Jon Matteo Church matr. 709752
Dalila Vescovi matr. 682063

SAR: autoregressione simultanea

10 luglio 2008

Statistica Applicata
Laurea Magistrale in Ingegneria Matematica

Indice

1

Introduzione

La regressione lineare è forse una delle tecniche più utilizzate in ambito statistico. Quando si trattano osservazioni distribuite spazialmente, se si sceglie di utilizzare un modello di regressione lineare, si rischia di perdere gran parte del potere predittivo del modello stesso poiché si ignora la correlazione spaziale dei dati. In particolare, l'errore di regressione ε è spazialmente autocorrelato, quindi non è verificata l'ipotesi fondamentale del modello di indipendenza tra gli errori sulle varie osservazioni del campione.

Un'alternativa alla regressione lineare è il modello di autoregressione simultanea (SAR). In questo modello per l'errore di regressione viene adottato un sotto modello di autocorrelazione spaziale, che permette di tenere conto del fatto che osservazioni vicine nello spazio sono tra loro correlate. In questo modo SAR risulta essere molto più efficace della regressione lineare non solo nel rappresentare il campione su cui si è costruita l'analisi, ma anche nella previsione su osservazioni future.

Purtroppo questa tecnica necessita la conoscenza di tutte le relazioni tra le varie osservazioni del campione, che è rappresentabile con una matrice delle distanze tra i punti di dimensione $n \times n$, dove n è il numero di osservazioni a disposizione. La stima dei parametri del modello richiede il calcolo del determinante di tale matrice, ovvero la risoluzione di n^3 operazioni, che per un campione numeroso risulta essere un costo computazionale molto elevato.

Fortunatamente, la correlazione spaziale di norma va diminuendo con l'aumentare della distanza spaziale tra i punti. Se si sceglie di fissare un numero limite m di osservazioni influenti per ogni individuo e di annullare l'effetto quindi delle $n - m$ osservazioni più lontane, il numero di relazioni necessarie per la stima dei parametri diminuisce notevolmente (elementi non nulli della matrice delle distanze). Allora diminuisce il costo computazionale e si accelerano i tempi di calcolo.
Anche se il numero di elementi non nulli della matrice delle distanze è sceso a $n \times m$, per campioni numerosi il metodo richiede comunque un costo elevato.

Nel capitolo 2 dopo una breve trattazione dei concetti fondamentali della regressione lineare, viene descritto il modello di autocorrelazione spaziale che verrà

applicato all'errore, e il modello definitivo di autoregressione simultanea (SAR).

Nel capitolo 3 viene riportata l'applicazione del modello ad un caso reale ed il relativo confronto con i risultati della regressione lineare.

Infine nel capitolo 4 sono illustrate le conclusioni ottenute dall'applicazione del modello SAR.

2

Il modello di autoregressione simultanea SAR

Nel Par. 2.1 viene brevemente trattato il modello di regressione lineare classico (OLS) e i risultati ad esso correlati. Nel Par. 2.2 viene descritto il modello di autoregressione spaziale per una variabile risposta i cui valori sono tra loro autocorrelati. Quindi nel Par. 2.3 viene presentato il modello SAR (Simultaneous AutoRegression) per l'errore di regressione e la relativa funzione di massima verosomiglianza. Tale funzione dipende dalla matrice spaziale dei pesi $\mathbf{W}$, la quale viene trattata nel Par. 2.4.

2.1 Regressione Lineare (OLS)

Siano $X_1, X_2, \ldots, X_r$ r variabili (predittori) ritenute correlate con una variabile Y (risposta). Allora il modello di regressione lineare assume la forma

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r + \varepsilon$$

Dove ε è l'errore di regressione ed è una variabile aleatoria indipendente dai predittori e tale che $\mathsf{E}(\varepsilon) = 0$ e $\mathsf{Var}(\varepsilon) = \sigma^2$.

Date n osservazioni indipendenti della variabile risposta e i corrispondenti valori dei predittori, il modello completo diventa

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

dove
$\mathbf{Y} \in \mathbb{R}^n$ vettore delle realizzazioni di Y
$\mathbf{Z} = [\mathbb{I}_{n\times 1}|\mathbf{X_1}|\ldots|\mathbf{X_r}] \in \mathbb{R}^{n\times(r+1)}$ matrice disegno
$\mathbf{X_i} \in \mathbb{R}^n$ vettore delle realizzazioni di X_i
$\boldsymbol{\varepsilon} \in \mathbb{R}^n$ vettore degli errori tale che $\mathsf{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ e $\mathsf{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbb{I}_n$

$\boldsymbol{\beta}$ e σ^2 sono i parametri del modello OLS. Con il metodo dei minimi quadrati si ottiene una stima di $\boldsymbol{\beta}$ minimizzando lo scarto quadratico medio $S(\mathbf{b}) = (\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})$:

$$\hat{\boldsymbol{\beta}} = \mathsf{argmin}_{\mathbf{b}\in\mathbb{R}^{r+1}} S(\mathbf{b})$$

In seguito vengono riportati alcuni risultati noti sulla regressione lineare (si omettono le dimostrazioni).

Teorema 1: *Sia* $\mathbf{Z}$ *di rango pieno* $r+1 \leq n$*, allora:*

$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$
$\hat{\mathbf{Y}} = \mathbf{Z}\hat{\boldsymbol{\beta}}$
$\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \hat{\mathbf{Y}}$

Teorema 2: *Sia* $\mathbf{Z}$ *di rango pieno* $r+1 \leq n$*, allora:*

$\mathsf{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$
$\mathsf{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{Z}'\mathbf{Z})-1$
$\mathsf{E}(\hat{\boldsymbol{\varepsilon}}) = \mathbf{0}$
$\mathsf{Cov}(\hat{\boldsymbol{\varepsilon}}) = \sigma^2(\mathbb{I}_n - \mathbf{H})$ dove $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$
$\mathsf{E}(\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}) = \sigma^2(n-r-1))$

Si definisce indice di determinazione:

$$R^2 = 1 - \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

R^2 è un indice di quanto il modello spieghi adeguatamente i dati.

Corollario 1:

$$S^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{n-(r+1)}$$

è uno stimatore non distorto di σ^2.

Teorema 3: *Sia* $\mathrm{rank}(\mathbf{Z}) = r+1 \leq n$ *ed inoltre* $\boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbb{I}_n)$*, allora:*

$\hat{\boldsymbol{\beta}}$ e $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/n$ sono stimatori di massima verosomiglianza per $\boldsymbol{\beta}$ e σ^2
$\hat{\boldsymbol{\beta}} \sim N_{r+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1})$
$\hat{\boldsymbol{\varepsilon}} \sim N_n(\mathbf{0}, \sigma^2(\mathbb{I}_n - \mathbf{H}))$
$\hat{\boldsymbol{\varepsilon}}$ e $\hat{\boldsymbol{\beta}}$ sono tra loro ortogonali
$\hat{\sigma}^2 \sim \sigma^2\chi(n-r-1)$

Corollario 2:

$$\frac{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{Z}'\mathbf{Z})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})}{S^2} \sim (r+1)F(r+1, n-r-1)$$

Da un punto di vista geometrico, i valori fittati trovati con OLS

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_r X_r$$

giacciono sull'iperpiano di $\mathbb{R}^{r+2}$ generato dalle colonne di $\mathbf{Z}$.

2.2 Autoregressione spaziale

Quando i valori assunti $\mathbf{Y} = (y_1, y_2, \ldots, y_n)'$ da una variabile Y sono in alcuni punti fortemente correlati ai valori che li precedono e seguono (autocorrelazione spaziale),

$$y_t = f(y_{t-1}, y_{t-2}, \ldots, y_{t-p}, e_t), \quad \forall \ p+1 \leq t \leq n$$

si adotta un modello di autoregressione spaziale, che può essere espresso come:

$$y_t = \beta_0 + \sum_{i=1}^{p} \beta_i y_{t-i} + e_t, \quad \forall \ p+1 \leq t \leq n$$

dove:
β_i sono i coeffcienti di autoregressione
$p \in \{1, \ldots, n\}$ è l'ordine di autoregressione
e_t è il t-esimo errore.

Ponendo
$\widetilde{\mathbf{Y}} = (y_{p+1}, y_{p+2}, \ldots, y_p)'$
$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$
$\mathbf{e} = (e_{p+1}, e_{p+2}, \ldots, e_n)'$

$$\mathbf{X} = \begin{pmatrix} 1 & y_p & y_{p-1} & \cdots & y_1 \\ 1 & y_{p+1} & y_p & \cdots & y_2 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & y_{n-1} & y_{n-2} & \cdots & y_{n-p} \end{pmatrix}$$

il modello può essere espresso come per la regressione lineare nella forma

$$\widetilde{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

ove evidentemente la matrice disegno $\mathbf{X}$ ha un significato completamente diverso, poiché in questo caso non ci sono regressori, l'unica variabile è Y, e i valori che assume su ogni osservazione dipendono solo dai valori che la stessa assume sulle altre osservazioni.

Per mettere in evidenza proprio la dipendenza del vettore $\mathbf{Y}$ solo da se stesso, una forma più generale del modello è:

$$\mathbf{Y} = \lambda \mathbf{W}\mathbf{Y} + \mathbf{e}$$

dove $0 \leq \lambda < 1$ è un parametro di autoregressione spaziale e $\mathbf{W} = w_{ij}$ è una matrice di pesi che soddisfa la condizione di non auto-predittibilità:

$$w_{ii} = 0 \quad \forall i \in \{1, \ldots, n\}$$

e di normalizzazione per righe:

$$\sum_{j=1}^{n} w_{ij} = 1 \quad \forall i \in \{1, \dots, n\}$$

2.3 Autoregressione simultanea SAR

Quando si trattano osservazioni distribuite spazialmente, se si sceglie di utilizzare un modello di regressione lineare del tipo

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

si rischia di perdere gran parte del potere predittivo del modello stesso poiché si ignora la correlazione spaziale dei dati.
In particolare in questi casi l'errore di regressione ε è spazialmente autocorrelato, quindi viene violata l'ipotesi fondamentale di indipendenza tra gli errori sulle varie osservazioni.

Si può pensare allora di adottare per ε il modello di autoregressione spaziale:

$$\boldsymbol{\varepsilon} = \lambda \mathbf{W}\boldsymbol{\varepsilon} + \mathbf{e}$$

e di sostituirlo nell'espressione della regressione lineare, si ottiene così il modello di autoregressione simultanea (SAR):

$$\begin{aligned} \mathbf{Y} &= \mathbf{Z}\boldsymbol{\beta} + \lambda \mathbf{W}\boldsymbol{\varepsilon} + \mathbf{e} \\ &= \mathbf{Z}\boldsymbol{\beta} + \lambda \mathbf{W}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}) + \mathbf{e} \end{aligned}$$

In pratica SAR corregge la forma classica di regressione aggiungendo una media pesata delle deviazioni $\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta}$ sulle osservazioni vicine.

* La matrice dei pesi $\mathbf{W}$ gode delle proprietà descritte al Par. 2.2:

$$\begin{aligned} w_{ii} &= 0, \quad \forall i \in \{1, \dots, n\} \\ \sum_{j=1}^{n} w_{ij} &= 1 \quad \forall i \in \{1, \dots, n\} \end{aligned}$$

Un valore non nullo del generico w_{ij} significa che la j-esima osservazione verrà usata per correggere la previsione sull'i-esimo individuo, ciò equivale a ritenere che gli errori sulle 2 osservazioni i e j siano correlati. Poiché si stanno trattando individui spazialmente correlati, i valori non nulli di $\mathbf{W}$ si trovano in corrispondenza di individui tra loro vicini nello spazio.

* λ è il parametro di autoregressione dell'errore e deve essere tale che

$$0 \leq \lambda < 1$$

* Infine deve risultare

$$\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2 \mathbb{I}_n)$$

λ, σ^2 e $\boldsymbol{\beta}$ sono i parametri del modello SAR e vengono stimati massimizzando il logaritmo della funzione di massima verosomiglianza:

$$L(\lambda, \sigma^2, \boldsymbol{\beta}) = \frac{1}{2} \ln |\mathbf{B}| - \frac{1}{2}[n \ln(2\pi\sigma^2) + \sigma^{-2}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})'\mathbf{B}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})]$$

dove $\mathbf{B} = (\mathbb{I}_n - \lambda\mathbf{W})'(\mathbb{I}_n - \lambda\mathbf{W})$.
Affinché la somma degli scarti quadratici dell'errore, $(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})'\mathbf{B}(\mathbf{Y} - \mathbf{Z}\boldsymbol{\beta})$ sia strettamente positiva, è necessario che $\mathbf{B}$ sia definita positiva. Per definizione di $\mathbf{W}$ ed essendo $0 \leq \lambda < 1$, risulta $0 < |\mathbf{B}| \leq 1$, e quindi $-\infty < \ln |\mathbf{B}| \leq 0$.

Supposta l'esistenza degli stimatori di massima verosomiglianza $\check{\lambda}, \check{\sigma}^2$ e $\check{\boldsymbol{\beta}}$, è allora possibile predire $\mathbf{Y}$ tramite

$$\check{\boldsymbol{Y}} = \mathbf{Z}\check{\boldsymbol{\beta}} + \check{\lambda}\mathbf{W}(\mathbf{Y} - \mathbf{Z}\check{\boldsymbol{\beta}})$$

e stimare l'errore

$$\check{\boldsymbol{e}} = \mathbf{Y} - \check{\boldsymbol{Y}}$$

Purtroppo non si riesce ad ottenere alcuna forma analitica a priori dei coefficienti o dei valori fittati, né delle loro distribuzioni, pertanto non è possibile svolgere test statistici o valutare intervalli di confidenza per i parametri stimati (che invece è possibile per la regressione lineare se l'errore ha distribuzione normale). Possiamo solo utilizzare i coefficienti stimati per fare previsione su nuovi campioni.

2.4 Matrice dei pesi W

Per ogni individuo del campione, sono noti i valori della variabile risposta Y, dei predittori $X_1, \ldots, X_r$ e le coordinate spaziali in termini di latitudine e longitudine, tramite le quali è possibile costruire la matrice dei pesi $\mathbf{W}$. Le componenti di $\mathbf{W}$ sono infatti funzione della vicinanza spaziale tra i vari individui.

Sia $d_{ij} = \sqrt{[(\text{latitudine})_i - (\text{latitudine})_j]^2 + [(\text{longitudine})_i - (\text{longitudine})_j]^2}$ la distanza Euclidea tra ogni coppia di individui i e j, sia inoltre $d_{max\,i}$ la distanza tra l'osservazione i e il suo m-esimo punto più vicino. Si pone allora

$$\widetilde{w}_{ij} = 1 \text{ se } d_{ij} \leq d_{max\,i}$$
$$\widetilde{w}_{ij} = 0 \text{ altrimenti.}$$

$\widetilde{\mathbf{W}} = (\widetilde{w}_{ij})$ non è ancora la matrice dei pesi poiché deve essere normalizzata per righe, e deve avere valori nulli sulla diagonale, quindi:

$$w_{ij} = \frac{\widetilde{w}_{ij}}{\sum_{j=1,\, i\neq j}^{n} \widetilde{w}_{ij}}$$
$$w_{ii} = 0 \quad \forall i \in \{1, \ldots, n\}$$

In questo modo per ogni previsione $\check{Y}_i$, le m osservazioni più vicine ad essa (eccetto lei stessa) hanno un peso pari a $\frac{\check{\lambda}}{m}$ nella sua valutazione, mentre le altre non contano, ovvero:

$$\check{Y}_i = \check{\beta}_0 + \check{\beta}_1 x_{i1} + \ldots + \check{\beta}_r x_{ir} + \frac{\check{\lambda}}{m} \sum_{k=1}^{m} (Y_k^* - \check{\beta}_0 - \check{\beta}_1 x_{k1}^* - \ldots - \check{\beta}_r x_{kr}^*)$$

dove * indica che l'individuo k è uno degli m del campione più vicini all'osservazione i in termini di distanza spaziale.

Se si sceglie $m = n$ la matrice $\mathbf{W}$ è piena, e questo significa che esistono n^2 potenziali correlazioni. Poichè per stimare i parametri della SAR è necessario valutare il determinante di tale matrice, sono necessarie n^3 operazioni, quindi il costo computazionale diventa veramente molto alto se il numero di osservazioni è elevato.

Fortunatamente l'autocorrelazione spaziale degli errori di norma va diminuendo all'aumentare della distanza, ecco perchè si sceglie di troncare l'influenza delle osservazioni ai soli m punti più vicini, con m non molto elevato. Scegliere m piccolo è quindi in accordo con il modello teorico, ed inoltre fa si che la matrice dei pesi non sia piena ma abbia solo un numero ridotto di elementi non nulli.

Da un punto di vista numerico è conveniente usare un matrice in forma sparsa, con questa tecnica vengono memorizzati solo gli elementi non nulli della matrice e si evita di calcolare ogni volta un'operazione con un elemento nullo. In questo modo diminuisce lo spazio di memoria necessario al calcolatore e si accelerano i tempi computazionali.

Non entriamo nei dettagli computazionali, ci limitiamo a dire che il modello SAR è implementato nella *spdep* libreria del software R da noi utilizzato per svolgere l'analisi, richiede che vengano fornite le informazioni: valori assunti dalla variabile risposta e dai predittori sulle osservazioni, matrice $\mathbf{W}$ in forma sparsa. Anche per la costruzione di $\mathbf{W}$ si può far riferimento alla stessa libreria, basta fornire una matrice contenente latitudine e longitudine delle varie osservazioni e il valore m di vicini influenti, per maggiori chiarimenti il codice è riportato in Appendice A.

3

Esempio su data set reale

In questo capitolo vedremo l'applicazione del modello di regressione lineare (OLS) e di autoregrassione simultanea (SAR) su un data set reale caratterizzato da distribuzione spaziale. Nel Par. 3.1 vengono illustrati data set e modello, nel Par. 3.2 sono riporati i risultati ottenuti con le 2 diverse analisi e nel Par. 3.3 tali risultati vengono rappresentati su grafici comodi per la visualizzazione del confronto, infine nel Par. 3.4 vengono utilizzati i parametri stimati per ottere previsioni su un secondo data set, del quale sono noti i valori della variabile risposta.

3.1 Data set reale

Il data set a nostra disposizione contiene informazioni relative a 20640 individui, che rappresentano tutti i condomini della California tratti da un censimento del 1990. Per ogni condominio si hanno a disposizione le seguenti osservazioni: numero di occupanti, rendita, età, numero totale di vani, numero totale di camere, numero di proprietari, valore,latitudine e longitudine.
Le variabili di interesse sono:

variabile risposta $Y = \ln(\text{valore di un condominio})$
$X_1 = \text{rendita}$
$X_2 = (\text{rendita})^2$
$X_3 = (\text{rendita})^3$
$X_4 = \ln(\text{età})$
$X_5 = \ln(\text{totale vani / numero occupanti})$
$X_6 = \ln(\text{totale camere / numero occupanti})$
$X_7 = \ln(\text{numero occupanti / numero proprietari})$
$X_8 = \ln(\text{numero proprietari})$

Il modello che si vuole analizzare è quindi:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8$$

Le variabili latitudine e longitudine verranno utilizzate per costruire la matrice dei pesi **W** nel modello SAR.

3.2 Coefficienti stimati per i 2 modelli

Per motivi di costo computazionale, si è scelto di ridurre notevolmente il campione a sole $n = 1000$ aree di interesse.
Si è fissato inoltre a $m = 4$ il numero di elementi non nulli su ogni riga di **W**. In questo modo si ritiene cioè che l'errore relativo ad ogni osservazione dipenda solo da altre 4 osservazioni.

La Tabella 3.1 contiene i valori stimati dei coefficienti $\boldsymbol{\beta}$ per i 2 modelli OLS e SAR, i valori di $S^2 = \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{n-r-1}$ come stimatore di σ^2 e quelli dell'indice di determinazione $R^2 = 1 - \frac{\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}}{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}$.

Tabella 3.1. Stime con OLS e SAR

	OLS	**SAR**
β_0	11.439828159	11.565929281
β_1	0.119867563	−0.035349870
β_2	0.022258423	0.025167169
β_3	−0.001689108	−0.001400044
β_4	0.062130790	−0.010339510
β_5	−0.148146463	0.231575036
β_6	−0.038564672	−0.255060934
β_7	−0.638793568	−0.234159068
β_8	0.083856199	0.049893059
λ		0.71442
σ^2	0.06612056	0.03514641
R^2	0.6448312	0.8112099

Si può notare come il coefficiente di determinazione della SAR sia molto migliore di quello di OLS, il che è significativo di come il modello SAR sia più adeguato a rappresentare il campione.

3.3 Rappresentazione grafica dei risultati

Poiché i regressori sono 8, la rappresentazione completa del campione necessita di uno spazio a 9 dimensioni. Si può pensare allora di proiettare i dati su 8 piani $X_i \times Y$, in modo da vedere come la variabile risposta vari lungo ogni regressore.
Se alle Y_i note si aggiungono i valori fittati $\hat{Y}_i$ per OLS e $\check{Y}_i$ per SAR, si può vedere quanto i valori predetti siano prossimi o meno a quelli noti. In Figura 3.1 sono riportati i valori della variabile risposta (in rosso) e dei valori fittati in corrispondenza delle osservazioni delle variabili X_1 e X_4, i punti verdi corrispondono al

modello OLS, quelli blu al SAR. Anche se è evidente come i valori fittati ottenuti con la SAR siano più prossimi ai valori reali di quelli ottenuti con la regressione lineare, queste rappresentazioni non sono molto efficaci.

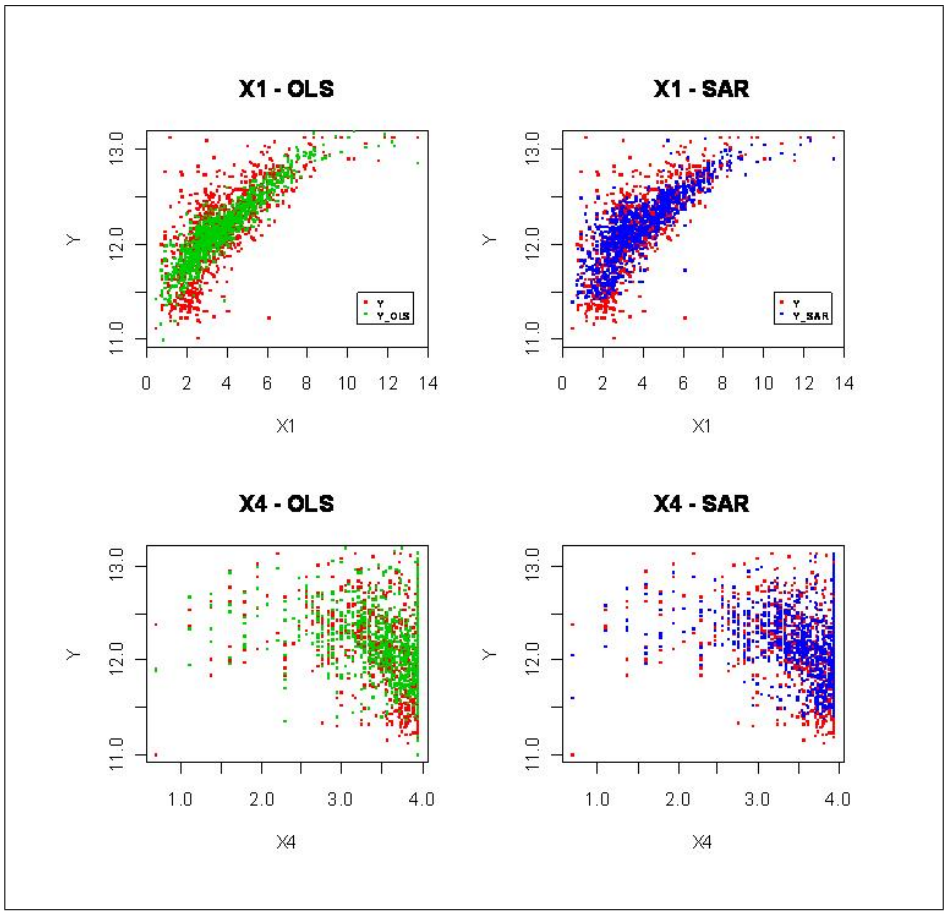

Figura 3.1. Proiezioni su $X_1 \times Y$ e $X_4 \times Y$ di Y, $\hat{Y}$, $\check{Y}$.

Abbiamo già detto al Par. 2.1 che le componenti di $\hat{\mathbf{Y}}$ ottenute con OLS giacciono su l'iperpiano affine di $\mathbb{R}^{r+1}$ di equazione

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_r X_r$$

Sarebbe quindi interessante poter visualizzare nei piani i valori fittati in modo da rappresentarne l'allineamento. A tal fine si può pensare di proiettare i dati sul generico piano $X_i \times Y$, invece che ortogonalmente al piano stesso, parallelamente

all'iperpiano generato dalle colonne di $\mathbf{Z}$.

Consideriamo per semplicità il caso $\mathbb{R}^3$ in modo da poter visualizzare lo spazio. In questo caso allora l'iperpiano OLS ha equazione:

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

ed è generato dai vettori ortogonali $(1, 0, -\hat{\beta}_1), (0, 1, -\hat{\beta}_2)$, quindi se si considera un generico punto (x_1, x_2, y), la sua proiezione su $X_1 \times Y$ parallelamente a questo piano ha per coordinate $(x_1, y - \hat{\beta}_2 x_2)$, mentre quella su $X_2 \times Y$ $(x_2, y - \hat{\beta}_1 x_1)$. La figura 3.2 chiarisce il processo.

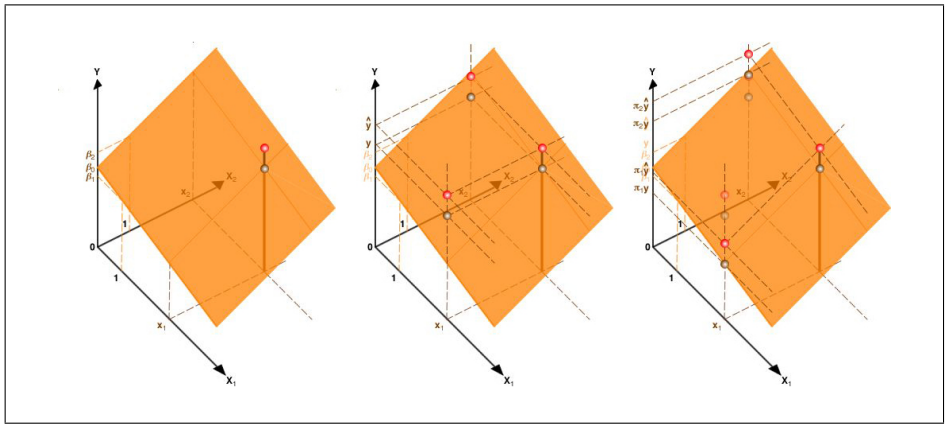

Figura 3.2. Generico punto proiettato ortogonalmente al piani $X_1 \times Y$ e $X_2 \times Y$ e poi parallelamente al piano generato con OLS.

Generalizzando al caso $\mathbb{R}^{r+1}$, la proiezione sul generico piano $X_i \times Y$ parallelamente all'iperpiano OLS $Y = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_r X_r$ del punto

$$(\mathbf{x}, y) = (x_1, x_2, \ldots, x_r, y)$$

ha per coordinate

$$(x_i, y - \mathbf{x} \cdot \hat{\boldsymbol{\beta}}^*)$$

dove $\hat{\boldsymbol{\beta}}^*$ è uguale al vettore $\hat{\boldsymbol{\beta}}$ tranne che per la i-esima coordinata che è nulla.

Proiettiamo in questo modo sia i valori reali assunti dalla variabile risposta che quelli fittati ottenuti dai 2 modelli. Ovviamente le proiezioni dei $\hat{\mathbf{Y}}$ sono allineate lungo una retta. In figura 3.3 e 3.4 sono riportati i grafici delle proiezioni su tutti i piani, in rosso le proiezioni dei dati reali, in verde quelle dei valori ottenuti con OLS e in blu quelli di SAR.

Questi grafici mettono in evidenza come i valori fittati stimati da SAR sono molto più simili a quelli reali che non quelli trovati da OLS.

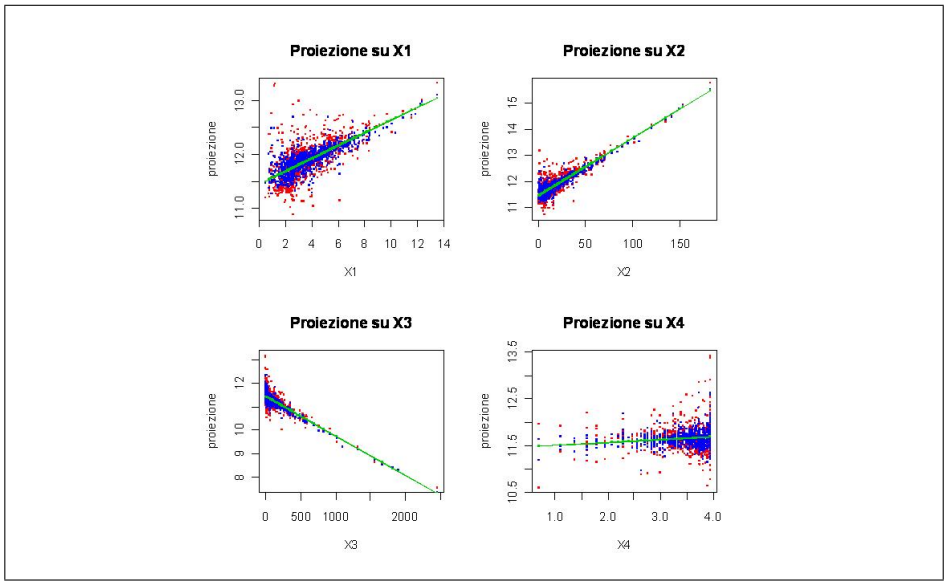

Figura 3.3. Proiezione dei dati reali e fittati con i 2 modelli parallelamente all'iperpiano OLS

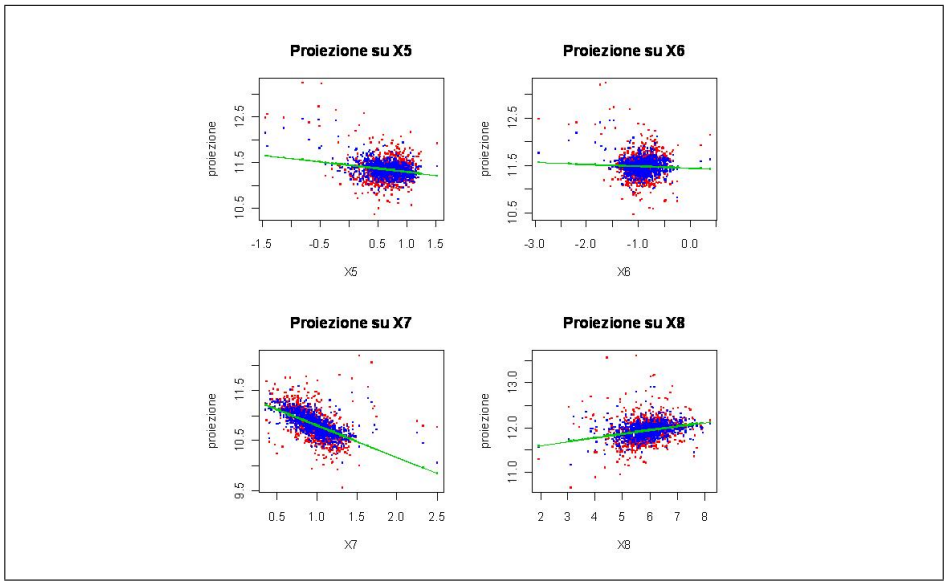

Figura 3.4. Proiezione dei dati reali e fittati con i 2 modelli parallelamente all'iperpiano OLS

3.4 Previsioni

Inizialmente si è specificato che si dispone in realtà di molte più osservazioni di quante utilizzate per le analisi. Si è scelto pertanto di utilizzare 100 delle osservazioni restanti per svolgere delle analisi di previsione e poi confrontare i valori fittati ottenuti con quelli reali.
Sia $\mathbf{Z}^0 = [\mathbb{I}_{n_0} | \mathbf{X}_1^0 | \dots | \mathbf{X}_r^0]$ il nuovo campione di osservazioni sui regressori composto da $n_0 = 100$ individui.

I valori fittati della variabile risposta ottenuti con OLS sono dunque:

$$\hat{\mathbf{Y}}^0 = \hat{\boldsymbol{\beta}}\mathbf{Z}^0$$

Per quanto riguarda la SAR invece:

$$\check{\mathbf{Y}}^0 = \check{\boldsymbol{\beta}}\mathbf{Z}^0 + \check{\lambda}\mathbf{W}(\mathbf{Y}^0 - \check{\boldsymbol{\beta}}\mathbf{Z}^0)$$

Ora però i valori $\mathbf{Y}^0$ non sono noti, ma sono proprio le quantità da stimare. L'unica stima che si può ottenere allora è:

$$\check{\mathbf{Y}}^0 = \check{\boldsymbol{\beta}}\mathbf{Z}^0$$

Ma come prevedibile i valori fittati così ottenuti sono molto distanti da quelli reali, e quindi le previsioni ottenute con il metodo SAR sono abbastanza deludenti, sicuramente peggiori di quelle ottenute con OLS.

Si può ovviare però a questo problema pensando di utilizzare l'eventuale correlazione tra le osservazioni del nuovo campione di dati $\mathbf{Z}^0$ e quelle del campione originale $\mathbf{Z}$ sulle quali è stata svolta la regressione.

In effetti abbiamo visto al Par. 2.4 che per ogni osservazione i-esima, solo m delle altre osservazioni influenzano la stima di $\check{Y}_i$, e di certo non se stessa per la proprietà di non autopredittibilità della matrice $\mathbf{W}$. Possiamo supporre allora che le m osservazioni da cui dipende il generico individuo del nuovo campione non appartengano a tale nuovo campione, ma al campione iniziale, del quale sono noti i valori reali della variabile risposta.

Il nuovo modello di previsione può allora essere espresso come:

$$\check{\mathbf{Y}}^0 = \check{\boldsymbol{\beta}}\mathbf{Z}^0 + \check{\lambda}\mathbf{W}^0(\mathbf{Y} - \check{\boldsymbol{\beta}}\mathbf{Z})$$

$\mathbf{W}^0 \in \mathbb{R}^{n_0 \times n_0}$ è la nuova matrice dei pesi, ed è tale che $w_{ij}^0 \neq 0$ se il j-esimo individuo di $\mathbf{Z}$ è uno degli m più vicini all'i-esimo individuo di $\mathbf{Z}^0$.

Questo modello di previsione ha senso se gli individui del nuovo campione sono correlati con quelli del campione iniziale, ovvero se ci sono almeno m punti di $\mathbf{Z}$ che sono abbastanza vicini in termini di distanza spaziale da i punti di $\mathbf{Z}^0$.

Disponendo dei veri valori della variabile risposta, siamo in grado di confrontare questi con quelli previsti ottenuti con i 2 modelli. I risultati ottenuti sono rappresentati in figura 3.5, con la solita convenzione rosso per i valori reali, verde per OLS e blu per SAR.

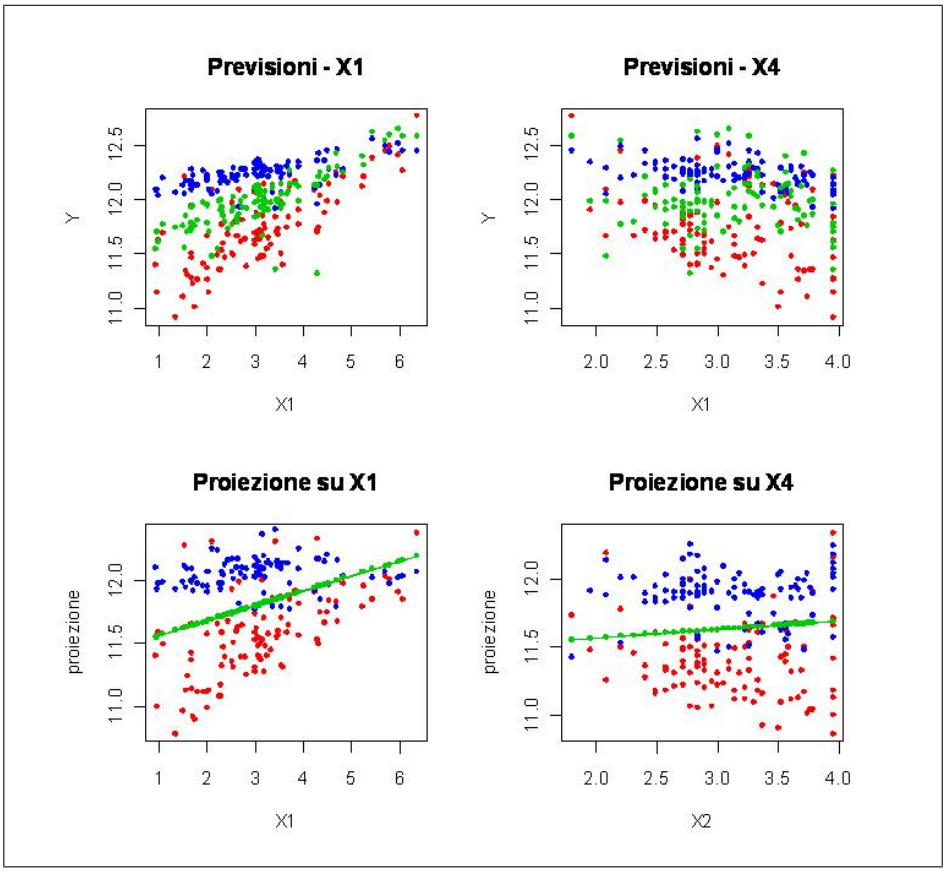

Figura 3.5. Proiezione delle previsioni su $X_1 \times Y$ e $X_4 \times Y$ ortogonalmente ai piani e parallelamente all'iperpiano OLS

Dai grafici sembra che le previsioni ottenute con SAR non siano buone.

A questo punto andiamo a vedere se effettivamente le correlazioni che sono state supposte tra il nuovo campione e quello iniziale esistono, ovvero se i 2 campioni sono vicini in termini di coordinate spaziali. In Figura 3.6 sono raffigurate le disposizioni spaziali del campione iniziale (in nero) e del secondo campione (in arancio).

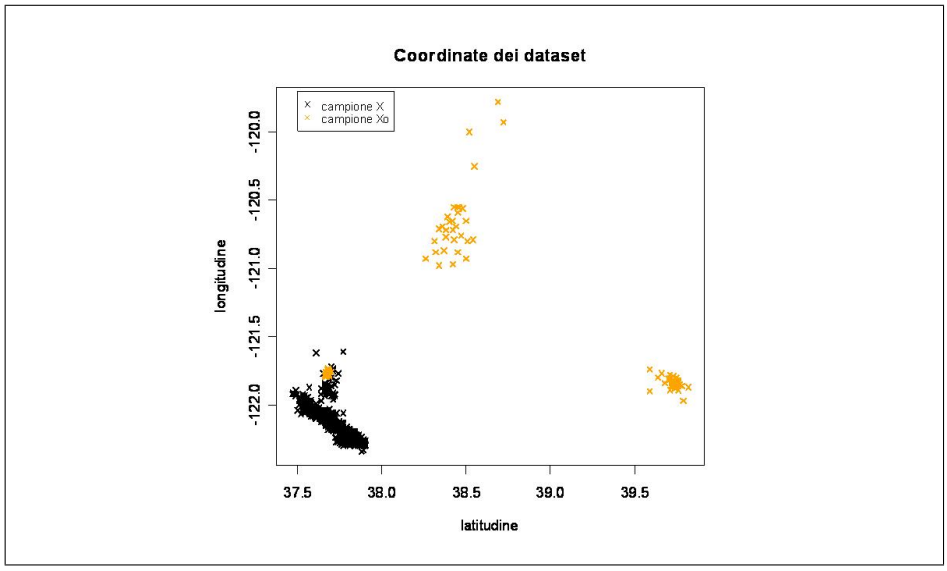

Figura 3.6. Distribuzione spaziale dei 2 campioni

È evidente che i punti del secondo campione sono quasi tutti molto distanti dal primo, il che spiega perchè le previsioni ottenute siano così scadenti. Ci sono però 22 punti del nuovo campione che cadono in prossimità del primo (Figura 3.7).

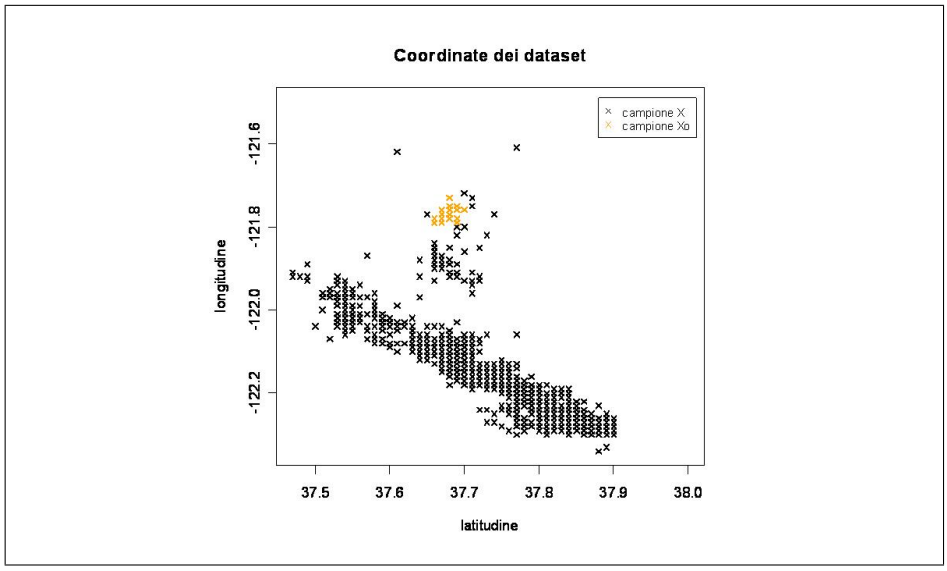

Figura 3.7. Distribuzione spaziale dei 2 campioni

Proviamo a vedere se per queste 22 osservazioni la SAR fallisce o fornisce delle stime migliori di OLS. In figura 3.8 sono riportati i valori stimati di Y con i 2 modelli e quelli reali, sui vari piani descritti precedentemente.

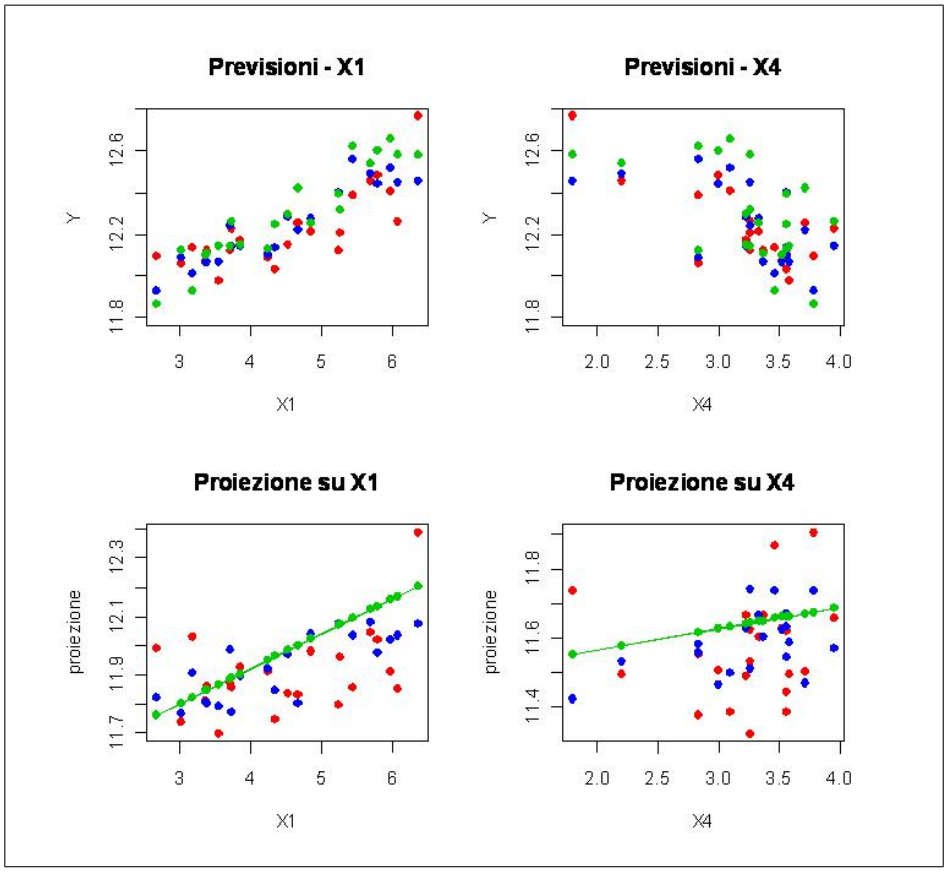

Figura 3.8. Proiezione delle previsioni ridotte su $X_1 \times Y$ e $X_4 \times Y$ ortogonalmente ai piani e parallelamente all'iperpiano OLS

Per questi 22 individui la SAR ottiene risultati di previsione molto soddisfacenti, migliori di OLS.

Tutto ciò è in perfetto accordo con il modello che prevede che ci sia correlazione solamente tra osservazioni vicine nello spazio.

4

Conclusioni

Una corretta analisi e un'accurata previsione in un campione caratterizzato da distribuzione spaziale, devono essere in grado di sfruttare le informazioni provenienti dagli errori delle osservazioni vicine.
Il metodo SAR è in grado di cogliere le correlazioni spaziali tra le osservazioni, almeno tra le più vicine, e questo lo rende molto più adeguato nella rappresentazione del campione, infatti abbiamo visto nel capitolo 3 quanto l'indice di determinazione R^2 sia notevolmente migliore di quello ottenuto con una regressione lineare classica.
Inoltre se si vuole svolgere un'analisi di previsione su individui che sono spazialmente prossimi a quelli del campione di riferimento, abbiamo visto come SAR ottenga risultati molto buoni sfruttando le informazioni sulla variabile risposta degli individui del campione iniziale.

Tuttavia per poter implementare l'analisi in tempi modesti, abbiamo dovuto scegliere di lavorare con un campione di dimensioni relativamente ridotte, a causa dell'elevato costo computazionale del metodo.

Appendice

CODICE **SAR**:

```
#####################################
## PROGETTO DI STATISTICA APPLICATA ##
 ########### A.A. 2007-2008###########
## Dalila Vescovi Jon Matteo Church ##
#####################################

## SAR: Spatial AutoRegresion ##

## data set:

setwd('C:/Documents and Settings/Dalila
Vescovi/Desktop/progetto_1-7-08/DALILA')

dati.comp <- read.table('dataset.txt', header = T) dati <-
dati.comp[1:1000,]

attach(dati)

Y <- log(median_house_value)
X1 <- median_income
X2 <- median_income^2
X3 <- median_income^3
X4 <- log(housing_median_age)
X5 <- log(total_rooms/population)
X6 <- log(total_bedrooms/population)
X7 <- log(population/households)
X8 <- log(households)
lat <- latitude
long <- longitude
detach(dati)
```

```
n <- length(Y)
p<-8
X <- cbind(X1,X2,X3,X4,X5,X6,X7,X8)
data <-data.frame(Y,X1,X2,X3,X4,X5,X6,X7,X8)
Z <-cbind(rep(1,n),X1,X2,X3,X4,X5,X6,X7,X8)
########################################################## ## Matrice
dei pesi W ## distanze con latitudine e longitudine:

library(spdep)

d <- knearneigh(cbind(lat,long), k=4)
## d matrice n x 4,
## la riga i_esima contiene le etichette dei 4 punti pi vicino al
## punto i sulla base della distanza euclidea calcolata tramite lat
## e long.

W. <- knn2nb(d)

W <- nb2listw(W.,  style="W")

#####################################
## simultaneous autoregression(SAR)
## Y = Z b + L*W (Y-Z b) + e

sar<-errorsarlm(Y ~ X1+X2+X3+X4+X5+X6+X7+X8 ,data,W)

summary(sar)
L <- sar$lambda
bh <- sar$coefficients
eh <- sar$residuals
yh<-sar$fitted
SSE <- t(eh)%*%eh
SST <- sum((Y-mean(Y))^2)
R2 <- 1-SSE/SST
S2 <- SSE/(n-r-1)

## matrice dei pesi espressa (non sparsa)
D <- matrix(0,n,n)
for(k in 1:n){ for(j in 1:4){
    D[k,d$nn[k,j] ]<-0.25}}
## W e D sono la stessa cosa...

##########################################
## regressione lineare (OLS):
## Y = Z B + EPS

OLS <- lm(Y ~ X1+X2+X3+X4+X5+X6+X7+X8)
```

```
summary(OLS)

B1 <- OLS$coefficients
e1 <- OLS$residuals
estd1 <- rstandard(OLS)
Yh1 <- OLS$fitted
hii1 <- hatvalues(OLS)


SSe1 <- t(e1)%*%e1
R2linreg <- 1-SSe1/SST
 S2.1<-SSe1/(n-r-1)

shapiro.test( estd1 )

############################################################

windows()
 par(mfrow=c(2,2))
 plot(X1, Y, col=2, pch=16, main='X1 -OLS',cex=0.5)
 points(X1, Yh1, col=3, pch=16,cex=0.3)
 legend(10.5,11.5,c('Y','Y_OLS'),col=c(2,3),pch=16,cex=0.5)
 plot(X1,Y, col=2, pch=16, main='X1 - SAR',cex=0.5)
 points(X1, yh, col=4,pch=16,cex=0.3)
 legend(10.5,11.5,c('Y','Y_SAR'),col=c(2,4),pch=16,cex=0.5)
 plot(X2,Y, col=2, pch=16, main='X2 - OLS',cex=0.5)
 points(X2, Yh1, col=3,pch=16,cex=0.3)
 legend(140,11.5,c('Y','Y_OLS'),col=c(2,3),pch=16,cex=0.5)
 plot(X2,Y, col=2, pch=16, main='X2 - SAR',cex=0.5)
 points(X2, yh, col=4,pch=16,cex=0.3)
 legend(140,11.5,c('Y','Y_SAR'),col=c(2,4),pch=16,cex=0.5)

#########################################################

 ## PREVISIONE ##

## Xo = matrice (8,q) le cui righe sono i valori dei
## regressori sui nuovi individui sui quali si vuole effettiare le
## previsione. Per i nuovi q individui sono noti i valori assunti
## dai regressoni (Xo) e anche latitudine e longitudine.

## In base al modello:
## fitted:   yh.o = Zo bh + L*D.o(yo - Zo bh)
## Zo_i = (1,X1_oi,X2_oi,...,X8_oi)
## D.o = matrice dei pesi per i
## nuovi dati yo (valori della variabile risposta in corrispondenza
```

```
## dei nuovi individui)
## yo non  noto ( proprio quello che si
## deve stimare) => l'unica stima dei fitted che si pu ottenere con
## le informazioni  del nuovo dataset:
## yh.o = Zo bh

## IDEA: se i nuovi dati sono 'vicini' ai dati del primo data set:
## fitted:   yh.o = Zo bh + L*D.o(Y - Z bh)
## D.o = matrice dei pesi per il nuovo data set (n,q)
## la riga i_esima di D.o contiene i 4 individui del data set
## originale pi vicini all'i_esimo individuo del nuovo dataset
## in questo modo i nuovi dati sono supposti dipendere dai 4
##individui pi vicini non dello stesso data set (per cui le y non
## sono note ma vanno stimate) ma del data set originale. Questo ha
## senso se gli individui di Xo sono abbastanza vicini ad almeno 4
## individui di X.

## uso q=100 dati per fare la previsione

dati.new <- dati.comp[1001:1100,]
attach(dati.new)

Yo <- log(median_house_value)

X1o <- median_income
X2o <- median_income^2
X3o <- median_income^3
X4o <- log(housing_median_age)
X5o <- log(total_rooms/population)
X6o <- log(total_bedrooms/population)
X7o <-log(population/households)
X8o <- log(households)
lato <- latitude
longo <- longitude
detach(dati.new)
q <- length(Yo)

Zo <- cbind(rep(1,q),X1o,X2o,X3o,X4o,X5o,X6o,X7o,X8o)
dati.prev <- data.frame(Zo[,2:9])
Xo <- as.matrix(dati.prev)
var <- c('X1','X2','X3','X4','X5','X6','X7','X8')
dimnames(dati.prev)[[2]]<-var

## creiamo la matrice dei pesi W per il nuovo data set Xo

## 1) valutazione per ogni individui i di Xo dei 4 individui del
## dataset iniziale X pi vicini a i (distanza euclidea valutata
```

```
## sulle coord. lat o long.)

g <- matrix(0,q,4) for(i in 1:q)
    g[i,] <- knearneigh(cbind(c(lat,lato[i]),c(long,longo[i])), k=4)$nn[1001,]

## 2) D.o: matrice dei pesi  (q, n)
## per ogni riga i, D[i,j] = 0.25 se j  il valore corrispondente ad
## uno dei 4 'vicini' di i, altrimenti D[i,j]=0.

Do <- matrix(0,q,n)

for(k in 1:q){ for(j in 1:4){
    Do[ k, g[k,j] ] <-0.25}}

## trend = Zo bh   ( quello che si ottiene anche con predict.sarlm)
trend <- Zo%*%bh

trend predict.sarlm(sar,newdata=dati.prev,W)

## signal = L*D.o(Y - Z bh)
signal <- L*Do%*%(Y-Z%*%bh)

## yh.o = trend + signal
yh.o <- trend+signal

## previsione con Regressione Lineare:
yh1.o <-Zo%*%B1

## distanze tra gli yo (noti) e quelli stimati da i 2 modelli
delta.sar <- Yo-yh.o
delta.ols <- Yo - yh1.o

oss <- c('DATI NOTI','SAR','OLS','delta SAR','delta OLS.')
LL <-cbind(Yo,yh.o,yh1.o,delta.sar,delta.ols)
dimnames(LL)[[2]]<-oss

windows()
par(mfrow=c(2,2))

plot(X1o,Yo,col=2,cex=0.7,ylim=c(min(Yo,yh.o,yh1.o),max(Yo,yh.o,yh1.o))
,pch=16,main='Previsioni- X1',xlab='X1',ylab='Y' )
points(X1o,yh.o,col=4,pch=16,cex=0.7)
points(X1o,yh1.o,col=3,pch=16,cex=0.7)

plot(X4o,Yo,col=2,ylim=c(min(Yo,yh.o,yh1.o),max(Yo,yh.o,yh1.o)),pch=16,
main='Previsioni- X4',xlab='X1',ylab='Y' ,cex=0.7)
points(X4o,yh.o,col=4,pch=16,cex=0.7)
```

```
points(X4o,yh1.o,col=3,pch=16,cex=0.7)


alpha <- B1[2:9]
alpha[1] <- 0
plot(X1o,Yo-Xo%*%alpha,type='p',pch=16,cex=0.7,col=2,
main=paste('Proiezione su X1'),ylab="proiezione",xlab='X1')
points(X1o,yh.o-Xo%*%alpha,pch=16,col=4,cex=0.7)
points(X1o, yh1.o-Xo%*%alpha,col=3,type='l')
points(X1o, yh1.o-Xo%*%alpha,col=3,pch=16,cex=0.7)

alpha <- B1[2:9] alpha[4] <- 0
plot(X4o,Yo-Xo%*%alpha,type='p',pch=16,cex=0.7,col=2,
main=paste('Proiezione su X4'),ylab="proiezione",xlab='X2')
points(X4o,yh.o-Xo%*%alpha,pch=16,col=4,cex=0.7)
points(X4o, yh1.o-Xo%*%alpha,col=3,type='l')
points(X4o, yh1.o-Xo%*%alpha,col=3,pch=16,cex=0.7)


## dal grafico sembra che siano migliori le previsioni fatte con
## la regressione lineare che non con la sar

## ma se analiziamo le distanze tra il data set iniziale X e Xo

## vediamo in effetti quanto i punti di X e Xo siano tra loro vicini

windows()
 plot(lat,long,pch=4,cex=0.8,lwd=2, col=1,main='Coordinate dei dataset',
 xlim=c(min(c(lat,lato)),max(c(lat,lato))), ylim=c(min(c(long,longo)),
 max(c(long,longo))),xlab='latitudine',ylab='longitudine'   )
points(lato,longo,pch=4,col='orange',lwd=2,cex=0.8)
legend(37.5,-119.7,col=c('black','orange'),c('campione X','campione
Xo'),pch=4,cex=0.8)


## vediamo che in effetti che i dati di Xo sono molto distanti da X
##  quindi in accordo con il modello che la sar dia delle cattive
## previsioni

## proviamo a selezionare gli unici punti di Xo che sono vicini ad
## almeno 4 di X lato < 38.0 sono i primi 22 dati di Xo

Xo2<-Xo[1:22,]
t <- 22
lato2 <- lato[1:t]
longo2 <- longo[1:t]
points(lato2,longo2,pch=16,col=6)
```

```
windows()

plot(lat,long,pch=4,cex=0.8,lwd=2, col=1,main='Coordinate dei
dataset',xlim=c(min(c(lat,lato)),38),ylim=c(min(c(long,longo)),-121.5),
xlab='latitudine',ylab='longitudine'   )
points(lato,longo,pch=4,col='orange',lwd=2,cex=0.8)
legend(37.88,-121.49,col=c('black','orange'),c('campione
X','campione Xo'),pch=4,cex=0.8)

## vediamo quanto i valori di y predetti dalla sar sono vicini a
## quelli veri

LL[1:t,]

windows()
par(mfrow=c(2,2))
plot(X1o[1:t],Yo[1:t],col=2,cex=1.2,ylim=c(11.8,max(Yo,yh.o,yh1.o)),
pch=16,main='Previsioni- X1',xlab='X1',ylab='Y' )
points(X1o[1:t],yh.o[1:t], col=4,pch=16,cex=1.2)
points(X1o[1:t],yh1.o[1:t], col=3,pch=16,cex=1.2)

plot(X4o[1:t],Yo[1:t],col=2,cex=1.2,ylim=c(11.8,max(Yo,yh.o,yh1.o)),
pch=16,main='Previsioni - X4',xlab='X4',ylab='Y' )
points(X4o[1:t],yh.o[1:t], col=4,pch=16,cex=1.2)
points(X4o[1:t],yh1.o[1:t], col=3,pch=16,cex=1.2)

alpha <- B1[2:9] alpha[1] <- 0
plot(X1o[1:t],Yo[1:t]-Xo2%*%alpha,type='p',pch=16,cex=1.2,col=2,
main=paste('Proiezione su X1'),ylab="proiezione",xlab='X1')
points(X1o[1:t],yh.o[1:t]-Xo2%*%alpha,pch=16,col=4,cex=1.2)
points(X1o[1:t], yh1.o[1:t]-Xo2%*%alpha,col=3,type='l')
points(X1o[1:t], yh1.o[1:t]-Xo2%*%alpha,col=3,pch=16,cex=1.2)

alpha <- B1[2:9] alpha[4] <- 0
plot(X4o[1:t],Yo[1:t]-Xo2%*%alpha,type='p',pch=16,cex=1.2,col=2,
main=paste('Proiezione su X4'),ylab="proiezione",xlab='X4')
points(X4o[1:t],yh.o[1:t]-Xo2%*%alpha,pch=16,col=4,cex=1.2)
points(X4o[1:t], yh1.o[1:t]-Xo2%*%alpha,col=3,type='l')
points(X4o[1:t], yh1.o[1:t]-Xo2%*%alpha,col=3,pch=16,cex=1.2)


#####################################################
#####################################################

## altri grafici:
```

```
## rosso = dati reali
## verde = regressione lineare
## blue = sar

## regressione lineare OLS
## 1)
## projection of fitted and response
vs regressors plot

windows()
layout(matrix(1:8,2,4,byrow=T))

for(i in 1:p){
    alpha <- B1[2:9]
    alpha[i] <- 0
    plot(X[,i],Y-X%*%alpha,type='p',pch=16,col=3,main=
    paste('Proiezione su',names(data)[i+1]),ylab="proiezione",
    xlab=names(data)[i+1])
    points(X[,i],Yh1-X%*%alpha, type='l',pch=16,col=2)
}

## 2)
## studentized residuals vs fitted plot
windows()
plot(Yh1,estd1,type='p',pch=16,col=3,main='Residui studentizzati vs
Fitted',ylab="residui studentizzati",xlab="fitted")

## 3)
## studentized residuals vs regressors plot
windows()
layout(matrix(1:8,2,4,byrow=T)) for(i in 1:p){
    plot(X[,i],estd1,type='p',pch=16,col=3,main=paste('Residui studentizzati
    vs',names(data)[i+1]),ylab="residui studentizzati",xlab=names(data)[i+1])
}

## 4)
## leverage vs fitted plot
avarage.lavarage <- (p+1)/n
windows()
plot(Yh1,hii1,type='p',pch=16,col=3,main='Leverage vs
Fitted',ylab="leverage",xlab="fitted")
abline(avarage.lavarage,0,col=2)

## 5)
## studentized residuals qqplot
windows()
qqnorm(estd1, main='Residui studentizzati QQplot',col=3, pch=16)
```

```
qqline(estd1,col=2)

## SAR

windows()
par(mfrow=c(2,2))
for(i in 1:4){
    alpha <- B1[2:9]
    alpha[i] <- 0
    plot(X[,i],Y-X%*%alpha,type='p',pch=16,cex=0.5,col=2,main=
    paste('Proiezione su',names(data)[i+1]),ylab="proiezione",
    xlab=names(data)[i+1])
    points(X[,i],yh-X%*%alpha,pch=16,col=4,cex=0.3)
    points(X[,i], Yh1-X%*%alpha,col=3,type='l')
      points(X[,i], Yh1-X%*%alpha,col=3,pch=16,cex=0.3)
} windows() par(mfrow=c(2,2)) for(i in 5:p){
    alpha <- B1[2:9]
    alpha[i] <- 0
    plot(X[,i],Y-X%*%alpha,type='p',pch=16,cex=0.5,col=2,main=
    paste('Proiezione su',names(data)[i+1]),ylab="proiezione",
    xlab=names(data)[i+1])
    points(X[,i],yh-X%*%alpha,pch=16,col=4,cex=0.3)
    points(X[,i], Yh1-X%*%alpha,col=3,type='l')
    points(X[,i], Yh1-X%*%alpha,col=3,pch=16,cex=0.3)
}



## PREVISIONI

windows()
layout(matrix(1:8,2,4,byrow=T))
for(i in 1:p){
    alpha <- B1[2:9]
    alpha[i] <- 0
    plot(X[,i], Yh1-X%*%alpha,col=3,type='l',main=
    paste('Proiezione su',names(data)[i+1]),ylab="proiezione",
    xlab=names(data)[i+1])
    points(Xo2[,i],Yo[1:22]-Xo2%*%alpha,col=2,pch=16)
    points(Xo2[,i],yh.o[1:22]-Xo2%*%alpha,col=4,pch=16)
    points(Xo2[,i],yh1.o[1:22]-Xo2%*%alpha,col=3,pch=16)
}

windows()
layout(matrix(1:8,2,4,byrow=T))
for(i in 1:p){
    alpha <- B1[2:9]
```

```
    alpha[i] <- 0
    plot(X[,i], Yh1-X%*%alpha,col=3,type='l',main=
    paste('Proiezione su',names(data)[i+1]),ylab="proiezione",
    xlab=names(data)[i+1])
    points(Xo[,i],Yo-Xo%*%alpha,col=2,pch=16)
    points(Xo[,i],yh.o-Xo%*%alpha,col=4,pch=16)
    points(Xo[,i],yh1.o-Xo%*%alpha,col=3,pch=16)
}
```

Riferimenti bibliografici

[1] Pace R. K. Barry R. *Sparse spatial autoregressions.* Statistics & Probability, Letters 33, 1997, 291-197.

[2] http://www.spatialanalysisonline.com/output/html/Spatialautoregressivemodelling.html.

[3] Troy A. *Spatial autoregressive methods.* University of Vermont, pubblicato sul sito: http://www.uvm.edu/envnr/gradgis/advanced.

[4] http://portal.wsiz.rzeszuw.pl/plik.aspx?id=3644.