Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche

Modelli bayesiani per l'analisi di dati ambientali: uno studio della velocità del vento nelle zone del Veneto colpite dalla Tempesta Vaia

Relatore: Prof. Antonio Canale
Dipartimento di Scienze Statistiche
Correlatore: Prof. Marco Marani
Dipartimento di Ingegneria Civile, Edile e Ambientale

Laureanda: Sara Ceschin
Matricola: 1179192

Anno Accademico 2018/2019

Indice

Introduzione

Tra il 27 e il 30 ottobre 2018 si è verificata una combinazione di piogge abbondanti e venti molto forti, definita dai meteorologi *Tempesta Vaia,* che ha portato alla caduta di numerosi alberi nelle zone montane del Veneto e del Trentino Alto-Adige. Questo ha avuto conseguenze tangibili sul territorio ed è stato molto sentito dalla popolazione.
L'anomalia dell'evento che si è verificato a fine ottobre è dunque, dal punto di vista ambientale, la caduta degli alberi e la domanda che ci si pone è capire cosa abbia portato a questo e se sia un fenomeno estremo che si può studiare e prevedere. Da un punto di vista statistico, tuttavia, formulare una risposta a queste domande richiede alcune accortezze. La caduta degli alberi è sicuramente dovuta alla combinazione di precipitazioni abbondanti e vento forte, ma come analizzare i due aspetti in maniera congiunta risulta complicato in quanto l'effetto della pioggia non può essere descritto semplicemente dalla quantità di precipitazione registrata ma si deve all'accumulo di acqua nel terreno. Queste domande risultano molto importanti ma troppo complesse per trovare risposta nei tempi di un progetto di tesi.

Questo progetto, dunque, si concentra esclusivamente su uno degli aspetti che hanno portato alle conseguenze di Vaia, la velocità del vento. Lo scopo iniziale è cercare di capire quanto si fosse rivelato estremo rispetto al passato tramite l'analisi dei dati di alcune stazioni di rilevazione situate nel territorio montano del Veneto, forniti dall'Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto (ARPAV).

Si è notato subito che i dati presentano un andamento bimodale, di conseguenza si è cercato di trovare una distribuzione che descriva adeguatamente questo comportamento. L'approccio qui seguito è di tipo metodologico con l'introduzione di nuovi modelli statistici e la loro implementazione tramite uno studio di simulazione. L'inferenza è condotta

secondo il paradigma bayesiano. La tesi si sviluppa studiando prima i metodi e modelli su campioni adeguatamente simulati e riportando poi come illustrazione le analisi sui dati reali.

Le analisi preliminari dei dati e lo studio di simulazione hanno permesso di avvicinarsi a quello che è l'obiettivo finale di questo progetto di tesi e che riguarda lo studio di un metodo innovativo per la modellazione di fenomeni ambientali volto alla descrizione dei comportamenti estremi. Il modello proposto è definito seguendo la struttura del modello bayesiano gerarchico introdotto da Zorzetto, Canale e Marani (2019). Esso si propone di superare alcune delle problematiche riscontrate nella teoria classica dei valori estremi. In un contesto dove lo studio di questi eventi estremi sta assumendo sempre maggiore importanza ha senso infatti chiedersi se la teoria classica possa essere integrata con metodi innovativi.

Nel Capitolo 1 verrà descritto il fenomeno della tempesta Vaia dal punto di vista ambientale. Nel Capitolo 2 saranno illustrati i principali metodi della teoria classica dei valori estremi per poi introdurre il modello bayesiano gerarchico di Zorzetto, Canale e Marani (2019). Le analisi dei dati relativi al vento inizieranno nel Capitolo 3. Queste danno chiare indicazioni relativamente alle caratteristiche di questo tipo di fenomeni. Nel Capitolo 4 viene introdotto e studiato un modello mistura adatto per descrivere il comportamento bimodale dei dati. Date le sue performance empiriche questo modello viene inserito all'interno del contesto di Zorzetto, Canale e Marani (2019) nel Capitolo 5.

Capitolo 1

La tempesta Vaia

La tempesta Vaia ha colpito soprattutto l'Italia settentrionale ed in particolare ha avuto effetti devastanti nelle zone alpine del Veneto e del Trentino Alto Adige. Il paesaggio di queste montagne è stato modificato in maniera evidente dal passaggio di venti molto forti che, in concomitanza all'accumulo di precipitazioni abbondanti nei giorni precedenti, ha abbattuto migliaia di ettari di boschi; ne è un esempio la foto in Figura 1.1. Le conseguenze di questo evento si sono estese poi all'intero Nord-Est, dove fiumi che nascono in quelle montagne e attraversano l'intero triveneto sono stati monitorati perché a rischio esondazione. Il Piave ha registrato livelli di piena superiori a quelli dell'alluvione del novembre 1966, ricordata come una delle peggiori ad interessare il corso d'acqua (Dariol, 2008).
In Europa la tempesta Vaia non è il primo evento che colpisce aree montane e boschive, con effetti devastanti sul territorio e sull'economia. Negli ultimi vent'anni sono state colpite anche Svezia, Norvegia, Germania, Francia, Austria e non solo (Cason, 2018).

A fine ottobre 2018 il maltempo era diffuso in tutta Italia e in parte dell'Europa, delimitando una zona di bassa pressione. I danni ambientali si sono verificati in particolare al Nord, dove oltre al Triveneto è stata colpita la Liguria, con violente mareggiate che hanno investito in particolare Portofino (GE), interrompendo la strada che lo collega agli altri paesi. Successivamente gli eventi si sono spostati lungo la costa tirrenica della penisola arrivando a causare alluvioni fino in Sicilia. Gli effetti maggiori si sono avuti però nel Nord-Est, dove le conseguenze di Vaia sono tuttora tangibili, si veda il panorama in Figura 1.3. Sono stati danneggiati risorse e panorami unici come i boschi del Cadore, riserva inestimabile di

Figura 1.1: Boschi del Trentino dopo la tempesta Vaia (`http://bit.ly/2oAeT2c`).

legname, e le Dolomiti, patrimonio dell'Unesco, meta turistica per eccellenza, frequentata sia in estate che in inverno.

In Veneto tutto è iniziato con un periodo di abbondanti precipitazioni che hanno colpito le aree montane e con un vento di scirocco molto forte soprattutto sulle Prealpi. Le precipitazioni hanno raggiunto valori record, in 96 ore a Soffranco (BL) si sono accumulati 715.8mm di pioggia e sono numerose le stazioni montane che hanno registrato un totale superiore ai 300mm nei 4 giorni, soprattutto nel Bellunese. Un valore, quello di Soffranco, che non si era mai registrato prima, come si può notare in Figura 1.2. I rovesci hanno interessato principalmente la zona prealpina e montuosa mentre nelle altre province le piogge sono state scarse. Inoltre le precipitazioni sono state più abbondanti nelle giornate di domenica 28 e lunedì 29, durante la quale tutto il Veneto è stato toccato, rimanendo però sempre Belluno la provincia più colpita. Soffranco ha registrato quasi sempre i valori più alti raggiungendo i 284.6mm nella sola giornata di domenica, ma non è l'unico luogo in cui sono stati superati i 200mm in 24 ore. Nella giornata di lunedì le precipitazioni sono state violente e abbondanti nel breve periodo, registrando valori di cumulate orarie paragonabili ai temporali estivi (ARPAV, 2018).

Figura 1.2: Serie storica delle precipitazioni cumulate a intervalli di 4 giorni a Soffranco. I giorni della tempesta Vaia (27-30 ottobre 2018) sono stati raggruppati insieme e sono quelli evidenziati dal punto nero.

Dalla zona montana del Veneto nascono molti fiumi di grande portata, di conseguenza i principali bacini sono stati tenuti sotto controllo per limitare i danni di una possibile esondazione. Hanno registrato i maggiori livelli di piena l'Adige, il Brenta e il Piave. Quest'ultimo ha toccato livelli critici nelle zone di Ponte di Piave e Nervesa della Battaglia, creando disagi alla circolazione e disattivando alcuni idrometri. I valori registrati hanno superato i massimi dell'alluvione del novembre 1966 e l'altezza dell'acqua ha raggiunto i principali ponti, come si vede in Figura 1.5.

Alle abbondanti precipitazioni si sono aggiunte le raffiche di vento, che hanno superato la velocità di 100km/h, con un massimo di 192km/h sul Monte Cesen (TV) a quota 1552m lunedì 29. Qui si è registrata l'intensità massima del vento da sabato 27 a martedì 30. Nella giornata di lunedì si è toccato il culmine dell'evento, con raffiche di intensità superiore ai 100km/h che si sono abbattute su buona parte del Bellunese e nelle vicine province di Treviso e Vicenza. I 192km/h sono risultati essere il valore più alto mai registrato da una stazione ARPAV negli ultimi 25 anni (ARPAV, 2018) e in Figura 1.4 si può notare quanto fosse estremo rispetto ai valori rilevati abitualmente.

Precipitazioni e raffiche di vento hanno raggiunto valori estremi, soprattutto in alcune zone, e questa combinazione è stata l'origine di un disastro ambientale ed economico che ha colpito profondamente la popolazione. Strade interrotte e boschi distrutti sono stati

Figura 1.3: Boschi del Veneto dopo la tempesta Vaia (`http://bit.ly/2n3hh0Y`).

i principali problemi da affrontare con la stagione sciistica alle porte. Sin da subito si è provveduto nelle regioni colpite a ripristinare la circolazione e a trovare soluzioni per arginare i danni arrecati alla popolazione dal maltempo. A questi si aggiungono i problemi portati dalla caduta degli alberi. La raccolta del legname risultava complicata ma doveva essere tempestiva per evitare che si marcissero i tronchi. Inoltre il rischio fitopatologico dell'aumento di parassiti delle piante metteva a repentaglio anche gli alberi rimasti, senza pensare ai danni all'ecosistema montano. In aggiunta i boschi svolgono anche una funzione di protezione da frane e valanghe e a seguito della perdita di questi aumentano i siti a rischio e la sicurezza ambientale è un'altra questione di cui tenere conto (Cason, 2018).
In Veneto la situazione era stata prevista da ARPAV, erano state prese tutte le precauzioni ed era stato emanato lo stato di allerta. Subito dopo la tempesta Vaia, è stato dichiarato lo stato di emergenza e attivato un conto corrente per ricevere donazioni a favore delle zone più colpite. Gli interventi per riparare i danni maggiori sono stati tempestivi e ad oggi si sta facendo ancora molto per migliorare la situazione del Bellunese. Non sono mancati i finanziamenti da parte degli enti pubblici, ne sono un esempio le Camere di Commercio del Veneto che si sono dimostrate attive nel sostenere la popolazione e le imprese colpite da Vaia emanando un bando per l'erogazione di contributi a fondo perduto (Treviso-Belluno, 2019a). Nel frattempo anche l'Europa ha approvato la richiesta di un aiuto economico, finanziando la ricostruzione delle zone interessate con il Fondo di solidarietà, a distanza di quasi un anno dall'accaduto (Treviso-Belluno, 2019b).

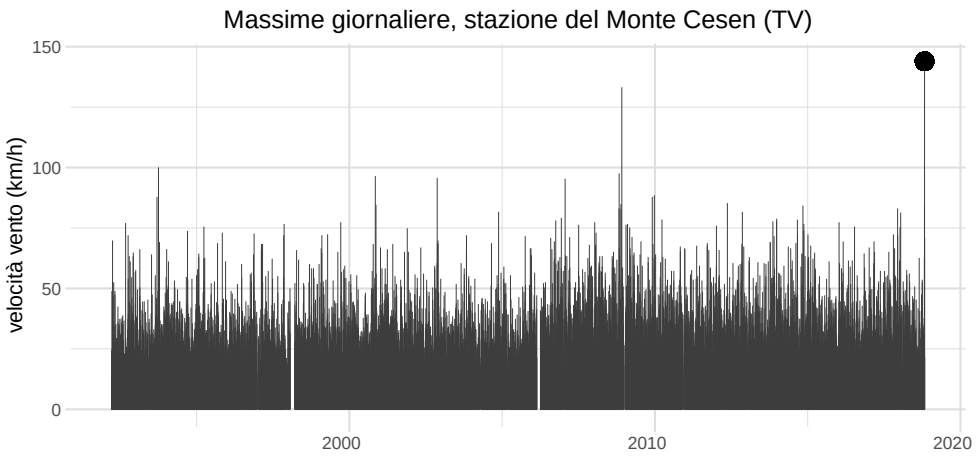

Figura 1.4: Serie storica della velocità del vento sul Monte Cesen. I valori si riferiscono ad una media oraria delle massime registrate ogni 10 minuti, di conseguenza il valore evidenziato dal punto nero non rappresenta il massimo assoluto 192km/h.

Figura 1.5: La piena del Piave a San Donà di Piave (VE) (`http://bit.ly/2n1TzCc`).

Capitolo 2

Metodi e modelli per l'analisi dei valori estremi

In ambito statistico i valori estremi sono definiti come gli eventi più rari, provenienti dalle code della distribuzione che descrive il fenomeno di interesse. Spesso la statistica si occupa di analizzare l'aspetto generale di un fenomeno, tralasciando lo studio degli eventi meno probabili. In alcuni contesti, però, è importante capire il comportamento delle code di una distribuzione per poter comprendere meglio il fenomeno nella sua interezza, soprattutto se eventi estremi portano a conseguenze rilevanti, come nel caso oggetto di questo studio. In ambito ambientale conoscere e prevedere correttamente comportamenti estremi, ad esempio del vento, e soprattutto la loro frequenza, può risultare in una prevenzione di disastri ambientali o in uno sfruttamento più efficiente di essi come risorse rinnovabili.
Nel seguito verranno illustrati l'approccio classico allo studio dei valori estremi ed uno alternativo basato sul paradigma bayesiano.

2.1 Teoria classica dei valori estremi

La teoria classica dei valori estremi si è sviluppata con lo scopo di descrivere gli eventi rari. Uno dei primi libri sull'argomento è *Statistics of extremes* (Gumbel, 1958). L'interesse è focalizzato sulla previsione di eventi estremi rispetto al comportamento standard di un fenomeno, non solo in un futuro prossimo, ma anche per periodi temporali molto ampi,

talvolta più estesi rispetto alla durata dei dati storici di cui si dispone. Per questo motivo è rilevante costruire modelli con basi solide che permettano di fare estrapolazioni valide nel lungo periodo. L'estrapolazione è infatti un'operazione molto delicata in ambito statistico e comporta una notevole quantità di incertezza in quanto prevedere l'andamento di un fenomeno in un contesto che non è stato ancora osservato, supponendo che il modello utilizzato sia adeguato, può portare a risultati completamente diversi dalla realtà se questa assunzione non è verificata (e non è possibile controllarne la validità).

2.1.1 Distribuzione generalizzata dei valori estremi

La teoria classica prevede due tipologie di modelli. Il primo approccio considera la variabile casuale $Y_n = \max\{X_1, \ldots, X_n\}$, dove $\{X_1, \ldots, X_n\}$ è una sequenza di variabili casuali indipendenti e identicamente distribuite con funzione di ripartizione F. Conoscendo la distribuzione F è facile calcolare la distribuzione del massimo

$$Pr(Y_n \leq z) = Pr(X_1 \leq z, \ldots, X_n \leq z) = Pr(X_1 \leq z) \cdot \ldots \cdot Pr(X_n \leq z) = F(z)^n. \quad (2.1)$$

Dovendo però stimare F, nell'ottenere la distribuzione di M_n si incorre in un problema di propagazione dell'incertezza non trascurabile. Di conseguenza negli anni si è preferito studiare direttamente uno stimatore di F^n. Un secondo problema è dovuto al fatto che F assume valori in [0,1], di conseguenza per ogni $z < z^+$, dove z^+ rappresenta l'estremo superiore del dominio di X, $F^n \to 0$ per $n \to \infty$. La soluzione è considerare $Y_n^* = (Y_n - b_n)/(a_n)$ per sequenze di valori appropriati $\{a_n > 0\}$ e $\{b_n\}$. Tutto ciò è specificato nel seguente teorema.

Teorema 1 *Se esiste una sequenza di costanti* $\{a_n > 0\}$ *e* $\{b_n\}$ *tali che*

$$Pr\left(\frac{Y_n - b_n}{a_n} \leq z\right) \to G(z) \qquad \textit{per } n \to \infty,$$

dove G *è una funzione di probabilità non degenere, allora* G *appartiene ad una delle seguenti famiglie:*

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

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

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

con $a > 0$ *e, per le famiglie II e III,* $\alpha > 0$.

Il Teorema 1 si deve a Fisher e Tippett (1928) e Gnedenko (1943). Le tre famiglie descritte nel Teorema 1 sono quelle delle distribuzioni di Gumbel, Frechét e Weibull. Esse possono essere unite in un'unica definizione: la distribuzione generalizzata dei valori estremi (GEV dall'inglese *generalized extreme value distribution*), in cui si distinguono in base al valore assunto dal parametro di forma ξ

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

dove per $\xi = 0$ Gumbel, $\xi > 0$ Frechét e $\xi < 0$ Weibull.
Il Teorema 1 garantisce che, qualsiasi sia la distribuzione F di X, $\lim_{n\to\infty} F^n$ è definito dalla distribuzione generalizzata dei valori estremi (2.2).

L'inferenza per questo tipo di modello, sia in ambito frequentista che bayesiano, è condotta utilizzando solo i valori considerati massimi (o minimi) all'interno di un blocco temporale o spaziale, scartando così tutto il resto dell'informazione veicolata dall'intero insieme di dati. Nonostante si possa considerare più di un evento estremo per blocco, resta il fatto che, se un periodo temporale presenta eventi più estremi di un altro, questi potrebbero non venire considerati e quindi verrebbe eliminata informazione utile. La scelta della divisione dei dati in blocchi è cruciale per ottenere risultati affidabili e deve essere raggiunto un compromesso tra varianza e distorsione. Infatti raggruppamenti troppo frequenti porterebbero ad avere una lunga sequenza di valori dei quali non tutti in realtà sono massimi e di conseguenza le stime risulterebbero distorte. Al contrario pochi blocchi limiterebbero le informazioni disponibili con l'effetto di aumentare la varianza delle stime. Inoltre, spesso la scelta dei blocchi temporali deve tenere conto della stagionalità di un fenomeno, ad esempio considerando per dati ambientali un intero anno invece di periodi

più brevi.

I parametri della distribuzione GEV possono essere stimati con il metodo della massima verosimiglianza. Tuttavia i risultati della teoria asintotica non possono essere applicati direttamente agli stimatori in quanto viene violata una delle condizioni di regolarità. Infatti è stato dimostrato che il dominio della distribuzione GEV è limitato da valori che sono funzione dei parametri.

Lo scopo di un'analisi degli eventi estremi è principalmente quello di comprendere la loro entità e la loro frequenza. Una volta stimati i parametri del modello, le quantità di interesse sono i quantili dei valori estremi v_p, ottenuti invertendo l'equazione (2.2)

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

dove $G(v_p) = 1 - p$. I v_p sono detti valori di ritorno e sono associati ai tempi di ritorno $\frac{1}{p}$ perché la probabilità che v_p sia ecceduto una volta è pari a p, ovvero è un evento che potrebbe accadere ogni $\frac{1}{p}$ periodi temporali. Spesso i tempi di ritorno vengono espressi come $t_p = -\log(1-p)$. L'aspetto principale dell'inferenza per un modello per valori estremi è dunque il grafico dei valori di ritorno, nel quale vengono rappresentati in ordinata i v_p e in ascissa $\log(t_p)$, dove la scala logaritmica ha lo scopo di enfatizzare le code. In Figura 2.1 vi è un esempio di come appare un grafico dei valori di ritorno al variare del parametro di forma ξ.

2.1.2 *Peak over threshold*

Il secondo metodo per l'analisi dei valori estremi proveniente dalla teoria classica prende in considerazione tutti i dati a disposizione e seleziona solo i valori oltre una certa soglia, definita dai dati stessi. In realtà questo approccio è collegato alla distribuzione generalizzata dei valori estremi e tale dualità è descritta nel seguente teorema.

Teorema 2 *Si consideri la variabile casuale*

$$Y_n = max\{X_1, \ldots, X_n\},$$

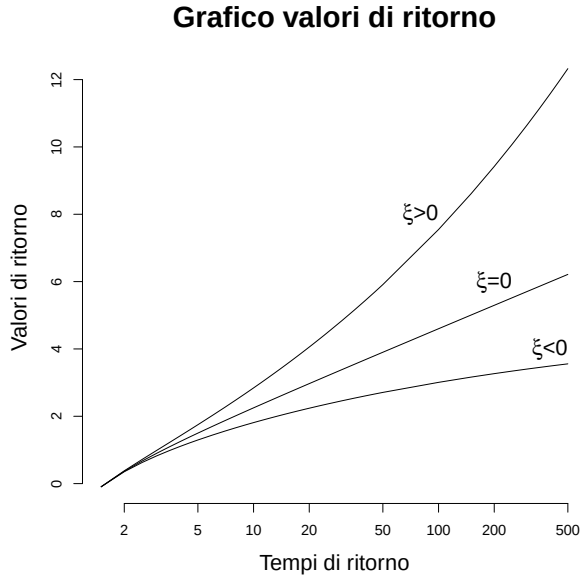

Figura 2.1: Esempio di come appare un grafico dei valori di ritorno con $\mu = 0$, $\sigma = 1$ e $\xi = (0.2, 0, -0.2)$.

dove $\{X_1, \ldots, X_n\}$ *è una sequenza di variabili casuali indipendenti e identicamente distribuite con funzione di ripartizione* F *e si consideri valido il Teorema 1. Allora, se la distribuzione che descrive il comportamento dei valori estremi di* X *è quella descritta nell'equazione* (2.2)*,* $W = X - u$ *condizionato a* $X > u$*, per* u *abbastanza grande (*u *è la soglia definita), segue approssimativamente la distribuzione di Pareto generalizzata*

$$H(w) = 1 - \left(1 + \frac{\xi w}{\widetilde{\sigma}}\right)^{-\frac{1}{\xi}}, \tag{2.4}$$

definita per $\{w : w > 0 \ e \ (1 + \frac{\xi w}{\widetilde{\sigma}}) > 0\}$*, dove* $\widetilde{\sigma} = \sigma + \xi(u - \mu)$*.*

Questo metodo è riconosciuto con il nome di *peak over threshold* (POT) (Balkema e De Haan, 1974; Pickands III, 1975, Davison e Smith, 1990). Anche in questo contesto, come in molti altri, per definire la soglia u bisogna trovare un compromesso tra varianza e distorsione (Coles, 2001). Infatti una soglia troppo bassa selezionerebbe dati che non sono valori estremi ed il modello fornirebbe stime distorte dei parametri, descrivendo un fenomeno che comprende sia valori estremi che regolari. Dall'altro lato, invece, una soglia troppo alta scarterebbe quasi tutti i dati costringendo a stimare il modello con poche osservazioni e di

conseguenza la varianza aumenterebbe ed il modello sarebbe poco affidabile.

Come nel primo caso, la stima dei parametri può essere ottenuta con il metodo della massima verosimiglianza. Di conseguenza si possono ottenere i valori di ritorno supponendo che $Pr(X > x) = p$. Dall'equazione (2.4) si ha che

$$\begin{aligned} Pr(X > x) =& Pr(X > u)Pr(X > x|X > u) \\ =& \zeta_u \left[1 + \xi \left(\frac{x-u}{\sigma}\right)\right]^{-\frac{1}{\xi}}, \end{aligned} \tag{2.5}$$

di conseguenza

$$\zeta_u \left[1 + \xi \left(\frac{v_p - u}{\sigma}\right)\right]^{-\frac{1}{\xi}} = p, \tag{2.6}$$

e si ottiene per $\frac{1}{p}$ sufficientemente grande tale che $v_p > u$

$$v_p = \begin{cases} u + \frac{\sigma}{\xi}\left[\left(\frac{\zeta_u}{p}\right)^{\xi} - 1\right], & \xi \neq 0; \\ u + \sigma \log\left(\frac{\zeta_u}{p}\right), & \xi = 0. \end{cases} \tag{2.7}$$

Entrambi gli approcci derivati dalla teoria classica dei valori estremi hanno in comune il fatto che, dalla stima del modello al calcolo di quantità come i valori e i tempi di ritorno, vengono usati solo i dati considerati estremi (in blocco od oltre una certa soglia). Tutta la parte della distribuzione a maggiore densità viene dunque esclusa dalle analisi. La conseguenza è una notevole perdita di informazione perché gli eventi estremi sono, per definizione, rari. Inoltre, i valori ritenuti regolari possono influenzare il comportamento estremo di un fenomeno e analizzarli può contribuire a comprenderlo meglio e garantire migliori previsioni ed estrapolazioni.

2.2 Un modello bayesiano gerarchico per i valori estremi

Il principale difetto della teoria classica è lo spreco di informazione derivante dai dati a maggiore densità di probabilità. È infatti inverosimile che il comportamento standard del fenomeno di interesse non influenzi l'andamento dei valori estremi, perciò risulta fondamentale riuscire a capire come sfruttare al meglio tutta l'informazione raccolta. Un ulteriore

aspetto negativo della teoria classica è dovuto al fatto che si basa sulla validità dell'approssimazione asintotica per $n \to \infty$. Ossia si assumono validi i risultati dei Teoremi 1 e 2 per campioni finiti. Tuttavia nel caso di serie storiche divise in blocchi temporali, n assume al massimo un valore finito noto, ad esempio se le serie sono giornaliere e vengono divise in periodi annuali $n = 365$.

Un nuovo approccio allo studio dei valori estremi viene da Zorzetto, Canale e Marani (2019). Per superare i limiti della teoria classica, l'attenzione viene focalizzata sull'espressione nell'equazione (2.11). Tramite la definizione gerarchica del modello e un approccio bayesiano all'inferenza si evita il ricorso all'argomentazione asintotica. Inoltre l'adottare un approccio bayesiano consente di sfruttare informazioni esterne. Infatti in un contesto come quello dei dati ambientali ci sono molti studi che consentono di avere informazioni dettagliate sui fenomeni di interesse. Queste conoscenze sono molto utili quando in ambito statistico si vuole modellare l'incertezza del fenomeno partendo da una solida base di assunzioni preliminari. L'inferenza bayesiana è ottima da questo punto di vista poiché permette di integrare in maniera naturale informazioni note a priori con l'informazione portata dai dati. Questo aspetto consente di ottenere previsioni a lungo termine affidabili anche in casi di poche osservazioni disponibili.

Una delle principali differenze tra il paradigma frequentista e quello bayesiano è come vengono considerate le quantità ignote, come i parametri di una distribuzione. Nel primo caso si assume che siano valori puntuali e ignoti, nel secondo invece sono variabili casuali a cui è associata una distribuzione che rappresenta la nostra incertezza rispetto al fenomeno. Nel paradigma bayesiano vengono dunque definite due quantità alla base di un qualsiasi modello. Si considerano infatti la distribuzione dei dati $p(y|\theta)$, o funzione di verosimiglianza, e la distribuzione a priori dei parametri $\pi(\theta)$. L'inferenza bayesiana si basa poi sulla ricerca della distribuzione a posteriori dei parametri del modello. Questa è definita come

$$\pi(\theta|y) = \frac{\pi(\theta)p(y|\theta)}{\int_{\Omega_\theta} \pi(\theta)p(y|\theta)d\theta} , \tag{2.8}$$

dove Ω_θ rappresenta il dominio di θ.

Rispetto ai modelli della teoria classica, la differenza più rilevante, però, è che si tratta di un modello che tiene conto dell'intera informazione portata dai dati. Questo è un punto

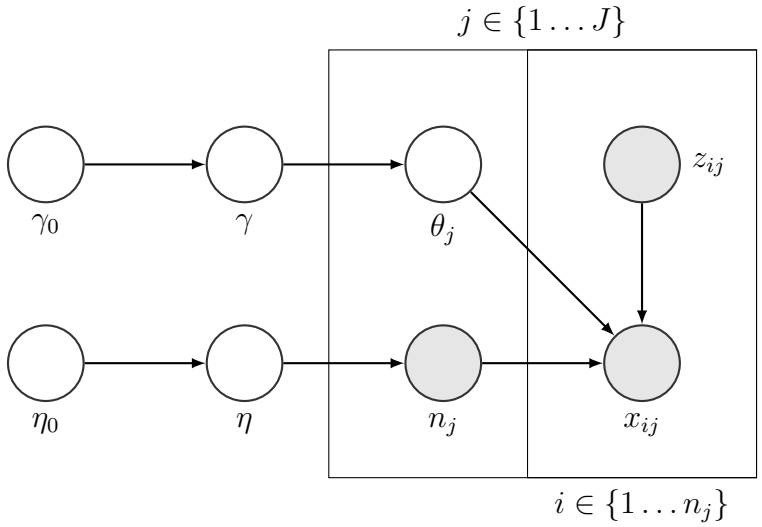

Figura 2.2: Struttura del modello bayesiano gerarchico per i valori estremi.

di forza perché, nonostante il contesto dei dati ambientali porti con sé un vasto background di conoscenze scientifiche, sfruttare l'osservazione del fenomeno nella sua interezza, anziché i soli valori considerati estremi, consente di ottenere risultati più precisi.

Un altro vantaggio di un modello bayesiano gerarchico per i valori estremi (HBEV dall'inglese *hierarchical bayesian extreme values*) è la flessibilità. In un modello gerarchico infatti, non esiste uno standard rigido da seguire ma in base al fenomeno di interesse si costruiscono i livelli scegliendo la distribuzione che meglio si adatta ai dati e le distribuzioni a priori per i parametri.
Nell'ambito di questo progetto di tesi è stato definito il modello descritto in Figura 2.2, dove x_{ij}, z_{ij} ed n_j rappresentano delle quantità osservate, rispettivamente la variabile di interesse (nel seguito sarà la velocità del vento), un'eventuale covariata (direzione del vento) ed il numero di valori osservati all'interno di un anno, definito come il blocco temporale di riferimento. In questo modo n_j permette di gestire sia i dati mancanti sia l'osservazione di un non evento in quanto registrare un valore pari a 0 significa, nel contesto di questo lavoro, che non c'è stato vento. θ_j raggruppa i parametri della distribuzione della velocità del vento x_{ij}. Sia per θ_j che per n_j vengono definite delle distribuzioni che consentono di modellare la variabilità tra i blocchi temporali e che dipendono da ulteriori parametri γ e η per i quali a loro volta vengono specificate delle distribuzioni a priori.

L'obiettivo principale dell'analisi dei valori estremi si ritrova nella stima della distribuzione dei massimi $Y_j = \max_i\{X_{ij}\}$ che, coerentemente con l'equazione 2.1, risulta

$$\zeta_j(y) = Pr(Y_j \leq y) = F(y|\theta_j)^{n_j}. \tag{2.9}$$

Essa dipende dai parametri incogniti θ e n. Se essi fossero noti, una stima a priori sarebbe data dal suo valore atteso

$$g(s, \eta, \gamma|x_{ij}, n_j) = \sum_{n=0}^{N_t} \int_{\Omega_\theta} F(s|\theta)^n p(\theta|\gamma) p(n|\eta) d\theta, \tag{2.10}$$

con N_t valore massimo assunto da n_j, $j = 1, \ldots, J$, Ω_θ dominio di θ. Essendo però sia θ che n variabili casuali, con distribuzioni che dipendono da ulteriori parametri incogniti, questa quantità non è direttamente calcolabile. L'equazione (2.10) descrive una funzione del parametro di interesse $F(y|\theta)^n$, pertanto per ottenere una stima di quest'ultimo si deve calcolare la media a posteriori di $g(\cdot)$

$$\begin{aligned} \hat{\zeta}(s) =& E(g(s, \eta, \gamma|x_{ij}, n_j)) \\ =& \int_{\Omega_\eta} \int_{\Omega_\gamma} P(\gamma|x_{ij}) P(\eta|n_j) g(s, \eta, \gamma|x_{ij}, n_j) d\eta d\gamma, \end{aligned} \tag{2.11}$$

dove Ω_η e Ω_γ sono i rispettivi domini dei parametri.
A partire da queste stime è possibile arrivare al calcolo dei valori e dei tempi di ritorno, questi ultimi definiti da

$$t_p(y) = \{1 - \zeta(y)\}^{-1}. \tag{2.12}$$

Nel grafico dei valori di ritorno vengono quindi rappresentati in ordinata i valori di ritorno e in ascissa i relativi tempi di ritorno. Scopo finale, infatti, come per i modelli della teoria classica, è quello di prevedere entità e frequenza di eventi estremi.

Zorzetto, Botter e Marani (2016) e Zorzetto, Canale e Marani (2019) mostrano una prima applicazione di questo modello su dati relativi a serie storiche di precipitazioni rilevate nel Nord Italia. I risultati mostrano che il modello bayesiano gerarchico riduce l'incertezza delle stime dei valori di ritorno rispetto ai modelli basati sulla distribuzione generalizzata dei valori estremi. Questo significa avere previsioni più precise nel lungo periodo.

2.3 Calcolo della distribuzione a posteriori tramite Hamiltonian Monte Carlo

Per le analisi presenti in questo progetto di tesi è stato utilizzato il software Stan tramite l'interfaccia presente in R con la libreria `rstan`. Stan è una piattaforma per la modellazione statistica, sviluppata per l'inferenza bayesiana, che usa algoritmi Markov Chain Monte Carlo (MCMC). In particolare la sua potenzialità è dovuta all'uso dell'algoritmo Hamiltonian Monte Carlo (HMC).
L'inferenza bayesiana si basa sulla ricerca della distribuzione a posteriori dei parametri del modello, ma spesso questa distribuzione non è riconoscibile o ha una forma complessa che porta a complicazioni nel calcolo di quantità inferenziali come il valore atteso o intervalli di credibilità. Di conseguenza è necessario ricorrere ad algoritmi di tipo MCMC, un tipo di approccio che permette di campionare valori dalla distribuzione a posteriori.
Quando ci si riferisce agli algoritmi MCMC, spesso si intende l'algoritmo di Metropolis-Hastings (Metropolis et al., 1953 e Hastings, 1970). Quest'ultimo può risultare in una lenta esplorazione del supporto della distribuzione perché si muove nel dominio con una passeggiata casuale rimanendo piuttosto vicino al valore proposto al passo precedente. Un algoritmo di tipo HMC, invece, migliora questa esplorazione riuscendo a spostarsi in un insieme del supporto a maggiore densità (Betancourt, 2017).
La base di un algoritmo HMC è una funzione $H(q, p)$ che dipende da un vettore di posizioni q e un vettore dei momenti p, entrambi di dimensione d. Le derivate parziali di $H(q, p)$ definiscono le equazioni di Hamilton

$$\begin{aligned} \frac{dq_i}{dt} &= \frac{\delta H}{\delta p_i} \\ \frac{dp_i}{dt} &= -\frac{\delta H}{\delta q_i} \end{aligned} \tag{2.13}$$

per $i = 1, \ldots, d$. Queste derivate definiscono delle funzioni $p(t)$ e $q(t)$ che permettono di calcolare i valori di (p, q) rispetto al tempo, ovvero di descrivere una dinamica hamiltoniana. In genere $H(q, p)$ viene definita come la somma di due quantità, $U(q)$ energia potenziale e $K(p)$ energia cinetica. Essa descrive il comportamento di un corpo in un sistema al variare del tempo t. Per poter implementare l'algoritmo è necessario prima discretizzare il tempo t.

In ambito statistico questo algoritmo viene utilizzato sostituendo ad $U(q)$ il logaritmo della funzione di densità da cui campionare (in un contesto bayesiano è la distribuzione a posteriori), cambiato di segno e a $K(p)$ il logaritmo della funzione di densità di una $N_d(0, M)$, con M matrice simmetrica definita-positiva, spesso scelta come multiplo della matrice identità.
Un algoritmo HMC ha due passi. Nel primo viene campionato un valore per p dalla $N_d(0, M)$. Nel secondo, partendo da (q, p) (nella prima iterazione q è inizializzato), viene simulata una dinamica hamiltoniana per L passi, alla fine dei quali viene proposto un valore (q^*, p^*) dove p^* è l'opposto del p ottenuto. A questo punto la proposta viene valutata come in un algoritmo di Metropolis e se rifiutata si copia il valore dell'iterazione precedente. Di conseguenza i valori ottenuti per q, rappresentano un campione dalla distribuzione a posteriori fornito dall'algoritmo (Neal, 2010).
HMC permette dunque di esplorare più velocemente il dominio di una funzione di densità rispetto ad una passeggiata casuale, richiedendo meno simulazioni e risultando in un campionamento più completo. Esso infatti sfrutta le informazioni del gradiente per ottenere la direzione lungo cui muoversi all'interno delle regioni a maggiore densità.

Capitolo 3

Analisi preliminare dei dati

I dati analizzati in questa tesi sono stati forniti da ARPA Veneto. Essi sono stati rilevati da 16 stazioni situate nelle zone alpine del Veneto, nelle quali tra il 27 e il 30 ottobre 2018 si è abbattuta la tempesta Vaia. La loro disposizione nel territorio si può vedere in Figura 3.1. I valori misurati si riferiscono alla velocità del vento, registrati come medie aritmetiche orarie delle massime rilevate ogni 10 minuti in m/s. Ad essi è associata la direzione del vento, espressa come l'ampiezza in gradi di un angolo con origine al nord.

Come si può notare in Tabella 3.1 l'installazione delle stazioni risale a momenti diversi e di conseguenza, dato che i valori terminano tutti il 31 ottobre 2018, ossia appena dopo il passaggio di Vaia, gli anni di osservazione a disposizione sono in numero diverso nelle 16 località. Un altro aspetto da considerare riguarda il fatto che per alcune stazioni il sensore che rileva la direzione del vento è stato attivato successivamente rispetto a quello della velocità, quindi nel caso in cui vengano studiate congiuntamente velocità e direzione, si farà riferimento alla data più recente come momento iniziale.

Inizialmente si sono considerati solo i dati relativi alla velocità del vento. In un successivo momento l'analisi di questo fenomeno ha mostrato delle particolarità dal punto di vista statistico. Ciò ha portato a considerare la velocità del vento in relazione alla direzione di provenienza e di conseguenza si sono integrate le nuove informazioni.

Tabella 3.1: Informazioni di base sulle stazioni.

Stazione	Altezza sensore	Inizio rilevazioni velocità	Anni velocità	Inizio rilevazioni direzione	Anni direzione
Asiago (VI)	10m	22/08/1997	22	26/06/2001	18
Cansiglio (BL)	5m	01/07/2003	16	15/07/2003	16
Caprile (BL)	5m	01/03/1984	35	03/05/1984	35
Longarone (BL)	5m	01/11/1991	28	01/11/1991	28
Lusiana (VI)	5m	15/11/1991	28	15/11/1991	28
Monte Cesen (TV)	10m	01/01/1992	27	19/03/1992	27
Monte Avena (BL)	5m	03/10/1985	34	08/11/1990	29
Monte Verena (VI)	10m	19/07/2004	15	19/07/2004	15
Passo Monte Croce (BL)	5m	07/11/1990	29	07/11/1990	29
Passo Pordoi (BL)	5m	10/04/1986	33	10/04/1986	33
Passo Valles (BL)	5m	01/11/1991	28	11/02/1992	27
Perarolo (BL)	5m	18/12/2002	17	18/12/2002	17
Piana di Marcesina (VI)	5m	01/06/1998	21	01/06/1998	21
Quero (BL)	5m	14/11/2002	17	14/11/2002	17
Rifugio la Guardia (VI)	5m	15/11/1991	28	15/11/1991	12
Sella Ciampigotto (BL)	5m	09/08/2005	14	09/08/2005	14

3.1 Velocità del vento

L'obiettivo di base di questo progetto di tesi è analizzare i valori storici relativi alla velocità del vento, uno dei due eventi che ha caratterizzato la tempesta Vaia. Al fine di valutare quanto i valori osservati a fine ottobre 2018 fossero anomali rispetto ad un andamento regolare del fenomeno, e perciò definibili estremi, il primo passo è stato quello di capire come si distribuissero i dati. Questo è stato fatto considerando separatamente ogni anno di ciascuna stazione. Studi precedenti hanno affermato che la velocità del vento segue una distribuzione di Weibull, ossia presenta un andamento asimmetrico con una coda

Figura 3.1: Localizzazione stazioni.

destra pesante (Li e Zhi, 2016). Se per alcune stazioni questa assunzione sembra soddisfatta, in alcuni casi i dati presentano due mode, diverse a seconda della stazione.

Nelle Figure 3.2 e 3.3 si possono notare gli andamenti differenti che si riscontrano in alcune stazioni (gli altri grafici sono riportati in Appendice). Ad esempio nel caso del Monte Cesen (TV) la distribuzione di Weibull sembra adattarsi perfettamente ai dati che presentano una sola moda e un'asimmetria positiva. Questo si verifica in misura minore anche per Longarone, Monte Avena, Monte Verena, Passo Pordoi e Sella Ciampigotto. La stazione di Caprile, invece, mostra una moda principale a sinistra, alla quale se ne affianca a destra un'altra meno marcata. Questo fa sì che una distribuzione unimodale non sia adeguata a descrivere il comportamento del fenomeno. Lo stesso si riscontra anche per Asiago, Cansiglio, Marcesina, Perarolo e Rifugio la Guardia. Anche Passo Valles e Lusiana mostrano una coda destra più pesante rispetto a quella di una distribuzione di Weibull. Infine le stazioni del Passo Monte Croce Comelico e di Quero si caratterizzano per una bimodalità che vede

il picco principale più spostato verso destra, accanto ad una moda più o meno marcata nei valori inferiori. Questi comportamenti apparentemente diversi tra loro potrebbero però essere originati da un meccanismo simile.

Un aspetto interessante è che all'interno di una stessa stazione, ma in anni diversi, il comportamento dei dati resta simile, con la seconda moda, dove presente, più o meno marcata. Si è cercato dunque di capire se questa bimodalità fosse dovuta ad una stagionalità ma si è visto che dividendo i dati in periodi semestrali i risultati non cambiavano analiticamente. Per evitare di cadere nella ricerca di periodi plurimensili che spiegassero la bimodalità, si è scelto di percorrere un'altra strada.
Pensando anche alla conformazione del territorio montano, si è ipotizzato che la motivazione di questo comportamento potesse essere data dalla direzione di provenienza del vento. Questo è stato ritenuto un aspetto interessante dell'analisi da un punto di vista statistico e di conseguenza si sono condotte delle analisi sui dati relativi alla direzione.

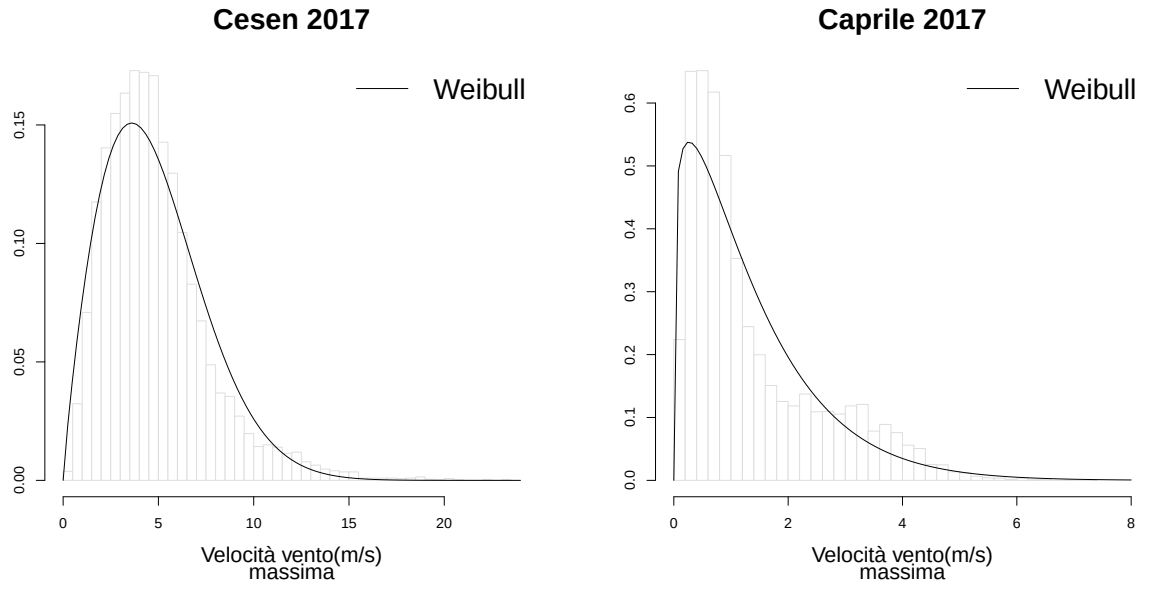

Figura 3.2: Esempi di comportamento della velocità del vento in diverse stazioni nell'anno precedente a Vaia. La linea continua rappresenta l'adattamento di una distribuzione di Weibull.

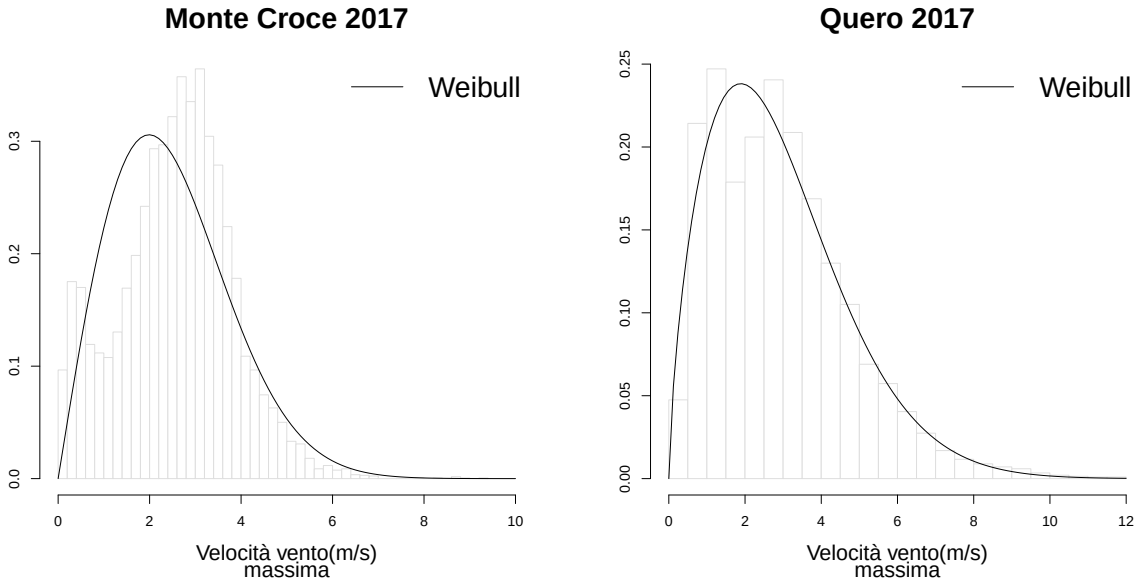

Figura 3.3: Esempi di comportamento della velocità del vento in diverse stazioni nell'anno precedente alla tempesta Vaia. La linea continua rappresenta l'adattamento di una distribuzione di Weibull.

3.2 Direzione di origine del vento

I dati relativi alla direzione del vento sono stati registrati come l'angolo tra il nord e la direzione osservata, dunque come una variabile quantitativa continua limitata in $(0, 360)$. Di conseguenza $0°, 90°, 180°, 270°$ rappresentano rispettivamente nord, est, sud e ovest (ARPAV, 2019). Nei dati a disposizione le direzioni rilevate sono 16 e corrispondono ai 4 punti cardinali, alle direzioni intermedie (nord-est, nord-ovest, sud-est, sud-ovest) e alle loro combinazioni (nord nord-est, est nord-est, ecc.). Nell'insieme di dati vengono rappresentate in una sequenza di valori da $0°$ a $337.5°$ a intervalli di $22.5°$, rendendo così discreta questa variabile.

La Figura 3.4 mostra quali sono le direzioni di origine del vento in alcune stazioni. Siccome i dati relativi alla velocità del vento di una singola stazione avevano un comportamento simile nei vari anni, in questo caso non si è fatta una distinzione per blocchi temporali ma si è considerato l'intero periodo di osservazione.

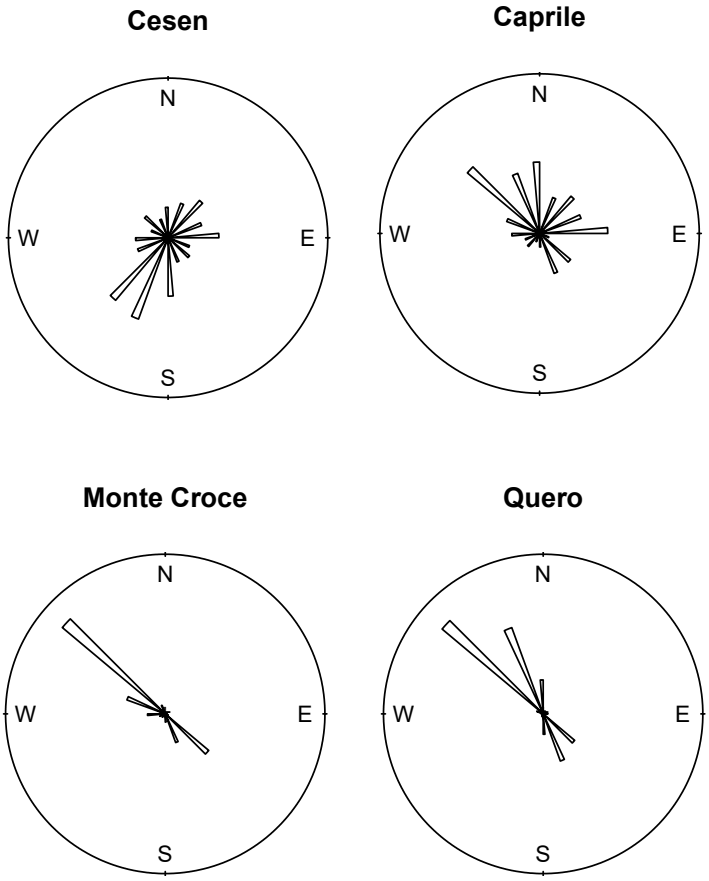

Figura 3.4: Esempi di distribuzione della direzione del vento.

Se si guardano anche le Figure 3.2 e 3.3 si può notare come nei casi in cui la velocità del vento mostri una marcata bimodalità, questa è presente anche nella distribuzione empirica della direzione. Da qui nasce l'ipotesi che la direzione possa avere una certa influenza sulla velocità del vento, sia perché i venti sono differenti a seconda della provenienza, sia perché la conformazione stessa del territorio montano può contribuire a limitare o enfatizzare l'effetto di alcuni venti. Anche in Figura 3.5 si può notare come la direzione di origine determini in alcune stazioni valori mediani differenti di velocità osservate. Nel caso invece del Monte Cesen i boxplot sono ben allineati tra di loro, a conferma che qualsiasi sia la provenienza del vento la velocità assume valori all'interno di uno stesso intervallo.

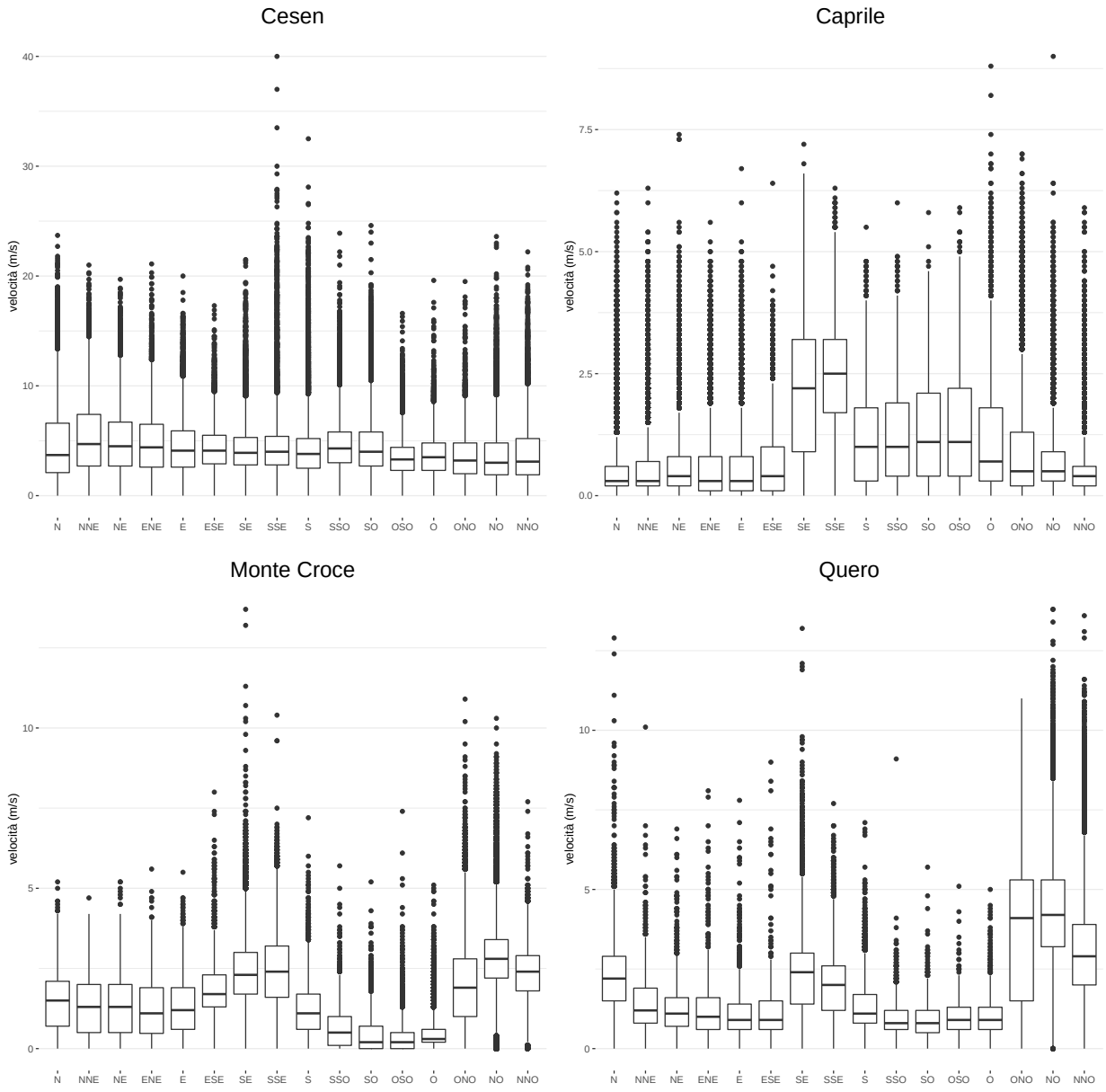

Figura 3.5: Esempi di come la distribuzione della velocità del vento varia in base alla direzione in alcune stazioni.

Nel seguito delle analisi sono state considerate solo le stazioni Monte Cesen, Caprile, Monte Croce e Quero. La scelta è stata effettuata valutando in primo luogo le differenze osservate nell'andamento della velocità del vento. Infatti si osserva l'unimodalità per il

Monte Cesen e delle bimodalità particolari nelle altre tre stazioni. Inoltre si è valutata la lunghezza delle serie storiche per avere abbastanza anni a disposizione. La stazione di Quero è più recente rispetto alle altre tre ma aveva un comportamento differente e questo è il motivo per cui è stata studiata. Infine si è considerata la percentuale di dati mancanti all'interno delle varie stazioni, combinando sia i valori della velocità del vento che quelli della direzione. La stazione di Caprile presenta percentuali di valori mancanti elevate fino all'anno 1991, di conseguenza si è preferito partire dal 1992, riducendo così il numero di anni di osservazione utilizzati. Un'eccezione è stata fatta per il Monte Cesen, che per l'anno 2013 presenta il 43.7% di dati mancanti, ma considerando l'intero insieme di dati la percentuale si abbassa al 10.5%. Dato che il comportamento di questa stazione è sembrato più fortemente unimodale rispetto alle stazioni con andamenti simili, si è preferito non sostituirla. Per quanto riguarda le altre località, la percentuale massima di dati mancanti in un anno risulta 23.2% per Caprile, con una percentuale sull'intero insieme di dati del 6.2%, per Monte Croce rispettivamente 29.8% e 4.9% e per Quero 11.8% e 2.3%.

Capitolo 4

Un modello per la distribuzione della velocità del vento

Nel seguito verranno proposti alcuni modelli mistura per valutare se sia possibile utilizzarli per descrivere il comportamento bimodale osservato nei dati. Per fare ciò, i modelli introdotti nella prossima sezione verranno testati su campioni simulati nella Sezione 4.1.1. Il modello più adeguato sarà poi inserito in un modello bayesiano gerarchico nel Capitolo 5. Quest'ultimo modello, insieme alla mistura dipendente proposta nella prossima sezione, sono i principali contributi metodologici di questo lavoro.

4.1 Modello mistura

Come anticipato nella Sezione 3.1, il fatto che i dati relativi alla velocità del vento non seguano la distribuzione di Weibull ma presentino code più pesanti o una bimodalità più o meno marcata ha suscitato un interesse modellistico. Una prima ipotesi è che si tratti di una distribuzione mistura nella quale due diverse distribuzioni contribuiscono ad evidenziare le mode. Questa idea può essere descritta dal modello

$$g(x) = \pi f_1(x) + (1 - \pi) f_2(x) \tag{4.1}$$

dove $g(x)$ è la funzione di densità di X, vento, e $f_1(\cdot)$ e $f_2(\cdot)$ sono delle opportune distribuzioni di Weibull con $\pi \in (0, 1)$. In letteratura ci sono stati degli studi per valutare la

differenza tra varie tipologie di misture riguardo alla descrizione della velocità del vento. Carta, Ramírez e Velázquez (2009) concludono che la mistura di due distribuzioni di Weibull è la più adatta (tra quelle indagate) nei casi di bimodalità ed inoltre fornisce risultati più accurati rispetto ad una singola componente Weibull anche in presenza di una sola moda. Anche i risultati ottenuti da Ouarda e Charron (2018) e Kollu et al. (2012) mostrano che una distribuzione mistura è una soluzione adeguata. Nel primo caso viene privilegiata una mistura con due componenti Gumbel mentre nel secondo una mistura gamma-Weibull, ma in entrambi gli studi la performance ottenuta usando due Weibull è molto simile se non equivalente alle soluzioni ritenute migliori.

Come si nota in Figura 4.1, le $f_1(\cdot)$ e $f_2(\cdot)$ scelte si differenziano nei parametri in modo da cogliere i diversi andamenti nei dati. Infatti sono state considerate una distribuzione di Weibull con una moda che prevale verso valori piccoli e un'altra con una moda meno accentuata e concentrata su valori più grandi.

In un secondo momento si è ipotizzato che un motivo alla base della bimodalità potesse essere la direzione di origine del vento. Ad esempio un vento che proviene da nord potrebbe avere una velocità media diversa rispetto ad un vento che proviene da sud, a causa anche della conformazione del territorio montano che può influenzare le raffiche di vento. Di conseguenza un modello più dettagliato è

$$\begin{aligned} g(x,z) &= \pi(z) f_1(x) + (1-\pi(z)) f_2(x) \\ \pi(z) &= \frac{e^{\tau(z)}}{1+e^{\tau(z)}} \\ \tau(z) &= \alpha + \beta_1 \sin(z) + \beta_2 \cos(z), \end{aligned} \tag{4.2}$$

dove z è una variabile che descrive la direzione di origine del vento. Utilizzare le funzioni trigonometriche garantisce che π assuma lo stesso valore in 0° e 360° perché rappresentano lo stesso angolo, come si nota nell'esempio in Figura 4.4. Si noti che considerare la direzione del vento come covariata ha un'implicazione non banale in ambito pratico. Infatti, per poter sviluppare un modello predittivo di questo tipo, occorre riuscire ad avere l'informazione riguardo alla direzione in anticipo rispetto alla velocità. Diversamente si potrebbe stimare congiuntamente la distribuzione del vento e la sua direzione. Questo studio esula tuttavia dagli scopi di questa tesi ma rappresenta un'interessante estensione futura del modello.

Prima di ottenere dei risultati sui dati a disposizione, si è voluto approfondire l'aspetto della stima dei parametri relativi ad una mistura. Il problema principale di questi modelli, infatti, è che quando le varie componenti non sono così differenziate tra loro, identificarle risulta complicato e le stime dei parametri potrebbero non essere accurate. Di conseguenza tramite un approccio bayesiano è auspicabile avere informazioni a priori molto informative. Il problema della non identificabilità è cruciale quando l'algoritmo di stima non riesce a separare le osservazioni nelle varie componenti e di conseguenza non riesce a stimare i parametri delle distribuzioni della mistura.

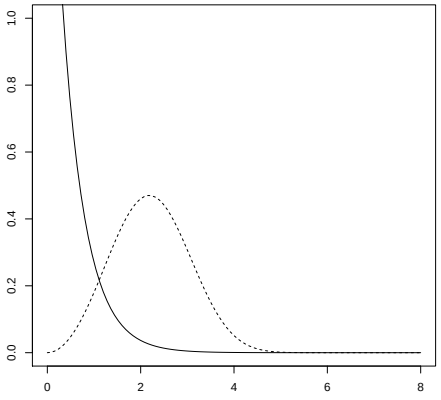

Figura 4.1: Rappresentazione delle distribuzioni Weibull(1,0.5) e Weibull(3,2.5) (linea tratteggiata) utilizzate nelle simulazioni.

In questo progetto di tesi la necessità di uno studio di simulazione, riportato nella prossima sezione, nasce anche dal fatto che nella letteratura citata l'inferenza era stata condotta secondo il paradigma frequentista.

4.1.1 Studio di simulazione

Per testare le capacità di identificare correttamente i parametri di una mistura è stato impostato uno studio di simulazione. Supponendo valida l'Equazione (4.1) si sono simulati 10 campioni di 10000 osservazioni per ogni valore di π ipotizzato in $0.1, 0.2, \ldots, 0.9$.

Le componenti della mistura sono due Weibull con parametri di forma e scala differenti: Weibull(3,2.5) e Weibull(1,0.5) (Figura 4.1). Al variare della proporzione π si riescono a cogliere tutti gli andamenti visti nelle serie originali, riassunti in Figura 4.2 (gli altri grafici sono riportati in Appendice).

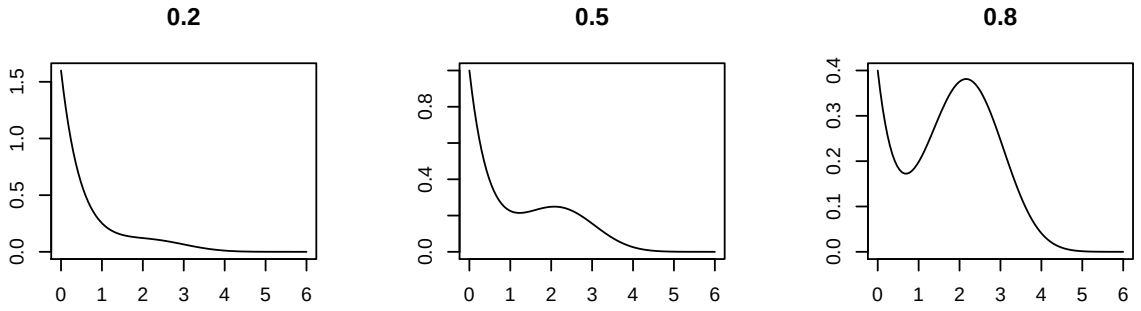

Figura 4.2: Esempi di comportamento di una mistura con due componenti Weibull per $\pi \in \{0.2, 0.5, 0.8\}$.

Inizialmente sono state considerate delle distribuzioni a priori molto informative per i parametri. Ossia, conoscendo il valore di π, si è scelta una beta(a, b) con moda nel vero valore di π, data da $(a-1)/(a+b-2)$. Per i parametri di forma e scala delle due Weibull, invece, sono state utilizzate delle lognormali(μ, σ) con mediana (e^{μ}) pari al vero valore dei parametri usati nella simulazione e σ piccola.

Successivamente, per vedere se il modello potesse fornire stime accurate anche senza

Tabella 4.1: A priori informative per i parametri di forma e scala delle due distribuzioni di Weibull.

Weibull	forma (k)	scala (λ)
Weibull(1,0.5)	lognormale(0,0.5)	lognormale(0,1)
Weibull(3,2.5)	lognormale(1.1, 0.25)	lognormale(0.9, 0.25)

precise informazioni a priori, sono state usate una distribuzione beta(1,1), che corrisponde ad una Uniforme(0,1), per π, mentre il parametro σ delle lognormali è stato raddoppiato, aumentando così la varianza.

Il software Stan è stato utilizzato per campionare dalla distribuzione a posteriori derivata dal modello mistura (4.1) tramite un approccio HMC discusso nella Sezione 2.3. La distribuzione a posteriori risulta

$$\begin{aligned} p(\pi, \boldsymbol{k}, \boldsymbol{\lambda}|x) &= p(\pi) \times p(k_1) \times p(k_2) \times p(\lambda_1) \times p(\lambda_2) \times g(x) \\ g(x) &= \prod_{i=1}^{n} \pi f_1(x_i, k_1, \lambda_1) + (1-\pi) f_2(x_i, k_2, \lambda_2). \end{aligned} \tag{4.3}$$

Sono state considerate 3000 iterazioni, di cui 1500 di *burn in*, ripetute per 4 catene indipendenti, per un totale di 6000 campioni MCMC.

In un secondo momento si è valutata l'ipotesi che il parametro π, che definisce la distribuzione a cui appartiene ciascuna osservazione, potesse dipendere dalla direzione del vento secondo l'Equazione (4.2). Di conseguenza è stato simulato un unico vettore con valori da 0° a 337.5° a intervalli di 22.5°, da usare in tutte le successive simulazioni della velocità del vento. Per rispecchiare il comportamento dei dati, la simulazione della direzione è stata fatta campionando i valori con probabilità proporzionale alla frequenza nell'intero insieme di dati e se ne è ottenuta una rappresentazione in Figura 4.3.

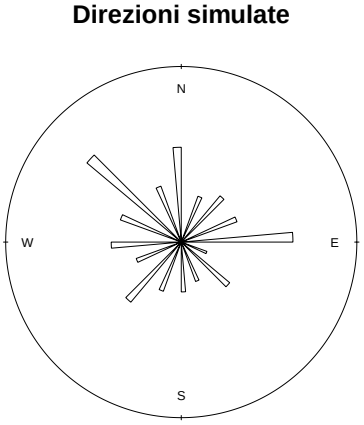

Figura 4.3: Rappresentazione della frequenza delle direzioni simulate per il modello mistura con covariata.

Con la stessa logica della simulazione precedente sono stati simulati 10 campioni di 10000 osservazioni provenienti dalla mistura con le due componenti Weibull utilizzate pre-

cedentemente, questa volta definite usando $\text{logit}(\pi(z_i)) = \alpha + \beta_1 \sin(z_i) + \beta_2 \cos(z_i)$. I valori di α e $\boldsymbol{\beta}$ scelti per le simulazioni sono elencati in Tabella 4.2 e sono tali per cui gli andamenti simulati ricordano il comportamento dei dati originali, alcuni riassunti in Figura 4.5 (gli altri grafici sono riportati in Appendice). Come distribuzioni a priori si sono mantenute le lognormali per i parametri di forma e scala delle Weibull, mentre per α e $\boldsymbol{\beta}$ si sono utilizzate rispettivamente Normali di media pari al vero valore dei parametri e deviazione standard pari a 0.5. Successivamente per rendere le a priori di α e $\boldsymbol{\beta}$ meno informative, le deviazioni standard sono state raddoppiate.

Tabella 4.2: Valori scelti per α e $\boldsymbol{\beta}$ negli scenari studiati per il modello mistura con covariata.

Scenario	1	2	3	4	5	6	7	8	9
α	-2.5	-2.5	-1	-0.5	0	0.5	1	1.5	2.5
β_1	1	-2.5	1	1	1	1	1	1	1
β_2	1	-2.5	1	1	1	1	1	1	1

Con Stan si è effettuato il campionamento dalla distribuzione a posteriori basata sull'Equazione (4.2)

$$
\begin{aligned}
p(\pi, \boldsymbol{k}, \boldsymbol{\lambda} | x, z) &= p(\alpha) \times p(\beta_1) \times p(\beta_2) \times p(k_1) \times p(k_2) \times p(\lambda_1) \times p(\lambda_2) \times g(x, z) \\
g(x, z) &= \prod_{i=1}^{n} \pi(z_i) f_1(x_i, k_1, \lambda_1) + (1 - \pi(z_i)) f_2(x_i, k_2, \lambda_2) \\
\pi(z_i) &= \frac{e^{\alpha + \beta_1 \sin(z_i) + \beta_2 \cos(z_i)}}{1 + e^{\alpha + \beta_1 \sin(z_i) + \beta_2 \cos(z_i)}}
\end{aligned}
\tag{4.4}
$$

dove x indica i valori della velocità del vento mentre z la direzione. Esattamente come nel caso precedente sono state considerate 3000 iterazioni, di cui 1500 di *burn in*, ripetute per 4 catene indipendenti.

4.1.2 Risultati

L'interesse di questo progetto di tesi è la corretta stima della probabilità della coda destra delle distribuzioni mistura considerate. I modelli mistura presentano un problema

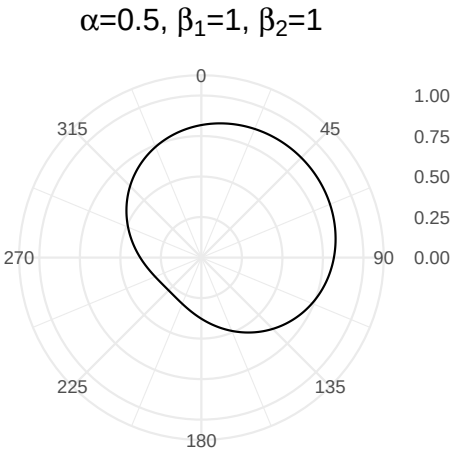

Figura 4.4: Esempio di $\pi(x) = \alpha + \beta_1 \sin(x) + \beta_2 \cos(x)$ per specifici α e $\boldsymbol{\beta}$.

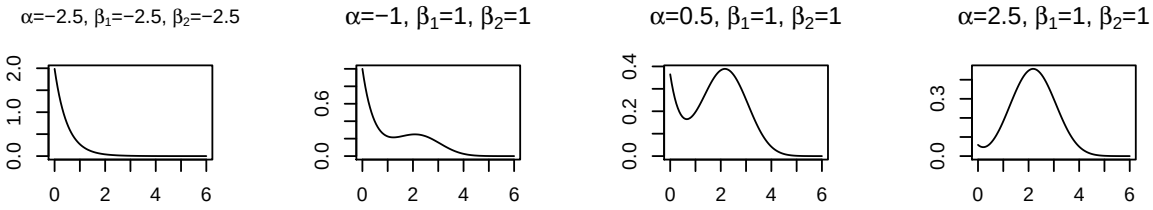

Figura 4.5: Esempi di comportamento di una mistura con due componenti Weibull definita da una covariata. Sono considerati diversi valori di α e $\boldsymbol{\beta}$ per la direzione nord.

di non identificabilità dei parametri che rende la stima di questi ultimi complicata. Questa situazione può portare, ad esempio, ad uno scambio delle componenti nel corso delle iterazioni, ossia non c'è un ordine fisso con cui entrano nella mistura ma possono essere invertite. Inoltre un altro aspetto importante è che diversi parametri, dunque in questo caso diverse distribuzioni di Weibull combinate con pesi appropriati, possono portare alla stima della stessa densità.

Le catene relative al campionamento dei parametri del modello dalla distribuzione a posteriori non convergono verso un unico valore ma i risultati non devono essere valutati in termini di convergenza di queste catene, bensì di quelle relative a funzionali statistici identificabili quali la probabilità nella coda destra. L'obiettivo di questo studio, infatti, non è identificare correttamente i parametri di una distribuzione mistura, ma ottenere una

descrizione accurata del comportamento della coda destra e pertanto dei valori 'anomali'.

A questo scopo sono stati calcolati i quantili di livello 0.85, 0.9, 0.95, 0.975 e 0.99 per ogni scenario, sia per quanto riguarda la mistura semplice sia per il modello che include la direzione del vento. Successivamente per ogni modello stimato si è calcolato $P(X > q_p)$ per ogni quantile. Questo è stato fatto per ogni valore dei parametri campionato nelle 4 catene dopo il *burn in*, ottenendo dunque altre 4 catene con un valore della probabilità della coda. Successivamente si sono riassunti i risultati in una media complessiva delle 4 catene.

I primi due grafici mostrati in Figura 4.6 si riferiscono al calcolo della probabilità della coda destra per ogni modello stimato in ogni scenario. I quantili sono stati ottenuti analiticamente dalla distribuzione mistura utilizzata per simulare i campioni. Da questi grafici si evince che, a prescindere dalla mancata convergenza delle catene e di conseguenza stime diverse dei parametri, il comportamento nella coda destra dei modelli stimati è compatibile con il vero modello. Questo suggerisce che la non identificabilità delle misture è un problema che non influenza le analisi proposte in questo progetto di tesi perché diverse combinazioni di parametri possono descrivere la stessa distribuzione mistura e in particolare la relativa coda destra. Anche nel caso in cui vengano utilizzate a priori meno informative il comportamento della coda destra ottenuto dalle stime è in linea con quello del processo generatore dei dati.

Per analizzare i risultati ottenuti dalla stima del un modello mistura definito da (4.2), nei campioni simulati con la direzione sono stati seguiti gli stessi passi utilizzati nel caso precedente. Per una visione più dettagliata si sono calcolate le probabilità della coda destra per ogni modalità assunta dalla direzione. Anche in questo contesto si può notare, nei grafici in basso in Figura 4.6, che le stime delle probabilità nelle code non si discostano dal vero valore.

Dopo aver valutato la performance del modello sotto condizioni di corretta specificazione, si è cercato di vedere se e quanto una errata specificazione potesse influire sulle stime delle probabilità delle code. Di conseguenza si sono stimati, per i campioni simulati con la direzione, i parametri di un modello mistura senza covariata e quelli di una distribuzione di Weibull, mentre per i campioni simulati dalla mistura semplice solo i parametri della distribuzione di Weibull. Dato che il modello da stimare non è in linea con il processo ge-

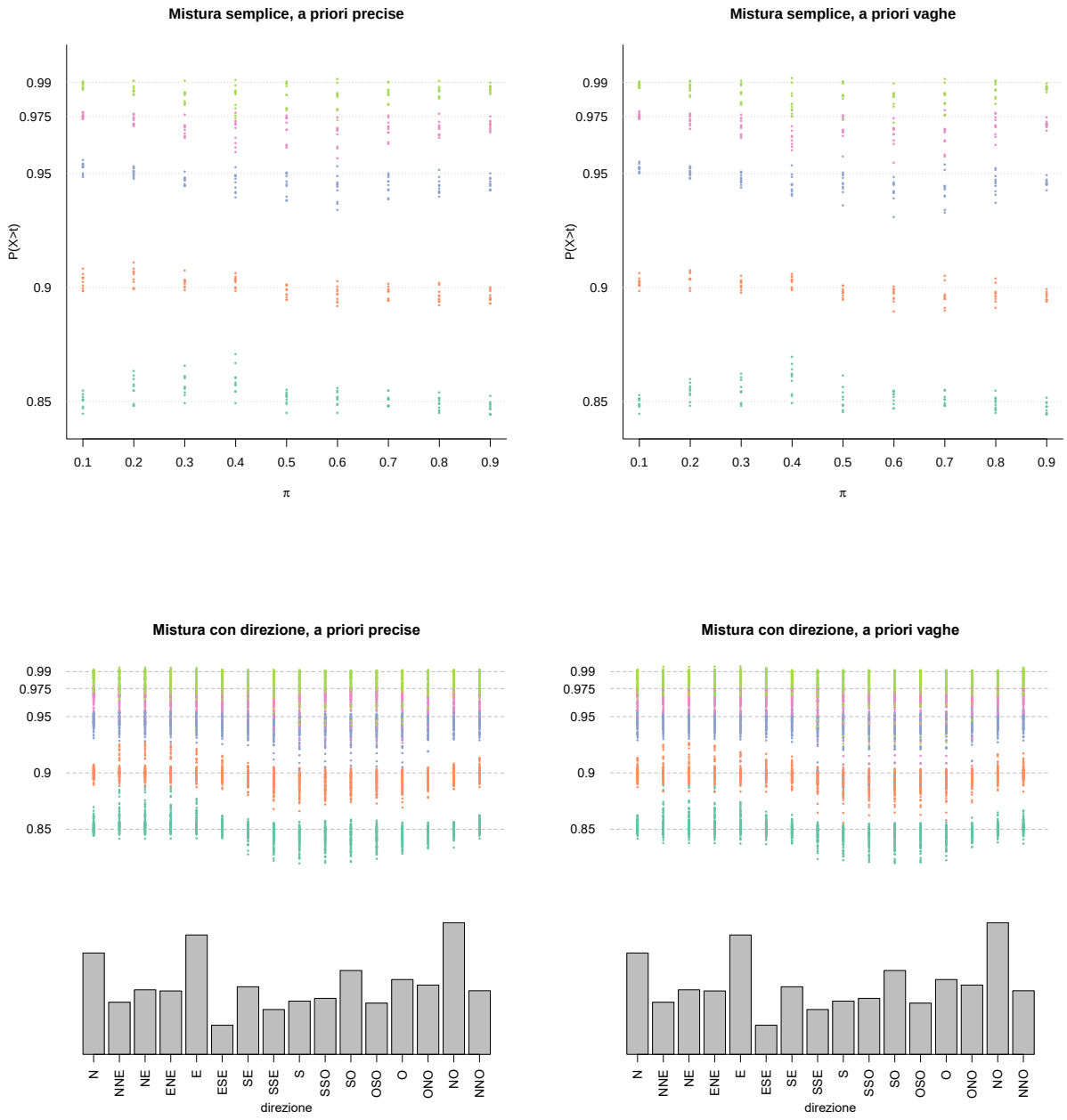

Figura 4.6: Stime delle probabilità nelle code per ognuno dei 10 campioni utilizzati nei diversi scenari. In alto per la mistura senza direzione, in basso per la mistura con la covariata. A sinistra i risultati relativi ad una stima con distribuzioni a priori precise, a destra con a priori meno informative. Nei grafici relativi alla mistura dipendente viene riportata la distribuzione empirica della direzione.

neratore dei dati, le a priori considerate sono state quelle meno informative. Le Figure 4.7 e 4.8 riassumono i risultati ottenuti in tutti i contesti per il quantile 0.95, per gli altri quantili i grafici sono riportati in Appendice (Figure 26 - 33).

Come si può ben notare, quando viene stimata solo una componente Weibull anziché la mistura, la probabilità della coda destra stimata si discosta nettamente da quella effettiva. L'andamento del secondo grafico in Figura 4.7 è particolare, questo perché gli scenari sono stati costruiti assegnando alla Weibull(3,2.5) una proporzione sempre maggiore, di conseguenza quando π è piccolo questa componente, che ha una moda meno spiccata rispetto alla Weibull(1,0.5) (Figura 4.1) tende a venire oscurata dall'altra e il comportamento è molto più simile ad una distribuzione unimodale.

La Figura 4.8 mostra i risultati ottenuti quando il vero modello si basa sulla direzione del vento. Per mostrare un parallelismo con i grafici in basso in Figura 4.6, sono stati considerati i risultati raggruppando secondo la direzione. Infatti calcolando $P(X > t)$, il quantile t, ottenuto analiticamente dalla distribuzione utilizzata per simulare i campioni, è diverso a seconda della direzione. Se non si inserisce la covariata nel modello le stime sembrano stabili per alcune direzioni, mentre per altre presentano una variabilità elevata.

Questo accade in misura maggiore se si sceglie come modello da stimare la sola distribuzione di Weibull. Gli effetti di una errata specificazione del modello sono quindi negativi per una stima corretta della probabilità nelle code.
Questo suggerisce che, se effettivamente la velocità del vento dipendesse dalla direzione, è meglio sfruttare questa informazione perché contribuisce ad ottenere stime più precise e coerenti delle probabilità nella coda destra.

Un ulteriore controllo è stato fatto confrontando il processo generatore dei dati con le densità stimate, sia sotto corretta che errata specificazione del modello. Nei grafici in Figura 4.9 si può notare come la singola distribuzione di Weibull possa essere adeguata solo nel caso in cui la bimodalità sia quasi inesistente. Già quando $\pi = 0.2$ l'adattamento non segue la vera funzione di densità e le differenze maggiori si ottengono quando π assume valori centrali rispetto al suo dominio e la bimodalità risulta più marcata.

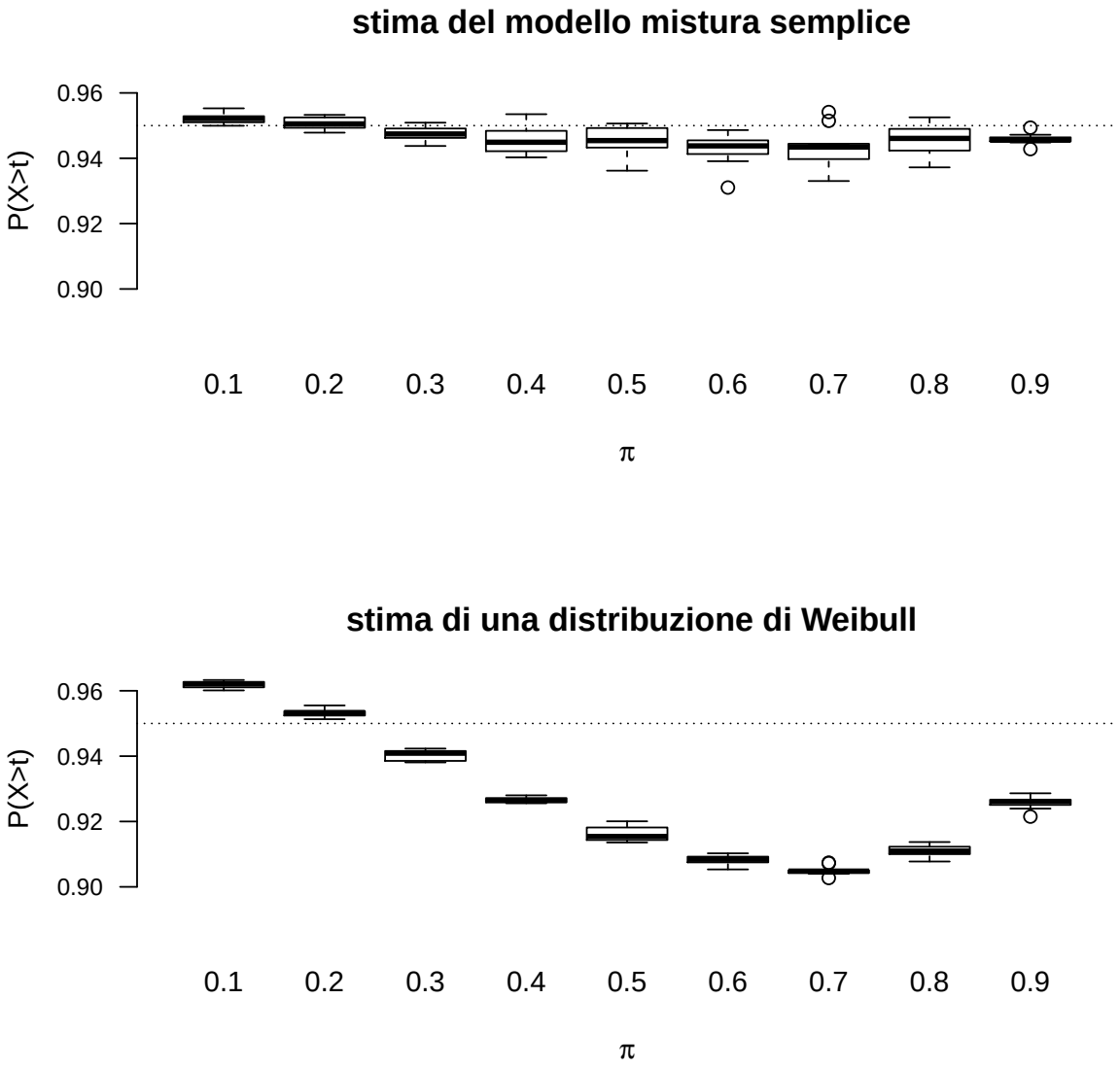

Figura 4.7: Stima delle probabilità nelle code per il quantile 0.95 per i campioni simulati dalla mistura semplice. I due grafici si riferiscono alle stime ottenute adattando il modello mistura corretto e una distribuzione di Weibull. Vengono distinti gli scenari definiti da π

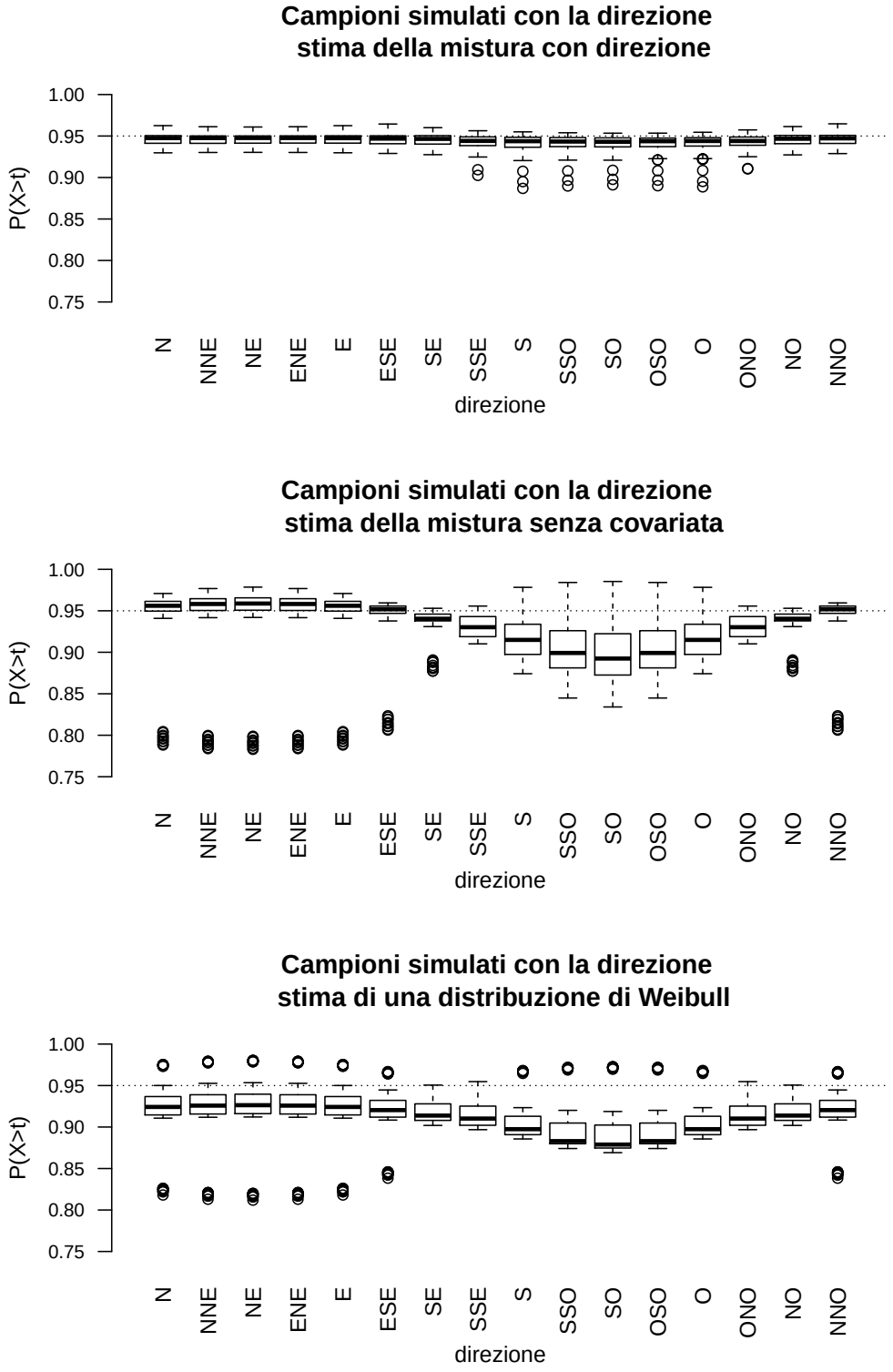

Figura 4.8: Stima della probabilità nelle code per il quantile 0.95 per i campioni simulati dalla mistura definita con la direzione. I grafici si riferiscono alle stime ottenute adattando il modello mistura corretto, il modello mistura senza la covariata e una distribuzione di Weibull. Vengono raggruppati i risultati ottenuti nei diversi scenari e distinti solo sulla base della direzione.

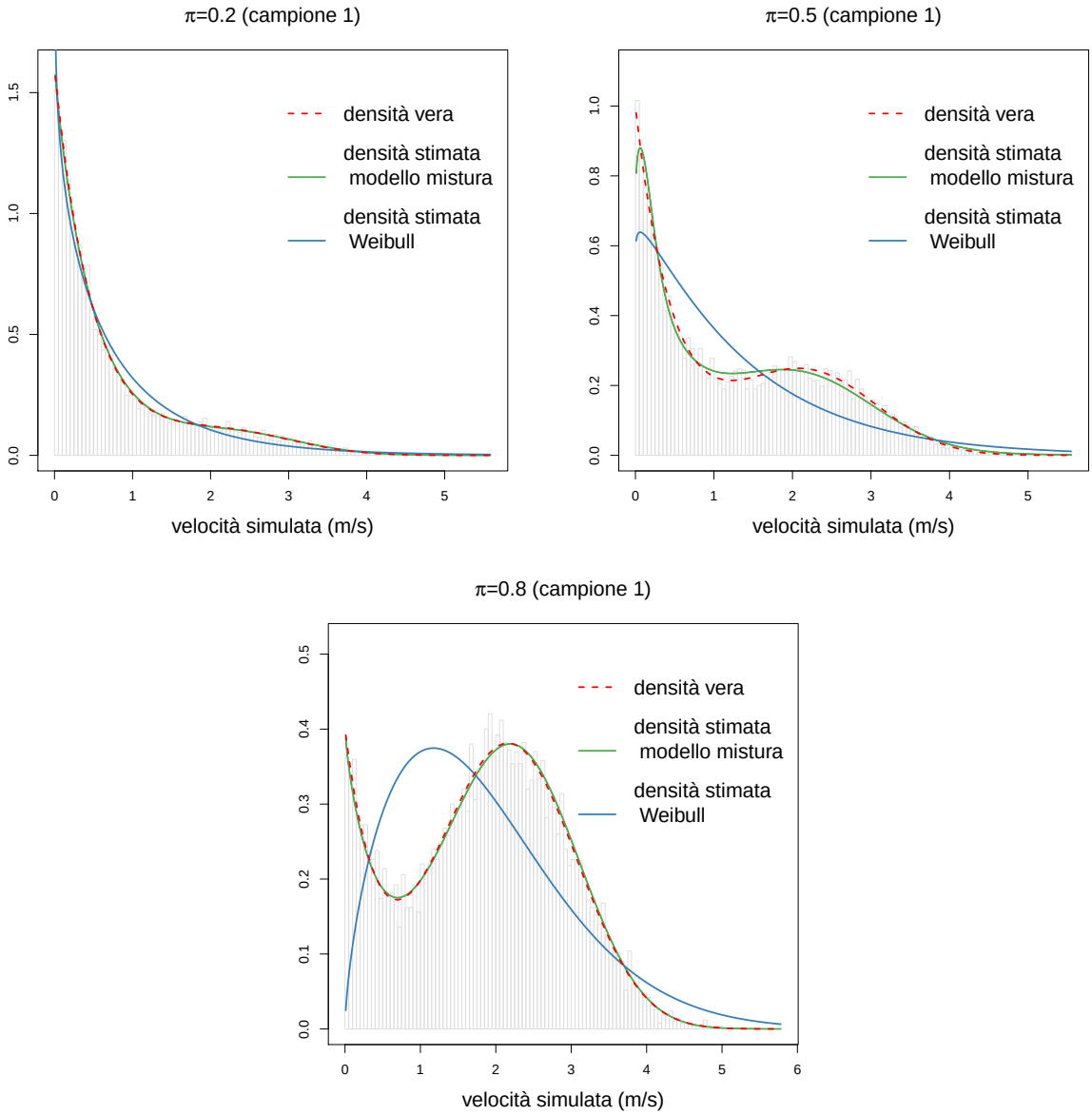

Figura 4.9: Densità sotto corretta ed errata specificazione per diversi scenari quando il vero modello è la mistura senza covariata.

La Figura 4.10 mostra i risultati ottenuti in alcuni scenari quando il processo generatore contempla anche la direzione del vento. Come ci si poteva aspettare, la distribuzione di Weibull fatica ad approssimare la vera densità. Per quanto riguarda invece il modello mistura senza covariata, questo non si discosta molto dalla vera densità quando la bimodalità è meno accentuata. La curva che rappresenta il modello corretto è stata ottenuta mediando le 16 possibili curve derivate ognuna da un valore diverso della direzione, con pesi proporzionali alla frequenza dei valori nel vettore delle direzioni simulate.

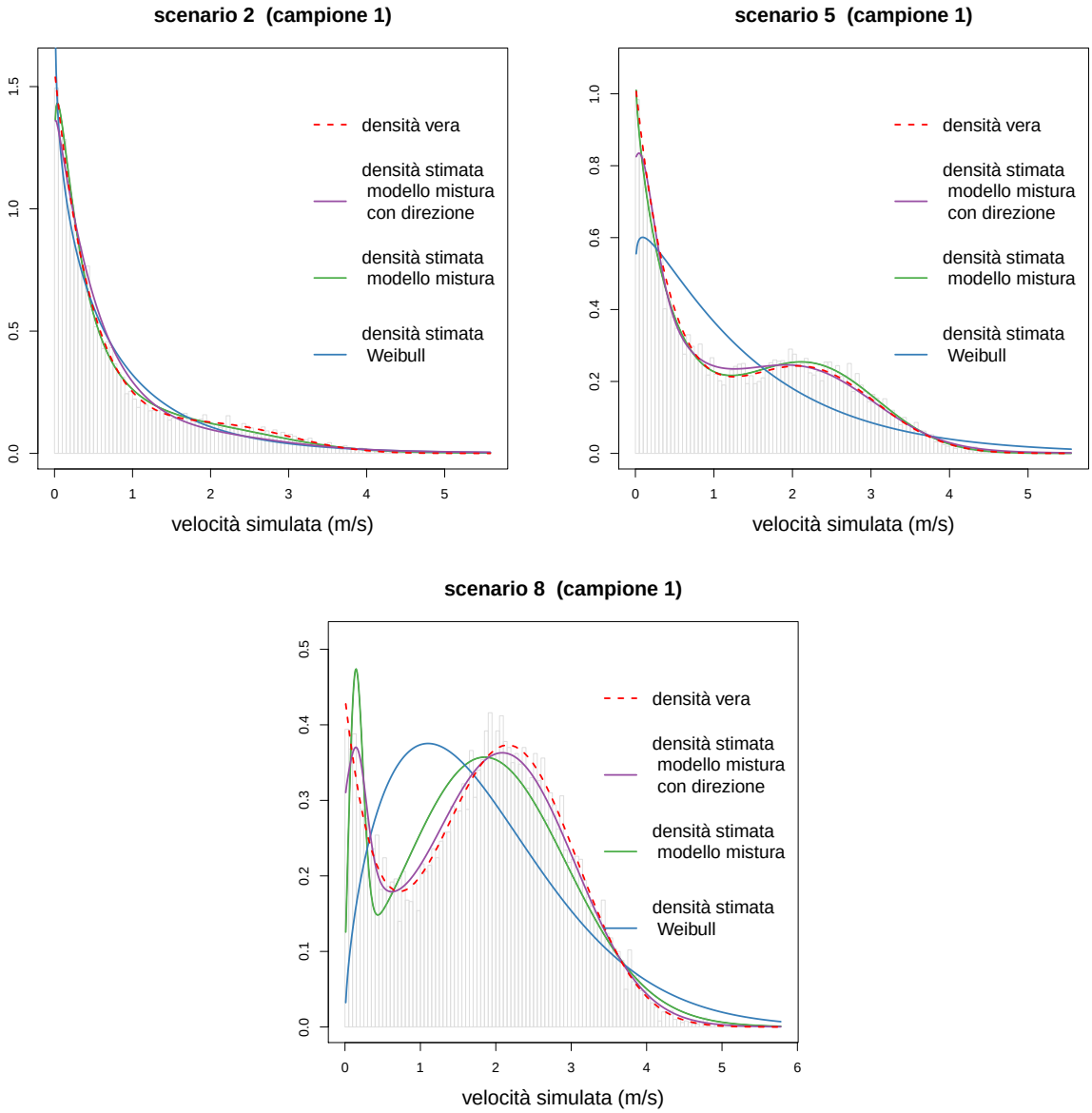

Figura 4.10: Densità sotto corretta ed errata specificazione per diversi scenari quando il vero modello è la mistura con la direzione del vento.

4.2 Applicazione ai dati del modello mistura dipendente

L'intenso studio di simulazione è servito per comprendere se effettivamente un modello mistura fosse una valida alternativa alla distribuzione di Weibull e per valutare quale mistura si adattasse meglio a descrivere una bimodalità. I risultati mostravano che inserire le informazioni riguardo alla direzione, se presenti, portava a risultati migliori.

Si è valutato dunque l'adattamento di un modello mistura con covariata, presentato nell'Equazione 4.2, ai dati relativi alla velocità del vento. In linea con le scelte effettuate nel Capitolo 3, vengono presentati nelle Figure 4.11 e 4.12 i risultati relativi alle quattro stazioni di riferimento per l'anno 2017.

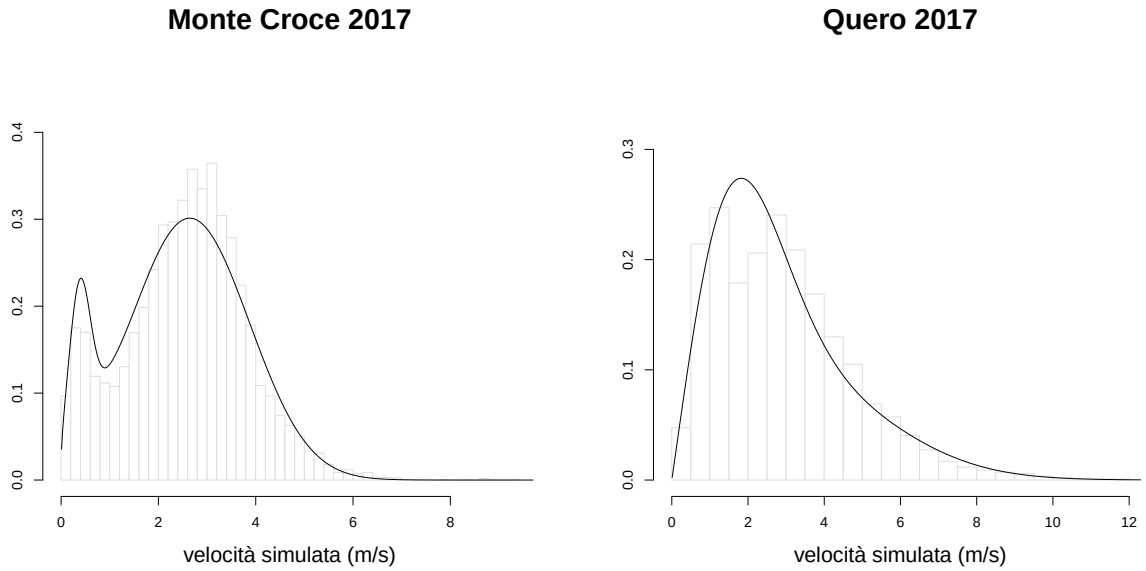

Figura 4.11: Densità stimata del modello mistura con la direzione del vento sui dati relativi alla velocità rilevata in due delle stazioni scelte.

Nei grafici viene riportata una singola curva per rappresentare la densità stimata. In realtà essa è una misura di sintesi delle 16 curve che si ottengono condizionandosi alla direzione. Di conseguenza quella nelle Figure 4.11 e 4.12 è una media pesata delle densità, dove i pesi sono proporzionali alla frequenza delle direzioni nel singolo anno in una stazione. Come si può notare, infatti, l'adattamento non risulta ottimale in tutto il dominio della velocità simulata, ma questo può essere dovuto appunto al modo in cui è stata ottenuta

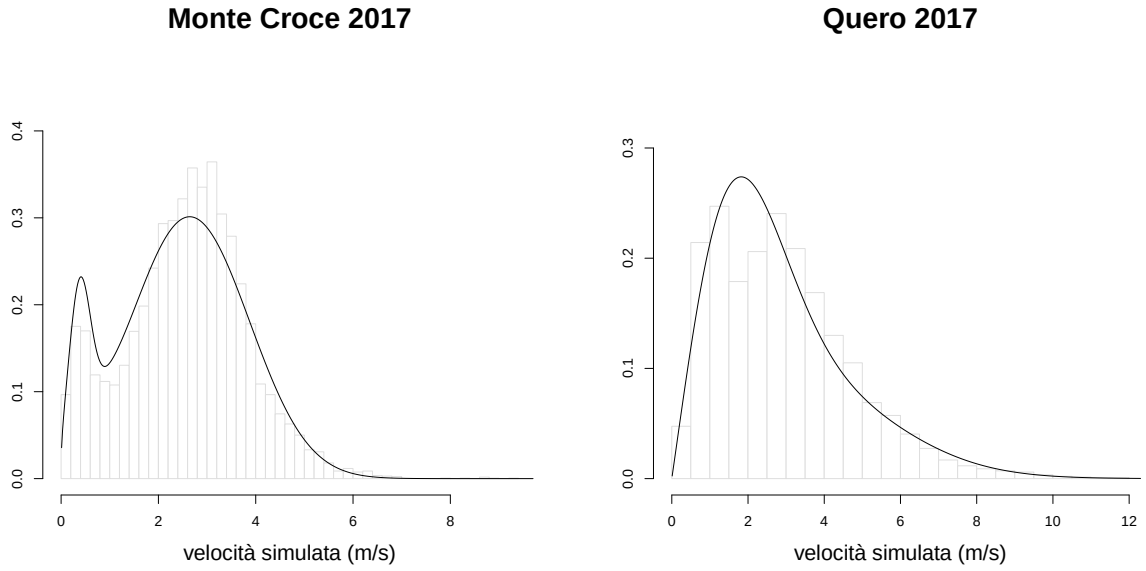

Figura 4.12: Densità stimata del modello mistura con la direzione del vento sui dati relativi alla velocità rilevata in due delle stazioni scelte.

la rappresentazione grafica. Tuttavia l'approssimazione è migliore rispetto a quella fornita dalla distribuzione di Weibull nelle Figure 3.2 e 3.3. Questo rende il modello mistura con l'introduzione della direzione il miglior candidato a entrare nella specificazione del modello bayesiano gerarchico di Zorzetto, Canale e Marani (2019).

Capitolo 5

Modello bayesiano gerarchico

I risultati presentati nel capitolo precedente hanno mostrato che l'informazione portata dalla direzione del vento è molto rilevante al fine di ottenere stime più precise delle quantità di interesse considerate in questo lavoro. Lo sviluppo del successivo modello, dunque, ha come base di partenza la distribuzione mistura dipendente. Il modello bayesiano gerarchico introdotto da Zorzetto, Canale e Marani (2019) non prevede l'inserimento di una variabile esplicativa e questa estensione rappresenta un altro importante contributo metodologico di questa tesi.

Per testare se effettivamente il modello più elaborato può essere adatto allo studio dei valori estremi, si è proceduto con uno studio di simulazione.

In Figura 5.1 viene descritto in dettaglio il modello bayesiano gerarchico introdotto in questo lavoro e utilizzato per analizzare i dati relativi alla velocità del vento. Esso è stato definito a partire da quello sintetizzato in Figura 2.2.

In questo modello x_{ij} rappresenta la velocità del vento mentre z_{ij} è la direzione ad essa associata con $i = 1, \dots, n_j$ numero di valori osservati e diversi da 0, dove $j = 1, \dots, J$ rappresenta il numero di anni di osservazione. Per n_j viene considerata una distribuzione beta-binomiale(N,a, b), ritenuta adeguata a descrivere il numero di eventi quando il totale è prefissato ma le probabilità possono variare e c'è sovradispersione. In questa simulazione N viene considerato fisso e pari a $24 \times 365 = 8760$, ossia il numero massimo di osservazioni orarie che possono essere registrate in un anno. Per i parametri della componente beta, invece, è necessario definire delle distribuzioni a priori. Si è supposto che i dati mancanti

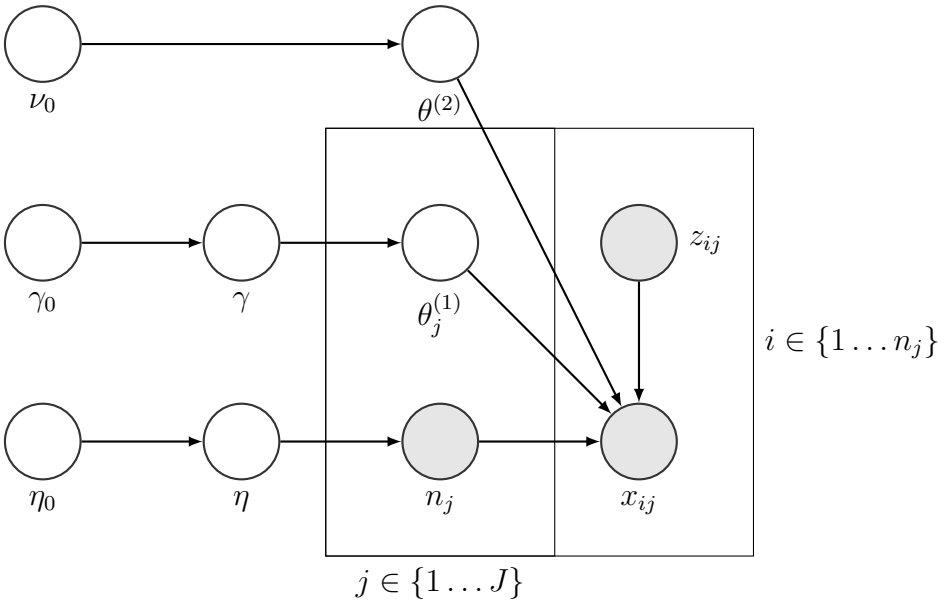

Figura 5.1: Struttura del modello bayesiano gerarchico per i valori estremi.

e le osservazioni orarie registrate come 0, ossia assenza di vento, fossero poche e perciò che le proporzioni di osservazioni effettive rispetto al totale possibile fossero concentrate su valori elevati tra 0.8 e 1. Questa assunzione rispecchia l'andamento riscontrato nelle 4 stazioni scelte (Figure 5.2 e 5.3). La moda di una distribuzione beta è data da $(a-1)/(a+b-2)$ e affinché sia concentrata verso il limite superiore del suo dominio a deve assumere valori più grandi e b piccoli. Di conseguenza è stato ritenuto adeguato utilizzare una lognormale(3,1) per a in quanto ha una varianza elevata e una gamma(2,2) per b poiché risulta più concentrata verso valori piccoli.

Il vettore $\boldsymbol{\theta}^{(2)} = (\alpha, \boldsymbol{\beta})$ contiene i parametri che definiscono la proporzione $\pi(z_{ij}) = e^{\tau(z_{ij})}/(1+e^{\tau(z_{ij})})$ con $\tau(z_{ij}) = \alpha + \beta_1 \sin(z_{ij}) + \beta_2 \cos(z_{ij})$. Si è deciso che i parametri che definiscono la proporzione della mistura sono costanti negli anni, ossia il meccanismo che a seconda della direzione associa alla velocità una componente della mistura è invariante. Di conseguenza si sono definite come a priori per α, β_1 e β_2 delle distribuzioni normali(0.5,2), in modo che la media fosse leggermente spostata rispetto allo 0 e la varianza non concentrasse troppo la distribuzione.

Il vettore $\boldsymbol{\theta}_j^{(1)} = (k_{1j}, k_{2j}, \lambda_{1j}, \lambda_{2j})$ contiene invece i parametri relativi alle due componenti Weibull del modello mistura che descrive x_{ij}. Per modellare la variabilità tra i

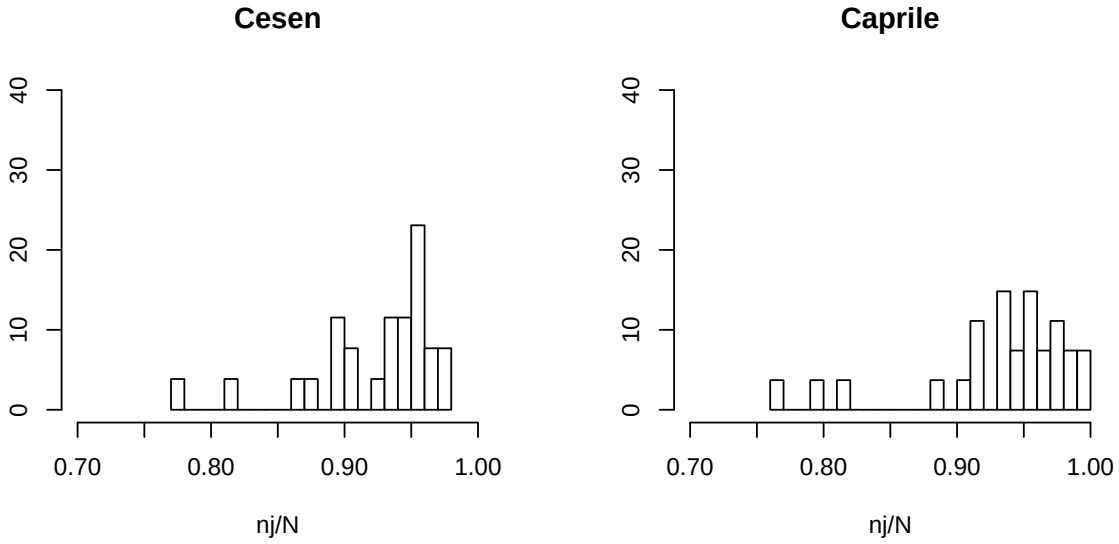

Figura 5.2: Proporzione di valori osservati sul totale di osservazioni annuali possibili in due delle stazioni scelte.

blocchi temporali, ai parametri delle distribuzioni di Weibull sono state associate delle distribuzioni lognormali con parametri incogniti. In una distribuzione lognormale(μ, σ), σ ha come dominio i numeri reali positivi, di conseguenza è stata associata ad una distribuzione lognormale(-2,0.25) che risulta avere la maggiore densità di probabilità per valori inferiori a 0.3. Infatti si è assunto che la variabilità tra i diversi anni non fosse elevata per le componenti di $\boldsymbol{\theta}^{(1)}$. Per quanto riguarda la scelta delle a priori di μ che assume valori sull'intero asse reale, si è pensato a una normale con deviazione standard pari a 0.25, sempre per enfatizzare la bassa variabilità, e media ottenuta in modo tale che il valore atteso di $k_1, \lambda_1, k_2, \lambda_2$ fosse pari ad un valore noto prefissato (1, 0.5, 3, 2.5). Questo perché si volevano conservare le distribuzioni di Weibull in Figura 4.1, ritenute abbastanza diverse tra loro da poter rappresentare bene diversi andamenti unimodali e bimodali.

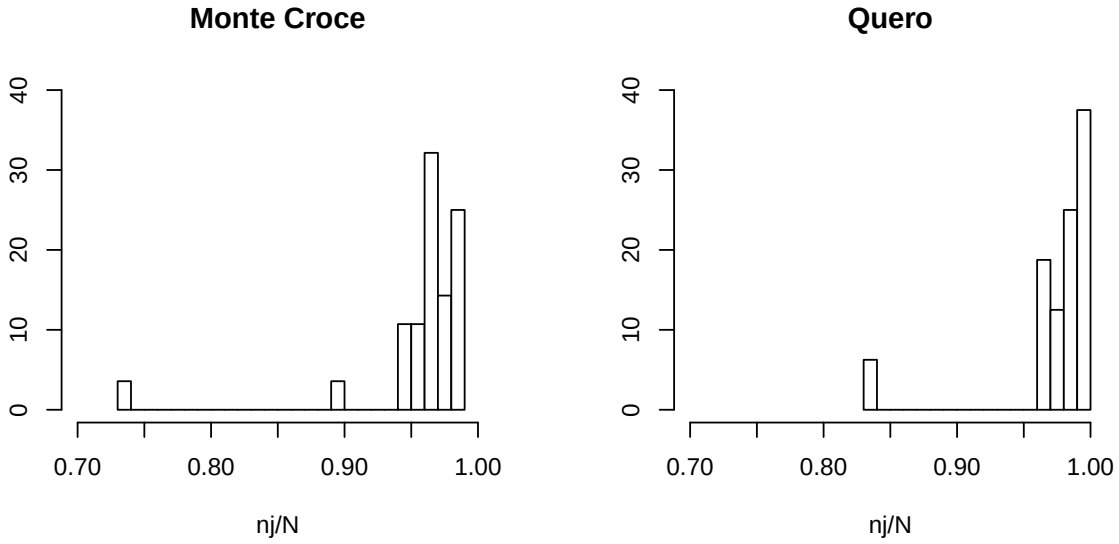

Figura 5.3: Proporzione di valori osservati sul totale di osservazioni annuali possibili in due delle stazioni scelte.

5.1 Studio di simulazione

A partire da questo modello si sono simulati 100 anni di osservazioni, dei quali i primi 20 sono stati utilizzati per stimare i parametri campionando dalla distribuzione a posteriori del modello. Dei rimanenti 80 si sono considerati solo i valori massimi per avere un confronto con le stime ottenute dal modello. Si è impostato l'algoritmo HMC per avere campioni di ampiezza 2000 ripetuti in 4 catene, considerando le prime 1000 iterazioni come *burn in*.

Il problema della non identificabilità delle misture, discusso nella Sezione 4.1.2, ha richiesto di stimare i parametri del modello in modo tale che non ci fosse uno scambio delle componenti Weibull. Si è dunque considerato il parametro di forma $k_{2j} = k_{1j} + \delta_j$ e di conseguenza è stata definita una distribuzione per δ anziché k_2, scelta come una lognormale il cui parametro di posizione segue sempre una distribuzione lognormale in modo tale da avere uno scarto positivo che garantisse $k_{1j} \leq k_{2j}$. Questo permette semplicemente di definire un ordinamento tra le componenti Weibull della mistura ed evita che vengano scambiate tra di loro nel corso delle iterazioni.

Dopo aver stimato i parametri è stata ottenuta una stima della funzione di ripartizione del massimo $\hat{\zeta}(y) = \hat{F}(y|\theta_j^{(1)})^{n_j}$. Per ogni campionamento dalla distribuzione a posteriori, vengono selezionati $\boldsymbol{\theta}^{(2)} = (\alpha, \beta_1, \beta_2)$, $\boldsymbol{\gamma}$ e $\boldsymbol{\eta}$ per generare 100 valori di $\boldsymbol{\theta}^{(1)} = (k_1, k_2, \lambda_1, \lambda_2)$ dalle rispettive lognormali e n dalla beta-binomiale(N, η_1, η_2). Questi sono considerati come i parametri di 100 blocchi temporali annuali futuri, con i quali calcolare la funzione di ripartizione. Riprendendo l'Equazione (2.11), la stima dell'integrale è data da una media della funzione di ripartizione rispetto all'intero campione di $\boldsymbol{\gamma}$ ed $\boldsymbol{\eta}$. Oltre alla media vengono calcolati anche i quantili 0.025 e 0.975, per avere un'indicazione della variabilità.
Nella stima dell'integrale dell'Equazione (2.11) si procede in due passi. Il primo è calcolare la media rispetto ai 100 anni generati dalla distribuzione predittiva. Successivamente si ottengono la media e i quantili rispetto all'intero campione. Ripetendo la procedura per diversi valori di y si ottiene la stima della funzione di ripartizione presentata in Figura 5.4. Dato che nel modello è inclusa la direzione, la funzione di ripartizione varia a seconda di quest'ultima e in Figura 5.4 è riportato un esempio (i grafici relativi alle altre direzioni sono riportati in Appendice). Come si può notare la stima della funzione di ripartizione è molto vicina alla funzione di ripartizione empirica osservata per i massimi dei rimanenti 80 anni, che è inclusa nell'intervallo di credibilità di livello 0.95. Questo dimostra che il modello bayesiano gerarchico introdotto in questo lavoro riesce a cogliere correttamente il comportamento dei massimi, oggetto di interesse primario in questo contesto.

La stessa procedura in più passi è stata seguita per stimare i valori di ritorno associati a specifici tempi di ritorno. Ad ogni iterazione essi vengono calcolati risolvendo

$$\hat{\zeta}(y) - 1 + \frac{1}{t_p} = 0, \tag{5.1}$$

ottenuta dall'Equazione (2.12), per ognuno dei 100 valori di $\boldsymbol{\theta}^{(1)}$ e n generati in ciascuna iterazione. Viene poi ottenuta una media complessiva ad ogni tempo di ritorno considerato. I risultati sono riassunti nel grafico a destra in Figura 5.4. Anche in questo caso, poiché i valori di ritorno derivano da una trasformazione della funzione di ripartizione, i risultati non cambiano di molto a seconda della direzione scelta. I grafici in Figura 5.4 sono relativi alla stessa direzione sud sud-ovest (i grafici relativi alle altre direzioni sono riportati in Appendice).

I risultati possono essere letti come una previsione del comportamento nei prossimi anni, con una variabilità che aumenta all'aumentare dell'orizzonte predittivo. Si potrebbe ad esempio dire che nei prossimi 200 anni ci si aspetta in media un valore massimo pari a 18.7m/s. Ad esso è associato un intervallo di credibilità di livello 0.95 con estremi 13.9 e 26.1m/s. I risultati ottenuti sono coerenti con i valori simulati negli 80 anni non considerati per la stima.

I risultati di quest'ultimo studio di simulazione mostrano che il modello bayesiano gerarchico introdotto da Zorzetto, Canale e Marani (2019) può essere ritenuto valido per la descrizione dei comportamenti estremi di fenomeni ambientali, in questo caso la velocità del vento. Inoltre, rispetto a Zorzetto, Canale e Marani (2019), il modello è stato ampliato considerando per i dati una distribuzione mistura, la cui stima dei parametri tramite algoritmi HMC è stata effettuata con particolari accorgimenti per arginare il problema della non identificabilità. Infine un'altra novità è rappresentata dall'inserimento di una variabile esplicativa nel modello.
Lo studio di simulazione ha permesso di valutare le modifiche apportate alla formulazione originale del modello e, considerando i risultati positivi ottenuti, ha posto una solida base per lo sviluppo di un'analisi sui dati reali, discussa brevemente nella prossima sezione.

5.2 Applicazione ai dati

Il punto di svolta di questo progetto di tesi è rappresentato dall'applicazione ai dati del modello bayesiano gerarchico finalizzato allo studio dei valori estremi. Lo studio di simulazione ha mostrato come una distribuzione mistura con due componenti Weibull definita da una covariata fosse adeguata a descrivere un comportamento bimodale. L'applicazione di questo modello mistura ai dati si è dimostrato coerente con i risultati ottenuti per le simulazioni. Il passo successivo è stato quello di testare il modello bayesiano gerarchico su dati simulati per poi estenderlo ai dati reali.
Lo sviluppo di questa tesi è stato condizionato dai tempi computazionali dell'algoritmo HMC che risultano elevati. Si è quindi riuscito ad ottenere le stime dei parametri del modello bayesiano gerarchico per una sola delle quattro stazioni considerate. L'applicazione, dunque, si riferisce alla stazione del Passo Monte Croce Comelico. Per effettuare la stima

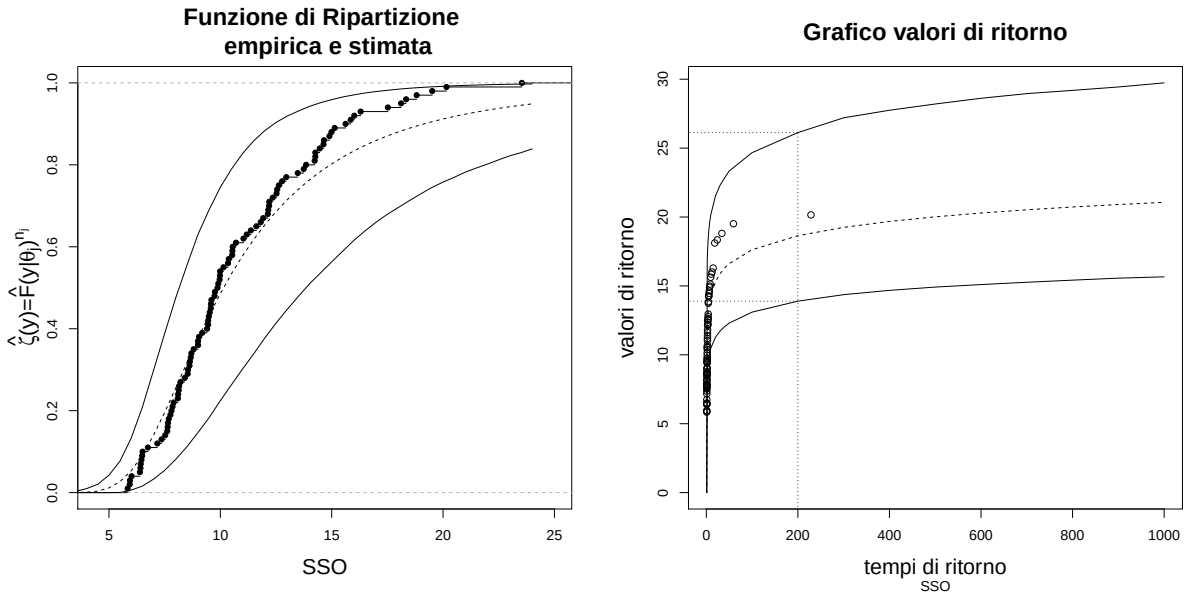

Figura 5.4: Stima della funzione di ripartizione e dei valori di ritorno per la direzione sud sud-ovest. La linea tratteggiata rappresenta la media mentre le linee continue i quantili 0.25 e 0.975. A sinistra viene raffigurata anche la funzione di ripartizione empirica.

si è escluso il 2018, in modo tale da avere un riscontro predittivo. Come a priori sono state utilizzate le distribuzioni definite per lo studio di simulazione, ovvero delle normali per $\boldsymbol{\theta}^{(2)}$; una beta-binomiale per n_j i cui parametri vengono descritti da una lognormale e da una gamma; della lognormali per $\boldsymbol{\theta}_j^{(1)}$ con parametri incogniti con distribuzione normale per quelli di posizione e lognormale per quelli di scala. Dopodiché si sono campionati valori dalla distribuzione a posteriori dei parametri, impostando 2000 iterazioni (di cui 1000 di *burn in*) per 3 catene indipendenti. L'algoritmo, come nel caso delle simulazioni, è stato impostato in modo da arginare il problema della non identificabilità della mistura con l'introduzione di δ_j tale che $k_{2j} = k_{1j} + \delta_j$. I risultati in Figura 5.5 sono i grafici della funzione di ripartizione e dei valori di ritorno calcolati come nello studio di simulazione, considerando la direzione di origine del vento rilevata nel momento in cui è stato osservato il valore massimo del 2018. Nel grafico dei valori di ritorno è stato rappresentato questo valore pari a 13.7m/s. Come c'era da aspettarsi, nel modello adattato, all'evento accaduto nell'ottobre 2018 sono associati tempi di ritorno elevati e ciò sottolinea la rarità dell'accaduto.

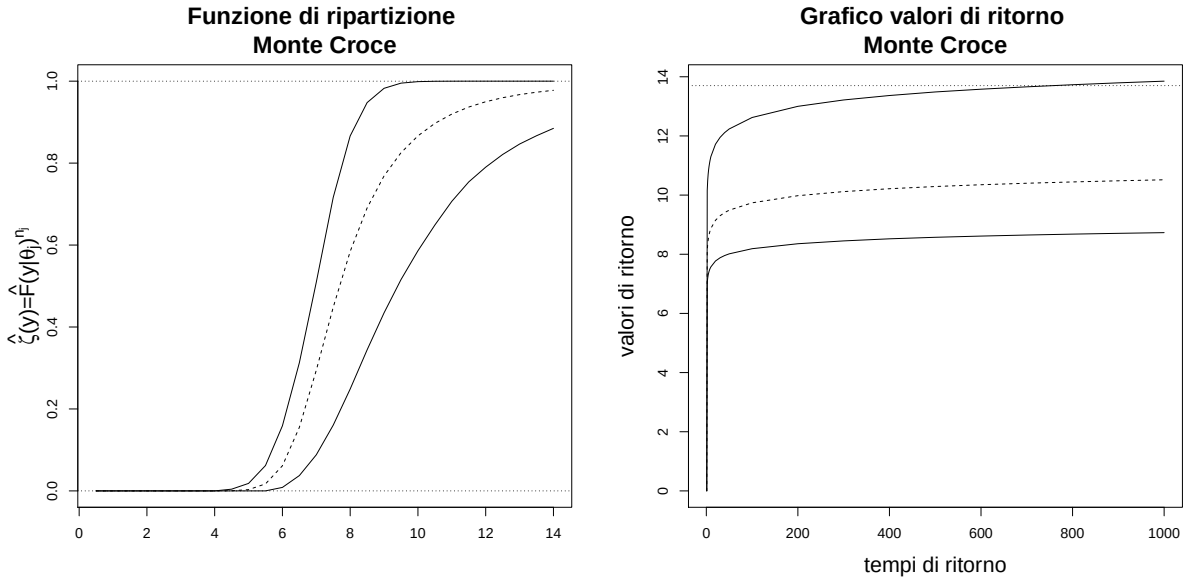

Figura 5.5: Stima della funzione di ripartizione e dei valori di ritorno per la direzione sud-est nella stazione del Passo Monte Croce Comelico. La linea tratteggiata rappresenta la media mentre le linee continue i quantili 0.25 e 0.975. Nel grafico dei valori di ritorno viene rappresentato il massimo del 2018, con linea punteggiata.

Conclusioni

In questo progetto di tesi, motivato dall'analisi della velocità del vento durante gli eventi che hanno caratterizzato la tempesta Vaia, sono stati affrontati diversi temi di carattere modellistico. Innanzitutto si è notato che il comportamento riscontrato nei dati non era quello tipicamente assunto per questo fenomeno, ossia quello di una distribuzione di Weibull (Li e Zhi, 2016). I dati infatti presentavano più o meno marcata bimodalità. Ulteriori studi avevano mostrato come fosse possibile descrivere questo andamento con delle distribuzioni mistura, in particolare con due componenti Weibull (Carta, Ramírez e Velázquez, 2009). Di conseguenza si è affrontato il problema della ricerca di una modello che descrivesse accuratamente il comportamento dei dati.

Poiché lo scopo di questo progetto di tesi era incentrato sull'analisi dei valori estremi, era fondamentale che il modello cogliesse in particolare l'andamento della coda destra. In un primo momento si è sperimentata la stima di una mistura con due componenti Weibull ma successivamente, avendo a disposizione delle informazioni riguardo alla direzione di origine del vento, è stato modificato il modello introducendo una variabile esplicativa per definire la proporzione della mistura. A partire da questi modelli sono stati simulati dei campioni sui quali effettuare una stima dei parametri. Si è scelto un approccio bayesiano e si è utilizzato il software Stan, che permette di condurre l'inferenza secondo il paradigma bayesiano tramite l'algoritmo HMC. I modelli proposti sono stati prima testati sui campioni simulati. Osservando i risultati sotto una corretta ed errata specificazione del modello, si è concluso che sfruttare l'informazione aggiunta della direzione, quando presente, portava a risultati migliori nella stima delle probabilità nella coda destra della distribuzione.
Dopodiché si è voluto ottenere un riscontro diretto sui dati e per quattro stazioni, scelte per la loro diversità, è stato stimato il modello mistura dipendente. L'applicazione ai dati ha fornito risultati positivi.

Ottenuto un modello per descrivere il comportamento bimodale della distribuzione della velocità del vento riscontrato nei dati, si è proceduto con uno studio di simulazione per testare la performance di un modello bayesiano gerarchico che utilizza il modello mistura all'interno della formulazione di Zorzetto, Canale e Marani (2019). Sono stati simulati dei campioni secondo la struttura del modello bayesiano gerarchico e si sono stimati i parametri con i quali calcolare la funzione di ripartizione del massimo e i valori di ritorno. La funzione di ripartizione approssima bene la funzione di ripartizione empirica dei valori massimi e, in ottica predittiva, anche il grafico dei valori di ritorno è coerente con il processo generatore dei dati. Questo ha posto le basi per una applicazione ai dati reali della stazione del Passo Monte Croce Comelico. I risultati confermano l'estrema eccezionalità di Vaia.

Riferimenti bibliografici

Bibliografia

ARPAV, Centro Funzionale Decentrato (nov. 2018). *RELAZIONE EVENTO 27/10/2018 – 01/11/2018*.

Balkema, A. A. e L. De Haan (1974). «Residual Life Time at Great Age». In: *The Annals of Probability* 2.5, pp. 792–804.

Betancourt, Michael (2017). «A Conceptual Introduction to Hamiltonian Monte Carlo». In: *arXiv e-prints*, arXiv:1701.02434, arXiv:1701.02434.

Carta, J.A., P. Ramírez e S. Velázquez (2009). «A review of wind speed probability distributions used in wind energy analysis: Case studies in the Canary Islands». In: *Renewable and Sustainable Energy Reviews* 13.5, pp. 933 –955.

Coles, Stuart (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer-Verlag.

Davison, A. C. e R. L. Smith (1990). «Models for Exceedances Over High Thresholds». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 52.3, pp. 393–425.

Fisher, R. A. e L. H. C. Tippett (1928). «Limiting forms of the frequency distribution of the largest or smallest member of a sample». In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.2, 180–190.

Gnedenko, B. (1943). «Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire». In: *Annals of Mathematics* 44.3, pp. 423–453.

Gumbel, E.J. (1958). *Statistics of Extremes*. Columbia University Press, New York.

Hastings, W. K. (apr. 1970). «Monte Carlo sampling methods using Markov chains and their applications». In: *Biometrika* 57.1, pp. 97–109.

Kollu, Ravindra, Srinivasa Rao Rayapudi, SVL Narasimham e Krishna Mohan Pakkurthi (2012). «Mixture probability distribution functions to model wind speed distributions». In: *International Journal of Energy and Environmental Engineering* 3.1, p. 27.

Li, Guojie e Jing Zhi (2016). «Chapter 2 - Analysis of Wind Power Characteristics». In: *Large-Scale Wind Power Grid Integration*. A cura di Ningbo Wang, Chongqing Kang e Dongming Ren. Oxford: Academic Press, pp. 19 –51.

Metropolis, Nicholas, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller e Edward Teller (1953). «Equation of State Calculations by Fast Computing Machines». In: *The Journal of Chemical Physics* 21.6, pp. 1087–1092.

Neal, Radford M. (2010). «MCMC Using Hamiltonian Dynamics». In: *Handbook of Markov Chain Monte Carlo* 54, pp. 113–162.

Ouarda, Taha B.M.J. e Christian Charron (2018). «On the mixture of wind speed distribution in a Nordic region». In: *Energy Conversion and Management* 174, pp. 33 –44.

Pickands III, James (1975). «Statistical Inference Using Extreme Order Statistics». In: *The Annals of Statistics* 3.1, pp. 119–131.

Zorzetto, E., G. Botter e M. Marani (2016). «On the emergence of rainfall extremes from ordinary events». In: *Geophysical Research Letters* 43.15, pp. 8076–8082.

Zorzetto, E., A. Canale e M. Marani (2019). *A Hierarchical Bayesian Model for Extreme Value Analysis*. Working paper.

Sitografia

ARPAV (2019). *Descrizione delle stazioni*. last accessed 23 luglio 2019. URL: http://www.arpa.veneto.it/bollettini/storico/Mappa_2019_VVENTO.htm?t=RG.

Cason, Diego (2018). *Conseguenze della tempesta Vaia*. last accessed 5 settembre 2019. URL: www.michelenardelli.it/commenti.php?id=4258.

Dariol, Carlo (2008). *Alluvione 1966*. last accessed 25 settembre 2019. URL: www.elevamentealcubo.it/Crocedipiave/1966_alluvione.htm.

Treviso-Belluno, Camera di Commercio (2019a). last accessed 17 settembre 2019. URL: www.trevisobellunosystem.com/tvsys/home/archivio-news/13599_pozza-ve-lo-avevamo-promesso-che-non-vi-avremo-lasciati-soli.html.

— (2019b). last accessed 17 settembre 2019. URL: www.trevisobellunosystem.com/tvsys/home/archivio-news/13855_fondi-europei-per-maltempo-2018-via-libera-della-commissione-ue-al-18-settembre.html.

Appendice

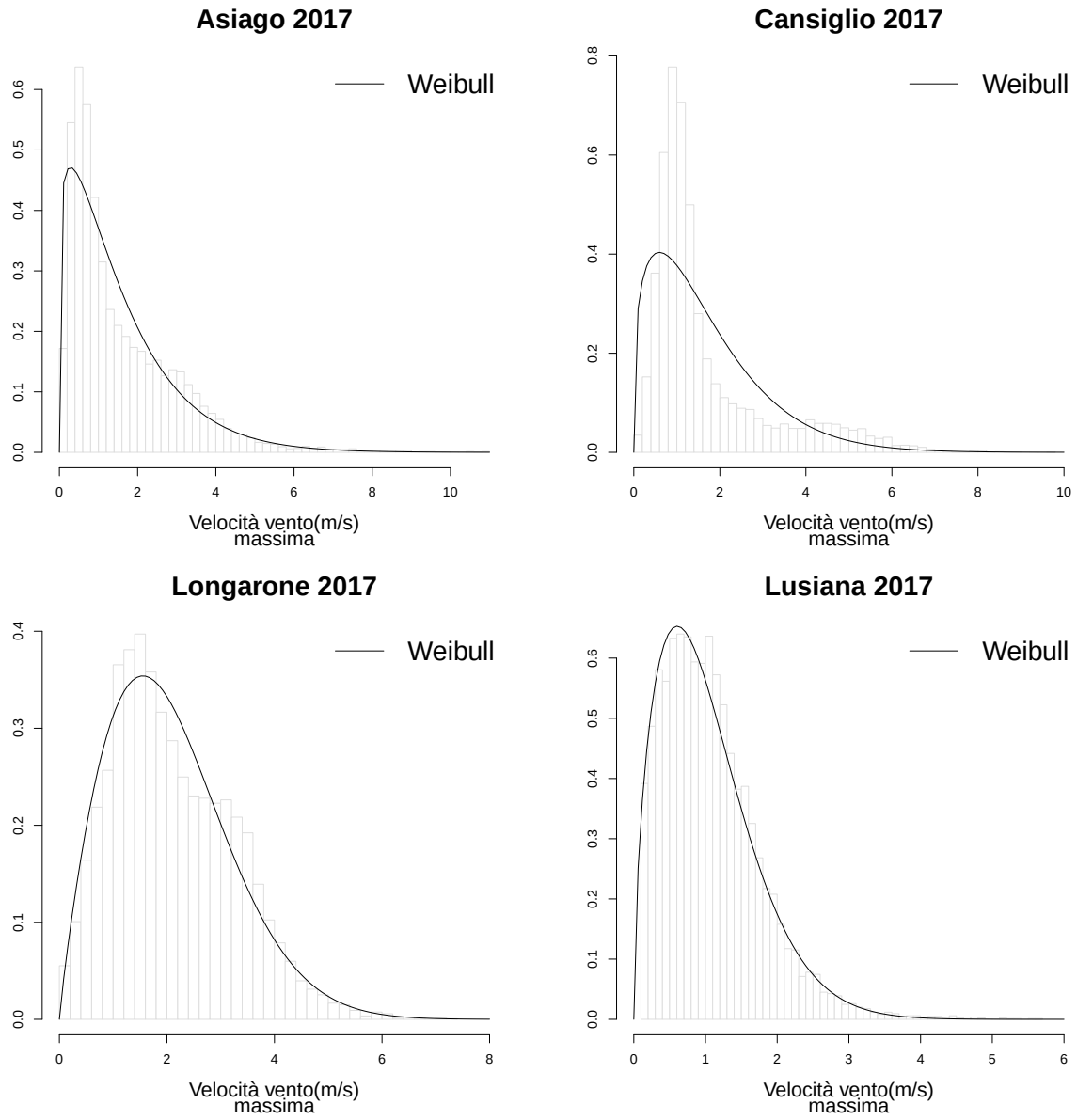

Figura 6: Esempi di comportamento della velocità del vento in diverse stazioni nell'anno precedente alla tempesta Vaia. La linea continua rappresenta l'adattamento di una distribuzione di Weibull.

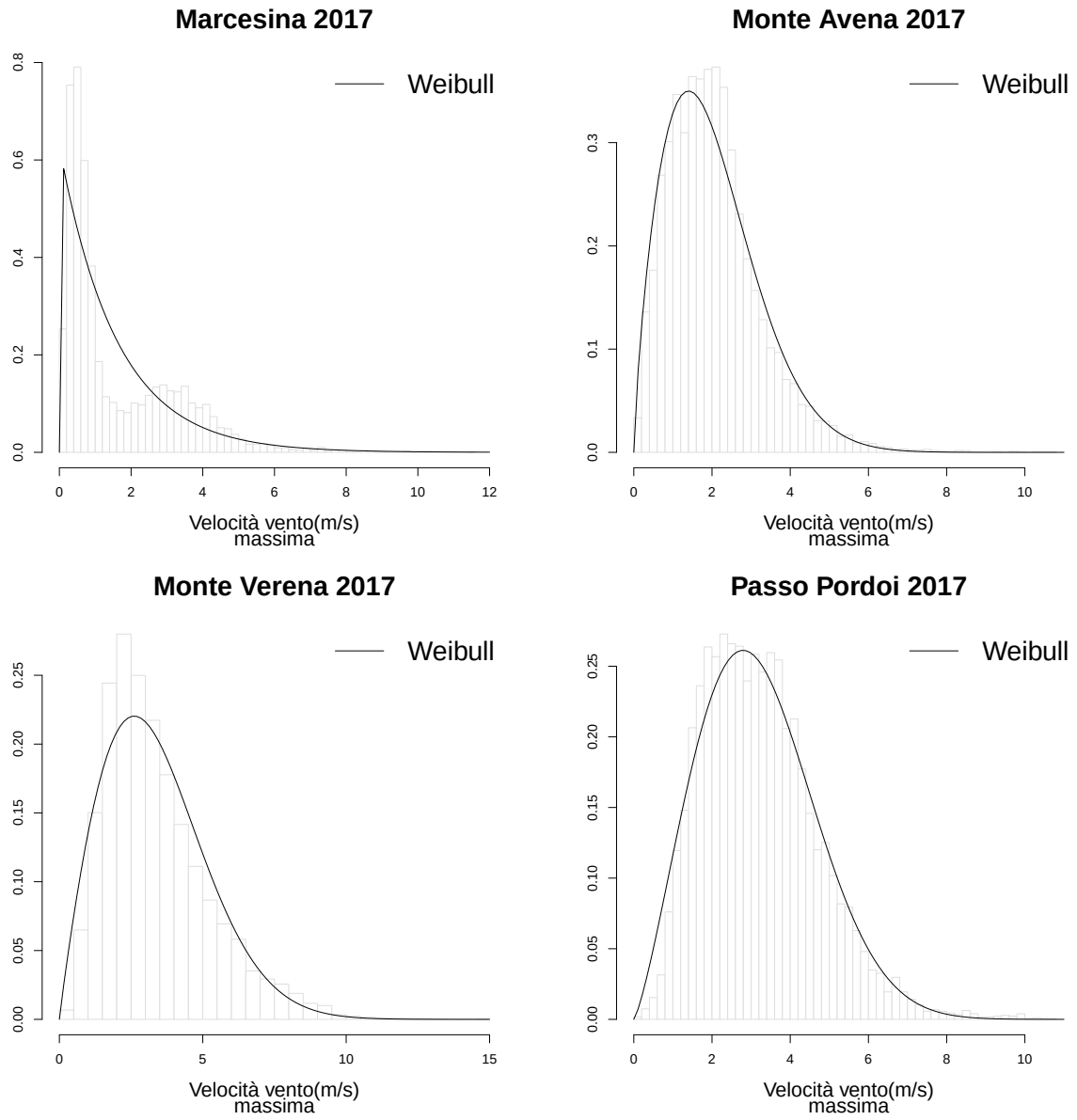

Figura 7: Esempi di comportamento della velocità del vento in diverse stazioni nell'anno precedente alla tempesta Vaia. La linea continua rappresenta l'adattamento di una distribuzione di Weibull.

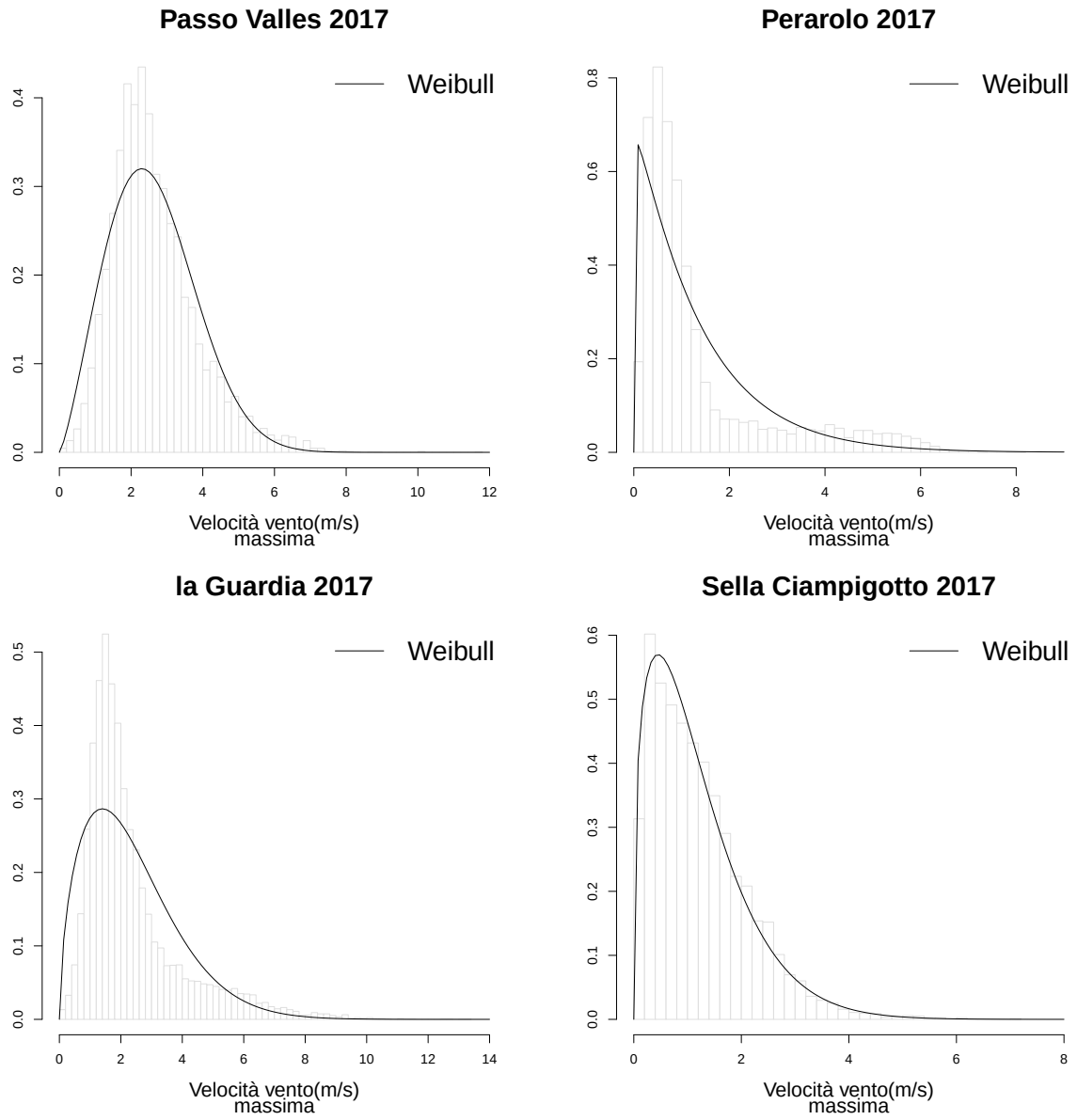

Figura 8: Esempi di comportamento della velocità del vento in diverse stazioni nell'anno precedente alla tempesta Vaia. La linea continua rappresenta l'adattamento di una distribuzione di Weibull.

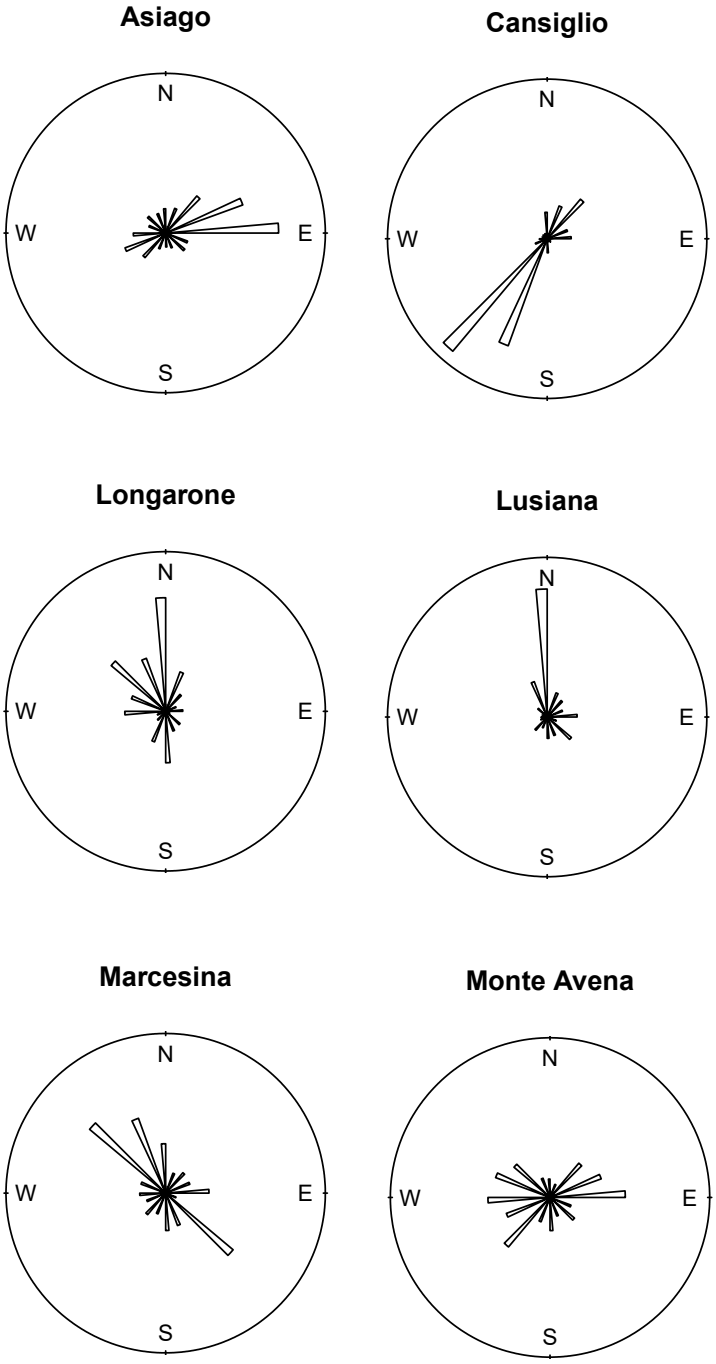

Figura 9: Esempi di distribuzione della direzione del vento.

Figura 10: Esempi di distribuzione della direzione del vento.

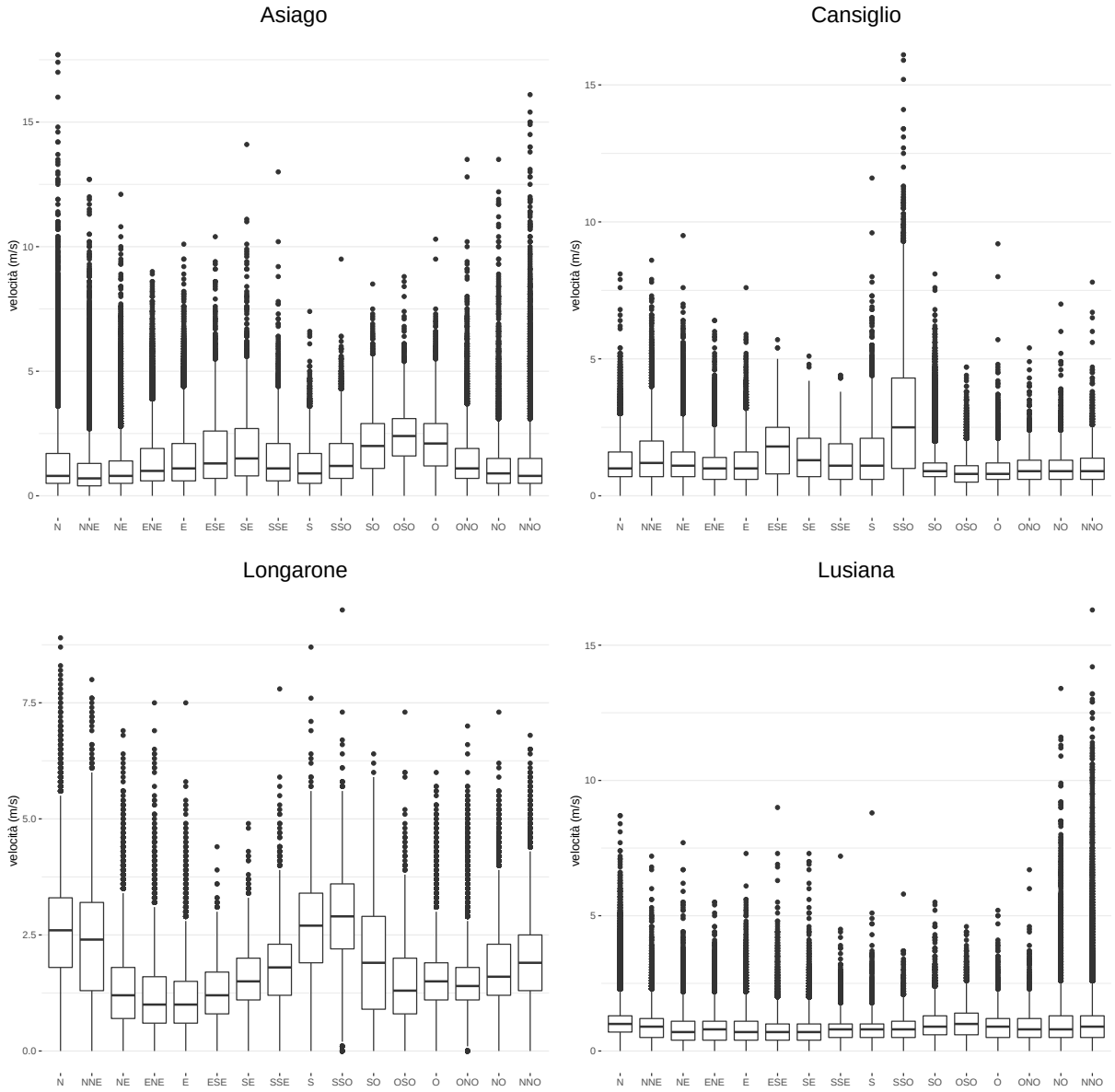

Figura 11: Esempi di come la distribuzione della velocità del vento varia in base alla direzione in alcune stazioni.

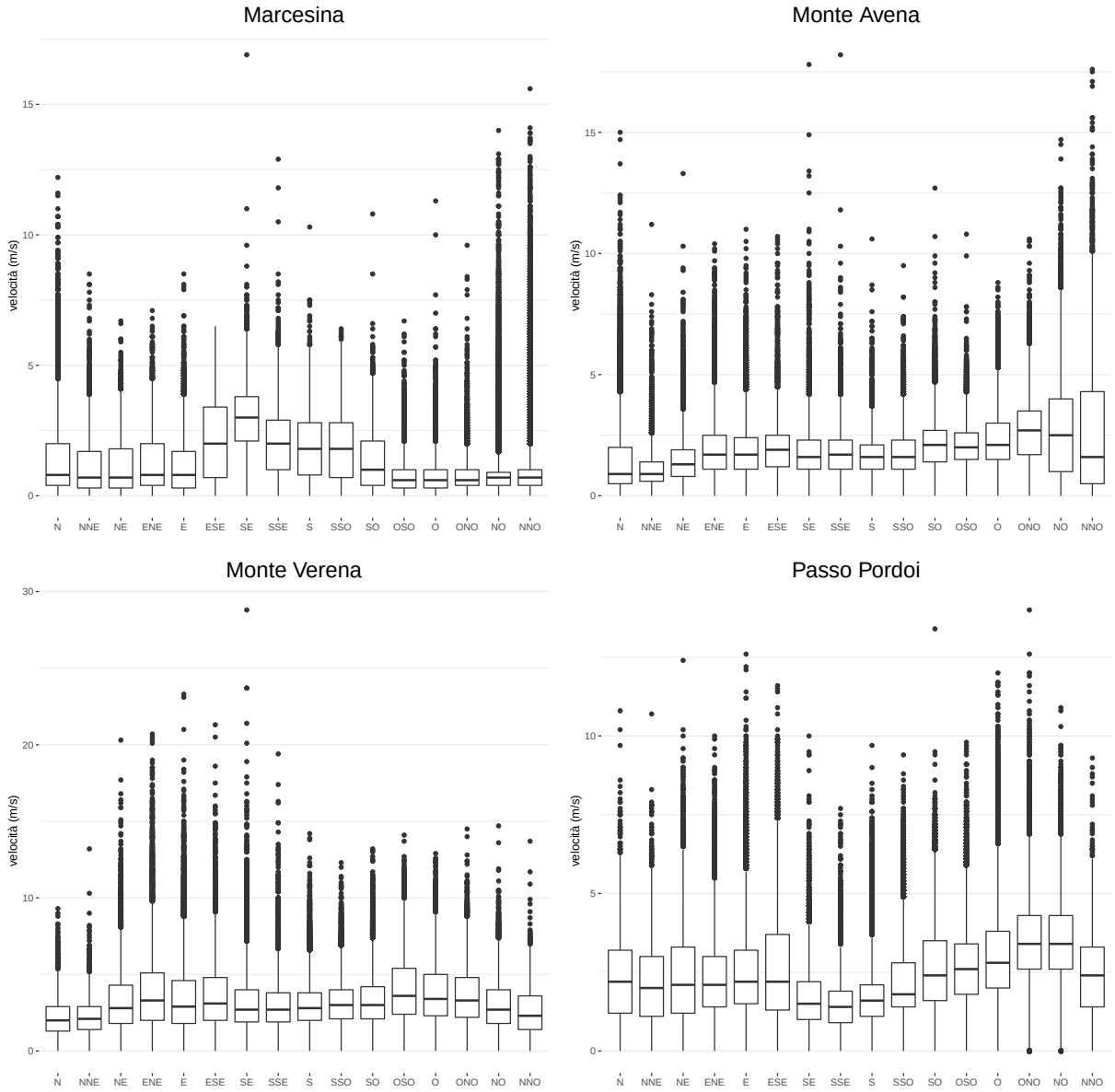

Figura 12: Esempi di come la distribuzione della velocità del vento varia in base alla direzione in alcune stazioni.

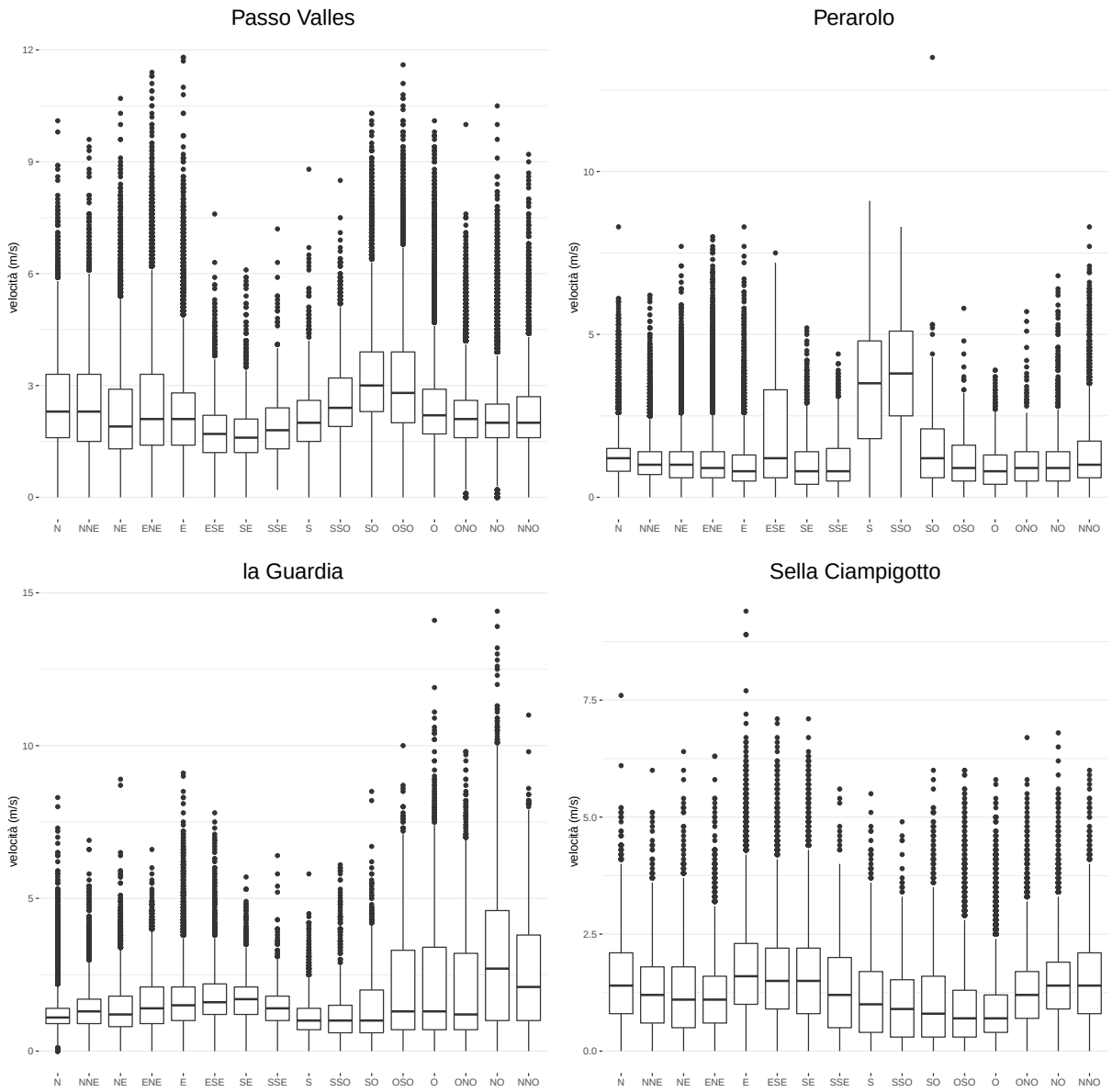

Figura 13: Esempi di come la distribuzione della velocità del vento varia in base alla direzione in alcune stazioni.

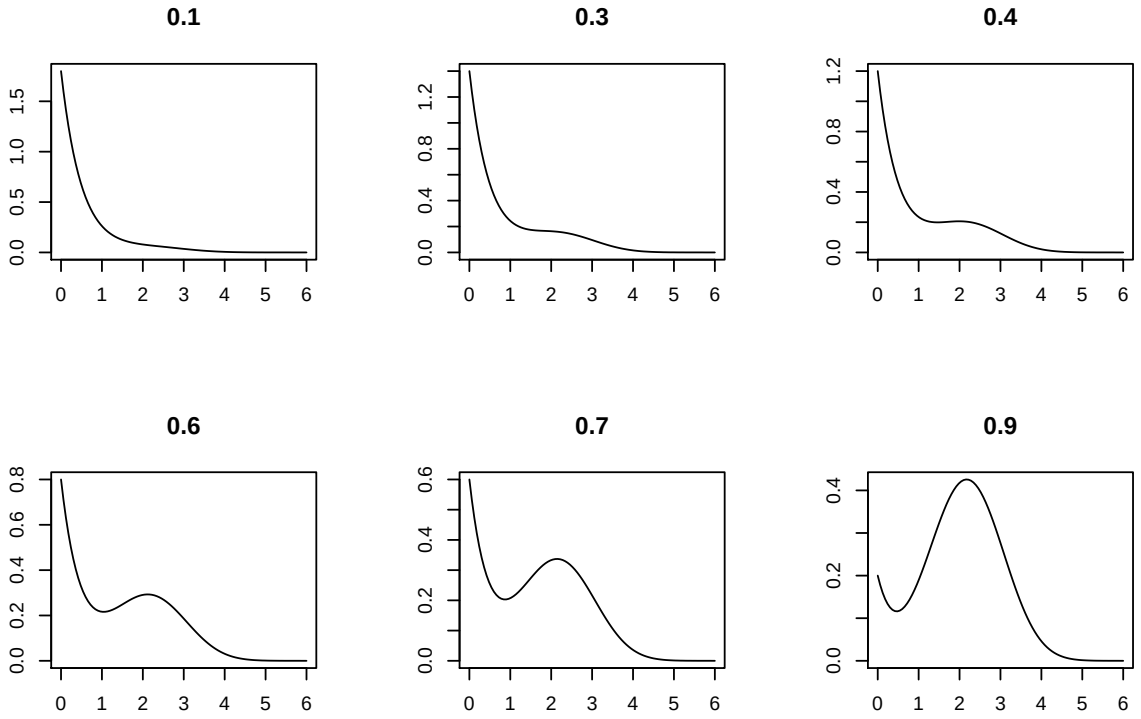

Figura 14: Esempi di comportamento di una mistura con due componenti Weibull al variare di π.

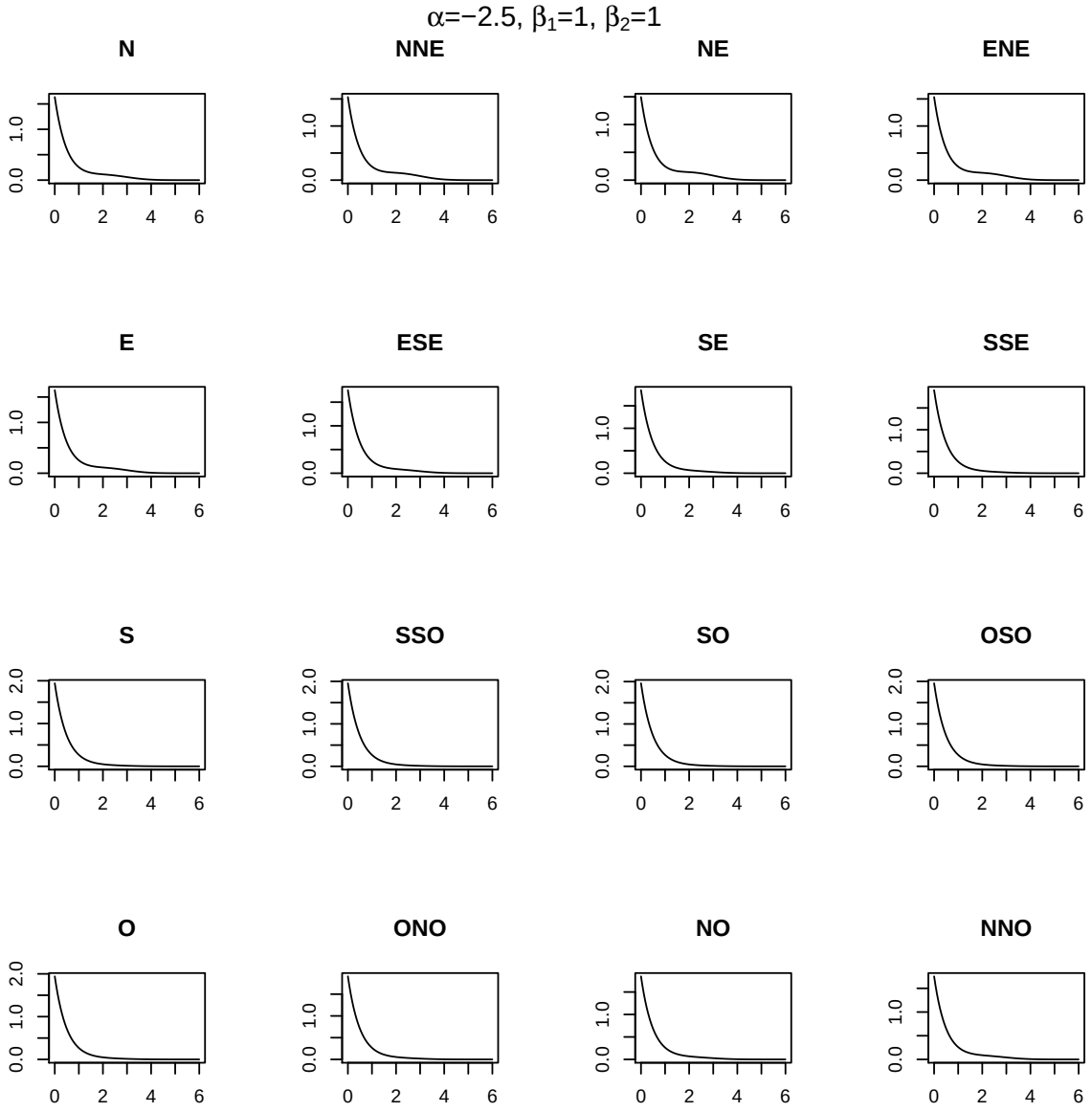

Figura 15: Esempi di comportamento di una mistura con due componenti Weibull definita da una covariata. Sono considerati specifici valori di α e $\boldsymbol{\beta}$ al variare della direzione.

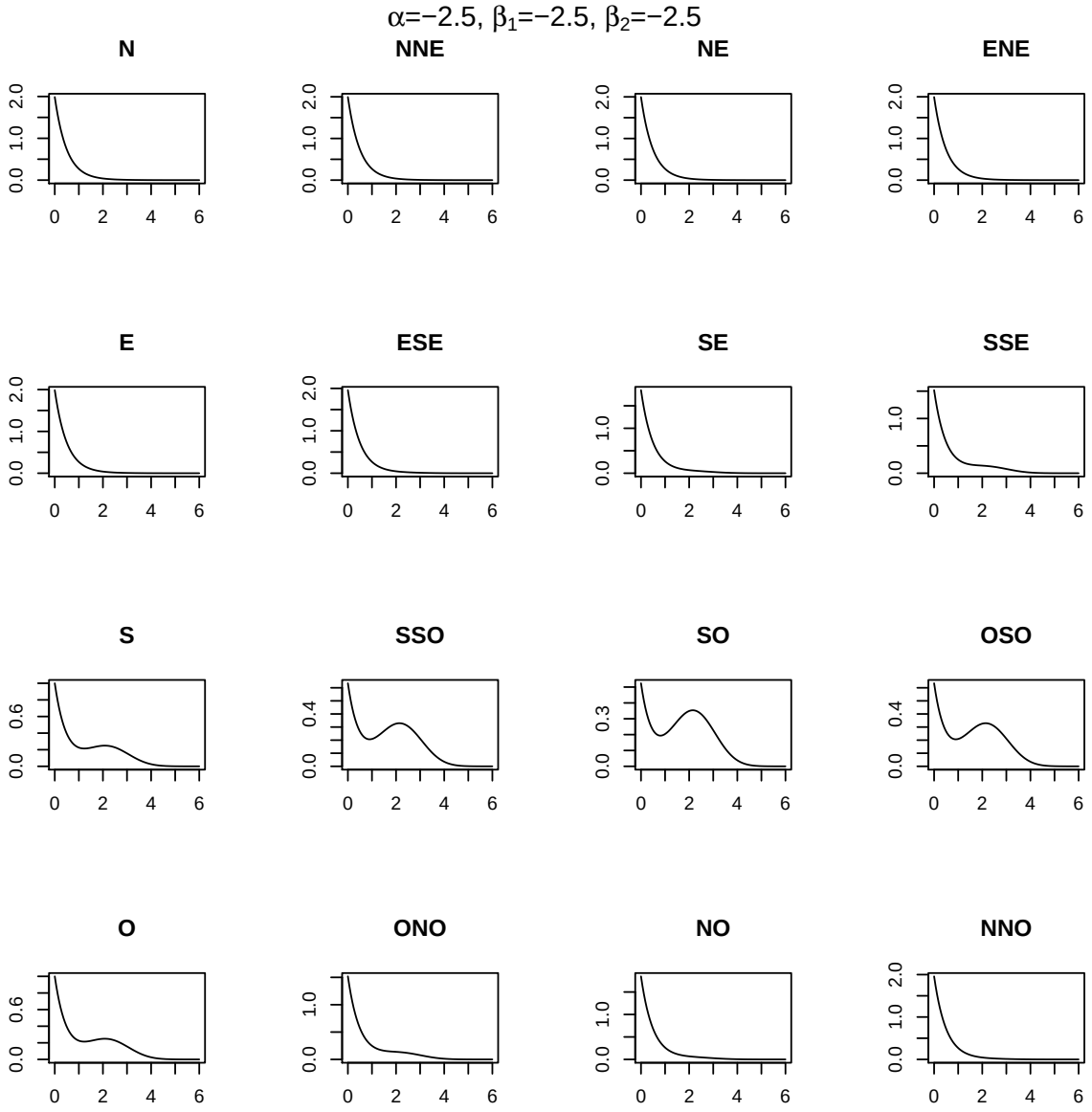

Figura 16: Esempi di comportamento di una mistura con due componenti Weibull definita da una covariata. Sono considerati specifici valori di α e $\boldsymbol{\beta}$ al variare della direzione.

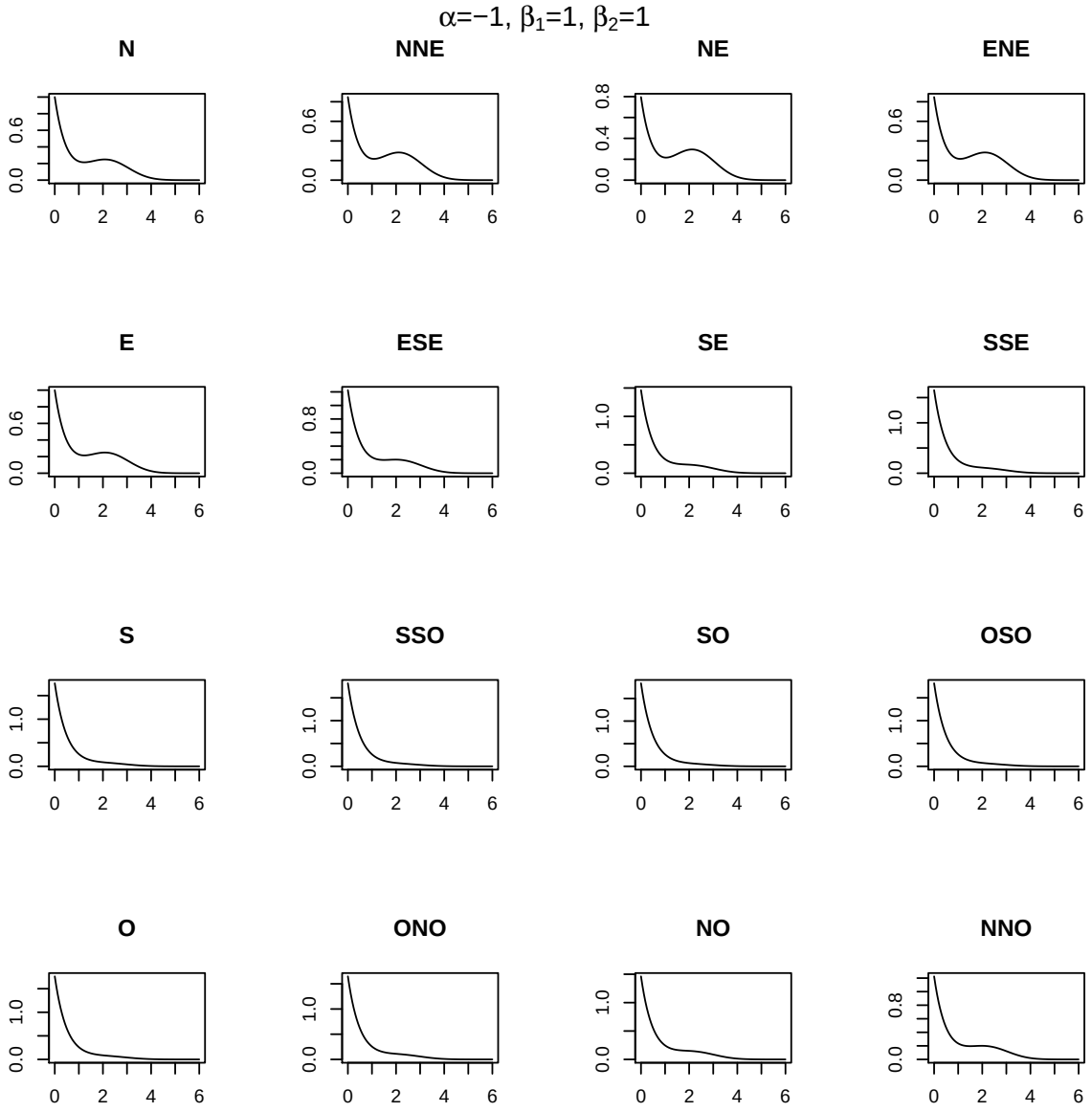

Figura 17: Esempi di comportamento di una mistura con due componenti Weibull definita da una covariata. Sono considerati specifici valori di α e $\boldsymbol{\beta}$ al variare della direzione.

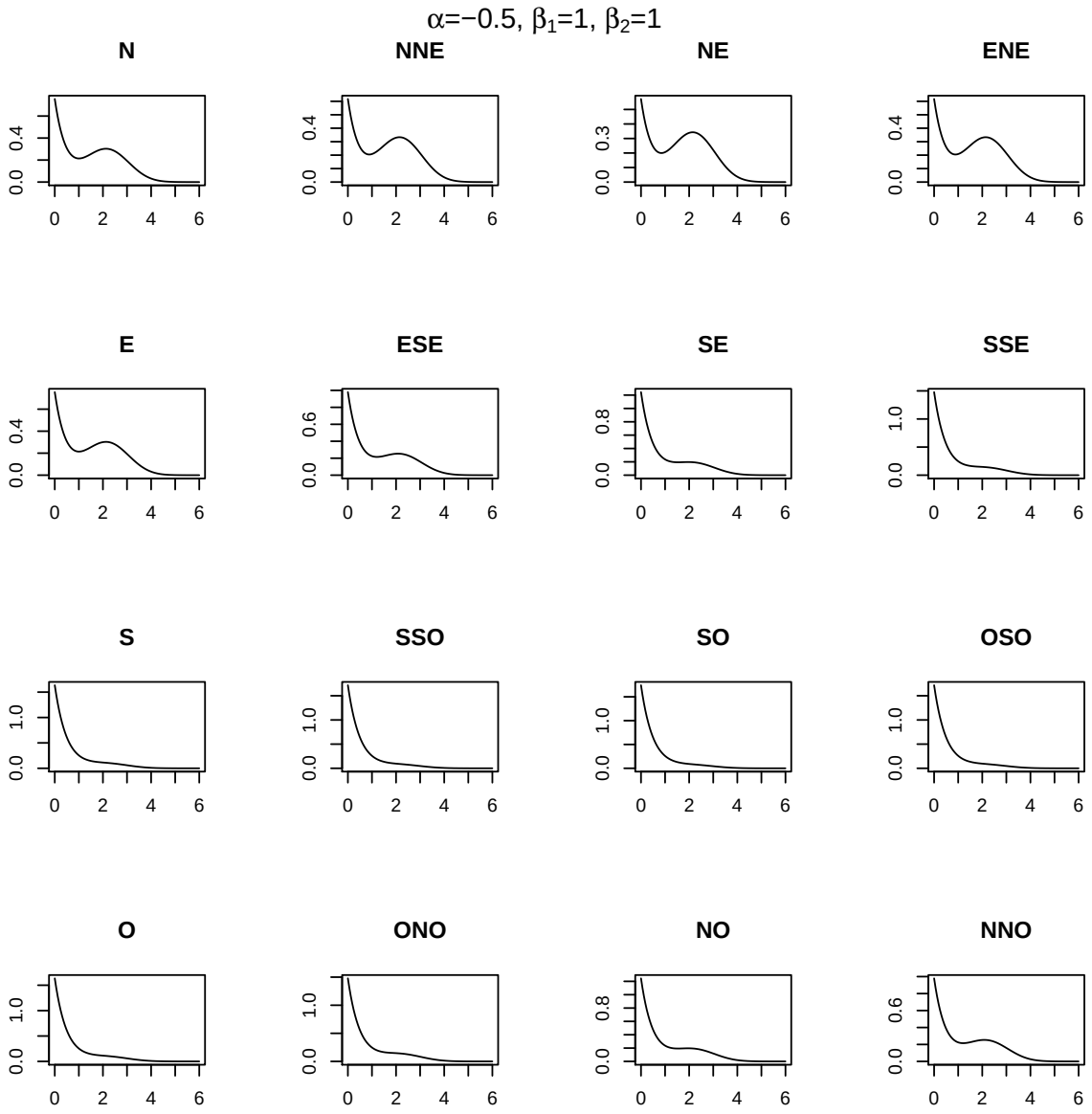

Figura 18: Esempi di comportamento di una mistura con due componenti Weibull definita da una covariata. Sono considerati specifici valori di α e $\boldsymbol{\beta}$ al variare della direzione.

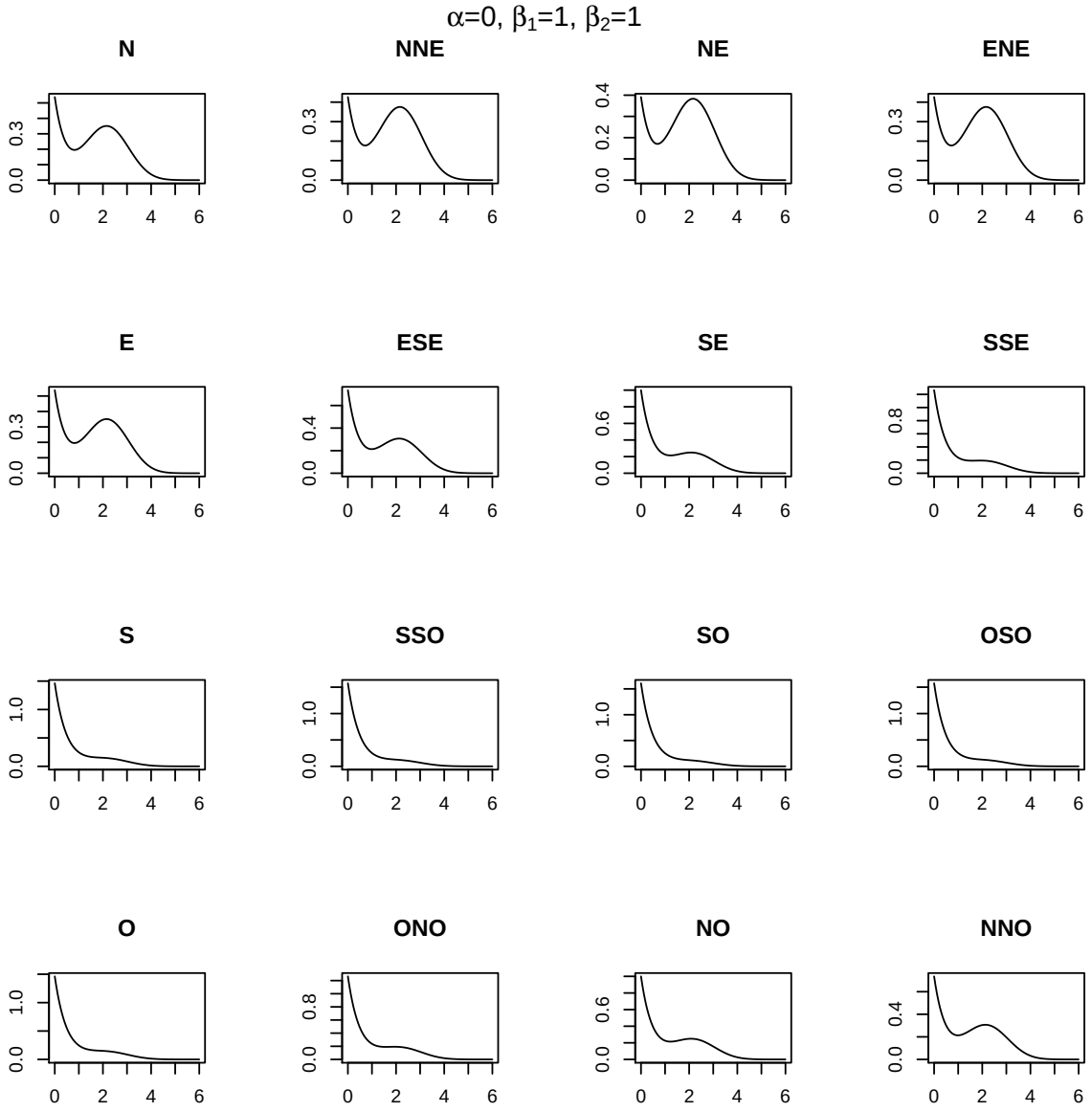

Figura 19: Esempi di comportamento di una mistura con due componenti Weibull definita da una covariata. Sono considerati specifici valori di α e $\boldsymbol{\beta}$ al variare della direzione.

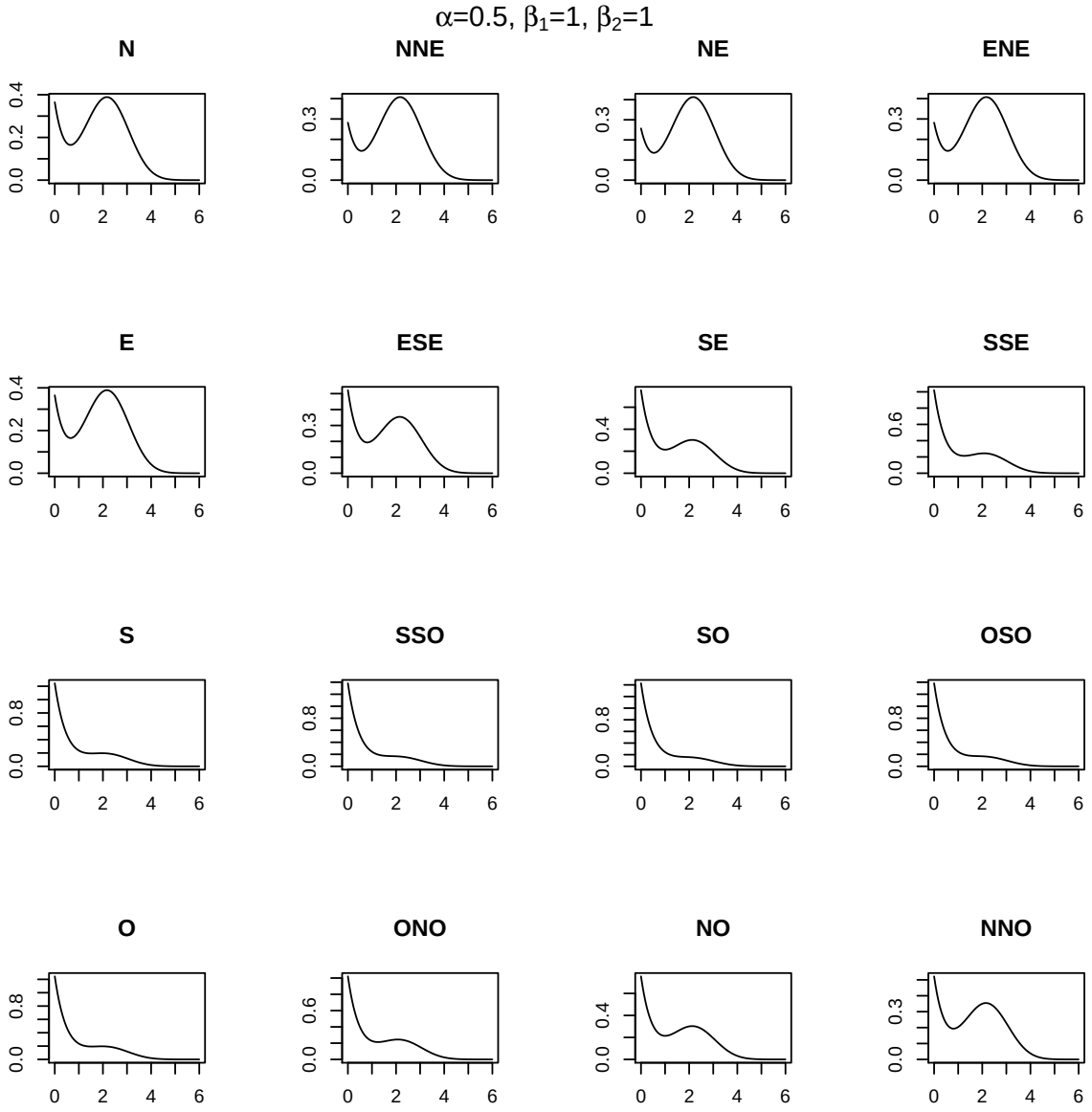

Figura 20: Esempi di comportamento di una mistura con due componenti Weibull definita da una covariata. Sono considerati specifici valori di α e $\boldsymbol{\beta}$ al variare della direzione.

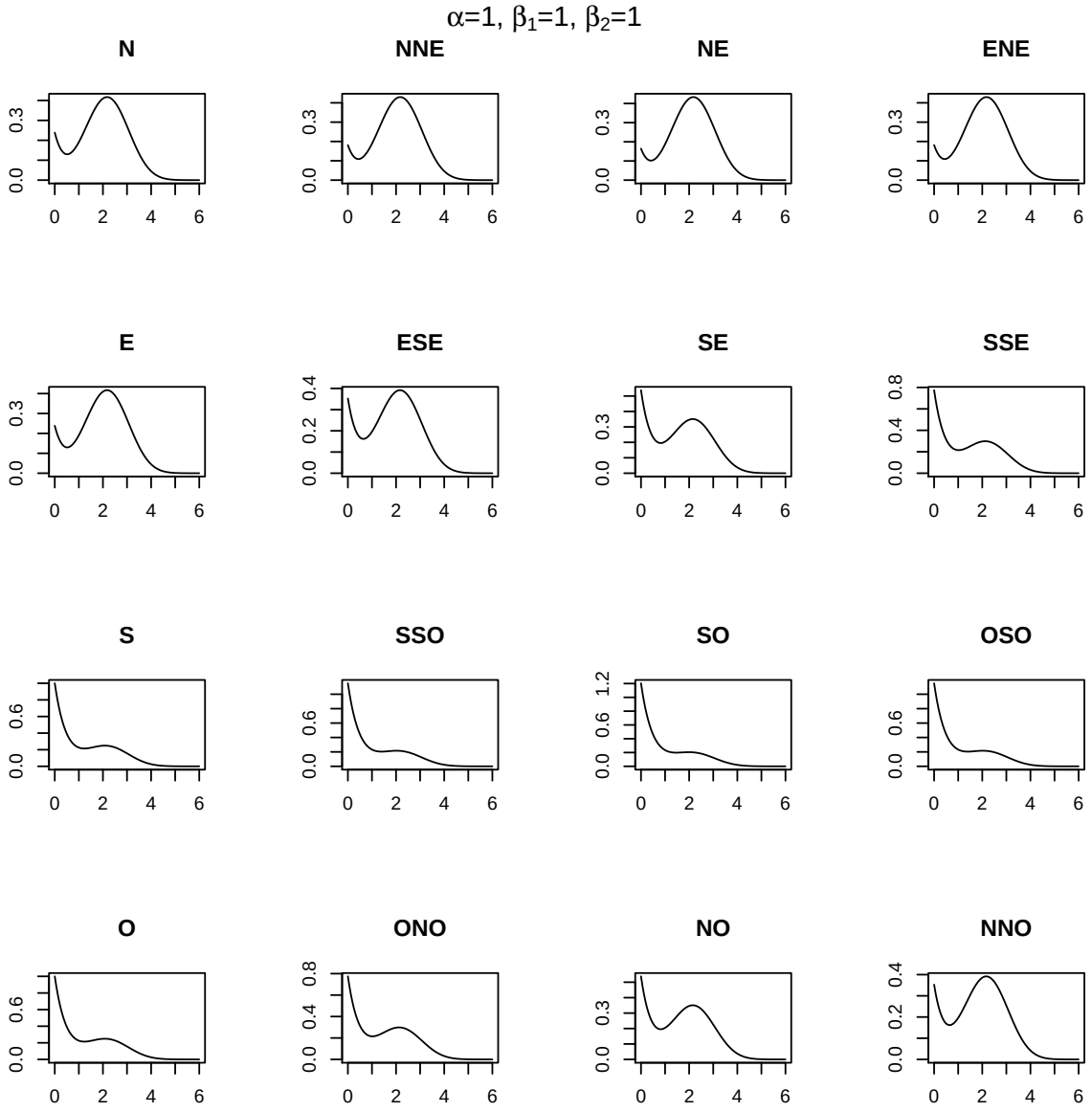

Figura 21: Esempi di comportamento di una mistura con due componenti Weibull definita da una covariata. Sono considerati specifici valori di α e $\boldsymbol{\beta}$ al variare della direzione.

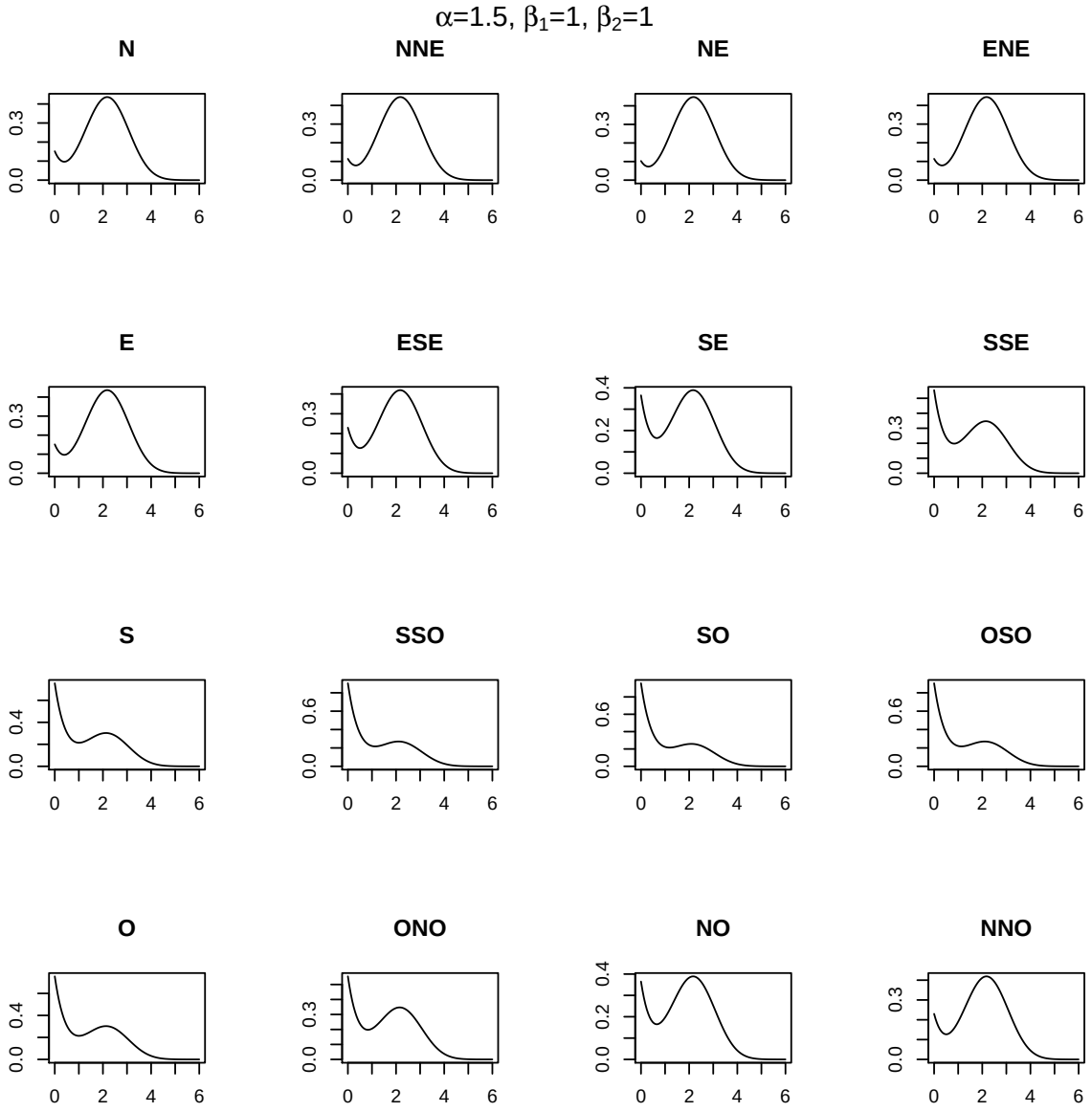

Figura 22: Esempi di comportamento di una mistura con due componenti Weibull definita da una covariata. Sono considerati specifici valori di α e $\boldsymbol{\beta}$ al variare della direzione.

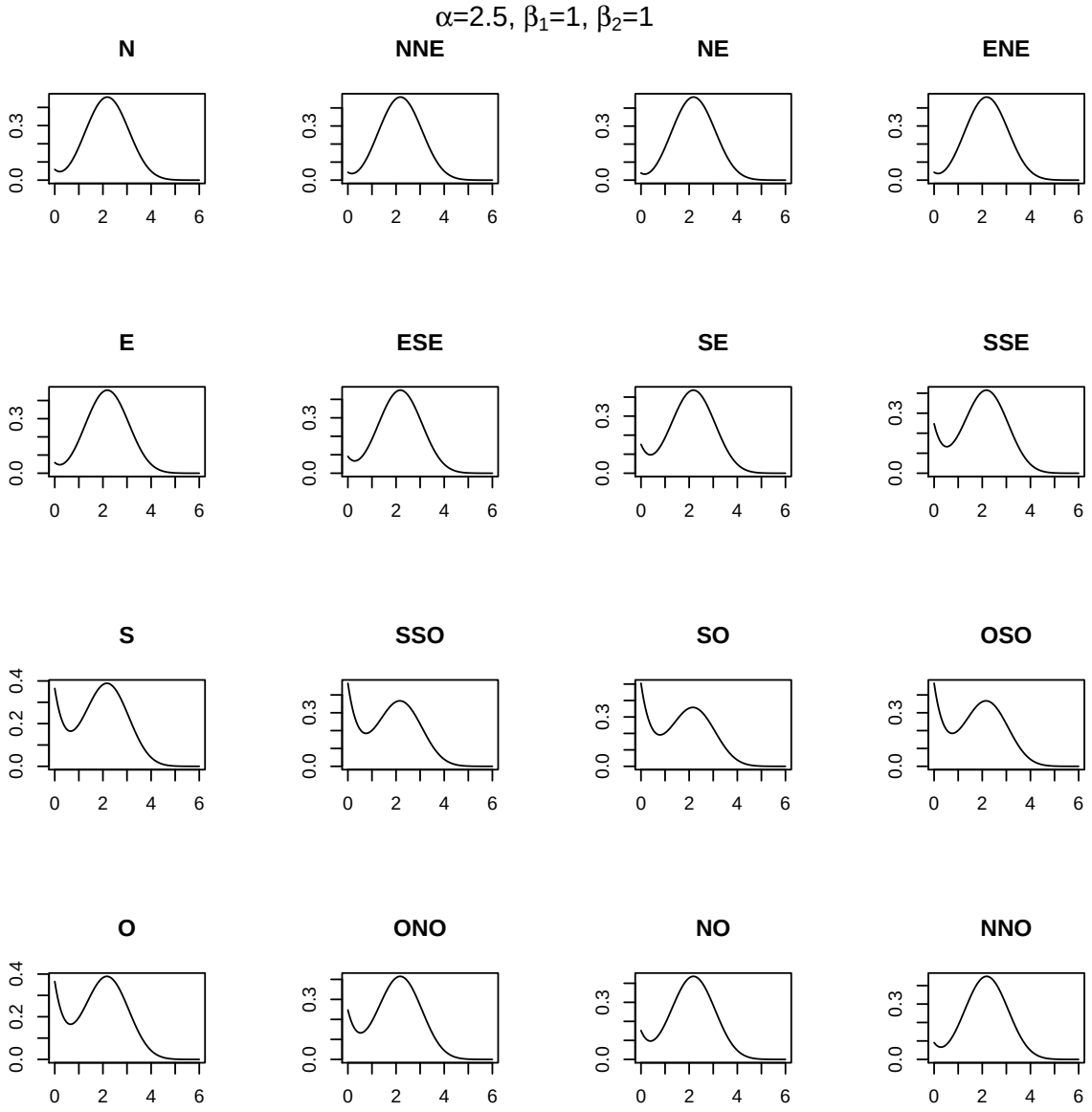

Figura 23: Esempi di comportamento di una mistura con due componenti Weibull definita da una covariata. Sono considerati specifici valori di α e $\boldsymbol{\beta}$ al variare della direzione.

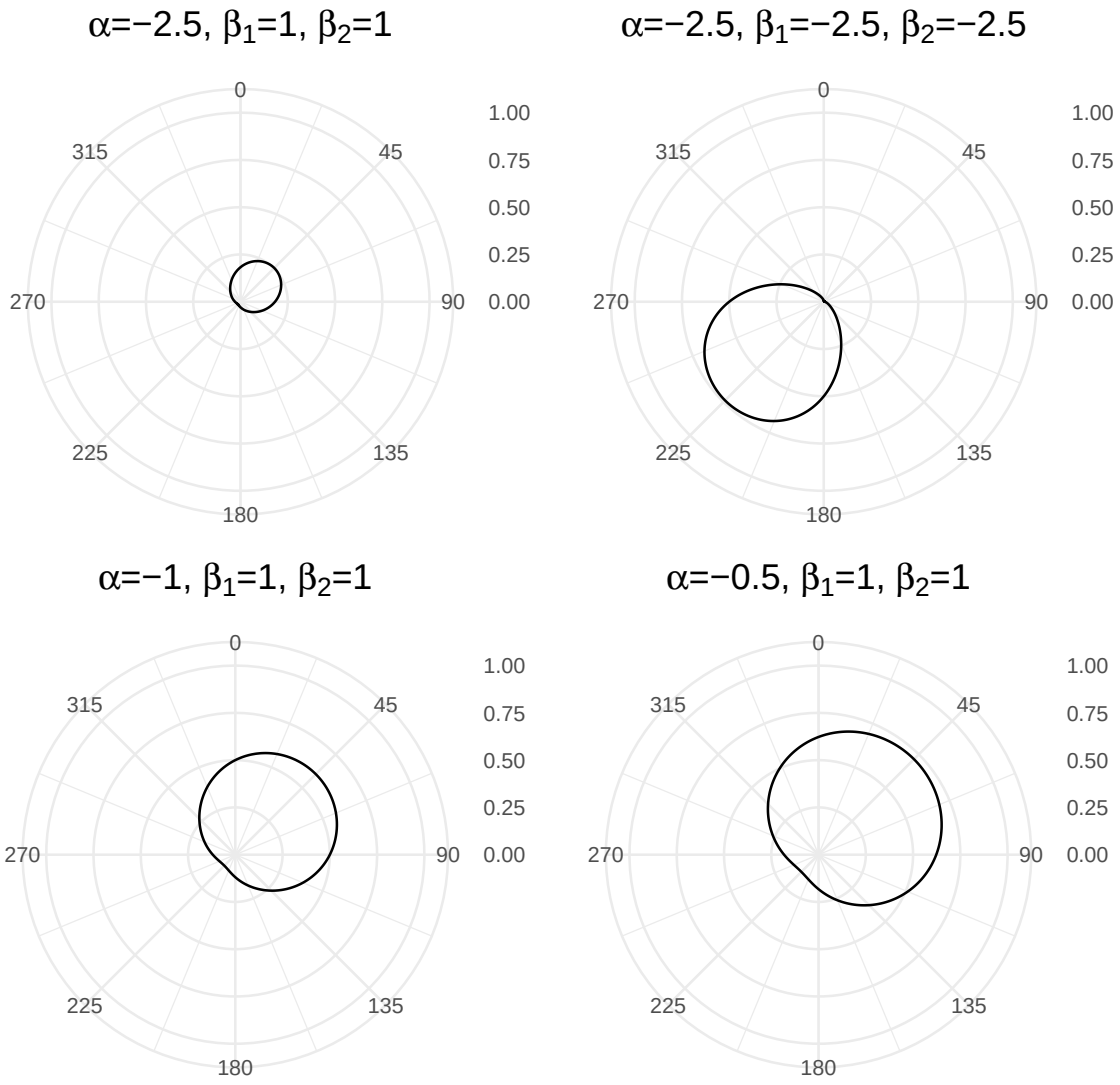

Figura 24: Esempio di $\pi(x) = \alpha + \beta_1 \sin(x) + \beta_2 \cos(x)$ per specifici α e $\boldsymbol{\beta}$.

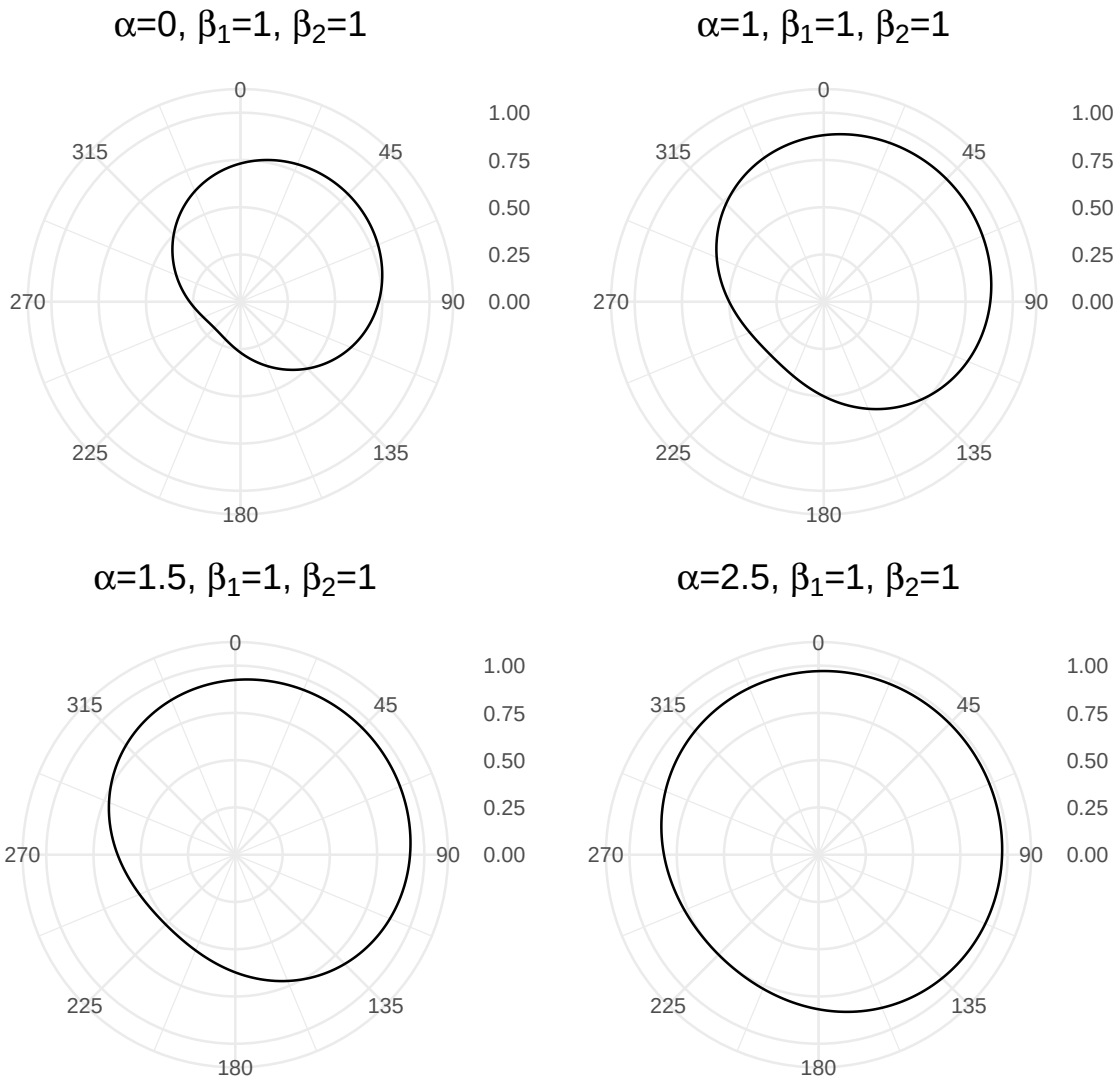

Figura 25: Esempio di $\pi(x) = \alpha + \beta_1 \sin(x) + \beta_2 \cos(x)$ per specifici α e $\boldsymbol{\beta}$.

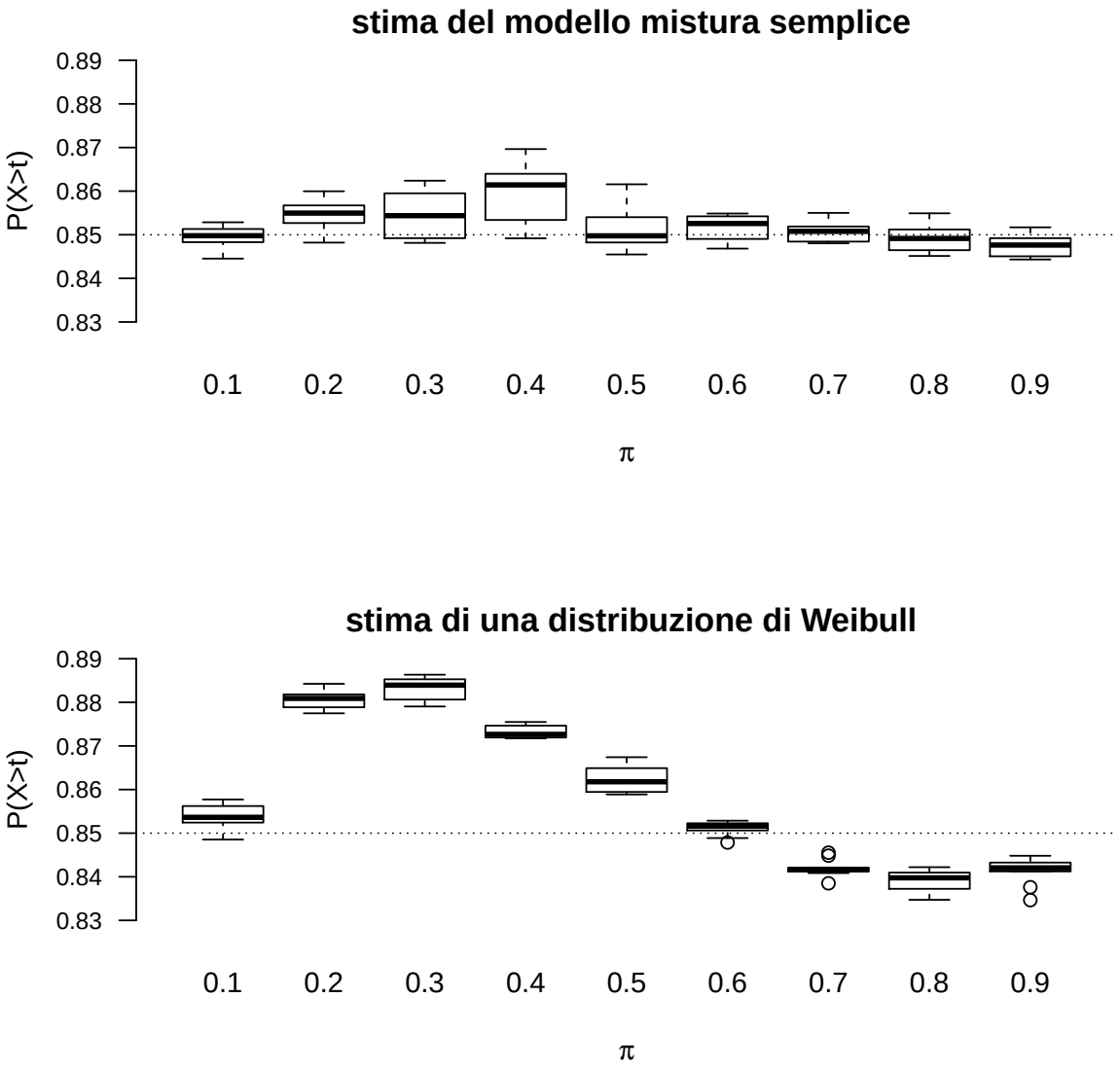

Figura 26: Stima delle probabilità nelle code per il quantile 0.85 per i campioni simulati dalla mistura semplice. I due grafici si riferiscono alle stime ottenute adattando il modello mistura corretto e una distribuzione di Weibull. Vengono distinti gli scenari definiti da π

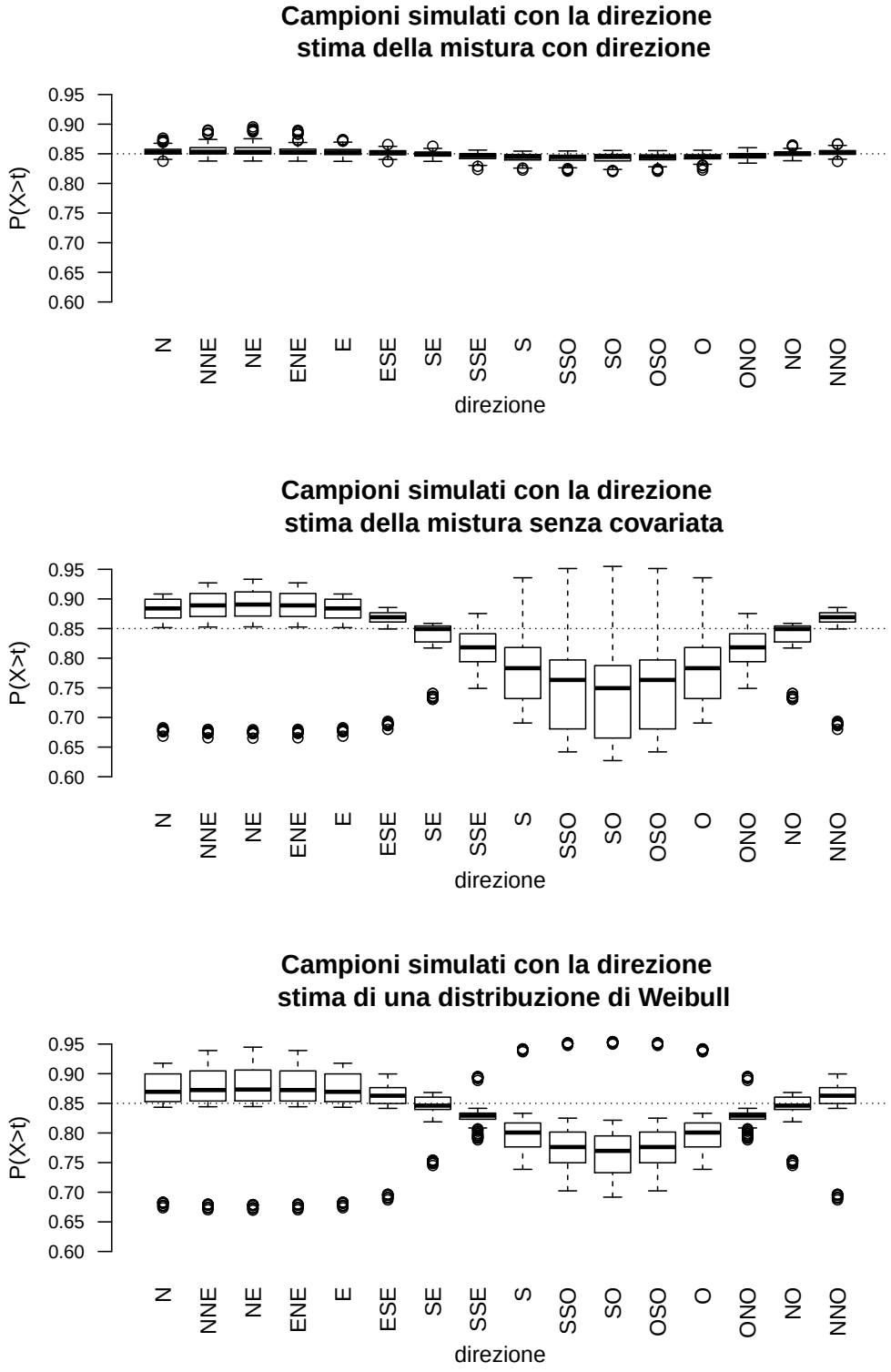

Figura 27: Stima della probabilità nelle code per il quantile 0.85 per i campioni simulati dalla mistura definita con la direzione. I grafici si riferiscono alle stime ottenute adattando il modello mistura corretto, il modello mistura senza la covariata e una distribuzione di Weibull. Vengono raggruppati i risultati ottenuti nei diversi scenari e distinti solo sulla base della direzione.

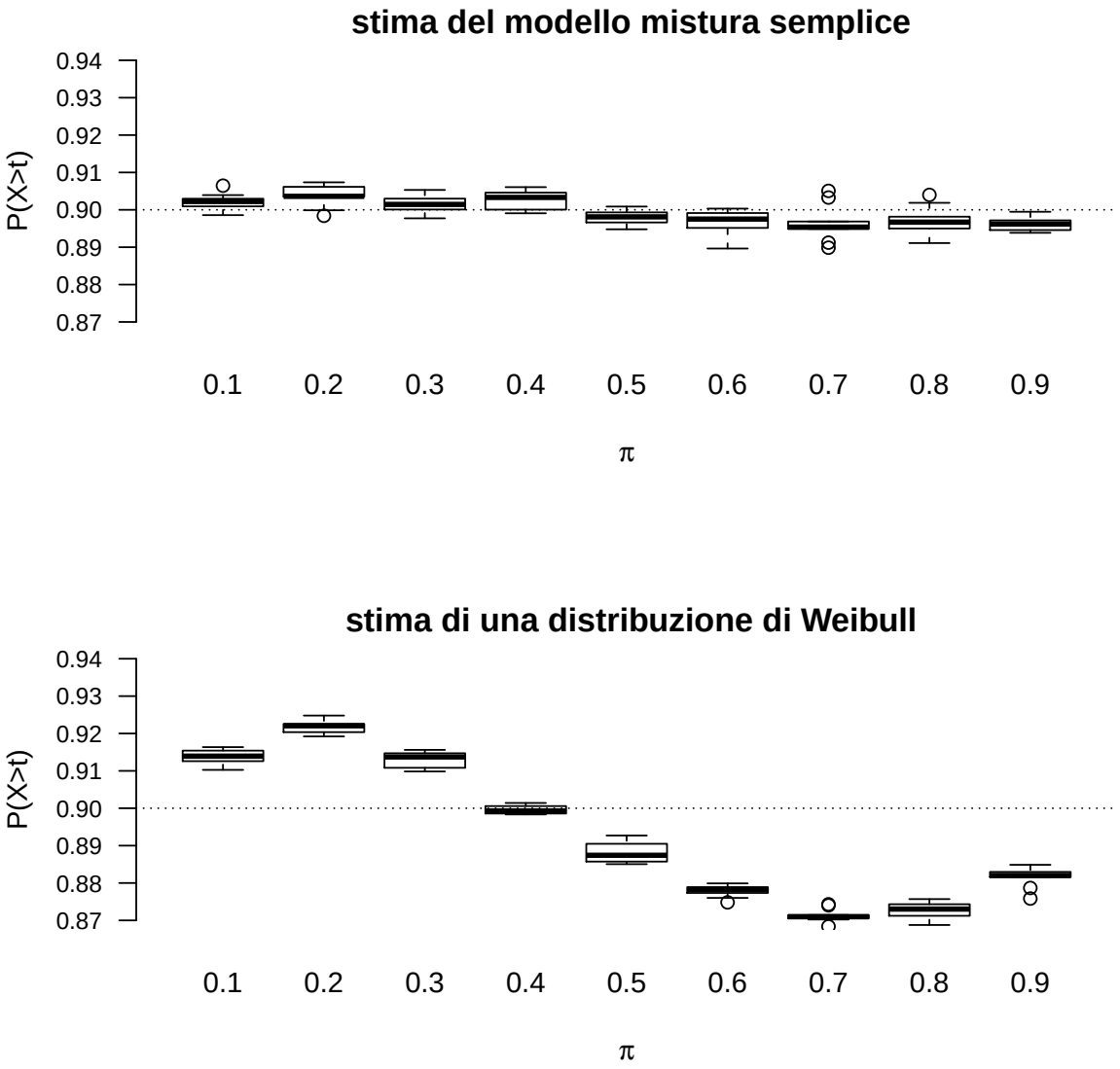

Figura 28: Stima delle probabilità nelle code per il quantile 0.90 per i campioni simulati dalla mistura semplice. I due grafici si riferiscono alle stime ottenute adattando il modello mistura corretto e una distribuzione di Weibull. Vengono distinti gli scenari definiti da π

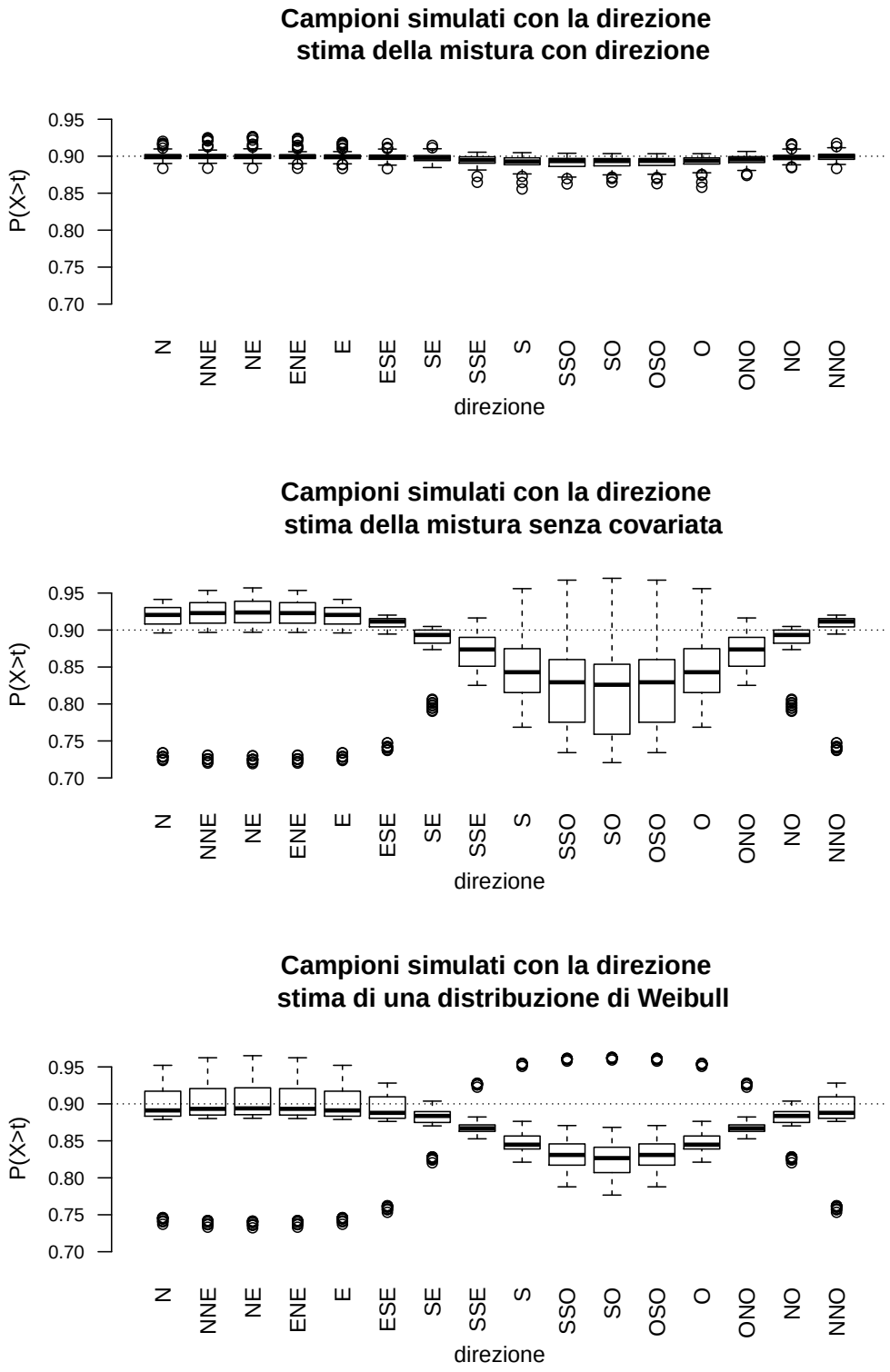

Figura 29: Stima della probabilità nelle code per il quantile 0.90 per i campioni simulati dalla mistura definita con la direzione. I grafici si riferiscono alle stime ottenute adattando il modello mistura corretto, il modello mistura senza la covariata e una distribuzione di Weibull. Vengono raggruppati i risultati ottenuti nei diversi scenari e distinti solo sulla base della direzione.

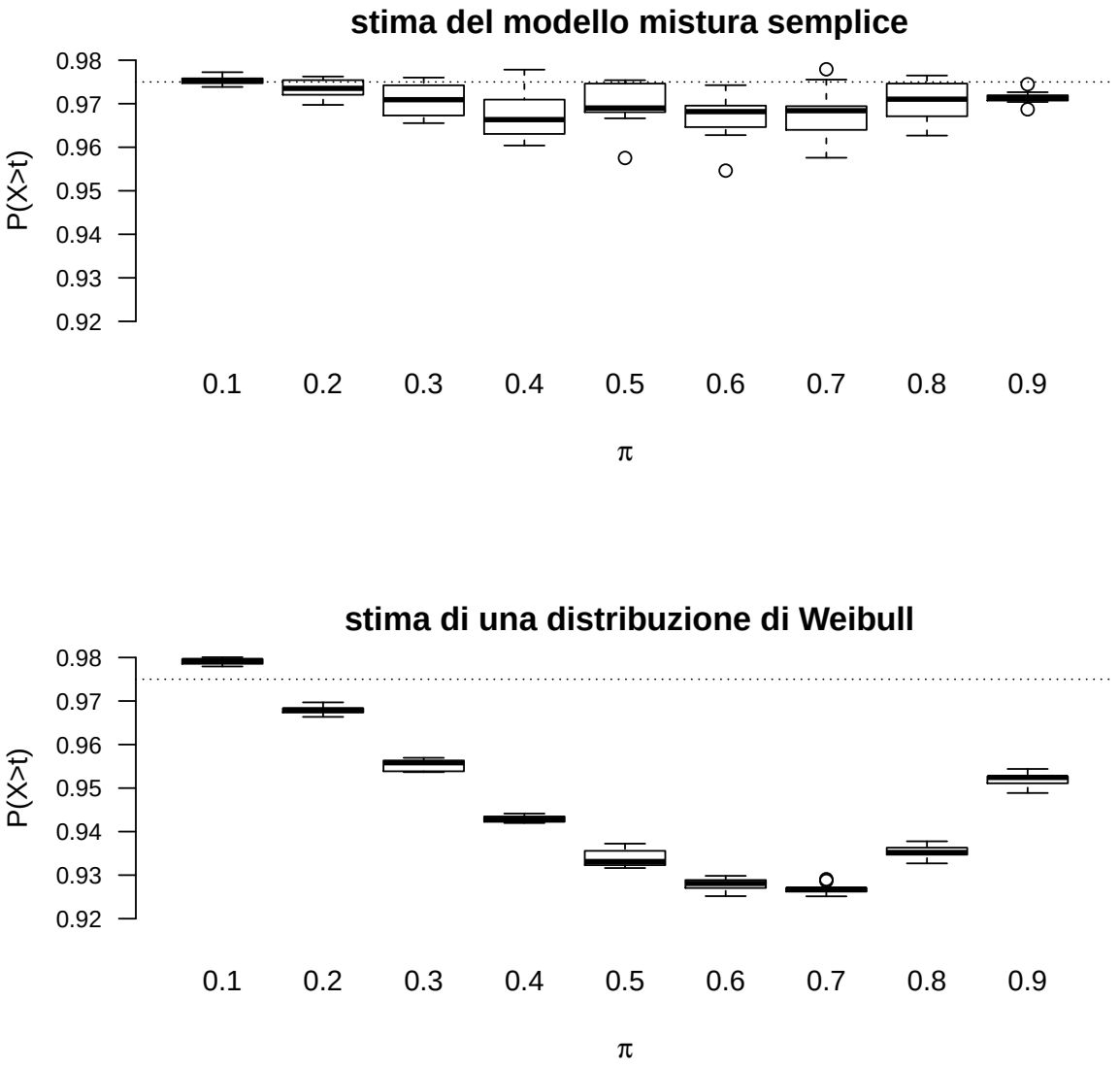

Figura 30: Stima delle probabilità nelle code per il quantile 0.975 per i campioni simulati dalla mistura semplice. I due grafici si riferiscono alle stime ottenute adattando il modello mistura corretto e una distribuzione di Weibull. Vengono distinti gli scenari definiti da π

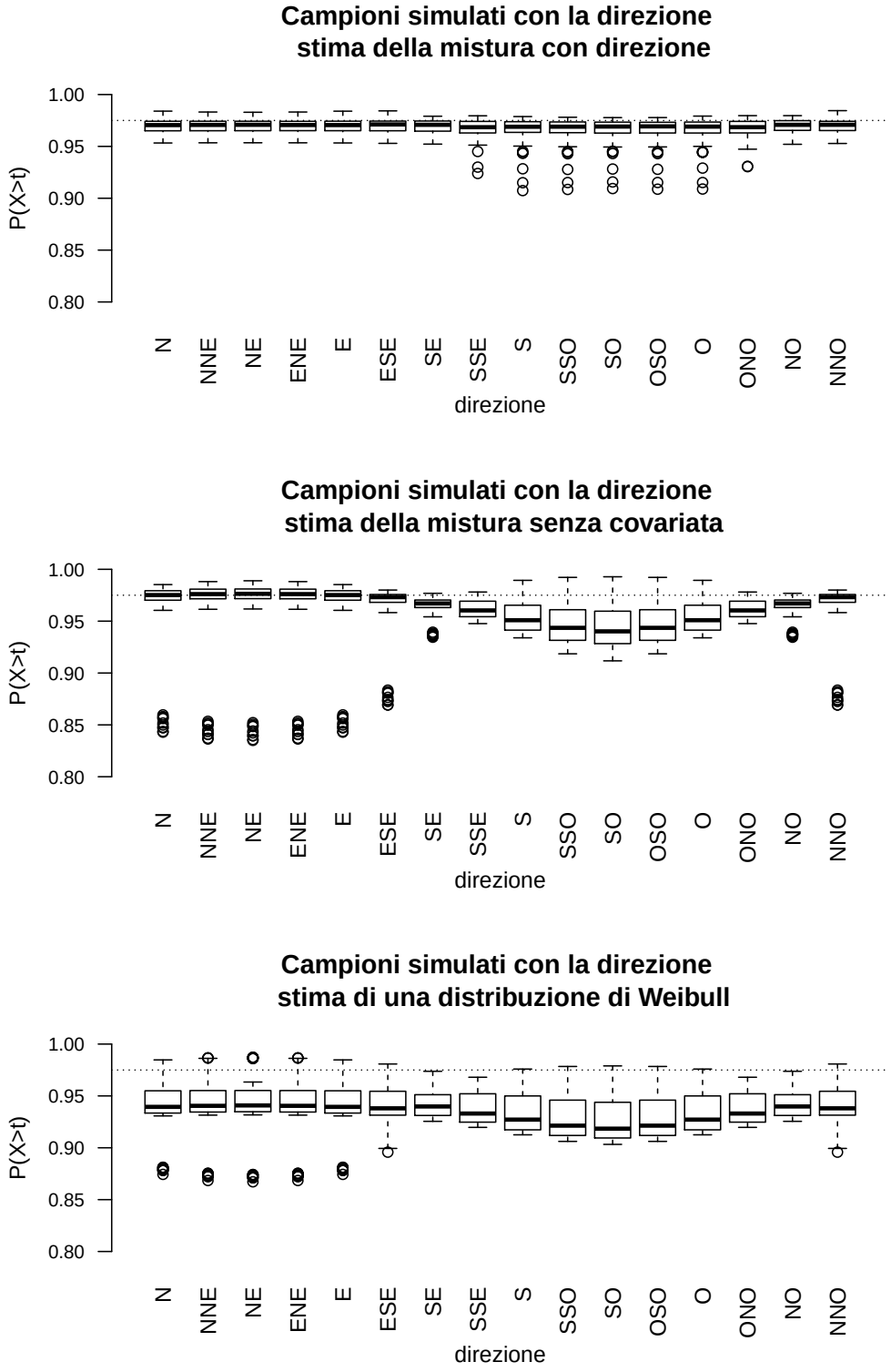

Figura 31: Stima della probabilità nelle code per il quantile 0.975 per i campioni simulati dalla mistura definita con la direzione. I grafici si riferiscono alle stime ottenute adattando il modello mistura corretto, il modello mistura senza la covariata e una distribuzione di Weibull. Vengono raggruppati i risultati ottenuti nei diversi scenari e distinti solo sulla base della direzione.

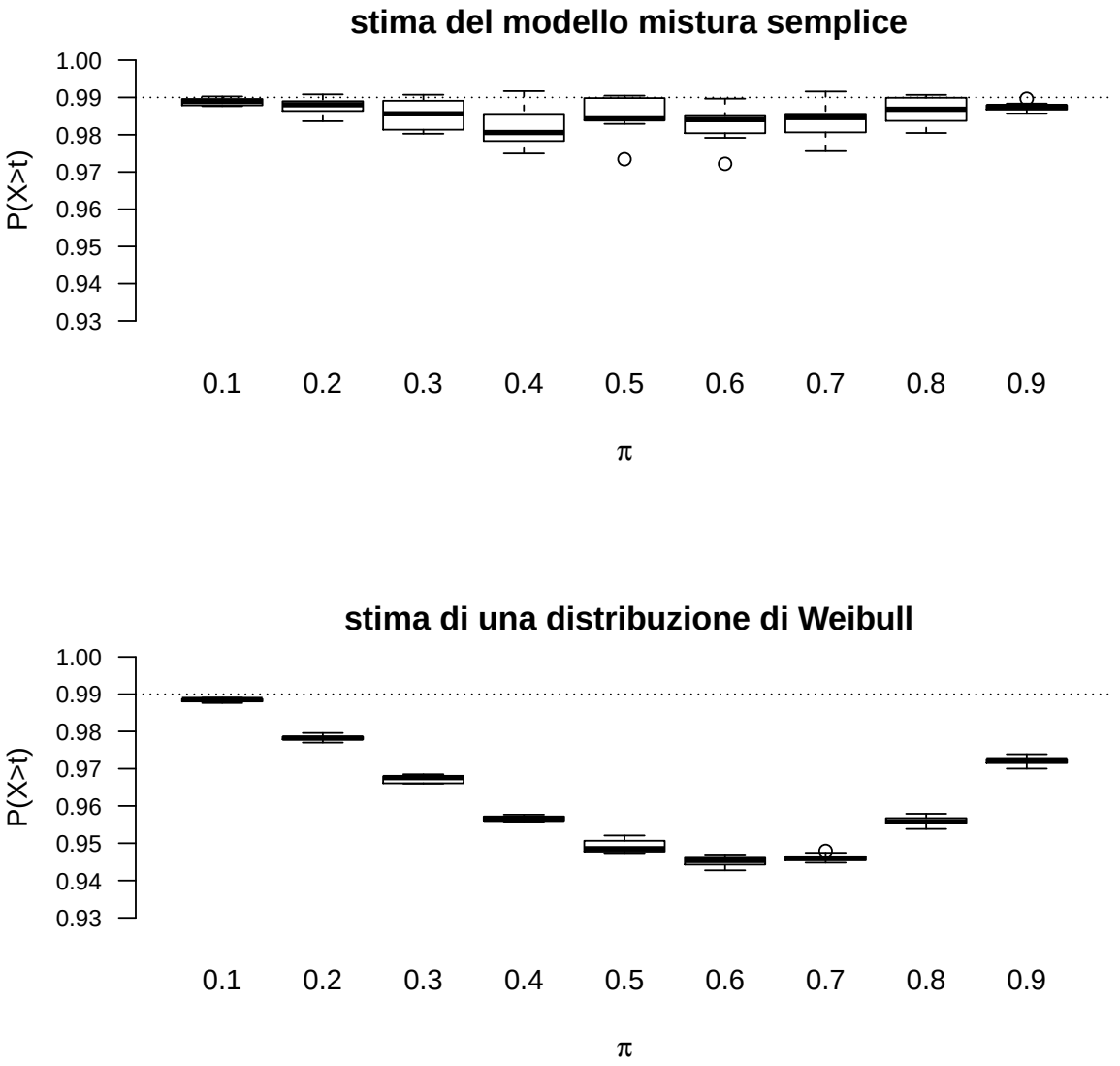

Figura 32: Stima delle probabilità nelle code per il quantile 0.99 per i campioni simulati dalla mistura semplice. I due grafici si riferiscono alle stime ottenute adattando il modello mistura corretto e una distribuzione di Weibull. Vengono distinti gli scenari definiti da π

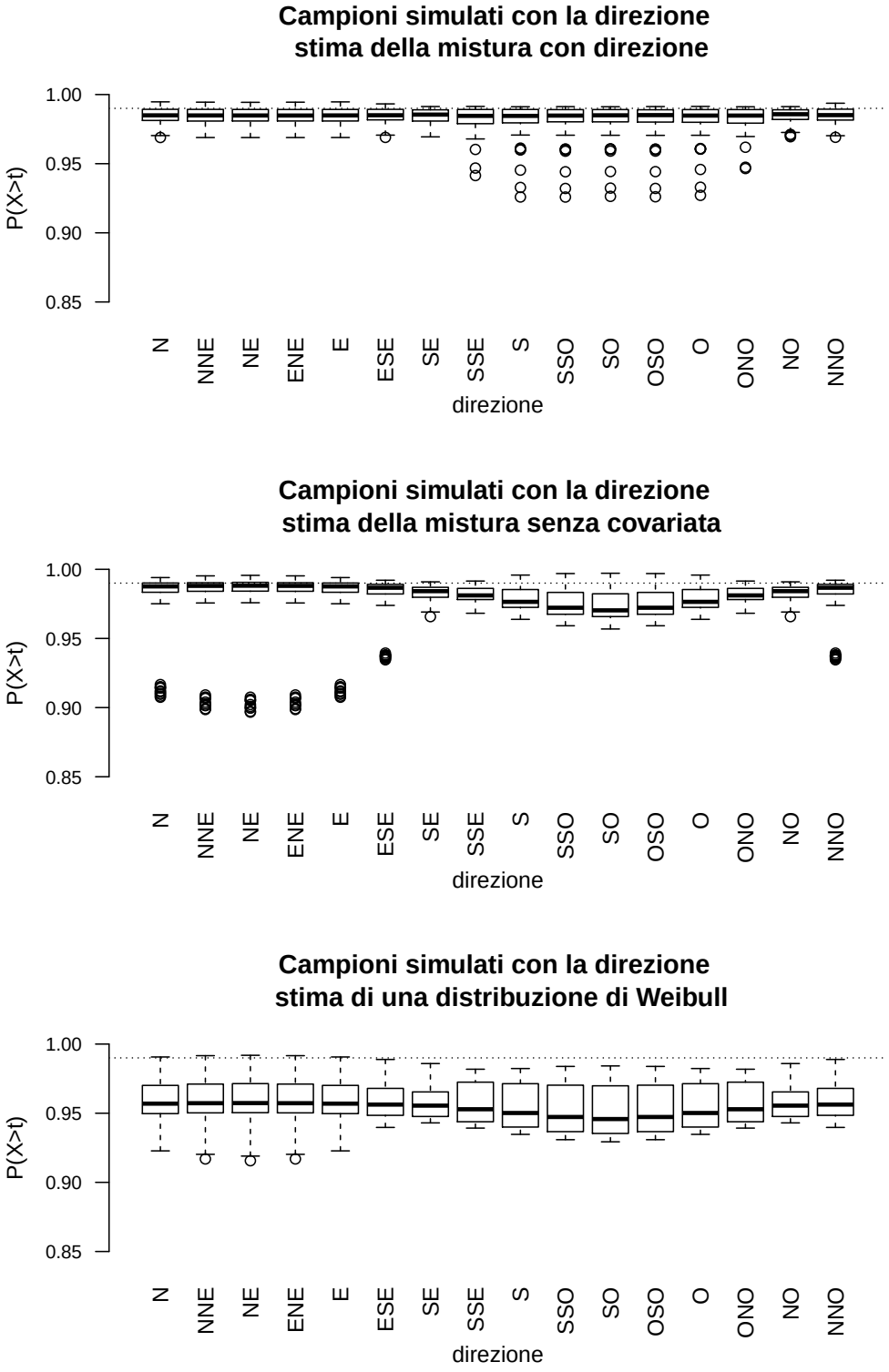

Figura 33: Stima della probabilità nelle code per il quantile 0.99 per i campioni simulati dalla mistura definita con la direzione. I grafici si riferiscono alle stime ottenute adattando il modello mistura corretto, il modello mistura senza la covariata e una distribuzione di Weibull. Vengono raggruppati i risultati ottenuti nei diversi scenari e distinti solo sulla base della direzione.

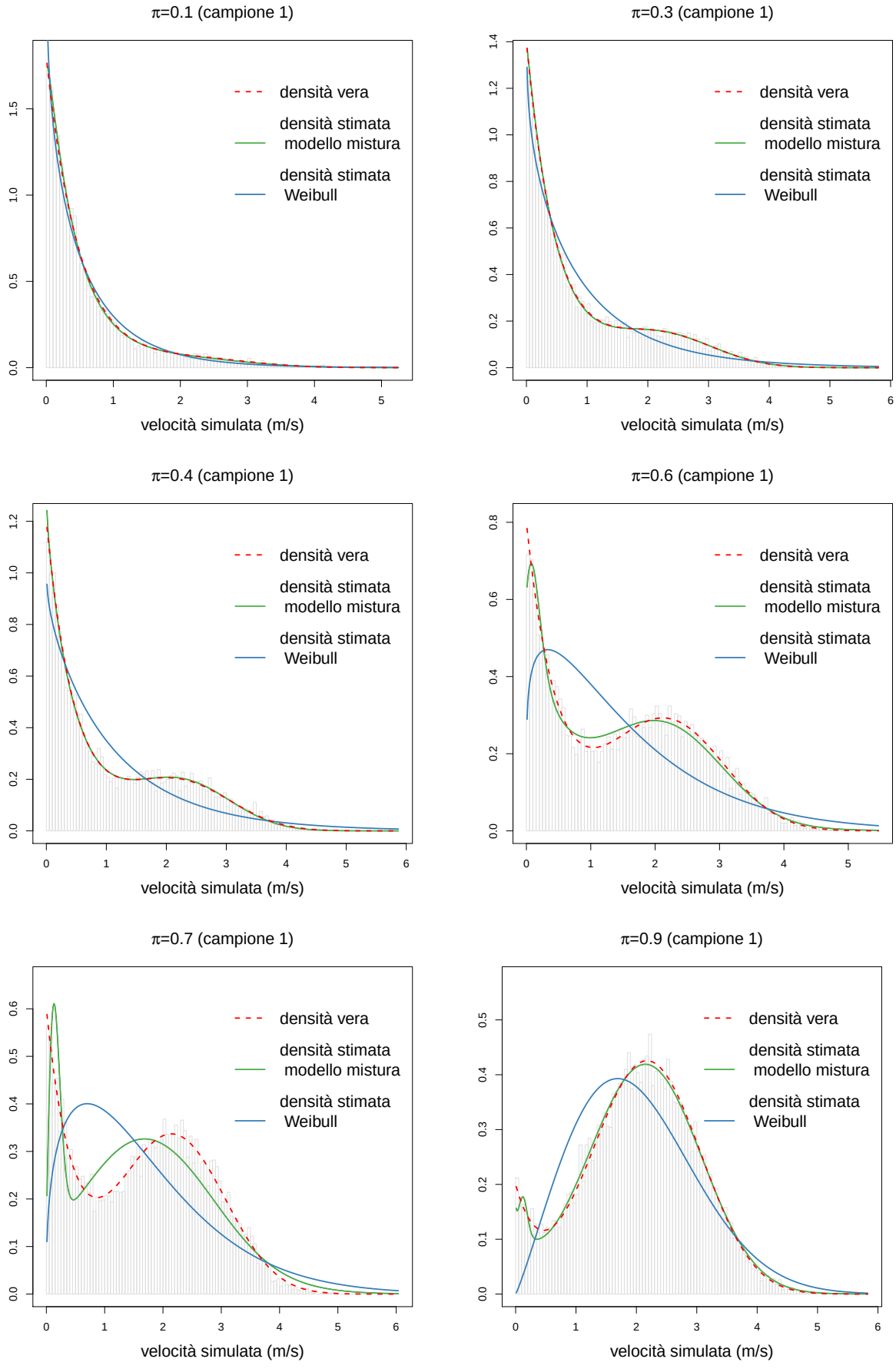

Figura 34: Densità sotto corretta ed errata specificazione per diversi scenari quando il vero modello è la mistura senza covariata.

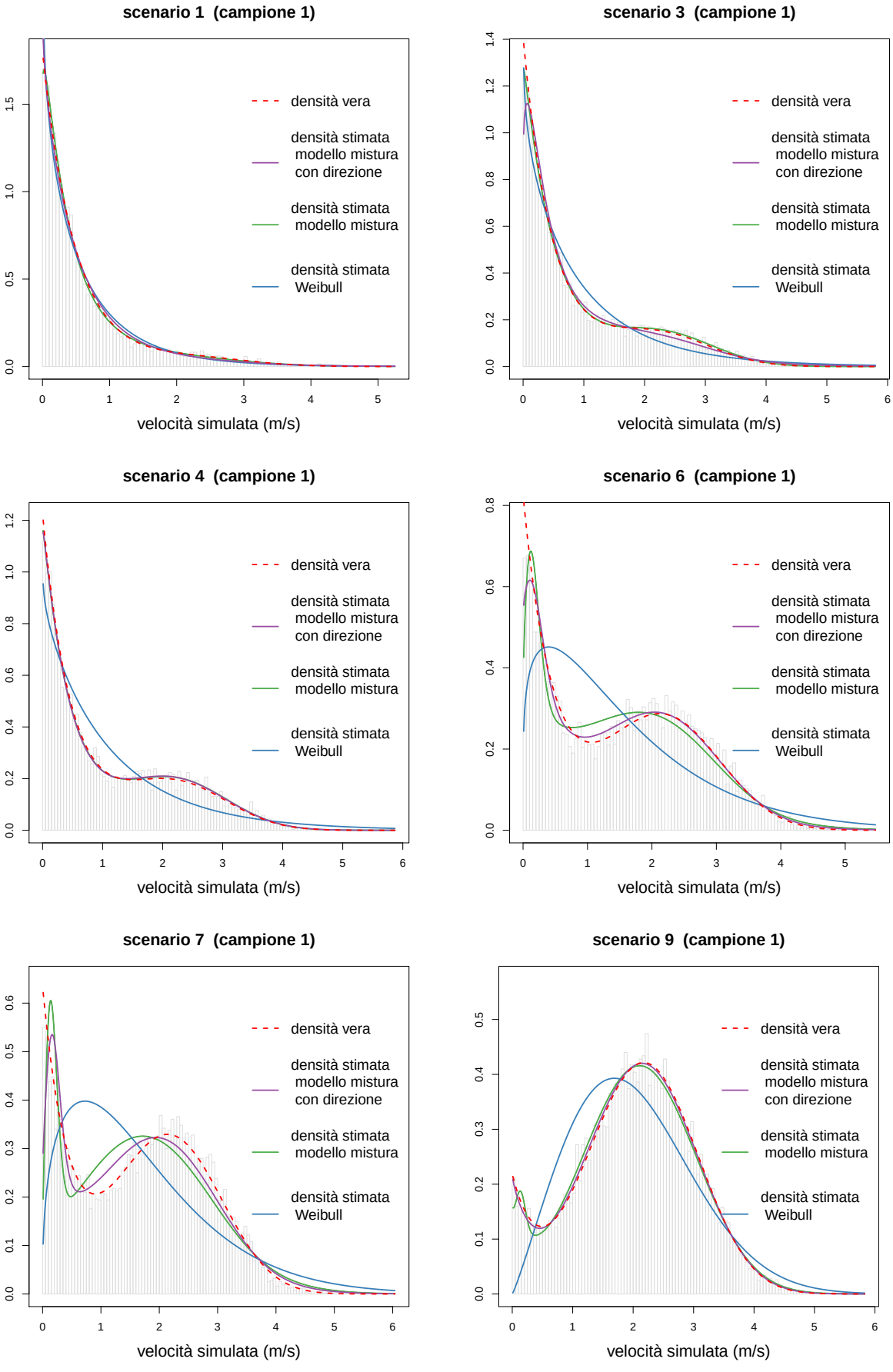

Figura 35: Densità sotto corretta ed errata specificazione per diversi scenari quando il vero modello è la mistura con la direzione del vento.

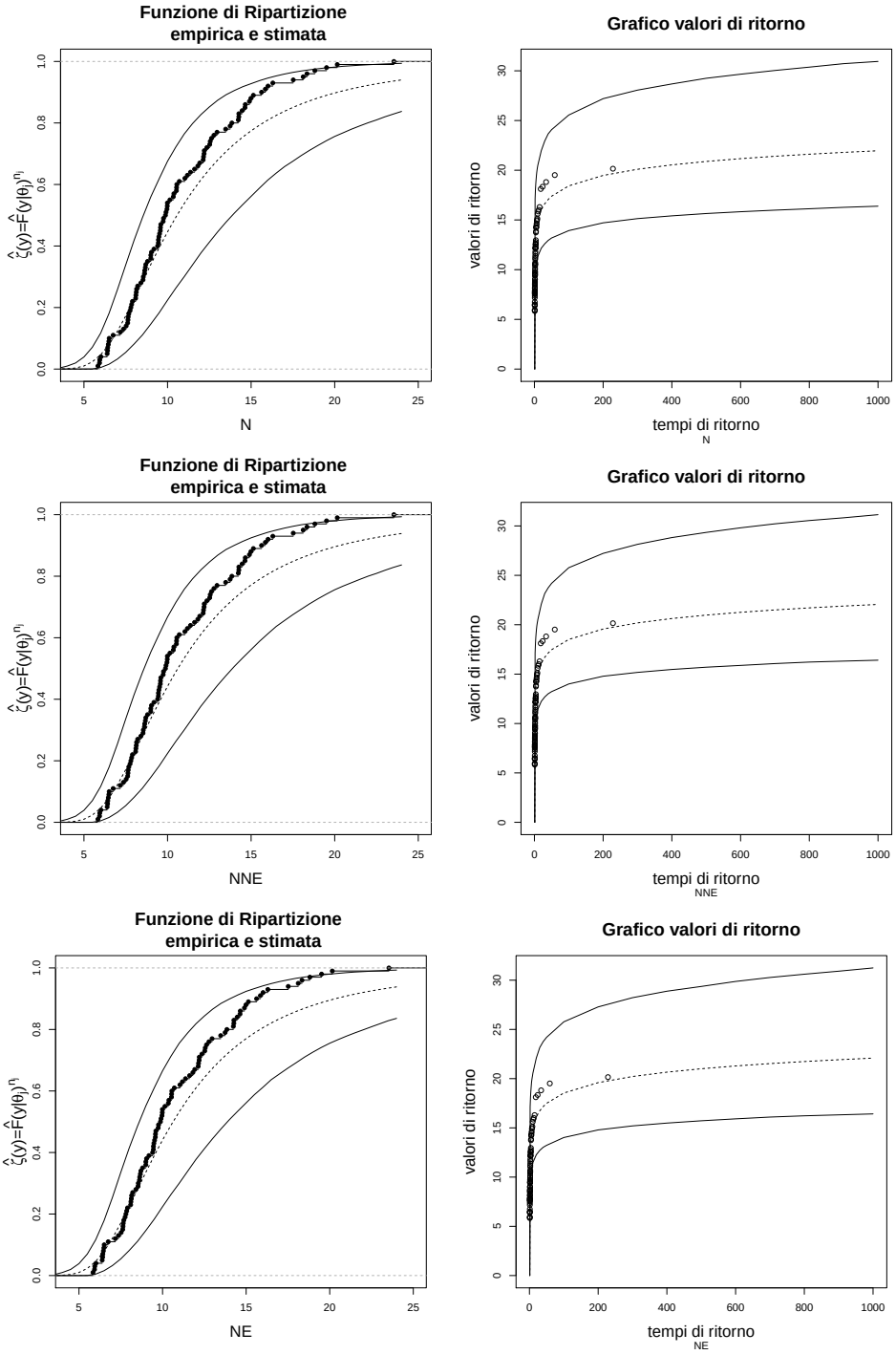

Figura 36: A sinistra funzione di ripartizione stimata sulla base del modello bayesiano gerarchico con intervalli di credibilità e funzione di ripartizione empirica dei massimi degli 80 anni non usati per la stima. A destra grafico dei valori di ritorno confrontato con i massimi degli 80 anni.

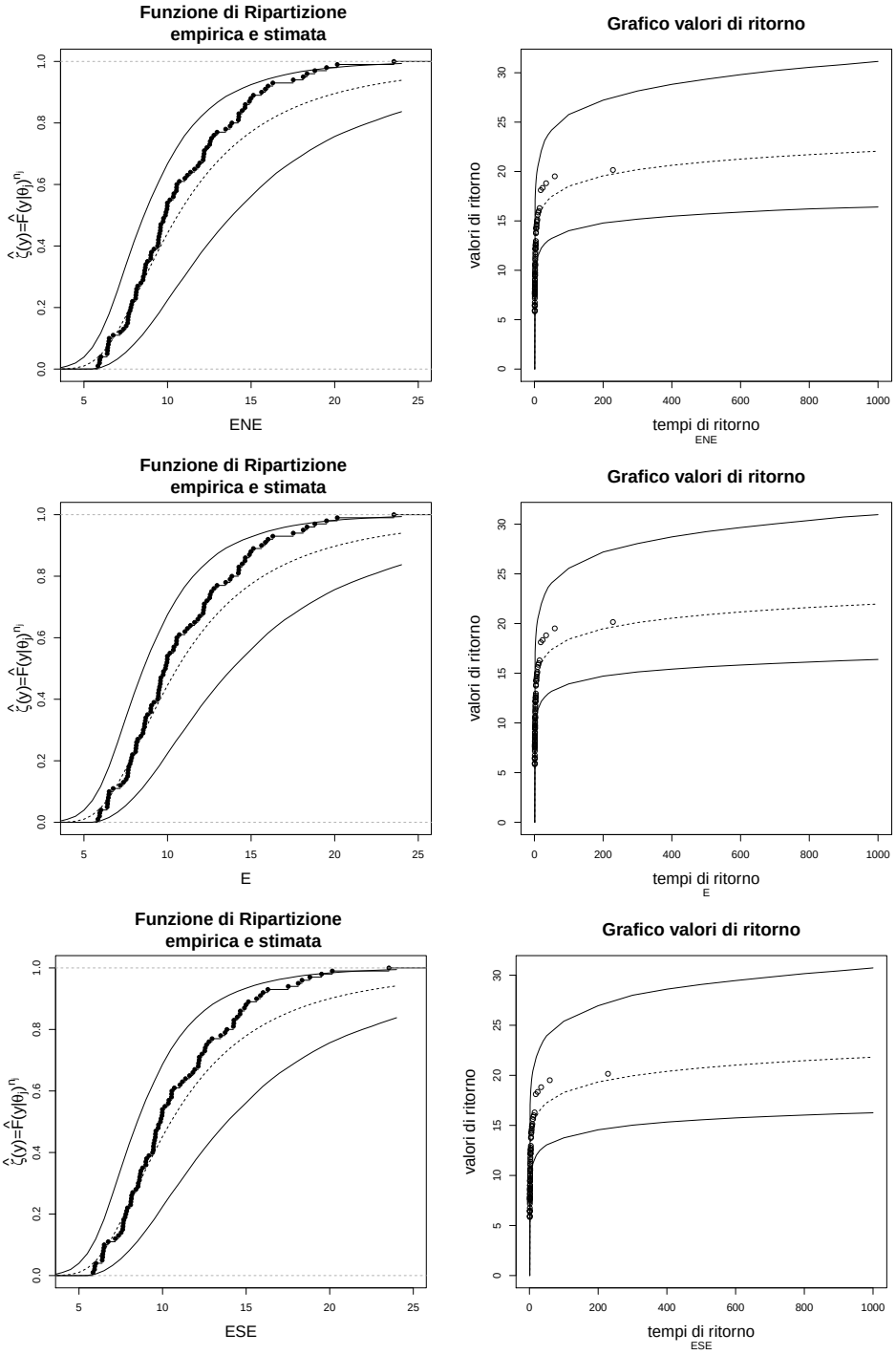

Figura 37: A sinistra funzione di ripartizione stimata sulla base del modello bayesiano gerarchico con intervalli di credibilità e funzione di ripartizione empirica dei massimi degli 80 anni non usati per la stima. A destra grafico dei valori di ritorno confrontato con i massimi degli 80 anni.

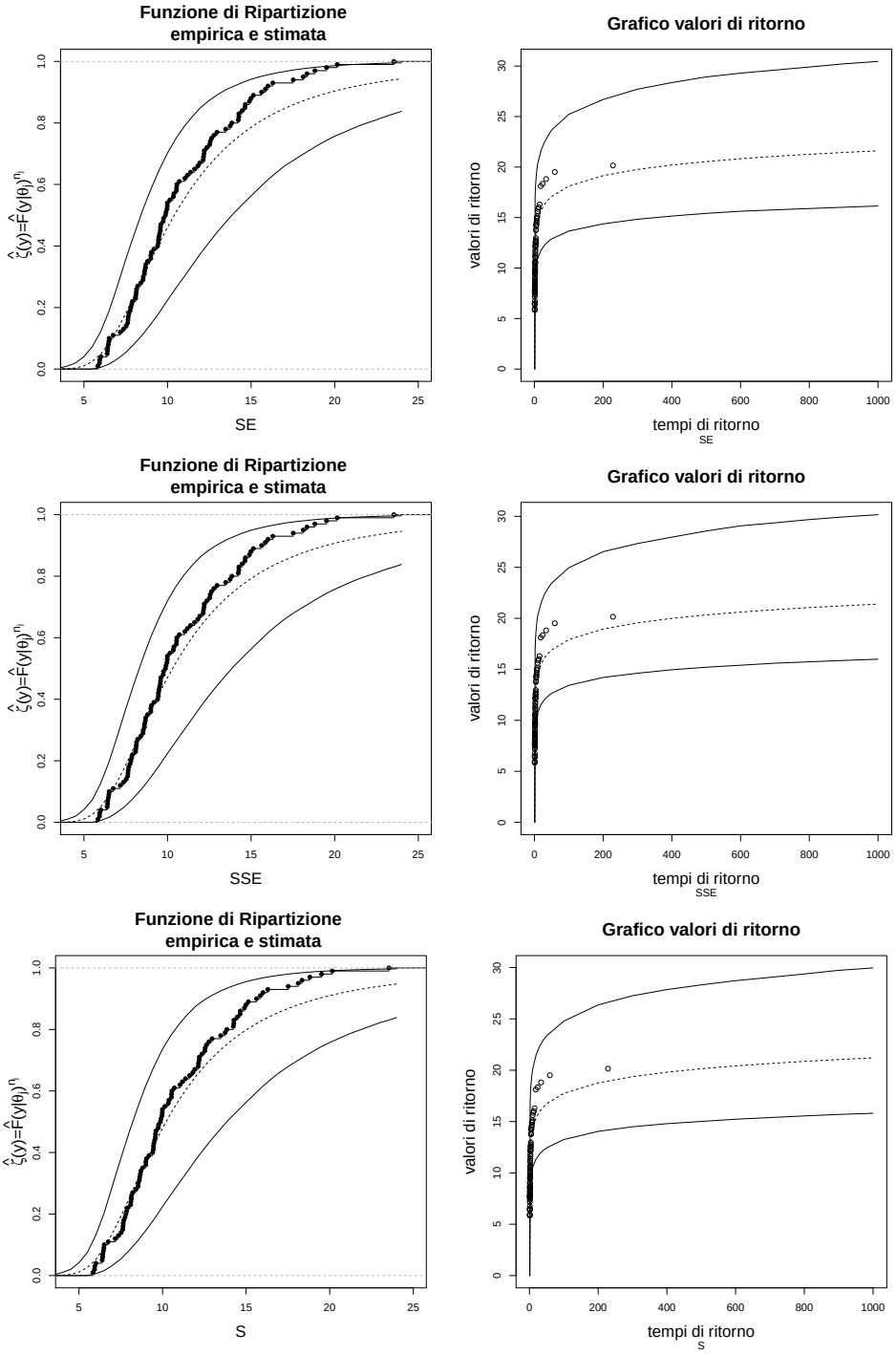

Figura 38: A sinistra funzione di ripartizione stimata sulla base del modello bayesiano gerarchico con intervalli di credibilità e funzione di ripartizione empirica dei massimi degli 80 anni non usati per la stima. A destra grafico dei valori di ritorno confrontato con i massimi degli 80 anni.

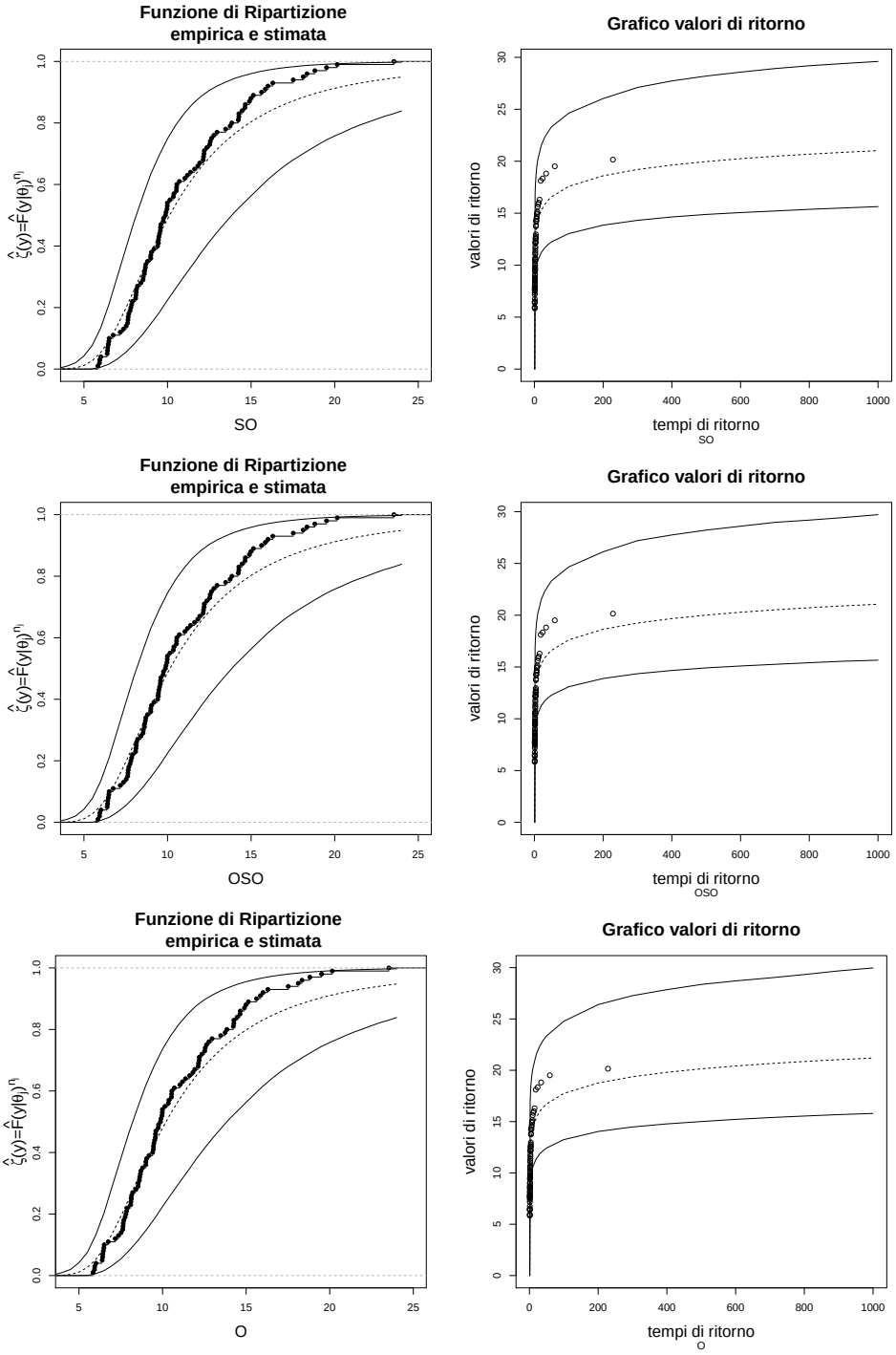

Figura 39: A sinistra funzione di ripartizione stimata sulla base del modello bayesiano gerarchico con intervalli di credibilità e funzione di ripartizione empirica dei massimi degli 80 anni non usati per la stima. A destra grafico dei valori di ritorno confrontato con i massimi degli 80 anni.

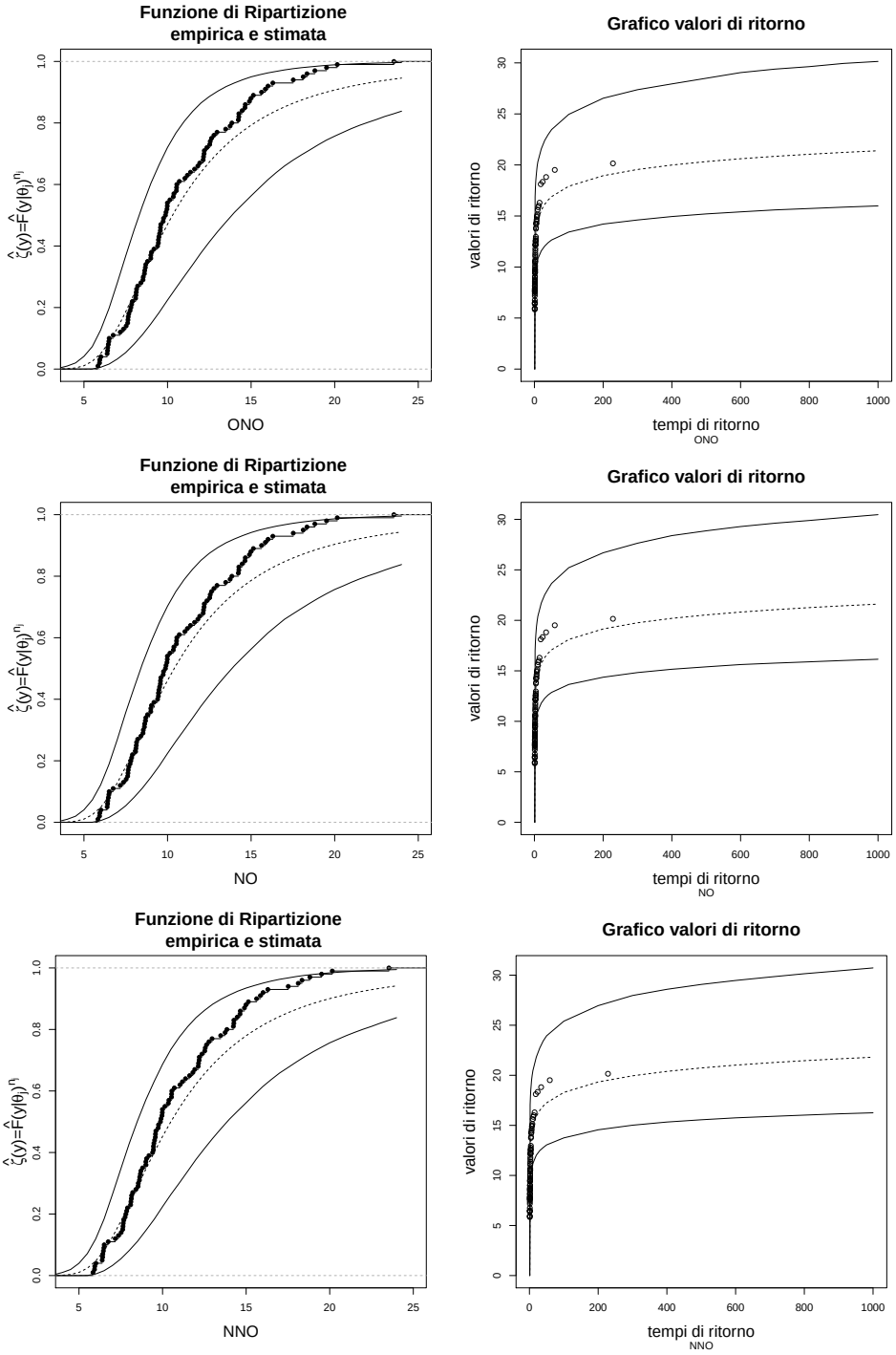

Figura 40: A sinistra funzione di ripartizione stimata sulla base del modello bayesiano gerarchico con intervalli di credibilità e funzione di ripartizione empirica dei massimi degli 80 anni non usati per la stima. A destra grafico dei valori di ritorno confrontato con i massimi degli 80 anni.