POLITECNICO DI TORINO

Corso di Laurea Magistrale in Ingegneria Matematica

Modelli Bayesiani per lo studio dell'interazione tra variabili ambientali e l'attività di Mustela Nivalis

Relatore
Professor Gianluca Mastrantonio

Candidato
Lorenzo Imarisio

Anno Accademico 2024-2025

Introduzione

In un mondo sempre più minacciato dagli effetti del cambiamento climatico [25], l'analisi statistica del comportamento animale si rivela uno strumento sempre più utile per l'ecologia comportamentale e le scienze ambientali.

Le applicazioni spaziano dal monitoraggio della biodiversità alla previsione degli effetti del mutamento delle condizioni ambientali, e sono fondamentali per intraprendere azioni per la conservazione e la gestione delle risorse ambientali.

Studiare il movimento e l'attività animale può essere un procedimento complesso, sia per via della natura schiva delle specie osservate, sia per le difficoltà ambientali presentate dall'habitat in cui esse si trovano. Una tecnica di monitoraggio, applicata per ricavare il dataset utilizzato, consiste nell'utilizzo di radio transmitter (tag) applicati per mezzo di collari sull'animale, con i quali è possibile raccogliere dati relativi al movimento e all'attività dell'animale [18].

La variabile su cui ci concentreremo maggiormente è quella legata al tipo di attività dell'animale, che sarà d'ora in avanti denominata Activity. Questa variabile binaria assume il valore 0 quando l'animale non sta svolgendo attività (l'animale è a riposo), mentre assume valore 1 qualora invece l'animale sia attivo, come durante la caccia. Il tipo di attività viene determinato dalla presenza di segnale intermittente rilevato dal tag. [18]

Gli animali presi in considerazione per questa tesi sono donnole *Mustela nivalis* tracciate nell'area in prossimità del Lago Lungo e del Lago di Ripasottile in un periodo che va da Febbraio 2003 a Giugno 2005.

A partire dai dati forniti, questa tesi si propone di modellizzare le osservazioni in modo da stimare i parametri legati a fattori ambientali che possono influenzare l'attività, quali la stagione, la temperatura, la fase lunare e le condizioni di luce.

L'analisi comincerà con la visualizzazione e l'esplorazione del dataset e si occuperà poi di stimare i parametri del modello tramite un approccio bayesiano su un modello di regressione. Questo tipo di approccio consente di incorporare esplicitamente l'incertezza legata alle stime dei parametri, di integrare eventuali informazioni a priori e di ottenere distribuzioni a posteriori per i parametri del modello. Per fare ciò utilizzeremo STAN e la libreria brms in R, che permettono di analizzare in maniera rigorosa e statisticamente accurata i modelli presi in esame, oltre che a confrontarli per via grafica e per mezzo di altre diagnostiche.

I risultati ottenuti serviranno a fornire un'interpretazione biologica dei

fattori ambientali significativi per la modellizzazione dell'attività animale, contribuendo ad aumentare la comprensione dei pattern di comportamento animale.
Analizzeremo anche, in una sezione dedicata, le motivazioni per cui modelli più complessi come gli Hidden Markov model non sono risultati applicabili ed esploreremo i motivi di tale inapplicabilità.

Indice

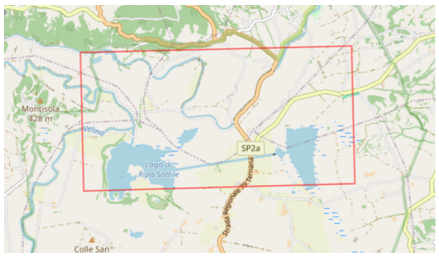
Capitolo 1

Esplorazione e descrizione del dataset

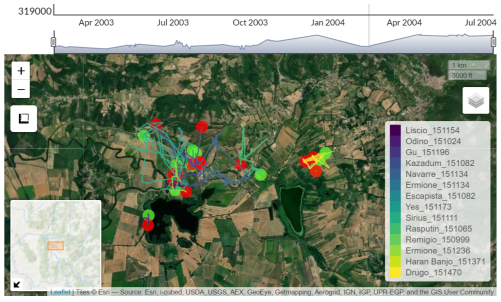
In questo capitolo ci occupiamo di introdurre il dataset utilizzato per la nostra analisi, le variabili che lo compongono e le modalità di raccolte, oltre che ad un'analisi esplorativa dei dati e delle loro proprietà, con lo scopo di fornire una panoramica approfondita delle caratteristiche del dataset.

1.1 Introduzione del dataset

Le osservazioni riportate nel dataset fanno riferimento al tracciamento di donnole *Mustela nivalis* nel Centro Italia. In particolare gli esemplari sono nativi di un'area del Lazio che si estende a nord del Lago Lungo e del Lago di Ripasottile, vicino a Rieti. La *Mustela nivalis*, comunemente nota come donnola (etimologia che deriva dal latino dominula "signorina", diminutivo di domina, "signora")[15], è il più piccolo carnivoro appartenente alla famiglia dei Mustelidi. Dispone di un corpo allungato e snello, con arti corti e una coda anch'essa relativamente breve. La lunghezza totale varia tra 15 e 30 cm, con un peso compreso tra 60 e 250 grammi.[22]
In Figura 1.1a possiamo osservare la zona di tracciamento degli animali, in particolare in 1.1a è rappresentato sulla mappa un riquadro rosso che racchiude al suo interno tutte le coordinate geografiche corrispondenti alle osservazioni all'interno del dataset. Nella sezione 1.1b della figura 1 troviamo invece una visione satellitare dell'area di interesse, che permette di osservare meglio la natura agricola del paesaggio considerato. E' ben evidente di come si tratti di un'area caratterizzata dalla presenza dell'uomo e delle sue attività. In sovraimpressione sono inoltre indicati schematicamente i percorsi seguiti dalle donnole durante il periodo di tracciamento, la legenda in basso a destra permette di associare al diverso colore del



(a) Mappa che raffigura la zona di tracciamento degli animali, corrispondente all'interno del quadrato in rosso



(b) Immagine satellitare della zona di tracciamento degli animali, in legenda il codice identificativo univoco per animale

Figura 1.1

percorso un particolare animale.

Il tracciamento degli animali, con le corrispondenti informazioni relative a posizione e tipo di attività, è stato possibile grazie all'utilizzo di trasmettitori radio applicati ai diversi esemplari. Tramite l'utilizzo di trappole in legno, contenenti come esca topi domestici (*Mus domesticus*) la donnola viene catturata e anestetizzata. In seguito si procede a dotarla di un radio trasmettitore (tag), applicato per mezzo di un collare di raso, e l'animale viene rilasciato dopo un periodo di osservazione di 12 ore, necessario per monitorarne la salute o eventuali reazioni all'anestetico[18]. Una volta rilasciato in natura, è possibile seguire gli spostamenti del soggetto tramite un'antenna direzionale, che permette di rilevare l'intensità del segnale trasmesso dal tag. Questo segnale ha un raggio che può variare dai 20 ai 200 metri, a seconda della vegetazione, pertanto per poter raccogliere i dati è necessario seguire fisicamente l'animale. La tecnica utilizzata per ricavare la posizione precisa dell'esemplare è chiamata "housing" e consiste nel seguire la crescente forza del segnale per determinare la posizione del soggetto[18]. Nonostante questa tecnica non presenti, in termini di misurazione, un errore ad essa associata, è importante notare che la scelta del sistema di mappatura (coordinate UTM chilometriche) ha introdotto un'errore di 10 metri per tutte le misurazioni. Questo errore sistematico all'interno del dataset è stato un fattore determinante nel rendere inapplicabili metodi come gli Hidden Markov Models.

La variabile Activity, che rappresenta la misura su cui ci siamo concentrati in questa analisi, è stata registrata per ogni osservazione determinando la presenza di segnali intermittenti[18]. Si tratta di una variabile binaria, che assume valore zero quando l'animale è considerato inattivo,

ad esempio quando è a riposo, e assume valore 1 quando invece è considerato attivo, come nel caso in cui è impegnato nella caccia.

Le osservazioni sono state raccolte, in generale, registrando la posizione dell'animale ogni 15 minuti per intervalli di 8 ore. Oltre alla sopracitata posizione e allo stato di attività, sono state raccolte informazioni relative alla fase lunare, le condizioni di luce, la stagione e l'habitat; quest'ultimo non è stato considerato nella nostra analisi.

Le osservazioni sono state raccolte in un periodo che va dal febbraio 2003 al giugno 2004, per un totale di 1882 osservazioni.

1.2 Preparazione del dataset

Per procedere all'analisi dei dati e alla definizione del modello, il dataset ha prima subito diversi step di preprocessing per garantirne la qualità e la coerenza. In particolare il dataset originario, fornito dai ricercatori Dott.ssa Caterina Magrini e Dott. Emiliano Manzo, è stato sottoposto a una serie di operazioni di preprocessing con lo scopo di garantire l'integrità dei dati, di arricchirli sia con informazioni derivate sia con dati ambientali provenienti da dataset esterni e di renderli compatibili con le tecniche di modellizzazione che si intendevano adottare, in particolare i modelli lineari generalizzati (GLM) e gli Hidden Markov Models (HMM).

La prima operazione è stata quella di eliminare righe duplicate o contenenti valori NA (not assigned) per informazioni che non potevano essere recuperate, in modo da garantire l'integrità dei dati e non portare in fase di analisi righe con valori mancanti che potrebbero compromettere l'implementazione dei modelli.
Successivamente abbiamo proceduto a uniformare i valori contenuti nella colonna denominata `date`, che si suppone contenere la data e l'orario di ciascuna osservazione. Tuttavia, in alcuni casi l'informazione relativa all'orario risulta mancante, compromettendo l'integrità del dataset. Per ovviare a tale problema, abbiamo proceduto col:

- convertire la colonna in formato $dd/mm/yyyy\ hh:mm$, coerente con lo standard ISO 8601 [14].

- assegnare un orario fisso convenzionale alle osservazioni prive di dati relativi all'orario, preservando in questo modo le altre informazioni relative a quell'osservazione.

Questa operazione ha permesso di mantenere una struttura temporale uniforme, condizione necessaria per il calcolo di successive informazioni derivate e per la corretta applicazione di modelli adottati in seguito.

Gli Hidden Markov Models (HMM) sono ampiamente utilizzati in letteratura per descrivere il comportamento animale a partire da una sequenza di osservazioni temporali[31].

Un HMM è definito come un processo stocastico doppio, composto da:

- una sequenza di stati latenti $\{S_t\}_{t=1}^{T}$, non osservabili direttamente, che evolve nel tempo secondo una catena di Markov
- una sequenza di osservazioni $\{Y_t\}_{t=1}^{T}$, condizionatamente indipendenti dato lo stato latente corrente, generate da distribuzioni dipendenti dallo stato.

La probabilità di transizione tra stati è descritta da una matrice di transizione $\Lambda = [\lambda_{ij}]$, dove $\lambda_{ij} = P(S_{t+1} = j \mid S_t = i)$. La distribuzione iniziale degli stati è invece rappresentata da un vettore $\iota = (\iota_1, \ldots, \iota_N)$, dove $\iota_n = P(S_1 = n)$.[16]

Nel contesto dell'analisi del movimento animale questi modelli permettono di collegare le osservazioni relative alla posizione ad un processo comportamentale non osservabile direttamente con esse, come la caccia o il riposo.

L'utilizzo degli HMM in questo ambito di ricerca offre numerosi vantaggi, tra i quali:

- consente di modellare il comportamento dell'animale e la sua tendenza a mantenere un certo comportamento per più osservazioni consecutive
- fornire una struttura flessibile per incorporare covariate che influenzano la probabilità di transizione tra stati.

Per utilizzare gli HMM è prassi comune in fase di preprocessing ricavare due variabili derivate:

- la step length, che indica la distanza euclidea tra due osservazioni consecutive
- il turning angle, che indica l'angolo tra due vettori di spostamento consecutivi [16]

La *step length*, come anticipato, rappresenta la distanza percorsa tra due osservazioni consecutive. Data una sequenza di posizioni (h_i, v_i), la

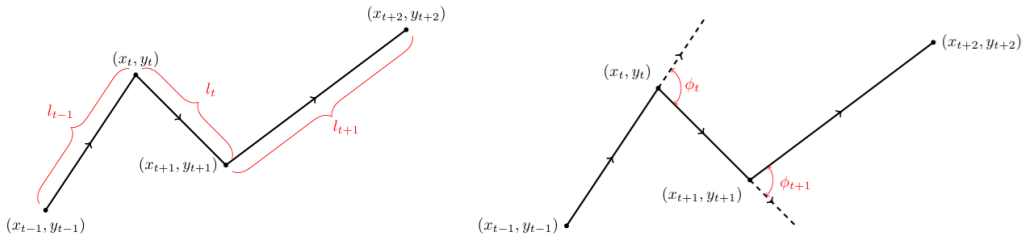


Figura 1.2: Visualizzazione delle variabili step length (l_t) e turning angle

lunghezza del passo tra le osservazioni i e $i+1$ è definita come:

$$l_i = \sqrt{(h_{i+1} - h_i)^2 + (v_{i+1} - v_i)^2}$$

[5] Il *turning angle* quantifica la variazione di direzione tra tre osservazioni consecutive, ovvero l'angolo tra due vettori di spostamento consecutivi. Formalmente, dato il vettore di movimento tra le osservazioni $i-1$ e i, e quello tra le osservazioni i e $i+1$, il turning angle α_i è definito come[6]:

$$\phi_i = \arccos\left(\frac{\vec{m}_{i-1,i} \cdot \vec{m}_{i,i+1}}{\|\vec{m}_{i-1,i}\| \cdot \|\vec{m}_{i,i+1}\|}\right)$$

dove:

- $\vec{m}_{i-1,i} = (h_i - h_{i-1}, v_i - v_{i-1})$
- $\vec{m}_{i,i+1} = (h_{i+1} - h_i, v_{i+1} - v_i)$
- $\cdot$ indica il prodotto scalare
- $\|\cdot\|$ è la norma euclidea.

Il turning angle consente di effettuare una distinzione tra movimenti diretti e movimenti più erratici, fornendo informazioni sullo spostamento dell'animale considerato.

In Figura 1.2 viene riportata a titolo esplicativo un'immagine tratta da [21] che permette di visualizzare queste due variabili.

Per verificare la correttezza dei valori di *step length* e *turning angle*, che sono stati ricavati per mezzo della funzione *prepData*() del pacchetto R *bayesmove*[4], è stata sviluppata una funzione in R che è in grado di ricostruire graficamente il percorso dell'animale a partire da tali variabili. Il percorso così costruito viene sovrapposto a quello originale (ricavato utilizzando invece le posizioni degli animali), permettendo una conferma

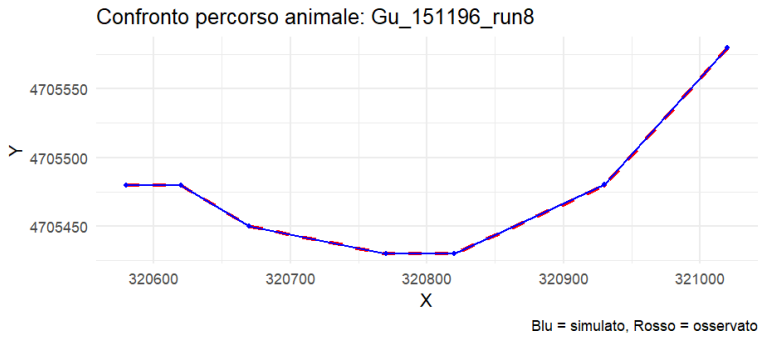


Figura 1.3: Ricostruzione del percorso seguito dall'esemplare identificato come Gu151196, in blu, sovrapposta al percorso ottenuto utilizzando le posizioni dell'animale nel dataset originale, in rosso

visiva della coerenza tra dati derivati e osservazioni reali. Si propone in Figura 1.3 un'immagine illustrativa del lavoro svolto dalla funzione.

A partire dalle coordinate spaziali e temporali, è stata calcolata una variabile derivata: la velocità (integrata nel dataset con la colonna speed), intesa come il rapporto tra la distanza percorsa dall'animale e il tempo impiegato per due osservazioni temporalmente consecutive. Le posizioni geografiche associate con ogni osservazione del dataset sono espresse attraverso il sistema di riferimento EPSG:23033, zona 33N UTM (Universal Transverse Mercator)[7] che comprende l'area di interesse. In questo formato, la posizione geografica di un punto sulla mappa viene definita attraverso la coppia (h, v), dove h rappresenta la coordinata orizzontale (est-ovest) e v la coordinata verticale (nord-sud). Questi due valori rappresentano la distanza della posizione geografica considerata rispetto all'origine del sistema UTM, e trattandosi di coordinate planari sono adatte per calcolare la distanza fra punti utilizzando la distanza euclidea.[23] Formalmente, per due osservazioni consecutive i e $i+1$, la velocità è definita come:

$$s_i = \frac{d_i}{\Delta t_i}$$

dove:

- d_i è la distanza euclidea tra le posizioni (h_i, v_i) e (h_{i+1}, v_{i+1}),
- Δt_i è l'intervallo temporale tra le due osservazioni, misurato considerando la differenza in termini di tempo tra i valori contenuti nella colonna `date`.

L'unità di misura della velocità così ricavata sono dunque i m/s (metri al secondo). Le osservazioni originali erano già ordinate temporalmente in ordine crescente, non è dunque stato necessario riorganizzare le righe del

dataset.
Questa variabile fornisce un'indicazione quantitativa del comportamento dell'animale, utile anche nell'interpretazione biologica per distinguere tra stati attivi e inattivi.

Per arricchire il dataset con ulteriori informazioni legate alle condizioni ambientali, è stata integrata la temperatura rilevata presso la stazione meteorologica di Rieti (Università degli studi di Perugia, centro appenninico del Terminillo "Carlo Jucci"), situata sul Monte Mario, a un'altitudine di 3800 metri s.l.m. I dati coprono l'intervallo di tempo tra gli anni 2003 e 2004 e riportano le temperature giornaliere in tre momenti diversi della giornata: le ore 8:00, 14:00 e 19:00. La temperatura è misurata in gradi centigradi.

A ciascuna osservazione del dataset è stata associata, come da prassi in questi casi, la temperatura più vicina temporalmente all'orario relativo all'osservazione. Nei casi in cui l'orario risultava mancante, che sono stati precedentemente gestiti, si è optato per l'assegnazione della media giornaliera delle tre rilevazioni disponibili.

La colonna Moon del dataset contiene informazioni relative alla fase lunare al momento dell'osservazione. Le categorie originali per questa variabile includono: *plenilunio*, *piena*, *calante*, *crescente* e *nuova*. Poiché lo scopo dell'inclusione di questa variabile è rilevare eventuali effetti dell'illuminazione lunare sul comportamento animale, si è proceduto a ristrutturare le categorie precedenti come segue:

- nuova: mantenuta come categoria autonoma, associata alla fase lunare di luna nuova
- CC: categoria aggregata che comprende *calante* e *crescente*, in quanto queste fasi lunari sono caratterizzate da livelli di illuminazione simili in generale
- piena: che unifica in un'unica categoria le precedenti categorie *plenilunio* e *piena*, poiché indicano la stessa fase lunare di luna piena.

Questa ristrutturazione ha permesso di migliorare l'analisi permettendo ai modelli di focalizzarsi sull'effetto dell'intensità luminosa, piuttosto che sulla nomenclatura specifica delle fasi lunari.

1.3 Esplorazione del dataset

Passiamo ora all'analisi esplorativa del dataset. In questa sezione ci occuperemo di presentare grafici e immagini che illustrano le caratteristiche

Tabella 1.1: *head*() del dataset

ID	**date**	**Moon**	**Season**	**Light**	**Activity**	**tempOK**
Drugo_151470	2004-05-23 00:00	CC	SP	giorno	0	14.23
Drugo_151470	2004-05-24 11:00	CC	SP	giorno	0	15.50
Drugo_151470	2004-05-24 11:15	CC	SP	giorno	0	20.50
Drugo_151470	2004-05-24 11:30	CC	SP	giorno	0	20.50
Drugo_151470	2004-05-24 11:45	CC	SP	giorno	0	20.50
Drugo_151470	2004-05-24 12:00	CC	SP	giorno	0	20.50

h	**v**	**dt**	**step**	**speed**	**angle_clean**
323700	4705970	126000	98.995	0.00079	NA
323630	4705900	900	0.000	0.00000	NA
323630	4705900	900	0.000	0.00000	NA
323630	4705900	900	0.000	0.00000	NA
323630	4705900	900	0.000	0.00000	NA
323630	4705900	900	0.000	0.00000	NA

Tabella 1.2: Tabelle di contingenza per Moon, Light e Season rispetto ad Activity

Variabile	**Livello**	**Activity = 0**	**Activity = 1**
Moon	CC	1292	484
	Nuova	54	25
	Piena	6	7
Light	Alba	38	7
	Giorno	743	407
	Notte	508	84
	Tramonto	64	18
Season	F (Fall)	305	124
	SP (Spring)	286	106
	SU (Summer)	422	189
	W (Winter)	340	97

del dataset, le distribuzioni dei valori e altre statistiche utili dei dati che andremo successivamente ad analizzare.

In Tabella 1.2 si possono osservare le prime righe del dataset considerato per la nostra analisi. Ogni riga caratterizza un'osservazione e le colonne rappresentano le seguenti variabili:

- ID: variabile che identifica univocamente ogni animale
- Moon: variabile categorica che esprime i livelli della fase lunare secondo le aggregazioni che abbiamo implementato in sede di preprocessing dei dati

Tabella 1.3: Conteggio delle osservazioni per ciascun individuo (ID)

ID	**Numero osservazioni**
Drugo_151470	76
Ermione_151134	55
Ermione_151236	315
Escapista_151082	2
Gu_151196	153
Haran Banjo_151371	121
Kazadum_151082	21
Liscio_151154	122
Navarre_151134	270
Odino_151024	24
Rasputin_151065	248
Remigio_150999	51
Sirius_151111	344
Yes_151173	67

- Activity: variabile binaria che indica il livello di attività dell'animale
- Light: variabile categorica che indica i diversi livelli di luce al momento dell'osservazione
- step: variabile continua positiva che indica la dimensione della step length ricavata in fase di preprocessing
- h: variabile che rappresenta la coordinata orizzontale per il sistema di riferimento UTM 33 N
- v: variabile che rappresenta la coordinata verticale per il sistema di riferimento UTM 33 N
- dt: variabile continua positiva, che indica l'intervallo di tempo trascorso, in secondi, tra due osservazioni temporalmente consecutive del dataset
- speed: variabile continua positiva che rappresenta la velocità dell'animale ricavata nella precedente parte di preprocessing
- angle_clean: variabile che rappresenta il turning angle ricavato nella precedente fase di preprocessing
- tempOK: variabile continua che indica la temperatura disponibile più vicina al momento dell'osservazione

Tabella 1.4: Statistiche descrittive per le variabili del dataset

Variabile	**Statistiche**
Variabili categoriche	
ID	Lunghezza: 1868, Classe: character
Moon	CC: 1776, Nuova: 79, Piena: 13
Activity	0: 1352, 1: 516
Season	F: 429, SP: 392, SU: 610, W: 437
Light	Alba: 45, Giorno: 1149, Notte: 592, Tramonto: 82
date	Lunghezza: 1868, Classe: character
Variabili continue	
step	Min: 0.00
	1° Quartile: 0.00
	Mediana: 0.00
	Media: 32.57
	3° Quartile: 0.00
	Max: 1216.22
speed	Min: 0.00000
	1° Quartile: 0.00000
	Mediana: 0.00000
	Media: 0.01738
	3° Quartile: 0.00000
	Max: 0.86781
angle_clean	Min: -3.1183
	1° Quartile: -0.6597
	Mediana: 0.0884
	Media: 0.2827
	3° Quartile: 1.4769
	Max: 3.1416
	NA: 1669
tempOK	Min: -4.90
	1° Quartile: 8.80
	Mediana: 16.50
	Media: 15.13
	3° Quartile: 22.03
	Max: 32.30

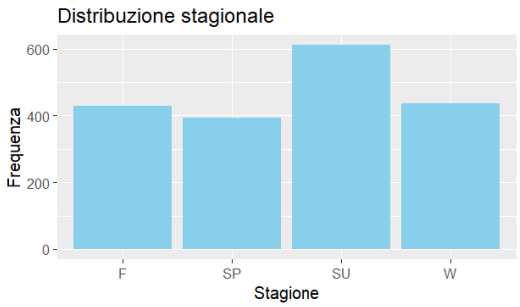
- date: timestamp del momento in cui è stata effettuata l'osservazione

In Tabella 1.4 sono riportate alcune statistiche descrittive iniziali per le colonne del dataset appena descritte.

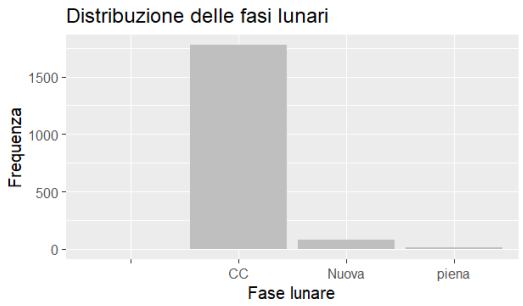
In totale sono stati tracciati 14 esemplari di donnola *Mustela nivalis*, 11 maschi e 3 femmine, per un totale, dopo il preprocessing, di 1882 osservazioni. Il numero di osservazioni per animale è raffigurato in Tabella 1.3.

Sempre dai risultati in Tabella 1.4 possiamo osservare che le variabili categoriche presenti nel dataset sono composte dalle seguenti categorie:

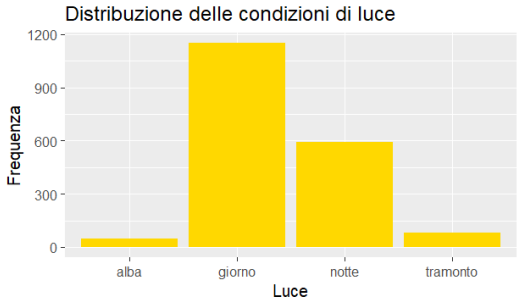
- Variabile Moon: "CC", "piena", "nuova"
- Variabile Light: "alba", "giorno", "notte", "tramonto"
- Variabile Season: "SP" (primavera), "SU" (estate), "W" inverno, "F" autunno



(a) Istogramma per le frequenze all'interno del dataset della variabile Season



(b) Istogramma per le frequenze all'interno del dataset della variabile Moon



(c) Istogramma per le frequenze all'interno del dataset della variabile Light

(d) Istogramma per le frequenze all'interno del dataset della variabile Activity

Figura 1.4: Istogrammi per le frequenze all'interno del dataset delle variabili d'interesse

In Figura 1.4b si possono osservare gli istogrammi che mostrano la distribuzione all'interno del dataset dei livelli delle variabili categoriche e della variabile Activity.
Per quanto riguarda l'istogramma relativo alla variabile Moon, notiamo come sia predominante il numero di osservazioni registrate durante momenti di fase lunare calante o crescente, che ancor prima di essere accorpati

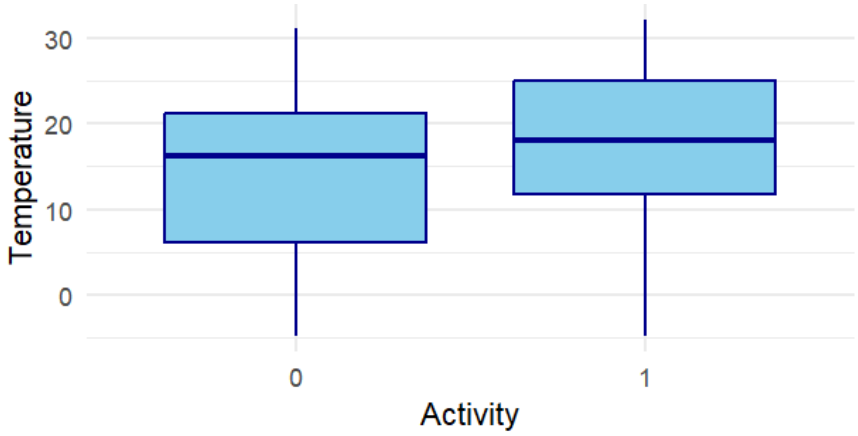


Figura 1.5: Box plot per la distribuzione della temperatura rispetto all'attività

in un'unica categoria rappresentavano una schiacciante maggioranza.
Nell'istogramma relativo alla variabile Light, possiamo anche qui osservare come una categoria, quella della luce diurna, domini per numero di osservazioni rispetto alle altre, seguita dalla notte che ha circa metà del conteggio. Ciò è coerente col fatto che non solo tramonto e alba sono due finestre di tempo molto più ristrette temporalmente, ma anche con il metodo di tracciamento usato, che richiede di seguire in maniera continuativa l'animale per tutta la durata del tracciamento.
La variabile Season rappresenta invece la distribuzione di valori più equa tra le variabili categoriche, anche qui coerentemente col fatto che la fase di raccolta dati si è svolta in un arco di tempo superiore ad un anno.
La variabile Activity presenta invece uno squilibrio di valori verso lo zero, che conta più del doppio delle osservazioni del valore 1.
Nella nostra analisi le variabili categoriche Moon, Light e Season saranno fondamentali per modellizzare la variazione dell'attività animale; in Tabella 1.2 sono riportate le tabelle di contingenza per queste variabili rispetto al livello di Activity.

In Figura 1.5, possiamo osservare invece i box plot relativi alle informazioni sulla temperatura, in particolare la distribuzione della temperatura in funzione del valore dell'attività. Sull'asse delle ascisse sono riportati i due valori possibili per la variabile Activity, 0 e 1, mentre sull'asse delle ordinate la temperatura misurata. Interpretiamo il boxplot ricordando che la linea centrale rappresenta la mediana, le linee inferiori e superiori che definiscono la box rappresentano il primo e terzo quantile, mentre infine i whiskers rappresentano i valori più estremi non ancora considerati come outliers. Si evince dalla Figura 1.5 che per quanto riguarda le osservazioni

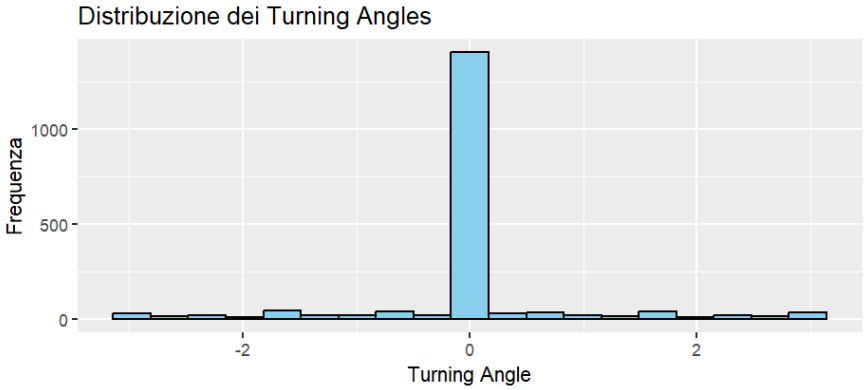


Figura 1.6: Istogramma per i valori del turning angle all'interno del dataset

con attività 1, c'è una tendenza ad avere valori leggermente più alti con temperatura più alta. Questo potrebbe intuitivamente far supporre che ci sia un legame tra temperature più alte e maggior propensione per gli animali ad essere attivi.

Dalla Figura 1.6, è ben evidente come la distribuzione dei turning angle sia caratterizzata da una maggioranza di zeri. Sull'asse delle ascisse sono rappresentati i valori che la variabile può assumere, sull'asse delle ordinate la frequenza di quante volte uno specifico valore compare nel dataset.
Questi zeri sono dovuti al fatto che la funzione *prepData*() del pacchetto *bayesmove*[4] assegna come valore di default lo zero per tutti i turning angle che si trovano tra osservazioni in cui la posizione del soggetto non cambia. Come accennato già nell'introduzione del dataset, la scelta del sistema di mappatura, in particolare l'utilizzo di una mappa in scala 1:10 000 e l'utilizzo delle coordinate UTM chilometriche, ha introdotto un'errore di misurazione pari a 10 metri[18]. Ciò significa che nel dataset tutte le posizioni sono approssimate di questa distanza, e nelle osservazioni per un singolo animale nella maggior parte delle occasioni si traduce nel non avere spostamenti apprezzabili con questa scala di misura. Questo produce numerose osservazioni consecutive in cui l'animale risulta nella stessa posizione geografica, anche se probabilmente si sta muovendo, ma a distanze inferiori al margine di errore. Abbiamo dunque eliminato i valori pari a zero per il turning angle ottenuti dall'assenza di spostamento, rimanendo con soli 199 valori della variabile che non sono NA.

Nelle figure 1.7 e 1.8 sono invece rappresentati i plot relativi alla velocità, variabile da noi ricavata in fase di preprocessing. In Figura 1.8 è rappresentato uno scatterplot dei valori della velocità per le osservazioni

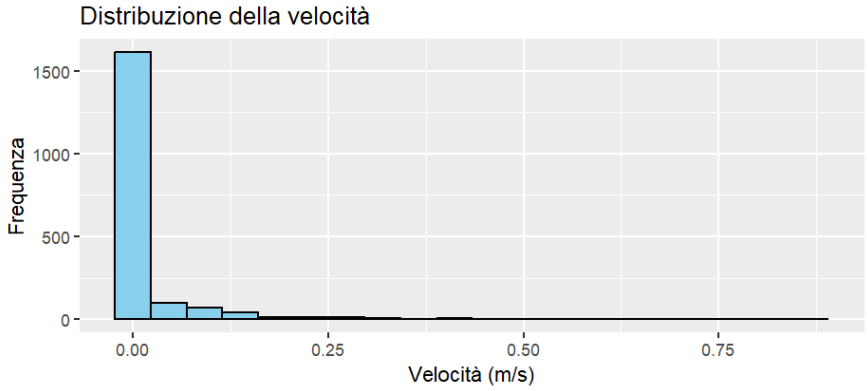


Figura 1.7: Istogramma per la frequenza dei valori della variabile speed all'interno del dataset

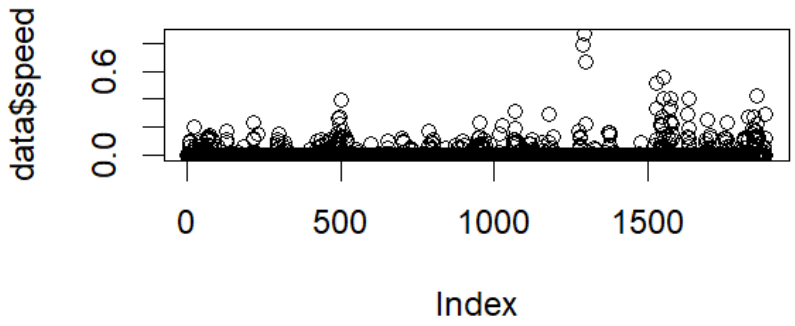


Figura 1.8: Scatterplot per i valori della variabile speed

nel dataset. Sull'asse delle ascisse sono indicati gli indici di tutte le osservazioni nel dataset, mentre sull'asse delle ascisse i relativi valori della velocità. Per le ragioni di cui sopra, chiaramente se l'animale non risulta muoversi per motivi di approssimazione, anche la velocità sarà uguale a zero.
Si può notare bene questo fenomeno in Figura 1.7, un istogramma dove sull'asse delle x sono riportati i possibili valori della velocità, mentre sull'asse delle y è riportata la frequenza all'interno del dataset. Si ha chiaramente un picco in corrispondenza del valore 0 per la velocità, coerentemente con quanto osservato in precedenza.

Analizziamo infine la Figura 1.9, in cui è raffigurato uno scatterplot dove l'asse delle ascisse rappresenta tutti gli indici delle osservazioni del

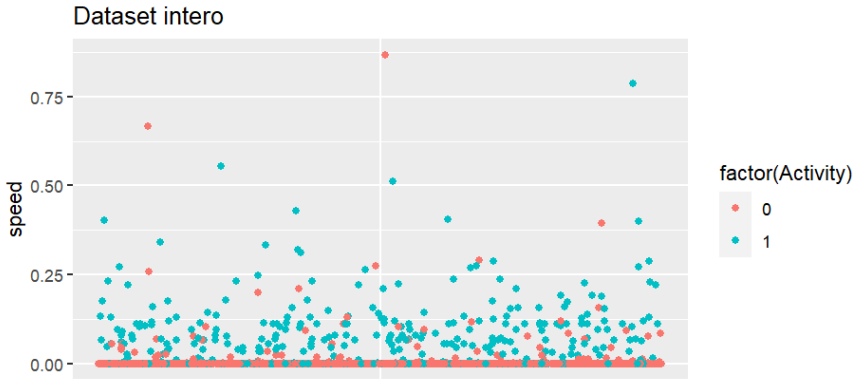


Figura 1.9: Scatterplot per la distribuzione dei valori della velocità a seconda del livello di attività, 0 in rosso e 1 in blu

dataset, mentre l'asse delle ascisse riporta i valori della velocità. I pallini all'interno dello scatter-plot sono colorati a seconda del valore di Activity corrispondente per l'osservazione che rappresentano. Si osserva come è ben evidente la tendenza ad avere velocità maggiori di zero quando l'animale è attivo. Tale forte ed evidente relazione tra l'attività (Activity $= 1$) e la velocità non è solo intuitiva, ma anche ben conosciuta dai ricercatori, motivo per cui nei modelli per analizzare l'attività non è stata inserita tra le covariate.

Capitolo 2

Approccio Bayesiano

In questo capitolo introduciamo gli strumenti teorici che costituiscono le fondamenta dei modelli utilizzati. Introdurremo per prima cosa i modelli lineari generalizzati, spiegando la motivazione della scelta di questo particolare modello. Illustreremo poi l'approccio bayesiano, utilizzato per ricavare le stime dei parametri del modello e infine forniremo un quadro teorico degli algoritmi che R utilizza per effettuare il campionamento.

2.1 Modelli GLM

I modelli lineari generalizzati, conosciuti con l'acronimo di GLM, rappresentano una generalizzazione dei classici modelli di regressione lineare. Mentre questi ultimi gestiscono variabili dipendenti distribuite secondo una distribuzione normale, i modelli GLM allargano l'area di interesse a tutte le variabili aleatorie con distribuzioni che appartengono alla famiglia esponenziale.
Le distribuzioni di probabilità appartenenti alla famiglia delle distribuzioni esponenziali sono quelle distribuzioni che, data una variabile aleatoria n, possono essere scritte come

$$distr(n, \psi, \upsilon) = exp(\frac{n\psi - b(\psi)}{a(\upsilon)} + c(n, \upsilon))$$

dove ψ è chiamato il parametro naturale e υ è chiamato parametro di dispersione.[20]

Tra le distribuzioni di probabilità appartenenti alla famiglia esponenziale, consideriamo la distribuzione di Bernoulli, che prende il nome dall'omonimo matematico svizzero. La distribuzione di Bernoulli è una distribuzione di probabilità discreta che descrive l'esito di un esperimento binario, ovvero che può ammettere solo due risultati possibili: successo (1)

o insuccesso (0). Questa distribuzione è parametrizzata dal solo parametro $\pi \in [0,1]$, che rappresenta la probabilità di successo dell'esperimento. La funzione di massa di probabilità di questa distribuzione è definita come:

$$f(y;\pi) = \begin{cases} \pi \text{ se } y = 1 \\ 1 - \pi \text{ se } y = 0 \end{cases} \qquad y \in \{0,1\}$$

Dove y è la realizzazione della variabile aleatoria.
In alternativa, può essere espressa in modo più compatto utilizzando:

$$f(y;\pi) = \pi^y (1-\pi)^{1-y} \quad \text{per } y \in \{0,1\}$$

La distribuzione di Bernoulli può essere anche considerata come un caso particolare della distribuzione binomiale, ottenuto quando il numero di prove $n = 1$. In tal caso, la variabile aleatoria che segue la distribuzione di Bernoulli rappresenta il numero di successi in una singola prova binomiale.[34]

Questa distribuzione è particolarmente utile nel contesto dei GLM per modellare variabili binarie, come nel caso preso in esame in cui si vuole analizzare come varia la probabilità di osservare la variabile Activity con valore 1 in relazione al variare delle altre variabili indipendenti presenti nel dataset.

Consideriamo dunque Activity come la variabile aleatoria y_i che segue una distribuzione di Bernoulli con parametro π_i che indica la probabilità di osservare il valore 1 per la realizzazione della variabile aleatoria.

$$y_i \sim Bernoulli(\pi_i)$$

dove i rappresenta l'i-esima osservazione

Lo scopo di questi tipi di modello è quello di modellizzare questa variabile y_i, chiamata anche variabile indipendente, per mezzo di una combinazione lineare di altre variabili, dette variabili dipendenti o covariate che si suppone possano influenzarne la distribuzione.
Per definire un modello GLM sono dunque necessarie tre componenti principali:

- la distribuzione della variabile dipendente y_i (Activity), nel caso in esame una Bernoulli. Esattamente come la variabile Activity per descrivere se l'animale è attivo o inattivo permette solo due esiti, 0 e 1, anche la Bernoulli ha lo stesso comportamento
- il predittore lineare η che rappresenta una combinazione lineare delle variabili dipendenti

- la funzione link, che come suggerisce il nome collega il valore atteso di y_i alla combinazione lineare delle variabili dipendenti. Nel caso in esame si tratta di una funzione *logit*.

Il valore atteso $\mu_i = E[y_i]$ della variabile dipendente y_i è dunque collegato al predittore lineare η_i tramite una funzione di collegamento $g(\cdot)$, chiamata link function, tale che si verifica:

$$g(\mu_i) = \eta_i$$

Formalizzando η_i come combinazione lineare delle covariate x_{ij} tramite la formula

$$\eta_i = \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij}$$

Dove β_j corrisponde ai parametri associati ai diversi livelli delle covariate. Il parametro β_0 rappresenta invece l'intercetta, ovvero è il parametro associato ai livelli di riferimento delle covariate. [20]

Si possono dunque unire le considerazioni precedenti e modellizzare la probabilità di successo π_i come funzione delle covariate tramite la funzione logit:

$$\mu_i = E[Y_i] = \pi_i, \quad logit(\pi_i) = \eta_i = \beta_0 + \sum_{j=1}^{P} \beta_j x_{ij}$$

Alternativamente, possiamo esprimere π_i tramite la funzione sigmoide, inversa della logit:

$$\pi_i = \frac{1}{1 + e^{-\eta_i}}$$

Questa rappresentazione sarà utile in fase di interpretazione per computare facilmente la probabilità dell'individuo di essere attivo dati i coefficienti stimati del modello utilizzando approcci Bayesiani.

2.2 Approccio Bayesiano ai GLM

Nell'approccio statistico Bayesiano, la probabilità viene interpretata come grado di fiducia che accada un determinato evento piuttosto che frequenza relativa di un evento. Al contrario dell'approccio frequentistico, in statistica bayesiana anche i parametri d'interesse sconosciuti sono considerabili variabili aleatorie. I parametri del modello β non sono dunque considerati fissi, ma come variabili aleatorie dotate di una distribuzione a priori $P(\beta)$. L'obiettivo in questa analisi è di aggiornare questa distribuzione a

partire dai dati osservati D, tramite il Teorema di Bayes, su cui si fonda la statistica Bayesiana[20].

$$P(\beta|D) = \frac{P(D|\beta) \cdot P(\beta)}{P(D)} \propto \underbrace{p(D \mid \beta)}_{\text{likelihood}} \times \underbrace{\pi(\beta)}_{\text{prior}} .$$

Dove:

- $P(\beta)$: distribuzione a priori dei parametri, che può essere non informativa (ad esempio una distribuzione uniforme o una distribuzione normale con varianza elevata) oppure informativa, se si dispone di conoscenze pregresse.
- $P(D|\beta)$: verosimiglianza o likelihood, ovvero la probabilità di osservare i dati D dato un certo valore dei parametri β. Questo termine valuta quanto bene i parametri spiegano i dati osservati e contribuisce ad aggiornare la posteriori.
- $P(\beta|D)$: distribuzione a posteriori, che rappresenta la conoscenza aggiornata sui parametri dopo aver osservato i dati. Questa distribuzione dunque contiene l'informazione relativa ai parametri ottenuta combinando i dati e le conoscenze a priori.
- $P(D)$: termine di normalizzazione, spesso ignorato in fase di inferenza.

Questo approccio consente di incorporare l'incertezza sui parametri e di ottenere intervalli credibili, oltre a facilitare la modellizzazione gerarchica e l'inclusione di conoscenze pregresse, date ad esempio da analisi passate.

Nel nostro caso in analisi, attraverso la conoscenza delle stime dei parametri per mezzo di questo approccio, si possono dunque utilizzare quest'ultime per calcolare i valori di π_i per varie combinazioni di covariate e stimare così le probabilità per la variabile indipendente, oltre a valutare per mezzo dei parametri, quali sono le covariate significative e la tipologia della loro influenza sulla variabile dipendente y_i.

Nel contesto dell'inferenza bayesiana, l'obiettivo principale è, come anticipato, ottenere la distribuzione a posteriori dei parametri $P(\beta|D)$ del modello, ovvero la distribuzione condizionata dei parametri rispetto ai dati osservati.
Raramente è però possibile ottenere questa distribuzione in forma chiusa, ed è per questo motivo che sono stati sviluppati i metodi Markov Chain

Monte Carlo, i quali sfruttano l'intuizione che non è necessario dover conoscere la distribuzione a posteriori in forma chiusa, ma è solo necessario ottenere dei campioni da questa distribuzione per poter fare inferenza.
La prima delle due parti fondamentali delle MCMC è quindi quella dei metodi Montecarlo, che si basano appunto sull'idea di approssimare le quantità di interesse tramite campioni casuali estratti da una distribuzione di probabilità. Se, ad esempio, vogliamo calcolare un valore atteso

$$E_q[r(\mathbf{x})] = \int q(x)r(x)\,dx$$

dove $q(x)$ è una distribuzione di probabilità e $r(x)$ una funzione d'interesse[24], questo può essere approssimato come media empirica dei campioni generati dalla distribuzione target

$$\frac{1}{n}\sum_{i=1}^{n} r(\mathbf{x}^{(i)})$$

Questa idea costituisce la parte centrale dell'aspetto Monte Carlo dell'MCMC: invece di calcolare analiticamente l'integrale che definisce l'attesa, esso viene stimato tramite la media dei campioni ottenuti dalla catena di Markov.

L'altra idea alla base dell'MCMC è costruire una catena di Markov che abbia come distribuzione stazionaria proprio la distribuzione a posteriori $P(\beta \mid D)$. Si definisce una catena di Markov (del primo ordine) una sequenza di variabili aleatorie $\beta^{(1)}, \beta^{(2)}, \ldots, \beta^{(T)}$ tale che:

$$P(\beta^{(t)} \mid \beta^{(1)}, \ldots, \beta^{(t-1)}) = P(\beta^{(t)} \mid \beta^{(t-1)})$$

Questa proprietà indica che per determinare la distribuzione del t-esimo campione, è necessario solo conoscere la distribuzione al tempo successivo. Costruendo la catena di Markov sulla sequenza di campioni consente di generarli in maniera iterativa, che dopo un certo numero di iterazioni raggiungono la convergenza alla distribuzione stazionaria della catena di Markov. I campioni ricavati in questo modo possono essere poi utilizzati per stimare medie, varianze, intervalli credibili e altre quantità di interesse nel modello in esame[26].

Tra i tipi di algoritmi MCMC, menzioniamo brevemente il Gibbs-Sampler e l'algoritmo Metropolis-Hastings, che sono entrambi algoritmi che generano campioni della distribuzione a posteriori. Il primo sfrutta la full-conditional e campiona da essa, mentre il secondo viene impiegato quando ciò non è possibile[20].

2.3 Hamiltonian Monte Carlo

Per implementare i modelli GLM in ambiente R è stata utilizzata la libreria *brms*[2], che permette di costruire modelli bayesiani complessi usando STAN per l'interferenza.

Il metodo di campionamento che viene utilizzato in questo caso è l'Hamiltonian Monte Carlo (HMC), un algoritmo di campionamento avanzato che utilizza la funzione Hamiltoniana per esplorare in maniera più efficiente lo spazio dei campioni e con bassa autocorrelazione tra di essi.

Consideriamo uno spazio dei parametri di dimensione T, $(\beta_1, \beta_2, ...\beta_T)$, il metodo Hamiltonian Monte Carlo utilizza T variabili addizionali $(m_1, m_2, ..., m_T)$ dette momenti, e definisce una funzione Hamiltoniana

$$H(\boldsymbol{\beta}, \boldsymbol{m}) = U(\boldsymbol{\beta}) + K(\boldsymbol{m})$$

dove:

- $U(\beta) = -\log S(\beta)$ è l'energia potenziale, con $S(\beta)$ proporzionale alla densità a posteriori del modello;
- $K(m) = \frac{1}{2} m^\top M^{-1} m$ è l'energia cinetica, con $m \sim \mathcal{N}(0, M)$.

Le equazioni di Hamilton che descrivono la regolarità del sistema sono

$$\frac{d\boldsymbol{\beta}}{dt} = \frac{\partial H}{\partial \boldsymbol{m}}, \quad \frac{d\boldsymbol{m}}{dt} = -\frac{\partial H}{\partial \boldsymbol{\beta}}$$

Queste equazioni sono integrate numericamente (tipicamente con il metodo *Leapfrog*) per simulare traiettorie nello spazio delle fasi (β, m).

Le proprietà fondamentali della dinamica hamiltoniana che rendono HMC efficace in questo tipo di applicazioni sono le seguenti:

- reversibilità: la traiettoria può essere invertita cambiando il segno del momento $m \to -m$
- conservazione dell'energia: il valore di $H(\beta, m)$ rimane costante lungo la traiettoria
- conservazione del volume: il volume nello spazio delle fasi è preservato.

Il metodo utilizzato per integrare le equazioni di Hamilton è simmetrico, con errore locale $\mathcal{O}(\epsilon^3)$ e errore globale $\mathcal{O}(\epsilon^2)$, dove ϵ è il passo di integrazione. La scelta dei parametri ϵ e L (numero di passi) è cruciale per bilanciare precisione e esplorazione dello spazio dei parametri. [12]

2.4 Modelli GLMM

Un'estensione dei modelli GLM è rappresentata dai Generalized Linear Mixed Models (GLMM), anche chiamati in italiano Modelli Lineari Misti. Questo tipo di modelli consente di modellare dati che hanno al loro interno dei gruppi tra loro "dipendenti".
Un esempio di questa situazione è il caso in esame, dove le osservazioni che compongono il dataset fanno riferimento ad un numero limitato di individui, ed è dunque legittimo aspettarsi che alcuni esemplari siano più propensi ad avere un comportamento più attivo nell'intervallo di tempo in cui sono stati osservati.
I modelli GLMM tengono conto di questo includendo oltre agli effetti fissi già presenti nei modelli GLM, anche degli effetti casuali, che non sono uguali per tutti i livelli delle covariate, ma assumono valori diversi a seconda di come scegliamo di modellizzarli.
Un modello lineare misto, analogamente ai modelli GLM, assume che la variabile risposta y_i segua una distribuzione appartenente alla famiglia esponenziale e che la sua media condizionata sia collegata linearmente alle covariate tramite una link function $g(\cdot)$. La formula generale [20] [28] è:

$$g(E[y_{ij} \mid u_i]) = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \mathbf{z}_{ij}^T \mathbf{u_i}$$

dove:

- $\mathbf{x}_{ij} \in R^p$ è il vettore delle covariate associate agli effetti fissi per l'osservazione j del gruppo i;
- $\boldsymbol{\beta} \in R^p$ è il vettore dei coefficienti degli effetti fissi;
- $\mathbf{z}_{ij} \in R^q$ è il vettore delle covariate associate agli effetti casuali per l'osservazione j del gruppo i;
- $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma_i})$ è il vettore degli effetti casuali specifici del gruppo i, distribuito secondo una normale multivariata con media nulla e matrice di covarianza Σ.

Questa struttura consente di modellare la variabilità tra gruppi, migliorando la flessibilità e la capacità predittiva del modello rispetto ai GLM.

Nel contesto dei GLMM, gli effetti fissi rappresentano il contributo sistematico delle covariate, che viene assunto costante per tutte le osservazioni. Gli effetti casuali, invece, modellano la variabilità non spiegata dagli effetti fissi, assumendo che tali effetti siano estratti da una distribuzione di probabilità.

L'inclusione di effetti casuali consente, in particolare, di:

- modellare la dipendenza tra osservazioni appartenenti allo stesso gruppo
- evitare la sovrastima della significatività degli effetti fissi.

Un caso particolarmente comune nei GLMM è quello dell'inclusione di un'intercetta casuale (random intercept), in cui si assume che ciascun gruppo abbia una propria deviazione rispetto all'intercetta globale. Questo è utile, ad esempio, quando si osservano più misurazioni per ciascun individuo, come nel caso preso in esame.

Il modello assume la forma[20]:

$$\mu_i = \mathbf{x}_i\beta + u_i\mathbf{1}_{n_i}$$

alternativamente, in versione matriciale

$$\mu = X\beta + Zu$$

dove:

- la colonna i-esima di Z assume valore 1 per le osservazioni appartenenti al gruppo i e zero altrove
- $U_i \sim N(0, \sigma_u^2)$
- $z_i = \mathbf{1}_{n_i}$

L'aggiunta di una singola intercetta può consentire di modellare la variabilità tra gruppi, assumendo che ciascun gruppo abbia un proprio livello di riferimento, ma che tali livelli siano distribuiti attorno a una media comune.

L'inclusione di un'intercetta casuale presenta numerosi vantaggi:

- controllo per la variabilità non osservata: consente di tenere conto di fattori latenti che influenzano la risposta, ma non sono esplicitamente modellati attraverso gli effetti fissi
- inferenza più robusta: migliora la stima degli effetti fissi, riducendo il rischio di confondere effetti sistematici con variazioni casuali tra gruppi.

Nel contesto bayesiano, i GLMM vengono stimati specificando distribuzioni a priori sui parametri $\boldsymbol{\beta}$ e σ_u, e ottenendo le distribuzioni a posteriori tramite metodi di campionamento come l'Hamiltonian Monte Carlo (HMC) descritto in precedenza. Questo approccio consente di:

- incorporare conoscenze pregresse tramite le distribuzioni a priori

- ottenere stime complete dell'incertezza tramite le distribuzioni a posteriori
- modellare strutture complesse in modo flessibile e coerente.

Capitolo 3

Modelli utilizzati

In questo capitolo introduciamo i modelli utilizzati per modellizzare la variabile Activity rispetto alle variabili d'interesse all'interno del dataset, ne definiamo la formulazione matematica e includiamo alcune nozioni teoriche sugli strumenti utilizzati per confrontare e costruire i modelli.

3.1 Modelli GLM

In questa sezione vengono descritti i modelli statistici GLM implementati per analizzare l'attività in funzione delle altre variabili rilevanti del dataset. Tutti i modelli sono stati stimati con approccio bayesiano tramite regressione logistica, assumendo una distribuzione Bernoulli per la variabile risposta y_i e utilizzando la funzione link *logit*.

Le variabili considerate rilevanti per la nostra analisi sono le seguenti:

- Activity: variabile binaria per rappresentare l'attività (0 = inattivo, 1 = attivo)
- Moon: variabile categorica con livelli `CC` (crescente e calante), `Nuova`, `Piena`, che rappresenta la fase lunare, in particolare il livello di illuminazione ad essa associata
- Season: variabile categorica con livelli `W` (inverno), `SP` (primavera), `SU` (estate), `F` (autunno), che rappresenta la stagione
- Light: variabile categorica con livelli `alba`, `giorno`, `tramonto`, `notte`, che rappresenta il momento della giornata, in particolare la quantità di luce ad essa associata
- tempOK: variabile continua che rappresenta la temperatura ambientale al momento dell'osservazione

- ID: identificativo individuale, utilizzato come effetto casuale nei modelli gerarchici per modellizzare la differenza tra individui non espressa dagli effetti fissi.

Sono state testate diverse combinazioni delle covariate, per verificare la loro influenza sulla variabile Activity e confrontare la performance dei diversi modelli. A seguire presentiamo la formulazione matematica di 4 TRA I modelli utilizzati.

3.1.1 Primo modello GLM, effetti principali

Per il primo modello (model_activity2) utilizziamo una semplice combinazione degli effetti principali delle covariate Season, Moon, Light e tempOK (temperatura).
Per legare il valore atteso della variabile dipendente y_i al predittore lineare η_i viene utilizzata, coerentemente alla teoria introdotta in precedenza, la funzione $logit()$. Il predittore lineare η_i, che rappresenta la combinazione lineare delle covariate contiene anche l'intercetta β_0, la quale rappresenta il valore di η_i quando tutti i valori delle covariate categoriche sono ai livelli di riferimento e la temperatura è zero.

Modellizziamo dunque la probabilità di osservare $Activity = 1$, che scriviamo come $\pi_i = P(y_i = 1)$ in funzione del predittore lineare η_i

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

dove il predittore lineare, per questo modello, rappresenta la combinazione degli effetti principali delle covariate selezionate.

$$\eta_i = \beta_0 + \sum_{j=2}^{3} \beta_{\text{Moon}_j} \cdot \text{Moon}_{ij} + \sum_{k=2}^{4} \beta_{\text{Season}_k} \cdot \text{Season}_{ik} + \sum_{l=2}^{4} \beta_{\text{Light}_l} \cdot \text{Light}_{il} + \beta_{\text{tempOK}} \cdot \text{tempOK}_i$$

dove:

- β_0 è l'intercetta del modello, che rappresenta il log-odds di attività nel caso dei livelli di riferimento
- Moon_{ij} sono variabili dummy che codificano i livelli 2 e 3 della fase lunare (rispettivamente "nuova" e "piena"), con "CC" come livello di riferimento

- Season_{ik} sono variabili dummy che codificano le stagioni 2 (primavera), 3 (estate) e 4 (autunno), con Season1 (inverno) come riferimento
- Light_{il} rappresentano i livelli di illuminazione ambientale (Light2–4) (rispettivamente "giorno", "tramonto" e "notte"), con Light1 (alba) come livello riferimento.
- tempOK_i è una variabile continua.

In particolare $Moon_{ij}$, $Season_{ik}$ e $Light_{il}$ assumono il valore 1 se l'i-esima osservazione è nella j-esima/k-esima/l-esima categoria.
L'indice delle variabili categoriali nelle sommatorie parte da 2 poiché il primo livello viene considerato come riferimento (corner), come anticipato in precedenza.

3.1.2 Secondo modello GLM, inclusione del termine di interazione tra stagione e temperatura

Per il secondo modello (model_activity3) utilizziamo la combinazione degli effetti principali delle covariate Season, Moon, Light e tempOK (temperatura), inoltre si aggiunge anche un termine di interazione tra la variabile categorica Season e la variabile continua tempOK (temperatura). Questo termine di interazione permette al modello di valutare come l'effetto della temperatura cambia con il variare della stagione.
Anche in questo caso, come anticipato, per legare il valore atteso della variabile dipendente y_i al predittore lineare η_i viene utilizzata la funzione $logit()$.

Modellizziamo dunque la probabilità di osservare $Activity = 1$, che scriviamo come $\pi_i = P(y_i = 1)$ in funzione del predittore lineare η_i

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

dove il predittore lineare, per questo modello, rappresenta la combinazione degli effetti principali delle covariate selezionate più il termine di interazione tra le covariate Season e tempOK (temperatura).

$$\eta_i = \beta_0 + \sum_{j=2}^{3} \beta_{\text{Moon}_j} \cdot \text{Moon}_{ij} + \sum_{k=2}^{4} \beta_{\text{Season}_k} \cdot \text{Season}_{ik} + \sum_{l=2}^{4} \beta_{\text{Light}_l} \cdot \text{Light}_{il}$$
$$+ \beta_{\text{tempOK}} \cdot \text{tempOK}_i \quad + \sum_{k=2}^{4} \beta_{\text{Season}_k:tempOK_i} \cdot \text{Season}_{ik} \cdot \text{tempOK}_i$$

3.1.3 Terzo modello GLM, inclusione del termine di interazione tra luce e temperatura

Nel terzo modello considerato (model_activity4.5) utilizziamo, oltre alla combinazione degli effetti principali delle covariate considerate in precedenza e un termine di interazione tra la variabile categorica Season e la variabile continua tempOK (temperatura), un nuovo termine di interazione tra le variabili. Questo termine di interazione aggiuntivo mette in relazione la variabile categorica Light con la variabile continua tempOK, e permette al modello di valutare come l'effetto della temperatura cambia con il variare delle condizioni di luce.
Anche in questo caso, come anticipato, per legare il valore atteso della variabile dipendente y_i al predittore lineare η_i viene utilizzata la funzione $logit()$.

Modellizziamo dunque la probabilità di osservare $Activity = 1$, che scriviamo come $\pi_i = P(y_i = 1)$ in funzione del predittore lineare η_i

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \tag{3.1}$$

dove il predittore lineare, per questo modello, rappresenta la combinazione degli effetti principali delle covariate selezionate più i termini di interazione tra le covariate Season e tempOK (temperatura) e le covariate Light e tempOK.

$$\begin{aligned}\eta_i = \beta_0 &+ \sum_{j=2}^{3} \beta_{\text{Moon}_j} \cdot \text{Moon}_{ij} + \sum_{k=2}^{4} \beta_{\text{Season}_k} \cdot \text{Season}_{ik} + \sum_{l=2}^{4} \beta_{\text{Light}_l} \cdot \text{Light}_{il} \\ &+ \beta_{\text{tempOK}} \cdot \text{tempOK}_i \quad + \sum_{k=2}^{4} \beta_{\text{Season}_k:tempOK_i} \cdot \text{Season}_{ik} \cdot \text{tempOK}_i \\ &+ \sum_{l=2}^{4} \beta_{\text{Light}_l} \cdot \text{Light}_{il} \cdot \text{tempOK}_i \end{aligned} \tag{3.2}$$

3.1.4 Quarto modello GLM, inclusione dell'interazione tra luce e temperatura

L'ultimo modello che presentiamo per i modelli GLM utilizzati (model_activity5), in aggiunta agli effetti principali delle covariate precedenti include un termine di interazione tra la variabile categorica Light e la variabile continua tempOK, oltre ad un termine di interazione tra le variabili categoriche Season e Light. Quest' ultimo termine di interazione è stato aggiunto per

cercare di comprendere se la luce ha un effetto diverso a seconda della stagione.

Modellizziamo dunque la probabilità di osservare $Activity = 1$, che scriviamo come $\pi_i = P(y_i = 1)$ in funzione del predittore lineare η_i

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

dove il predittore lineare, per questo modello, rappresenta la combinazione degli effetti principali delle covariate selezionate più i termini di interazione tra le covariate Season e tempOK (temperatura) e le covariate Season e Light.

$$\begin{aligned}\eta_i = \beta_0 &+ \sum_{k=2}^{4} \beta_{\text{Season}_k} \cdot \text{Season}_{ik} + \sum_{l=2}^{4} \beta_{\text{Light}_l} \cdot \text{Light}_{il} + \sum_{j=2}^{3} \beta_{\text{Moon}_j} \cdot \text{Moon}_{ij} \\ &+ \beta_{\text{tempOK}} \cdot \text{tempOK}_i \quad + \sum_{k=2}^{4} \beta_{\text{Season}_k:\text{tempOK}} \cdot \text{Season}_{ik} \cdot \text{tempOK}_i \\ &+ \sum_{k=2}^{4}\sum_{l=2}^{4} \beta_{\text{Season}_k:\text{Light}_l} \cdot \text{Season}_{ik} \cdot \text{Light}_{il}\end{aligned}$$

3.2 Modelli GLMM

In questa sezione vengono descritti i modelli statistici GLMM implementati per analizzare l'attività in funzione delle altre variabili rilevanti del dataset. A differenza dei modelli precedenti, quelli in questa sezione includono sia effetti fissi sia effetti casuali, con lo scopo di tenere conto della presenza di osservazioni ripetute per ciascun esemplare considerato. Tutti i modelli sono stati stimati con approccio bayesiano tramite regressione logistica, assumendo una distribuzione Bernoulli per la variabile risposta y_i e utilizzando la funzione link *logit*. L'inclusione di effetti casuali consente di modellare la variabilità tra i soggetti non spiegata dagli effetti fissi utilizzati in precedenza, con l'obiettivo di migliorare la capacità predittiva e fornire una migliore interpretazione dei risultati.

3.2.1 Primo modello GLMM

Per il primo modello GLMM sfruttiamo una semplice combinazione degli effetti principali delle covariate Season, Moon, Light e tempOK (temperatura) analogamente a quanto effettuato per il primo modello GLM.

Viene però aggiunta un'intercetta random u_{ID_i} distribuita secondo una distribuzione normale $u_{ID_i} \sim N(0, \tau^2)$, dove a sua volta il termine τ è distribuito secondo una distribuzione Inverse Gamma $\tau \sim InvGamma(\alpha, \beta)$
Per legare il valore atteso della variabile dipendente y_i al predittore lineare η_i viene utilizzata, coerentemente alla teoria introdotta in precedenza, la funzione $logit()$. Il predittore lineare η_i, che rappresenta la combinazione lineare delle covariate contiene anche l'intercetta β_0, la quale rappresenta il valore di η_i quando tutti i valori delle covariate categoriche sono ai livelli di riferimento e la temperatura è zero. Il termine di intercetta random u_{ID_i} è anch'esso inserito nella combinazione lineare delle covariate.
Per tutti i modelli GLMM implementati i livelli di riferimento sono analoghi a quelli dei modelli GLM.
Modellizziamo dunque la probabilità di osservare $Activity = 1$, che scriviamo come $\pi_i = P(y_i = 1)$ in funzione del predittore lineare η_i

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

dove il predittore lineare, per questo modello, rappresenta la combinazione degli effetti principali delle covariate selezionate più l'intercetta random u_{ID_i}.

$$\eta_i = \beta_0 + \sum_{j=2}^{3} \beta_{\text{Moon}_j} \cdot \text{Moon}_{ij} + \sum_{k=2}^{4} \beta_{\text{Season}_k} \cdot \text{Season}_{ik} + \sum_{l=2}^{4} \beta_{\text{Light}_l} \cdot \text{Light}_{il} + \beta_{\text{tempOK}} \cdot \text{tempOK}_i + u_{\text{ID}_i}$$

dove $u_{\text{ID}_i} \sim \mathcal{N}(0, \tau^2)$ è il termine casuale associato all'individuo ID_i, che cattura la variabilità non spiegata tra soggetti.

Per l'effetto casuale associato all'identificatore dell'individuo ID, è stato specificato un prior di tipo Inverse Gamma:

$$\tau \sim \text{Inv-Gamma}(2, 1)$$

3.2.2 Secondo modello GLMM

Per il secondo GLMM, analogamente a quanto effettuato per il primo modello GLMM, andiamo ad aggiungere un'intercetta casuale oltre alle covariate utilizzate in un modello GLM precedente.
Alle covariate utilizzate per il modello GLM model_activity3 viene dunque aggiunta un'intercetta random u_{ID_i} distribuita secondo una distribuzione normale $u_{ID_i} \sim N(0, \tau^2)$, dove a sua volta il termine τ è distribuito secondo una distribuzione Inverse Gamma $\tau \sim InvGamma(\alpha, \beta)$. Per

legare il valore atteso della variabile dipendente y_i al predittore lineare η_i utilizziamo la funzione $logit()$.

Modellizziamo dunque la probabilità di osservare $Activity = 1$, che scriviamo come $\pi_i = P(y_i = 1)$ in funzione del predittore lineare η_i

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i$$

dove il predittore lineare, per questo modello, rappresenta la combinazione degli effetti principali delle covariate selezionate più il termine di interazione tra temperatura e stagione e, infine, l'intercetta random u_{ID_i}.

$$\eta_i = \beta_0 + \sum_{j=2}^{3} \beta_{\text{Moon}_j} \cdot \text{Moon}_{ij} + \sum_{k=2}^{4} \beta_{\text{Season}_k} \cdot \text{Season}_{ik} + \sum_{l=2}^{4} \beta_{\text{Light}_l} \cdot \text{Light}_{il}$$
$$+ \beta_{\text{tempOK}} \cdot \text{tempOK}_i + \sum_{k=2}^{4} \beta_{\text{Season}_k:\text{tempOK}} \cdot \text{Season}_{ik} \cdot \text{tempOK}_i + u_{\text{ID}_i}$$

Per l'effetto casuale associato all'identificatore dell'individuo ID, è stato specificata una prior uguale al modello GLMM precedente.

3.2.3 Terzo modello GLMM

Per il terzo modello GLMM viene utilizzata una combinazione degli effetti principali delle covariate Season, Moon, Light e tempOK (temperatura), a cui si aggiungono le interazioni tra stagione e temperatura e quelle tra luce e temperatura, oltre a un effetto casuale per l'ID (che identifica l'animale). Abbiamo dunque aggiunto alle stesse covariate utilizzate nel modello GLM model_activity4.5, un'intercetta random che varia a seconda dell'esemplare, con lo scopo di valutare se il singolo animale considerato ha una tendenza ad essere più o meno attivo, non catturata per gli effetti fissi comuni a tutti gli individui analizzati.
Come nei modelli precedenti, per legare il valore atteso della variabile dipendente y_i al predittore lineare η_i viene utilizzata la funzione $logit()$.

Modellizziamo dunque la probabilità di osservare $Activity = 1$, che scriviamo come $\pi_i = P(y_i = 1)$, in funzione del predittore lineare η_i:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i \tag{3.3}$$

dove il predittore lineare, per questo modello, è definito come:

$$\eta_i = \beta_0 + \sum_{j=2}^{3} \beta_{\text{Moon}_j} \cdot \text{Moon}_{ij} + \sum_{k=2}^{4} \beta_{\text{Season}_k} \cdot \text{Season}_{ik} + \sum_{l=2}^{4} \beta_{\text{Light}_l} \cdot \text{Light}_{il}$$
$$+ \beta_{\text{tempOK}} \cdot \text{tempOK}_i \quad + \sum_{k=2}^{4} \beta_{\text{Season}_k:\text{tempOK}} \cdot \text{Season}_{ik} \cdot \text{tempOK}_i +$$
$$\sum_{l=2}^{4} \beta_{\text{Light}_l:\text{tempOK}} \cdot \text{Light}_{il} \cdot \text{tempOK}_i + u_{\text{ID}_i} \tag{3.4}$$

3.3 Strumenti utilizzati per l'approccio Bayesiano

Tutti i modelli sono stati poi implementati seguendo un approccio bayesiano, questo per fare inferenza sui parametri campionando dalla distribuzione a posteriori dei parametri, in modo da sfruttare una combinazione dell'informazione a priori che si ha a disposizione e l'evidenza fornita dai dati osservati.
Nel caso di studio, se non diversamente specificato, per ciascun parametro si è assunta una distribuzione a priori debole o non informativa, in modo da lasciare che sia l'evidenza dei dati a determinare i risultati.

Per stimare la distribuzione a posteriori dei parametri, è stato impiegato il metodo Monte Carlo Markov Chain (MCMC) in particolare l'algoritmo Hamiltonian Monte Carlo (HMC) nella versione NUTS. Il No-U-Turn Sampler (NUTS)[13], come suggerisce il nome, è una variante adattiva di HMC che si ferma autonomamente quando l'esplorazione dello spazio dei parametri inizia a tornare verso il punto di partenza. Inoltre determina automaticamente la lunghezza ottimale delle traiettorie simulate (leapfrog steps).

Un aspetto particolarmente importante di brms è la sua capacità di generare automaticamente il codice Stan[27] sottostante a partire dalla formula del modello, traducendo sintassi di alto livello. Questo consente allo user di focalizzarsi sulla struttura teorica del modello e sull'interpretazione dei risultati, lasciando a Stan le operazioni di campionamento e ottimizzazione numerica. In Figura 3.1 [2] includiamo uno schema high level della procedura utilizzata da *brms*.

Tra le caratteristiche principali della libreria citiamo[3]:

- la possibilità di specificare modelli gerarchici e multilevel, includendo effetti casuali (random effects)

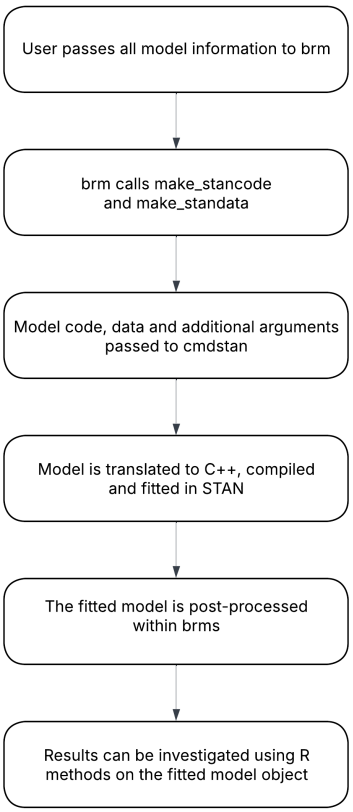


Figura 3.1: Descrizione ad alto livello della procedura di stima del modello utilizzata da *brms*

- l'uso di priori informative, specificabili per ogni parametro del modello (ad esempio, per le varianze delle intercette casuali o per i coefficienti della regressione)
- la compatibilità con diversi backend di calcolo, tra cui `rstan` e `cmdstanr`, quest'ultimo caratterizzato da maggiore efficienza e stabilità computazionale
- Un'ampia serie di funzioni per l'analisi delle distribuzioni a posteriori.

3.3.1 Specificazione del terzo modello GLM

In questa analisi, la libreria brms è stata utilizzata per stimare la probabilità di osservare il valore della variabile binaria Activity pari a 1 in funzione di variabili rappresentanti fattori ambientali specificati nelle sezioni precedenti. Il modello che ha meglio performato secondo l'indice WAIC[33] è stato definito in R come segue:

```
model_activity4.5 <- brm(
  Activity ~ Moon + Season + Light + tempOK +
             Season:tempOK + Light:tempOK,
  data = data,
  family = bernoulli(link = "logit"),
  chains = 4, iter = 8000, warmup = 4000,
  cores = 4, seed = 1234,
  backend = "cmdstanr",
  file = "model_activity4.5",
  save_pars = save_pars(all = TRUE)
)
```

La formula del modello implementa la formulazione matematica del modello, in cui la variabile dipendente y_i (Activity) viene influenzata dalle covariate Moon + Season + Light + tempOK + Season:tempOK + Light:tempOK (dove ":" rappresenta il termine di interazione tra le covariate interessate). L'utilizzo della funzione logit come link function e la distribuzione della variabile dipendente y_i sono state specificate nelle sezione `family = bernoulli(link = "logit")`.

Il modello è stato stimato mediante quattro catene MCMC indipendenti, ciascuna composta da 8.000 iterazioni totali, di cui 4.000 destinate al *warm-up* (fase di adattamento), per un totale di 16.000 campioni a posteriori. L'opzione `save_pars(all = TRUE)` consente di memorizzare tutti i parametri del modello, facilitando successive analisi e visualizzazioni.

3.3.2 Estensione GLMM del terzo modello, inclusione degli effetti casuali

La libreria *brms* consente anche di estendere il modello GLM precedente introducendo effetti casuali, come nel caso illustrato a seguire, dove si aggiunge un termine di intercetta casuale per la variabile che identifica l'individuo.

```
model_activity_conID <- brm(
  Activity ~ Moon + Season + Light + tempOK +
             Season:tempOK + Light:tempOK + (1 | ID),
  data = data,
  family = bernoulli(link = "logit"),
  prior = prior,
  chains = 4, iter = 8000, warmup = 4000,
  cores = 4, seed = 1234,
```

```
  backend = "cmdstanr",
  file = "model_activity_conID",
  save_pars = save_pars(all = TRUE)
)
```

In questa scrittura del modello sono ancora valide le considerazioni precedenti per il modello GLMM. In addizione citiamo il termine `(1 | ID)` che introduce l'intercetta random specificata nella formulazione matematica per il terzo modello GLMM presentato nella rispettiva sezione. La funzione `inv_gamma(2, 1)` (come specificato nella formulazione matematica precedente) viene utilizzata come una prior debole, ma regolarizzante per prevenire stime spurie dovute a campioni di dimensione limitata, come nel caso del nostro dataset. Questa scelta di prior riflette la credenza iniziale che ci possa essere una variabilità tra individui, ma permette comunque all'inferenza dei dati di dimostrare il contrario qualora questo sia supportato.

Come vedremo nella sezione dedicata ai risultati, quest'ultimo modello è stato il migliore in termini di valutazione utilizzando il criterio WAIC.

Capitolo 4

Risultati

In questo capitolo si espongono i risultati del fitting dei modelli GLM e GLMM introdotti nei capitoli precedenti. Inoltre si mettono a confronto le prestazioni dei modelli per mezzo del criterio WAIC, di cui si fornisce un'introduzione teorica. Infine forniamo un'interpretazione dei risultati dei modelli migliori.

4.1 Il criterio WAIC

Il criterio WAIC, anche detto Watanabe-Akaike information criterium, è un'estensione del classico criterio AIC[1], entrambi utilizzati per la selezione di modelli tipici del contesto di inferenza bayesiana.

L'idea alla base di questo criterio è di correggere stime ottimistiche di altre misure che valutano l'accuratezza predittiva di un modello, come la log pointwise predictive density (lppd) attraverso la sottrazione di un termine di penalizzazione che è basato sulla complessità del modello.[9]

WAIC parte computando la lppd e aggiunge il termine di correzione che dipende dal numero effettivo di parametri per correggere per l'overfitting.

Si tratta quindi di un approccio per stimare la out-of-sample expectation, ovvero la capacità del modello di predire nuovi dati. La *log pointwise predictive density* (lppd) viene computata in pratica come[9]:

$$\text{lppd} \approx \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i \mid \theta^{(s)}) \right)$$

dove chiamiamo le simulazioni della distribuzione a posteriori $\theta^s, s = 1, ..., S$ e $p(y_i \mid \theta^s)$ è la funzione di verosimiglianza. Siccome lppd è una

Tabella 4.1: Confronto dei modelli tramite WAIC

Modello	**ELPD diff**	**SE diff**
model_activity4.5	0.0	0.0
model_activity5	-0.4	2.4
model_activity3	-9.7	5.5
model_activity2	-14.0	7.0

misura che tende a sovrastimare la capacità predittiva del modello, per correggere questo bias, WAIC introduce, come anticipato, una penalizzazione basata sul numero effettivo di parametri. Per la funzione *waic*() del pacchetto R *loo*[8] questa ha la forma:

$$p_{\text{WAIC}} = \sum_{i=1}^{n} \text{Var}_{\theta} \left[\log p(y_i \mid \theta) \right]$$

Questa quantità rappresenta la somma delle varianze, calcolate punto per punto, della log-verosimiglianza rispetto alla distribuzione a posteriori. In questo modo più il modello cambia in base ai parametri, maggiore sarà la varianza e di conseguenza anche il termine di penalizzazione.

Si definisce dunque il WAIC come:

$$\text{WAIC} = -2 \left(\text{lppd} - p_{\text{WAIC}} \right)$$

Per interpretare quale modello è migliore rispetto ad un altro, basta osservare il valore di WAIC più basso, il quale indica una migliore capacità predittiva corretta per la complessità del modello[9].

4.2 Risultati per il miglior modello GLM

Per il criterio WAIC, il GLM migliore è il modello model_activity4.5, presentato nel capitolo precedente. Si può osservare ciò nella tabella 4.1, dove il modello migliore è elencato per primo e fa da riferimento per gli altri che lo seguono. Nella tabella 1 sono presenti, oltre alle denominazioni dei modelli, anche altre due colonne: nella prima è riportata l'*elpd* diff (Expected log predictive density difference) e nella seconda l'SD diff. L'Expected log predictive density difference è una misura che aiuta a determinare quanto bene predice il modello e viene stimata da WAIC. Maggiore è questo valore, migliore è il modello nelle previsioni out-of-sample, dunque i modelli che hanno nella tabella un valore negativo sono peggiori rispetto al modello di riferimento, ed *elpd* quantifica questa differenza. SD diff è invece l'errore standard associato alla differenza di *elpd* tra i modelli[30].

Tabella 4.2: Stime dei coefficienti di regressione del modello bayesiano (GLM).

Parametro	**Est.**	**Est. Error**	**l-95% CI**	**u-95% CI**	**Rhat**	**Bulk ESS**	**Tail ESS**
Intercept	-1.58	0.56	-2.73	-0.57	1.00	5479	8096
Moon2	0.37	0.28	-0.19	0.92	1.00	20099	11942
Moon3	1.17	0.60	-0.00	2.38	1.00	20966	11729
Season2	-1.26	0.58	-2.44	-0.16	1.00	10166	10564
Season3	-0.35	0.57	-1.46	0.75	1.00	10615	10636
Season4	-0.20	0.30	-0.78	0.38	1.00	11503	12149
Light2	0.40	0.59	-0.68	1.62	1.00	5309	8013
Light3	0.89	0.75	-0.59	2.38	1.00	6364	9134
Light4	0.25	0.58	-0.82	1.44	1.00	5431	8011
tempOK	-0.06	0.06	-0.18	0.05	1.00	4900	8028
Season2:tempOK	0.08	0.04	0.00	0.16	1.00	7482	9529
Season3:tempOK	0.03	0.03	-0.03	0.10	1.00	7446	9979
Season4:tempOK	0.06	0.03	0.00	0.12	1.00	7948	10571
Light2:tempOK	0.07	0.06	-0.03	0.18	1.00	4569	7631
Light3:tempOK	-0.01	0.06	-0.12	0.12	1.00	5141	8204
Light4:tempOK	-0.01	0.06	-0.12	0.10	1.00	4892	7547

A seguire, in Tabella 4.2 vediamo le stime dei parametri β del modello GLM, ottenuti con il campionamento dalla distribuzione a posteriori per mezzo dell'algoritmo MCMC Hamiltonian Monte Carlo:

Per i risultati relativi ai coefficienti della regressione, presentati in Tabella 4.2, possiamo notare 8 diverse colonne:

- colonna per i nomi dei parametri β, questa colonna riporta i nomi dei livelli delle covariate a cui sono associati i coefficienti stimati
- Est. (Estimate): colonna per i valori stimati dei parametri; nel contesto della funzione link logit questo valore indica l'effetto additivo della covariata sulle log-odds per la probabilità di osservare Activity = 1
- Est. Error: contiene il valore che indica l'errore standard per la stima dei parametri
- l-95% CI: rappresenta il limite inferiore dell'intervallo di credibilità al 95%
- u-95% CI: rappresenta il limite superiore dell'intervallo di credibilità al 95%

- Rhat: indica il Potential Scale Reduction, una misura diagnostica per verificare la convergenza delle catene MCMC
- Bulk_ESS: mostra il valore dell'Effective Sample Size (ESS) nella parte centrale della distribuzione a posteriori. Valori elevati di questo parametro suggeriscono una buona efficienza del campionamento e stime robuste[32]
- Tail_ESS: Effective Sample Size nelle code della distribuzione a posteriori. Misura la qualità del campionamento nelle regioni estreme della distribuzione. Un valore elevato garantisce che anche le stime degli intervalli di credibilità siano affidabili, in quanto anche le code sono ben esplorate dal campionamento[32].

L'indice diagnostico Rhat [27] [10] viene definito come:

$$\hat{R} = \sqrt{\frac{\widehat{\text{var}}^{+}(\beta \mid D)}{W}}$$

dove:

- β è il parametro
- D sono i dati osservati
- W è la varianza within-chain, considerata come media di tutte le catene. E' necessaria per misurare quanto le catene differiscono al loro interno
- $\widehat{\text{var}}^{+}(\beta \mid D)$ rappresenta invece il variance estimator e valuta la varianza totale del parametro β condizionata ai dati D, combinando la varianza tra le catene e la varianza all'interno delle catene.

Segue dunque dalla definizione che quando il parametro Rhat è vicino o pari a uno significa che le catene di Markov sono arrivate a convergenza. Ciò significa che il metodo MCMC sta campionando i valori del parametro d'interesse in modo affidabile dalla sua distribuzione a posteriori[32].

Nel nostro caso di studio il valore di Rhat è esattamente 1, ciò indica che le catene hanno molto probabilmente raggiunto la convergenza: i campioni generati possono essere considerati rappresentativi della distribuzione a posteriori dei parametri d'interesse e le stime inferenziali ottenute sono per questo ritenute affidabili.

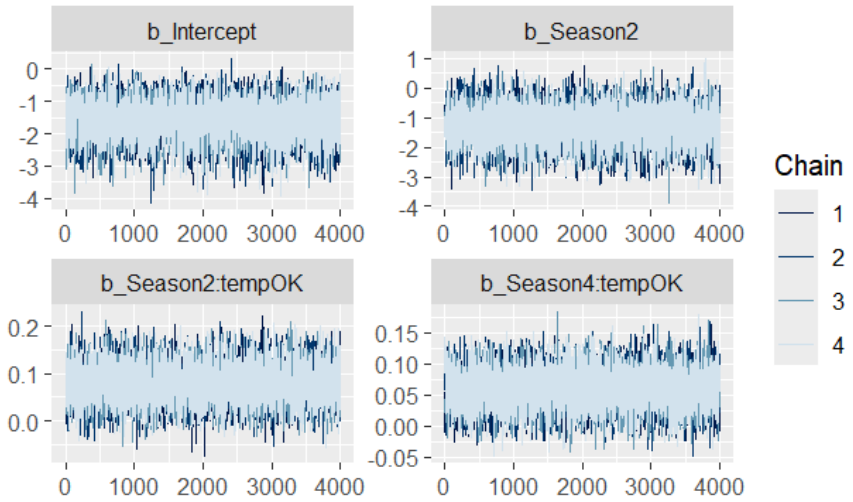


Figura 4.1: Traceplot delle catene MCMC per i parametri stimati significativi del modello GLM

In Figura 4.1 sono riportati i traceplot per le catene MCMC utilizzate in quest' analisi.

L'intervallo di credibilità[19], in questo caso al 95%, è un riferimento importante per il nostro studio, perchè fornisce un'informazione su quali parametri del modello siano significativi. Questo intervallo indica, con un probabilità al 95%, che il valore stimato del parametro β d'interesse risiede al suo interno. Definiamo dunque come "significativi" quei parametri il cui intervallo di credibilità esclude il valore zero, in quanto altrimenti significherebbe che esiste una probabilità non trascurabile che il valore stimato del parametro sia zero e dunque il livello della covariata ad esso associato non ha effetto sulla variazione della probabilità di osservare Activity = 1

Con lo scopo di analizzare ora le stima dei parametri significativi, ricordiamo i livelli di riferimento. Questi sono rappresentati dall'intercetta β_0, che corrisponde con il valore di η_i quando tutti i valori delle covariate categoriche sono ai livelli di riferimento e la temperatura è zero. I livelli di riferimento per questo modello sono:

- covariata Moon: "CC" (luna crescente/calante)
- covariata Light: "alba"
- covariata Season: "W" (inverno).

Al fine di comprendere l'output del modello contenuto in Tabella 4.2, riportiamo i corrispondenti livelli delle covariate associati alle stime dei parametri

- **Moon**
  - `Moon2` → livello `nuova`
  - `Moon3` → livello `piena`
- **Light**
  - `Light2` → livello `giorno`
  - `Light3` → livello `tramonto`
  - `Light4` → livello `notte`
- **Season**
  - `Season2` → livello `SP` (primavera)
  - `Season3` → livello `SU` (estate)
  - `Season4` → livello `F` (autunno)

Considerando gli intervalli di credibilità riportati in tabella 4.2, le stime dei parametri che possiamo ritenere statisticamente significative sono le seguenti:

- l'intercetta, con valore stimato -1.58
- Season 2, che corrisponde al livello "primavera" della variabile Season, con valore -1.26
- Season2:tempOK, che corrisponde all'influenza della temperatura durante la primavera, con valore 0.08
- Season4:tempOK, che corrisponde all'influenza della temperatura durante l'autunno, con valore 0.06.

L'output delle stime dei coefficienti tende ad approssimare per ragioni di spazio i valori inseriti nelle tabelle, per cui nonostante compaia il valore 0.00 per i lower bound dell'intervallo di credibilità al 95%, questo numero è appunto un'approssimazione dei valori

- 0.001658 per Season2:tempOK
- 0.001878 per Season4:tempOK

che quindi escludono lo zero e sono considerabili significativi.

Per stime dei parametri come $Moon3$ e $Season3$: $tempOK$ si può ritenere plausibile l'effetto sulla probabilità, in quanto lo zero è marginalmente incluso nell'intervallo di credibilità, ma questo effetto non può ovviamente essere ritenuto certo.

Ricordiamo la formulazione matematica del modello model_activity4.5 in 3.1 e 3.2. Possiamo dunque esprimere π_i tramite la funzione sigmoide, inversa della logit:

$$\pi_i = \frac{1}{1 + e^{-\eta_i}}$$

Sfruttiamo questa formulazione matematica per fornire l'interpretazione dei valori delle stime. Siccome l'intercetta rappresenta il valore delle log-odds nel caso di riferimento, il valore indicato nella colonna Estimate rappresenta la variazione per quel coefficiente nei log-odds per Activity=1, quando i livelli delle altre covariate rimangono costanti. Dunque un valore di Estimate positivo significa un aumento della probabilità di osservare Activity=1, mentre, al contrario, un valore negativo rappresenta al contrario una diminuzione della stessa probabilità.

Per le stime significative dei parametri, otteniamo dunque i seguenti effetti:

- Season 2, che corrisponde al livello "primavera" della variabile Season, causa una forte diminuzione della probabilità di osservare l'animale in attività (rispetto all'inverno)
- Season2:tempOK, che corrisponde all'influenza della temperatura durante la primavera: aumenta leggermente la probabilità con l'aumentare della temperatura in primavera
- Season4:tempOK, che corrisponde all'influenza della temperatura durante l'autunno: aumenta leggermente la probabilità con l'aumentare della temperatura in autunno.

In Tabella 4.8 sono riportati i valori delle stime dei parametri considerati significativi, a titolo esplicativo riportiamo quindi i valori di π_i in alcuni casi specifici.

Nel caso in cui consideriamo solo l'intercetta β_0, η_i corrisponde a -1.58, pertanto

$$\pi_i = \frac{1}{1 + e^{1.58}} = 0.1708$$

Tabella 4.3: Parametri significativi del modello GLM%.

Parametro	**Estimate**	**l-95% CI**	**u-95% CI**
Intercept	-1.58	-2.73	-0.57
Season2	-1.26	-2.44	-0.16
Season2:tempOK	0.08	0.00	0.16
Season4:tempOK	0.06	0.00	0.12

Ovvero nelle condizioni di riferimento la probabilità di osservare l'animale in stato di attività sono circa del 17%.

Mantenendo costanti tutte le covariate e passando dal livello di riferimento per Season (inverno) alla primavera, il parametro η_i diventa

$$\eta_i = -1.58 - 1.26 = -2.84$$

E si ricava

$$\pi_i = \frac{1}{1 + e^{2.84}} = 0.0552$$

Ovvero questo cambiamento di condizioni ha fatto scendere la probabilià di osservare l'animale in attività al 5%, in questa particolare situazione.

In Figura 4.4, osserviamo invece una heatmap che riflette come il passare dal livello di una covariata ad un altro modifichi le log-odds di Activity = 1, per avere una panoramica degli effetti causati dalla variazione dei parametri ambientali. Nelle colonne dell'heatmap abbiamo il livello di "partenza" (ad esempio W (inverno)) mentre nelle righe quello di "arrivo" (ad esempio F (autunno)). Il numero riportato all'interno della cella rappresenta la variazione, se negativo (colore rosso) influenza negativamente la probabilità di osservare Activity = 1, al contrario se positivo (colore blu).

4.3 Risultati per il miglior modello GLMM

Anche per i GLMM, il criterio WAIC indica come migliore il modello model_activity_conID, riportato in Tabella 4.7 e presentato come terzo nel capitolo precedente. Si tratta infatti della versione meglio performante dei modelli GLM con l'aggiunta di un intercetta random per l'individuo. Si può osservare la prestazione del modello per WAIC nella tabella 4.7, la cui interpretazione è analoga alla Tabella 4.2 precedente

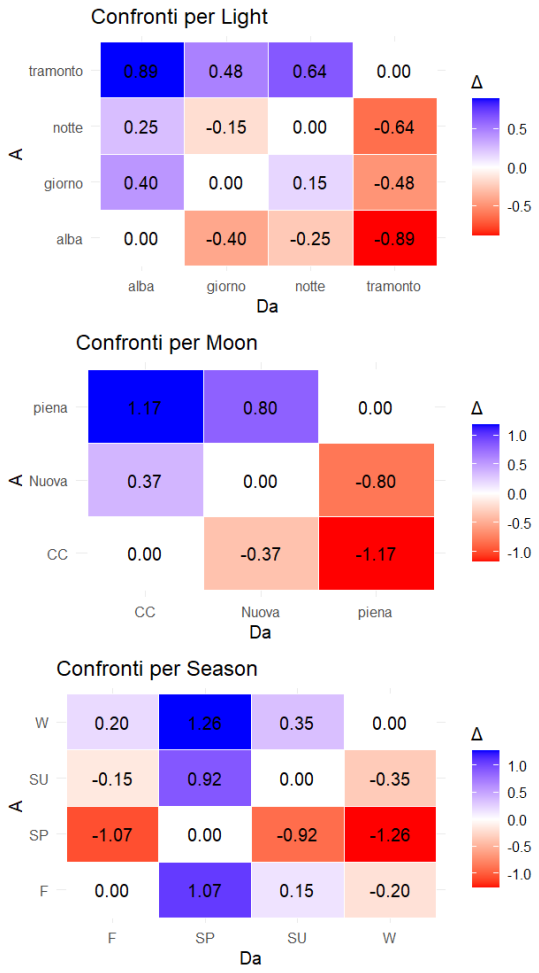


Figura 4.2: Confronto per le variazioni di log-odds di Activity per i vari livelli delle covariate Light, Moon e Season

A seguire, in Tabella 4.4, vediamo le stime dei parametri β del modello GLMM migliore, ottenute con il campionamento dalla distribuzione a posteriori per mezzo dell'algoritmo MCMC Hamiltonian Monte Carlo.

Per i risultati relativi ai coefficienti della regressione, presentati in Tabella 4.4, l'interpretazione dei valori delle colonne è analoga a quella presentata nella sezione precedente per la tabella 4.2, a cui si aggiunge però un altro elemento importante. Trattandosi di un modello GLMM viene anche riportata la deviazione standard dell'intercetta rispetto ai livelli di ID (che identifica univocamente l'animale).

Distinguiamo tra

- β_0:l'intercetta media del modello

Tabella 4.4: Risultati del modello GLMM con intercetta random per individuo

Parametro	**Estimate**	**Est. Error**	**l-95% CI**	**l-95% CI**	**Rhat**	**Bulk ESS**	**Tail ESS**
Deviazione standard dell'intercetta							
sd(Intercept)	0.96	0.27	0.55	1.61	1.00	6066	8965
Coefficienti di regressione							
Intercept	-1.79	0.91	-3.64	-0.04	1.00	6939	8746
Moon2	0.66	0.30	0.08	1.23	1.00	20571	12405
Moon3	1.44	0.65	0.17	2.72	1.00	18039	11844
Season2	-0.44	1.08	-2.54	1.74	1.00	8392	9217
Season3	-0.15	1.05	-2.27	1.84	1.00	7087	8263
Season4	-0.97	0.96	-2.92	0.90	1.00	7160	8409
Light2	0.47	0.59	-0.62	1.69	1.00	7328	7864
Light3	1.03	0.76	-0.43	2.56	1.00	8954	8952
Light4	0.30	0.58	-0.75	1.51	1.00	7373	8449
tempOK	-0.06	0.06	-0.17	0.05	1.00	6565	8806
Season2:tempOK	0.07	0.04	-0.01	0.16	1.00	12654	11438
Season3:tempOK	0.01	0.04	-0.06	0.08	1.00	12570	10594
Season4:tempOK	0.08	0.04	0.01	0.15	1.00	12190	12168
Light2:tempOK	0.07	0.05	-0.03	0.19	1.00	5963	8805
Light3:tempOK	-0.01	0.06	-0.13	0.11	1.00	6535	8286
Light4:tempOK	-0.02	0.05	-0.12	0.10	1.00	5965	8534

- u_{ID_i}:l'intercetta specifica per il livello i del gruppo `ID`
- τ è la deviazione standard dell'intercetta casuale.

Questo ci consente di modellare la variabilità non spiegata tra i diversi esemplari rispetto alla probabilità di osservare Activity = 1. Nello specifico, ogni livello di ID ha una propria intercetta, che riflette la propensione individuale dell'animale all'attività, in maniera indipendente dalle altre covariate. Siccome in questo caso τ corrisponde al valore 0.96, con intervallo di credibilità che esclude lo zero, abbiamo ragione di supporre che sia presente una variabilità significativa tra le intercette associate ai vari individui e che l'intercetta casuale sia importante ai fini del modello. In Tabella 4.5 sono riportate le stime dell'intercetta casuale per ogni individuo, mentre in Tabella 4.6 quelle significative. Si può notare in quest'ultima come alcuni degli individui abbiamo una tendenza personale all'attività rispetto alla media dei soggetti.

Anche in questa implementazione il valore di Rhat è 1, quindi nuovamente ciò indica che le catene hanno molto probabilmente raggiunto la

Tabella 4.5: Stime dell'intercetta per ciascun livello del fattore `ID`.

ID	Stima	Errore Std	CI 2.5%	CI 97.5%
Drugo_151470	-0.86	0.52	-1.97	0.08
Ermione_151134	-0.53	0.48	-1.47	0.47
Ermione_151236	0.26	0.72	-1.20	1.71
Escapista_151082	-0.49	0.90	-2.44	1.17
Gu_151196	-0.73	0.48	-1.77	0.14
HaranBanjo_151371	-0.92	0.51	-2.00	0.01
Kazadum_151082	1.63	0.63	0.47	2.95
Liscio_151154	-0.29	0.72	-1.77	1.15
Navarre_151134	0.82	0.44	0.00	1.75
Odino_151024	-0.05	0.62	-1.33	1.11
Rasputin_151065	1.05	0.48	0.16	2.05
Remigio_150999	0.03	0.43	-0.77	0.94
Sirius_151111	0.63	0.48	-0.26	1.65
Yes_151173	0.63	0.48	-0.26	1.65

Tabella 4.6: Stime significative dell'intercetta per livelli del fattore `ID`.

ID	Stima	CI 2.5%	CI 97.5%
Kazadum_151082	1.63	0.47	2.95
Navarre_151134	0.82	0.00	1.75
Rasputin_151065	1.05	0.16	2.05

convergenza: i campioni generati possono essere considerati rappresentativi della distribuzione a posteriori dei parametri d'interesse e le stime inferenziali ottenute sono per questo affidabili.

In Figura 4.3 sono riportati i traceplot per le catene MCMC utilizzate in quest'analisi.

Ricordiamo che definiamo come "significativi" quei parametri il cui intervallo di credibilità esclude il valore zero, in quanto altrimenti significherebbe che esiste una probabilità non trascurabile che il valore stimato del parametro sia zero e dunque il livello della covariata ad esso associato non ha effetto sulla variazione della probabilità di osservare Activity = 1

Con lo scopo di analizzare ora le stima dei parametri significativi, ricordiamo che i livelli di riferimento utilizzati sono analoghi a quelli dei modelli GLM.

Tabella 4.7: Confronto tra modelli GLMM utilizzando WAIC

Modello	**elpd_diff**	**se_diff**
model_activity_conID	0.0	0.0
model_activity_conID3	-11.3	5.7
model_activity_conID2	-14.9	7.1

Tabella 4.8: Parametri significativi del modello GLMM.

Parametro	**Stima**	**CI inferiore**	**CI superiore**
Intercept	-1.79	-3.64	-0.04
Moon2	0.66	0.08	1.23
Moon3	1.44	0.17	2.72
Season4:tempOK	0.08	0.01	0.15

Considerando gli intervalli di credibilità riportati in Tabella 4.4, le stime dei parametri che possiamo ritenere statisticamente significative sono le seguenti:

- l'intercetta, con valore stimato -1.79
- Moon2, che corrisponde al livello "nuova" della variabile Moon, con valore 0.66
- Moon3, che corrisponde al livello piena della variabile Moon, con valore 1.44
- Season4:tempOK, che corrisponde all'influenza della temperatura durante l'autunno, con valore 0.08.

Anche in questo caso, per stime dei parametri come $tempOK$ e $Season2$: $tempOK$ si può ritenere plausibile l'effetto sulla probabilità di osservare Activity = 1, in quanto lo zero è marginalmente incluso nell'intervallo di credibilità, ma questo effetto non può ovviamente essere ritenuto certo.

Ricordiamo la formulazione matematica del modello GLMM in 3.3 e 3.4. Possiamo dunque esprimere π_i tramite la funzione sigmoide, inversa della logit:

$$\pi_i = \frac{1}{1 + e^{-\eta_i}}$$

Sfruttiamo questa formulazione matematica per fornire l'interpretazione dei valori delle stime. Siccome l'intercetta rappresenta il valore delle log-odds nel caso di riferimento, il valore indicato nella colonna Estimate

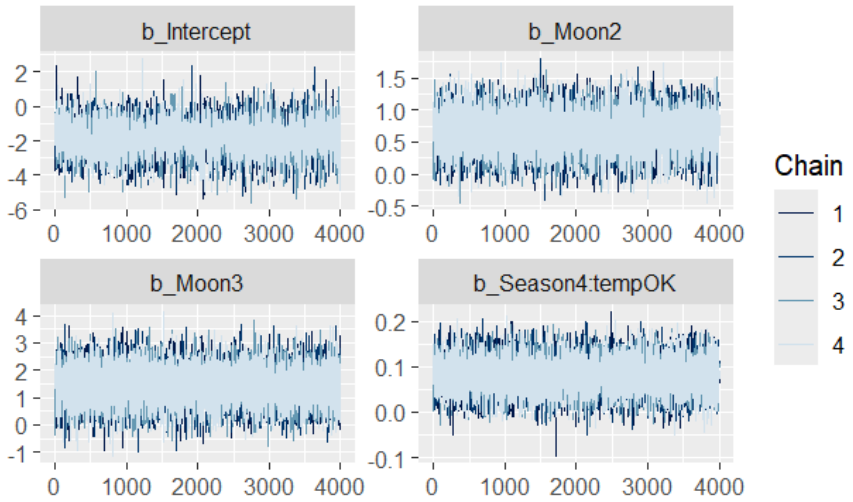


Figura 4.3: Traceplot delle catene MCMC per i parametri stimati significativi del modello GLMM

rappresenta la variazione per quel coefficiente nei log-odds per Activity=1, quando i livelli delle altre covariate rimangono costanti. Dunque un valore di Estimate positivo significa un aumento della probabilità di osservare Activity=1, mentre, al contrario, un valore negativo rappresenta al contrario una diminuzione della stessa probabilità.

Per le stime significative dei parametri, otteniamo dunque i seguenti effetti:

- Moon2, che corrisponde al livello "nuova" della variabile Moon, causa un moderato aumento della probabilità di osservare l'animale in attività (rispetto alla luna calante o crescente)
- Moon3, che corrisponde al livello "piena" della variabile Moon, causa un significativo aumento della probabilità di osservare l'animale in attività (rispetto alla luna calante o crescente)
- Season4:tempOK, che corrisponde all'influenza della temperatura durante l'autunno: aumenta leggermente la probabilità con il crescere della temperatura in autunno.

In Tabella 4.8 sono riportati i valori delle stime dei parametri considerati significativi, a titolo di esplicativo riportiamo quindi i valori di π_i in alcuni casi specifici.

Nel caso in cui consideriamo solo l'intercetta β_0, η_i corrisponde a -1.79, se aggiungiamo l'effetto casuale per l'esemplare Kazadum_151082, che

corrisponde a 1.63, otteniamo

$$\eta_i = -1.79 + 1.63 = -0.16$$
$$\pi_i = \frac{1}{1 + e^{0.16}} = 0.4601$$

Ovvero nelle condizioni di riferimento la probabilità di osservare questo particolare esemplare in condizioni di attività sono circa del 46%, percentuale molto più alta di quella corrispondente alle condizioni di riferimento.

Mantenendo costanti tutte le covariate, considerando sempre lo stesso individuo e passando dal livello di riferimento per Moon (CC) alla luna piena, il parametro η_i diventa

$$\eta_i = -1.79 + 1.63 + 1.44 = 1.28$$

E si ricava

$$\pi_i = \frac{1}{1 + e^{-1.28}} = 0.7824$$

Ovvero questo cambiamento di condizioni ha fatto aumentare la probabilità di osservare l'animale in attività al 78%, nelle condizioni descritte.

In Figura 4.3, osserviamo invece una heatmap che riflette come il passare dal livello di una covariata ad un altro modifichi le log-odds di Activity = 1, per avere una panoramica degli effetti causati dalla variazione dei parametri ambientali. L'interpretazione della Figura 4.3 è analoga a Figura 4.1.

Infine operiamo un confronto tra i modelli con e senza intercetta, sempre per mezzo del criterio WAIC. I risultati sono presenti in Tabella 4.9.

Tabella 4.9: Confronto tra modelli tramite `loo_compare()`

Modello	$\widehat{\text{elpd}}_{\text{diff}}$	SE_{diff}
model_activity_conID	0.0	0.0
model_activity_conID3	-11.3	5.7
model_activity_conID2	-14.9	7.1
model_activity4.5	-37.7	8.3
model_activity5	-38.1	8.6
model_activity3	-47.3	9.6
model_activity2	-51.7	10.3

Il modello chiaramente migliore è il GLMM, di cui a seguire diamo un'interpretazione biologica.

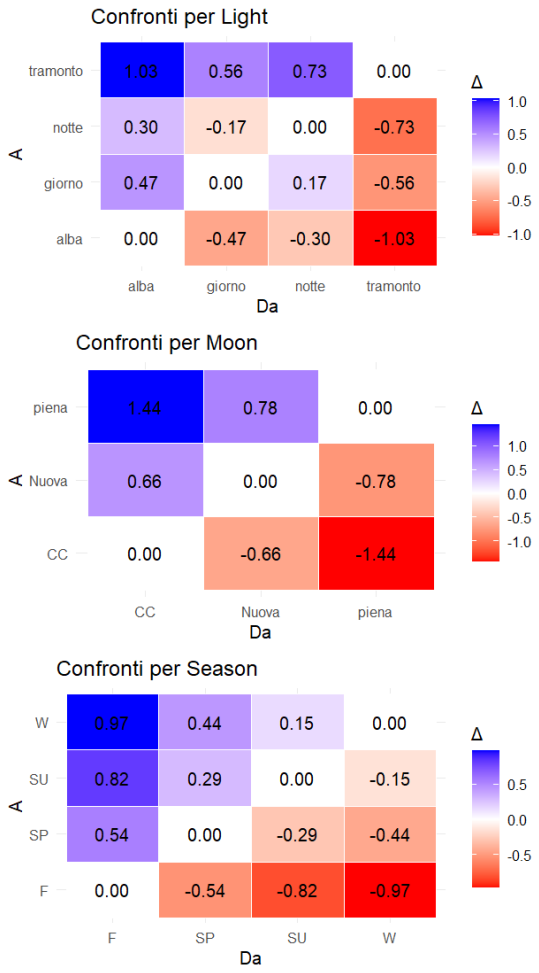


Figura 4.4: Confronto per le variazioni di log-odds di Activity per i vari livelli delle covariate Light, Moon e Season

La fase lunare di luna piena, che ha un forte effetto sulla probabilità di osservare l'animale in fase di attività, corrisponde alla situazione in cui la luna illumina maggiormente l'ambiente in cui le donnole si muovono. Questo fatto suggerisce che un predatore come la donnola tenda ad essere più attiva nelle notti in cui le condizioni ambientali garantiscano una visione migliore per la caccia.
L'effetto stagionale, che nel modello GLM sembrava avere un'influenza negativa su Activity, si è mostrato, nel modello con migliori capacità predittive, poco significativo.
Questo potrebbe significare che l'influenza sulla probabilità di attività degli individui, prima attribuita a fattori stagionali sia più probabilmente da ricondurre alla variabilità tra individui.
L'effetto della stagione può essere visto nel suo rapporto con la temperatura, l'effetto di quest'ultima infatti è positivo su Activity in maniera

significativa nella differenza tra inverno e autunno. Un risultato simile plausibile (lo zero è presente nell'estremità dell'intervallo di confidenza) vale anche nel caso della primavera.

Capitolo 5

Modelli Zero Inflated per la velocità

In questo capitolo concentriamo la nostra analisi sui modelli zero inflated per modellizzare la velocità. In particolare definiamo prima la formulazione matematica, per poi proseguire esponendo i risultati dei due modelli migliori analizzati e offrire un'interpretazione dei valori ottenuti. Infine terminiamo con una sezione in cui introduciamo teoricamente gli Hidden Markov Model e forniamo una spiegazione delle cause che ci hanno portato a non applicarli in questo studio.

5.1 Formulazione matematica dei modelli zero inflated

Come osservato in fase di esplorazione del dataset, la velocità presenta un elevato numero di osservazioni con valore zero. In Figura 5.1 viene presentato un istogramma che illustra il numero di frequenze per fasce di valori.

E' interessante, vista la relazione conosciuta tra Activity e la velocità dell'animale, provare a modellizzare quest'ultima per vedere se i fattori ambientali che la influenzano hanno effetto simile a quello sull'attività.

Per tenere conto dell'elevato numero di zeri, si è scelto di modellare la velocità tramite un modello Zero-Inflated Gamma (ZIG), che consente di distinguere tra due meccanismi generativi diversi per ciascuna

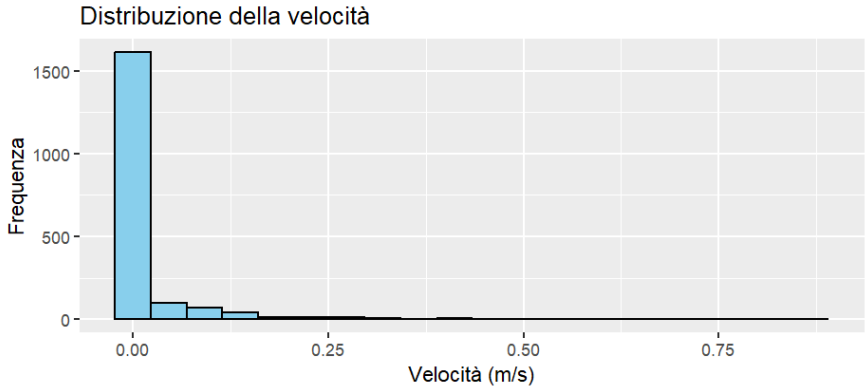


Figura 5.1: Istogramma per le frequenze dei valori della velocità all'interno del dataset

realizzazione della variabile aleatoria s_i:

$$s_i \sim \begin{cases} 0 & \text{con probabilità } \pi_i \\ \text{Gamma}(\alpha_i, \lambda_i) & \text{con probabilità } 1 - \pi_i \end{cases}$$

dove:

- s_i è una variabile aleatoria che rappresenta la velocità
- $\lambda_i = \frac{\alpha_i}{\mu_i}$ (parametrizzazione con shape-rate [29])
- μ_i è il valore atteso della componente Gamma
- α_i è il parametro di forma della componente Gamma
- π_i è la probabilità di osservare uno zero strutturale.

In questo contesto, la probabilità π_i di osservare uno zero strutturale è collegata alle covariate tramite una regressione logistica:

$$\text{logit}(\pi_i) = \mathbf{z}_i^T \boldsymbol{\gamma}$$

dove z_i^T è l'i-esima riga della matrice delle covariate Z_π, che contiene i valori delle covariate per le tutte osservazioni. Il vettore di γ è il vettore dei coefficienti stimati.
Il valore atteso della distribuzione Gamma condizionata a $s_i > 0$ (quando il processo di Bernoulli con probabilità $1-\pi$ assume valore 1) è modellato come:

$$\log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

dove x_i^T è l'i-esima riga della matrice delle covariate X_μ, che contiene i valori delle covariate per le tutte osservazioni. Il vettore di β è il vettore

dei coefficienti stimati. Infine, il parametro di forma della distribuzione Gamma è modellato come:

$$\log(\alpha_i) = \mathbf{w}_i^T \boldsymbol{\delta}$$

dove w_i^T è l'i-esima riga della matrice delle covariate W_α, che contiene i valori delle covariate per le tutte osservazioni. Il vettore di δ è il vettore dei coefficienti stimati.

E' stata anche implementata, analogamente a quanto svolto per la variabile Activity, una versione che include un'intercetta random $u_{\text{ID}_i} \sim \mathcal{N}(0, \tau^2)$. Anche in questo caso τ è distribuito secondo una distribuzione Inverse Gamma. L'intercetta random è stata inclusa nella modellizzazione del predittore lineare per il valore atteso della componente Gamma.

Questi modelli sono stati implementati su STAN[27], analogamente ai modelli per Activity è stato seguito un approccio bayesiano per le stime dei parametri dei modelli. In particolare attraverso la variante NUTS dell'Hamiltonian Monte Carlo già illustrato in precedenza.

Sono state testate diverse combinazioni delle condizioni ambientali come covariate; a seguire riportiamo i due modelli migliori secondo il criterio WAIC.

Per il primo modello le matrici di design utilizzate per ciascuna componente sono le seguenti:

```
X_mu    <- model.matrix(~ tempOK + Season + Light + Moon, data = df)
Z_pi    <- model.matrix(~ Light + Moon + Season + tempOK, data = df)
W_alpha <- model.matrix(~ Season + tempOK, data = df)
```

Questa specificazione implica che:

- μ_i è modellata in funzione della temperatura (tempOK), della stagione (Season), del livello di luce (Light) e della fase lunare (Moon)
- la probabilità π_i di osservare uno zero sistematico è modellata dalle stesse covariate;
- la forma della distribuzione Gamma α_i è modellata esclusivamente in funzione della stagione e della temperatura.

I tre predittori lineari per questo modello sono dunque definiti come:

$$\log(\mu_i) = \eta_i^{(\mu)} = \beta_0 + \sum_{k=2}^{4} \beta_{\text{Season}_k} \cdot \text{Season}_{ik} + \sum_{l=2}^{4} \beta_{\text{Light}_l} \cdot \text{Light}_{il} + \sum_{j=2}^{3} \beta_{\text{Moon}_j} \cdot \text{Moon}_{ij} + \beta_{\text{tempOK}} \cdot \text{tempOK}_i$$

che nella versione con intercetta random diventa

$$\log(\mu_i) = \eta_i^{(\mu)} = \beta_0 + \sum_{k=2}^{4} \beta_{\text{Season}_k} \cdot \text{Season}_{ik} + \sum_{l=2}^{4} \beta_{\text{Light}_l} \cdot \text{Light}_{il} + \sum_{j=2}^{3} \beta_{\text{Moon}_j} \cdot \text{Moon}_{ij} + \beta_{\text{tempOK}} \cdot \text{tempOK}_i + u_{\text{ID}_i}$$

$$\text{logit}(\pi_i) = \eta_i^{(\pi)} = \gamma_0 + \sum_{k=2}^{4} \gamma_{\text{Season}_k} \cdot \text{Season}_{ik} + \sum_{l=2}^{4} \gamma_{\text{Light}_l} \cdot \text{Light}_{il} + \sum_{j=2}^{3} \gamma_{\text{Moon}_j} \cdot \text{Moon}_{ij} + \gamma_{\text{tempOK}} \cdot \text{tempOK}_i$$

$$\log(\alpha_i) = \eta_i^{(\alpha)} = \delta_0 + \sum_{k=2}^{4} \delta_{\text{Season}_k} \cdot \text{Season}_{ik} + \delta_{\text{tempOK}} \cdot \text{tempOK}_i$$

Ricordiamo l'intercetta casuale e le prior non informative

$$u_{\text{ID}_i} \sim \mathcal{N}(0, \tau^2)$$

$$\tau \sim InvGamma(2,1)$$

$$\beta_k \sim \mathcal{N}(0, 3),$$
$$\gamma_k \sim \mathcal{N}(0, 3),$$
$$\delta_k \sim \mathcal{N}(0, 2).$$

Il secondo modello è una versione con meno covariate, diverse per ciascuna componente.

```
X_mu <- model.matrix(~ Season + tempOK, data = df)
Z_pi <- model.matrix(~ Light + Moon, data = df)
W_alpha <- model.matrix(~ Season, data = df)
```

Questa specificazione implica che:

- μ_i è modellata in funzione della temperatura (tempOK) e della stagione
- la probabilità π_i di osservare uno zero sistematico è modellata dalla luce e dalla fase lunare
- la forma della distribuzione Gamma α_i è modellata esclusivamente in funzione della stagione.

I tre predittori lineari per questo modello sono dunque definiti come:

$$\log(\mu_i) = \eta_i^{(\mu)} = \beta_0 + \sum_{k=2}^{4} \beta_{\text{Season}_k} \cdot \text{Season}_{ik} + \beta_{\text{tempOK}} \cdot \text{tempOK}_i$$

che nella versione con intercetta random diventa

$$\log(\mu_i) = \eta_i^{(\mu)} = \beta_0 + \sum_{k=2}^{4} \beta_{\text{Season}_k} \cdot \text{Season}_{ik} + \beta_{\text{tempOK}} \cdot \text{tempOK}_i + u_{\text{ID}_i}$$

$$\text{logit}(\pi_i) = \eta_i^{(\pi)} = \gamma_0 + \sum_{l=2}^{4} \gamma_{\text{Light}_l} \cdot \text{Light}_{il} + \sum_{j=2}^{3} \gamma_{\text{Moon}_j} \cdot \text{Moon}_{ij}$$

$$\log(\alpha_i) = \eta_i^{(\alpha)} = \delta_0 + \sum_{k=2}^{4} \delta_{\text{Season}_k} \cdot \text{Season}_{ik}$$

L'intercetta casuale e le prior non informative sono analoghe a quelle del modello precedente.

Risultati dei modelli

Come anticipato, le prestazioni dei modelli sono state confrontate utilizzando il criterio WAIC. A seguire mostriamo i risultati delle stime dei campioni a posteriori ottenute con inferenza bayesiana per mezzo dell'algoritmo NUTS, metodo di default utilizzato da STAN.

Ricordiamo che, come nella presentazione dei risultati dei GLM

- Estimate: colonna per i valori stimati dei parametri
- l-95% CI: rappresenta il limite inferiore dell'intervallo di credibilità al 95%
- u-95% CI: rappresenta il limite superiore dell'intervallo di credibilità al 95%

L'intervallo di credibilità, in questo caso al 95%, è un riferimento importante in quanto fornisce un'informazione su quali parametri del modello siano significativi. Questo intervallo indica, con un probabilità al 95%, che il valore stimato del parametro d'interesse risiede al suo interno. Definiamo dunque come "significativi" quei parametri il cui intervallo di credibilità esclude il valore zero, in quanto altrimenti significherebbe che esiste una probabilità non trascurabile che il valore stimato del parametro

sia zero e dunque il livello della covariata ad esso associato non ha effetto.

Si sottolinea inoltre che, ai fini dell'interpretazione, i livelli di riferimento delle covariate sono i seguenti:

- covariata Moon: "CC" (luna crescente/calante)
- covariata Light: "alba"
- covariata Season: "F" (autunno).

Al fine di comprendere i risultati dei modelli contenuti nelle tabelle successive, riportiamo i corrispondenti livelli delle covariate associati alle stime dei parametri

- Moon
  - `Moon2` → livello `nuova`
  - `Moon3` → livello `piena`
- Light
  - `Light2` → livello `giorno`
  - `Light3` → livello `notte`
  - `Light4` → livello `tramonto`
- Season
  - `Season2` → livello `SP` (primavera)
  - `Season3` → livello `SU` (estate)
  - `Season4` → livello `W` (inverno.

In Tabella 5.1 sono riportati i risultati delle stime dei parametri per il primo modello introdotto, mentre in Tabella 5.2 sono riportati solo i valori considerati significativi.

Come riportato in Tabella 5.2, le stime statisticamente significative dei coefficienti per il primo modello sono:

- Intercetta per parte Gamma, con valore -2.110. Questo risultato esprime un valore atteso della velocità di base (quando ci si trova in condizioni di riferimento) molto basso, infatti se ricaviamo il valore di μ_i per mezzo della funzione esponenziale (l'inversa del logaritmo) otteniamo $\exp(-2.110) = 0.1212$
- Season2 per parte Gamma, riferito al livello della covariata "SP" (primavera), con valore -0.503, significa che rispetto all'autunno la velocità osservata tende a decrescere

Tabella 5.1: Coefficienti stimati per tutte le componenti del primo modello

Parte del modello	**Termine**	**Estimate**	**l-95% CI**	**u-95% CI**
Gamma ($\log(\mu)$)	(Intercept)	-2.110	-3.260	-0.764
Gamma ($\log(\mu)$)	tempOK	-0.015	-0.043	0.013
Gamma ($\log(\mu)$)	Season2	-0.503	-0.963	-0.041
Gamma ($\log(\mu)$)	Season3	0.275	-0.185	0.706
Gamma ($\log(\mu)$)	Season4	-0.956	-1.540	-0.377
Gamma ($\log(\mu)$)	Light2	0.093	-1.178	1.150
Gamma ($\log(\mu)$)	Light3	0.262	-1.023	1.305
Gamma ($\log(\mu)$)	Light4	-0.117	-1.522	1.130
Gamma ($\log(\mu)$)	Moon2	-0.254	-0.911	0.501
Gamma ($\log(\mu)$)	Moon3	0.502	-0.460	1.689
Zero-inflation ($\text{logit}(\pi)$)	(Intercept)	2.475	1.577	3.499
Zero-inflation ($\text{logit}(\pi)$)	Season2	-0.250	-0.633	0.124
Zero-inflation ($\text{logit}(\pi)$)	Season3	0.025	-0.402	0.448
Zero-inflation ($\text{logit}(\pi)$)	Season4	-0.287	-0.711	0.141
Zero-inflation ($\text{logit}(\pi)$)	Light2	-0.798	-1.831	0.056
Zero-inflation ($\text{logit}(\pi)$)	Light3	0.445	-0.587	1.339
Zero-inflation ($\text{logit}(\pi)$)	Light4	-0.131	-1.291	0.929
Zero-inflation ($\text{logit}(\pi)$)	Moon2	-0.097	-0.678	0.541
Zero-inflation ($\text{logit}(\pi)$)	Moon3	-1.427	-2.615	-0.302
Zero-inflation ($\text{logit}(\pi)$)	tempOK	-0.028	-0.051	-0.005
Alpha ($\log(\alpha)$)	(Intercept)	-0.500	-1.013	-0.003
Alpha ($\log(\alpha)$)	Season2	-0.237	-0.618	0.146
Alpha ($\log(\alpha)$)	Season3	0.261	-0.199	0.725
Alpha ($\log(\alpha)$)	Season4	0.066	-0.469	0.608
Alpha ($\log(\alpha)$)	tempOK	0.001	-0.026	0.028

Tabella 5.2: Coefficienti significativi per le tre componenti del primo modello

Componente	**Termine**	**Media**	**l-95% CI**	**u-95% CI**
Gamma ($\log(\mu)$)	(Intercept)	-2.110	-3.260	-0.764
Gamma ($\log(\mu)$)	Season2	-0.503	-0.963	-0.041
Gamma ($\log(\mu)$)	Season4	-0.956	-1.540	-0.377
Zero-inflation ($\text{logit}(\pi)$)	(Intercept)	2.475	1.577	3.499
Zero-inflation ($\text{logit}(\pi)$)	Moon3	-1.427	-2.615	-0.302
Zero-inflation ($\text{logit}(\pi)$)	tempOK	-0.028	-0.051	-0.005
Alpha ($\log(\alpha)$)	(Intercept)	-0.500	-1.013	-0.003

Parte del modello	**Termine**	**Estimate**	**l-95% CI**	**u-95% CI**
Gamma ($\log(\mu)$)	Intercept	-1.923	-2.414	-1.410
Gamma ($\log(\mu)$)	Season2	-0.511	-0.939	-0.085
Gamma ($\log(\mu)$)	Season3	0.296	-0.155	0.736
Gamma ($\log(\mu)$)	Season4	-1.032	-1.587	-0.482
Gamma ($\log(\mu)$)	tempOK	-0.019	-0.045	0.007
Zero-inflation ($\text{logit}(\pi)$)	Intercept	2.104	1.286	3.079
Zero-inflation ($\text{logit}(\pi)$)	Light2	-1.026	-2.004	-0.195
Zero-inflation ($\text{logit}(\pi)$)	Light3	0.366	-0.635	1.250
Zero-inflation ($\text{logit}(\pi)$)	Light4	-0.302	-1.436	0.728
Zero-inflation ($\text{logit}(\pi)$)	Moon2	-0.147	-0.714	0.476
Zero-inflation ($\text{logit}(\pi)$)	Moon3	-1.521	-2.712	-0.409
Alpha ($\log(\alpha)$)	Intercept	-0.501	-0.795	-0.224
Alpha ($\log(\alpha)$)	Season2	-0.208	-0.575	0.166
Alpha ($\log(\alpha)$)	Season3	0.293	-0.055	0.651
Alpha ($\log(\alpha)$)	Season4	0.088	-0.318	0.499

Tabella 5.3: Coefficienti stimati per tutte le componenti del secondo modello

- Season4 per parte Gamma, riferito al livello della covariata "W" (inverno), con valore -0.956, che comporta una diminuzione della velocità attesa rispetto all'autunno ancora più grande che per la primavera
- Intercetta per parte Zero-Inflation, con valore 2.475, implica che si ha un'altissima probabilità di osservare uno zero strutturale, di circa il 92% (usando la funzione sigmoide, inversa della logit, si ha $\frac{1}{1+e^{-(2.475)}}$)
- Moon3 per parte Zero-Inflation, riferito al livello della covariata "piena" , con valore -1.427. Questo implica che la luna piena diminuisce di molto, rispetto ai livelli di riferimento, la probabilità di osservare uno zero strutturale
- tempOK per parte Zero-Inflation, riferito all'osservazione della temperatura, con valore -0.028. Anche la temperatura dunque, con il crescere del suo valore fa diminuire la probabilità di osservare uno zero strutturale
- Intercetta per parte Alpha, con valore -0.500, che indica una dispersione moderata della velocità.

Le stime dei parametri per il secondo modello introdotto sono invece consultabili in Tabella 5.3.

Parte del modello	**Termine**	**Estimate**	**l-95% CI**	**u-95% CI**
Gamma ($\log(\mu)$)	Intercept	-1.923	-2.414	-1.410
Gamma ($\log(\mu)$)	Season2	-0.511	-0.939	-0.085
Gamma ($\log(\mu)$)	Season4	-1.032	-1.587	-0.482
Zero-inflation ($\text{logit}(\pi)$)	Intercept	2.104	1.286	3.079
Zero-inflation ($\text{logit}(\pi)$)	Light2	-1.026	-2.004	-0.195
Zero-inflation ($\text{logit}(\pi)$)	Moon3	-1.521	-2.712	-0.409
Alpha ($\log(\alpha)$)	Intercept	-0.501	-0.795	-0.224

Tabella 5.4: Coefficienti significativi per il secondo modello

Come riportato in Tabella 5.4, le stime statisticamente significative dei coefficienti sono:

- Intercetta per parte Gamma, con valore -1.923. Rispetto al primo modello dunque, si ha una velocità attesa più alta nelle condizioni di riferimento ($\exp(-1.923) = 0{,}1461$)
- Season2 per parte Gamma, riferito al livello della covariata "SP" (primavera), con valore -0.511, analogamente al primo modello la primavera rispetto all'autunno comporta una diminuzione del valore atteso della velocità
- Season4 per parte Gamma, riferito al livello della covariata "W" (inverno), con valore -1.032, risultato analogo al parametro precedente, con effetto ancora più forte
- Intercetta per parte Zero-Inflation, con valore 2.104, dunque nel secondo modello rimane una probabilità di osservare uno zero strutturale alta, ma comuque leggermente più bassa rispetto al primo modello
- Light2 per parte Zero-Inflation, riferito al livello "giorno" della covariata Light, con valore -1.026. La luce diurna dunque contribuisce alla diminuzione della probabilità di avere zeri sistematici, in maniera meno forte rispetto a Moon3
- Moon3 per parte Zero-Inflation, riferito al livello della covariata "piena", con valore -1.521. Analogamente al primo modello, la fase di luna piena diminuisce la probabilità di osservare zeri sistematici.
- Intercetta per parte Alpha, con valore -0.501. Risultato praticamente identico a quello del primo modello.

Parte del modello	**Termine**	**Estimate**	**l-95% CI**	**u-95% CI**
Gamma ($\log(\mu)$)	Intercept	-2.391	-3.811	-0.884
Gamma ($\log(\mu)$)	tempOK	-0.021	-0.052	0.011
Gamma ($\log(\mu)$)	Season2	-0.308	-1.149	0.699
Gamma ($\log(\mu)$)	Season3	0.479	-0.092	1.025
Gamma ($\log(\mu)$)	Season4	-0.728	-1.783	0.486
Gamma ($\log(\mu)$)	Light2	0.138	-1.148	1.224
Gamma ($\log(\mu)$)	Light3	0.314	-0.979	1.404
Gamma ($\log(\mu)$)	Light4	-0.071	-1.459	1.204
Gamma ($\log(\mu)$)	Moon2	-0.455	-1.137	0.316
Gamma ($\log(\mu)$)	Moon3	0.672	-0.354	1.879
Zero-inflation ($\text{logit}(\pi)$)	Intercept	2.473	1.584	3.477
Zero-inflation ($\text{logit}(\pi)$)	Season2	-0.249	-0.630	0.129
Zero-inflation ($\text{logit}(\pi)$)	Season3	0.026	-0.396	0.439
Zero-inflation ($\text{logit}(\pi)$)	Season4	-0.283	-0.722	0.150
Zero-inflation ($\text{logit}(\pi)$)	Light2	-0.799	-1.775	0.059
Zero-inflation ($\text{logit}(\pi)$)	Light3	0.445	-0.558	1.331
Zero-inflation ($\text{logit}(\pi)$)	Light4	-0.129	-1.257	0.933
Zero-inflation ($\text{logit}(\pi)$)	Moon2	-0.103	-0.685	0.515
Zero-inflation ($\text{logit}(\pi)$)	Moon3	-1.436	-2.648	-0.284
Zero-inflation ($\text{logit}(\pi)$)	tempOK	-0.028	-0.050	-0.005
Alpha ($\log(\alpha)$)	Intercept	-0.471	-0.985	0.031
Alpha ($\log(\alpha)$)	Season2	-0.175	-0.550	0.214
Alpha ($\log(\alpha)$)	Season3	0.375	-0.091	0.854
Alpha ($\log(\alpha)$)	Season4	0.040	-0.498	0.569
Alpha ($\log(\alpha)$)	tempOK	-0.001	-0.028	0.026

Tabella 5.5: Coefficienti stimati per tutte le componenti del primo modello con aggiunta di intercetta casuale

Parte del modello	**Termine**	**Estimate**	**l-95% CI**	**u-95% CI**
Gamma ($\log(\mu)$)	Intercept	-2.391	-3.811	-0.884
Zero-inflation ($\text{logit}(\pi)$)	Intercept	2.473	1.584	3.477
Zero-inflation ($\text{logit}(\pi)$)	Moon3	-1.436	-2.648	-0.284
Zero-inflation ($\text{logit}(\pi)$)	tempOK	-0.028	-0.050	-0.005

Tabella 5.6: Coefficienti significativi per il primo modello con aggiunta di intercetta casuale

Per quanto riguarda il terzo modello implementato, ovvero l'aggiunta al

primo modello presentato di un'intercetta casuale per lo stimatore legato al valore atteso della velocità μ_i, in Tabella 5.5 si possono consultare le stime complete di tutti i parametri stimati. Come riportato in Tabella 5.6, le stime statisticamente significative dei coefficienti sono :

- Intercetta per parte Gamma, con valore -2.391. Il valore atteso della velocità in condizioni di riferimento
- Intercetta per parte Zero-Inflation, con valore 2.473. Stessa interpretazione del primo modello
- Moon3 per parte Zero-Inflation, riferito al livello della covariata "piena", con valore -1.436. Stessa interpretazione del primo modello
- tempOK per parte Zero-Inflation, riferito all'osservazione della temperatura, con valore -0.028. Stessa interpretazione del primo modello.

Si può osservare come l'inserimento di un'intercetta random ha reso poco significativi i contributi dei livelli della covariata Season rispetto al valore atteso μ_i della velocità. Da questi risultati sembra dunque che gran parte delle differenze tra velocità siano dovute a differenze tra individui osservati piuttosto che alle condizioni ambientali. L'effetto casuale ha incorporato parte della variabilità che prima veniva attribuita agli effetti fissi. Di conseguenza, questo modello suggerisce che la media della velocità sia spiegata più efficacemente da caratteristiche individuali non osservate (catturate dall'intercetta casuale) piuttosto che dalle condizioni ambientali esplicitamente modellate.

Per quanto riguarda il quarto modello esposto, ovvero l'aggiunta al secondo modello presentato di un'intercetta casuale per lo stimatore legato al valore atteso della velocità μ_i, in Tabella 5.8 si possono consultare le stime complete di tutti i parametri stimati. Come riportato in Tabella 5.9, le stime statisticamente significative dei coefficienti sono:

- Intercetta per parte Gamma, con valore -2.188
- Intercetta per parte Zero-Inflation, con valore 2.089
- Light2 per parte Zero-Inflation, riferito al livello "giorno" della covariata Light, con valore -1.011. Interpretazione identica al secondo modello
- Moon3 per parte Zero-Inflation, riferito al livello della covariata "piena", con valore -1.526. Interpretazione identica al secondo modello

Parametro / ID	**Estimate**	**Errore Std**	**l-95% CI**	**u-95% CI**
sd(Intercept)	0.485	0.188	0.228	0.956
u[1]	-0.043	–	-0.733	0.611
u[2]	-0.310	–	-0.970	0.367
u[3]	0.087	–	-0.691	0.879
u[4]	-0.235	–	-1.548	0.716
u[5]	0.448	–	-0.182	1.142
u[6]	-0.247	–	-1.002	0.436
u[7]	-0.222	–	-1.262	0.686
u[8]	-0.110	–	-0.902	0.671
u[9]	-0.216	–	-0.744	0.409
u[10]	-0.344	–	-1.288	0.432
u[11]	0.284	–	-0.373	1.103
u[12]	-0.003	–	-1.046	1.029
u[13]	0.459	–	-0.033	1.128
u[14]	0.390	–	-0.161	1.127

Tabella 5.7: Deviazione standard degli effetti casuali (τ) e intercette casuali per individuo (u_j) per il primo modello con aggiunta di intercetta casuale

Parte del modello	**Termine**	**Estimate**	**l-95% CI**	**u-95% CI**
Gamma ($\log(\mu)$)	Intercept	-2.188	-3.119	-1.368
Gamma ($\log(\mu)$)	Season2	-0.283	-1.096	0.719
Gamma ($\log(\mu)$)	Season3	0.543	-0.021	1.081
Gamma ($\log(\mu)$)	Season4	-0.759	-1.808	0.429
Gamma ($\log(\mu)$)	tempOK	-0.025	-0.055	0.005
Zero-inflation ($\text{logit}(\pi)$)	Intercept	2.089	1.259	3.071
Zero-inflation ($\text{logit}(\pi)$)	Light2	-1.011	-2.000	-0.162
Zero-inflation ($\text{logit}(\pi)$)	Light3	0.380	-0.654	1.267
Zero-inflation ($\text{logit}(\pi)$)	Light4	-0.287	-1.426	0.755
Zero-inflation ($\text{logit}(\pi)$)	Moon2	-0.144	-0.701	0.453
Zero-inflation ($\text{logit}(\pi)$)	Moon3	-1.526	-2.724	-0.416
Alpha ($\log(\alpha)$)	Intercept	-0.508	-0.811	-0.216
Alpha ($\log(\alpha)$)	Season2	-0.162	-0.539	0.218
Alpha ($\log(\alpha)$)	Season3	0.387	0.032	0.760
Alpha ($\log(\alpha)$)	Season4	0.088	-0.345	0.517

Tabella 5.8: Coefficienti stimati per tutte le componenti del secondo modello con aggiunta di intercetta casuale

Parte del modello	Termine	Estimate	l-95% CI	u-95% CI
Gamma ($\log(\mu)$)	Intercept	-2.188	-3.119	-1.368
Zero-inflation ($\text{logit}(\pi)$)	Intercept	2.089	1.259	3.071
Zero-inflation ($\text{logit}(\pi)$)	Light2	-1.011	-2.000	-0.162
Zero-inflation ($\text{logit}(\pi)$)	Moon3	-1.526	-2.724	-0.416
Alpha ($\log(\alpha)$)	Intercept	-0.508	-0.811	-0.216
Alpha ($\log(\alpha)$)	Season3	0.387	0.032	0.760

Tabella 5.9: Coefficienti significativi per il secondo modello con aggiunta di intercetta casuale

Parametro / ID	Estimate	Errore Std	l-95% CI	u-95% CI
sd(Intercept)	0.485	0.188	0.228	0.956
u[1]	-0.043	–	-0.733	0.611
u[2]	-0.310	–	-0.970	0.367
u[3]	0.087	–	-0.691	0.879
u[4]	-0.235	–	-1.548	0.716
u[5]	0.448	–	-0.182	1.142
u[6]	-0.247	–	-1.002	0.436
u[7]	-0.222	–	-1.262	0.686
u[8]	-0.110	–	-0.902	0.671
u[9]	-0.216	–	-0.744	0.409
u[10]	-0.344	–	-1.288	0.432
u[11]	0.284	–	-0.373	1.103
u[12]	-0.003	–	-1.046	1.029
u[13]	0.459	–	-0.033	1.128
u[14]	0.390	–	-0.161	1.127

Tabella 5.10: Deviazione standard degli effetti casuali (τ) e intercette casuali per individuo (u_j) per il secondo modello con aggiunta di intercetta casuale

- Intercetta per parte Alpha, con valore -0.508, Interpretazione identica al secondo modello

- Season3 per parte Alpha, riferito al livello della covariata "SU" estate, con valore 0.387. L'effetto del livello "estate" poteva già essere considerato plausibile nel secondo modello, ma essendo lo zero compreso nell'estremità dell'intervallo non poteva essere considerato significativo. Alla luce di questo nuovo intervallo il passaggio dal livello di riferimento autunno all'estate comporta una maggiore dispersione per la velocità.

Tabella 5.11: Confronto tra modelli basato su WAIC

Modello	**Differenza ELPD**	**Errore Std**
model1	0.0	0.0
model2	-1.4	3.4
model3	-7.8	3.0
model4	-7.8	3.0

Le considerazioni che riguardano l'aggiunta dell'intercetta casuale sono invece analoghe a quelle del modello precedente. In Tabella 5.11 troviamo invece le valutazioni basate sul WAIC, che indicano come model1, ovvero l'ultimo modello di cui abbiamo esposto i risultati, sia considerato il migliore tra quelli implementati. Si osserva come anche in questo caso l'aggiunta di un'intercetta casuale si traduce in generale in una performance migliore per il modello.
Questi risultati per l'interpretazione della velocità mostrano coerenza con quelli ottenuti per la modellizzazione dell'attività. In entrambi i casi infatti si nota un forte effetto dell'intercetta casuale per il singolo individuo piuttosto che degli effetti stagionali, sia per la probabilità di osservare l'animale in attività che per il valore atteso della velocità. Inoltre, l'effetto della fase lunare è sia significativo in termini di maggiore illuminazione per favorire l'attività, sia in termini di riduzione della probabilità di osservare zeri strutturali quando studiamo la velocità.

5.2 I modelli HMM e i motivi dell'impossibilità della loro implementazione

5.2.1 I modelli HMM

In questo capitolo forniamo il quadro teorico in cui si sviluppano gli Hidden Markov Models e discutiamo le motivazioni che ci hanno impedito di utilizzarli per l'analisi dei dati a disposizione.

Gli Hidden Markov Models (HMM) sono modelli ampiamente utilizzati in letteratura per descrivere il comportamento animale a partire da una sequenza di osservazioni temporali. Nel contesto dell'analisi del movimento animale questi modelli permettono di collegare le osservazioni relative alla posizione ad un processo comportamentale non osservabile direttamente con esse, come la caccia o il riposo.

Un HMM è definito come un processo stocastico doppio, composto da:

- una sequenza di stati latenti $\{S_t\}_{t=1}^T$, non osservabili direttamente, che evolve nel tempo secondo una catena di Markov
- una sequenza di osservazioni $\{Y_t\}_{t=1}^T$, condizionatamente indipendenti per lo stato latente corrente, generate da distribuzioni dipendenti esclusivamente dallo stato corrente.

La probabilità di transizione tra stati è descritta da una matrice di transizione $\Lambda = [\lambda_{ij}]$, dove $\lambda_{ij}^{(t)} = P(S_{t+1} = j \mid S_t = i)$. La distribuzione iniziale degli stati è invece rappresentata da un vettore $\iota^{(1)} = (\iota_1, \ldots, \iota_N)$, dove $\iota_n^{(1)} = P(S_1 = n)$. [21]

Un HMM è dunque completamente definito da:

- Una sequenza di stati nascosti $S_t \in \{1, \ldots, N\}$
- Una sequenza di osservazioni Y_t, condizionatamente indipendenti dato S_t
- Una matrice di transizione Λ con elementi $\lambda_{ij} = P(S_t = j | S_{t-1} = i)$
- Una distribuzione iniziale $\iota = P(S_1)$
- Una famiglia di distribuzioni osservabili $f_j(y_t) = P(Y_t = y_t | S_t = j)$

L'utilizzo degli HMM in questo ambito di ricerca offre numerosi vantaggi, tra i quali:

- modellare il comportamento dell'animale e la transizione tra i diversi stati di attività
- stimare la probabilità di un certo tipo di comportamento in un dato momento
- modellizzare e stimare la quantità di tempo trascorsa in un singolo stato
- fornire una struttura flessibile per incorporare covariate che influenzano la probabilità di transizione tra stati.

Siccome un'animale in natura tende a mantenere un certo tipo di attività in maniera persistente nel tempo, ovvero non alterna il riposo alla caccia in spazi di tempo poco distanti l'uno dall'altro, i modelli HMM permettono di catturare questa persistenza nella matrice di transizione, oltre a sottolineare le probabilità di transizione da un tipo di stato ad un altro. Inoltre, essendo le osservazioni dipendenti esclusivamente dallo stato corrente, gli HMM permettono di utilizzare tipologie diverse di famiglie di distribuzioni per modellare variabili differenti.

Per implementare correttamente i modelli HMM sono necessarie dunque, tra le altre, alcune condizioni:

- intervalli di tempo costanti: le osservazioni devono essere raccolte a intervalli regolari. Questo garantisce che la dinamica temporale del processo latente sia ben definita[11].
- catena di Markov di primo ordine: lo stato corrente dipende solo dallo stato precedente.

5.2.2 Motivazioni per il mancato utilizzo degli HMM

Per quanto riguarda i dati a nostra disposizione, alla luce di quanto introdotto dal punto di vista teorico, sono stati riscontrati diversi fattori che hanno impedito l'utilizzo di questi modelli:

- l'intervallo di tempo tra le singole osservazioni
- l'approssimazione della posizione geografica delle osservazioni
- in generale il numero di dati a disposizione.

Per quanto riguarda il primo punto, ovvero l'intervallo di tempo trascorso tra la raccolta delle informazioni, la problematica riscontrata riguarda la disponibilità di osservazioni che rispettino tale criterio.

Tabella 5.12: *head*() del dataset

ID	h	v
Drugo_151470	323700	4705970
Drugo_151470	323630	4705900
Drugo_151470	323630	4705900
Drugo_151470	323630	4705900
Drugo_151470	323630	4705900
Drugo_151470	323630	4705900

Sebbene i dati siano stati raccolti con l'obiettivo di effettuare misurazioni ogni 15 minuti per la durata di 8 ore seguite da 8 ore di riposo, diverse problematiche hanno impedito che le sequenze di osservazioni raggiungessero un numero ideale per l'analisi.
In prima battuta sottolineiamo come essendo l'animale di piccola taglia, la dimensione della tecnologia utilizzata per i radio transmitter (radio tag) che forniscono l'informazione deve essere conseguentemente in scala per non essere di disturbo all'animale. Pertanto il peso dei radio tag utilizzati rappresenta circa lo 0,5% del peso di un maschio adulto di donnola[18].
Questo introduce problemi legati alla durata della batteria, senza contare i problemi legati al radio tag che viene perso o presenta malfunzionamenti. Numerosi dispositivi infatti sono risultati offline dopo in media 10 giorni, su una durata di batteria prevista di 30-40 giorni.
La presenza di pause tra le osservazioni, dati intermedi ottenuti da altri strumenti (come fototrappole) e informazioni mancanti per via dei missing value hanno fatto in modo che la sequenza di osservazioni per un singolo individuo non presenti regolarità tra le osservazioni.
Abbiamo, in fase di esplorazione e preparazione dei dati, provato a separare i vari periodi di tracciamento con intervalli di tempo costanti da 15 minuti, ma questo creava sequenze di osservazioni comunque troppo corte.

Per quanto riguarda invece il secondo punto, ovvero la posizione geografica approssimata dell'animale, ricordiamo che questa è stata introdotta in fase di registrazione dei dati.
Infatti la mappa utilizzata per la zona di ricerca aveva una scala 1:10 000 e sono state utilizzate le coordinate UTM chilometriche per registrare le singole posizioni. Ciò ha introdotto un errore di 10 metri per il sistema di mappatura.

Come evidenziato già nel Capitolo 2 in fase di introduzione e presentazione del dataset, questo comporta che in numerose osservazioni consecutive dell'animale le posizioni geografiche risultino identiche. Osserviamo in Tabella 5.12, che rappresenta le prime 6 righe del dataset, quanto questo sia un fenomeno diffuso. Ricordiamo per la Tabella 1.2 che la coppia (h, v) determina la posizione geografica dell'animale in formato UTM 33-N.
Quanto sopra citato è stata la problematica più significativa tra quelle presentate, in quanto l'animale tracciato non risulta muoversi nello spazio e rende impossibile ricavare variabili come i turning angles.
Delle 1882 osservazioni ricavate dopo il preprocessing, solo 199 di essere sono caratterizzate dalla presenza di un turning angle valido, rendendo inapplicabile l'utilizzo di modelli che sfruttano questa variabile per descrivere il movimento dell'animale nello spazio, come nel caso degli HMM.

Per concludere, sottolineiamo come il numero scarseggiante di dati non sia una caratteristica esclusiva del dataset in analisi, ma è un problema comune nell'ambito della ricerca e lo studio dei movimenti e pattern di comportamento animali.

Bibliografia

[1] Hirotugu Akaike. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov and F. Csaki, editors, *Proceedings of the 2nd International Symposium on Information Theory*, pages 267–281, Budapest, 1973. Akademiai Kiado.

[2] Paul-Christian Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, 80(1):1–28, 2017.

[3] Paul-Christian Bürkner. *brms: Bayesian Regression Models using Stan*. R Project, 2025. R package version 2.23.0, pubblicato il 9 settembre 2025.

[4] T. M. Cullen et al. *bayesmove: Non-Parametric Bayesian Analyses of Animal Movement*. CRAN, October 2025. R package version 0.2.3.

[5] Joseph M. Eisaguirre, Perry J. Williams, and Mevin B. Hooten. Rayleigh step-selection functions and connections to continuous-time mechanistic movement models. *Movement Ecology*, 12(1):14, 2024.

[6] Encyclopædia Britannica. Law of cosines. https://www.britannica.com/science/law-of-cosines, 2024.

[7] EPSG.io. Epsg:23033 – ed50 / utm zone 33n. https://epsg.io/23033, 2025.

[8] Jonah Gabry, Aki Vehtari, and Stan Development Team. *loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models*, 2025. R package version 2.8.0, pubblicato il 3 luglio 2024.

[9] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and Computing*, 24(6):997–1016, 2014.

[10] Andrew Gelman and Donald B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–511, 1992.

[11] Zoubin Ghahramani. An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 15(1):9–42, 2001.

[12] Abraham Granados and Isaías Bañales. Understanding the hamiltonian monte carlo through its physics fundamentals and examples.

arXiv preprint arXiv:2501.13932, 2025. Submitted on 8 Jan 2025.

[13] Matthew D. Hoffman and Andrew Gelman. The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *arXiv preprint arXiv:1111.4246*, 2011.

[14] International Organization for Standardization. Iso 8601 — date and time format. https://www.iso.org/iso-8601-date-and-time-format.html, 2017.

[15] Istituto della Enciclopedia Italiana. Donnola. https://www.treccani.it/vocabolario/ricerca/donnola/, 2025.

[16] Vianey Leos-Barajas and Théo Michelot. An introduction to animal movement modeling with hidden markov models using stan for bayesian inference. *arXiv preprint arXiv:1806.10639*, 2018. Submitted on 27 Jun 2018, [q-bio.QM].

[17] Sandro Lovari, Francesco Ferretti, Anna Bocci, and Isabelle Minder. Protocollo di monitoraggio radio-telemetrico degli individui marcati: Azioni a.5, c.2, c.4 e c.5. Technical report, Università degli Studi di Siena, 2025. Progetto LIFE NAT/IT/000183, "Development of coordinated protection measures for Apennine chamois (Rupicapra pyrenaica ornata)".

[18] Caterina Magrini, Emiliano Manzo, Livia Zapponi, Francesco M. Angelici, Luigi Boitani, and Michele Cento. Weasel mustela nivalis spatial ranging behaviour and habitat selection in agricultural landscape. *Acta Theriologica*, 54(2):137–146, 2009.

[19] Dominique Makowski, Daniel Lüdecke, Mattan S. Ben-Shachar, Henrik Singmann, and Paul-Christian Bürkner. bayestestr: Credible intervals vignette. https://easystats.github.io/bayestestR/articles/credible_interval.html, 2025.

[20] Gianluca Mastrantonio. Slide del corso modelli statistici / apprendimento statistico. Materiale didattico, Politecnico di Torino., 2025.

[21] Théo Michelot, Roland Langrock, and Toby Patterson. *moveHMM: an R package for the analysis of animal movement data*. CRAN, April 2025.

[22] Parco Nazionale d'Abruzzo, Lazio e Molise. Donnola – mustela nivalis. https://www.parcoabruzzo.it/fauna.schede.dettaglio.php?id=326, 2018.

[23] Proj Contributors. Universal transverse mercator (utm) projection — proj documentation. https://proj4.org/en/stable/operations/projections/utm.html, 2025.

[24] T. S. Richardson. Monte carlo integration — computational statistics with r. https://cswr.nrhstat.org/mci, 2025.

[25] William J. Ripple, Christopher Wolf, Jillian W. Gregg, and Erik Joaquín Torres-Romero. Climate change threats to earth's wild animals. *BioScience*, 75(6):519–523, 2025. Advance access publication date: 20 May 2025. Special Report.

[26] Joshua S. Speagle. A conceptual introduction to markov chain monte carlo methods. *arXiv preprint arXiv:1909.12313*, 2019. Submitted on 26 Sep 2019, last revised 7 Mar 2020.

[27] Stan Development Team. *Stan Reference Manual, Version 2.18*. Stan Development Team, 2018.

[28] Walter W. Stroup. *Generalized Linear Mixed Models: Modern Concepts, Methods and Applications*. CRC Press, 2012.

[29] R Core Team. Gamma distribution — r documentation. https://rdrr.io/r/stats/GammaDist.html, 2025.

[30] Stan Development Team. *loo_compare function — R package 'loo'*, 2025. Consultato il 4 novembre 2025.

[31] Peter R. Thompson, Peter D. Harrington, Conor D. Mallory, Subhash R. Lele, Erin M. Bayne, Andrew E. Derocher, Mark A. Edwards, Mitch Campbell, and Mark A. Lewis. Simultaneous estimation of the temporal and spatial extent of animal migration using step lengths and turning angles. *Movement Ecology*, 12(1), 2024.

[32] Aki Vehtari, Andrew Gelman, Daniel Simpson, Bob Carpenter, and Paul-Christian Bürkner. Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of mcmc. *Bayesian Analysis*, 16(2):667–718, 2021.

[33] Sumio Watanabe. A widely applicable bayesian information criterion. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

[34] Eric W. Weisstein. Bernoulli distribution. https://mathworld.wolfram.com/BernoulliDistribution.html, 2025.