Università degli Studi di Padova
Facoltà di Scienze Statistiche

Corso di Laurea Specialistica in
Scienze Statistiche Demografiche e Sociali

Tesi di Laurea

I terremoti in Italia: un'analisi dei valori estremi

Relatore: Prof. Stuart G. Coles

Laureanda: Ilaria Prosdocimi

ANNO ACCADEMICO 2005-2006

Indice

Introduzione

Questo lavoro nasce dall'idea di riproporre quanto fatto da Coles e Sparks nel loro *Extreme values Methods for modelling historical series of large volcanic magnitudes*, per un dataset riguardante eventi sismici, piuttosto che fenomeni vulcanici. L'intento è quello di verificare se la funzione di censura identificata nell'articolo potesse essere valida anche per dei fenomeni sismici, date le similitudini che intercorrono tra questi eventi. L'interesse per i fenomeni sismici è inoltre nato dal fatto che l'Italia è da sempre soggetta a terremoti, che hanno nel corso del tempo lasciato ferite profonde: riuscire a capire meglio come tali eventi sorgono e comprenderne meglio il loro comportamento nel tempo, è di grande interesse per il nostro paese. I dati utilizzati sono stati recuperati tramite internet, e corrispondono ad un dataset riguardante i sismi avvenuti in Italia dal 217 a.C. ad oggi (capitolo 2).

Il lavoro si propone di trattare i dati a livello puramente statistico, andando ad utilizzare tecniche di analisi dei valori estremi, ed in particolare la modellazione tramite processi di punto (sezione 4.1). Poichè tali processi richiedono che il fenomeno in esame possa essere considerato stazionario, viene condotta in prima istanza un'analisi che prende in considerazione solo l'ultima parte del dataset, in cui i dati possono appunto essere considerati stazionari (sezione 4.2). Infine viene condotta l'analisi su tutto il dataset, in cui i dati riguardanti sismi avvenuti nel passato sono molti meno di quelli raccolti nel presente, e si comportano in maniera diversa da quelli più recenti, a causa di problemi tecnici che rendevano impossibile nel passato la stessa precisione

che abbiamo ai giorni nostri nel raccogliere informazioni riguardanti i sismi. Venendo quindi a cadere l'ipotesi di stazionarietà dei dati raccolti, a causa di una sorta di censura avvenuta sui dati, si è reso necessario inserire nel modello una funzione che potesse stimare il processo di censura avvenuto nel tempo, così come si trova nell'articolo di Coles e Sparks.

Nel provare ad applicare una funzione di censura simile a quella utilizzata da Coles e Sparks, si sono tuttavia verificati problemi di tipo numerico (sezione 5.1), e si sono cercate nuove funzioni che potessere stimare la funzione di censura (sezioni 5.2 e 5.3). Queste ultime mostrano come il processo di censura dipenda principalmente dal tempo, e non dalla magnitudo del terremoto: la probabilità che un terremoto sia stato registrato o meno dipende solo dal tempo in cui esso è avvenuto e non dalla magnitudo momento che esso ha generato.

Infine nel capitolo 6 si è andati a verificare se i diversi modelli fossero stimati bene attraverso uno studio di simulazione. E' stato simulato un processo con parametri noti, ed in primo luogo si è fatta un'analisi dei dati stazionari tramite processi di punto. Sucessivamente ai dati è stato applicato un processo di censura tramite diverse funzioni, andando poi a stimare i modelli. Si è andati così a verificare se i diversi modelli riuscissero a stimare in maniera soddisfacente le diverese classi di funzioni di censura.

Capitolo 1

I terremoti

I terremoti sono da sempre stati per l'uomo un evento, inaspettato e terribile, che non si riesce a spiegare e prevedere.

Fin dall'antichità la vita dell'uomo è stata costellata da fenomeni di tipo sismico, inaspettati e terribili, che difficilmente trovavano una spiegazione e che era impossibile prevedere. Fu solo dal XVIII secolo che la scienza iniziò a studiare i terremoti, per cercare di capire cosa li originasse e poterli prevedere, in modo da potersi proteggere almeno in parte i loro effetti distruttivi. Da allora la scienza ha imparato molto su questi eventi, sebbene essi provochino ancora, in maniera a volte non prevedibile, terribili distruzioni, ovunque nel mondo. In realtà piccoli sismi avvengono continuamente nel tempo, tuttavia solo alcuni di questi sono percepiti dell'uomo, mentre molti vengono registrati solo dalle macchine, che hanno raggiunto ai nostri giorni dei livelli di sensibilità elevatissimi. In particolare però i sismi tendono a verificarsi sempre nelle stesse zone, che sono dette appunto sismicamente attive. In queste zone il sisma è percepito in maniera più forte, le onde del sisma poi si dissipano rapidamente e gli effetti del terremoto vengono percepiti sempre meno dall'uomo, mentre possono essere ancora registrati dalle macchine.

Un terremoto è in definitiva una vibrazione più o meno forte della Terra prodotta da una rapida liberazione di energia meccanica sotto forma di onde sferiche in qualche punto al suo interno, detto ipocentro.

Il punto sulla superficie della verticale dell'ipocentro è invece detto epicentro.

In particolare in un terremoto l'energia che si libera dall'ipocentro, dopo aver passato tutti gli strati della terra fino a raggiungere l'epicentro, viene registrata e riconosciuta attraverso diversi tipi di onde: onde di compressione, onde di taglio e onde superficiali.

Le onde di compressione, o longitudinali, sono le più veloci, e le prime ad arrivare: per questo sono dette onde primo o onde P. Al passaggio delle onde P le rocce si dilatano e si comprimono nella direzione dell'onda; queste onde sono in grado di propagarsi in ogni mezzo, compresa l'aria, e sono infatti esse a creare il rombo cupo che accompagna l'inizio del terremoto propagandosi nell'aria con frequenze percepibili dall'orecchio umano

Le onde di taglio, o trasversali, provocano invece deformazioni di taglio, che vanno a cambiare la forma, ma non il volume delle rocce attraverso cui si propagano.Queste onde non si possono propagare attraverso i fluidi, quali acqua, aria o magma, e, essendo più lente delle onde primarie, vengono dette onde secondarie o onde S.

Le onde P e S sono generate direttamente nell'ipocentro, e vengono per questo dette onde interne, tuttavia quando esse raggiungono la superficie si propagano dall'epicentro lungo la superficie terrestre sotto forma di onde superficiali.

Nel tempo si è cercato inoltre di trovare un metodo per misuare la "forza" di un terremoto e sono state proposte diverse scale e diverse misure. Alcune di queste si basano sugli effetti che il sisma ha sull'ambiente esterno, in particolare sulle strutture costruite dall'uomo, e vengono dette scale di intensità: la più nota tra queste è la scalla detta di Mercalli, dal nome del sismologo che la ha proposta. Un'altra celebre scala, quella di Richter, calcola la cosiddetta Magnitudo, basandosi sul logaritmo del rapporto tra le ampiezze delle onde del terremoto che si sta considerando e un terremoto standard. Esistono in realtà diversi tipi di magnitudo a seconda di quale tipo di onde si

usano per fare il calcolo, tuttavia la più comune si basa sulle onde superficiali, che dominano i sismogrammi. Un'ulteriore possibile misura della grandezza di un terremoto è il cosiddetto momento sismico, che, a differenza della magnitudo non dipende dal tipo di onda, ma cerca di misurare l'effettiva sorgente del terremoto, sotto forma di energia.

Esistono in definitiva diversi modi per misurare la grandezza di un terremoto, e la trattazione di essi esula dallo scopo di questo lavoro. Si vuole però sottolineare come sia complicata la misura di tali eventi, e come sia varia la risposta che si è data al problema.

Capitolo 2

I dati

I dati utilizzati per l'analisi sono quelli contenuti nel Catalogo Parametrico dei terremoti Italiani, versione 2004 (CPTI04), che è un aggiornamento della versione del 1999. I dati contenuti vanno dal 217 a.C. al 2002, e nel catalogo non sono contenute repliche, cioè eventi sismici legati ad un sisma già registrato nel catalogo. Sono stati inseriti solamente eventi avvenuti sul territorio italiano, e quelli che, pur essendo avvenuti in zone limitrofe, siano stati risentiti sul suolo italiano.

Il catalogo contiene una lunga serie di informazioni per ognuno dei sismi registrati: luogo e data di avvenimento, e una serie di informazioni geofisiche che sono state utilizzate solo in parte per questo lavoro.

L'attenzione è posta sulla variabile Maw, la Magnitudo momento del sisma registrato, che è una misura che non dipende dal tipo di onde registrate e dal sismometro, ma è una misura diretta della sorgente del terremoto.

In figura 2.1 si mostra l'andamento della magnitudo momento per anno di accadimento. Si nota come ci siano dei valori di Maw che si ripetono negli anni più frequentemente di altri e che il numero di eventi registrato in Italia ha subito negli anni nette variazioni.

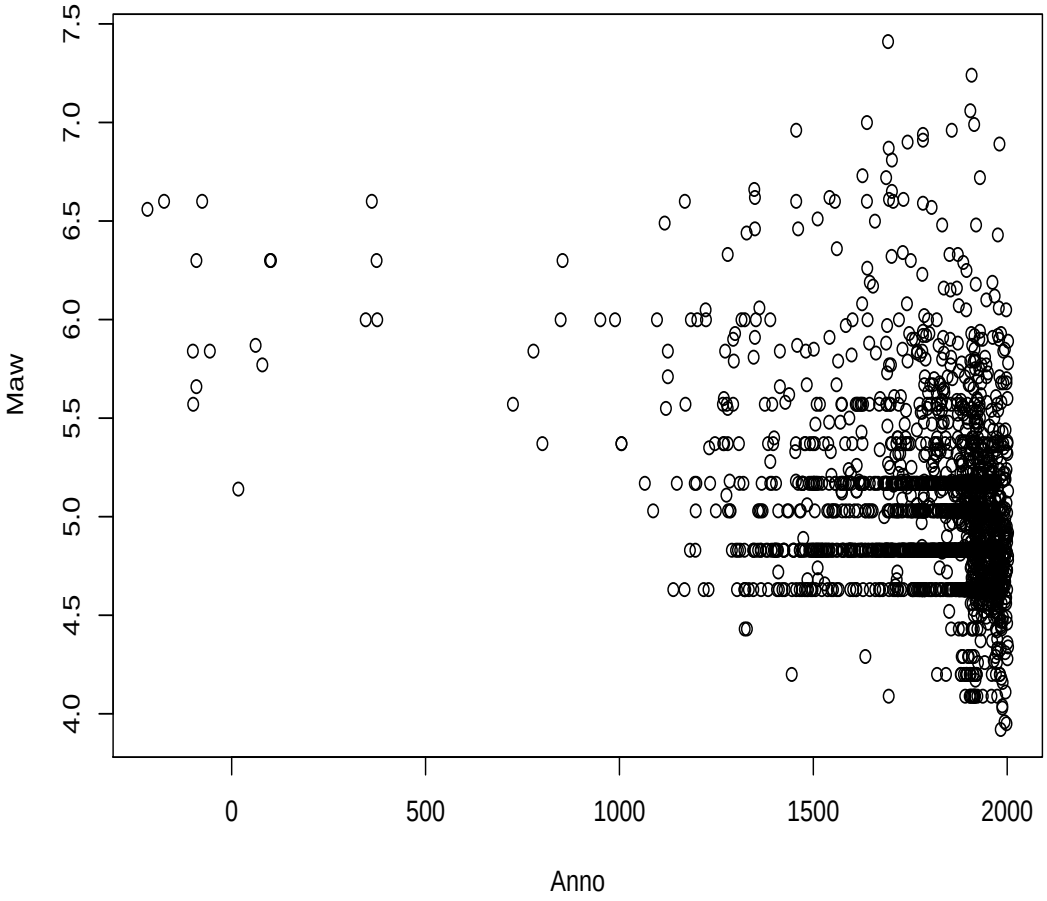


Figura 2.1: La variabile Maw per anno di accadimento

La variabile Maw ha un campo di variazione tra il 3.920 e il 7.41, e nel grafico 2.1 saltano subito all'occhio delle linee orizzontatali, i valori 4.83, 4.63, 5.17, 5.03, 5.57 e 5.37 compaiono rispettivamente 700 343, 266, 188, 48 e 40 volte nel dataset. I motivi di questo fenomeno non ci sono del tutto chiari, probabilmente questo è legato a come sono stati ottenuti i valori numerici e alle attrezzature in uso, tuttavia ciò non toglie che questo potrebbe causare seri problemi sia nel computo che nella valutazione della bontà delle stime, e si è quindi deciso di correggere queste osservazioni, inserendo del rumore. Le osservazioni coincidenti con i valori suddetti sono state cioè cambiate con valori che appartengono ad una uniforme di ampiezza 0.2 centrata nel valore ripetuto, in modo da ottenere più casualità nel processo.

Il risultato di questa operazione viene mostrato nell'istogramma della nuova variabile Maw, che verrà utilizzata da adesso poi (figura 2.2). Ora non si notano più valori che predominano in maniera forte sugli altri, la variabile assume più valori diversi tra loro, e nessun valore viene ripetuto un numero eccessivo di volte.

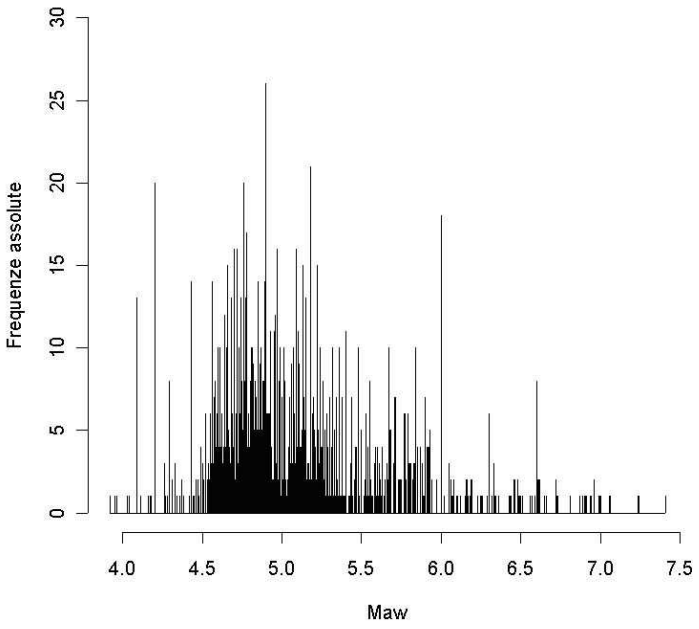


Figura 2.2: La variabile Maw dopo aver modificato i valori che si ripetevano

Come si vede nella figura 2.2 nella parte centrale della distribuzione sono stati registrati la parte più consistente dei terremoti, il 90% delle osservazioni ha un valore tra il 4.54 e il 5.85, sebbene spicchino dei valori anche nella code, sia a destra che a sinistra. La stima kernel della distribuzione in figura 2.3 evidenzia ancora di più la forma con una doppia punta nella parte della distribuzione che comprende buona parte delle osservazioni.

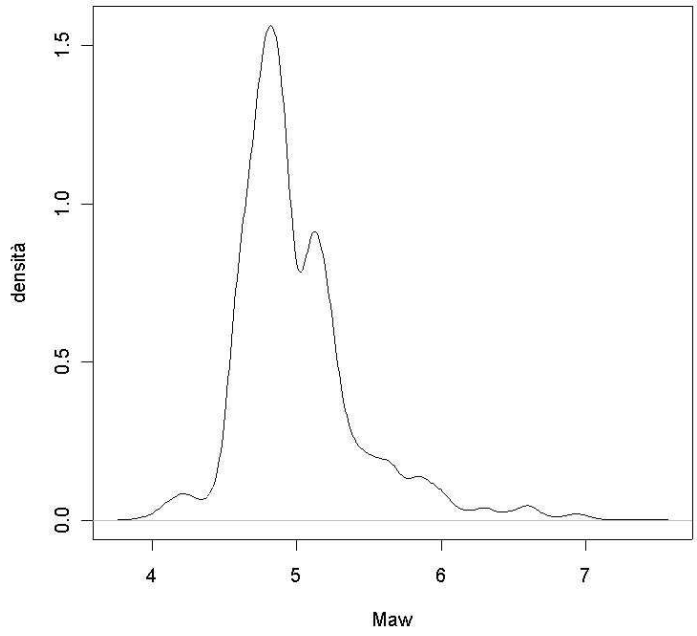


Figura 2.3: Densità stimata con metodo kernel della nuova variabile Maw

In figura 2.4 si presenta il grafico della Magnitudine momento dopo la correzione per l'anno di accadimento dell'evento.

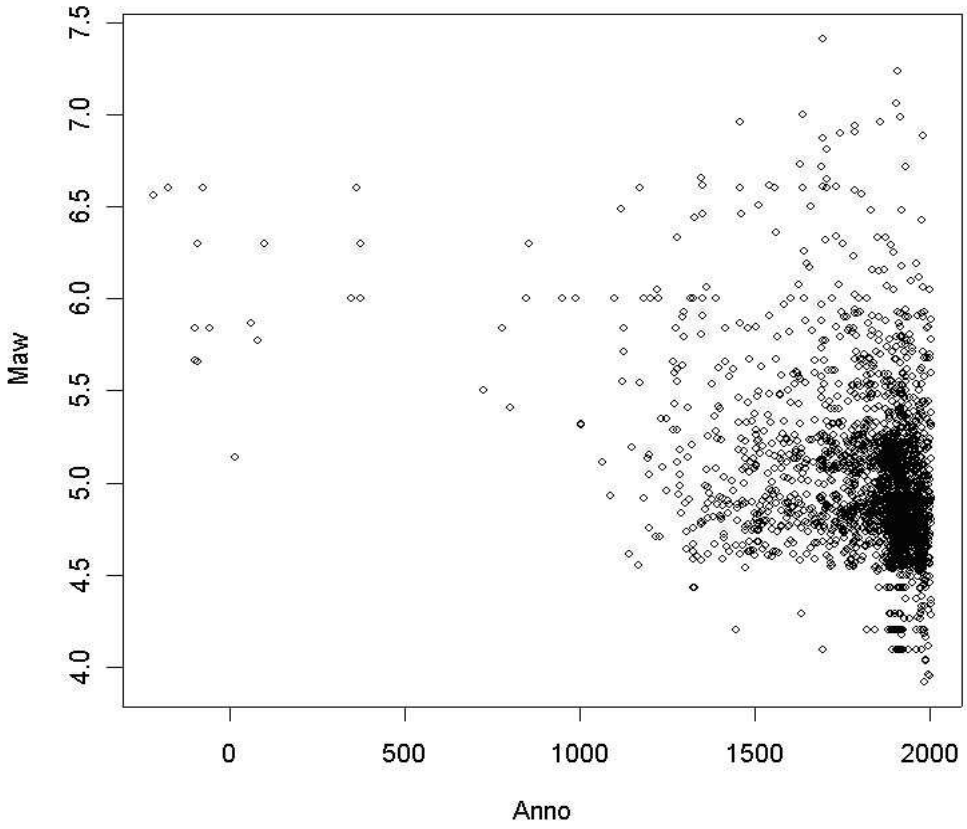


Figura 2.4: Magnitudo momento dei terremoti registrati per anno di avvenimento

Si nota come sia notevole la differenza del numero di sismi registrati negli ultimi tre secoli rispetto a quelli registrati in tutti i precedenti duemila anni, ma questo si evidenzia meglio osservando l'istogramma che mostra il numero di sismi registrati per ogni anno (Figura 2.5).

Le osservazioni sono distribuite in maniera massiccia negli anni più recenti. Il 75% dei dati è stato registrato dopo il 1784 e il 50 % dopo il 1897, nel 1908 si è registrato il massimo numero di terremoti: 28.

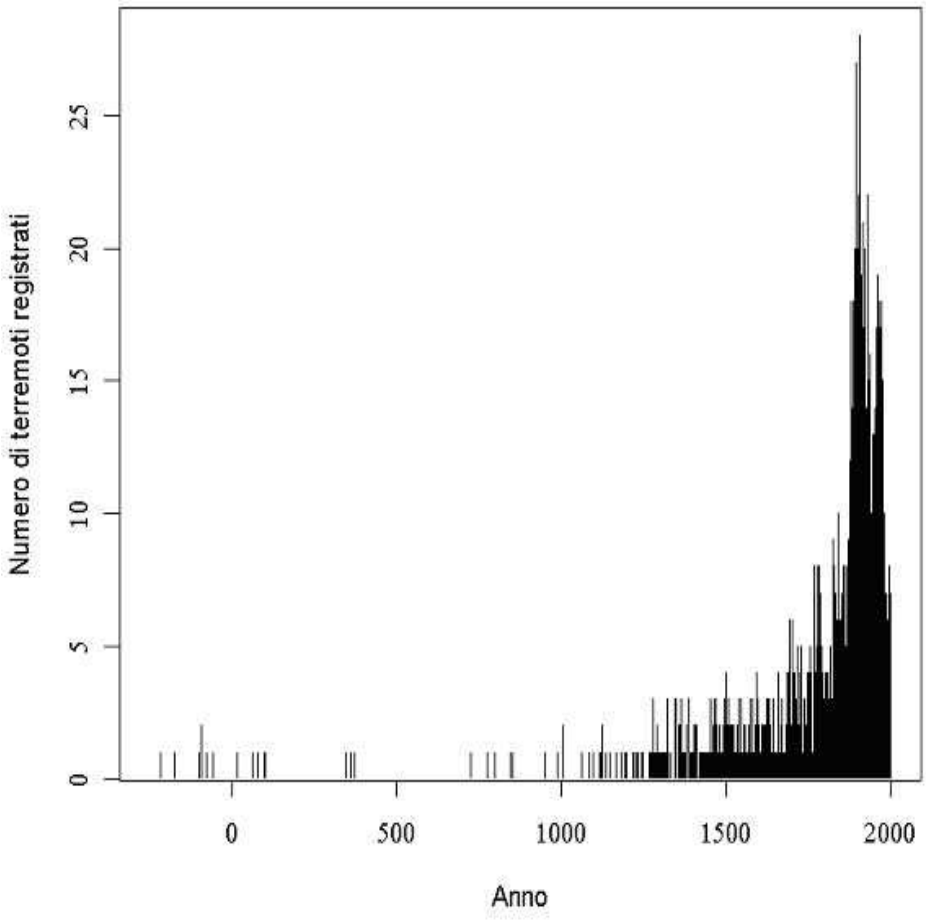


Figura 2.5: Numero di eventi registrati per anno di avvenimento

Questo ovviamente non può rispecchiare l'effettivo andamento dell'attività sismica in Italia, ma rispecchia piuttosto la capacità di registrare gli eventi e di riportarli all'interno del dataset. Inoltre si nota che i sismi registrati negli anni lontani dal presente hanno in genere un valore di Magnitudo momento se non elevato, certamente non basso. Questo perché è ipotizzabile che siano stati registrati nel passato solo eventi percepibili in maniera forte dall'uomo e avvenuti in zone abitate (un terremoto avvenuto in mare, sebbene di una certa entità potrebbe essere non stato registrato dalle cronache del passato).

L'analisi dovrà quindi tener conto di questa diversa distribuzione degli eventi nel tempo. Infatti verranno svolte due diversi tipi d'analisi.

La prima analisi condotta è basata sull'ultima parte della serie, in cui i dati sono stati registrati in maniera regolare e sembrano soddisfare l'ipotesi di stazionarietà, mentre una seconda analisi, che tiene conto della minore probabilità di aver registrato eventi del passato, verrà svolta su tutto il dataset.

In entrambi i casi l'approccio usato per l'analisi dei dati è basato sui processi di punto, ma nella seconda parte, oltre ai parametri necessari a descrivere il processo, sarà necessario definire dei parametri che descrivano le diverse probabilità che un evento ha di essere registrato.

Capitolo 3

Il software utilizzzato

Per condurre le analisi sui dati è stato utilizzato in massima parte il software R. Alcune applicazioni sono state scritte *ex novo* appositamente per i problemi che si ponevano di volta in volta, ma in larga parte sono state utilizzate le normali funzioni di R e delle sue librerie.

Nella parte di calcolo delle stime del primo modello basato solo su una parte del daset si è fatto uso della libreria ismev per R. La libreria *ismev* 1.2 infatti permette di svolgere diverse analisi su dati con valori estremi, ed è inoltre possibile svolgere analisi attraverso processi di punto. Per la stima dei parametro nella seconda parte, riguardante l'intero dataset, è stato invece necessario scrivere delle funzioni che fossero in grado di massimizare la funzione di verosimiglianza. Per questa delicata fase si è preferito utilizzare S-plus rispetto ad R, dato che la funzione *optim*, necessaria per massimizzare la verosimiglianza del modello, in quest'ultimo pare aver dato dei problemi.

Parte I

Processi di Punto per valori estremi: un esempio con i dati sui terremoti italiani degli ultimi 130 anni

Capitolo 4

I Processi di Punto per valori estremi

4.1 La teoria dei processi di punto

Per l'analisi dei dati si è deciso di utilizzare un approccio basato sui processi di punto, secondo le linee indicate da Pickands nel 1971: egli fu il primo a proporre di utilizzare metodiche dei processi di punto per l'analisi dei valori estremi.

I processi di punto sono infatti dei metodi che, sotto l'ipotesi di stazionarietà dei dati, possono essere utilizzati anche per la modellazione di valori estremi, e sono caratterizzati dalla flessibilità e dalla possibilità di poter usare strumenti di inferenza e di modellazione. Tali metodi valgono solo nel caso in cui il processo generatore dei dati possa essere considerato stazionario: l'idea di base di questa modellazione è che il processo di interesse, nel nostro caso il processo sisomgenetico, possa essere visto come la realizzazione di un insieme aleatorio di punti indipendenti tra loro, distribuiti nello spazio bidimensionale definito da tempo e magnitudo momento: di ogni punto i possiamo registrare un valore t_i, momento di avvenimento, e m_i, valore della Magnitudo Momento registrato in modo da ottenere un set di punti $\{ (t_i, m_i), i = 1, ..., n\}$. In particolare, trasformando il tempo in modo che vari tra 0 e 1, possiamo definire i dati come una regione $\mathcal{A}_u = \{(t, m) : 0 < t < 1, m > u\}$, con u il valore di soglia al di sopra

del quale stanno i valori di interesse. Il valore minimo di u osservato, nel caso di tutto il datset, è pari a 3.92. Al crescere dei valori di u si stringe l'attenzione su regioni che sono caratterizzate da eventi sempre più estremi. I processi di punto permettono proprio di modellare il numero e la posizione dei punti nella regione $\mathcal{A}_u$ attraverso un processo di Poisson, che nel nostro caso sarà bidimensionale. In particolare si può dividere $\mathcal{A}_u$ in sottoinsiemi A in cui vengono osservati *N(A)* punti. Nel caso più semplice di un processo di Poisson omogeneo, la media μ_A dei punti in A corrisponde al valore $\lambda\,|A|$, con λ parametro della distribuzione di Poisson: il numero medio dei punti presenti nell'intervallo cioè è proporzionale alla dimensione dell'intervallo stesso e per intervalli che coprono lo stesso lasso temporale t ci si aspetta il medesimo numero medio di punti.

Nel caso invece di un processo di Poisson non omogeneo il numero di punti presenti nel sottoinsieme A si distribuisce sempre come una Poisson di parametro λ, ma quest'ultimo varia nel tempo. Il numero di punti presenti in intervalli aventi la stessa ampiezza temporale varia cioè all'interno della regione secondo la funzione di intensità $\lambda(t, m)$.

Un'ulteriore caratteristica dei processi di Poisson è che il numero di punti presente in un intervallo è indipendente dal numero di punti presente negli altri intervalli dell'insieme, e questo li rende ancora più adatti a modellare dati disposti casualmente su due o più dimensioni.

In definitiva, per qualunque intervallo A di $\mathcal{A}$, abbiamo:

$$N(A) \sim Pois(\Lambda(A))$$

con $\Lambda(A)$ misura di intensità del processo:

$$E[N(A)] = \Lambda(A) = \int_A \lambda(t, m)\, dt\, dm.$$

La misura di intensità $\Lambda(A)$ dunque permette che il numero medio di punti cambi nelle diverse regioni dello spazio, e questa diversità è accettabile per il fatto che i numeri dei punti presenti nelle regioni sono indipendenti gli uni dagli altri. Si suppone infatti che i dati osservati

siano stati generati dalla serie di variabili $X_1, ..., X_n$ che si ipotizzano essere indipendenti e identicamente distribuite tra loro. Inoltre, poichè si lavora nell'ambito della modellazione di valori estremi al di sopra di determinati valori di soglia, si suppone che le X_i si comportino bene da un punto di vista dei valori estremi, cioè che dato $M_n = max\{X_1, ..., X_n\}$, ci siano una serie di costanti $\{a_n\}$ e $\{b_n\}$ tali che

$$Pr\left[\frac{(M_n - b_n)}{a_n} \leq z\right] \rightarrow G(z)$$

con

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

che corrisponde alla classica famiglia delle distribuzioni per valori estremi generalizzati (GEV). Il Processo di Poisson, sotto condizioni molto generali sulla distribuzione dei dati è quindi uno dei metodi più eleganti per modellare i valori estremi. In particolare, dato un valore di soglia u sufficientemente grande, è ragionevole modellare i dati della regione $\mathcal{A}_u = \{(t, m) : 0 < t < 1, m > u\}$ con un processo di Poisson non omogeneo avente una funzione di intensità appartenente alla famiglia:

$$\lambda(t, m) = \frac{1}{\sigma}\left[1 + \xi\frac{(m-\mu)}{\sigma}\right]_+^{\frac{-1}{\xi}-1}, \sigma > 0$$

dove '+' è definito come segue:

$$a_+ = \begin{cases} a & \text{se } a > 0 \\ 0 & \text{altrimenti} \end{cases}.$$

Da ciò deriva che

$$E[N(\mathcal{A}_u)] = \int_{\mathcal{A}_u} \lambda(t, m)\, dt\, dm = \left[1 + \xi\frac{(u-\mu)}{\sigma}\right]_+^{-1/\xi}$$

e che quindi la modellazione del numero di valori estremi al di sopra di una determinata soglia nella regione è descritta da i tre parametri μ, σ e ξ. A livello asintotico la modellazione tramite processi di Poisson

porta comunque a dover fare inferenza sui tre parametri che, così come nei modelli GEV, descrivono rispettivamente la posizione (μ), la scala (σ) e la forma della coda (ξ) della distribuzione. Il parametro ξ in particolare controlla la velocità con cui la coda della distribuzione va (eventualmente) a zero. Se $\xi < 0$ la distribuzione è limitata da $\mu - \sigma/\xi$, nel caso in cui $\xi = 0$ la distribuzione va a zero come se fosse un'esponenziale, mentre se $\xi > 0$ la coda è più pesante di quella di un'esponenziale.

I parametri μ, σ e ξ sono quindi il fulcro dell'inferenza, che viene fatta attraverso il metodo della verosimiglianza sulla base delle osservazioni fatte su $\mathcal{A}_u$. Si ricorda ancora una volta che perché sia possibile utilizzare i metodi della massima verosimiglianza per un metodo di Poisson devono valere le assunzioni di indipendenza e di stazionarietà sulle X_i specificate in precedenza.

Rimane ora il problema di calcolare la funzione di vermosimiglianza per un processo di Poisson nello spazio bidimensionale A_u, ed essa viene dimostrata essere (Coles, 2001):

$$L(\mu, \sigma, \xi; x_1, ..., x_n) = \exp\{-\Lambda(\mathcal{A}_u; \mu, \sigma, \xi)\} \prod_{i=1}^{n} \lambda(x_i; \mu, \sigma, \xi).$$

La funzione di verosimiglianza viene quindi massimizzata per i tre parametri in maniera numerica.

L'ulteriore dato da considerare nel calcolo della verosimiglianza è u, il valore di soglia oltre il quale i dati vengono considerati estremi. La scelta del valore di soglia è infatti una parte centrale della modellazione: un valore di soglia troppo alto sarà superato da poche osservazioni e darebbe luogo a stime poco distorte ma ad una grande variabilità campionaria; se il valore scelto è invece troppo basso le stime saranno generalmente distorte, ma la variabilità campionaria sarà tenuta sotto controllo.

In definitiva descrivere i dati, i punti che possiamo tracciare sul grafico, come la realizzazione di un processo di Poisson permette di sfruttare le caratteristiche asintotiche di questo, e di poter costruire

modelli statistici che permettano di fare inferenza sui parametri che regolano questo processo, che viene assunto stazionario.

Una volta effettuate le stime, esse, per quanto ottenute attraverso un metodo che è in grado di descrivere il processo che ha generato i dati, sono di difficile interpretazione. Generalmente si ricorre quindi ad una riparemetrizzazione e, sfruttando il fatto che il processo di punto è stato descritto su base annuale, viene calcolato η un tasso annuale di volte in cui il valore di soglia u viene superato.

$$\eta = \int_{m=u}^{\infty} \int_{t_0}^{t=t_n} \lambda(m,t)\, dm\, dt = \left[1 + \xi \frac{(u-\mu)}{\sigma}\right]_+^{-\frac{1}{\xi}-1}.$$

Inoltre, concentrando sucessivamente l'attenzione sulla distribuzione delle magnitudo registrate al di sopra della soglia u, viene stimata la funzione di ritorno, che mi dice ogni quanti anni si stima che un determinato valore x di magnitudo venga superato. Poichè, per un qualunque x maggiore di u:

$$P(X > x | X > u) = \left[1 + \xi \frac{(x-\mu)}{\tilde{\sigma}}\right]_+^{\frac{1}{\xi}}, \tilde{\sigma} = \sigma + \xi(u-\mu).$$

si può definire con *r(x)* la funzione di ritorno, che mi dice ogni quanti anni ci si aspetta che il livello $x > u$ sia superato:

$$r(x) = \eta^{-1} \left[1 + \xi \frac{(x-\mu)}{\tilde{\sigma}}\right]_+^{\frac{1}{\xi}}$$

E' infine possibile fare un grafico della funzione di ritorno, e generalmente in ascisse troviamo il logaritmo degli anni che rende più agevole la lettura del grafico.

I dati riguardanti i sismi in Italia permettono di presentare un'applicazione di queste metodiche.

4.2 I Processi di Punto per valori estremi applicati ai dati sui terremoti in Italia

4.2.1 Prime scelte per l'analisi dei dati

Come già preannunciato verrà condotta un'analisi attraverso i normali processi di punto basata solamente su una parte dei dati, poichè il dataset presenta delle diversità nella registrazione dei dati nel tempo, che vanno a minare le ipotesi di stazionarietà sotto cui vale la teoria esposta in 4.1. Si procederà quindi ad una selezione di un sottodataset in cui si possa ipotizzare che i dati sono stati registrati con probabilità uniforme, e in cui sussista l'ipotesi di stazionarietà. I dati sono ipotizzati derivare da $X_1, ..., X_n$ indipendenti e identicamente distribuite.

La prima domanda che ci si pone per analizzare i dati attraverso metodi di processi di punto è quindi: che parte utilizzare? Che anno prendere per dividere i dati in un sottodataset che possa essere analizzato attraverso processi di punto?

Per poter utilizzare i processi di punto i dati devono essere stazionari, e in particolar modo devono essere stazionari i valori estremi. Inoltre è auspicabile aver osservato un numero più o meno uniforme di avvenimenti negli anni.

Per verificare in maniera sommaria la stazionarietà del processo andiamo a osservare i grafici riguardanti i terremoti degli ultimi anni, e l'andamento di medie e varianze dei dati osservati in quel periodo.

Presentiamo per gli anni dopo il 1770 il grafico della Magnitudo momento misurata negli anni, il numero di eventi registrati, l'andamento delle medie e l'andamento della varianza. Questi ultimi due grafici sono ottenuti andando a calcolare le medie e le varianze di finestre temporali quinquennali che si sovrappongono per un anno con la finestra precedente. In ascissa al grafico troviamo quindi il numero di finestra temporale, ed ogni punto nel grafico corrisponde alla media e alla varianza calcolata negli anni corrispondenti a quella finestra temporale. Il punto nel grafico 4.3 che corrisponde alla seconda finestra

temporale sarà quindi la media della Magnitudo momento calcolata per gli anni tra il il 1774 e il 1780, quello corrispondente alla terza finestra temporale la media calcolata negli anni tra il 1779 e il 1785, e così via.

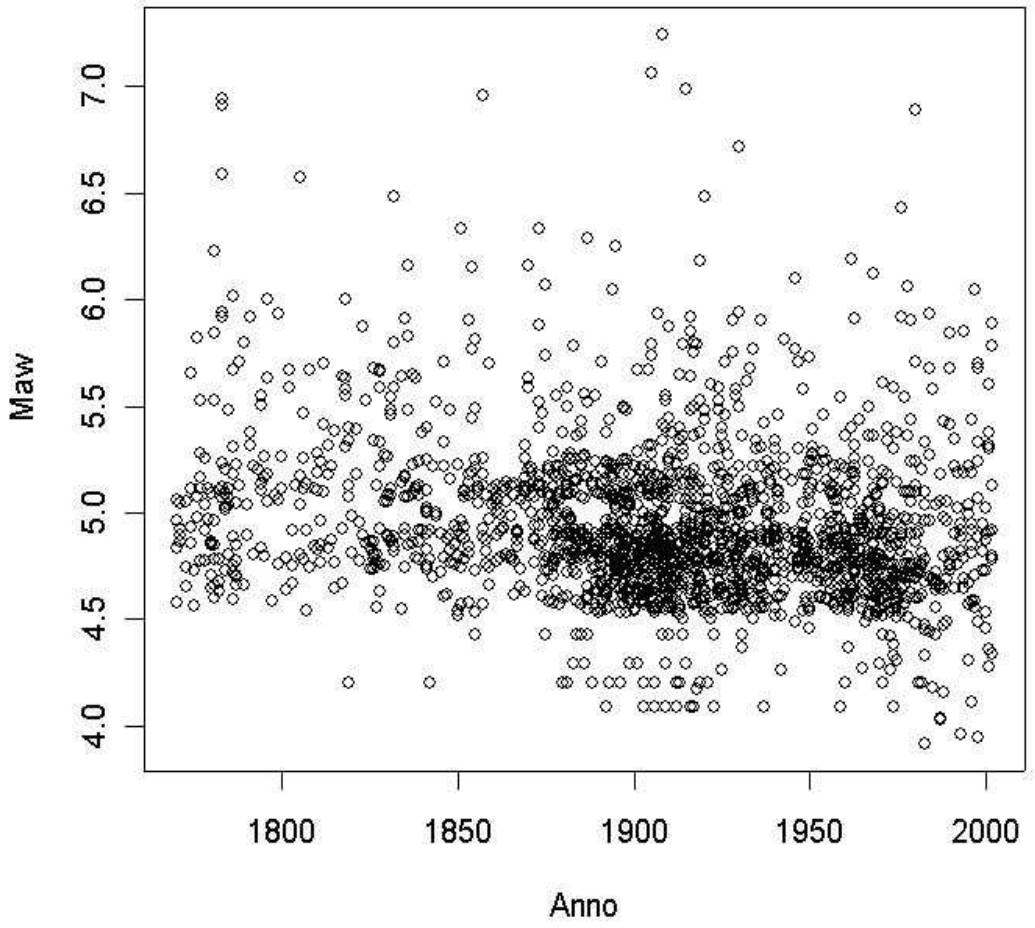


Figura 4.1: La variabile Maw per anno di avvenimento del sisma dopo il 1770

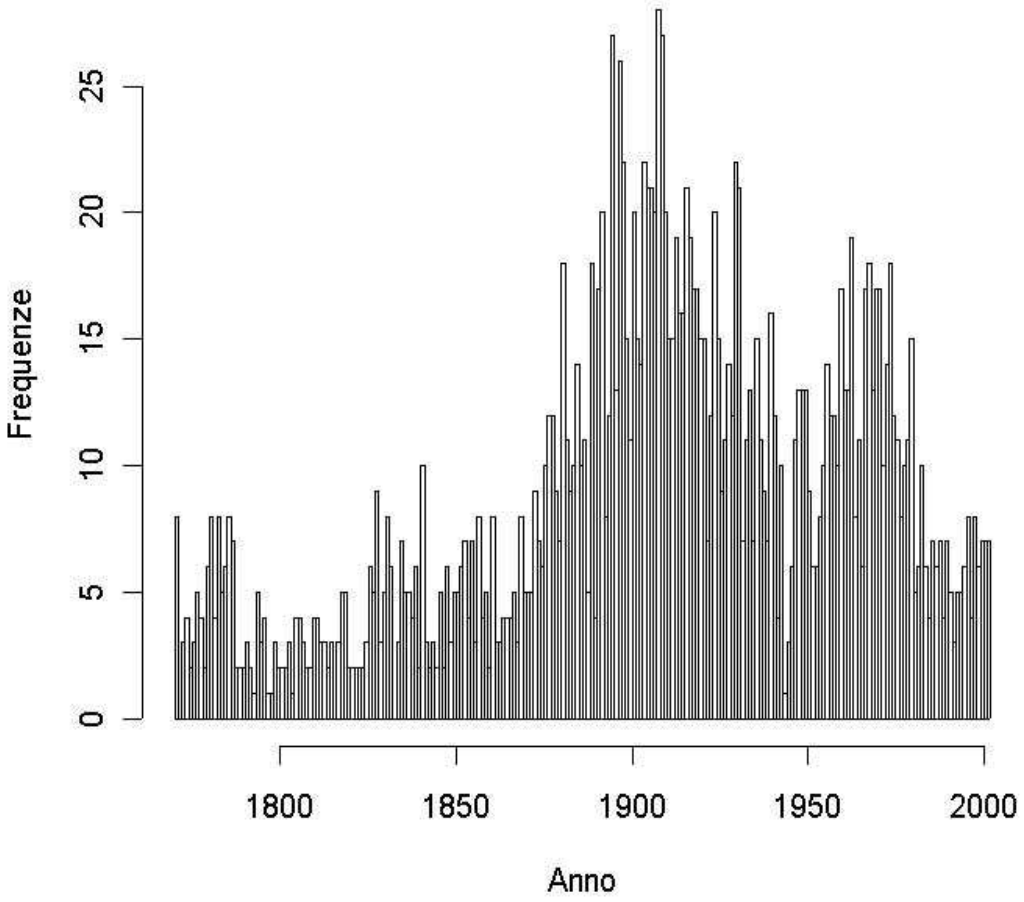


Figura 4.2: Numero di terremoti avvenuti dopo il 1770

Il grafico sembra presentare una certa forma di ciclicità, come se il numero di terremoti nel corso del tempo, variasse in base ad una forma che rimane costante. Si vede infatti che il numero di terremoti, cambiando nel tempo tende ad avere una forma "ad onda". Sale il numero di terremoti, che poi scende ancora, e risale fino ad un nuovo livello. Questo fenomeno è stato osservato in diversi dataset di sismi, per diverse parti del pianeta, e le motivazioni potrebbero essere legate al fenomeno delle macchie solari, sebbene questa teoria non sia stata

confermata. Questo porta ad avere un diverso numero di osservazioni nei diversi periodi, periodi con molti terremoti registrati tendono ad alternarsi a periodi con meno sismi.

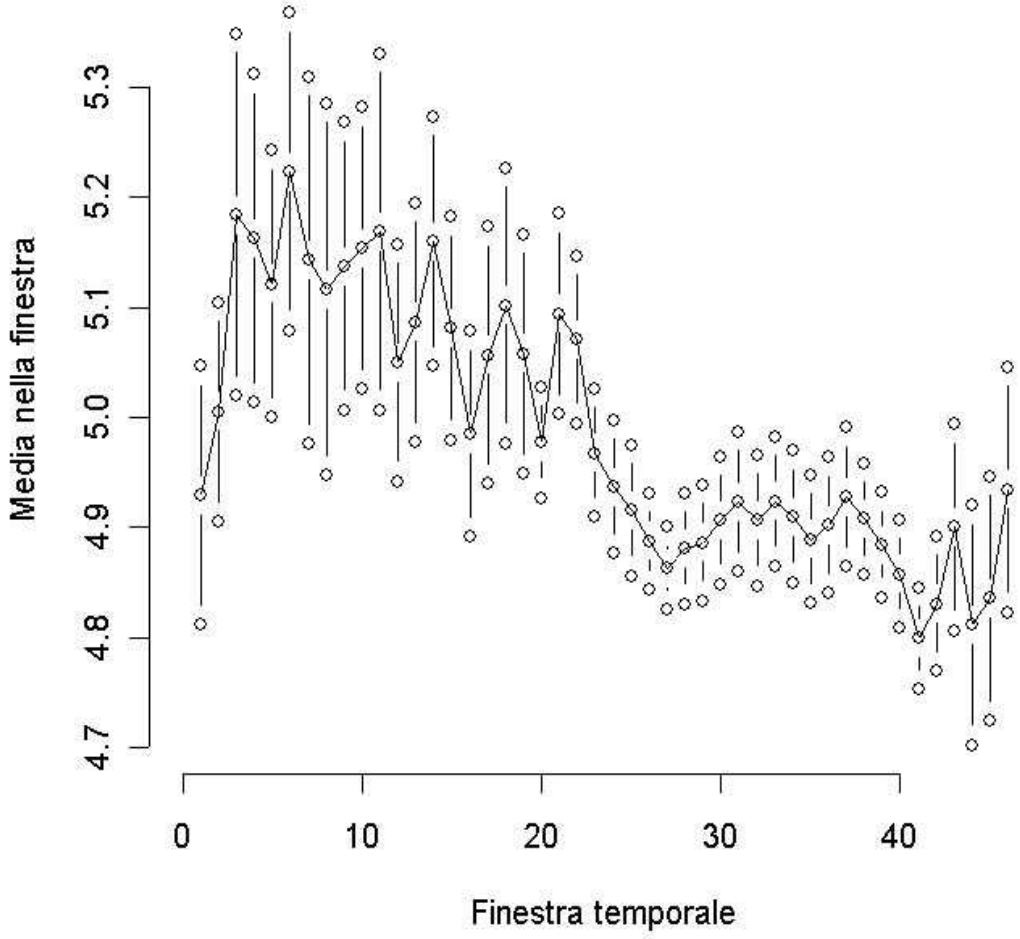


Figura 4.3: Andamento della media della variabile Maw dopo il 1770

La media sembra stabilizzarsi attorno agli anni 1880-1890, corrispondenti alle finestre temporali tra la 22 e la 23. Si noti inoltre come il campo di variazione delle medie sia comunque molto più stretto di quello della variabile stessa.

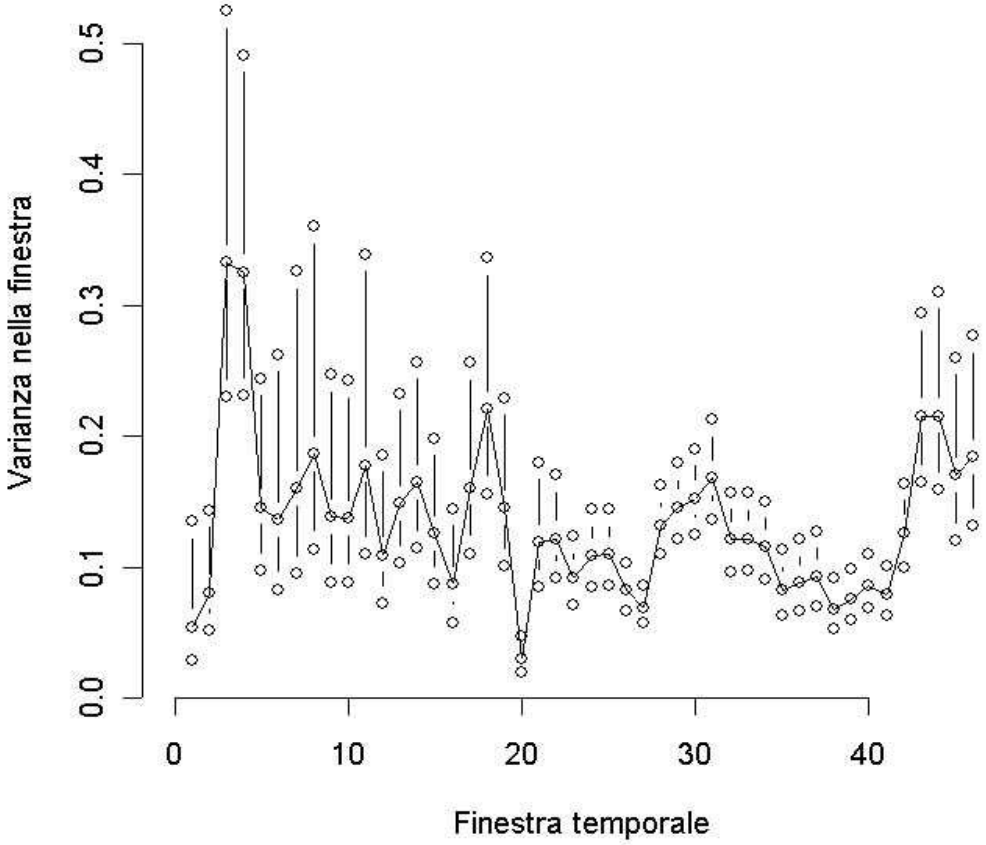


Figura 4.4: Andamento della varianza della variabile Maw dopo il 1770

La varianza sembra stabilizzarsi attorno all'anno 1875 corrispondente alla 21° finestra temporale.

Sembra quindi sensato focalizzarsi sul sottodataset che parte dall'anno 1885, in modo da avere medie e varianze piuttosto stabili nel tempo. Il numero di osservazioni inoltre è piuttosto alto: 1449 terremoti sono stati osservati dal 1885 in poi. Il grafico (figura 4.5) del sottodataset ora mostra un buon numero di osservazioni, relativamente sparpagliate nel tempo.

La variabile Maw nel sottodataset varia tra 3.920 (anno 1983) e

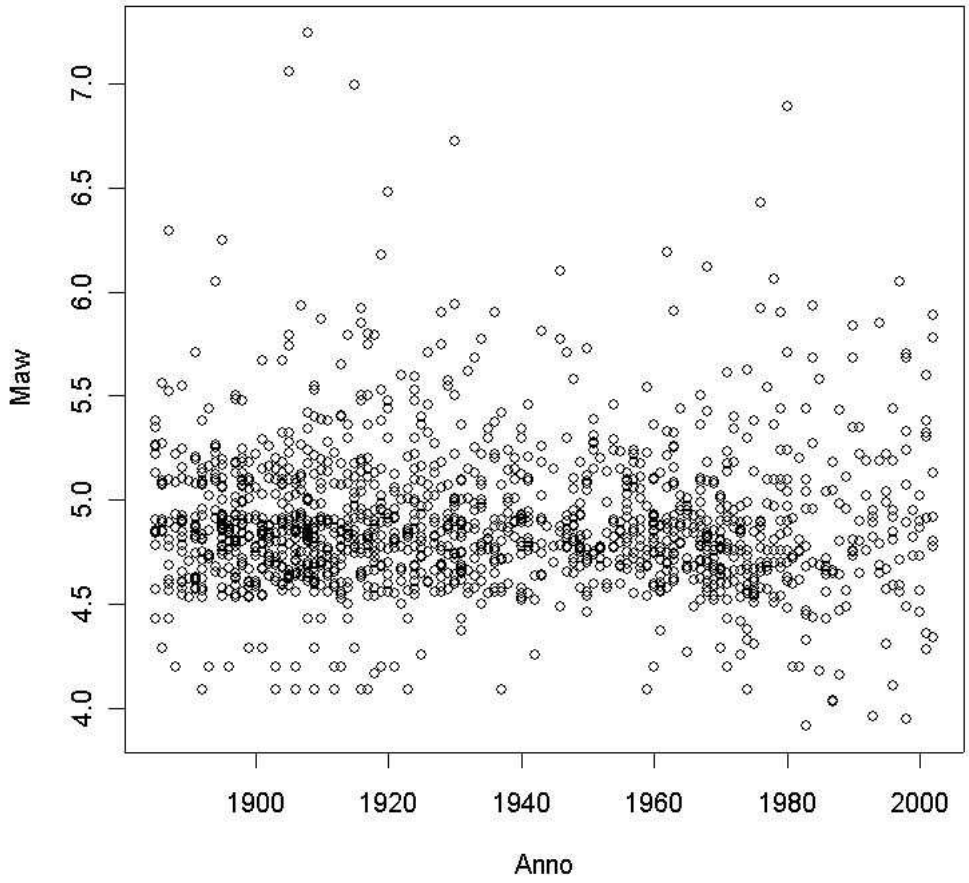


Figura 4.5: La variabile Maw per anno di avvenimento del sisma dopo il 1885

7.240 (anno 1908). è stata quindi esclusa dall'analisi l'osservazione più alta pari a 7.41, registrata nel 1693. La media di Maw calcolata dopo il 1885 è pari a 4.889 e la mediana è pari a 4.841; entrambi i valori sono più bassi a quelli calcolati su tutta l'arco temporale (5 il valore della media, 4.898 il valore della mediana) a ulteriore dimostrazione del fatto che nel passato vengono registrati valori più alti, e che per lo meno negli anni più recenti vengono registrati anche terremoti di entità meno forte.

Negli anni successivi al 1885 inoltre i valori estremi sembrano non

discostarsi dall'ipotesi di stazionarietà, come indicano i grafici di figura 4.6 in cui vengono mostrati i valori al di sopra di determinate soglie, con indicato il numero di osservazioni eccedenti la soglia stessa.

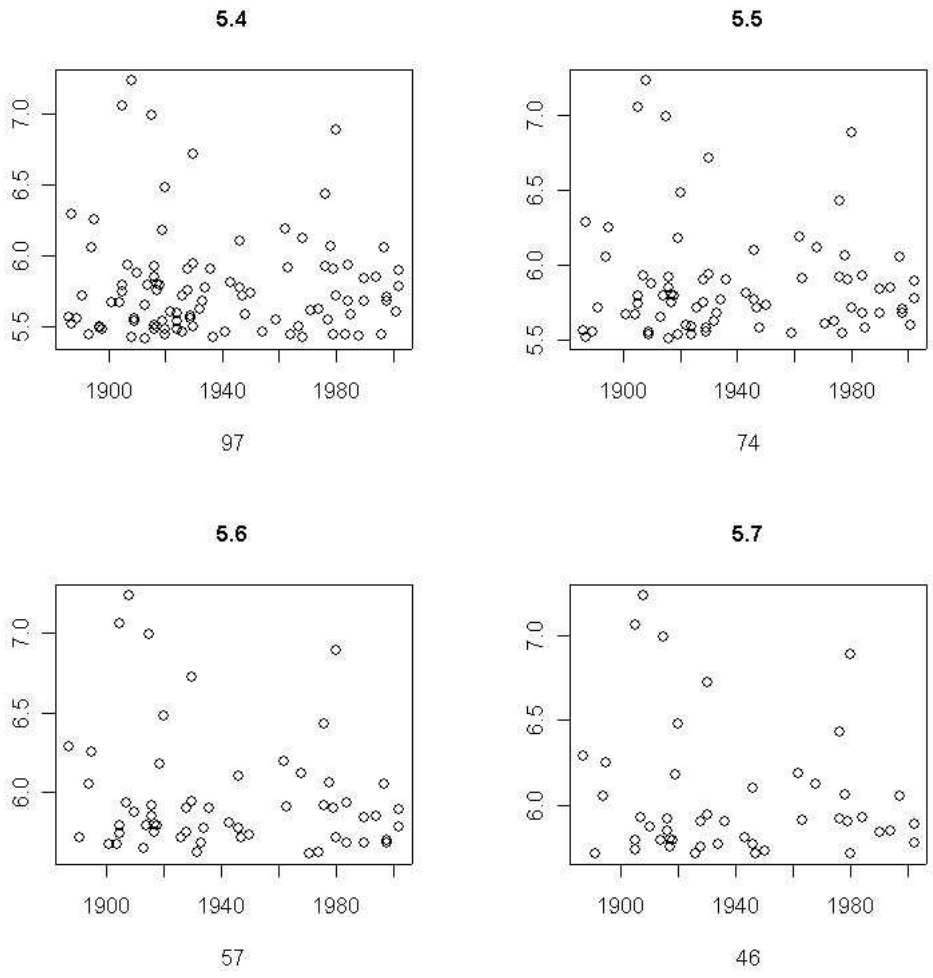


Figura 4.6: Osservazioni al di sopra della soglia indicata con il numero di osservazioni

Da notare che, così come la numerosità dei sismi registrati subisce un rallentamento negli anni tra gli anni '30 e gli anni '50, anche per i valori al di sopra di una certa soglia si nota che essi avvengono in particolar modo all'inizio del '900 e alla fine del secolo.

Si noti come diminuisce il numero di valori estremi a seconda della soglia: il problema di decidere quale sia infine il valore di soglia per

cui un terremoto venga considerato un valore estremo è l'ulteriore punto su cui soffermarsi. Si presenta il problema del bilanciamento tra variabilità campionaria e distorsione. Per scegliere quale sia un buon valore di soglia, che dia buone stime per i parametri di posizione μ, scala σ e forma ξ si confronta l'andamento delle stime dei parametri in relazione al livello di soglia adottato. Questo tipo di analisi non è purtroppo preciso, ma si basa più sull'intuito e la sensibilità.

Utilizzando la funzione pp.fitrange della libreria ismev otteniamo i grafici dell'andamento delle stime di massima verosimiglianza dei parametri a seconda dei valori di soglia (figura 4.7).

L'idea alla base di queste figure è che se un modello è valido per un valore di soglia u, esso dovrà essere valido anche per un valore $u^* > u$, quindi il grafico delle stime dei parametri σ e ξ per diversi valori di soglia dovrebbe variare attorno agli stessi valori, ed essere cioè di forma lineare, a meno di errori dovuti al campionamento. Al variare dei valori di u quindi le stime dei parametri dovrebbero rimanere stabili. Un valore attorno a 5.8 sembra dare buoni risultati per μ e σ, sebbene le stime per il parametro ξ non siano particolarmente soddisfacenti per nessun valore di soglia: le stime oscillano attorno allo 0 e gli intervalli di confidenza sono molto ampi.

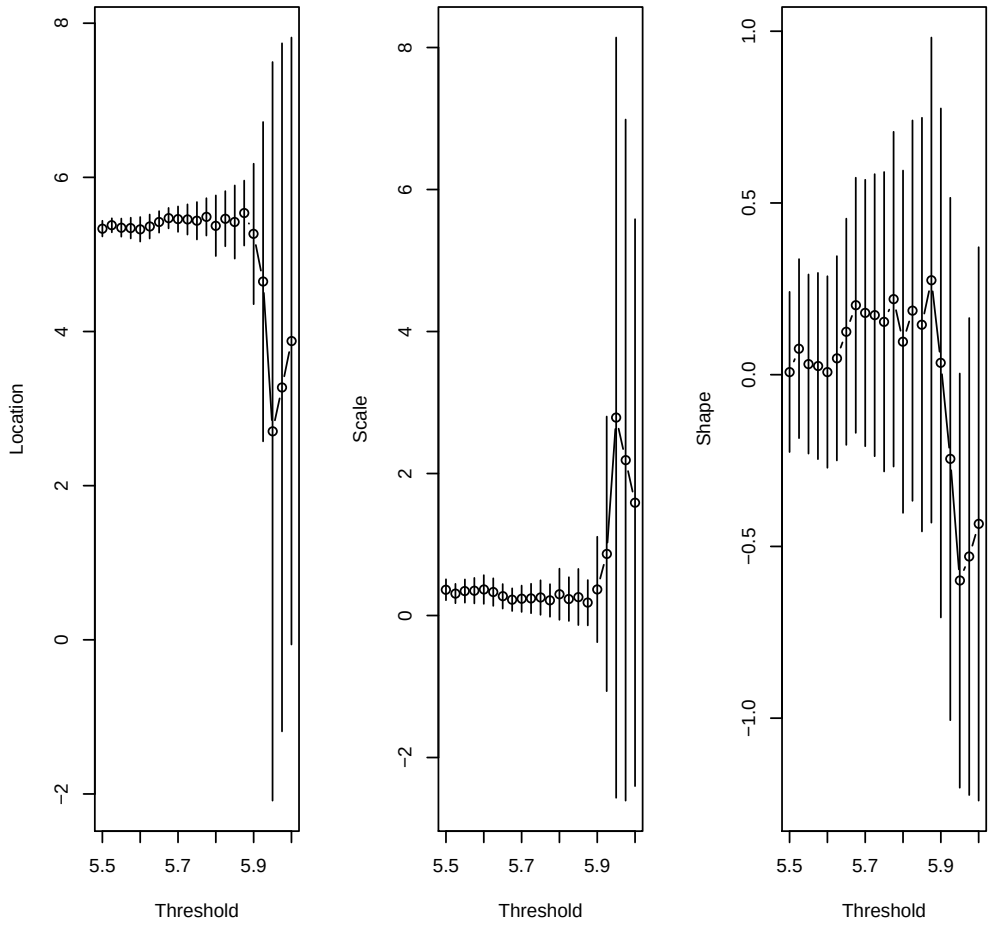


Figura 4.7: I valori stimati dei parametri del modello al variare del valore di soglia

Un altro strumento utile alla selezione del valore di soglia è il Mean Residual Life Plot, un metodo esplorativo non basato sulla stima del modello. L'idea alla base di questo grafico è legata alla media di funzione generalizzata di Pareto, che viene utilizzata per modellare dati al di sopra di un determinato valore di soglia. Si può dimostrare come la modellazione attraverso i processi di punto e attraverso la funzione generalizzata di Pareto siano similari, tuttavia questo esula dal nostro scopo, e si rimanda alla letteratura per questo (Coles, 2001). La media di una variabile Y che si distribuisce come una variabile generalizzata di Pareto, per $\xi < 0$ è:

$$E[Y] = \frac{\sigma}{1-\xi}.$$

Quindi, se il modello di Pareto è valido per i dati al di sopra del valore di soglia u la media delle osservazioni $X > u$ sarà, per $\xi < 0$:

$$E[X - u|X > u] = \frac{\sigma_u}{1-\xi}.$$

Tuttavia se il modello è valido per i dati al di sopra della soglia u, esso deve esser valido anche per i dati al di sopra di un livello $u^* > u$, per cui la media sarà:

$$E[X - u^*|X > u^*] = \frac{\sigma_{u^*}}{1-\xi} = \frac{\sigma_u + \xi u^*}{1-\xi},$$

per le proprietà della funzione generalizzata di Pareto. $E[X - u^*|X > u^*]$ è quindi una funzione lineare in u^*, e può essere stimata empiricamentre dalla media campionaria dell'eccesso rispetto al valore di soglia. Ci si aspetta quindi che, al variare di u, per valori per cui il modello generalizzato di Pareto è valido e per cui $\xi < 0$, le stime di $E[X - u^*|X > u^*]$ cambino in maniera lineare. Il luogo dei punti

$$\left\{\left(u, \frac{1}{n}\sum\right)\right\} E[X - u^*|X > u^*] = \frac{\sigma_{u^*}}{1-\xi} = \frac{\sigma_u + \xi u^*}{1-\xi},$$

è appunto il Mean Residual Life Plot, e ci si aspetta che dopo un certo valore di u il grafico diventi lineare.

Osservando il Mean Residual Life Plot di figura 4.8 tuttavia non si vede un punto in cui la forma del grafico si fa decisamente lineare, tuttavia si nota che il valore 5.3 sembra essere un punto in cui la forma del grafico inzia a cambiare in maniera sensibile.

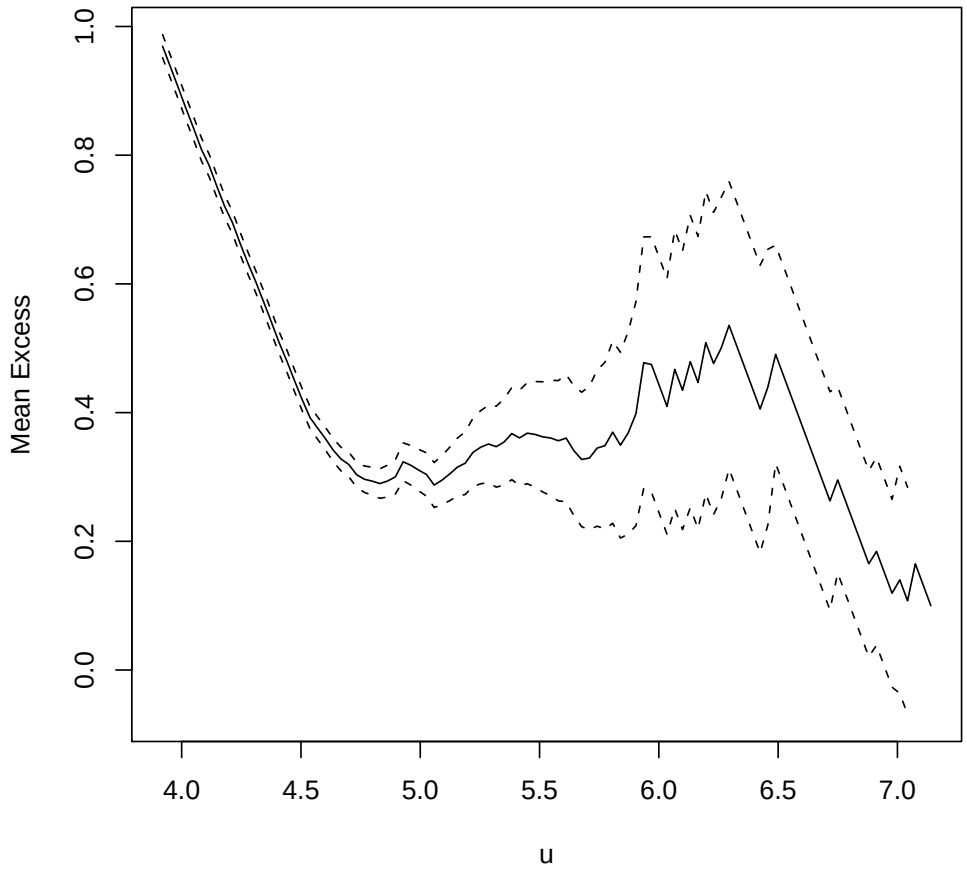


Figura 4.8: Mean Residual Life Plot della variabile Maw dopo il 1885

Per questa difficoltà di scelta del valore di soglia da considerare, saranno presentati i risultati sia per il valore 5.3 che per il valore 5.8, il primo individua una parte consistente dei dati, mentre il secondo separa più fortemente valori estremi.

4.2.2 Stima dei modelli attraverso processi di punto

Di seguito vengono presentate le tabelle con le stime di massima verosimiglianza dei parametri per entrambi i valori di soglia calcolate secondo le procedure descritte in precedenza (4.1).

	n. eccedenze	μ(s.e.)	σ(s.e.)	ξ(s.e.)
u=5.3	133	5.34(0.034)	0.35(0.045)	0.018(0.082)
u=5.8	31	5.37(0.200)	0.29(0.183)	0.096(0.254)

Tabella 4.1: Stime di massima verosimiglianza per il modello stimato dopo il 1885

I due valori identificano parti molto diverse dello spazio, 133 valori sono al di sopra del livello 5.3 (9% delle osservazioni dal 1885 in poi), mentre sono solo 31 i terremoti per cui è stato registrato un valore di Magnitudo momento maggiore di 5.8 (2% delle osservazioni nel sottodatset considerato). Inoltre i valori stimati tramite il valore di soglia pari a 5.8 sono meno precisi come dimostrano gli standard error, che sono ben più grandi di quelli stimati con un valore di soglia pari a 5.3.

Da notare come in entrambi i casi il limite inferiore di un intervallo di confidenza per ξ calcolato come $\hat{\xi} - 2 * se(\xi)$ va ad individuare valori negativi di ξ che indicano che la funzione generatrice dei dati è limitata a sinistra dal valore $\mu - \sigma/\xi$.

Tramite le stime di massima verosimiglianza andiamo a valutare il valore del parametro η che stima il tasso annuale di volte in cui il valore di soglia u viene superato. Per il valore di soglia 5.3 $\hat{\eta}$ è pari a 1.12, mentre per il modello stimato con un livello di soglia pari a 5.8 $\hat{\eta}$ è stimato essere 0.26.

Inoltre vengono disegnate delle curve di ritorno su scala logaritmica per entrambi i valori di soglia. Le curve di ritorno, come già detto, sono una previsione di quanti anni sarà necessario aspettare perchè un terremoto di una data magnitudo m si ripeta. Si presentano prima i grafici delle curve di ritorno per entrambe le soglie.

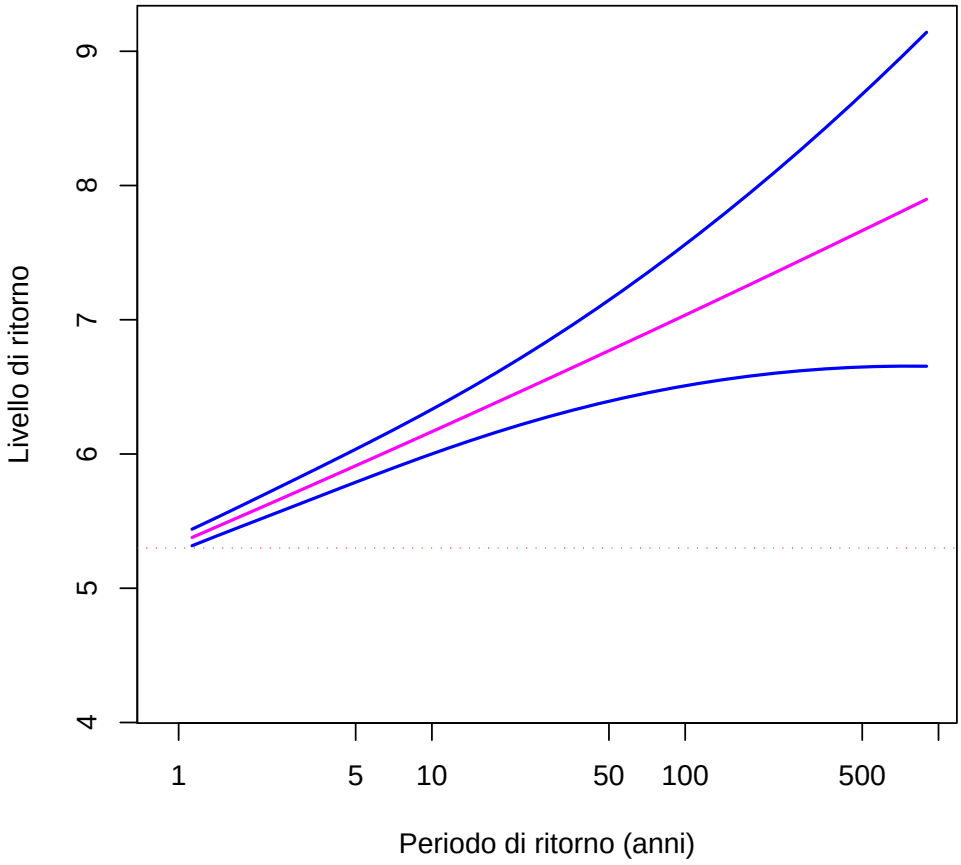


Figura 4.9: Curva di ritorno ed intervalli di confidenza al 95%, soglia pari a 5.3

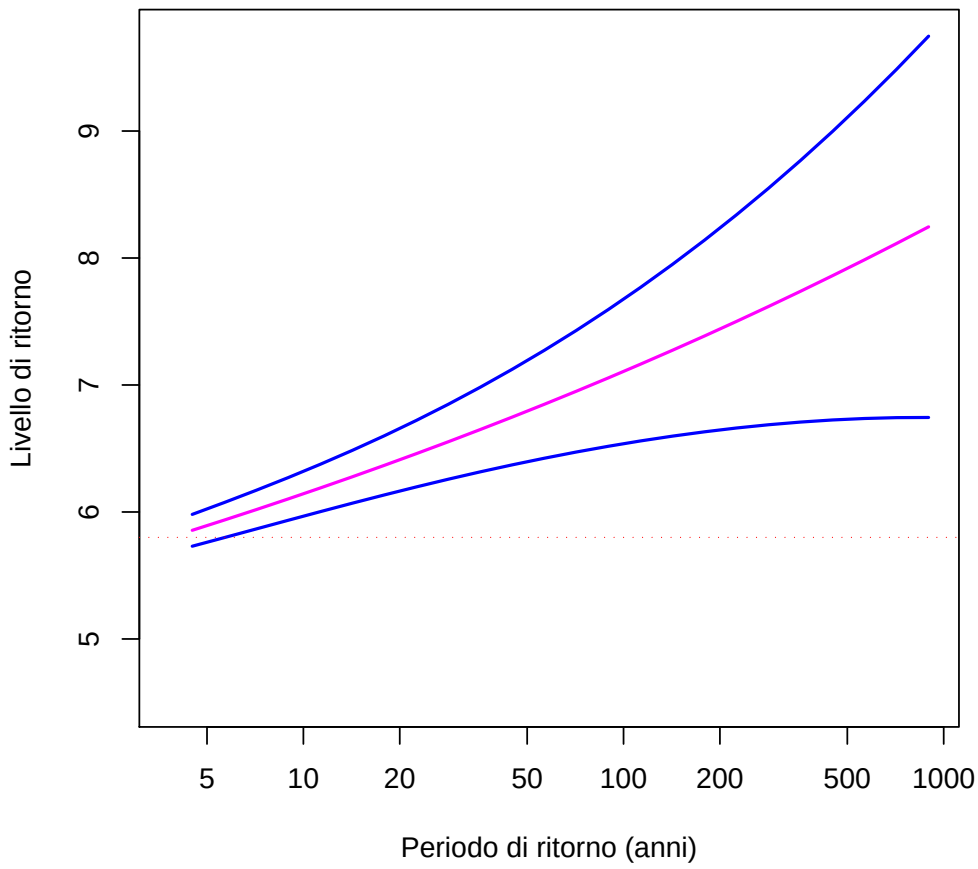


Figura 4.10: Curva di ritorno ed intervalli di confidenza al 95%, soglia pari a 5.8

Si nota subito che la forma stimata della funzione di ritorno per entrambi i valori di soglia è pressochè lineare. Questo è duvuto al fatto che il parametro ξ è stimato positivo e molto prossimo allo zero: la forma stimata per la curva è non limitata a destra, e pressochè esponenziale nella coda. Questo spiega anche il perchè del divergere degli intervalli di confidenza: poichè la funzione è non limitata anche valori molto alti di magnitudo non possono essere considerati come un evento impossibile. Da notare però come sia maggiore la variabilità nel caso si stimi il modello con un valore di soglia pari a 5.8: gli intervalli sono molto più ampi, mentre nel caso del modello calcolato con una soglia pari a 5.3 gli intervalli sono inizialmente molto stretti e divergono in maniera meno veloce.

Parte II

Processi di Punto con censura per valori estremi: un esempio con tutti i dati sui terremoti italiani

Capitolo 5

Processi di punto con dati censurati

La struttura dei processi punto esposta nel capitolo 4.1 è valida, come detto, per processi che siano stazionari, i cui momenti, come media e varianza, non dipendano cioè dal tempo. L'ipotesi di stazionarità tuttavia è palesemente non riscontrata nei dati riguardanti i terremoti in Italia negli ultimi 2300 anni, e per questo nel capitolo precendente si era selezionato un sotto dataset in cui l'ipotesi di stazionarietà potesse valere. Le cause della non stazionarietà dei dati sono però da ricercare nel modo di raccolta dei dati e non nel processo che li ha generati. Non è cioè credibile che siano i processi sismogenetici ad esseree non stazionari: essi rimangono immutati nel tempo (specie in un arco di anni lungo appena 2300 anni), mentre è cambiata nella storia la capacità dell'uomo di misurare i sismi. Si ipotizza quindi la presenza di un processo stazionario in cui è avvenuto nel tempo un processo di censura dei dati; in assenza di tale processo di censura i dati potrebbero essere modellati tramite processi di punto in maniera del tutto similare a quanto fatto nella sezione 4.2. Sarà quindi necessario modellare anche il processo di censura che ha selezionato i dati nel tempo. In definitiva il processo sismogentico degli ultimi 2300 anni viene modellato attraverso un processo di Poisson, avente una funzione di intensità $\lambda(m, t)$ che è la funzione di intensità relativa al

vero processo sottostante i terremoti degli ultimi 2300 anni, ma questa funzione verrà moltiplicata per una funzione $p(m,t)$, che descriva il processo di censura a cui i dati sono stati di fatto sottoposti. I dati che sono giunti fino a noi quindi dovranno essere modellati tramite una funzione $\lambda_M(m,t) = \lambda(m,t)p(m,t)$, in cui $\lambda(m,t)$ descrive il vero processo sottostante ai terremoti in Italia, e $p(m,t)$ descrive l'effettivo processo di censura avvenuto nel tempo sui dati.

5.1 Un'analisi analoga a quella dei vulcani

Una possibile proposta di funzione che soddisfi i vincoli sopra esposti, e che permetta di considerare una vasta gamma di possibili comportamenti è quella proposta per lo studio dei fenomeni vulcanici da Coles e Sparks (2004):

$$p(m,t) = \left(1 - \frac{v}{x^w}\right) + \frac{v}{x^w}t^b,$$

in cui per i parametri (v, w, b) vengono posti i seguenti vincoli: $b \geq 0$, $w \geq 0$ e $v \leq u^w$. In particolare con questa specificazione i parametri acquistano un significato particolare: v determina la misura nella quale gli eventi sono censurati nel tempo (se $v = 0$ vuol dire che non c'è stata una censura nel tempo), w determina la misura in cui i dati sono censurati a seconda dei livelli di magnitudo ($w = 0$ implica che la censura sia avvenuta in misura costante per i diversi livelli) e b determina il tasso di cambiamento nella censura (b=1, ad esempio, indicherebbe un cambiamento lineare nel tempo del tasso di censura).

In particolare, avendo riscalato il tempo su un intervallo $[0, 1]$, la funzione $p(m,t)$ soddisfa i seguenti vincoli:

- $p(m,t)$ è funzione crescente in t per ogni m fissato; per un determinato valore di Magnitudo è più probabile che vengano registrati gli eventi più recenti piuttosto che quelli avvenuti nel passato più remoto.

- $p(m,t)$ è, per ogni t fissato, funzione non descrescente per m; qualunque sia il momento t in cui avviene il terremoto, è più difficile che sfuggano alla registrazione eventi caratterizzati da livelli più alti di Magnitudo.
- $p(m,1) = 1$ per qualunque m; se avvenisse cioè un evento nel presente esso verrebbe registrato con probabilità 1, qualunque sia la sua Magnitudo.

La funzione di verosimiglianza da massimizzare risulta quindi essere:

$$
\begin{aligned}
&L(\mathbf{z}, \mathbf{t}; \mu, \sigma, \xi, v, w, b) = \\
&\quad = \exp\left\{ - \int_{\mathcal{A}_u} \lambda(m,t) p(m,t)\, dm\, dt \right\} \prod_{i=1}^{n} \lambda(m,t) p(m,t) = \\
&\quad = \exp\left\{ - \int_{\mathcal{A}_u} \frac{1}{\sigma} \left[1 + \xi \frac{(m-\mu)}{\sigma} \right]_+^{-\frac{1}{\xi}-1} \left[\left(1 - \frac{v}{m^w} \right) + \frac{v}{m^w} t^b \right] dm\, dt \right\} \\
&\qquad \prod_{i=1}^{n} \left\{ \frac{1}{\sigma} \left[1 + \xi \frac{(m_i-\mu)}{\sigma} \right]_+^{-\frac{1}{\xi}-1} \left[\left(1 - \frac{v}{m_i^w} \right) + \frac{v}{m_i^w} t_i^b \right] \right\}
\end{aligned}
$$

L'uso di questa funzione per i dati riguardanti i terremoti si è rivelata però inefficace. Sono stati svolti molti tentativi, ma non si è riusciti a massimizzare la funzione di verosimiglianza. Vengono qui mostrati alcuni dei risultati ottenuti nel tentativo di cercare valori che minimizzassero la log-verosimiglianza negativa per il modello calcolato con un valori di soglia pari a 5.8, tenendo fisso di volta in volta uno dei parametri *(v,w,b)*.

Lasciando invece liberi di variare tutti i parametri si ottengono valori similari a quelli mostrati in tabella 5.2

μ	σ	ξ	v	w	b	-log-lik
4.870	0.65	-0.190	1	0.01	2	493.903
4.792	0.670	-0.189	1	0.05	2	493.550
4.757	0.675	-0.188	1	0.07	2	496.078
4.710	0.682	-0.188	1	0.1	2	500.310
5.042	0.616	-0.186	1	0.07	4.9	483.346
5.177	0.595	-0.188	1	0.03	4.9	474.166
5.260	0.583	-0.190	1	0.01	4.9	476.053
5.246	0.581	-0.188	1.03	0.03	4.9	474.375
5.222	0.586	-0.188	1.02	0.03	4.9	473.500
5.199	0.591	-0.188	1.01	0.03	4.9	473.565
5.134	0.604	-0.188	0.98	0.03	4.9	476.237

Tabella 5.1: Valori della log-verosimiglianza negativa in corrispondenza del valore fissato di un parametro

μ	σ	ξ	v	w	b	-log-lik
5.49	0.59	-0.10	0.99	0.004	4.49	431.372
5.01	0.75	-0.26	0.96	6.8e-010	4.88	392.415
5.17	0.70	-0.26	0.96	2.8e-010	5.81	393.431
5.05	0.70	-0.23	0.96	6.1e-009	4.67	392.618
4.91	0.86	-0.31	0.96	2.8e-007	4.85	392.373

Tabella 5.2: Valori della log-verosimiglianza negativa: tentativo di minimizzazione

La funzione come si vede tende a cercare valori molto bassi per il parametro w, e dà risultati di volta in volta molto diversi tra loro, qualunque siano i valori iniziali che vengono sottoposti.

Il motivo di questa mancata convergenza è da cercare probabilmente più nei dati che nella funzione stessa. Perchè una funzione che ha dato dei risultati nello stimare il processo sottostante all'attività vulcanica non riesce a trovare delle stime per un processo sismico? La risposta è scritta nei dati, nella differenza che c'è tra i dati dei terremoti in Italia e i dati utilizzatio da Coles e Sparks. Nei dati dei sismi italiani infatti non sembra riscontarsi la condizione per cui i terremoti del passato vengono registrati solo se sono di forte entità. I terremoti di magnitudo maggiori sono stati registrati negli anni dopo il 1500, e gli eventi dell'antichità sebbene non siano di bassa entità non si caratterizzano per elevati valori di magnitudo momento. La probabilità che un evento venga registrato nel passato non sarebbe quindi fortemente dipendente dal valori di magnitudo, il che spiegherebbe la difficoltà computazionale nel trovare una stima per il parametro w.

5.2 Un'analisi attraverso una diversa funzione di censura

Avendo osservato come la stima dei parametri nel caso della funzione proposta nella precedente sezione sia difficile a causa di problemi computazionali, e avendo notato quali siano i problemi all'interno del dataset che non hanno reso possibile la convergenza della funzione, si è pensato di modellare la probabilità che un evento venga registrato ad un determinato tempo con una funzione diversa, che ponesse meno vincoli ai parametri. La funzione proposta è:

$$p(m,t) = \exp\left\{\frac{b(t-1)}{m^w}\right\}.$$

con $b \geq 0$ unico vincolo posto ai parametri, viene richiesto infatti che per tempi più recenti la probabilità di registrare un terremoto aumenti. In particolare, il parametro b indica la misura della censura nel tempo,

e il parametro w regola la censura avvenuta a causa della magnitudo dell'evento.

In particolare, avendo riscalato il tempo su un intervallo $[0, 1]$,la funzione ha le seguenti proprietà:

- $p(m, t)$ è funzione crescente in t per ogni m fissato: per un determinato valore di Magnitudo è più probabile che vengano registrati gli eventi più recenti piuttosto che quelli avvenuti nel passato più remoto.
- $p(m, 1) = 1$ per qualunque m; se avvenisse cioè un evento nel presente esso verrebbe registrato con probabilità 1, qualunque sia la sua Magnitudo.

La funzione di verosimiglianza verrà dunque ad essere:

$$
\begin{aligned}
&L(\mathbf{z}, \mathbf{t}; \mu, \sigma, \xi, w, b) = \\
&\quad = \exp\left\{ - \int_{\mathcal{A}_u} \lambda(m,t) p(m,t)\, dm\, dt \right\} \prod_{i=1}^{n} \lambda(m,t) p(m,t) = \\
&\quad = \exp\left\{ - \int_{\mathcal{A}_u} \frac{1}{\sigma} \left[1 + \frac{\xi(m-\mu)}{\sigma} \right]_+^{-\frac{1}{\xi}-1} \exp\left\{ \frac{b(t-1)}{x^w} \right\} dm\, dt \right\} \\
&\qquad \prod_{i=1}^{n} \frac{1}{\sigma} \left[1 + \frac{\xi(m-\mu)}{\sigma} \right]_+^{-\frac{1}{\xi}-1} \exp\left\{ \frac{b(t-1)}{x^w} \right\}
\end{aligned}
$$

Il calcolo numerico delle stime di massima verosimiglianza per questa equazione è possibile e i risultati trovati vengono mostrati nella tabella 5.3.

	μ	σ	ξ	w	b
u=5.3	5.39	0.37	-0.06	5.11	51445.36
u=5.8	5.13	0.60	-0.19	0.03	4.59

Tabella 5.3: Stime di massima verosimiglianza per il modello stimato su tutti i dati

Salta subito agli occhi come il valore di b e w cambino al variare del valore di soglia. Tuttavia, nel caso del modello stimato per

$u = 5.3$ l'altissimo valore di b viene compensato da un valore altrettanto grande di w. Il grande valore di b al numeratore dell'esponente infatti schiaccia la funzione verso lo 0, mentre il grande valore di w al denominatore, fa si che l'esponente non raggiunga valori troppo grandi in valore assoluto, e che la funzione possa quindi arrivare a valori prossimi ad uno. L'andamento della funzione è quindi il risultato di un delicato equilibrio tra i due parametri.

Le funzioni di probabilità stimate risulterebbero quindi essere quelle che si vedono nei grafici 5.1 e 5.2.

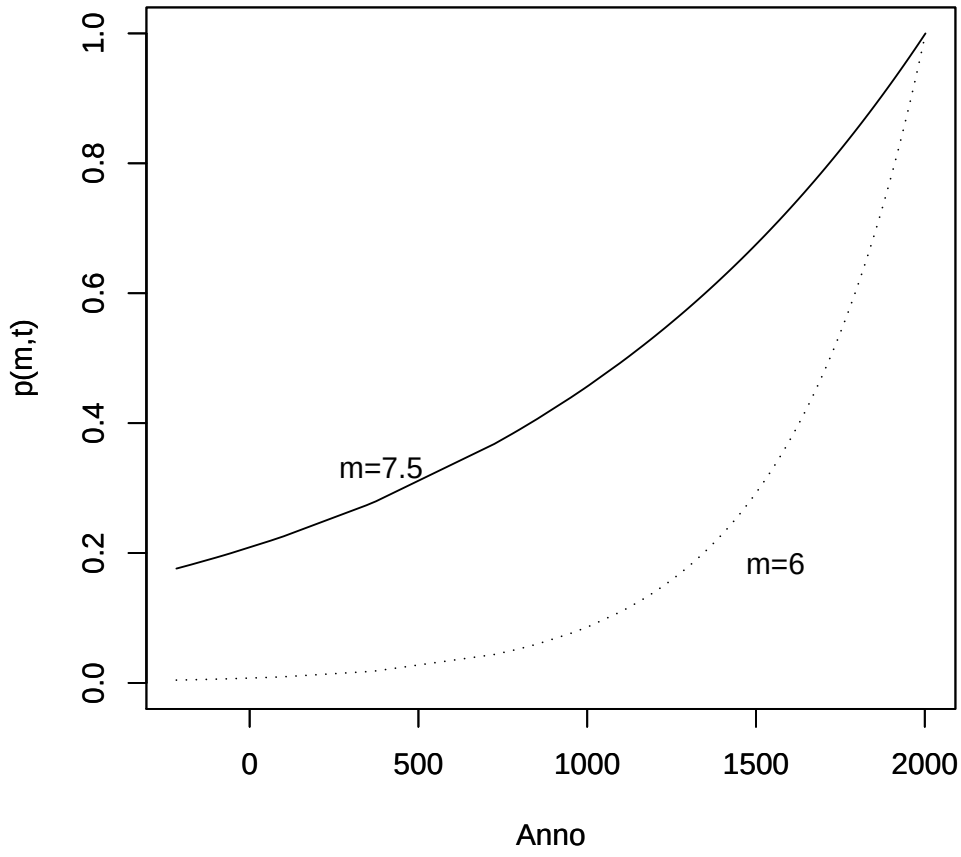


Figura 5.1: Probabilità di registare un evento calcolata per un valore di soglia pari a 5.3

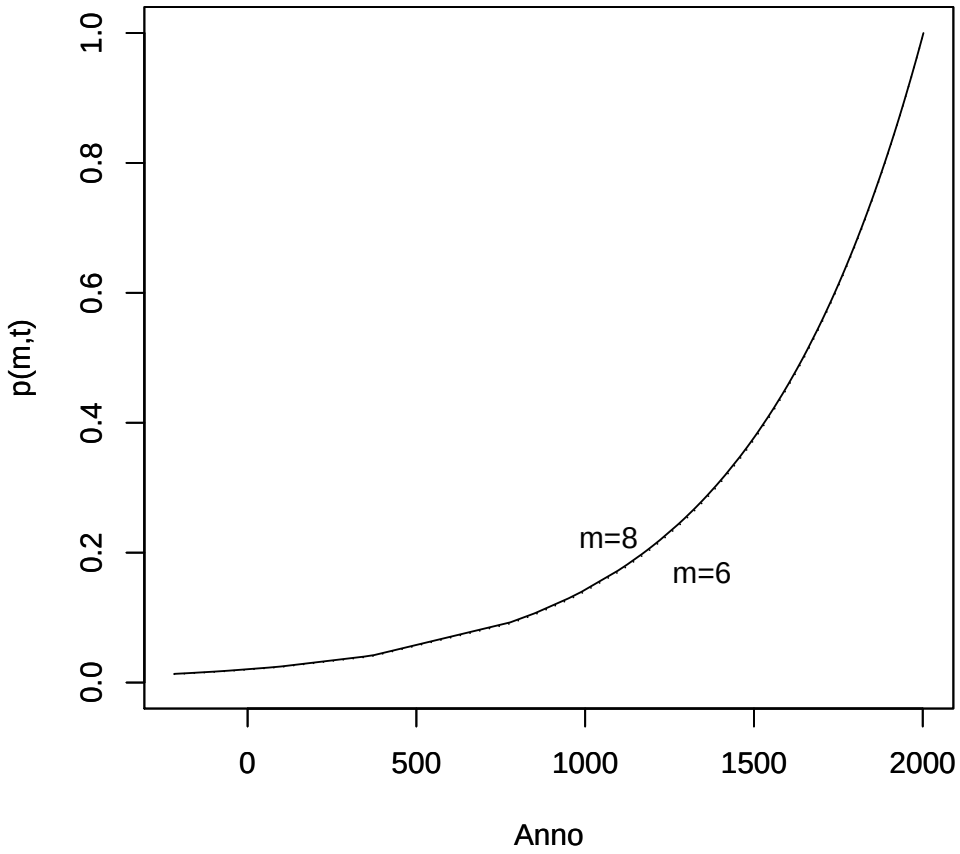


Figura 5.2: Probabilità di registare un evento calcolata per un valore di soglia pari a 5.8

Come si vede anche dai grafici, la funzione proposta sembra dare risultati soddisfacenti: al crescere del valore della magnitudo cresce la probabilità di registrare il dato, così come la funzione cresce in maniera regolare rispetto al tempo.

Da notare come per entrambi i valori di soglia il parametro ξ viene stimato negativo quando si vada a considerare tutto il dataset. I valori limite della distribuzione risulterebbero così essere pari a 8.32, stimando il modello per u=5.8, e pari a 10.87 se il modello viene stimato per u=5.3.

Tuttavia la funzione proposta ha dato dei nuovi problemi numerici quando si è andati a calcolare la varianza legata agli stimatori. Per fare questo si è cercato di calcolare la derivata seconda della funzione per i diversi parametri, la cui inversa, nell'ambito della stima di massima verosimiglianza, può essere usata come stima della varianza dei parametri. La derivata seconda è stata calcolata in maniera numerica, e il calcolo ha presentato una grande sensibilità alla scelta di precisione richiesta. In particolare i valori calcolati per la varianza di w e b sembrano dare indicazioni di una forte correlazione tra i due parametri, come dimostrerebbe anche lo strano risultato ottenuto per le stime in corrispondenza del valore di soglia pari a 5.3, così lontani da quelli ottenuti per il valore di soglia pari a 5.8. Il problema numerico inoltre non viene risolto andando a modificare la scala dei parametri, anzi i tentativi danno un'indicazione ancora maggiore di un legame tra b e w.

Per cercare di capire meglio come i parametri si comportino e per cercare di avere una valutazione della varianza dei parametri si sono calcolate anche le log-verosimiglianze profilo di b e w. In particolare si ricorda che una zona di confidenza di livello $1 - \alpha$ basata sulla verosimiglianza profilo per un generico parametro scalare θ, con τ parametro di disturbo, è definita come:

$$\hat{\Theta}(y) = \left\{ \theta \in \Theta : l_P(\theta) > l_P(\hat{\theta}) - \frac{1}{2}\chi^2_{1;1-\alpha} \right\}.$$

Idealmente la log-verosimiglianza profilo dovrebbe avere una forma parabolica simmetrica attorno al punto di massimo, corrispondente alla stima di massima verosimiglianza e gli intervalli di confidenza stimati attraverso la distribuzione χ^2dovrebbero non differire troppo da quelli stimati attraverso la stima della varianza attraverso l'inversa della derivata seconda della log-verosimiglianza.

Di seguito si mostrano i grafici della log-verosimiglianza profilo negativa per i parametri b e w, con, ove possibile, i quantili della distribuzione χ^2 ad un livello del 95%.

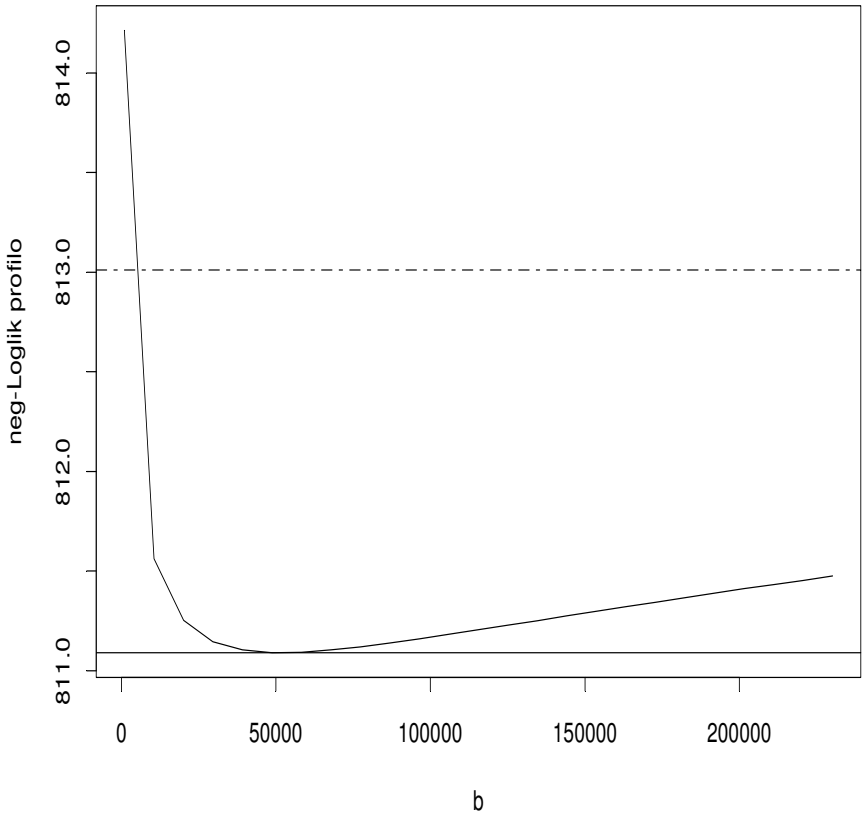


Figura 5.3: Verosimiglianza profilo per il parametro b in un modello con u=5.3

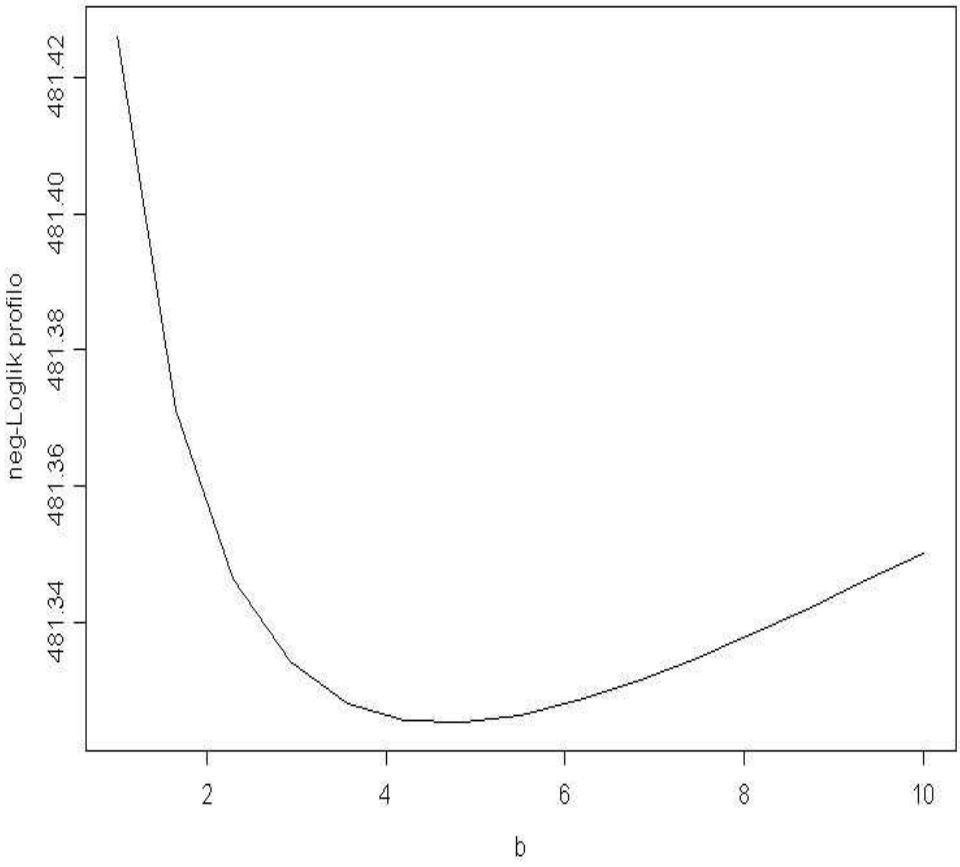


Figura 5.4: Verosimiglianza profilo per il parametro b in un modello con u=5.8

Le verosimiglianze profilo calcolate per il parametro b mostrano nel caso di entrambi i valori di soglia un figura non simmetrica rispetto al punto di massimizzazione della verosimiglianza. Inoltre in entrambi i casi si evidenzia una varianza molto grande, in particolare per quanto riguarda il parametro stimato nel caso in cui il valore di soglia sia pari a 5.3.

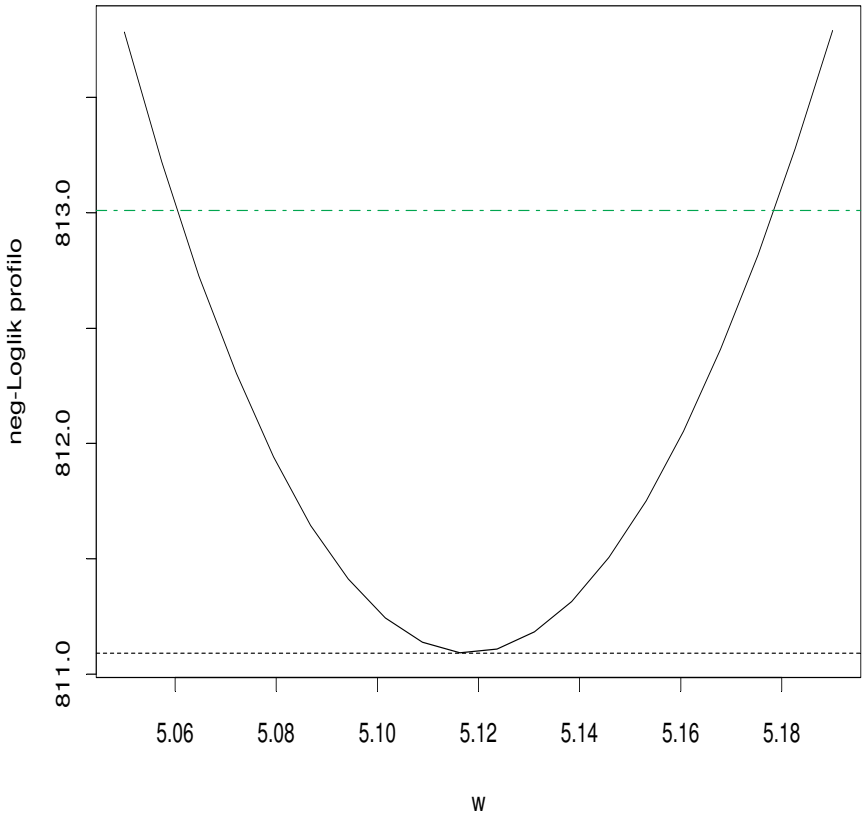


Figura 5.5: Verosimiglianza profilo per il parametro w in un modello con u=5.3

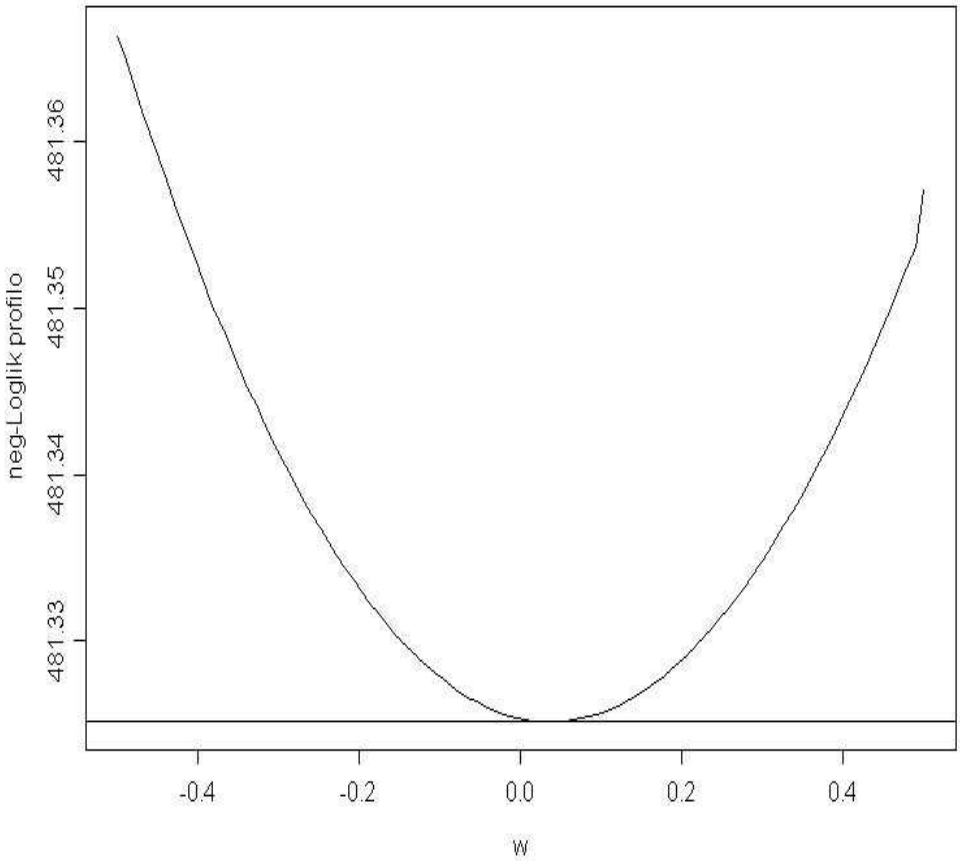


Figura 5.6: Verosimiglianza profilo per il parametro w in un modello con u=5.8

Per quanto rigurda il parametro w invece le verosimiglianze profilo sembrano assumere forma parabolica per entrambi i valori di soglia, tuttavia, per il modello stimato in corrispondenza di $u = 5.8$ l'intervallo di confidenza al 95% comprende anche il valore 0, nel modello cioè verrebbe a mancare il legame tra valore della variabile Maw e probabilità di aver registrato l'evento. Questo non avviene però nel modello con u pari a 5.3, sebbene il valore così elevato e la varianza spropositata del parametro b per questo modello, segnalino come siano necessarie delle modifiche al modello stesso per migliorare.

In definitiva il modello stimato sembra non dare i risultati sperati. Sebbene sia possibile attraverso le verosimiglianza profilo dare una stima della varianza dei parametri b e w, si evidenziano dei problemi che, oltre a rendere più difficile ed imprecisa la ricerca tramite metodi numerici delle quantità ignote, fanno dubitare della bontà della funzione scelta per modellare la probabilità di censura dei dati nel tempo. In particolare si evidenzia una sorta di correlazione tra i due parametri, e il parametro w, che descrive come la probabilità di censura di un evento sia legata alla Magnitudo momento, sembra essere il più problematico da stimare. Questo è ancora una volta legato al fatto che i terremoti che sono stati registrati nel passato non sono caratterizzati unicamento da valori molto alti di magnitudo momento, ma anzi, gli eventi con un valori più alti per la variabile Maw, sono avvenuti negli ultimi anni.

Si procede quindi ad una stima di un nuovo modello in cui la funzione di probabilità di censura di un evento è del tutto similare a quella usata in questo paragrafo, con l'accortezza di eliminare la dipendenza dal valore della Magnitudo dell'evento, alle luce del fatto che l'ipotesi che w sia pari a 0 viene accettata per un modello stimato con $u = 5.8$.

5.3 Un'analisi attraverso una funzione di censura che non dipenda dal valore della Magnitudo

Visti i problemi riscontrati nella modellazione della probabilità di registrare un terremoto, si è provato ad utilizzare una funzione avente un solo parametro, che leghi la probabilità di registrare un terremoto unicamente al momento in cui questo terremoto è avvenuto. E' stata cioè eliminata la dipendenza dal valore della magnitudo, poichè come si è visto nel paragrafo 5.2 l'ipotesi che w sia pari a 0 può essere accettata ad un livello del 95%. La funzione proposta è quindi la medesima funzione usata nel paragrafo 5.2 con $w = 0$:

$$p(t) = \exp\{b(t-1)\}.$$

con $b \geq 0$, poichè si immagina che per tempi più recenti la probabilità di registrare un terremoto aumenti.

Avendo riscalato il tempo su un intervallo $[0, 1]$, la funzione $p(t)$ è crescente e, per $t = 1$, corrispondente al tempo attuale, essa è pari ad 1: registrare un sisma al momento attuale è cioè un evento quasi certo.

La funzione di verosimiglianza da massimizzare risulta quindi essere:

$$\begin{aligned} L(\mathbf{z}, \mathbf{t}; \mu, \sigma, \xi, b) = \\ = \exp\left\{-\int_{\mathcal{A}_u} \{\lambda(m,t)p(t)\}\, dt\, dm\right\} \prod_{i=1}^{n} \lambda(m,t)p(t) = \\ = \exp\left\{-\int_{\mathcal{A}_u} \frac{1}{\sigma}\left[1 + \frac{\xi(m-\mu)}{\sigma}\right]_+^{-\frac{1}{\xi}-1} \exp\{b(t-1)\}\, dt\, dm\right\} \\ \prod_{i=1}^{n} \frac{1}{\sigma}\left[1 + \frac{\xi(m-\mu)}{\sigma}\right]_+^{-\frac{1}{\xi}-1} \exp\{b(t-1)\}. \end{aligned}$$

I valori delle stime di massima verosimiglianza per i parametri del modello calcolati numericamente vengono mostrati nella tabella 5.4.

	n. eccedenze	μ	σ	ξ	b
u=5.3	413	5.39	0.52	-0.15	6.48
u=5.8	148	5.13	0.61	-0.19	4.31

Tabella 5.4: Stime di massima verosimiglianza per il modello stimato su tutti i dati, con probabilità di censura dipendente esclusivamente dal tempo

La matrice di varianze e covarianze per $u = 5.3$ è

$$\begin{bmatrix} 0.001 & -10^{-5} & -2 \times 10^{-4} & 0.008 \\ -10^{-5} & 0.001 & -0.001 & -0.001 \\ -2 \times 10^{-4} & -0.001 & 0.002 & -1.3 \times 10^{-8} \\ 0.008 & -0.001 & -1.3 \times 10^{-8} & 0.11 \end{bmatrix}$$

mentre per $u = 5.8$ è

$$\begin{bmatrix} 0.019 & -0.018 & 0.010 & 0.022 \\ -0.018 & 0.020 & -0.012 & -0.004 \\ 0.010 & -0.012 & 0.009 & -4.3 \times 10^{-8} \\ 0.022 & -0.004 & -4.3 \times 10^{-8} & 0.169 \end{bmatrix}$$

I risultati sembrano finalmente essere concordi per i modelli stimati per i due valori di soglia diversi, e le varianze delle stime non sembrano eccessive. Inoltre le figure della verosimiglianza profilo per il parametro b hanno la forma sperata, e gli intervalli di confidenza basati su quest'ultima corrispondono con quelli stimati attraverso l'approssimazione alla normale della stima di massima verosimiglianza. Nei grafici 5.7 e 5.8 è possibile vedere le verosimiglianze profilo per il parametro b per i due diversi valori di soglia, in cui si evidenzia come un valore nullo di b sia non accettablie a livello statistico: la capacità di registrare un terremoto è cioè effettivamente cambiata nel tempo.

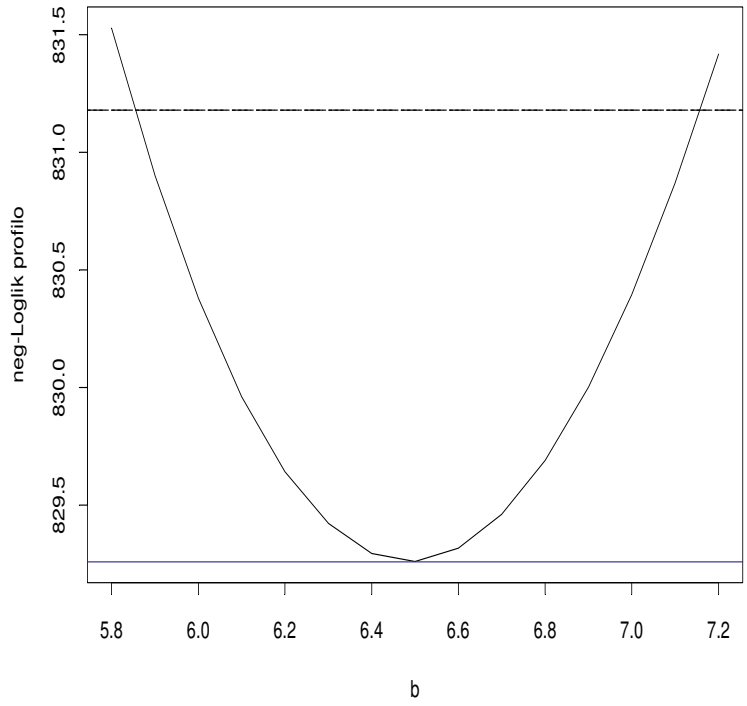
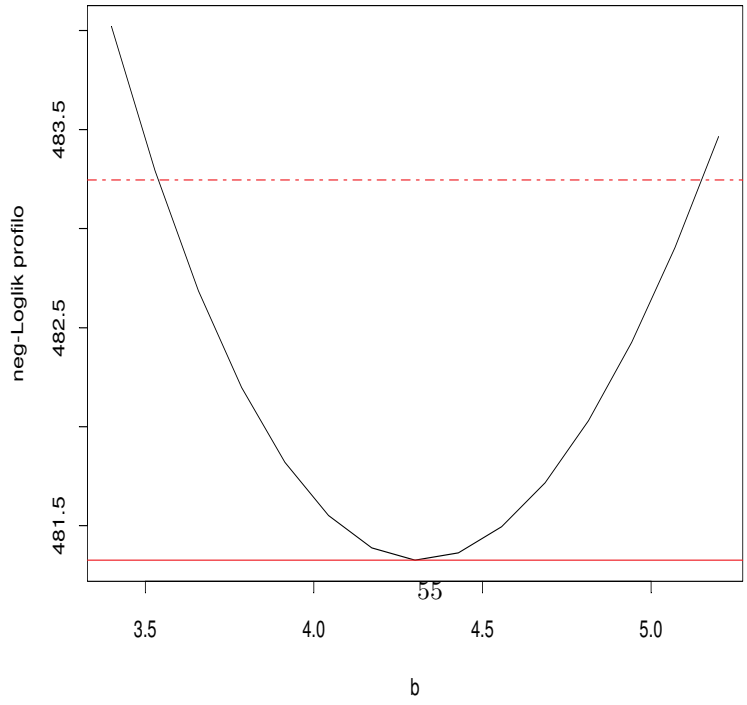


Figura 5.7: Verosimiglianza profilo per il parametro b in un modello con u=5.3



La funzione di probabilità stimata può essere vista nelle figure 5.9 e 5.10.

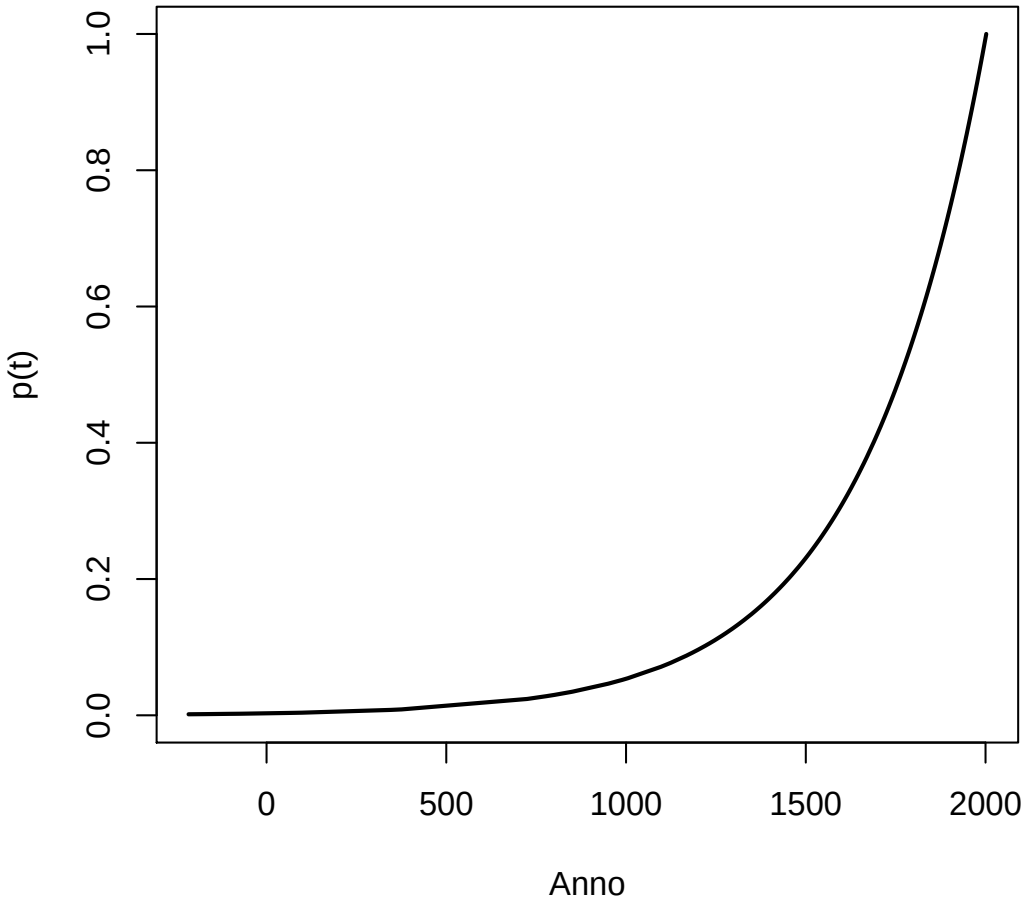


Figura 5.9: La funzione di probabilità di registrare un evento con u=5.3

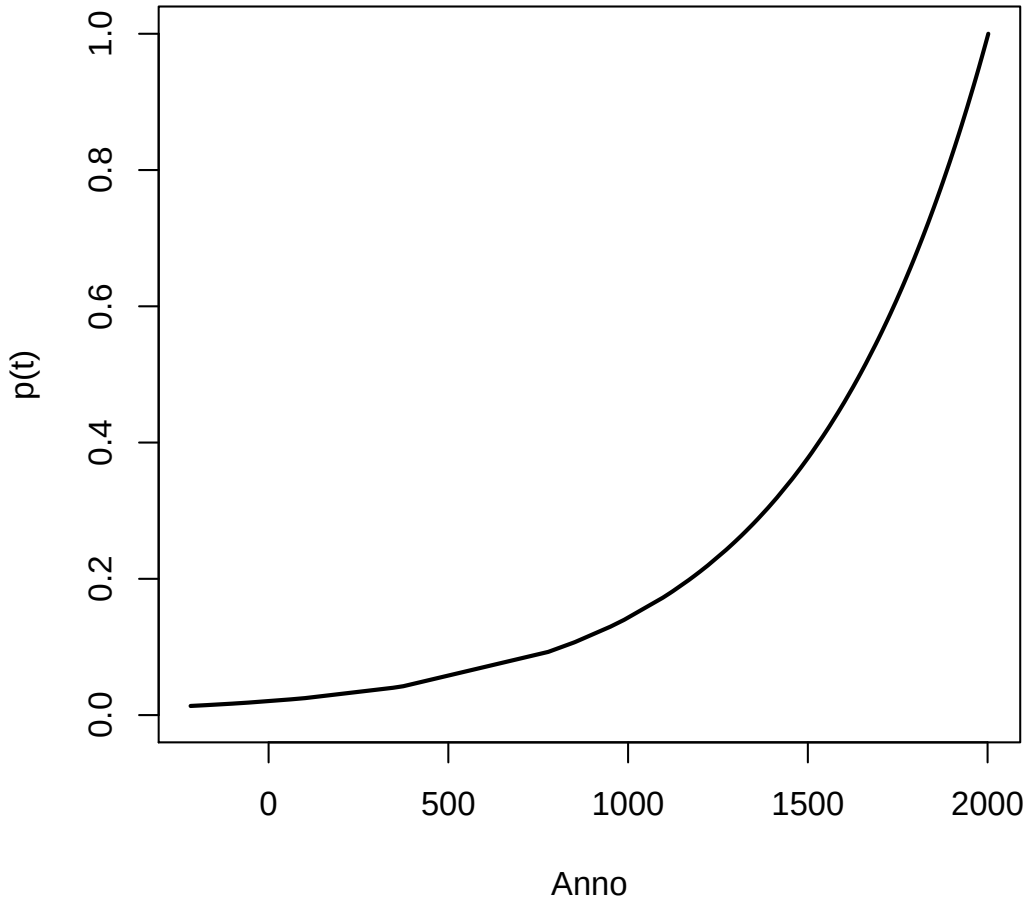


Figura 5.10: La funzione di probabilità di registrare un evento con u=5.8

Per i parametri (μ, σ, ξ) così stimati, le stime di η risultano essere 1.18 per $u = 5.3$ e 0.29 per $u = 5.8$. Le curve di ritorno, ancora una volte in scala logaritmica, sono quelle che si possono vedere in figura 5.11 e 5.12.

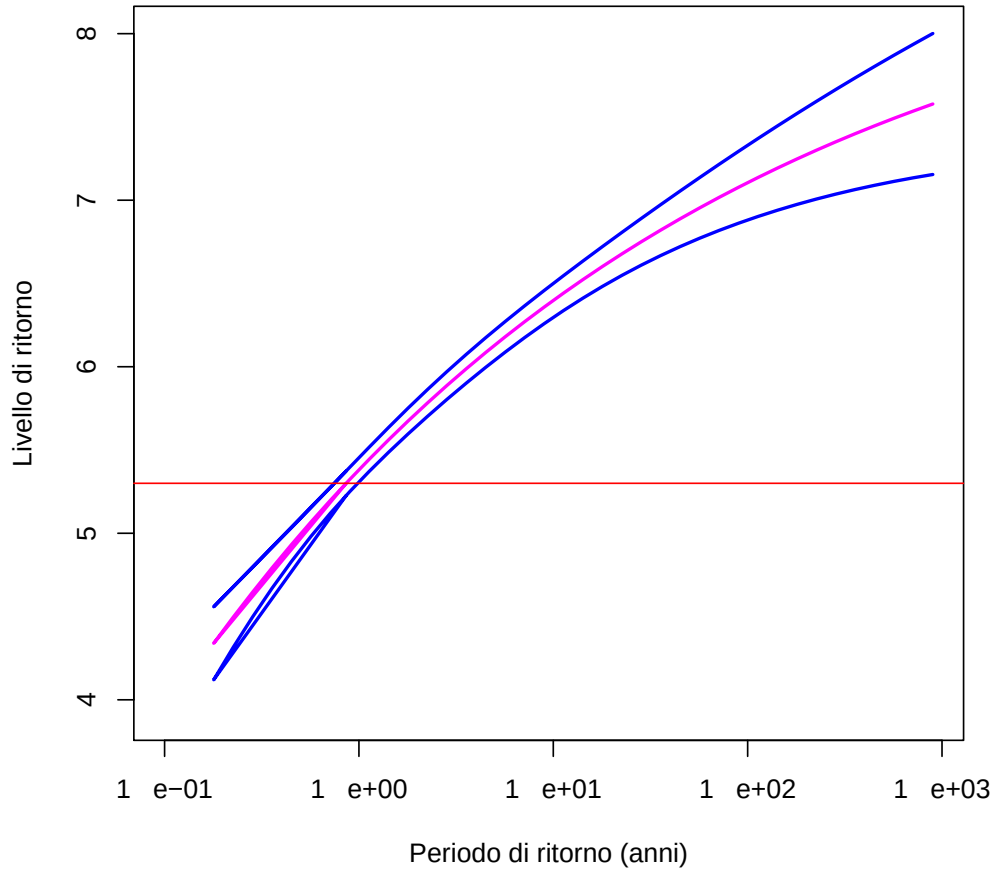


Figura 5.11: Curva di ritorno ed intervalli di confidenza, soglia pari a 5.3 e modello basato su tutto il dataset

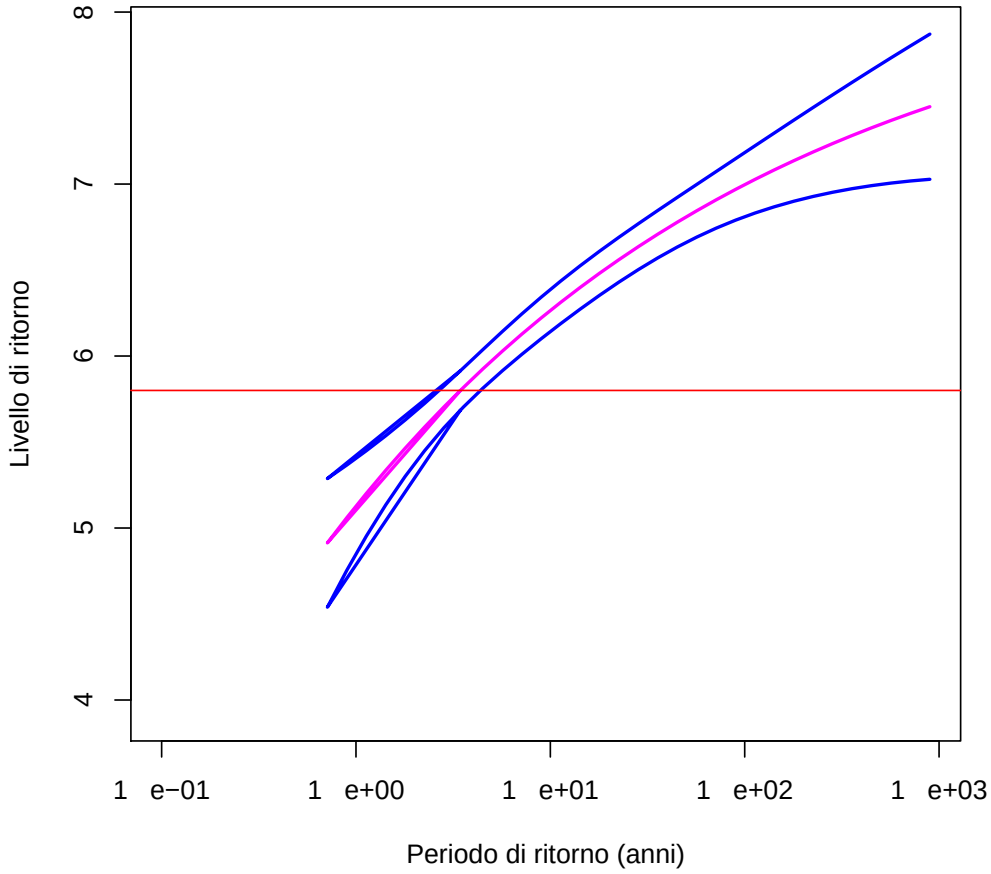


Figura 5.12: Curva di ritorno ed intervalli di confidenza, soglia pari a 5.8 e modello basato su tutto il dataset

Come si vede in questo caso la forma stimata per la curva di ritorno non è più lineare, poichè il parametro ξ è stimato essere negativo, e la curva risulta quindi essere limitata a destra. Inoltre gli intervalli di confidenza sono molto stretti, e al contrario di quanto avveniva per gli intervalli di confidenza delle figure 4.9 e 4.10 non mostrano tendenza a divergere così velocemente.

Capitolo 6

Uno studio di simulazione

Date le difficoltà sorte nel processo di stima dei parametri nei diversi modelli usati finora, ci si propone di esplorare in maniera più approfondita la relazione che intercorre tra la specifiaczione del modello e l'inferenza per i le diverse classi di modelli considerate fin'ora.

Le difficoltà incontrate nella stima dei preocessi sismogenetici in Italia, infatti, fanno sorgere dubbi sull'effettiva bontà ed utilià dei modelli usati fin qui. Tuttavia, dal lavoro svolto in questa e in altre sedi (cfr. Coles and Sparks,2004), non è chiaro se le difficoltà nascono da un'errata specificazione del modello, o dall'inferenza svolta sul modello, quand'anche questo fosse corretto.

In questo capitolo si simulano dati aventi struttura stocastica di processi di punto con valori estremi censurati, così come è stato descritto nei capitoli precedenti (parte II), e si esamina la capacità di stimare in maniera corretta questi modelli tramite l'inferenza di verosimiglianza.

Inizialmente sono stati simulati 10000 dati z provenienti da una funzione GEV avente funzione di ripartzione $G(z; \mu, \sigma, \xi)$

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

di parametri $\mu = 3$, $\sigma = 1$ e $\xi = -0.2$ (figura 6.1). Idealmente ognuna delle realizzazioni della variabile è avvenuta in un istante di tempo diverso dalle altre, come se avessimo registrato un'osservazione

ogni anno, o ogni giorno. Il tempo t è quindi diviso in 10000 istanti diversi l'uno dall'altro.

Inizialmente a questi dati simulati non è stato applicato nessun processo di censura, e sono stati analizzati tramite processi di punto così come è descritto in 4. Il valore di soglia u che separa i valori estremi è stato scelto in maniera arbitraria pari a 5.25, in modo da avere il 5% dei dati al di sopra del valore di soglia. (figura 6.2).

Le stime del processo attraverso il modello sono riporate nella tabella 6.1

	n. eccedenze	μ(s.e.)	σ(s.e.)	ξ(s.e.)
u=5.25	500	2.9(0.25)	1.04(0.16)	-0.21(0.03)

Tabella 6.1: Stime di massima verosimiglianza su tutti i dati simulati

Come si vede i parametri vengono stimati in maniera corretta dal modello, come ci si auspica che sia.

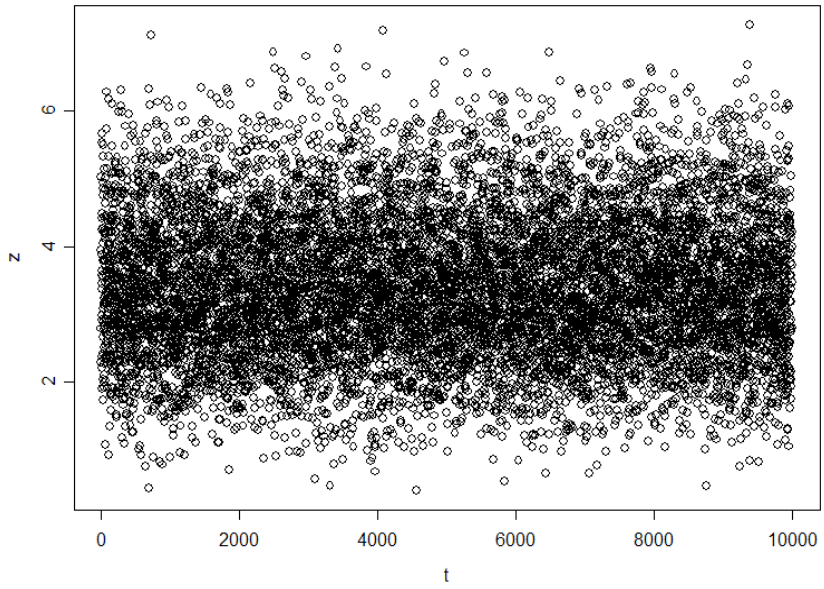


Figura 6.1: I dati simulati

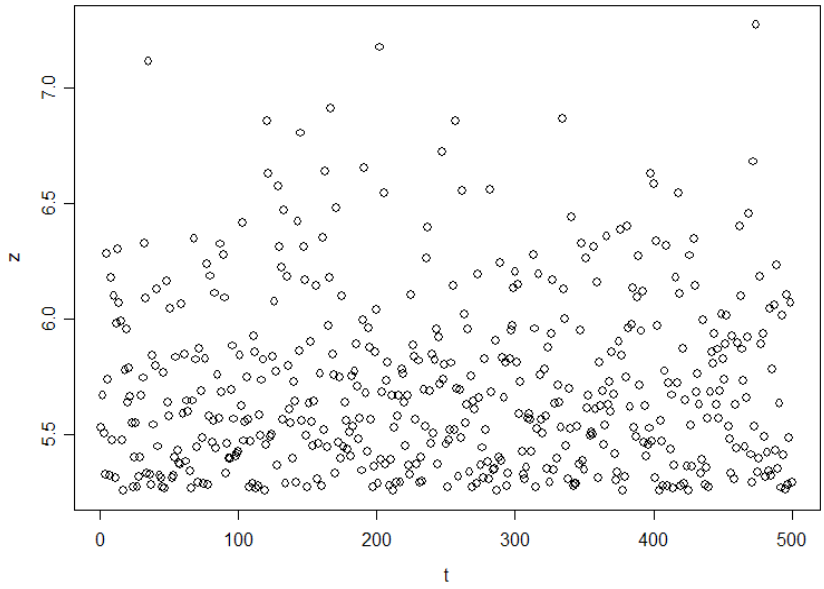


Figura 6.2: I dati simulati al di sopra della soglia $u = 5.25$

Si procede ora ad una verifica della bontà delle stime nel caso in cui i dati vengano sottoposti ad una censura. L'idea che è alla base di questa parte è quella di simulare il processo che si è stimato nella parte II. Dai dati simulati il cui valore supera la soglia di *u=5.25*, tramite un funzione di probabilità $p(z,t)$ viene estratto un campione, in cui ogni punto (z_i, t_i) viene inserito con probabilità $p_i = p(z_i, t_i)$. La funzione di intensità del processo non risultarà quindi esere $\lambda(z,t)$, ma una $\lambda_M z, t = \lambda(z,t)p(z,t)$ e la verosimiglianza da massimizzare attraverso i dati campionati sarà:

$$L(\mathbf{z}, \mathbf{t}; \mu, \sigma, \xi, v, w, b) = \\ = \exp\Big\{-\int_{\mathcal{A}_u} \lambda(z,t)p(z,t)\, dz\, dt\Big\} \prod_{i=1}^{n} \lambda(z,t)p(z,t).$$

L'interesse è ancora una volta quello di andare a verificare se le stime di massima verosimiglianza corrispondono ai veri parametri del modello, avendo simulato il processo di censura dei dati attraverso diverse funzioni di probabilità simili a quelle utilizzate per l'analisi dei dati dei terremoti in Italia.

Il tempo t nelle diverse funzioni di probabilità è da considerarsi da qui in avanti riscalato su una scala da 0 ad 1.

La prima funzione di probabilità, già descritta in 5.1, è $a(z,t;v,w,b)$, di forma polinomiale:

$$a(z,t,v,w,b) = \left(1 - \frac{v}{x^w}\right) + \frac{v}{x^w} t^b.$$

con $b \geq 0$, $w \geq 0$ e $v \leq u^w$.

In figura 6.3 vediamo l'andamento della funzione $a'(z,t)$ di parametri $v = 11$,$w = 1.5$ e $b = 3$,

$$a'(z,t) = \left(1 - \frac{11}{x^{1.5}}\right) + \frac{11}{x^{1.5}} t^3.$$

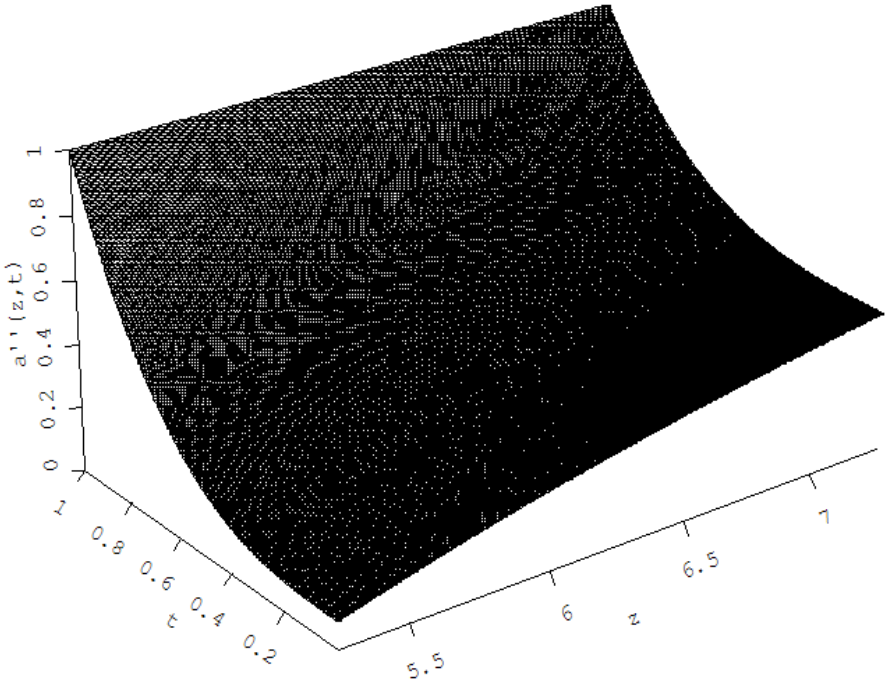


Figura 6.3: La funzione di probablità $a'(z,t)$

In figura 6.4 vediamo invece un campione estratto, attraverso il quale sono state ottenute le stime di massima verosimiglianza in tebella 6.2 e la matrice di varianze e covarianze $\Sigma_{a'}$:

	n. eccedenze	μ	σ	ξ	v	w	b
u=5.25	199	4.16	0.79	-0.2	4.29	0.98	2.41

Tabella 6.2: Stime di massima verosimiglianza basate sul campione estratto tramite funzione di probabilità $a'(z,t)$

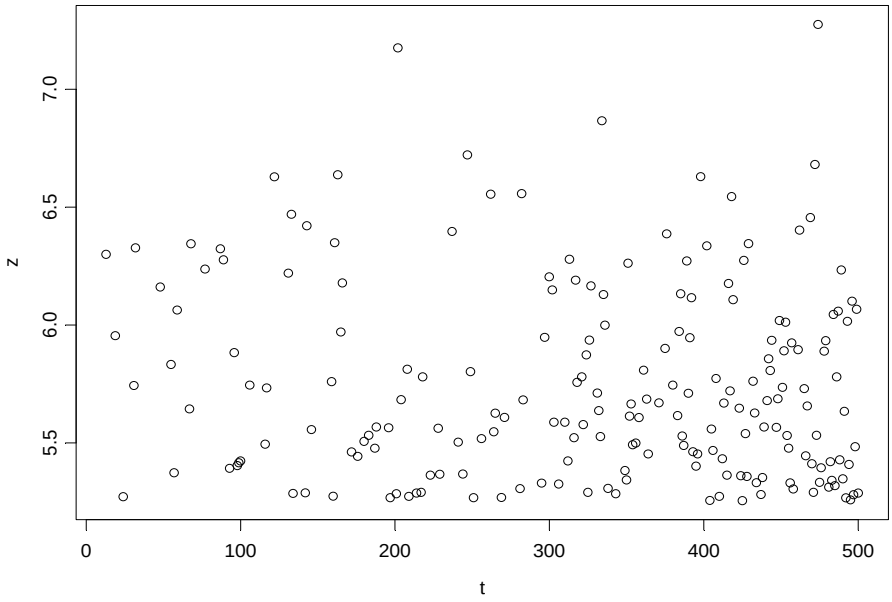


Figura 6.4: Il campione selezionato tramite $a'(z,t)$

$$\Sigma_{a'} = \begin{bmatrix} 0.04 & -0.03 & 0.00 & 0.37 & 0.04 & 0.08 \\ -0.03 & 0.02 & -0.00 & -0.55 & -0.07 & -0.01 \\ 0.00 & -0.00 & 0.00 & 0.18 & 0.02 & -0.00 \\ 0.37 & -0.55 & 0.18 & 48.02 & 6.47 & -2.48 \\ 0.04 & -0.07 & 0.02 & 6.47 & 0.87 & -0.32 \\ 0.08 & -0.01 & -0.00 & -2.48 & -0.32 & 1.00 \end{bmatrix}$$

Come si vede le stime di massima verosimiglianza stimano abbastanza bene i parametri del modello: le stime puntuali di μ, σ, ξ si avvicinano ai veri valori dei parametri, mentre per quanto riguarda i parametri v, w, b le stime puntuali sembrano non essere così precise,tuttavia il vero valore del parametro è compreso in un intervallo di confidenza al 95%, e si può quindi essere soddisfatti del risultato.

Si prova ora a vedere come viene stimato il modello nel caso in cui la funzione di probabilità $a''(z,t)$ censuri in maniera meno forte i dati,

con parametri $v = 2.3$,$w = 1$ e $b = 3$:

$$a''(z,t) = \left(1 - \frac{2.3}{x}\right) + \frac{2.3}{x}t^3.$$

In figura 6.5 troviamo le probabilità assegnate ad ogni punto (z_i, t_i) e in figura 6.6 il campione selezionato. Salta subito agli occhi la differenza rispetto al campione selezionato tramite $a'(z,t)$: in figura 6.6 abbiamo molti più dati, e in particolare, troviamo molti punti anche nella parte sinistra del grafico, a causa della probabilità piuttosto alta di inclusione che hanno anche dati registrati nei primi istanti di tempo.

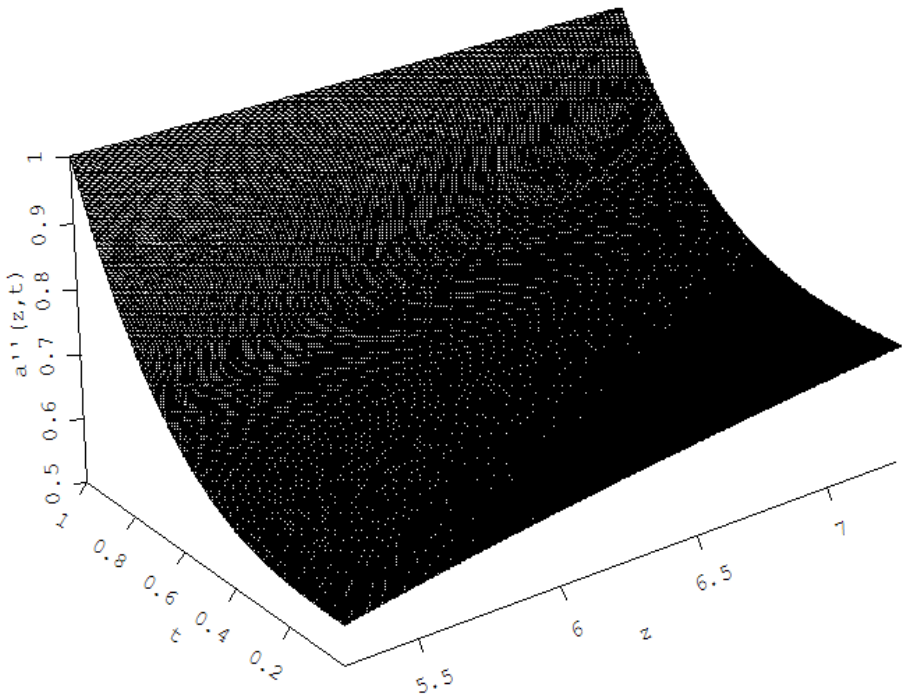


Figura 6.5: La funzione di probablità $a''(z,t)$

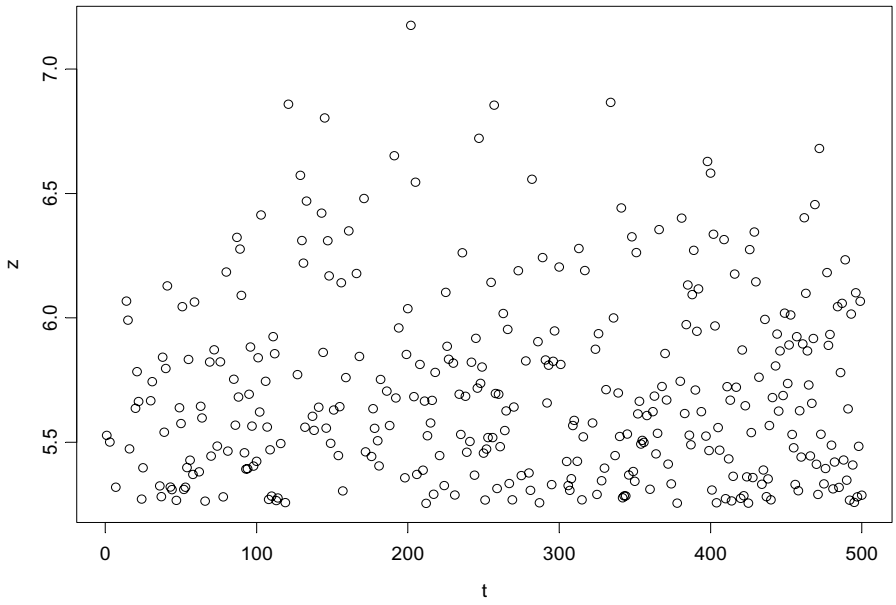


Figura 6.6: Il campione selezionato tramite $a''(z,t)$

Attraverso la massimizzazione della funzione di verosimiglianza, individuiamo le stime di tabella 6.3 e la matrice delle varianze e covarianze $\Sigma_{a''}$

	n. eccedenze	μ	σ	ξ	v	w	b
u=5.25	339	4.16	0.79	-0.2	4.29	0.98	2.41

Tabella 6.3: Stime di massima verosimiglianza basate sul campione estratto tramite funzione di probabilità $a''(z,t)$

$$\Sigma_{a''} = \begin{bmatrix} 0.04 & -0.03 & 0.00 & 0.37 & 0.04 & 0.08 \\ -0.03 & 0.02 & -0.00 & -0.55 & -0.07 & -0.01 \\ 0.00 & -0.00 & 0.00 & 0.18 & 0.02 & -0.00 \\ 0.37 & -0.55 & 0.18 & 48.02 & 6.47 & -2.48 \\ 0.04 & -0.07 & 0.02 & 6.47 & 0.87 & -0.32 \\ 0.08 & -0.01 & -0.00 & -2.48 & -0.32 & 1.00 \end{bmatrix}$$

Le stime sembrano essere piuttosto soddisfacenti, anche in questo caso, sebbene si noti ancora che le stime per quanto riguarda i parametri v, w, b sono caratterizzate da varianze decisimente più grandi di quelli dei parametri μ, σ e ξ.

Bisogna tuttavia aggiungere che la massimizzazione della funzione ha dato, in alcuni casi dei problemi: a volte l'algoritmo ha trovato risultati tra loro diversi a seconda dei valori iniziali dati o ha trovato, in alcuni casi, stime intervallari che non contenevano il vero valore del parametro. La funzione infatti è piuttosto complicata, e l'esistenza di un vincolo sul parametro v legato a w fa si che la stima dei due parametri sia a volte piuttosto complicata, mostrando anche una certa correlazione tra i due il più delle volte.

In generale però la funzione stima i veri valori dei parametri del processo μ, σ e ξ in maniera sodddisfacente.

Tuttavia andando ad utilizzare funzioni di probabilità con scarsa capacità di censura, in cui ad ogni punto fosse attribuita una probabilità piuttosto alta di essere incluso nel campione, l'algoritmo non riesce a trovare i punti che massimizzano la verosimiglianza. L'uso di una funzione di probabilità del tipo $a(z,t)$ risulta quindi efficace nel caso in cui il processo di censura sia abbstanza forte, ma può dare difficoltà se il processo di censura è debole.

Provando invece ad utilizzare una forma diversa per la funzione di probabilità quale una forma esponenziale con due parametri w e b, si ottiene la funzione $b(z,t;w,b)$ (cfr: 5.2):

$$b(z,t;w,b) = \exp\left\{\frac{b(t-1)}{z^w}\right\}.$$

La prima scelta dei parametri è di $w = 1.4$ e $b = 22$:

$$b'(z,t) = \exp\left\{\frac{22(t-1)}{z^{1.4}}\right\}.$$

Nelle figure 6.7 e 6.8 troviamo la funzione di probabilità, e il campione effettivamente estratto.

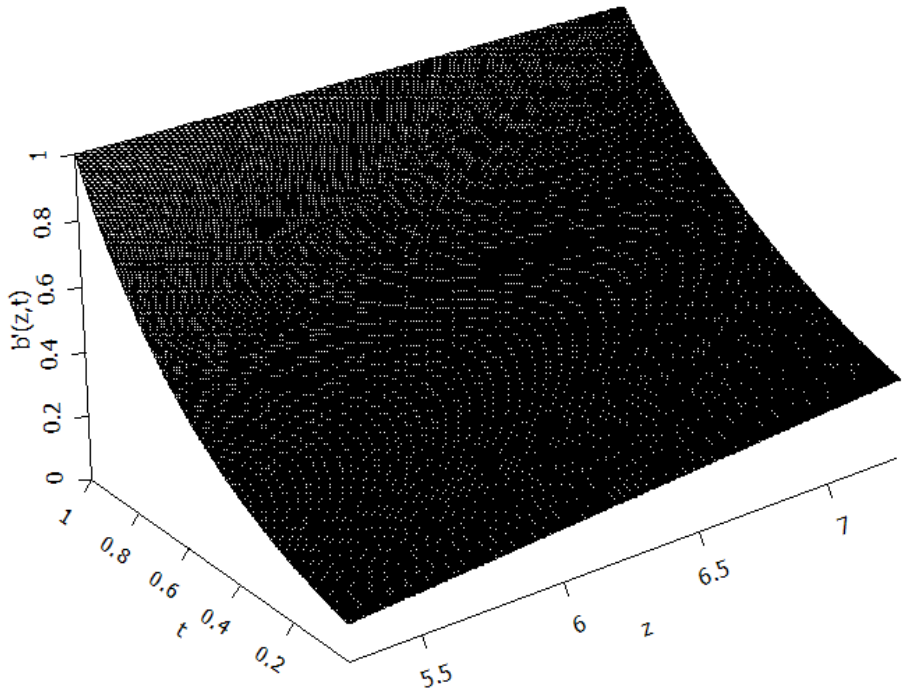


Figura 6.7: La funzione di probablità $b'(z,t)$

In tabella 6.4 vediamo le stime di massima verosimiglianza, con matrice di varianze e covarianze $\Sigma_{b'}$

	n. eccedenze	μ	σ	ξ	w	b
u=5.25	227	4.3	0.71	-0.18	0.77	6.93

Tabella 6.4: Stime di massima verosimiglianza basate sul campione estratto tramite funzione di probabilità $b'(z,t)$

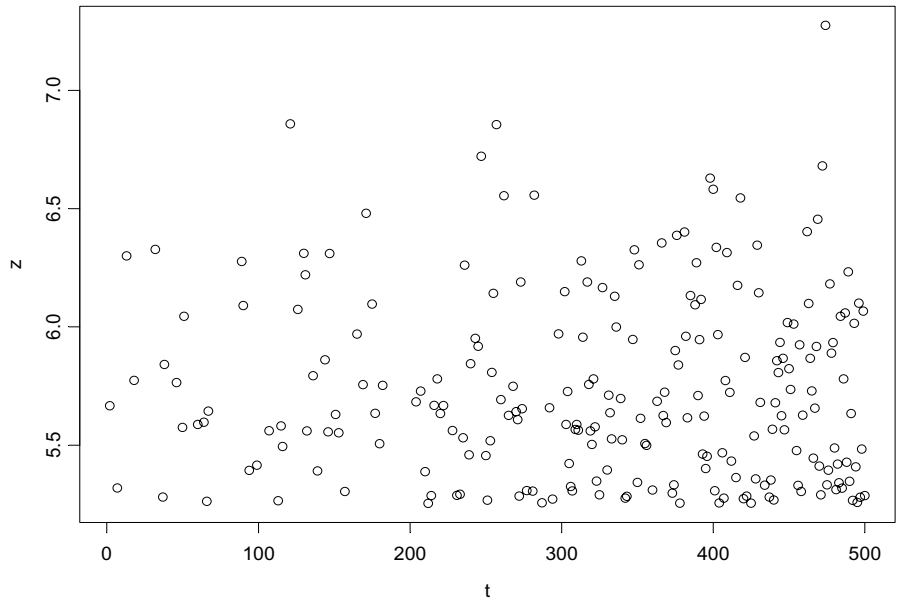


Figura 6.8: Il campione selezionato tramite $b'(z,t)$

$$\Sigma_{b'} = \begin{bmatrix} 0.03 & -0.02 & 0.00 & 0.23 & 2.85 \\ -0.02 & 0.02 & -0.00 & -0.20 & -2.53 \\ 0.00 & -0.00 & 0.00 & 0.06 & 0.75 \\ 0.23 & -0.20 & 0.06 & 4.18 & 50.36 \\ 2.85 & -2.53 & 0.75 & 50.36 & 607.64 \end{bmatrix}$$

Le stime, come si vede, sono abbastanza soddisfacenti, sebbene si noti che i parametri b e w hanno varianze molto grandi. Si noti inoltre la forte correlazione tra i due parametri b e w, che era già stata sottolineata in 5.2.

Utilizzando un funzione che attribuisca probabilità più alte ad ogni punto di essere inserito nel campione con parametri $w = 1.5$ e b=8,

$$b''(z,t) = \exp\left\{\frac{8(t-1)}{z^{1.5}}\right\},$$

otteniamo le figure 6.9 e 6.10, in cui vediamo la funzione di probabilità ed un campione estratto.

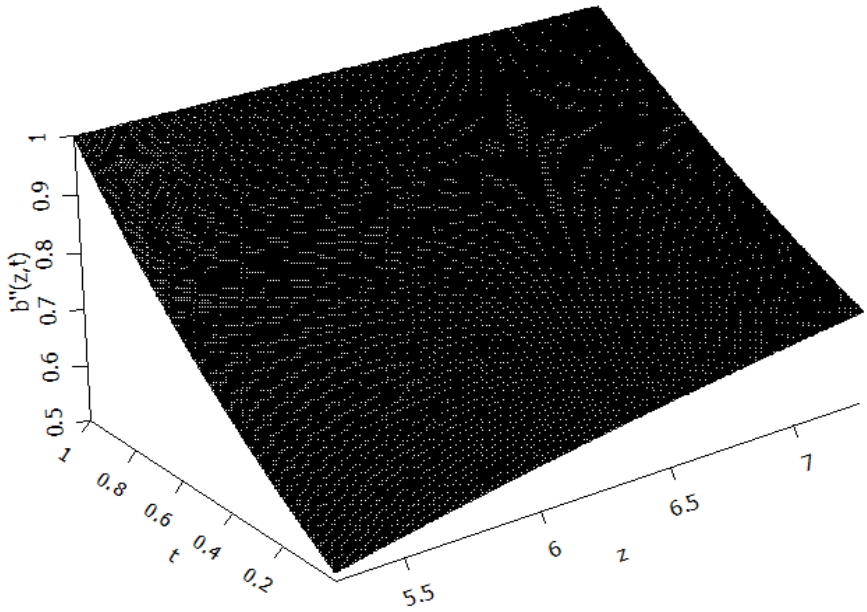


Figura 6.9: La funzione di probablità $b''(z,t)$

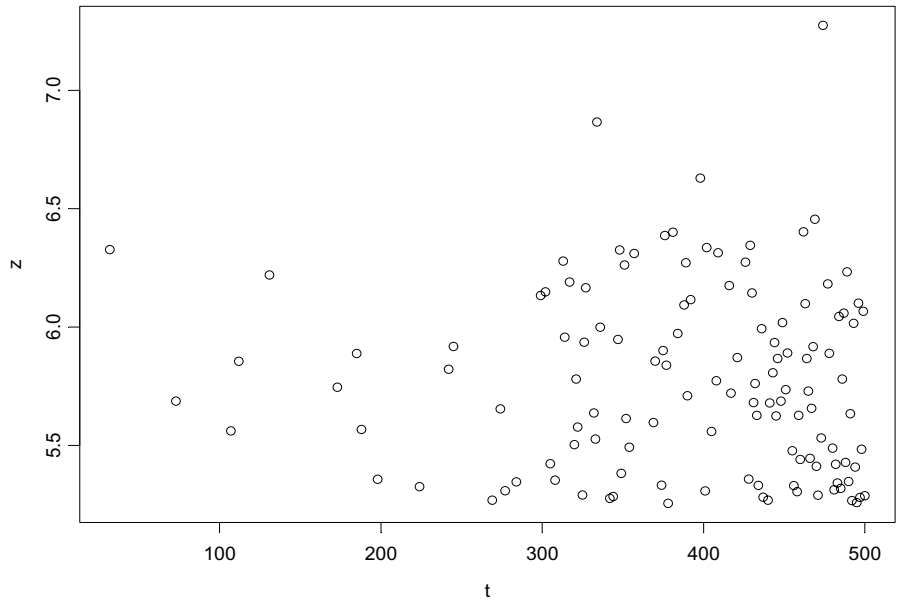


Figura 6.10: Il campione selezionato tramite $b''(z,t)$

Le stime di massima verosimiglianza vengono mostrate in tabella 6.5:

	n. eccedenze	μ	σ	ξ	w	b
u=5.25	355	4.32	0.69	-0.17	1.11	4.42

Tabella 6.5: Stime di massima verosimiglianza basate sul campione estratto tramite funzione di probabilità $b''(z,t)$

con matrice di varianze e covarianze $\Sigma_{b''}$

$$\Sigma_{b''} = \begin{bmatrix} 0.02 & -0.01 & 0.00 & 0.48 & 3.79 \\ -0.01 & 0.01 & -0.00 & -0.44 & -3.39 \\ 0.00 & -0.00 & 0.00 & 0.13 & 1.06 \\ 0.48 & -0.44 & 0.13 & 21.14 & 161.85 \\ 3.79 & 3.39 & 1.06 & 161.854 & 1240.64 \\ . & & & & \end{bmatrix}$$

Le stime trovate sono anche in questo caso soddisfacenti, sebbene

le varianze di b e w e la covarianza tra i due parametri siano ancora una volta molto grandi. Inoltre la funzione ha dato risultati molto diversi a seconda del campione estratto, andando a volte ad individuare stime intervallari che non contengono i veri valori dei parametri b e w. Piuttosto corrette invece le stime ottenute per i parametri μ, σ e ξ.

Inoltre utilizzando funzioni di probabilità che attribuiscono probabilità di entrare nel campione piuttosto alte ad ogni punto, l'algoritmo di massimizzazione della verosimiglianza non riesce,il più delle volte, a trovare un risultato. Anche in questo caso quindi è bene essere molto cauti nella stima nel caso in cui il processo di censura dei dati sia debole.

Viene infine utilizzata la funzione $c(t; b)$ di forma esponenziale con un unico parametro b che regola la dipendenza dal tempo (cfr. 5.2):

$$c(t; b) = \exp\{b(t-1)\}.$$

Inizialmente il parametro b viene fissato pari a 4,

$$c'(t) = \exp\{4(t-1)\},$$

e in figura 6.11 e 6.12 vediamo rispettivamente le probabilità associate ad ogni punto e il campione estratto.

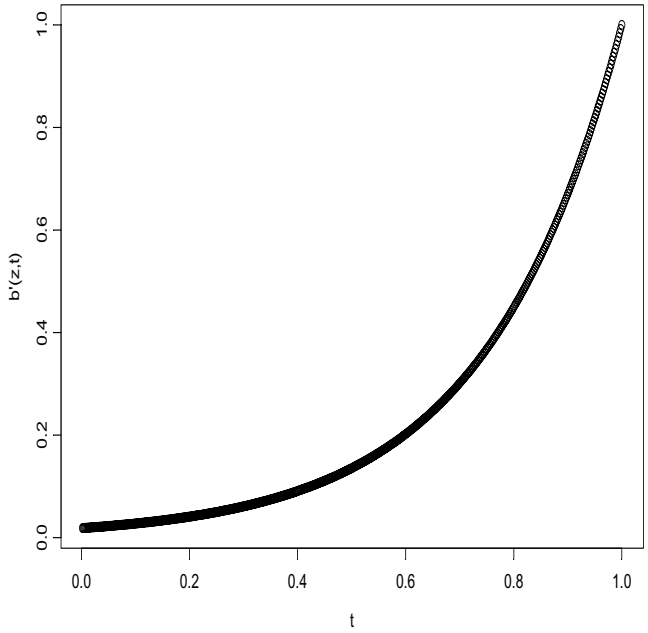


Figura 6.11: La funzione di probablità $c'(z,t)$

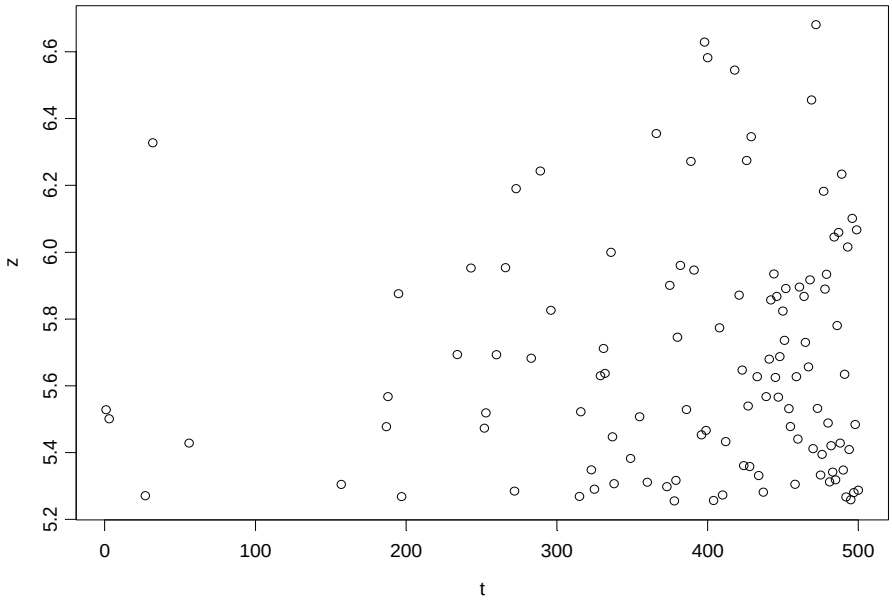


Figura 6.12: Il campione selezionato tramite $c'(z,t)$

Le stime di massima verosimiglianza ottenute vengono mostrate in tabella 6.6:

	n. eccedenze	μ	σ	ξ	b
u=5.25	114	4.1	0.9	-0.30	4.11

Tabella 6.6: Stime di massima verosimiglianza basate sul campione estratto tramite funzione di probabilità $c'(t)$

con matrice di varianze e covarianze $\Sigma_{c'}$:

$$\Sigma_{c'} = \begin{bmatrix} 0.06 & -0.05 & 0.02 & 0.04 \\ -0.05 & 0.06 & -0.02 & -0.01 \\ 0.02 & -0.02 & 0.01 & -0.00 \\ 0.04 & -0.01 & -0.00 & 0.2 \end{bmatrix}$$

Le stime sono piutosto soddisfacenti, sia per i parametri specifici del processo sismiogenetico, sia per quanto riguarda il parametro b che regola la funzione di probabilità.

Si è utilizzata anche un funzione $c''(t)$ che censurasse in maniera nettamente meno forte il processo con parametro $b = 0.2$

$$c''(t) = \exp\left\{\frac{(t-1)}{5}\right\}.$$

In figura 6.13 vediamo la funzione $c''(t)$, e in figura 6.14 il campione estratto attraverso cui vengono trovate le stime di massima verosimiglianza (tabella 6.7) e la matrice di varianze e covarianze $\Sigma_{c''}$.

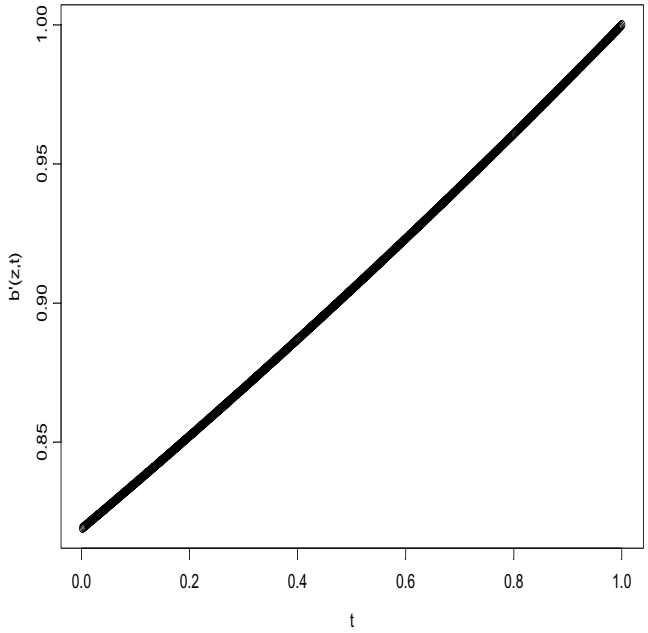
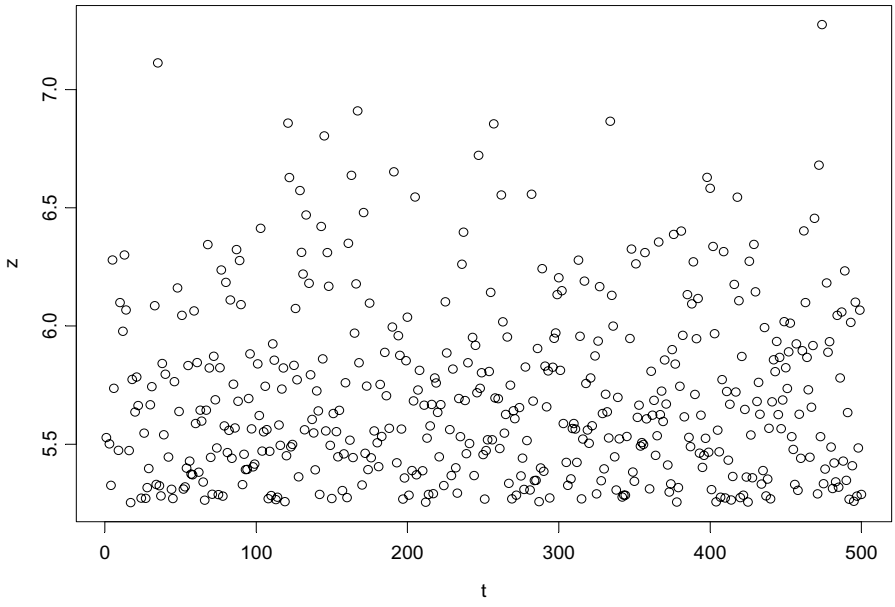


Figura 6.13: La funzione di probablità $c''(z,t)$



0

Figura 6.14: Il campione selezionato tramite $c''(z,t)$

	n. eccedenze	μ	σ	ξ	b
u=5.25	458	4.29	0.74	-0.2	0.18

Tabella 6.7: Stime di massima verosimiglianza basate sul campione estratto tramite funzione di probabilità $c''(t)$

con matrice di varianze e covarianze $\Sigma_{c''}$:

$$\Sigma_{c''} = \begin{bmatrix} 0.0114 & -0.0077 & 0.0029 & 0.0095 \\ -0.0077 & 0.0071 & -0.0030 & -0.0019 \\ 0.0029 & -0.0030 & 0.0015 & -0.0000 \\ 0.0095 & -0.0019 & -0.0000 & 0.0264 \end{bmatrix}$$

In definitiva per tutte e tre le funzioni di censura utilizzate il modello viene stimato abbastanza bene tramite la massima verosimiglianza. In alcuni casi si possono trovare dei problemi numerici di cui bisogna tener conto nel momento in cui si va a stimare un modello; in generale però le stime corrispondenti ai parametri relativi all'effettivo processo generatore dei dati sono corrette. Nel caso in cui il processo di censurata sui dati sia poco forte, l'uso di funzioni di probabilità del tipo $a(z,t)$ e $b(z,t)$ può portare a stime poco corrette o addiruttura potrebbe risultare impossibile trovare il massimo della verosimiglianza.

La funzione $c(t;b)$ non presenta in alcun caso problemi numerici, e dà stime corrette anche nel caso in cui il processo di censura sia poco forte, tuttavia essa non da quella ricchezza di informazioni che si può avere utilizzando le prime due fuzioni.

Analizzando quindi dei dati simili a quelli dei terremoti in Italia, si può essere relativamente sicuri che nel caso in cui le verosimiglianza possa essere massimizzata i risultati siano corretti, qualunque sia la classe di modelli che si decide di utilizzare. Tuttavia si consiglia di verificare che almeno due dei modelli diano risultati tra di loro non discordanti, poichè essi stimano in maniera corretta i parametri se il modello ipotizzato è vero. Inoltre è bene ricordarsi che le stime

puntuali riguardanti i parametri che regolano il processo di censura, usando funzioni di probabilità del tipo $a(z,t)$ e $b(z,t)$, possono spesso essere lontane dal vero valore del parametro, ed è quindi necessario verificare attentamente le stime intervallari.

Conclusioni

La modellazione dei dati riguardanti i terremoti in Italia ha infine portato a risultati diversi da quelli che si erano ipotizzati inizialmente. La funzione di censura utilizzata da Coles e Sparks nel loro articolo riguardante i fenomeni vulcanici si è rivelata inefficacie per modellare i terremoti in Italia e si è provato a ricorrere ad una funzione di forma non più polinomiale ma esponenziale, in cui il numero di parametri da stimare è passato da tre a due. Questa funzione di censura tuttavia ha dato ancora problemi, in praticolare si è notata un sorta di correlazione tra i due parametri, e si è evidenziato come, nei dati riguardanti i sismi degli ultimi 2300 anni in Italia, la censura sia avvenuta solo in funzione del tempo trascorso dal momento del sisma ad oggi e non in funzione della magnitudo momento del sisma. La funzione di censura di forma esponenziale, avente un unico parametro legato al tempo trascorso dalla data del sisma, ha infine dato dei risultati coerenti e chiaramente leggibili, rendendo possibile la stima dei parametri del modello sottostante il processo sismogenetico.

In particolare si sono riscontrate delle differenze nelle stime ottenute basando l'analisi sui dati riguardanti i sismi avvenuti dal 1885 in poi, e l'analisi basata sulla serie completa dei dati. Nel primo caso il parametro ξ è stato stimato positivo e prossimo a zero, segno che la coda destra della curva della distribuzione dei dati scende in maniera pressochè esponenziale, mentre nel secondo caso il parametro ξ è stato stimato negativo, segno che la curva è limitata a destra. Questo influisce anche le curve di ritorno, e cioè il tempo stimato affinchè un

sisma di determinata magnitudo momento torni a verificarsi nel territorio italiano, e in particolare sugli intervalli di confidenza legati a tali curve.

Infine, data l'incertezza sull'effettiva validità dei modelli derivata dalle difficoltà incontrate nel trovare delle stime di massima verosimiglianza, si è verificata la capacità di stima dei diversi modelli attraverso uno studio di simulazione. Dapprima i dati, simulati attraverso un processo con parametri noti, non sono stati sottoposti ad alcuna censura, e i valori stimati si avvicinano molto ai veri valori dei parametri.

In un secondo momento i dati sono stati sottoposti a processi di censura più o meno forti attraverso diverse classi di funzioni di probabilità di censura. Le stime ottenute tramite la massima verosimiglianza per i diversi modelli portano generalmente a individuare in maniera piuttosto soddisfacente i veri valori dei parametri. Le stime tuttavia sembrano più corrette e più stabili nel caso in cui il processo di censura a cui vengono sottoposti i dati sia piuttosto forte. Infine si è notato che la classe delle funzioni di probabilità di tipo esponenziale dipendenti unicamente dal momento di avvenimento dell'evento riesce a stimare anche funzioni di censura molto deboli, e dà risultati generalamente corretti. Tutti e tre le classi di funzioni di censura utilizzati, comunque, danno stime relativamente soddisfacenti per quanto riguarda i veri parametri sottostanti al processo generatore dei dati, mentre a volte le stime dei parametri relativi al processo di censura sono distanti dai veri valori dei parameteri, e danno alcuni problemi numerici nella massimizzazione della verosimiglianza.

Un' eventuale futura analisi potrebbe cercare di valutare l'esistenza di altre classi di funzioni di censura, che non diano problemi numerici nella massimizzazione della verosimiglianza, pur mantenendo la dipendenza dai dati, oltre che dal momento di avvenimento di un evento.

Inoltre, basandosi sui dati ottenuti in questo lavoro, potrebbe essere di interesse uno studio dell'attività di regioni più limitate della

penisola italiana, le cui condizioni geomorfologiche e sismiche siano più omogenee.

Bibliografia

[1] BOSCHI E., DRAGONI M. (2000) *Sismologia*, UTET, Torino

[2] BOSCHI E., GUIDOBONI E., FERRARI G., VALENSISE G., GASPERINI P. (1997) *Catalogo dei Forti Terremoti in Italia dal 461 a.C. al 1990*, ING e SGA, Bologna

[3] COLES S. G.(2001) *An Introduction to Statistical Modeling of Extreme Values*, Springer Verlag, London

[4] COLES S. G., SPARKS R. S. J. (2006)*Extreme value methods for modelling historical series of large volcanic magnitudes*, in MADER H. M., COLES S. G., CONNOR C. B., CONNOR L. J. *Statistics and Volcanology*, GSL, to appear

[5] FURLAN C. (2005) *Processi di punto parametrici e non parametrici per la modellazione di eventi vulcanici estremi in presenza di censura*, Working Paper Series, **6**, Dipartimento di Scienze Statistiche, Università di Padova

[6] Gruppo di lavoro CPTI (2004) *Catalogo Parametrico dei Terremoti Italiani, versione 2004 (CPTI04)*, INGV, Bologna. http://emidius.mi.ingv.it/CPTI/

[7] PACE L., SALVAN A. (2001)*Introduzione alla Statistica. Inferenza, verosimiglianza, modelli*, CEDAM, Padova

[8] PICKANDS J. (1971) *The two-dimensional Poisson process and extremal process*, Journal of applied Probability, **8**, 745-756

[9] SMITH R. L. (1989) *Extreme value analysis of enviromental time series: An application to trend detectuion in ground-level ozone (with discussion)*, Statistical Science, **4**, 367-393