POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

Modelli di Ecologia Spaziale con INLA: Inferenza su Processi Puntuali e Campionamento Differenziato della Popolazione di Oloturie nell'Area Marina dell'Isola del Giglio.

Relatore
Professore Mastrantonio Gianluca

Candidato
Poggio Daniele

Anno Accademico 2023-2024

Sommario

I processi di punto spaziali sono utilizzati per modellare fenomeni dove i punti indicano la posizione di oggetti in due o tre dimensioni. Alcuni dei campi di applicazioni sono, per esempio: epidemiologia spaziale, analisi immagini e ecologia. In particolare, per i modelli ecologici, l'utilizzo di modelli di processi di punti permette lo studio della dinamica di popolazione delle specie in habitat ecologici, potendo valutare l'influenza di fattori esterni, per esempio: le caratteristiche geografiche dell'habitat, come l'altitudine, la pendenza del terreno, fattori climatici, come la temperatura e l'umidità, la presenza di altre specie, come la flora presente nell'habitat o la presenza umana nel territorio. Nella raccolta dati per studi ecologici, spesso le caratteristiche del territorio sono un fattore limitante per una ricerca estensiva e una soluzione che viene spesso adottata e di utilizzare diverse tecniche di campionamento per la raccolta dati, in modo che una tecnica di campionamento riesca a coprire aree dell'habitat dove non sarebbe possibile utilizzare le altre tecniche disponibili per lo studio. Da un punto di vista statistico questo approccio permette una maggior raccolta di dati, ma è necessario sviluppare i modelli tenendo conto della diversa origine dei dati raccolti. Per esempio, all'interno di uno stesso habitat, l'analisi di immagini potrebbe essere particolarmente efficiente ad idenficare campioni in spazi aperti, come per esempio praterie, ma avere difficoltà a registrare campioni in aree boschive; al contrario la registrazioni manuale da parte di persone preposte potrebbe permettere di identificare facilmente campioni in zone boschive di dimensioni ridotte e facilmente esplorabili, ma potrebbe essere troppo impegnativo da utilizzare.

Ringraziamenti

Ringrazio tutte le persone che mi sono state vicine e mi hanno supportato durante questi anni.
Ringrazio prima di tutto i miei genitori, per avermi dato la possibilità di seguire i miei interessi e la mia curiosità e di aver cercato di supportarmi come meglio hanno potuto in questo percorso.
Ringrazio Annalisa, per avere reso momenti difficili più leggeri, momenti belli indimenticabili, per tutto quello che abbiamo condiviso assieme, per avermi aiutato a guardare dentro me stesso, imparando a conoscermi e capire meglio me stesso.
Ringrazio i miei amici Serena, Martina, Samu, Gaia, Francesco, Marco, Domenico e Gianluca per i momenti divertenti passati assieme che mi hanno alleggerito le giornate, per avermi ascoltato e cercato di capirmi quando mi chiudevo in me stesso.
Infine ringrazio il Professore Mastrantonio per l'aiuto nello sviluppo di questo lavoro di Tesi, per aver stimolato la mia passione e la mia curiosità per la ricerca e avermi aiutato ad essere, forse, un po' più consapevole dei miei mezzi.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Point Pattern Process

1.1 Introduzione

I processi di punto, nell'ambito della geostatistica, sono strumenti matematici e statistici utilizzati per modellare e analizzare la distribuzione spaziale di eventi o oggetti che possono essere rappresentati come punti su una superficie o nello spazio tridimensionale.
I processi di punto trovano applicazione in una vasta gamma di settori, tra cui la geologia, l'ecologia, l'epidemiologia, l'ingegneria civile e le scienze ambientali. Ad esempio, possono essere impiegati per studiare la distribuzione dei giacimenti minerari, la localizzazione delle piante in un'area boschiva, la diffusione di malattie infettive in una popolazione o la disposizione delle infrastrutture urbane.
Uno degli obiettivi principali dell'analisi dei processi di punto è determinare se la distribuzione spaziale degli eventi sia casuale o se presenti qualche forma di struttura, come aggregazione o dispersione. Per fare ciò, vengono utilizzati diversi strumenti statistici, come le funzioni di intensità, le funzioni di autocorrelazione spaziale e i test di completezza spaziale.
Dal punto di vista matematico, un processo di punto $\mathbf{S}$ è un sottoinsieme casuale e numerabile di punti in un insieme limitato D. Durante la trattazione, assumeremo sempre che $D \subseteq \mathbb{R}^d$.
Per coerenza con i casi di studio reali che andremo a considerare, ci concentriamo sui processi di punto $\mathbf{S}$ le cui realizzazioni sono sottoinsiemi localmente finiti di D. Formalmente, per ogni sottoinsieme $A \subseteq D$, $n(A)$ denota il cardinalità di A, ponendo $n(A) = \infty$ se A non è finito.
In questo capitolo introdurremo definizioni, proposizione, teoremi e concetti della teoria dei processi di punto che saranno utili in seguito. Partiremo dalla definizione di processo di punto, per poi definire alcuni casi specifici come i processi di punto binomiali, dai quali poi arriveremo a definire i processi di punto di Poisson. Faremo

poi una breve digressione sui Campi Aleatori Gaussiani, necessari per la definizione dei processi di punto di Cox, dove la funzione di intensità è un campo aleatorio, concentrandoci in particolare sul caso Log-Gaussian Cox Process, dove il logaritmo della funzione di intensità è un campo gaussiano. Infine introdurremo alcuni concetti sulle misure dei momenti per un processo di punto utilizzando il Teorema di Campbell. Le nozioni definite in questo capitolo sono prese in misura maggiore dai libri Mollet et al. [1] e Banerjee et al. [2].

1.2 Definizione dei Processi di Punto

Sia (D, d) uno spazio metrico. Sia $\mathcal{B}$ la σ-algebra di Borel associata a D e $\mathcal{B}_0 \subset \mathcal{B}$ la classe degli insiemi di Borel limitati in $\mathcal{B}$.
Per ogni sottoinsieme $\mathbf{s} \subset D$, definiamo la cardinalità di $\mathbf{s}$ come $n(\mathbf{s})$, che poniamo $n(\mathbf{s}) = \infty$ se $\mathbf{s}$ è non finito. Quindi avremo che $n(\mathbf{s}) < \infty$ se e solo se $\mathbf{s} = \{s_1, \ldots, s_k\}$, dove $k \in \mathbb{N}$ e $s_i \in D$, per $i = 1, \ldots, k$.
Definiamo allora:

$$N_{lf} = \{\mathbf{s} \subset D : n(\mathbf{s} \cap B) < \infty, \forall B \subset \mathbf{S} \textit{ limitato}\}$$

A questo punto possiamo definire una σ-algebra su N_{lf}:

$$\mathcal{N}_{lf} = \sigma(\{\mathbf{s} \in N_{lf} : n(\mathbf{s} \cap B) = m\} : B \in \mathcal{B}_0, m \in \mathbb{N}_0)$$

Definizione 1.2.1 (Processo di Punto) *Un processo di punto* $\mathbf{S}$ *definito su* D *è una funzione misurabile definita su uno spazio di probabilità* $(\Omega, \mathcal{F}, P)$ *che assume valori in* $(\mathcal{N}_{lf}, \mathcal{N}_{lf})$*. La distribuzione* P_S *di* $\mathbf{S}$ *è data da*

$$P_S(F) = P(\{\omega \in \Omega : \mathbf{S}(\omega) \in F\})$$

per $F \in \mathcal{N}_{lf}$*. .*

Sia $N(B) = n(\mathbf{S} \cap B)$, allora per la definizione precedente, la misurabilità di $\mathbf{S}$ implica che $N(B)$ è una variabile aleatoria per ogni $B \in \mathcal{B}$.
In particolare, la distribuzione di un point process $\mathbf{S}$ è determinata dalla distribuzione congiunta di $N(B_1), \ldots, N(B_m)$ per ogni $B_1, \ldots, B_m$ e per ogni $m \in \mathbb{N}_0$.
Per definire un processo di punto $\mathbf{S}$ sono necessari due ingredienti: una è la distribuzione per $N(D)$ e una collezione di funzioni di densità multivariata $\{f(s_1, \ldots, s_n)\}_{n=0}^{+\infty}$.
La prima è il numero di punti in D, che è una variabile aleatoria che assume valori nell'insieme $n \in \{0, 1, \ldots, \infty\}$. La seconda è, per ogni n, una densità multivariata su D^n, diciamo $f(s_1, s_2, \ldots, s_n)$. Chiameremo f una densità di posizione e, poiché i punti non sono ordinati/etichettati, f deve essere simmetrica nei suoi argomenti. Una conseguenza della simmetria di f è che possiamo scrivere la funzione di verosomiglianza come: [2]

$$L(\mathbf{S}) = P(N(D) = n)n!\ f(s_1, s_2, \ldots, s_n) \tag{1.1}$$

Definizione 1.2.2 (Eventi vuoti e probabilità di evento vuoto) *Insiemi del tipo* $F_B = \{s \in N_{lf} : n(s \cap B) = 0\}$*, dove* $B \in \mathcal{B}_0$ *sono detti* eventi vuoti.
Definiamo probabilità di evento vuoto *la funzione* $v : \mathcal{B}_0 \to [0,1]$ *tale che:*

$$v(B) = \mathbb{P}(N(B) = 0) \tag{1.2}$$

per ogni $B \in \mathcal{B}_0$.

Teorema 1.2.1 *Un Point Process* $\mathbf{S}$ *è unicamente determinato dalla sua probabilità di evento vuoto.*

1.3 Processo di Punto di Poisson

I Processi di Punto di Poisson sono un particolare tipo di processi di punto, utili a modellizzare i casi di *non dipendenza* tra i punti della realizzazione $\mathbf{s}$ e di *Complete Spatial Randomness* (CSR).
Illustriamo il concetto portando come esempio un processo di punto spaziale in cui mancano queste condizioni: gli epicentri dei terremoti in sismologia. Essi sono dei fenomeni naturali che sembrano perfetti per essere modellizzati come dei processi di punto spaziali: siamo in grado di definirne una posizione spaziale ben precisa e, guardando lo storico dei terremoti in una particolare zona geografica, sembrano seguire pattern precisi. Se andiamo a considerare un intero evento sismico (sia le scosse principali che quelle di assestamento) il processo non sarà poissoniano, poichè alle scosse principali seguono delle scosse più deboli di intensità, dette scosse di assestamento, con epicentri vicini alla posizione della scossa principale. Tuttavia, se nel modello non si tenesse in conto delle scosse principali, allora l'approssimazione di processo di punto poissoniano sarebbe corretta, come dimostrato da J. K. Gardner & L. Knopoff [3]. Iniziamo introducendo il processo di punto binomiale per il quale conosciamo il numero di punto presenti all'interno di una regione di interesse B.

Definizione 1.3.1 (Processo di punto binomiale) *Sia* f *una funzione di densità definita su* $B \subset D$ *e sia* $n \in \mathbb{N}$*. Un processo di punto* $\mathbf{S}$ *consistente di* n *punti i.i.d. con densità* f *si dice* Processo di Punto Binomiale *e si indica come* $\mathbf{S} \sim Bin(B, n, f)$

Quindi, dato $\mathbf{S} \sim Bin(D, n, f)$, con D regione di interesse limitata, allora avremo che per ogni $A \subset D$ [4]:

$$N(A) = \sum_{i=1}^{n} \mathbf{1}\{S_i \in A\} \tag{1.3}$$

dove S_i sono i punti appartenenti all'insieme aleatorio $\mathbf{S}$. Ma allora la probabilità dell'evento $N(A) = k$, sarà data da:

$$\mathbb{P}(N(A) = n) = \sum_{i=1}^{n} \mathbb{P}(S_i \in A) \tag{1.4}$$

dove

$$\mathbb{P}(S_i \in A) = \int_A f(\eta)d\eta \tag{1.5}$$

e nel caso più semplice, in cui la funzione di densità è uniforme, $f(x) = \frac{1}{|B|}\mathbf{1}\{s \in B\}$ otteniamo [4] [1]:

$$\mathbb{P}(N(A) = n) = \frac{|A|}{|B|} \tag{1.6}$$

Un ingrediente fondamentale per la definizione di un processo di punto di Poisson è la funzione di intensità $\lambda : D \to [0, \infty)$, la quale è localmente integrabile per ogni $B \subset D$ limitato. In particolare possiamo andare a definire la misura di intensità (*intensity measure*), definita come:

$$\mu(A) = \int_A \lambda(\eta)d\eta \tag{1.7}$$

per ogni $B \subset S$ limitato [1].

Definizione 1.3.2 (Processo di punto di Poisson) *Un processo di punto* $\mathbf{S}$ *su* D *è un processo di punto di Poisson con funzione di intensità* $\lambda(s)$ *se soddisfa le seguenti proprietà:*

- *per ogni* $B \subset D$ *con* $\mu(B) < \infty$ *allora* $N(B) \sim Poisson(\mu(B))$*;*
- *per ogni* $n \in \mathbb{N}$ *e* $B \subset D$ *con* $0 < \mu(B) < \infty$ *allora* $\mathbf{S}_B \sim Binomial(B, n, f)$ *dove* $f(\eta) = \frac{\lambda(\eta)}{\mu(B)}$.

Il processo così descritto si indica con $\mathbf{S} \sim Poisson(D, \lambda)$. [1][2]

Rispetto a questa definizione è interessante osservare i legami tra λ, $N(B)$ e f. Infatti, fissato un sottinsieme $B \subset D$, si ha:

$$\mathbb{E}[N(B)] = \int_B \lambda(\eta)d\eta = \int_B \mu(B)f(\eta)d\eta = \mu(B)\int_B f(\eta)d\eta = \mu(B)$$

In pratica conoscendo il valore di $\lambda(x), \forall x \in D$, siamo in grado di valutare le area della regione di interesse S dove è più probabile osservare un punto s_i della realizzazione $\mathbf{s}$ del processo di Poisson. Questo risultato è interessante dal punto di vista applicativo, perchè nella costruzione di modelli statistici per fenomeni descritti da processi di punto, una delle domande tipiche a cui rispondere è *"Dove mi aspetto di trovare il maggior numero di campioni?"*
Successivamente potremmo chiederci *quali sono i fattori che contrubuiscono ad aumentare il valore della funzione di intensità in determinati punti della regione di interesse?*. Questa domanda ci porterà poi allo sviluppo di modelli in cui la funzione di densità è una funzione dipendenti da campi spaziali definiti nelle regione

di interesse (ad esempio la temperatura, l'illuminazione, ecc..), utilizzando la forma dei modelli GLM.
In particolare, conoscendo l'espressione di λ per ogni sottoinsieme B possiamo calcolare la probabilità di vuoto [1]:

$$v(B) = exp(-\mu(B))$$

Sia $F \subset N_{lf}$ allora avremo che:

$$\mathbb{P}(\mathbf{S}_B \in F) = \sum_{n=0}^{\infty} \frac{exp(-\mu(B)}{n!} \int_B \cdots \int_B \mathbf{1}\{\{s_1, \ldots, s_n\} \in F\} \prod_{i=1}^{n} \lambda(s_i) ds_1 \ldots ds_n \tag{1.8}$$

Questa espressione tuttavia risulta molto complicata e non sembra essere particolarmente utile per poter fare inferenza bayesiana. La seguente proposizione può dare un aiuto nel superare questo ostacolo, oltre a giustificare il perchè il processo di punto di Poisson sia un processo *senza interazione*: [1]:

Proposizione 1.3.1 *Sia* $\mathbf{S}$ *un processo di punti di Poisson su* $\mathbf{S}$*, allora i processi* $\mathbf{S}_{B_1}, \mathbf{S}_{B_2}, \ldots$ *sono indipendenti per insieme disgiunti* $B_1, B_2, \cdots \in D$

Una conseguenza immediata di questa proposizione è che, grazie a 1.3.1, è possibile riscriverla la congiunta n-dimensionale $f(s_1, \ldots, s_n)$ come [2]:

$$f(s_1, \ldots, s_n) = \prod_{i=1}^{n} f(s_i) = \prod_{i=1}^{n} \frac{\lambda(s_i)}{\mu(S)} \tag{1.9}$$

Ma allora come possiamo scrivere la funzione di verosomiglianza del processo di punto di Poisson $\mathbf{S}$ conoscendone una realizzazione $\mathbf{S}$ di n punti?
Data la definizione 1.3.2 e l'equazione 1.9, allora avremo che [2]:

$$L(\lambda(\mathbf{s}); \mathbf{s} \in D, s_1, \ldots, s_n) = e^{-\mu(S)} \prod_{i=1}^{n} \lambda(x_i) \tag{1.10}$$

Abbiamo così trovato un'espressione più semplice e maneggevole per la likehood di un processo di punto di Poisson.
Introduciamo ora i concetti di *Thinning* e di *Superposition*. Sono due operazioni base che possono essere sfruttare per la manipolazione dei modelli di questi processi. In particolare sono utili per definire la "somma" di diversi processi di punto sullo stesso spazio D e per generare simulazioni di processi di punto di Poisson non omogenei. Queste proprietà sono simili alle analoghe proprietà delle variabili aleatorie di Poisson e dei Processi di Poisson temporali. Le definizione e i teoremi illustrati in seguito sono tratti da [1].

Definizione 1.3.3 (Superposition di Processi di Punto) *Un'unione disgiunta* $\mathbf{S} = \cup_{i=1}^{n} \mathbf{S}_i$ *di processi di punto è detta* Superposition.

Definizione 1.3.4 (Thinning Poisson Process) *Sia* $p : D \to [0,1]$ *una funzione e sia* $\mathbf{S}$ *un processo di punto su* D*. Il processo di punto* $\mathbf{S}_{thin}$ *ottenuto includendo il punto* $\mathbf{s} \in boldS$ *in* $\mathbf{S}_{thin}$ *con probabilità* $p(\mathbf{s})$*. Il processo* $\mathbf{S}_{thin}$ *è definisce* indipendent thinning *si* $\mathbf{S}$*. Definiamo la funzione* p *come* probabilità di accettazione*.*

Illustreremo ora due proposizioni che garantiscono che i Processi di Punto di Poisson sono chiusi rispetto alla Superposition e all'Indipendent Thinning.

Proposizione 1.3.2 *Siano* $\mathbf{S}_i \sim Poisson(D, \lambda_i), \quad i = 1, \ldots, n$ *processi di punto di Poisson indipendenti tali che* $\lambda = \sum_i^n \lambda_i$ *sia localmente integrabile. Allora, con probabilità uno, il processo di punto* $\mathbf{S} = \cup_{i=1}^n \mathbf{S}_i$ *è un'unione disgiunta e* $\mathbf{S} \sim Poisson(D, \lambda)$*.*

Proposizione 1.3.3 *Il processo di punto* $\mathbf{S}_{thin}$ *ottenuto come Indipendent Thinng da un processo di punto di Poisson* $\mathbf{S} \sim Poisson(D, \lambda)$ *con funzione di probabilità di accettazione* $p(\cdot)$ *è un processo di punto di Poisson con funzione di intesità* $\lambda_{thi.n}(s) = \lambda(s)p(s), \quad s \in D$*.*

1.4 Gaussian Random Field

Prima di continuare la trattazione sui processi spaziali di punto è necessario introdurre i concetti di *Random Fields* e *Gaussian Random Fields*, facendo riferimento a [5] e l'appendice di [6].

Definizione 1.4.1 (Random Field) *Dato uno spazio di probabilità* $(\Omega, \mathcal{F}, P)$*, un random field* T *definito su* $\mathbb{R}^n$ *è una funzione tale che, per ogni* $s \in \mathbb{R}^n$*,* $T(s)$ *è una variabile aleatoria sullo spazio di probabilità* $(\Omega, \mathcal{F}, P)$*.*

Definizione 1.4.2 (Funzione di covarianza di un Random Field) *Sia* T *un campo aleatorio definito su* $\mathbb{R}^n$*.*
La funzione di covarianza $R_T(x, y)$ *di un campo aleatorio* T *è definita come*

$$R_T(x, y) = \mathbb{E}\left[(T(x) - \mathbb{E}[T(x)])\,(T(y) - \mathbb{E}[T(y)])\right].$$

Consideriamo ora un campo aleatorio T. Se la distribuzione congiunta

$$F_{x_1,\ldots,x_m}(z_1, \ldots, z_m) = P\left(T(x_1) \leq z_1, \ldots, T(x_m) \leq z_m\right)$$

è invariante sotto la traslazione

$$(x_1, \ldots, x_m) \to (x_1 + \tau, \ldots, x_m + \tau),$$

allora T si dice *stazionario* o omogeneo.
Per un campo aleatorio stazionario T, possiamo dimostrare che

$$\mathbb{E}[T(x)] = \mathbb{E}[T(0)]$$

e successivamente

$$R_T(x, y) = f(x - y)$$

per qualche funzione f.

Definizione 1.4.3 (Campo Aleatorio Gaussiano) *T è un campo aleatorio gaussiano se la distribuzione congiunta finita $F_{x_1,\dots,x_m}(z_1, \dots, z_m)$ è una multivariata normale per ogni $(x_1, \dots, x_m) \in \mathbb{R}^m$.*

Possiamo quindi costruire una campo aleatorio gaussiano (GF) definendo una funzione $\mu(x) : \mathbb{R}^n \to \mathbb{R}$ che indica il valore del campo in ogni punto $x \in \mathbb{R}^n$ e una funzione di correlazione $R : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$.

Nel nostro caso, come funzione di correlazione per il campo aleatorio gaussiano T, consideriamo la *funzione di correlazione di Matèrn*:

$$R_T(s_i, s_j) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\kappa \parallel s_i - s_j \parallel)^{\nu} K_{\nu}(\kappa \parallel s_i - s_j \parallel) \tag{1.11}$$

dove $\nu, \kappa > 0$. Nel 2011 Lindgren, Rue, and Lindström [6] hanno dimostrato che i campi aleatori gaussiani con funzione di covarianza di Matèrn sono soluzione della equazione differenziale stocastica a derivate parziali:

$$(\kappa^2 - \Delta)^{\frac{\alpha}{2}} T(s) = W(s), \quad x \in \mathbb{R}^d, \quad \alpha = \nu + \frac{d}{2}, \quad \kappa > 0, \quad \nu > 0 \tag{1.12}$$

dove $(\kappa^2 - \Delta)^{\frac{\alpha}{2}}$ è un operatore pseudodifferenziale. Il processo aleatorio W è rumore bianco gaussiano spaziale con varianza unitaria, Δ è il laplaciano

$$\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial x_i^2},$$

e la varianza marginale è

$$\sigma^2 = \frac{\Gamma(\nu)}{\Gamma\left(\nu + \frac{d}{2}\right)(4\pi)^{d/2}\kappa^{2\nu}}.$$

I Campi Gaussiani sono utili nella definizioni dei processi di punto di Cox, dove la funzione di intensità $\lambda(s)$ è è una funzione aleatoria per ogni punto $s \in D$

1.5 Log-Gaussian Cox Process

Un processo Log-Gaussian Cox (LGCP) è un modello stocastico utilizzato nell'analisi dei processi di punto in geostatistica. È particolarmente utile per descrivere distribuzioni spaziali di eventi che mostrano variabilità e dipendenze spaziali complesse. Il LGCP combina due componenti principali:

1. **Processo di Cox**: Una classe di processi di punto in cui l'intensità dei punti varia spazialmente secondo una funzione aleatoria [7].
2. **Campo Gaussiano Logaritmico**: La funzione aleatoria che determina l'intensità è modellata come un campo gaussiano, i cui valori sono poi trasformati tramite un'esponenziale, per garantire che l'intensità sia sempre positiva.

In pratica, il LGCP permette di modellare l'intensità dei punti come una funzione log-normale spazialmente variabile. Questo approccio è particolarmente potente perché può catturare sia la media sia la variabilità della densità dei punti, fornendo una rappresentazione flessibile delle strutture spaziali.

Formalmente, sia $Z(s)$ il campo aleatorio gaussiano che descrive l'intensità logaritmica del processo, dove s rappresenta una posizione nello spazio. L'intensità del processo di punto $\lambda(s)$ è data da:

$$\lambda(s) = \exp(Z(s)).$$

L'utilizzo del LGCP è rilevante in molte applicazioni, come la geologia per la modellizzazione dei giacimenti minerari, l'ecologia per la distribuzione delle specie, e l'epidemiologia per l'analisi della diffusione delle malattie. Grazie alla sua capacità di rappresentare complessi pattern spaziali, il LGCP è uno strumento prezioso per l'analisi geostatistica avanzata. Iniziamo introducendo dal punto di vista matematico i Cox Process:

Definizione 1.5.1 (Cox Process) *Supponiamo che $Z = \{Z(s) : s \in D\}$ sia un campo casuale non negativo tale che con probabilità uno, $s \to Z(s)$ sia una funzione localmente integrabile. Se la distribuzione condizionale di* $\mathbf{S}$ *data Z è un processo di Poisson su D con funzione di intensità Z, allora* $\mathbf{S}$ *è detto essere un processo di Cox guidato da Z [8].*

Andiamo ora a specificare la struttura della funzione di intensità, che verrà descritta dall'esponenziale di Campo Gaussiano.

Definizione 1.5.2 (Log-Gaussian Cox Process) *Sia* $\mathbf{S}$ *un processo di Cox su $\mathbb{R}^d$ guidato da $\Lambda = \exp(Z)$, dove Z è un campo gaussiano. Allora* $\mathbf{S}$ *è Log-Gaussian Cox Process(LGCP).*

Per generare un processo di Log Gaussian Cox, prima si ottiene una realizzazione $\lambda(s)$ del campo aleatorio $\Lambda(s)$ e poi si ottiene il processo di Poisson $\mathbf{S}$ condizionato a $\lambda(s)$ [2].
La distribuzione di $(\mathbf{S}, \Lambda)$ è completamente determinata dalla media e dalla funzione di covarianza:

$$\begin{aligned} m(s) &= \mathbb{E}[\Lambda(s)] \\ c(s_i, s_j) &= \mathrm{Cov}(\Lambda(s_i), \Lambda(s_j)) \end{aligned} \tag{1.13}$$

Supponiamo che $c(s_i, s_j) = c(s_i - s_j)$ sia invariante per traslazione e abbia la forma

$$c(\xi) = \sigma^2 r\left(\frac{\xi}{\alpha}\right) \tag{1.14}$$

dove $\sigma^2 > 0$ è la varianza, r è una funzione di correlazione e $\alpha > 0$ è un parametro di correlazione. Una funzione $r : \mathbb{R}^d \to [-1, 1]$ con $r(0) = 1$ per ogni $\xi \in \mathbb{R}^d$ è una funzione di correlazione per un campo gaussiano se e solo se r è positiva semidefinita. [1].
Una possibile implementazione è che la funzione di intensità sia un campo gaussiano con funzione di covarianza di Matèrn. Questa combinazione verrà utilizzata nel capitolo successivo per sviluppare i modelli che saranno proposti, studiati e analizzti.

1.6 Misure dei momenti per processi di punto

In questa sezione presenteremo alcuni momenti per i processi di punto, un risultato notevolmente elegante per il calcolo di queste quantità è il Teorema di Campbell. Per il caso di misura del primo ordine, supponiamo che la caratteristica sia indicata con $g(s)$, potrebbe interessarci il valore di questa caratteristica sommato sui punti in D, cioè $\sum_{s_i \in \mathbf{S}} g(s_i)$. Il Teorema di Campbell fornisce il valore atteso di questa variabile, cioè

$$\mathbb{E}_S\left[\sum_{s_i \in \mathbf{S}} g(s_i)\right] = \int_S g(x)\lambda(x)\,dx \tag{1.15}$$

Supponiamo che g sia una caratteristica che è una funzione di due argomenti, $g(s, s')$, ad esempio, la distanza tra s e s'. Supponiamo che siamo interessati al valore di questa caratteristica sommata su coppie di punti in S, allora, il Teorema di Campbell fornisce l'aspettativa di questa variabile su S, cioè

$$\mathbb{E}_{\mathbf{S}}\left[\sum_{\substack{s_i, s_j \in \mathbf{S} \\ i \neq j}} g(s_i, s_j)\right] = \iint g(x, x')\gamma(x, x')\,dx\,dx' \tag{1.16}$$

come mostrato in [2].
Le proprietà del primo e secondo ordine sono descritte dalla cosiddetta misura di intensità e dalla misura di momento fattoriale del secondo ordine.

Definizione 1.6.1 *La misura di intensità μ su $\mathbb{R}^d$ è data da*

$$\mu(B) = \mathbb{E}[N(B)], \quad B \subseteq \mathbb{R}^d,$$

e la misura di momento fattoriale del secondo ordine $\alpha(2)$ su $\mathbb{R}^d \times \mathbb{R}^d$ da

$$\alpha^{(2)}(C) = \mathbb{E}\left[\sum_{\substack{\xi,\eta\in\mathbf{S}\\ \xi\neq\eta}} 1\left[(\xi,\eta)\in C\right]\right], \quad C \subseteq \mathbb{R}^d \times \mathbb{R}^d.$$

Definizione 1.6.2 *Se la misura di intensità μ può essere scritta come*

$$\mu(B) = \int_B \rho(\xi)\, d\xi, \quad B \subseteq \mathbb{R}^d,$$

dove ρ è una funzione non negativa, allora ρ è chiamata funzione di intensità. Se ρ è costante, allora X è detto omogeneo o stazionario del primo ordine con intensità ρ; altrimenti X è detto non omogeneo.

In modo euristico, $\rho(\xi)\, d\xi$ è la probabilità dell'occorrenza di un punto in una sfera infinitesimale con centro ξ e volume $d\xi$. Per un processo puntiforme omogeneo, ρ è il numero medio di punti per unità di volume.

Definizione 1.6.3 *Se la misura di momento fattoriale del secondo ordine $\alpha(2)$ può essere scritta come*

$$\alpha^{(2)}(C) = \iint_C 1[(\xi,\eta)\in C]\rho^{(2)}(\xi,\eta)\, d\xi d\eta, \quad C \subseteq \mathbb{R}^d \times \mathbb{R}^d,$$

dove $\rho^{(2)}$ è una funzione non negativa, allora $\rho^{(2)}$ è chiamata densità di prodotto del secondo ordine.

In modo intuitivo, $\rho^{(2)}(\xi,\eta)\, d\xi d\eta$ è la probabilità di osservare una coppia di punti da X che si verifica congiuntamente in due sfere infinitesimali con centri ξ e η e volumi $d\xi$ e $d\eta$. Per studiare se un processo di punti si discosta dal processo di Poisson, è utile normalizzare la densità di prodotto del secondo ordine $\rho^{(2)}$ dividendo per $\rho(\xi)\rho(\eta)$.

Definizione 1.6.4 *Se sia ρ che $\rho^{(2)}$ esistono, la funzione di correlazione di coppia è definita da*

$$g(\xi, \eta) = \frac{\rho^{(2)}(\xi, \eta)}{\rho(\xi)\rho(\eta)}$$

Per un processo di Poisson possiamo prendere $\rho^{(2)}(\xi, \eta) = \rho(\xi)\rho(\eta)$ in modo che $g(\xi, \eta) = 1$. Se ad esempio $g(\xi, \eta) > 1$, questo indica che le coppie di punti hanno più probabilità di occorrere insieme alle posizioni ξ, η rispetto a un processo di Poisson con la stessa funzione di intensità di $\mathbf{S}$. [1]
Il modo più elementare per investigare la casualità spaziale completa (CSR) procede dalla definizione di un processo puntiforme omogeneo (HPP), in particolare, sotto questa ipotesi, le caratteristiche del processo hanno forme esplicite semplici [2].
Consideriamo il processo sull'intero $\mathbb{R}^2$, cioè un modello di punti infinitamente numerabili. Consideriamo la variabile casuale $N(s, d; \mathbf{S})$, dove $s \in \mathbf{S}$, $\partial_d s$ è un cerchio di raggio d centrato in s, e N conta il numero di punti nel cerchio da S, escludendo s. Per stazionarietà, $N(s, d; \mathbf{S})$ segue $N(0, d; \mathbf{S} - s)$, dove $\mathbf{S} - s$ è la traslazione di $\mathbf{S}$ per s. Questo risultato distribuzionale è equivalente a dire che ogni punto in $\mathbf{S}$ è un punto tipico, nel senso che ciascuno può essere considerato equivalente a 0 sotto la traslazione. Sotto l'ipotesi di un insieme limitato D, consideriamo

$$\mathbb{E}_X \left[\sum_{s_i \in \mathbf{S}, X \subset D} 1(N(s_i, d; \mathbf{S}) > 0) \right] = \lambda |D| P(N_D(s, d; \mathbf{S}) > 0)$$

dove $P(N(s, d; \mathbf{S}) > 0)$ è indicato con $G(d)$ e può essere calcolata tramite il teorema di Campbell. Vediamo che $G(d)$ aumenta con d e, infatti, può essere considerato come una funzione di distribuzione cumulativa nella distanza d. Un'alternativa a $G(d)$ nella letteratura è $F(d)$, dove ora $N(s, d; \mathbf{S})$ assume che s non sia in $\mathbf{S}$. Sì può pensare a $G(d)$ come alla distribuzione del "punto nell'intorno più vicino", cioè la funzione di distribuzione cumulativa della distanza più vicina tra eventi, cioè in un evento osservato, $G(d) = \mathbb{P}(\text{vicino più vicino} \leq d)$. Allo stesso modo, si può pensare a $F(d)$ come alla distribuzione dello "spazio vuoto", cioè per una posizione arbitraria, la funzione di distribuzione cumulativa della distanza più vicina tra un punto e un evento, $F(d) = \mathbb{P}(\text{vicino più vicino} \leq d)$. Sotto CSR, $G(d) = F(d) = 1 - \exp(-\lambda \pi d^2)$. A questo due misure possiamo associare una terza, $K(d)$ che considera il numero medio di punti nella regione di raggio d da un punto arbitrario, che è facilmente calcolabile sotto ipotesi CSR. [2]

$$K(d) = \frac{1}{\lambda} \mathbb{E}[\text{numero di punti entro } d \text{ da un punto arbitrario}] = \pi d^2 \qquad (1.17)$$

Quindi, da un punto di vista statistico-descrittivo, per valutare se per il processo di punto di interesse è valida l'ipotesi CSR, possiamo confrontare le misure empiriche di F, G e K come quelle attese. [2]

Capitolo 2

Inferenza di Log-Gaussian Cox Process con INLA

Nell'inferenza bayesiana l'algoritmo classico, il più conosciuto e il più utilizzato è *Monte Carlo Markov Chain* (MCMC), che grazie all'aumento della potenza di calcolo dei computer, ha permesso un sempre maggiore utilizzo e sviluppo dei modelli di statistica bayesiana, in particolare grazie a modelli classici come Gibbs Sampler o Metropolis-Hastings. In particolare gli algoritmi MCMC vengono utilizzati per ottenere le distribuzioni a posteriori tramite approssimazione numerica degli integrali. Gli algoritmi MCMC sono molto flessibili e potenzialmente sono in grado di trattare qualsiasi tipo di dato e di modello. In pratica però spesso risultano molto lenti: in particolare quando il numero di componenti da stimare sono molti (*course of dimensionality*); inoltre, essendo un metodo stocastico, soffre di problemi di convergenza.
INLA è un algoritmo di statistica bayesiana per *Latent Gaussian Models* (LGM) presentato da Rue, Martino e Chopin [9] nel 2009. A differenza di MCMC, INLA è un algoritmo deterministico e invece di restituire la distribuzione a posteriori congiunta del modello, restituisce delle distribuzione a posteriori marginali per i parametri e gli iperparametri di interesse.
In questo capitolo illustremo brevemente la struttura e le caratteristiche dei LGM e perchè siano particolarmente agli approcci numerici diretti. Successivamente illustreremo lo schema computazionale di INLA e termineremo presentando due approcci utilizzati per fare inferenza sui LGCP con INLA.

2.1 Latent Gaussian Models

I Latent Gaussian Models sono la classe di modelli bayesiani di nostro interesse, in particolare, come illustrato da Martino e Riebler [10], un modello LGM per INLA

consiste di tre elementi:

1. un modello statistico per la likehood;
2. un Latent Gaussian Field (LGF);
3. un vettore di iperparametri

Come nei modelli lineari generalizzati supponiamo di avere qualche predittore η_i legato all'osservazione y_i tramite una link function:

$$\eta_i = \mu + \sum_j \beta_j z_{ij} + \sum_k w_k f_k(u_{ik}) \tag{2.1}$$

dove: μ è l'intercetta generale, $\mathbf{z}$ sono le covariate con effetto lineare β e $\mathbf{w}$ sono dei pesi conosciuti associati agli effetti random $\mathbf{f^k}$ associati alle covariate $\mathbf{u}$. Il latent Gaussian field sarà allora $\mathbf{x} = \{\eta, \mu, \mathbf{f^1}, \mathbf{f^2}, \dots\}$. Il vettore di iperparametri $\mathbf{\Theta}$ può apparire sia nel modello della likehood sia nel latent field. In particolare il modello può essere riscritto come:

$$\begin{aligned} y|x,\theta &\sim Y_\pi(y_i|\eta_i, \mathbf{Q}(\theta))), \\ x|\theta &\sim \mathcal{N}(0, (\mathbf{Q^{-1}}(\theta)))), \\ \theta &\sim \pi(\theta). \end{aligned} \tag{2.2}$$

dove $\mathbf{Q}(\theta))$ è la matrice di precisione per il latent Gaussian field. In Rue et all. [9] e Martino et all. [10] viene suggerito che la matrice di precisione $\mathbf{Q}(\theta))$ debba essere sparsa. In particolare, la struttura della matrice di precisione $\mathbf{Q}(\theta)$ dipenderà dalla scelta delle priori per $\mathbf{f^k}$.
Molti dei modelli comunemente usati come priori per i termini f_k appartengono alla classe dei cosiddetti campi aleatori di Markov gaussiani (GMRF). I GMRF possono essere utilizzati per modellare effetti di una covariata, effetti casuali, errori di misura, dipendenze temporali e così via. La dipendenza spaziale continua può essere specificata utilizzando il cosiddetto approccio SPDE [6], creando una rappresentazione GMRF approssimata del campo di covarianza di Matérn basata su equazioni differenziali parziali stocastiche. I GMRF sono modelli gaussiani dotati di proprietà di Markov. Queste, a loro volta, sono collegate alla struttura degli elementi non nulli della matrice di precisione nel senso che, se due elementi del campo sono *condizionatamente indipendenti* dati tutti gli altri, allora l'elemento corrispondente della matrice di precisione è uguale a zero, cioè:

$$y_i|\mathbf{x} \perp y_j|\mathbf{x}, \quad i \neq j$$

In pratica, la scelta di un GMRF per i termini f^k comporta che la matrice di precisione $\mathbf{Q}(\theta)$ sia sparsa. La distruzione posteriore congiunta sarà:

$$\pi(x, \theta|y) \propto \exp\left(-\frac{1}{2}x^T Q(\theta)x + \sum_i \log(\pi(y_i|\eta_i, \theta)) + \log \pi(\theta)\right) \tag{2.3}$$

Per riassumere, INLA può essere applicato ai LGM che soddisfano le seguenti ipotesi [10]:

1. Ogni punto dato dipende solamente dal predittore lineare corrispondente, in modo che la verosimiglianza possa essere scritta come:
$$y|x,\theta \sim \prod_i \pi(y_i|\eta_i,\theta).$$
2. Il campo latente x deve essere dotato di alcune proprietà di indipendenza condizionale (Markov) in modo che la matrice di precisione $Q(\theta)$ sia sparsa.
3. Il predittore lineare dipende linearmente dalla funzione liscia sconosciuta delle covariate.
4. Il risultato finale saranno le marginali posteriori univariati $\pi(x_i|y)$ e $\pi(\theta_j|y)$ piuttosto che nel posteriore congiunto $\pi(x,\theta|y)$.

2.2 Schema computazionale di INLA

Prima di illustrare lo schema computazionale utilizzato in INLA, è utile illustrare due risultati, che possono essere considerati gli ingredienti chiave di questo algoritmo. Il primo ingrediente è la definizione di *probabilità condizionata*; partendo dal Teorema di Bayes:
$$\pi(x|z) = \frac{\pi(x,z)}{\pi(z)} \Rightarrow \pi(z) = \frac{\pi(x,z)}{\pi(x|z)}$$
Se aggiungiamo una terza variabile w rispetto alla quale condizionare z otteniamo:
$$\pi(z|w) = \frac{\pi(x,z|w)}{\pi(x|z,w)} \tag{2.4}$$
Il secondo ingrediente è l'approssimazione di Laplace, descritta in [11]. Dato il seguente integrale per una funzione f integrabile:
$$\int f(x)\,dx = \int \exp[\log f(x)]\,dx$$
mediante uno sviluppo in serie di Taylor attorno al massimo:
$$x^* = \arg\max_x \log f(x)$$
Si ottiene:
$$\int f(x)\,dx \approx \int \exp\left(\log f(x^*) + \frac{(x-x^*)^2}{2}\left.\frac{\partial^2 \log f(x)}{\partial x^2}\right|_{x=x^*}\right) dx \tag{2.5}$$

Impostando $\sigma_*^2 = -\left.\frac{1}{\frac{\partial^2 \log f(x)}{\partial x^2}}\right|_{x=x^*}$, possiamo riscrivere:

$$\int f(x)\,dx \approx f(x^*) \int \exp\left(-\frac{(x-x^*)^2}{2\sigma_*^2}\right) dx \tag{2.6}$$

Quindi, sotto l'approssimazione di Laplace, $f(x) \approx \text{Normale}(f(x^*), \sigma_*^2)$.
Concentriamoci ora sullo schema computazionale di INLA. Partendo dalle equazioni per ricavare le marginali di interesse:

$$\pi(\theta_j|y) = \int\int \pi(x,\theta|x)\,dx\,d\theta_{-j} = \int \pi(\theta|y)\,d\theta_j \tag{2.7}$$

$$\pi(x_i|y) = \int\int \pi(x,\theta|x)\,dx_{-i}\,d\theta = \int \pi(x_i|\theta,y)\pi(\theta|y)\,d\theta \tag{2.8}$$

Concentriamo sul termine $\pi(\theta|y)$, usando (2.4), possiamo riscriverlo come:

$$\pi(\theta|y) = \frac{\pi(x,\theta|y)}{\pi(x|\theta,y)} \propto \frac{\pi(y|x,\theta)\pi(x|\theta)\pi(\theta)}{\pi(x|\theta,y)} \tag{2.9}$$

Valutiamo l'espressione per specifici valori θ^k, sfruttando l'approssimazione di Laplace introdotta in [11] ed illustrata in precedenza, per l'approssimazione gaussiana di $x|\theta^k, y$, sfruttando la moda, otteniamo:

$$\pi(\theta^k|y) \propto \frac{\pi(y|x,\theta^k)\pi(x|\theta^k)\pi(\theta^k)}{\pi_G(x|\theta^k,y)} \tag{2.10}$$

Per l'equazione (2.8) è necessario ottenere un'approssimazione della marginale $\pi(x_i|\theta^k, y)$. In INLA sono presenti tre possibili opzioni di approssimazione [10]:

- marginalizzare partendo dall'approssimazione $\pi_G(x|\theta^k, y)$ ottenuta in precedenza, ma è sconsigliata dagli autori;
- riscrivere la marginale come:

  $$\pi(x_i|\theta_k, y) = \frac{\pi(x|\theta,y)}{\pi(x_{-i}|x_i,\theta,y)} \propto \frac{\pi(y|x,\theta)\pi(x|\theta)\pi(\theta)}{\pi(x_{-i}|x_i,\theta,y)}$$

  e riscrivere il denominatore utilizzando l'approssimazione di Laplace già sfruttata in precedenza;
- utilizzare l'approssimazione di Laplace semplificata proposta in [9], la quale consiste nell'applicazioni di termini correttivi tramite un'espansione in serie di Taylor al terzo ordine dell'approssimazione proposta da [11]

In conclusione, il metodo di calcolo INLA è il seguente:

1. Esplora lo spazio di θ attraverso l'approssimazione di $\pi_e(\theta|y)$. Trova la moda di $\pi(\theta|y)$ e individua una serie di punti $\{\theta^1, ..., \theta^K\}$ nell'area di alta densità di $\pi(\theta|y)$.

2. Per i K punti di supporto selezionati, calcola $\pi(\theta^1|y), ..., \pi(\theta^K|y)$ utilizzando l'equazione (2.10).

3. Per ciascun punto θ^k selezionato, approssima la densità di $x_i|\theta, y$ come $\pi(x_i|\theta_k, y)$ per $k = 1, ..., K$ utilizzando una delle tre possibili approssimazioni: Laplace, Laplace Semplificata o Gaussiana.

4. Risolvi l'equazione (2.8) tramite integrazione numerica come:

$$\pi_e(x_i|y) = \sum_{k=1}^{K} \pi_e(x_i|\theta_k, y)\pi_e(\theta_k|y)\Delta_k$$

2.3 Inferenza bayesiana di Processi di Punto di Cox Log-Gaussiani

I metodi standard per adattare i processi di Cox sono computazionalmente dispendiosi e i metodi di catena di Markov Monte Carlo comunemente usati sono difficili da regolare per questo problema.

2.3.1 Regular Lattice Approach

Un metodo comune per fare inferenza su un LGCP potrebbe essere adottare la soluzione proposta nel 2012 da Illian et al. [12], i quali hanno sviluppato un framework rapido e flessibile per adattare i processi di Cox log-normali utilizzando INLA.
Si discretizza l'area di osservazione $D \subset \mathbb{R}^2$ in $N = n_{row} \times n_{cols}$ celle $\{A_{i,j}\}$, di area $|A_{i,j}|$, per $i = 1, \ldots, n_{row}$ e $j = 1, \ldots, n_{col}$. Indichiamo con $y_{i,j}$ il numero di punti della realizzazione processo $\mathbf{S}$ che si trovano nella cella $A_{i,j}$. Detta la quantità:

$$\lambda_{i,j} = \int_{A_{i,j}} \lambda(\mathbf{s})d\mathbf{s} \tag{2.11}$$

Possiamo allora definire:

$$y_{i,j}|\lambda_{i,j} \sim Poisson(\lambda_{i,j}) \tag{2.12}$$

È impossibile calcolare l'intensità totale per ogni cella e pertanto utilizziamo l'approssimazione $\Lambda_{ij} \approx |A_{ij}| \exp(z_{ij})$, dove z_{ij} è un valore rappresentativo del campo

aleatorio $Z(s)$ all'interno della cella A_{ij} e $|A_{ij}|$ è l'area della cella s_{ij}.
Simpson el all. [13] mostrano varie criticità di questo approcio, in particolare riguardo alla problematica del reticolo che viene usato nella discretizzazione del dominio.
La sfida computazionale è che, se $Z(s)$ è un campo aleatorio gaussiano, il vettore gaussiano multivariato z che contiene i z_{ij} avrà una matrice di covarianza densa. La complessità computazionale risultante limita questo metodo a reticoli piuttosto piccoli.
Questi metodi sono insoddisfacenti poiché il reticolo computazionale ha due ruoli fondamentalmente diversi. Il primo è quello di approssimare il campo aleatorio gaussiano latente $Z(s)$. Il secondo è quello di approssimare le posizioni dei punti. Chiaramente, più fine è il reticolo, meno informazione si perde, quindi la qualità dell'approssimazione della verosimiglianza dipende principalmente dalla dimensione della griglia.
Si parte costruendo un'approssimazione di Poisson alla funzione di verosimiglianza del Log-Gaussian Cox Process. Se la griglia è abbastanza fine e il campo gaussiano latente è discretizzato in modo appropriato, questa approssimazione è abbastanza buona, ma può essere dispendiosa dal punto di vista computazionale, specialmente quando l'intensità del processo è alta o la finestra di osservazione è grande o di forma strana.

2.3.2 SPDE approach

Un altro metodo, più efficiente, è stato proposto da Simpson et all. [13]. L'idea chiave della loro proposta è di individuare una equazione differenziale stocastica la cui soluzione sia il campo aleatorio $Z(s)$, che identifica la funzione di intensità nei processi di Cox.
Partiamo dall'equazione della funzione di verosomiglianza per un LGCP condizionata alla funzione d'intensità $\lambda(x)$, come in (1.10):

$$L(\mathbf{S}|\lambda) = Ce^{-\mu(D)} \prod_{s_i \in \mathbf{S}}^{n} \lambda(s_i) \tag{2.13}$$

Per un LGCP $\mathbf{S}$, come visto in precedenza, la funzione di intensità è un campo aleatorio gaussiano, cioè $log(\Lambda(s)) = Z(s)$, dove $Z(s)$ è un GF, per cui il processo condizionato $\mathbf{S}|\lambda$ è un NHPP.
Partendo dal lavoro in [6], si costruisce un GF con la seguente struttura:

$$Z(s) = \sum_{i=1}^{n} z_i \varphi_i(s) \tag{2.14}$$

come visto nella sezione (1.4), dove $\mathbf{Z} = \{z_1, \ldots, z_n\}$ provengono da una normale multivariata. La log-verosomiglianza $logL(\mathbf{S}|Z) = -\int_\Omega \exp Z(s)ds + \sum_{i=1}^{N} Z(s_i)$ è composta da due termini: l'integrale stocastico e la valutazione del campo nei punti dati. Mentre i modelli di equazioni differenziali parziali stocastiche specificati in modo continuo ci consentono di calcolare il termine di somma esattamente, dobbiamo approssimare l'integrale con una somma. Consideriamo una regola di integrazione deterministica della forma generale:

$$\int_\Omega f(s)ds \approx \sum_{i=1}^{p} \tilde{\alpha}_i f(\tilde{s}i) \tag{2.15}$$

per nodi fissi e deterministici $\{\tilde{s}_i\}_{i=1}^p$ e pesi $\{\tilde{\alpha}_i\}_{i=1}^p$. Utilizzando questa regola di integrazione, possiamo costruire l'approssimazione

$$\log L(\mathbf{S}|Z) \approx C - \sum_{i=1}^{p} \tilde{\alpha}_i \exp\left(\sum_{j=1}^{n} z_j \varphi_j(\tilde{s}_i)\right) + \sum_{i=1}^{N}\sum_{j=1}^{n} z_j \varphi_j(s_i) \tag{2.16}$$

$$= C - \tilde{\alpha}^T exp(A_1 z) + 1^T A_2 z \tag{2.17}$$

dove C è una costante dipendente dall'area di ricerca, $[A_1]_{ij} = \phi_j(\tilde{s}_i)$ è una matrice che contiene i valori del modello gaussiano latente nei nodi di integrazione $\{\tilde{s}_i\}$, e $[A_2]_{ij} = \phi_j(s_i)$ valuta il campo gaussiano latente nei punti osservati $\{s_i\}$.
Il vantaggio dell'approssimazione appena ottenuta è che essa è di forma Poisson. In particolare, dati z e θ, la verosimiglianza approssimata consiste in $N + p$ variabili casuali di Poisson indipendenti. Per vedere questo, scriviamo $\log \eta = (exp(z^T A_1^T), z^T A_2^T)^T$ e $\alpha = (\tilde{\alpha}^T, 0_{N\times 1}^T)^T$. Poi, se costruiamo alcune pseudo-osservazioni $y = (0_{p\times 1}^T, 1_{N\times 1}^T)^T$, la verosimiglianza approssimata si fattorizza come

$$\pi(y|z) \approx C \prod_{i=1}^{N+p} \eta_i^{y_i} e^{-\alpha_i \eta_i}, \quad (4)$$

che è simile alla verosimiglianza per l'osservazione di $N + p$ variabili casuali di Poisson condizionatamente indipendenti con medie $\alpha_i \eta_i$ e valori osservati y_i [13].
Riusciamo così ad ottenere un'espressione della likehood della forma utile all'utilizzo di INLA.
Questa approssimazione funziona con qualsiasi campo aleatorio finito dimensionale del tipo (2.14). E' possibile trovare un'approssimazione finito dimensionale di un qualsiasi GRF in modo da poter sfruttare l'approssimazione illustrata prima per la likehood di Gaussian Cox Process?
In Lindgren et al. [6], i ricercatori sono riusciti a ottenere importanti risultati per l'approssimazione di un GRF e rielaborata poi successivamente in [13].
L'idea di base è ottenere una soluzione approssimata nella forma (2.14) della SPDE

di Matèrn (1.12).
Se supponiamo che le funzione $\{\phi_i(\cdot)\}$ siano funzione lineari a tratti e che definiscano una base in $L^2(D)$ e che $\alpha = 2$, allora otteremo come soluzione debole della SPDE:

$$(\kappa^2 C + G - B)z \sim N(0, C) \tag{2.18}$$

dove B, C e G sono matrici sparse con elementi

$$C_{ii} = \int_\Omega \phi_i(s)\, ds, \quad G_{ij} = \int_\Omega \nabla\phi_i(s)\nabla\phi_j(s)\, ds, \quad B_{ij} = \int_{\partial\Omega} \phi_i(s)\partial_n\phi_j(s)\, ds.$$

Il bordo di Ω è $\partial\Omega$, mentre $\partial_n\phi_j(s)$ è la derivata normale di $\phi_j(s)$ e C è diagonale.
Un aspetto interessante dell'uso di modelli a equazioni differenziali parziali stocastiche su un campo aleatorio gaussiano a dimensione finita è che la distribuzione a priori converge man mano che la griglia viene affinata [6]. Questo è distinto dai processi predittivi o dagli approcci di kriging a rango fisso, dove il modello a dimensione finita è considerato il vero modello sottostante.

Teorema 2.3.1 (Risultato su Convergenza per approccio SPDE) *Supponiamo che la finestra di osservazione sia un poligono convesso in $\mathbb{R}^2$ e sia h la lunghezza massima del lato nella mesh. Sia $Z(\cdot)$ la soluzione di* (1.12) *con $\alpha = 2$. Se $G(\cdot)$ è una funzione uniformemente lipschitziana continua, allora l'errore nella speranza a posteriori $\mathbb{E}\{G(Z) \mid y\}$ dovuto all'approssimazione dell'equazione differenziale stocastica parziale è dell'ordine $h^{1-\epsilon}$ per qualsiasi $\epsilon > 0$.*

Come conseguenza abbiamo che al raffinamento della mesh, indicato dal parametro h, la soluzione approssimata tramite FEM converge alla soluzione dell'equazione (1.12).

Capitolo 3

Oloturie

Le oloturie di mare sono un gruppo di specie marine diffuse nelle regioni tropicali e subtropicali, dove loro pesca ha comportato una drastica riduzione della disponibilità e l'interesse si è spostata verso le specie che abitano il Mediterraneo [14].
Di conseguenza è aumentato l'interesse su queste specie marine, in modo da poter pianificare un'adeguata gestione delle popolazioni e dagli ambienti marini di interesse. Lo scopo dei modelli implementati è valutare l'influenza di alcune covariate spaziali (profondità e pendenza del fondale, temperatura e presenza di posidonia sul fondale) sulla popolazione di oloturie. In [15] sono stati proposti dei modelli di processi di punto utilizzando INLA, sfruttando i dati raccolti nelle prime 5 campagne di esplorazione. I dati ottenuti in queste prime campagne di esplorazione sono stati limitati dalla tecnica utilizzata di campionamento utilizzata, non essendo possibile valutare la presenza di esemplari di oloturie nascoste dalla presenza di Posidonia. Successivamente sono state effettuate altre due campagne di esplorazione in cui i ricercatori hanno raccolto dati sulla presenza di oloturie solamente nelle zone in cui erano presenti campi di posidonia, spostando con cura le foglie di Posidonia identificando le specie marine nascoste.
La differente tecnica di campionamento delle ultime due campagne rende necessario modificare i modelli presentati in [15]. I modelli implementati proposti il fatto che le aree di ricerca tra le due metodologie sono diverse. In particolare, l'integrazione dei dati provenienti dai due tipi di campionamento, permette di valutare l'influenza della presenza di Posidonia sulla popolazione di Oloturie, che non era stato possibile nello studio di Mastroantonio et al. [15].
In questo capitolo inizieremo introducendo *inalbru* il wrapper di RINLA progettato per applicazione di statistica spaziale, che dovrebbe rendere la scrittura di modelli meno complessa. Successivamente viene presentata un'analisi sui dati raccolti durante le campagne di campionamento, l'interpolazione delle covariate spaziali di interesse e la generazione della mesh necessaria ad approssimare la funzione di intensità del processo descritta come un campo di Matèrn. Illusteremo poi la tecnica

utilizzata per generare delle realizzazione di campi gaussiani e processi di punto di Poisson. Nelle due sezioni successive vengono presentati i due modelli proposti: per ciascuno di essi viene brevemente presentata la struttura della funzione di intensità, successivamente si analizzino i risultati ottenuti da dei test per valutare il comportamento del modello e successivamente si analizzano i risultati ottenuti applicando il modello ai dati reali. Infine si conclude riassumendo i risultati ottenuti e facendo le considerazioni finali.

3.1 Introduzione al wrapper inlabru

Come detto in precedenza, INLA è un algoritmo che permette di fare inferenza bayesiana molto più velocemente rispetto agli algoritmo MCMC per i Latent Gaussian Models. Il pacchetto R in cui viene implementato INLA, detto RINLA richiede all'utente di conoscere lo schema di approssimazione della likehood e non permette di fare inferenza quando la *detection probability* è sconosciuta, fatto comune nel caso di studi ecologici. Il pacchetto *inlabru* rende più lo sviluppo di modelli spaziali utilizzando INLA, semplificando la sintassi utilizzata [16].
inlabru basa la costruzione del modello sul concetti di *componenti*: come visto nel Capitolo (2), osserviamo che le variabili latenti del modello sono legate alle osservazioni y_i tramite un qualche predittore η_i. In *inlabru*, si mette in evidenza che, date le variabili latenti $\mathbf{x}$, possiamo definire una matrice non stocastica $\mathbf{A}$, detta *model design matrix* tale che $\eta_i = \sum_j A_{i,j} x_j$, che possiamo riscrivere in notazione vettoriale, mettendo in evidenza la dipendenza tra il predittore lineare e le variabili latenti:

$$\eta(\mathbf{x}) = \mathbf{A}\mathbf{x} \tag{3.1}$$

Le osservazioni $\mathbf{x}$ si suppone siano condizionatamente dipendenti da η e θ.
La matrice $\mathbf{A}$ può essere decomposta come $\mathbf{A} = [\mathbf{X}\mathbf{Z}]$, dove $\mathbf{X}$ contiene informazioni sugli effetti fissi, mentre la matrice $\mathbf{Z}$ contiene informazioni sugli effetti random delle osservazioni.
Nel contesto dei modelli lineari gerarchici bayesiani, la distinzione tra effetti fissi e casuali è in qualche modo arbitraria. Possiamo pensare alla matrice di design come scomposta in sotto-matrici separate per ciascun componente del modello, cioè

$$\mathbf{A} = \begin{pmatrix} A^{(1)} & \dots & A^{(d)} \end{pmatrix}$$

per un modello con d componenti, considerando ciascun componente del modello come avente la propria "matrice di design del componente", denotata qui come $A^{(j)}$ per $j = 1, \dots, d$. Il vettore dei parametri può essere partizionato in

$$\mathbf{u} = \begin{pmatrix} u^{(1)} \\ \vdots \\ u^{(d)} \end{pmatrix},$$

dove ciascun $u^{(j)}$ è il vettore dei parametri per il componente j. Data questa decomposizione, l'espressione del predittore per la i-esima osservazione può essere scritta come

$$\eta(u)_i = \delta_i + \sum_{j=1}^{d} A_i^{(j)} u^{(j)},$$

dove $A_i^{(j)}$ è la i-esima riga di $A^{(j)}$ e δ_i è un termine opzionale di offset costante [17].

Ad esempio, la matrice di design del componente per un parametro di intercetta è un vettore colonna di uni e la matrice di design del componente per l'effetto di una covariata nota è un vettore colonna contenente le informazioni della covariata per ogni osservazione. La matrice di design del componente per un effetto gaussiano iid è la matrice identità (o una matrice diagonale a blocchi composta da sottomatrici identità della dimensione corretta per ciascun gruppo). La matrice di design del componente per un effetto SPDE contiene le funzioni base degli elementi finiti valutate in ciascuna posizione di osservazione. L'idea chiave è che ogni componente del predittore del modello può essere scritto come una matrice nota moltiplicata per un vettore di parametri [17].

Partendo da queste assunzioni, *inlabru* riesce a costruire automaticamente la *Model Design Matrix*. Invece utilizzando RINLA, dovrebbe essere l'utente stesso a costruire la matrice del modello. Da un punto di vista di nomenclatura, la submatrice legate ad effetti SPDE viene chiamata *proiettore*, perchè l'idea è che quella matrice proietti i valori del GRMF sui punti dell'osservazioni [17].

In particolare, in Bach et al. [16], gli autori illustrano come utilizzare *inlabru* per modelli di conteggi spaziali e per processi di punto Log-Gaussian Cox.

Per l'utilizzo di *inlabru* è necessario disporre dei seguenti:

1. dataset contenente le posizioni degli esemplari di oloturie osservati, che rappresentano le realizzazione del processo di punto che si vuole modellizzare;

2. mappe dei campi spaziali che descrivono le covariate spaziali utili a modellizzare la funzione di intensità del processo di punto;

3. una mesh sull'area di ricerca di interesse, che permetta di ottenere un'approssimazione del GRF descrivente la funzione di intensità in un processo di Cox;

4. ulteriori dati necessari allo sviluppo di modelli (tipo di campionamento utilizzato e istante temporale delle osservazioni).

Il parametro ν nella funzione di covarianza di Matérn determina la differenziabilità quadratica media del campo, il che può influenzare le previsioni fatte dal modello. Poiché ν è difficile da identificare dai dati, è comune fissare ν o adattare modelli per valori seminteri di ν e scegliere il suo valore tramite selezione del modello [18].

Il parametro κ è un parametro di scala, collegato al range ρ dalla relazione empiricamente derivata $\rho = \sqrt{8/\kappa}$. Per un GRF con covarianza di Matérn con parametri κ e $\nu = 1$, la correlazione spaziale è 0.1 a distanza ρ . Pertanto, possiamo pensare a κ come un parametro di range che governa la struttura di dipendenza spaziale del GRF. Aggiungiamo più flessibilità alla struttura di dipendenza riscalando il campo $x(u)$ con un parametro di varianza τ, che fornisce la varianza marginale $\sigma^2 = \frac{1}{4\pi\kappa^2\tau^2}$ [19].

In RINLA e *inlabru*, sia possibile definire il processo di Matèrn non utilizzando i parametri κ e ν utilizzati in (1.12), ma i parametri ρ e σ, legati ai precedenti tramite le seguenti relazioni [19]:

$$\begin{array}{rcl} \log(\tau) & = & \frac{1}{2}\log\left(\frac{\Gamma(\nu)}{\Gamma(\alpha)(4\pi)^{d/2}}\right) - \log(\sigma) - \nu\log(\kappa) \\ \log(\kappa) & = & \frac{\log(8\nu)}{2} - \log(\rho). \end{array} \tag{3.2}$$

In particolare, nella creazione del processo di Matèrn all'interno del codice, verranno indicate le informazioni relative alle distribuzione dei parametri ρ e σ, indicati come *range* e *sigma*.

Per ρ andremo a definire i valori ρ_p e p_ρ tale che:

$$\mathbb{P}(\rho < \rho_p) = p_\rho \tag{3.3}$$

mentre per σ andremo a definire i parametri σ_p e p_σ tale che:

$$\mathbb{P}(\sigma > \sigma_p) = p_\sigma \tag{3.4}$$

3.2 Raccolta, Analisi e Rielaborazione dei dati

I dati si distinguono per le tecniche di campionamento utilizzate:

1. *Fotogrammetria SfM* per le prime 5 campagne di campionamento, svolte tra il 20 febbrario 2022 e il 20 ottobre 2022, per un totale di 4891 osservazioni;

2. *Ispezione manuale subacquea* per ultime due campagne di campionamento, svolte tra agosto e dicembre 2023, per un totale di 201 osservazioni.

Durante le prime 5 campagne di campionamento è stata utilizzata la tecnica della fotogrammetria *Structure from Motion*. Questa tecnica permette di estrarre una vasta quantità di informazioni dagli ambienti marini con la generazione di modelli 3D dalla sovrapposizione di immagini 2D scattate da prospettive differenti. Le oloturie di mare sono facilmente riconoscibili sul fondale, grazie al forte contrasto con il chiaro fondale sabbioso.
Le immagini sono state acquisite tramite un veicolo subacqueo dotato di una a GoPro Hero10 Black action camera, con la quale sono state raccolta all'incirca 6000 immagini. L'elaborazione delle immagini è stata effettuata utilizzando il software di ricostruzione 3D Agisoft Metashape v 1.6.2 (Agisoft LLC, Russia), che ha generato modelli di superficie digitale (DSM) e mosaici ortofotografici. Ogni oloturia visibile nell'ortomosaico è stata digitalizzata manualmente utilizzando ArcGIS 10.6 in ciascun evento di campionamento (stagione), risultando nella rilevazione di 984, 1478, 938, 769 e 668 oloturie per ciascuna delle cinque stagioni, rispettivamente. I DSM sono stati utilizzati per stimare altre caratteristiche del fondale marino come pendenza, esposizione, rugosità vettoriale (VRM) e rugosità.
Ad agosto e dicembre 2023 è stato condotto un nuovo campionamento su piccola scala nelle zone dell'area di ricerca coperte di Posidonia oceanica, per raccogliere informazioni sulla presenza di animali sotto le piante. Due subacquei con autorespiratore hanno registrato l'abbondanza di oloturie (Holothuria spp.) all'interno della Posidonia oceanica ispezionando manualmente le chiazze. Le foglie sono state spostate con cura per identificare gli esemplari. Per garantire una copertura completa dell'area di ricerca, è stato adottato un modello di ricerca circolare controllato da corde con una larghezza di spazzata di 1 metro. Nelle campagne del 2023 sono state trovate rispettivamente 110 e 91 oloturie [15].
Nella descrizione dei modelli: $c = 0{,}1$ indica la tecnica di campionamento, dove $c = 0$ corrisponde all'utilizzo dell'approccio SfM e $c = 1$ all'ispezione manuale da parte di subacquei, mentre $t = 1, \ldots, 7$ indica la campagna di raccolta dati. Da notare che nel caso di studio $t = 1, \ldots, 5$ se e solo se $c = 0$ e $t = 6{,}7$ se e solo se $c = 1$.
Un'altra osservazione importante è che le due tecniche di campionamento, vengono applicate in regione spaziali tra loro non sovrapponibili, perchè l'ispezione manuale è avvenuta nei punti dove non era possibile utilizzare la fotogrammetria a causa della presenza di Posidonia. Un'approssimazione della distribuzione spaziale della due aree di campionamento può essere osservata nella Figura 3.4.

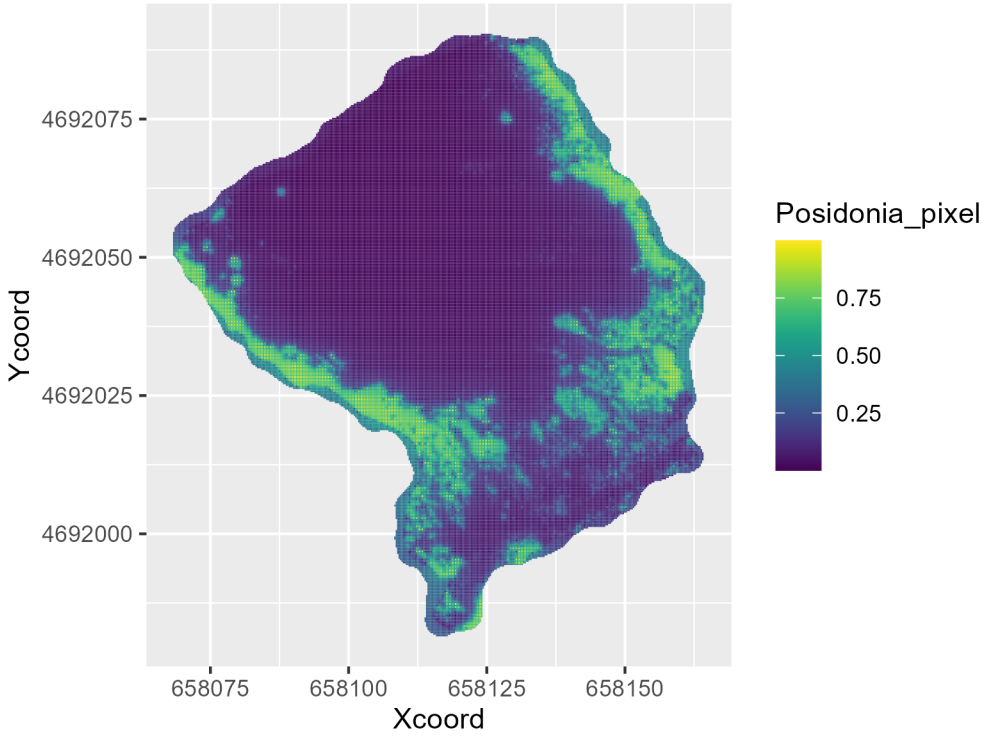

Figura 3.1: Rappresentazione grafica della presenza di Posidonia nell'area di ricerca.

Oltre alle informazioni legate alla posizione degli individui, sono stati raccolti dati riguardo le covariate spaziali dell'area di ricerca eseguendo campionamenti equispaziati a griglia, all'incirca ogni metro. Questi dati sono stati utilizzati per generare una mappa continua delle covariate nell'area di ricerca tramite l'*Inverse Distance Weighting* (IDW), introdotta da Shepard [20].
Dato un insieme di punti campione $\{\mathbf{s}_i, u_i \mid \text{per } \mathbf{s}_i \in \mathbb{R}^n, u_i \in \mathbb{R}\}_{i=1}^N$, la funzione di interpolazione IDW $u(\mathbf{s}) : \mathbb{R}^n \to \mathbb{R}$ è definita come:

$$u(\mathbf{x}) = \begin{cases} \dfrac{\sum_{i=1}^N w_i(\mathbf{s}) u_i}{\sum_{i=1}^N w_i(\mathbf{s})}, & \text{se } d(\mathbf{s}, \mathbf{s}_i) \neq 0 \text{ per tutti } i, \\ u_i, & \text{se } d(\mathbf{s}, \mathbf{s}_i) = 0 \text{ per qualche } i, \end{cases}$$

dove

$$w_i(\mathbf{s}) = \frac{1}{d(\mathbf{s}, \mathbf{s}_i)^p}$$

è un peso inversamente proporzionale alla distanza tra i due punti $\mathbf{s}$ e $\mathbf{s_i}$, dove $\mathbf{s}$ denota un punto interpolato (arbitrario), $\mathbf{s}_i$ è un punto interpolante (noto), d è un

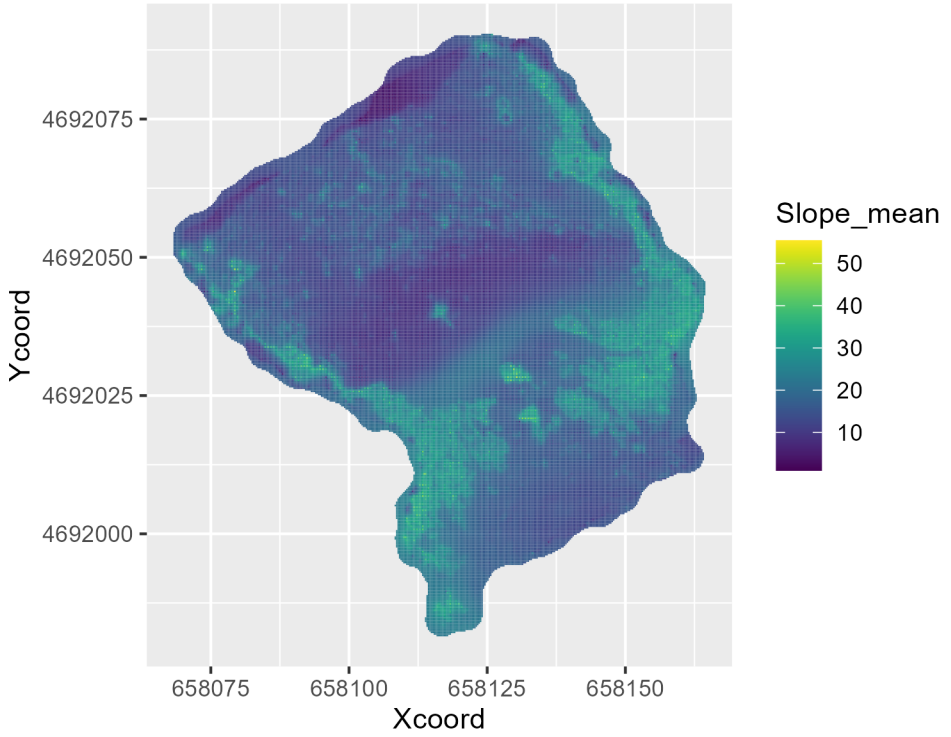

Figura 3.2: Rappresentazione grafica della pendenza del fondale nell'area di ricerca.

operatore metrico dato dalla distanza dal punto noto $\mathbf{s}_i$ al punto sconosciuto $\mathbf{s}$, N è il numero totale di punti noti usati nell'interpolazione.
Sono state ricavate delle mappe continue di alcune covariate spaziali, ritenute utili per lo sviluppo del modello:

1. *Posidonia*, valore tra 0 e 1 che indica il numero di pixel all'interno dell'immagine occupati dalla posidonia Fig. 3.1;

2. *Slope*, indicante la pendenza del fondale, Fig 3.2;

3. *Depth*, indicante la profondità del fondale, Fig 3.3;

Oltre a questi campi spaziali, sono state create altre mappe scalari descriventi: le coordinate locali e un campo per valutare se il valore di Posidonia pixel fosse superiore ad un valore soglia. Quest'ultima mappa è stata utilizzata, insieme alle informazioni raccolte dalla campagna di campionamento, per stimare le aree in cui sono state utilizzate le due tecniche di campionamento.

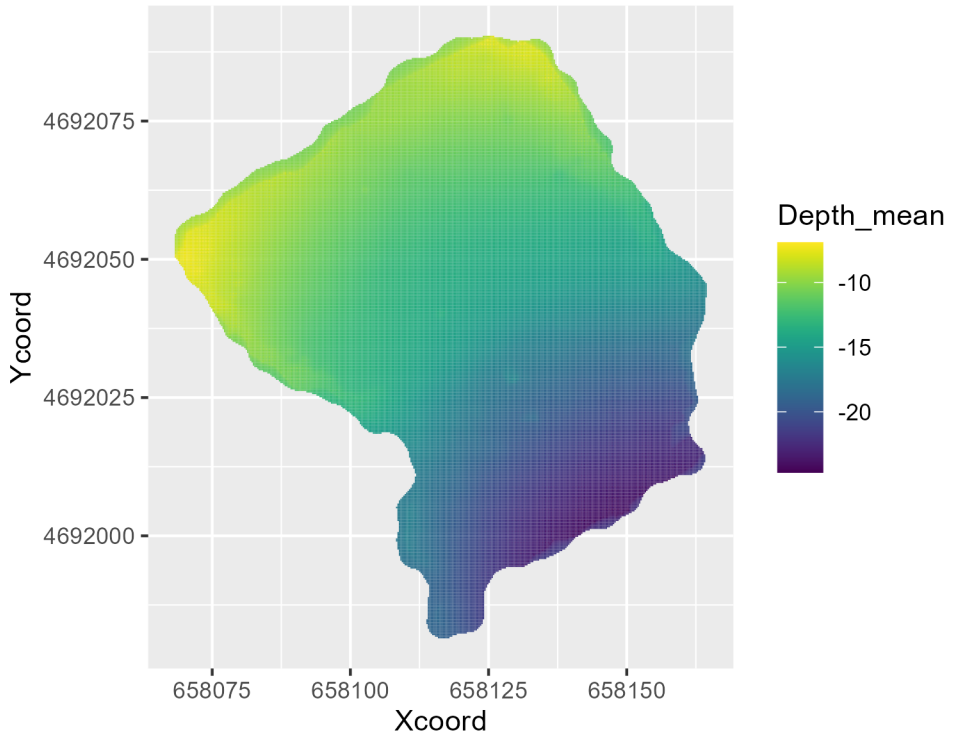

Figura 3.3: Rappresentazione grafica della profondità del fondale nell'area di ricerca.

Tra le varie covariate spaziali registrate durante le prime 5 campagne, è stata registrata anche la tempearatura media del fondale. Nella figura 3.5 è possibile osservare il valore della temperatura in data 20.02.2022 nell'area di ricerca, ottenuto interpolando i dati raccolti con *IDW*. In particolare, nel fase di campionamento sono stati registrati solamente tre diversi valori di temperatura: 15.10, 15.18 e 15.20.
Dalla figura 3.5 ed eseguendo un modello lineare su R, si osserva che sembrerebbe esserci una forte correlazione positiva tra la temperatura e la coordinate y. Per semplicità dunque, nei modelli successivi useremo come covariata spaziale la coordinata Y, interpretandola come un'indicazione delle temperatura.

Per approssimare il campo di Matèrn è stata utilizzata una Mesh di Delaunay, visibile in Figura 3.6. La regione interna linea chiusa blu è area di ricerca dove è stato effettuato il conteggio degli individui di oloturie. Nella regione esterna non sono state effettuate conteggi e nemmeno misure delle covariate spaziali. La necessità di estendere la mesh oltre alla regione in cui avviene il processo è necessario per non avere effetti distorcenti dal bordo. Nell'area esterna alla linea blu, le celle della mesh sono state generate settando come massimo lato 20 m, l'angolo

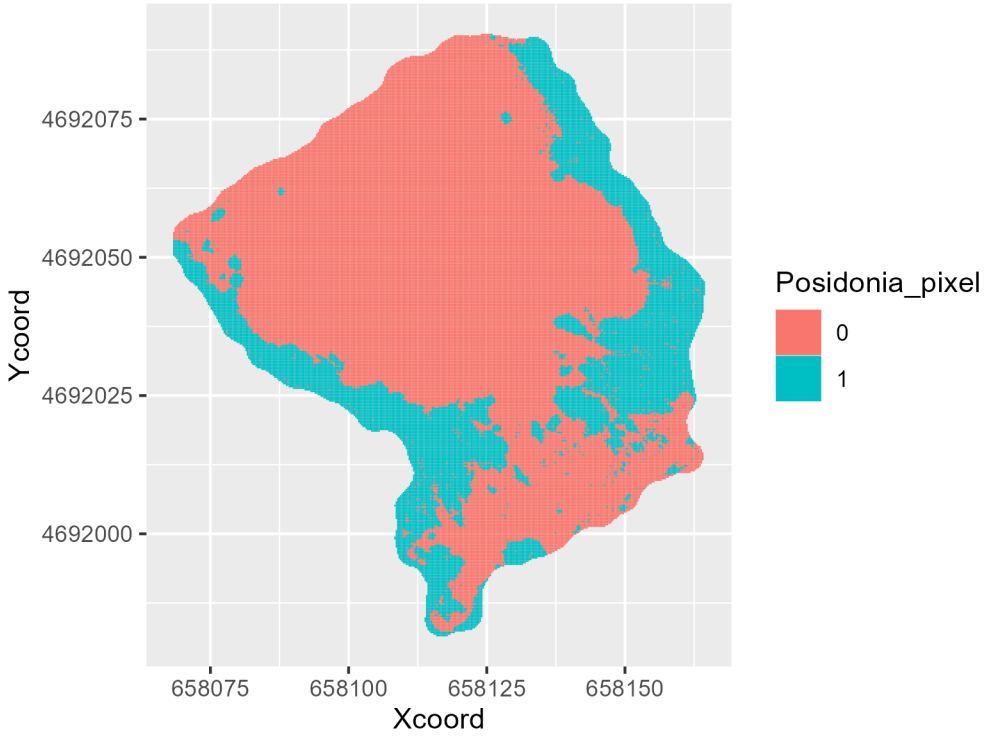

Figura 3.4: Divisione dell'area di ricerca in base alla tecnica di campionamento utilizzata.

minimo per ciascuna cella poteva essere di 30 gradi. Nella regione interna, dove vogliamo una maggiore risoluzione, la lunghezza massimo del lato è di 3 metri, mentre l'angolo minimo è di 20 gradi. La mesh ottenuta è composta da 2077 vertici e verrà utilizzata per l'inferenza di tutti e due i modelli proposti.

3.3 Simulazione di Log-Gaussian Cox Process

Prima di applicare i modelli proposti ai dati reali, è interessante testarli su dati simulati. Sia l'area di ricerca $D \subset \mathbb{R}^2$ sia $R \subset \mathbb{R}^2$ regione rettangolare tale che $D \subset R$ e definiamo gli iperparametri ρ e σ^2 del campo gaussiano w, la struttura della funzione di intensità $\lambda_t(\cdot)$ in base al modello scelto.
Per semplicità e coerenza con il caso reale, simuliamo solamente 3 istanti di tempo e supponiamo che i primi due siano stati effettuati con la prima tecnica di campionamento ($c = 0$) e l'ultimo con la seconda tecnica di campionamento ($c = 1$). Per brevità, il pedice del tipo di campionamento non sarà indicato, siccome

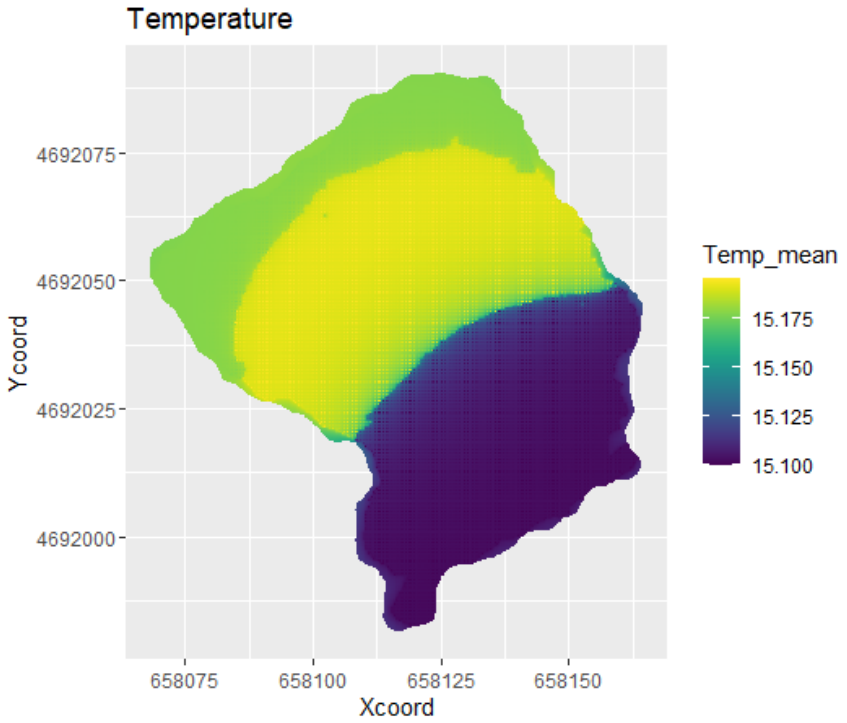

Figura 3.5: Valori di temperatura media registrati in data 20.02.2022.

è implicito dall'istante di tempo osservato.
A questo punto, generiamo una realizzazione di un LGCP, sul quale proviamo a fare inferenza per ricavare i parametri di partenza. Per generare una realizzazione del processo di punto, è necessario prima generare una simulazione del campo gaussiano di Matèrn w e poi generare una realizzazione dal processo di punto di punto di Poisson non omogeneoo con funzione di intensità $\lambda_t(\cdot)|w$ Il nostro focus di interesse è che i parametri e gli iperparametri ricadono nell'intervallo di confidenza del 95% per le loro rispettive distribuzioni posteriori marginali ricavate con INLA. Nelle seguenti sottosezioni, viene illustrato brevemente come simulare un campo gaussiano con covarianza di matern 3.3 e come simulare un processo di punto di Poisson non Omogeneo 3.3.

Simulazione di Campo Gaussiano con Covarianza di Matèrn

Sia Z campo gaussiano in $\mathbb{R}^2$ con media nulla e funzione di covarianza di Matèrn con parametri σ^2 e ρ. Siano $\{s_i\}_{i=1}^n$ una collezione di punti in $\mathbb{R}^2$ per i quali ottenere una realizzazione del campo Z. Definiamo la matrice $\Sigma_{\sigma^2,\rho} = (\Sigma_{\sigma^2,\rho})_{i,j} = Cov(Z(s_i), Z(s_j))$. Ma allora $\mathbf{z} = (Z(s_i)_{i=1}^n \sim \mathcal{N}_n(\mathbf{0_n}, \Sigma_{\rho,\sigma^2})$. Quindi per simulare

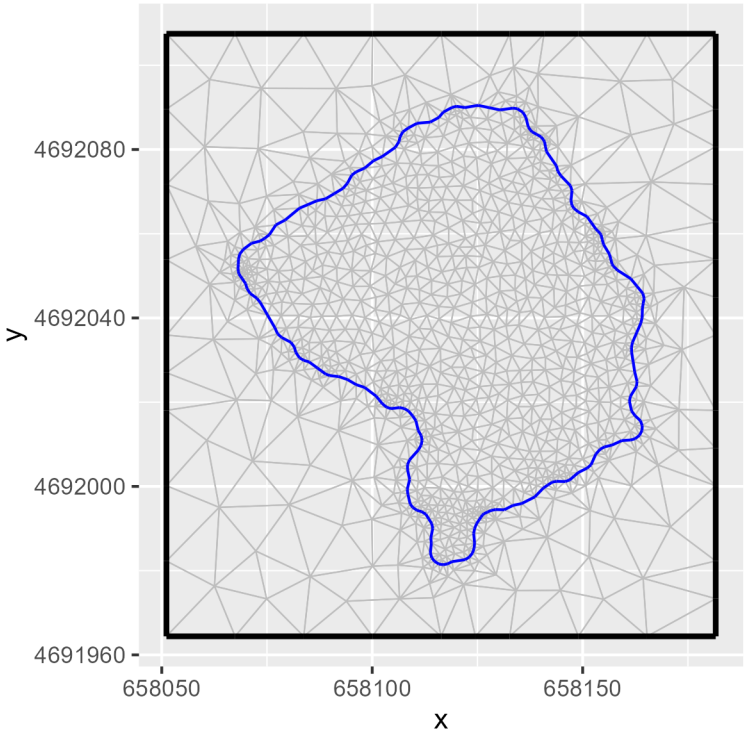

Figura 3.6: Mesh di Delaunay utillizata per l'approssimazione agli elementi finiti del campo di Matèrn.

un campo gaussiano con media nulla e funzione di Covarianza di Matèrn Z, possiamo applicare la seguente procedura:

1. Definiamo i parametri ρ e σ^2;
2. Definiamo una collezione di punti $\{s_i\}_{i=1}^n$ per i quali vogliamo conoscere il valore del campo Z;
3. Utilizziamo l'equazione (1.11) per calcolare Σ_{ρ,σ^2};
4. Simuliamo un vettore $\mathbf{z} \sim \mathcal{N}_n(\mathbf{0_n}, \Sigma_{\rho,\sigma^2})$;
5. Definiamo $Z(s_i) = z_i$.

Nel nostro caso, la collezione di punti è generata partendo da una griglia regolare di punti sulla regione rettangolare R contenente D. I valori del campo negli altri punti possono essere approssimati utilizzando procedure di interpolazione spaziale, per esempio IDW.

Simulazione di Processi di Punto di Poisson Non Omogenei

Supponiamo di voler simulare la realizzazione di un processo di punto di Poisson non omogeneo con funzione di intensità $\lambda(\cdot)$ nella regione $R \subset \mathbb{R}^2$. Per farlo possiamo utilizzare la proprietà di thinning dei processi di Poisson, per poter ricavare una processo di Poisson non omogeneo da un processo di Poisson omogeneo iniziale:

1. Ricaviamo il valore massimo della funzione di intensità di base $\lambda_{max} = max_{s \in R}\lambda(s)$;
2. Otteniamo il numero massimo di punti N_{max} come una variabile di Poisson con parametro $\lambda_{max}R$;
3. Detti x_{min}, x_{max}, y_{min}, y_{max} le coordinate minime e massime di R, simuliamo $x_i \sim Uniform(x_{min}, x_{max})$ e $y_i \sim Uniform(y_{min}, y_{max})$, per $i = 1, \ldots, N_{max}$;
4. Definiamo la funzione di probabilità di accettazione come $p(s) = \frac{\lambda_t(s)}{\lambda_{max}}$;
5. Conserviamo il punto $s^{(i)} = (x_i, y_i)$ con probabilità $p(s^{(i)})$.

Per la proprietà di thinning, il processo di punto ottenuto come insieme dei punti accettati, sarà un processo di punto di Poisson non omogeneo.

3.4 Modello Intercetta

Introduciamo il primo modello proposto in questo lavoro. Il seguente modello viene denominato *Intercetta* perchè è stato sviluppato partendo dall'idea iniziale di definire un'intercetta che distinguesse le due diffenti tecniche di campionamento. Vista l'ipotesi di aree di campionamento distinte, è utile definire una nuova covariata spaziale con valori discreti $\{0, 1\}$.
Definiamo un campo $P(s, c) : D \times \{0, 1\} \to \{0, 1\}$ tale che:

$$P(s, p) = \begin{cases} p, & \text{se } x \in D_p \\ 1 - p, & \text{se } x \notin D_p \end{cases} \tag{3.5}$$

dove D_c, con $c = 0{,}1$ indicano le rispettive aree in cui è stato effettuata una tecnica di campionamento. L'idea è che a ciascun punto $s \in D$ sia associato valore 1 quando è nella regione esplorata dalla tecnica di campionamento c che si sta analizzando, mentre sia 0 altrimenti.
Il modello *Intercetta* sarà caratterizzato dal logaritmo della funzione di intensità $\lambda_{t,c}(s)$ con la seguente forma:

$$log(\lambda_{t,c}(s)) = \beta_{0,t} + \sum_{i=1}^{m} \beta_i F(s) + \beta_c P(s, c) + w(s) \tag{3.6}$$

Parameter	Value
Posidonia	0.1
Slope	−0.2
Depth	0.3
SamplingSfM	1.0
SamplingSub	−1.0
Intercept1	−2.0
Intercept2	−1.5
Intercept3	−1.0

Tabella 3.1: Parametri utilizzati per la generazione dei processi di punto nel modello Intercetta durante i test eseguiti.

Istante di tempo	Numero di punti
$t = 1$	15
$t = 2$	22
$t = 3$	1

Tabella 3.2: Numero di punti generati per ciascun istante di tempo nel test del modello Intercetta.

dove $F_i(s)$ per $i = 1, \ldots, m$ sono le covariate spaziali, $\beta_{0,t}$ è un termine costante dipendente dal tempo e w_t è un GF.
Possiamo riscrivere la funzione di intensità come:

$$\lambda_{t,c}(s) = e^{\beta_c P(s,c)} Z_t(s)$$

con $Z_t(s) = e^{\beta_{0,t} + \sum_{i=1}^{m} \beta_i F(s)}$ come *funzione di intensità base*, ossia indipendente dal tipo di campionamento selezionato. Questa forma mette in evidenza la caratteristica di questo modello di separare la componente dovuta dalla tecnica di campionamento dalla componente legata al processo naturale sottostante su quale si sta facendo inferenza.
E' possibile interpretare il termine e^{β_c} come la distorsione portata dalla tecnica di campionamento c sul processo di base descritto dalla funzione di intensità $Z_t(s)$.

3.4.1 Risultati Test Modello Intercetta

Utilizziamo gli algoritmi introdotti nella Sezione 3.3 per generare una realizzazione di un LGCP e valutare la qualità delle predizione del modello.
Nella Tabella 3.1 sono definiti i valori puntuali associati ai coefficienti di ciascuna covariata spaziale e delle intercette temporali. Nella Tabella 3.2 sono indicati i

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
Posidonia	1.556	1.023	-0.449	1.556	3.561	1.556	0
Slope	-0.054	0.041	-0.135	-0.054	0.028	-0.054	0
Depth	0.338	0.075	0.191	0.338	0.484	0.338	0
SamplingSfM	-7.043	11.160	-28.916	-7.043	14.831	-7.043	0
SamplingSub	-5.921	12.057	-29.552	-5.921	17.710	-5.921	0
Intercept1	-1.256	0.953	-3.123	-1.256	0.612	-1.256	0
Intercept2	-0.561	0.922	-2.369	-0.561	1.246	-0.561	0
Intercept3	-2.799	1.432	-5.606	-2.799	0.007	-2.799	0

Tabella 3.3: Tabella dei risultati ottenuti per il test del modello intercept.

	Media	Precisione
Intercetta	0	0.01
Effetto Fisso	0	0.01

Tabella 3.4: Tabella informazioni a priori per componenti modello Intercetta

numeri di punti ottenuti per ciascuna delle 3 realizzazione temporali del processo. In particolare osserviamo che il numero di realizzazione ottenute è particolarmente basso e questo potrebbe aver influenzato i risultati ottenuti nella procedura di inferenza.
Nella Tabella 3.3 osserviamo come il valore reale della coefficiente lineare associato alla covariata spaziale *Slope* non ricade nell'intervallo del 95% di confidenza, mentre invece le posteriori relative ai termini di campionamento presentano intervalli di confidenza molto ampli, non permettendo di ricavare particolari informazioni sui termini iniziali.
Probabilmente cio è dovuta allo scarso numero di punto, in particolare per il terzo istante di tempo, che è l'unico ad essere associato alla tecnica di campionamento qua chiamata *Sub*.

3.4.2 Analisi Risultati su Dati Reali

RINLA permette di definire le distribuzione a priori sulle variabili latenti gaussiane del modello, definendo la media e la precisione per ciascuna componente degli effetti fissi e delle intercette, disponibili nella Tabella 3.4.
Le informazioni per descrivere il processo di Matèrn invece sono disponibili nella Tabella 3.5.
Questi valori sono stati ottenuti eseguendo più simulazioni e conservando i valori ottenuti nella simulazione con i risultati migliori.
Impostato il modello, possiamo utilizzare *inlabru* per valutare i risultati ottenuti, nella Tabelle 3.6 e 3.8.
Nella tabella 3.6 possiamo osservare le informazioni sulle distrubuzioni a posteriori dei regressori lineari per ciascuna componente, nonchè le distribuzione a posteriori

	parametro	probabilità
ρ	150	0.95
σ	10	0.05

Tabella 3.5: Iperparamentri per il processo di Matèrn nel modello Intercetta.

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
Posidonia	-0.820	0.310	-1.429	-0.820	-0.212	-0.820	0
Slope	-0.069	0.007	-0.083	-0.069	-0.054	-0.069	0
Depth	-0.014	0.124	-0.257	-0.014	0.229	-0.014	0
Ycoord	0.036	0.063	-0.087	0.036	0.159	0.036	0
InterceptSamplingPos	-0.910	0.118	-1.142	-0.910	-0.678	-0.910	0
InterceptSamplingNoPos	-6.542	1.004	-8.510	-6.542	-4.575	-6.542	0
Intercept1	-3.535	5.107	-13.544	-3.535	6.474	-3.535	0
Intercept2	-3.112	5.107	-13.121	-3.112	6.897	-3.112	0
Intercept3	-3.571	5.107	-13.580	-3.571	6.438	-3.571	0
Intercept4	-3.761	5.107	-13.770	-3.761	6.248	-3.761	0
Intercept5	-3.918	5.107	-13.927	-3.918	6.091	-3.918	0
Intercept6	-3.906	5.109	-13.920	-3.906	6.108	-3.906	0
Intercept7	-4.096	5.109	-14.110	-4.096	5.918	-4.096	0

Tabella 3.6: Statistiche descrittive effetti fissi modello Intercetta

delle intercette temporali.
Per interpretare in maniera più veloce ed intuitiva l'effetto delle varie covariate, conoscendo la distribuzione a posteriori di ciascun coefficiente di regressione, possiamo valutare la probabilità che esso sia maggiore di zero, permettendo di valutare l'influenza della rispettiva componente sul processo di punto, come mostrato nella tabella 3.7.
Osserviamo in particolare come i valori relativi alle tecniche di campionamento siano quasi certamente minore di 0, come conseguenza quindi avremo $e^{\beta_c}, \quad c = 0{,}1$ saranno compresi tra 0 e 1. Un'interpretazione di questo risultato potrebbe essere che entrambe le tecniche di campionamento non riescono a catturare completamente la distribuzione della popolazione. Dunque è come se i processi catturati dalla tecniche di campionamento sia dei processi *thinning* del processo di base naturale, dunque possiamo interpretare e^{β_c} come la probabilità di osservazione della tecnica c nell'area D_c.
Analogamente, anche il coefficiente lineare relativo alla pendenza media del fondale (*Slope* nella tabella) è quasi certamente negativo, dunque la popolazione viene favorita da un fondale con pendenze più dolci.
Per il coefficiente relativo alla profondità otteniamo invece come le probabilità che siano negativo o positivo siano abbastanza equilibrate, indicando che probabilmente la profondità media del fondale non è un fattore che influenza la dinamica della popolazione di oloturie. Un'interpretazione più corretta potrebbe essere che il range di profondità dei campioni non sia sufficientemente ampio per poter osservare

Parameter	**Mean**	**SD**	**Prob > 0**	**Prob < 0**
Posidonia	-0.820	0.310	4.083e-03	0.9959
Slope	-0.069	0.007	0.0000	1.0000
Depth	-0.014	0.124	0.4551	0.5449
Ycoord	0.036	0.063	0.7161	0.2839
SamplingSfM	-0.910	0.118	6.217e-15	1.0000
SamplingSub	-6.542	1.004	3.612e-11	1.0000

Tabella 3.7: Probabilità che ciascuna variabile aleatoria sia maggiore o minore di zero modello intercetta.

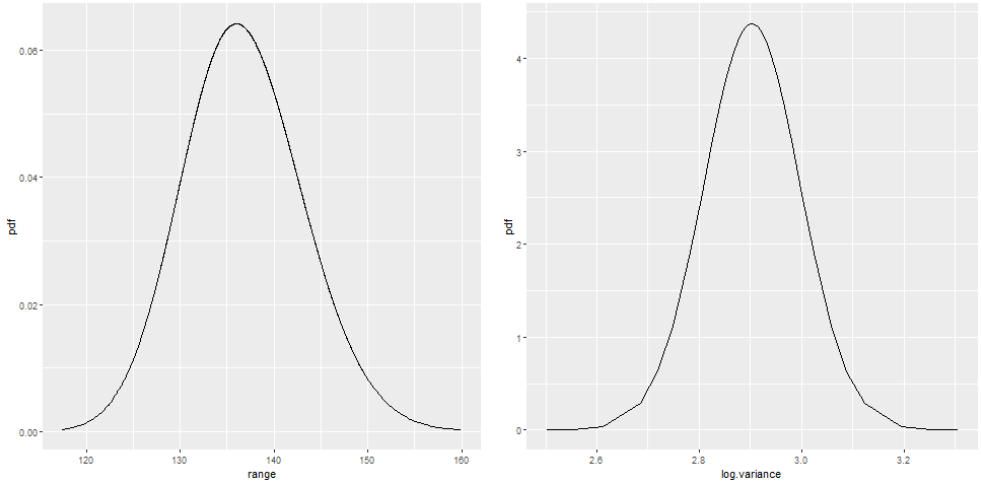

Figura 3.7: Statistiche descrittive delle distribuzioni per i parametri del campo gaussiano $w_t(x)$: range e il logaritmo della varianza nel modello intercetta.

l'effetto di questa covariata spaziale (l'intervallo della profondità è compreso tra -4 m ai -25 m). Il modello dunque riesce a spiegare a i risultati raccolti senza l'informazione della profondità media.
Per la presenza di Posidonia e la coordinata y (che ricordiamo essere legata alla temperatura media registrata nelle acque sul fondale) non siamo in grado di fare considerazioni nette come in precendenza, tuttavia la tabella 3.7 suggerisce che entrambi la presenza di Posidonia possa avere un effetto negativo sulla popolazione di oloturie, mentre invece una temperatura maggiore sembra avere un effetto positivo.
Nella Tabella 3.8 sono presenti le statistiche descrittive del campo gaussiano w_t. In Figura 3.7 possiamo osservare le distribuzioni dei parametri range e logvar della campo gaussiano. Nella figura 3.8 è presente una rappresentazione del valor medio del campo nella regione d'interesse, rispetto ai campioni osservati in ciascuna delle camoagne di campionamento, rappresentate con colore diverso in base alla tecnica

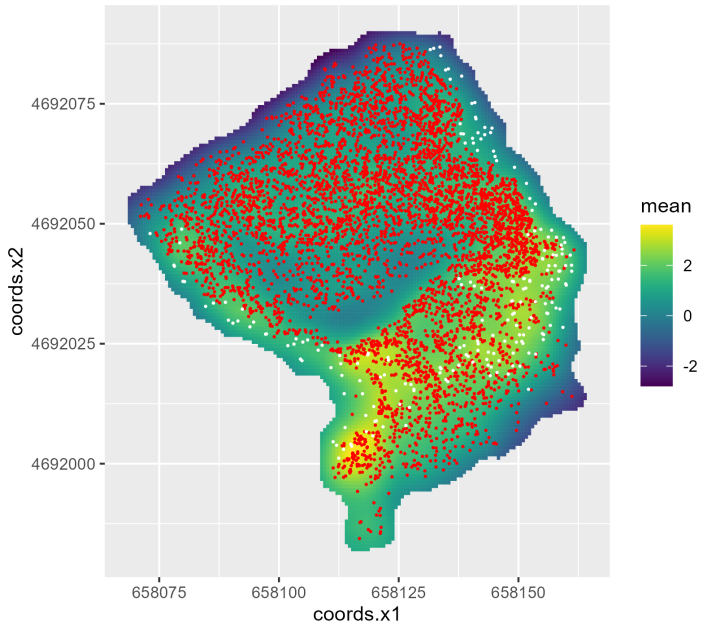

Figura 3.8: Distribuzionale spaziale della media del campo gaussiano w_t per il modello con intercetta. I punti rossi indicano esemplari di oloturie raccolti con la tecnica di campionamento con fotogrammetria, i punti bianchi sono esemplari osservati tramite ispezione manuale.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
Range for Random1	136.77	6.43	124.75	136.55	150.05	135.97
Stdev for Random1	4.28	0.20	3.89	4.27	4.68	4.26

Tabella 3.8: Statistiche descrittive della distrinbuzione a posteriori per Random1 modello Intercetta

di campionamento utilizzata. Infine, in Figura 3.9 viene presentato l'andamento della correlazione tra due punti al variare della distanza, ottenuto tramite l'inferenza dei parametri κ e σ del campo.
Nella Figura 3.10 possiamo confrontare il valore delle intercette e il loro intervallo di confidenza. Da questa figura sembrerebbe non esserci stagionalità nelle variazioni della funzioni di intensità del processo, o per lo meno queste variazioni non riescono ad essere catturate dal modello corrente.

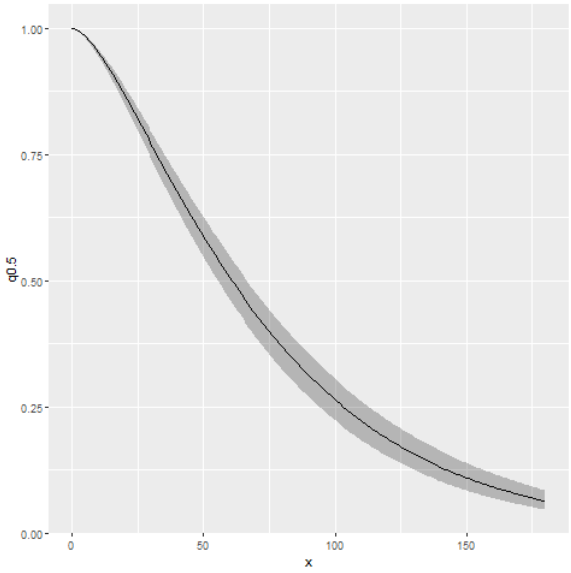

Figura 3.9: Funzione di correlazione tra due punti del campo gaussiano w_t al variare della distanza modello intercetta.

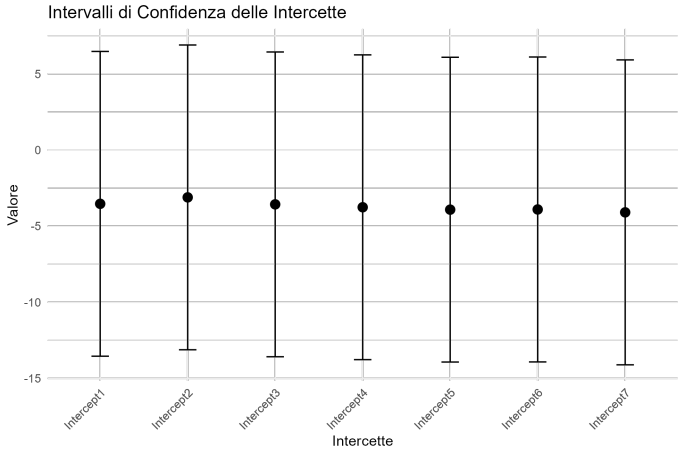

Figura 3.10: Confronto tra i valori delle intercette temporali e i loro intervalli di confidenza nel modello intercetta.

3.5 Modello Masking

Il seguente modello si basa sull'idea di "mascherare" o "silenziare" i valori delle covariate spaziali nell'area dove la tecnica di campionamento non viene utilizzata, lasciando in quei punti solamente l'effetto dell'intercetta temporale e del campo gaussiano.

Parameter	Value
Posidonia	0.1
Slope	−0.2
Depth	0.3
Intercept1	−2.0
Intercept2	−1.5
Intercept3	−1.0

Tabella 3.9: Parametri utilizzati per la generazione dei processi di punto nel modello masking.

Definiamo una funzione h tale che:

$$h(F(s), c) = \begin{cases} F(s), & \text{se } s \in D_c \\ 0, & \text{se } s \notin D_c \end{cases} \tag{3.7}$$

detta *funzione di masking*. Descriviamo quindi il logaritmo della funzione di intensità del processo come:

$$log(\lambda_{t,c}(s)) = \beta_{0,t} + \sum_{i=1}^{m} \beta_i h(F(s), c) + w(s) \tag{3.8}$$

da cui:

$$log(\lambda_{t,c}(s)) = \beta_{0,t} + \sum_{i=1}^{m} \beta_i F(s) + w(s), \text{ se } x \in D_c$$
$$log(\lambda_{t,c}(s)) = \beta_{0,t} + w(s), \text{ se } s \notin D_c$$

A differenza del modello precedente, in questa caso non è possibile separare l'effetto della tecnica di campionamento da quello del processo naturale sottostante. In questo modello vengono perse le informazioni sulla efficacia della tecnica di campionamento utilizzata, riducendo allo stesso tempo la quantità di parametri da dover stimare nella fase di inferenza del modello.

3.5.1 Risultati Test

Nella Tabella 3.9 sono definiti i valori puntuali associati ai coefficienti di ciascuna covariata spaziale e delle intercette temporali. Nella Tabella 3.10 sono indicati i numeri di punti ottenuti per ciascuna delle 3 realizzazione temporali del processo; in questo il numero di realizzazione è decisamente più numeroso rispetto al test precedente.

Istante di tempo	Numero di punti
$t = 1$	34
$t = 2$	68
$t = 3$	98

Tabella 3.10: Numero di punti generati per ciascun istante di tempo nel test del modello masking.

	Mean	**sd**	**0.025quant**	**0.5quant**	**0.975quant**	**Mode**	**kld**
Posidonia	-0.144	0.676	-1.468	-0.144	1.181	-0.144	0
Slope	0.090	0.016	0.059	0.090	0.122	0.090	0
Depth	-0.134	0.018	-0.169	-0.134	-0.098	-0.134	0
Intercept1	-8.503	0.334	-9.158	-8.503	-7.848	-8.503	0
Intercept2	-7.803	0.312	-8.413	-7.803	-7.192	-7.803	0
Intercept3	-7.224	0.342	-7.895	-7.224	-6.554	-7.224	0

Tabella 3.11: Tabella dei risultati ottenuti per il test del modello masking.

Tuttavia, come possiamo notare osservando la Tabella 3.11, i parametri reali non cadono negli intervalli di confidenza al 95% (eccetto che per il termine *Posidonia*). Questa difficoltà potrebbe essere legata alla definizione della masking function utilizzata per modellare l'interazione tra il processo di punto e la tecnica di campionamento.

3.5.2 Analisi risultati

Come per il modello precedente, i risultati dell'analisi sono state raccolte per la descrizione degli effetti fissi nella Tabella 3.12, per la probabilità a posteriori che l'effetto di ciascuna covariata sia positivo o negativo nella Tabella (3.13) e per le distribuzioni a posteriori degli iperparametri del campo gaussiano nella Tabella (3.14).
Come nel modello precedente, osserviamo nella Tabella 3.12 che le intercette temporali sembrano assumere valori molto simili tra di loro, mantenendo un valore della varianza pressochè identico per ciascun istante di tempo (2.821 per i primi cinque istanti temporali e 2.823 per gli ultimi due).
Nella Tabella 3.13 osserviamo come riusciamo a descrivere in maniera certa l'effetto delle pendenza media, della profondità media e della coordinata Y (legata alla temperatura) hanno sulla popolazion; infatti le prime due hanno un effetto quasi certamente negativo (a differenza del modello precedente che non riusciva a valutare l'effetto della profondità). Invece non siamo in grado di descrivere in maniera accurata l'effetto della Posidonia: le probabilità che sia negativo o positivo sono troppo equilibrate, non riuscendo ad avere valenza statistica.

	mean	sd	0.025quant	0.5quant	0.975quant	mode	kld
Posidonia	-0.042	0.370	-0.766	-0.042	0.682	-0.042	0
Slope	-0.067	0.008	-0.082	-0.067	-0.052	-0.067	0
Depth	-0.172	0.010	-0.192	-0.172	-0.152	-0.172	0
Ycoord	0.042	0.004	0.034	0.042	0.049	0.042	0
Intercept1	-5.854	2.821	-11.382	-5.854	-0.326	-5.854	0
Intercept2	-5.431	2.821	-10.959	-5.431	0.098	-5.431	0
Intercept3	-5.890	2.821	-11.418	-5.890	-0.362	-5.890	0
Intercept4	-6.080	2.821	-11.608	-6.080	-0.552	-6.080	0
Intercept5	-6.237	2.821	-11.766	-6.237	-0.709	-6.237	0
Intercept6	-6.743	2.823	-12.275	-6.743	-1.211	-6.743	0
Intercept7	-6.938	2.823	-12.471	-6.938	-1.405	-6.938	0

Tabella 3.12: Statistiche descrittive effetti fissi modello Masking.

	mean	sd	$P > 0$	$P < 0$
Posidonia	-0.042	0.370	0.4548	0.54519
Slope	-0.067	0.008	0.0000	1.0000
Depth	-0.172	0.010	0.0000	1.0000
Ycoord	0.042	0.004	1.0000	0.0000

Tabella 3.13: Probabilità che ciascuna variabile aleatoria sia maggiore o minore di zero modello masking.

Nella Tabella 3.14 possiamo osservare i quantili delle distribuzioni a posteriori dei parametri ρ e σ. Osserviamo in particolare come il range presenti un valore minore rispetto al modello precedente, con una varianza della distribuzione ridotta.
Come nel modello precedente, anche questo modello sembra non riuscire a modellare l'evoluzione temporale della funzione di intensità del processo tramite un'intercetta temporale. Nella Figura 3.14 è possibile osservare infatti che i valori medi della intercette sono simili e che gli intervalli di confidenza del 0.95 sono pressochè sovrapponibili per tutti gli istanti di tempo analizzati.

3.6 Conclusioni

In questo capitolo sono stati proposti, discussi, testati e utilizzati due modelli per la funzione di intensità di un LGCP che permettessero di integrare dati raccolti con tecniche di campionamento diverse. Lo scopo finale era cercare di inferire l'effetto di alcune caratteristiche spaziali dell'ambiente marino sulla popolazione di Oloturie nei fondali nei pressi dell'Isola del Giglio, come mostrato nelle Tabelle 3.7 e 3.13. Di particolare interesse era valutare l'effetto della presenza di Posidonia sulla popolazione, tuttavia solamente per il modello Intercetta (3.6) siamo stati in grado di dare una risposta con certo grado di sicurezza, poichè per il modello (3.8) Masking le probabilità di un effetto positivo o negativo erano simili.

	mean	sd	0.025quant	0.5quant	0.975quant	mode
Range for Random1	102.68	4.931	93.31	102.57	112.72	102.35
Stdev for Random1	3.05	0.147	2.78	3.05	3.35	3.03

Tabella 3.14: Statistiche descrittive dalla distribuzione a posteriori dei parametri per Random1 modello masking.

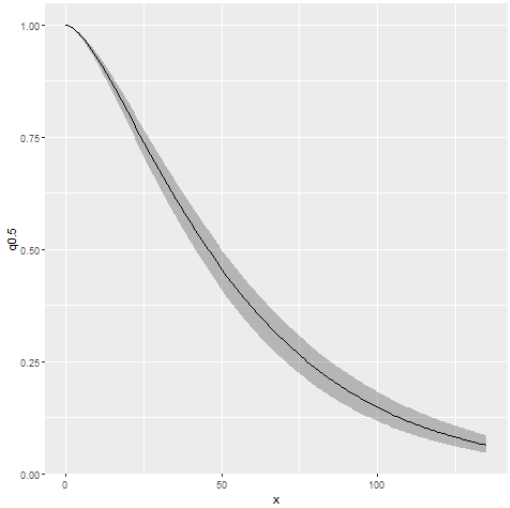

Figura 3.11: Funzione di correlazione tra due punti del campo gaussiano w_t al variare della distanza modello masking.

I modelli proposti hanno presentato dei problemi durante la fase di test, tuttavia i risultati proposti con il primo modello sembrano poter essere considerati affidabili. Sarebbe necessario un confronto con esperti di dominio, che possano analizzare i risultati ottenuti e confrontarle con le conoscenze attualmente disponibili sulla specie marina in questione.
La maggiore efficacia del modello Intercetta (3.6) è probabilmente legata alla separazione dell'effetto della tecnica di campionamento rispetto alla funzione di intensità del processo base: infatti, per ciascun istante di tempo t e ciascuna tecnica di campionamento c, possiamo riscrivere la funzione di intensità del processo osservato come $\lambda_{t,c}(s) = e^{\beta_c P(s,c)} Z_t(s)$, con $Z_t(s)$ funzione di intensità del processo base. Questa struttura sfrutta le proprietà di thinning dei Processi di Punto di Poisso ed è la stessa che viene presentata nelle tecniche di inferenza che fanno usando della *Sampling Effort Function*, presentata in [21], in cui la funzione di intensità del processo di punto osservata è data dal prodotto del processo di punto di base $\lambda(s)$ e una funzione $q(s)$ con immagine tra [0,1] che misura l'efficacia del campionamento nel punto spaziale s.
Futuri studi potrebbero investigare altri modelli per descrivere l'efficacia delle due

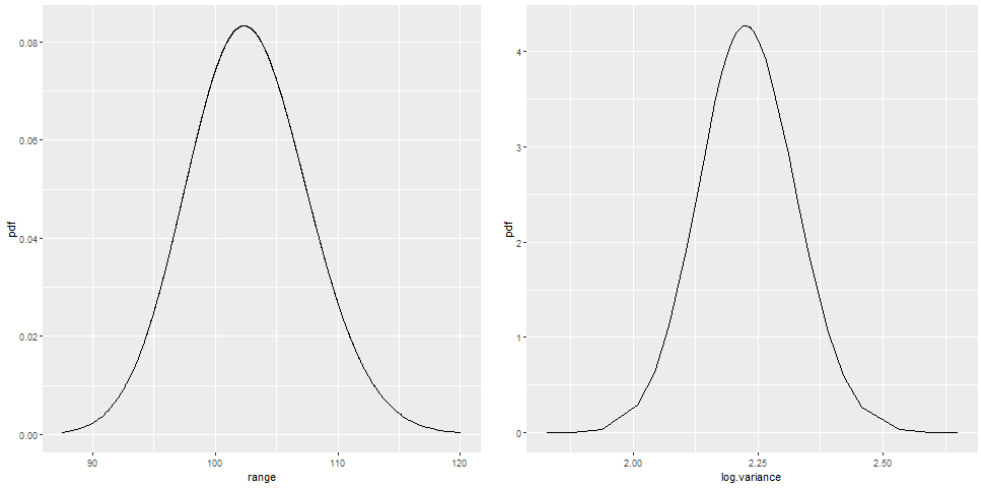

Figura 3.12: Statistiche descrittive delle distribuzioni per i parametri del campo gaussiano $w_t(x)$: range e il logaritmo della varianza nel modello masking.

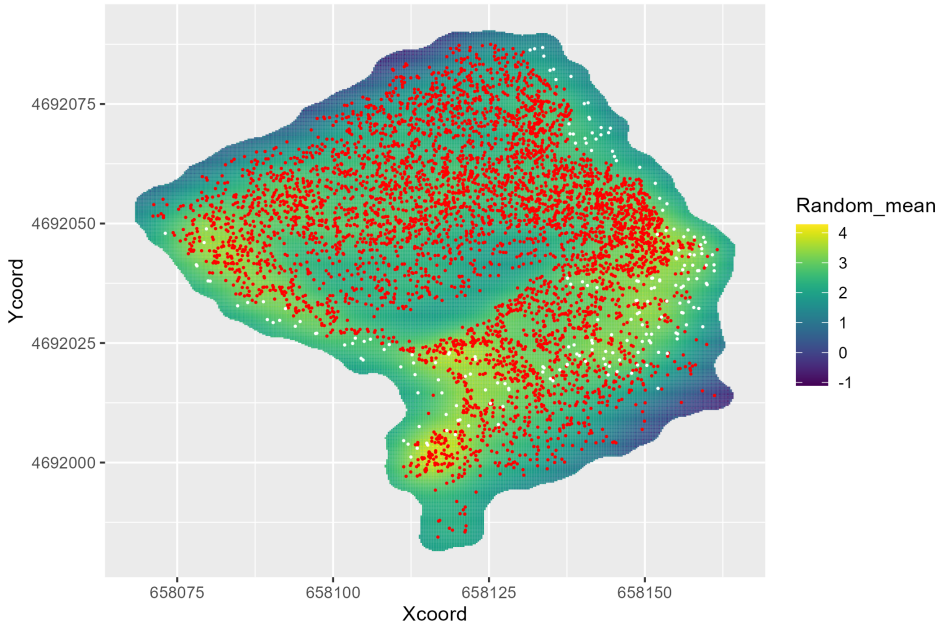

Figura 3.13: Distribuzionale spaziale della media del campo gaussiano w_t per il modello con Masking. I punti rossi indicano esemplari di oloturie raccolti con la tecnica di campionamento con fotogrammetria, i punti bianchi sono esemplari osservati tramite ispezione manuale.

tecniche di campionamento utilizzate: in particolare la tecnica SfM presenta un

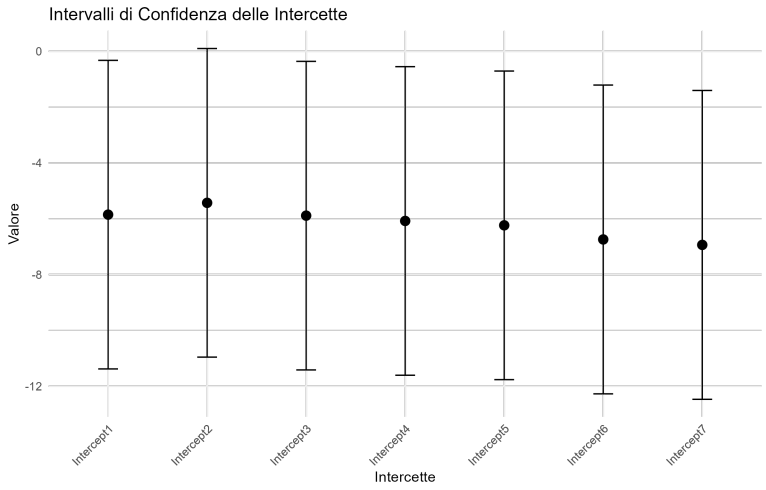

Figura 3.14: Confronto tra i valori delle intercette temporali e i loro intervalli di confidenza nel modello masking.

certo grado di efficacia anche nelle zone del fondale marino ricoperte da Posidonia, richiedendo probabilmente l'utilizzo di una funzione di campionamento che tenga conto della quantità di Posidonia presente; invece la funzione di campionamento per la seconda tecnica, di osservazione diretta da parte di sub, per come è stata impostata può probabilmente essere descritta come una funzione a valori discreti $\{0, 1\}$, con valore 1 nella zona dove è stata effettuata l'osservazione (quindi dov'era presente la posidonia) e 0 altrove.

Bibliografia

[1] J. Moller e R.P. Waagepetersen. *Statistical Inference and Simulation for Spatial Point Processes*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press, 2003. ISBN: 9780203496930. URL: https://books.google.it/books?id=dBNOHvElXZ4C (cit. alle pp. 2, 4, 5, 9, 11).

[2] S. Banerjee, B.P. Carlin e A.E. Gelfand. *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis, 2014. ISBN: 9781439819173. URL: https://books.google.it/books?id=zNLhAwAAQBAJ (cit. alle pp. 2, 4, 5, 9–11).

[3] JK Gardner. «Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian?» In: *Bulletin of the seismological society of America* 64.5 (1974), pp. 1363–1367 (cit. a p. 3).

[4] A.E. Gelfand, P. Diggle, P. Guttorp e M. Fuentes. *Handbook of Spatial Statistics*. ISSN. CRC Press, 2010. ISBN: 9781420072884. URL: https://books.google.it/books?id=Xf4leslPDzsC (cit. alle pp. 3, 4).

[5] Moo K. Chung. «Introduction to Random Fields». In: *arXiv* (2020). eprint: 2007.09660 (math.ST). URL: https://arxiv.org/abs/2007.09660 (cit. a p. 6).

[6] Finn Lindgren, Håvard Rue e Johan Lindström. «An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach». In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4 (2011), pp. 423–498. DOI: https://doi.org/10.1111/j.1467-9868.2011.00777.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2011.00777.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.00777.x (cit. alle pp. 6, 7, 13, 17–19).

[7] D. R. Cox. «Some Statistical Methods Connected with Series of Events». In: *Journal of the Royal Statistical Society: Series B (Methodological)* 17.2 (dic. 2018), pp. 129–157. ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.

1955.tb00188.x. eprint: https://academic.oup.com/jrsssb/article-pdf/17/2/129/49094096/jrsssb_17_2_129.pdf. URL: https://doi.org/10.1111/j.2517-6161.1955.tb00188.x (cit. a p. 8).

[8] Jesper Møller, Anne Randi Syversveen e Rasmus Plenge Waagepetersen. «Log Gaussian Cox Processes». In: *Scandinavian Journal of Statistics* 25.3 (1998), pp. 451–482. ISSN: 03036898, 14679469. URL: http://www.jstor.org/stable/4616515 (visitato il 05/05/2024) (cit. a p. 8).

[9] Håvard Rue, Sara Martino e Nicolas Chopin. «Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations». In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71.2 (2009), pp. 319–392. DOI: https://doi.org/10.1111/j.1467-9868.2008.00700.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2008.00700.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2008.00700.x (cit. alle pp. 12, 13, 15).

[10] Sara Martino e Andrea Riebler. *Integrated Nested Laplace Approximations (INLA)*. 2019. arXiv: 1907.01248 [stat.CO] (cit. alle pp. 12–15).

[11] Luke Tierney e Joseph B. Kadane. «Accurate Approximations for Posterior Moments and Marginal Densities». In: *Journal of the American Statistical Association* 81.393 (1986), pp. 82–86. DOI: 10.1080/01621459.1986.10478240. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1986.10478240. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1986.10478240 (cit. alle pp. 14, 15).

[12] Janine B. Illian, Sigrunn H. Sørbye e Håvard Rue. «A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA)». In: *The Annals of Applied Statistics* 6.4 (2012), pp. 1499–1530. DOI: 10.1214/11-AOAS530. URL: https://doi.org/10.1214/11-AOAS530 (cit. a p. 16).

[13] D. Simpson, J. B. Illian, F. Lindgren, S. H. Sørbye e H. Rue. «Going off grid: computationally efficient inference for log-Gaussian Cox processes». In: *Biometrika* 103.1 (feb. 2016), pp. 49–70. ISSN: 0006-3444. DOI: 10.1093/biomet/asv064. eprint: https://academic.oup.com/biomet/article-pdf/103/1/49/7084356/asv064.pdf. URL: https://doi.org/10.1093/biomet/asv064 (cit. alle pp. 17, 18).

[14] Jane E. Williamson, Stephanie Duce, Karen E. Joyce e Vincent Raoult. «Putting sea cucumbers on the map: projected holothurian bioturbation rates on a coral reef scale». English. In: *Coral Reefs* 40.2 (apr. 2021), pp. 559–569. ISSN: 0722-4028. DOI: 10.1007/s00338-021-02057-2 (cit. a p. 20).

[15] Gianluca Mastrantonio, Daniele Ventura, Edoardo Casoli, Arnold Rakaj, Giovanna Jona Lasinio, Daniele Poggio, Cecilia Vitiello e Crescenza Calculli. «Species distribution models with masking: the case of holothurians in a Posidonia rich area». In: *SIS24* (2024) (cit. alle pp. 20, 24).

[16] Fabian E. Bachl, Finn Lindgren, David L. Borchers e Janine B. Illian. «inlabru: an R package for Bayesian spatial modelling from ecological survey data». In: *Methods in Ecology and Evolution* 10.6 (2019), pp. 760–766. DOI: https://doi.org/10.1111/2041-210X.13168. eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13168. URL: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13168 (cit. alle pp. 21, 22).

[17] Finn Lindgren, Fabian Bachl, Janine Illian, Man Ho Suen, Håvard Rue e Andrew E. Seaton. *inlabru: software for fitting latent Gaussian models with non-linear predictors*. 2024. arXiv: 2407.00791 [stat.ME]. URL: https://arxiv.org/abs/2407.00791 (cit. a p. 22).

[18] P. Diggle e P.J. Ribeiro. *Model-based Geostatistics*. Springer Series in Statistics. Springer New York, 2007. ISBN: 9780387485362. URL: https://books.google.it/books?id=qCqOm39OuFUC (cit. a p. 23).

[19] Rikke Ingebrigtsen, Finn Lindgren e Ingelin Steinsland. «Spatial models with explanatory variables in the dependence structure». In: *Spatial Statistics* 8 (2014). Spatial Statistics Miami, pp. 20–38. ISSN: 2211-6753. DOI: https://doi.org/10.1016/j.spasta.2013.06.002. URL: https://www.sciencedirect.com/science/article/pii/S2211675313000377 (cit. a p. 23).

[20] Donald Shepard. «A two-dimensional interpolation function for irregularly-spaced data». In: *Proceedings of the 1968 23rd ACM National Conference*. ACM '68. New York, NY, USA: Association for Computing Machinery, 1968, pp. 517–524. ISBN: 9781450374866. DOI: 10.1145/800186.810616. URL: https://doi.org/10.1145/800186.810616 (cit. a p. 25).

[21] Avishek Chakraborty, Alan E. Gelfand, Adam M. Wilson, Andrew M. Latimer e John A. Silander. «Point pattern modelling for degraded presence-only data over large regions». In: *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 60.5 (2011), pp. 757–776. ISSN: 00359254, 14679876. URL: http://www.jstor.org/stable/41262305 (visitato il 17/09/2024) (cit. a p. 41).