POLITECNICO DI TORINO

Corso di Laurea in Ingegneria Matematica

Tesi di Laurea Magistrale

Modelli di spostamento animale:
Simulazione e stima di movimento animale in base a coordinate GPS.

Relatore:
Gianluca Mastrantonio

Candidata:
Lucia Raucci

Anno Accademico 2024/2025

Un sentito ringraziamento va alla mia famiglia e a tutti coloro

che mi hanno sempre sostenuta in questo lungo percorso.

Indice

Introduzione

Da sempre il movimento animale è stato oggetto d'interesse, basti pensare al fatto che ne parla Aristotele nel *"De motu Animalium"*.
Inizialmente lo studio era principalmente rivolto alla ricerca di fonti di cibo, mentre adesso è fondamentale per comprendere dinamiche ecologiche e comportamentali.
Studiare come e perché gli animali si spostano nello spazio consente non solo di approfondire aspetti legati alla biologia e all'etologia, ma anche di monitorare gli effetti dei cambiamenti climatici e delle trasformazioni ambientali sugli ecosistemi, al fine di adottare provvedimenti atti al mantenimento delle risorse naturali.
Inoltre, negli ultimi anni, le tecnologie per la raccolta dei dati di spostamento, come il GPS o i dispositivi biologgers, hanno conosciuto uno sviluppo notevole, fornendo dati sempre più dettagliati e ad alta risoluzione. Di pari passo, anche i modelli statistici utilizzati per analizzare tali dati si sono evoluti, diventando sempre più complessi e raffinati.
Vi sono molti modelli che si possono prendere in considerazione.
Limitandosi a considerare il dominio temporale, si distinguono modelli a **tempo continuo** e modelli a **tempo discreto**. Nella realtà, il movimento degli animali avviene ovviamente in modo continuo, ma nella pratica è osservato a intervalli discreti, determinati dalla frequenza di registrazione del dispositivo.
Sebbene i modelli a tempo continuo siano, da un punto di vista teorico, più coerenti con la natura del movimento, quelli a tempo discreto risultano spes-

so più intuitivi e semplici da implementare.
Un altro aspetto rilevante nello studio del movimento animale è la presenza di diversi *stati comportamentali*, che riflettono strategie differenti adottate dall'animale in base al contesto ecologico.
L'animale può compiere diversi tipi di movimento in base alle sue necessità. Può permanere in uno stato stazionario, risultando fermo o muovendosi entro un'area ristretta; può assumere un comportamento tipico dell'esplorazione, effettuando spostamenti dalla direzione variabile e velocità moderata; oppure può manifestare una locomozione diretta, cioè uno spostamento più rapido e orientato verso una destinazione. Altri stati comportamentali tra i più frequentemente osservati includono il foraggiamento, associato alla ricerca attiva di cibo, la dispersione o la migrazione, che implicano movimenti su larga scala, ed infine comportamenti di fuga o evitamento, messi in atto in risposta a pericoli o ostacoli ambientali. Alcuni comportamenti possono anche essere indotti da condizioni ambientali, come disponibilità di risorse, ostacoli topografici o anche dall'andamento stagionale.

Oltre alla posizione, i modelli possono incorporare variabili derivate come la *step-length*, la *velocità*, il *turning angle* (l'angolo di deviazione tra due movimenti successivi), o il *bearing angle* (l'angolo rispetto a una direzione di riferimento).
L'approccio modellistico può essere ulteriormente arricchito includendo le dipendenze tra le traiettorie effettuate da individui della stessa specie o di specie diverse, mediante l'introduzione di termini di correlazione spaziale o strutture gerarchiche condivise.

Per descrivere queste dinamiche di movimento animale, si deve fare una distinzione tra il processo di movimento reale, continuo, e spesso non direttamente osservabile, ed il meccanismo di osservazione, discreto, che può essere affetto da errori ed irregolarità. Si utilizzano dunque modelli statistici che tengono conto dell'evoluzione temporale di variabili latenti, quali la posizione e la velocità, come gli *State-Space Models* (SSMs), che offrono una struttura

che ben si adatta a considerare separatamente il processo reale di movimento ed il meccanismo di osservazione.
In questa tesi verranno implementati modelli a tempo discreto e modelli a tempo continuo, basati su equazioni differenziali stocastiche; tra questi modelli sarà implementato il modello di Ornstein–Uhlenbeck (OU), un processo stocastico che descrive dinamiche tendenti alla regressione verso una media o un centro d'attrazione, con stima bayesiana dei parametri tramite metodi Monte Carlo via catene di Markov (MCMC).
Questi metodi generano campioni dalla distribuzione a posteriori, permettendo di stimare medie, intervalli credibili e probabilità.

Il processo OU verrà studiato inizialmente in un contesto discreto, successivamente in un contesto continuo, applicandolo dapprima a dati simulati e poi, limitatamente al caso discreto, verrà applicato a dati reali ottenuti dal monitoraggio GPS di lupi e orsi nel Parco Nazionale d'Abruzzo, Lazio e Molise. In un secondo momento si testerà la bontà del modello a tempo continuo su una traiettoria simulata a tempo discreto, aggiungendo un fattore di rumore o distorsione sull'intervallo temporale di acquisizione del dato.
Infine il modello a tempo discreto verrà utilizzato per stimare dei parametri a partire da dati reali, precedentemente sottoposti a una fase di analisi esplorativa e di pre-processing.
L'obiettivo finale è quello di ottenere inferenze robuste e affidabili sul comportamento spaziale degli animali analizzati e trarre delle conclusioni al fine di confrontare il comportamento di specie differenti, individuando dei trend stagionali.

Capitolo 1

Inferenza bayesiana in modelli di spostamento animale

L'inferenza bayesiana rappresenta un approccio particolarmente adatto per la stima dei parametri nei modelli di spostamento animale, grazie alla sua capacità di gestire l'incertezza in modo esplicito e flessibile.
Utilizzando il linguaggio probabilistico, nello specifico `Stan`, è possibile specificare in maniera diretta la struttura del modello e ottenere stime a posteriori affidabili tramite campionamento MCMC (Markov Chain Monte Carlo).

Il cuore dell'approccio bayesiano consiste nella combinazione tra una **funzione di verosimiglianza** $\mathcal{L}(\theta \mid \mathbf{y})$, che esprime quanto è probabile osservare i dati $\mathbf{y}$, dato un insieme di parametri θ e una **distribuzione a priori** $p(\theta)$, che rappresenta la conoscenza disponibile sui parametri prima di osservare i dati.

Questa combinazione fornisce la **distribuzione a posteriori**:

$$p(\theta \mid \mathbf{y}) = \frac{\mathcal{L}(\theta \mid \mathbf{y}) \cdot p(\theta)}{p(\mathbf{y})}$$

dove $p(\mathbf{y})$ è la costante di normalizzazione o evidenza.

Nel contesto dei modelli OU, questa formulazione permette di incorporare conoscenza a priori sui parametri di interesse, ovvero su una gamma di

valori plausibili, nel nostro caso specifico per θ, σ^2, μ, e consente di ottenere **intervalli di credibilità**, che quantificano l'incertezza delle stime, e di derivare **indicatori di copertura**, fondamentali per valutare l'affidabilità dell'inferenza in simulazioni.
Il modello è descritto in termini matematici, mentre l'esplorazione dei parametri è lasciata al campionamento MCMC. Il risultato è una distribuzione a posteriori da cui si possono derivare media, varianza, quantili ed altre statistiche di interesse.
L'approccio bayesiano risulta particolarmente efficace in caso di dati rumorosi o raccolti irregolarmente, in presenza di modelli non lineari o modelli con dipendenza temporale complessa o ancora con più livelli di incertezza, quali errori di misura o variabilità individuale.
Nell'acquisizione dei dati di movimento animale, queste problematiche risultano facilmente riscontrabili, dunque l'approccio utilizzato risulta essere sicuramente tra i più indicati.
Parte della teoria riportata di seguito è stata presa dall'Hooten [1].

1.1 Notazione

Convenzionalmente, i dati consistono in un set finito di localizzazioni geografiche $S = s_1, ..., s_n$, che rappresentano le osservazioni di un individuo in determinati istanti di tempo.
μ_i è la reale posizione dell'animale, che può differire dalla posizione misurata. Gli istanti t_i in cui viene osservato l'animale possono essere fissi o variabili. Se $\Delta t_i = t_i - t_{i-1}$ sono fissi, ci troviamo nel caso discreto e le posizioni osservate all'istante t possono essere indicizzate come s_t con $t \in (1, ..., n)$, tuttavia è più realistico che l'intervallo di tempo sia irregolare, anche perché il movimento animale è un processo a tempo continuo.

1.2 Concetti statistici e definizioni

Nel contesto dell'analisi statistica, i modelli più comunemente utilizzati sono i **modelli statistici parametrici**.

Definizione 1.1. Modello statistico parametrico. Un modello statistico parametrico specifica una famiglia di distribuzioni di probabilità note, che dipendono da un insieme finito di parametri sconosciuti. Questi parametri vengono stimati nel processo di adattamento del modello ai dati.

Un modello matematico generico può essere rappresentato come:

$$y_i \sim [y_i \mid \theta],$$

dove le y_i sono le osservazioni per $i = 1, \ldots, n$, e θ è il vettore dei parametri incogniti. L'espressione $[y_i \mid \theta]$ denota la distribuzione di probabilità nota di y_i condizionata da θ. Nel contesto frequentista, la **verosimiglianza** è la funzione:

$$L(\theta; \mathbf{y}) = [\mathbf{y} \mid \theta],$$

dove $\mathbf{y} = (y_1, \ldots, y_n)$ è il vettore di tutte le osservazioni. Se si assume che le osservazioni siano condizionatamente indipendenti, la verosimiglianza si fattorizza nel seguente modo:

$$[\mathbf{y} \mid \theta] = \prod_{i=1}^{n} [y_i \mid \theta].$$

L'approccio della **massima verosimiglianza (MLE)** consiste nel trovare il valore dei parametri che massimizza la funzione di verosimiglianza:

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta} [\mathbf{y} \mid \theta].$$

Nel paradigma **bayesiano**, si introduce una distribuzione a priori sui parametri:

$$\theta \sim [\theta],$$

dove $[\theta]$ è la distribuzione a priori, che può dipendere da iperparametri assunti noti. Tale distribuzione rappresenta la conoscenza o le assunzioni sui

parametri prima dell'osservazione dei dati. In alcuni casi, come nella selezione di modelli con regolarizzazione, la distribuzione a priori può essere scelta o ottimizzata tramite cross-validation.

L'obiettivo dell'analisi bayesiana è ottenere la **distribuzione a posteriori** dei parametri a partire dai dati osservati, secondo la formula di Bayes:

$$[\theta \mid \mathbf{y}] = \frac{[\mathbf{y} \mid \theta][\theta]}{[\mathbf{y}]},$$

dove il denominatore

$$[\mathbf{y}] = \int [\mathbf{y} \mid \theta][\theta]\, d\theta$$

è una costante di normalizzazione (nota anche come *evidenza*), che garantisce che la distribuzione a posteriori sia una vera distribuzione di probabilità.

Per modelli complessi, il calcolo dell'integrale al denominatore può non essere fattibile analiticamente. In questi casi si ricorre a metodi numerici o a tecniche di simulazione, come i metodi **Monte Carlo a catena di Markov (MCMC)**, che permettono di generare campioni dalla distribuzione a posteriori senza la necessità di calcolare esplicitamente la costante di normalizzazione.

Gli algoritmi MCMC sono flessibili e relativamente semplici da implementare, ma possono richiedere tempi di esecuzione elevati e attenzione nella diagnosi della convergenza.

Vi sono anche i modelli gerarchici, caratterizzati da una sequenza di distribuzioni di probabilità annidate per i dati, i processi e i parametri.

Un modello bayesiano gerarchico può essere espresso nel seguente modo:

$$y_{i,j} \sim [y_{i,j} \mid z_i, \theta]$$

$$z_i \sim [z_i \mid \beta]$$

$$\theta \sim [\theta]$$

$$\beta \sim [\beta]$$

dove z_i rappresenta un processo latente per l'individuo i e $y_{i,j}$ sono osservazioni ripetute per ciascun individuo $j = 1, \ldots, J$. Se il modello è bayesiano,

anche i parametri del processo β richiedono una distribuzione a priori. La distribuzione a posteriori del modello sarà del tipo:

$$[z, \theta, \beta \mid y] = \frac{[y \mid z, \theta][z \mid \beta][\theta][\beta]}{\int\int\int [y \mid z, \theta][z \mid \beta][\theta][\beta]\, dz\, d\theta\, d\beta}.$$

Il numeratore contiene la verosimiglianza dei dati e le distribuzioni a priori sui parametri, mentre il denominatore è la costante di normalizzazione, nota come *evidenza bayesiana*. Molti modelli gerarchici complessi risultano più facili da implementare in ambito bayesiano, ma ciò non sempre è necessario. Infatti non verranno utilizzati in questo elaborato, in quanto non saranno considerati stati latenti del processo.

1.3 Il metodo Monte Carlo a catena di Markov (MCMC)

Nell'inferenza statistica sono ricorrenti due principali classi di problemi numerici, quelli di integrazione e quelli di ottimizzazione. Spesso non è possibile calcolare analiticamente gli stimatori associati a un determinato paradigma inferenziale (massima verosimiglianza, Bayes, metodo dei momenti, ecc.) e si è dunque costretti a ricorrere a soluzioni numeriche.
Alla base di questo approccio vi è sicuramente la generazione di campioni di variabili casuali da distribuzioni arbitrarie. La simulazione permette di calcolare quantità di interesse integrando su distribuzioni di probabilità e il fatto di poter generare un numero arbitrario di variabili casuali da una certa distribuzione consente di sfruttare risultati frequentisti e asintotici, come la **legge dei grandi numeri** e il **teorema del limite centrale**, che permettono di valutare la convergenza dei metodi di simulazione con maggiore flessibilità rispetto ai contesti inferenziali classici, dove la dimensione campionaria è fissa.

Prima di introdurre i metodi Monte Carlo, è utile notare che un'alternativa naturale per l'approssimazione di integrali del tipo

$$\int_X h(x)f(x)\,dx,$$

dove f è una densità di probabilità, consiste nell'uso di tecniche numeriche deterministiche, ma in questo contesto, i metodi di simulazione, in particolare quelli Monte Carlo, rappresentano una strategia flessibile e potente per affrontare problemi complessi in inferenza statistica.

Integrazione Monte Carlo Classica

Si consideri il generico integrale

$$\mathbb{E}_f[h(X)] = \int h(x)f(x)\,dx,$$

dove f è una funzione di densità di una distribuzione di probabilità e h(x) è una funzione di cui si vuole calcolare il valore atteso sotto f. Il metodo Monte Carlo fornisce una soluzione pratica attraverso la generazione di un campione $X_1, \ldots, X_n \sim f$, e l'approssimazione del valore atteso mediante la media campionaria:

$$h_n = \frac{1}{n}\sum_{j=1}^{n} h(X_j)$$

Questa tecnica si basa sulla Legge dei Grandi Numeri, che garantisce la convergenza quasi certa di h_n a $\mathbb{E}_f[h(X)]$, e sul Teorema del Limite Centrale, che consente di valutare l'incertezza associata alla stima. Si richiamano di seguito la legge dei grandi numeri e il teorema del limite centrale.

Teorema 1.1. *Legge dei grandi numeri forte*

Sia $X_1, X_2, \ldots$ una successione di variabili aleatorie i.i.d. con valore atteso finito $\mu = \mathbb{E}[X_i]$. Allora con probabilità 1,

$$\frac{1}{n}\sum_{i=1}^{n} X_i \longrightarrow \mu \quad \textit{quando } n \to \infty,$$

ovvero la media campionaria converge al valore atteso.

Teorema 1.2. *Teorema del limite centrale*
Sia $X_1, X_2, \ldots$ *una successione di variabili aleatorie i.i.d. con* $\mathbb{E}[X_i] = \mu$ *e* $\mathrm{Var}(X_i) = \sigma^2$*. Allora:*

$$\frac{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1) \quad \textit{quando } n \to \infty$$

ovvero, la media campionaria standardizzata converge in distribuzione a una normale standard.
Quindi per $-\infty < a < +\infty$

$$\mathbb{P}\left\{\frac{X_1 + ... + X_N - N\mu}{\sigma\sqrt{n}}\right\} \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} dx \quad \textit{quando} \quad n \to \infty$$

Nel contesto della modellazione del movimento animale tramite un processo di Ornstein-Uhlenbeck (OU), l'integrazione Monte Carlo ha un ruolo implicito, perché essa è interamente gestita dal software utilizzato, ma è essenziale. L'inferenza Bayesiana sui parametri del modello richiede il calcolo di medie e probabilità condizionate rispetto a distribuzioni posteriori molto complesse.

Definizioni e concetti preliminari

Per comprendere al meglio come funzionano gli MCMC, riporto di seguito alcune utili definizioni.

Definizione 1.2. Processo stocastico
Un processo stocastico è una collezione di variabili casuali$\{x_i\}_{i \in I}$ per un insieme di indici ordinati I e con le variabili definite sullo stesso spazio degli stati S.

Per quanto riguarda la definizione delle catene di Markov, esse possono essere considerate sia a tempo discreto sia continuo.

Definizione 1.3. Proprietà di Markov a tempo discreto
Sia $(X_n)_{n \geq 0}$ un processo stocastico definito sullo spazio degli stati S.

$(X_n)_{n\geq 0}$ è una **catena di Markov a tempo discreto (DTMC)** se, per ogni $i_0, i_1, \ldots, i_{n-1}, i, j \in S$, vale la **proprietà di Markov**:

$$\mathbb{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathbb{P}(X_{n+1} = j \mid X_n = i) \stackrel{\text{def}}{=} p(i, j)$$

Questo significa che la probabilità associata a uno stato futuro dipende solo dallo stato attuale e **non** dalla storia passata del processo.

Definizione 1.4. Proprietà di Markov per processi continui
Diciamo che un processo continuo $(X_t)_{t\geq 0}$ gode della **proprietà di Markov** se, per ogni tempo $s, t > 0$ e ogni sequenza $0 \leq s_0 < s_1 < \cdots < s_n < s$ e stati $i_0, i_1, \ldots, i_n, i, j$, si ha:

$$\mathbb{P}(X_{t+s} = j \mid X_s = i, X_{s_n} = i_n, \ldots, X_{s_0} = i_0) = \mathbb{P}(X_{t+s} = j \mid X_s = i)$$

Definizione 1.5. Markov Chain Monte Carlo
I metodi Monte Carlo sono tecniche computazionali basate sulla simulazione casuale, utilizzate per risolvere problemi matematici complessi, quali integrali o ottimizzazioni, attraverso l'uso di campioni casuali.
La sequenza delle variabili casuali $X_1, ...X_n$ è una catena di Markov.

Definizione 1.6. Processo stazionario Un processo di Markov si dice **stazionario** se esiste una distribuzione di probabilità π tale che

$$\pi = \pi P,$$

dove P è la matrice di transizione. In tal caso, se $X_0 \sim \pi$, allora $X_n \sim \pi$ per ogni $n \in \mathbb{N}$, e la distribuzione del processo rimane invariata nel tempo.

Definizione 1.7. Processo reversibile Un processo di Markov con distribuzione stazionaria π si dice **reversibile** se, per ogni coppia di stati i, j, vale la condizione di bilanciamento dettagliato:

$$\pi(i)\, p(i, j) = \pi(j)\, p(j, i),$$

dove $p(i, j)$ è la probabilità di transizione da i a j (oppure $q(i, j)$ nel caso di catene di Markov a tempo continuo). In tal caso, il processo osservato all'indietro nel tempo ha la stessa distribuzione del processo diretto.

Capitolo 2

Modello a tempo discreto e continuo

In questo capitolo si parlerà dei modelli a tempo discreto con spazio degli stati continuo, partendo dal modello di passeggiata aleatoria, per poi arrivare a parlare di modelli a tempo continuo, in cui il processo è definito per ogni t in $[0, T]$, anche se a livello di implementazione il modello a tempo continuo verrà poi discretizzato.
Parte della teoria riportata di seguito è stata presa dall'Hooten [1], dal Casella [4] e da materiale didattico del corso di Processi Stocastici.

2.1 Modelli a tempo discreto

Alla base del modello a tempo discreto è essenziale supporre che sia nota la distanza temporale, Δt, tra due diverse osservazioni e che essa sia costante. Inoltre non si tenga in considerazione l'errore di misurazione, in modo da considerare direttamente la posizione μ_t al tempo t.
Per parlare di passeggiata aleatoria, o random walk (RW), è necessario richiamare le dinamiche Markoviane, assumendo quindi che la posizione al tempo t, μ_t dipenda solo dalla posizione all'istante precedente, $\mu_{t-\Delta t} = \mu_{t-1}$, visto che Δt è costante.

Se consideriamo una random walk al primo ordine, AR(1), ovvero un modello di serie temporale autoregressivo del primo ordine, possiamo scrivere

$$\mu_t = \mu_{t-1} + \epsilon_t, \tag{2.1}$$

per $t = 1, ..., T$, dove gli errori sono supposti essere indipendenti e normalmente distribuiti, $\epsilon_t \sim N(0, \Sigma)$. Nel caso più semplice $\Sigma \equiv \sigma^2 I$, con I matrice identità.
Essendo μ_t multidimensionale, questo modello viene chiamato anche VAR(1), ovvero modello vettore autoregressivo di ordine uno, nel quale lo stato attuale è influenzato dal passo precedente. In caso il modello fosse di ordine due, lo stato attuale sarebbe influenzato fino a due passi precedenti del processo e così via dicendo.
Nel modello VAR(1) si suppone che lo spostamento dell'individuo ad ogni istante avvenga in una direzione casuale e con la step-length distribuita come una Weibull [1] univariata. In questo caso la varianza σ^2 controlla la step-length tra una posizione e l'altra.
In Figura 2.1 sono riportate le distribuzioni empiriche e teoriche delle step-length ottenute da una traiettoria in 2-D simulata usando l'equazione 2.1, con T = 10000 tempi di osservazione e $\sigma^2 = (0.5, 1, 2)$. Si può notare come al crescere di σ^2 aumentino la dispersione e la variabilità attorno alla media.
La formulazione espressa dall'equazione 2.1 è spesso descritta come modello autoregressivo intrinseco (ICAR), in quanto lo stato attuale è influenzato unicamente dal precedente indipendentemente dagli altri stati. Inoltre, il processo ICAR è non stazionario e irreversibile, in quanto non è possibile

[1] La distribuzione di Weibull con parametri di scala $\lambda > 0$ e di forma $k > 0$ è definita sui reali positivi. La sua funzione di ripartizione è:

$$F(x) = 1 - e^{-(x/\lambda)^k}$$

e la corrispondente funzione di densità di probabilità è:

$$f(x) = \frac{k}{\lambda^k} x^{k-1} e^{-(x/\lambda)^k}$$

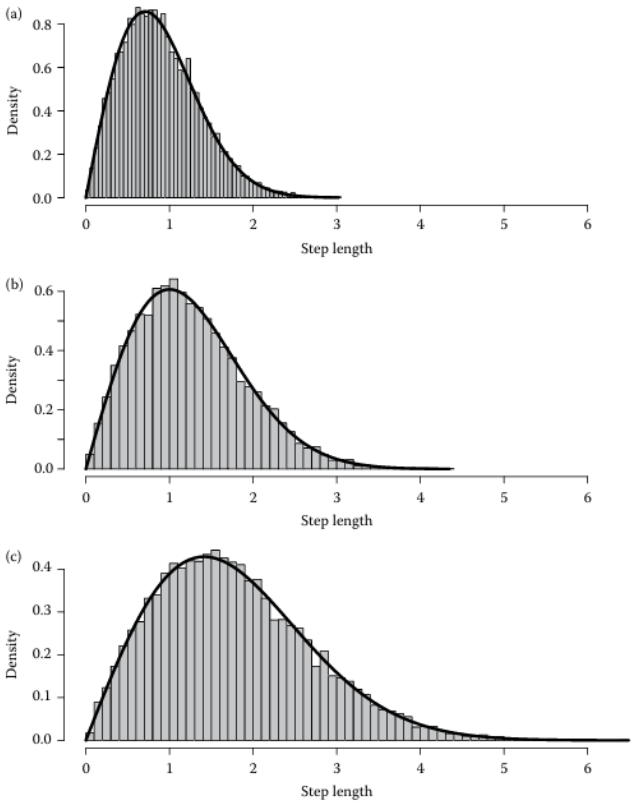


Figura 2.1: Distribuzioni empiriche (istogrammi) e teoriche (linea continua) della lunghezza del passo, basate su traiettorie simulate utilizzando l'Equazione (2.1), con $T = 10000$ e **(a)** $\sigma^2 = 0.5$, **(b)** $\sigma^2 = 1$, **(c)** $\sigma^2 = 2$. La distribuzione teorica della lunghezza del passo per una random walk 2D con incrementi gaussiani è una Weibull con forma 2 e scala $\sqrt{2}\sigma$, cioè Weibull$(k = 2, \lambda = \sqrt{2}\sigma)$.

tornare esattamente nello stato precedente e non vi sono altri vincoli nel processo.

Infine uno degli aspetti rilevanti dei modelli condizionali autoregressivi è che è facile passare da un modello al primo ordine, in questo caso dipendente dalle medie, ad uno del secondo ordine, ovvero dipendente dalle covarianze. Infatti se consideriamo $\mu \equiv (\mu'_1, ..., \mu'_T)$, è possibile scrivere la probabilità congiunta di ogni posizione come $\mu \sim N(\mathbf{I} \otimes \hat{\mu}, \Sigma_\mu \otimes \mathbf{I})$.

Relativamente al modello RW, verranno considerate le seguenti generaliz-

zazioni, che saranno utili alla trattazione anche di modelli continui ed, in particolar modo, del processo di Ornstein-Uhlenbeck.

Attrazione

Un'estensione del modello VAR(1) permette di includere la trattazione di punti di attrazione o di una zona centrale.
Si può imporre un attrattore rendendo il processo stazionario nel modo che segue:

$$\mu_t - \mu^* = M(\mu_{t-1} - \mu^*) + \varepsilon_t, \tag{2.2}$$

dove $\mu*$ è il centroide del processo di movimento e M è la matrice di propagazione che controlla la dinamica del processo. $M \equiv \rho\mathbf{I}$, con ρ parametro che controlla la regolarità del processo.
Il propagatore controlla indipendentemente sia la latitudine sia la longitudine.
Possiamo riscrivere l'equazione di prima nel seguente modo

$$\mu_t = M\mu_{t-1} + (I - M)\mu^* + \varepsilon_t, \tag{2.3}$$

dove nel caso in cui ρ fosse uguale a 1, si tornerebbe al modello ICAR, mentre per avere un'attrazione verso il punto μ esso dovrebbe essere compreso tra 0 e 1.
Imponendo il seguente vincolo su ρ

$$-1 < \rho < 1,$$

esso può essere interpretato come coefficiente di correlazione.
In Figura 2.2 si vedono due traiettorie simulate dall'equazione 2.3, con attrattore $\mu* = (1, 1)'$, $\sigma^2 = 1$ e $\rho = 0.5$ per le figure a sinistra e $\rho = 0.95$ per quelle a destra. Entrambi i processi sono stazionari attorno a $\mu*$ e il parametro ρ vicino a 1, forza la traiettoria a essere più liscia e regolare.

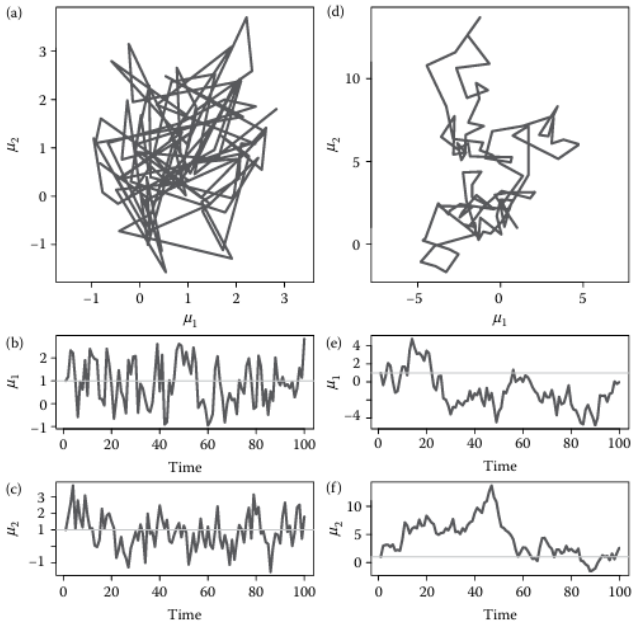


Figura 2.2: Grafici congiunti (a, d) e marginali (b, c, e, f) di serie temporali VAR(1) simulate a partire dall'Equazione (2.3), con $\mu^* = (1, 1)$ e $\sigma^2 = 1$ in entrambi i casi. I pannelli (a–c) mostrano μ_t, $\mu_{1,t}$ e $\mu_{2,t}$ con $\rho = 0{,}5$, mentre i pannelli (d–f) mostrano μ_t, $\mu_{1,t}$ e $\mu_{2,t}$ con $\rho = 0{,}95$.

Dati irregolari, allineamento temporale

Sino ad ora sono stati considerati dati equidistanti nel tempo.

Poiché il processo di movimento può essere incorporato come componente latente in un modello gerarchico, la risoluzione temporale diventa una scelta dell'utente. La scelta effettiva del passo temporale t è direttamente connessa all'inferenza che si ottiene dal modello. Ad esempio, se $t = 1$ ora, i parametri che controllano la dinamica del movimento sono interpretabili sulla scala temporale di un'ora. In tal caso, una quantità come il turning angle rappresenta l'angolo associato al vettore di spostamento complessivo nel periodo di un'ora. Solitamente si sceglie la scala temporale del processo corrispondente alla scala più regolare a cui i dati appaiono, e si sviluppa un modello di errore di misura che si adatti alla scala del processo. Un approccio potrebbe essere

costituito dall'allineare le scale dei dati e del processo tramite interpolazione lineare.

Sia $\mathbf{s}_i$ la posizione osservata dell'animale al tempo i, per $i = 1, \ldots, n$ osservazioni e siano t_i i tempi di osservazione.
Si possono collegare i tempi di osservazione con i tempi del processo tramite una media pesata nel seguente modo:

$$\mathbf{s}_i \sim \mathcal{N}\left((1 - w_i)\boldsymbol{\mu}_{t^- \Delta t} + w_i \boldsymbol{\mu}_t,\, \sigma_s^2 \mathbf{I}\right), \tag{2.4}$$

dove $\boldsymbol{\mu}_{t^-}$ e $\boldsymbol{\mu}_t$ rappresentano rispettivamente i valori del processo nei tempi immediatamente prima e dopo t_i. Il peso w_i è funzione della distanza temporale tra t_i e i tempi del processo, ed è definito come:

$$w_i = \frac{t - t_i}{\Delta t}. \tag{2.5}$$

Questo modello è sufficientemente generale da garantire che, quando t_i coincide con un tempo del processo, il dato osservato venga associato esattamente alla posizione del processo. Nei casi in cui t è piccolo rispetto alla frequenza di movimento dell'animale, questo tipo di interpolazione lineare offre buone prestazioni. Tuttavia, quando t aumenta, essa può non essere più appropriata.

2.2 Modelli a tempo continuo

Vi sono due principali approcci ai modelli di movimento animale. Uno di tipo Lagrangiano, basato sul monitoraggio di un singolo individuo a tempo discreto, ed uno Euleriano, basato sul monitoraggio di una popolazione di individui a tempo continuo. Turchin, uno scienziato russo-americano noto per i suoi studi in ambito di cliodinamica, illustra il passaggio da modelli lagrangiani a tempo discreto a modelli euleriani a tempo continuo, partendo da un'equazione di ricorrenza sino ad arrivare a una nota equazione alle derivate parziali.

Alla base dello sviluppo di un modello lagrangiano vi è l'idea di partire da semplici principi. Si parta ad esempio da un dominio spaziale unidimensionale, dove l'individuo preso in analisi potrà muoversi di un'unità di spazio a destra o a sinistra o restare nella posizione corrente per l'intervallo di tempo previsto.
Si consideri dunque un animale nella posizione μ con probabilità di muoversi a destra $\phi_R(\mu,t)$, a sinistra $\phi_L(\mu,t)$ e di restare nella stessa posizione $\phi_N(\mu,t)$, tali che $\phi_R(\mu,t)+\phi_L(\mu,t)+\phi_N(\mu,t=1)$.
La probabilità che l'animale occupi la posizione μ al tempo t è

$$\begin{aligned} p(\mu,t) = {} & \phi_L(\mu+\Delta\mu,t-\Delta t)\,p(\mu+\Delta\mu,t-\Delta t) \\ & +\phi_R(\mu-\Delta\mu,t-\Delta t)\,p(\mu-\Delta\mu,t-\Delta t) \\ & +\phi_N(\mu,t-\Delta t)\,p(\mu,t-\Delta t), \end{aligned} \tag{2.6}$$

dove il $\Delta\mu$ rappresenta lo spostamento a destra o a sinistra.
Da questa equazione Turchin ha sviluppato ogni probabilità in serie di Taylor, troncando i termini di ordine superiore e sostituendo i termini nell'equazione 2.6.
L'espansione in serie di Taylor porta ad un'equazione di ricorrenza alle derivate parziali:
(2.7)

$$\begin{aligned} p = {} & (\phi_L+\phi_N+\phi_R)\,p - \Delta t\,(\phi_L+\phi_N+\phi_R)\,\frac{\partial p}{\partial t} - \Delta t\,p\,\frac{\partial}{\partial t}(\phi_L+\phi_N+\phi_R) \\ & - \Delta\mu(\phi_R-\phi_L)\,\frac{\partial p}{\partial \mu} - \Delta\mu p\,\frac{\partial}{\partial\mu}(\phi_R-\phi_L) + \frac{\Delta\mu^2}{2}(\phi_L+\phi_R)\,\frac{\partial^2 p}{\partial\mu^2} \\ & + \Delta\mu^2\,\frac{\partial p}{\partial\mu}\,\frac{\partial}{\partial\mu}(\phi_L+\phi_R) + p\,\frac{\Delta\mu^2}{2}\,\frac{\partial^2}{\partial\mu^2}(\phi_L+\phi_R) + \cdots \end{aligned}$$

dove $p \equiv p(\mu,t)$, $\phi_L \equiv \phi_L(\mu,t)$, $\phi_N \equiv \phi_N(\mu,t)$, e $\phi_R \equiv \phi_R(\mu,t)$. Combinando i termini simili e troncando i termini di ordine superiore nell'Equazione 2.7, otteniamo un'equazione differenziale alle derivate parziali (PDE) della forma:

$$\frac{\partial p}{\partial t} = -\frac{\partial}{\partial\mu}(\beta p) + \frac{\partial^2}{\partial\mu^2}(\delta p), \tag{2.8}$$

dove $\beta = \frac{\mu(\phi_R - \phi_L)}{\Delta t}$ e $\delta = \frac{(\Delta\mu)^2(\phi_R + \phi_L)}{2\Delta t}$. Il modello risultante dell'Equazione 2.8 è di tipo euleriano e l'equazione è nota come equazione di Fokker-Planck o di Kolmogorov.
Si può estendere il modello a livello di popolazione e considerare l'intensità spaziale $u(\mu, t)$ per un numero totale di animali N, ponendo $u(\mu, t) \equiv Np(\mu, t)$. In questo contesto, assumendo momentaneamente che non ci sia una componente di deriva o bias, cioè $\beta = 0$, otteniamo l'equazione di diffusione ecologica:

$$\frac{\partial u}{\partial t} = \frac{\partial^2}{\partial \mu^2}(\delta u), \tag{2.9}$$

dove il processo di interesse è $u \equiv u(\mu, t)$, e $\delta \equiv \delta(\mu, t)$ rappresenta i coefficienti di diffusione che possono variare nello spazio e nel tempo. Nel contesto del movimento animale, il parametro di diffusione δ rappresenta la *motilità* degli animali.

Una riduzione alternativa dell'equazione di Fokker-Planck si può ottenere assumendo $\delta = 0$, implicando che il movimento animale sia guidato unicamente dall'advezione. Sebbene meno intuitiva, tale situazione potrebbe verificarsi in popolazioni trasportate da vento o acqua (ad esempio, dispersione di uova in un sistema fluviale), oppure in casi in cui sia presente una forte attrazione o repulsione verso caratteristiche spaziali.

Equazioni differenziali stocastiche

Dopo aver dimostrato come convertire un modello stocastico a tempo discreto ad un modello a tempo continuo, torniamo alla definizione di random walk a tempo discreto 2.1 e sia b la posizione al posto di μ:

$$b(t_i) = b(t_{i-1} + \epsilon(t_i), \tag{2.10}$$

dove il passo temporale è supposto essere $\Delta_i = t_i - t_{i-1}$ e $\epsilon(t_i)$ dipende da Δ_i nel seguente modo:

$$\epsilon(t_i) \sim N(0, \Delta_i \mathbf{I}).$$

Per grandi intervalli di tempo, la step length sarà maggiore della media. Per semplicità si considera il caso in cui tutti gli intervalli temporali siano uguali, $\Delta_i = \Delta t \ \forall i$.
Un altro modo per scrivere il modello per la posizione corrente $b(t_i)$ è il seguente:

$$b(t_i) = \sum_{j=i}^{i} b(t_j) - b(t_{j-1}) = \sum_{j=i}^{i} \epsilon(t_j), \tag{2.11}$$

partendo da una posizione iniziale $b(t_0) = (0, 0)'$ con $t_0 = 0$. Così si arriva ad avere una somma cumulativa di incrementi gaussiani indipendenti $\epsilon(t_i)$. Se si considera il limite per Δt che tende a zero, si ottiene il passaggio da modello discreto a continuo ed in particolare ci si trova in presenza di un moto Browniano:

$$b(t_i) = \lim_{\Delta t \to 0} \sum_{j=1}^{i} \varepsilon(t_j), \tag{2.12}$$

che è noto principalmente come integrale stocastico di Itô, mentre la sequenza di $b(t) \ \forall t$ è nota come processo di Wiener o moto Browniano.

Definizione 2.1. Moto browniano Un processo stocastico $\{B_t\}_{t \geq 0}$ si dice *moto browniano standard* se soddisfa le seguenti proprietà:

- $B_0 = 0$;
- per ogni $0 \leq s < t$, gli incrementi $B_t - B_s$ sono indipendenti dal passato, ossia da $\{B_u : u \leq s\}$;
- gli incrementi sono stazionari: $B_t - B_s \sim \mathcal{N}(0, t - s)$;
- le traiettorie del processo sono continue quasi sicuramente, ma non derivabili in nessun punto.

Il moto browniano gode di molte altre proprietà:

- è un processo di **Markov**, ossia la probabilità condizionata futura dipende solo dallo stato corrente e non dal passato.

- È una **martingala** con media nulla: $\mathbb{E}[B_t \mid \mathcal{F}_s] = B_s$ per $s < t$.
- Ha **incrementi stazionari**, cioè la distribuzione di $B_t - B_s$ dipende solo da $t - s$.
- Per ogni $t \geq 0$, si ha: $B_t \sim \mathcal{N}(0, t)$.

Riportiamo di seguito definizioni utili alla comprensione del moto browniano.

Definizione 2.2 (**Filtrazione**)**.** Sia $(\Omega, \mathcal{F}, \mathbb{P})$ uno spazio di probabilità. Una famiglia crescente di σ-algebre $\{\mathcal{F}_n\}_{n\geq 0}$, tale che $\mathcal{F}_n \subseteq \mathcal{F}_{n+1} \subseteq \mathcal{F}$ per ogni n, si dice una *filtrazione*. Essa rappresenta l'informazione disponibile fino al tempo n.

Definizione 2.3 (**Martingala**)**.** Sia $\{M_n\}_{n\geq 0}$ una successione di variabili aleatorie reali definite su uno spazio di probabilità $(\Omega, \mathcal{F}, \mathbb{P})$ e adattate a una filtrazione $\{\mathcal{F}_n\}_{n\geq 0}$. La successione $\{M_n\}$ si dice una *martingala* rispetto alla filtrazione $\{\mathcal{F}_n\}$ se, per ogni $n \geq 0$, valgono le seguenti condizioni:

1. M_n è $\mathcal{F}_n$-misurabile;
2. $\mathbb{E}[|M_n|] < \infty$;
3. $\mathbb{E}[M_{n+1} \mid \mathcal{F}_n] = M_n$ quasi sicuramente.

Definizione 2.4. Integrale stocastico di Itô

Poiché le traiettorie del moto browniano B_t non sono derivabili in senso classico e presentano variazione infinita su intervalli finiti, l'integrazione tradizionale (di Riemann o Lebesgue) non è applicabile. Per affrontare questa difficoltà, si introduce l'*integrale di Itô*, definito come il limite in probabilità di somme stocastiche:

$$\int_0^t f(s)\, dB_s := \lim_{n\to\infty} \sum_{i=0}^{n-1} f(t_i)(B_{t_{i+1}} - B_{t_i}),$$

dove $\{t_i\}$ è una partizione di $[0, t]$, e la funzione $f(s)$ è adattata alla filtrazione generata dal processo B_s, ovvero non dipende dal futuro.

Proprietà fondamentali dell'integrale di Itô:

- L'integrale di Itô è una *martingala* se $f(s)$ è un processo prevedibile e quadrato sommabile.

- La sua *variazione quadratica* è:

$$\left\langle \int_0^t f(s)\, dB_s \right\rangle = \int_0^t f^2(s)\, ds.$$

- La regola di Itô (o formula di Itô) generalizza la regola della catena per processi stocastici.

Gli integrali di Itō vengono scritti solitamente come

$$b(t) = \int_0^t db(\tau) = \int_0^t \frac{db(\tau)}{d\tau} d\tau, \tag{2.13}$$

dove $db(t) = \varepsilon(t)$, implica che i vettori di spostamento individuali siano la "derivata" di $b(t)$ quando $t \to 0$. In modo informale, possiamo pensare a

$$db(t) = b(t) - b(t - \Delta t) \quad \text{con} \quad \Delta t \to 0.$$

Quindi, per semplicità, si usa spesso la notazione integrale standard nei modelli di equazioni differenziali stocastiche (ad esempio, il moto browniano), mentre in realtà andrebbe usata la notazione sommatoria come in 2.12.

É anche comune scrivere l'integrale di $b(t)$ come

$$b(t) = \int_0^t db(\tau), \tag{2.14}$$

perché l'integrale di una funzione costante rispetto al processo browniano $b(t)$ è collegato all'integrale di $\varepsilon(t)$ rispetto al tempo.

I vettori di spostamento originali $\varepsilon(t)$ sono casuali, quindi il processo browniano $b(t)$ è anch'esso casuale. Infatti, nel tipo di moto browniano descritto, il valore atteso di $b(t)$ è zero e la varianza è t. La covarianza del processo tra i tempi t_i e t_j è $\min(t_i, t_j)$, mentre la correlazione è $\frac{\min(t_i,t_j)}{\max(t_i,t_j)}$, e la covarianza tra

due differenze separate del processo browniano è zero. Il processo browniano ha anche la proprietà utile che

$$b(t_i) - b(t_j) \sim \mathcal{N}(0, |t_i - t_j|I),$$

dove $|t_i - t_j|$ rappresenta il tempo tra t_i e t_j.

Per generalizzare il processo di moto browniano affinché possa essere localizzato e scalato per un processo di posizione specifico $\mu(t_i)$, riprendiamo l'equazione del modello discreto:

$$\mu(t_i) = \mu(t_{i-1}) + \varepsilon(t_i), \tag{2.15}$$

assumendo che la posizione iniziale sia $\mu(0)$ e scalando il processo con $\varepsilon(t_i) \sim \mathcal{N}(0, \sigma^2 tI)$, dove σ^2 allarga o restringe la traiettoria nello spazio.

Usando la notazione del moto browniano, questo modello diventa

$$\mu(t) = \mu(0) + \sigma b(t), \tag{2.16}$$

per ogni t.

Ponti Browniani

In questo ambito sono molto utilizzati i ponti browniani, in quanto utili a modellare i dati di telemetria.

Un ponte Browniano è un processo di moto browniano con tempi e posizioni di partenza e arrivo noti e fissi.

Tornando alla notazione della posizione $\mu(t)$ utilizzata in precedenza, Horne et al. (2007) descrivono il Brownian bridge come un processo casuale multivariato normale tale che

$$\mu(t) \sim \mathcal{N}\left(\mu(t_{i-1}) + \frac{t - t_{i-1}}{t_i - t_{i-1}}\big(\mu(t_i) - \mu(t_{i-1})\big), \frac{(t - t_{i-1})(t_i - t)}{t_i - t_{i-1}}\sigma^2\right), \tag{2.17}$$

per $t_{i-1} < t < t_i$, dove $\mu(t_{i-1})$ e $\mu(t_i)$ sono note. L'equazione 2.17 definisce una distribuzione normale multivariata centrata su una distanza scalata tra

gli endpoint $\mu(t_{i-1})$ e $\mu(t_i)$. La varianza del processo al tempo t diminuisce al crescere della vicinanza temporale agli istanti iniziale t_{i-1} o finale t_i.

Per situazioni con errori gaussiani di misura, le posizioni osservate $s(t_i)$ possono essere modellate come

$$s(t_i) \sim \mathcal{N}(\mu(t_i), \sigma_s^2 I), \quad i = 1, \ldots, n,$$

introducendo così una struttura gerarchica naturale al modello. Molti metodi comuni per l'implementazione di questi modelli integrano sul processo browniano $\mu(t)$ per adattare il modello tramite metodi di massima verosimiglianza.

Horne et al. (2007) propongono un approccio che condiziona ogni altra osservazione come punto finale, utilizzando le posizioni intermedie come dati per adattare il modello Brownian bridge, sfruttando la proprietà di indipendenza tramite triplette di dati (x_i, x_{i+1}, x_{i+2}), in cui l'osservazione intermedia x_{i+1} è trattata come dato osservato da confrontare con la distribuzione del Brownian Bridge condizionata su x_i e x_{i+2}. Dopo una prima scansione dei dati, l'algoritmo ripete il ciclo con uno shift delle triplette, ottenendo una "dimensione del campione" di circa $n/2$.

Nonostante l'efficienza computazionale di questo metodo, Pozdnyakov et al. (2014) evidenziano diversi problemi:il metodo di Horne et al. (2007) produce un bias nella stima della varianza del movimento σ^2 che aumenta con la varianza dell'errore di misura σ_s^2; le varianze del movimento e dell'errore di misura non sono identificabili nel modello di verosimiglianza, specialmente con intervalli temporali equidistanti; solo circa la metà dei dati viene utilizzata per adattare il modello.

Pozdnyakov et al. (2014) mostrano che la varianza delle posizioni osservate è

$$\operatorname{var}(s(t_i)) = \sigma^2 t_i I + \sigma_s^2 I, \tag{2.18}$$

e la covarianza è

$$\operatorname{cov}(s(t_i), s(t_j)) = \sigma^2 \min(t_i, t_j) I. \tag{2.19}$$

La matrice di covarianza per i dati di telemetria congiunti è quindi densa, completamente popolata da elementi diversi da zero. Tuttavia, la matrice di covarianza per le velocità osservate (cioè, $s(t_i) - s(t_{i-1})$) è tri-diagonale, ma non diagonale, poiché la varianza dell'errore di misura si manifesta anche fuori dalla diagonale. Questo implica che la natura non diagonale della matrice di covarianza diventa sempre più importante all'aumentare dell'errore di misura. Gli elementi diagonali della matrice di covarianza per le velocità osservate sono

$$\sigma^2(t_i - t_{i-1}) + \sigma_s^2.$$

Pozdnyakov et al. (2014) suggeriscono di utilizzare la distribuzione congiunta di tutte le velocità, che è multivariata normale, come funzione di verosimiglianza per adattare il modello di moto browniano, invece dei metodi basati sul Brownian bridge di Horne et al. (2007), affermando che il loro approccio è altrettanto semplice da implementare.

Pertanto, anziché condizionare su una sequenza incrementale di punti finali, è più vantaggioso modellare il processo di movimento animale come un vero processo dinamico continuo nel tempo.

Segue quindi un ritorno alla modellazione della covarianza proposta da Pozdnyakov et al. (2014) per una classe più ampia di modelli di movimento basati su processi stocastici in tempo continuo.

Attrazione e drift

Il moto browniano $b(t)$ produce traiettorie più regolari rispetto al rumore bianco $\varepsilon(t)$ perché è una quantità integrata. Questa è la ragione per cui il moto browniano è spesso scelto come modello per descrivere lo spostamento degli animali nel tempo continuo. Tuttavia, come mostrato nell'Equazione 2.12, il moto browniano non è un modello molto flessibile per lo spostamento, poiché manca di componenti di deriva e di attrazione.

Ricordiamo che possiamo convertire un processo a tempo discreto in un processo a tempo continuo seguendo questi passaggi:

1. Specificare la ricorrenza stocastica

$$\mu(t_i) = \mu(0) + \sum_{j=1}^{i} \big(\mu(t_j) - \mu(t_{j-1})\big). \tag{2.20}$$

2. Specificare il modello parametrico condizionale a tempo discreto per $\mu(t_j)$.

3. Sostituire il modello per $\mu(t_j)$ nel membro di destra dell'Equazione 2.20.

4. Prendere il limite di $\mu(t_i)$ per $\Delta t \to 0$ per ottenere la rappresentazione in forma integrale di Itô del processo a tempo continuo.

5. Se desiderato, riscrivere il modello in termini della derivata di Itô di $\mu(t)$.

Per dimostrare questa procedura, supponiamo di voler aggiungere un'attrazione puntuale al processo di moto browniano. In questo caso, ricordiamo il modello a tempo discreto per l'attrazione dal Capitolo 5:

$$\mu(t_i) = M\mu(t_{i-1}) + (I - M)\mu^* + \varepsilon(t_i),$$

dove M è la matrice di propagazione VAR(1), μ^* è la posizione di attrazione e $\varepsilon(t) \sim N(0, \sigma^2 tI)$.

Sostituendo questo modello condizionale a tempo discreto nell'Equazione 2.20 per $\mu(t_j)$ otteniamo:

$$\begin{aligned}
\mu(t_i) = \mu(0) + \sum_{j=1}^{i} \big(\mu(t_j) - \mu(t_{j-1})\big) &= \mu(0) + \sum_{j=1}^{i} \big(M\mu(t_{j-1}) + (I - M)\mu^* + \varepsilon(t_j) - \mu(t_{j-1})\big) = \\
&= \mu(0) + \sum_{j=1}^{i} \big((M - I)(\mu(t_{j-1}) - \mu^*) + \varepsilon(t_j)\big) \\
&= \mu(0) + \sum_{j=1}^{i} (M - I)(\mu(t_{j-1}) - \mu^*) + \sum_{j=1}^{i} \varepsilon(t_j).
\end{aligned}$$

Riconosciamo nell'ultimo termine $\sum_{j=1}^{i} \varepsilon(t_j)$ l'elemento cardine del moto browniano. Quindi, prendendo il limite del membro di destra per $t \to 0$ si

ottiene l'equazione integrale di Itô

$$\mu(t) = \mu(0) + \int_0^t (M - I)(\mu(\tau) - \mu^*)d\tau + \int_0^t \sigma db(\tau). \tag{2.21}$$

L'equazione integrale contiene tre componenti: la quantità $\mu(0)$, che fornisce la posizione iniziale corretta, il processo di attrazione

$$\int_0^t (M - I)(\mu(\tau) - \mu^*)d\tau,$$

e il processo di moto browniano scalato $\sigma b(t)$.

Derivando secondo Itô entrambi i membri, otteniamo l'equazione differenziale stocastica (SDE) per il moto browniano con attrazione

$$d\mu(t) = (M - I)(\mu(t) - \mu^*)dt + \sigma db(t) = (M - I)(\mu(t) - \mu^*)dt + \varepsilon(t).$$

Si noti che $\varepsilon(t) \sim N(0, \sigma^2 dtI)$ e la forma sopra è comune nella letteratura sulle SDE, ma può anche essere scritta come

$$\frac{d\mu(t)}{dt} = (M - I)(\mu(t) - \mu^*) + \frac{\varepsilon(t)}{dt}. \tag{2.22}$$

Questa equazione rappresenta un'equazione differenziale con un termine additivo corrispondente al moto browniano derivato. Questo è ciò che distingue le SDE dalle equazioni differenziali deterministiche con errore additivo: il termine di errore $\varepsilon(t)$ è "incorporato" nella derivata del processo di posizione $\mu(t)$.

Possiamo riscrivere l'equazione integrale stocastica in parole come

Posizione = posizione iniziale + deriva cumulativa + diffusione cumulativa.

La deriva cumulativa integra il processo di deriva, che sono gli spostamenti propagati dal punto di attrazione μ^*. La diffusione cumulativa integra gli incrementi non correlati o "errori" per ottenere un processo di movimento correlato, il moto browniano. Insieme, questi due componenti forniscono un modello realistico di movimento continuo per animali come i foraggiatori centrali. L'espressione fornisce inoltre un modo generale per caratterizzare molti modelli SIE modificando le componenti di deriva e diffusione.

Abbiamo iniziato con un semplice processo di moto browniano senza attrazione e abbiamo aggiunto un termine di deriva, ottenendo un modello più flessibile per la posizione reale del processo. La SIE risultante non è browniana, ma contiene una componente browniana; infatti, l'equazione SDE rappresenta un modo per descrivere un processo di Ornstein-Ulhenbeck.

Modelli di Ornstein-Uhlenbeck

Nella sezione precedente, abbiamo osservato che il moto browniano con attrazione 2.22 è denominato processo di Ornstein-Uhlenbeck (OU). In effetti, lo abbiamo derivato differenziando un'equazione integrale stocastica (SIE) 2.21 che ha origine da una sequenza di argomentazioni euristiche basate sulla somma di infiniti passi. Tuttavia, il processo OU è spesso espresso in notazione esponenziale (ad esempio, Dunn e Gipson 1977; Blackwell 2003; Johnson et al. 2008a).

Per arrivare all'espressione OU che coinvolge esponenziali, notiamo che è più comune nella modellazione matematica partire dall'equazione differenziale stocastica (SDE) che coinvolge il processo velocità e poi "risolverla" per trovare il processo posizione $\mu(t)$. Per dimostrare come tipicamente si derivano le soluzioni alle SDE, iniziamo con una SDE semplificata basata sull'Equazione 2.22 in uno spazio unidimensionale, con attrattore $\mu^* = 0$, varianza browniana $\sigma^2 = 1$, e parametro di autocorrelazione θ, tale che

$$d\mu(t) = -\theta\mu(t)dt + db(t). \tag{2.23}$$

Una tecnica di soluzione coinvolge un metodo di variazione delle costanti. In questo caso, moltiplichiamo entrambi i membri dell'Equazione 2.23 per $e^{\theta t}$ e poi integriamo entrambi i membri da 0 a t. Il termine $e^{\theta t}$ semplifica l'integrazione richiesta e permette una soluzione analitica. Quindi, moltiplicando entrambi i membri dell'Equazione 2.23 per $e^{\theta t}$, otteniamo

$$e^{\theta t}d\mu(t) = -\theta e^{\theta t}t\mu(t)dt + e^{\theta t}db(t). \tag{2.24}$$

Integrando entrambi i membri dell'Equazione 2.24 da 0 a t, otteniamo

$$\int_0^t e^{\theta\tau} d\mu(\tau) d\tau = -\theta \int_0^t e^{\theta\tau} \mu(\tau) d\tau + \int_0^t e^{\theta\tau} db(\tau). \tag{2.25}$$

L'integrale a sinistra nell'Equazione 2.25 può essere risolto usando l'integrazione per parti:

$$\int_0^t e^{\theta\tau} d\mu(\tau) d\tau = e^{\theta t}\mu(t) - \mu(0) - \int_0^t \mu(\tau)\theta e^{\theta\tau} d\tau. \tag{2.26}$$

Sostituendo l'Equazione 2.26 nell'Equazione 2.25 otteniamo

$$e^{\theta t}\mu(t) - \mu(0) - \int_0^t \mu(\tau)\theta e^{\theta\tau} d\tau = -\theta \int_0^t e^{\theta\tau} \mu(\tau) d\tau + \int_0^t e^{\theta\tau} db(\tau), \tag{2.27}$$

che, dopo qualche passaggio algebrico, si semplifica in

$$\mu(t) = \mu(0)e^{-\theta t} + \int_0^t e^{-\theta(t-\tau)} db(\tau). \tag{2.28}$$

La soluzione risultante ha diverse proprietà interessanti. Primo, si noti che, per $t \to \infty$, il primo termine a destra dell'Equazione 2.28 si annulla (cioè $\mu(0)e^{-\theta t} \to 0$). Questo risultato implica che, all'aumentare del tempo, la posizione iniziale ha un effetto minore sulla soluzione $\mu(t)$. Secondo, l'integrale a destra è una convoluzione di $\exp(-\theta(t-\tau))$ con un processo di rumore bianco. Per determinare media e varianza di questa variabile casuale, torniamo alla rappresentazione come somma infinita dell'integrale di Itô. Così,

$$\int_0^t e^{-\theta(t-\tau)} db(\tau) = \lim_{\Delta t \to 0} \sum_{j=1}^{i} e^{-\theta(t-t_j)} \big(b(t_j) - b(t_{j-1})\big), \tag{2.29}$$

dove $t_0 = 0$ e $t_i = t$. Per ogni Δt, la somma pesata $\sum_j e^{-\theta(t-t_j)}(b(t_j) - b(t_{j-1}))$ è una somma di variabili normali indipendenti con media zero e varianze $\sigma^2 e^{-2\theta(t-t_j)}\Delta t$; pertanto, la varianza dell'Equazione 2.29 è

$$\lim_{\Delta t \to 0} \sum_j \sigma^2 e^{-2\theta(t-t_j)} \Delta t = \int_0^t \sigma^2 e^{-2\theta(t-\tau)} d\tau = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t}). \tag{2.30}$$

Un altro modo comune per esprimere il processo OU è usando la notazione di distribuzione condizionata. Dunn e Gipson (1977) utilizzano questa notazione nel loro lavoro sui processi OU come modelli per il movimento animale. Nel contesto del nostro semplice processo OU unidimensionale, per $t > \tau$, possiamo scrivere

$$\mu(t)|\mu(\tau) \sim \mathcal{N}\left(\mu(\tau)e^{-\theta(t-\tau)}, \frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta(t-\tau)}\right)\right). \tag{2.31}$$

Quindi, al crescere dell'intervallo temporale tra $\mu(t)$ e $\mu(\tau)$, il processo condizionato tende a zero e la varianza converge a σ^2. Tuttavia, per valori piccoli di $|t-\tau|$, $\mu(t)$ sarà vicino a $\mu(\tau)$. Si considerino ad esempio due processi stocastici univariati condizionati in 1-D, simulati a partire dall'Equazione 2.31 con diversi valori del parametro θ. Essi mostrano comportamenti distinti. In particolare, un valore relativamente grande di θ, pari a 1, produce un processo che presenta una debole dipendenza dal valore di riferimento $\mu(\tau) = 1$, indicando una memoria breve. Al contrario, un valore molto piccolo di θ, pari a 0.001, genera un processo che mantiene una forte dipendenza da esso, suggerendo una memoria lunga rispetto alla media condizionata, come si vede in Figura 2.3.

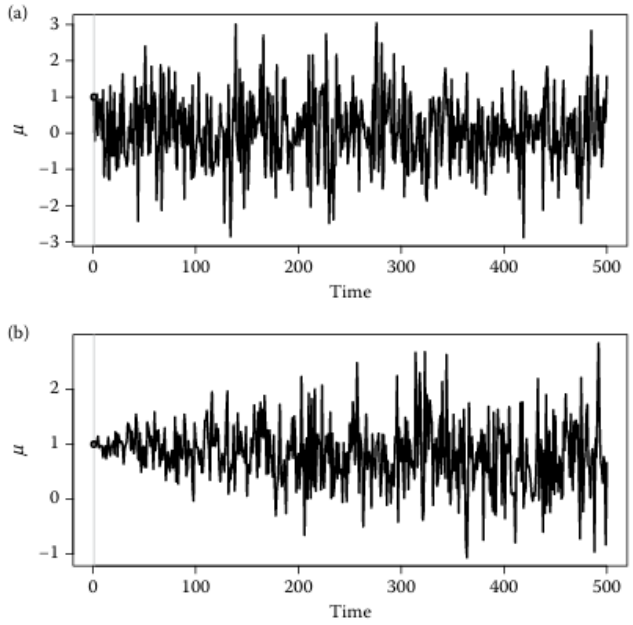


Figura 2.3: Due processi 1-D, simulati a partire dall'Equazione 2.31 con $\sigma^2 = 1$, $\tau = 1$ $\mu(\tau) = 1$. (a) $\theta = 1$ e in (b) $\theta = 0.001$.

Capitolo 3

Modello di Ornstein-Uhlenbeck

Come riportato nel capitolo precedente, Hooten et al. derivano il processo di Ornstein–Uhlenbeck (OU) nella forma scalare unidimensionale come:

$$d\mu(t) = -\theta\mu(t)\,dt + \sigma\,dB(t), \tag{3.1}$$

oppure, in forma più generale:

$$d\mu(t) = (M - I)(\mu(t) - \mu^*)\,dt + \sigma\,db(t).$$

L'equazione OU seguente

$$d\mathbf{X}(t) = \theta\,(\boldsymbol{\mu} - \mathbf{X}(t))\;dt + \sigma\,d\mathbf{W}(t)$$

è una forma equivalente, ma utilizza $\theta(\boldsymbol{\mu}-\mathbf{X}(t))$ invece di $(M-I)(\boldsymbol{\mu}-\mathbf{X}(t))$. Si tratta della stessa struttura, dove θ può essere interpretato come uno scalare o una matrice che regola l'intensità dell'attrazione verso $\boldsymbol{\mu}$. Invece di $db(t)$ impiega $d\mathbf{W}(t)$: entrambi rappresentano incrementi di moto browniano, ma $d\mathbf{W}(t)$ è la notazione più comune nei modelli multidimensionali (spazio d-dimensionale)

dove:

- $\mathbf{X}(t)$ è il vettore posizione al tempo t,
- μ è il valore medio a lungo termine,

- $\theta > 0$ è il tasso di ritorno verso μ,
- $\sigma > 0$ è il coefficiente di diffusione,
- $d\mathbf{W}(t)$ è un moto browniano bidimensionale.

Gli intervalli temporali $\Delta t_i = t_i - t_{i-1}$ sono generati casualmente e la posizione successiva è ricavata usando:

$$\mathbf{X}_i \sim \mathcal{N}\left(\mu + (\mathbf{X}_{i-1} - \mu)e^{-\theta\Delta t}, \sigma^2 \frac{1 - e^{-2\theta\Delta t}}{2\theta}\right). \tag{3.2}$$

3.1 Simulazioni dei modelli di Ornstein-Uhlenbeck: continuo e discreto

Sono state simulate diverse traiettorie, a tempo sia discreto sia continuo, al fine di ottenere i dati per stimare i parametri d'interesse tramite l'inferenza bayesiana.
I dati simulati sono quindi trattati come dati osservati e l'obiettivo dell'inferenza è stimare i parametri originali del processo OU, verificando la capacità del modello di recuperarli a partire dalle osservazioni.
Dopo aver definito un errore, si mira a individuare le combinazioni di parametri migliori tramite la simulazione e la stima con combinazioni di più parametri.
Una volta testata la bontà del metodo inferenziale su dati simulati, esso verrà applicato a dati reali.

Simulazione dei dati

Per entrambi i modelli, le traiettorie sono simulate secondo la dinamica del processo OU bidimensionale.

Dinamica del processo.

$$X(t_i) \mid X(t_{i-1}) \sim \mathcal{N}\left(\mu + (X(t_{i-1}) - \mu)e^{-\theta\Delta t},\ \frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta\Delta t}\right)\right) \tag{3.3}$$

Tempi di osservazione.

- Nel caso **discreto**, i tempi sono fissi e regolari.
- Nel caso **continuo**, i tempi sono ottenuti come somma cumulativa di intervalli campionati da distribuzioni uniformi con media pari al corrispondente Δt discreto, per mantenere un parallelismo tra le condizioni simulate.

Inferenza bayesiana. In entrambi i modelli, per ogni traiettoria vengono stimati i parametri $\mu_1, \mu_2, \theta, \sigma^2$ tramite campionamento MCMC, con 1500 iterazioni totali, di cui 500 di warm-up iterazioni vengono scartate, perché troppo vicine alla distribuzione a posteriori dei parametri, e 1000 effettive. Si calcolano:

- la media a posteriori dei parametri;
- l'errore assoluto rispetto ai valori veri;
- intervalli di credibilità al 95%;
- indicatori booleani di copertura.
- frequenza di copertura.

Per ogni traiettoria simulata, l'inferenza bayesiana restituisce una distribuzione posteriore per ciascun parametro stimato. Da questi risultati vengono calcolate le seguenti metriche:

- **Media a posteriori dei parametri**: per ogni parametro $\theta \in \{\mu_1, \mu_2, \theta, \sigma^2\}$, si calcola la media dei campioni MCMC ottenuti dalla distribuzione posteriore:

$$\hat{\theta} = \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)}$$

dove S è il numero totale di campioni post-burn-in.

- **Errore assoluto**: misura la distanza tra la stima a posteriori e il valore reale utilizzato nella simulazione. È calcolato come:

$$\text{Errore}(\theta) = \left|\hat{\theta} - \theta_{\text{true}}\right|$$

- **Errore totale:** Per ogni configurazione di simulazione, fissati i valori veri dei parametri, l'errore totale è la somma degli errori assoluti medi per ciascun parametro stimato, calcolati su tutte le ripetizioni della simulazione.

$$\text{Errore Totale} = \overline{|\mu_1 - \hat{\mu}_1|} + \overline{|\mu_2 - \hat{\mu}_2|} + \overline{|\theta - \hat{\theta}|} + \overline{|\sigma^2 - \hat{\sigma}^2|}$$

dove il simbolo $\overline{|\cdot|}$ indica la media degli errori assoluti su 10 ripetizioni della stessa configurazione. L'errore totale consente di ordinare le configurazioni in termini di accuratezza complessiva nella stima dei parametri.
Verranno estratte le migliori 100 combinazioni di parametri con minore errore totale.

- **Intervalli di credibilità al 95%**: per ogni parametro stimato, si calcolano i percentili 2.5 e 97.5 della distribuzione a posteriori per ottenere l'intervallo credibile:

$$\text{IC}_{95\%}(\theta) = [\theta_{2.5\%}, \theta_{97.5\%}]$$

dove $\theta_{2.5\%}$ e $\theta_{97.5\%}$ rappresentano i quantili empirici della distribuzione posteriore.

- **Indicatore booleano di copertura:** Si verifica per ciascun parametro stimato se il valore reale del parametro, utilizzato per simulare la traiettoria, è compreso o meno nell'intervallo di credibilità. La copertura viene codificata come una variabile booleana:

$$\text{Copertura}(\theta) = \begin{cases} 1 & \text{se } \theta_{\text{true}} \in [\theta_{2.5\%}, \theta_{97.5\%}] \\ 0 & \text{altrimenti} \end{cases}$$

Questo indicatore viene calcolato per ogni ripetizione e parametro, e successivamente aggregato per ottenere la **frequenza di copertura**, ovvero la proporzione di volte in cui il valore reale è incluso nell'intervallo credibile, utile per valutare l'affidabilità dell'inferenza bayesiana.

Condizioni simulate. Le simulazioni sono ripetute per:

- $\mu \in \{(0.0,\ 0.0)\}$;
- $\theta \in \{0.01,\ 0.03,\ 0.05,\ 0.08,\ 0.1,\ 0.2,\ 0.5\}$;
- $\sigma \in \{0.1,\ 0.3,\ 0.6,\ 0.9,\ 1.2,\ 2.0\} \Rightarrow \sigma^2 \in \{0.01,\ 0.09,\ 0.36,\ 0.81,\ 1.44,\ 4.0\}$;
- $n \in \{50,\ 100,\ 200,\ 400,\ 1000\}$;
- 10 ripetizioni per ciascuna combinazione.
- tempi di osservazione descritti in precedenza:
  - $\Delta t \in \{0.5,\ 1.0,\ 3.0,\ 5.0\}$ per il caso discreto;
  - $(a, b) \in \{(0.1), (0, 2), (2, 4), (4, 6)\}$ per il caso continuo, dove (a, b) sono i parametri di una distribuzione uniforme.

Otterremo quindi alla fine, 8400 simulazioni per ciascun modello.
I valori dei parametri sono stati scelti in modo da individuare combinazioni di essi utili a comprendere l'accuratezza e l'affidabilità del modello in molteplici situazioni.
Al fine di analizzare le prestazioni del modello OU, sia discreto sia continuo, è stato quindi implementato un codice python che esegue un'analisi di

clustering. L'obiettivo è quello di individuare tre gruppi di combinazioni di parametri tramite i quali si ottengono buone, intermedie o pessime.
L'algoritmo di clustering utilizzato è il **K-means**, applicato agli errori assoluti dei parametri stimati.
Le fasi principali dell'analisi includono:

- **Standardizzazione**: gli errori vengono standardizzati per garantire che ciascun parametro contribuisca in modo equilibrato alla distanza euclidea usata nel clustering.
- **Clustering K-means**: viene eseguito con un numero prefissato di cluster ($k = 3$).
- **Riduzione dimensionale**: viene effettuata un'analisi **PCA** (Principal Component Analysis) per visualizzare la distribuzione delle osservazioni nei primi due componenti principali.
- **Visualizzazione**: vengono generati grafici PCA colorati per cluster, boxplot degli errori per ciascun parametro e heatmap dei parametri medi per ogni cluster.

Questa procedura consente di individuare insiemi omogenei di configurazioni in base all'accuratezza delle stime bayesiane e, in particolare, è utile per capire quali condizioni sperimentali siano associate a stime più o meno affidabili.

Modello Stan per il processo OU discreto e continuo

Una volta preparati i dati, si procede con l'implementazione del modello inferenziale, sia nel caso a tempo discreto sia in quello a tempo continuo.
La sostanziale differenza tra le due versioni consiste nell'intervallo temporale, il quale nel caso discreto è costante e specificato a priori, mentre nel caso continuo è irregolare e ottenuto dalla differenza tra tempi osservati successivi. Inoltre, gli intertempi sono ottenuti da una distribuzione uniforme con media

pari al corrispondente Δt discreto.
Il modello assume che, per ogni $n = 2, \ldots, N$, la posizione osservata (x_n, y_n) segua una distribuzione normale bidimensionale, condizionata alla posizione precedente (x_{n-1}, y_{n-1}):

$$x_n \sim \mathcal{N}\left(\mu_1 + (x_{n-1} - \mu_1)e^{-\theta\Delta t},\ \sigma^2 \frac{1 - e^{-2\theta\Delta t}}{2\theta}\right)$$
$$y_n \sim \mathcal{N}\left(\mu_2 + (y_{n-1} - \mu_2)e^{-\theta\Delta t},\ \sigma^2 \frac{1 - e^{-2\theta\Delta t}}{2\theta}\right)$$

Per quanto riguarda le distribuzioni a priori per i parametri del modello, sono state utilizzate le seguenti:

$$\mu_i \sim \mathcal{N}(0, 100^2) \quad \forall i \in (1, 2)$$
$$\theta \sim \text{Uniform}(0.01, 0.75)$$
$$\sigma^2 \sim \text{Inv-Gamma}(1, 1)$$

Per la media a lungo termine μ è stata scelta una distribuzione normale con varianza elevata, che riflette una conoscenza a priori debole e permette alla media a lungo termine del processo di spostamento di essere appresa dai dati senza vincoli troppo restrittivi.

La scelta della distribuzione a priori per il parametro θ è guidata dall'esigenza di garantire una sufficiente dipendenza tra osservazioni consecutive del processo.
Invece per σ^2, che indica la variabilità del movimento, è stata scelta una distribuzione a priori Inv-Gamma, utilizzata per garantire che la varianza sia positiva e per assegnare una maggiore probabilità a valori piccoli o moderati. Inoltre, la distribuzione Inverse-Gamma è caratterizzata da una coda lunga, che permette di esplorare anche valori elevati di σ^2. Inoltre, quando il parametro di forma $\alpha = 1$, la media della distribuzione non è definita, rendendola una prior particolarmente poco informativa.

$$\mathbb{E}[X] = \frac{\beta}{\alpha - 1} \quad \forall \alpha > 1$$

La scelta di questa distribuzione non impone forti vincoli sulla varianza e permette al modello di apprendere liberamente dalle osservazioni.
Tuttavia, poiché l'Inverse-Gamma(1,1) assegna una probabilità elevata a valori molto piccoli di σ^2, potrebbe essere necessario prestare attenzione alla stabilità numerica del modello. In particolare, se i valori campionati risultano troppo piccoli, si potrebbe considerare una distribuzione più informativa come una Inverse-Gamma(2,1) o una Gamma(2,2), che riducono la probabilità di ottenere varianze troppo basse. Oppure, per evitare instabilità numeriche, è anche possibile imporre un vincolo inferiore a σ^2 direttamente nel modello:

$$\sigma^2 > 10^{-6} \tag{3.4}$$

per assicurarsi che la varianza non collassi su valori estremamente piccoli, migliorando la stabilità del campionamento. Quest'ultima precauzione verrà adottata nel modello.

Nel processo di Ornstein-Uhlenbeck discreto, la dinamica è modellata sempre dall'Equazione ricorsiva 3.3 con Δt costante.

Il parametro θ rappresenta la velocità con cui il processo ritorna verso la media μ a lungo termine. Tuttavia, θ influenza anche il grado di dipendenza temporale tra osservazioni successive. Tale dipendenza è formalizzata attraverso la funzione di autocorrelazione:

$$\rho(\Delta t) = \exp(-\theta \cdot \Delta t)$$

Per garantire che le osservazioni successive (cioè a distanza Δt) siano significativamente correlate, si richiede che $\rho(\Delta t)$ non sia né troppo vicino a 1, indicando un processo quasi statico, né troppo vicino a 0, indicando un processo quasi privo di memoria.

Scegliendo un intervallo $\theta \in [0.01, 0.75]$ e considerando $\Delta t = 2.5$, intervallo temporale medio tra quelli considerati, si ottiene:

$$\rho(2.5) \in [\exp(-0.75 \cdot 2.5), \exp(-0.01 \cdot 2.5)] = [0.153, 0.975]$$

Tale range permette di modellare un processo che mantiene una memoria sufficiente tra osservazioni consecutive, riflettendo la natura temporale del fenomeno analizzato. Per questi motivi e per essere allineati al modello continuo dove l'intervallo è variabile, si adotta la seguente distribuzione a priori:

$$\theta \sim \text{Uniform}(0.01, 0.75).$$

Analisi dei cluster nel modello OU discreto

Tabella 3.1: Statistiche riassuntive dei parametri nei tre cluster del modello OU discreto

Cluster	$\theta_{\mathbf{true}}$			$\sigma^2_{\mathbf{true}}$			Δt			$n_{\mathbf{obs}}$		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
0	0.14	0.01	0.50	0.87	0.01	4.00	2.41	0.5	5.0	374.93	50	1000
1	0.03	0.01	0.50	2.31	0.09	4.00	1.37	0.5	5.0	176.48	50	1000
2	0.18	0.01	0.50	3.46	0.81	4.00	2.86	0.5	5.0	163.23	50	1000

In Tabella 3.1, sono riportate le condizioni medie dei parametri per ciascun cluster. Nel **cluster 0** vi sono i casi in cui il modello OU discreto ha fornito stime accurate, mentre negli altri due cluster vi sono condizioni più critiche, con basse forze di attrazione, elevata diffusione e numero ridotto di osservazioni.

- **Cluster 0**: vi sono situazioni associate a stime più precise del modello. In queste simulazioni, il valore medio di θ è intorno a 0.14 e la varianza del processo è contenuta,$\sigma^2 \approx 0.87$. L'intervallo temporale tra le osservazioni è mediamente pari a 2.4, e il numero medio di osservazioni è superiore a 370. Queste condizioni sembrano offrire un buon equilibrio

tra quantità di dati, forza di ritorno verso la media e livello di rumore, permettendo al modello di lavorare con dati informativi. Non a caso, l'errore totale medio in questo gruppo è il più basso tra tutti, circa 1.29.

- **Cluster 1**: in questo cluster vi sono i casi più problematici.θ è molto basso ≈ 0.03, quindi siamo in presenza di processo poco attratto verso la media e la varianza è elevata, $\sigma^2 \approx 2.31$, evidenziando un processo fortemente rumoroso. Il numero di osservazioni è basso, in media 176, e i tempi tra le osservazioni sono più ravvicinati, $\Delta t \approx 1.37$. In queste condizioni il modello presenta un errore totale medio molto alto pari a 16.32 e fornisce stime pessime.

- **Cluster 2**: in questo cluster sono racchiuse le situazioni intermedie. La forza di attrazione verso la media è maggiore rispetto al cluster 1, $\theta \approx 0.18$, la varianza è ancora più elevata, $\sigma^2 \approx 3.46$, e il numero di osservazioni resta contenuto, circa 163 in media. L'errore totale medio è pari a 3.48: migliore rispetto al cluster 1, ma comunque ben più alto rispetto al cluster 0. Questo suggerisce che un buon θ può aiutare, ma senza abbastanza dati o con σ elevata, la stima rimane imprecisa.

Dunque, l'analisi dei cluster mostra che per ottenere buone stime nel modello discreto è indispensabile avere:

- un numero sufficiente di osservazioni;

- una distanza tra osservazioni né troppo breve né troppo ampia;

- un livello di rumore (cioè una σ^2 non troppo alto) che permetta al modello di individuare una traiettoria chiara del processo.

Visualizzazione dei cluster - Modello OU discreto

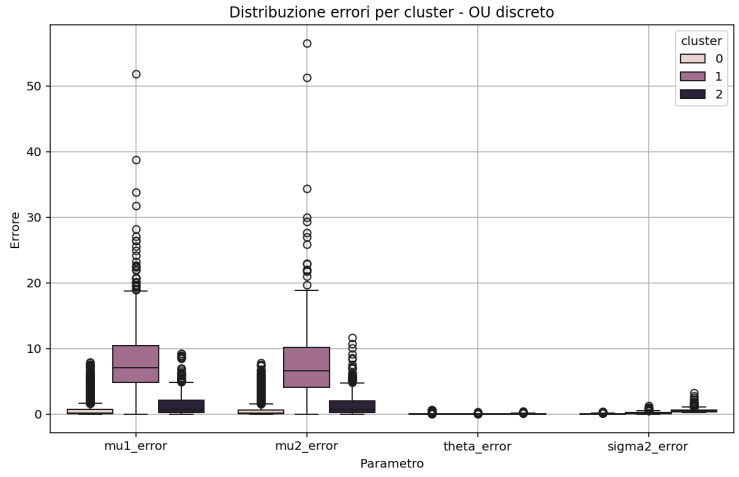


Figura 3.1: Distribuzione degli errori per ciascun parametro stimato, suddivisi per cluster. Il cluster 0 mostra chiaramente gli errori più contenuti su tutti i parametri, mentre i cluster 1 e 2 evidenziano difficoltà nella stima, in particolare su μ_1 e μ_2.

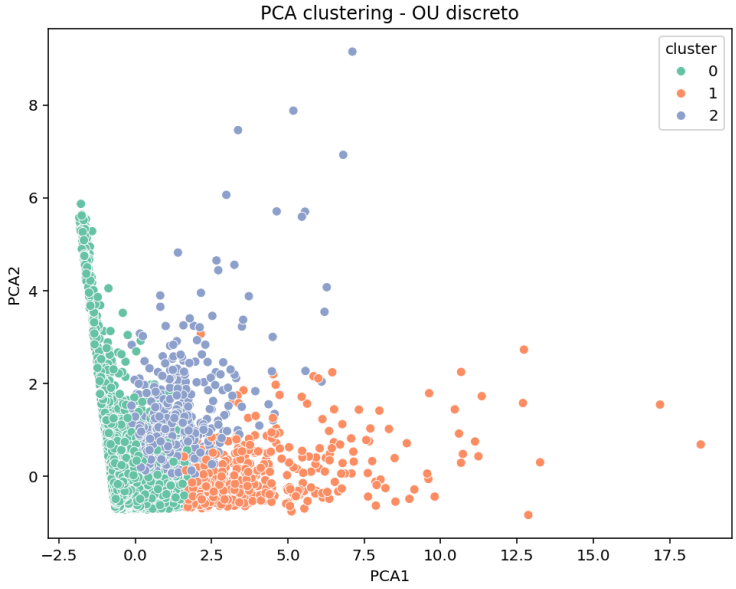


Figura 3.2: Proiezione PCA dei dati di errore nei cluster individuati. I tre cluster sono ben separati nello spazio delle componenti principali, indicando che le variabili di errore sono sufficientemente discriminanti.

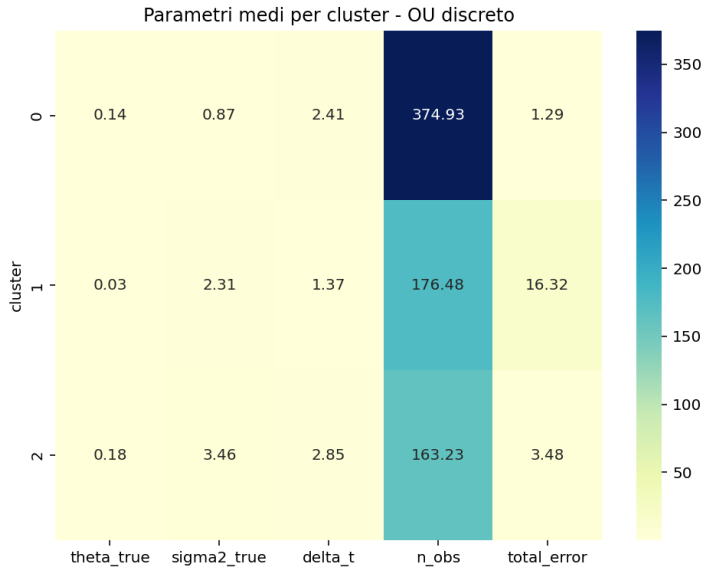


Figura 3.3: Heatmap delle medie dei parametri reali e dell'errore totale nei tre cluster. Il cluster 0 si caratterizza per parametri più favorevoli e errore medio inferiore rispetto agli altri.

Analisi dei risultati del modello OU discreto

L'analisi dei risultati ottenuti tramite simulazioni del modello di Ornstein-Uhlenbeck (OU) è stata svolta considerando diverse configurazioni sperimentali. Ogni combinazione di parametri è stata ripetuta 10 volte per garantire la stabilità statistica delle stime.

Errore assoluto di stima. Per ogni parametro stimato $(\mu_1, \mu_2, \theta, \sigma^2)$ è stato calcolato l'errore assoluto tra la media a posteriori ottenuta tramite MCMC e il valore reale utilizzato per la simulazione. I boxplot (Figura 3.4) mostrano la distribuzione di questi errori per ciascun parametro. Si osserva che θ e σ^2 presentano una distribuzione degli errori mediamente contenuta, mentre μ_1 e μ_2 mostrano una maggiore variabilità dovuta alla presenza di numerosi outlier, pur mantenendo un errore mediano basso.

Heatmap degli errori. Le heatmap relative all'errore medio di stima sono state costruite per μ_1, μ_2, θ e σ^2 in funzione del numero di osservazioni (n) e del passo temporale (Δt). Dai grafici emerge che:

- all'aumentare del numero di osservazioni l'errore tende a ridursi;
- per θ e μ l'errore diminuisce anche con Δt più ampi;
- al contrario, per σ^2 l'errore può aumentare con Δt, indicando una minore informazione sul rumore se le osservazioni sono troppo distanti.

Frequenza di copertura. La frequenza di copertura rappresenta la proporzione di volte in cui il valore reale di ciascun parametro è stato incluso nell'intervallo di credibilità al 95% calcolato sulla distribuzione a posteriori. È stata calcolata:

- su tutte le simulazioni globali (*copertura media totale*);
- limitatamente alle top 100 combinazioni con errore totale minimo.

I valori medi di copertura osservati sono i seguenti:

Parametro	**Globale**	**Top 100**
μ_1	0.92	0.93
μ_2	0.92	0.93
θ	0.73	0.75
σ^2	0.77	0.52

Si osserva che la copertura per i parametri μ_1, μ_2 è ottima, anche quella di θ è soddisfacente, mentre quella di σ^2 è piuttosto variabile e più problematica tra le combinazioni ottimali. Gli intervalli di credibilità posteriori non riescono spesso a includere il valore reale di σ^2.
Inoltre, la frequenza di copertura media di σ^2 nella top 100 è inferiore a quella globale, quindi non è detto che a un errore assoluto basso corrisponda una buona frequenza di copertura.
Infatti, in corrispondenza di un errore basso, la stima è molto vicina al valore

Tabella 3.2: Top 10 combinazioni con errore totale minimo e frequenze di copertura (OU discreto)

θ	σ^2	Δt	n	Errore Totale	μ_1	μ_2	θ	σ^2
0.1000	0.0100	5.0000	1000	0.0366	0.9000	1.0000	0.3000	0.0000
0.0800	0.0100	3.0000	1000	0.0395	0.9000	1.0000	1.0000	0.1000
0.1000	0.0100	3.0000	1000	0.0407	1.0000	1.0000	0.5000	0.0000
0.0800	0.0100	5.0000	1000	0.0472	0.8000	0.9000	0.4000	0.0000
0.0500	0.0100	5.0000	1000	0.0488	0.9000	1.0000	0.7000	0.0000
0.2000	0.0900	5.0000	1000	0.0499	1.0000	0.9000	1.0000	0.8000
0.2000	0.0100	3.0000	1000	0.0526	0.9000	0.9000	0.3000	0.0000
0.2000	0.0900	3.0000	1000	0.0528	0.9000	1.0000	0.9000	0.8000
0.5000	0.0900	0.5000	1000	0.0544	1.0000	1.0000	1.0000	1.0000
0.2000	0.0100	5.0000	1000	0.0601	0.9000	1.0000	0.1000	0.0000

reale, l'intervallo di credibilità è molto stretto, e quindi è facile che il parametro reale non sia compreso in esso. In Tabella 3.2 sono riportate le prime dieci combinazioni ottimali.

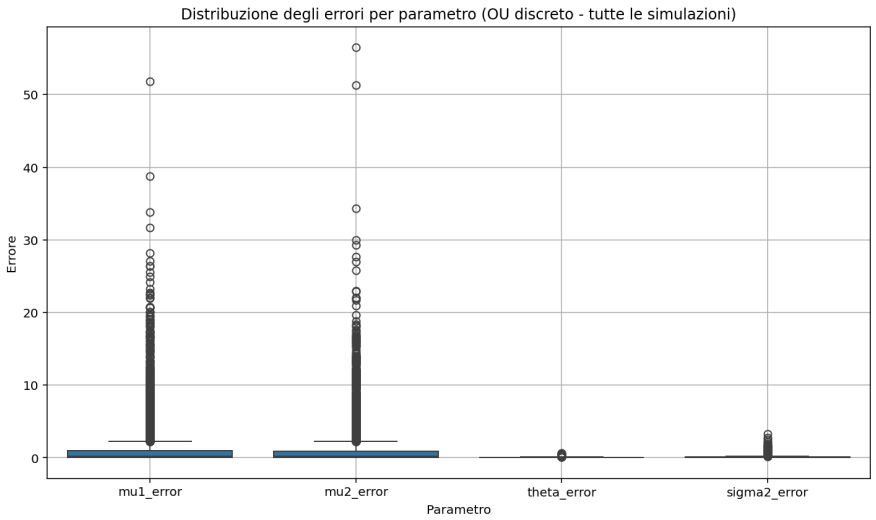


Figura 3.4: Distribuzione degli errori assoluti per ciascun parametro stimato nel modello OU discreto.

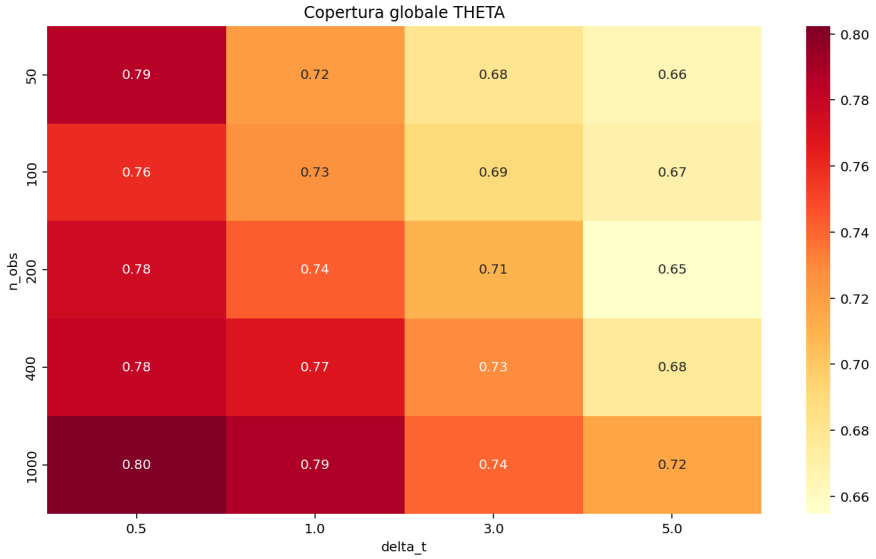


Figura 3.5: Heatmap della copertura globale per il parametro θ, in funzione di Δt e n.

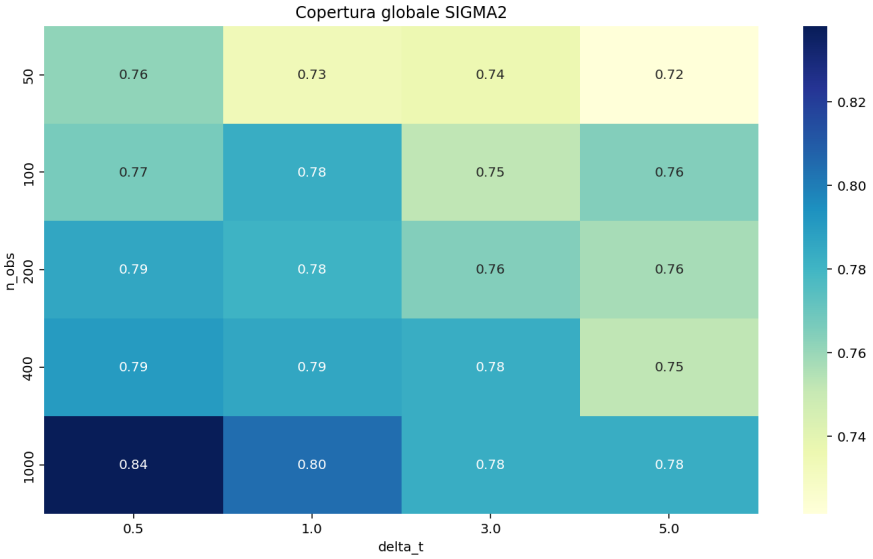


Figura 3.6: Heatmap della copertura globale per il parametro σ^2, in funzione di Δt e n.

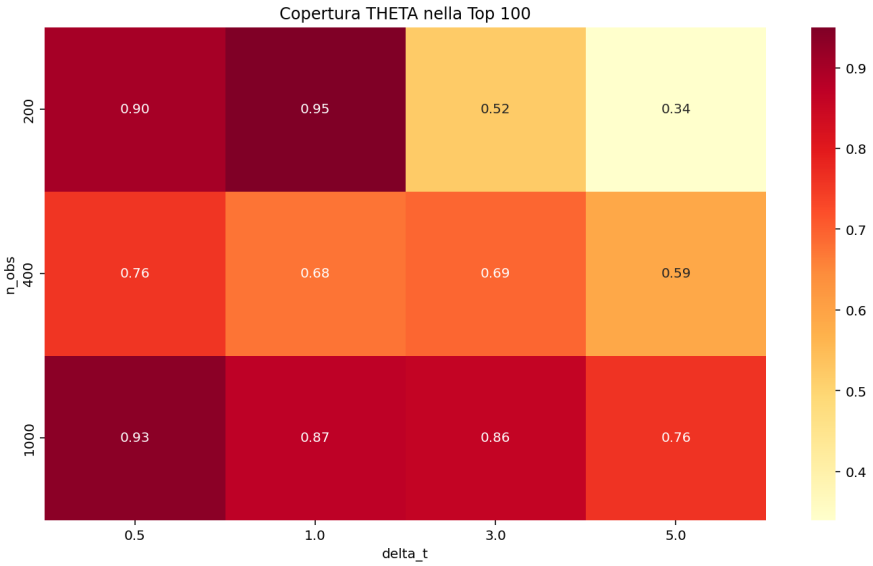


Figura 3.7: Heatmap della copertura per θ nelle 100 combinazioni con errore totale più basso.

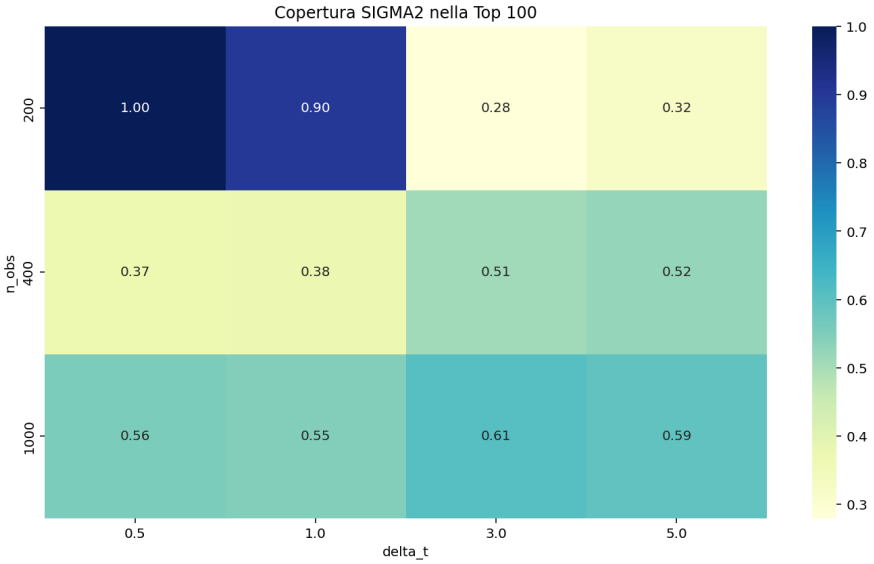


Figura 3.8: Heatmap della copertura per σ^2 nelle 100 combinazioni con errore totale più basso.

Analisi dei cluster nel modello OU continuo

Tabella 3.3: Statistiche riassuntive dei parametri nei tre cluster del modello OU continuo

Cluster	$\theta_{\mathbf{true}}$			$\sigma^2_{\mathbf{true}}$			a			b			$n_{\mathbf{obs}}$		
	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
0	0.14	0.01	0.50	0.91	0.01	4.00	1.50	0.0	4.0	3.30	0.0	4.0	370.41	50	1000
1	0.03	0.01	0.20	2.45	0.09	4.00	0.60	0.0	4.0	2.10	0.0	4.0	182.21	50	1000
2	0.18	0.01	0.50	3.68	0.81	4.00	1.90	0.0	4.0	3.70	0.0	4.0	137.83	50	1000

In Tabella 3.3, sono riportate le condizioni medie dei parametri per ciascun cluster nel caso continuo. Come per il modello discreto, anche qui il **cluster 0** corrisponde alle configurazioni in cui il modello riesce a stimare i parametri in modo accurato, mentre i cluster 1 e 2 rappresentano situazioni più complesse.

- **Cluster 0**: è composto dalle simulazioni con stime più affidabili. I valori medi dei parametri sono: $\theta \approx 0.14$, $\sigma^2 \approx 0.91$, intervallo medio uniforme tra $a = 1.5$ e $b = 3.3$, e un numero medio di osservazioni pari a circa 370. Anche l'errore totale medio è contenuto, pari a 1.34, segno che queste condizioni forniscono informazioni sufficienti per una stima efficace.

- **Cluster 1**: questo cluster raccoglie i casi più problematici per la stima. Il parametro θ è molto basso,≈ 0.03, indicando un processo poco attratto alla media. La varianza è alta, pari a $\sigma^2 \approx 2.45$, e le osservazioni sono meno numerose, 182 in media. L'intervallo temporale $[a, b]$ è concentrato su valori bassi compresi tra 0.6 e 2.1, suggerendo osservazioni ravvicinate. L'errore totale medio è molto elevato, 16.49, ed evidenzia una scarsa qualità delle stime.

- **Cluster 2**: rappresenta una situazione intermedia. La forza di attrazione è più alta, $\theta \approx 0.18$, rispetto al cluster 1, ma il livello di rumore è massimo, $\sigma^2 \approx 3.68$. Anche qui il numero medio di osservazioni, 138, è limitato, e la finestra temporale media $[1.9, 3.7]$ è leggermente più

ampia. L'errore totale medio è pari a 3.87, migliore rispetto al cluster 1 ma comunque peggiore rispetto al cluster 0.

Nel complesso, anche nel modello continuo si confermano i fattori critici per una buona inferenza, ovvero numerose osservazioni, n_{obs} alto, una distribuzione temporale sufficientemente ampia, una diffusione del processo contenuta, σ^2 non troppo elevata e una dinamica sufficientemente attrattiva θ maggiore di 0.1.

Analisi grafica dei cluster nel modello OU continuo

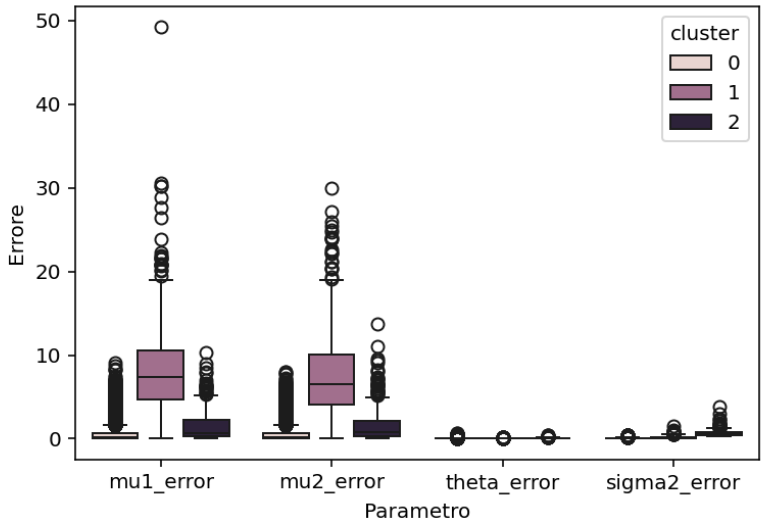


Figura 3.9: Distribuzione degli errori per ciascun parametro nel modello continuo. Il cluster 0 contiene le configurazioni con le stime più accurate, mentre μ_1 e μ_2 mostrano una maggiore variabilità.

Nel modello continuo, a differenza del caso discreto, non è possibile costruire una tabella aggregata dei parametri veri per ciascun cluster, poiché le simulazioni sono generate a partire da intervalli temporali irregolari definiti dalla coppia (a, b), e non da un Δt fisso.

Per questa ragione, si procede con un'analisi attraverso una serie di heatmap, che mostrano l'errore medio per ciascun parametro e l'errore totale in funzione della coppia (a, b).

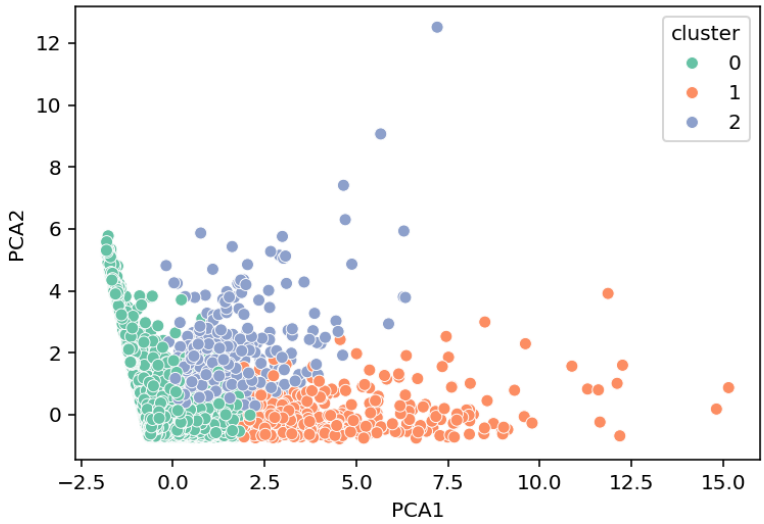


Figura 3.10: PCA dei dati di errore nel modello continuo. I cluster risultano più compatti e meno separati rispetto al modello discreto, ma mantengono comunque una distinzione coerente con le caratteristiche delle stime.

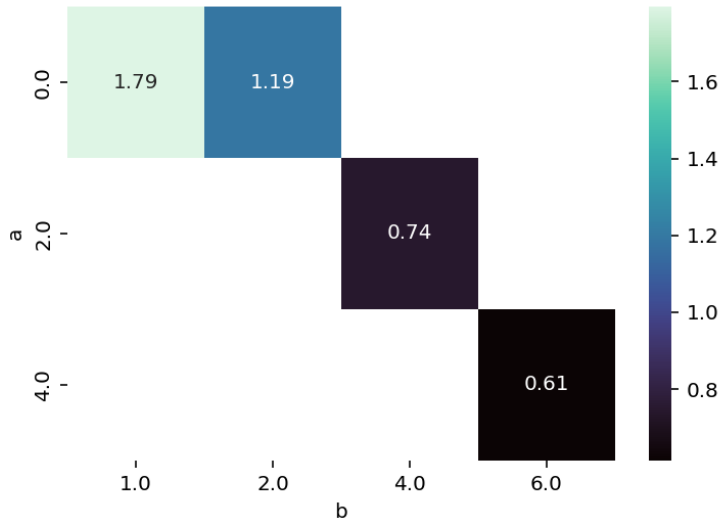


Figura 3.11: Heatmap dell'errore medio su μ_1 in funzione dei parametri (a, b). Si osserva che l'errore tende a diminuire con intervalli temporali più ampi (valori maggiori di a e b).

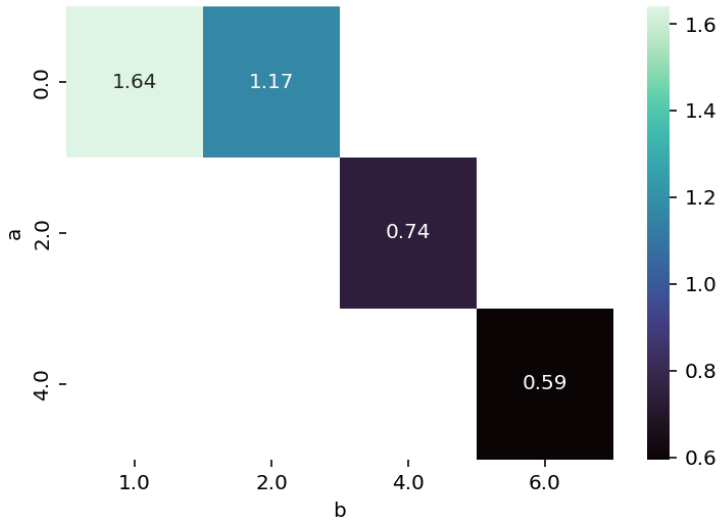


Figura 3.12: Heatmap dell'errore medio su μ_2. Analogamente a μ_1, l'errore si riduce con l'aumentare dell'intervallo di osservazione.

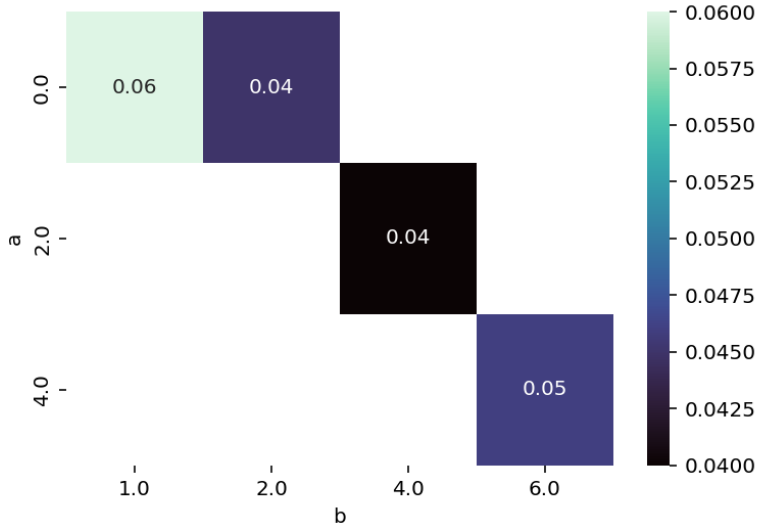


Figura 3.13: Errore di stima medio su θ rispetto a (a, b). L'errore si mantiene contenuto e tende a ridursi all'aumentare dell'ampiezza temporale.

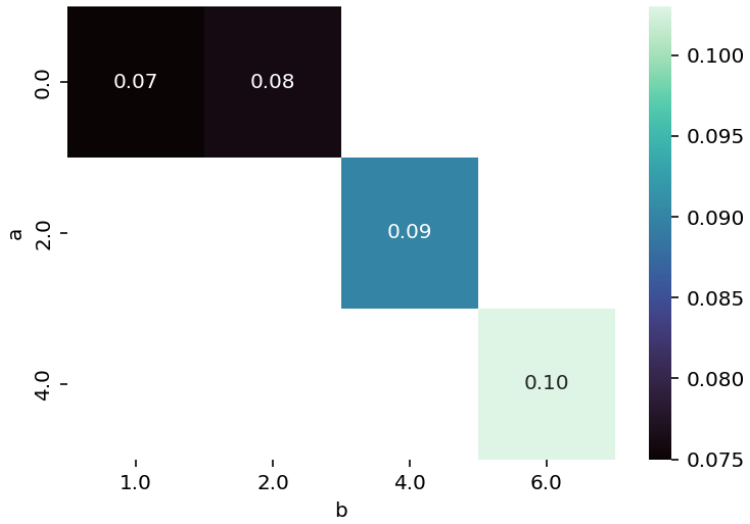


Figura 3.14: Heatmap dell'errore medio su σ^2. La stima della varianza del processo migliora sensibilmente con (a, b) elevati, cioè quando le traiettorie hanno maggiore variabilità temporale.

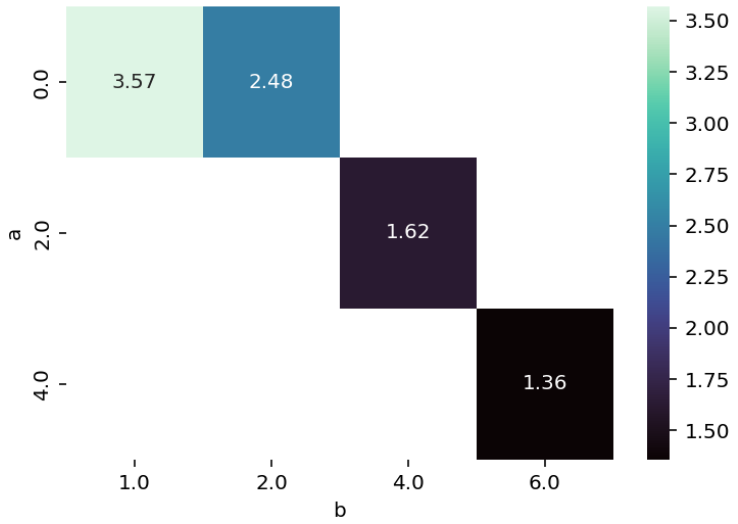


Figura 3.15: Errore totale medio in funzione di (a, b). Le combinazioni con $a = 4$, $b = 6$ risultano ottimali, mentre valori bassi di entrambi portano a stime meno accurate.

Analisi dei risultati del modello OU continuo

L'analisi dei risultati ottenuti dal modello di Ornstein-Uhlenbeck (OU) continuo con tempi di osservazione irregolari è stata condotta allo stesso modo del caso discreto, fatta eccezione per gli intervalli temporali, i quali preservano una media pari ai tempi del modello discreto.

Errore assoluto di stima. Come nel caso discreto, per ciascun parametro stimato è stato calcolato l'errore assoluto tra la media a posteriori e il valore reale usato per la simulazione. I boxplot mostrano che:

- gli errori relativi a μ_1 e μ_2 sono in generale contenuti e con bassa variabilità;
- la stima di θ presenta maggiore variabilità, specialmente per piccoli numeri di osservazioni;

- la stima di σ^2 è la più incerta.

Heatmap degli errori. Sono state realizzate heatmap dell'errore medio assoluto per ciascun parametro, in funzione del numero di osservazioni (n) e dell'intervallo temporale $[t_{\min}, t_{\max}]$. L'analisi visiva suggerisce che:

- l'aumento di n riduce sistematicamente l'errore per tutti i parametri;
- una finestra temporale più ampia (cioè un $t_{\max} - t_{\min}$ elevato) migliora la stima di μ e θ;
- per σ^2, una finestra più ampia non sempre aiuta: l'errore può peggiorare con intervalli temporali troppo ampi.

Frequenza di copertura. La copertura a posteriori al 95% è stata calcolata per ciascun parametro, considerando:

- tutte le simulazioni (*copertura globale*);
- le migliori 100 configurazioni (*top 100*) in base all'errore totale.

I risultati medi sono i seguenti:

Tabella 3.4: Confronto tra copertura globale e top 100 per ciascun parametro (OU continuo)

Parametro	**Globale**	**Top 100**
μ_1	0.92	0.94
μ_2	0.93	0.95
θ	0.73	0.72
σ^2	0.77	0.50

Come per il modello discreto, anche nel continuo si osserva che le stime di μ_1, μ_2 e θ sono ben calibrate. Tuttavia, la copertura per σ^2 risulta più

Tabella 3.5: Top 10 combinazioni con errore totale minimo e frequenze di copertura (OU continuo)

μ_1	μ_2	θ	σ^2	$t_{\min}$	$t_{\max}$	n	μ_1	μ_2	θ	σ^2
0.0000	0.0000	0.1000	0.0100	2.0000	4.0000	1000	1.0000	1.0000	0.7000	0.0000
0.0000	0.0000	0.1000	0.0100	4.0000	6.0000	1000	1.0000	0.9000	0.5000	0.0000
0.0000	0.0000	0.0800	0.0100	2.0000	4.0000	1000	0.9000	0.9000	0.7000	0.1000
0.0000	0.0000	0.2000	0.0900	2.0000	4.0000	1000	0.9000	1.0000	1.0000	1.0000
0.0000	0.0000	0.0500	0.0100	4.0000	6.0000	1000	1.0000	1.0000	0.6000	0.0000
0.0000	0.0000	0.1000	0.0100	0.0000	2.0000	1000	0.9000	1.0000	0.7000	0.0000
0.0000	0.0000	0.5000	0.0900	0.0000	2.0000	1000	1.0000	0.9000	1.0000	1.0000
0.0000	0.0000	0.0800	0.0100	4.0000	6.0000	1000	0.8000	0.9000	0.3000	0.0000
0.0000	0.0000	0.5000	0.0900	2.0000	4.0000	1000	0.9000	1.0000	0.9000	1.0000
0.0000	0.0000	0.2000	0.0100	2.0000	4.0000	1000	1.0000	1.0000	0.3000	0.1000

variabile, indicando che anche nel caso continuo la stima della componente stocastica è la più incerta.

La Tabella riporta le 10 migliori combinazioni di parametri in termini di errore totale medio, calcolato come la somma degli errori medi assoluti sui quattro parametri del modello. Le configurazioni migliori corrispondono generalmente a situazioni con un numero elevato di osservazioni ($n = 1000$) e intervalli temporali ampi, indicando che queste condizioni favoriscono la stima accurata dei parametri.

Le frequenze di copertura degli intervalli credibili al 95% per ciascun parametro, riportate in tabella, mostrano che μ_1 e μ_2 tendono a essere stimati con buona affidabilità. Presentano coperture superiori o pari a 0.9 nella maggior parte dei casi. Tuttavia, la copertura per σ^2 presenta una maggiore variabilità, con valori che talvolta scendono sotto la soglia desiderata.

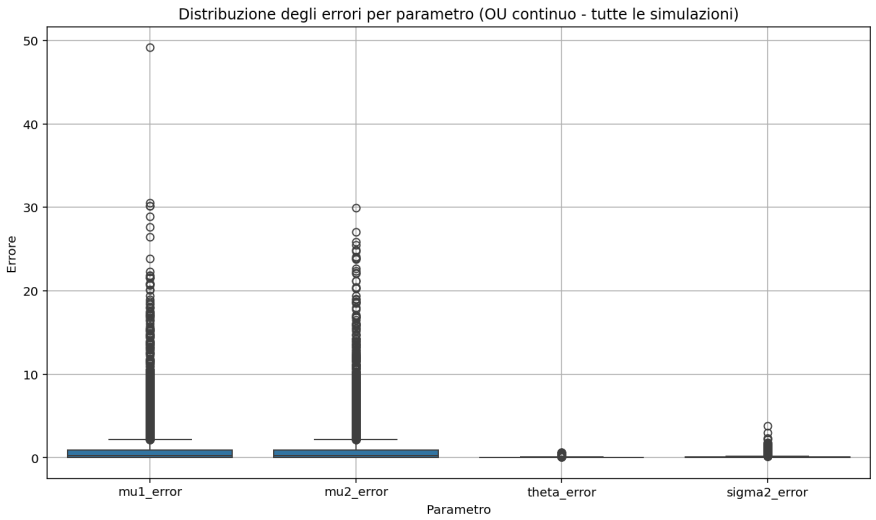


Figura 3.16: Distribuzione degli errori assoluti per ciascun parametro nel modello OU continuo.

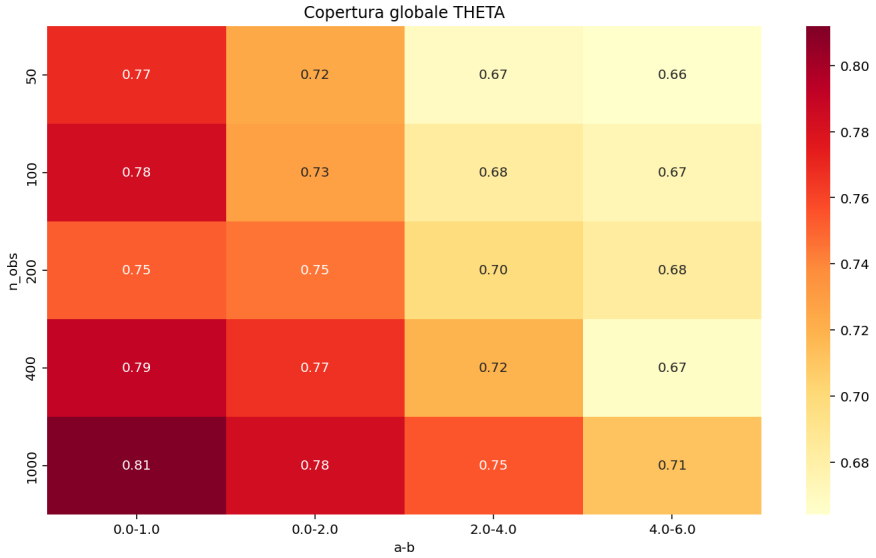


Figura 3.17: Heatmap dell'errore medio assoluto per μ_1 in funzione di n e intervallo $[t_{\min}, t_{\max}]$.

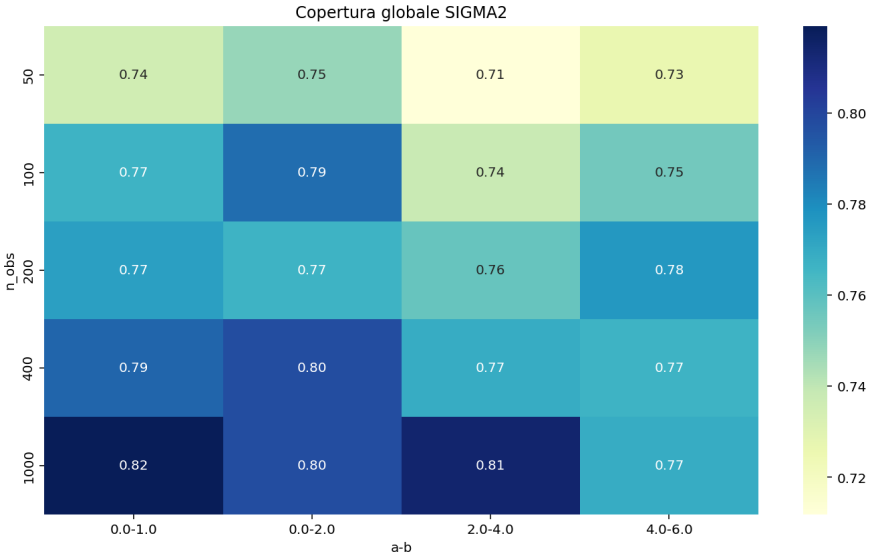


Figura 3.18: Heatmap dell'errore medio assoluto per μ_2 in funzione di n e intervallo $[t_{\min}, t_{\max}]$.

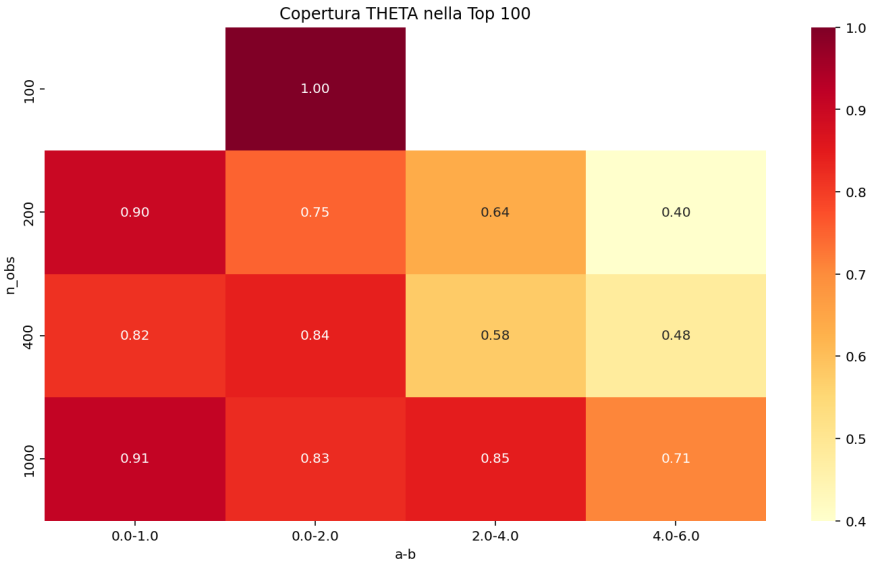


Figura 3.19: Heatmap dell'errore medio assoluto per θ in funzione di n e intervallo $[t_{\min}, t_{\max}]$.

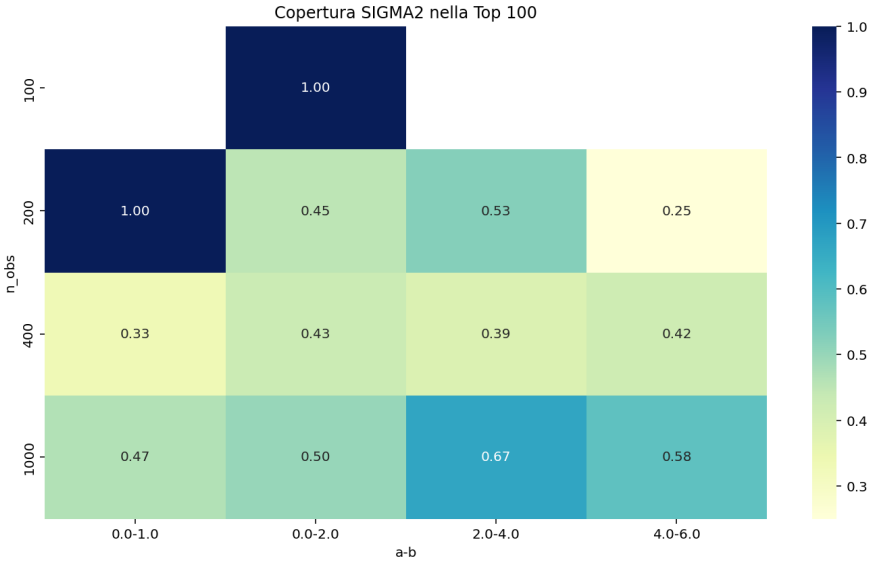


Figura 3.20: Heatmap dell'errore medio assoluto per σ^2 in funzione di n e intervallo $[t_{\min}, t_{\max}]$.

3.2 OU discreto e continuo con parametri significativi

In seguito alle simulazioni con le diverse combinazioni, sono stati scelti dei parametri per affrontare un'analisi più dettagliata del modello e per confrontare i modelli a tempo discreto e continuo.
La combinazione è tra le prime cento migliori, non è quella con errore assoluto minore, ma è quella che ritengo possa essere più simile alla realtà. La simulazione delle posizioni del processo è stata ottenuta imponendo i seguenti parametri:

- $\theta = 0.03$;
- $\mu = [0.0, 0.0]$;
- $\sigma = 0.9$.

I valori sono stati scelti in modo da rappresentare un processo con un ritorno alla media relativamente lento: il parametro θ basso implica una tendenza graduale a ritornare verso il centro del dominio.

La diffusione, invece, è contenuta, limitando l'ampiezza delle fluttuazioni casuali. Si ottiene quindi una traiettoria che si sviluppa lentamente attorno alla media, con movimenti più regolari, correlati e moderatamente vincolati nello spazio.
Il processo tende a rimanere abbastanza confinato intorno alla media: la bassa diffusione limita la dispersione e l'ampiezza dei movimenti casuali, mentre il ritorno alla media, seppur lento, contribuisce a mantenere una certa coerenza spaziale. I punti risultano quindi discretamente correlati.

Sono state generate $n = 1000$ osservazioni.

Per quanto riguarda gli intervalli temporali, per analizzare il modello a tempo discreto, essi sono stati considerati a intervalli costanti di 3 unità.
Nel caso continuo gli intervalli temporali irregolari, definiti da:

$$\Delta t = t_i - t_{i-1}, \quad t_i \sim \text{Uniform}(2, 4).$$

e avvengono con una media di 3 unità temporali.
L'uso di intervalli di tempo casuali preserva la natura del processo OU continuo, anche se i dati sono osservati a istanti discreti.

La dinamica del processo è descritta come anticipato da:

$$\mathbf{x}_i \sim \mathcal{N}\left(\mu + (\mathbf{x}_{i-1} - \mu)e^{-\theta\Delta t}, \sigma^2 \frac{1 - e^{-2\theta\Delta t}}{2\theta}\right).$$

Le posizioni simulate (x, y) sono tracciate in un piano bidimensionale, come si vede in Figura 3.21.

Sia μ il valore medio e σ la deviazione standard. Il processo viene generato come segue:

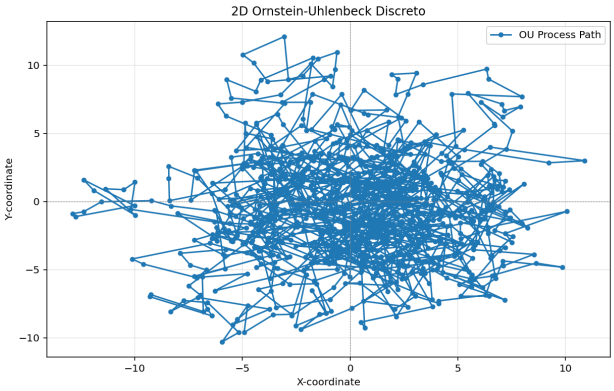


Figura 3.21: Percorso simulato del processo OU in 2D a intervalli regolari, con $\theta = 0.03$, $\mu = [0.0, 0.0]$, $\sigma = 0.9$.

$$
\begin{aligned}
\boldsymbol{X}_0 &= \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I) \quad \text{(Posizione iniziale)} \\
&\text{Per } i = 1, \ldots, n_{\text{obs}} - 1 : \\
\Delta t &= t_i - t_{i-1} \\
\boldsymbol{\mu}_i &= \boldsymbol{\mu} + (\boldsymbol{X}_{i-1} - \boldsymbol{\mu}) e^{-\theta \Delta t} \\
\sigma_i^2 &= \frac{\sigma^2 (1 - e^{-2\theta \Delta t})}{2\theta} \\
\boldsymbol{X}_i &\sim \mathcal{N}(\boldsymbol{\mu}_i, \sigma_i^2 I)
\end{aligned}
$$

dove I è la matrice identità 2×2.

Autocorrelazione e memoria del processo:
Il parametro θ controlla il tasso di ritorno alla media. Valori alti causano una perdita rapida di memoria, mentre un valore basso, come nel nostro caso, implica una dipendenza più duratura tra osservazioni.

La funzione di autocorrelazione del processo è:

$$\rho(\Delta t) = e^{-\theta \Delta t}$$

Con $\theta = 0.03$ e $\mathbb{E}[\Delta t] = 3$, otteniamo:

$$\mathbb{E}[\rho(\Delta t)] = e^{-0.03\times 3} = e^{-0.09} \approx 0.914.$$

Questo garantisce una forte dipendenza tra le osservazioni, mantenendo coerenza spaziale nel tempo.

In sintesi, le scelte effettuate su θ, σ e sulla distribuzione degli intervalli temporali consentono di simulare un processo OU 2D che rispecchia più fedelmente un comportamento animale moderatamente vincolato, con movimenti regolari e dipendenti nel tempo.

Risultati del Campionamento OU discreto

Il modello è stato eseguito con 2000 iterazioni di campionamento e 500 iterazioni di warm-up su una singola catena. I risultati principali includono:

Tabella 3.6: Riassunto dei parametri stimati – Modello OU Discreto 2D

Parametro	Media	MCSE	Dev.Std	N_Eff	N_Eff/s	$\hat{R}$
lp__	-1786.06	0.0527	1.5548	870.64	225.03	1.0035
μ_1	0.1576	0.0150	0.6463	1861.05	481.02	1.0020
μ_2	-0.5584	0.0147	0.6395	1897.45	490.42	0.9995
θ	0.0261	0.0001	0.0032	1666.82	430.81	1.0012
σ	0.8877	0.0004	0.0150	1718.08	444.06	1.0000
σ^2	0.7882	0.0006	0.0267	1711.33	442.32	1.0000

N_Eff: è l'efficienza dell'estrazione, che indica quanta parte delle informazioni è stata effettivamente utilizzata dai campioni. Più alto è il valore, più i campioni sono indipendenti tra loro.

R_hat: è un parametro diagnostico che misura la convergenza del campionamento. Valori vicini a 1 indicano una buona convergenza.
Gli intervalli di credibilità al 95% per i parametri stimati sono i seguenti:

Tabella 3.7: Intervalli di Credibilità al 95% – OU Discreto 2D

Parametro	IC 95%
θ	(0.0197, 0.0321)
μ_1	(-1.1415, 1.4442)
μ_2	(-1.8275, 0.7156)
σ^2	(0.7377, 0.8424)

Questi valori rappresentano gli intervalli in cui ci si aspetta che il vero valore del parametro cada con il 95% di probabilità.

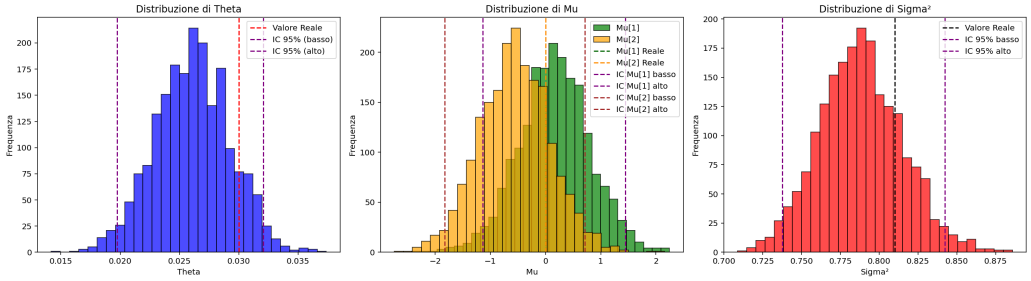


Figura 3.22: Distribuzioni a posteriori per θ, μ e σ^2.

Possiamo notare che le stime sono buone, in quanto gli intervalli di credibilità comprendono tutti i valori reali.

In Figura 3.22 sono riportate le distribuzioni a posteriori con i loro intervalli di credibilità e i valori reali, che rispecchiano quanto si è appena detto analizzando gli intervalli di credibilità.
Il modello è in grado di stimare i parametri con buona precisione.

In Figura 3.23 invece con la funzione arviz.plot_trace() è stato generato automaticamente un tracciato della catena Markov Chain Monte Carlo (MCMC).
A sinistra vi sono le distribuzioni a posteriori per ogni parametro, mentre a destra vi sono i valori campionati per ogni iterazione della catena (time series plot).
Questa funzione è utile per diagnosticare la convergenza del campionamento, in quanto nel caso in cui le catene siano ben miscelate (senza trend evidenti), il modello è convergente; diversamente, potrebbero esserci problemi di campionamento.

I traceplot non mostrano segni di autocorrelazione eccessiva o problemi di convergenza, confermando che l'MCMC ha esplorato bene lo spazio dei parametri.

Tuttavia è riduttivo riportare i risultati di una sola simulazione, quindi, avendo già affrontato in precedenza la simulazione multiparametrica con 10

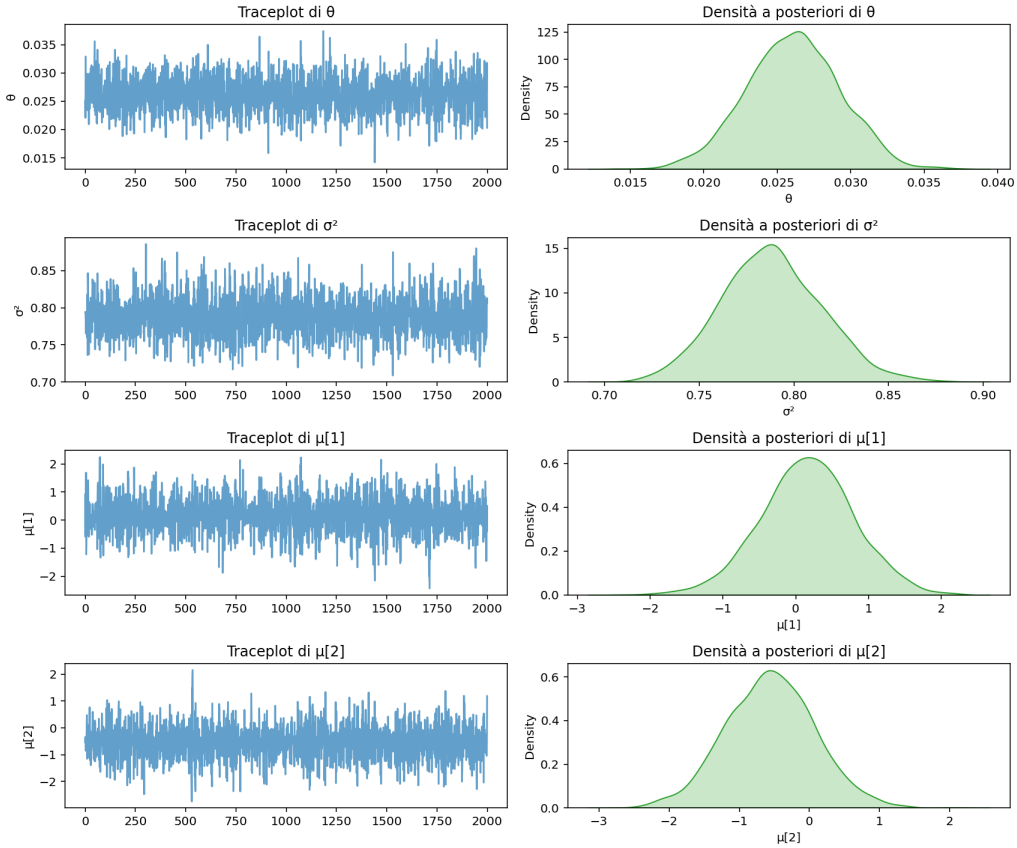


Figura 3.23: Densità a posteriori e tracce campionate

ripetizioni, riportiamo di seguito il dettaglio dei risultati ottenuti con questa combinazione di parametri:

Parametro	Errore Medio	Dev. Std	Copertura 95%
μ_1	0.571	0.451	80%
μ_2	0.447	0.249	100%
θ	0.003	0.003	90%
σ^2	0.020	0.017	100%
Errore Totale	1.041	0.423	

Tabella 3.8: Statistiche di stima per $\theta = 0{,}03$, $\sigma = 0{,}9$, $\mu = (0, 0)$, $\Delta t = 3$, $n = 1000$, su 10 simulazioni.

Risultati del Campionamento OU Continuo

Il modello è stato eseguito con 2000 iterazioni di campionamento e 500 iterazioni di warm-up su una singola catena. I risultati principali includono:

Tabella 3.9: Riassunto dei parametri stimati – Modello OU Continuo 2D

Parametro	Media	MCSE	Dev.Std	N_Eff	N_Eff/s	$\hat{R}$
lp__	-1746.02	0.0478	1.4139	875.06	240.20	0.9997
θ	0.0288	0.0001	0.0033	1636.15	449.12	0.9999
μ_1	0.4636	0.0109	0.5270	2328.08	639.06	0.9995
μ_2	-0.1139	0.0122	0.5706	2174.98	597.03	0.9996
σ	0.8807	0.0003	0.0143	1855.69	509.38	0.9998
σ^2	0.7759	0.0006	0.0251	1854.40	509.03	0.9998

N_{Eff}: è l'efficienza dell'estrazione, che indica quanta parte delle informazioni è stata effettivamente utilizzata dai campioni. Più alto è il valore, più i campioni sono indipendenti tra loro.
R_hat: è un parametro diagnostico che misura la convergenza del campionamento. Valori vicini a 1 indicano una buona convergenza.
Gli intervalli di credibilità al 95% per i parametri stimati sono in Tabella 3.10

e indicano l'intervallo in cui ci si aspetta che il vero valore del parametro cada con il 95% di probabilità.

Tabella 3.10: Intervalli di Credibilità al 95% – OU Continuo 2D

Parametro	IC 95%
θ	(0.0225, 0.0352)
μ_1	(-0.5908, 1.4778)
μ_2	(-1.2054, 1.0344)
σ^2	(0.7293, 0.8273)

Nella Figura 3.26 sono riportate le distribuzioni a posteriori per θ, μ e σ, con il relativo valore reale e l'intervallo di credibilità, e notiamo che il valore reale è compreso per tutti i parametri nell'intervallo di credibilità.

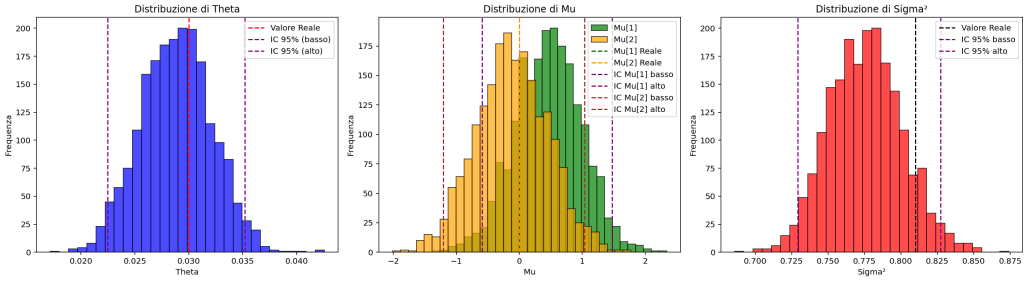


Figura 3.24: Distribuzioni posteriori per θ, μ e σ^2.

In Figura 3.25 sono tracciati nuovamente i traceplot, che anche in questo caso non mostrano segni di autocorrelazione eccessiva o problemi di convergenza, confermando che l'MCMC ha esplorato bene lo spazio dei parametri. Come per il caso discreto, riportiamo i risultati ottenuti dalla simulazione parametrica eseguita 10 volte:

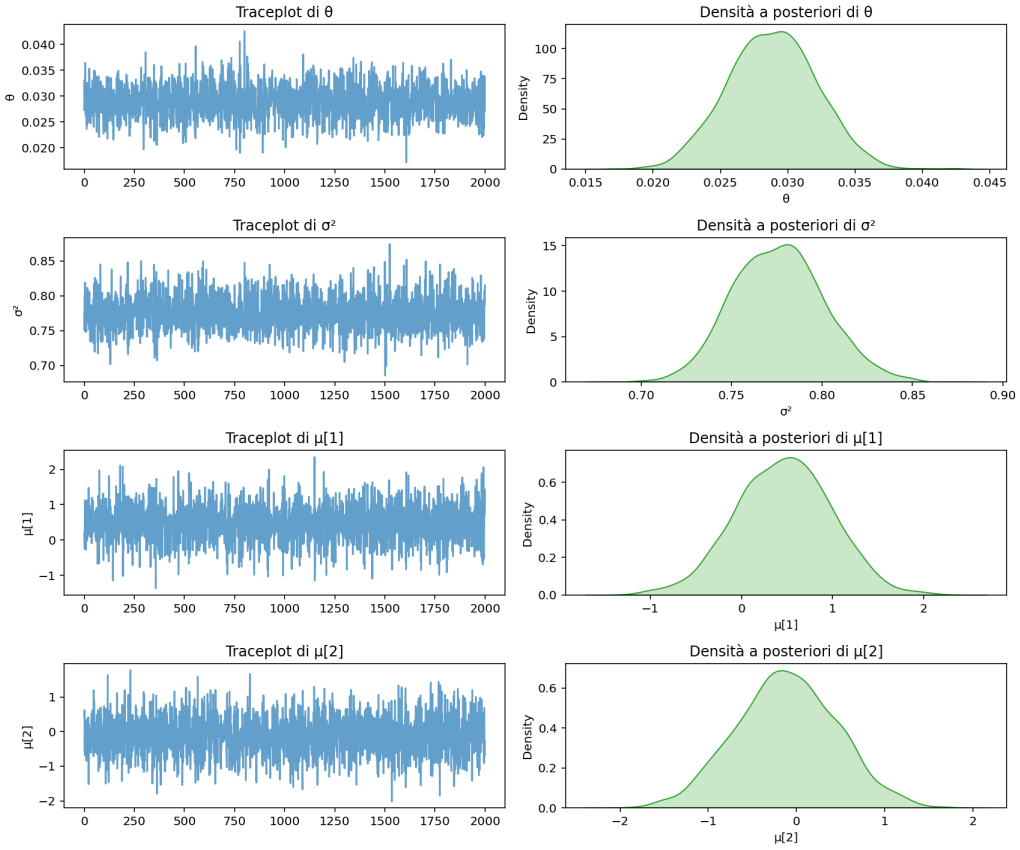


Figura 3.25: Densità a posteriori e tracce campionate

3.3 Simulazione con tempi aleatori e correzione del modello discreto

Nello studio del modello di Ornstein-Uhlenbeck (OU) a tempo discreto, come più volte anticipato, una delle ipotesi fondamentali riguarda il fatto che le osservazioni avvengano ad intervalli di tempo regolari. Tuttavia, questa assunzione non è sempre realistica, in quanto anche i tempi di osservazione sono affetti da rumore o incertezza.

Quindi per provare a modellare questa incertezza e ad avvicinare il modello a tempo discreto a quello continuo per renderlo più realistico, invece di assumere che le osservazioni avvengano a tempi regolari, si ipotizza che il tempo

Parametro	Errore Medio	Dev. Std	Copertura 95%
μ_1	0.298	0.226	100%
μ_2	0.532	0.317	90%
θ	0.0029	0.0026	90%
σ^2	0.0178	0.0112	100%
Errore Totale	0.851	0.293	

Tabella 3.11: Statistiche di stima per $\theta = 0{,}03$, $\sigma = 0{,}9$, $a = 2$, $b = 4$, $n = 1000$, su 10 simulazioni (modello continuo).

reale t_i^* sia una variabile aleatoria distribuita uniformemente nell'intervallo

$$t_i^* \sim \mathcal{U}(t_i - \varepsilon,\, t_i + \varepsilon)$$

dove $\varepsilon > 0$ rappresenta l'ampiezza della fluttuazione attorno al tempo previsto.
Per rendere possibile il confronto tra i vari modelli, si mantiene $\Delta t = 3$, ovvero di tre ore, e $\varepsilon = 0.5$, in modo che l'incertezza sul monitoraggio sia di circa mezz'ora.

$$t_i^* \sim \mathcal{U}(t_i - 0{,}5,\, t_i + 0{,}5)$$

Così facendo la media dei tempi rimane inalterata $\mathbb{E}[t_i^*] = t_i$, ma ogni osservazione avviene a un tempo leggermente perturbato.

Questa correzione permette di simulare un processo stocastico che riflette la variabilità reale nel campionamento temporale, preservando in media un passo temporale fisso.

Per quanto riguarda la simulazione, i tempi sono ottenuti nel seguente modo:

$$t_i^* = t_i + \eta_i, \quad \eta_i \sim \mathcal{U}(-\varepsilon, \varepsilon)$$

Le posizioni vengono quindi simulate con dinamica continua 2D utilizzando tali tempi t_i^*, mantenendo la struttura del modello OU nel continuo.

Tabella 3.12: Riassunto dei parametri stimati – OU Discreto 2D (tempo aleatorio)

Parametro	Media	MCSE	Dev.Std	N_Eff	N_Eff/s	$\hat{R}$
lp__	-1811.54	0.0552	1.5474	787.20	194.90	0.9996
θ	0.0264	0.0001	0.0032	1835.39	454.42	0.9997
μ_1	0.5931	0.0149	0.6693	2010.51	497.77	0.9995
μ_2	0.5973	0.0153	0.6336	1725.88	427.30	0.9997
σ	0.9033	0.0004	0.0150	1665.54	412.37	0.9995
σ^2	0.8162	0.0007	0.0271	1661.07	411.26	0.9995

Le osservazioni (x_i, y_i) dipendono quindi da $t_i^* - t_{i-1}^*$ tramite la funzione:

$$\mathbb{E}[X_i] = \mu + (X_{i-1} - \mu)e^{-\theta\Delta t_i}, \quad \text{Var}[X_i] = \frac{\sigma^2}{2\theta}\left(1 - e^{-2\theta\Delta t_i}\right)$$

Questa simulazione discreta corretta verrà confrontata poi con il modello a tempo discreto e continuo e si valuterà se il modello continuo, che tiene esplicitamente conto di $\Delta t_i = t_i - t_{i-1}$, riesca a recuperare correttamente i parametri del processo OU con dati provenienti da un discreto "rumoroso".

Risultati campionamento OU discreto con rumore

Dalle tabelle e dai grafici anche in questo caso vediamo che i valori reali sono compresi nei rispettivi intervalli di credibilità, che sono più ampi rispetto agli altri, comportamento che riflette l'aumento di incertezza dovuto alla variabile temporale aleatoria.
Rispetto agli altri modelli σ è leggermente più elevato e vi è più dispersione attorno al centro.
Questo modello fornisce traiettorie più realistiche e, nonostante non mostri miglioramenti netti in termini di accuratezza rispetto ai modelli discreto e continuo, esso fornisce comunque stime affidabili e coerenti.

Tabella 3.13: Intervalli di Credibilità al 95% – OU Discreto 2D (tempo aleatorio)

Parametro	IC 95%
θ	(0.0201, 0.0326)
μ_1	(-0.7341, 1.9534)
μ_2	(-0.6002, 1.9203)
σ^2	(0.7654, 0.8730)

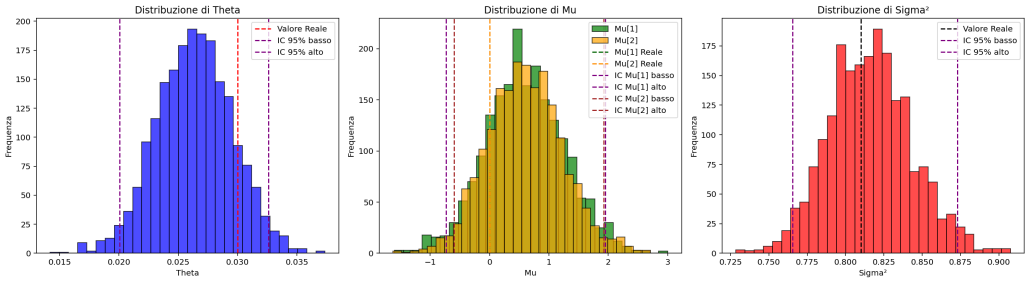


Figura 3.26: Distribuzioni posteriori per θ, μ e σ^2.

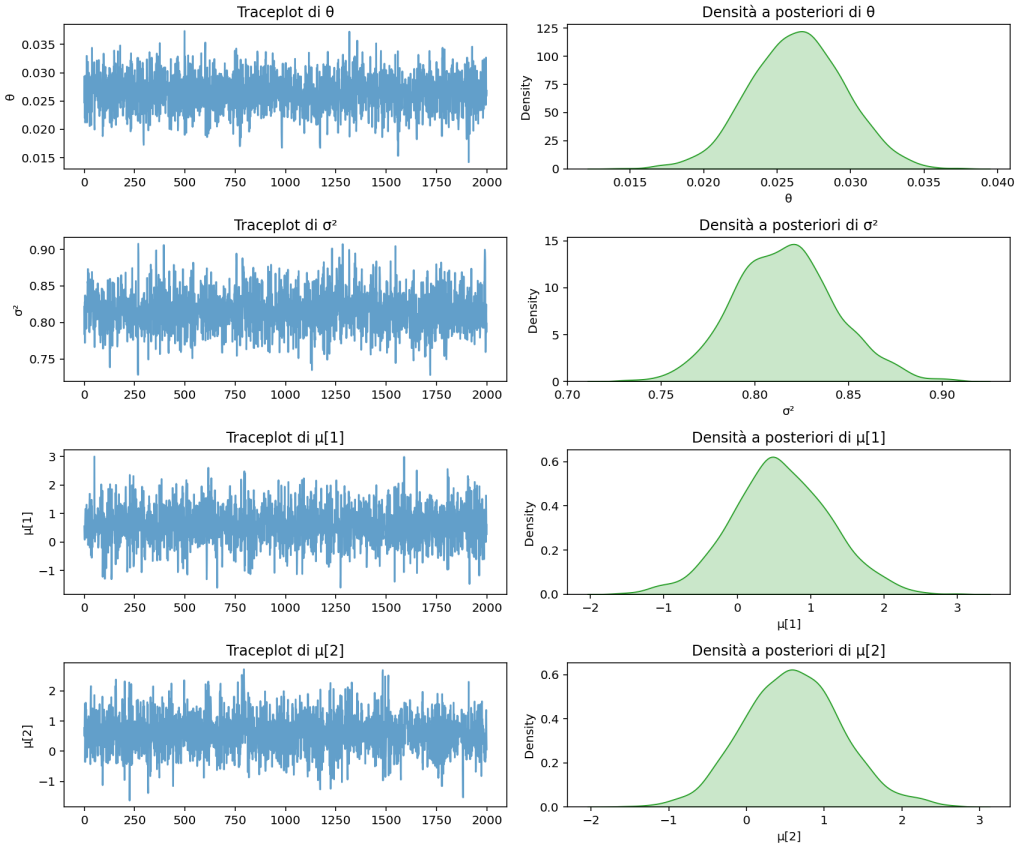


Figura 3.27: Densità a posteriori e tracce campionate

Capitolo 4

Analisi dati reali e applicazione modello OU

4.1 Descrizione dati reali

I dati utilizzati per questa analisi derivano da localizzazioni GPS raccolte su orsi e lupi, successivamente sottocampionate in modo da conservare una localizzazione ogni tre ore. Le localizzazioni di ciascun individuo o branco sono state suddivise in quattro stagioni distinte: *spring*, *early summer*, *late summer* e *autumn*.

Ogni punto nel dataset è associato a un identificativo, id, che nel caso degli orsi corrisponde al codice individuale ed è accompagnato dal sesso F o M, mentre per i lupi rappresenta la sigla del branco di appartenenza.

La colonna `Pres` indica se un punto è reale, "USE", o disponibile, "AVAILABLE". Nel primo caso è posta uguale a 1, mentre nel secondo è posta uguale a 0 ed è generato artificialmente.
I punti disponibili sono stati campionati casualmente all'interno dell'home range, ovvero all'interno dell'area usualmente utilizzata dall'individuo o dal branco durante le attività quotidiane, quali il riposo o la ricerca di cibo, con una densità di 10 localizzazioni per chilometro quadrato.

Ad ogni punto, sia di USE, sia disponibile, sono associate diverse variabili ambientali:

- **Agriculture**: percentuale di aree agricole;
- **NoVeg**: percentuale di aree rocciose prive di vegetazione;
- **PasturesGrasslands**: percentuale di pascoli e praterie;
- **Shrubs**: percentuale di arbusteti
- **TD_Beech**: densità media di alberi nei boschi di faggio (numero di alberi per pixel);
- **TD_OakHop**: densità media di alberi nei boschi di rovere e carpino nero, numero di alberi per pixel;
- **Dist_ForEdge**: distanza dai margini del bosco, valori positivi per l'esterno e negativi per l'interno;
- **Dist_Settl_R1**: distanza dai centri abitati e dalle strade principali;
- **Dist_R2**: distanza da strade secondarie;
- **Hillshade**: ombreggiatura;
- **TRI**: Terrain Ruggedness Index, cioè indice di asperità del del terreno
- **mappa_RSF_SPECIE_STAGIONE**: probabilità di occorrenza dell'altra specie, derivata esclusivamente da modelli di Resource Selection Function (RSF) specifici per specie e stagione, costruiti considerando esclusivamente variabili ambientali e non l'altra specie

Alcune variabili sono state trasformate, in questo caso standardizzate, al fine di ridurre problemi di collinearità e inoltre per ogni variabile è disponibile una versione standardizzata (z-score), indicata con il suffisso `.s` e utilizzata nei modelli di Resource Selection Function (RSF).

Esplorazione del dataset

Come già anticipato, sono presenti i monitoraggi di diversi individui e branchi di lupi e orsi raccolti in quattro diverse stagionalità.
in Tabella 4.1 sono riportati il numero di orsi e branchi monitorati per stagione, assieme al numero totale di punti totali, use, available e anche il rapporto tra available e use.

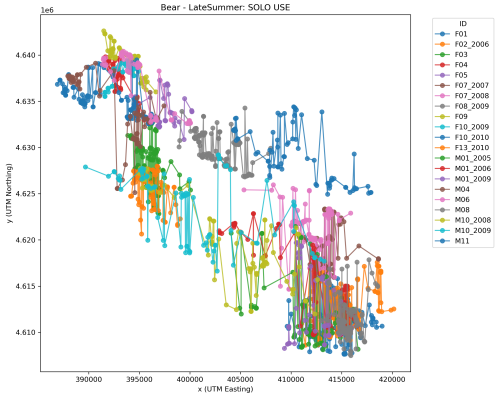
Tabella 4.1: Dati riepilogativi per specie e stagione.

Stagione	Specie	Individui	Pt totali	Pt use	Punti available	Available/Use
Spring	Bear	15	26509	1876	24633	13.13
Autumn	Bear	20	28037	2840	25197	8.87
EarlySummer	Bear	16	31961	1285	30676	23.87
LateSummer	Bear	21	38381	2949	35432	12.01
EarlySummer	Wolf	4	8039	1355	6684	4.93
LateSummer	Wolf	3	6906	921	5985	6.50
Spring	Wolf	5	11047	1619	9428	5.82
Autumn	Wolf	5	10989	1843	9146	4.96

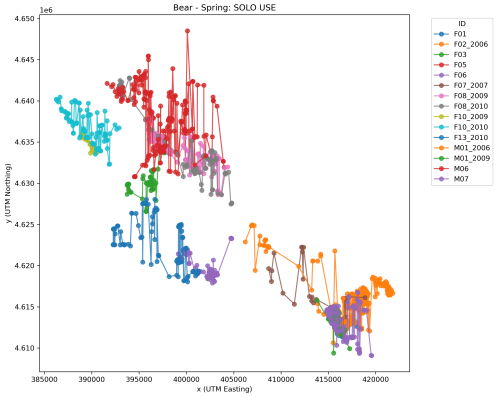
Tuttavia verranno considerati solo i punti USE, ovvero quelli reali. Vi sono degli orsi e dei branchi di lupi che presentano dati per tutte e quattro le stagioni prese in analisi.

- **Orsi (Bear):** F01, F02_2006, F03, F05, F07_2007, F08_2009, F10_2010, F13_2010.
- **Lupi (Wolf):** COLP, CP, VP.

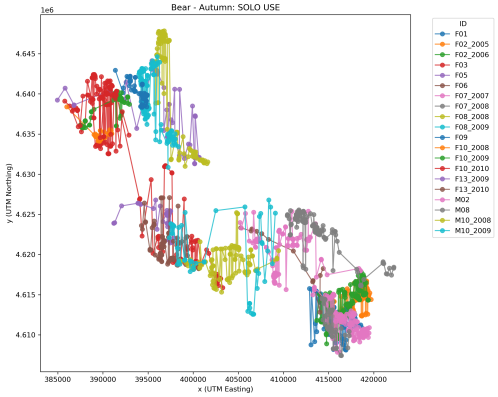
Si nota che non sono presenti dati relativi a tutte le stagioni per tutti gli esemplari di orso maschio, ma ve ne sono alcuni di cui sono disponibili gli spostamenti per tre stagioni: M01_2006, M01_2009, M06, M10_2008.
Di seguito sono riportate tutte le traiettorie suddivise per specie e stagione.



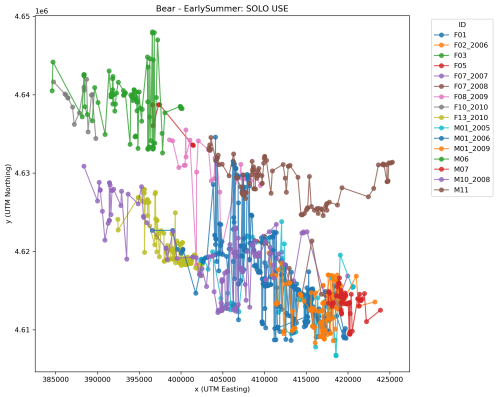
(a) Bear - Late Summer



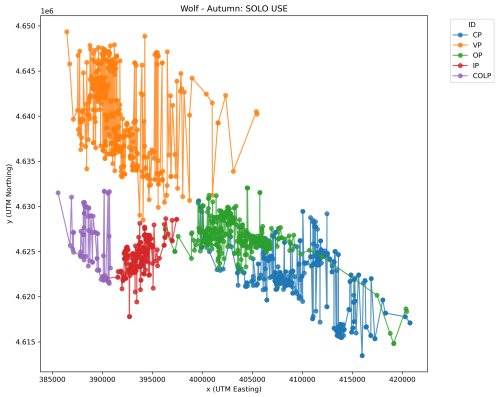
(b) Bear - Spring



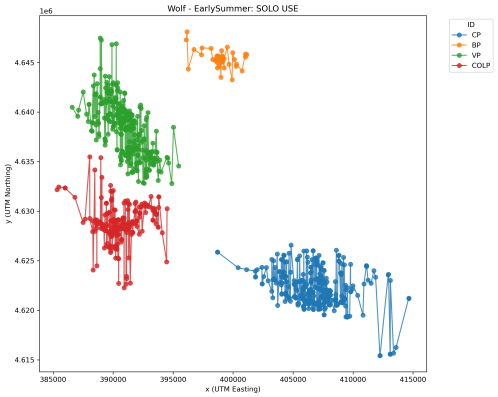
(c) Bear - Autumn



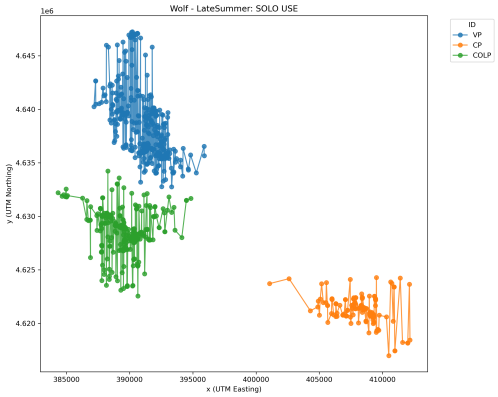
(d) Bear - Early Summer



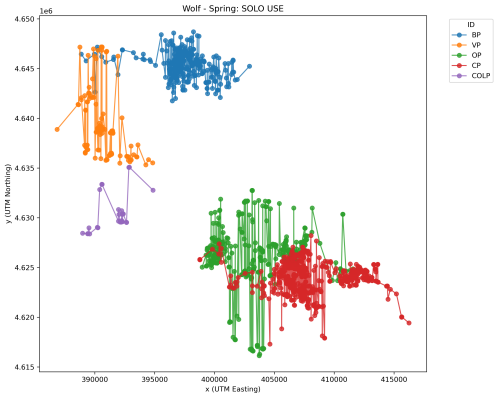
(e) Wolf - Autumn



(f) Wolf - Early Summer



(g) Wolf - Late Summer



(h) Wolf - Spring

Analisi della distanza totale degli spostamenti per stagione e specie

Per ogni individuo e branco in ciascuna stagione, sono state ricavate la distanza totale percorsa, la massima distanza, la media e la varianza di ogni spostamento.
Vengono considerati solo i punti USE, ovvero quelli in cui è stato realmente localizzato l'esemplare.

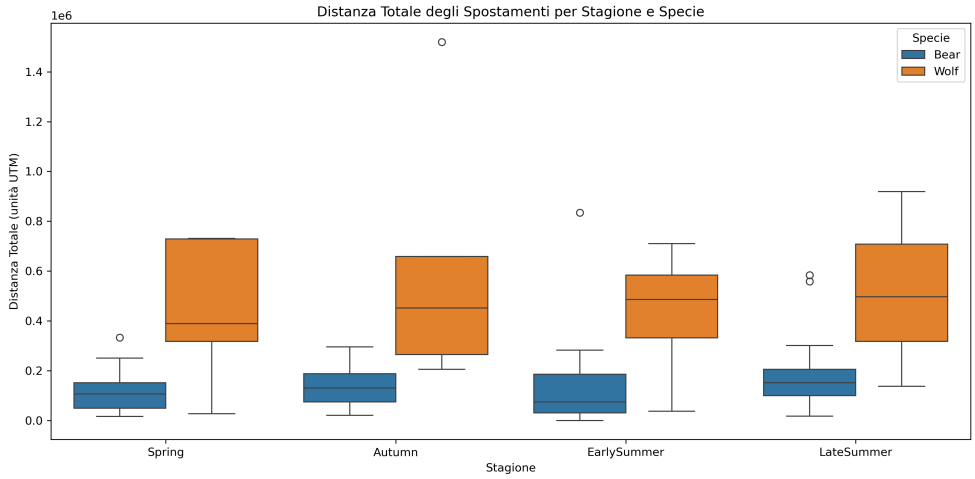


Figura 4.2: Boxplot della distanza totale percorsa (in unità UTM) da orsi e lupi nelle diverse stagioni.

In Figura 4.2 i boxplot riportano la distribuzione della distanza totale percorsa, in unità UTM, dagli individui di due specie animali, l'orso e il lupo, nel corso delle quattro stagioni.
Questa analisi non può essere considerata valida in quanto il dataset è sbilanciato sia per il numero di individui per specie, sia per le misurazioni presenti; tuttavia si possono trarre delle conclusioni separatamente per ciascuna specie.
É evidente che i lupi percorrono distanze totali significativamente maggiori rispetto agli orsi, nonostante siano presenti pochi branchi monitorati, e le distribuzioni relative ai lupi presentino una maggiore variabilità, evidenziata

da boxplot più ampi e dalla presenza di numerosi outlier, in particolare durante l'Autumn e l'Early Summer.

Per gli orsi, la distanza totale percorsa tende ad aumentare dalla Spring all'Early Summer, con una leggera stabilizzazione nella Late Summer, e la variabilità è contenuta e le distribuzioni risultano abbastanza simmetriche. I lupi, invece, mostrano una notevole mobilità già dalla primavera, con un'ulteriore espansione della variabilità nei mesi estivi e autunnali. Si può comunque notare, anche con dataset sbilanciati, che i lupi, in quanto predatori attivi, percorrono distanze maggiori per cercare di inseguire le prede. Invece, gli orsi adottano una strategia più localizzata, con una mobilità nettamente inferiore.

Calcolo home-range e centrodi per specie e stagione

Per ogni individuo monitorato sono stati stimati il centroide e l'home range stagionale, utilizzando la tecnica del Minimum Convex Polygon, MCP, che rappresenta il più piccolo poligono convesso, che racchiude tutti i punti GPS dell'animale per una data stagione. Si può così stimare l'area utilizzata da ciascun individuo per stagione.
I centroidi, invece, sono il baricentro dei punti utilizzati, USE.

Lupo (Wolf)

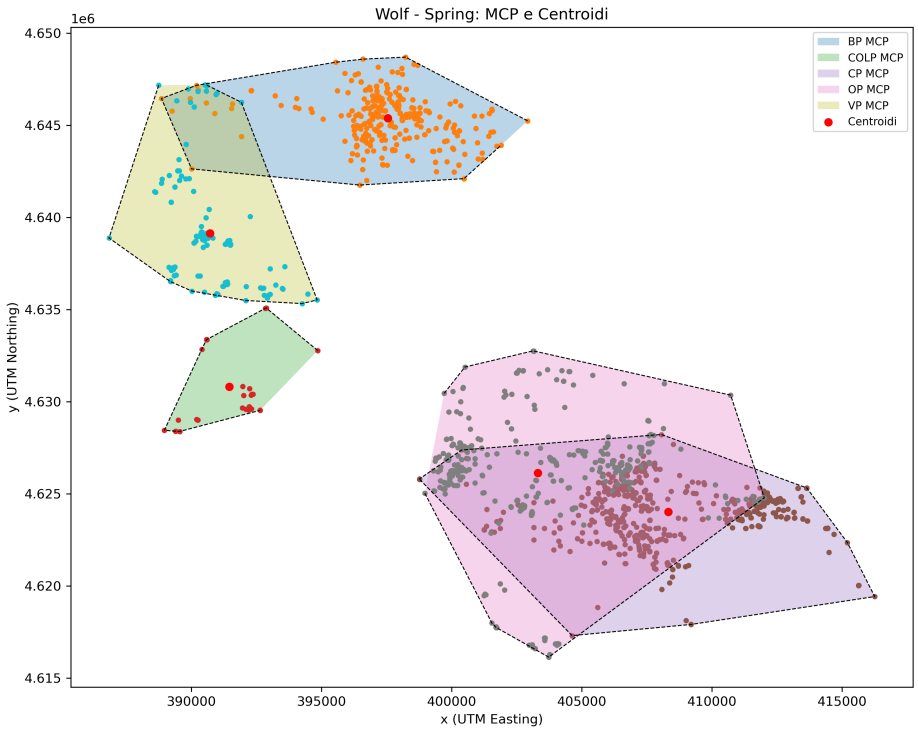


Figura 4.3: Lupo - Primavera: home range stagionali (MCP) e centroidi

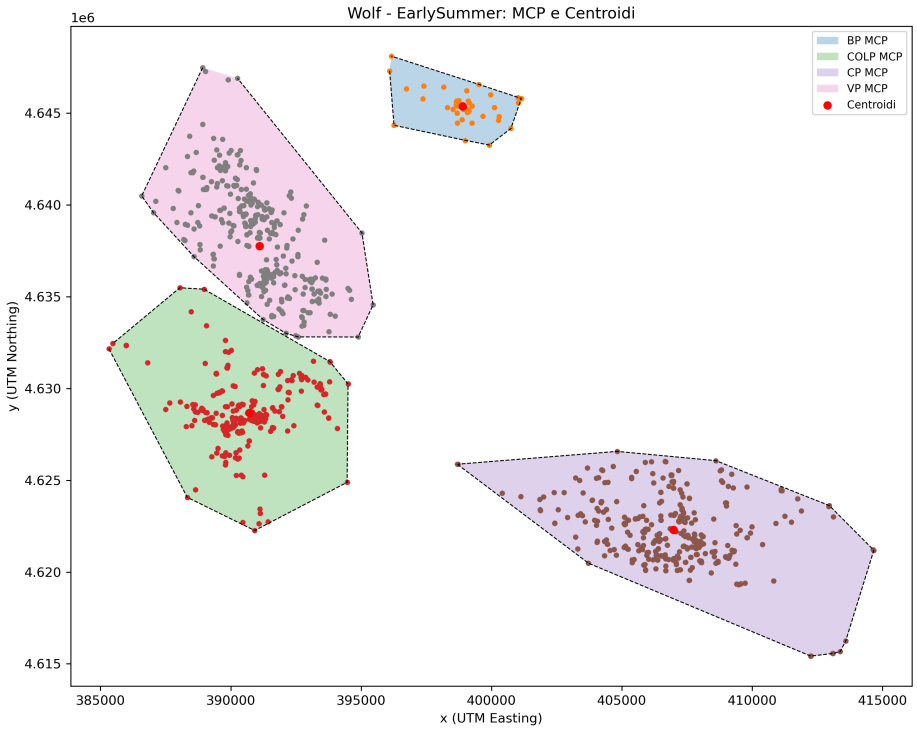


Figura 4.4: Lupo - Inizio estate: home range stagionali (MCP) e centroidi

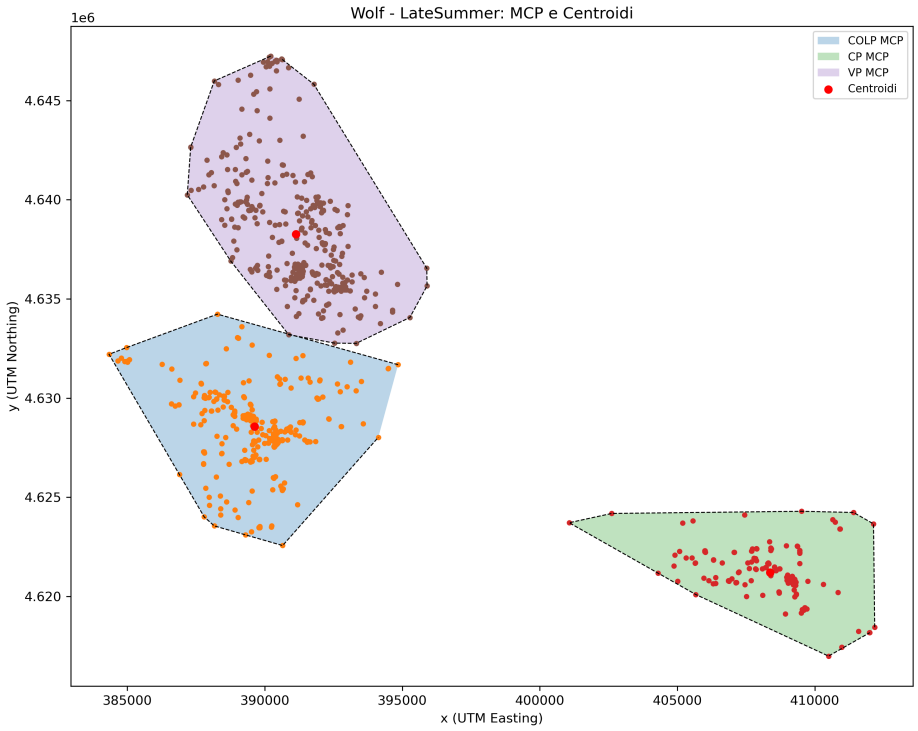


Figura 4.5: Lupo - Fine estate: home range stagionali (MCP) e centroidi

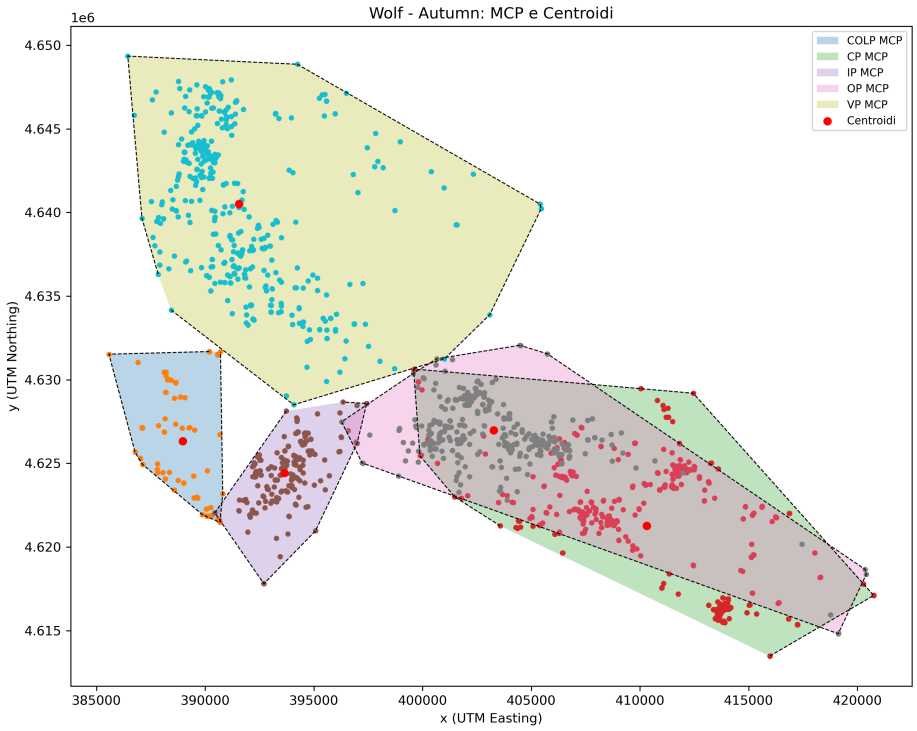


Figura 4.6: Lupo - Autunno: home range stagionali (MCP) e centroidi

Orso (Bear)

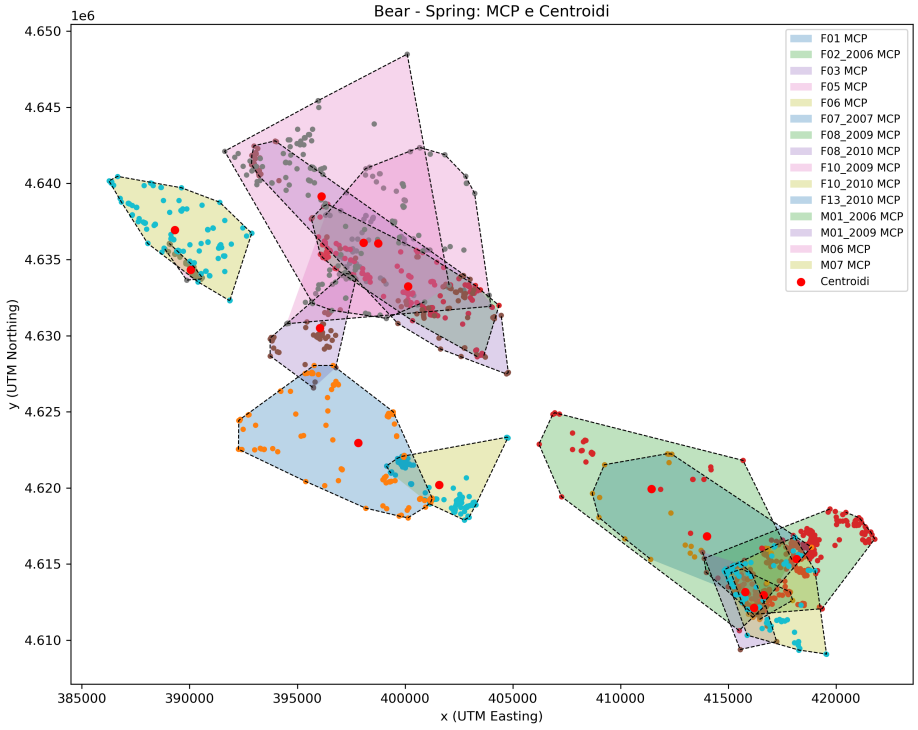


Figura 4.7: Orso - Primavera: home range stagionali (MCP) e centroidi

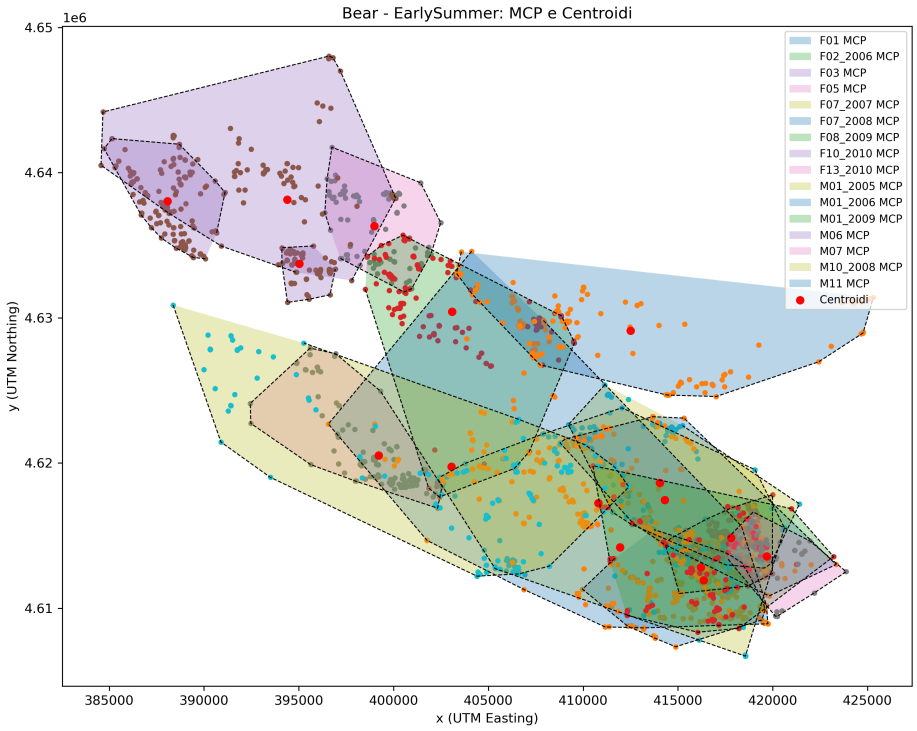


Figura 4.8: Orso - Inizio estate: home range stagionali (MCP) e centroidi

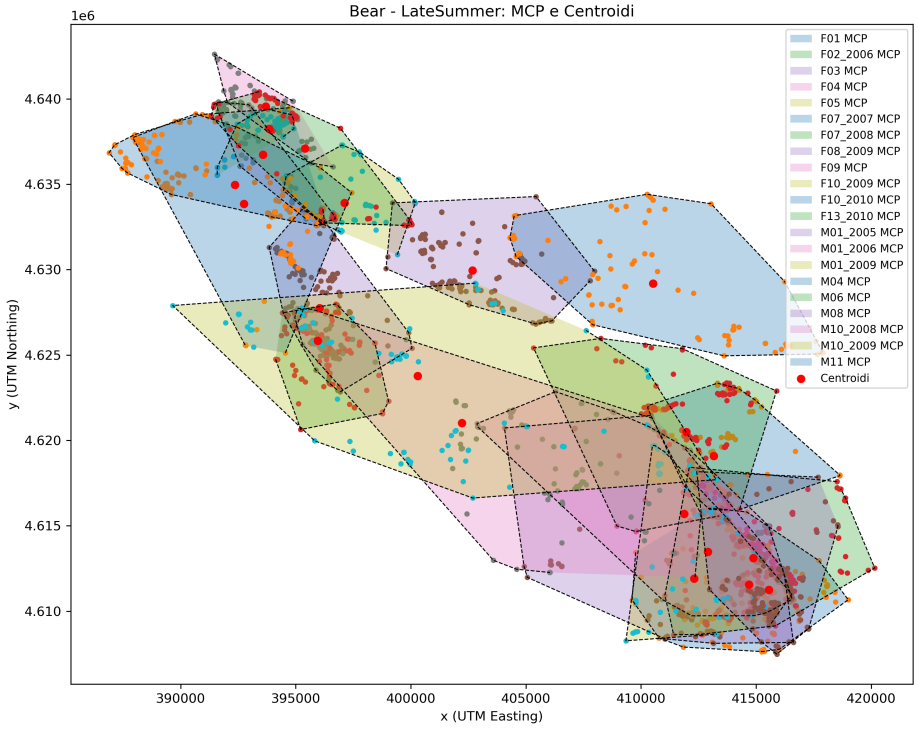


Figura 4.9: Orso - Fine estate: home range stagionali (MCP) e centroidi

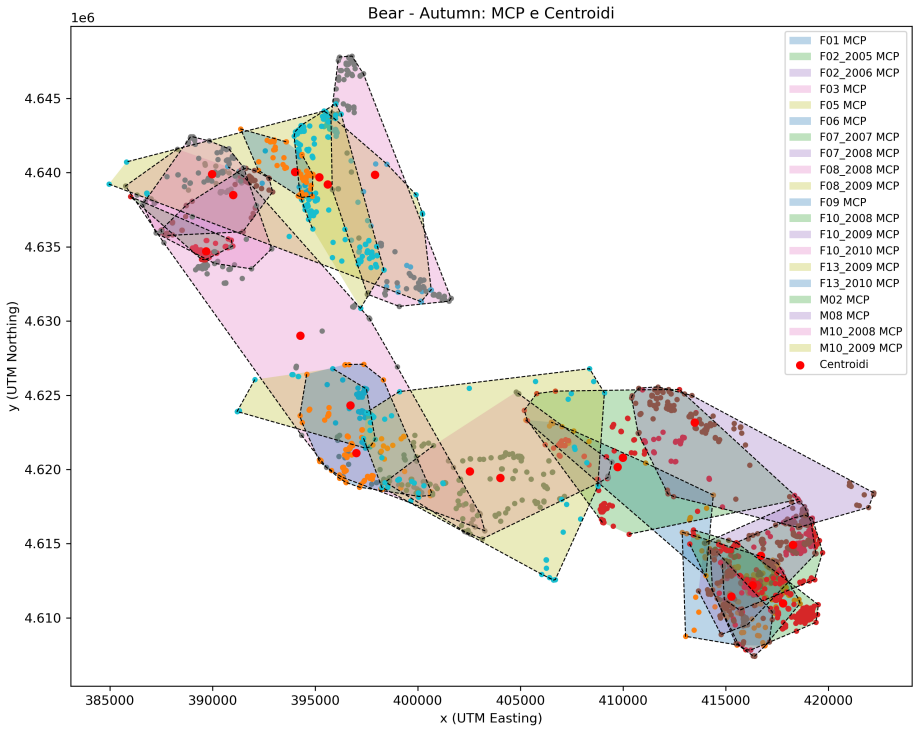


Figura 4.10: Orso - Autunno: home range stagionali (MCP) e centroidi

Osservazioni: Nei grafici dell'orso, si osserva una notevole variazione nella dimensione e nella posizione degli MCP tra le stagioni. In estate, alcuni individui mostrano un'espansione dell'area di utilizzo, potenzialmente legata a una maggiore disponibilità di risorse alimentari o all'attività pre-ibernazione. I centroidi si distribuiscono su un'ampia porzione del territorio, suggerendo una diversificazione delle aree frequentate.
Nei lupi, l'estensione degli MCP appare più costante tra le stagioni, suggerendo una territorialità più stabile rispetto agli orsi. I centroidi mostrano una minore dispersione spaziale, coerente con un comportamento di uso del territorio più focalizzato, probabilmente legato alla struttura sociale del branco e al controllo del territorio.
In seguito è riportato anche un test di analisi della varianza, ANOVA, intraspecie, per vedere se la stagionalità influisce sulla distanza totale media percorsa.

L'ipotesi nulla:
$H0$: "Le medie delle distanze totali nelle stagioni sono uguali".
Se il p-value è < 0.05, si rifiuta $H0$, quindi c'è evidenza di una differenza significativa tra stagioni. Tuttavia, essendoci un valore di P_value alto, la stagionalità sembra non essere influente sulle distanze totali percorse da lupi e orsi nei periodi diversi, possiamo pertanto accettare l'ipotesi nulla.

Tabella 4.2: Risultati dell'ANOVA intraspecie sulla distanza totale tra stagioni.

Specie	**Metrica**	**F-value**	**p-value**	**Stagioni confrontate**
Bear	Distanza totale	0.7305	0.5374	Autumn, EarlySummer, LateSummer, Spring
Wolf	Distanza totale	0.2360	0.8697	Autumn, EarlySummer, LateSummer, Spring

4.2 Preprocessing

Le variabili che verranno poste in analisi saranno solo le coordinate spaziali x e y, e l'intervallo temporale, fisso a 3 ore.
Ad una prima vista i dati delle coordinate GPS sembrano essere proiettati in coordinate UTM, ma sono presenti alcuni outlier, ovvero valori fuori scala, probabilmente dovuti a qualche errore di misurazione o trascrizione nel dataset, che compromettono la visualizzazione dei grafici di spostamento, l'analisi delle metriche di spostamento e quelle successive di tipo inferenziale.

Pulizia dei dati spaziali

I dati grezzi delle coordinate spaziali (`x`, `y`) associati agli spostamenti degli individui contenevano alcune osservazioni errate o anomale, ovvero valori fuori dal dominio coerente con il sistema di riferimento UTM, Universal Transverse Mercator, zona 33T[1], che copre l'area di studio. Nel sistema UTM, la coordinata x, Easting, rappresenta la distanza in metri rispetto al meridiano

[1]https://coordinates-converter.com/it/decimal/43.383416,13.013151?karte=OpenStreetMap&zoom=5

centrale della zona di riferimento. Per la zona 33T, il meridiano centrale è posto a 15° Est. La coordinata y, Northing misura la distanza in metri dall'equatore verso nord.

Per garantire l'accuratezza dell'analisi, è stato applicato un filtro preliminare alle coordinate, mantenendo solo i punti che rispettano i seguenti criteri:

$$200{,}000 < x < 800{,}000 \quad \text{e} \quad 4{,}500{,}000 < y < 4{,}900{,}000$$

Tale intervallo include le coordinate valide per la zona UTM di interesse, escludendo righe che presentavano errori di formattazione, conversione o valori estremi derivanti da esportazioni errate o assenza di dati.
Il filtraggio ha permesso di eliminare le righe non coerenti con la zona di interesse, come si vede in tabella 4.3:

Tabella 4.3: Riepilogo righe valide e rimosse per file dopo filtraggio coordinate UTM zona 33T

File	**Totale**	**Valide**	**Rimosse**
Bear_Spring	26509	26509	0
Bear_Autumn	28037	28037	0
Bear_EarlySummer	33374	31961	1413
Bear_LateSummer	38381	38381	0
Wolf_EarlySummer	8039	8039	0
Wolf_LateSummer	6906	6906	0
Wolf_Spring	11047	11047	0
Wolf_Autumn	10989	10989	0

Il numero di righe conservate per ciascun file è stato tracciato e confrontato con il numero di osservazioni originali per quantificare l'impatto della pulizia.

L'unico file che presenta outlier è quello di "Bear_EarlySummer".

Dopo una fase iniziale di inferenza con il modello OU discreto bidimensionale adottato in precedenza, nonostante il modello fosse corretto e il processo di campionamento si effettuasse senza errori, sono emersi valori di σ^2 molto elevati e, viceversa, di μ molto bassi, non coerenti con i reali valori delle coordinate UTM date in input.
X si aggira sempre attorno ai 400.000, mentre Y attorno ai 4.000.000, mentre i valori medi stimati di μ sono inferiori a 10, motivo per cui ci troviamo in presenza di una varianza molto alta per compensare questo grande divario dimensionale.
Tutto questo è dovuto al fatto che la prior imposta su μ_i, $\mathcal{N}(0, 100)$, è non informativa se i dati sono compresi in un intervallo tra 0 e 10. In questo caso la prior risulterebbe molto più larga della variabilità dei dati e avrebbe un'influenza minima.
Essendo però i dati reali nell'ordine delle centinaia di migliaia, la prior è totalmente fuori scala, perché concentra tutta la sua massa vicino allo 0, mentre i dati sono distanti centinaia di migliaia di unità. Quindi essa è fortemente informativa in senso negativo, perché forza l'inferenza lontano dai valori plausibili.
Al fine di lavorare quindi con delle prior che abbiano senso con il dataset reale, si possono standardizzare le coordinate spaziali. Questa trasformazione ha lo scopo di centrare i dati rispetto alla media e di normalizzare la loro varianza, portandola ad unità:

$$x_i^{\text{std}} = \frac{x_i - \mu_x}{\sigma}, \quad y_i^{\text{std}} = \frac{y_i - \mu_y}{\sigma}$$

dove μ_x, σ_x, μ_y, e σ rappresentano la media delle coordinate x e y e la deviazione standard, rispettivamente, che sarà mantenuta uguale per entrambe le coordinate al fine di rendere confrontabili le stime senza cambiare i rapporti.

4.2.1 Standardizzazione delle Coordinate

Un'elevata dispersione può compromettere la stabilità numerica del campionamento MCMC e la standardizzazione migliora la condizione numerica del modello e rende più un confronto tra individui o periodi stagionali diversi.

Sia la rimozione di outlier sia la standardizzazione non alterano la struttura del movimento, ma garantiscono l'affidabilità delle stime dei parametri. In particolare, la standardizzazione permette una migliore interpretabilità di μ e della varianza σ^2, riducendo l'effetto dell'unità di misura originale e concentrando l'analisi sul comportamento dinamico piuttosto che sulla scala assoluta, al fine di poter confrontare i comportamenti tra diverse specie, individui e stagionalità, penalizzando tuttavia l'interpretazione geografica.

4.3 Inferenza bayesiana con modello OU discreto

L'inferenza è stata condotta su tutti i dataset disponibili per ciascun individuo o branco e ciascuna stagionalità.
Ho deciso però di restringere l'analisi dei risultati ottenuti dall'inferenza ad un esemplare per specie. In particolare su un esemplare di orso femmina, F01, e su un branco di lupi, COLP.
La scelta degli individui è stata guidata dal fatto che essi abbiano dati presenti su tutte e quattro le stagionalità.
Mi soffermerò sui risultati ottenuti per un'esemplare di orso femmina, F01 nella stagione Late-Summer.
In seguito alla standardizzazione, si nota che le medie μ_1 e μ_2 risultano comprese tra circa $-0.0001.5$ e 0.319, indicando un centro di attrazione localizzato attorno allo zero.
Si riscontra una netta riduzione della varianza da ordini di grandezza di 10^9 a valori attorno a 10^{-2}, il che testimonia l'efficacia della trasformazione.

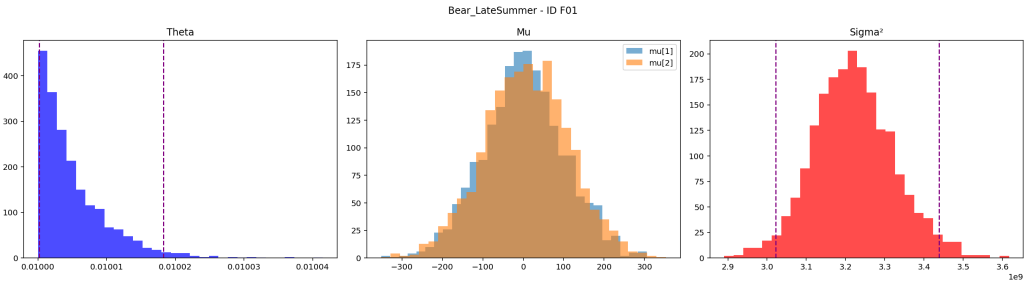


Figura 4.11: Distribuzioni posteriori dei parametri per l'individuo F01

Anche θ presenta una distribuzione a posteriori più simmetrica rispetto alla versione non standardizzata, che era schiacciata verso lo zero e fortemente asimmetrica.

Tabella 4.4: Confronto tra parametri stimati non standardizzati e standardizzati (F01, fine estate)

Parametro	**Non standardizzato**	**Standardizzato**
μ_1 (media)	~ -1.34	-0.0001
μ_2 (media)	~ 8.589	0.0319
θ (media)	0.01001	0.0445
σ^2 (media)	3.2×10^9	0.0846

In sintesi, la standardizzazione è essenziale a rendere le analisi più robuste e comparabili. Tuttavia, nel caso in cui si desideri interpretare i parametri stimati in termini di unità originali, è necessario effettuare la trasformazione inversa utilizzando i parametri di media e deviazione standard salvati durante il preprocessing.

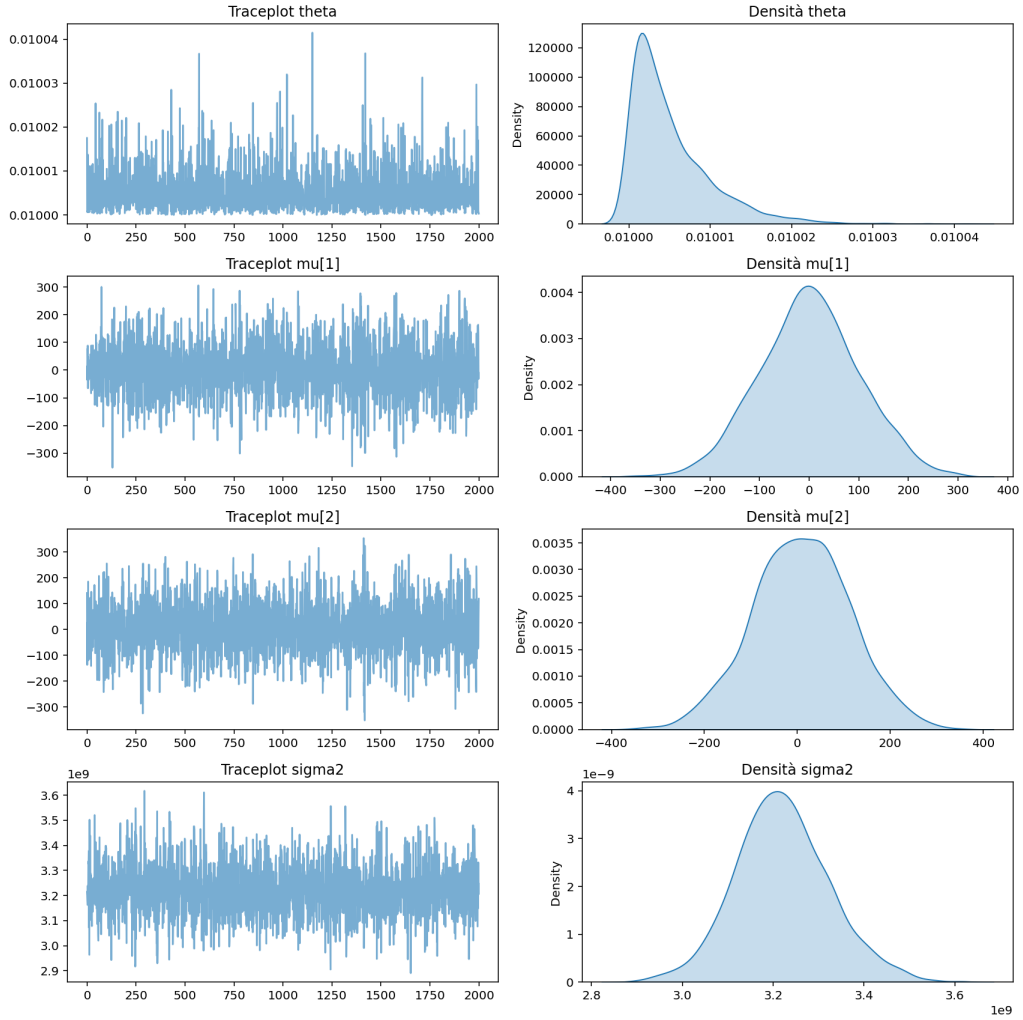


Figura 4.12: Traceplot e densità a posteriori dei parametri per l'individuo F01

Analisi stagionale dei parametri stimati per F01

Dati originali

Stagione	theta	$\mu[1]$	$\mu[2]$	$\sigma 2$
Primavera	0.0100	-1.0193	1.0021	3221944065.0000
Estate Inizio	0.0100	-1.1267	3.0261	3221986440.0000
Estate Fine	0.0100	1.3426	8.5893	3220460875.0000
Autunno	0.0100	1.9236	3.1107	3215982620.0000

Tabella 4.5: Valori medi stimati dei parametri per l'individuo F01 (orso), coordinate originali, per ciascuna stagione

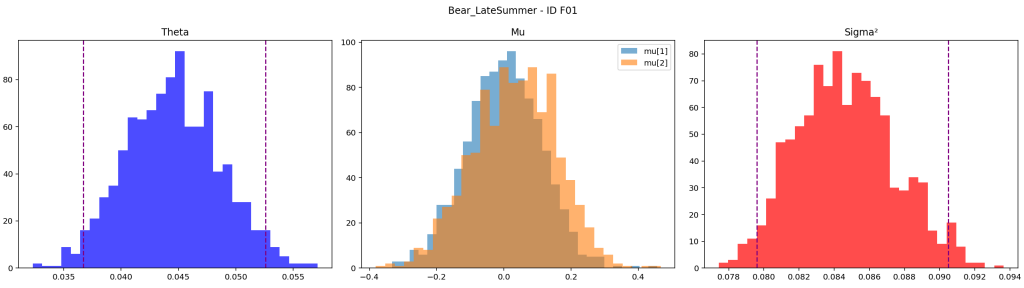


Figura 4.13: Distribuzioni posteriori dei parametri per l'individuo F01 dopo standardizzazione (tarda estate)

L'analisi evidenzia una variazione stagionale nel centro del movimento dell'orso F01.

Il parametro μ_1 passa da valori negativi in primavera ed estate, con un minimo di circa -1.13, a valori positivi in autunno di circa 1.92, suggerendo uno spostamento verso est. Il parametro μ_2 mostra un massimo in fine estate, pari a 8.59, seguito da una riduzione in autunno, circa 3.11, indicando un picco dell'attività a nord durante l'estate.

Il parametro θ risulta costante e basso, pari a 0.0100 in tutte le stagioni, denotando una dinamica di movimento poco vincolata a un centro fisso oppure una limitazione nella stima del parametro.

La dispersione spaziale σ^2 rimane anch'essa stabile tra le stagioni, tra circa 3.22×10^9 e 3.21×10^9, il che suggerisce che l'estensione dell'area esplorata non cambia eccessivamente nel tempo, nonostante lo spostamento del centro sia considerevole.

Dati standardizzati

Passiamo ora all'analisi dei dati con coordinate standardizzate.

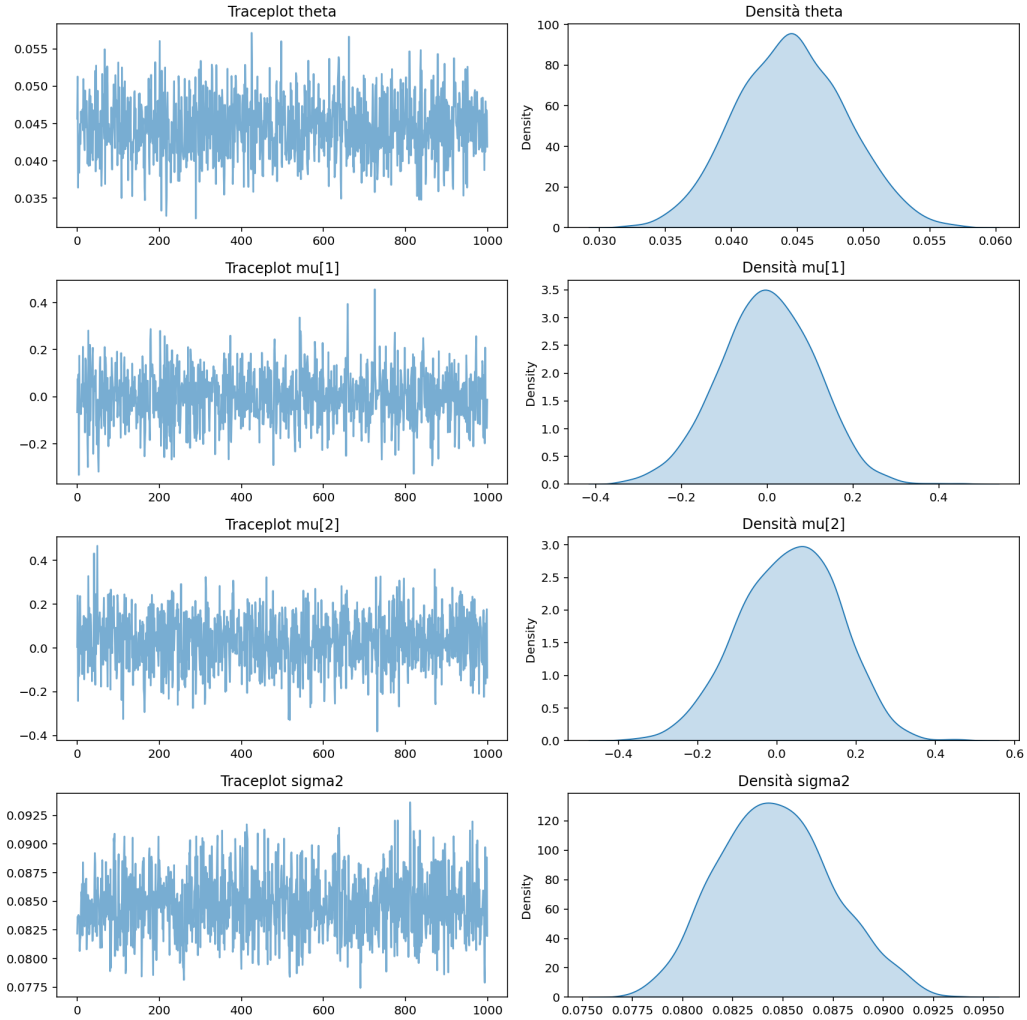


Figura 4.14: Traceplot e densità a posteriori dei parametri per l'individuo F01 dopo standardizzazione (tarda estate)

Tabella 4.6: Valori medi stimati dei parametri per l'individuo F01 (orso), coordinate standardizzate, per ciascuna stagione

Stagione	θ	μ_1	μ_2	σ^2
Primavera	0.0436	-0.0113	0.0011	0.0744
Estate Inizio	0.0462	-0.0069	0.0194	0.0801
Estate Fine	0.0445	-0.0002	0.0319	0.0846
Autunno	0.0392	0.0073	-0.0012	0.0669

L'analisi dei parametri stimati per l'individuo F01 nel corso delle diverse stagioni rivela variazioni significative che riflettono cambiamenti comportamentali e ambientali.
Il parametro θ, mostra valori relativamente stabili tra le stagioni, con una leggera diminuzione in autunno, evidenziando una maggiore regolarità nei movimenti durante la primavera e l'estate, mentre in autunno potrebbe esserci una maggiore variabilità nei percorsi seguiti.
Per quanto riguarda i parametri μ_1 e μ_2, si nota che in primavera, i valori sono prossimi allo zero e suggeriscono un'attività distribuita attorno alla posizione media di riferimento. Durante l'estate, si osserva un incremento di entrambi i parametri, soprattutto di μ_2, indicando uno spostamento dell'attività verso nord-est. In autunno, μ_2 si riduce sensibilmente, suggerendo un possibile ritorno verso sud, mentre μ_1 rimane positivo, mantenendo una tendenza verso est.
Il parametro σ^2 mostra un aumento durante l'estate, raggiungendo il massimo in fine estate. Questo indica una maggiore dispersione nei movimenti, probabilmente dovuta a una ricerca più ampia di risorse. In autunno, la varianza diminuisce, evidenziando una concentrazione dell'attività in un'area più ristretta.
Le variazioni stagionali nei parametri stimati indicano che l'individuo F01 adatta il proprio comportamento spaziale in risposta ai cambiamenti ambientali. L'espansione dell'area di attività durante l'estate potrebbe essere legata alla disponibilità di risorse, mentre il cambiamento riguardante l'autunno potrebbe riflettere una preparazione per l'inverno o una risposta a condizioni ambientali più restrittive.

Stagione	theta	mu[1]	mu[2]	sigma2
Primavera	0.0100	-1.7639	1.3867	3243370760.0000
Estate Inizio	0.0100	-2.3193	11.4586	3236807510.0000
Estate Fine	0.0100	-0.0110	6.5225	3238047755.0000
Autunno	0.0100	2.4003	4.3279	3241422185.0000

Tabella 4.7: Valori medi stimati dei parametri per il branco COLP (lupi), coordinate originali, per ciascuna stagione

Analisi stagionale dei parametri stimati per COLP

Dati originali

Anche in questo caso si riscontra lo stesso problema, evidenziato in precedenza nello studio di F01, della stima del parametro μ.
I parametri μ_1 e μ_2, che rappresentano il centro del movimento del branco COLP, mostrano variazioni stagionali compatibili con un comportamento esplorativo o con cambiamenti d'uso del territorio. In particolare, μ_1 cresce progressivamente da valori negativi in primavera ed estate, quali -1.76 e -2.32 relativamente, a valori positivi in autunno, pari a 2.40, suggerendo un lieve spostamento verso est.

Parallelamente, μ_2 mostra un andamento non monotono: parte da 1.39 in primavera, raggiunge un picco a inizio estate, pari a 11.46 e poi decresce fino a 4.33 in autunno. Questo comportamento potrebbe indicare variazioni stagionali nelle risorse o negli schemi di utilizzo dell'habitat, con un massimo estivo di attività verso nord.

Il parametro θ resta costante a 0,01 in tutte le stagioni, come osservato anche per l'orso. Questo potrebbe indicare una dinamica del movimento poco vincolata a un punto centrale stabile, oppure una difficoltà del modello a stimare θ con precisione, come già detto in precedenza.

La variabilità spaziale σ^2 resta anch'essa stabile attorno a valori di $3{,}2 \times 10^9$, suggerendo che l'ampiezza dell'area di attività del branco COLP non

cambia significativamente nel corso delle stagioni. La coerenza tra le stagioni indica una strategia di movimento stabile in termini di estensione spaziale.

Dati standardizzati

Passiamo ora a trattare le coordinate standardizzate.

Stagione	θ	μ_1	μ_2	σ^2
Primavera	0.0253	-0.0100	-0.0010	0.0734
Estate Inizio	0.0330	0.0107	0.0249	0.0953
Estate Fine	0.0340	-0.0008	0.0182	0.0996
Autunno	0.0260	0.0142	0.0504	0.0827

Tabella 4.8: Valori medi stimati dei parametri per il branco COLP (lupo), coordinate standardizzate, per ciascuna stagione

L'analisi dei parametri stimati per il branco di lupi COLP evidenzia variazioni stagionali che riflettono strategie di movimento e utilizzo dello spazio probabilmente legate a disponibilità ambientali.
Il parametro θ mostra valori leggermente più bassi in primavera e autunno pari a circa 0.025, con un incremento durante l'estate, raggiungendo il valore massimo in fine estate, 0.0340. Questo suggerisce una maggiore coerenza direzionale nei movimenti estivi, probabilmente dovuta a percorsi più lineari verso risorse specifiche o a spostamenti territoriali più marcati.
Per quanto riguarda i parametri μ_1 e μ_2, In primavera, entrambi i valori sono negativi o prossimi allo zero, suggerendo un'attività centrata attorno a un'area di riferimento. Con l'arrivo dell'estate, i valori aumentano progressivamente, in particolare μ_2, che passa da -0.0010 a 0.0504 in autunno. Questo indica uno spostamento netto verso nord, mentre l'aumento di μ_1 in autunno suggerisce anche una componente di spostamento verso est. Tali cambiamenti possono riflettere un ampliamento dell'area frequentata o un cambio stagionale nella posizione delle risorse.

Il parametro σ^2, che misura la dispersione attorno al centro dell'area di attività, aumenta progressivamente dalla primavera all'estate, raggiungendo un picco di 0.0996 a fine estate. Questo indica un ampliamento di utilizzo dell'area durante i mesi estivi, potenzialmente legato alla ricerca di cibo o all'esplorazione del territorio. In autunno, la varianza si riduce leggermente, segnalando una moderata contrazione dell'attività spaziale.
Nel complesso, i risultati indicano che il branco COLP adatta la propria strategia spaziale alle stagioni, con una maggiore espansione e direzionalità nei mesi estivi e una tendenza alla contrazione e maggiore variabilità nei mesi primaverili e autunnali, comportamento che potrebbe riflettere una risposta collettiva a cambiamenti nella disponibilità di risorse o a pressioni ambientali stagionali.

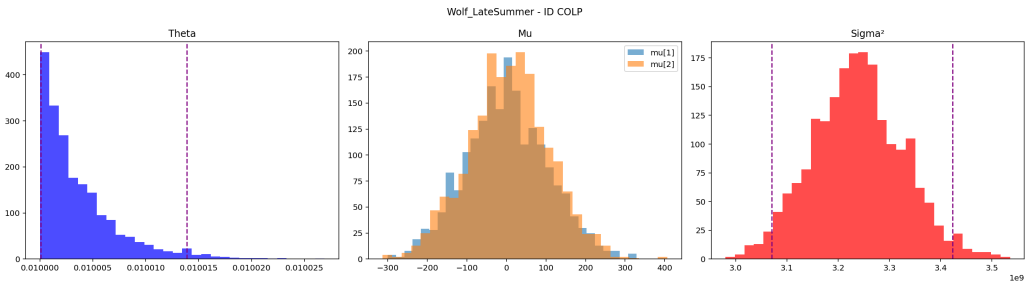


Figura 4.15: Distribuzioni posteriori dei parametri per il branco COLP

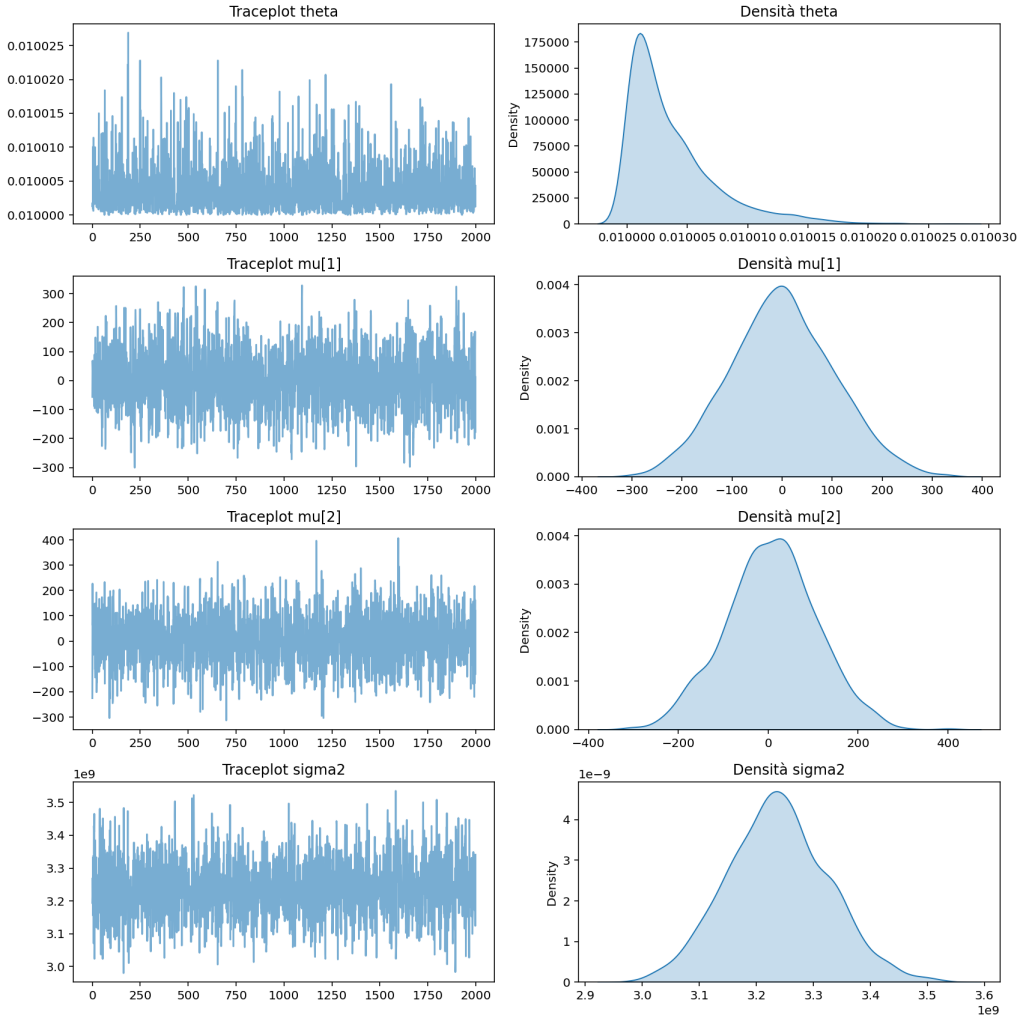


Figura 4.16: Traceplot e densità a posteriori dei parametri per il branco COLP

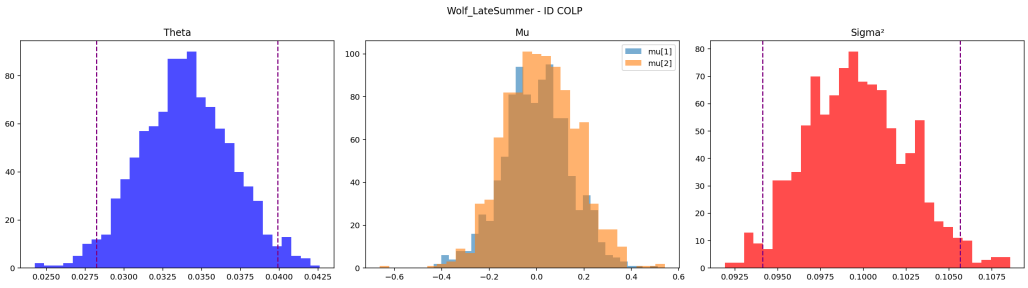


Figura 4.17: Distribuzioni posteriori dei parametri per branco COLP dopo standardizzazione (estate tarda)

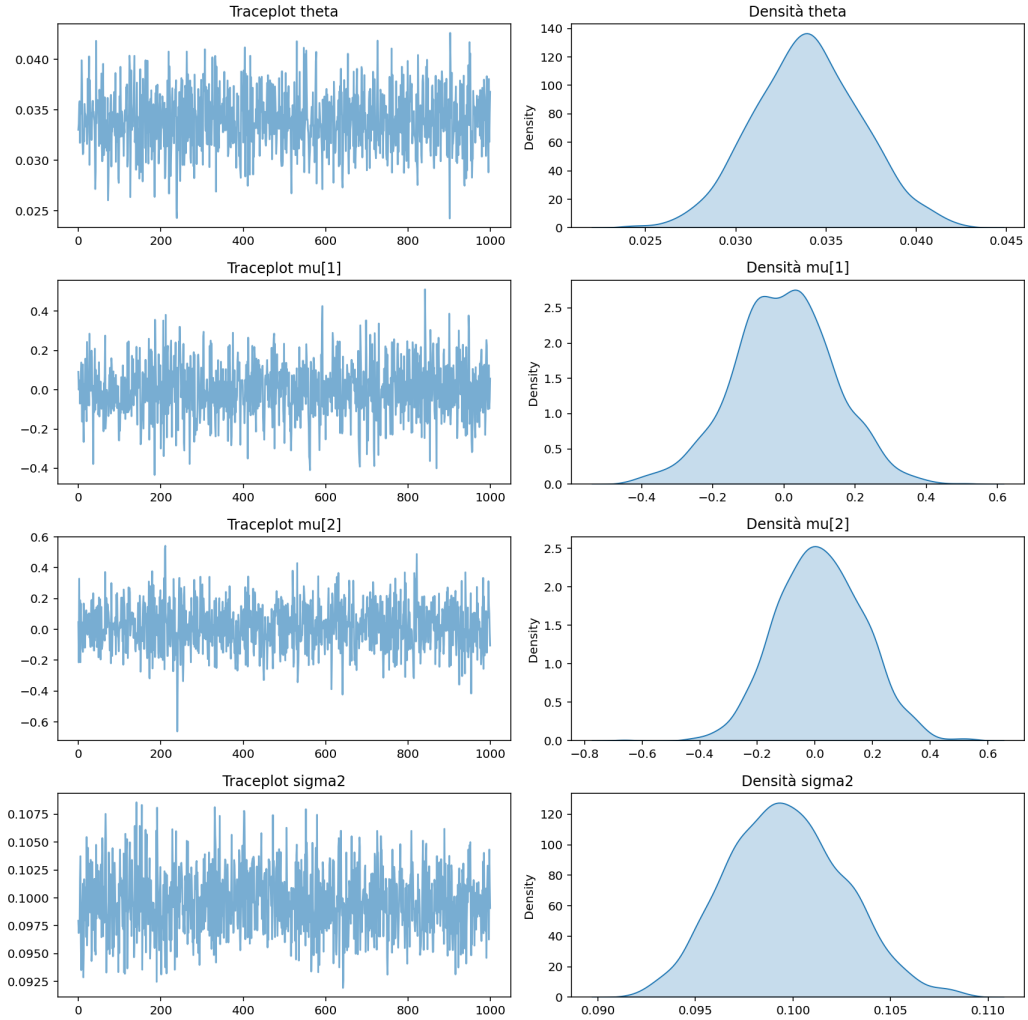


Figura 4.18: Traceplot e densità a posteriori dei parametri per branco COLP dopo standardizzazione

4.3.1 Confronto tra lupo e orso: COLP vs F01

Dati originali

Confrontando i parametri stimati per il branco di lupi COLP e l'orso femmina F01, si notano differenze comportamentali e spaziali rilevanti tra le due specie. Per quanto riguarda il centro del movimento, rappresentato da μ_1 e μ_2, si osserva che COLP presenta valori di μ_1 sempre negativi e compresi tra

-1,35 e -0,96, mentre F01 ha valori positivi compresi tra 0,99 e 1,26. Questo suggerisce un'occupazione di aree diverse del dominio.
μ_2 è anch'esso sistematicamente negativo per F01, mentre per COLP mostra una transizione da valori negativi a positivi nel corso dell'anno. Questo potrebbe indicare una maggiore dinamicità negli spostamenti nord-sud del branco rispetto all'orso.

Invece θ mostra valori maggiori per F01, compresi tra 0.047 e 0.061, rispetto a COLP, che presenta valori tra 0.028 e 0.039. Questo potrebbe indicare un comportamento più centrato e meno esplorativo dell'orso rispetto al branco di lupi, che potrebbe avere maggiore flessibilità nei propri spostamenti.

Anche σ, che rappresenta la dispersione spaziale del movimento, mostra differenze interessanti. Mentre COLP mantiene valori molto stabili attorno a 0.084 in tutte le stagioni, F01 presenta una maggiore variabilità: da 0.072 a 0.081, con un picco primaverile. Questo suggerisce che l'orso può avere variazioni stagionali più marcate nella scala del proprio movimento.

In sintesi, il modello OU discreto evidenzia differenze coerenti con i comportamenti tipici delle due specie: l'orso femmina F01 mostra un comportamento più centralizzato e stagionalmente variabile, mentre il branco COLP si muove in modo più stabile e distribuito nello spazio, con minore attrazione verso un centro fisso e direzioni di spostamento più dinamiche.

Dati standardizzati

Entrambi i soggetti mostrano un incremento di θ durante l'estate. Tuttavia, l'orso F01 presenta valori più elevati in tutte le stagioni rispetto al branco di lupi COLP, suggerendo una maggiore regolarità nei movimenti individuali rispetto a quelli collettivi. In particolare, F01 mantiene θ intorno a 0.04, mentre COLP varia tra 0.025 e 0.034.
Per entrambi i soggetti, i parametri μ_1 (est) e μ_2 (nord) aumentano durante l'estate, indicando uno spostamento del centro di attività verso nord-est.

L'orso F01 mostra un picco in μ_2 in fine estate, 0.0319, mentre COLP raggiunge un massimo di 0.0504 in autunno, evidenziando un'estensione più marcata verso nord. Anche lo spostamento verso est è più pronunciato per i lupi in autunno, $\mu_1 = 0.0142$, rispetto all'orso, $\mu_1 = 0.0073$.
Entrambi mostrano un aumento di σ^2 durante l'estate, indicativo di una maggiore dispersione spaziale. Il massimo per F01 si osserva in fine estate e raggiunge 0.0846, mentre per COLP il valore più alto è leggermente superiore, 0.0996, suggerendo una maggiore estensione dell'area utilizzata dal branco. In autunno entrambi mostrano una contrazione, evidente soprattutto negli spostamenti dell'orso.

Precisazione

Si sono analizzati i dati sia standardizzati sia non, ma i trend dei parametri sono sempre uguali in quanto viene applicata ad essi una trasformazione lineare. Tuttavia sono state riportate le analisi di entrambi per completezza e per fornire una lettura più interpretabile dal punto di vista geografico, ovvero in scala reale, e una più interpretabile dal punto di vista statistico, in scala standardizzata, più coerente con il modello matematico OU utilizzato.

Conclusioni

Durante tutto l'elaborato, si è cercato di capire come un modello stocastico applicato a dati simulati in diverse condizioni potesse fornire stime accurate e affidabili una volta applicato a dati reali.
Si è quindi adottato un approccio inferenziale e di simulazione per cercare di analizzare il comportamento di animali selvatici.
Le prime difficoltà sono state riscontrate nella costruzione del modello statistico inferenziale, dove è stato cruciale comprendere quali distribuzioni a priori adottare e in seguito con quali combinazioni di parametri il modello potesse fornire stime accurate. L'analisi dei cluster sulle prestazioni ha fornito indicazioni sulle condizioni ottimali e critiche del modello.
Chiaramente il tutto si è complicato nel passaggio da dati simulati a dati reali, ma con alcune accortezze si sono state tratte conclusioni accettabili, seppur minimali, in quanto limitate alla stima dei parametri medi del modello.
Lo studio potrebbe essere esteso e approfondito in molte maniere, ma il tutto si complica molto velocemente.
Il modello potrebbe essere esteso allo studio di più individui e su periodi di tempo più lunghi.
Si potrebbero integrare nel modello le covariate ambientali, quali quelle rese disponibili nei dati reali come la percentuale di aree rocciose prive di vegetazione, distanza da centri abitati e così via.
Inoltre si potrebbe ricorrere all'uso dell'inferenza predittiva per valutare scenari futuri di spostamento o di risposta a cambiamenti ambientali.

Limitatamente a quanto fatto, l'inferenza bayesiana ha comunque consentito di stimare i parametri chiave del comportamento spaziale degli animali presi in considerazione, un orso femmina e un branco di lupi con osservazioni per tutte le stagioni: posizione media del centro di attrazione μ, forza di attrazione verso il centro θ e dispersione σ^2.

L'analisi ha evidenziato differenze comportamentali significative tra le due specie.
L'uso dell'inferenza bayesiana ha offerto numerosi vantaggi, quali l'integrazione di incertezze e variabilità nei dati; stime credibili dei parametri e delle loro distribuzioni posteriori; confronto tra individui, stagioni e specie su basi statisticamente fondate.

Inoltre, la procedura di standardizzazione si è dimostrata cruciale per migliorare la stabilità numerica e l'efficienza del campionamento MCMC, rendendo i risultati interpretabili e robusti anche in presenza di scale spaziali molto diverse.

Il quadro emerso dai risultati consente di trarre indicazioni utili per la comprensione ecologica del comportamento animale.
La variazione stagionale nei parametri suggerisce un adattamento attivo all'ambiente, alle risorse e alle pressioni esterne, principalmente si suppone clima e ricerca di cibo, ma anche antropizzazione.
Le differenze interspecifiche riflettono strategie ecologiche distinte, come il comportamento solitario dell'orso rispetto alla struttura collettiva del branco di lupi.

In conclusione, questa tesi dimostra come un modello stocastico ben calibrato e supportato da inferenza bayesiana possa fornire una rappresentazione coerente, interpretabile e utile dei movimenti animali, unendo rigore statistico e rilevanza ecologica.

Bibliografia

[1] Mevin B. Hooten, Devin S. Johnson, Brett T. McClintock, Juan M. Morales, *Animal Movement: Statistical Models for Telemetry Data*, CRC Press, ISBN 978-1-4665-8214-9.

[2] P.G. Blackwell, *Random Diffusion Models for Animal Movement*, Ecological Modelling, Elsevier.

[3] Brett T. McClintock, Devin S. Johnson, Mevin B. Hooten, Jay M. Ver Hoef, Juan M. Morales, *When to Be Discrete: The Importance of Time Formulation in Understanding Animal Movement*, Movement Ecology, 2014.

[4] Christian P. Robert, George Casella, *Introducing Monte Carlo Methods with R*, Springer, 2009.