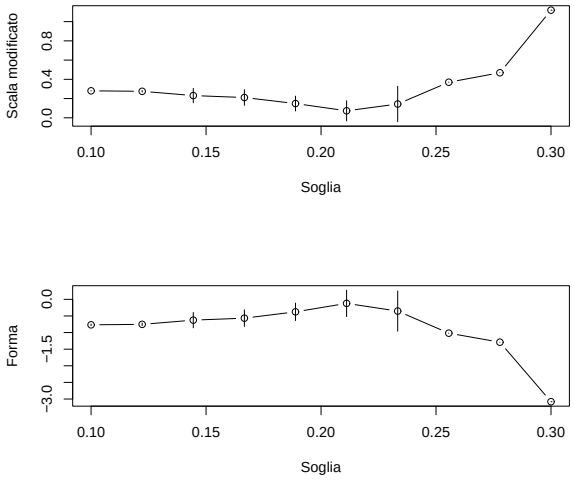

Figura 4.2: Stime dei parametri contro la soglia per dati sul Larice

distribuzione,

$$x_0 = u - \frac{\sigma}{\xi} = 0.25 - 0.0728/(-0.585) = 0.3744.$$

Tornando alla scala originale quindi, il massimo possibile della dimensaione cellulare del Larice è 3744 micromentri quadrati.

Fissiamo ora l'attenzione sui livelli di ritorno. Per poter calcolare i quantili x_m, stimiamo ζ_C, probabilità dei valori estremi,

$$\hat{\zeta}_C = \frac{13}{53} = 0.245,$$

dove 13 sono le osservazioni che eccedono la soglia e 53 il totale delle osservazioni. La varianza è Var(ζ_C)=0.0034. Per calcolare il 100-campione return level abbiamo $m = 100 \cdot n_C$ (n_C numerosità del campione). Non abbiamo più, come prima un solo valore per campione, ma n_C valori per campione. Sostituendo in (4,6) troviamo $\hat{x}_m = 0.369$ con varianza 0.0207. Dunque in ogni 100 campioni si stima che, in media, almeno una cellula abbia un diametro superiore a 3690 micrometri quadrtati. L'intervallo di confidenza al 95% è [0.328,0.409].

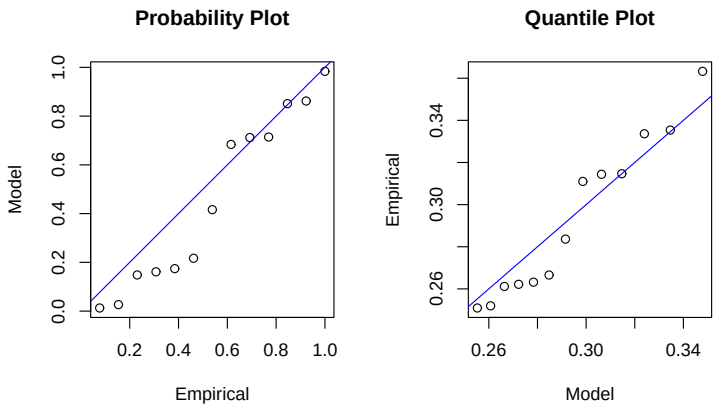

Figura 4.3: Grafici diagnostici per DGP sui dati di Larice

Infine, i grafici diagnostici per il modello generalizzato di Pareto stimato per il Larice sono mostrati in fig. 4.3. La linearità dei punti non è ottima, ma dobbiamo tenere in considerazione che si basa su poche osservazioni.

Pioppo Nero In fig. 4.4 è riportato il Mean residual life plot per i dati del Pioppo Nero. Decidiamo di scegliere $u = 0.12$, perchè il grafico risulta curvo fino a circa 1.2 e poi scende abbastaza linearmente.

La scelta trova conferma anche dal grafico in fig. 4.5, dopo la soglia 1.2 infatti, le curve non risultano più costante. Questa scelta porta ad avere 41 valori eccedenti su un totale di 236 osservazioni. La stima di ζ_C, probabilità dei valori estremi, è $\hat{\zeta}_C = \frac{13}{53} = 0.245$. Le stime di massima verosimiglianza dei parametri della distribuzione di Pareto sono,

$$(\sigma, \xi) = (0.251, -0.573),$$

con corrispondente valore di verosimiglianza -39.173. Notiamo che anche in questo caso ξ è leggermente maggiore di -0.5. Gli errori standar per le due stime sono,

$$S.E = (0.048, 0.143),$$

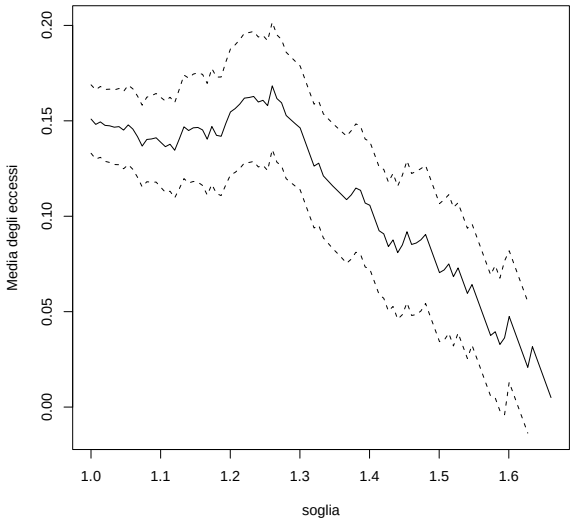

Figura 4.4: Mean Residual life plot del Pioppo Nero

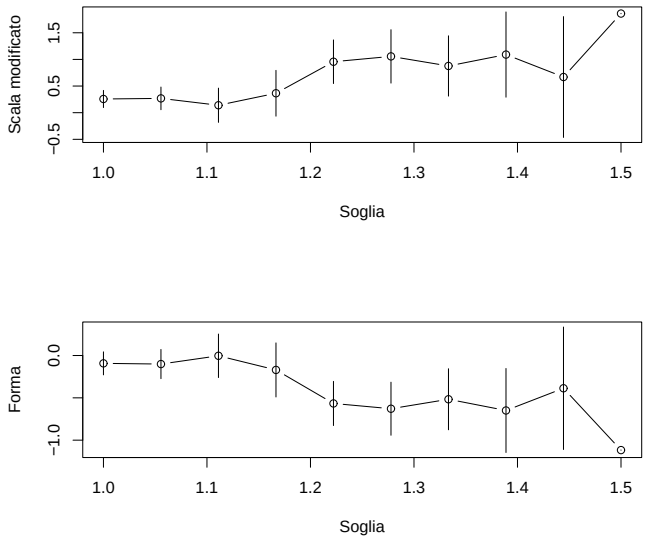

Figura 4.5: Stime dei parametri contro la soglia per dati sul Pioppo nero

quindi l'intervallo al 95% per ξ è [-0.855 , -0.291], per cui siamo certi dell'esistenza di un punto massimo,

$$x_0 = u - \frac{\sigma}{\xi} = 1.25 - 0.251/(-0.573) = 1.687.$$

Possiamo affermare che il massimo possibile della dimensione cellulare del Pioppo nero è 16870 micromentri quadrati. I grafici di fig. 4.6, confermano l'ottima affidabilità dei dati al modello, i punti sono perfettamente allineati.

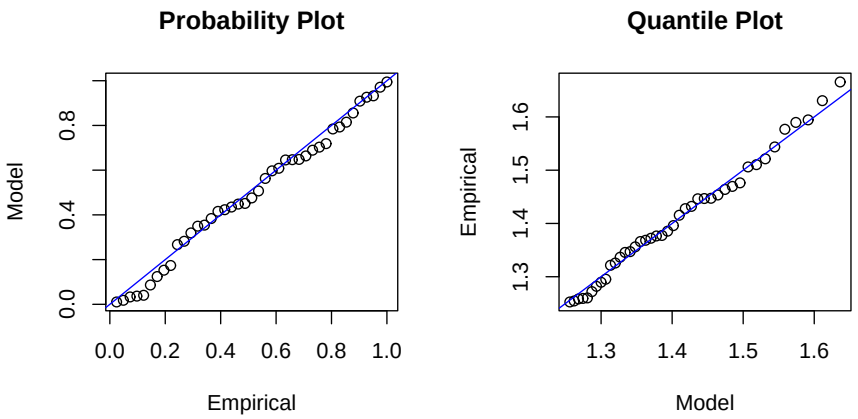

Figura 4.6: Grafici diagnostici per GPD sui dati di Pioppo nero

Robinia Osservando la fig. 4.7, risulta molto difficile scegliere la soglia u; il grafico si presenta completamente lineare, fin da soglie basse. Il grafico 4.8 però ci suggerisce che u potrebbe essere pari a 2.5 visto che, da questa soglia in poi, l'intervallo di confidenza tende ad aumentare notevolmente.

Questa scelta porta ad avere 41 valori eccedenti su un totale di 236 osservazioni. La stima di ζ_C, probabilità dei valori estremi, è $\hat{\zeta}_C = \frac{13}{53} = 0.245$. Le stime di massima verosimigliuanza dei parametri della distribuzione di Pareto sono,

$$(\sigma, \xi) = (1.198, -0.348).$$

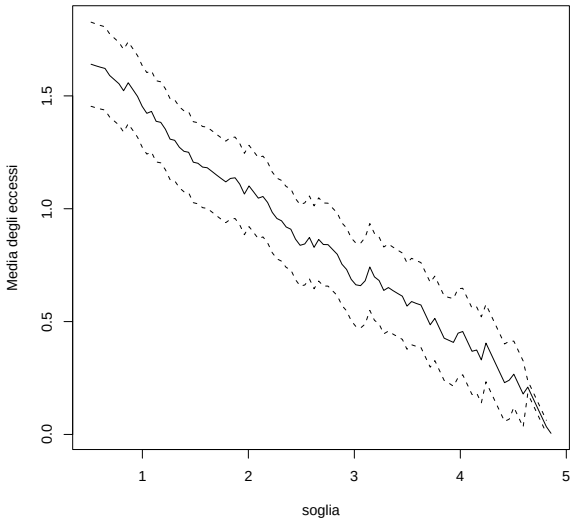

Figura 4.7: Mean Residual life plot della Robinia

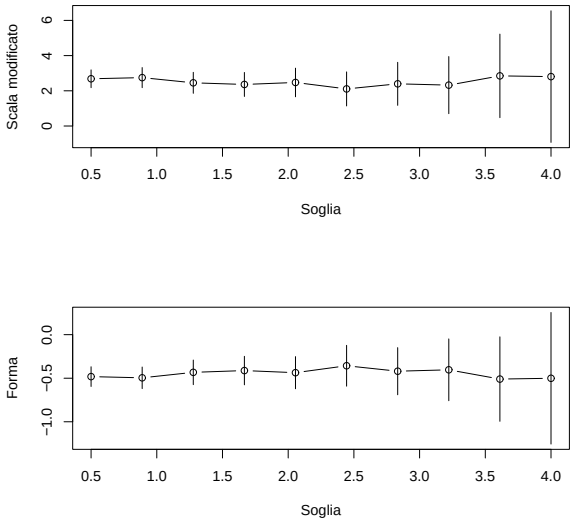

Figura 4.8: Stime dei parametri contro la soglia per dati sulla Robinia

Gli errori standard sono,

$$S.E. = (0.215, 0.124),$$

quindi l'intervallo al 95% per ξ è [-0.593 , -0.104], per cui siamo certi dell'esistenza di un punto massimo,

$$x_0 = u - \frac{\sigma}{\xi} = 2.5 - 1.198/(-0.348) = 5.938.$$

Il massimo cellulare della Robinia è 59380 micromentri quadrati.

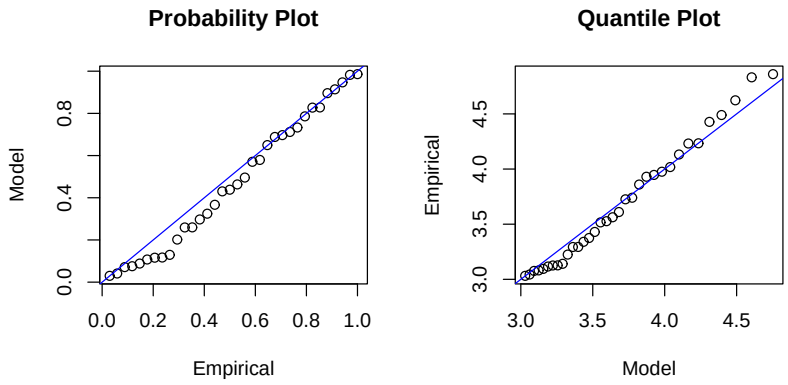

Figura 4.9: Grafici diagnostici per GPD sui dati di Robinia

I grafici di fig. 4.6, confermano l'ottima affidabilità dei dati al modello, i punti sono tutti perfettamente allineati.

4.7 Confronto tra i 3 alberi

	Larice	Pioppo nero	Robinia
σ	0.073	0.251	1.198
ξ	-0.585	-0.573	-0.348
x_0	0.374	1.687	5.938

Tabella 4.1: Confronto tra le stime dei parametri dei 3 alberi. (σ è il parametro di scala, ξ il parametro di forma e x_0 è il massimo cellulare possibile).

Come abbiamo visto, il modello DGP dunque si adatta abbastanza bene ai dati dei tre alberi. Nella tabella sopra sono riportate le principali stime ottenute. Anche qui, le stime dei massimi sono molto diverse a seconda dell'albero e, minori rispetto a quelle calcolate con il modello GEV. La stima del parametro ξ è quasi uguale per il Larice e il Pioppo nero, un pò diversa per la Robinia.

Capitolo 5

Proposta di analisi Bayesiana

Abbiamo fin qui analizzato e commentato due modelli della teoria dei valori estremi, il Modello Classico basato su GEV e il Modello della Soglia basato su DGP. Entrambi, sono stati uno strumento valido per determinare la dimensione cellulare massima cercata. A volte abbiamo avuto dei problemi, con la massima verosimiglianza, nel trovare l'intervallo di confidenza delle stime. Un'analisi più completa si potrebbe ottenere attraverso il Metodo Bayesiano, basato anch'esso sulla funzione di verosimiglianza, però con una procedura del tutto diversa. In questa tesina non andremmo a rianalizzare i dati secondo questo metodo, ma come proposta di approfondimento, introduciamo soltanto brevemente la Teoria generale e i relativi vantaggi.

5.1 Teoria Generale

Ipotizziamo che i dati $X_1 ... X_n$ sono realizzazioni di una variabile aleatoria la cui densità è una funzione compresa nell'insieme F

$$F = \{f(x;\theta) : \theta \in \Theta\},$$

famiglia di distribuzioni parametriche. La novità rispetto alle precedenti analisi stà nella possibilità di includere delle caratteristiche di θ senza far riferimento ai dati, caratteristiche che vengono espresse da una distribuzione di probabilità.

Diversamente dal solito, θ non viene più considerata una costante da stimare, ma una realizzazione di una determinata distribuzione di probabilità. Ad esempio, se siamo sicuri che $0 < \theta < 1$ e che ogni valore in questo rango è ugualmente possibile, il comportamento di θ può essere espresso dalla distribuzione di probabilità della Uniforme,

$$\theta \sim U(0, 1).$$

Questa distribuzione di probabilità per θ, fissata senza far riferimento ai dati X, è chiamata *Prior distribution*. Da costante ignota θ diventa una variabile aleatoria con densità $f(\theta)$. Se gli X_i sono indipendenti , la verosimiglianza per θ si definisce

$$f(x|\theta) = \prod_{i=1}^{n} f(x_i; \theta).$$

Le informazioni di base su cui si basa la scelta della Prior distribution rappresenta allo stesso tempo il punto di forza ma anche di debolezza dell'inferenza Bayesiana. Da una parte perché aumentano l'informazione, dall'altra perché queste informazioni risultano soggettive e quindi non attendibili.

Teorema di Bayes: Stabilisce come convertire un iniziale insieme di caratteristiche su θ, rappresentate dalla Prior distribution $f(\theta)$, in una *Posteriror distribution* $f(\theta|x)$ che include le informazioni addizionali fornite dai dati X. La seguente funzione,

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{\int_\Theta f(\theta)f(x|\theta)d\theta},$$

definisce il comportamento di θ condizionato al fatto di conoscere i dati X. Riassumendo in uno schema:

$$\underset{\text{di } \theta}{\text{Primitive caratteristiche}} \rightarrow f(\theta) \rightarrow \underset{\text{dei dati}}{\text{informazioni}} \rightarrow f(\theta|x).$$

Predizioni Anche i valori predetti hanno una loro densità,

$$f(\theta|x) = \int_\Theta f(z|\theta)f(\theta|x)d\theta,$$

dove $f(Z|\theta)$ è la funzione di probabilità di z, le osservazioni future, e $f(\theta|x)$ è la Posterior distribution di θ sulla base dei dati osservati.

Diversamente da altri approcci, la densità dei valori predetti riflette l'incertezza del modello -attraverso il termine $f(\theta|x)$- e l'incertezza dovuta alla variabilità degli z -attraverso il termine $f(Z|\theta)$. Quindi, affrontate le difficoltà per specificare una Prior distribution, ci sono forti ragioni pragmatiche per preferire l'analisi Bayesiana.

Vantaggi dell'inferenza Bayesiana:

- Il risultato di un'analisi Bayesiana non dà la stima dei parametri, ma fornisce una distribuzione, la Posterior distribution, che permette una analisi più compiuta rispetto alla massima verosimiglianza.

Ostacoli all'inferenza Bayesiana:

- Soggettività dell'uso della prior distbution.
- Difficoltà computazionali: è difficile calcolare l'integrale normalizzatore della $f(\theta|x)$. Quando θ ha un elevato numero di parametri il calcolo del denominatore può essere problematico anche utilizzando sofisticate tecniche numeriche. Il recente sviluppo di numerose tecniche di simulazione ha semplificato il problema e sviluppato l'uso delle tecniche Bayesiane. Una tecnica in particolare è la catena Markov chain Monte Carlo (MCMC).

5.2 Analisi bayesiana per valori estremi

Ci sono molte ragioni per preferire l'analisi bayesiana nello studio di dati estremi.

1. La facilità di risolvere il problema della scarsità dei dati, includendo, attraverso la Prior distribution, altre fonti di informazioni utili.
2. Maggiore precisione nel fare inferenza, sopratutto nel prevedere con quale probabilità eventi futuri possano superare i livelli estremi, che è lo scopo principale della Teoria dei Valori Estremi.

Per esempio, il modello adatto ai massimi M_L del Larice era $M_L \sim GEV(\mu, \sigma, \xi)$ e sulla base delle osservazioni $\{x_1, ..., x_{25}\}$, attarverso la log-verosimiglianza, abbiamo trovato la stima di $\theta = (\mu, \sigma, \xi)$. Se avessimi adottato l'analisi bayesiana il risultato ottenuto sarebbe stato la Posterior distribution, e quindi la probabilità dei valori predetti Z sarebbe stata,

$$P(Z \leq z|x_1, ..., x_n) = \int_\Theta P(Z \leq z|\theta) f(\theta|x) d\theta.$$

Questa distribuzione, dei i futuri massimi, include sia l'incertezza parametrica che l'errore casuale delle osservazioni future.

3. L'indipendenza dell'analisi da assunzioni di regolarità, che sono richieste nella teoria asintotica della massima verosimiglianza. In particolare quando $\xi < -0.5$ e la teoria della massima verosimiglianza crolla, l'inferenza Bayesiana offre una valida alternativa.

5.3 Conclusioni

L'obbiettivo principale della nostra tesina era quello di stimare con buona affidabilità la Massima Dimensione Cellulare di ogni albero. Applicando il modello GEV alla serie dei massimi campionari abbiamo ottenuto le seguenti stime (l'unità di misura sono i micrometri quadrati):

$$Larice, Pioppo, Robinia = (4110, 17890, 58410).$$

L'affidabilità di questi valori non risultava pero molto buona. Basti pensare che gli intervalli di confidenza al 95% ottenuti dal profilo di verosimiglianza erano molto grandi e addirittura per il Pioppo e la Robinia erano superiormente illimitati. Per raggirare il problema abbiamo svolto l'analisi basandoci su più dati, tutti i valori maggiori di una ceta soglia, e ad essi abbiamo applicato il modello DGP. Le stime che abbiamo ottenuto sono le seguenti:

$$Larice, Pioppo, Robinia = (3744, 16870, 59380).$$

In questo caso le stime della dimensione massima risultavano minori rispetto a prima. Però per il Larice e il Pioppo nero il parametro di formaξ risultava minore di -0.5, per cui la stima non era del tutto regolare. L'analisi Bayesiana è dunque una valida alternativa che offre maggiori vantaggi per risolvere questi problemi di inferenza.

Bibliografia

[1] Stuart Coles: *An introduction to Statistical modeling of Extreme Values*, Springer Series instatistics, 2001.

[2] Azzalini Adelchi: *Inferenza statistica : una presentazione basata sul concetto di verosimiglianza*, Springer-Italia, 2^a ed., Milano 2001.

[3] Freedman D., Pisani R., Purves R.: *Statistica*, McGraw, Milano 1998.