Report Progetto Dati Funzionali

Matteo Ceola, Paolo Magagnato, Marco Piccolo, Pietro Stangherlin

Indice

1 Introduzione

2 Obbiettivi

- Analisi esplorative funzionali
- ANOVA funzionale: confronto tra spettri di frequenze per diverse specie di uccelli

- Modello funzione su funzione: si è interessati a valutare se esistano delle relazioni tra ciascun suono emesso ed il suono precedente

3 Dati

I dati considerati sono presenti sul portale xeno-canto. Per ogni audio è disponibile la specie di uccello e le coordinate geografiche del rilevamento.

4 Operazioni preventive

- Passaggio al dominio della frequenza tramite spettro medio
- Normalizzazione delle ampiezze

4.1 Rappresentazione funzionale

4.1.1 Spline penalizzate e vincolate

Per le curve di Ampiezza in funzione della frequenza si è scelta una rappresentazione in base bspline di grado 3. Inizialmente si sono considerate due penalizzazioni: la prima sul numero di basi e la seconda sull'integrale del quadrato della derivata seconda (per un numero di basi abbastanza alto fissato), per entrambi i casi si è considerato come riferimento il parametro che minizzasse il criterio di GCV. Tuttavia questi due criteri non permettono il rispetto dei vincoli: 1) di non negatività della curva 2) di ampiezza non superiore a 1 (a causa della normalizzazione).

Per ciascuno dei due criteri sopra menzionati si sono quindi introdotti i vincoli nel problema di ottimizzazione che può essere scritto come un programma di programmazione quadratica (A1) per cui sono disponibili delle routine.

4.1.2 Risultati

Introdurre i vincoli non dà luogo ad uno stimatore lineare in y, non è quindi possibile usare GCV come criterio per la selezione dei parametri di regolazione, si impiega invece una procedura di convalida incrociata "Leave One Out" (LOO-CV). A titolo esemplificativo si riportano le curve relative ai gufi con i quattro metodi: GCV senza vincolo e LOOCV con vincolo; in @tab-representation-selection-df sono riportate le specifiche di ciascun metodo.

Tabella 1: Tabella le cui colonne da sinistra verso destra contengono rispettivamente l'animale considerato, se il metodo di stima è vincolato in (0,1) (TRUE) oppure no (FALSE), il tipo di penalità: numero intero di basi (INT) o penalità sull'integrale della derivata seconda (DIFF), il parametro ottimo selezionato (numero di basi per penalità INT e lambda per penalità DIFF) e il numero di punti (unici) del dominio.

animal	constraint	penalty.type	min.error.parameter	domain.unique.points
falchi	FALSE	INT	9.0e+01	125
falchi	FALSE	DIFF	1.2e-06	125
falchi	TRUE	INT	4.8e+01	125
falchi	TRUE	DIFF	2.1e-06	125
gufi	FALSE	INT	9.7e+01	111
gufi	FALSE	DIFF	2.0e-07	111
gufi	TRUE	INT	2.3e+01	111
gufi	TRUE	DIFF	1.0e-07	111
gabbiani	FALSE	INT	9.0e+01	111
gabbiani	FALSE	DIFF	4.0e-07	111
gabbiani	TRUE	INT	2.0e+01	111
gabbiani	TRUE	DIFF	6.2e-06	111

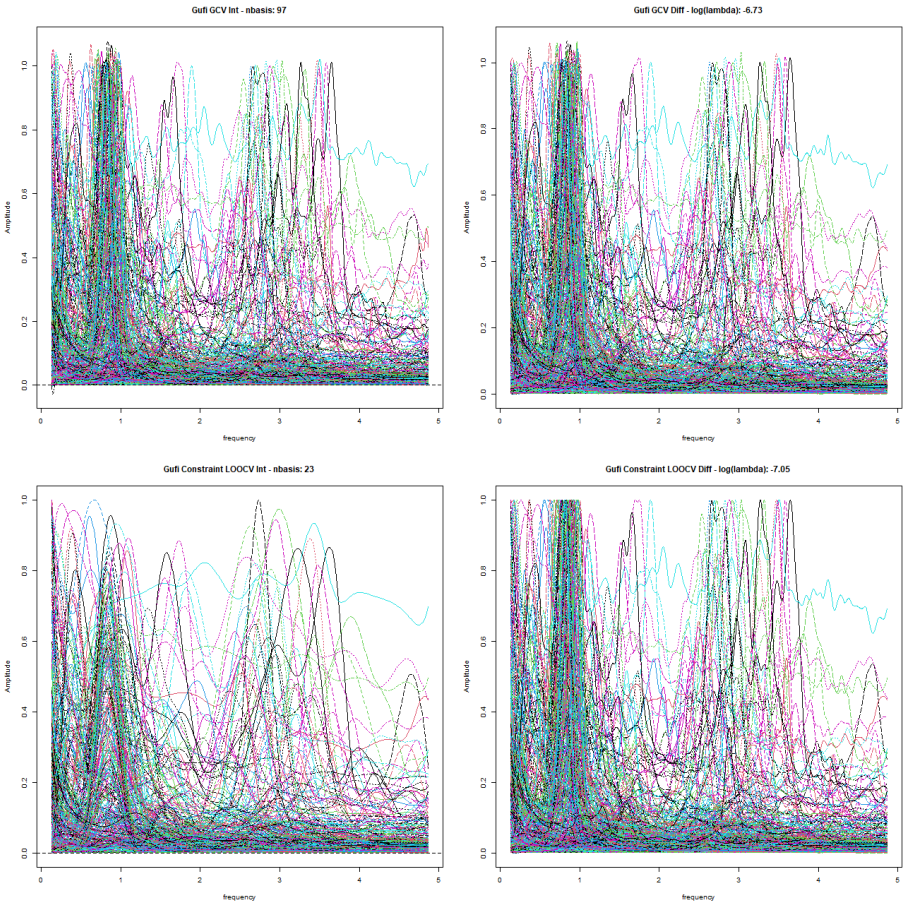

Figura 1: Rappresentazioni funzionali tramite funzioni di base degli spettrogrammi medi dei suoni dei gufi. In alto i criteri non vincolati: a sinistra con il numero di basi selezionate tramite GCV per non penalizzato, a destra con il lambda selezionato con penalità sulla derivata seconda. In basso i criteri vincolati: a sinistra con il numero di basi selezionate tramite convalida incrociata LOOCV per il criterio non penalizzato e a destra con il lambda ottimo con penalità sulla derivata seconda.

Esaminando la Figura 1 si evidenziano diverse problematiche:

- nel caso senza vincoli e con penalizzazione solo sul numero di basi (grafico in alto a sinistra) si osserva che non sono rispettati i vincoli di non negatività e di ampiezza inferiore ad uno, problema presente anche con penalizzazione sull'integrale della derivata seconda al quadrato (grafico in alto a destra).

- i vincoli migliorano chiaramente la rappresentazione funzionale, tuttavia, il numero di basi che minimizza LOOCV (grafico in basso a sinistra) è probabilmente troppo piccolo in quanto alcune funzioni hanno dei picchi troppo bassi, anche qui si potrebbe pensare di aumentare il numero di basi; nell'ultimo caso (penalizzaione sull'integrale della derivata seconda al quadrato) (grafico in basso a destra) una possibile critica è che le funzioni non siano abbastanza lisce.

Risultati simili si hanno anche con le altre due specie. Dato che i criteri di selezione automatica proposti hanno mostrato le problematiche sopra descritte si è deciso di adottare un'euristica in maniera tale che le curve mostrassero un discreto adattamento e al contempo fossero abbastanza lisce. Dopo una serie di prove si sono considerate le funzioni vincolate con penalità sulla derivata seconda riducendo il numero di basi a 70 per tutte le specie e adottando il criterio vincolato con penalità sulla derivata seconda selezionando il parametro di regolazione tramite LOOCV, in Figura 2 i grafici delle rappresentazioni in base.

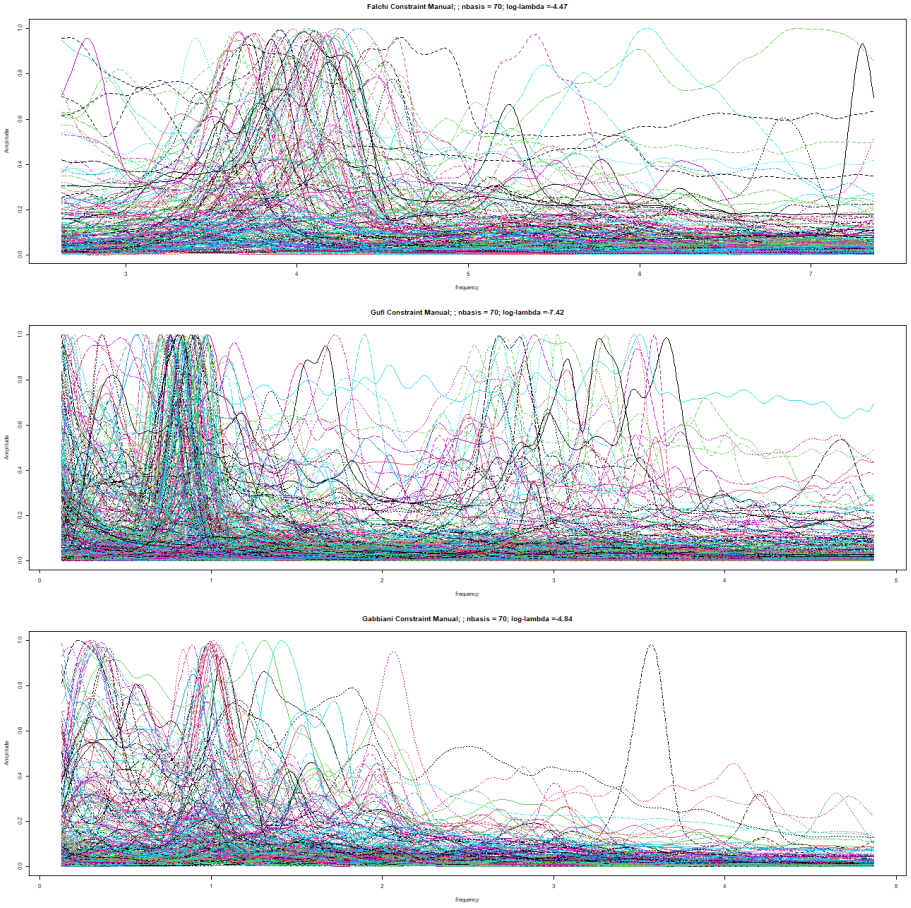

Figura 2: Rappresentazione in basi splines vincolate in (0,1) in cui per ogni specie è indicato il numero di basi scelto euristicamente e il parametro di penalità sull'integrale del quadrato della derivata seconda selezionato tramite convalida incrociata leave-one-out (LOOCV).

5 Medie funzionali

In Figura 3 si riportano le medie e le deviazioni standard funzionali per tutti i gruppi di ciascun animale. Le deviazioni standard sono circa dello stesso ordine delle medie per quasi tutti i gruppi e tutti i punti.

Per i falchi le medie e le deviazioni standard dei gruppi hot e temperate appaiono vicine nel range di frequenze tra 3khz e 4.5khz, mentre fuori da questo intervallo presentano differenze sia in media che in varianza, rimandendo comunque sempre sopra la curva media del gruppo cold che mostra un primo picco d'ampiezza ad una frequenza più bassa rispetto agli altri due gruppi.

Come per i falchi anche per i gufi le curve medie nei tre gruppi sembrano seguire un andamente comune, così come le deviazioni standard, qui le due curve più simili sono invece quelle dei gruppi cold e temperate, in queso caso la curva media che presenta un picco di ampiezza "anticipato" rispetto alle altre due è quella relativa al gruppo Hot.

Nei gabbiani la curva media che si discosta dalle altre è quella relativa al gruppo delle Canarie che domina le altre curve nelle frequenze basse, mentre per quelle più alta è sotto tutte le altre curve medie. Si nota inolte un picco nella funzione media e nella deviazione standard in prossimità dei 3.5 khz per il gruppo Centre.

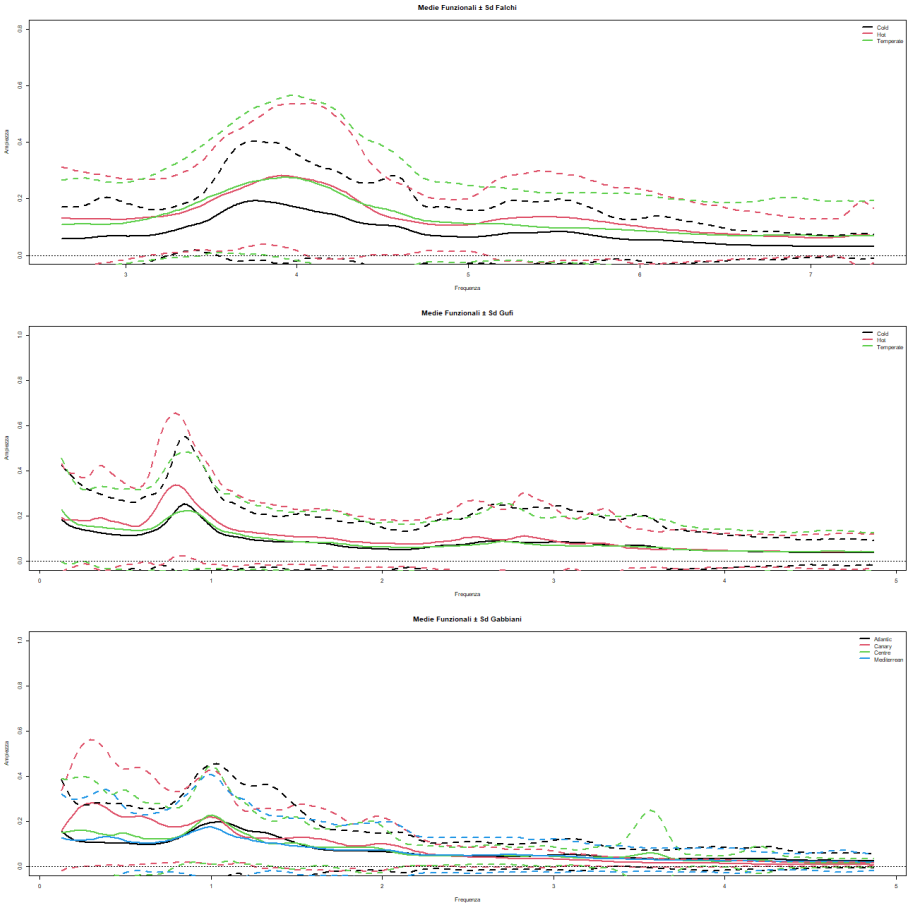

Figura 3: Medie Funzionali degli spettrogrammi medi dei suoni degli animali per i diversi gruppi a cui è aggiunto e sottrato l'errore standard funzionale. Dall'alto verso il basso falchi, gufi e gabbiani.

6 PCA funzionale

Si effettua un'analisi delle componenti principali funzionali, sia per avere delle ulteriori informazioni descrittive sia per vedere se i punteggi di tali componenti individuano o meno dei cluster di osservazioni potenzialmente diversi da quelli scelti in questa sede.

Inizialmente si fissa un numero di armoniche pari a 10 (Figura 4), per i falchi le prime 3 componenti spiegano circa l'80% della varianza, per i gufi sono necessarie 4 componenti per spiegare circa l'80% della varianza, mentre per i gabbiani l'85% della varianza è spiegato dalle prime tre componenti.

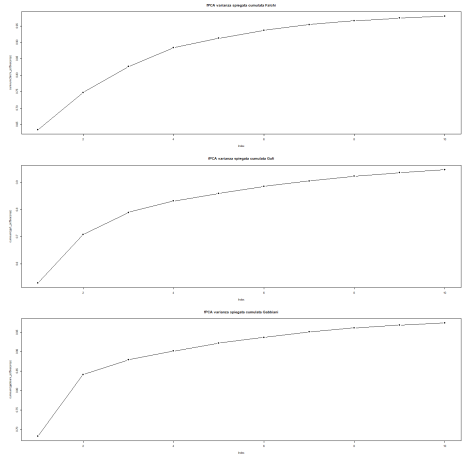

Figura 4: Varianza spiegata cumulata per le prime 10 componenti funzionali principali degli spettrogrammi dei suoni (dall'alto in basso) di falchi, gufi e gabbiani.

Focalizzandosi sulle prime tre armoniche (Figura 5) si può vedere che, per tutti gli animali, la prima armonica segue abbastanza regolarmente la forma della media funzionale, la seconda presenta un andamento opposto alla media e la terza varia molto tra i diversi animali e non è di facile intepretazione.

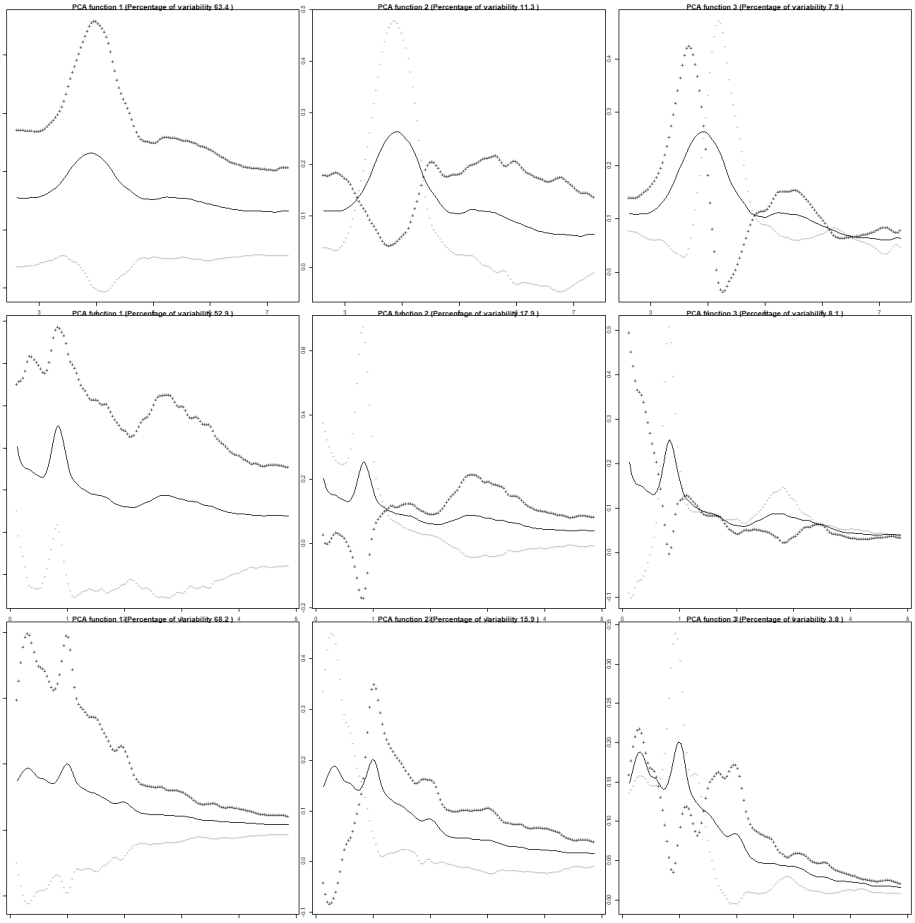

Figura 5: Prime tre armoniche (da sinistra verso destra) con piccoli scostamenti dalla media relative alle componenti principali funzionali degli spettrogrammi dei suoni (dall'alto in basso) di falchi, gufi e gabbiani.

Considerando le prime due armoniche (ma un risultato analogo si ottiene anche con le prime tre) e rappresentando graficamente i punteggi (Figura 6) non si distingue nessun cluster di punti.

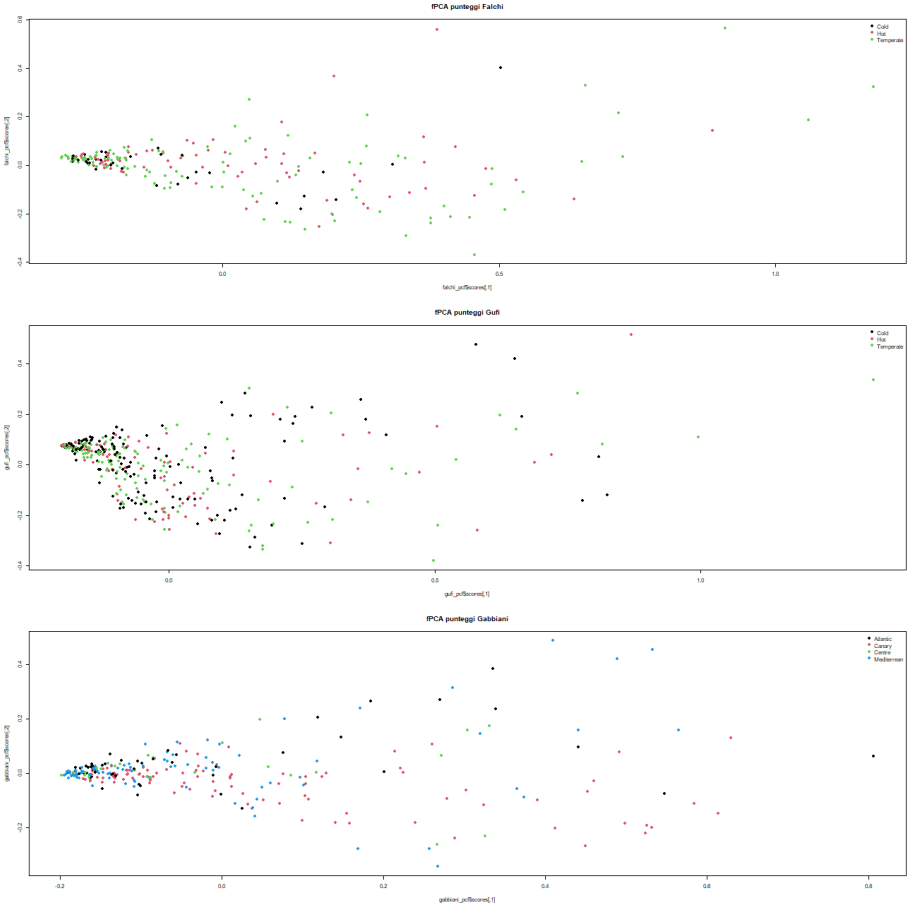

Figura 6: Punteggi delle prime due armoniche relative alle componenti principali funzionali degli spettrogrammi dei suoni (dall'alto in basso) di falchi, gufi e gabbiani.

7 ANOVA funzionale

Per ciascun animale si considera un modello di ANOVA funzionale:

$$y(f) = \alpha(f) + \sum_{g=1}^{G} \beta_g(f) + \epsilon(f) \text{ s.t. } \sum_{g=1}^{G} \beta_g(f) = 0 \ \forall f \tag{1}$$

Dove $y(f)$ è l'ampiezza in funzione della frequenza, $\alpha(f)$ è l'intercetta funzionale, $\beta_g(f)$ è il coefficiente associato al gruppo g-esimo, si impone il vincolo di somma a zero per l'identificabilità e $\epsilon(f)$ è il termine di errore omoschedasti-

co. Per la rappresentazione funzionale di intercetta e coefficienti si utilizzano le stesse basi usate per rappresentare $y(f)$ e descritte nei punti precedenti, qui non si impongono vincoli ma utilizza una penalità sull'integrale del quadrato della derivata seconda per tutti i coefficienti ad eccezione dell'intercetta; per la selezione del parametro di regolazione λ (per semplicità comune a tutti i gruppi) si considera il valore minimo di errore integrato di convalida incrociata a 5 fold (ogni insieme di convalida è campionato per strati in modo tale da mantenere la medesima proporzione di osservazioni per ciascun gruppo presente nei dati completi).

Per una valutazione dell'incertezza sui coefficienti, senza assumere l'omoschedasticità del termine d'errore, si impiega una procedura di bootstrap non parametrico (stratificato per gruppo) i cui risultati sono visibili in Figura 7 (per completezza si riporta anche la Figura 9 in cui le bande intorno ai coefficienti sono ottenute tramite la stima analitica dell'errore standard funzionale, il risultato appare in buon accordo con quello ottenuto con le bande bootstrap); per i falchi il criterio induce dei coefficienti molto lisci rispetto agli altri due animali, ciò è probabilmente dovuto ad una maggiore variabilità delle curve all'interno di ciascun gruppo dei falchi in confronto agli altri due casi.

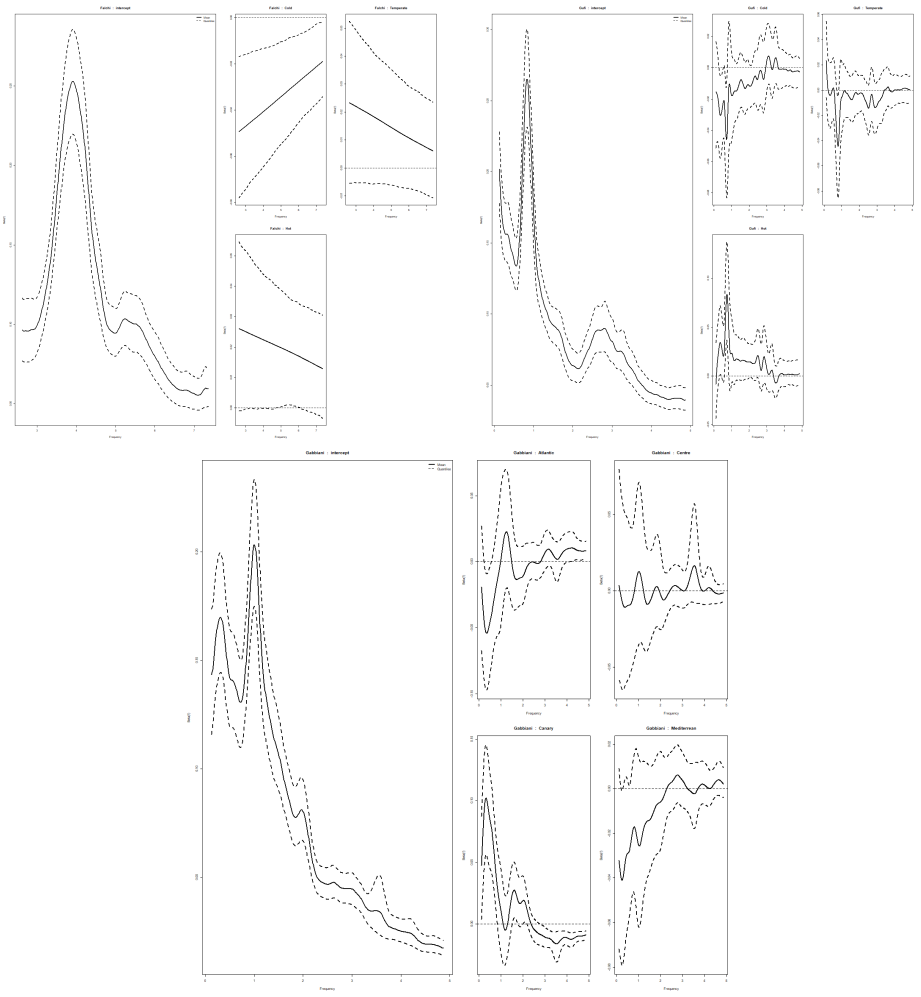

Figura 7: Coefficienti del modelli di ANOVA funzionale, le bande intorno a ciascun coefficiente funzionale sono intervalli puntuali bootstrap (stratificato per gruppi) (1000 campioni bootstrap) relativi ai percentili 2.5% e 97.5%. In alto da sinistra verso destra i modelli per falchi e gufi, in basso per gabbiani.

Per saggiare l'ipotesi nulla di nullità di tutti i coefficienti associati ai gruppi si eseguo un test di permutazione basato sulla statistica F funzionale considerando i percentili al 95% (Figura 8).

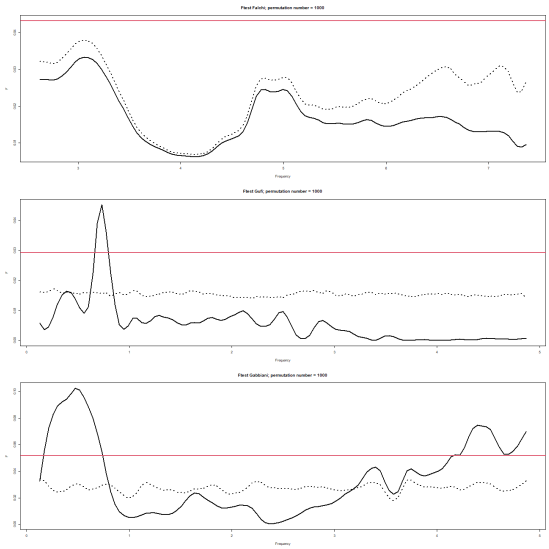

Figura 8: Grafici associati ai test F funzionali per ciascun modello di ANOVA funzionale. La linea continua rappresenta la statistica osservata, quella trattettagia il percentile puntuale al 95% della statistica F sotto l'ipotesi nulla stimato tramite permutazione (1000 permutazioni) e la linea rossa è il percentile al 95% dei massimi delle statistiche F sotto l'ipotesi nulla.

Per i falchi i coefficienti relativi ai climi Temperate e Hot sono positivi e mostrano degli andamenti decrescenti (nella frequenza) con delle bande che comprendono lo zero per tutte le frequenze relativamente a Temperate, e lo incrociano in alcuni punti del dominio per Hot, quello relativo a Cold invece è sempre negativo crescente e le sue bande sono sempre sotto lo zero; ciò è in accordo con le analisi esplorative: le curve del gruppo Cold presentano ampiezze inferiori a quelle degli altri due gruppi e la differenza è più marcate per basse frequenze. Sembra quindi che climi freddi siano correlati con una minore presenza di versi a frequenze basse (rispetto alla media), tuttavia questa ipotetica differenza non trova riscontro nel test F che non rifiuta l'ipotesi nulla di uguaglianza dei coefficienti.

Nei gufi, per frequenze medio-alte tute le bande relative ai coefficienti comprendono lo zero, poco sotto 1khz invece i gruppi Cold e Temperate presentano dei picchi negativi in dove le bande non includono lo zero (con maggiore evidenza per il gruppo Cold), il coefficiente per il gruppo Hot presenta invece un picco positivo in cui le bande non includono lo zero, questa differenza è evidenziata dal rifiuto del test F proprio in prossimità dei picchi. Secondo il modello dunque climi caldi sono correlati con una maggiore presenza di frequenze basse.

Nel caso dei falchi, per il coefficiente relativo al gruppo Centre non si evidenziano andamenti particolari e le bande racchiudono sempre lo zero; per i gruppi

Atlantic e Mediterrean si hanno delle curve inizialmente negative e crescenti con bande che presentano una regione totalmente negativa per Atlantic e vicina a zero per Meditterean, le due curve presentano poi dei picchi positivi e un successivo calo seguito da una fase di stabilizzazione, il tutto con una differenza di fase, per frequenze medio alte le bande racchiudono sempre lo zero ad eccezione degli estremi relativi ad Atlantic. L'andamento più singolare è quello del coefficiente per Canary che presenta un intervallo iniziale in cui è positivo decrescente con relative bande sopra lo zero, per poi presentare un successivo picco sempre con bande di poco positive ed infine un'ultima fase con bande sotto lo zero. Il test F rifiuta l'ipotesi di uguaglianza dei coefficienti per frequenze sia basse che alte ad dovuti a Canary e probabilmente ad Atlantic. Il modello descrive, per i gabbiani delle Canarie, un impiego maggiore di frequenze basse e uno minore di frequenze alte rispetto ai gabbiani delle altre regioni.

8 Modello funzione su funzione

9 Conclusioni

10 Appendice

10.1 A1: Splines vincolate penalizzate

Il programma di ottimizzazione quadratica nella sua forma più generale è definito come

$$\min_b(-d^T b + 1/2 b^T D b) \text{ s.t. } A^T b \geq b_0 \tag{2}$$

Per una singola osservazione funzionale, per il lisciamento tramite scrittura in funzioni di base la funzione da minimizzare rispetto a b è

$$(y - \Phi b)^T (y - \Phi b) + \lambda b^T P b \tag{3}$$

dove y è il vettore dei punti osservati Φ è la matrice delle funzioni di base valutate nei punti osservati del dominio della curva e b è il vettore dei coefficienti, P è una generica matrice di penalità e $\lambda > 0$ indica l'entità della penalizzaione (per un criterio non penalizzato è sufficiente porre $\lambda = 0$). Minizzare l' Equazione 3 equivale a minizzare

$$-y^T \Phi b + b^T \frac{1}{2}(\Phi^T \Phi + \lambda P) b \tag{4}$$

Da cui $d = y^T \Phi$ e $D = \Phi^T \Phi + \lambda P$. Usando la definizione di scrittura in basi il vincolo è $0 \leq \phi^T(f) b \leq 1 \quad \forall f$, in pratica si discretizza $\phi(f_j)$ per $j = 1, .., J$. Sia Φ_J la matrice di funzioni di base valuate su griglia discretizzata: deve valere $\Phi_J b \leq \mathbb{1}$ o equivalentemente $-\Phi_J b \geq -\mathbb{1}$ dove con questa scrittura si intende

che la disuguaglianza deve valere per ogni elemento dei vettori. Similmente, per vincolo di positività si ha $\Phi_J b \geq \mathbb{0}$, tuttavia, poichè per costruzione le basi bspline sono sempre non negative è sufficiente imporre $\mathbb{I}_J b \geq \mathbb{0}$ con $\mathbb{I}_J$ matrice identità. Combinando le due espressioni si ottiene

$$\begin{pmatrix} -\Phi \\ \mathbb{I}_J \end{pmatrix} b \geq \begin{pmatrix} -\mathbb{1} \\ \mathbb{0} \end{pmatrix}$$

chiaramente $A = (-\Phi^T \quad \mathbb{I}_j)$ e $b_0 = (-\mathbb{1} \quad \mathbb{0})^T$, è dunque conclusa la scrittura del problema vincolato come programma quadratico.

10.2 A2: Grafici coefficienti funzionali

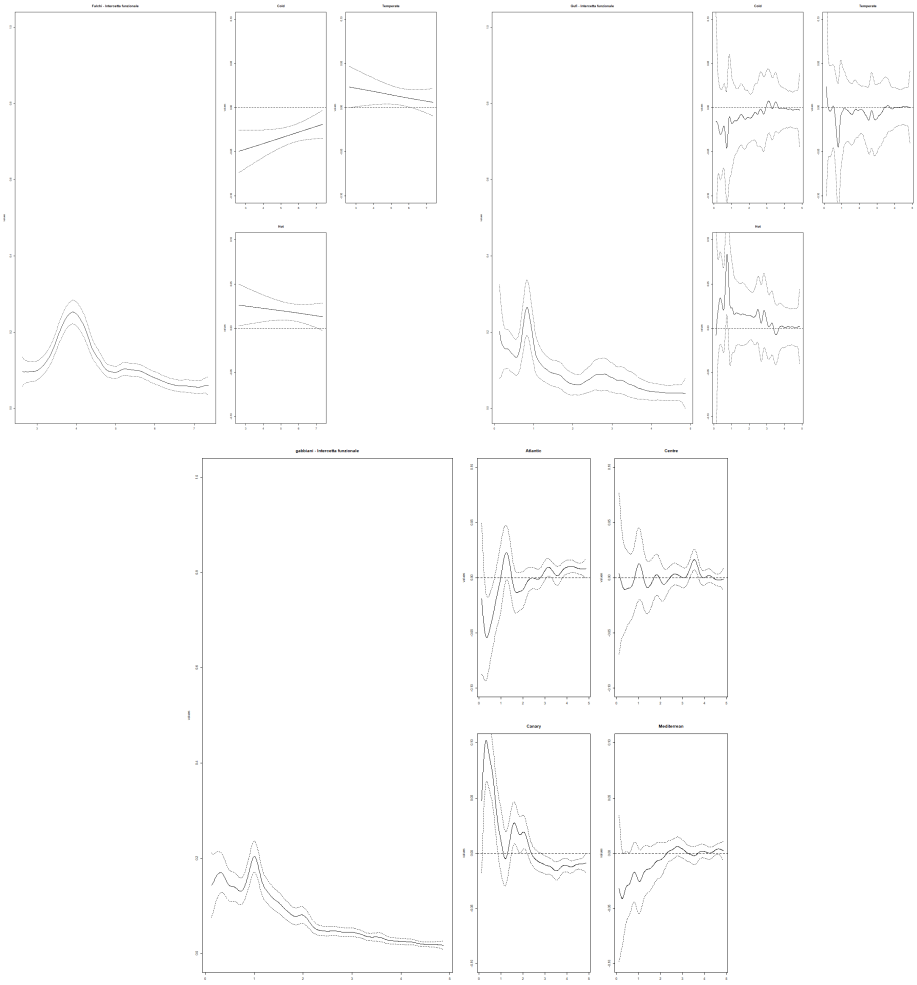

Figura 9: Coefficienti del modelli di ANOVA funzionale, le bande intorno a ciascun coefficiente funzionale sono ottenute considerando il coefficiente funzionale aggiungendo e sottrendo due volte l'errore standard funzionale. In alto da sinistra verso destra i modelli per falchi e gufi, in basso per gabbiani.