Working Paper Series, N. 6, May 2005

Processi di punto parametrici e non parametrici per la modellazione di eventi vulcanici estremi in presenza di censura

Claudia Furlan
Department of Statistical Sciences
University of Padua
Italy

Abstract: Extreme value theory provides a class of models for the behaviour of stochastic processes at extreme levels. Since volcanic eruptions are (at least in a non-scientific sense) extreme events, it might be hoped that there is some role for the extreme value models in the science of volcanology. In this article we explore such a possibility through a particular catalogue of extreme eruptions registered in the last two millennia. The analysis is based on a particular point process characterization of extremes: it takes into account that historical events in the dataset are less likely to have been recorded than recent events and that this effect seems especially pronounced for events of relatively low magnitude. Ignoring these aspects can lead to a biased estimate of extremal behaviour.

La teoria dei valori estremi fornisce una classe di modelli per il comportamento dei processi stocastici ai livelli estremi. Visto che le eruzioni vulcaniche sono eventi estremi (almeno da un punto di vista non prettamente scientifico), ci si può aspettare che i modelli dei valori estremi possano essere di una qualche utilità nella scienza della vulcanologia. In questo articolo viene esplorata tale possibilità attraverso lo studio di un particolare catalogo di eruzioni estreme registrate negli ultimi due millenni. Le analisi sono basate sulla caratterizzazione degli eventi estremi attraverso un particolare processo di punto: esso tiene conto del fatto che gli eventi più storici del dataset sembrano avere una più bassa probabilità di essere registrati nel catalogo di quelli più recenti e che tale aspetto è più pronunciato per quegli eventi di magnitudine relativamente più bassa. Ignorando tali considerazioni si potrebbe ottenere una stima distorta del comportamento delle eruzioni estreme.

Keywords: Extreme values, Bayesian techniques, Censored data, Volcano eruptions.

Final version (2005-05-05)

Indice

Department of Statistical Sciences
Via Cesare Battisti, 241
35121 Padova
Italy

tel: +39 049 8274168
fax: +39 049 8274170
http://www.stat.unipd.it

Corresponding author:
Claudia Furlan
tel: +39 049 827 4192
furlan@stat.unipd.it

Processi di punto parametrici e non parametrici per la modellazione di eventi vulcanici estremi in presenza di censura

Claudia Furlan
Department of Statistical Sciences
University of Padua
Italy

Abstract: Extreme value theory provides a class of models for the behaviour of stochastic processes at extreme levels. Since volcanic eruptions are (at least in a non-scientific sense) extreme events, it might be hoped that there is some role for the extreme value models in the science of volcanology. In this article we explore such a possibility through a particular catalogue of extreme eruptions registered in the last two millennia. The analysis is based on a particular point process characterization of extremes: it takes into account that historical events in the dataset are less likely to have been recorded than recent events and that this effect seems especially pronounced for events of relatively low magnitude. Ignoring these aspects can lead to a biased estimate of extremal behaviour.

La teoria dei valori estremi fornisce una classe di modelli per il comportamento dei processi stocastici ai livelli estremi. Visto che le eruzioni vulcaniche sono eventi estremi (almeno da un punto di vista non prettamente scientifico), ci si può aspettare che i modelli dei valori estremi possano essere di una qualche utilità nella scienza della vulcanologia. In questo articolo viene esplorata tale possibilità attraverso lo studio di un particolare catalogo di eruzioni estreme registrate negli ultimi due millenni. Le analisi sono basate sulla caratterizzazione degli eventi estremi attraverso un particolare processo di punto: esso tiene conto del fatto che gli eventi più storici del dataset sembrano avere una più bassa probabilità di essere registrati nel catalogo di quelli più recenti e che tale aspetto è più pronunciato per quegli eventi di magnitudine relativamente più bassa. Ignorando tali considerazioni si potrebbe ottenere una stima distorta del comportamento delle eruzioni estreme.

Keywords: Extreme values, Bayesian techniques, Censored data, Volcano eruptions.

1 Introduzione

La vulcanologia è una scienza indispensabile, oltre che per lo studio della composizione terrestre, per quantificare il rischio nelle zone propense alle eruzioni vulcaniche. Infatti, in tali regioni, è necessario sviluppare dei piani di protezione civile efficienti, tenendo conto di alcuni criteri quali la direzione del flusso della lava, la verosimiglianza di una futura eruzione e i valori plausibili per l'intensità degli eventi. Dato che le eruzioni vulcaniche sono il processo naturale più esplosivo che accade nella terra, c'è anche un interesse scientifico a capire quale potrebbe essere lo scenario

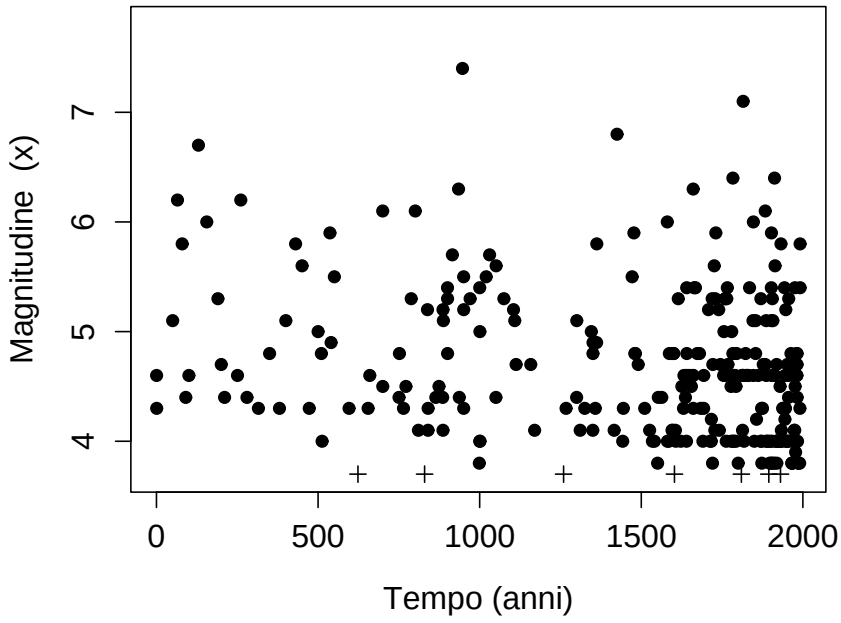
peggiore che possa capitare.
Sebbene sia difficile interpretare da un punto di vista statistico queste problematiche, la loro natura suggerisce l'utilizzo dei modelli dei valori estremi, rispetto a quelli di altre aree statistiche. I modelli vulcanologici di previsione, invece, sono per tradizione deterministici.
In Coles and Sparks (2004) si trova uno studio sulle eruzioni estreme degli ultimi due millenni, in cui vengono modellati i dati di un catalogo di eruzioni vulcaniche degli ultimi due millenni. L'obiettivo era quello di analizzare un problema su dati di eruzioni vulcaniche e mostrare i benefici che si ottengono utilizzando modelli con una parte stocastica. Si proponeva, quindi, di sviluppare un ipotetico modello che fungesse da metafora per la possibile integrazione della conoscenza scientifica all'interno di un modello statistico. Il modello proposto, che verrà illustrato nelle Sezioni 3 e 4, suscitò molto entusiasmo da parte della comunità scientifica dei vulcanologi, nell'ambito del workshop *Statistics in Volcanology* all'Università di Bristol (UK), nel marzo 2004, inducendo ad approfondirlo e svilupparlo.

2 Descrizione dei dati

Il catalogo di eruzioni vulcaniche che verrà utilizzato in questo studio, può essere scaricato dalla pagina web `http://www.edu.gunma-u.ac.jp/~hayakawa/-catalog/2000W` ed è rappresentato nella Figura 1.
Questo catalogo aveva lo scopo di raccogliere tutte le eruzioni vulcaniche, accadute negli ultimi due millenni, di magnitudine maggiore a $x = 3.7$, dove $x = \log_{10} m - 7$ e m è la massa di magma fuoriuscita misurata in kg (Simkin and Siebert, 1994).
Visto che per alcuni eventi non viene indicato il mese o il giorno di accadimento, il tempo è stato casualizzato all'interno del periodo conosciuto (mese o anno). Questo aggiustamento non ha peso sufficiente ad influire sulle analisi che verrranno svolte.
Ci sono poi alcuni eventi, disegnati nella Figura 1 con un +, di cui si ignora la magnitudine, ma di cui si conosce la collocazione temporale e l'eccesso della magnitudine dalla soglia del catalogo. Per semplicità di interpretazione del Grafico, sono state fissate al valore 3.7. Queste poche osservazioni non sono state considerate e non verranno più rappresentate graficamente, anche se in linea di principio si sarebbe potuta utilizzare l'informazione censurata che forniscono.
Ci sono 280 eruzioni che superano la soglia di $x = 3.7$ e 221 la soglia di $x = 4$.

Dall'osservazione delle eruzioni della Figura 1 si potrebbe capire che il tasso di attività vulcanica sia diventato molto più alto negli ultimi 400-500 anni, soprattutto per eventi di bassa magnitudine. Tuttavia, questa considerazione è in netto contrasto con la conoscenza della vulcanologia, che indica un tasso di attività piuttosto costante nel tempo.
La spiegazione sembra stare nel fatto che i vulcanologi hanno una varietà di strumenti per misurare il valore della magnitudine x, anche per quelle eruzioni datate nei secoli, sebbene sia naturale che la loro capacità identificativa decresca con la distanza temporale e la magnitudine dell'evento. Inoltre, le esplorazioni geografiche hanno permesso di raggiungere la copertura totale della Terra solo negli ultimi due secoli: di conseguenza alcune terre, tipo le isole del Pacifico, non possono contare

Figura 1: Catalogo bimillenario di eruzioni vulcaniche eccedenti una magnitudine di $x = 3.7$, nel pianeta Terra. Le osservazioni disegnate con + indicano alcuni eventi di cui si conosce la collocazione temporale ma non la magnitudine (comunque superiore a 3.7).



sulla memoria storica di tali eventi come l'Europa, per esempio.
È ormai noto, infatti, il problema della distorsione nella registrazione delle eruzioni, ipotesi pure sostenuta da Simkin and Siebert (1994).
Di conseguenza, le eruzioni a disposizione sono il risultato di due processi: il verificarsi dell'evento e la sua registrazione nel catalogo. Se non si tenesse conto di quest'ultimo aspetto di identificazione si potrebbe andare incontro a distorsioni non trascurabili nell'interpretazione dell'attività vulcanica della Terra.

3 Caratterizzazione dei Valori Estremi e della censura

Le eruzioni vulcaniche sono il risultato di un'attività sotterranea che diventa a un certo punto estrema. Tale processo è articolato e non completamente prevedibile nei suoi due aspetti principali: la collocazione temporale e la magnitudine degli eventi. Tuttavia, tale imprevedibilità lascia lo spazio alla statistica per cercare di identificare possibili strutture e comportamenti dell'attività vulcanica. La natura

estrema di questo processo suggerisce che la teoria dei valori estremi può fornire degli appropriati strumenti di analisi.

3.1 Processo di Punto

Ci sono vari modi per caratterizzare i comportamenti estremi di un fenomeno, ma il processo di Punto è senz'altro il più flessibile e quello che permette di utilizzare più strumenti inferenziali e di modellazione. Fu Pickands (1971) che propose per la prima volta la teoria seguendo questo approccio, mentre Smith (1989) fu il primo a creare strumenti inferenziali in tale contesto.
Si supponga che $X_1, \ldots, X_n$ sia una sequenza di variabili casuali indipendenti con la stessa funzione di ripartizione F, di cui si vuole modellare la coda. Il processo di Punto viene definito come $P_n = \{(i/(n+1), X_i) : \ i = 1, \ldots, n\}$. Sotto alcune condizione molto generali di F, è ragionevole modellare il processo P_n sulla regione $A_n = [0,1] \times [u, \infty)$, per una soglia u sufficientemente grande, come un processo di Poisson non-omogeneo[1] con una funzione di densità appartenente alla famiglia:

$$\lambda(t,x) = \frac{1}{\sigma}\left[1 + \xi\frac{(x-\mu)}{\sigma}\right]_+^{-1/\xi-1} \tag{1}$$

con $\sigma > 0$ and $a_+ = \max(a, 0)$. Questa rappresentazione delle eccedenze da una soglia è coerente con la rappresentazione classica dei valori estremi basata sui massimi a blocchi; si veda Coles (2001, ch. 7) per i collegamenti e una discussione generale tra le varie rappresentazioni. L'inferenza riguarda la stima dei parametri (μ, σ, ξ) sulla base delle osservazioni nella regione A_u, che si possono denominare come $\{(t_1, x_1), \ldots, (t_m, x_m)\}$. La funzione di verosimiglianza, grazie alle assunzioni del processo di Poisson, risulta:

$$L(\mu, \sigma, \xi; (t_1, x_1), \ldots, (t_n, x_n)) = n_y \exp\left\{-\int_{A_u} \lambda(t,x)\ dtdx\right\} \prod_{i=1}^{n} \lambda(t_i, x_i), \tag{2}$$

dove la costante di proporzionalità n_y, definita come il numero di anni di osservazioni, opera una scala nella parametrizzazione del modello. La funzione di verosimiglianza (2) può essere utilizzata per un'inferenza sia classica che bayesiana.

Il processo di attività vulcanica in analisi, però, non è omogeneo nel tempo, se non forse negli ultimi 400 anni circa. Di conseguenza, non si può applicare la rappresentazione appena descritta all'intera serie storica.

3.2 Processo di Punto con censura

La struttura del processo di Punto permette facilmente di essere estesa in modo da poter modellare anche il processo di identificazione degli eventi, di cui ci si è accorti guardando il Grafico 1.
Si supponga che un evento avvenuto al tempo t con magnitudine x sia registrato nel

[1] Il processo non è costante in x; è costante in t.

catalogo con probabilità $p(t, x)$ (*funzione di presenza*). Allora, il modello di Poisson resta ancora valido, però la funzione di densità diventa:

$$\lambda_M(t, x) = p(t, x)\lambda(t, x) \tag{3}$$

Questa è la metafora a cui si riferiva Stuart Coles e di cui si è parlato nella Sezione 1. Senza una conoscenza proveniente dall'esterno, i dati da soli sarebbero stati insufficienti a formulare un modello: $p(\cdot, \cdot)$ è formulata dalla conoscenza scientifica del processo, mentre $\lambda(\cdot, \cdot)$ è determinata da considerazioni statistiche.

4 Modello parametrico

In questa sezione si espongono le motivazioni e i risultati principali della particolarizzazione della funzione di presenza proposta da Stuart Coles, i cui dettagli possono essere trovati in Coles and Sparks (2004). Questo materiale permetterà di capire lo sviluppo della ricerca intrapresa.

4.1 Specificazione della funzione di presenza

Sempre nell'ambito del workshop di Bristol, di cui si parlava nella Sezione 1, Stuart Coles propose le caratteristiche per le possibili famiglie parametriche per $p(t, x)$:

1. $p(1, x) = 1$ per ogni x: ogni eruzione vulcanica con magnitudine superiore alla soglia, al tempo presente, verrebbe registrata con certezza nel catalogo.

2. $p(t, x)$ è una funzione non-decrescente di t, per ogni fissato x. Cioè, un'eruzione vulcanica di magnitudine x, per qualsiasi x sopra la soglia, ha probabilità maggiore di essere registrata se è avvenuta recentemente e diminuisce con l'aumentare degli anni di distanza.

3. $p(t, x)$ è una funzione non-decrescente di x, per ogni fissato t. Questo significa che ad ogni tempo t, è meno probabile che eventi di magnitudine più grande vengano perduti.

La proposta che fece fu la seguente e a questa ci si riferirà (insieme con $\lambda(t, x)$) con il termine *modello parametrico*[2]:

$$p(t, x) = \left(1 - \frac{v}{x^w}\right) + \frac{v}{x^w} t^b \tag{4}$$

in cui i parametri soddisfano le restrizioni: $b \geq 0$, $w \geq 0$, $v < u^w$ e t è riscalato in modo da appartenere all'intervallo $[0, 1]$.
Ogni parametro del modello ha un particolare significato:

- v determina il grado di censura storica a cui gli eventi sono sottoposti ($v = 0$ implica che non c'è stata censura);

[2] Anche gli altri modelli che verranno proposti in seguito saranno parametrici, ma si intende solo dare un nome di riferimento al modello che si sta per presentare, per semplicità nell'esposizione.

- w indica come il processo di registrazione degli eventi sia differente per diversi livelli della magnitudine ($w = 0$ implica che la censura è uguale per ogni livello di x);

- b indica il tasso di cambiamento del processo di censura a differenti istanti temporali ($t = 1$ implica che il cambiamento è lineare).

Quindi, dei sei parametri del modello tre (μ, σ, ξ) corrispondono alle proprietà dei valori estremi del processo di attività vulcanica (perfettamente registrato), e tre (v, w, b) corrispondono al processo di registrazione degli eventi.

Per procedere, occorre fissare il valore della soglia che si intende utilizzare. È noto che le argomentazioni asintotiche per l'utilizzo del processo di Poisson sono valide per grandi valori della soglia. Ma quanto grandi? Ci si trova di fronte alla scelta di un compromesso tra la varianza e la distorsione di stima. Infatti, per bassi valori di u si dispongono di più punti a disposizione, facendo così diminuire la varianza di stima; tuttavia, l'affidabilità delle argomentazioni asintotiche diventa discutibile, introducendo una distorsione. Ovviamente, per valori grandi di u, si ottiene il contrario. Quello che si fa in pratica, è di scegliere quel valore di u in modo da ottenere una certa stabilità tra le stime dei parametri ottenute con u e quelle ottenute con valori di $u^* > u$ (tenendo conto anche della variabilità di stima). In Coles and Sparks (2004) viene stimato, con la massima verosimiglianza, un processo di Poisson omogeneo negli ultimi 400 anni, per vari valori di u, in quanto in tale periodo sembra che il processo non subisca l'effetto di censura nella registrazione degli eventi (si veda Figura 1). Poi, si confrontò l'andamento delle stime dei parametri (μ, σ, ξ) in relazione al livello di soglia adottato. Purtroppo, come spesso succede in questo tipo di analisi, i risultati non inducono a delle scelte precise. I parametri corrispondenti ad una soglia di 4 sono coerenti con quelli corrispondenti a valori di $u \geq 5.1$; inoltre, fino al valore $u = 5.1$ i cambiamenti dei parametri sembrano sistematici. Di conseguenza, fu proposto il valore $u = 5.1$ come quello più plausibile, ma si continuò ad utilizzare entrambi i valori $u = 4$ e $u = 5.1$ per confrontare i risultati.
In questa Sezione verranno riportati i risultati di Coles and Sparks (2004) solo per $u = 4$ (nella Sezione 7 saranno riportati quelli per $u = 5.1$), in quanto la discussione della scelta della soglia viene rimandata per dare spazio alla discussione della scelta della funzione di presenza; infatti, per studiare più in dettaglio il processo di identificazione degli eventi è meglio utilizzare una soglia (tra le plausibili) bassa, per aver eventi di più varia intensità.

4.2 Risultati

In questo modello i parametri μ, σ, ξ corrispondono all'attuale processo di attività vulcanica, che si assume essere perfettamente registrato.
Nella Tabella 1 sono riportati le stime di massima verosimiglianza e gli errori standard di tali parametri, ad una soglia di 4, sia nel caso di un processo di Poisson omogeneo negli ultimi 400 anni, che di un processo di Poisson con funzione di presenza sull'intera serie storica.

Si considerino, ora, le analisi dell'intera serie storica. I parametri v, w, b sono significativamente diversi da 0, di cui v e b in maniera decisa. Il fatto che v e w

siano risultati statisticamente diversi da 0 è molto importante perché confermano le ipotesi di sotto-identificazione degli eventi nel tempo ($v \neq 0$) e che tale processo sia più forte per quegli eventi di bassa magnitudine ($w \neq 0$). Queste conclusioni sono ulteriormente supportate dal confronto delle relative massime log-verosimiglianze, che sono riportate nella Tabella 2.

Nella Figura 2, a sinistra, sono state disegnate le stime della funzione di presenza per 2 differenti valori di magnitudine: 4.5 e 7. Tali stime suggeriscono che, per i primi 1500 anni circa, la probabilità di registrare un evento è costante a differenti livelli, crescendo poi rapidamente. Il rapido incremento è probabilmente dovuto all'espansione geografica e allo sviluppo scientifico, cioè alle nuove scoperte geografiche e alla disponibilità di nuove attrezzature che hanno avuto luogo negli ultimi 500 anni.
La stima della funzione di presenza all'anno 0 è circa 0.09 e 0.24 rispettivamente per le magnitudine $x = 4.5$ e $x = 7$: questa diversità del fenomeno di identificazione degli eventi per diverse magnitudini era già stato preannunciato dal risultato di non-nullità statistica del parametro w.
Tuttavia, una crescita convessa, continua e monotona nel tempo per la funzione di presenza non è forse la scelta più adatta, in quanto permette l'omogeneità del processo solo al tempo dell'ultima osservazione, mentre i vulcanologi asseriscono che negli ultimi tempi sono in grado di identificare tutti gli eventi. Per capire i limiti di questa funzione di presenza, in relazione ai dati osservati, si osservi il Grafico 2 di destra, in cui la stima di detta funzione, per magnitudine pari a 4.5 e a 7, è stata riscalata in modo da permetterne la sovrapposizione al grafico originale dei dati. Si può notare come il processo risulti stazionario (o quasi), per lo meno ad un primo sguardo, negli ultimi 400 anni circa. La formulazione (4) di $p(t, x)$, invece, forza la probabilità di presenza ad aumentare sempre fino a raggiungere quota 1 (che nella Figura 2 di destra corrisponde a 8) nell'ultimo istante considerato. Quindi, sarebbe

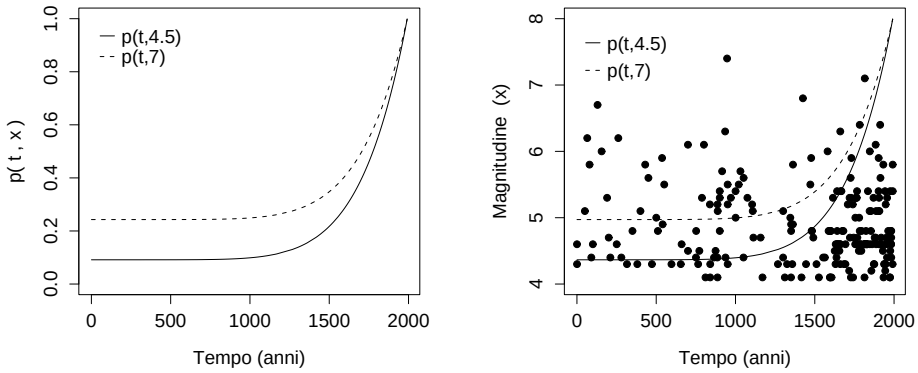
	μ	σ	ξ	v	w	b
Ultimi 400	2.444	1.587	-0.317	-	-	-
anni	(0.247)	(0.251)	(0.0052)	-	-	-
Tutti gli	3.289	1.124	-0.239	1.691	0.413	6.971
anni	(0.183)	(0.158)	(0.047)	(0.552)	(0.219)	(1.25)

Tabella 1: Stime di massima verosimiglianza ed errori standard (tra parentesi) del processo di Punto omogeneo (stimato negli ultimi 400 anni) e con censura (stimato su tutta la serie storica), applicato ai dati delle eruzioni vulcaniche con una soglia di $x = 4$.

	Senza vincoli	$v = 0$	$w = 0$
log-ver	-820.96	-890.81	-823.39

Tabella 2: Valori di massima log-verosimiglianza per diversi sottomodelli, per una soglia di $x = 4$.

Figura 2: Funzione di presenza $p(t, x)$ per $x = 4.5$ e $x = 7$, a sinistra, e riscalata per permettere la sovrapposizione ai dati, a destra.



più opportuna una funzione che permettesse un andamento costante durante l'ultimo periodo della serie storica. Tale parte verrà sviluppata nella Sezione 5.

Sebbene il processo di Poisson sia la caratterizzazione migliore per i valori estremi, ciò non implica che la sua formulazione sia la più facile da essere interpretata. Per esempio, si può considerare la trasformazione η dei parametri μ, σ, ξ, che indica, per le proprietà del processo di Poisson, il tasso annuo con cui la soglia u viene superata:

$$\eta = \int_{x=u}^{+\infty} \int_{t=0}^{1992} \lambda(t, x)\ dx\ dt = \left[1 + \xi \frac{(u - \mu)}{\sigma}\right]_+^{-1/\xi}. \tag{5}$$

Sostituendo ai parametri le relative stime di massima verosimiglianza si ottiene $\hat{\eta} = 0.309$ considerando gli ultimi 400 anni e $\hat{\eta} = 0.504$ considerando il modello sull'intera serie storica con la funzione di presenza. Questi due valori sono piuttosto differenti: infatti, con il primo si fissa il periodo di non censura negli ultimi 400 anni, mentre con il secondo solo nel momento dell'ultima osservazione.
Inoltre, si può considerare la distribuzione delle eruzioni X la cui magnitudine supera la soglia u:

$$P(X > x) = [1 + \xi(x - u)/\sigma]_+^{-1/\xi}. \tag{6}$$

Combinando le equazioni (5) e (6), si ottiene che il livello $x > u$ viene superato una volta ogni $r(x)$ anni, dove:

$$r(x) = \eta^{-1}[1 + \xi(x - u)/\sigma]_+^{1/\xi}. \tag{7}$$

Nella terminologia comune, si dice che $r(x)$ è il periodo di ritorno associato al livello x.
Si rimanda a Coles and Sparks (2004) per le curve del livello di ritorno sia per il processo di Poisson omogeneo sugli ultimi 400 anni che per il processo di Poisson con la funzione di presenza su tutta la serie storica.

Si noti che, nel caso in cui $\xi < 0$, la distribuzione in (6) è limitata superiormente dal valore

$$X_{max} = \mu - \sigma/\xi. \tag{8}$$

In questo caso, $\hat{\xi} < 0$ e, quindi, sostituendo le stime di massima verosimiglianza, si ottiene $\hat{X}_{max} = 7.45$ per il processo di Poisson omogeneo sugli ultimi 400 anni, e $\hat{X}_{max} = 7.992$ nel caso del processo di Poisson non omogeneo sull'intera serie storica: questo vuol dire che, secondo questi due modelli, valori dopo 7.45 e 7.992, rispettivamente, hanno probabilità nulla di accadere.
Dalle differenti curve dei livelli di ritorno e dei limiti superiori delle distribuzioni in (6), si può capire (anche se la funzione di presenza utilizzata non è, forse, la più adeguata) che la distorsione nel non considerare l'effetto della censura è piccola ai livelli estremi delle eruzioni ma decisiva per i livelli più bassi, e che l'utilizzo della funzione di presenza legittima valori di eruzioni più estreme.

5 Modello con punto di cambio

Le tecniche bayesiane offrono una valida alternativa alla stima di massima verosimiglianza nell'ambito dei valori estremi. Esse permettono, infatti, vista la scarsità dei dati, di includere dell'informazione tramite una distribuzione a priori. Poi, il risultato di un'analisi bayesiana -la distribuzione a posteriori- dà luogo a un'inferenza più completa di quella che si ottiene con la stima di massima verosimiglianza. In particolare, visto che l'obiettivo di un'analisi dei valori estremi è quello di stimare la probabilità che eventi futuri raggiungano certi livelli estremi, è naturale utilizzare la distribuzione predittiva. Inoltre, le tecniche MCMC permettono di stimare modelli con strutture parametriche più articolate, anche quando il numero di parametri è esso stesso un parametro.

5.1 Specificazione della funzione di presenza

Si è visto nella Sezione 4.2 come la funzione di presenza proposta abbia insite delle limitazioni, in quanto non riesce a rappresentare efficacemente l'evoluzione temporale dell'effetto di censura nella registrazione delle eruzioni vulcaniche. Alcune proposte per una formulazione alternativa di $p(t, x)$ potrebbero essere una funzione monotona non decrescente a forma di S (qualcosa del tipo costante-crescente-1 rispetto al tempo t) o una funzione monotona non decrescente con 2 gradini (costante-1). Le due funzioni devono necessariamente terminare con la costante 1 nell'ultimo periodo per la definizione di probabilità.
Si è preso in considerazione la seconda proposta, perché permette di stimare un modello con un punto di cambio e idealmente è adatta a possibili generalizzazioni (ad esempio la stima di più punti di cambio o di un numero ignoto di punti di cambio), di cui si parlerà nella Sezione 6.

A questo punto non resta che formulare $p(t, x)$. La scelta fatta è la seguente:

$$p(t,x) = \begin{cases} \dfrac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} & t \leq k \\ 1 & t > k, \end{cases} \tag{9}$$

con $k \in [0, 1992]$. Le proprietà di (9) sono:

1. $p(t, x)$ è una funzione a gradini, rispetto al tempo t;

2. non è prevista alcuna censura nell'identificazione degli eventi dall'anno k;

3. per ogni livello della magnitudine x, la funzione di presenza è costante, prima dell'anno k;

4. per $\beta > 0$, la probabilità che una eruzione venga registrata nel catalogo cresce con la magnitudine, prima dell'anno k.

Ora, nel modello con un punto di cambio, i parametri sono:

1. (μ, σ, ξ) per $\lambda(t, x)$,

2. (α, β, k) per $p(t, x)$.

A questo punto, sembra necessaria la specificazione delle distribuzioni a priori dei parametri e a tal proposito viene colta l'occasione per riscrivere il *modello con un punto di cambio*, suddividendolo in due livelli: la distribuzione delle osservazioni e quella dei parametri:

1. $(t, x) \sim \lambda_M = \lambda(t, x)p(t, x)$,

2. $\mu \sim N(0, 10^3) \quad \log \sigma \sim N(0, 10^3) \quad \xi \sim N(0, 10^3)$
   $\alpha \sim N(0, 10^3) \quad \beta \sim N(0, 10^3)$
   $k \propto k(1992 - k)$

dove $\lambda(t, x)$ è formulata in (1) e $p(t, x)$ in (9); la distribuzione a priori di k è costruita in modo da non preferire un valore di k troppo vicino agli estremi del dominio (Green, 1995). Le distribuzioni a priori sono vaghe.
La verosimiglianza diventa:

$$
\begin{aligned}
&L(\mu, \sigma, \xi, \alpha_1, \beta_1; (t_1, x_1), \ldots, (t_n, x_n)) = \\
&\exp\left\{ \int_{x=u}^{+\infty} \int_{t=0}^{k_1} \lambda(t, x) \frac{\exp(\alpha_1 + \beta_1 x)}{1 + \exp(\alpha_1 + \beta_1 x)} \, dt \, dx \right\} \times \\
&\exp\left\{ \int_{x=u}^{+\infty} \int_{t=k_1+1}^{1992} \lambda(t, x) \, dt \, dx \right\} \times \\
&\prod_{i:0<t_i\leq k_1} \lambda(t_i, x_i) \frac{\exp(\alpha_1 + \beta_1 x_i)}{1 + \exp(\alpha_1 + \beta_1 x_i)} \prod_{i:k_1<t_i\leq 1992} \lambda(t_i, x_i)
\end{aligned}
$$

Come strumenti per le procedure inferenziali sono stati presi in considerazione gli algoritmi MCMC (si veda Gilks *et al.* (1996) per i dettagli).

5.2 Studi di simulazione

Prima di passare a stimare il modello occorre svolgere degli studi di simulazione per appurare se gli algoritmi MCMC che si utilizzeranno riescano a stimare correttamente i parametri del modello con un punto di cambio. Per fare ciò, si fissano i valori dei parametri, si simula il processo di generazione dei dati, si applica la procedura di stima e, infine, si controlla se le catene di Markov convergono ai relativi valori fissati dei parametri.

5.2.1 Simulazione dei dati

I dati oggetto di studio sono, come già detto, il risultato di due processi: quello di eruzione vulcanica e quello di identificazione degli eventi (censura dell'informazione). La simulazione si articola in due fasi:

1. generazione di un processo di eruzioni vulcaniche omogeneo nel tempo, per esempio: $z_i \sim Ga(1/2, 3)$, dove $i = 1, \ldots, 1992$ e di cui si tengono solo quelle eccedenti una soglia fissata a $u = 4$ (si veda il Grafico 3 a sinistra);

2. omissione di una parte delle eruzioni secondo la funzione di presenza (9), con valori, ad esempio, $\alpha = -4,\ \beta = 0.6,\ k = 1400$. La procedura utilizzata è la seguente:

   - ad ogni eruzione simulata z_i è stata calcolata la probabilità di comparire nel catalogo tramite la funzione di presenza (9);
   - è stato condotto un esperimento casuale che ha riprodotto il processo di registrazione-perdita degli eventi tramite la distribuzione $Bi(1, p(t, z_i))$.

   Le eruzioni che hanno avuto esito 1 nell'esperimento binomiale sono state riportate nel Grafico 3, a destra.

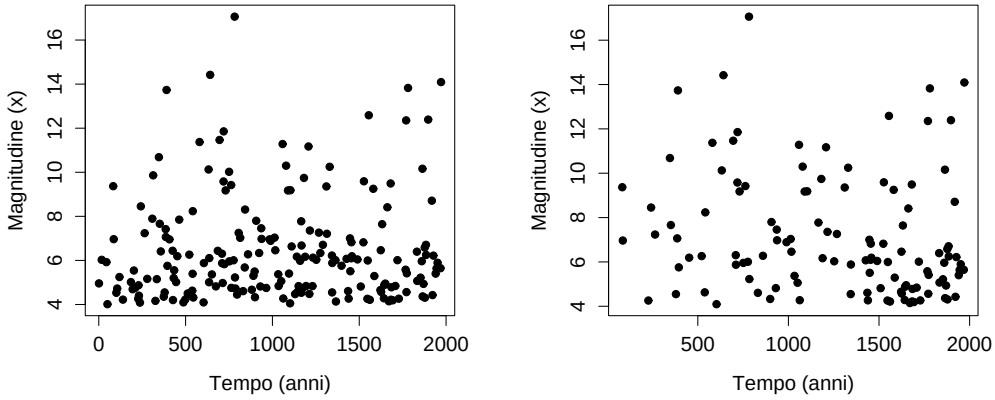
5.2.2 Valori di riferimento tramite stime di massima verosimiglianza

Questo studio di simulazione non è condotto tradizionalmente perché non si conoscono i veri valori di μ, σ, ξ. Si è pensato, quindi, di stimarli nel dataset completo (prima di applicare la funzione di presenza) tramite la massima verosimiglianza (di un processo di Poisson omogeneo).
Le stime e i relativi intervalli di confidenza sono rappresentati nella Tabella 3.

D'ora in avanti, si considereranno questi tre valori come i veri valori per μ, σ, ξ, ma per ricordare che sono valori di "riferimento" verranno riportati con il simbolo *.

STIMA CON ALGORITMO MCMC
A questo punto, avendo tutti i valori veri o di riferimento a cui devono tendere le catene markoviane, si è utilizzato un algoritmo Metropolis-Hastings (si veda Gilks *et al.* (1996) per i dettagli), che viene presentato qui di seguito, per stimare i parametri del modello $(\mu, \sigma, \xi, \alpha, \beta, k)$.

Figura 3: A sinistra, simulazione di un processo di eruzioni vulcaniche omogeneo nel tempo, eccedenti la soglia di 4; a destra, le eruzioni che vengono registrate nel catalogo dopo aver applicato la funzione di presenza (9), con $\alpha = -4,\ \beta = 0.6,\ k = 1400$.



Visto che i parametri (μ, σ, ξ) in un processo di punto sono piuttosto correlati tra loro, occorre trovare uno stratagemma per ottenere delle catene markoviane che si comportino correttamente. Per esempio, al posto dei tre parametri, se ne possono aggiornare due di essi più una loro trasformazione, che non sia correlata con essi. Alla fine, si potrà ottenere la catena markoviana per il terzo parametro escluso, invertendo la trasformazione utilizzata.
Vengono scelti, allora, due parametri tra i tre del processo di punto, ad esempio μ, ξ che variano in tutto $\mathbb{R}$ e, quindi, sono agevoli da aggiornare.
Si introduca la seguente quantità:

$$\rho = P[Y > u] = 1 - \exp\left\{-\frac{1}{n}\left(1 + \frac{\xi}{\sigma}(u - \mu)\right)^{-1/\xi}\right\}$$

dove ρ è la probabilità che un'eruzione superi la soglia u. Questa quantità viene, in genere, stimata con una certa fiducia; non è agevole, però, da aggiornare in un

Tabella 3: Stime di massima verosimiglianza, standard error e intervalli di confidenza al 95% per μ, σ, ξ, per i dati simulati.

Par.	stima	st. error	int. conf.
μ	-2.376	1.141	$(-4.612, -0.139)$
σ	2.923	0.744	$(1.465, 4.381)$
ξ	-0.051	0.073	$(-0.194, 0.092)$

algoritmo MCMC perché può variare solo tra 0 e 1. Si consideri allora la seguente trasformata che varia su tutto $\mathbb{R}$:

$$\zeta = \log \frac{\rho}{(1-\rho)}.$$

I valori inziali dei parametri vengono fissati a $\mu_0 = 1, \sigma_0 = 1, \xi_0 = -0.001$. I parametri vengono aggiornati uno alla volta tramite una passeggiata casuale. Ad ogni passo le proposte per i parametri μ, ζ, ξ hanno distribuzione normale, di media il valore corrente dei parametri e di varianza fissata:

$$\begin{array}{rcl} \mu_p & \sim & N(\mu_t, 1) \\ \zeta_p & \sim & N(\zeta_t, 1) \\ \xi_p & \sim & N(\xi_t, 0.1) \end{array} \tag{10}$$

dove t è il numero di iterazione appena compiuta (quindi sta a indicare il valore attuale nella catena markoviana) e il pedice $_p$ sta a indicare il valore proposto.
Alla fine, si riottiene la catena di Markov per σ operando la trasformazione inversa a ζ.

Si considerino ora i parametri della funzione di presenza. Per ridurre la correlazione tra α e β è stata operata la seguente trasformazione:

$$\alpha^+ = \alpha + \beta\bar{y}$$

dove $\bar{y}$ è la media delle simulazioni y_i. I valori inziali sono stati fissati a $\alpha_0 = 3$ e $\beta_0 = -0.6$. Per aggiornare α^+ e β ci si è avvalsi, ancora, di una passeggiata casuale governata da una normale, di media il valore corrente del parametro e di varianza prefissata:

$$\begin{array}{rcl} \alpha_p^+ & \sim & N(\alpha_t^+, 0.1) \\ \beta_p & \sim & N(\beta_t, 0.3). \end{array} \tag{11}$$

Per k, invece, il valore iniziale è stato fissato a $k_0 = 600$ e la passeggiata casuale è descritta da una Uniforme discreta su tutto il periodo temporale:

$$k_p \sim U[1, 1992]. \tag{12}$$

L'efficienza dell'algoritmo verrà diminuita, perché dovendo attingere i valori di k da un ampio spazio, che comprende anche valori completamente inaccettabili, si avrà una catena con la tendenza a soffermarsi sui valori accettati. Inoltre, visto che la scelta del punto di cambio influenza tutti gli altri parametri, si può avere difficoltà ad abbandonare i massimi locali. Tuttavia, in questo modo, non si corre il rischio che la catena non visiti delle zone dello spazio parametrico, dove si potrebbe trovare un punto di cambio più plausibile.

Le probabilità di accettazione coinvolte nella scelta dei valori proposti, sono il minimo tra 1 e il rapporto dei prodotti della verosimiglianza per la distribuzione a priori, nel caso rispettivamente del valore proposto e del valore corrente. Infatti, come si è descritto, sono state usate delle funzioni proposta simmetriche (normali) rispetto al valore corrente per i parametri $(\mu, \zeta, \xi, \alpha^+, \beta)$ e uniforme per k; tali quantità si semplificano nel rapporto delle probabiltà di accettazione.

Le catene di Markov derivanti sono rappresentate in Figura 4. Si può affermare, che i valori veri di α, β, k e quelli di riferimento per μ, σ, ξ vengono stimati soddisfacentemente: si veda la Tabella 4. Confrontando gli intervalli di credibilità al 95% della Tabella 4 con gli intervalli di confidenza al 95% della Tabella 3, si può notare come essi siano simili per σ e ξ, mentre sia più grande quello di credibilità per μ.

5.3 Risultati

Una volta appurata l'affidabilità delle tecniche discusse nella Sezione 5.2, sono state applicate ai dati reali. Il numero di iterazioni svolte dall'algoritmo è stato 20000 e il periodo di burn-in è stato fissato a 10000.
Nel Grafico 5 sono rappresentate le catene markoviane per i parametri $(\mu, \sigma, \xi, \alpha, \beta)$ e l'istogramma per k. Le catene hanno un buon mixing e sembrano aver raggiunto facilmente la stabilità.
Nella Tabella 5 vengono presentati le stime dei valori attesi a posteriori di $\mu, \sigma, \xi, \alpha, \beta, k$ e gli intervalli di credibilità al 95%. La log-verosimiglianza in corrispondenza alle stime dei valori attesi a posteriori dei parametri è -820.074. I risultati sono coerenti con quelli ottenuti con l'approccio di massima verosimiglianza (per $u = 4$) e presentati nelle Tabelle 1 e 2.

Si passi, ora, a considerare la trasformazione η dei parametri μ, σ, ξ indicata in

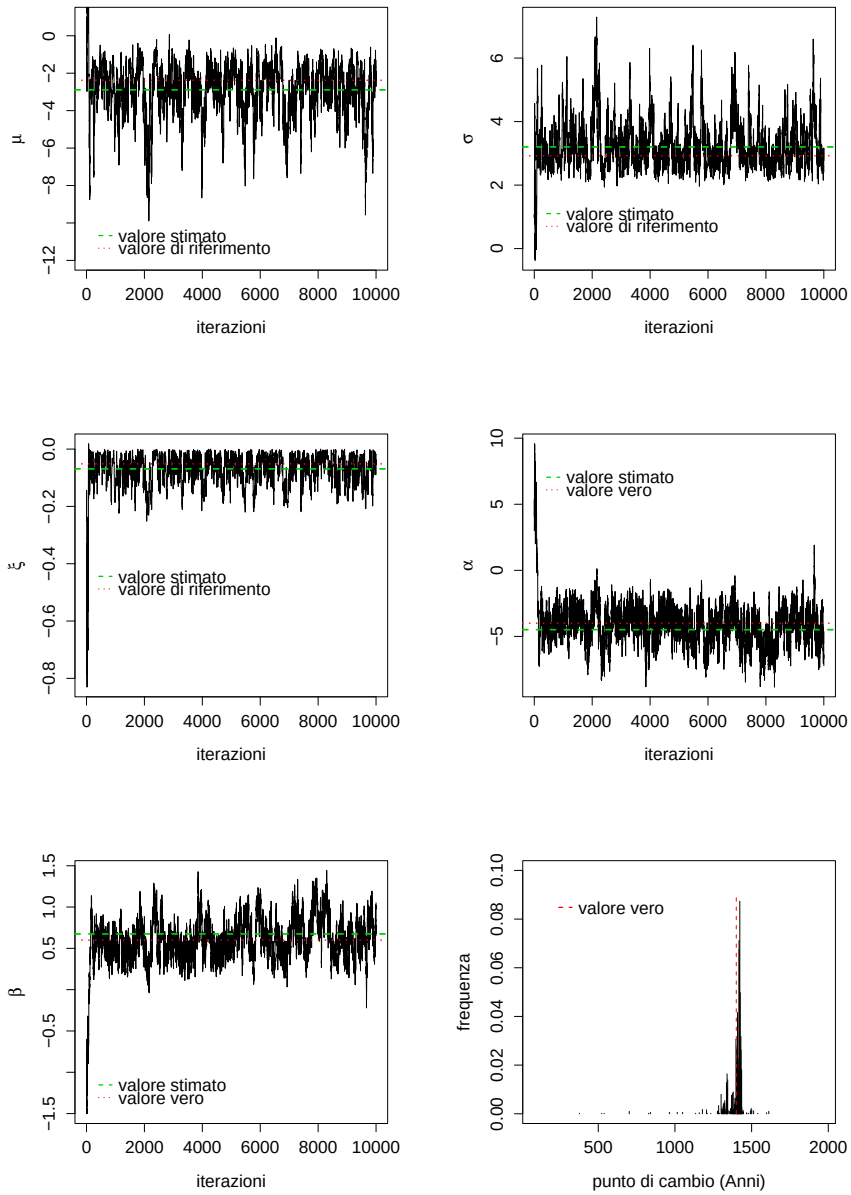
	vero	val. att.	int. cred.
μ	-2.376^*	-2.889	$(-6.001, -0.905)$
σ	2.923^*	3.201	$(2.271, 5.153)$
ξ	-0.051^*	-0.069	$(-0.187, -0.002)$
α	-4	4.489	$(-7.185, -1.875)$
β	0.6	0.674	$(0.224, 1.158)$
k	1400	1403	$(1305, 1435)$

Tabella 4: Stima dei valori attesi a posteriori dei parametri $\mu, \sigma, \xi, \alpha, \beta, k$ e loro intervalli di credibilità al 95% (sono riportati i quantili al 2.5% e al 97.5%). Il simbolo * sta ad indicare che il valore vero del parametro è in realtà un valore di riferimento, perché stimato con la massima verosimiglianza.

Parametro	val. att.	int. cred.
μ	2.522	$(2.079, 2.914)$
σ	1.424	$(1.067, 1.852)$
ξ	-0.258	$(-0.337, -0.174)$
α	-3.249	$(-6.081, -0.796)$
β	0.388	$(-0.121, -0.799)$
k	1584	$(1560, 1596)$

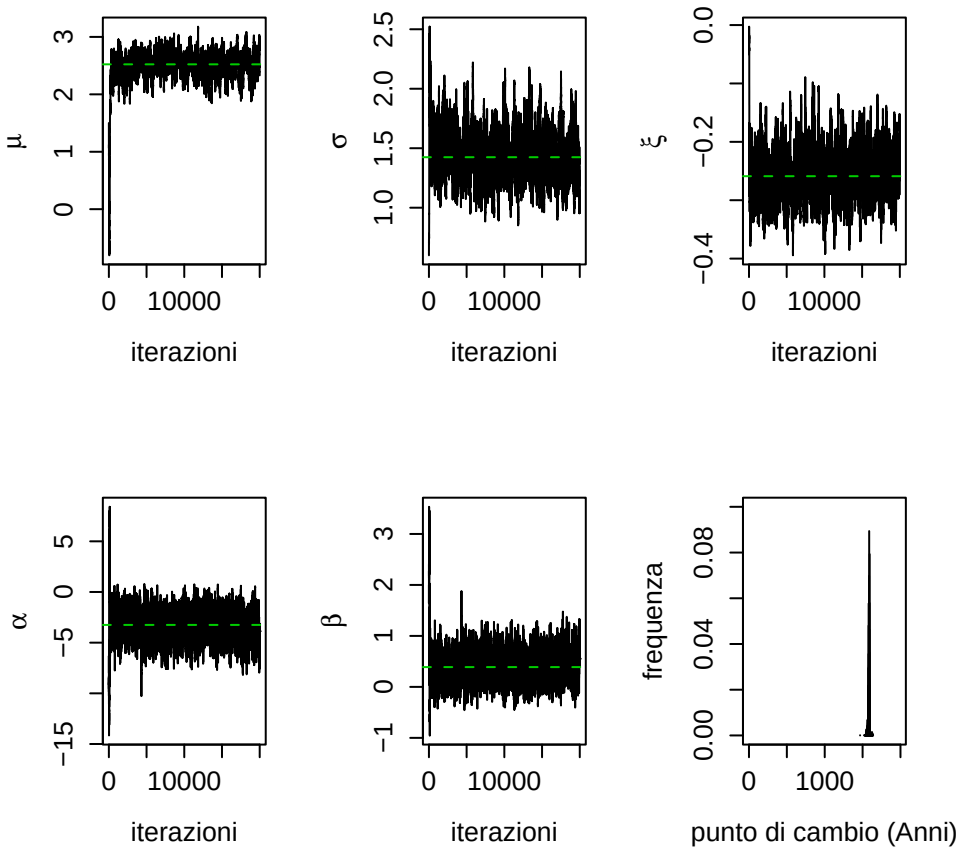
Tabella 5: Stime dei valori attesi a posteriori ed intervalli di credibilità al 95% per $\mu, \sigma, \xi, \alpha, \beta, k$, per dati corrispondenti a $x > 4$.

Figura 4: Studio di simulazione. Catene di Markov generate dall'algoritmo MCMC per i parametri (da sinistra a destra e dall'alto al basso) $\mu, \sigma, \xi, \alpha, \beta$. L'ultimo grafico in basso a destra è la densità a posteriori stimata di k.



(5) che sta ad indicare il tasso annuo con cui la soglia u viene superata. Grazie agli output delle catene di Markov ottenute per ogni parametro, si può stimare η tenendo conto della variabilità di queste catene, cioè della variabilità delle densità

Figura 5: Applicazione al catalogo. Catene markoviane per i parametri $(\mu, \sigma, \xi, \alpha, \beta)$ e l'istogramma per k.



stimate a posteriori dei parametri:

$$\hat{\eta} = \frac{1}{m - m_0} \sum_{j=m_0+1}^{m} \left[1 + \xi_j \frac{(u - \mu_j)}{\sigma_j}\right]_+^{-1/\xi_j},$$

dove m_0 indica la fine del burn-in e m il numero totale delle iterazioni. Si è ottenuto $\hat{\eta} = 0.3$ che si discosta un pò dal valore ottenuto con il modello parametrico (0.528), ma che è in accordo con il valore ottenuto con la stima del processo omogeneo negli ultimi 400 anni (0.309). Infatti, nel modello con punto di cambio, il periodo con assenza di censura viene stimato essere proprio quello degli ultimi 400 anni.
Essendo $\hat{\xi}$ e il relativo intervallo di credibilità negativo, ha senso calcolare il valore massimo della distribuzione stimata di X, secondo l'equazione (8), attraverso

l'output MCMC:

$$\hat{X}_{max} = \frac{1}{m - m_0} \sum_{j=m_0+1}^{m} [\mu_j - \sigma_j/\xi_j]$$

e risulta $\hat{X}_{max}$=8.072. Tale valore risulta simile a quello ottenuto con il modello parametrico (7.992).
Nel Grafico 6 viene riportata la curva del livello di ritorno per il modello con un punto di cambio. Le linee tratteggiate corrispondono agli intervalli di credibilità al 95% per ogni valore del periodo di ritorno. I pallini neri corrispondono alle stime empiriche di (7); dall'evidente relazione, infatti, tra (7) e (6) si può scrivere:

$$\begin{aligned} \hat{r}(x) = \hat{\eta}^{-1}[\hat{P}(X > x)]^{-1} &= \left(\frac{\#\{X_i : X_i > u\}}{1992}\right)^{-1} \left(\frac{\#\{X_i : X_i > x\}}{\#\{X_i : X_i > u) + 1\}}\right)^{-1} \\ &= \left(\frac{221}{1992}\right)^{-1} \left(\frac{x}{221+1}\right)^{-1} \end{aligned}$$

per $x = 1, \ldots, 221$, in quanto ci sono 1992 anni. Tale grafico è in accordo con la relativa curva del livello di ritorno del modello parametrico, presentato in Coles and Sparks (2004), per il modello con la funzione di presenza: solo per valori alti di magnitudine (circa $x > 6.5$) non si hanno discrepanze tra le stime empiriche e quelle previste dal modello. Di conseguenza, il modello coglie la presenza di una censura per valori di $x < 6.5$ (circa). Il modello con un punto di cambio ha un intervallo di credibilità un pò più ampio, superiormente, dell'intervallo di confidenza del modello parametrico.

Gli intervalli di credibilità (al 95%) dei parametri μ, σ, ξ, α non contengono lo 0, quindi si possono considerare rilevanti nella spiegazione del fenomeno. Il parametro β, invece, ammette lo 0 nell'intervallo di credibilità al 95%, anche se la stima della probabilità che β sia maggiore di 0 è pari a 0.93. Nella Tabella 6 vengono presentate le log-verosimiglianze in corrispondenza delle stime dei valori attesi a posteriori dei parametri: la presenza di β ne comporta un aumento di 1.371.
Come nel modello con un punto di cambio β è il parametro in corrispondenza della magnitudine, nel modello parametrico è w a coprire questo ruolo: entrambi hanno una significatività debole.

Nel Grafico 7, infine, vengono presentate le funzioni di presenza per il modello con punto di cambio, per alcuni valori della magnitudine. Si noti come, all'aumentare della magnitudine di un evento cresca, nel primo periodo, la probabilità di essere registrato. Le differenze non sono, comunque, troppo marcate, come testimonia la debole significatività del parametro β.

	Senza vincoli	$\beta = 0$
log-ver	-820.074	-821.391

Tabella 6: Log-verosimiglianze (in corrispondenza delle stime dei valori attesi a posteriori dei parametri) per il modello completo e il sub-modello con $\beta = 0$, per dati corrispondenti a $x > 4$.

Figura 6: Curva del livello di ritorno per il modello con un punto di cambio. Le linee tratteggiate corrispondono agli intervalli di credibilità al 95% per ogni valore del periodo di ritorno. I pallini neri corrispondono alle stime empiriche.

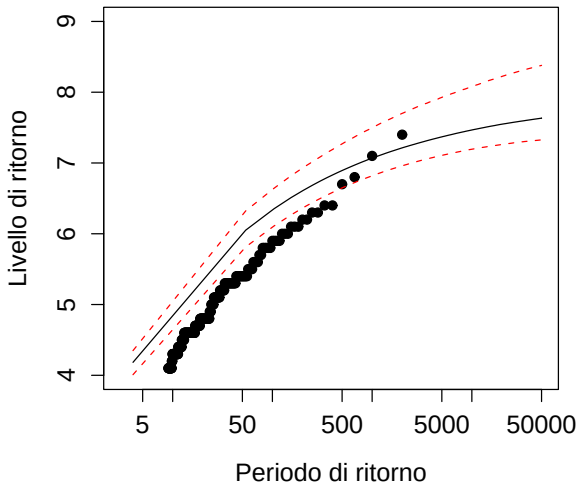
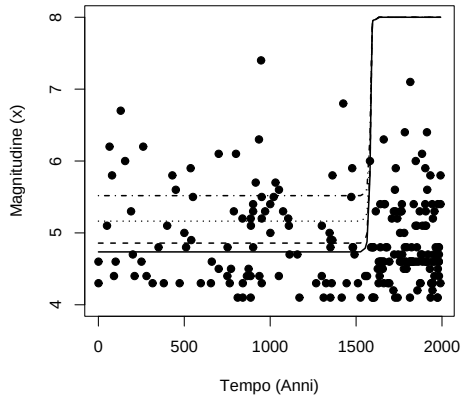
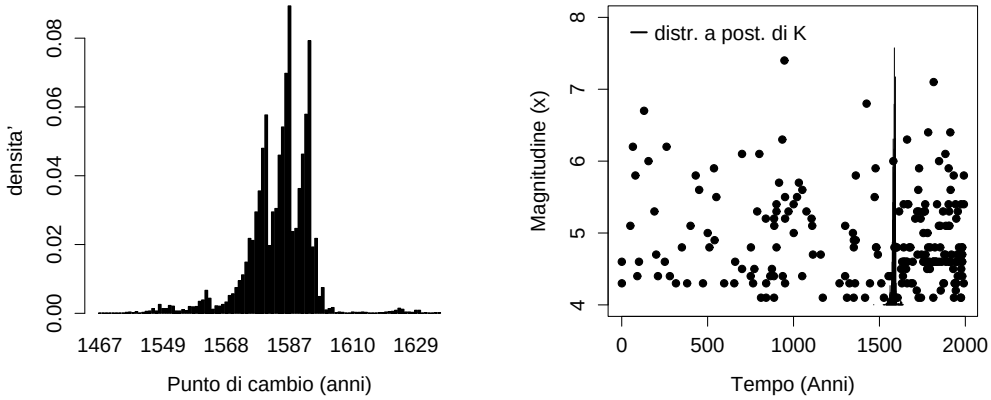


Figura 7: Funzioni di presenza $p(t, x)$ del modello con punto di cambio, corrispondenti a $x = 4.5$ (—), $x = 5$ (- -), $x = 6$ ($\cdots$), $x = 7$ (- $\cdot$ -). I dati utilizzati sono corrispondenti a $x > 4$.



In definitiva, il tempo assume più importanza nel processo di censura delle informazioni, però tale censura risente anche dell'intensità dei fenomeni da registrare.

Figura 8: Stima della densità a posteriori di k, da sola (a sn) e riscalata per essere inserita nel Grafico dei dati originari (a ds). I dati utilizzati sono corrispondenti a $x > 4$.



5.3.1 Considerazioni

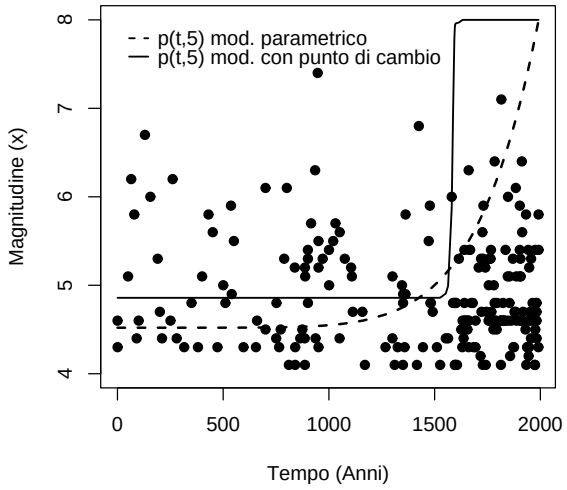
È interessante comprendere un pò più in dettaglio le caratteristiche della distribuzione di k, in quanto esso rappresenta il punto di forza della scelta per il modello con punto di cambio. Nel Grafico 8 si trova la stima della densità a posteriori di k, da sola e riscalata per essere inserita nel grafico dei dati originari. La moda della distribuzione stimata di k è 1587, mentre l'intervallo di credibilità per k (al livello del 95%) è $(1560, 1596)$ (si veda la Tabella 5): la sua ampiezza è di circa 35 anni, un periodo piuttosto breve rispetto ai 1992 possibili valori. La localizzazione del punto di cambio sembra essere, quindi, piuttosto evidente.

Il valore della log-verosimiglianza (in corrispondenza alle stime dei valori attesi a posteriori dei parametri) è -820.786 con un numero di parametri pari a 6 $(\mu, \sigma, \xi, v, w, b)$; il valore massimo della log-verosimiglianza del modello parametrico è, invece, -820.96 con un numero di parametri pari a 6 $(\mu, \sigma, \xi, \alpha, \beta, k)$. Questo sta a indicare che i due approcci, per il momento si equivalgono nella spiegazione dei dati. Sembra opportuno, quindi, andare a vedere da vicino l'andamento della stima di $p(t, x)$, tra i due approcci.

Nel Grafico 9 sono rappresentate le due funzioni, di cui sopra, valutate alla magnitudine 5 e riscalate in modo da poter essere sovrapposte ai dati originari[3]. Entrambe sono, in un certo senso, costanti per i primi 1500 anni circa e successivamente hanno una rapida crescita nello stesso periodo. Tuttavia, hanno delle differenze concettuali.

[3]Dato che k ha una distribuzione a posteriori, bisogna operare una scelta per la rappresentazione grafica di $p(t, x)$. Si sarebbe potuto disegnare il salto in corrispondenza della moda di k oppure, come è stato riportato, disegnare la funzione in modo che ad ogni anno prendesse come valore la

Figura 9: Grafico di $p(t, x)$ ad una magnitudine pari a $x = 5$, per il modello parametrico e il modello con punto di cambio.



Quella del modello parametrico ha una rapida crescita che, per costruzione, arriva fino ai giorni d'oggi: la censura dell'informazione, cioè, è prevista che si annulli solo al tempo corrente. La funzione di presenza del modello con punto di cambio, invece, dopo la rapida crescita, torna a essere costante e pari a 1: impone, quindi, che non ci possa essere censura nella registrazione degli eventi negli ultimi 400 anni circa. Questa ipotesi è stata preferita dai vulcanologi durante il workshop *Statistics in Volcanology* all'Università di Bristol (UK), nel marzo 2004, in cui si presentarono tali problematiche. C'è una ragione geografica per questo comportamento: prima del 1600 non erano disponibili monitoraggi per vaste zone della terra, per esempio l'Africa settentrionale, l'Africa meridionale e la Nuova Zelanda.

Dopo queste valutazioni, sembrano piuttosto evidenti i vantaggi che porta il modello con punto di cambio rispetto al modello parametrico:

- in una situazione di equivalenza nella spiegazione dei dati, la funzione di presenza sembra essere concettualmente coerente con le dinamiche fisiche del processo di censura dell'informazione;

media di tutte le traiettorie valutate in quell'anno, cioè:

$$p(t^*, 5) = \frac{1}{m - m_0} \sum_{i=m_0+1}^{m} p(t^*, 5|\mu_i, \sigma_i, \xi_i, \alpha_i, \beta_i, k_i)$$

per ogni $t^* \in [0, 1992]$, dove m è la lunghezza delle catene di Markov e m_0 è l'iterazione corrispondente al burn-in. Il grafico risultante è molto simile a un salto in corrispondenza della moda: d'altra parte la distribuzione di k è ha un dominio molto ristretto.

- Il modello con un punto di cambio lascia lo spazio a possibili generalizzazioni: si può immaginare, infatti, che la probabilità di un'eruzione di essere registrata nel catalogo, possa cambiare in più di un periodo. Tuttavia, se verranno identificati altri punti di cambio, essi avranno sicuramente meno influenza di quello attorno al 1600, altrimenti la stima della densità a posteriori di k non sarebbe stata concentrata in un così breve dominio.

6 Generalizzazione del modello con punto di cambio

Osservando i dati del catalogo (per $x > 4$) nel Grafico 10, si possono scorgere altri possibili punti di cambio e cioè zone con differenti concentrazioni di punti. Per esempio, attorno all'anno 1000 sembra esserci un aumento di eventi registrati sia a livelli bassi (prima comparsa di eventi vicino a 4) sia medio-alti (tra 5 e 6). Cronologicamente, poi, c'è la fortissima evidenza di un cambio di densità dei punti a tutti i livelli attorno al 1600 (già identificata con il modello con un punto di cambio), seguita da una macchia bianca a livello di magnitudine tra 4.2 e 4.5. Negli ultimi 100 anni, invece, il processo di eruzioni vulcaniche sembra essere registrato con omogeneità a tutti i livelli.
Si potrebbe, quindi cercare di inserire la presenza di 2 punti di cambio e poi di 3, per vedere quanti sono plausibili e le loro localizzazioni.

6.1 Due punti di cambio

Quando si ipotizzano due punti di cambio, la funzione di presenza viene generalizzata come segue:

$$p(t,x) = \begin{cases} \dfrac{\exp(\alpha_1 + \beta_1 x)}{1+\exp(\alpha_1 + \beta_1 x)} & t \le k_1 \\ \dfrac{\exp(\alpha_2 + \beta_2 x)}{1+\exp(\alpha_2 + \beta_2 x)} & k_1 < t \le k_2 \\ 1 & t > k_2, \end{cases} \tag{13}$$

con $k_1, k_2 \in [0, 1992]$. Si ha, di conseguenza, per ogni livello di magnitudine x, un livello di censura antecedente al primo punto di cambio k_1 (in funzione di α_1 e β_1), un altro livello di censura tra k_1 e il secondo punto di cambio k_2 (in funzione di α_2 e β_2) e poi, dopo k_2, un ultimo periodo di non censura.
La funzione di verosimiglianza diventa:

$$\begin{aligned} &L(\mu, \sigma, \xi, \alpha_1, \beta_1, \alpha_2, \beta_2; (t_1, x_1), \ldots, (t_n, x_n)) = \\ &\exp\left\{ \int_{x=u}^{+\infty} \int_{t=0}^{k_1} \lambda(t,x) \frac{\exp(\alpha_1 + \beta_1 x)}{1+\exp(\alpha_1 + \beta_1 x)} \, dt \, dx \right\} \times \\ &\exp\left\{ \int_{x=u}^{+\infty} \int_{t=k_1+1}^{k_2} \lambda(t,x) \frac{\exp(\alpha_2 + \beta_2 x)}{1+\exp(\alpha_2 + \beta_2 x)} \, dt \, dx \right\} \times \\ &\exp\left\{ \int_{x=u}^{+\infty} \int_{t=k_2+1}^{1992} \lambda(t,x) \, dt \, dx \right\} \times \\ &\prod_{i:0<t_i\le k_1} \lambda(t_i, x_i) \frac{\exp(\alpha_1 + \beta_1 x_i)}{1+\exp(\alpha_1 + \beta_1 x_i)} \prod_{i:k_1 t_i \le k_2} \lambda(t_i, x_i) \frac{\exp(\alpha_2 + \beta_2 x_i)}{1+\exp(\alpha_2 + \beta_2 x_i)} \\ &\prod_{i:k_2<t_i\le 1992} \lambda(t_i, x_i) \end{aligned}$$

6.1.1 Generalizzazione dell'algoritmo MCMC

L'algoritmo descritto nella Sezione 5.2 deve essere generalizzato per procedere all'aggiornamento di tre parametri supplementari. Per quanto riguarda (μ, σ, ξ) le distribuzioni a priori rimangono identiche, mentre vengono introdotte le seguenti distribuzioni vaghe:

1. $\alpha_1 \sim N(0, 10^3)$
2. $\beta_1 \sim N(0, 10^3)$
3. $\alpha_2 \sim N(0, 10^3)$
4. $\beta_2 \sim N(0, 10^3)$
5. $(k_1, k_2) \propto k1(k2 - k1)(1992 - k2)$,

dove la distribuzione congiunta di (k_1, k_2) tende a sfavorire punti di cambio troppo vicini tra essi o ai limiti.

Per le funzioni proposta viene utilizzata la stessa struttura dell'algoritmo con un solo punto di cambio per (μ, σ, ξ), semplicemente duplicando la procedura per i parametri $(\alpha_1, \beta_1, \alpha_2, \beta_2)$. Per i punti di cambio, invece, si è ipotizzato delle distribuzioni discrete uniformi nei domini autorizzati:

1. $k_{1p} \sim U[1, k_2]$
2. $k_{2p} \sim U[k_1 + 1, 1992]$

Dato che le funzioni proposta si semplificano, le probabilità di accettazione, per ogni parametro, si ottengono calcolando il minimo tra 1 e il rapporto del prodotto tra la verosimiglianza e la densità a priori calcolate, rispettivamente, per il valore proposto e corrente.

6.1.2 Risultati

Le stime delle densità a posteriori di k_1 e k_2 sono rappresentate nel Grafico 11; le stesse vengono riscalate per essere inserite insieme ai dati del catalogo nel Grafico 12. Il primo punto di cambio è risultato lo stesso di quello identificato con il modello con un solo punto di cambio. Il secondo, invece, si colloca attorno al 1928 e, secondo il modello, dovrebbe limitare a sinistra il periodo in cui l'identificazione delle eruzione è totale.

Le stime dei valori attesi a posteriori e gli intervalli di credibilità al 95% dei parametri del modello con due punti di cambio[4], sono riportati nella Tabella 7.

Le stime di μ, σ, ξ sono cambiate da quelle del modello con un solo punto di cambio (riportate nella Tabella 5): esse rappresentano, infatti, la parametrizzazione di un processo di Poisson, la cui espressione senza censura si colloca in un periodo

[4]Nelle catene markoviane i parametri risultano un pò più correlati, rispetto al modello con un solo punto di cambio: la convergenza sembra restare soddisfacente, mentre si perde un pò in mixing per alcuni parametri.

Figura 10: Dati del catalogo corrispondenti a $x > 4$.

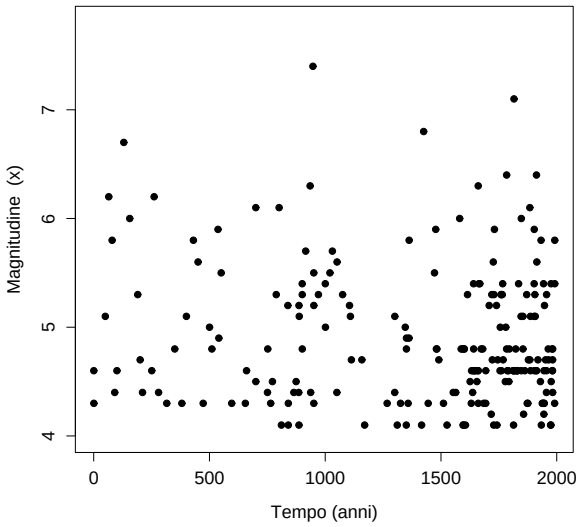
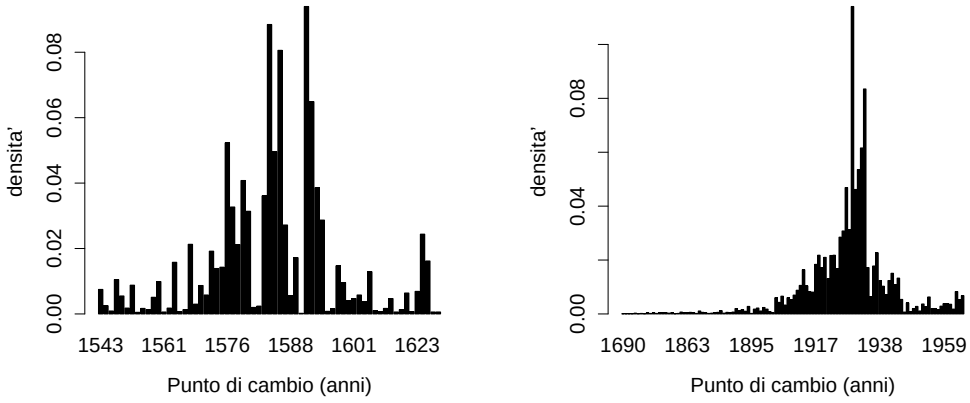
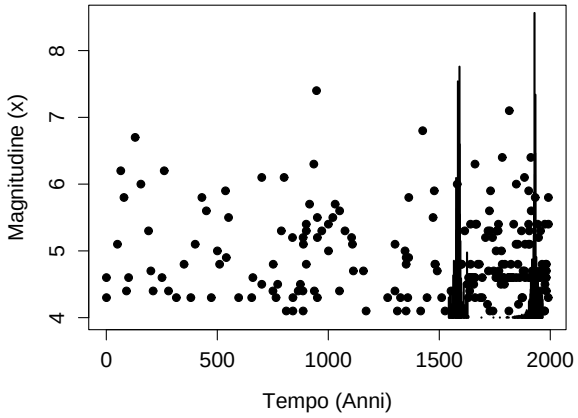


Figura 11: Stima della densità a posteriori di k_1 e k_2. I dati utilizzati sono corrispondenti a $x > 4$.



molto più breve del precedente (ultimi 100 anni *versus* ultimi 400 anni). Di conseguenza, il modello con due punti di cambio identifica un periodo di omogeneità delle osservazioni molto più breve di quello identificato dal modello con un punto di cambio. Osservando i dati del catalogo, si capisce come la macchia bianca di osservazioni a livello di magnitudine tra 4.2 e 4.5 dopo il 1600 si sia rivelata di una

Figura 12: Dati del catalogo corrispondenti a $x > 4$, con sovrapposte le densità stimate di k_1 e k_2, riscalate opportunamente.



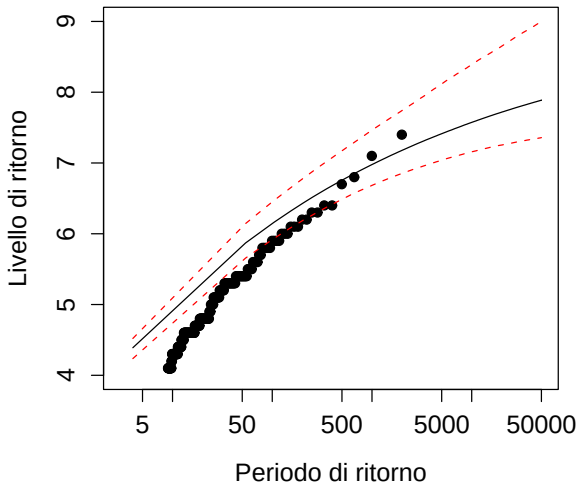
certa importanza nell'identificazione del secondo punto di cambio: esso si colloca proprio alla sua fine. Sembra, quindi, che gli eventi di magnitudine poco superiore a 4 vengano registrati senza pericolo di censura solo dopo circa la I guerra mondiale. Un'altra interpretazione può essere che tale macchia bianca sia la conseguenza di un'approssimazione sistematica nell'identificazione dell'intensità delle eruzioni. Se fosse così, questo modello non avrebbe molto senso perché il secondo punto di cambio perderebbe l'interpretazione originaria.

In questo modello, si ha $\hat{\eta} = 0.436$: questo valore si avvicina al valore trovato con il modello parametrico. Infatti, il periodo di non censura per il modello con due

Parametro	val. att.	int. cred.
μ	3.325	$(2.839, 3.591)$
σ	0.855	$(0.607, 1.233)$
ξ	-0.144	$(-0.242, -0.046)$
α_1	-6.559	$(-10.033, -3.201)$
β_1	1.001	$(0.324, 1.732)$
α_2	-35.531	$(-61.795, -16.042)$
β_2	8.222	$(3.668, 14.621)$
k_1	1585	$(1549, 1625)$
k_2	1928	$(1896, 1961)$

Tabella 7: Stime dei valori attesi a posteriori e intervalli di credibilità al 95% per $\mu, \sigma, \xi, \alpha_1, \beta_1, \alpha_2, \beta_2, k_1, k_2$. Sono riportati i quantili corrispondenti al 2.5% e al 97.5%.

Figura 13: Curva del livello di ritorno per il modello con due punti di cambio. Le linee tratteggiate corrispondono agli intervalli di credibilità al 95%per ogni valore del periodo di ritorno. I pallini neri corrispondono alle stime empiriche.

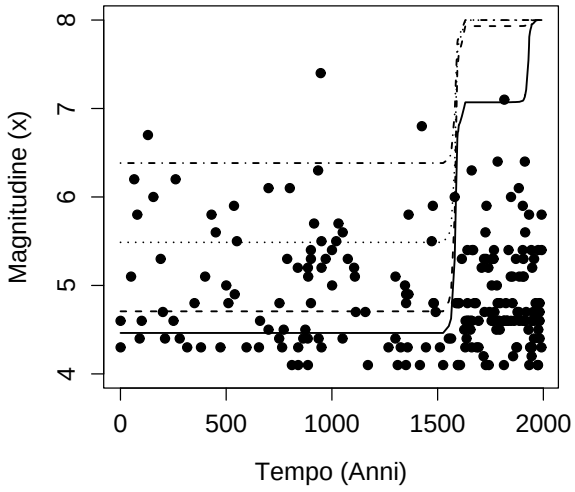


punti di cambio viene stimato essere circa nell'ultimo secolo e di conseguenza è più vicino alla situazione proposta dal modello parametrico. Qui, il limite superiore della densità stimata di X, risulta essere 10.182, ottenendo un valore più alto di quello relativo sia al modello parametrico, sia al modello con un solo punto di cambio. Il modello con due punti di cambio permette, quindi, il verificarsi di eruzioni di magnitudine maggiore e risulterebbe perciò più verosimile del modello con un punto di cambio: è noto, infatti, che nella storia si sia verificata un'eruzione di magnitudine superiore a 9. Per esempi di eruzioni di magnitudine superiori a 8 si veda Crisp (2004) e Mason *et al.* (2004).

Si veda il grafico 13 per la rappresentazione della curva del livello di ritorno per il modello con due punti di cambio. Si noti come l'intervallo di credibilità sia più ampio, per valori alti, del relativo intervallo per il modello con un solo punto di cambio. Inoltre, il modello con due punti di cambio prevede un effetto di censura per valori bassi di x più forte di quello identificato dal modello con un solo punto di cambio. Il Grafico 14 spiega tale comportamento: le eruzioni di bassa entità (circa inferiori a $x = 5$) trovano una differente registrazione tra il secondo periodo di censura e il terzo periodo (omogeneo nella registrazione).

I parametri β_1, β_2 non hanno lo 0 nei loro intervalli di credibilità (riportati nella Tabella 7): la loro significatività sembra essere aumentata in questo modello con due punti di cambio. Qui, la magnitudine sembra avere un ruolo più importante nel processo di registrazione degli eventi. Nel Grafico 14, sono state riportate le

Figura 14: Funzioni di presenza $p(t, x)$ del modello con 2 punti di cambio, corrispondenti a $x = 4.5$ (—), $x = 5$ (- -), $x = 6$ ($\cdots$), $x = 7$ (- $\cdot$ -). I dati utilizzati sono corrispondenti a $x > 4$.



funzioni di presenza per il modello con due punti di cambio, per vari livelli di magnitudine. A conferma della maggiore significatività dei parametri β_1, β_2, si noti come le traiettorie cambino sostanzialmente per alcuni livelli della magnitudine: quella corrispondente a $x = 4.5$ e cioè ad alcuni eventi appartenenti alla macchia bianca, di cui si è parlato prima, hanno una probabilità di essere registrata nel secondo periodo sostanzialmente più piccola di quelli di magnitudine superiore a 5. Questi ultimi sembra vengano registrati senza censura anche nel secondo periodo, in quanto le loro traiettorie sono pressocché sovrapponibili alla costante 1 del terzo periodo (che nel Grafico 14 corrisponde a 8).
Nel primo periodo, poi, le traiettorie sono più distanziate, per magnitudine, delle corrispondenti del modello con un solo punto di cambio (Grafico 7 e Tabella 8). Questo può essere imputato al fatto che i periodi di non censura, con cui ci si paragona per stimare i parametri responsabili della probabilità di registrazione degli eventi, cambiano tra il modello con un punto e due punti di cambio.

La log-verosimiglianza, in corrispondenza delle stime dei valori attesi a posteriori dei parametri, è -806.69. Confrontandola con quella del modello con un solo punto di cambio (si veda la Tabella 6), si nota un incremento di 13.384. La generalizzazione con due punti di cambio sembra essere soddisfacente, tenendo che conto che questa ha 3 parametri in più. Tuttavia, si cercherà di capire con degli esperti di vulcanologia se tale modello ha fondamento o se è la conseguenza di un'approssimazione sistematica nelle misurazioni, che ha dato luogo alla macchia bianca di cui si è parlato.

6.2 Tre punti di cambio

All'inizio della Sezione 6 si erano avanzate delle ipotesi su possibili punti di cambio, attraverso un'analisi visiva dei dati del catalogo. Fino a questo punto si è riusciti ad individuare con decisione due dei tre punti di cambio ipotizzati; ma nei risultati non c'è mai stata evidenza per quello corrispondente all'area attorno all'anno 1000. Evidentemente le differenze che si notano nella densità dei punti non sono forti come quelle del periodo dell'espansione geografica e quella dell'ultimo secolo. Si è provato, quindi, a stimare un modello con 3 punti di cambio[5], per vedere se questo periodo attorno all'anno 1000 avesse comunque della probabilità ad essere identificato in un modello che impone la presenza di 3 periodi con diversa censura.
Quello che è risultato è che sono stati individuati i due periodi del modello con due punti di cambio e il terzo (imposto dal modello) si collassava alternativamente su uno dei due periodi. Quindi, il periodo attorno l'anno 1000 non ha differenze significative nel processo di registrazione per essere considerato periodo a sè.
Tuttavia, nel caso in cui si imponesse il terzo punto di cambio prima dell'anno 1500, il periodo attorno all'anno 1000 veniva identificato con decisione. Quindi, la sensazione dell'analisi visiva per questo terzo punto di cambio ha avuto un riscontro solo in questo particolare contesto vincolato. Evidentemente, quando la spazio d'azione viene liberato su tutti e due i millenni tale evidenza perde di importanza.

6.3 Numero ignoto di punti di cambio

Si potrebbe provare a ipotizzare un numero ignoto di punti di cambio e utilizzare un algoritmo RJMCMC (si veda Green (1995) per i dettagli) per stimarne la densità a posteriori.
Sembra però evidente che, per $u = 4$, non ci sia evidenza di un numero di punti di cambio superiori a due. Infatti, si è visto nella Sezione 6.1 che il secondo punto di cambio viene accettato per il diverso tasso di registrazione delle eruzioni di più bassa intensità e che un eventuale terzo punto non viene identificato.

[5] È una facile generalizzazione del modello con due punti di cambio.

magnitudine (x)	1 p. di cambio	2 p. di cambio
4.5	0.184	0.116
5	0.214	0.177
6	0.291	0.372
7	0.379	0.596

Tabella 8: Stima delle funzioni di presenza al tempo $t = 0$ per i modelli con uno e due punti di cambio.

7 Studio di sensibilità sulla soglia

Si era già accennato nella Sezione 4, che in questa sede si sarebbe terminato di riportare le tappe salienti del lavoro di Coles and Sparks (2004). Nella Tabella 9 si presentano le stime di massima verosimiglianza e i loro errori standard per il processo omogeneo stimato negli ultimi 400 anni e il modello parametrico sull'intera serie storica, per eruzioni superiori a $x = 5.1$. Confrontandoli con i relativi valori presentati nella Tabella 1, si può capire come l'utilizzo di una soglia più alta porti all'estrapolazione di valori più alti. Infatti, il valore di ξ, che determina la forma della coda della distribuzione, è più vicino allo zero. Confrontando le curve dei livelli di ritorno Coles vide che l'intervallo di confidenza per $u = 4$ non giustificava eventuali eruzioni di magnitudine attorno a 9, come sono giustificate, invece, dal modello con una soglia di 5.1. Infatti, la distribuzione risulta essere limitata superiormente a 7.45, per una soglia pari a 4, e a 9.17 per una soglia pari a 5.1. Eruzioni così estreme non sono presenti nel dataset ed è per questo che considerando troppi eventi si induce una distorsione nella determinazione della coda della distribuzione; tuttavia, si possono trovare numerosi esempi di eruzioni di magnitudine superiore a $x = 8$ in Mason *et al.* (2004). Il valore massimo della log-verosimiglianza per il modello parametrico, per $u = 5.1$ è -809.30.
Infine, Coles trovò che utilizzando la soglia $u = 5.1$ l'effetto di sottoidentificazione degli eventi non sembrava dipendere dal livello della magnitudine degli stessi. Quindi, è più utile utilizzare una soglia bassa per studiare il fenomeno della sottoidentificazione degli eventi e una soglia più alta per stimare più correttamente la coda della distribuzione delle eruzioni. Il primo punto è già stato sviluppato nella Sezione 5, mentre ora si cerca di sviluppare il secondo tramite l'uso di modelli con punti di cambio.

Nella Figura 15 vengono raffigurate le eruzioni di magnitudine superiori a $x = 5.1$. A un primo sguardo sembra che sia ancora plausibile un punto di cambio attorno al 1600; quello nell'ultimo secolo, invece, non sembra più realistico perché avendo alzato la soglia sono state escluse quelle eruzioni di magnitudine bassa che lo avevano fatto identificare. Poi, sembra ci siano delle macchie di punti contornate da spazi bianchi in tutto il periodo prima del 1600.

	μ	σ	ξ	v	w	b
Ultimi 400	3.416	0.805	-0.140	-	-	-
anni	(0.722)	(0.483)	(0.161)	-	-	-
Tutti gli	4.004	0.628	-0.102	5.529	1.125	9.232
anni	(0.498)	(0.334)	(0.125)	(2.026)	(1.263)	(3.096)

Tabella 9: Stime di massima verosimiglianza ed errori standard (tra parentesi) del processo di punto omogeneo (stimato negli ultimi 400 anni) e con censura (stimato su tutta la serie storica), applicato ai dati delle eruzioni vulcaniche con una soglia di $x = 5.1$.

7.1 Modello con un punto di cambio

Come primo passo è stato ristimato il modello con un solo punto di cambio, per il nuovo valore della soglia. L'effetto della funzione proposta per k (uniforme discreta sui 1992 anni), faceva in modo che il punto di cambio stesse per la maggior parte del tempo attorno al 1600 e ogni tanto cadesse attorno al 700. Mentre nel primo caso la log-verosimiglianza era molto simile al valore massimo con il modello parametrico, nel secondo diminuiva di circa 5 punti. Questo vuol dire, come si era già preannunciato, che tale funzione proposta ostacola l'abbandono di probabili massimi locali. Visto che tale cambiamento di zona per k induceva dei forti cambiamenti nei parametri (α, β) si è deciso di utilizzare come funzione proposta un'uniforme discreta su una finestra di 200 anni, simmetrica rispetto al valore corrente del punto di cambio:

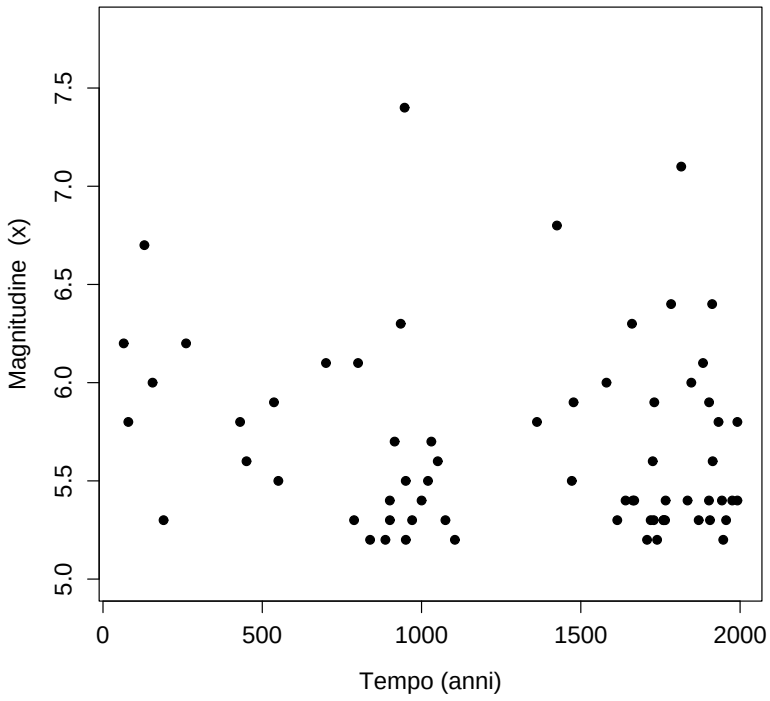
$$k_p \sim U[\max(1, k_t - 100), \min(k_t + 100, 1992)],$$

dove k_t è il valore corrente della catena e k_p è il valore proposta.
Il valore di log-verosimiglianza calcolata in corrispondenza delle stime dei valori attesi a posteriori è -308.879: l'approccio del modello parametrico e del modello di punto di cambio sono piuttosto equivalenti da un punto di vista della log-verosimiglianza. Nella Tabella 10 vengono riportate le stime e gli intervalli di credibilità al 95% dei valori attesi a posteriori dei parametri. Si noti come, a differenza del modello parametrico, la differente funzione presenza, con punto di cambio, identifica una certa differenza nel processo di sottoidentificazione degli eventi per livello di magnitudine. Infatti, la stima di β è diventata da quasi zero (con $u = 4$) a superiore ad 1, anche se l'intervallo di credibilità al 95% contiene lo zero. L'intervallo di confidenza per k, poi, è situato grossomodo nello stesso periodo del caso per $u = 4$, però è più ampio e infatti copre circa 100 anni.
Nella Figura 16 viene riportata la curva del livello di ritorno; considerando che il valore limite della distribuzione viene stimato essere pari a 9.15 si può dire che questo modello è coerente con la memoria storica delle eruzioni più estreme di quelle registrate nel catalogo. Di conseguenza, l'approccio di massima verosimiglianza di Coles and Sparks (2004) e quello bayesiano, adottato in questa tesi, sembrano stimare la stessa forma della coda della distribuzione delle eruzioni. Tuttavia, guardando l'intervallo di confidenza della curva del livello di ritorno (approccio di massima verosimiglianza), si vede come la banda di confidenza inferiore sia prima crescente e poi decrescente, assegnando a stesse eruzioni livelli di ritorno diversi: questo aspetto non è verosimile e viene superato dall'approccio bayesiano.
Nella Figura 17 vengono presentate le funzioni di presenza (riscalate per essere sovrapposte ai dati) per $x \in \{5.2, 6, 7\}$. Il tasso annuo di eruzioni eccedenti la soglia viene stimato pari a $\hat{\eta} = 0.083$.

7.2 Modello con due punti di cambio

Come descritto nella Sezione precedente, si era osservata dell'evidenza per un probabile secondo punto di cambio attorno al 700. Si è pensato che imponendo due punti di cambio (k_1 e k_2) si sarebbe ottenuto facilmente una stabilizzazione attorno al 700 oltre a quella del 1600 (circa). Invece, non è successo così: a sinistra del 1600

Figura 15: Catalogo bimillenario di eruzioni vulcaniche eccedenti una magnitudine di $x = 5.1$, nel pianeta Terra.

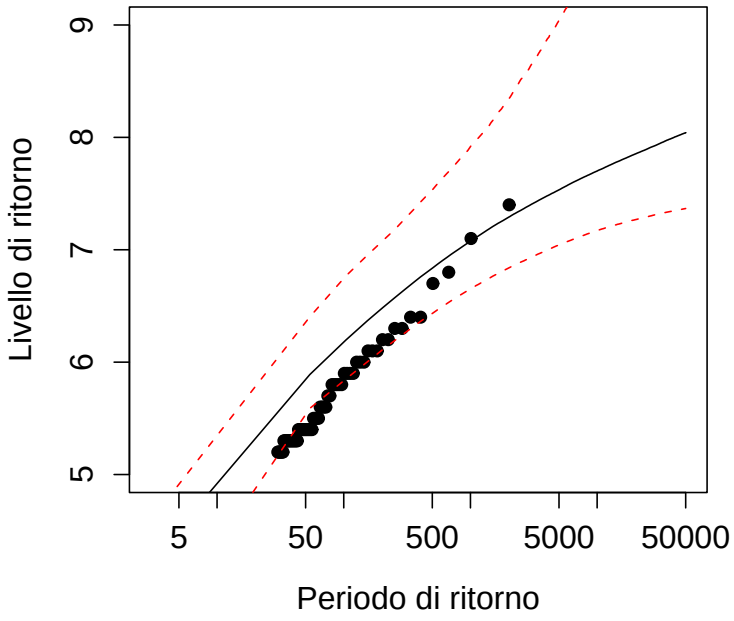


ci sono varie macchie di punti e il punto di cambio K_1 non è riuscito a trovare una collocazione. Tale fatto giustifica l'incertezza nel capire il motivo della presenza di tali macchie: si ipotizza, infatti, che queste non siano sufficienti per creare nuovi periodi di sottoidentificazione degli eventi.
La stima della densità a posteriori di k_1 risulta piatta sui primi 1000 anni e que-

Parametro	val. att.	int. cred.
μ	3.178	$(1.199, 4.395)$
σ	0.898	$(0.234, 2.367)$
ξ	-0.131	$(-0.364, 0.136)$
α	-7.362	$(-30.164, 3.723)$
β	1.151	$(-0.831, 5.323)$
k	1624	$(1536, 1709)$

Tabella 10: Stime dei valori attesi a posteriori ed intervalli di credibilità al 95% per $\mu, \sigma, \xi, \alpha, \beta, k$, per dati corrispondenti a $x > 5.1$.

Figura 16: Curva del livello di ritorno per il modello con un punto di cambio. Le linee tratteggiate corrispondono agli intervalli di credibilità al 95% per ogni valore del periodo di ritorno. I pallini neri corrispondono alle stime empiriche. Eruzioni di magnitudine superiore a $x = 5.1$.

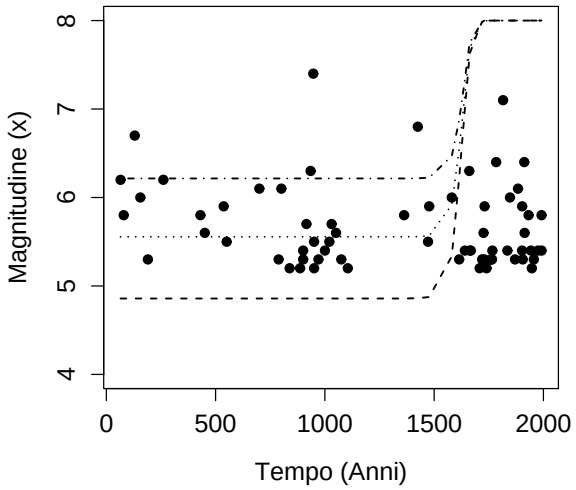


sto continuo cambiamento sul dominio ha impedito la stabilizzazione dei parametri relativi alla funzione di presenza. Inoltre, i valori della log-verosimiglianza si sono sempre mantenuti, nei migliori casi, attorno a quelli osservati per il modello con un punto di cambio.
In definitiva, non sembra esserci un'adeguata evidenza per l'identificazione di due punti di cambio.

8 Introduzione di una componente spaziale

Esistono alcune zone della Terra che sono più soggette a eruzioni di intensità estreme. Nel Grafico 18 sono riportate con i pallini le eruzioni delle isole dell'Oceania e con le crocette le eruzioni del resto del mondo. Si può notare come le eruzioni dell'Oceania coprano tutti i valori più estremi di $x = 6.5$ e confermino le ipotesi fatte, sulla dipendenza dell'effetto di identificazione degli eventi in relazione all'intensità: gli esperti sono riusciti, infatti, ad identificare gli eventi più estremi anche a distanza temporale di 2000 anni, ma non è stato così per le eruzioni di più bassa intensità

Figura 17: Funzioni di presenza $p(t, x)$ del modello con 1 punto di cambio, corrispondenti a $x = 5.2$ (- -), $x = 6$ ($\cdots$), $x = 7$ (- $\cdot$ -). I dati utilizzati sono corrispondenti a $x > 5.1$.

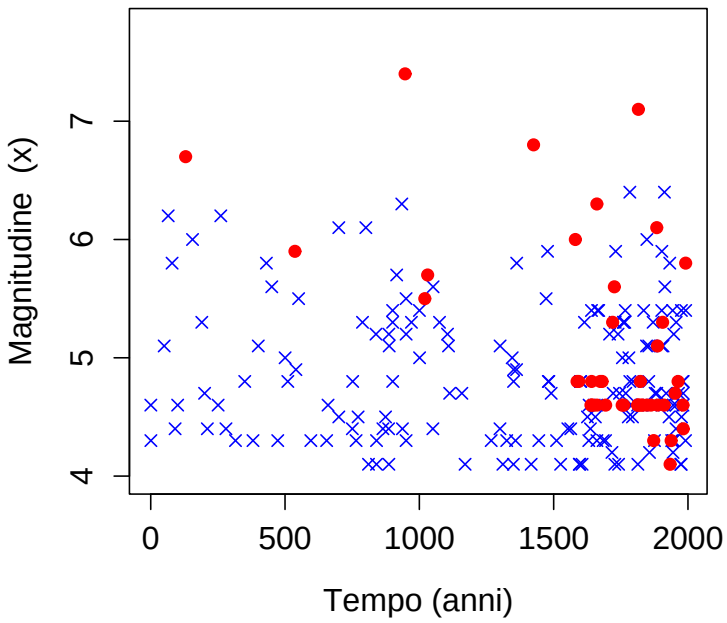


(gli eventi inferiori a $x = 5$ sono disponibili solo dopo il 1600). Le nuove scoperte geografiche del XVII secolo hanno contribuito moltissimo alla riduzione della censura fornendo la possibilità di monitorare nuove terre.
Si noti, poi, come ci sia una uniformità di eruzioni, per l'Oceania dopo il 1600, a livello $x = 4.6$. Questo fenomeno potrebbe essere la conseguenza di un'approssimazione nella quantificazione dell'intensità delle eruzioni in quelle zone oppure essere dovuto agli strumenti utilizzati. Infatti, sotto tale linea si trova quella macchia bianca di osservazioni, che ha generato l'identificazione del secondo punto di cambio, nel caso di $u = 4$ (si veda la Sezione 6). Tale comportamento sembra un pò strano e dovrebbe essere approfondito insieme a dei geologi: potrebbe, infatti, essere un motivo per non ritenere realistico il modello con due punti di cambio.

Ci potrebbe essere lo spazio, quindi, per stimare un processo di Poisson che tenga conto di eventuali differenze nel comportamento delle eruzioni estreme, dovute a diverse collocazioni territoriali. Tale idea è ancora in una fase iniziale, però sembrava opportuno delineare le principali idee di un possibile modello. Per semplicità, si continui a considerare i due gruppi già presentati; l'effetto della censura viene tenuto comune, mentre vengono stimati parametri diversi per il processo di Poisson. Siano x le eruzioni del catalogo, che è l'unione delle eruzioni dell'Oceania x_O e del resto del mondo x_{RM}; ripetendo il cambio di parametrizzazione e introducendo ζ al posto di σ (si veda la Sezione 5.2), allora:

$$(t, x) \sim \begin{cases} \lambda(t, x|\mu_O, \zeta_O, \xi_O) p_C(t, x|\alpha, \beta, k), & x \in X_O \\ \lambda(t, x|\mu_{RM}, \zeta_{RM}, \xi_{RM}) p_C(t, x|\alpha, \beta, k), & x \in X_{RM} \end{cases}$$

Figura 18: Catalogo bimillenario di eruzioni vulcaniche eccedenti una magnitudine di $x = 4$, nel pianeta Terra. Le osservazioni disegnate con il pallino indicano le eruzioni registrate nelle isole dell'oceano Pacifico, mentre quelle con x sono del resto del Mondo.



dove t è la relativa collocazione temporale, $\lambda(\cdot, \cdot)$ viene definita in (1) e $p(t, x)$ è la funzione di presenza con un punto di cambio. Si può, inoltre, decidere quale soglia si voglia utilizzare a seconda degli obiettivi che ci si propone.
Le distribuzioni dei parametri del processo di Poisson per i due gruppi sono:

$$\begin{aligned} \mu_i &\sim N(\mu, \delta_\mu) \\ \zeta_i &\sim N(\zeta, \delta_\zeta) \\ \xi_i &\sim N(\xi, \delta_\xi) \end{aligned}$$

dove $i = \{O, RM\}$ e $(\delta_\mu,\ \delta_\zeta,\ \delta_\xi)$ rappresentano le precisioni delle relative distribuzioni. Infine si scrivino le distribuzioni a priori degli iperparametri:

$$\begin{array}{lllllllll} \mu & \sim & N(0, 10^3) & \zeta & \sim & N(0, 10^3) & \xi & \sim & N(0, 10^3) \\ \alpha & \sim & N(0, 10^3) & \beta & \sim & N(0, 10^3) & k & \propto & k(1992 - k) \\ \mu & \sim & N(0, 10^3) & \zeta & \sim & N(0, 10^3) & \xi & \sim & N(0, 10^3) \\ \delta_\mu & \sim & Ga(10^{-2}, 10^{-2}) & \delta_\zeta & \sim & Ga(10^{-2}, 10^{-2}) & \delta_\xi & \sim & Ga(10^{-2}, 10^{-2}) \end{array}$$

In questo modo, possono venire stimate le code per ognuno dei due gruppi e poi, tramite gli iperparametri, la coda della distribuzione di tutte le eruzioni del catalogo.

Di conseguenza, se le code della distribuzione dei singoli gruppi sono diverse tra loro, questo indurrà una maggiore variabilità nella coda comune.

Come ulteriori sviluppi, si potrebbe eventualmente pensare di stimare diversi parametri della funzione presenza per i due gruppi. Oppure, si potrebbe cambiare il numero di gruppi, fino al limite di ipotizzare un numero ignoto di essi (utilizzando il RJMCMC).

9 Conclusioni

L'approccio bayesiano sembra essere preferibile all'approccio di massima verosimiglianza, in quanto è più naturale interpretare il rischio tramite la distribuzione predittiva. Poi, la funzione di presenza con punti di cambio si è rivelata preferibile a quella presentata in Coles and Sparks (2004), per interpretare il processo di identificazione degli eventi.

Entrambe le scelte della soglia si sono rivelate interessanti, perché la soglia più bassa ($u = 4$) permette di studiare più a fondo il fenomeno di identificazione degli eventi, mentre quella più alta ($u = 5.1$) permette di introdurre meno distorsione nella stima della coda della distribuzione delle eruzioni.
Nel primo caso, il modello con un punto di cambio sembra essere soddisfacente; quello con due punti di cambio è risultato migliore di quest'ultimo, perché permette di stimare la limitazione superiore della coda coerentemente con eruzioni più estreme di quelle riportate nel catalogo, di cui si ha memoria storica, già a un valore basso della soglia. Tuttavia, c'è la possibilità che il secondo punto di cambio venga identificato per un'approssimazione sistematica nella misurazione delle eruzioni di intensità tra 4 e 5 tra il 1600 e il 1900 (circa) soprattutto nelle isole dellOceania: questi aspetti devono essere approfonditi con dei geologi per valutare l'effettiva validità logica del secondo punto di cambio.
Nel secondo caso, si è riusciti a identificare un solo punto di cambio; però la stima della coda della distribuzione delle eruzioni permette il verificarsi di quelle eruzioni estreme che non sono presenti nel catalogo, differentemente dal modello con un punto di cambio con la soglia più bassa.

Si pensa possa essere utile sviluppare una componente spaziale nel modello per tenere conto di eventuali differenze di comportamento delle eruzioni estreme tra diversi territori.

Riferimenti bibliografici

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

Coles S.G. and Sparks R.S.J. (2004) Extreme value methods for modelling historical series of large volcanic magnitudes, in: *Statistics in volcanology*, Mader H., Coles S.G. and Connor C., eds., Forthcoming..

Crisp J.A. (2004) Rates of magma emplacement and volcanic output, *Jnl. Volcanology and Geothermal Research*, 20, 177–211.

Gilks W.R., Richardson S. and Spiegelhalter D.J. (Eds.) (1996) *Markov chain Monte Carlo in practice*, Chapman & Hall, London.

Green P. (1995) Reversible jump markov chain monte carlo computation and bayesian model determination, *Biometrica*, 82, 711–732.

Mason B.G., Pyle D.M. and Oppenheimer C. (2004) The size and frequency of the largest explosive eruptions on earth, *Bulletin of Volcanology*, doi:10.1007/s00445–004–0335–9.

Pickands J. (1971) The two-dimensional poisson process and extremal processes, *Journal of Applied Probability*, 8, 745–756.

Simkin T. and Siebert L. (1994) *Volcanoes of the World*, Geoscience Press, Tucson.

Smith R.L. (1989) Extreme value analysis of environmental time series: An application to trend detection in ground-level ozone (with discussion), *Statistical Science*, 4, 367– 393.

Acknowledgements

This work was supported by the University of Padova (Italy) grant CPDA037217: "Methods for the analysis of extreme sea levels and for coastal erosion".

Working Paper Series
Department of Statistical Sciences, University of Padua

You may order paper copies of the working papers by emailing wp@stat.unipd.it
Most of the working papers can also be found at the following url: http://wp.stat.unipd.it

UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI SCIENZE STATISTICHE

Department of Statistical Sciences
University of Padua
Italy