Analisi esplorativa di dati spaziali relativi alla Biomassa e alla Temperatura nel Mare Adriatico

Francesco Chelli Pietro Muliere Piercesare Secchi

Dipartimento di Economia Politica e Metodi Quantitativi
Univeristà degli Studi di Pavia

7 Giugno 1996

Indice

Capitolo 1

Analisi esplorativa di dati spaziali univariati

L'analisi esplorativa dei dati (EDA) riguarda l'organizzazione, la sintesi e la presentazione dei dati osservati allo scopo di renderli più comprensibili e per meglio scoprirne la struttura sottostante e determinarne le tendenze.

In questo capitolo descriveremo in dettaglio alcuni strumenti grafici ed analitici opportuni per la EDA di dati spaziali univariati.

1.1 Individuazione dei valori anomali della distribuzione empirica dei dati e dei valori anomali spaziali

In un insieme grande di dati, un *valore anomalo* è un dato che appare 'diverso' dagli altri rispetto ad uno o più criteri stabiliti dal ricercatore. Dal punto di vista dell'analisi di dati univariati distribuiti nello spazio sembra interessante considerare i seguenti tipi di anomalia:

- Valori che risultano estremi (troppo grandi o troppo piccoli) rispetto agli altri e che quindi influenzano il calcolo di indici sintetici di posizione quali, per esempio, la media.
- Valori che risultano anomali rispetto al modello statistico che il ricercatore ritiene abbia generato i dati. Molti dei modelli utilizzati nella statistica spaziale assumono, ad esempio, che i dati siano generati da un modello gaussiano; è in questo caso importante individuare le osservazioni che meno si adattano a questa ipotesi.
- Valori che risultano eccessivamente diversi da quelli che a loro sono vicini nello spazio. Ovviamente il ricercatore dovrà in questo caso descrivere una opportuna topologia in base alla quale decidere quando due dati provengono da luoghi vicini.

Se un dato risulta anomalo rispetto ad uno o più criteri di analisi, non per questo esso non va considerato nelle analisi successive. Al contrario la presenza di dati anomali da una parte induce il ricercatore ad utilizzare tecniche di analisi che siano resistenti (per esempio, la mediana al posto della media come indicatore di posizione della distribuzione) e dall'altra

lo deve spingere ad individuare le ragioni non statistiche che hanno eventualmente prodotto le anomalie (ad esempio, per i dati del mare Adriatico, le piene del Po, i venti di Bora, etc.)

In questa sezione presentiamo alcuni strumenti atti all'individuazione di dati anomali.

1.1.1 Box plot

Il *box plot* è uno strumento grafico che permette una semplice rappresentazione dei principali indicatori di posizione di una distribuzione di dati nonchè dei suoi valori estremi.

Il box plot si costruisce considerando i *quartili* della distribuzione. Il primo quartile di una distribuzione è quel valore $q_{.25}$ tale che il 25% delle osservazioni ha un valore inferiore o al più uguale a $q_{.25}$. Il terzo quartile è quel valore $q_{.75}$ tale che il 75% delle osservazioni ha un valore inferiore o al più uguale a $q_{.75}$. Il secondo quartile è la mediana. La differenza $IQR = q_{.75} - q_{.25}$ prende il nome di *differenza interquartile*.

Il box plot individua graficamente l'intervallo nel quale cadono le osservazioni comprese tra $q_{.25}$ e $q_{.75}$ per mezzo di un rettangolo. Dai lati orizzontali del rettangolo si estendono due segmenti che raggiungono il valore osservato più grande e quello più piccolo compresi nell'intervallo $(q_{.25} - 1.5IQR, q_{.75} + 1.5IQR)$. Le osservazioni, se ne esistono, che cadono all'esterno di questo intervallo sono considerate anomale in quanto troppo grandi o troppo piccole e visualizzate nel grafico per mezzo di un punto o di un asterisco.

Il ricercatore ha così a disposizione un semplice strumento grafico per rappresentare una distribuzione di dati che risulta particolarmente utile per i confronti tra distribuzioni di dati diversi (Figura 1.1).

1.1.2 Rankit plot

Il *rankit plot* permette di valutare graficamente l'ipotesi secondo la quale i dati osservati sono stati generati da una distribuzione Normale.

Il rankit plot si ottiene riportando su un piano cartesiano i punti di coordinate (u_i, y_i) dove $y_1 \leq y_2 \leq \ldots \leq y_n$ sono i dati osservati ordinati in ordine non decrescente mentre $u_1 \leq u_2 \leq \ldots \leq u_n$ sono i valori attesi delle statistiche d'ordine per un campione casuale di ampiezza n estratto da una popolazione Normale di media 0 e varianza 1.

Quando è vera l'ipotesi di normalità, ci aspetta che il legame tra i valori y_i e i valori u_i sia dato da una relazione lineare del tipo

$$y_i = \mu + \sigma u_i.$$

Tanto più il grafico ottenuto si discosta da quello di una retta tanto più si ritiene che l'ipotesi di normalità non sia soddisfatta dai dati osservati (Figura 1.2). Inoltre particolari forme assunte dal rankit plot suggeriscono talvolta al ricercatore opportune trasformazioni (ad esempio, il logaritmo o, più in generale, una trasformazione di Box-Cox) tali che i valori trasformati possano essere considerati come un campione da una popolazione normale.

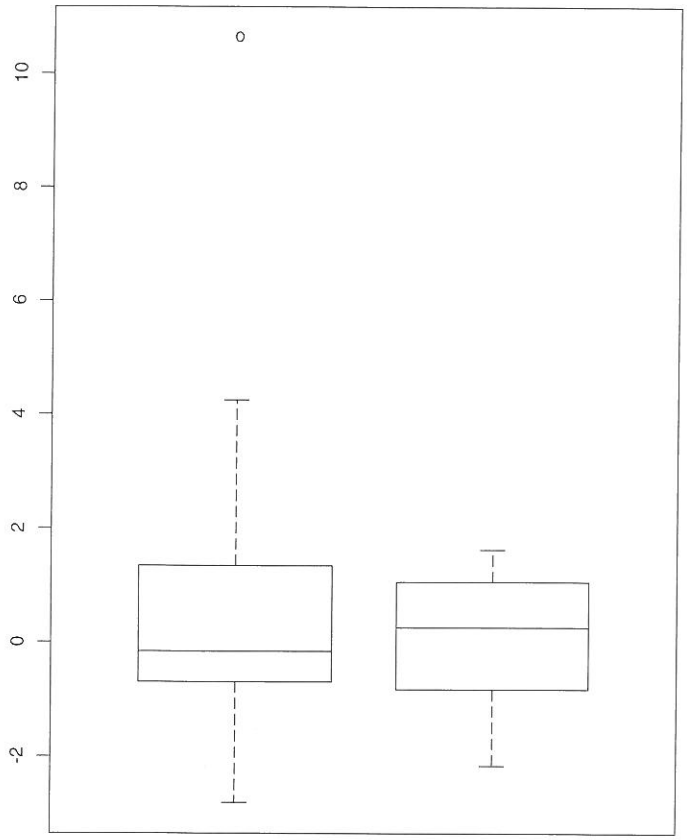

Figura 1.1: Box plot di due insiemi di dati a confronto: il boxplot di sinistra è relativo a 20 osservazioni generate da una Cauchy(0,1), quello di destra a 20 osservazioni generate da una Normale(0,1).

1.1.3 Analisi dei valori anomali rispetto a quelli contigui mediante grafici

Obiettivo di questo tipo di analisi è l'individuazione di dati che sono anomali rispetto a quelli spazialmente contigui. Questi valori possono avere forti effetti distorsivi sugli indicatori delle proprietà spaziali della distribuzione qualora essi non siano resistenti.

Quando i dati sono osservati su una griglia indichiamo con $y(s_1, s_2)$ il valore osservato in corrispondenza della riga s_1 e della colonna s_2 della griglia. Allora, per esempio, il grafico dell'insieme di punti $\{(y(s_1, s_2), y(s_1 + 1, s_2))\}$ o quello dei punti $\{(y(s_1, s_2), y(s_1, s_2 + 1))\}$ possono indicare la presenza di valori anomali locali rispetto alle direzioni Nord-Sud o Est-Ovest.

Più in generale, se $y = (y_1, \ldots, y_n)$ è il vettore dei dati osservati, si riportano su un piano

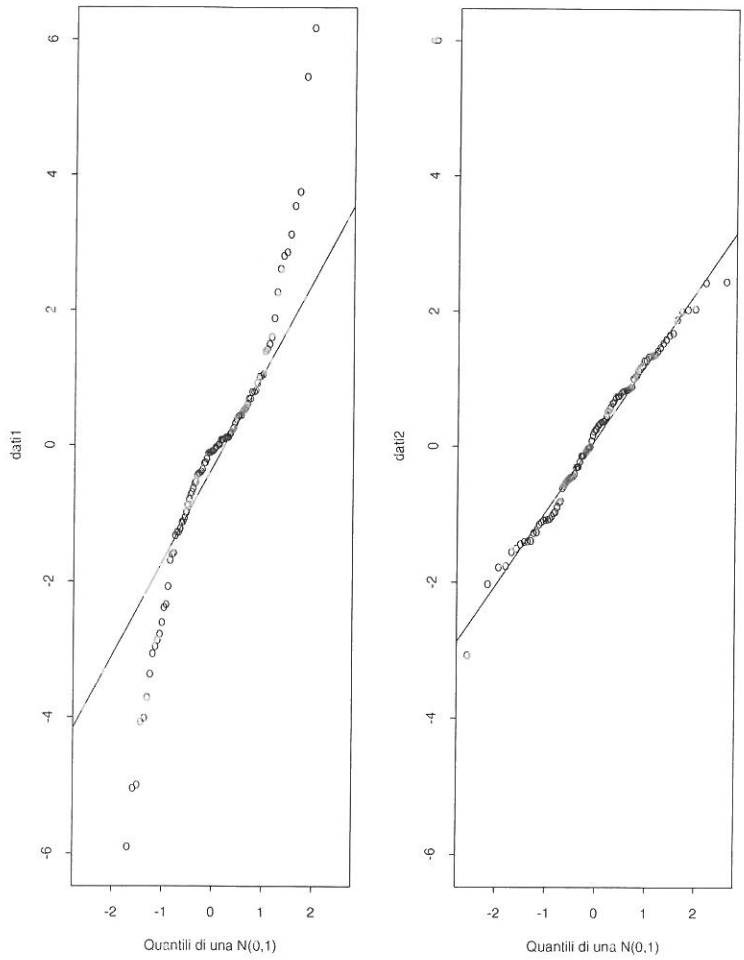

Figura 1.2: Due rankit plot a confronto: dati1 sono 100 osservazioni generate da una Cauchy(0,1) mentre dati2 sono 100 osservazioni generate da una Normale(0,1).

cartesiano i punti dell'insieme $\{(y_i, (Wy)_i) : i = 1, \ldots, n\}$ dove W è una opportuna matrice di contiguità.

Possibili esempi per la matrice di contiguità sono i seguenti. Indichiamo con i il sito da cui proviene l'osservazione y_i e con j quello da cui proviene l'osservazione y_j e sia w_{ij} l'elemento di ordine (i, j) della matrice W che misura il grado di contiguità tra il sito i e il sito j :

1.

$$w_{ij} \begin{cases} \neq 0 & \text{se } i \text{ è contiguo a } j \\ = 0 & \text{altrimenti} \end{cases}$$

2.

$$w_{ij} = d_{ij}^{-\gamma}$$

con $\gamma \geq 0$, e d_{ij} misura della distanza (euclidea, temporale, etc.) tra il sito i e il sito j.

3.

$$w_{ij} = \exp(-d_{ij}^{\gamma})$$

4.

$$w_{ij} = (l_{ij}/l_i)^{\gamma}$$

con $\gamma \geq 0$, dove l_{ij} è la lunghezza del bordo che i e j hanno in comune e l_i è il perimetro del sito i.

Spesso gli elementi delle righe o delle colonne della matrice W sono standardizzati in modo da sommare a 1: così facendo il valore $(Wy)_i$ assume il significato di media ponderata dei valori osservati in siti contigui al sito i secondo la topologia specificata da W.

Osserviamo che la matrice W permette al ricercatore di tener conto di informazioni relative alla natura spaziale dei dati di carattere anche diverso da quelli legati alla sola distanza o alla contiguità geografica dei siti. Per esempio, per i dati del Mare Adriatico, possiamo così codificare informazioni relative alle piene del Po, ai venti di Bora, alla profondità batimetrica, alla salinità dell'acqua etc.

Dopo aver riportato su un piano cartesiano i punti dell'insieme $\{(y_i, (Wy)_i) : i = 1, \ldots, n\}$ il ricercatore interpola un modello lineare quale, per esempio, la retta

$$(Wy)_i = \beta y_i + \text{residuo}, i = 1, \ldots, n.$$

Valori y_i ai quali corrispondono residui elevati individuano siti anomali rispetto ai contigui (Figura 1.3).

1.1.4 Outliers per righe o colonne

Quando i dati sono riferiti ad una griglia, la presenza di valori anomali da un punto di vista spaziale può essere segnalata da strumenti grafici quali il box plot o il rankit plot applicati ai dati relativi ad ogni riga o ad ogni colonna separatamente.

Nella Figura 1.4 compaiono i box plot delle righe e delle colonne della matrice dei dati *area1* descritta nella Tabella 1.1 e relativa ai valori di riflettanza delle acque inquinate di una zona costiera misurati via satellite. Se supponiamo che i dati che appaiono nelle prime righe della matrice corrispondano a locazioni più vicine alla costa, il confronto dei box plot, oltre ad individuare i dati anomali, fornisce una immediata ipotesi di lavoro: allontanandosi dalla costa il fenomeno osservato prima cresce debolmente di intensità e poi decresce rapidamente. Un analogo fenomeno non si osserva invece facendo variare l'indice di colonna ovvero spostandosi 'parallelamente' alla costa.

1.1.5 Test di Cressie

Cressie (1984) ha suggerito di provare l'ipotesi

$$H_0 : \text{assenza di valori anomali spaziali}$$

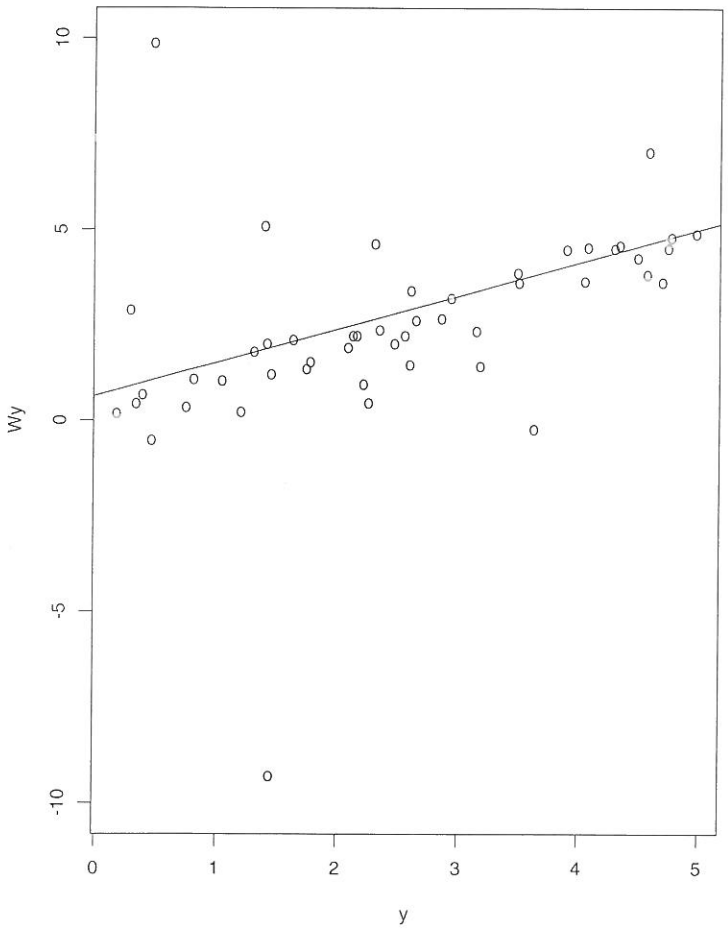

Figura 1.3: Esempio di un grafico (Wy,y) con retta interpolante; i punti cerchiati corrispondono a valori anomali rispetto ai contigui.

contro l'ipotesi

$$H_1 : \text{presenza di di valori anomali spaziali}$$

per mezzo del calcolo, per ogni riga e per ogni colonna, della statistica

$$D = \sqrt{n}\frac{\bar{y} - \text{med}(y)}{IQR\sqrt{0.5708}}(1.349)$$

dove $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ indica la media aritmetica delle osservazioni $y = (y_1, \ldots, y_n)$ e $\text{med}(y)$ è la loro mediana. Quando l'ipotesi H_0 è vera e le osservazioni sono indipendenti, D si distribuisce come una Normale di media 0 e varianza 1. E' da notare che, se i dati non sono indipendenti, il test tende a sovrastimare la presenza di valori anomali.

32	35	36	37	38	47	34	35	31
38	39	43	41	55	42	38	34	37
50	62	46	39	55	37	40	32	28
45	50	43	33	24	38	44	42	39
40	36	16	18	31	37	52	30	24
37	14	10	21	26	30	35	41	19
10	12	5	12	17	18	20	24	23
50	62	19	6	14	17	17	5	6
46	35	0	4	5	5	6	0	0

Tabella 1.1: Dati *area1*: valori di riflettanza (reflectance values) rilevati per mezzo di un satellite per un'area di acque costiere inquinate (Haining, 1987).

1.2 Individuazione di trend spaziali

La presenza di valori anomali nonchè, più in generale, le proprietà della distribuzione dei dati rilevate dai metodi descritti nella sezione precedente possono dipendere da una possibile eterogeneità spaziale dei dati medesimi. In questa sezione prendiamo in esame metodi esplorativi il cui scopo è quello di mettere in luce la presenza di trends spaziali ovvero di valori medi non stazionari dei dati.

In questa fase della ricerca l'obiettivo non è tanto quello della stima degli eventuali trends spaziali rispetto ad uno specificato modello statistico, quanto quello della eliminazione di una possibile non stazionarietà del primo ordine dai dati al fine di analizzare, per esempio, la struttura della variabilità spaziale degli stessi per mezzo di funzioni statistiche che assumono la stazionarietà del primo ordine come ipotesi di partenza; per esempio, per mezzo del semi-variogramma.

I metodi considerati sono quelli del median polishing e dell'interpolazione di superfici polinomiali con il metodo dei minimi quadrati.

1.2.1 Median polishing

Supponiamo che i dati siano osservati su una griglia con p righe e q colonne. Sia $y(s_1, s_2)$ il dato osservato nella cella corrispondente alla s_1-esima riga e alla s_2-esima colonna. Il metodo di *median polishing* assume un modello additivo tale che il valore numerico osservato nella cella (s_1, s_2) è rappresentato come

$$y(s_1, s_2) = \text{effetto comune} + \text{effetto riga} + \text{effetto colonna} + \text{residuo}. \qquad (1.1)$$

Le componenti del modello vengono così stimate:

$$\begin{aligned}
\text{effetto comune} &= y(.,.) \\
\text{effetto riga} &= y(s_1,.) - y(.,.) \\
\text{effetto colonna} &= y(.,s_2) - y(.,.) \\
\text{residuo} &= y(s_1,s_2) - y(.,.) - [y(s_1,.) - y(.,.)] - [y(.,s_2) - y(.,.)]
\end{aligned}$$

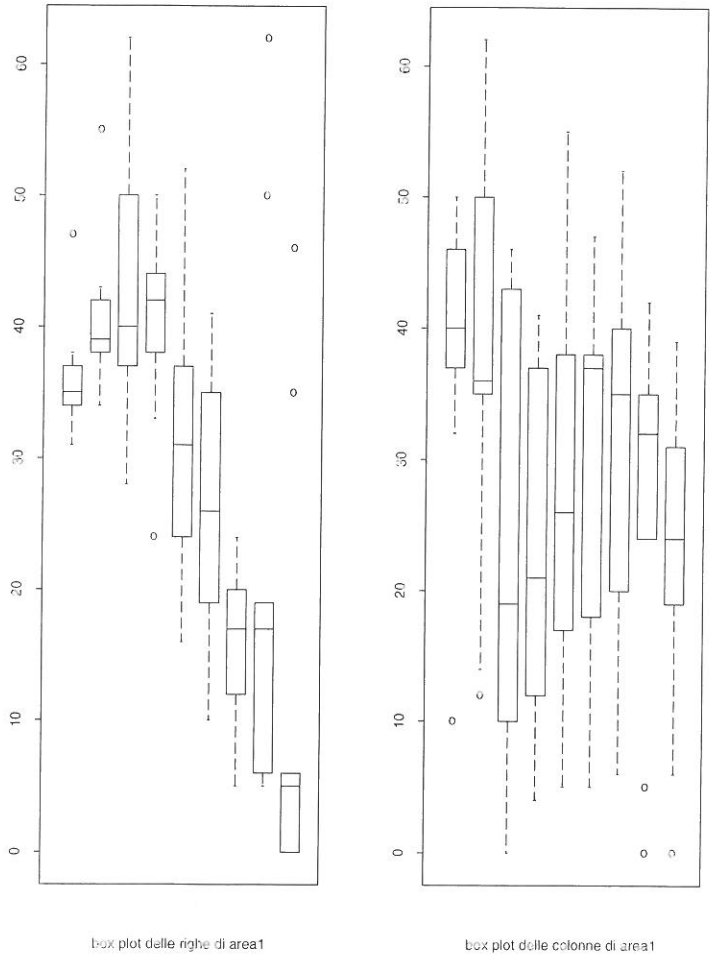

Figura 1.4: Confronto dei box plot per i dati delle righe e delle colonne della matrice dei dati *area1*.

dove il simbolo "." indica una media sui valori corrispondenti rispettivamente alla riga o alla colonna. La presenza di possibili valori anomali consiglia di usare medie troncate; spesso il troncamento viene fissato pari a 0.5, ovvero si fa uso delle mediane al posto delle medie.

Il modello è sufficientemente flessibile da poter essere usato anche quando vi siano dati mancanti in corrispondenza di alcune celle. E' inoltre possibile tenere conto di eventuali interazioni righe-colonna modificando il modello (1.1) nel seguente:

$$y(s_1, s_2) = \text{e. comune} + \text{e. riga} + \text{e. colonna} + k \left\{ \frac{\text{e.riga} \times \text{e.colonna}}{\text{e.comune}} \right\} + \text{residuo}. \qquad (1.2)$$

Il valore k viene così stimato: si calcolano le componenti del modello (1.1) e quindi si interpola una retta per i punti la cui ordinata è data dai valori residui del modello e la cui ascissa è data dai valori corrispondenti all'espressione che appare tra parentesi graffa nel modello (1.2). Se il coefficiente angolare di questa retta è diverso da zero, esso viene assunto come

effetti riga:	5	10	10	12	0	-5	-14	-15	-26
effetti colonna:	9	5	0	-5	0	1	1	-1	-5
effetto comune:	31								

Tabella 1.2: Median polishing dei dati *area1*; stime degli effetti (Haining, 1987).

Riga	1	2	3	4	5	6	7	8	9
D	-0.47	-0.52	0.89	-3.77	0.0	-0.71	-0.84	3.49	6.07
Colonna	1	2	3	4	5	6	7	8	9
D	0.46	1.13	-1.61	0.0	2.67	0.89	3.12	-0.39	-0.79

Tabella 1.3: Valore della statistica D per i residui dopo median polishing dei dati *area1* (Haining, 1987).

stima del valore k. Nel caso invece il coefficiente angolare sia approssimativamente nullo, si ritiene valido il modello (1.1).

Valori grandi per gli effetti riga o colonna sono a sostegno della presenza di trends spaziali. Il suggerimento è in questo caso quello di effettuare le analisi statistiche, soprattutto quelle relative allo studio sulla struttura della variabilità spaziale dei dati, sui residui (dati detrendizzati) e non sui dati originari. Sarà talvolta opportuno riesaminare la presenza di valori anomali applicando i metodi descritti nella sezione precedente ai dati detrendizzati.

Nella Tabella 1.2 sono riportati i risultati relativi al median polishing dei dati *area1* che appaiono nella Tabella 1.1. La presenza di effetti riga è molto evidente e ciò conferma l'ipotesi già formulata sulla base delle osservazioni dei box plot per riga.

La presenza di valori anomali dipende spesso anche da eventuali trend spaziali. Nella Tabella 1.3 compaiono i valori della statistica D di Cressie per le righe e le colonne di *area1* dopo che i dati sono stati detrendizzati per mezzo di di median polishing ovvero calcolata rispetto ai residui. Sono evidenti valori grandi di D in corrispondenza delle righe 4, 8 e 9 e delle colonne 5, 7 che fanno pensare alla presenza di valori estremi nelle locazioni corrispondenti.

1.2.2 Interpolazione di superfici polinomiali

Un approccio analogo al precedente per lo studio della presenza di trends spaziali è quello che interpola i dati osservati su una griglia per mezzo di una superficie polinomiale.

Se la griglia ha p righe e q colonne, indichiamo con $y = (y_1, \ldots, y_n)$, $n = p \times q$, il vettore dei dati osservati. Per $i = 1, \ldots, n$, siano (s_{i1}, s_{i2}) le coordinate relative al dato y_i. Assumiamo che il vettore y possa essere rappresentato come

$$y = A\theta + \text{vettore dei residui}$$

dove

$$A = \begin{bmatrix} 1 & s_{11} & s_{12} & s_{11}s_{12} & s_{11}^2 & s_{12}^2 & \dots & s_{11}^h s_{12}^k \\ 1 & s_{21} & s_{22} & s_{21}s_{22} & s_{21}^2 & s_{22}^2 & \dots & s_{21}^h s_{22}^k \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \dots & \cdot \\ 1 & s_{n1} & s_{n2} & s_{n1}s_{n2} & s_{n1}^2 & s_{n2}^2 & \dots & s_{n1}^h s_{n2}^k \end{bmatrix}$$

è una matrice e

$$\theta = (\theta_0, \theta_1, \theta_2, \theta_{11}, \dots, \theta_{hk})$$

è un vettore di coefficienti.

La somma $h + k$ corrisponde all'ordine della superficie interpolante: così se $h + k = 1$ la superficie interpolante è un piano, se $h + k = 2$ abbiamo una superficie quadratica etc.

Scelta la matrice A, i coefficienti del vettore θ vengono calcolati con il metodo dei minimi quadrati ovvero in modo tale che sia minima la distanza

$$\sum_{i=1}^{n} [y_i - (A\theta)_i]^2$$

Poichè la scala secondo la quale si misurano le coordinate sulla griglia è arbitraria, è consigliabile scegliere A in modo tale che nel polinomio risultante compaiono tutti i termini di grado minore o uguale a $h + k$.

Per la scelta del grado $h + k$ si può seguire una procedura a passi (*stepwise regression*) nel modo seguente: si interpola innanzitutto una superficie di ordine zero, a cui corrsiponde una matrice $A = A_0$, e si calcola il coefficiente di correlazione lineare R_0^2 tra y e $A_0\theta$. Poi si considera una superficie di ordine 1, a cui corrisponde una matrice $A = A_1$, e si calcola il coefficiente di correlazione lineare R_1^2 tra y e $A_1\theta$, e così via. Al crescere dell'ordine $h+k$ della superficie interpolata, cresce il valore del coefficiente R_{h+k}^2. Si arresta la procedura quando la crescita di R^2 non è più apprezzabile.

1.3 Analisi della stazionarietà del secondo ordine

Lo scopo di questa sezione è quello di introdurre ad alcuni metodi esplorativi per l'analisi della stazionarietà del secondo ordine di dati spaziali osservati su di una griglia. Assumeremo che la griglia abbia molte celle e che queste abbiano aree pressochè uguali.

Supponiamo che $s = (s_1, s_2)$ e $t = (t_1, t_2)$ indichino due celle sulla griglia e siano $y(s_1, s_2)$ e $y(t_1, t_2)$ rispettivamente i valori osservati nelle due celle. Assumiamo che il processo aleatorio che ha generato le osservazioni sia stazionario del primo ordine ovvero che il valor medio di y sia costante in ogni cella; i metodi sviluppati nella sezione precedente permettono di controllare questa ipotesi. Il processo aleatorio che ha generato i dati si dice *debolmente stazionario del secondo ordine* se, per ogni s e t, la *covarianza*

$$C(s,t) = \text{Media}\{[y(s_1, s_2) - \text{Media(y)}][y(t_1, t_2) - \text{Media(y)}]\}$$

non dipende da s e t, ma solo dalla loro posizione relativa ovvero dal vettore distanza $h = s - t$. Nel caso in cui $C(s,t)$ dipenda solo dalla lunghezza r di h allora il processo viene detto *isotropo*.

Quando non è soddisfatta l'ipotesi di stazionarietà del primo ordine, una più debole assunzione di stazionarietà del secondo ordine è quella che prende in considerazione gli scarti dai valori medi di y. Avremo allora che il processo aleatorio delle differenze dal valore medio è debolmente stazionario del secondo ordine se, per ogni per ogni s e t, la covarianza

$$C(s,t) = \text{Media}\{[y(s_1,s_2) - \text{Media}(y(s_1,s_2))][y(t_1,t_2) - \text{Media}(y(t_1,t_2))]\}$$

non dipende da s e t, ma solo dal vettore distanza $h = s - t$.

1.3.1 Stime della struttura della varianza e covarianza spaziale per diverse partizioni della regione studiata. Correlogramma e semi-variogramma

Una analisi esplorativa dei dati osservati per la verifica delle proprietà di stazionarietà del secondo ordine è sempre difficile e incompleta. Un possibile approccio, valido soprattutto quando si abbia a che fare con griglie con molte celle e di uguali dimensioni, è quello che suddivide l'area di studio in diverse sottoregioni e, per ogni sottoregione, stima la covarianza C per diversi vettori distanza h. Un confronto tra le stime ottenute nelle diverse sottoregioni fornisce importanti informazioni relative alla struttura della variabilità spaziale del fenomeno in esame e può essere a sostegno dell'ipotesi di stazionarietà del secondo ordine.

Supponiamo che la griglia abbia dimensioni $p \times q$ e fissiamo come distanza il vettore $h = (j,k)$, con j e k interi relativi. Allora la covarianza $C(j,t)$ tra tutte le celle s e t tali che $s - t = h = (j,k)$ è stimata dalla statistica

$$\hat{C}(j,t) = \frac{1}{pq} \sum_{s_1=m_1}^{M_1} \sum_{s_2=m_2}^{M_2} [y(s_1,s_2) - \text{Media}(y(s_1,s_2))][y(s_1+j,s_2+k) - \text{Media}(y(s_1+j,s_2+k))]$$

dove $m_1 = \max(1, 1-j), m_2 = \max(1, 1-k), M_1 = \min(p, p-j)$ e $M_2 = \min(q, q-k)$. Una diversa statistica per la medesima stima che tiene conto degli effetti 'bordo' è invece la seguente:

$$C^*(j,k) = \frac{pq}{(p-j)(q-k)} \hat{C}(j,k).$$

Si osservi che nelle espressioni precedenti compare il valore $\text{Media}(y(s_1,s_2))$ ovvero il valor medio del processo che ha generato le osservazioni calcolato in corrispondenza della cella di coordinate (s_1,s_2). Questa quantità è di solito sconosciuta e andrà stimata con gli strumenti introdotti nella sezione precedente ossia per mezzo del median polishing oppure interpolando una superficie polinomiale.

Se il processo che ha generato le osservazioni è isotropo, le precedenti statistiche devono dipendere solo dalla lunghezza r del vettore h e non dalla sua direzione. Sarà in questo caso possibile rappresentare graficamente le statistiche $\hat{C}$ e C^* come funzioni di r ossia per mezzo di un *correlogramma*. Per esplorare l'ipotesi di isotropia confronteremo i correlogrammi ottenuti fissando alcune direzioni per h e facendo variare il suo modulo; per esempio confronteremo il correlogramma ottenuto fissando per h la direzione Nord-Sud con quello ottenuto fissando la direzione Est-Ovest. Quando il processo è isotropo questi grafici risultano approssimativamente uguali. In caso contrario otteniamo delle importanti informazioni relative alla variabilità spaziale del fenomeno in esame.

Un analogo approccio allo studio della variabilità spaziale è quello che fa uso del semi-variogramma. Siano j e k due interi relativi e consideriamo la quantità

$$\gamma(j,k) = \frac{1}{2} Var[y(s_1 + j, s_2 + k) - y(s_1, s_2)]$$

Quando, per ogni $h = (j,k)$, $\gamma(j,k)$ dipende solo da h e non da $s = (s_1, s_2)$ essa prende il nome di *semi-variogramma*. Il semi-variogramma si dice *isotropo* nel caso in cui $\gamma(j,k)$ dipenda solo dalla lunghezza del vettore h.

Se il processo aleatorio che genera le osservazioni è debolmente stazionario del secondo ordine, si può dimostrare che

$$\gamma(j,k) = C(0,0) - C(j,k).$$

Il semi-variogramma può essere stimato per mezzo della statistica

$$\hat{\gamma}(j,k) = \frac{1}{2(p-j)(q-k)} \sum_{s_1=1}^{p-j} \sum_{s_2=1}^{q-k} [y(s_1, s_2) - y(s_1 + j, s_2 + k)]^2$$

quando si assuma che i dati siano stati generati da un processo a valori medi stazionari. Pertanto, prima di calcolare la statistica $\hat{\gamma}$ è opportuno rimuovere ogni tipo di trend per mezzo del median polishing o di interpolazione di superfici polinomiali.

Grafici della statistica $\hat{\gamma}$ per diverse direzioni del vettore $h = (j,k)$ e per diversi valori del suo modulo forniscono importanti informazioni relative alla struttura della variabilià spaziale del fenomeno oggetto di analisi; il loro confronto permette, per esempio, una immediata verifica empirica dell'ipotesi di isotropia. Inoltre la forma del grafico di $\hat{\gamma}$ spesso è di aiuto nella scelta di un modello per il processo aleatorio che si assume abbia generato i dati osservati.

Se i dati osservati y sono riferiti ad una griglia $p \times q$ è possibile rappresentare i valori delle statistiche $\hat{C}, C^*$ o $\hat{\gamma}$, al variare di $h = (j,k)$, per mezzo di una matrice di dimensioni $2p \times 2q$. Questa matrice si presta ad interessanti visualizzazioni grafiche.

Per esempio; possiamo rappresentare per mezzo di un *countour plot* la superficie che vale $\hat{C}(i,j)$ in corrispondenza della cella (i,j), con $i \in \{-p, -p+1, \ldots, -1, 1, \ldots, p\}$ e $j \in \{-q, -q+1, \ldots, -1, 1, \ldots, q\}$. Si osservi che le linee di contorno di questa superficie saranno necessariamente simmetriche poichè $\hat{C}(i,j) = \hat{C}(-i,-j)$. Se è soddisfatta l'ipotesi di isotropia ci aspettiamo che le linee di contorno siano approssimativamente delle circonferenze con centro nell'origine. Anche quando ciò non si verifichi si avrà comunque una semplice rappresentazione grafica della struttura della correlazione spaziale; questa rappresentazione può essere molto utile per confrontare la struttura della variabilità di dati univariati diversi con ugual riferimento spaziale (per esempio, Biomassa e Temperatura).

1.4 Analisi dell'autocorrelazione spaziale

Lo studio dell'*autocorrelazione* di dati con riferimento spaziale parte dalla constatazione che spesso il valore di un dato fornisce informazioni relative ai valori dei dati contigui. Pertanto in presenza di autocorrelazione l'informazione fornita da una osservazione relativa al

fenomeno in esame si sovrappone a quella fornita dalle osservazioni contigue e il ricercatore dovrà tenere conto di questo fatto nelle sue valutazioni. Diremo di essere in una situazione di assenza di autocorrelazione quando la variazione tra valori contigui non è mediamente diversa dalla variazione tra tutte le possibili coppie di valori osservati. Siamo invece in presenza di autocorrelazione positiva quando valori contigui sono mediamente simili. Viceversa l'autocorrelazione è negativa quando valori contigui sono mediamente diversi tra loro.

Lo studio della autocorrelazione spaziale verrà condotto in questa sezione per mezzo del calcolo di alcune statistiche tra le quali spiccano l'indice di Geary e quello di Moran.

Nel seguito indichiamo con con $y = \{y_1, \ldots, y_n\}, n = p \times q$, il vettore dei dati osservati su di una griglia di dimensioni $p \times q$ e assumiamo che la struttura di contiguità spaziale tra le celle della griglia sia descritta dalla matrice W.

1.4.1 L'indice di Geary

Supponiamo che la matrice W sia tale che

$$w_{ij} = \begin{cases} 1 & \text{se la cella } i\text{-esima è contigua alla cella} j\text{-esima} \\ 0 & \text{altrimenti.} \end{cases} \tag{1.3}$$

Siano

$$v_1 = \frac{\sum_{i \neq j} w_{ij}(y_i - y_j)^2}{\sum_{i \neq j} w_{ij}}$$

e

$$v_0 = \frac{\sum_{i \neq j}(y_i - y_j)^2}{n(n-1)}.$$

L'*indice di Geary* è definito come

$$c = \frac{v_1}{v_0}. \tag{1.4}$$

Per dirla con Lebart (1969), l'indice c misura il rapporto tra due stime della varianza la prima delle quali tiene conto della struttura di contiguità territoriale, mentre la seconda assume l'indipendenza delle osservazioni.

L'indice c assume valori non negativi; in particolare c assume valore uguale ad 1 in assenza di autocorrelazione spaziale, valori maggiori di 1 crescenti all'aumentare della autocorrelazione negativa e minori di 1 decrescenti all'aumentare della autocorrelazione positiva.

1.4.2 L'indice di Moran

Assumiamo ancora che la matrice W sia descritta dalla (1.3) e definiamo l'indice di Moran per mezzo della

$$I = \frac{\frac{\sum_{i \neq j} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i \neq j} w_{ij}}}{\frac{\sum_i (y_i - \bar{y})}{n}} \tag{1.5}$$

dove $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$ è la media campionaria delle osservazioni.

In assenza di autocorrelazione I assume valore $-(n-1)^{-1}$; si ha invece autocorrelazione positiva per valori di I maggiori di $-(n-1)^{-1}$ e negativa per valori minori.

Confrontando l'indice di Geary con quello di Moran, si osserva che c è più sensibile agli scostamenti assoluti tra coppie di valori mentre I è più sensibile ad eventuali valori anomali estremi.

1.4.3 Indici di autocorrelazione con ordine di contiguità superiore al primo

I valori assunti dagli indici c ed I dipendono strettamente dal tipo di matrice di contiguità considerata nelle loro espressioni. E' immediato pensare ad una generalizzazione di c e di I che tenga conto di ordini di contiguità superiore al primo modificando opportunamente l'espressione di W. Per esempio, per $r > 0$, possiamo definire

$$w_{ij}(r) = \begin{cases} 1 & \text{se la cella } i\text{-esima dista } r \text{ dalla cella} j\text{-esima} \\ 0 & \text{altrimenti.} \end{cases} \tag{1.6}$$

Sostituendo nella (1.4) e nella (1.5) le quantità w_{ij} con le $w_{ij}(r)$ si ottengono rispettivamente gli indici $c(r)$ e $I(r)$ che misurano l'autocorrelazione spaziale con ordine di contiguità r. Al variare di r, possiamo rappresentare $c(r)$ e $I(r)$ per mezzo di un correlogramma.

Osserviamo infine che gli indici c ed I sono casi particolari di statistiche prodotto, studiate da Huber.et.al. (1981), del tipo

$$\Gamma = \sum_{i,j} G_{ij} C_{ij}.$$

dove G_{ij} misura la prossimità spaziale tra la cella i-esima e quella j-esima mentre C_{ij} valuta la differenza tra y_i e y_j.

1.4.4 Prova dell'ipotesi

Consideriamo ora il problema relativo alla prova della ipotesi

$$H_0 : \text{assenza di autocorrelazione spaziale}$$

contro l'ipotesi bilaterale

$$H_1 : \text{autocorrelazione positiva o negativa.}$$

Cliff e Ord (1981) e Sen (1976) dimostrano che, quando l'ipotesi H_0 è vera e per valori sufficientemente grandi di n, la distribuzione di c e di I è Normale con media e varianza dati da

$$\begin{aligned} \text{Media}(I) &= -\frac{1}{n-1} \\ \text{Var}(I) &= \frac{n^2 S_1 - n S_2 + 3 S_0^2}{(n^2-1) S_0^2} - \frac{1}{(n-1)^2} \\ \text{Media}(C) &= 1 \\ \text{Var}(C) &= \frac{(2S_1 + S_2)(n-1) - 4 S_0^2}{2(n+1) S_0^2} \end{aligned}$$

dove

$$\begin{aligned} S_0 &= \sum_{i,j} w_{ij} \\ S_1 &= \frac{1}{2}\sum_{ij}(w_{ij} + w_{ji})^2 \\ S_2 &= \sum_{i}(w_{i\cdot} + w_{\cdot j})^2 \end{aligned}$$

e $w_{i\cdot} = \sum_j w_{ij}$ mentre $w_{\cdot j} = \sum_i w_{ij}$.

Cliff e Ord considerano anche situazioni ove n è piccolo e non sembra ragionevole assumere l'ipotesi di Normalità per c ed I quando H_0 è vera. In questi casi una possibile alternativa è quella di approssimare la distribuzione di queste statistiche, così come quella di più complicate statistiche prodotto, per mezzo di metodi di simulazione *Monte Carlo* o di metodi di ricampionamento quale il *bootstrap*.

Capitolo 2

Costruzione di mappe confrontabili per la Temperatura e per la Biomassa

L'analisi statistica congiunta delle osservazioni campionarie di due variabili aleatorie spazialmente distribuite su una stessa area, quali la Biomassa e la Temperatura nel Mare Adriatico, presuppone spesso che il sistema di riferimento spaziale rispetto al quale i dati sono stati osservati sia lo stesso per entrambe le variabili. Ciò è, per esempio, vero per le analisi della correlazione e dell'associazione spaziale tra Biomassa e Temperatura che descriveremo nel prossimo capitolo.

Quando i due insiemi di dati non hanno il medesimo riferimento spaziale si pone il problema relativo al modo di aggregare i dati di almeno un insieme, per esempio quelli relativi alla Temperatura, al fine di creare mappe, una per ogni insieme di dati, che avendo lo stesso riferimento spaziale siano tra loro confrontabili.

Non esiste un modo univoco di procedere e la corretta soluzione a questo problema di aggregazione dipende verosimilmente dalle caratteristiche dei dati in esame. Caratteristiche note solo dopo aver compiuto le analisi esplorative descritte nel precedente capitolo.

2.1 Una metodo di aggregazione per i dati della Temperatura

Nel caso particolare dei dati relativi alla Biomassa e alla Temperatura nel Mare Adriatico, la griglia di riferimento spaziale per i dati della Temperatura è più fine di quella per i dati della Biomassa. Chiamiamo G_B la griglia di riferimento per i dati della Biomassa e G_T quella per i dati della Temperatura. Poichè ad ogni cella di G_B corrispondono $k \geq 2$ celle di G_T, una prima soluzione al problema dell'aggregazione dei dati per la costruzione di mappe confrontabili sembra essere quella per cui ad ogni cella di G_B si fa corrispondere un valore della Temperatura che sia una media ponderata dei valori registrati nelle k celle di G_T in essa incluse. La scelta dei pesi di ponderazione dipenderà, in generale, dalla autocorrelazione dei dati medesimi.

Una proposta più articolata è quella che ora descriviamo.

Identifichiamo la locazione geografica di ogni dato di Temperatura con il centro della

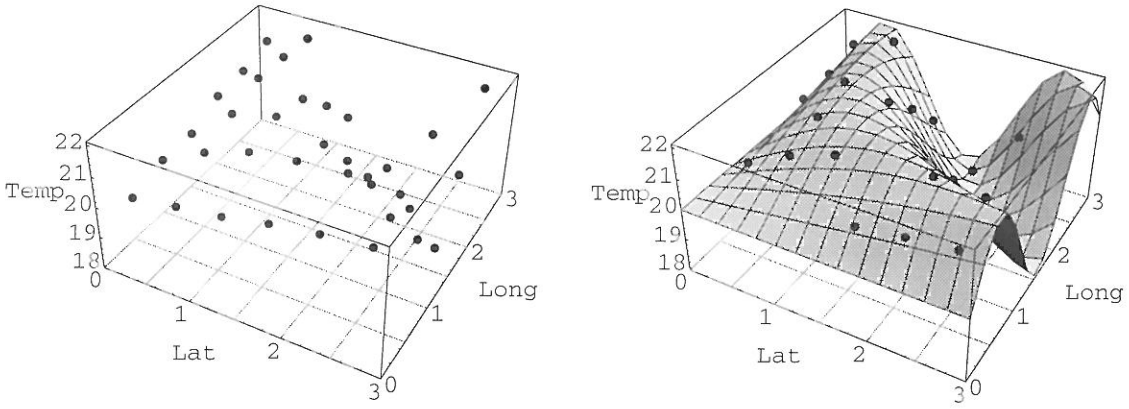

Figura 2.1: Nel grafico di sinistra sono riportati i punti nello spazio che identificano le temperature rilevate nelle celle di una griglia G_T; l'asse delle ascisse s_1 è stato indicato con *Lat*, quello delle ordinate s_2 con *Long*. Nel grafico di destra è mostrata una superficie interpolante.

cella di G_T alla quale corrisponde; ovvero, trasformiamo i dati della Temperatura in punti nello spazio di coordinate $(s_1(i), s_2(i), T(i))$, per $i = 1, \ldots, N_T$, dove $T(i)$ è la temperatura della cella i-esima di G_T, $(s_1(i), s_2(i))$ sono le coordinate del centro della medesima cella e N_T è il numero totale di celle della griglia G_T.

A questo punto interpoliamo una superficie di equazione

$$T = f(s_1, s_2)$$

per i punti nello spazio così ottenuti. La funzione f potrà essere un polinomio di grado opportuno, e in tal caso i metodi di interpolazione sono stati descritti nella Sezione 1.2.2, oppure il nucleo di una densità bidimensionale o, più in generale, un modello che tenga conto della struttura di autocorrelazione osservata per i dati della Temperatura o di altre caratteristiche che si vuole conservate dalla superficie di equazione f. La Figura 2.1 illustra il risultato di una siffatta interpolazione su dati simulati.

Infine, per ogni cella c_j della griglia G_B relativa ai dati della Biomassa, con $j = 1, \ldots, N_B$ e N_B uguale al numero di celle della griglia G_B, si considera un valore della Temperatura pari a

$$\hat{T}(j) = \frac{1}{|c_j|} \int_{c_j} f(s_1, s_2)\, ds_1 ds_2$$

dove abbiamo indicato con $|c_j|$ l'area della cella c_j.

Diverse scelte della funzione f forniranno stime alternative per la temperatura $\hat{T}(j)$ ognuna delle quali è una opportuna *media alla Chisini* dei dati della Temperatura rilevati nelle celle della griglia G_T contenute in c_j. Per esempio; se scegliamo come f la funzione che, in ogni cella c_j, determina il piano che minimizza l'errore quadratico medio e che interpola i punti $(s_1(i), s_2(i), T(i))$ al variare di $(s_1(i), s_2(i))$ in c_j, allora

$$\hat{T}(j) = \frac{1}{\#\{T(i) : (s_1(i), s_2(i)) \in c_j\}} \sum_{(s_1(i), s_2(i)) \in c_j} T(i)$$

risulterà essere uguale alla media aritmetica dei valori $T(i)$.

Capitolo 3

Analisi delle relazioni spaziali tra Temperatura e Biomassa

In questo capitolo ci proponiamo di analizzare alcuni strumenti atti all'analisi statistica congiunta di due caratteri, quali la Biomassa e la Temperatura, rilevati per unità statistiche che abbiano un comune riferimento spaziale. In particolare nelle prossime sezioni definiremo alcuni indici, funzioni delle osservazioni campionarie, che permettono di misurare il grado di correlazione oppure quello di associazione tra la Biomassa e la Temperatura tenendo conto della distribuzione spaziale dei dati osservati.

Supponiamo, per il momento, che $I(B,T)$ sia uno degli indici di correlazione o di associazione tra Biomassa e Temperatura definiti nelle sezioni che seguono. L'informazione fornita da I può essere utilizzata per indagare problemi tra loro diversi. Per esempio:

(a) Data una certa regione del mare Adriatico, il Golfo di Trieste, potremmo voler verificare l'ipotesi

$$H_0 : \text{assenza del fenomeno misurato da } I \text{ (correlazione o associazione)}$$

contro l'ipotesi alterniva di presenza del fenomeno misurato da I. Per risolvere questo problema dobbiamo conoscere la distribuzione di $I(B,T)$, per lo meno quando l'ipotesi H_0 è vera e quando si abbia a disposizione un numero elevato di osservazioni. Per questo motivo proporremo nel seguito statistiche di cui è nota in letteratura la distribuzione asintotica. Quando la distribuzione di I non sia derivabile analiticamente, è spesso possibile approssimarla tramite metodi di ricampionamento o di simulazione.

(b) Possiamo confrontare diverse regioni o zone del Mare Adriatico in base al diverso grado di correlazione o di associazione spaziale tra la Biomassa e la Temperatura misurato da I in un istante di tempo fissato (in un anno dato). Il confronto può avvenire lungo direzioni di interesse (quella Nord-Sud, per esempio) e in questo caso i risultati dell'analisi possono essere sinteticamente rappresentati per mezzo di un grafico dove vengano riportati i valori di I al variare di un indicatore geografico (la latitudine). Oppure si possono mettere a confronto rispetto al valore di I aree del Mare Adriatico dove si ritiene che le piene del Po o i venti di Bora abbiano un maggiore influenza sulla Biomassa o sulla Temperatura con aree dove si ritiene che la loro influenza sia nulla.

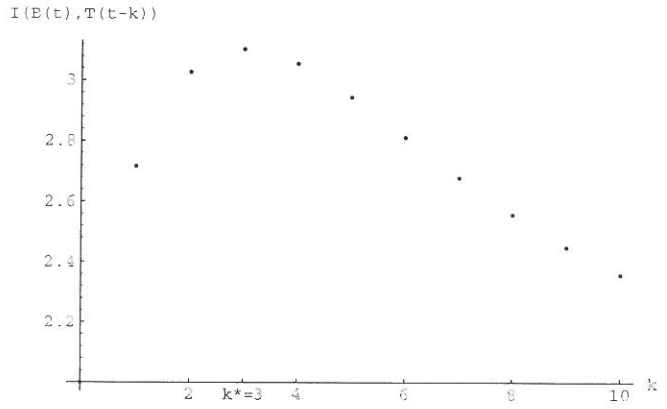

Figura 3.1: Grafico di $I(B(t), T(t-k))$ al variare di k. Il *lag temporale* di 3 unità di tempo è quello che rende massima la correlazione, o la associazione, tra Biomassa e Temperatura.

(c) Una ipotesi di lavoro nell'analisi della Biomassa e della Temperatura nel Mare Adriatico è quella secondo la quale la distribuzione spaziale della Biomassa in un dato istante temporale dipende dalla distribuzione spaziale della Temperatura in periodi precedenti. Un approccio esplorativo alla verifica di questa ipotesi potrebbe essere il seguente; si calcola un indice di correlazione, o di associazione, $I(B(t), T(t-k))$ tra i valori della Biomassa in un dato istante t (oggi) e i valori della Temperatura osservata nell'istante $t-k$, con $k = 0, 1, 2, \ldots$ (ovvero, oggi, ieri, l'altroieri, etc.). Se rappresentiamo su un grafico i valori di $I(B(t), T(t-k))$ al variare di $k = 0, 1, 2, \ldots$ possiamo sperare di individuare un valore $k^*(t)$ in corrispondenza del quale l'indice I è massimo (Figura 3.1). Se ciò accade possiamo concludere che, in prima approssimazione, i valori della Biomassa all'istante t sono massimamente correlati, o associati, con quelli della Temperatura osservata $k^*(t)$ istanti prima.

Ovviamente la stesso tipo di analisi può essere fatta considerando non i valori $T(t-k)$ della Temperatura k istanti prima, ma il valor medio della Temperatura nei k istanti precedenti. Oppure il valor medio di T negli r istanti che precedevano l'istante $t-k$, con r intero e finito.

Il confronto tra le curve così ottenute per diversi valori di t (ovvero nei diversi anni, se la Biomassa viene misurata una sola volta nell'anno), e più in particolare delle quantità $k^*(t)$, fornisce utili indicazioni su come e quanto la correlazione, o l'associazione , tra Temperatura e Biomassa sia cambiata nel tempo per mezzo di un'analisi che tiene anche conto della distribuzione nello spazio delle due variabili in questione.

Altrettanto utile potrebbe essere il confronto dei valori di $k^*(t)$ relativi a diverse zone del Mare Adriatico in un istante t fissato (in un dato anno): per esempio lungo particolari direzioni di riferimento, o tra aree soggette o meno ai venti di Bora, alle piene del Po etc.

3.1 Misura della correlazione spaziale tra Biomassa e Temperatura

Gli strumenti classici per lo studio della correlazione tra due caratteri sono l'indice di Pearson e quello di Spearman. Dopo averli definiti nella prossima sezione, descriveremo una procedura, dovuta a Clifford e Richardson (1985), che rende possibile l'uso di queste statistiche per la verifica dell'ipotesi di assenza di correlazione spaziale.

3.1.1 Indici di Pearson e di Spearman

Consideriamo una popolazione di n unità statistiche per ognuna delle quali sono osservate due quantità B e T. Per esempio, le unità statistiche potrebbero essere le n celle di una griglia $p \times q$ secondo la quale si è suddiviso il mare Adriatico e B e T potrebbero rappresentare, rispettivamente, i valori della Biomassa e della Temperatura registrati in corrispondenza di ogni cella.

Quando si suppone che i dati provenienti da celle diverse sono indipendenti, la misura della correlazione tra B e T è solitamente condotta per mezzo del calcolo dell'*indice di Pearson*

$$\hat{r} = \frac{\sum_{i=1}^{n}(b_i - \bar{b})(t_i - \bar{t})}{[\sum_{i=1}^{n}(b_i - \bar{b})^2 \sum_{i=1}^{n}(t_i - \bar{t})^2]^{\frac{1}{2}}},$$

dove

$$\bar{b} = \frac{1}{n}\sum_{i=1}^{n} b_i, \quad \bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i$$

oppure per mezzo dell'*indice di Spearman* basato sui ranghi

$$r_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n} d_i^2$$

dove d_i è la differenza tra il rango della osservazione b_i e il rango dell'osservazione t_i. Il *rango* di una osservazione è il posto che questa occupa nella successione delle osservazioni ordinate dalla più piccola alla più grande.

Inoltre, se le osservazioni che provengono da celle diverse sono indipendenti e selezionate da una popolazione con distribuzione approssimativamente Normale, l'ipotesi

$$H_0 : \text{assenza di correlazione tra } B \text{ e } T$$

può essere verificata per mezzo della funzione test

$$(n-2)^{\frac{1}{2}} \frac{|\hat{r}|}{(1-\hat{r}^2)^{\frac{1}{2}}}$$

che, quando H_0 è vera, ha distribuzione t di Student con $n-2$ gradi di libertà. Sono anche di uso corrente test analoghi che si basano su $\hat{r}_S$ anzichè $\hat{r}$.

Sia $\hat{r}$ che $\hat{r}_S$ misurano la correlazione punto a punto tra B e T senza tener conto, nella loro espressione, delle relazioni di tipo spaziale tra i dati osservati. Ma queste relazioni rendono manifestamente discutibile l'ipotesi di indipendenza tra le osservazioni relative a celle diverse e, in generale, comportano una sottostima della varianza delle statistiche in questione.

3.1.2 Modifica di Clifford e Richardson

Lo scopo di questa sezione è quello di descrivere una procedura proposta da Clifford e Richardson (1985) che permette di verificare l'ipotesi

$$H_0 : \text{assenza di correlazione tra } B \text{ e } T$$

con l'uso di $\hat{r}$ tenendo conto della distribuzione spaziale dei dati.

Supponiamo che per ogni cella di una griglia $p \times q$ siano stati osservati i valori di due variabili B e T. Innanzitutto i dati relativi a B e a T vengono detrendizzati per mezzo dei metodi descritti nella Sezione 1.2, in modo da ottenere insiemi di residui indipendenti per B e T rispettivamente. La statistica $\hat{r}$ viene calcolata in relazione a questi residui. Infine l'ipotesi H_0 di assenza di correlazione tra B e T viene rifiutata quando la quantità

$$(N'-2)^{\frac{1}{2}} \frac{|\hat{r}|}{(1-\hat{r}^2)^{\frac{1}{2}}}$$

supera il valore critico ottenuto per mezzo di una distribuzione t di Student con $N'-2$ gradi di libertà. Il valore di N' è posto approssimativamente uguale a

$$N' \approx 1 + \left[\frac{\sum N_k C_1^*(k) C_2^*(k)}{n^2 S_b^2 S_t^2} \right]^{(-1)}$$

dove N_k è il numero di celle separate da una distanza k, le quantità C^* sono le stime delle covarianze spaziali di distanza k definite nella Sezione 1.3.1 , mentre

$$S_b^2 = \frac{1}{n} \sum_{i=1}^{n} (b_i - \bar{b})^2, S_t^2 = \frac{1}{n} \sum_{i=1}^{n} (t_i - \bar{t})^2.$$

3.2 Misure dell'associazione

Le misure di associazione tra due variabili B e T con egual riferimento spaziale hanno come obiettivo quello di quantificare il grado secondo il quale valori 'simili' di B e di T sono spazialmente vicini.

3.2.1 L'indice di Tiøstheim

Consideriamo una griglia $p \times q = n$ e ordiniamone le celle prima secondo i valori crescenti di B e poi secondo quelli crescenti di T. Ovvero indichiamo con

$$l_B(i) = (s_1^B(i), s_2^B(i))$$

le coordinate cartesiane della cella alla quale corrisponde l'i-esimo valore osservato di B in ordine crescente. Analogamente indichiamo con

$$l_T(i) = (s_1^T(i), s_2^T(i))$$

le coordinate cartesiane della cella alla quale corrisponde l'i-esimo valore osservato di T in ordine crescente.

Tiøstheim (1978) ha proposto di quantificare l'associazione spaziale tra B e T per mezzo di un indice del tipo

$$\Lambda = \sum_{i=1}^{n} d(l_B(i), l_T(i)),$$

eventualmente normalizzato, dove $d(\cdot,\cdot)$ è una opportuna distanza tra le celle. In altre parole: si misura quanto siano spazialmente distanti celle che hanno ugual posto nella successione ordinata secondo valori crescenti rispettivamente di B e di T.

Tiøstheim propone di usare come distanza tra le celle di indice i la

$$d(l_B(i), l_T(i)) = s_1^B(i)s_1^T(i) + s_2^B(i)s_2^T(i)$$

e definisce di conseguenza l'indice di associazione normalizzato

$$A = \frac{\sum_{i=1}^{n} s_1^B(i)s_1^T(i) + s_2^B(i)s_2^T(i)}{\sum_{i=1}^{n}[(s_1^B(i))^2 + (s_2^B(i))^2]} = (0.5)[r(s_1^B, s_1^T) + r(s_2^B, s_2^T)]$$

dove $r(s_1^B, s_1^T)$ e $r(s_2^B, s_2^T)$ sono rispettivamente i coefficienti di correlazione di Pearson (definiti nella sezione precedente) tra i valori delle ascisse e tra quelli delle ordinate delle celle $l_B(i)$ e $l_T(i)$. Si dimostra che, se B e T sono indipendenti e le n osservazioni sia di B che di T sono spazialmente indipendenti tra loro, allora la statistica A ha media nulla e varianza

$$Var[A] = \frac{2n^2}{n-1}(1 + r(s_1, s_2))$$

dove $r(s_1, s_2)$ è il coefficiente di correlazione lineare tra le ascisse e le ordinate delle n locazioni che compongono la griglia di dimensioni $p \times q = n$.

Per provare l'ipotesi

$$H_0 : \text{assenza di associazione spaziale}$$

possiamo far uso della statistica A assumendo che, per n grande e quando H_0 è vera, essa abbia distribuzione Normale.

Per piccoli valori di n possiamo invece approssimare la distribuzione di A quando H_0 è vera tramite ricampionamento seguendo la procedura che ora descriviamo. Si assegnano in modo casuale le n osservazioni relative ad una variabile (per esempio B) nelle n celle che compongono la griglia, mentre si tengono fisse le osservazioni relative all'altra variabile e poi si calcola la statistica A. L'operazione viene ripetuta un numero grande di volte e la funzione di ripartizione empirica dei valori di A così ottenuti viene considerata come una approssimazione della distribuzione di A quando H_0 è vera.

3.2.2 Indici generalizzati di Huber

Huber et al. (1985) hanno fatto notare che il modo di misurare l'associazione tra due variabili proposto da Tiøstheim non tiene conto della loro osservata distribuzione spaziale. Per ovviare

a questo problema essi propongono di misurare l'associazione spaziale tra due variabili B e T per mezzo di una *statistica prodotto* (si veda anche la Sezione 1.4.3) del tipo

$$\Gamma = \sum_{i,j} c_{ij} d_{ij}$$

dove d_{ij} è una misura della distanza spaziale tra la cella i e la cella j mentre c_{ij} quantifica quanto sono diversi i valori di B osservati nella cella i dai valori di T osservati nella cella j; per esempio,

$$c_{ij} = \frac{1}{2}(|B_i - T_j| + |B_j - T_i|).$$

Sotto opportune ipotesi e per griglie con un gran numero di celle, le statistiche Γ hanno distribuzione Normale quando sia vera l'ipotesi H_0 che prevede l'assenza di associazione spaziale. Ad una approssimazione della distribuzione della statistica Γ si può spesso pervenire anche con l'uso di metodi di ricampionamento, quali il *bootstrap*, o per mezzo di *metodi Monte Carlo*.

Bibliografia

[1] CLIFF, A. D. e ORD. J. K. (1981). *Spatial processes: models and applications.* Pion. London

[2] CLIFF, A. D. e RICHARDSON, S. (1985). Testing the association between two spatial processes. *Statistics and Decisions*, Suppl. 2, 155-160.

[3] CRESSIE, N. (1984). Towards resistant geostatistics. *Geostatistics for Natural Resources Characterization*, ed. G. Verley et. al., 21–24, Reidel, Dordrecth

[4] HAINING, R.P. (1987). Trend surface analysis with regional and local scales of variation with an application to aerial survey data. *Technometrics*, 29, 461–469.

[5] HUBER, L.J., GOLLEDGE, R.G. e COSTANZO, C.M. (1981). Generalized procedures for evaluating spatial autocorrelation. *Geographical Analysis*, 14, 273–278.

[6] HUBER, L.J., GOLLEDGE, R.G, COSTANZO, C.M. e GALE, N. (1985). Measuring association between spatially defined variables: an alternative procedure. *Geographical Analysis*, 17, 36–46.

[7] LEBART, L. (1969). Analyse statistique de la contiguité. *Pubbl. Inst. Univ. Paris*, XVIII, 81–112.

[8] SEN, A. K. (1976). Large sample size distribution of statistics used in testing for spatial autocorrelation. *Geographical Analysis*, 8, 175–184.

[9] TIOSTHEIM, D. (1978). A measure of association for spatial variables. *Biometrika*, 65, 109-114.