UNIVERSITA' DEGLI STUDI DI PADOVA

Facoltà di Scienze Statistiche

Corso di laurea in statistica e gestione delle imprese

Tesi di Laurea

BIODIVERSITA' VEGETALE : UNO STUDIO TRAMITE MODELLI LINEARI GENERALIZZATI

VEGETABLE BIODIVERSITY :

A STUDY TROUGH GENERALIZED LINEAR MODELS

Relatore: Ch.ma Dott.ssa Monica Chiogna

Laureando: Nicola Borella

N° Matricola 545270-GEI

Anno Accademico 2008-2009

Indice

1 Introduzione 1

1.1 scopo 1
1.2 La biodiversità 2
1.3 La ricchezza di specie 3

2 Estrapolazione dei dati 5

3 La fase esplorativa e la costruzione delle variabili 9

3.1Inroduzione 9
3.2 La variabile risposta 9
3.3 Le variabili esplicative 15
3.3.1 Variabile anno 15
3.3.2 Variabile temp.summer 16
3.3.3 Variabile ppt.summer 17
3.3.4 Variabile Spazio 18
3.3.5 Variabile Freddo 19
3.3.6 Variabile Caldo 19
3.3.7Variabile Indice posto 20
3.3.8 Variabile quadX,quadY 21

3.3.9 Variabile Caldo.mean 21

4 Una prima analisi tramite modelli lineari 23

4.1 Introduzione 23
4.2 Modello 1 23
4.3 Modello 2 25

5 Modelli lineari generalizzati 27

5.1 Introduzione 27
5.2 Modello 1 27
5.3 Modello 2. 29
5.4 Modello 3 31
5.5 Conclusioni 32

6 Un ulteriore approfondimento 33
7 Conclusioni 37
Appendice A 41

Appendice B 43

Appendice C 47

Bibliografia 51

Capitolo 1

Introduzione

1.1 Scopo

Scopo del presente lavoro è utilizzare alcuni strumenti statistici per l'analisi di problemi rilevanti in campo ambientale, concentrandosi in particolare sullo studio della biodiversità misurata tramite la "ricchezza di specie".

Verranno utilizzati modelli lineari generalizzati che saranno applicati allo studio del numero di specie vegetali, per i quali daremo, ogni volta che risulti necessario, le spiegazioni opportune.

Scopo dell'analisi è verificare in che modo alcune variabili esplicative spiegano al ricchezza di specie.

Una volta esaminato il fenomeno e analizzati i modelli opportuni, si cercherà di interpretare in che modo le variabili, che ci siamo creati, spiegano la ricchezza di biodiversità e in che modo intervenire sulla biodiversità agendo sulle variabili che noi possiamo controllare.

1.2 La biodiversità

La biodiversità esprime la varietà delle forme viventi in un ambiente. Essa indica una misura della varietà di specie animali e vegetali nella biosfera, ed è il risultato di lunghi processi evolutivi. L'evoluzione è il meccanismo che da oltre tre miliardi di anni permette alla vita di adattarsi al variare delle condizioni sulla terra e che deve continuare a operare perché questa possa ancora ospitare forme di vita in futuro.

La biodiversità viene in genere studiata a tre diversi livelli, che corrispondono a tre livelli di organizzazione del mondo vivente: quello dei geni, quello delle specie e quello degli ecosistemi. Nel nostro studio analizzeremo la biodiversità a livello di specie. La specie viene definita come l'insieme degli organismi che possiedono caratteri simili, capaci di incrociarsi fra loro e di dare luogo a una prole fertile. Essa rappresenta l'unità di base della classificazione tassonomica e viene presa come riferimento in molti studi biologici e in molte ricerche condotte dagli ecologi.

La conoscenza della biodiversità rappresenta un punto di riferimento fondamentale per gli studi ecologici e per la pianificazione degli interventi di conservazione della diversità biologica.

Nel corso delle indagini sulla biodiversità, infatti, viene spesso contato il numero delle specie maggiormente conosciute (ad esempio di mammiferi, di uccelli e di determinate specie vegetali), presenti in un'area prescelta come campione. Il conteggio delle specie prende il nome di censimento.

Esistono diverse modalità per effettuare il censimento delle specie: il conteggio assoluto di tutti gli esemplari che vivono in una certa zona "perlopiù impraticabile", perché spesso le aree da considerare sono molto vaste, oppure particolarmente impervie e, ancora, non si può avere la certezza di conteggiare tutti gli individui, soprattutto in specie terricole di

piccole dimensioni (come insetti, molluschi o vermi). In genere, si procede con conteggi parziali, dai quali si effettua una stima del numero totale di individui per specie: ad esempio, per valutare la presenza di una certa pianta in un'area, si può contare il numero di esemplari presenti entro un metro quadrato, scelto a caso, dai quali si desumono le specie presenti. Il criterio di scelta casuale è importante perché i risultati delle osservazioni fatte in alcune aree-campione devono essere validi da un punto di vista statistico.
In questo lavoro noi abbiamo un conteggio assoluto delle specie vegetali presenti nell'area campionaria, definita da un predeterminato appezzamento di terreno dell'università americana di Fort Knox. I ricercatori di tale università per più di quarant'anni si sono alternati nel conteggio e nella raccolta di dati relativi alle specie presenti,svolgendo un immenso lavoro, condotto secondo il rispetto di certi criteri campionari e avente come esito un censimento di specie vegetali osservate.

1.3 La ricchezza di specie

Il numero delle specie viene definito con l'espressione "ricchezza di specie", e costituisce una delle possibili misure della biodiversità di un luogo.
La ricchezza di specie viene considerata come la misura generale di biodiversità più semplice e facile da valutare, anche se non può che rappresentare una stima approssimativa e incompleta della variabilità presente tra i viventi. Una cosa interessante da capire è come varia la ricchezza di specie nel tempo e nello spazio.
La ricchezza di specie varia geograficamente: nei climi caldi in genere vive un maggior numero di specie, rispetto a quelli freddi, così come nelle zone più umide vi sono più specie che in quelle più secche.

CAPITOLO 1. INTRODUZIONE

Le zone tropicali, in cui crescono le foreste pluviali, sono tra le regioni a più alto indice di biodiversità.

Le zone con minori variazioni stagionali accolgono più specie di quelle con stagioni ben definite nelle zone con topografia e clima diversificati albergano più specie di quelle uniformi.

Alcune particolari specie possono, per il ruolo che esse rivestono nell'ecosistema, aumentare la biodiversità. Ad esempio, specie arboree che raggiungono elevate altezze offrono una grande quantità di risorse utili a molte specie diverse (come uccelli nidificanti, piante epifite, parassiti, erbivori frugivori) e contribuiscono, quindi, alla biodiversità di quell'ambiente. Tuttavia, non esiste ancora un metodo per valutare in modo quantitativo l'importanza di questi ruoli e confrontare i valori ottenuti nei diversi gruppi.

Consideriamo quanto detto come un punto di partenza per la costruzione successiva delle variabili esplicative nello studio seguente.

Capitolo 2

Estrapolazione dei dati

I dati relativi al numero di specie provengono dal sito internet www.esapubs.org che li ha messi pubblicamente a disposizione. Questo è il sito di un ente americano che effettua vari studi sull'ecologia americana e mondiale che pubblica molti dataset.

I dati, si presentano in un formato non direttamente leggibile da usuali programmi di manipolazione di dati come Excel o OpenOffice, ma data la loro dimensione sono leggibili solo con un editor di testo.

In una prima fase è stato necessario copiare tutti i dati dall'editor di testo in due fogli elettronici di lavoro in modo da poterli rimodellare con semplicità e velocità. A questo punto i dati si presentavano in diverse colonne di variabili, come si può vedere in *Figura 1.* Nella prima colonna è esplicitato il numero progressivo dei dati, la seconda riporta la sigla "quadrante-anno" che rappresenta sia il quadrante di rilevazione, indicato con i primi cinque caratteri alfanumerici, sia l'anno di raccolta indicato dalle ultime due cifre. Quindi ad esempio, la sigla "e1q1-132" sta a dire che il dato è relativo al 1932 e fa riferimento al quadrante e1q1-1.

La terza colonna contiene l'identificativo del rilevatore, mentre la quarta contiene il nome della specie rilevata, spesso ripetuta più di una volta all'interno dello stesso anno. Le altre variabili: Area, x, y, rappresentano rispettivamente l'area fisica occupata dalla pianta sulla superficie, la posizione assunta in quel quadrante della pianta nell'asse x e nell'asse y.

CAPITOLO 2. ESTRAPOLAZIONE DEI DATI

.

Figura 1
Disposizione dei dati in OpenOffice

Numero	Plotyear	ID	Species	Area	x	y
1	e1q1-132	1	Short grass	9588.16.00	1100.55.00	452.35.00
2	e1q1-132	2	Bare ground	782.32.00	541.57.00	126.47.00
3	e1q1-132	3	Bare ground	372.19.00	274.26.00	708.41.00
4	e1q1-132	4	Bare ground	350.41.00	762.37.00	1104.44.00
5	e1q1-132	6	Ambrosia psilostachya	462.49.00	391.07.00	389.16.00
6	e1q1-132	7	Bare ground	560.29.00	512.07.00	1116.59.00
7	e1q1-132	8	Ambrosia psilostachya	93.24.00	937.00.00	365.57.00
8	e1q1-132	9	Bare ground	589.30.00	1103.05.00	838.27.00
9	e1q1-132	10	Gutierrezia sarothrae	39.44.00	290.03.00	673.09.00
10	e1q1-132	11	Bare ground	862.55.00	873.13.00	591.23.00
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
65533	e1q4a-232	21	Schizachyrium scoparium	226.30.00	454.15.00	27.25.00
65534	e1q4a-232	22	Schizachyrium scoparium	1207.42.00	908.54.00	215.38.00
65535	e1q4a-232	23	Schizachyrium scoparium	984.40.00	1069.52.00	628.57.00
65536	e1q4a-232	26	Schizachyrium scoparium	957.58.00	826.28.00	889.06.00
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
82119	e1q5-472	72	Bouteloua curtipendula	424.36.00	77.57.00	844.46.00
82120	e1q5-472	69	Cirsium undulatum	305.24.00	311.29.00	282.38.00
82121	e1q5-472	2	Bare ground	5217.18.00	117.21.00	382.54.00
82122	e1q5-472	78	Andropogon gerardii	253.44.00	184.03.00	248.23.00

Sigla indicante il quadrante

Sigla indicante l'anno

Per creare la ricchezza di specie, era necessario contare quanti tipi diversi di specie erano presenti in ogni singolo quadrante e in ogni singolo anno ricavandoci così una serie di conteggi.

Per fare questo, prima di tutto sono stati riordinati i dati dello stesso anno e stesso quadrante, usando come discriminante la variabile specie, in modo da mettere tutte le specie in ordine alfabetico. A questo punto sono state contate le diverse specie presenti in quell'anno all'interno di un singolo quadrante.

Se il nome di una specie era presente più di una volta essa veniva conteggiata sempre un'unica volta. Contate le specie i dati sono stati inseriti

direttamente all'interno di una matrice fatta come quella rappresentata nella *Figura2*, la quale conteneva il quadrante nell'asse verticale e l'anno nell'asse orizzontale.

Figura2
Conteggi della ricchezza di specie

Anno\ Quadrante	1932	....	.....	1972
e1q1-1	N° specie	.....	.....	N° specie
.....	.....	.....	.....	.....
e1q5-4	N° specie	.....	.....	N° specie

La fase successiva è stata quella di portare i dati dalla forma sopra indicata , utile per fare le prime analisi sia temporali che spaziali, ad una matrice di dati che contenesse in ogni riga una sola osservazione e che quindi potesse essere utilizzata ai fini dell'analisi statistica. La matrice dei dati assume la forma rappresentata nella *Figura 3.*

Ora i dati appaiono in forma utile per essere importati in uno dei programmi statistici utilizzabili per le analisi.

Figura 3

Matrice dei dati

Specie	anno	ppt.summer	.....	.....	spazio
5	32	21.22167	...	...	I
8	33	22.05333	...	...	I
...	...	.....	...	...	...
7	72	19.57500	...	...	I
...	32	21.22167	...	...	II
12	...	.....	...	...	...
...	...	.....	...	...	...
10	72	19.57500	...	...	IV

Tra i vari pacchetti statistici, è stato utilizzato R-project, scaricabile dal sito (www.R-project.org). R-project è un programma OpenOffice ed e quindi utilizzabile da tutti coloro che lo vogliono in modo freeware.

Capitolo 3

La fase esplorativa e la costruzione delle variabili

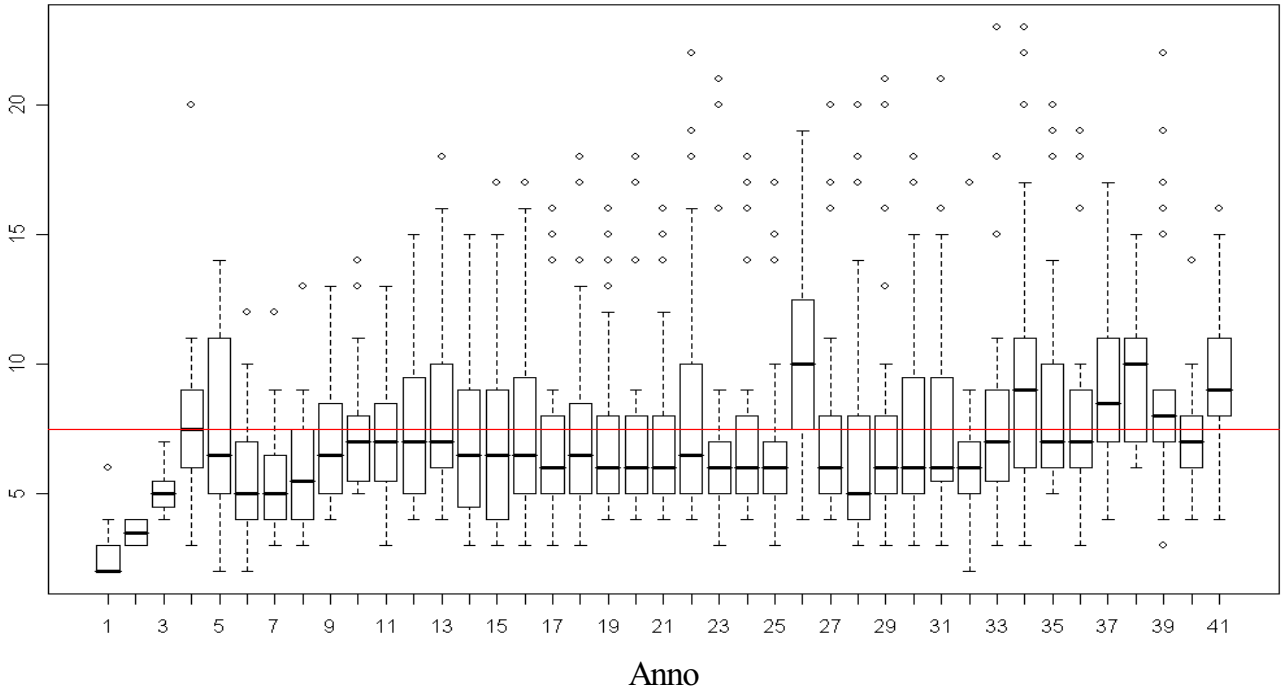
3.1 Introduzione

Questo capitolo contiene l'analisi esplorativa preliminare. Questa fase ci serve per farci una prima idea su come le variabili spiegano il fenomeno, la ricerca di possibili fattori che influenzano il fenomeno e su come i possibili fattori influenzano il fenomeno. Oltre a ciò, in questo capitolo analizzeremo l'andamento temporale della ricchezza di specie e accerteremo l'eventuale presenza di dati anomali.

3.2 La variabile risposta

La variabile risposta è la ricchezza di specie o biodiversità. Un primo grafico importante da analizzare è quello che contiene i vari diagrammi a scatola riferiti al numero di specie nel variare degli anni, rappresentato graficamente nella *Figura 4.*

Figura 4
Distribuzione del numero di specie per anno
(1 = 1932 , 41 = 1972)



Osservando i diagrammi a scatola del numero di specie esistenti in ogni diverso anno si può notare marcata asimmetria; oltre a questo, si osserva anche che alcuni particolari anni hanno media e varianza diversa dagli altri.
Nella *Figura 4,* la linea rossa passa ad una altezza pari alla media globale delle specie vegetali presenti in ogni anno, pari a 7,5 e aiuta a vedere subito qual'è l'andamento generale. Le medie, con il passare degli anni, mostrano un trend crescente, soprattutto nei primi anni, ma bisogna tenere presente che in questi anni si hanno poche osservazioni per anno.
La tabella di dati riportati in *Figura 5* mette in evidenza quanti quadranti sono disponibili per ogni anno sul totale di 48.

Figura 5

Relazione tra l'anno di raccolta e i quadranti presi in considerazione

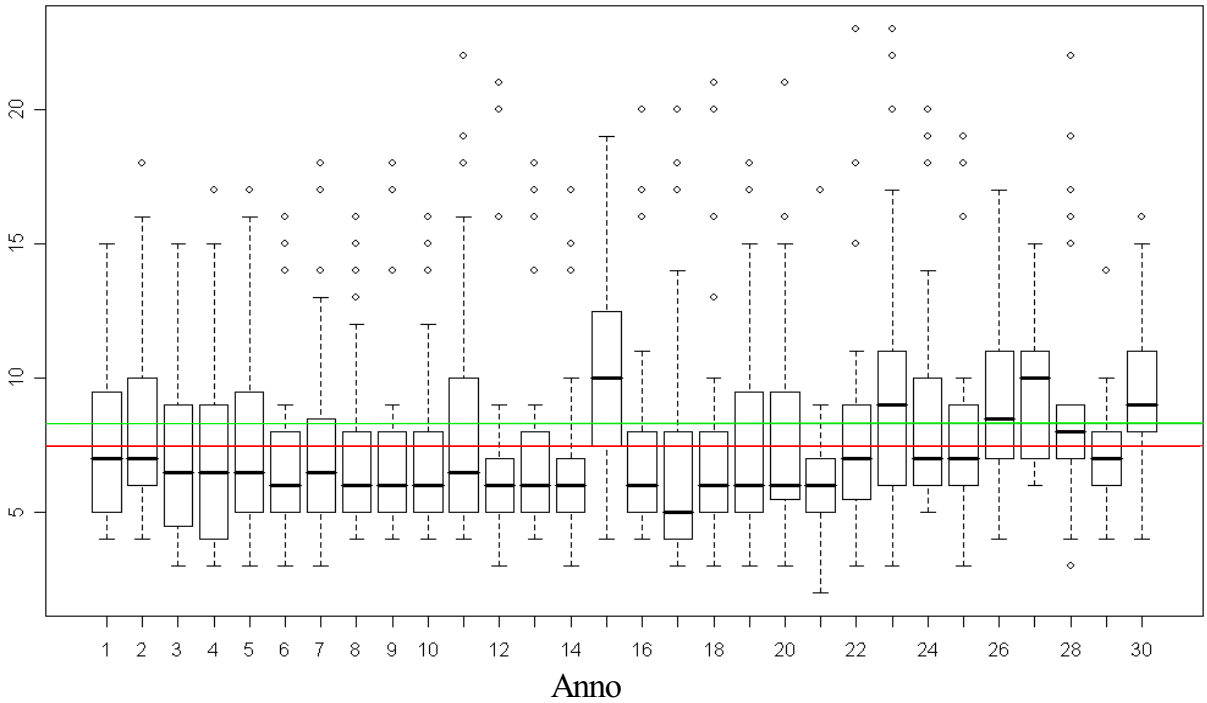
1932	1933	1934	1935	1936	1937	1938	1939	1940	1941	
14	6	20	20	18	19	20	20	20	20	
1942	1943	1944	1945	1946	1947	1948	1949	1950	1951	
20	28	28	28	28	28	28	28	28	28	
1952	1953	1954	1955	1956	1957	1958	1959	1960	1961	
27	28	28	28	28	28	28	28	28	28	
1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972
27	25	27	27	28	28	28	16	26	27	28

Visto che nei primi anni non sono disponibili tutti i quadranti, è stata riprodotta la *Figura 4* partendo dall'undicesimo anno nel quale la situazione dei quadranti conteggiati si stabilizza e, per quasi tutti gli anni, abbiamo il loro completto conteggio. Il nuovo grafico è rappresentato in *Figura 6.*

A questo fatto bisognerà prestare attenzione in futuro quando si analizzerà il fenomeno.

Figura 6

Distribuzione del numero di specie per anno
(1 = 1943 , 31 = 1972)

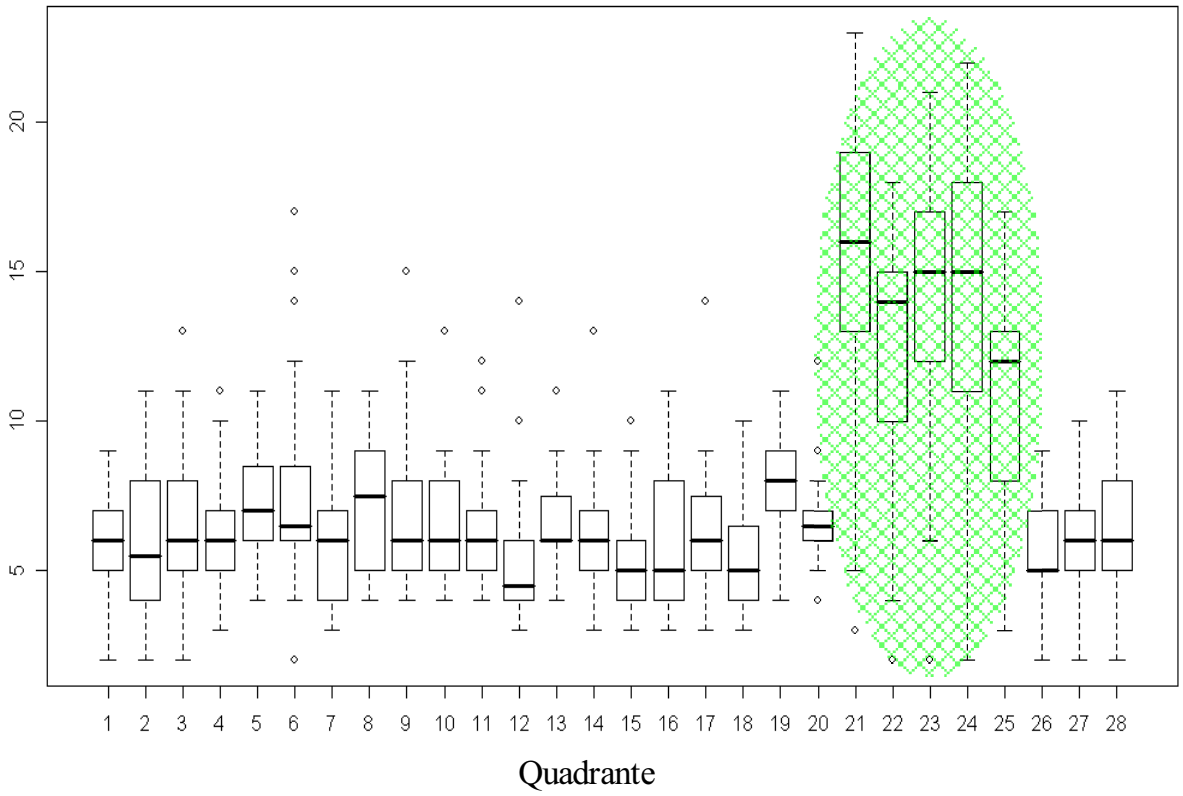


Rispetto alla *Figura 4* si può notare un andamento più regolare, sia in media che in varianza, anche se continuano ad esserci osservazioni segnalate come troppo lontane dalla media. La media del conteggio calcolata solo per questi ultimi 30 anni è pari a 8,3 ed è rappresentata della linea verde e come si vede è più alta rispetto alla media che prendeva in considerazione tutti gli anni, pari a 7,5.

Col passare degli anni, rimane un trend crescente, anche se meno evidente di prima. In *Figura 7* si considera i diagrammi a scatola dei vari quadranti, che permette di analizzare la media e la varianza dei vari quadranti da cui è stato estratto il numero di specie. Tale grafico consente di vedere se al variare dei vari quadranti le medie sono uguali o se ci sono degli andamenti o cambiamenti in funzione dei quadranti, mettendo così in gioco nell'analisi anche una componente spaziale.

Figura 7

Distribuzione del numero di specie per quadrante
(1 = e1q1-1 , 28 = e1q5-4)



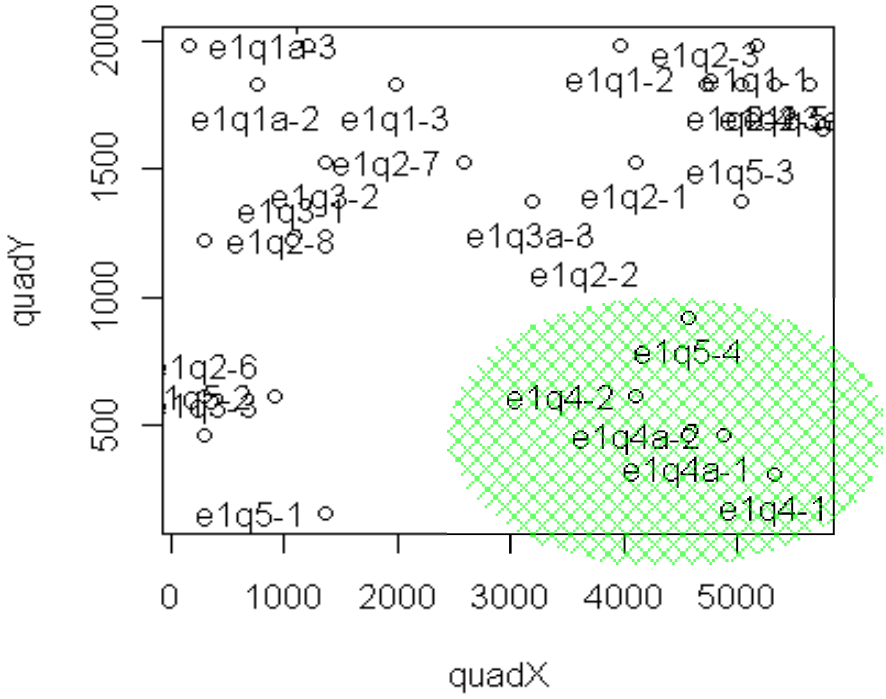
Si può notare una distribuzione abbastanza comparabile del numero di specie al variare dei quadranti, tranne che per i quadranti 21-25, compresi nell'ellisse tratteggiato, all'interno del grafico, e corrispondenti ai quadranti di terreno nell'area campionaria specificati con il nome : e1q4-1, e1q4-2, e1q4a-1, e1q4a-2,e1q5-1.

Questo grafico offre lo spunto per andare a vedere come questi quadranti, di media e varianza così diversa, siano disposti all'interno dello spazio.

Un modo per fare questo è costruire il diagramma di dispersione delle coordinate dei nostri 28 quadranti, riportato in *Figura 8,* che mostra la loro disposizione all'interno del terreno.

Figura 8

Diagramma di dispersione delle coordinate dei vari quadranti rispetto all'asse X e Y



I cinque quadranti, sotto esame con media e varianza diversa, sono vicini tra di loro, ad indicare la presenza di un effetto spaziale all'interno dall'analisi: la media e la varianza cambiano in base alla collocazione all'interno dell'area di campionamento.

Le considerazioni che emergono da queste analisi grafiche sono che la variabile risposta sembra essere influenzata dal passare degli anni e sembra ci possa essere anche una componente spaziale.

3.3 Le variabili esplicative

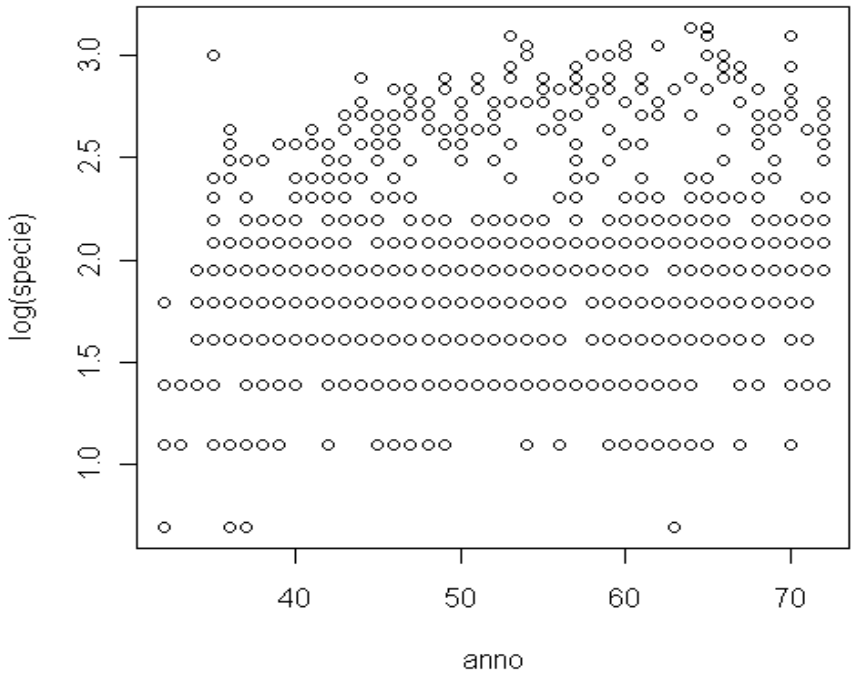
Nella parte successiva del capitolo vengono descritte nel dettaglio le variabili esplicate a disposizione.

3.3.1 Anno → Variabile che tiene in considerazione il passare degli anni, direttamente ricavata dal data set iniziale, in cui era presente sotto il nome di plotyear nella seconda colonna. Come visto nelle figure 4 e 5, sembra ci sia un trend positivo, quindi ci può essere un ruolo giocato dall'anno nel modellare la ricchezza di specie.

Nella *Figura10* riportiamo il diagramma di dispersione tra il logaritmo della specie e l'anno. Si è scelto di illustrare il logaritmo della variabile risposta perché fa risaltare maggiormente la tendenza crescente delle osservazioni.

Figura 10

Diagramma di dispersione tra la variabile Anno e il logaritmo della Specie

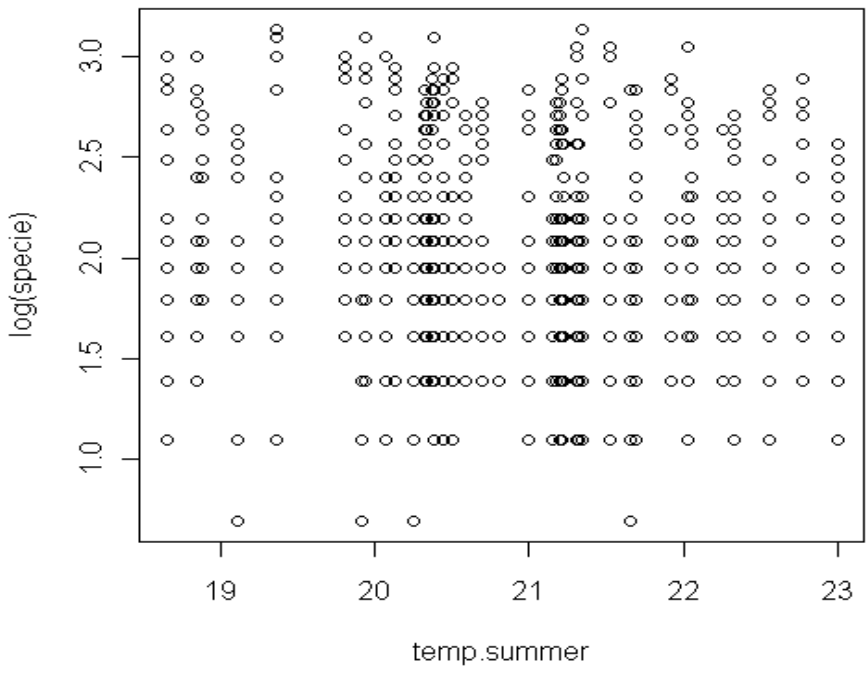


3.3.2 temp.summer → La variabile temp.summer è stata ricavata dalle temperature medie di tutti i mesi dell'anno per il periodo considerato, disponibili pubblicamente. Sono stati considerati i mesi estivi (da Aprile a Settembre) e per questi si è fatta la media tramite un comando in R, riportato in *appendice B*.

Si è costruita questa variabile perché, come detto prima, potrebbero esserci relazioni tra la ricchezza di specie e la temperatura. Nella *Figura 11* è riportato il diagramma di dispersione tra la variabile temp.summer e il logaritmo della variabile risposta .

Figura 11

Diagramma di dispersione tra la variabile temp.summer e il logaritmo della specie



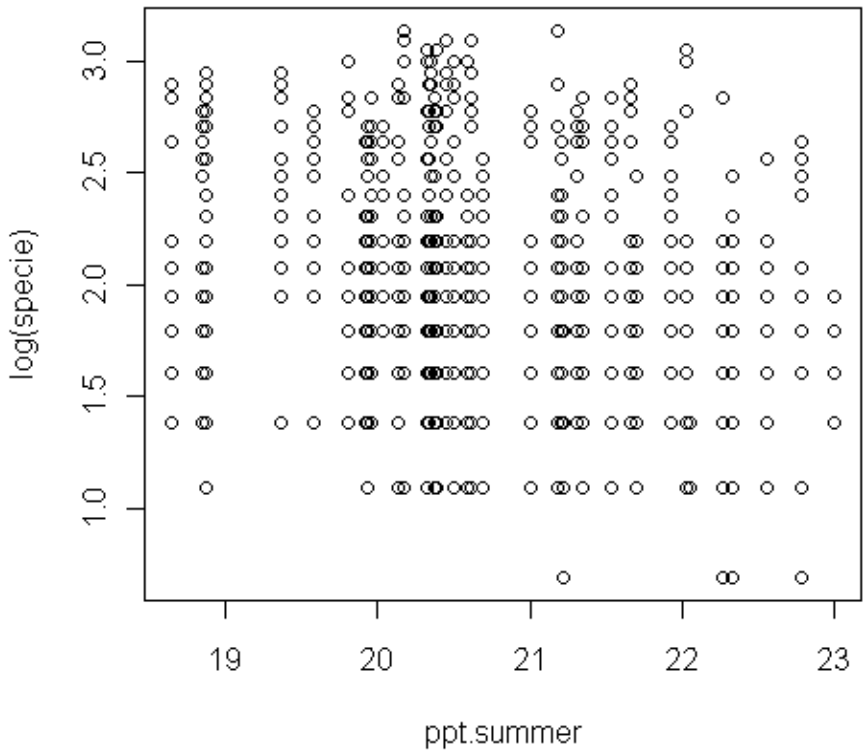
Graficamente questa variabile non è di facile lettura quindi lasciamo alle analisi il compito dirci come essa spiega il fenomeno.

3.3.3 ppt.summer → Variabile ricavata da dati pubblicati sulle precipitazioni. Per dare maggiore significato a questa variabile, si è deciso di prendere le precipitazioni estive che vanno (dal mese Aprile al mese di Settembre).

Nella *Figura 12* è riportato il diagramma di dispersione tra la variabile ppt.sumer e il logaritmo della variabile risposta.

Figura 12

Diagramma di dispersione tra la variabile ppt.summer e il logaritmo della variabile risposta



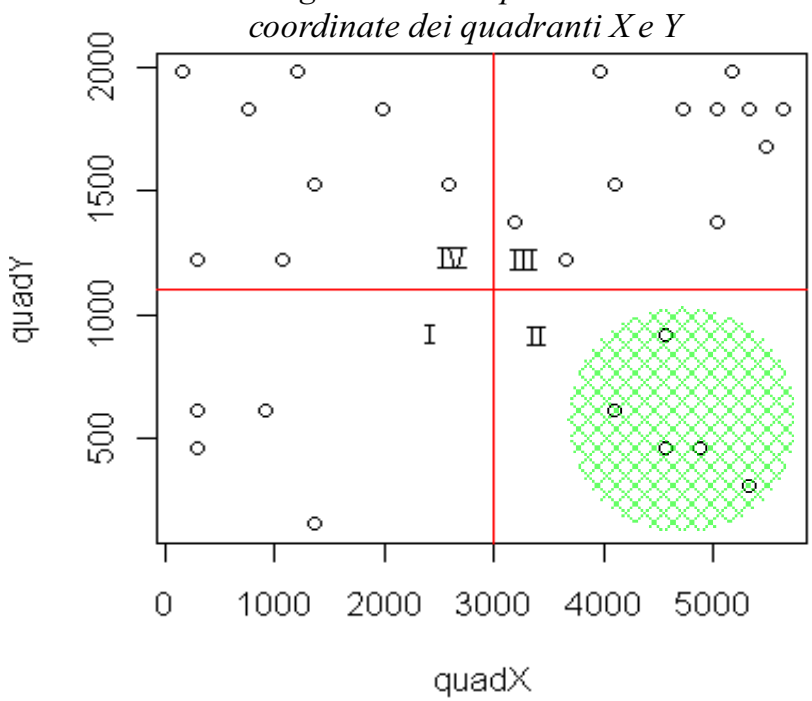
La lettura di questo diagramma non è semplice ma sembra ci possa essere una relazione leggermente negativa tra le precipitazioni estive e il numero di specie; le analisi avranno il compito di confermare o smentire questa considerazione.

3.3.4 Spazio → E' una variabile associata alle coordinate del quadX e quadY, che suddivide le osservazioni in quattro sotto quadranti, riportati nella *Figura 9.* Ha la funzione di inserire nell'analisi componenti spaziali da spiegare, permettendo di effettuare in quattro sotto quadranti analisi specifiche.

L'idea di costruire questa variabile, nasce dalle analisi esplorative che evidenziavano un andamento diverso per alcuni settori.

Figura 13

Diagramma di dispersione delle coordinate dei quadranti X e Y



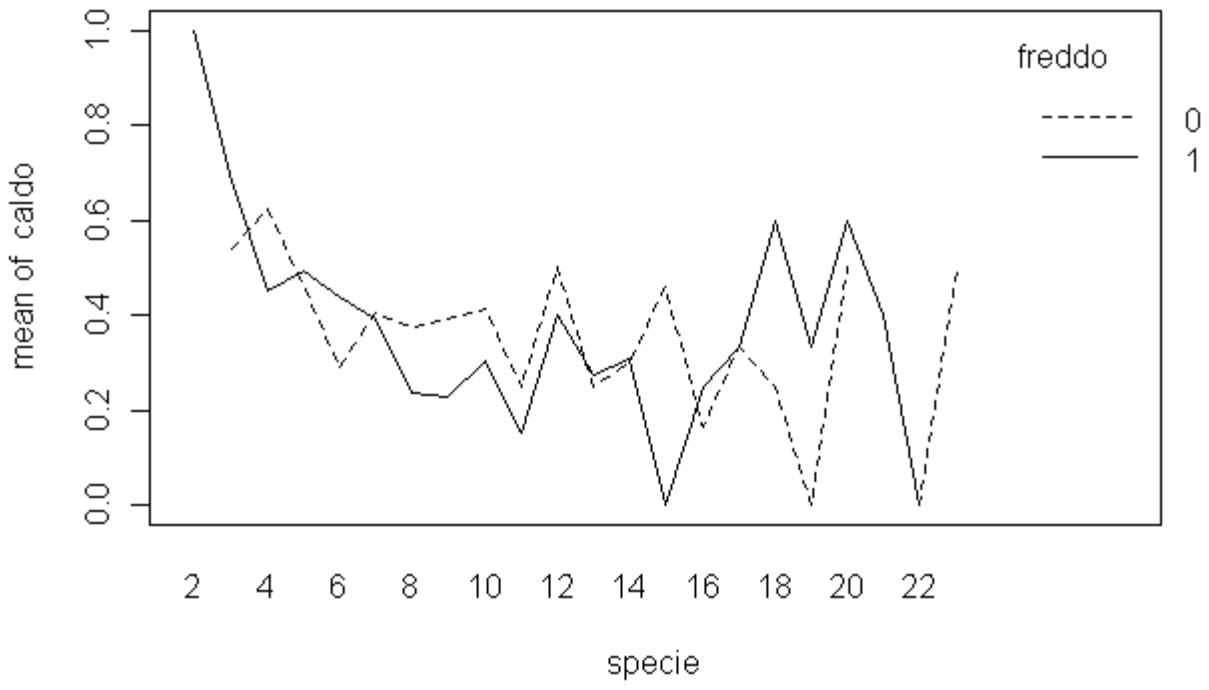
Con la suddivisione fatta, i quadranti con media e varianza più elevate cadono tutti all'interno del settore II.

3.3.5 Freddo → Questa variabile è stata estrapolata dallo stesso data set della variabile temp.summer. Freddo è una variabile dicotomica che prende in considerazione solo la temperatura media del mese più freddo per ogni anno e restituisce il valore 1 se in quell'anno c'è stato un freddo di un grado inferiore rispetto alla media dei mesi più freddi di tutti gli anni messi assieme.
Questa variabile ha la funzione di segnalare se in quell'anno c'è stato un particolare freddo.

3.3.6 Caldo → Questa variabile dicotomica è stata estrapolata dallo stesso data set della variabile temp.summer. Caldo prende in considerazione solo la media dei due mesi più caldi per ogni anno e restituisce il valore 1 se in quell'anno c'è stata una temperatura massima mensile superiore di un grado rispetto alla media di tutti gli altri anni messi assieme. Si è scelto di mettere i due mesi più caldi perché la temperatura massima non si raggiungeva sempre nello stesso mese. Questa variabile ha il compito di segnalare se in quel determinato anno c'è stato un caldo superiore alla norma. La figura 14 rappresenta il diagramma d'interazione tra la variabile caldo freddo e la variabile esplicativa.

Figura 14

Diagramma di interazione tra il numero di specie e le variabili dicotomiche caldo e freddo



3.3.7 Indice.posto → E' la variabile che permette di ricondurre e riconoscere ogni osservazione al proprio quadrante, rendendo così possibile attribuire a ogni quadrante una intercetta e un coefficiente angolare diversi. Questa variabile è rappresentata graficamente nel diagramma di dispersione quadX, quadY ed indicizza ogni singolo punto, rappresentata graficamente in *Figura 8.*

3.3.8 quadX → quadY → Sono le variabile associate alla variabile quadrante e sono rispettivamente le coordinate spaziali X e Y dei quadranti.

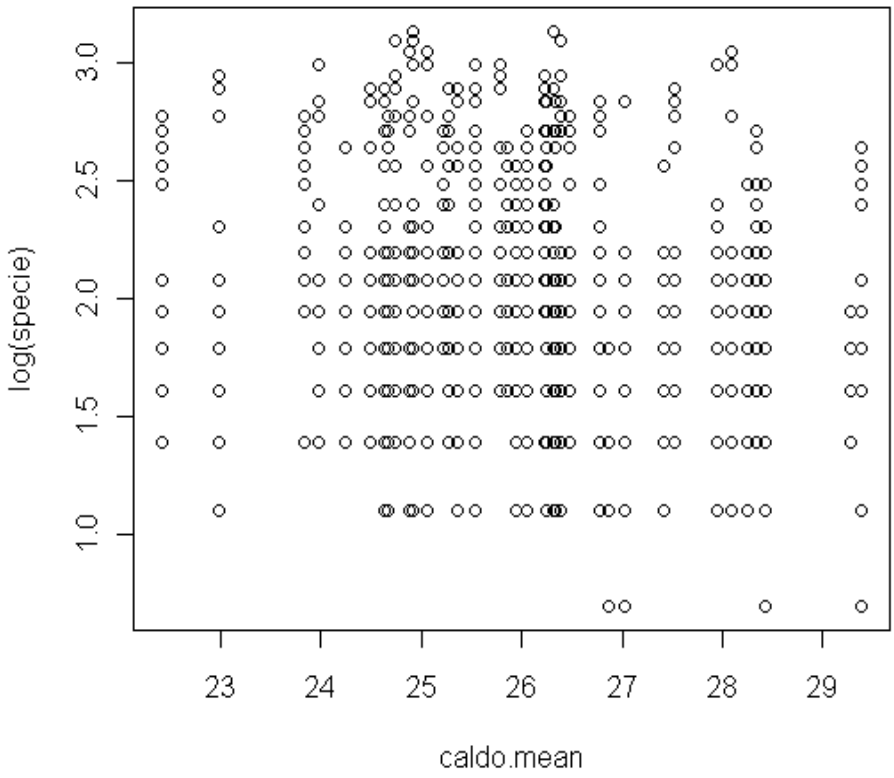
3.3.9 Caldo.mean → Variabile che contiene la media della temperatura estiva nei due mesi più caldi dell'anno.

I due mesi più caldi risultano essere Luglio e Agosto, quindi la nuova variabile caldo.mean è, in pratica la media delle temperature di questi due mesi.

La *Figura 15* riporta il diagramma di dispersione tra la variabile caldo.mean e il logaritmo della variabile risposta.

Figura 15

Diagramma di dispersione tra la variabile esplicativa calco.mean e il logaritmo del numero di specie



Capitolo 4

Una prima analisi tramite modelli lineari

4.1 Introduzione

In questo capitolo si proverà a spiegare la ricchezza di specie attraverso opportuni modelli di regressione lineare normale, nei quali viene utilizzata un'adeguata trasformata della variabile risposta. La decisione di utilizzare come primo strumento un modello lineare normale, discende dal fatto che il modello è di semplice interpretazione e spesso riesce a interpretare le relazioni tra le variabili di un conteggio se trasformate opportunamente.

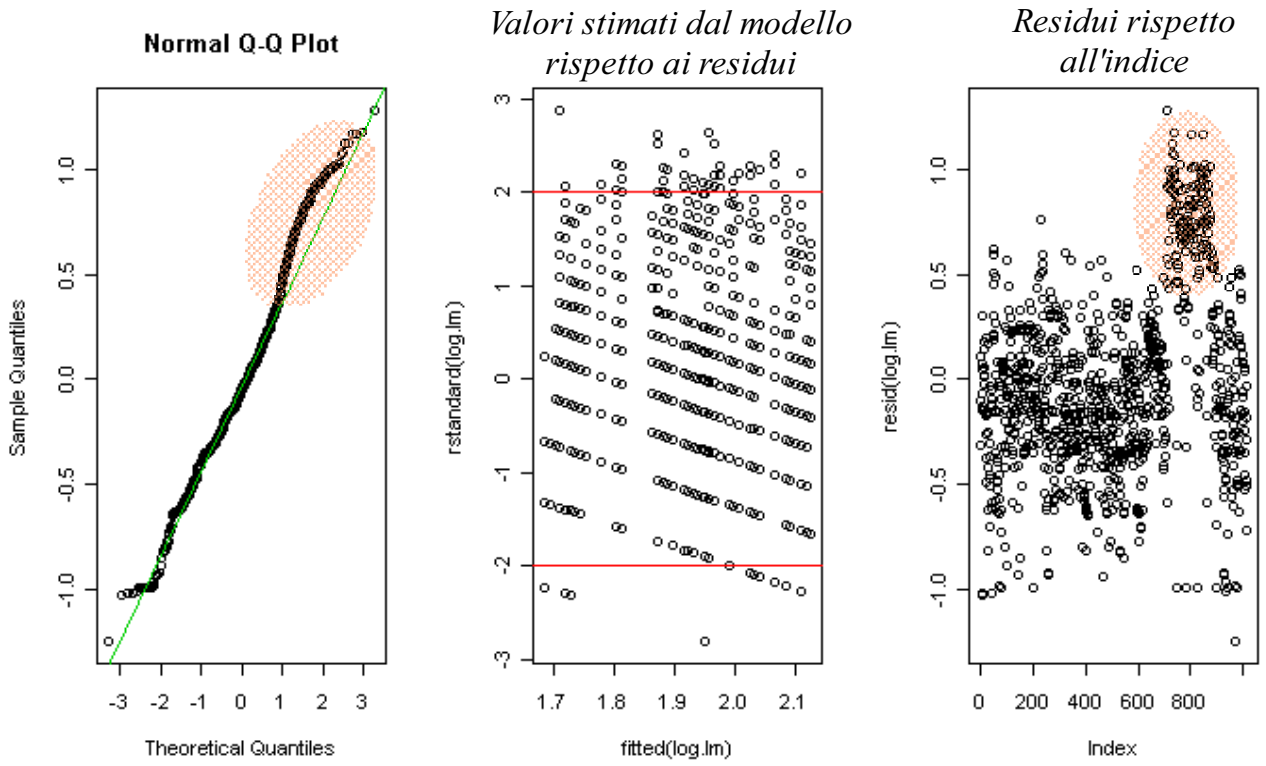
4.2 Modello 1

Il primo presentato contiene il logaritmo della ricchezza di specie e solo due variabili esplicative, che sono state scelte tra tutte le altre perché risultavano essere le più significative sia in termini di R^2, che confrontando i vari grafici dei residui prodotti dai vari modelli.

I coefficienti stimati e l'analisi statistica sono riportati nella *Appendice A* sotto il nome di *Modello 1*.

Il *modello 1* che considera il tempo e le ondate di calore, sembra non si adatti male ai dati, perché in esso tutti i coefficienti sono significativi. Analizzando i grafici dei residui per verificare la linearità della relazione, l'omoschedasticità degli errori e l'indipendenza degli errori da relazioni che potrebbero essere ancora presenti in essi, mostrati in *Figura 16,* si può dire

Figura 16



la linearità è tendenzialmente rispettata seppure si nota un distaccamento evidente di una parte di residui osservati rispetto al comportamento atteso. L'omoschedasticità è buona, come si può vedere dal grafico centrale, ma guardando il valore dei residui standardizzati posti rispetto all'indice si nota un andamento sistematico per un gruppo di dati i quali hanno i residui posti regolarmente tutti sopra allo zero. Questo gruppo di residui sono evidenziati

nel grafico. Concludendo il modello non sembra ancora in grado di interpolare bene i dati.

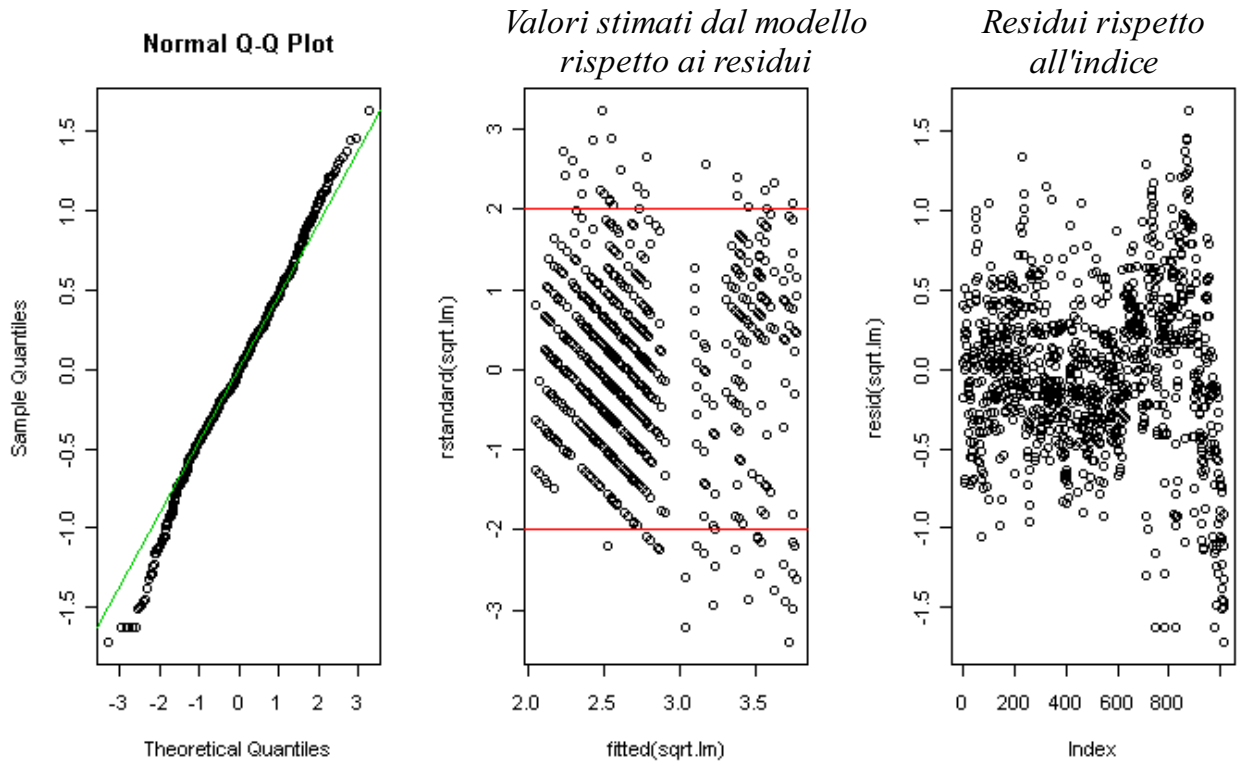
4.3 Modello 2

In seguito a quanto visto per il *Modello 1* si è inserita la variabile spazio all'interno del modello per vedere se fosse possibile eliminare i residui con comportamento anomalo.

Gli esiti di questa nuova analisi sono riportati in *Appendice A.*

Il *modello 2* utilizza la trasformazione della variabile risposta sotto radice quadrata. Grazie all'inserimento di questa variabile si è ottenuto un notevole miglioramento del coefficiente R^2 e la *Figura 18,* riporta l'analisi dei residui.

Figura 18



Osservando questi grafici si nota un miglioramento. La linearità espressa nel qq-plot è buona tranne che per la coda sinistra, la quale devia rispetto al comportamento atteso dei residui. I residui rispetto a valori stimati sono quasi tutti compresi all'interno di +/- 1,96 che è il quantile che contiene il 95% della probabilità per la distribuzione normale, anche se è possibile vedere che per valori stimati piccoli abbiamo errori positivi mentre per valori stimati grandi tendiamo ad avere errori negativi. Il grafico dei residui rispetto all'indice ci mostra che i residui sono disposti bene, anche se si può notare che quelli più a destra sono maggiormente dispersi rispetto a quelli di sinistra.

Concludendo, si può affermare che questo modello interpola i dati abbastanza bene.

Circa l'interpretazione del modello, l'analisi suggerisce che:

1) al variare dell'anno il numero di specie presenti aumenta;
2) sia un freddo che un caldo particolarmente forti fanno diminuire la ricchezza di specie;
3) lo spazio gioca un ruolo importante nella spiegazione dei dati evidenziando la diversità dei settori II e III.

Capitolo 5

Modelli lineari generalizzati

5.1 Introduzione

I dati in esame si riferiscono ad un conteggio di specie vegetali, quindi, per il rispetto delle assunzioni di base, si configura come opportuno utilizzare un modello di Poisson.
I dati osservati sono realizzazioni di una variabile casuale con distribuzione di Poisson di media λ_i

$$\mathbf{Y_i \sim Poi(\lambda_i) \text{ dove } \lambda_i > 0}$$

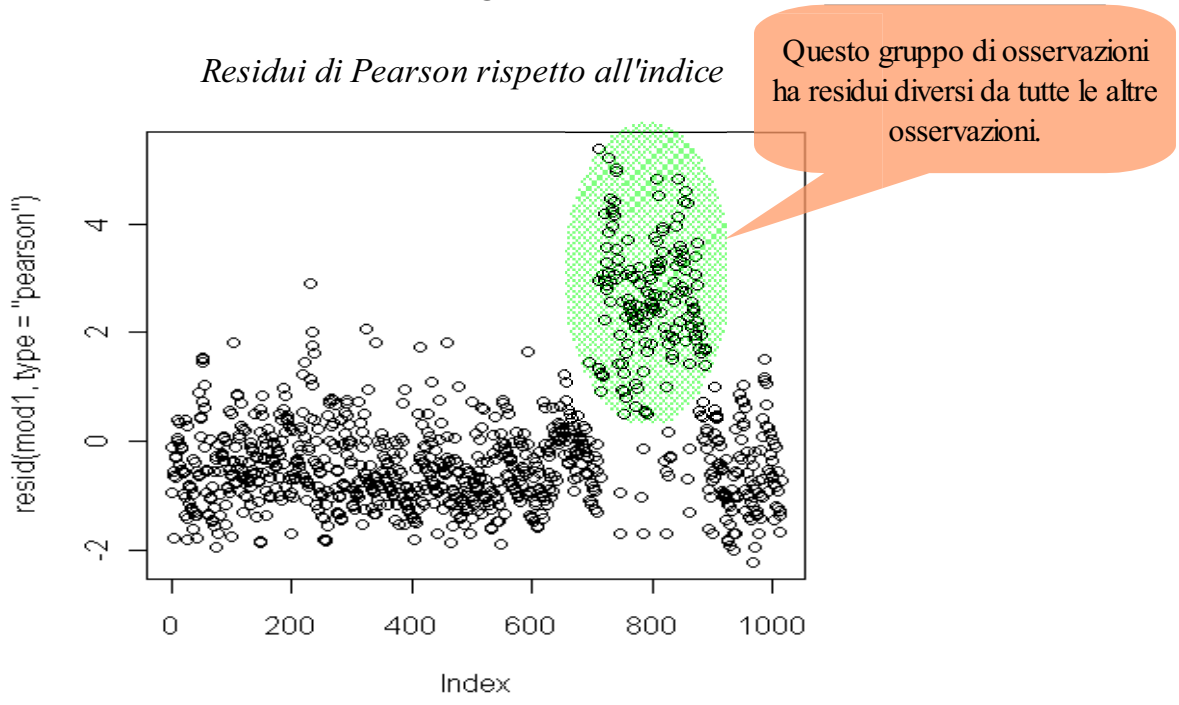
5.2 Modello 1

Volendo spigare la ricchezza di specie con una sola variabile esplicativa, le analisi suggeriscono che questa sia il tempo.
Il modello è riportato nella *Appendice B* sotto il nome di *Modello 1*.
Per questo modello sia l'intercetta che la variabile anno risultano altamente

significative ciò significa che al cambiare dell'anno cambia anche il numero di specie.

Dal risultato riportato in *appendice B*, si può capire che questo modello non riesce a spiegare adeguatamente la variabilità dei dati.
Per capire il problema si analizzi il grafico dei residui di Pearson[1] riportati in *Figura 19*.

Figura 19



La situazione di questi residui è buona per quasi tutte le osservazioni tranne che per il gruppo evidenziato nel grafico.

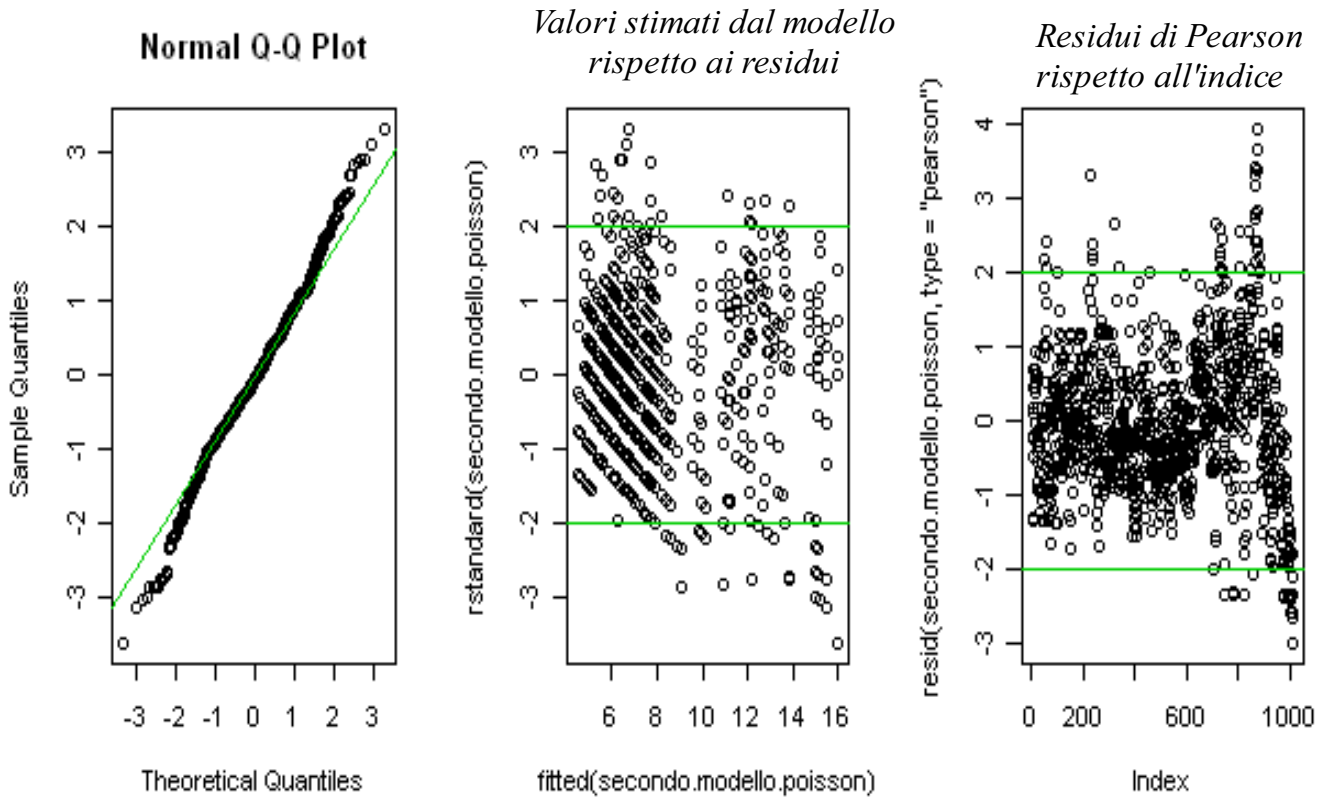
Questo ci fa pensare che per interpretare meglio i dati sia opportuno inserire delle altre variabili.

1 Il residuo di Pearson è l'estensione diretta del concetto di residuo standardizzato per il modello di Poisson

5. Modello 2

*I*l *Modello 2* riportato in *appendice B,* include le variabili anno, spazio, caldo, freddo, ppt.summer e caldo.mean, le quali risultano essere tutte significative. Prima di dare una interpretazione, è opportuno effettuare alcune analisi grafiche sui residui del modello riportate nella *Figura 20.*

Figura 20

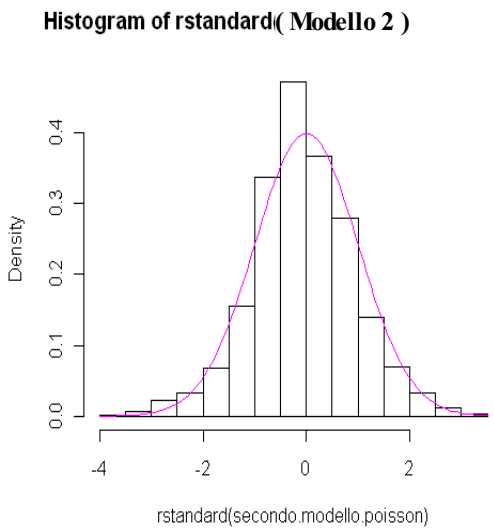


Iniziando a commentare la normalità dei residui si può affermare che, seppure sono presenti deviazioni dei residui osservati rispetto a quelli attesi in entrambe le code, l'andamento è buono. Il secondo grafico, che relaziona

i valori stimati rispetto ai residui, evidenzia che per valori piccoli abbiamo residui positivi mentre per valori grandi tendiamo ad avere residui negativi, anche se la maggior parte dei residui è contenuta nell'intervallo +/-2; il grafico dei residui rispetto all'indice mostra che i residui degli ultimi dati sono più dispersi rispetto ai primi.

Figura 21

Istogramma dei residui standardizzati per il modello 2



Un'ultima analisi da effettuare è quella che riguarda la distribuzione dei residui standardizzati rispetto alla forma teorica della densità di una normale.

Come si può vedere in *Figura 21* l'andamento dei residui alla distribuzione è buono.

Concludendo si può dire che il modello si adatta bene ai dati, tranne che nelle code. Dalle analisi si desume che con il passare degli anni abbiamo un incremento significativo del numero di specie; essere in una locazione spaziale comporta un cambiamento in media e in varianza della ricchezza di specie; i valori estremi di caldo e freddo comportano una diminuzione del numero di specie; all'aumentare delle precipitazioni estive abbiamo una diminuzione del numero di specie e in fine se la temperatura media estiva aumenta, aumenta anche la ricchezza di specie.

5.4 Modello 3

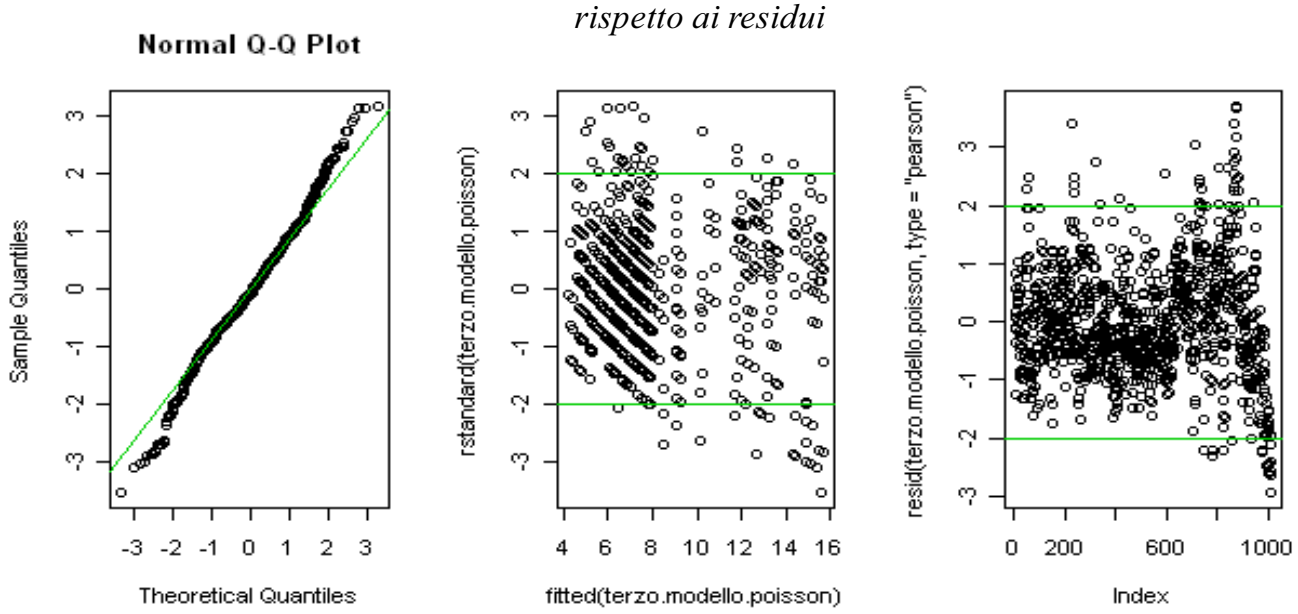
Interessante ora risulta analizzare un modello che contiene le interazioni tra le variabili.

Il modello corrispondente è riportato *nell'appendice B. S*i può notare subito che esso non ha un intercetta significativa, ma arriva a spigare il 76% della variabilità. Questo è il miglior modello che è stato possibili costruire con le informazioni a disposizione. Nella *Figura 22* si riportano le analisi grafiche effettuate su questo modello.

Figura 22

Valori stimati dal modello rispetto ai residui

Residui di Pearson rispetto all'indice



Osservando i grafici, si può subito vedere che la situazione dei residui è molto simile a quella vista nel *Modello 2* quindi le considerazioni da fare sono le stesse.

Le considerazioni fatte per le variabili del *Modello 2* continuano a essere valide anche per questo modello, come si può capire dal fatto che i coefficienti rimangono dello stesso segno e significativi. Utile però in questo modello è considerare il supporto che danno le interazioni nella spiegazione della ricchezza di specie.

La prima interazione tra la variabile anno e la variabile caldo ci dice che in ogni combinazione con l'anno e la variabile caldo si ha una intercetta e un coefficiente angolare diversi.

La seconda interazione tra le variabili caldo.mean e spazio invece dice che, in ogni settore, oltre all'intercetta diversa si ha anche un coefficiente angolare per la variabile caldo.mean diverso.

5.5 Conclusioni

Le prime conclusioni che si possono già trarre da questi modelli è che la componente spaziale abbia un ruolo fondamentale nello spiegare l'andamento delle osservazioni. Oltre a questo, si osserva che col passare degli anni c'è stato un incremento del numero di specie, che è comparabile con l'andamento della temperatura nel periodo di studio.

Inoltre, la biodiversità vegetale è maggiormente sensibile alle temperature estreme di caldo e di freddo, come si può capire dalla significatività delle variabili associate, piuttosto che essere sensibile alle temperature medie.

Capitolo 6

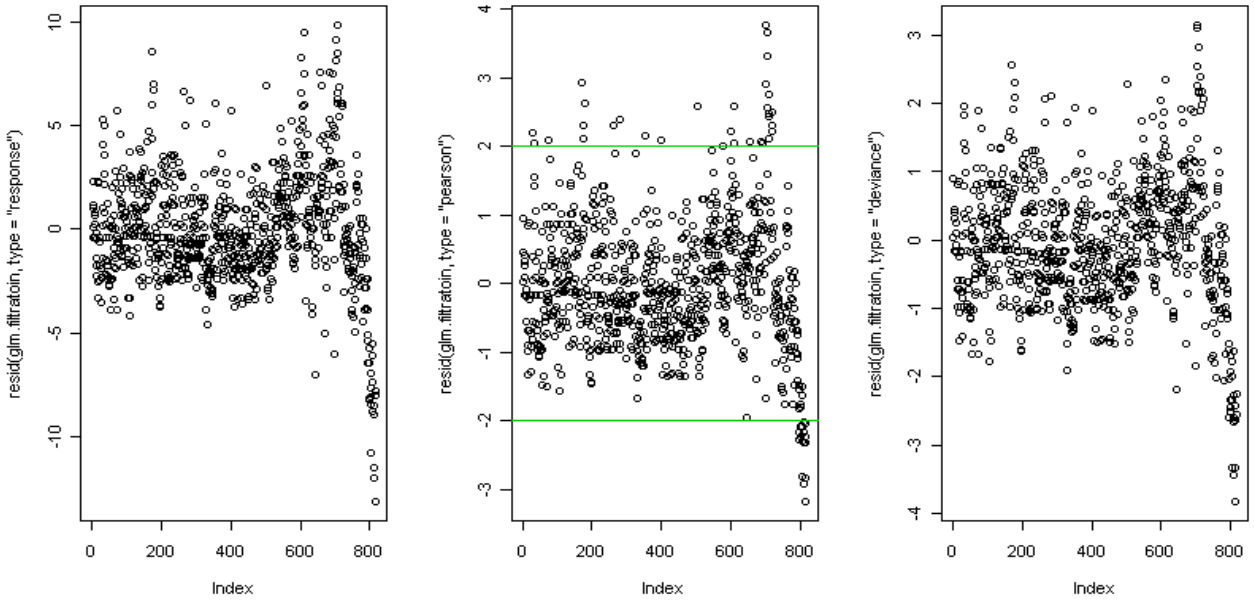
Un ulteriore approfondimento

La decisione di effettuare uno studio sul data set depurato dei dati appartenerti ai primi undici anni discende dal fatto che in quegli anni si dispone di un numero di osservazioni inferiore.

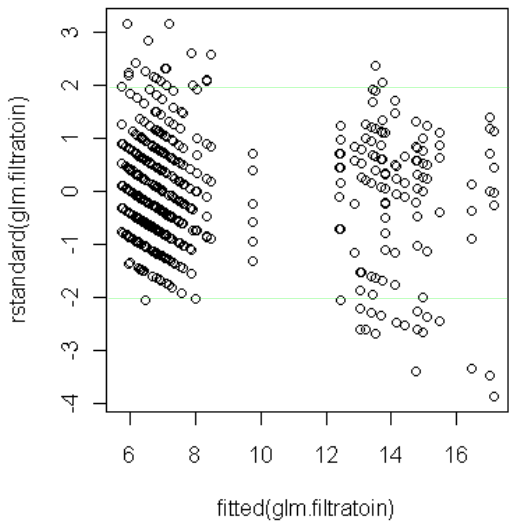
Lo scopo di questa ulteriore analisi è vedere se in questo nuovo periodo di riferimento cambia il ruolo giocato dalle variabili nello spiegare i dati.
Si consideri quidi il miglior modello ottenuto precedentemente, riportato in *appendice C*.
In *Figura 23* è riportata l'analisi dei residui.

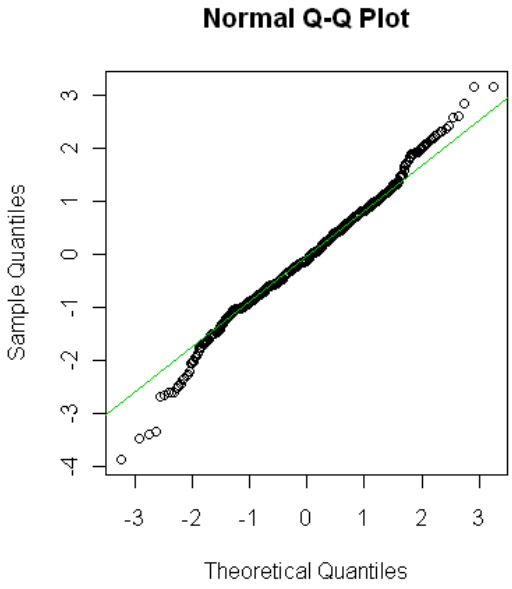
Figura 23

Diagrammi di dispersione dei residui in varie scale rispetto all'indice



Valori stimati dal modello rispetto ai residui





CAPITOLO 6 UN ULTERIORE APPROFONDIMENTO

Le analisi grafiche mostrano la stesse debolezze precedentemente riscontrate. Come si può vedere, il qq-plot continua ad avere tutte e due le code che si allontanano dall'andamento atteso dai residui, e per quanto riguarda sia il grafico dei residui rispetto ai valori stimati che ai residui di Pearson non abbiamo, come detto, grandi miglioramenti rispetto all'analisi condotta su tutti i dati completi.

La cosa più significativa da far notare per questo modello è che togliendo dall'analisi i primi 11 anni, cambia il significato che le variabili hanno nei confronti del fenomeno. In particolare si nota che la variabile temp.summer acquista una significatività mai avuta prima mentre la variabile freddo e la variabile ppt.summer perdono entrambe di significatività.

In questo modello le variabili in gioco dicono che con il passare degli anni il numero di specie presenti aumenta, all'aumentare della temperatura media abbiamo un aumento del numero di specie mentre se in presenza di un caldo superiore alla media, il numero di specie diminuisce. Lo spazio continua ad avere significato.

Per quanto riguarda le interazioni, tra temp.summer e l'anno comporta un cambiamento in termini di coefficiente angolare in ogni combinazione anno temp.summer mentre l'altra interazione tra la variabile spazio e temp.summer attribuisce un nuovo coefficiente angolare e una nuova intercetta ad ogni combinazione di spazio temperatura.

Capitolo 8

Conclusioni

Ripercorrendo l'analisi svolta tramite i modelli lineari generalizzati si può dire che la ricchezza di specie vegetali all'interno della nostra analisi è aumentata negli anni e ciò è stato evidenziato da tutti i modelli costruiti.

Un ruolo fondamentale nello spiegare la biodiversità è assegnato alla locazione spaziale della vegetazione, che fa variare in modo più che significativo il numero di specie presenti in un determinato quadrante.

La temperatura media estiva (da Aprile a Settembre) risulta altamente significativa negli ultimi 30 anni di osservazioni.

Le precipitazioni estive influenzano in modo negativo la ricchezza di specie. La variabile, però perde di significatività nell'analisi condotta nell'analisi senza i primi 11 anni.

La variabile Caldo in tutti i modelli mostra la sua influenza negativa sul numero di specie; quindi si può affermare che la ricchezza di specie soffre le temperature più calde della media.

CAPITOLO 8 CONCLUSIONI

La variabile Freddo mostra la sua influenza negativa sulla maggior parte dei modelli; quindi possiamo affermare che la ricchezza di specie soffre anche le temperature più fredde della media.

Infine la temperatura media dei due mesi più caldi, mostra un effetto positivo in quasi tutti i modelli stimati. Questo sta a significare che, mediamente, se abbiamo un mese caldo contribuisce positivamente alla ricchezza di specie, ma se si registrano temperature troppo elevate, si osserva una diminuzione del numero di specie.

L'incremento della temperatura nei mesi caldi è visto come un effetto positivo molto probabilmente perché il luogo dov'è stata condotta la ricerca ha una latitudine abbastanza a Nord. Quindi nell'aumentare delle temperature comporta un incremento della ricchezza ambientale. La stessa variabile, ad altre latitudini potrebbe comportare una riduzione significativa della ricchezza di specie.

Da questa semplice considerazione si può capire come sia molto difficile formulare un modello generale per spiegare la ricchezza di specie su una scala più ampia, visto che al cambiare delle condizioni climatiche cambia sia la tipologia delle specie presenti e, anche in presenza delle stesse specie, queste, durante la loro evoluzione, hanno modificato il loro modo di essere rispetto a quelle presenti in un altro territorio.

Appendice A

Modello 1

> modello.1<-lm(log(specie)~anno+caldo)

> summary(modello.1)

```
Call:
lm(formula = log(specie) ~ anno + caldo)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.25523 -0.30796 -0.04605  0.23941  1.28532 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.513576   0.081344  18.607  < 2e-16 ***
anno         0.008499   0.001389   6.117 1.36e-09 ***
caldo       -0.100624   0.031412  -3.203   0.0014 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4473 on 1011 degrees of freedom
(134 observations deleted due to missingness)
Multiple R-squared: 0.07022,    Adjusted R-squared: 0.06838 
F-statistic: 38.17 on 2 and 1011 DF,  p-value: < 2.2e-16
```

Modello 2

> modello.2<-lm(sqrt(specie)~anno+freddo+caldo+spazio)
> summary(modello.2)

```
Call:
lm(formula = sqrt(specie) ~ anno + freddo + caldo + spazio)

Residuals:
     Min       1Q   Median       3Q      Max
-1.722889 -0.309314  0.001244  0.308370  1.633939

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.869926   0.103094  18.138  < 2e-16 ***
anno         0.015384   0.001743   8.827  < 2e-16 ***
freddo      -0.094721   0.035199  -2.691  0.00724 **
caldo       -0.103804   0.036186  -2.869  0.00421 **
spazioII     0.870809   0.055750  15.620  < 2e-16 ***
spazioIII   -0.115888   0.049262  -2.352  0.01884 *
spazioIV    -0.090996   0.051136  -1.779  0.07546 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5071 on 1007 degrees of freedom

(134 observations deleted due to missingness)
Multiple R-squared: 0.3946,     Adjusted R-squared: 0.391

F-statistic: 109.4 on 6 and 1007 DF,  p-value: < 2.2e-16
```

Appendice B

Modello 1

modello.1<-glm(formula = specie ~ anno, family = poisson)

summary(primo.modello.poisson)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.6482	-0.9821	-0.3800	0.4312	4.2848

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.522452	0.057396	26.525	<2e-16 ***
anno	0.009558	0.001031	9.267	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 1782.2 on 1013 degrees of freedom

Residual deviance: 1695.8 on 1012 degrees of freedom

(134 observations deleted due to missingness)

AIC: 5544.7

Number of Fisher Scoring iterations: 4

La variabile anno è molto importante. Però il p-value:

1-pchisq(1695.8,1012) [1] 0 → *ci suggerisce che il modello non è perfetto.*

APPENDICE B

Modello 2

modello.2<-glm(specie~anno+spazio+caldo+freddo+ppt.summer +caldo.mean, poisson)
> summary(secondo.modello.poisson)

```
Call:
glm(formula = specie ~ anno + spazio + caldo + freddo + ppt.summer +

caldo.mean, family = poisson)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-3.60014  -0.64865  -0.08904   0.52487   3.28822


Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.214677   0.370270   3.281 0.001036 **
anno         0.012129   0.001373   8.836  < 2e-16 ***
spazioII     0.587950   0.037192  15.809  < 2e-16 ***
spazioIII   -0.106588   0.037031  -2.878 0.003998 **
spazioIV    -0.085676   0.038469  -2.227 0.025937 *
caldo       -0.094957   0.034079  -2.786 0.005329 **
freddo      -0.062509   0.025426  -2.458 0.013954 *
ppt.summer  -0.065609   0.019854  -3.305 0.000951 ***
caldo.mean   0.058280   0.014936   3.902 9.54e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1782.2  on 1013  degrees of freedom
Residual deviance:  989.2  on 1005  degrees of freedom
  (134 observations deleted due to missingness)
AIC: 4852.2
Number of Fisher Scoring iterations: 4
```

APPENDICE B

Le variabili sono tutte significative. Il p-value segnala che il modello spiega abbastanza bene i dati anche se non è ancora perfetto.

```
> 1-pchisq(989.2,1005)
[1] 0.632843
```

Modello 3

```
> modello.3<-glm(specie~anno*caldo+freddo+ppt.summer+
caldo.mean*spazio,poisson)
> summary(terzo.modello.poisson)

Call:
glm(formula = specie ~ anno * caldo + freddo + ppt.summer + caldo.mean
*
spazio, family = poisson)


Deviance Residuals:
    Min      1Q  Median      3Q     Max
-3.4982  -0.6323  -0.1025   0.5537   3.1158

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -0.526308   0.622331  -0.846 0.397718
anno                0.009540   0.001712   5.571 2.53e-08 ***
caldo              -0.411208   0.130428  -3.153 0.001617 **
freddo             -0.055695   0.025594  -2.176 0.029550 *
ppt.summer         -0.069791   0.019876  -3.511 0.000446 ***
caldo.mean          0.134702   0.024567   5.483 4.18e-08 ***
spazioII            2.905835   0.635454   4.573 4.81e-06 ***
spazioIII           1.805660   0.642190   2.812 0.004928 **
spazioIV            1.867593   0.662353   2.820 0.004808 **
anno:caldo          0.005799   0.002313   2.507 0.012189 *
caldo.mean:spazioII  -0.089570   0.024477  -3.659 0.000253 ***
caldo.mean:spazioIII -0.073989   0.024777  -2.986 0.002825 **
```

APPENDICE B

```
caldo.mean:spazioIV  -0.075524   0.025538  -2.957 0.003104 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1782.18  on 1013  degrees of freedom
Residual deviance:  968.78  on 1001  degrees of freedom
  (134 observations deleted due to missingness)
AIC: 4839.7

Number of Fisher Scoring iterations: 4
```

Le variabili e tutte le interazioni risultano significative. Il p-value segnala che il modello spiega abbastanza bene i dati anche se non è ancora perfetto.

```
> 1-pchisq(968.78,1001)
[1] 0.762088
```

Appendice C

Modello 1

modello.1<-
glm(specie~anno*temp.summer+caldo+spazio*temp.summer+
caldo.mean,poisson)
> summary(glm.filtratoin)

Call:
glm(formula = specie ~ anno * temp.summer + caldo + spazio *
temp.summer + caldo.mean, family = poisson)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.8225	-0.6103	-0.1122	0.5371	3.1377

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	-7.144030	2.001895	-3.569	0.000359	***
anno	0.130327	0.031872	4.089	4.33e-05	***
temp.summer	0.380797	0.093272	4.083	4.45e-05	***
caldo	-0.094252	0.035449	-2.659	0.007842	**
spazioII	2.480740	0.818739	3.030	0.002446	**
spazioIII	0.916837	0.804244	1.140	0.254287	
spazioIV	2.264150	0.829599	2.729	0.006349	**
caldo.mean	0.026849	0.012493	2.149	0.031630	*
anno:temp.summer	-0.005868	0.001501	-3.909	9.28e-05	***
temp.summer:spazioII	-0.084225	0.039171	-2.150	0.031540	*
temp.summer:spazioIII	-0.045859	0.038465	-1.192	0.233165	
temp.summer:spazioIV	-0.108997	0.039742	-2.743	0.006095	**

APPENDICE C

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
47
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 1460.36  on 816  degrees of freedom
Residual deviance:  720.48  on 805  degrees of freedom
  (23 observations deleted due to missingness)
AIC: 3881.8

Number of Fisher Scoring iterations: 4


> 1-pchisq(720.5,805)
[1] 0.98491
```

Le variabili sono tutte significative. Il p-value segnala che il modello spiega abbastanza bene i dati.

Bibliografia

- Pace, L. e Salvan, S. (2001). Introduzione alla statistica - II. Inferenza, verosimiglianza, modelli. Cedam, Padova.
- Web site www.esapubs.org
- Manuale del software R-project
- Materiale didattico per i laboratori di Modelli statistici I
  M. Chiogna, A Salvan, N. Sartori
- Materiale didattico per il corso di modelli statistici II
  L. Ventura

RINGRAZIAMENTI:

A conclusione di questo ciclo di studio mi sento in dovere di ringraziare alcune persone. Inizio con il ringraziare tutta la mia famiglia che mi è sempre stata vicina e mi ha sempre aiutato.
In seguito devo ringraziare il mio relatore la dottoressa Monica Chiogna per la disponibilità e la gentilezza mostrata nei miei confronti.
Durante il periodo di studio ho avuto la possibilità di svolgere due anni di servizio civile presso le Pro Loco del consorzio Padova sud-est, e ho avuto il modo di relazionarmi con delle persone splendide e quindi mi sento in dovere di ringraziarli.
Devo ringraziare tutti i miei amici per per avermi sempre aiutato e dimostrato di esserci quando ne avevo bisogno.
Infine ringrazio il mio maestro di arti marziali che grazie a lui ho avuto modo di vincere, negli ultimi due anni, tre titoli di campione D'Italia.

Grazie a tutti Nicola.