# POLITECNICO DI TORINO

**Corso di Laurea Magistrale**
**in Ingegneria Matematica**

Tesi di Laurea Magistrale

# Preferential sampling

**Relatore**
prof. Gianluca Mastrantonio

**Candidato**
Sonia Vittone

Anno Accademico 2024-2025

*A mia madre, la mia ricchezza*

# Sommario

La tesi analizza il fenomeno del campionamento preferenziale in diversi contesti evidenziandone le implicazioni. Il framework teorico è quello dell'approccio Bayesiano. Il lavoro si apre con un'introduzione alla statistica bayesiana, delineandone le fondamenta teoriche e le differenze rispetto all'approccio frequentista.

Successivamente, vengono approfonditi i metodi computazionali per l'implementazione di modelli bayesiani, con particolare attenzione agli algoritmi di campionamento e alle tecniche di inferenza. Viene inoltre discusso il ruolo delle distribuzioni a priori e la loro influenza sui risultati dell'analisi.

Attraverso l'analisi di casi studio e simulazioni, si dimostra come il campionamento preferenziale possa fornire stime più accurate e rappresentative in scenari complessi, contribuendo ad una comprensione più approfondita delle dinamiche sottostanti ai dati osservati.

# Indice

# Capitolo 1

# Introduzione alla Statistica Bayesiana

## 1.1 La statistica bayesiana: un nuovo paradigma

Lo scopo di questa introduzione non è quello di fornire un'introduzione approfondita all'inferenza bayesiana, ma di introdurre alcune notazioni ed il contesto per i successivi capitoli.

La statistica bayesiana rappresenta un approccio distinto e complementare rispetto alla statistica frequentista, comunemente insegnata e applicata in ambito accademico e professionale. Questo approccio prende il nome dal matematico Thomas Bayes, una delle due figure chiave, insieme a Pierre-Simon Laplace, che hanno contribuito alla nascita del pensiero bayesiano.

Bayes ha iniziato a considerare la probabilità come uno strumento per spiegare matematicamente le relazioni di causa-effetto. Il suo interesse era dimostrare che, conoscendo un effetto, si poteva determinare la probabilità delle cause che avrebbero potuto generarlo, fondando così la teoria della probabilità inversa, alla base del pensiero bayesiano. Pierre-Simon Laplace, indipendentemente da Bayes, formulò il principio fondamentale del bayesianesimo e della teoria della probabilità inversa, affermando che la probabilità di una causa è proporzionale alla probabilità di un evento dato quella causa, fornendo così la prima versione di quello che oggi chiamiamo teorema di Bayes.

La statistica bayesiana si distingue per il suo approccio nel trattare l'incertezza e le probabilità. La probabilità viene considerata come una misura soggettiva della nostra conoscenza o incertezza rispetto a un evento. Questa visione rende così la statistica bayesiana particolarmente utile in contesti dove le informazioni sono incomplete o dove è necessario aggiornare costantemente le stime alla luce di nuove evidenze.

I metodi bayesiani consentono di ottenere stime dei parametri con solide proprietà statistiche, di descrivere i dati osservati in modo efficiente, di prevedere sia dati mancanti che futuri, e di fornire un quadro computazionale per la stima, la selezione e la validazione dei modelli.

Questo capitolo fornisce un'introduzione agli elementi fondamentali dell'apprendimento bayesiano. Per chiarimenti e ulteriori approfondimenti, si può consultare il testo "A First Course in Bayesian Statistical Methods" di PD. Hoff [6] e i capitoli 3, 4 e 5 del libro "Spatial and Spatio-temporal Bayesian Models with R-INLA" di M. Blangiardo e M. Cameletti [1].

## 1.2 Il teorema di Bayes

Il cuore della statistica bayesiana è il teorema di Bayes, che fornisce una regola matematica per aggiornare le probabilità alla luce di nuove evidenze. Il teorema deriva dalla probabilità condizionata:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \tag{1.1}$$

da cui riarrangiando i termini, si può scrivere che la probabilità dell'intersezione tra gli eventi A e B é uguale a

$$P(A \cap B) = P(A \mid B) \times P(B). \tag{1.2}$$

Se vogliamo sapere la probabilità dell'evento B, condizionatamente ad A, possiamo applicare nuovamente la formula 1.1 e otteniamo:

$$P(B \mid A) = \frac{P(A \cap B)}{P(A)}, \tag{1.3}$$

dalla sostituzione della formula 1.2 nella 1.3 ne discende il seguente teorema di Bayes [1].

**Teorema 1.2.1 (Teorema di Bayes)**

$$P(B \mid A) = \frac{P(A \mid B) \times P(B)}{P(A)}, \tag{1.4}$$

*dove:*

- $P(B \mid A)$ *è la probabilità a posteriori, cioè la probabilità dell'evento B dato l'evento A.*
- $P(A \mid B)$ *è la probabilità condizionale di A dato B, chiamata anche verosimigliaza.*
- $P(B)$ *è la probabilità a priori di B, ovvero la nostra conoscenza iniziale prima di osservare A.*
- $P(A)$ *è la probabilità totale di A, spesso ottenuta come somma pesata delle probabilità condizionali di A su tutti i possibili eventi B.*

L'interpretazione di questo teorema si basa sulla teoria della probabilità inversa, sviluppata da Thomas Bayes e successivamente ampliata da Pierre-Simon Laplace. Se consideriamo il teorema di Bayes in un contesto sperimentale, possiamo affermare che, prima di condurre l'esperimento, il ricercatore possiede alcune informazioni sull'evento B, il cui grado di incertezza è rappresentato da $P(B)$ (la probabilità a priori). Questa informazione viene poi combinata con il risultato dell'esperimento, rappresentato da $P(A \mid B)$ (la verosimiglianza), per ottenere una probabilità aggiornata riguardo a B, ossia $P(B \mid A)$ (la probabilità a posteriori).

## 1.3 L'importanza delle distribuzioni a priori

Uno degli aspetti caratterizzanti è l'utilizzo delle distribuzioni a priori, le quali riflettono le credenze iniziali riguardo ad un parametro o ad un evento prima di osservare i dati. La scelta della distribuzione a priori può influenzare significativamente i risultati dell'analisi, specialmente quando i dati osservati sono limitati. Due elementi fondamentali devono essere considerati:

- il tipo di distribuzione, che deve riflettere la natura del parametro in esame, e
- gli iperparametri, che determinano quanto la distribuzione sia informativa, modulando il livello di conoscenza o ignoranza sui parametri.

In genere, esiste una distribuzione "naturale" per ogni tipo di parametro. Ad esempio, se si sta analizzando una proporzione (come la probabilità di decesso in una popolazione o la proporzione di risposte a un farmaco), l'incertezza su questo parametro dovrebbe essere rappresentata da una distribuzione che varia tra 0 e 1. Le distribuzioni a priori possono essere informative o non informative. Le prime incorporano conoscenze o ipotesi precedenti, che possono derivare da studi passati, esperienze o teorie consolidate. Al contrario, quelle non informative, sono utilizzate quando non si ha conoscenza specifica sul parametro, cercando di mantenere un'impostazione neutrale per far sì che i dati osservati guidino l'analisi.

## 1.4 Verosimiglianza e distribuzioni a posteriori

La verosimiglianza è una componente cruciale nel processo bayesiano, poiché esprime la probabilità di osservare i dati, dati i parametri del modello. In pratica, essa ci dice quanto è probabile che i dati osservati emergano sotto una specifica ipotesi.

Per comprendere come si ottiene la distribuzione a posteriori, consideriamo una variabile casuale Y, la cui incertezza è rappresentata da una distribuzione di probabilità o una funzione di densità, a seconda che Y sia discreta o continua. Questa distribuzione è parametrizzata da un parametro generico $\theta$. Ad esempio, in un contesto reale, la variabile casuale Y potrebbe essere il numero di decessi per malattie respiratorie. Osserviamo una realizzazione $Y = y$ e siamo interessati a studiare il tasso di mortalità $\theta$ nella popolazione. La funzione di verosimiglianza che specifica la distribuzione dei dati $y$ sotto il modello definito da $\theta$ è:

$$L(\theta) = p(Y = y \mid \theta), \tag{1.5}$$

dove $p(\cdot)$ è usato per indicare la distribuzione di probabilità o la funzione di densità di una variabile casuale e per semplicità denoteremo la funzione di verosimiglianza come $p(y \mid \theta)$. Assumiamo che i dati siano un campione casuale della popolazione in studio e l'incertezza deriva dal fatto che osserviamo solo quel campione invece di tutti gli altri possibili. Mentre, il parametro $\theta$ è una quantità sconosciuta, modellata attraverso un'opportuna distribuzione di probabilità a priori $p(\theta)$ prima di osservare qualsiasi realizzazione $y$ della variabile casuale Y e riflette la nostra conoscenza su $\theta$. Una volta definite le distribuzioni

a priori e calcolata la verosimiglianza, possiamo applicare il teorema di Bayes per ottenere la distribuzione a posteriori:

$$p(\theta \mid y) = \frac{p(y \mid \theta) \times p(\theta)}{p(y)}, \tag{1.6}$$

che rappresenta la nostra conoscenza aggiornata sul parametro del modello dopo aver osservato i dati. In altre parole, la distribuzione a posteriori è il risultato finale del processo bayesiano e incorpora sia le informazioni a priori che le nuove evidenze. Si noti che $p(y)$, nel denominatore dell'equazione 1.6, è la distribuzione marginale dei dati ed è considerata una costante di normalizzazione in quanto non dipende da $\theta$. Per ottenere la distribuzione marginale $p(y)$ dobbiamo applicare la legge delle probabilità totali per eventi mutuamente esclusivi ed esaustivi. Per spiegare questo punto, assumiamo che $\theta$ sia un parametro discreto che assume i valori 0 e 1. Consideriamo innanzitutto la probabilità condizionata $p(y \mid \theta = 0)$, pesandola con la probabilità che $\theta$ assuma il valore 0, $p(\theta = 0)$; consideriamo poi la probabilità condizionata $p(y \mid \theta = 1)$, pesandola con la probabilità che $\theta$ assuma il valore 1, $p(\theta = 1)$; infine $p(y)$ sarà semplicemente:

$$p(y) = p(y \mid \theta = 0) \times p(\theta = 0) + p(y \mid \theta = 1) \times p(\theta = 1). \tag{1.7}$$

Questo processo può essere facilmente esteso al caso in cui $\theta$ può assumere valori discreti in $\Theta$, il che porta a

$$p(y) = \sum_{\theta \in \Theta} p(y \mid \theta) p(\theta). \tag{1.8}$$

Quando $\theta$ è una variabile continua, la somma nell'equazione precedente è sostituita dall'integrale

$$p(y) = \int_{\theta \in \Theta} p(y \mid \theta) p(\theta) d\theta. \tag{1.9}$$

Inoltre, la distribuzione a priori e quella a posteriori si dicono coniugate quando la distribuzione a posteriori appartiene alla stessa famiglia della distribuzione a priori, dopo aver aggiornato le informazioni con i dati osservati. In altre parole, se scegliamo una prior coniugata, il processo bayesiano di aggiornamento mantiene la stessa forma della distribuzione, facilitando i calcoli. Tuttavia questa proprietà non si applica nella maggior parte dei casi pratici.

## 1.5 Esempio pratico: HIV test

Questo esempio è tratto da Blangiardo e Cameletti [1] e riguarda il calcolo della probabilità che un paziente abbia l'HIV, dato un risultato positivo al test, utilizzando il teorema di Bayes. Siamo in presenza di un test con una sensibilità del 95%, che indica la probabilità che il test risulti positivo se una persona ha l'HIV, e una specificità del 98%, che si traduce nella probabilità che il test risulti negativo se una persona non ha l'HIV. Nella popolazione inglese, la prevalenza dell'HIV (percentuale di popolazione con HIV) è pari a 0.0015, mentre il complemento di questa prevalenza, ossia la probabilità che una persona non abbia l'HIV, è pari a 1 - 0.0015 = 0.9985. Lo scopo è calcolare la probabilità che una persona

abbia l'HIV, dato che il test è risultato positivo. Definiamo B come l'evento che il paziente sia veramente positivo all'HIV e $B^c$ l'evento che sia veramente negativo all'HIV. Poiché B o $B^c$ si verificheranno definitivamente, ma non potranno accadere contemporaneamente, possiamo concludere che B e $B^c$ sono eventi esaustivi e mutuamente esclusivi. Se A è l'evento che il paziente risulta positivo, vogliamo valutare la probabilità dell'evento B dato A: $P(B \mid A)$. Dalle informazioni disponibili sul test, la "sensibilità del 95%" può essere tradotta in termini probabilistici come $P(A \mid B) = 0.95$, mentre la "specificità del 98%" può essere scritta come $P(A \mid B^c) = 0.02$. Applicando quindi il teorema di Bayes nella sua formulazione presentata nell'equazione 1.4, otteniamo

$$P(B \mid A) = \frac{P(A \mid B)P(B)}{P(A \mid B)P(B) + P(A \mid B^c)P(B^c)} =$$

$$= \frac{0.95 \times 0.0015}{0.95 \times 0.0015 + 0.02 \times 0.9985} = 0.067.$$

Possiamo quindi concludere che circa il 6.7% dei soggetti positivi al test è effettivamente affetto da HIV. Questo risultato indica che, nonostante il test abbia un'alta sensibilità e specificità, a causa della bassa prevalenza dell'HIV nella popolazione (0.0015), solo una piccola percentuale delle persone con un test positivo avrà effettivamente l'HIV.
Supponendo ora che la percentuale di sieropositivi nella popolazione sia del 20%, la probabilità che una persona abbia l'HIV, dato un risultato positivo al test, aumenterebbe significativamente. Applicando il teorema di Bayes, la probabilità $P(B \mid A)$ si calcola come

$$P(B \mid A) = \frac{0.95 \times 0.2}{0.95 \times 0.2 + 0.02 \times 0.8} = 0.92.$$

Quando la prevalenza dell'HIV è molto bassa (ad esempio $P(A) = 0.0015$), l'osservazione di un risultato positivo porta la probabilità a posteriori a circa $P(B \mid A) = 0.067$, ma se la prevalenza è più alta (ad esempio $P(A) = 0.2$), la probabilità a posteriori sale fino a $P(B \mid A) = 0.92$. Questo evidenzia come le informazioni a priori influenzano i risultati delle probabilità a posteriori.

## 1.6 Metodi computazionali

Un aspetto importante della statistica bayesiana è che il calcolo delle distribuzioni a posteriori spesso richiede tecniche numeriche avanzate. Questo perché, nella maggior parte dei casi, l'integrazione necessaria per calcolare $P(B)$ nel teorema di Bayes non è fattibile analiticamente. Uno dei metodi più utilizzati per affrontare questo problema è il metodo Markov chain Monte Carlo, che permette di approssimare la distribuzione a posteriori campionando iterativamente da essa.

### 1.6.1 Metodo Monte Carlo

Oltre a stimare parametri come la media e la varianza della distribuzione a posteriori, spesso si è interessati ad altri aspetti della distribuzione. Ad esempio, si potrebbe voler

calcolare $p(\theta \in A \mid y_1, ..., y_n)$ per un insieme arbitrario A, oppure stimare medie e deviazioni standard di una funzione di $\theta$, o ancora ottenere la distribuzione predittiva di dati mancanti o non osservati. In molti casi, il calcolo esatto di queste quantità è complesso o impossibile, ma possono essere approssimate con precisione arbitraria usando il metodo Monte Carlo. Consideriamo $\theta$ come un parametro di interesse e $y_1, ..., y_n$ come un campione osservato dalla distribuzione $p(y_1, ..., y_n \mid \theta)$. Se possiamo generare S campioni indipendenti $\{\theta^{(1)}, ..., \theta^{(S)}\}$ dalla distribuzione a posteriori $p(\theta \mid y_1, ..., y_n)$, allora la distribuzione empirica dei campioni $\{\theta^{(1)}, ..., \theta^{(S)}\}$ approssima la distribuzione a posteriori $p(\theta \mid y_1, ..., y_n)$. Questa approssimazione migliora all'aumentare di S ed è nota come approssimazione Monte Carlo di $p(\theta \mid y_1, ..., y_n)$. Se $g(\theta)$ è una qualsiasi funzione di $\theta$, la legge dei grandi numeri ci assicura che, per un campione sufficientemente grande, la media dei valori $g(\theta^{(s)})$, calcolata sui campioni $\{\theta^{(1)}, ..., \theta^{(S)}\}$, converge al valore atteso $E[g(\theta) \mid y_1, ..., y_n]$, cioè:

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^s) \longrightarrow E[g(\theta) \mid y_1, ..., y_n] = \int g(\theta)p(\theta \mid y_1, ..., y_n)d\theta, \tag{1.10}$$

per $S \longrightarrow \infty$.
Di conseguenza, qualsiasi aspetto della distribuzione a posteriori può essere approssimato con un campione Monte Carlo di dimensioni adeguate.

### 1.6.2 Markov chain Monte Carlo

Per la maggior parte degli esempi, tranne i casi più semplici, è complicato ottenere un campione Monte Carlo indipendente ed identicamente distribuito direttamente dalla distribuzione a posteriori. Questo avviene, ad esempio, quando il vettore dei parametri $\theta$ ha una dimensione elevata o quando la distribuzione a posteriori è una funzione di densità non standard, rendendo difficile il campionamento. Infatti, i metodi Monte Carlo assumono che la forma della distribuzione a posteriori sia nota, ma ciò non è sempre vero. In questi casi, si utilizza il metodo Markov Chain Monte Carlo, una procedura che unisce l'integrazione Monte Carlo con l'uso delle catene di Markov, noto come MCMC. L'idea è di costruire una catena di Markov che abbia come distribuzione stazionaria la distribuzione a posteriori. Vengono riportate di seguito alcune definizioni [1].

**Definizione 1.6.1 (Catena di Markov)** *Una collezione di variabili casuali dipendenti $\{X(0), X(1), ..., X(t), ...\}$, definite sullo spazio degli stati $\chi$, costituisce una catena di Markov se la distribuzione condizionale di $X(t)$, dato $X(0), X(1), ..., X(t-1)$, dipende solo dal valore immediatamente precedente $X(t-1)$ ed è indipendente dalle altre variabili passate. Grazie a questa proprietà, possiamo scrivere*

$$p(X(t) \mid X(0), X(1), ..., X(t-1)) = p(X(t) \mid X(t-1)),$$

*dove $p(X(t) \mid X(t-1))$ è comunemente chiamata probabilità di transizione.*

Inoltre, se $\pi$ rappresenta la distribuzione stazionaria, significa che $X(t) \sim \pi$ e anche $X(t+1) \sim \pi$; in altre parole, una volta che la catena raggiunge la distribuzione stazionaria $\pi$, la distribuzione marginale di $X(t)$ rimane invariata al variare di t.

**Definizione 1.6.2 (Irriducibilità)** *Una catena di Markov a stati discreti si dice irriducibile se, partendo da qualsiasi stato $s_i$, esiste una probabilità positiva di raggiungere qualsiasi altro stato $s_j$. Formalmente, una catena di Markov è irriducibile se:*

$$\forall s_i, s_j \in S, \forall m \in \mathbb{N}, \exists n \in \mathbb{N} : P(X_{m+n} = s_j \mid X_m = s_i) > 0.$$

Questo significa che, per ogni coppia di stati $s_i$ e $s_j$, c'è un numero finito di passi $n$ tale che la probabilità di passare da $s_i$ a $s_j$ sia maggiore di zero.

**Definizione 1.6.3 (Aperiodicità)** *Il periodo di uno stato $s_i \in S$ in una catena di Markov a stati discreti, con $S$ finito o numerabile, è il numero minimo di passi temporali necessari affinché ci sia una probabilità positiva di ritornare allo stato $s_i$ partendo da esso al tempo $t_m$. Formalmente, il periodo è*

$$d(s_i, t_m) = MCD\{n \geq 1 : P(X_{m+n} = s_i \mid X_m = s_i) > 0\},$$

*dove MCD rappresenta il massimo comune divisore.*
*Uno stato $s_i$ si definisce aperiodico se il suo periodo è uguale a 1 e una catena di Markov è detta aperiodica se tutti i suoi stati sono aperiodici.*

**Definizione 1.6.4 (Ricorrenza)** *Uno stato $s_i$ è detto ricorrente se, partendo da esso, la probabilità di ritornarvi in un numero finito di passi è pari a 1. Formalmente, uno stato $s_i$ è detto ricorrente se*

$$P(\textit{ritorno a } s_i \mid X_0 = s_i) = 1.$$

*In altre parole, lo stato verrà visitato infinite volte nel lungo periodo. Un insieme di stati è detto ricorrente se ogni stato al suo interno è ricorrente.*

L'esistenza di una distribuzione stazionaria unica richiede che la catena di Markov sia irriducibile, aperiodica e ricorrente.

## 1.7 Approccio frequentista o bayesiano?

L'approccio frequentista e quello bayesiano rappresentano due scuole di pensiero distinte in statistica, ciascuna con un'interpretazione diversa del concetto di probabilità e metodi differenti per l'inferenza statistica.

Nell'approccio frequentista, la probabilità è interpretata come una frequenza relativa a lungo termine. In altre parole, la probabilità di un evento è la proporzione di volte che questo evento si verifica su un numero elevato di ripetizioni di un esperimento. Ad esempio, se si lancia una moneta un numero infinito di volte, la probabilità frequentista di ottenere "testa" è il limite della proporzione di "testa" sui lanci totali. Questa interpretazione assume che i parametri sono fissi ma sconosciuti, e le probabilità riflettono la variabilità dovuta al campionamento. Le stime dei parametri e le decisioni inferenziali sono fondate su campioni teorici ripetuti all'infinito. Ad esempio, un intervallo di confidenza al 95% per un parametro significa che, se l'esperimento fosse ripetuto un numero infinito di volte, il 95% di questi intervalli conterrebbe il vero valore del parametro.

Nell'approccio bayesiano, la probabilità è vista come una misura della credenza o dell'incertezza rispetto a un evento o a un parametro. È una rappresentazione soggettiva della conoscenza, che può essere aggiornata man mano che nuove informazioni diventano disponibili. Ad esempio, la probabilità che una persona abbia una certa malattia può essere aggiornata con nuovi risultati medici, modificando la nostra conoscenza del caso specifico. I parametri del modello sono considerati variabili aleatorie, caratterizzate da distribuzioni di probabilità. Queste distribuzioni riflettono la nostra conoscenza iniziale (a priori) che viene aggiornata alla luce dei nuovi dati (a posteriori). L'inferenza consiste nel calcolare la distribuzione a posteriori dei parametri, che combina la distribuzione a priori e la verosimiglianza dei dati. La distribuzione a posteriori riflette la nostra conoscenza aggiornata sui parametri del modello dopo aver osservato i dati. Un intervallo di credibilità al 95% per un parametro significa che, dato i dati e la distribuzione a priori, c'è una probabilità del 95% che il parametro si trovi all'interno di quell'intervallo.

Entrambi gli approcci hanno i loro punti di forza e debolezza, e la scelta tra i due dipende spesso dal contesto applicativo e dagli obiettivi dell'analisi. In questa tesi, sfrutteremo la statistica bayesiana per svolgere le analisi che verranno presentate nei prossimi capitoli.

# Capitolo 2

# INLA

Per molto tempo, l'inferenza bayesiana si è affidata ai metodi Markov chain Monte Carlo (MCMC) per calcolare la distribuzione a posteriore congiunta dei parametri di un modello. Havard Rue, Martino e Chopin [12] hanno proposto un approccio innovativo che rende l'inferenza bayesiana più efficiente. Invece di stimare l'intera distribuzione a posteriore congiunta, suggeriscono di concentrarsi sulle distribuzioni marginali dei singoli parametri del modello. In molti casi, l'inferenza marginale è sufficiente per valutare i parametri e gli effetti latenti, evitando così la complessità delle distribuzioni a posteriori multivariate, spesso difficili da calcolare. Inoltre, il loro approccio si applica a modelli che possono essere descritti come campi casuali gaussiani di Markov (GMRF), il che comporta notevoli vantaggi computazionali e una riduzione dei tempi necessari per l'adattamento del modello. Hanno sviluppato una nuova tecnica basata sull'approssimazione di Laplace per calcolare le distribuzioni marginali a posteriori. Questo metodo, noto come INLA (Integrated Nested Laplace Approximation), si distingue per la sua efficienza computazionale e la capacità di affrontare un'ampia varietà di modelli, specialmente quelli appartenenti alla classe dei modelli gerarchici additivi gaussiani, che includono i GMRF. Rispetto ai metodi tradizionali come MCMC, che possono essere computazionalmente costosi per modelli complessi, INLA offre una soluzione più rapida e accurata per l'inferenza approssimata, consentendo di ottenere risultati affidabili con un carico computazionale ridotto. Per comprendere a fondo il funzionamento di INLA, è necessario avere familiarità con:

- inferenza bayesiana;
- modelli gaussiani latenti;
- campi casuali gaussiani di Markov;
- approssimazione di Laplace.

## 2.1 Modelli Gaussiani Latenti

I modelli gaussiani latenti (Latent Gaussian Models, LGM) rappresentano una classe di modelli statistici che utilizzano variabili latenti con distribuzione gaussiana per descrivere

la struttura dei dati osservati. Questi modelli sono particolarmente efficaci nel catturare relazioni complesse e dipendenze tra le variabili, rendendoli adatti a numerose applicazioni, tra cui la statistica spaziale. Un LGM si compone di tre elementi principali: la componente osservata, la componente latente e la componente degli iperparametri.

- La componente osservata (la likelihood) è definita come

$$y \mid x, \theta_1 \sim \prod_i p(y_i \mid x_i, \theta_1),$$

  dove le osservazioni $y$ sono considerate condizionatamente indipendenti dato il campo casuale gaussiano latente $x$ e gli iperparametri $\theta_1$.

- La componente latente $x$, modellata come un campo casuale gaussiano, è caratterizzata dalla distribuzione

$$x \mid \theta_2 \sim N(0, \Sigma(\theta_2)),$$

  che è un vettore multidimensionale con distribuzione normale, avente media zero e una matrice di covarianza $\Sigma(\theta_2)$. La scelta di $\Sigma(\theta_2)$ determina la struttura delle dipendenze tra le variabili latenti.

- Gli iperparametri $\theta = (\theta_1, \theta_2)^T$, distribuiti secondo una legge a priori

$$\theta \sim p(\theta),$$

  governano l'intero modello, influenzando sia la distribuzione delle variabili latenti che quella delle osservazioni. La distribuzione a priori può essere, ad esempio, una distribuzione gamma o normale, e il numero di iperparametri è ridotto, tipicamente da 2 a 5, ma non superiore a 20 [13].

In generale, la media $\mu_i$ dell'osservazione $y_i$ può essere espressa in termini di un predittore additivo strutturato $\eta_i$ mediante una funzione link $g(\cdot)$, tale che $g(\mu_i) = \eta_i$. Il predittore lineare additivo $\eta_i$ è definito come segue:

$$\eta_i = \alpha + \sum_j \beta_j z_{ij} + \sum_k f^k(u_{ik}), \tag{2.1}$$

dove $\alpha$ rappresenta un'intercetta generale e $z$ sono le covariate fisse con effetti lineari rappresentati dai coefficienti $\{\beta_j\}$. Il restante termine è la somma di funzioni che dipendono da un insieme di covariate $u$. Le funzioni $f^k(\cdot)$ possono assumere diverse forme, effetti lineari o non lineari delle covariate, trend temporali e stagionali, effetti casuali per intercette e pendenze, così come effetti spaziali e spaziotemporali strutturati. Questa flessibilità rende i modelli gaussiani latenti particolarmente versatili, in grado di rappresentare un'ampia varietà di modelli, dai modelli lineari generalizzati e dinamici a quelli spaziali e spaziotemporali. È importante notare che la formulazione del modello in 2.1 e i modelli gaussiani latenti appartengono alla stessa classe di modelli, quando si assume

una distribuzione a priori gaussiana per l'intercetta e i parametri degli effetti fissi. La distribuzione congiunta di

$$x = (\eta, \alpha, \beta, f^1, f^2, ...) \tag{2.2}$$

è quindi gaussiana e non singolare, se si aggiunge un piccolo termine di rumore all'equazione 2.1. È evidente che la dimensione di $x$ può diventare molto grande, poiché include il numero di osservazioni, l'intercetta, gli effetti fissi e la somma delle dimensioni di tutte le componenti del modello.

## 2.2 Campi casuali Gaussiani di Markov

I campi casuali gaussiani di Markov (Gaussian Markov Random Fields, GMRF) sono modelli statistici che descrivono strutture di dipendenza spaziale o temporale tra variabili. Un GMRF si caratterizza per due aspetti fondamentali: è un campo casuale, che rappresenta un insieme di variabili casuali dipendenti, ed è un campo di Markov, nel quale la dipendenza tra le variabili segue la proprietà di Markov. Questa proprietà implica che una variabile è condizionatamente indipendente da tutte le altre, tranne che da un sottoinsieme di variabili vicine. Esempi tipici di questa struttura sono i processi autoregressivi spaziali e temporali. Le variabili di un campo casuale gaussiano seguono una distribuzione normale multivariata, in cui ogni variabile $x_i$ è associata a una posizione nello spazio o nel tempo, e la dipendenza tra le variabili è descritta dalla matrice di precisione Q. Grazie alla proprietà di Markov, i GMRF risultano particolarmente efficienti dal punto di vista computazionale, poiché la matrice di precisione è sparsa, ovvero molti elementi della matrice sono pari a zero, indicando l'assenza di connessioni dirette tra variabili lontane. Inoltre, i GMRF sono molto utili per modellare una vasta gamma di fenomeni spaziali e temporali, come effetti casuali, errori di misura, e dipendenze strutturate nel tempo e nello spazio.

La matrice di precisione è l'inverso della matrice di covarianza $\Sigma$ e quindi, per un campo casuale gaussiano $x \sim N(0, \Sigma)$, si ha $Q = \Sigma^{-1}$. Tuttavia, l'inversione di una matrice densa è un'operazione computazionalmente costosa, con complessità $O(n^3)$, dove $n$ è il numero di variabili. Un aspetto cruciale dei GMRF è che, mentre ottenere matrici di covarianza sparsa richiederebbe forti ipotesi di indipendenza marginale tra le variabili,

$$x_i \perp x_j \iff \Sigma_{ij} = 0,$$

l'indipendenza condizionale (tipica della proprietà di Markov) è un'ipotesi molto più ragionevole e realistica. Rue e Held [11] hanno mostrato come le proprietà di indipendenza condizionale siano codificate nella matrice di precisione, e come queste possano essere sfruttate per migliorare i calcoli. Un risultato fondamentale che hanno ottenuto è che, per un GMRF,

$$x_i \perp x_j \mid x_{-ij} \iff Q_{ij} = 0.$$

Inoltre, l'indipendenza condizionale garantisce la possibilità di definire matrici di covarianza definite positive.
Questa struttura sparsa permette un uso efficiente di algoritmi di fattorizzazione di Cholesky, dove la matrice di precisione può essere scomposta come $Q = LL^T$, facilitando calcoli

veloci e stabili.
Nella costruzione di modelli additivi che includono i GMRF, una delle caratteristiche che ne rendono l'uso così efficiente, specialmente nell'approccio INLA, è che la distribuzione congiunta delle variabili latenti in 2.2 può essere considerata anch'essa come un GMRF [13]. La matrice di precisione congiunta è data dalla somma delle matrici di precisione delle diverse componenti del modello, incluse quelle degli effetti fissi e delle variabili latenti. Poiché questa distribuzione deve essere calcolata ripetutamente, dato che dipende dagli iperparametri $\theta$, la possibilità di trattarla come un GMRF con una matrice di precisione sparsa riduce drasticamente il carico computazionale.

## 2.3 Approssimazione di Laplace

L'approssimazione di Laplace è una tecnica usata per approssimare integrali complessi. Si basa sull'idea che molte funzioni possono essere ben approssimate da una distribuzione gaussiana attorno al loro massimo. Questo rende possibile effettuare calcoli più semplici e veloci, soprattutto nel contesto dell'inferenza bayesiana, dove gli integrali coinvolti possono essere particolarmente onerosi da risolvere esattamente.

Immaginiamo di voler calcolare il seguente integrale:

$$\int f(x)dx = \int \exp(\log(f(x)))dx, \tag{2.3}$$

dove $f(x)$ è la funzione di densità della variabile casuale X. Espandiamo $\log(f(x))$ utilizzando una serie di Taylor centrata in $x = x_0$:

$$\log(f(x)) \approx \log(f(x_0)) + (x - x_0) \left.\frac{\partial \log(f(x))}{\partial x}\right|_{x=x_0} + \frac{(x - x_0)^2}{2} \left.\frac{\partial^2 \log(f(x))}{\partial x^2}\right|_{x=x_0}. \tag{2.4}$$

Ponendo $x_0$ uguale al valore massimo $x^* = argmax_x \ \log(f(x))$, otteniamo che la derivata prima valutata in $x^*$ è nulla e quindi l'approssimazione in 2.4 si semplifica in:

$$\log(f(x)) \approx \log(f(x^*)) + \frac{(x - x^*)^2}{2} \left.\frac{\partial^2 \log(f(x))}{\partial x^2}\right|_{x=x^*}. \tag{2.5}$$

Quindi l'integrale di partenza si può approssimare nel modo seguente:

$$\begin{aligned} \int f(x)dx &\approx \int \exp\left(\log(f(x^*)) + \frac{(x - x^*)^2}{2} \left.\frac{\partial^2 \log(f(x))}{\partial x^2}\right|_{x=x^*}\right) dx \\ &= \exp(\log(f(x^*))) \int \exp\left(\frac{(x - x^*)^2}{2} \left.\frac{\partial^2 \log(f(x))}{\partial x^2}\right|_{x=x^*}\right) dx. \end{aligned} \tag{2.6}$$

Definendo $\sigma^{2*} = -1/ \left.\frac{\partial^2 \log(f(x))}{\partial x^2}\right|_{x=x^*}$ possiamo riscrivere l'integrale 2.6 come

$$\int f(x)dx \approx \exp(\log(f(x^*))) \int \exp\left(-\frac{(x - x^*)^2}{2\sigma^{2*}}\right) dx, \tag{2.7}$$

dove l'integranda rappresenta il kernel di una distribuzione normale con media $x^*$ e varianza $\sigma^{2*}$. In particolare,

$$\int_a^b f(x) \approx f(x^*)\sqrt{2\pi\sigma^{2*}}(\Phi(b) - \Phi(a)), \tag{2.8}$$

dove $\Phi$ esprime la funzione di ripartizione della normale con media $x^*$ e varianza $\sigma^{2*}$ [1].

## 2.4 The Integrated Nested Laplace Approximation

INLA (Integrated Nested Laplace Approximation) è un metodo per l'inferenza bayesiana nei modelli gerarchici additivi gaussiani, mirato all'approssimazione delle distribuzioni marginali. L'approccio si basa sull'integrazione delle seguenti distribuzioni marginali:

$$\begin{aligned} p(x_i \mid y) &= \int p(x_i \mid \theta, y)p(\theta \mid y)d\theta, \\ p(\theta_j \mid y) &= \int p(\theta \mid y)d\theta_{-j}. \end{aligned} \tag{2.9}$$

Il processo inizia calcolando un'approssimazione della distribuzione marginale degli iperparametri, indicata come $\tilde{p}(\theta \mid y)$, utilizzando l'approssimazione di Laplace. Successivamente, vengono scelti alcuni punti di valutazione per $\theta$ attorno al massimo della funzione di verosimiglianza, seguendo il metodo descritto da Rue et al. [12]. Per ciascun valore di $\theta$ selezionato, si procede quindi a calcolare l'approssimazione di Laplace della distribuzione condizionata $p(x_i \mid y, \theta)$. Le approssimazioni ottenute per i diversi valori di $\theta$ vengono successivamente combinate per produrre una stima più accurata della distribuzione marginale $p(x_i \mid y, \theta)$. Infine, si uniscono le approssimazioni $\tilde{p}(\theta \mid y)$ e $\tilde{p}(x_i \mid y, \theta)$ per ottenere la stima definitiva della distribuzione marginale $p(x_i \mid y)$. Questo passaggio implica l'integrazione numerica per tenere conto dell'incertezza sugli iperparametri $\theta$.
Le tipologie di approssimazione utilizzate per $p(x_i \mid \theta, y)$ sono le seguenti:

- L'approssimazione gaussiana, una delle forme più semplici di approssimazione. Si assume che la distribuzione condizionata $p(x_i \mid \theta, y)$ possa essere approssimata da una distribuzione normale, centrata attorno al massimo della verosimiglianza.
- L'approssimazione di Laplace, un metodo più sofisticato che migliora l'approssimazione gaussiana. Si basa sull'idea di espandere la funzione di verosimiglianza attorno al massimo trovato, considerando anche la forma della distribuzione.
- L'approssimazione semplificata di Laplace, una variante dell'approssimazione di Laplace che cerca di ridurre il costo computazionale mantenendo un buon livello di accuratezza. Si basa su una serie di espansioni per correggere l'approssimazione gaussiana.

Eravamo interessati anche all'approssimazione di $p(\theta_j \mid y)$, che può essere ricavata direttamente da $\tilde{p}(\theta \mid y)$ tramite integrazione numerica. Tuttavia, questo procedimento risulta oneroso dal punto di vista computazionale, poiché richiede la valutazione di $\tilde{p}(\theta \mid y)$

per molte configurazioni diverse. Un metodo più efficiente consiste nell'utilizzare i punti già selezionati in precedenza per costruire un'interpolante della funzione $\log(\tilde{p}(\theta \mid y))$, e poi calcolare le marginali attraverso l'integrazione numerica di tale interpolante.

Nonostante i suoi vantaggi, INLA ha alcune limitazioni:

- è particolarmente efficace per i modelli che rientrano nella classe dei GMRF, ma può non essere altrettanto efficiente per modelli che non seguono queste strutture condizionali;
- poiché è un metodo di approssimazione, esistono scenari in cui l'accuratezza delle approssimazioni potrebbe non essere sufficiente, specialmente per modelli con distribuzioni a posteriori altamente non gaussiane.

In conclusione, il nome di questo metodo deriva da tre componenti principali:

- Integrated: poiché si utilizza l'integrazione numerica per calcolare le distribuzioni marginali.
- Nested: poiché è necessario conoscere $p(\theta \mid y)$ per calcolare $p(x_i \mid y)$.
- Laplace Approximations: poiché viene impiegato il metodo di approssimazione di Laplace per ottenere i parametri per l'approssimazione normale.

# Capitolo 3

# Preferential sampling

La geostatistica è un ramo della statistica che si occupa dell'analisi e della modellazione di dati spaziali o georeferenziati, cioè dati che sono associati a posizioni geografiche. È utilizzata per studiare fenomeni che variano nello spazio e talvolta nel tempo, come la distribuzione di risorse naturali (petrolio, minerali, acqua), la qualità del suolo, o la concentrazione di inquinanti nell'ambiente. L'obiettivo principale della geostatistica è descrivere e predire come una certa variabile cambia in uno spazio continuo, basandosi su un insieme di misurazioni limitate prese in specifici punti. Questo consente di fare previsioni su aree non misurate, migliorare il processo decisionale e ottimizzare il campionamento. In alcuni casi, il processo di raccolta dei dati è influenzato dalla stessa variabile che si vuole misurare. In altre parole, la posizione o il momento in cui i dati vengono raccolti non è casuale, ma è correlato al fenomeno che si sta studiando. Questo tipo di campionamento è chiamato campionamento preferenziale. Per esempio, se immaginiamo di voler studiare la concentrazione di un inquinante in un fiume e scegliamo di misurare la concentrazione solo nei punti in cui si sospetta che l'inquinamento sia alto, ovvero vicino a una fabbrica, le osservazioni saranno condizionate da una conoscenza preesistente. Quindi, non stiamo campionando casualmente il fiume, ma stiamo preferendo i punti in cui si crede che l'inquinamento sia più elevato. Questo è un tipico caso di preferential sampling, dove il campionamento è influenzato dalla variabile di interesse, cioè il livello di inquinamento.
Formalmente, consideriamo un insieme di dati $y_i$ con $i = 1, ..., n$, ottenuti campionando un fenomeno spazialmente continuo $S(x)$ con $x \in \mathbb{R}^2$, dove $X$ è un insieme discreto di località $x_i$ con $i = 1, ...n$, all'interno di una regione spaziale di interesse $A \subset \mathbb{R}^2$. In questo scenario, si può assumere che i dati siano generati dal modello

$$Y_i = \mu + S(x) + Z_i, \quad i = 1, ..., n, \tag{3.1}$$

dove le $Y_i$ rappresentano il valore misurato nelle località $x_i$, $\mu$ è un effetto medio costante, $S(x)$ è un processo gaussiano stazionario con $E[S(x)] = 0$, e le $Z_i$ sono termini di errore mutuamente indipendenti con distribuzione $N(0, \tau^2)$. In molti casi, $S(x)$ non può essere misurato senza errori e quest'ultimi nei dati geostatistici sono tipicamente assunti come additivi. Una formulazione equivalente del modello 3.1 implica che, condizionatamente a $S(\cdot)$, le $Y_i$ siano mutuamente indipendenti con distribuzione normale:

$$Y_i \mid S(x_i) \sim N(\mu + S(x_i), \tau^2). \tag{3.2}$$

Nel modello standard 3.1 si trattano le posizioni di campionamento $x_i$ come fisse o comunque stocasticamente indipendenti dal processo $S(x)$, come accade comunemente nella letteratura geostatistica. In questo caso, la distribuzione congiunta di $S$, $X$ e $Y$ ha la struttura

$$p(S, X, Y) = p(S)p(X)p(Y \mid S(X)), \tag{3.3}$$

da cui è chiaro che per le inferenze su $S$ o $Y$ si può legittimamente condizionare su $X$ e utilizzare i metodi geostatistici classici. In questa circostanza, si parla di campionamento non preferenziale dei dati geostatistici. Al contrario, il campionamento preferenziale si riferisce a qualsiasi situazione in cui $p(S, X) \neq p(S)p(X)$. La definizione di preferential sampling proposta da Diggle et al. [3] implica una dipendenza stocastica tra il processo $S$ e il disegno di campionamento $X$. Quindi se $S$ e $X$ sono stocasticamente dipendenti e se $Y$ è osservato nei punti di $X$, la fattorizzazione appropriata della distribuzione congiunta di $S$, $X$ e $Y$ è

$$p(S, X, Y) = p(S)p(X|S)p(Y \mid S(X)). \tag{3.4}$$

Il preferential sampling può introdurre bias nei risultati e portare a conclusioni distorte. Poiché i dati raccolti non rappresentano in modo adeguato l'intera popolazione o il processo sottostante, le inferenze geostatistiche convenzionali sono quindi potenzialmente fuorvianti. Dunque, possiamo desumere che la natura stocastica di $X$ non può essere ignorata.

## 3.1 Il modello preferenziale

Per tenere conto del campionamento preferenziale, un approccio consiste nel costruire un modello congiunto utilizzando un modello log-gaussiano di Cox per le località e un modello per la variabile risposta $Y$ dei valori osservati. Il modello congiunto proposto da Diggle et al. [3] descrive il fenomeno $S(x)$, le posizioni di campionamento $X$ e i valori osservati $Y$, e può essere formulato come:

$$p(S, X, Y) = p(S)p(X \mid S)p(Y \mid S(X)). \tag{3.5}$$

In questa formulazione, $p(S)$ rappresenta la distribuzione del processo spaziale $S(x)$, che viene tipicamente modellato come un processo gaussiano stazionario con media zero e una funzione di covarianza che descrive la dipendenza spaziale tra i valori di $S(x)$ in punti diversi.
La componente $p(X \mid S)$ descrive la dipendenza delle posizioni di campionamento $X$ dal processo spaziale $S(x)$. Un modo comune per modellare questa dipendenza è tramite un processo di Poisson non omogeneo, in cui la densità delle posizioni di campionamento è legata al processo sottostante $S(x)$, seguendo la legge:

$$\lambda(x) = \exp\left(\alpha + \beta S(x)\right), \tag{3.6}$$

dove $\lambda(x)$ rappresenta la densità del campionamento in funzione della posizione $x$, e $\beta$ misura il grado di dipendenza tra il campionamento e il processo $S(x)$. Se $\beta = 0$ il campionamento è non preferenziale, ma casuale. Mentre se $\beta \neq 0$ le posizioni di campionamento

sono influenzate dai valori del processo spaziale, indicando campionamento preferenziale. In questo contesto, è più probabile che i campioni vengano raccolti in regioni dove i valori di $S(x)$ sono particolarmente elevati o bassi. Una generalizzazione dei processi di Poisson con intensità variabile e aleatoria, è il processo di Cox, il quale è un tipo di processo puntuale in cui l'intensità (cioè la densità media con cui si verificano gli eventi in uno spazio) non è fissa, ma essa stessa è un processo stocastico. Di conseguenza, è particolarmente adatto a modellare fenomeni in cui si sospetta che la distribuzione degli eventi sia influenzata da una struttura spaziale o temporale sottostante, o da altre variabili nascoste.
Infine, $p(Y \mid S(X))$ rappresenta la distribuzione dei dati osservati $Y$, condizionati sui valori del processo $S(x)$ nei punti di campionamento $X$. Di solito, questa parte del modello è un classico modello geostatistico, come quello 3.1, in cui i valori osservati $Y$ includono errori rispetto al processo spaziale $S(x)$ ed, eventualmente, un effetto medio costante.
In seguito, faremo riferimento al modello congiunto appena descritto con "modello preferenziale".

### 3.1.1 Log-gaussian Cox process

Un processo log-gaussiano di Cox (LGCP [10]) è un modello statistico utilizzato per descrivere processi puntuali spaziali o spazio-temporali. È un tipo di modello gerarchico, dove la densità del processo puntuale dipende da un processo latente, che a sua volta è modellato tramite un processo gaussiano. Questo modello è ampiamente utilizzato per rappresentare fenomeni dove si verificano eventi distribuiti nello spazio o nello spazio-tempo, ma con una densità variabile che è funzione di una struttura spaziale sottostante. Quindi un processo log-gaussiano di Cox è definito da un processo gaussiano latente e da un processo di Cox. In particolare, il logaritmo dell'intensità del processo di Cox segue la dinamica del processo gaussiano.
Nella prima componente del processo si assume l'esistenza di un processo gaussiano latente $S(x)$ su uno spazio continuo $x \in \mathbb{R}^d$, dove $S(x)$ è un campo aleatorio a valori reali e $d = 2$ per i dati spaziali bidimensionali. Questo processo è stazionario e completamente descritto dalla sua media e dalla sua covarianza, che definiscono le correlazioni spaziali tra i punti.
Mentre la seconda componente, dato il processo latente $S(x)$, descrive la densità di eventi che si verificano nello spazio. La densità $\lambda(x)$ del processo puntuale è determinata da un'esponenziale del processo gaussiano $S(x)$, ovvero

$$\lambda(x) = \exp(S(x)).$$

Questa funzione di intensità $\lambda(x)$ descrive la probabilità con cui gli eventi si verificano in diverse località $x$ dello spazio. In sostanza, nei punti $x$ dove il valore del campo gaussiano latente $S(x)$ è più alto, ci si aspetta una densità di eventi più alta.
Poiché l'intensità $\lambda(x)$ può variare in modo continuo nello spazio, il processo può facilmente modellare situazioni in cui gli eventi tendono a concentrarsi in alcune regioni. Alle aree con valori alti di $S(x)$ corrispondono regioni con alta densità di eventi.

## 3.2 Premesse

I seguenti paragrafi forniscono le basi necessarie per comprendere meglio e affrontare con maggiore rapidità l'analisi dell'applicazione del modello preferenziale.

### 3.2.1 Funzione di correlazione di Matérn

L'unico requisito affinché $\Sigma$ sia una funzione di covarianza valida per un processo gaussiano è che sia definita positiva. Questo significa che, per qualsiasi insieme di punti nello spazio o nel tempo, la matrice di covarianza risultante deve essere semidefinita positiva. Una funzione di covarianza ampiamente utilizzata nelle analisi spaziali è la covarianza di Matérn.
La funzione di correlazione di Matérn è una famiglia di funzioni utilizzata per modellare la dipendenza spaziale o temporale tra punti in un processo stocastico, particolarmente utile per descrivere la correlazione in un campo gaussiano. È molto flessibile e offre un controllo diretto sulla regolarità del processo tramite un parametro specifico $\nu$. La forma della funzione di correlazione di Matérn tra due punti distanti $u$ è:

$$\rho(u;\phi,\nu) = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{u}{\phi}\right)^{\nu} K_{\nu}\left(\frac{u}{\phi}\right), \qquad u > 0 \text{ [3]}. \tag{3.7}$$

Dove:

- $u$ è la distanza tra due punti;
- $\nu$ è il parametro di forma e se $\nu = 0.5$, si ottiene la funzione di correlazione esponenziale;
- $K_{\nu}(\cdot)$ è la funzione di Bessel modificata del secondo tipo, di ordine $\nu > 0$;
- $\phi$ è il parametro di scala che influenza il range spaziale della correlazione, cioè la distanza oltre la quale la correlazione tende a zero. Il range $r$ è pari a

  $$r = \frac{\sqrt{8*\nu}}{\kappa}, \text{ dove } \kappa = \frac{1}{\phi};$$

- $\Gamma(\cdot)$ è la funzione Gamma.

### 3.2.2 Variogramma

Il variogramma è uno strumento statistico utilizzato per analizzare la variabilità spaziale o temporale di un fenomeno. È largamente impiegato in geostatistica per descrivere come la dipendenza spaziale di un attributo (ad esempio, la concentrazione di piombo), cambia in funzione della distanza tra i punti di misurazione. Il variogramma è una funzione che mette in relazione la varianza della differenza tra i valori di una variabile in due punti con la distanza che li separa, ovvero

$$2\gamma(x_1 - x_2) = var(Z(x_1) - Z(x_2)).$$

In altre parole, ci dice quanto i valori di una variabile cambiano man mano che i punti di campionamento sono più distanti tra loro. Il concetto centrale è che punti più vicini tendono ad avere valori più simili, mentre punti più lontani sono generalmente meno correlati. Se i dati $Z(x_i)$ possono essere modellati con un processo stazionario con media costante, la funzione che descrive il variogramma è definita come:

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{i=1}^{N(h)} (Z(x_i) - Z(x_i + h))^2,$$

dove $|N(h)|$ è il numero di coppie distinte in

$$N(h) = \{(x_i, x_j) : x_i - x_j = h \ con \ i, j = 1, ..., n\} \ [2].$$

Il variogramma presenta diverse componenti fondamentali per la sua interpretazione. Il nugget, o nugget effect, corrisponde al valore del variogramma quando la distanza tra i punti è zero o molto ridotta, e rappresenta variazioni su scale inferiori alla minima distanza di campionamento, spesso dovute a rumore o errori di misurazione.
La soglia (o sill) è il valore massimo che il variogramma raggiunge man mano che aumenta la distanza tra i punti, indicando la varianza totale del fenomeno una volta che non esiste più correlazione tra i punti.
Il range, invece, è la distanza oltre la quale la correlazione spaziale si perde completamente; oltre questa distanza, il valore del variogramma rimane costante e pari al sill.
Infine, il concetto di anisotropia si applica quando il variogramma varia a seconda della direzione. In questo caso, la correlazione spaziale del fenomeno dipende non solo dalla distanza, ma anche dalla direzione lungo la quale viene misurata. Nell'applicazione trattata in questo lavoro, la correlazione spaziale dipende esclusivamente dalla distanza, il che indica che ci troviamo in un contesto isotropo.

Esistono diverse funzioni teoriche per descrivere un variogramma, a seconda del comportamento spaziale dei dati [2]. Le più comuni sono:

- lineare dove il valore del variogramma aumenta linearmente con la distanza;
- sferico dove il valore aumenta fino a raggiungere un plateau (il sill) dopo una certa distanza (il range);
- esponenziale dove il variogramma aumenta esponenzialmente, avvicinandosi al sill, ma non raggiungendolo mai;
- gaussiano dove il variogramma cresce lentamente inizialmente, ma aumenta rapidamente a distanze maggiori.

Il variogramma è spesso rappresentato graficamente, con la distanza tra i punti sull'asse delle ascisse e il valore del variogramma sull'asse delle ordinate. La forma della curva può fornire molte informazioni sulla struttura spaziale dei dati e sulla loro variabilità.

### 3.2.3 Approccio SPDE

Nella modellazione spaziale, i campi aleatori gaussiani (Gaussian Random Field, GRF) sono spesso utilizzati per descrivere fenomeni spazialmente correlati. Tuttavia, il metodo classico di rappresentazione dei GRF, basato sulla matrice di covarianza, ha un costo computazionale proibitivo quando il numero di punti nello spazio è molto grande. Questo rende difficile applicare tali metodi a problemi spaziali su larga scala o con domini complessi. Una soluzione più efficiente è offerta dall'approccio basato sulle Stochastic Partial Differential Equations (SPDE) proposto da Lindgren et al. [8]. Questo metodo innovativo consente di modellare campi aleatori spaziali, come i GRF, utilizzando equazioni differenziali alle derivate parziali stocastiche. L'idea di base è quella di:

- utilizzare un GRF su un insieme di località $x_i$, costruendo un GRF discretizzato con una matrice di covarianza $\Sigma$;
- identificare un Gaussian Markov Random Field con matrice di precisione $Q$, che approssimi al meglio il GRF, tale che $Q^{-1}$ sia il più vicino possibile a $\Sigma$ in qualche norma;
- effettuare i calcoli numerici usando la rappresentazione GMRF e sfruttare metodi efficienti per matrici sparse, consentendo l'uso dell'approccio INLA.

Lindgren et al. [8] dimostrano che un campo casuale gaussiano $S(x)$, definito su un dominio spaziale $A \subseteq \mathbb{R}^d$ e con una struttura di covarianza di Matérn, può essere approssimato dalla soluzione di un'equazione differenziale stocastica alle derivate parziali del tipo:

$$(\kappa^2 - \Delta)^{\alpha/2}(\tau S(x)) = W(x), \tag{3.8}$$

dove $\Delta$ è il Laplaciano, $\kappa$ è un parametro che controlla il range spaziale, $\alpha$ è un parametro che controlla la regolarità del campo, $\tau$ controlla la varianza e per ultimo $W(x)$ è un rumore bianco [1].
La soluzione esatta e stazionaria di questa SPDE è un campo gaussiano stazionario $S(x)$, con una funzione di covarianza di Matérn espressa come:

$$C(u;\phi,\nu,\sigma^2) = \sigma^2\rho(u;\phi,\nu) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}\left(\kappa u\right)^{\nu} K_{\nu}\left(\kappa u\right), \qquad u > 0.$$

Il legame tra i parametri della SPDE e quelli della covarianza di Matérn è dato dalle seguenti relazioni [1], che coinvolgono il parametro di smoothness $\nu$ e la varianza marginale $\sigma^2$:

$$\begin{aligned} \nu &= \alpha - \frac{d}{2} \\ \sigma^2 &= \frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}\kappa^{2\nu}\tau^2}. \end{aligned} \tag{3.9}$$

La soluzione della SPDE, rappresentata dal campo stazionario e isotropo di tipo Matérn $S(x)$, può essere approssimata numericamente tramite il metodo degli elementi finiti, utilizzando una base di funzioni su una triangolazione del dominio spaziale. Questo approccio porta a una rappresentazione del campo nella forma:

$$S(x) = \sum_{g=1}^{G} \tilde{S}_g \varphi_g(x),$$

dove $\varphi_g(x)$ sono le funzioni base e $\tilde{S}_g$ sono pesi distribuiti secondo una distribuzione gaussiana con media zero [1].

## 3.3 Applicazione del modello preferenziale al biomonitoraggio dei metalli pesanti in Galizia

L'applicazione che esamineremo in questa sezione riguarda il biomonitoraggio dell'inquinamento da piombo nella regione della Galizia, nel nord della Spagna, utilizzando campioni di muschio come bioindicatori [3]. Le concentrazioni di piombo sono state misurate in microgrammi per grammo di peso secco. Una prima indagine preliminare, condotta nel 1995 da Fernández et al. [4], aveva lo scopo di identificare le specie di muschio e i siti di raccolta più adeguati. Successivamente, sono stati realizzati due studi specifici sulla specie di muschio Scleropodium purum nel 1997 e nel 2000. I siti di campionamento per queste due indagini sono riportate nella Figura 3.1, dove si può osservare che alcune località sembrano trovarsi al di fuori della Galizia, questo è dovuto a un'interpretazione imprecisa del confine regionale sulla mappa, che tuttavia non influisce sull'analisi dei dati.

Nel 1997, il campionamento è stato intensificato nelle aree con previste variazioni significative nelle concentrazioni di piombo. Di conseguenza, il disegno di campionamento risultante era irregolare e potenzialmente preferenziale. Nel 2000, invece, è stato adottato un disegno di campionamento a reticolo regolare e non preferenziale, con alcune lacune dovute alla raccolta di specie di muschio diverse da quella presa in esame. Il nostro scopo è stimare e confrontare le mappe delle concentrazioni di piombo per il 1997 e il 2000.
Per cominciare nella Tabella 3.1 e nella Tabella 3.2 vengono riassunte le statistiche dei dati ottenuti in entrambi gli anni sia su scala non trasformata che su scala log-trasformata. Si nota una concentrazione media di piombo più alta nel 1997, risultato che può essere spiegato sia dal campionamento preferenziale vicino a fonti di inquinamento, sia da una riduzione complessiva dei livelli di piombo nei tre anni successivi [3]. Oltre alla diminuzione del valore medio della concentrazione di piombo, si ha una riduzione sia nei valori massimi che nei valori minimi. Inoltre, la trasformazione logaritmica dei dati ha eliminato una correlazione apparente tra varianza e media, rendendo le distribuzioni dei valori misurati più simmetriche (Figura 3.2). Infatti, guardando i dati nella Tabella 3.1, si può notare che ci sono forti differenze nella media e nella varianza tra i due anni e osservando i dati trasformati nella Tabella 3.2, la differenza tra i due anni diventa molto meno marcata.

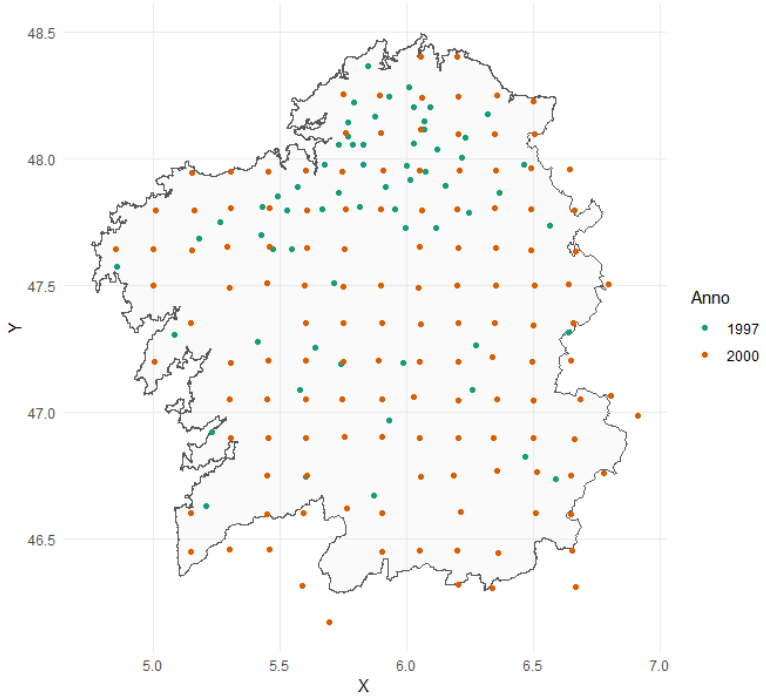



Figura 3.1: Campionamenti della concentrazione di piombo dal muschio del 1997 e del 2000.

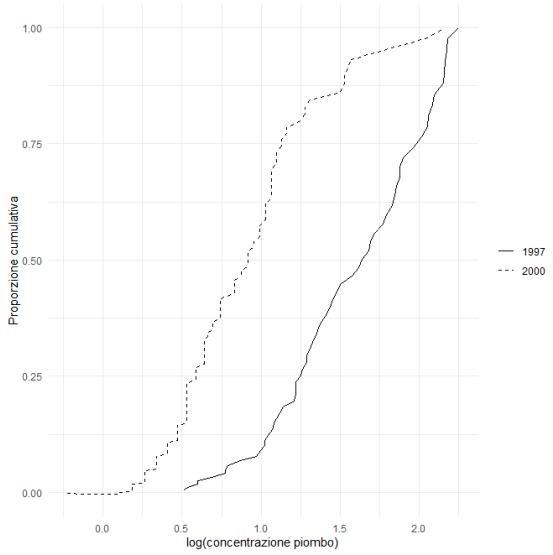



Figura 3.2: Distribuzioni cumulative dell'anno 1997 e del 2000.

| | **1997** | **2000** |
|---|---|---|
| Numero di siti | 63 | 132 |
| Media | 4.72 | 2.15 |
| Deviazione standard | 2.21 | 1.18 |
| Valore minimo | 1.67 | 0.80 |
| Valore massimo | 9.51 | 8.70 |

Tabella 3.1: Statistiche di sintesi per i livelli di inquinamento da piombo misurati nel 1997 e nel 2000 per la scala non trasformata.

| | **1997** | **2000** |
|---|---|---|
| Numero di siti | 63 | 132 |
| Media | 1.44 | 0.66 |
| Deviazione standard | 0.48 | 0.43 |
| Valore minimo | 0.52 | -0.22 |
| Valore massimo | 2.25 | 2.16 |

Tabella 3.2: Statistiche di sintesi per i livelli di inquinamento da piombo misurati nel 1997 e nel 2000 per la scala log-trasformata.

Per un'analisi preliminare, abbiamo sfruttato il modello gaussiano standard 3.1, assumendo che il campo latente $S(x)$ segua un processo gaussiano stazionario con media zero, varianza $\sigma^2$ e una funzione di correlazione di Matérn $\rho(u;\phi,\nu)$. Ricordiamo anche che gli errori di misura sono stati modellati come variabili gaussiane, $Z_i \sim N(0,\tau^2)$. Questo modello è stato applicato separatamente ai dati raccolti nel 1997 e nel 2000.
Confrontando i valori nelle Tabelle 3.3 e 3.4, si può vedere che il valore stimato per $\mu$ è più alto nell'anno del 1997, come già noto dalle statistiche di sintesi dei dati a disposizione. La stima per $\log(\tau)$ è più bassa nel 2000 e questo suggerisce che la variabilità non strutturata dei dati è minore in quell'anno. Tuttavia, per altri parametri come il range, $\log(\phi)$ e $\log(\sigma)$, le differenze sono minime e gli intervalli di credibilità si sovrappongono, indicando che queste differenze potrebbero non essere particolarmente significative e quindi suggerendo una struttura spaziale relativamente simile nei due anni.

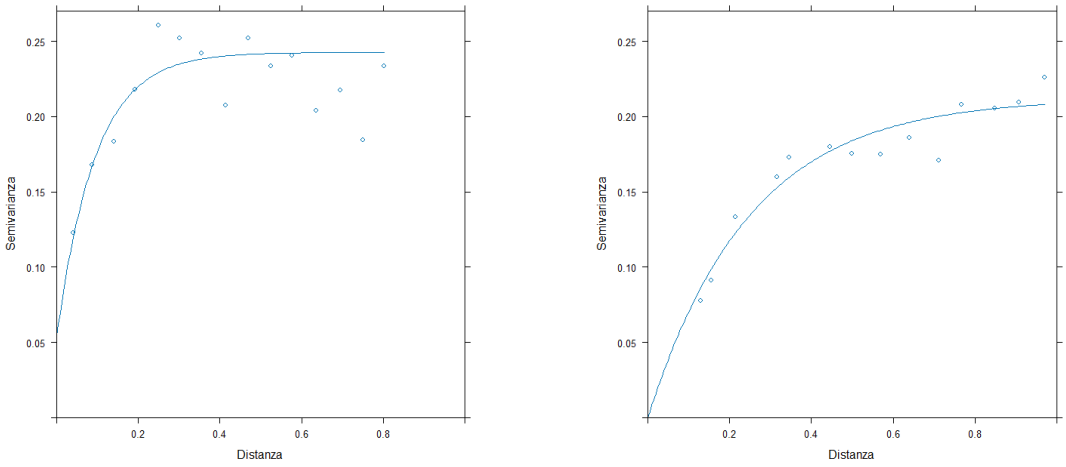
Inoltre, è possibile analizzare la distribuzione spaziale della concentrazione di piombo e la sua variabilità nei due anni, utilizzando i variogrammi empirici e teorici relativi alla concentrazione di piombo del 1997 (Figura 3.3a) e del 2000 (Figura 3.3b), dai quali emergono alcune osservazioni importanti.

| | **Media** | **Quantile 0.025** | **Quantile 0.975** |
|---|---|---|---|
| $\mu$ | 1.532 | 1.254 | 1.832 |
| $\log(\tau)$ | -1.186 | -1.482 | -0.828 |
| range | 0.499 | 0.090 | 1.654 |
| $\log(\phi)$ | -1.665 | -3.091 | -0.191 |
| $\log(\sigma)$ | -0.886 | -1.314 | -0.462 |

Tabella 3.3: Stime delle medie ed intervalli di credibilità dei parametri del modello standard relativo ai dati del 1997.

| | **Media** | **Quantile 0.025** | **Quantile 0.975** |
|---|---|---|---|
| $\mu$ | 0.704 | 0.543 | 0.873 |
| $\log(\tau)$ | -4.997 | -5.679 | -3.634 |
| range | 0.259 | 0.160 | 0.394 |
| $\log(\phi)$ | -2.071 | -2.523 | -1.624 |
| $\log(\sigma)$ | -0.662 | -0.781 | -0.541 |

Tabella 3.4: Stime delle medie ed intervalli di credibilità dei parametri del modello standard relativo ai dati del 2000.




(a) 1997 (b) 2000

Figura 3.3: Variogrammi empirici e teorici per i dati di concentrazione di piombo del 1997 (a) e del 2000 (b), su scala log-trasformata.

Si può notare che il range è leggermente maggiore per il 2000 rispetto al 1997 e che il sill (cioè $\sigma^2 + \tau^2$) è simile in entrambi i variogrammi, con valori compresi tra 0.21 e 0.25. Nel variogramma stimato dai dati del 2000 non è presente la componente nugget $\tau^2$, questo parametro è scarsamente identificato a causa della disposizione a reticolo del

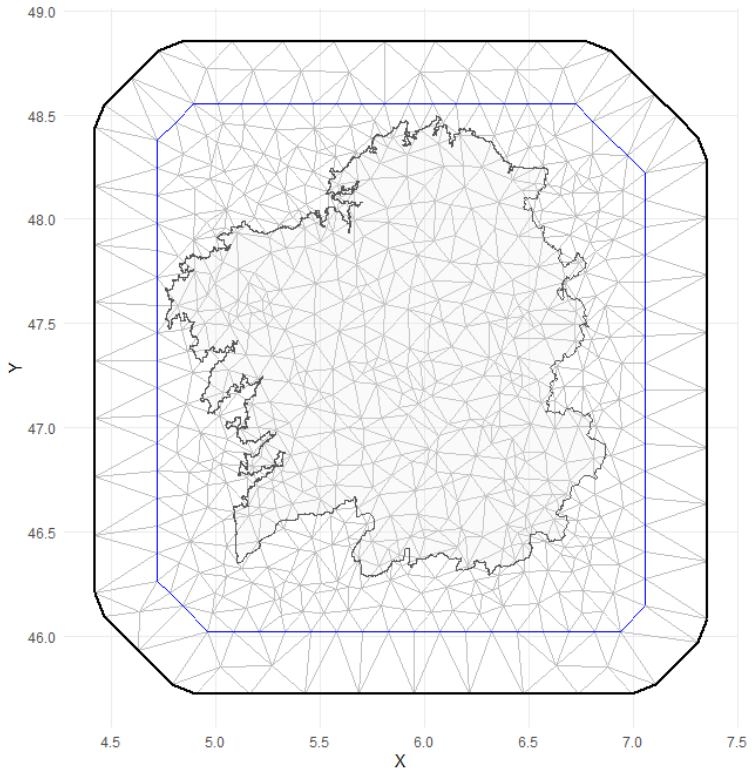



Figura 3.4: Mesh con confine della regione Galizia.

disegno di campionamento di quell'anno [3]. Nel complesso, queste osservazioni, insieme a quelle ottenute applicando il modello 3.1 separatamente per il 1997 e il 2000, provano che un modello congiunto per le due serie di dati potrebbe permettere di avere in comune alcuni parametri tra i due anni. La condivisione dei parametri $\sigma$, il range e $\tau$ consentirebbe al modello congiunto di stimarli in modo più efficiente [3].

### 3.3.1 Implementazione del modello

La prima fase dell'implementazione riguarda la costruzione di una mesh triangolare per rappresentare il dominio spaziale. Questa mesh consente di discretizzare lo spazio in un numero finito di nodi e triangoli, che costituiscono la base per il calcolo del campo spaziale. Per evitare effetti di bordo, la mesh è estesa oltre il dominio dei dati osservati. Nella Figura 3.4 si può osservare la mesh usata per la stima del campo spaziale.

In questo modello è stato usato l'approccio SPDE, questo metodo permette di rappresentare il campo spaziale come una combinazione lineare dei valori nei nodi della mesh. Questo riduce notevolmente la complessità del problema rispetto ai classici approcci geostatistici, rendendo l'inferenza computazionalmente fattibile anche per dataset spaziali di grandi dimensioni.
Nel modello è necessario specificare le prior utilizzate per ogni parametro. I parametri

| | **Media** | **Quantile 0.025** | **Quantile 0.975** |
|---|---|---|---|
| $\mu_{97}$ | 1.546 | 1.219 | 1.874 |
| $\mu_{00}$ | 0.739 | 0.459 | 1.037 |
| $\log(\tau)$ | -1.329 | -1.506 | -1.132 |
| range | 0.641 | 0.342 | 1.118 |
| $\log(\phi)$ | -1.183 | -1.765 | -0.582 |
| $\log(\sigma)$ | -0.896 | -1.130 | -0.662 |

Tabella 3.5: Stime delle medie ed intervalli di credibilità dei parametri del modello standard relativo ai dati del 1997 e del 2000.

fissi e il parametro $\beta$ seguono una distribuzione normale con media 0 e precisione 0.001. Mentre, per la precisione delle osservazioni gaussiane è stata impiegata una distribuzione log-gamma con parametro di forma pari a 1 e parametro di scala pari a $5 \times 10^{-5}$.

### Analisi classica

Oltre ad aver applicato il modello standard separatamente per gli anni 1997 e 2000, poiché possiamo assumere un modello congiunto per entrambi gli anni, possiamo applicare lo stesso modello standard considerando i due anni insieme. Questo approccio ci permette di sfruttare l'informazione combinata dei due periodi temporali, migliorando potenzialmente la stima dei parametri e la predizione del fenomeno studiato. Nella Tabella 3.5, si osserva che la media stimata per $\mu_{97}$ è significativamente più alta rispetto a $\mu_{00}$, questo risultato è in linea con le analisi precedentemente condotte sui dati, confermando che la concentrazione media di piombo nel 1997 era superiore a quella del 2000. Inoltre, gli intervalli di credibilità del range, di $\log(\phi)$ e di $\log(\sigma)$ sono contenuti nei rispettivi intervalli di credibilità ottenuti applicando il modello standard ai dati relativi al 1997 (Tabella 3.3) e si intersecano con quelli trovati considerando i dati del 2000 (Tabella 3.4). Questo rafforza l'idea che l'uso di un modello congiunto sia appropriato per catturare la struttura spaziale condivisa dei dati.

### Analisi preferenziale

Questa analisi considera la natura preferenziale del campionamento, il quale dipende stocasticamente dal campo gaussiano latente. Pertanto, utilizzeremo il modello preferenziale descritto nel paragrafo 3.1, applicandolo congiuntamente alle due serie di dati e trattando i dati del 1997 come campionati in modo preferenziale, mentre quelli del 2000 come campionati in modo non preferenziale. In particolare, i dati del 1997 sono descritti dai processi:

$$\begin{aligned} Y_{97}|X_{97}, S &\sim N(\mu_{97} + S, \tau^2), \\ X_{97}|S &\sim PP(\exp(\alpha + \beta S)); \end{aligned} \tag{3.10}$$

dove $X_{97}|S$ è un processo puntuale, nello specifico è un processo puntuale di Cox, ed $S$ è un campo casuale gaussiano.

| | **Media** | **Quantile 0.025** | **Quantile 0.975** |
|---|---|---|---|
| $\mu_{97}$ | 1.734 | 1.384 | 2.110 |
| $\mu_{00}$ | 0.743 | 0.417 | 1.087 |
| $\log(\tau)$ | -1.167 | -1.366 | -0.968 |
| range | 0.851 | 0.417 | 1.597 |
| $\log(\phi)$ | -1.914 | -1.567 | -0.225 |
| $\log(\sigma)$ | -1.030 | -1.358 | -0.699 |
| $\beta$ | -2.011 | -3.120 | -0.926 |

Tabella 3.6: Stime delle medie ed intervalli di credibilità dei parametri del modello preferenziale relativo ai dati del 1997 e del 2000.

Invece, i dati del 2000 seguono, semplicemente, il seguente processo:

$$Y_{00}|S \sim N(\mu_{00} + S, \tau^2). \tag{3.11}$$

Le stime trovate applicando il modello preferenziale sono riportate nella Tabella 3.6 e si nota che la media di $\mu_{97}$ continua ad essere più alta di quella di $\mu_{00}$. Il parametro del range ha una media di 0.851, con un intervallo di credibilità piuttosto ampio che contiene l'intervallo di credibilità del range usando il modello standard (Tabella 3.5). Mentre, gli intervalli di credibilità di $\log(\phi)$, $\log(\sigma)$ e $\log(\tau)$, si sovrappongono parzialmente con i rispettivi intervalli di credibilità nella Tabella 3.5. La stima del parametro $\beta$, che rappresenta l'effetto del campo latente $S$ sul campionamento preferenziale, è negativa, con un valore pari a -2.011. Questo risultato è controintuitivo, perché la metà settentrionale della regione sovracampionata è più industrializzata della metà meridionale sottocampionata [3]. La stima di $\beta$ è inoltre strettamente legata alla differenza tra i valori medi stimati per il 1997 e il 2000. Il fatto che il livello medio di inquinamento osservato è sostanzialmente più alto nel 1997 che nel 2000, porterebbe a interpretare il campionamento del 1997 come preferenziale, con un valore positivo di $\beta$. Invece, se dividiamo i dati del 1997 in due gruppi in base alla loro posizione geografica, osserviamo che i livelli di inquinamento osservati nei 47 punti nella parte settentrionale sono più bassi, con media delle concentrazioni in scala logaritmica pari a 1.38 e deviazione standard di 0.49. Mentre, nei restanti 16 punti nella metà meridionale, la media è di 1.62 con deviazione standard uguale a 0.40. Questa osservazione è coerente con un valore stimato di $\beta$ negativo [3].

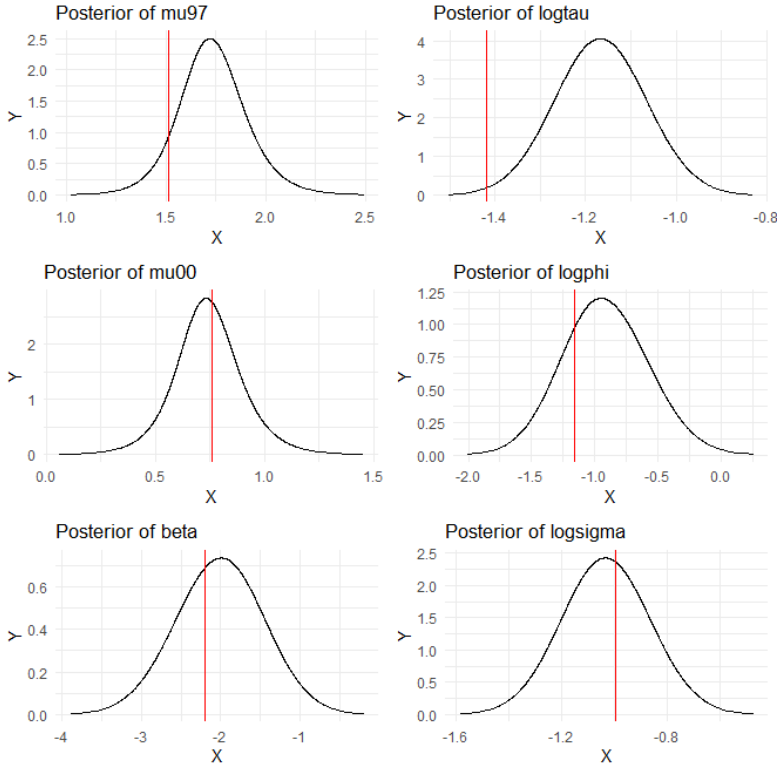



Figura 3.5: Distribuzioni a posteriori dei parametri, dove le linee verticali rappresentano le medie stimate nel lavoro di Diggle et al. [3].

Le medie stimate dei parametri nel lavoro di Diggle et al. [3], trovate tramite la massima verosimiglianza e il metodo Monte Carlo, sono contenute nei rispettivi intervalli di credibilità che sono stati ottenuti applicando il modello preferenziale (Figura 3.5).

# Capitolo 4

# Simulazioni e risultati

## 4.1 Struttura delle simulazioni

Le simulazioni sono state condotte generando un campo casuale e cercando di ricostruire il campo originale dalle osservazioni in alcuni punti, usando il software R.

Inizialmente, è stata creata una mesh ad alta risoluzione (vedi Figura 4.3) per il campo reale in una regione $[0,1]^2$. Su questa mesh è stato costruito un modello spaziale di Matérn, dal quale è stato generato un campione del campo casuale $S$. Successivamente, sono stati selezionati casualmente $N = 1000$ punti, a ciascuno dei quali è stata associata una variabile risposta $Y_i$, calcolata come $Y_i = \mu + S(x_i) + \epsilon_i$ con $i = 1, ..., N$, che nel caso dell'applicazione vista nel capitolo precedente, rappresentava la concentrazione di piombo nel muschio nei diversi punti. A partire da questo dataset, sono stati poi effettuati due tipi di campionamento senza reimmissione: un campionamento casuale e un campionamento preferenziale. Quest'ultimo utilizza una funzione di probabilità data da

$$p(S(x)) = \frac{\exp(\beta S(x))}{1 + \exp(\beta S(x))}, \tag{4.1}$$

dove $S(x)$ è il valore che il campo può assumere e $\beta$ controlla la pendenza della funzione, ovvero per valori maggiori la transizione tra 0 e 1 è più rapida.
La Figura 4.1 mostra i grafici di questa funzione di probabilità in relazione ai valori del campo 1 (Figura 4.4a), considerando tutti i valori di $\beta$. Analogamente, la Figura 4.2 presenta i grafici della funzione di probabilità per i valori del campo 2 (Figura 4.4b), sempre per tutti i valori di $\beta$. Entrambi i campi spaziali verranno presentati successivamente nel capitolo, insieme agli altri campi considerati nelle simulazioni. Tuttavia, anche le funzioni di probabilità associate agli altri campi delle simulazioni mostrano comportamenti simili. Quando $\sigma = 1.5$, il comportamento della funzione di probabilità ricorda quello osservato per il campo 1; mentre, per $\sigma = 0.371$, esso rispecchia quello del campo 2. In generale, si osserva che un valore elevato di $\sigma$ rende la funzione di probabilità più estrema rispetto ad un valore basso.

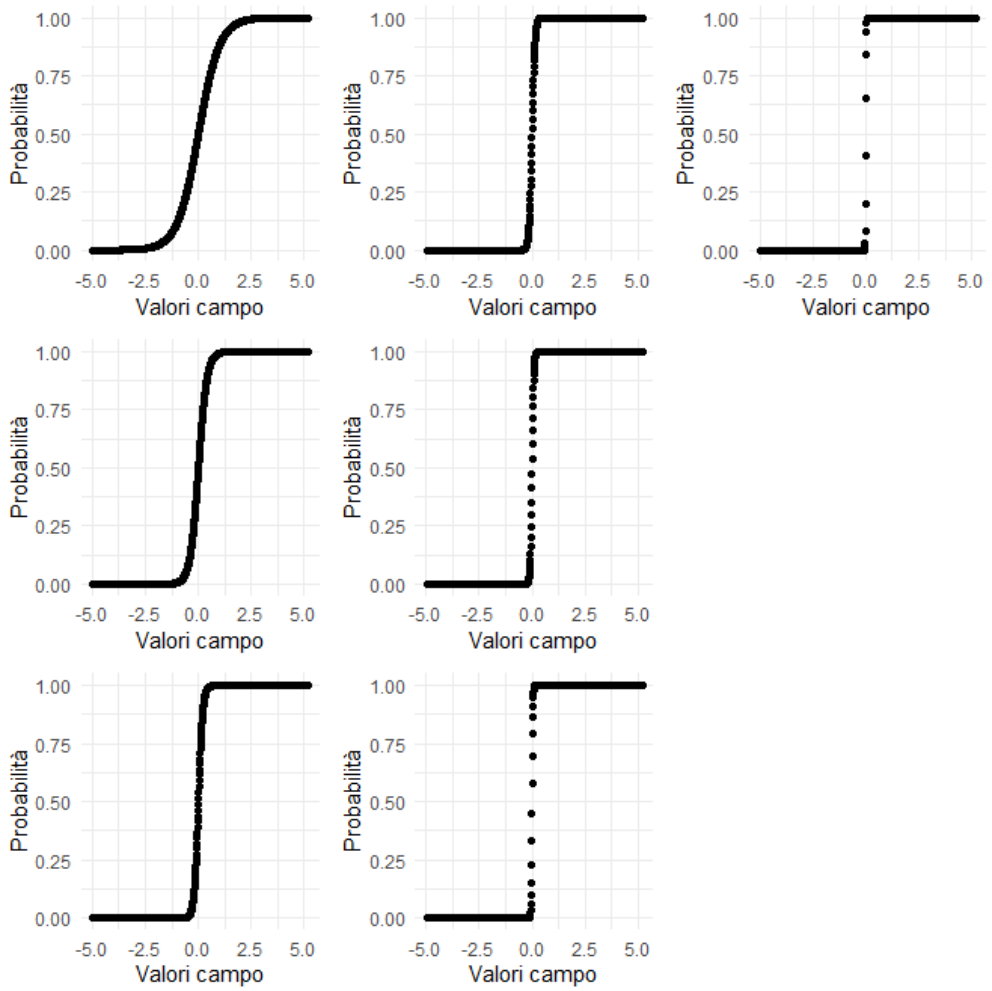



Figura 4.1: Grafici della funzione di probabilità in relazione ai valori del campo generato nella simulazione 1, con i rispettivi valori di $\beta$: nella prima colonna sono riportati $\beta = 2$, 5 e 10; nella seconda colonna $\beta = 15$, 25 e 50; infine, nell'ultima colonna, $\beta = 100$.

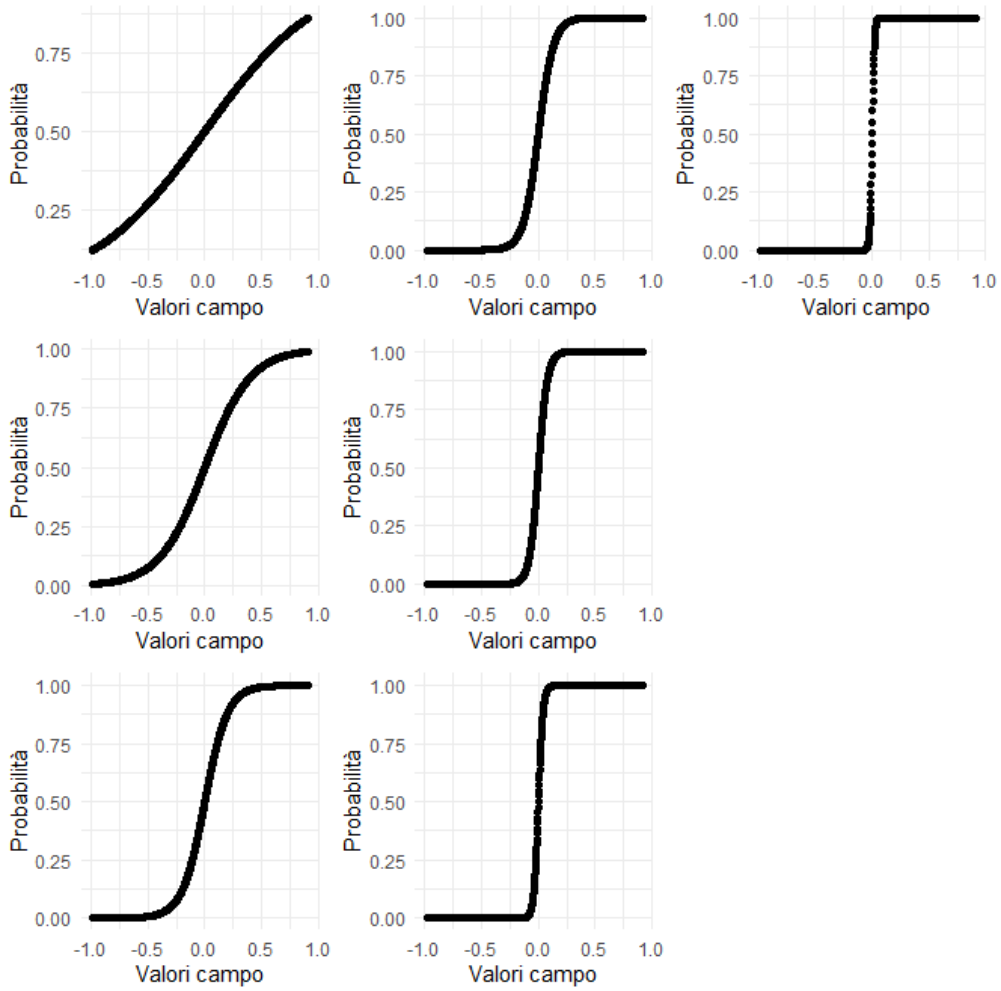



Figura 4.2: Grafici della funzione di probabilità in relazione ai valori del campo generato nella simulazione 2, con i rispettivi valori di $\beta$: nella prima colonna sono riportati $\beta = 2$, 5 e 10; nella seconda colonna $\beta = 15$, 25 e 50; infine, nell'ultima colonna, $\beta = 100$.

Dai due campionamenti si ottengono due dataset che verranno usati per stimare il campo $S$. Inoltre, per entrambi i tipi di campionamento, sono state considerate diverse numerosità campionarie, rispettivamente di 30, 50 e 100 punti.

Una volta definiti i diversi dataset in base al tipo di campionamento e alla numerosità campionaria, è necessario costruire una nuova mesh per la stima del campo $S$. Si può osservare nella Figura 4.3 che questa mesh è meno fine rispetto a quella utilizzata per generare il campo spaziale e che si estende un po' oltre il dominio iniziale. Analogamente a quanto fatto nell'applicazione precedente, anche in queste simulazioni è stato adottato un approccio SPDE e sono state utilizzate le stesse prior. I parametri fissi e il parametro $\beta$ seguono una distribuzione normale con media pari a 0 e precisione pari a 0.001. Per

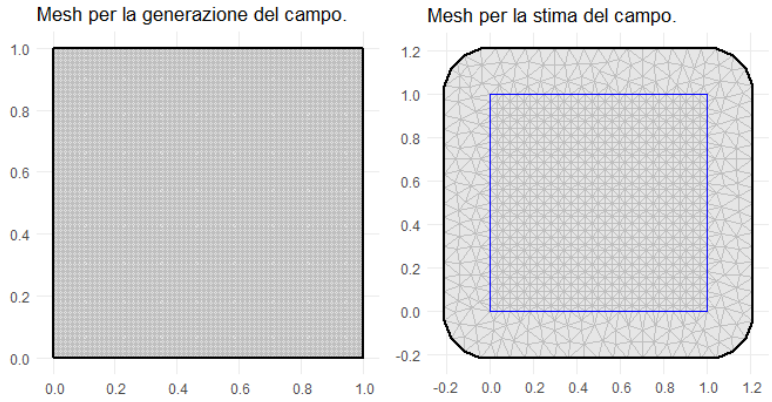



Figura 4.3: Mesh utilizzate rispettivamente per la generazione e per la stima del campo spaziale $S$.

la precisione delle osservazioni gaussiane, è stata invece impiegata una distribuzione log-gamma con parametro di forma 1 e parametro di scala $5 \times 10^{-5}$. Infine, sono stati applicati ai vari dataset sia il modello standard che quello preferenziale.

### 4.1.1 Scelta dei parametri

Per partire con le simulazioni, è necessario impostare i parametri per la generazione del campo casuale $S$ e per la creazione dei dataset. Nella generazione del campo di Matérn, i parametri fondamentali sono: il range, la deviazione standard $\sigma$ e il parametro di smoothness $\nu$. Sono stati scelti tre valori per il range: uno elevato, pari a 0.626, uno medio, pari a 0.4, e uno ridotto, pari a 0.2. Per la deviazione standard $\sigma$ sono stati scelti due valori: uno alto, pari a 1.5, e uno basso, pari a 0.371. Infine, il parametro $\nu$ è stato impostato a 0.5, generando così una funzione di correlazione di tipo esponenziale. Per quanto riguarda la formazione del dataset, il parametro $\mu$ è stato fissato a 0 senza perdita di generalità, poiché un valore diverso produrrebbe semplicemente una traslazione dei dati. Per la deviazione standard $\tau$ delle osservazioni del campo $S$ sono stati considerati un valore alto pari a 0.243 e uno basso pari a 0.05. Si può notare che alcuni valori, rispettivamente 0.626,

0.371 e 0.243, sono molto particolari, questo perché la loro combinazione simula le condizioni che si trovano nel contesto dell'applicazione della concentrazione di piombo nella Galizia. Infine, il parametro $\beta$ assume i valori 2, 5, 10, 15, 25, 50 e 100, per esaminare l'influenza della pendenza della funzione di probabilità 4.1 sul campionamento e determinare i casi in cui il modello preferenziale offre un vantaggio nei campionamenti non casuali.

## 4.2 Risultati

Prima di analizzare i risultati delle simulazioni effettuate, precisiamo la notazione che utilizzeremo. Ad esempio, useremo $\mu_{30}$ per indicare il parametro $\mu$ relativo a una numerosità campionaria pari a 30. Più in generale, denoteremo con $\mu_n$ il parametro $\mu$ per una numerosità campionaria $n$, con $n = 30, 50, 100$.

### 4.2.1 Campionamento casuale

Un primo scenario delle simulazioni prevede il campionamento casuale dei punti, applicando sia il modello standard sia il modello preferenziale.

In questo contesto, gli intervalli di credibilità relativi a tutti i parametri sono riportati nelle Figure presenti nell'Appendice A. Nel caso del modello preferenziale, l'analisi degli intervalli di credibilità stimati per il parametro $\beta$ mostra che il valore zero rientra in essi. Questo risultato è coerente con il campionamento casuale, poiché un valore di $\beta = 0$ rappresenta, appunto, un campionamento non preferenziale. In particolare, considerando le diverse numerosità campionarie, si osserva che il valore zero è contenuto nell'intervallo di credibilità di $\beta_{30}$ e $\beta_{100}$ nel 99% delle simulazioni, mentre l'intervallo di $\beta_{50}$ include sempre lo zero (Figure A.9, A.18 e A.27).
Al contrario, un parametro che risulta più difficile da stimare correttamente con entrambi i modelli è $\tau$. Con il modello standard, l'intervallo di credibilità per $\tau_{30}$ contiene il valore vero nel 5% delle simulazioni, per $\tau_{50}$ nel 14%, e per $\tau_{100}$ nel 49%. Utilizzando il modello preferenziale, le percentuali sono simili: rispettivamente 4%, 15% e 49%. L'aumento della numerosità campionaria migliora le stime medie, aumentando la frequenza con cui gli intervalli di credibilità contengono il valore vero. Infatti, la percentuale per cui l'intervallo di credibilità di $\tau_{100}$ contiene il valore vero è la più alta (Figure A.2, A.6, A.11, A.15, A.20 e A.24).
Per gli altri parametri, le stime medie e gli estremi degli intervalli di credibilità risultano comparabili (Figure A.1, A.5, A.10, A.14, A.19 e A.23 per $\mu$; Figure A.3, A.7, A.12, A.16, A.21 e A.25 per il range; Figure A.4, A.8, A.13, A.17, A.22 e A.26 per $\sigma$).

Per commentare e valutare le mappe dei campi stimati, sono state prodotte immagini che rappresentano la differenza al quadrato tra il campo vero generato e il campo stimato. In questo contesto di campionamento casuale, si osserva che le figure ottenute applicando il modello standard sono comparabili a quelle generate quando impieghiamo il modello

| | **Modello standard** | **Modello preferenziale** |
|---|---|---|
| $\mu_{30}$ | 32% | 76% |
| $\mu_{50}$ | 37% | 71% |
| $\mu_{100}$ | 38% | 74% |
| $\tau_{30}$ | 10% | 23% |
| $\tau_{50}$ | 38% | 38% |
| $\tau_{100}$ | 60% | 61% |
| $range_{30}$ | 80% | 77% |
| $range_{50}$ | 73% | 69% |
| $range_{100}$ | 69% | 75% |
| $\sigma_{30}$ | 82% | 73% |
| $\sigma_{50}$ | 80% | 75% |
| $\sigma_{100}$ | 74% | 75% |

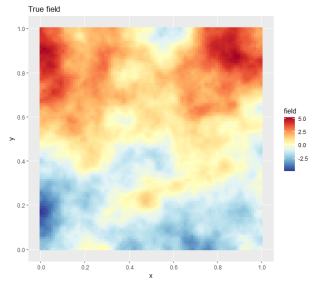
Tabella 4.1: Percentuali delle simulazioni in cui gli intervalli di credibilità contengono il valore vero, differenziando tra i due modelli.

preferenziale. Analogamente a quanto avviene per gli intervalli di credibilità, le differenze al quadrato tra i campi non presentano variazioni significative quando il valore di $\tau$ è pari a 0.243 o 0.05. Inoltre, si nota che quando $\sigma$ è 1.5, i valori delle differenze al quadrato risultano più alti rispetto a quelli osservati con $\sigma$ pari a 0.371.

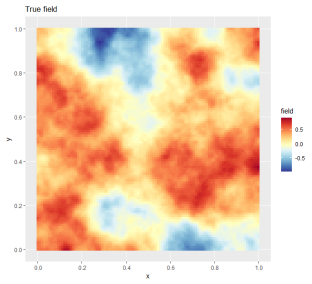
Possiamo quindi concludere che, in caso di campionamento casuale, i due modelli producono risultati simili, rendendo indifferente quale modello applicare.

### 4.2.2 Campionamento preferenziale

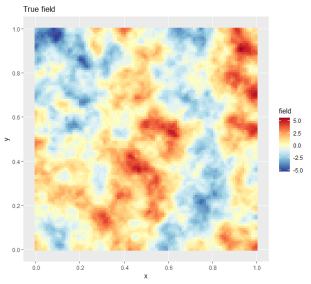
I due modelli sono stati testati anche in uno scenario di campionamento preferenziale, il quale segue la funzione di probabilità 4.1, definita in precedenza. In questa circostanza, la Tabella 4.1 mostra come variano le percentuali di simulazioni in cui il valore vero dei parametri è contenuto nei rispettivi intervalli di credibilità, in funzione del numero di punti campionati e del modello utilizzato. Il modello preferenziale si dimostra superiore al modello standard nella stima del parametro $\mu$, con percentuali significativamente più alte. Sebbene le prestazioni migliorino leggermente all'aumentare della numerosità campionaria $n$, il modello standard continua a essere meno efficace rispetto al preferenziale per questo parametro.
Come nel primo scenario, dove il campionamento è casuale, entrambi i modelli mostrano difficoltà nella stima di $\tau$, con percentuali molto simili. Tuttavia, le prestazioni migliorano con l'aumento della numerosità campionaria, indicando una maggiore accuratezza per campioni più grandi.
Le prestazioni dei due modelli per il parametro range e $\sigma$ sono comparabili, senza evidenti differenze tra i due approcci.
In questa seconda situazione, gli intervalli di credibilità per i parametri sono disponibili



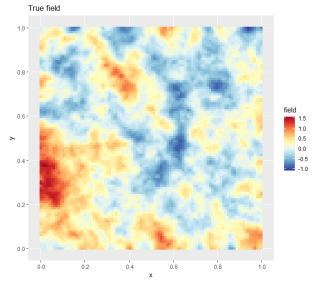

(a) Campo 1: *range* = 0.626 e $\sigma = 1.5$.



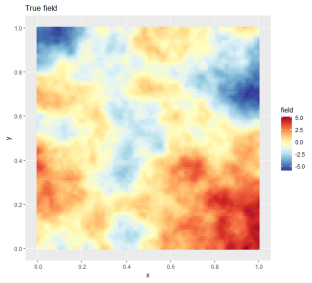

(b) Campo 2: *range* = 0.626 e $\sigma = 0.371$.



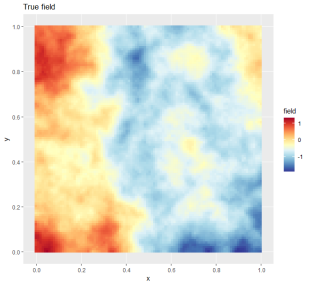

(c) Campo 3: *range* = 0.2 e $\sigma = 1.5$.




(d) Campo 4: *range* = 0.2 e $\sigma = 0.371$.




(e) Campo 5: *range* = 0.4 e $\sigma = 1.5$.




(f) Campo 6: *range* = 0.4 e $\sigma = 0.371$.

Figura 4.4: Campi spaziali considerati nelle simulazioni.

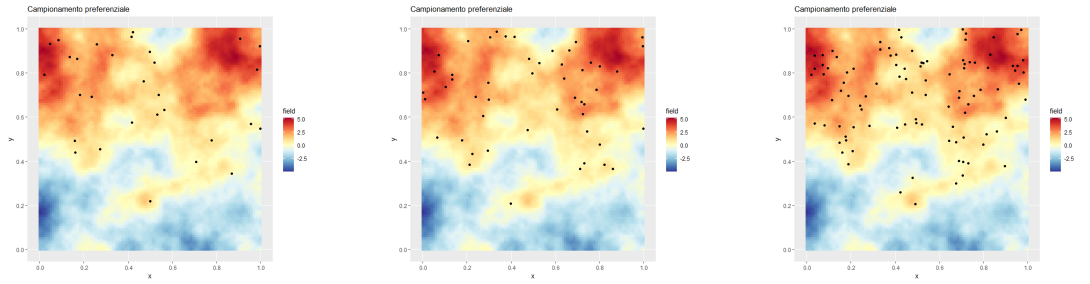
nelle Figure riportate nell'Appendice B.

Passiamo ora ad analizzare, campo per campo, gli intervalli di credibilità e le differenze al quadrato tra il campo vero e quello stimato per ogni simulazione, con l'obiettivo di trarre conclusioni sull'applicazione del modello preferenziale.
Poiché è stato già osservato che, nel caso di campionamento casuale, l'uso del modello standard o di quello preferenziale produce risultati confrontabili, ci concentreremo esclusivamente sul campionamento preferenziale.

### Campo 1

Il primo campo generato nelle simulazioni è riportato nella Figura 4.4a, dove il range è pari a 0.626 e $\sigma$ assume il valore di 1.5. La Figura 4.5 presenta tre immagini che rappresentano il campo spaziale, con un numero di punti campionati pari a 30, 50 e 100. In questo caso, i punti sono stati campionati in modo preferenziale con $\beta = 100$. Il campionamento viene effettuato in maniera analoga anche per gli altri valori di $\beta$.

Gli intervalli di credibilità per $\mu_{30}$, ottenuti applicando il modello preferenziale con




(a) n = 30 (b) n = 50 (c) n = 100

Figura 4.5: Campionamento preferenziale considerando un numero di punti pari a $n$, dove $n = 30, 50, 100$.

$\tau = 0.243$, includono il valore vero anche nei casi in cui ciò non accade utilizzando il modello standard. Inoltre, le medie stimate con il modello preferenziale risultano più vicine al valore vero rispetto a quelle ottenute con il modello standard. Quando $\tau = 0.05$, gli intervalli di credibilità mostrano un comportamento simile a quello osservato con $\tau = 0.243$. Considerazioni analoghe possono essere fatte anche per $\mu_{50}$ e $\mu_{100}$ (Figura B.1).
È evidente che il parametro $\tau$ risulta difficile da stimare con precisione per entrambi i modelli considerati, sia quando assume il valore di 0.243 sia quando è pari a 0.05. All'aumentare della numerosità campionaria, aumenta la frequenza con cui il valore vero è contenuto negli intervalli di credibilità ottenuti. Tuttavia, non si osserva un vantaggio specifico nell'uso del modello preferenziale, poiché i risultati trovati dai due modelli sono equivalenti (Figura B.2).
Analizzando il range, si osserva che gli intervalli di credibilità ottenuti con il modello preferenziale contengono il valore vero anche nei casi in cui il modello standard non lo fa, indipendentemente dal valore di $\tau$ o dal numero di punti considerati (Figura B.3).
Lo stesso vale per gli intervalli di credibilità relativi a $\sigma$ (Figura B.4).
Infine, gli intervalli di credibilità di $\beta$ risultano tutti positivi, ossia non includono lo zero, a prescindere dal valore di $n$ (Figura B.9).

Il 98% delle differenze al quadrato tra il campo vero e i campi stimati ottenuti utilizzando il modello preferenziale presenta valori massimi inferiori rispetto a quelli rilevati con il modello standard. Inoltre, si osserva che, con un valore di $\tau$ pari a 0.05, le differenze tra i due modelli diventano più evidenti nei campi stimati corrispondenti a valori di $\beta$ pari a 25, 50 e 100.

### Campo 2

La Figura 4.4b mostra il campo generato con un range pari a 0.626 e $\sigma$ uguale a 0.371. Il campionamento dei punti segue le stesse modalità descritte per il campo 1.

Gli intervalli di credibilità ottenuti per il parametro $\mu$, usando il modello preferenziale, migliorano, in alcuni casi, le stime fornite dal modello standard, includendo il valore vero all'interno dell'intervallo. Questo risultato si osserva per entrambi i valori di $\tau$, ma con il modello preferenziale si registra una maggiore frequenza di inclusione del valore vero quando consideriamo un numero di punti campionati pari a 100 (Figura B.5).
Anche per questo campo, le stime degli intervalli di credibilità di $\tau$ risultano insoddisfacenti (Figura B.6).
Quando $\tau$ è pari a 0.243, gli intervalli di credibilità per il range non includono il valore vero nella maggior parte dei casi, per entrambi i modelli. Tranne quando la numerosità campionaria è 100, dove gli intervalli ottenuti con il modello preferenziale includono il valore vero. Diminuendo $\tau$ a 0.05, gli intervalli di credibilità dei due modelli non mostrano differenze significative, ad eccezione delle ultime due simulazioni con $n = 100$, corrispondenti ai valori di $\beta$ pari a 50 e 100, in cui il modello preferenziale mostra una performance migliore (Figura B.7).
Nella maggior parte delle simulazioni, i due modelli producono risultati confrontabili per il parametro $\sigma$, con differenze minime nella lunghezza degli intervalli di credibilità. Tuttavia, per $n = 100$, il modello preferenziale tende a mostrare una maggiore precisione rispetto al modello standard, evidenziando una performance leggermente superiore nelle stime (Figura B.8).
Gli intervalli di credibilità di $\beta$ risultano più ampi rispetto a quelli del campo precedente e non sono tutti positivi. Alcuni di essi includono il valore zero, suggerendo che il campionamento potrebbe essere di natura casuale (Figura B.9).

In questa seconda situazione, i valori delle differenze al quadrato risultano simili per entrambi i modelli, con i valori massimi che rimangono inferiori a 1, diversamente da quanto osservato per il campo precedente, dove superavano ampiamente tale soglia. Per questo campo, nel 55% dei casi, i valori massimi delle differenze al quadrato sono inferiori quando si utilizza il modello preferenziale.

### Campo 3

Per il terzo campo, viene utilizzato un valore di range ridotto pari a 0.2 e un valore di $\sigma$ uguale a 1.5. Il campo generato è rappresentato nella Figura 4.4c, e il campionamento dei punti segue lo stesso schema descritto per il campo 1.

Per quanto riguarda il parametro $\mu$, quando $n$ è pari a 30, gli intervalli di credibilità risultano ampi. In generale, gli intervalli ottenuti dal modello standard non includono il valore vero, ad eccezione di una simulazione. Al contrario, utilizzando il modello preferenziale, alcuni intervalli di credibilità riescono a includere il valore vero, anche se le medie stimate rimangono significativamente distanti da esso (Figura B.10).
Anche in questo caso, il parametro $\tau$ è difficile da stimare e gli intervalli ottenuti non forniscono informazioni utili (Figura B.11).
Per quanto concerne il parametro range, gli intervalli di credibilità non sempre includono il valore vero del parametro, per entrambi i modelli. Tuttavia, l'ampiezza degli intervalli

si riduce con l'aumentare della numerosità campionaria $n$. Il modello preferenziale mostra lievi differenze rispetto al modello standard per campioni piccoli ($n = 30$) e quando $n = 100$, dove gli intervalli derivanti dal modello preferenziale contengono il valore vero più spesso rispetto a quelli trovati col modello standard (Figure B.12).
In generale, i due modelli producono risultati comparabili per $\sigma$. Il modello preferenziale sembra mostrare un vantaggio nelle simulazioni con $n = 100$ e $\tau = 0.05$, che corrispondono ad un valore di $\beta$ di 50 e 100. Questo vantaggio è osservabile anche in una simulazione con $n = 100$, $\tau = 0.243$ e $\beta$ pari a 50 (Figura B.13).
Infine, gli intervalli di credibilità di $\beta$, quando $n$ è 30 o 50, nella maggior parte delle simulazioni, contiene il valore zero. Invece, per $n = 100$, gli intervalli sono quasi tutti positivi. Inoltre, l'ampiezza degli intervalli diminuisce all'aumentare del numero di punti campionati (Figura B.18).

I risultati ottenuti per le differenze al quadrato tra il campo vero e i campi stimati nelle simulazioni, non evidenziano differenze significative tra l'applicazione del modello standard e quella del modello preferenziale. Nel 69% dei casi, i valori massimi risultano più bassi utilizzando il modello preferenziale, ma le differenze osservate nei valori massimi sono di entità ridotta rispetto a quelle rilevate nel campo 1.

### Campo 4

Il campo generato è illustrato nella Figura 4.4d, e il campionamento dei punti segue lo stesso schema adottato per il campo 1. I valori utilizzati per i parametri range e $\sigma$ sono rispettivamente 0.2 e 0.371.

Per il parametro $\mu$, gli intervalli di credibilità ottenuti dai due modelli risultano confrontabili. Tuttavia, il modello standard produce intervalli che raramente includono il valore vero, mentre il modello preferenziale riesce ad aumentare il numero di intervalli che contengono il valore vero, sebbene questo spesso si trovi molto vicino al bordo dell'intervallo (Figura B.14).
I risultati relativi a $\tau$ rimangono invariati rispetto a quelli osservati nei campi precedenti, mostrando difficoltà nella stima del parametro e scarsa informatività degli intervalli (Figure B.15).
Passando al parametro range, si osserva che entrambi i modelli producono intervalli di credibilità simili tra loro, indipendentemente dai valori di $\tau$ o dalla numerosità campionaria. Alcune eccezioni sono presenti, ma risultano poco rilevanti (Figura B.16).
Considerazioni analoghe valgono per il parametro $\sigma$, per il quale i due modelli generano intervalli di credibilità comparabili (Figura B.17).
In ultimo, gli intervalli di credibilità di $\beta$ includono spesso lo zero per $n = 30$ e $n = 50$. Invece, se $n = 100$, si ha una maggioranza di intervalli che non contengono lo zero (Figura B.18).

Analogamente al campo precedente, non si osserva un particolare vantaggio nell'uso del modello preferenziale per ottenere le mappe dei campi stimati. In questo caso, la percentuale di situazioni in cui i valori massimi risultano più bassi applicando il modello

preferenziale scende al 40%. Inoltre, le differenze tra i valori massimi ottenuti con i due modelli restano strettamente inferiori a 1.

### Campo 5

Questo campo è generato con un valore di range intermedio rispetto ai due precedenti, pari a 0.4, e un valore di $\sigma$ fissato a 1.5. La rappresentazione del campo è mostrata nella Figura 4.4d, e il campionamento dei punti segue lo stesso schema descritto per il campo 1.

Gli intervalli di credibilità per il parametro $\mu$ ottenuti applicando il modello preferenziale contengono quasi sempre il vero valore, con un'unica eccezione. Al contrario, il modello standard mostra maggiore difficoltà, con la maggior parte degli intervalli che non includono il valore vero. Questa tendenza è osservabile per entrambi i valori di $\tau$, senza differenze significative (Figura B.19).
La stima del parametro $\tau$ continua ad essere problematica: gli intervalli di credibilità prodotti da entrambi i modelli non risultano significativi e non forniscono informazioni utili (Figura B.20).
Relativamente al parametro range, il modello preferenziale mostra intervalli più ampi nei casi di campionamento ridotto ($n = 30$), ma riesce a contenere il valore vero più frequentemente rispetto al modello standard. Con l'aumentare della numerosità campionaria ($n = 100$), le differenze tra i due modelli diminuiscono, anche se il modello preferenziale continua a presentare una lieve superiorità (Figura B.21).
Gli intervalli di credibilità relativi al parametro $\sigma$, utilizzando il modello preferenziale, includono sempre il valore vero, garantendo stime più affidabili rispetto al modello standard, che presenta invece intervalli in cui il valore vero non è compreso. La variazione del parametro $\tau$ non sembra influenzare significativamente i risultati ottenuti da entrambi i modelli (Figura B.22).
Alla fine, gli intervalli relativi al parametro $\beta$, quando $n = 30$, sono ampi e talvolta il valore zero non è fuori dall'intervallo. Aumentando il numero di punti campionati, l'ampiezza degli intervalli diminuisce e lo zero non è più al suo interno (Figura B.27).

Nel 90% dei casi, le differenze al quadrato tra il campo vero e i campi stimati indicano che l'utilizzo del modello preferenziale offre un vantaggio, poiché i valori massimi osservati risultano inferiori rispetto a quelli ottenuti con il modello standard.

### Campo 6

L'ultimo campo viene generato con un valore di range pari a 0.4 e un valore di $\sigma$ uguale a 0.371. Il campo è riportato nella Figura 4.4f, e il campionamento dei punti segue le stesse modalità applicate per il campo 1.

Per il parametro $\mu$, gli intervalli di credibilità ottenuti dal modello preferenziale risultano più ampi rispetto a quelli prodotti dal modello standard, ma includono sempre il vero valore. Anche gli intervalli generati con il modello standard riescono a contenere il

| | |
|---|---|
| $\mu_{30}$ | 90% |
| $\mu_{50}$ | 83% |
| $\mu_{100}$ | 9% |

| | |
|---|---|
| $\tau_{30}$ | 46% |
| $\tau_{50}$ | 35% |
| $\tau_{100}$ | 3% |

| | |
|---|---|
| $range_{30}$ | 42% |
| $range_{50}$ | 31% |
| $range_{100}$ | 5% |

| | |
|---|---|
| $\sigma_{30}$ | 39% |
| $\sigma_{50}$ | 24% |
| $\sigma_{100}$ | 10% |

Tabella 4.2: Percentuali di volte in cui l'intervallo di credibilità del parametro ($\mu_n$, $\tau_n$, $range_n$ e $\sigma_n$) non contiene il vero valore e allo stesso tempo il rispettivo intervallo di credibilità di $\beta_n$ contiene il valore zero.

valore vero, sebbene in alcuni casi questo si trovi vicino all'estremo dell'intervallo (Figura B.23).
Gli intervalli di credibilità relativi a $\tau$ ottenuti non sono utili e risultano essere difficili da stimare bene (Figura B.24).
Per quanto riguarda il parametro range, gli intervalli di credibilità derivanti dal modello preferenziale sono generalmente più ampi, ma non riescono a contenere il vero valore nella maggior parte dei casi, indipendentemente dal valore di $\tau$ o dalla numerosità campionaria (Figura B.25).
Un comportamento simile si osserva per il parametro $\sigma$, per il quale gli intervalli di credibilità ottenuti con il modello preferenziale non includono frequentemente il vero valore (Figura B.26).
Per ultimo, il parametro $\beta$ ha intervalli di credibilità positivi e, inoltre, le medie stimate hanno valori alti (Figura B.27).

Per quest'ultimo campo, il 98% delle differenze al quadrato tra il campo vero e quelli stimati mostrano valori massimi inferiori quando si utilizza il modello preferenziale. Le differenze nei valori massimi tra i due modelli sono spesso superiori a 1, ma rimangono comunque contenute rispetto a quelle osservate nel campo precedente e nel primo campo analizzato.

### 4.2.3 Osservazioni

Un dettaglio importante da osservare riguarda il comportamento del parametro $\beta$ nel caso di campionamento preferenziale. Quando il valore vero dei parametri non è contenuto nei rispettivi intervalli di credibilità stimati, si osserva che il valore zero è incluso nell'intervallo di credibilità di $\beta$. Questo risultato suggerisce che, in tali situazioni, il campionamento si comporta come se fosse casuale, rendendo quindi il modello preferenziale non più vantaggioso rispetto al modello standard. Nella Tabella 4.2 sono riportate le percentuali di simulazioni in cui gli intervalli di credibilità di $\beta_n$ contengono lo zero e, contemporaneamente, gli intervalli di credibilità dei parametri non includono il valore vero. Ad esempio, quando l'intervallo di credibilità di $\mu_{30}$ non include il valore vero, nel 90% dei casi l'intervallo di $\beta_{30}$ contiene lo zero. Per $\mu_{50}$ e $\mu_{100}$, queste percentuali scendono rispettivamente all'83% e al 9%. Un andamento comune a tutti i parametri è che le percentuali decrescono all'aumentare del numero di punti campionati.

| $\beta\sigma$ | $\sigma = 1.5$ | $\sigma = 0.371$ |
|---|---|---|
| $\beta = 2$ | 3 | 0.742 |
| $\beta = 5$ | 7.5 | 1.855 |
| $\beta = 10$ | 15 | 3.71 |
| $\beta = 15$ | 22.5 | 5.565 |
| $\beta = 25$ | 37.5 | 9.275 |
| $\beta = 50$ | 75 | 18.55 |
| $\beta = 100$ | 150 | 37.1 |

Tabella 4.3: Prodotti dei valori di $\beta$ con i due valori di $\sigma$.

Per sintetizzare i risultati di tutte le simulazioni e confrontarle, introduciamo, come Diggle et al. [3], una misura del grado di preferenzialità, definita come il prodotto tra $\beta$ e $\sigma$. La Tabella 4.3 riporta i valori di questo prodotto per le diverse combinazioni di $\beta$ e $\sigma$ utilizzate nelle simulazioni. Si osserva che, quando il prodotto assume valori bassi, le differenze tra i risultati ottenuti con i due modelli sono meno marcate. In particolare, nel caso di campionamento preferenziale con $\sigma = 0.371$, che comporta un grado di preferenzialità ridotto, i risultati ottenuti applicando i due modelli risultano pressoché simili.

Nel nostro studio, il campionamento preferenziale segue la funzione di densità 4.1 e il grado di preferenzialità influisce direttamente sulla forma della funzione: più la funzione è estrema, maggiore è il grado di preferenzialità. I grafici della funzione di densità per tutti i valori di $\beta$ sono nelle Figure 4.1 e 4.2, rispettivamente per $\sigma = 1.5$ e $\sigma = 0.371$, viste all'inizio del capitolo quando abbiamo definito la funzione per campionare.

Quando il range è pari a 0.626 e $\sigma = 1.5$, gli intervalli di credibilità di $\beta_n$ nelle simulazioni risultano sempre positivi e non includono lo zero. Al contrario, riducendo $\sigma$ a 0.371, si osserva che l'intervallo di credibilità di $\beta_{30}$ contiene lo zero in 10 casi su 14, quello di $\beta_{50}$ in 6 casi su 14, e per $\beta_{100}$ solo in 2 casi su 14. In queste situazioni, le stime dei parametri ottenute con il modello preferenziale, nonostante il campionamento non casuale, sono comparabili ai risultati ottenuti applicando il modello standard. In generale, quando $\sigma = 0.371$, le differenze tra i due modelli non sono significative, poiché $\sigma$ controlla la variabilità del campo: con una bassa variabilità, l'effetto del campionamento preferenziale risulta meno evidente. Inoltre, su un totale di 42 simulazioni con $\sigma = 0.371$, l'intervallo di credibilità di $\beta_{30}$ contiene lo zero nel 43% dei casi, quello di $\beta_{50}$ nel 31% e quello di $\beta_{100}$ solo nel 7%.

Quando il range assume un valore piccolo, come nelle nostre simulazioni pari a 0.2, il modello preferenziale non mostra prestazioni superiori rispetto a quello standard. Inoltre, con questo valore di range e su un totale di 28 simulazioni, l'intervallo di credibilità di $\beta_{30}$ contiene lo zero nel 68% dei casi, quello di $\beta_{50}$ nel 71% e quello di $\beta_{100}$ solo nel 7%. Questi risultati evidenziano che il modello preferenziale offre vantaggi solo in determinate condizioni: il range e $\sigma$ non devono essere troppo piccoli, e l'effetto del campionamento

preferenziale risulta più evidente quanto più la funzione di probabilità è estrema, assumendo una forma simile a una funzione gradino. In queste circostanze, le differenze tra l'applicazione del modello standard e quello preferenziale diventano significative.

## 4.3 Ultima analisi: e se simulassimo anche dal modello?

Alla fine, ci siamo chiesti cosa accadrebbe se simulassimo i dati seguendo il modello preferenziale descritto al paragrafo 3.1 che, nelle simulazioni precedenti, abbiamo utilizzato esclusivamente in fase di stima. In questo contesto, analizzeremo cosa succede quando ci troviamo di fronte ad un campo spaziale generato con le stesse caratteristiche delle simulazioni precedenti, Figura 4.4. Anche i valori usati per $\tau$ rimangono gli stessi, ovvero 0.243 e 0.05. Per quanto riguarda il campionamento preferenziale, questa volta seguirà effettivamente il modello proposto: i punti saranno campionati tramite un processo di Poisson non omogeneo con intensità $\lambda(x)$ data da 3.6, ovvero

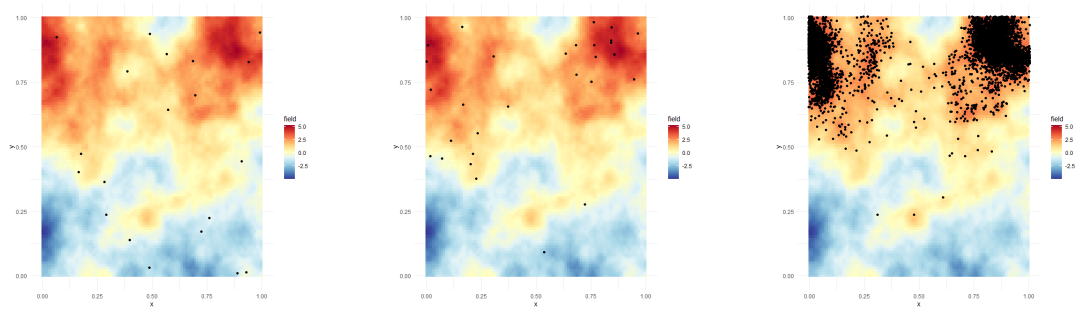
$$\lambda(x) = \exp(\alpha + \beta S(x)).$$

Il parametro $\alpha$ viene fissato a 3, basandoci sulle stime ottenute nelle simulazioni precedenti. Per il parametro $\beta$, esploreremo diversi scenari:

- assumeremo $\beta = 0$ per simulare un campionamento casuale e
- considereremo anche i valori $\beta = 0.5$ e $\beta = 2$, che rappresentano i valori medi stimati più frequentemente nelle simulazioni precedenti.

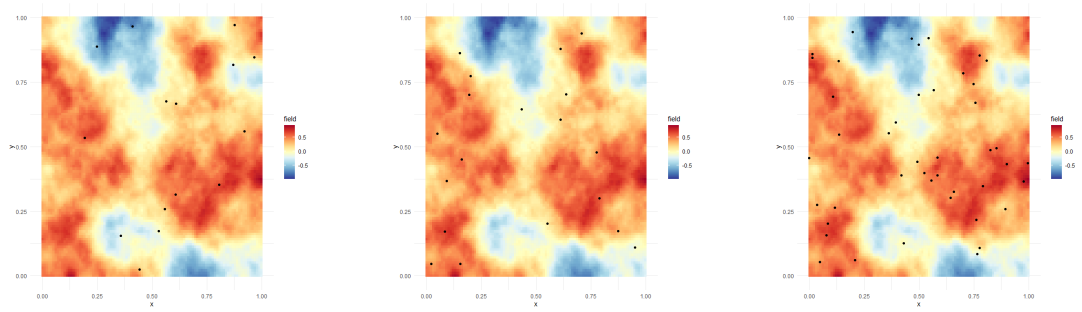
Nelle Figure 4.6 e 4.7 sono illustrate le realizzazioni dei processi di Poisson non omogenei per diversi valori di $\beta$, con un valore di range uguale a 0.626 e, rispettivamente, con $\sigma = 1.5$ e $\sigma = 0.371$. Si nota che, quando $\beta = 0$, il campionamento risulta casuale. Si può anche osservare che, nella Figura 4.6c, il numero di punti campionati è molto elevato. Questo accade perché, con $\sigma = 1.5$ (valore alto), i valori del campo spaziale possono raggiungere livelli elevati, in questo caso intorno a 5. Di conseguenza, poiché l'intensità del processo $\lambda(x)$ dipende direttamente da $\beta S(x)$, otteniamo che valori di campo alti moltiplicati per un valore di $\beta$ alto porta a un'intensità elevata, risultando in un numero atteso maggiore di punti campionati.
I processi di Poisson per i restanti campi mostrano un comportamento simile in funzione del valore di $\sigma$. In particolare, i campi con $\sigma = 1.5$ si comportano in modo analogo al campo 1 nella Figura 4.6, mentre quelli con $\sigma = 0.371$ presentano un andamento simile al campo 2 nella Figura 4.7.

Come nelle simulazioni descritte nel paragrafo precedente, analizziamo gli intervalli di credibilità ottenuti applicando sia il modello standard che quello preferenziale, riportati nelle Figure presenti nell'Appendice C.
Per il parametro $\mu$, nella Figura C.1, si osserva che nei campi in cui $\sigma = 1.5$ e $\beta = 2$ (corrispondenti a $mu3$, $mu6$, $mu15$, $mu18$, $mu27$ e $mu30$), gli intervalli di credibilità ottenuti



(a) $\beta = 0$ (b) $\beta = 0.5$ (c) $\beta = 2$

Figura 4.6: Realizzazione dei point process sul campo 1 con $\beta = 0, 0.5, 2$ rispettivamente.



(a) $\beta = 0$ (b) $\beta = 0.5$ (c) $\beta = 2$

Figura 4.7: Realizzazione dei point process sul campo 2 con $\beta = 0, 0.5, 2$ rispettivamente.

con il modello preferenziale includono il vero valore, mentre ciò non avviene con il modello standard. Invece, nei campi con $\sigma = 0.371$, entrambi i modelli producono intervalli di credibilità comparabili.
Analogamente alle simulazioni precedenti, dalla Figura C.2 si vede che i risultati relativi a $\tau$ non forniscono informazioni particolarmente utili in nessuno degli scenari considerati. Nella Figura C.3, gli intervalli di credibilità ottenuti per il range seguono l'andamento di quelli trovati per $\mu$. In particolare, quando $\sigma = 1.5$ e $\beta = 2$ (corrispondenti a *range*3, *range*6, *range*15 e *range*18), gli intervalli di credibilità ottenuti con il modello preferenziale contengono il vero valore. Inoltre, nei casi di *range*27 e *range*30 (dove range $= 0.4$, $\sigma = 1.5$ e $\beta = 2$), gli intervalli risultano differenti tra i due modelli, e quello derivante dal modello preferenziale non include il vero valore. Invece, quando $\sigma = 0.371$, entrambi i modelli producono intervalli di credibilità simili.
Per quanto riguarda il parametro $\sigma$, dalla Figura C.4 si osserva che il modello preferenziale non offre un vantaggio significativo rispetto al modello standard. In particolare, nel caso di *sigma*30 (corrispondente a range $= 0.4$, $\sigma = 1.5$, $\tau = 0.05$ e $\beta = 2$), l'intervallo di credibilità ottenuto applicando il modello standard include il valore vero, mentre quello derivante dal modello preferenziale non lo contiene.
Infine, dalla Figura C.5 si osserva che il parametro $\beta$ presenta intervalli di credibilità più

ampi quando $\sigma$ è pari a 0.371. Quando il valore vero di $\beta$ è zero, gli intervalli di credibilità lo contengono, ad eccezione di quelli relativi a *beta*31 e *beta*34 (dove range = 0.4 e $\sigma = 0.371$). Per $\beta = 0.5$, solo l'intervallo di credibilità di *beta*32 (corrispondente a range = 0.4, $\sigma = 0.371$ e $\tau = 0.243$) non include il valore vero. Infine, quando $\beta = 2$, il valore vero spesso non rientra negli intervalli di credibilità e le stime risultano generalmente più basse.

Inoltre, anche in questo quadro descriviamo i campi stimati attraverso le immagini che rappresentano la differenza al quadrato tra il campo vero e i campi stimati.
In generale, quando $\beta = 0$, il campionamento è casuale e le differenze al quadrato ottenute dai due modelli risultano confrontabili. Questo è coerente con i risultati delle simulazioni precedenti, in cui l'uso del modello standard o preferenziale porta a risultati simili in caso di campionamento casuale.
Quando $\beta = 0.5$, emergono differenze tra i due modelli nel campo 1, dove i valori massimi ottenuti con il modello preferenziale sono inferiori. Nei campi 3 e 5, per $\tau = 0.243$, si osservano miglioramenti nei valori massimi con il modello preferenziale, ma nel complesso le differenze tra i due modelli rimangono confrontabili. In tutti gli altri casi, non si hanno miglioramenti significativi derivanti dall'uso del modello preferenziale.
Per ultimo, per $\beta = 2$, nei campi 1 e 5, l'applicazione del modello preferenziale comporta una riduzione nei valori massimi e nelle zone campionate in modo preferenziale le differenze al quadrato risultano prossime allo zero. Si osserva, inoltre, una lieve riduzione nei valori massimi per i campi 6 e 2, ma per quest'ultimo solo quando $\tau = 0.243$. In tutte le altre simulazioni, le differenze al quadrato non presentano variazioni significative rispetto a quelle ottenute con il modello standard.

### 4.3.1 Osservazioni

Il parametro $\tau$ risulta difficile da stimare accuratamente, anche quando i dati sono simulati seguendo il modello preferenziale proposto. Si osserva che, analogamente al caso in cui il campionamento seguiva la funzione di probabilità 4.1, emergono differenze significative tra l'applicazione del modello standard e quella del modello preferenziale quando il grado di preferenzialità, definito come $\beta\sigma$, è elevato.

# Capitolo 5

# Conclusioni

In questa tesi, abbiamo analizzato il campionamento preferenziale attraverso un approccio bayesiano. Le simulazioni condotte hanno fornito risultati significativi che supportano l'ipotesi che, in caso di campionamento non casuale, il modello preferenziale possa migliorare l'accuratezza delle stime rispetto ai metodi geostatistici standard.

I risultati delle simulazioni hanno mostrato che, variando il parametro $\beta$, il modello preferenziale ha superato il modello standard nella stima del parametro $\mu$, con percentuali di simulazioni in cui il valore vero dei parametri era contenuto negli intervalli di credibilità significativamente più alte. Questo suggerisce che, in scenari di campionamento non casuale, l'adozione di un modello preferenziale può portare a stime più affidabili e informative.

Tuttavia, è emerso che entrambi i modelli, sia quello preferenziale che quello standard, hanno mostrato difficoltà nella stima del parametro $\tau$, sebbene le prestazioni siano migliorate con l'aumentare della numerosità campionaria. Questo risultato sottolinea l'importanza di considerare la dimensione del campione nella progettazione degli studi e nella scelta del modello statistico.

Inoltre, le simulazioni hanno dimostrato che, mentre il campionamento casuale produce risultati comparabili tra i due modelli, il campionamento preferenziale offre vantaggi tangibili in situazioni specifiche: quando il range del campo spaziale non è troppo ridotto e quando il grado di preferenzialità è alto.

In conclusione, questo lavoro non solo contribuisce a una migliore comprensione del campionamento preferenziale e della sua applicazione nella statistica bayesiana, ma fornisce anche evidenze empiriche che supportano l'adozione di modelli preferenziali in scenari pratici.

# Appendice A

# Intervalli di credibilità: Campionamento casuale

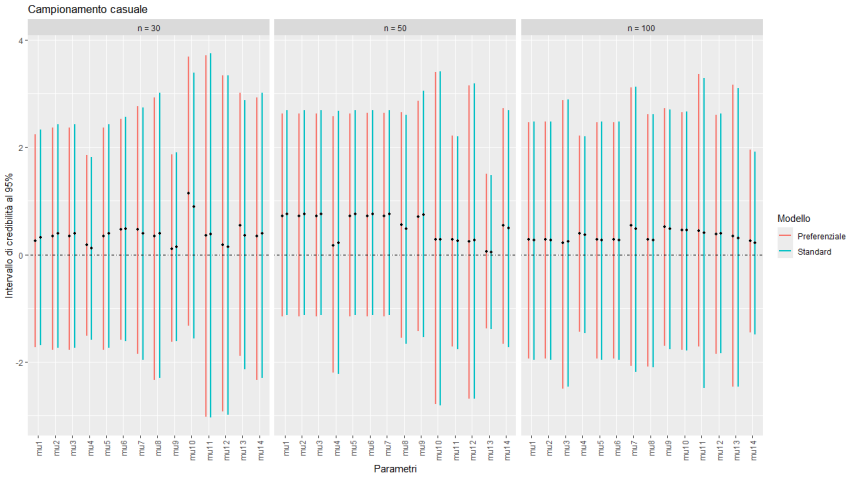



Figura A.1: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 1, ..., 14$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

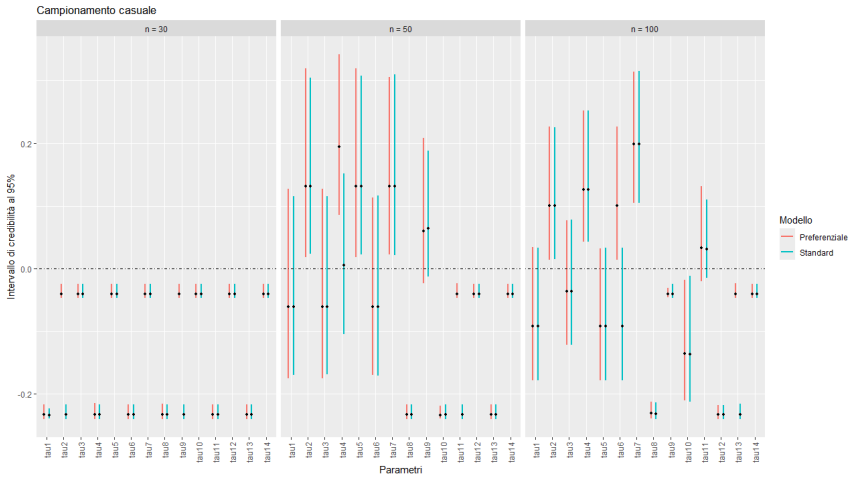



Figura A.2: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 1, ..., 14$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

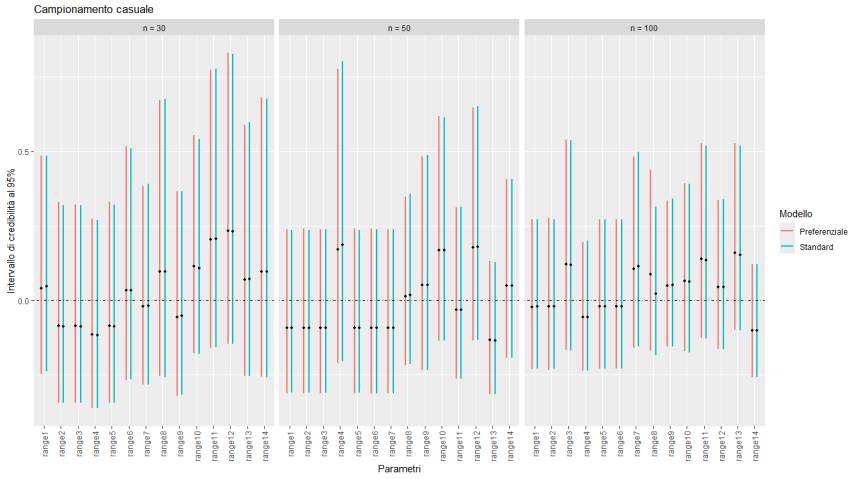



Figura A.3: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 1, ..., 14$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

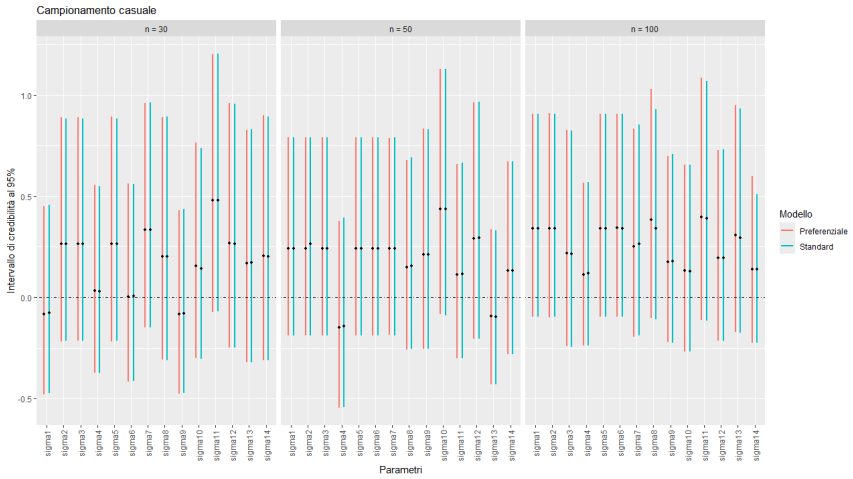



Figura A.4: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 1, ..., 14$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.


Figura A.5: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 15, ..., 28$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

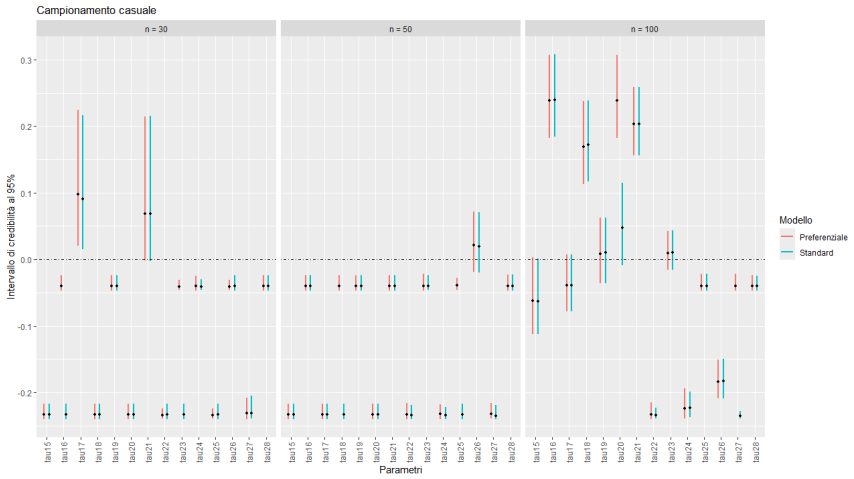



Figura A.6: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 15, ..., 28$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

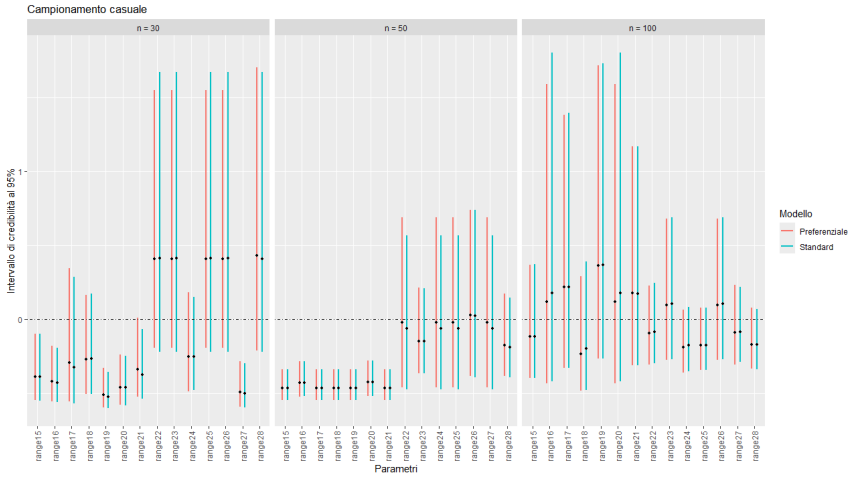



Figura A.7: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 15, ..., 28$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

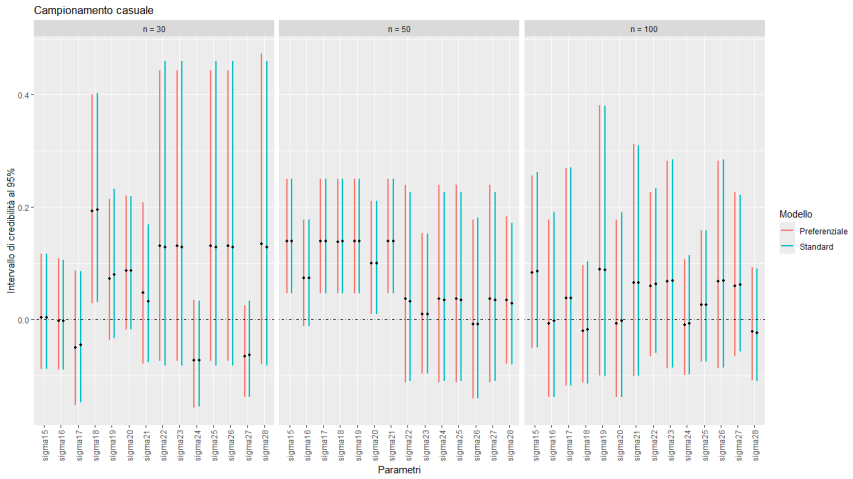



Figura A.8: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 15, ..., 28$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.


Figura A.9: Intervalli di credibilità al 95% della distribuzione a posteriori del parametro $\beta_i$ (con $i = 1, ..., 28$). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ assume due valori: 1.5 (linee verdi) e 0.371 (linee rosa). I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello e per entrambi i valori di $\sigma$, i primi 7 intervalli sono relativi alle simulazioni effettuate con $\tau = 0.243$, mentre i successivi 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.


Figura A.10: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 29, ..., 42$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

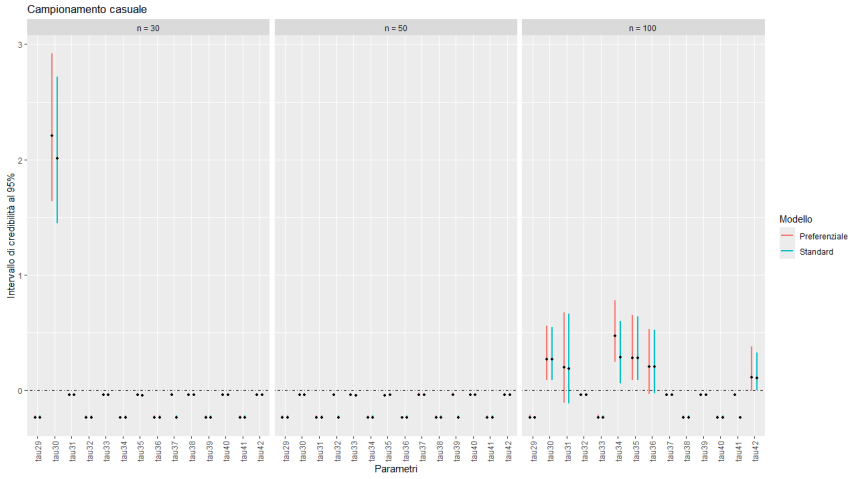



Figura A.11: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 29, ..., 42$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

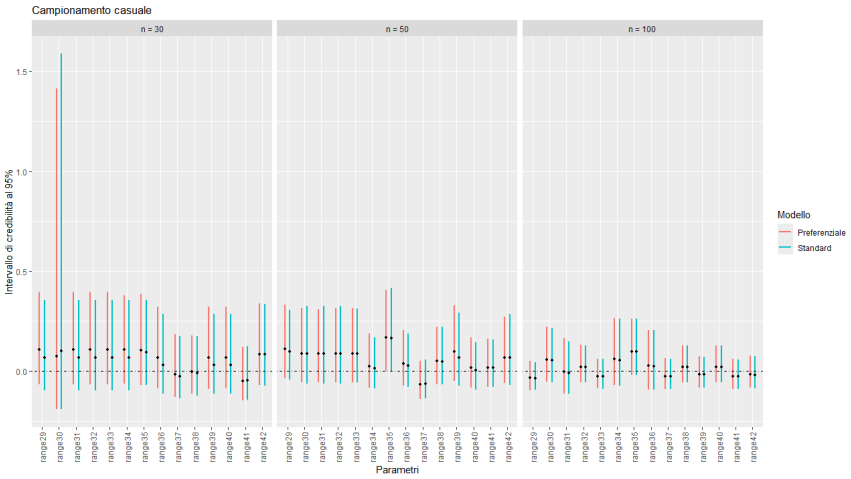



Figura A.12: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 29, ..., 42$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

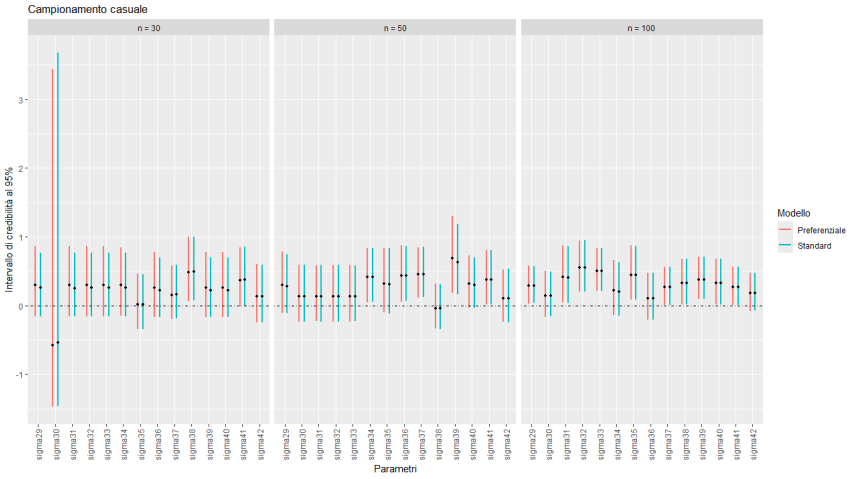



Figura A.13: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 29, ..., 42$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

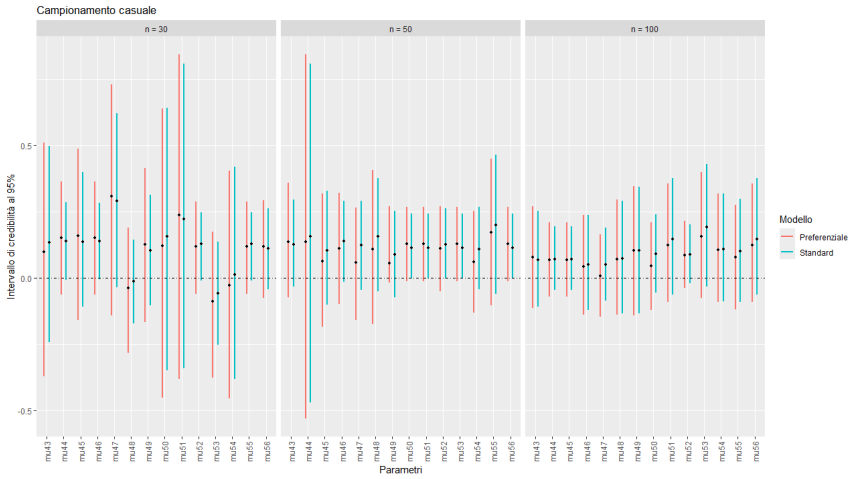



Figura A.14: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 43, ..., 56$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

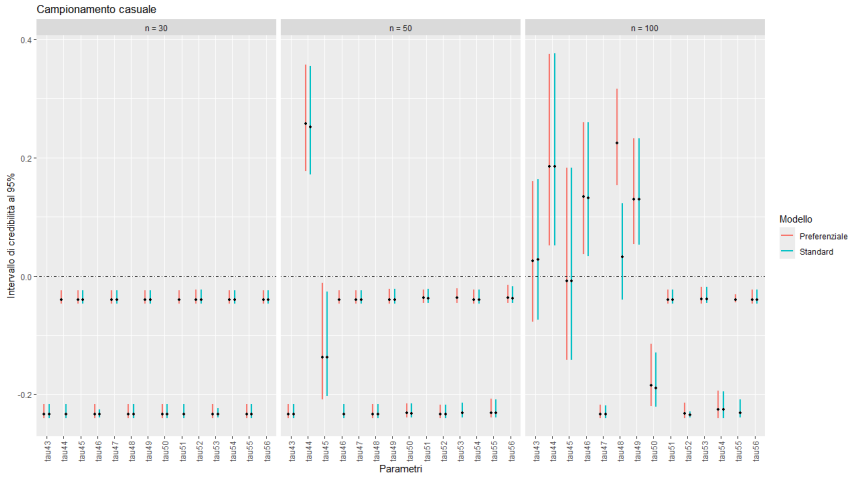



Figura A.15: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 43, ..., 56$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

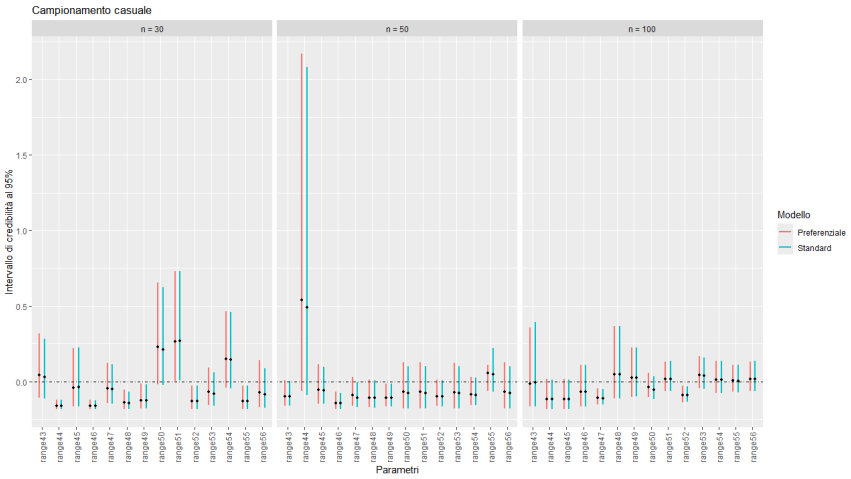



Figura A.16: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 43, ..., 56$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.


Figura A.17: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 43, ..., 56$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

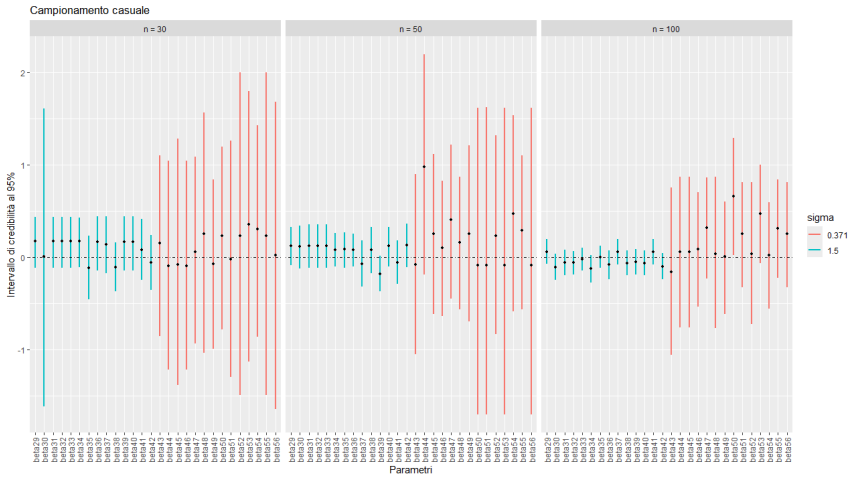



Figura A.18: Intervalli di credibilità al 95% della distribuzione a posteriori del parametro $\beta_i$ (con $i = 29, ..., 56$). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ assume due valori: 1.5 (linee verdi) e 0.371 (linee rosa). I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello e per entrambi i valori di $\sigma$, i primi 7 intervalli sono relativi alle simulazioni effettuate con $\tau = 0.243$, mentre i successivi 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.


Figura A.19: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 57, ..., 70$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

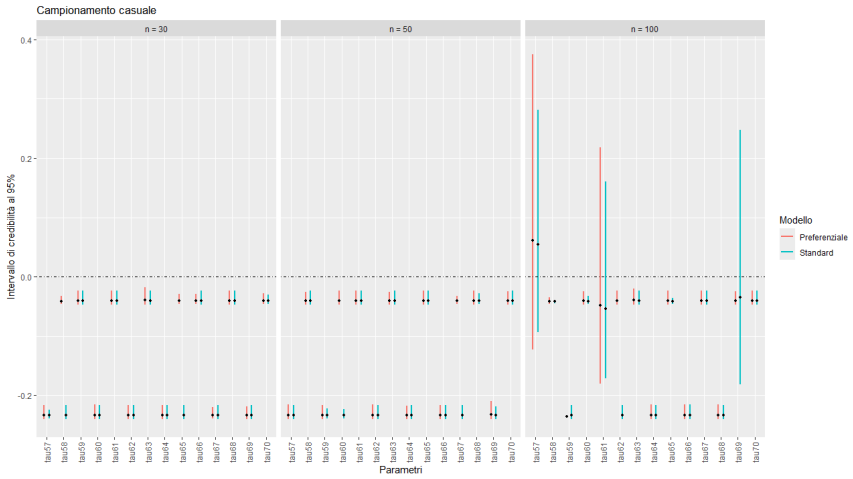



Figura A.20: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 57, ..., 70$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

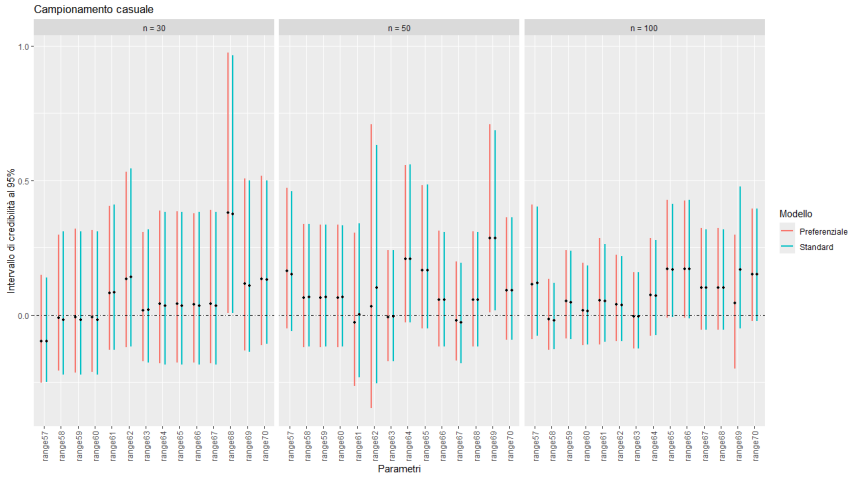



Figura A.21: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 57, ..., 70$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.


Figura A.22: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 57, ..., 70$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.


Figura A.23: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 72, ..., 84$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

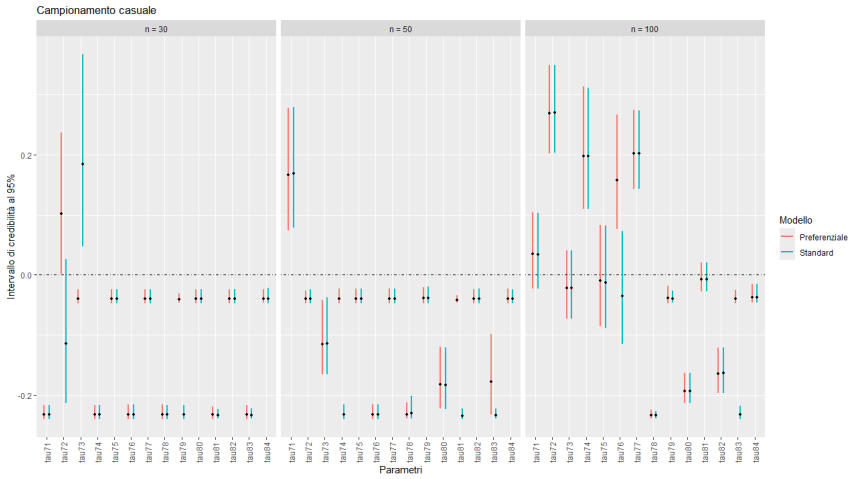



Figura A.24: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 72, ..., 84$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

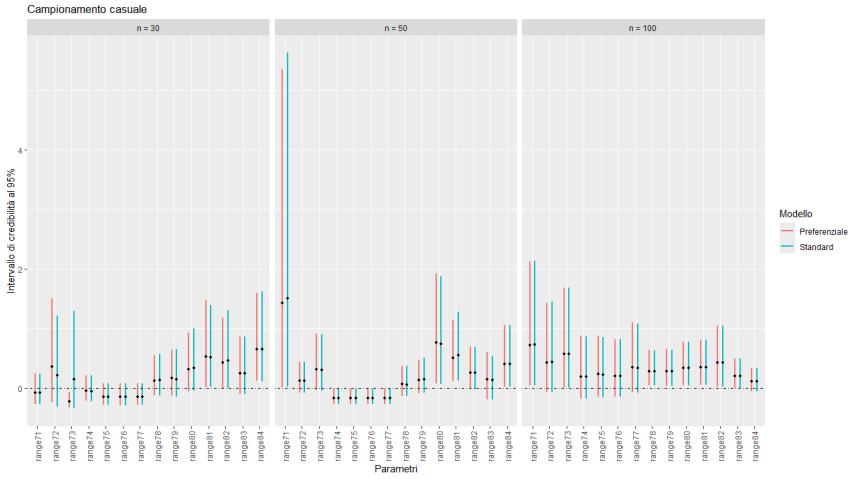



Figura A.25: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 72, ..., 84$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

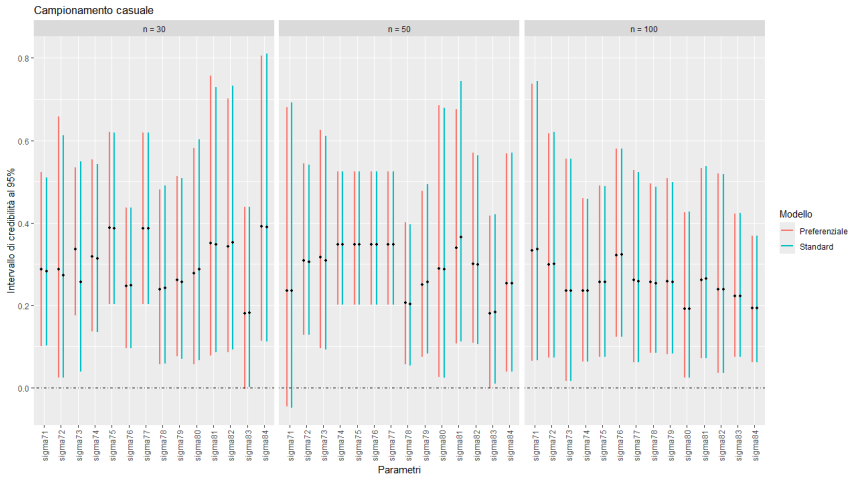



Figura A.26: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 72, ..., 84$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

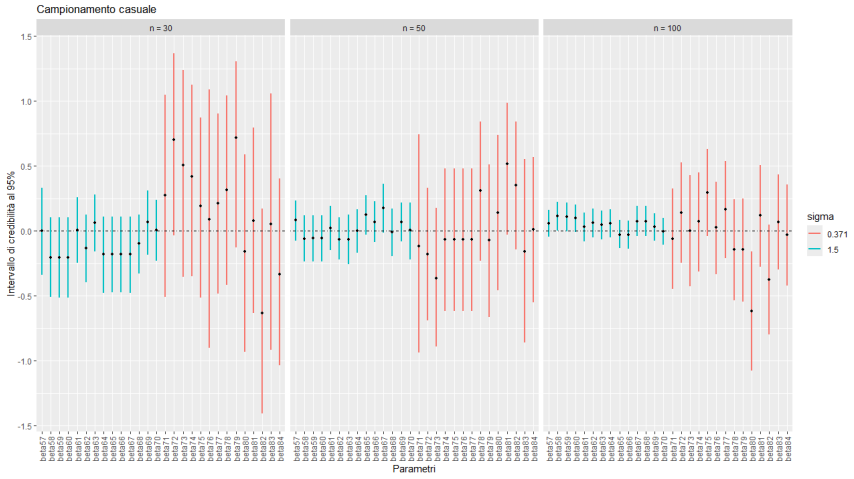



Figura A.27: Intervalli di credibilità al 95% della distribuzione a posteriori del parametro $\beta_i$ (con $i = 56, ..., 84$). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ assume due valori: 1.5 (linee verdi) e 0.371 (linee rosa). I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello e per entrambi i valori di $\sigma$, i primi 7 intervalli sono relativi alle simulazioni effettuate con $\tau = 0.243$, mentre i successivi 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

# Appendice B

# Intervalli di credibilità: Campionamento preferenziale

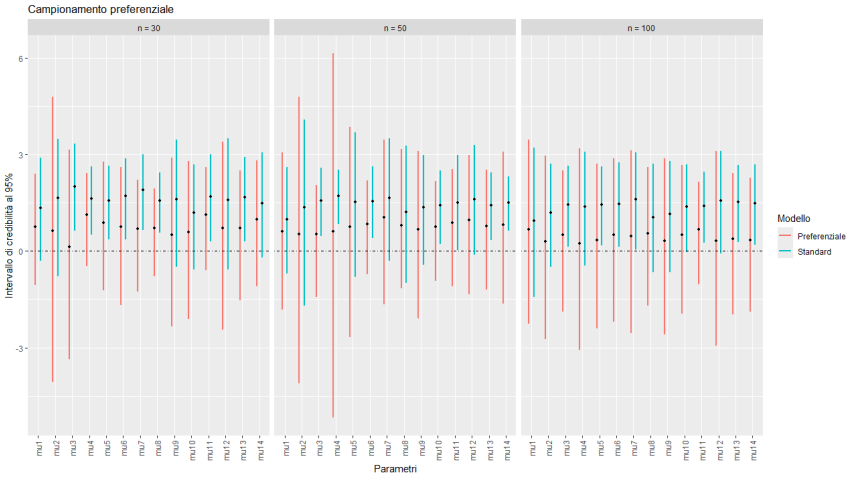



Figura B.1: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 1, ..., 14$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.


Figura B.2: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 1, ..., 14$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

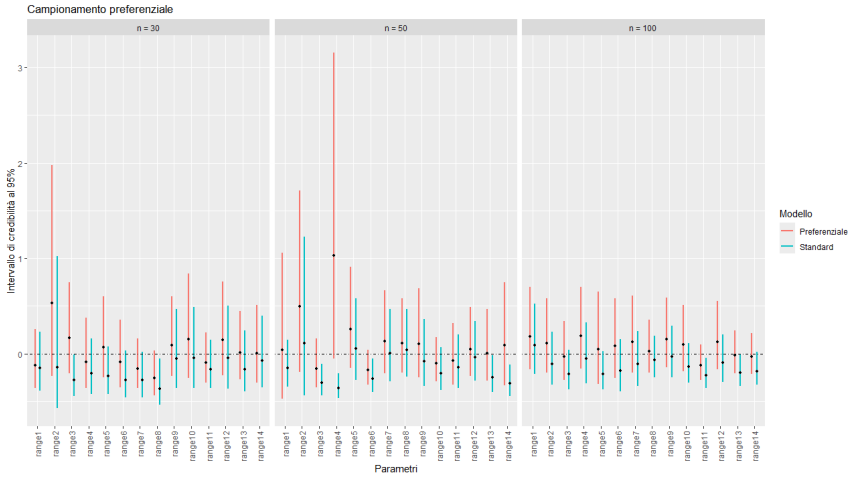



Figura B.3: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 1, ..., 14$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

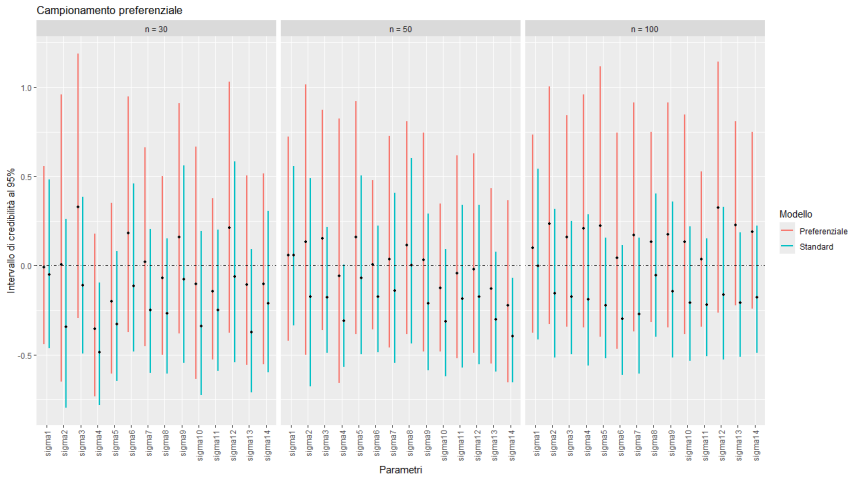



Figura B.4: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 1, ..., 14$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

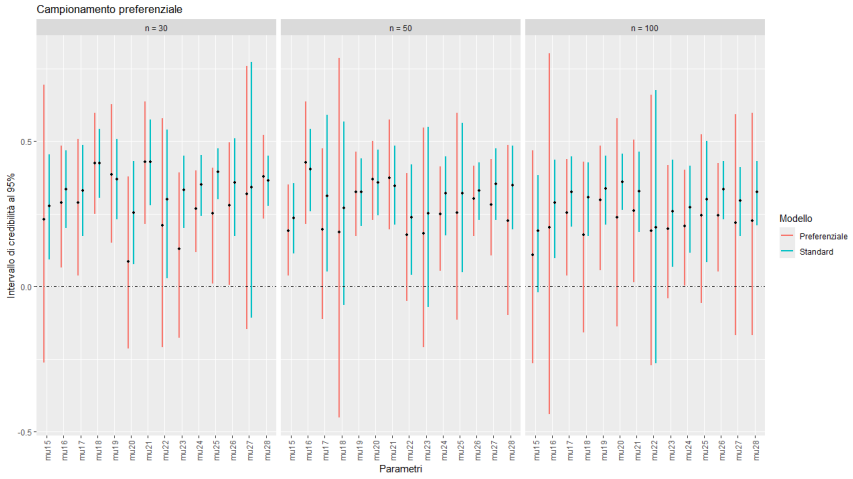



Figura B.5: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 15, ..., 28$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

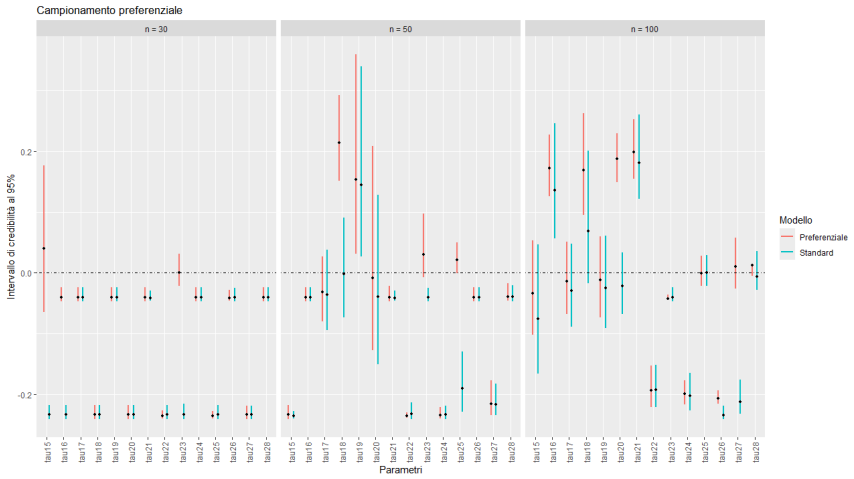



Figura B.6: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 15, ..., 28$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

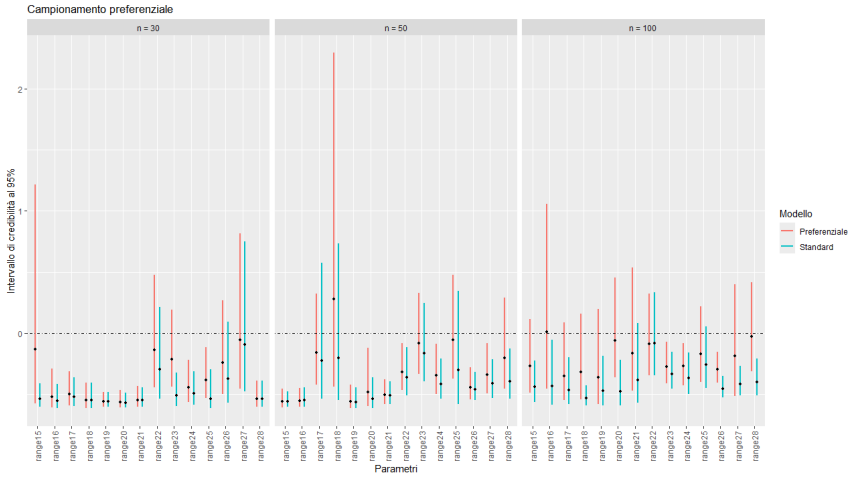



Figura B.7: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 15, ..., 28$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

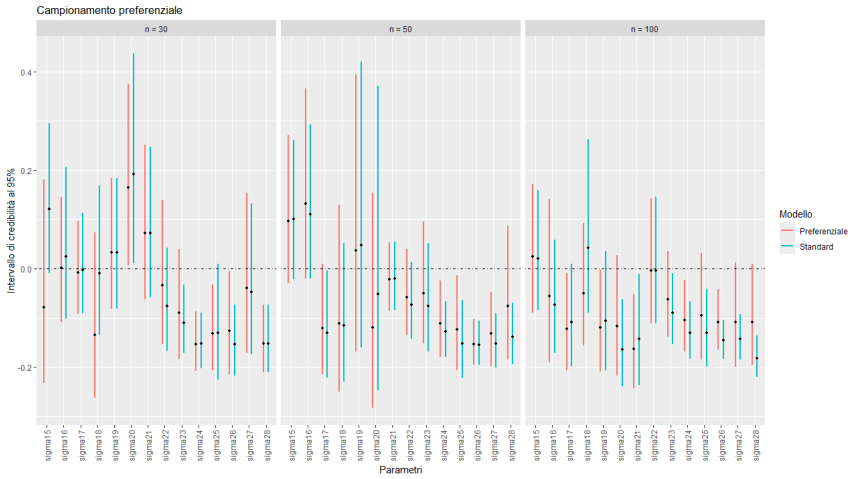



Figura B.8: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 15, ..., 28$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

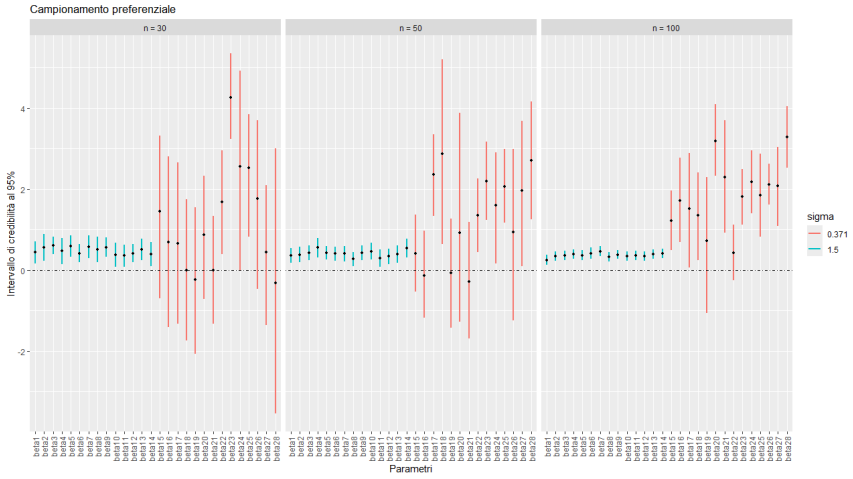



Figura B.9: Intervalli di credibilità al 95% della distribuzione a posteriori del parametro $\beta_i$ (con $i = 1, ..., 28$). I punti rappresentano le medie a posteriori. Il range è 0.626 e $\sigma$ assume due valori: 1.5 (linee verdi) e 0.371 (linee rosa). I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello e per entrambi i valori di $\sigma$, i primi 7 intervalli sono relativi alle simulazioni effettuate con $\tau = 0.243$, mentre i successivi 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

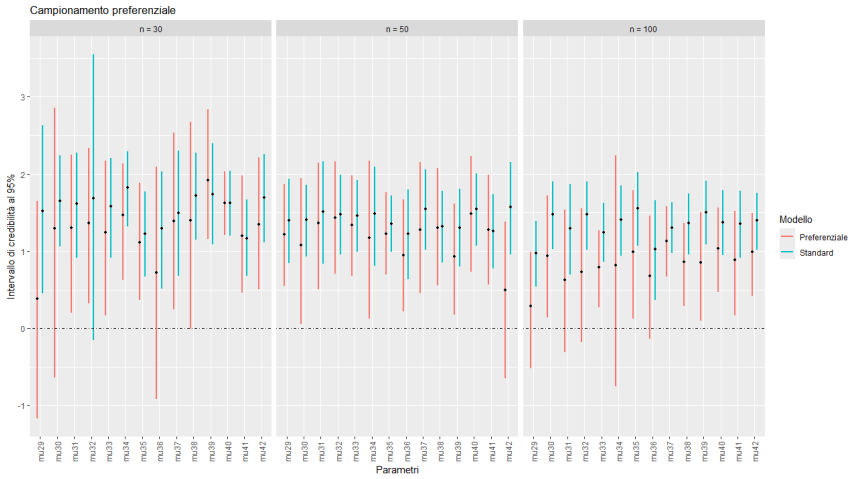



Figura B.10: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 29, ..., 42$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

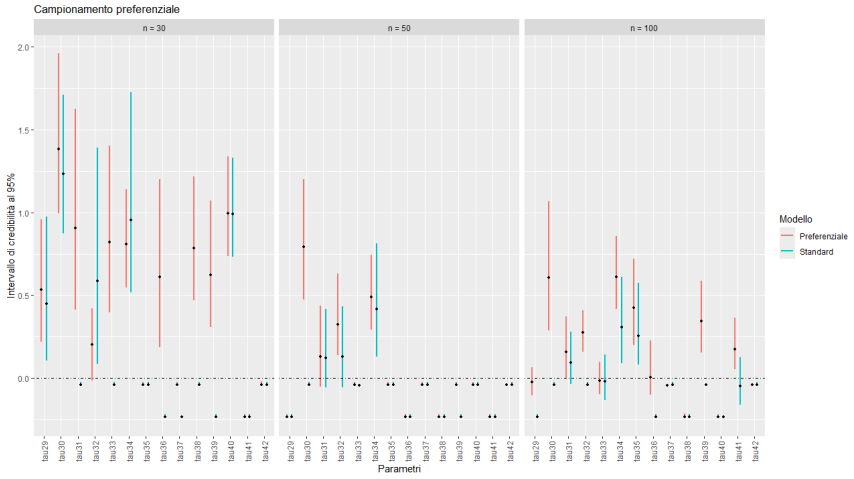



Figura B.11: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 29, ..., 42$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

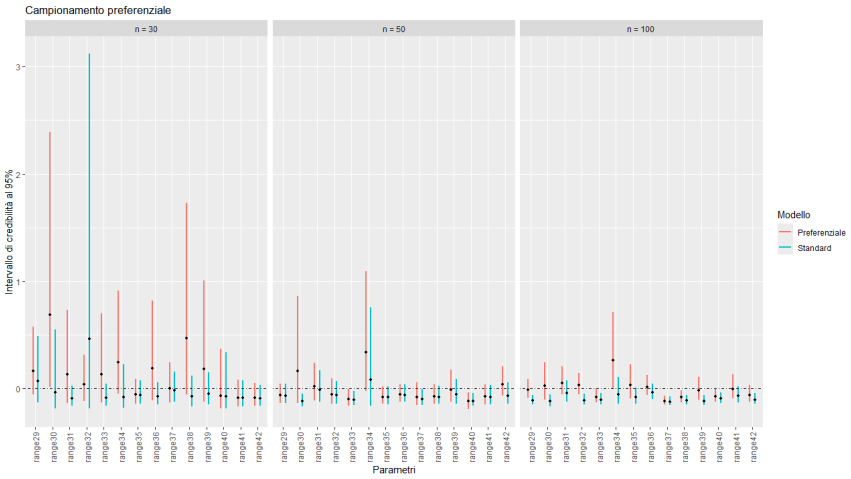



Figura B.12: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 29, ..., 42$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

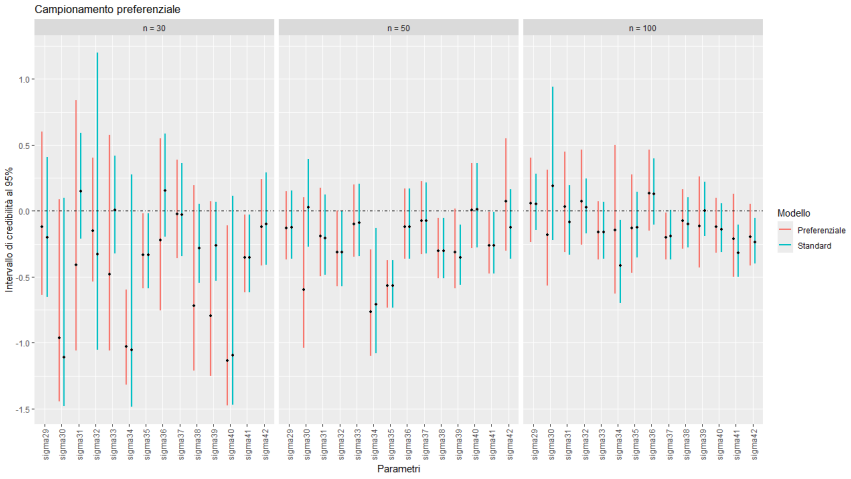



Figura B.13: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 29, ..., 42$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

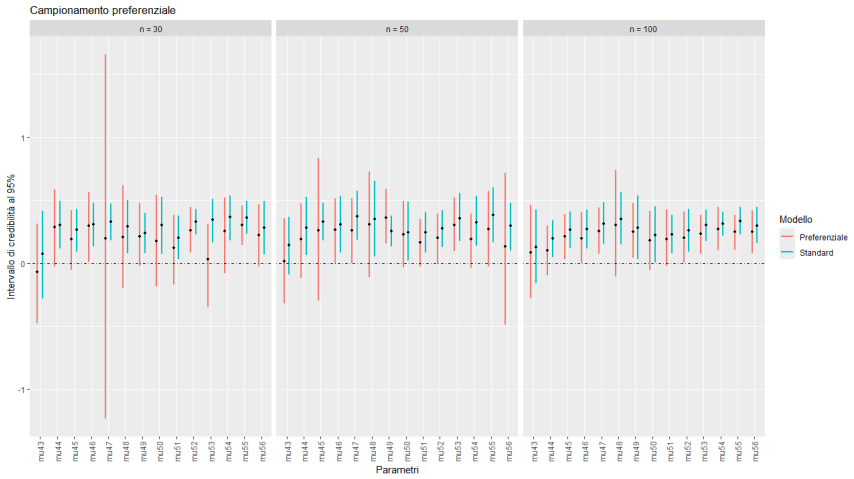



Figura B.14: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 43, ..., 56$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

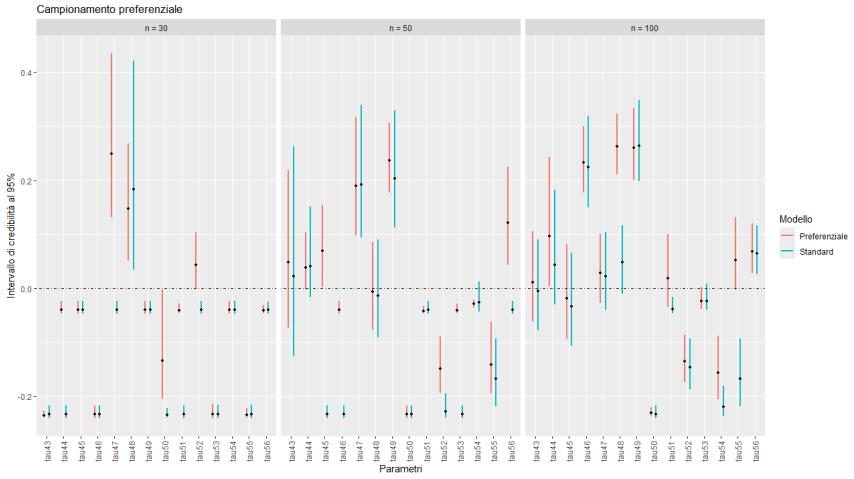



Figura B.15: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 43, ..., 56$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

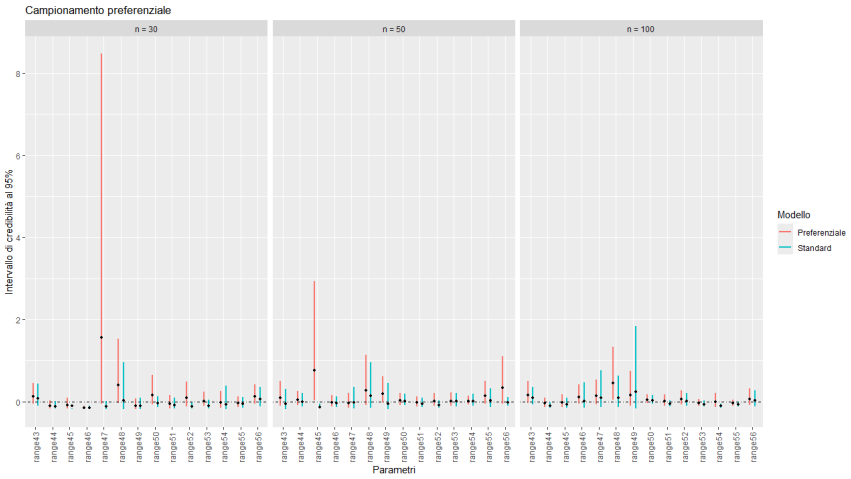



Figura B.16: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 43, ..., 56$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

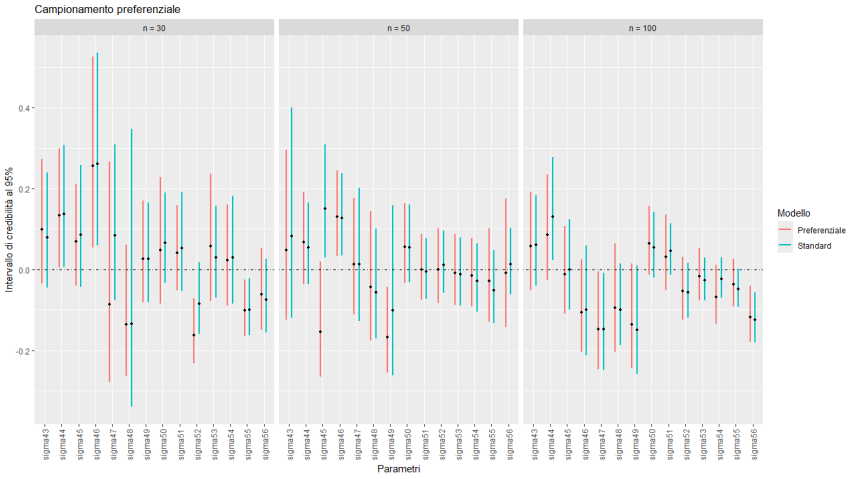



Figura B.17: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 43, ..., 56$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

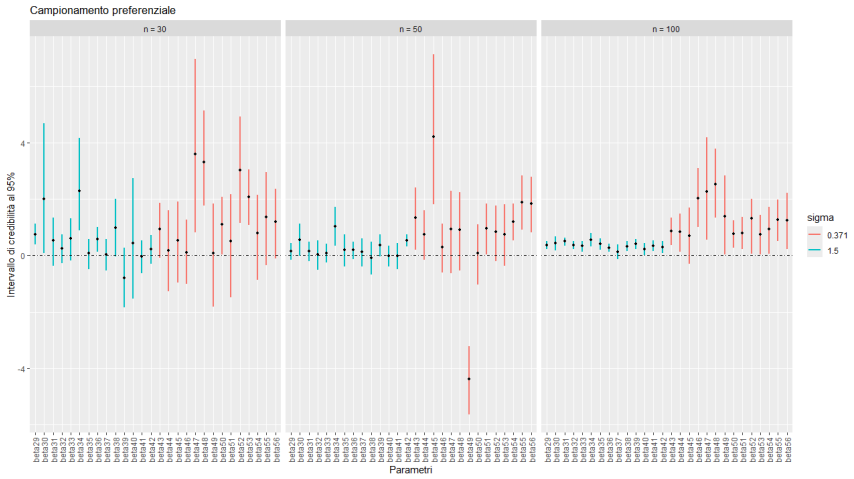



Figura B.18: Intervalli di credibilità al 95% della distribuzione a posteriori del parametro $\beta_i$ (con $i = 29, ..., 56$). I punti rappresentano le medie a posteriori. Il range è 0.2 e $\sigma$ assume due valori: 1.5 (linee verdi) e 0.371 (linee rosa). I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello e per entrambi i valori di $\sigma$, i primi 7 intervalli sono relativi alle simulazioni effettuate con $\tau = 0.243$, mentre i successivi 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

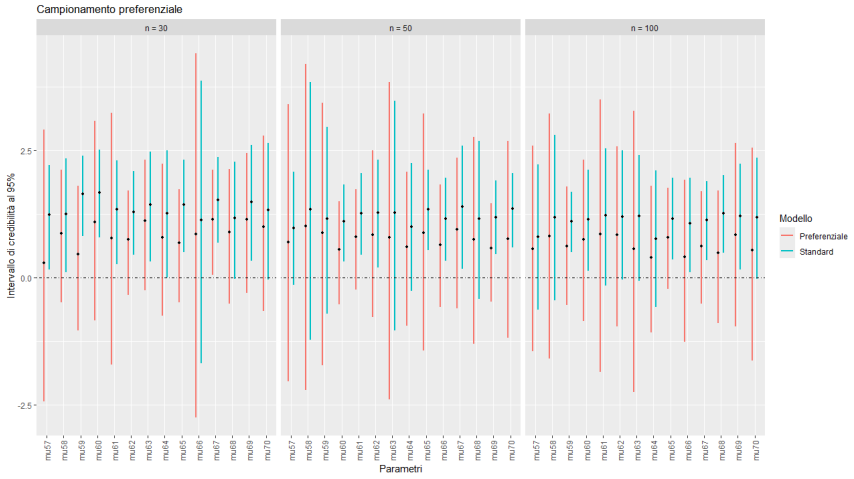



Figura B.19: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 57, ..., 70$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

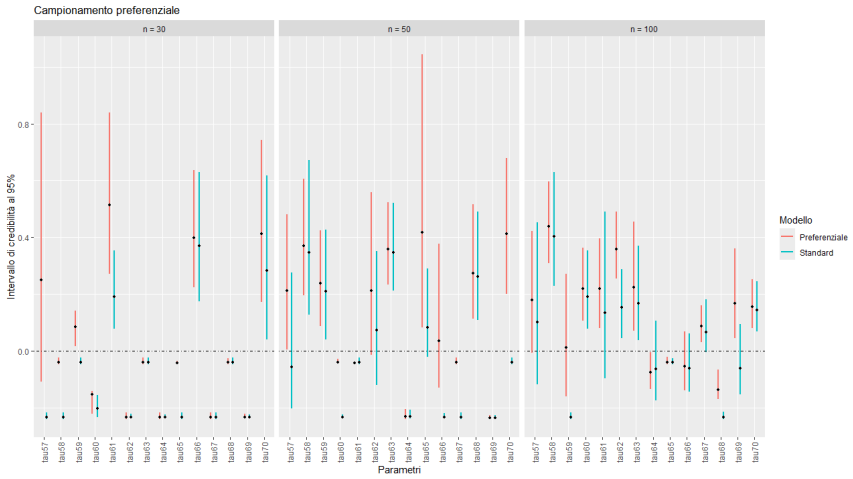



Figura B.20: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 57, ..., 70$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

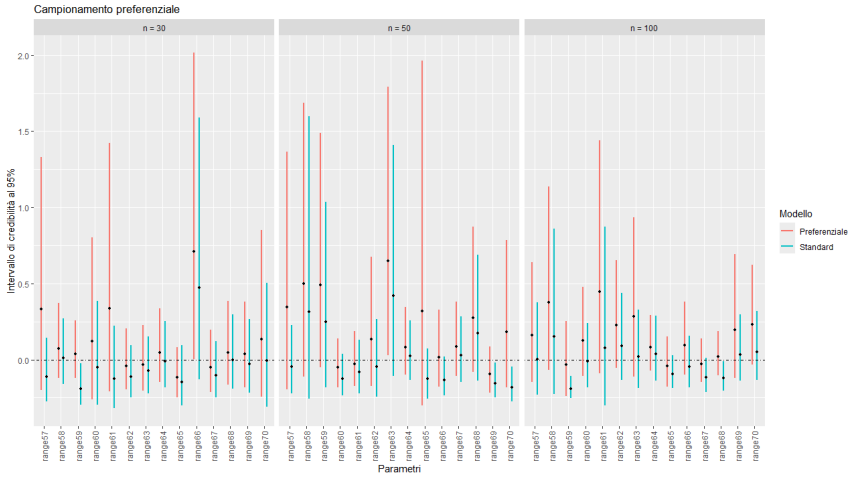



Figura B.21: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 57, ..., 70$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

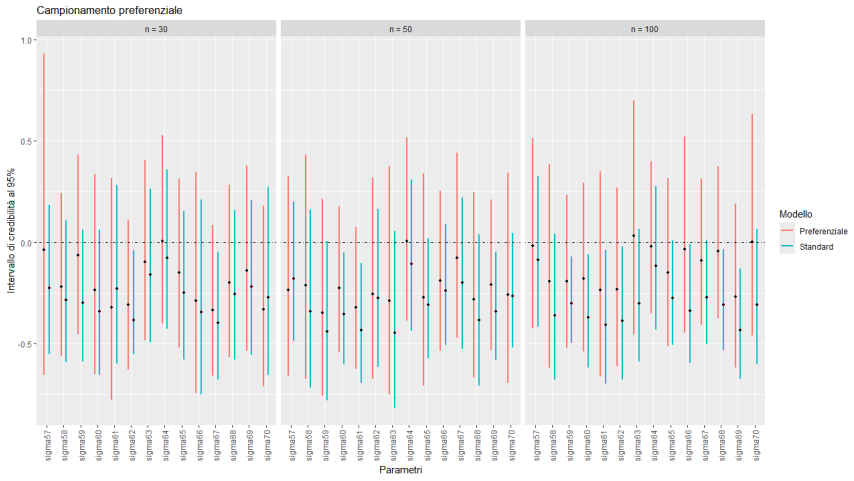



Figura B.22: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 57, ..., 70$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 1.5. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

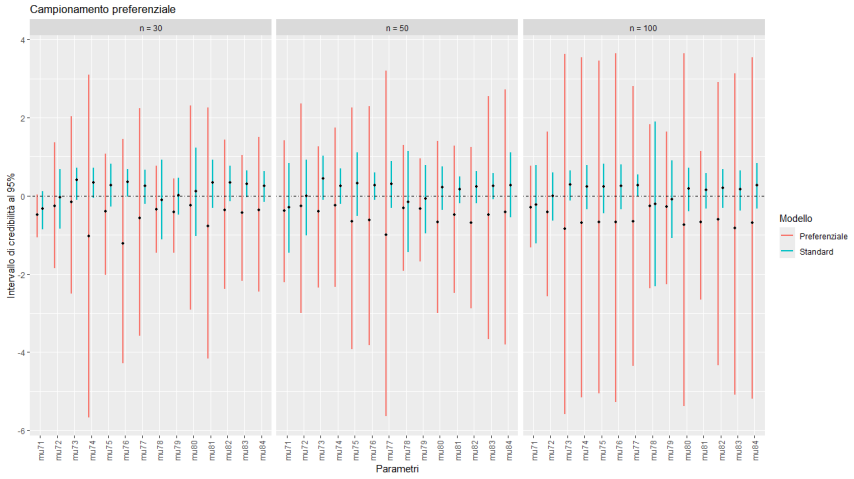



Figura B.23: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu_i$ (con $i = 72, ..., 84$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

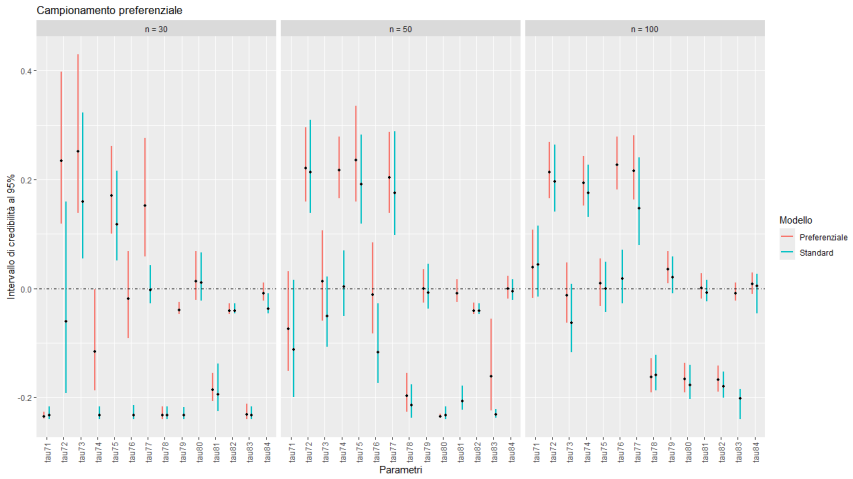



Figura B.24: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau_i$ (con $i = 72, ..., 84$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

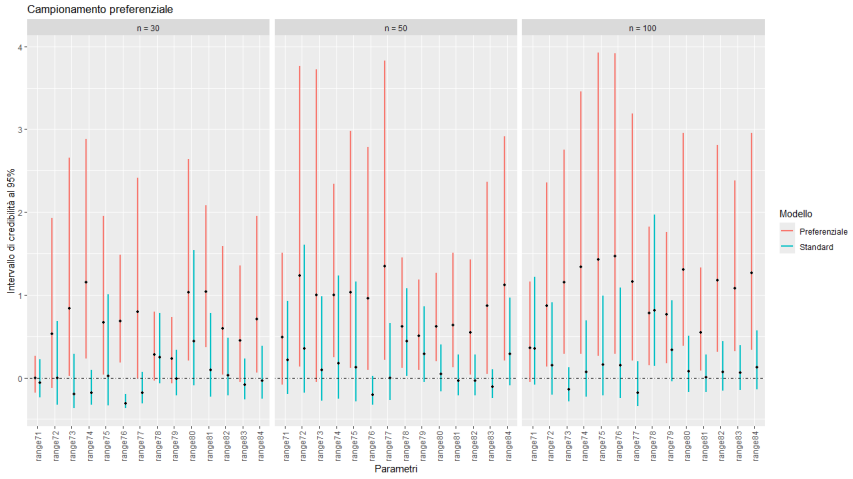



Figura B.25: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $range_i$ (con $i = 72, ..., 84$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

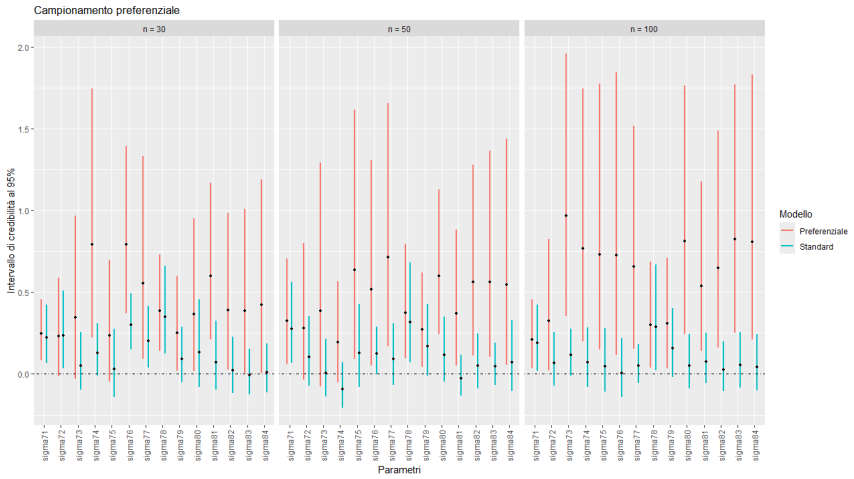



Figura B.26: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma_i$ (con $i = 72, ..., 84$) e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ è 0.371. I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello, le prime 7 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

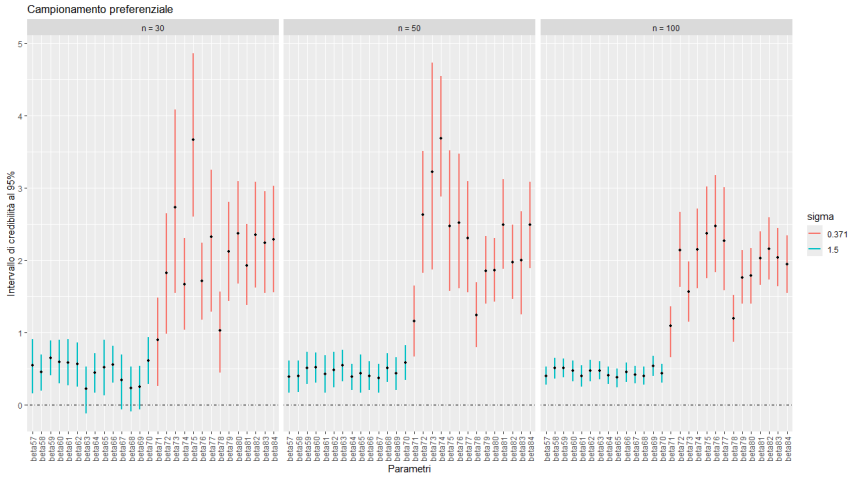



Figura B.27: Intervalli di credibilità al 95% della distribuzione a posteriori del parametro $\beta_i$ (con $i = 56, ..., 84$). I punti rappresentano le medie a posteriori. Il range è 0.4 e $\sigma$ assume due valori: 1.5 (linee verdi) e 0.371 (linee rosa). I 3 pannelli mostrano i risultati per diverse numerosità campionarie, pari a 30, 50 e 100. In ciascun pannello e per entrambi i valori di $\sigma$, i primi 7 intervalli sono relativi alle simulazioni effettuate con $\tau = 0.243$, mentre i successivi 7 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

# Appendice C

# Intervalli di credibilità: Point Process

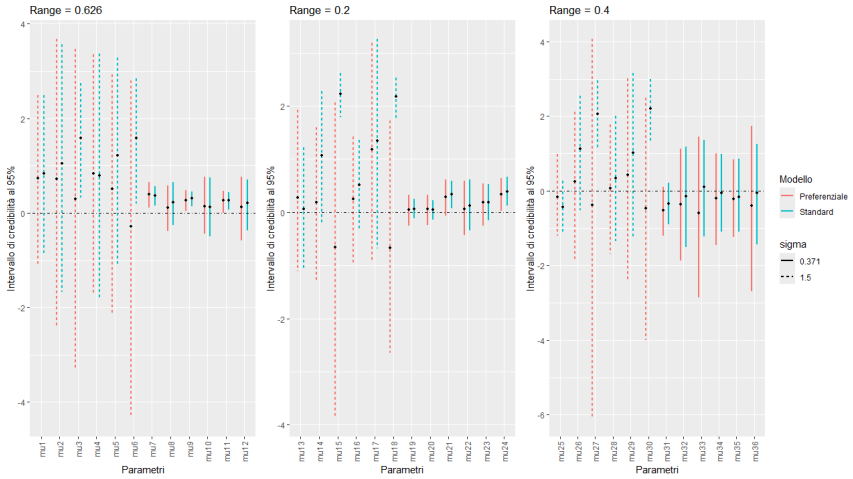



Figura C.1: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\mu$ e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. I 3 pannelli mostrano i risultati per i diversi valori di range. In ciascun pannello, le linee tratteggiate si riferiscono al campo con $\sigma = 1.5$, mentre quelle continue al campo con $\sigma = 0.371$. Per ogni range e per ogni valore di $\sigma$ le prime 3 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 3 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

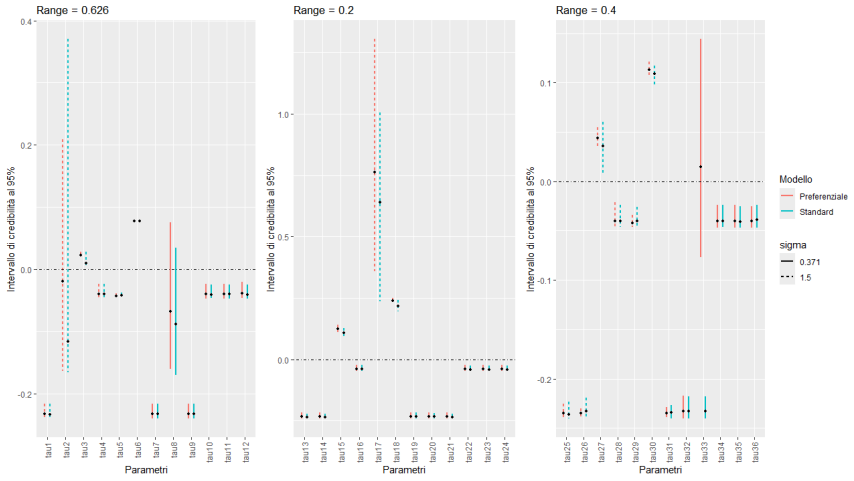



Figura C.2: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\tau$ e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. I 3 pannelli mostrano i risultati per i diversi valori di range. In ciascun pannello, le linee tratteggiate si riferiscono al campo con $\sigma = 1.5$, mentre quelle continue al campo con $\sigma = 0.371$. Per ogni range e per ogni valore di $\sigma$ le prime 3 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 3 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

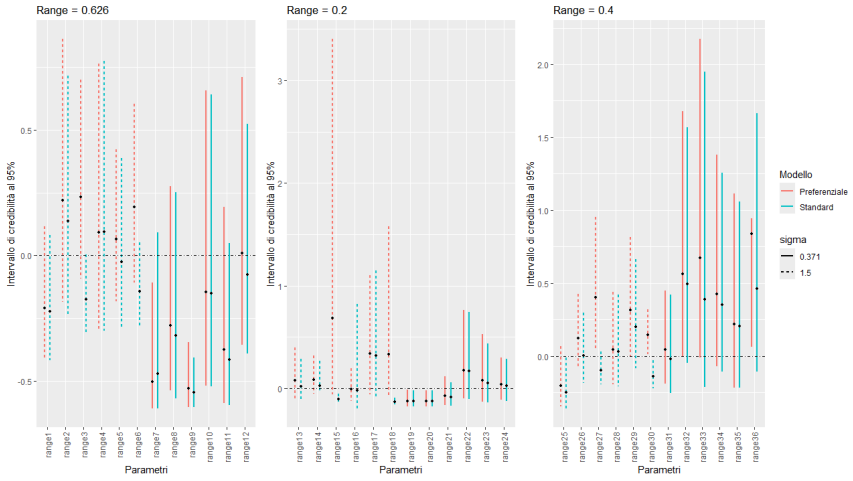



Figura C.3: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro *range* e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. I 3 pannelli mostrano i risultati per i diversi valori di range. In ciascun pannello, le linee tratteggiate si riferiscono al campo con $\sigma = 1.5$, mentre quelle continue al campo con $\sigma = 0.371$. Per ogni range e per ogni valore di $\sigma$ le prime 3 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 3 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

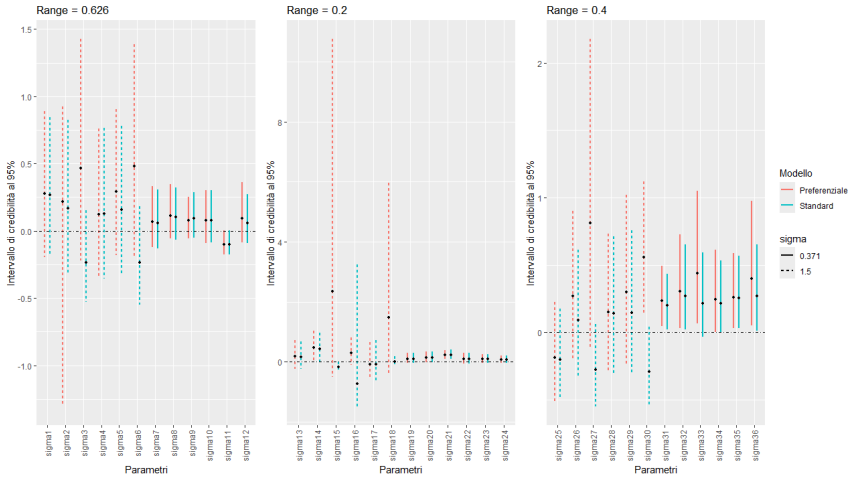



Figura C.4: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\sigma$ e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. I 3 pannelli mostrano i risultati per i diversi valori di range. In ciascun pannello, le linee tratteggiate si riferiscono al campo con $\sigma = 1.5$, mentre quelle continue al campo con $\sigma = 0.371$. Per ogni range e per ogni valore di $\sigma$ le prime 3 coppie di intervalli sono relative alle simulazioni effettuate con $\tau = 0.243$, mentre le successive 3 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

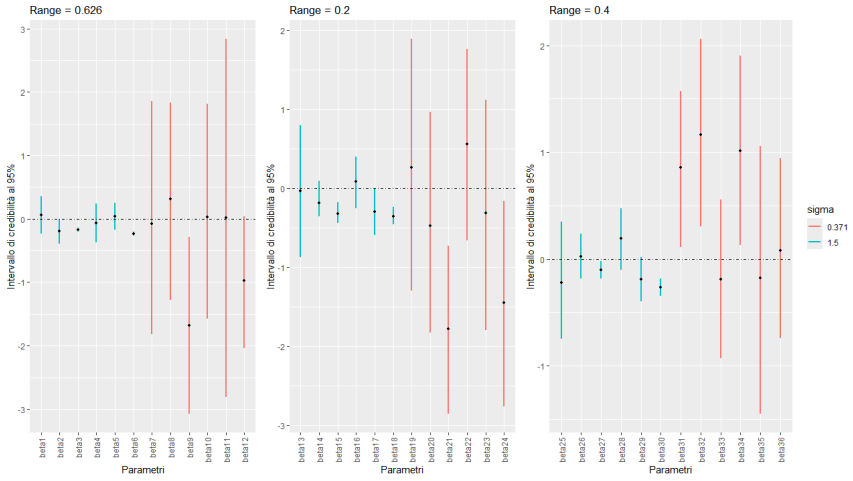



Figura C.5: Intervalli di credibilità al 95% della distribuzione a posteriori per la differenza tra il parametro $\beta$ e il suo vero valore, confrontando l'applicazione del modello standard (linee verdi) con quella del modello preferenziale (linee rosa). I punti rappresentano le medie a posteriori. I 3 pannelli mostrano i risultati per i diversi valori di range. In ciascun pannello, le linee tratteggiate si riferiscono al campo con $\sigma = 1.5$, mentre quelle continue al campo con $\sigma = 0.371$. Per ogni range e per ogni valore di $\sigma$ i primi 3 intervalli sono relativi alle simulazioni effettuate con $\tau = 0.243$, mentre i successivi 3 con $\tau = 0.05$. I valori di $\beta$ sono in ordine crescente per ogni valore di $\tau$.

# Bibliografia


[1] Marta Blangiardo and Michela Cameletti. *Spatial and spatio-temporal Bayesian models with R-INLA*. John Wiley & Sons, 2015.

[2] Noel Cressie. *Statistics for spatial data*. John Wiley & Sons, 2015.

[3] Peter J Diggle, Raquel Menezes, and Ting-li Su. Geostatistical inference under preferential sampling. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 59(2):191–232, 2010.

[4] JA Fernández, A Rey, and A Carballeira. An extended study of heavy metal deposition in galicia (nw spain) based on moss analysis. *Science of the Total Environment*, 254(1):31–44, 2000.

[5] Virgilio Gómez-Rubio. *Bayesian inference with INLA*. Chapman and Hall/CRC, 2020.

[6] Peter D Hoff. *A first course in Bayesian statistical methods*, volume 580. Springer, 2009.

[7] Elias Krainski, Virgilio Gómez-Rubio, Haakon Bakka, Amanda Lenzi, Daniela Castro-Camilo, Daniel Simpson, Finn Lindgren, and Håvard Rue. *Advanced spatial modeling with stochastic partial differential equations using R and INLA*. Chapman and Hall/CRC, 2018.

[8] Finn Lindgren, Håvard Rue, and Johan Lindström. An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(4):423–498, 2011.

[9] Sara Martino and Andrea Riebler. Integrated nested laplace approximations (inla). *arXiv preprint arXiv:1907.01248*, 2019.

[10] Jesper Møller, Anne Randi Syversveen, and Rasmus Plenge Waagepetersen. Log gaussian cox processes. *Scandinavian journal of statistics*, 25(3):451–482, 1998.

[11] Havard Rue and Leonhard Held. *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC, 2005.

[12] Håvard Rue, Sara Martino, and Nicolas Chopin. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 71(2):319–392, 2009.

[13] Håvard Rue, Andrea Riebler, Sigrunn H Sørbye, Janine B Illian, Daniel P Simpson, and Finn K Lindgren. Bayesian computing with inla: a review. *Annual Review of Statistics and Its Application*, 4(1):395–421, 2017.