

Autoregressione simultanea (SAR)

Jon Matteo Church - 709752
Dalila Vescovi - 682063

11 luglio 2008

Modelli di regressione

- Regressione lineare (OLS)
- Autoregressione
- Autoregressione simultanea (SAR)

Applicazione su dataset reale

- Stima dei parametri
- Previsione

Conclusioni

Regressione lineare (OLS)

Regressione Lineare (OLS)

Siano X_1, X_2, \dots, X_r r variabili (predittori) ritenute correlate con una variabile Y (risposta) allora il modello di regressione lineare per una singola risposta assume la forma

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_r X_r + \epsilon$$

Date n osservazioni indipendenti

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

dove

$$\begin{aligned}\mathbf{Z} &= (\mathbb{I}_{n \times 1}, \mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_r) \\ \mathbf{E}(\boldsymbol{\epsilon}) &= \mathbf{0} \\ \text{Cov}(\boldsymbol{\epsilon}) &= \sigma^2 \mathbb{I}_n\end{aligned}$$

Stime ai minimi quadrati

Con il metodo dei minimi quadrati si stima β minimizzando lo scarto quadratico $S(\mathbf{b}) = (\mathbf{Y} - \mathbf{Z}\mathbf{b})^T(\mathbf{Y} - \mathbf{Z}\mathbf{b})$

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

Proposizione

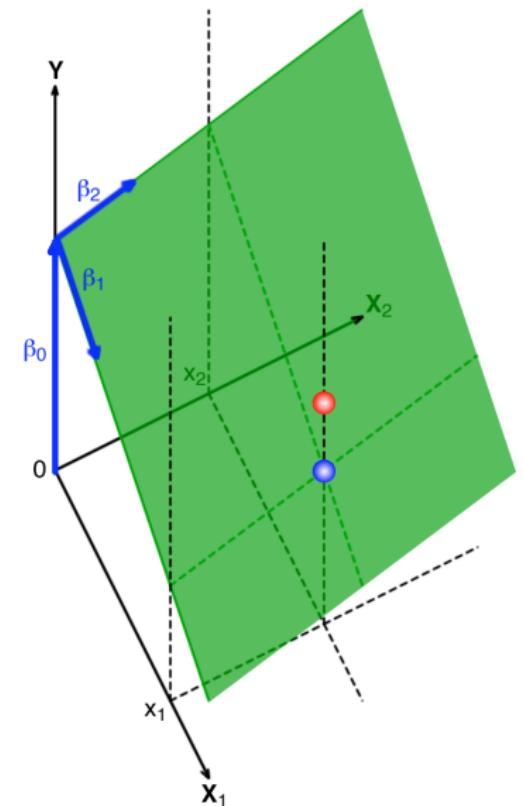
Sia Z di rango pieno $r + 1 \leq n$ allora:

$$\begin{aligned}\hat{\beta} &= (Z^T Z)^{-1} Z^T \mathbf{Y} \\ \hat{\mathbf{Y}} &= Z \hat{\beta} \\ \hat{\epsilon} &= \mathbf{Y} - \hat{\mathbf{Y}}\end{aligned}$$

Proiezione dei dati

OLS individua un iperpiano affine
dello spazio \mathbb{R}^{r+1}

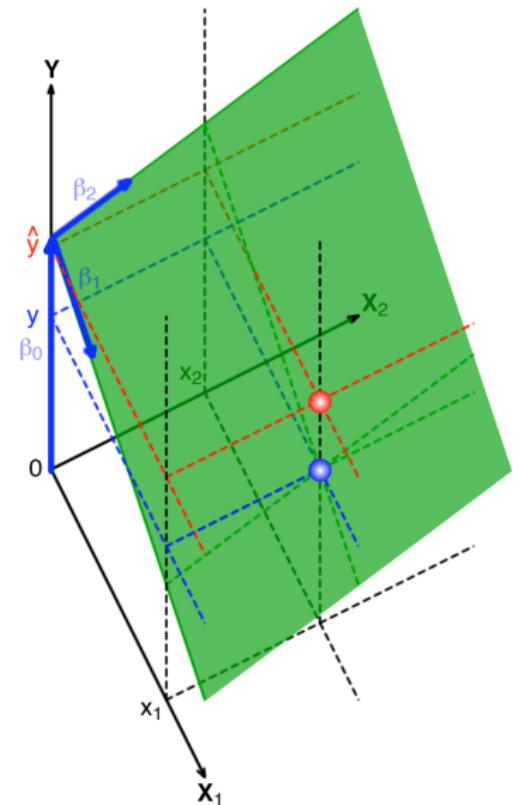
$$OLS = \hat{\beta}_0 + \text{span}\{\hat{\beta}_i\}_{i=1}^r$$



Proiezione dei dati

Proiezione ortogonale sul generico piano X_i, Y

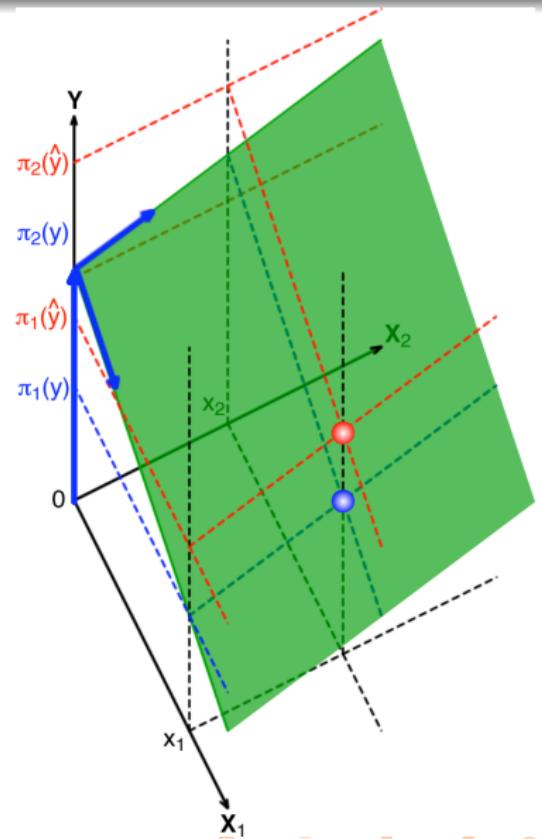
$$\begin{aligned} l_i : \mathbb{R}^{r+1} &\rightarrow \mathbb{R}^2 \\ (\mathbf{x}, y)^T &\mapsto (x_i, y)^T \end{aligned}$$



Proiezione dei dati

Proiezione lungo la direzione parallela
all'iperpiano OLS

$$\begin{aligned}\pi_i : \mathbb{R}^{r+1} &\rightarrow \mathbb{R}^2 \\ (\mathbf{x}, y)^T &\mapsto (x_i, y - (\hat{\beta} \cdot \mathbf{x} - \hat{\beta}_i x_i))^T\end{aligned}$$



Autoregressione

Quando i valori assunti $\mathbf{Y} = (y_1, y_2, \dots, y_n)^\top$ da una variabile Y sono in alcuni punti **fortemente correlati** ai valori che li precedono e seguono (autocorrelazione), si adotta un modello di autoregressione

Autoregressione

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

la cui versione lineare è

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

dove:

- β_i sono i coefficienti di autoregressione
- $p \in \{1, \dots, n\}$ è l'ordine di autoregressione
- e_t è il t -esimo errore

Autoregressione

Se poniamo

$$\begin{aligned}\tilde{\mathbf{Y}} &= (y_{p+1}, y_{p+2}, \dots, y_n) \\ \boldsymbol{\beta} &= (\beta_0, \beta_1, \dots, \beta_p)^T \\ \mathbf{e} &= (e_{p+1}, e_{p+2}, \dots, e_n)^T \\ \mathbf{X} &= \begin{pmatrix} 1 & y_p & y_{p-1} & \cdots & y_1 \\ 1 & y_{p+1} & y_p & \cdots & y_2 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & y_{n-1} & y_{n-2} & \cdots & y_{n-p} \end{pmatrix}\end{aligned}$$

il modello di autoregressione lineare si scrive

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

Autoregressione

Un modello più generale di autoregressione è

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

dove

$$\begin{aligned}0 &\leq \lambda < 1 \\ \mathbf{e} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbb{I}_n)\end{aligned}$$

e $W = (w_{ij})_{i,j=1}^n$ è una matrice di pesi che soddisfa le condizioni di non auto-predittività

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

e di normalizzazione per righe

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

Autoregressione simultanea (SAR)

Quando si trattano **osservazioni distribuite spazialmente**, se si sceglie di utilizzare un modello di regressione lineare si rischia di perdere potere predittivo del modello stesso poiché si ignora la correlazione spaziale dei dati.

In particolare nelle osservazioni con distribuzione spaziale, l'errore di regressione ϵ è spazialmente autocorrelato.

Per ϵ addottiamo pertanto il modello di autoregressione

$$\epsilon = \lambda W\epsilon + e$$

ottenendo così il modello

Autoregressione simultanea (SAR)

$$Y = Z\beta + \lambda W(Y - Z\beta) + e$$

Stima dei parametri

λ , β e σ^2 vengono stimati massimizzando il logaritmo della funzione di verosimiglianza della SAR

$$L(\lambda, \beta, \sigma^2) = \frac{1}{2} \ln |(\mathbb{I}_n - \lambda W)^T (\mathbb{I}_n - \lambda W)| - \frac{1}{2}(n \ln(2\pi\sigma^2) + \sigma^{-2}(\mathbf{Y} - Z\beta)^T (\mathbb{I}_n - \lambda W)^T (\mathbb{I}_n - \lambda W)(\mathbf{Y} - Z\beta))$$

pertanto, supposta l'esistenza degli stimatori di massima verosimiglianza $\check{\lambda}$, $\check{\beta}$ e $\check{\sigma}^2$

$$\check{\mathbf{Y}} = Z\check{\beta} + \check{\lambda}W(\mathbf{Y} - Z\check{\beta})$$
$$\check{\mathbf{e}} = \mathbf{Y} - \check{\mathbf{Y}} = (\mathbb{I}_n - \check{\lambda}W)(\mathbf{Y} - Z\check{\beta})$$

Nota

Non c'è una formula esplicita per il calcolo dei parametri.
Si devono stimare numericamente.

Matrice dei pesi

Se per ogni individuo sono note le coordinate spaziali si possono calcolare

- d_{ij} = distanza euclidea tra gli individui i e j
- $d_{max} i$ = distanza tra i e l' m -esimo primo vicino

e porre

$$w_{ij} = \begin{cases} 1/m & \text{se } d_{ij} < d_{max} i \\ 0 & \text{altrimenti} \end{cases}$$

Nota

W è sparsa, ciò nonostante costi computazionali alti.

Dataset

Per applicare i modelli di regressione presentati abbiamo estratto da un dataset più ampio $n = 1000$ individui rappresentanti condomini della California. Le variabili osservate per ciascun condominio sono

Variabili

- valore
- rendita
- età
- numero di vani totali
- numero di camere totali
- numero di occupanti
- numero di proprietari
- latitudine
- longitudine

Variabili

Abbiamo poi considerato i seguenti regressori

Regressori

X1 rendita

X2 $(\text{rendita})^2$

X3 $(\text{rendita})^3$

X4 $\ln(\text{età})$

X5 $\ln(\text{numero di vani totali}/\text{numero di occupanti})$

X6 $\ln(\text{numero di camere totali}/\text{numero di occupanti})$

X7 $\ln(\text{numero di occupanti}/\text{numero di proprietari})$

X8 $\ln(\text{numero di proprietari})$

per la variabile risposta

Risposta

Y $\ln(\text{valore})$

Stime dei parametri per OLS e SAR

Fissato $m = 4$

	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	0.049893059

Stime dei parametri per OLS e SAR

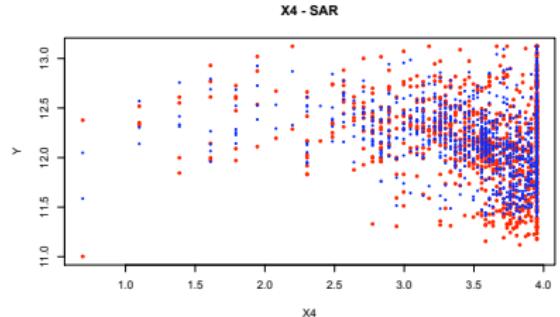
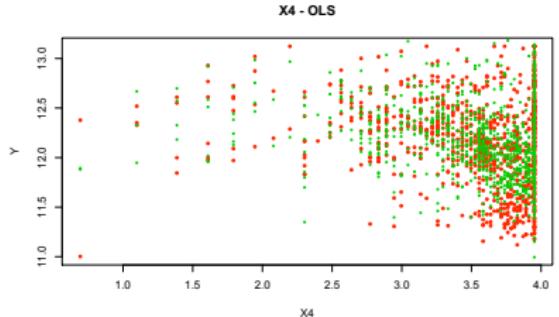
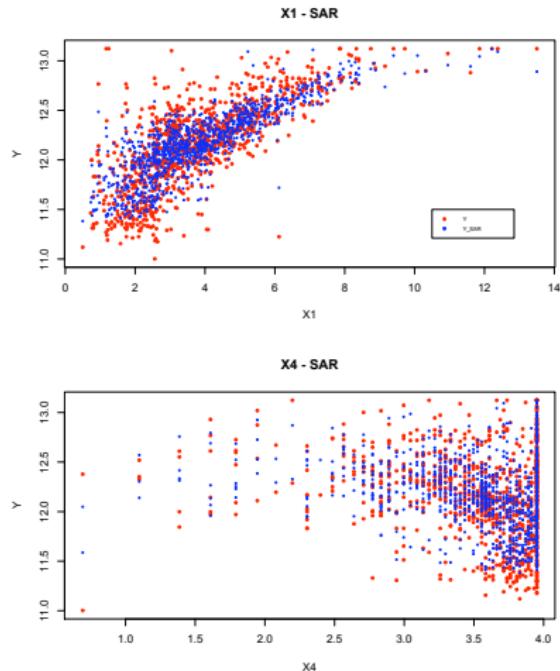
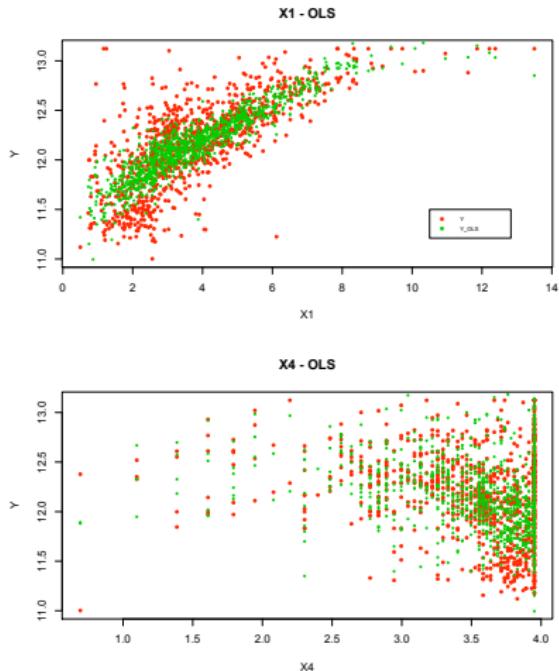
Fissato $m = 4$

	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	0.049893059

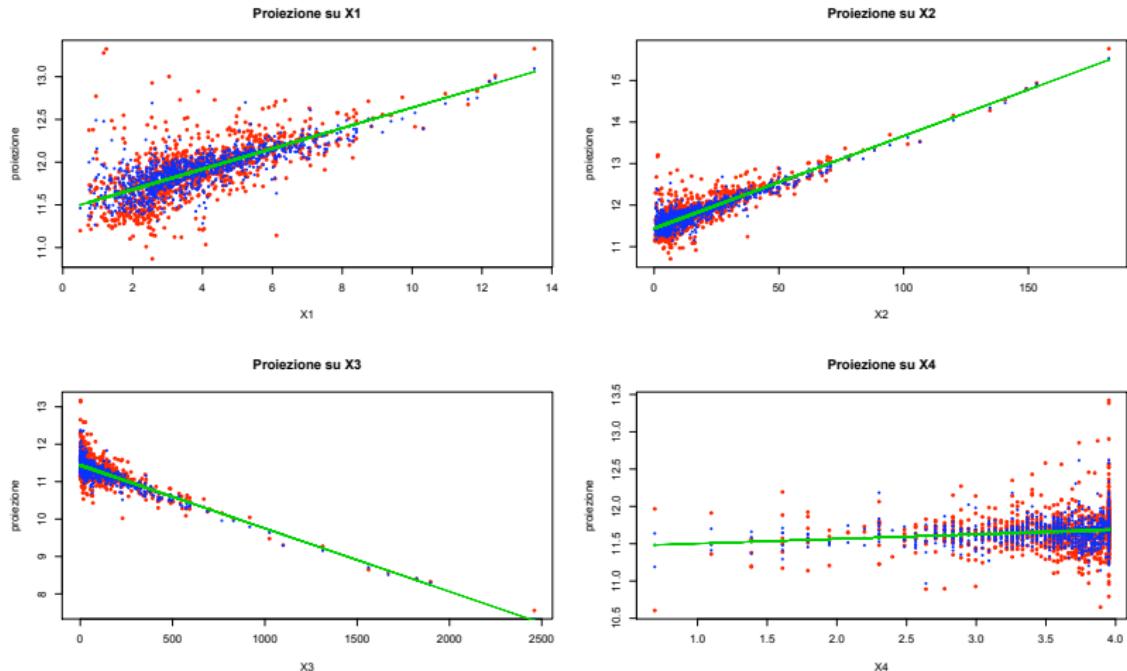
Indice di determinazione

	OLS	SAR
R^2	0.6448312	0.8112099

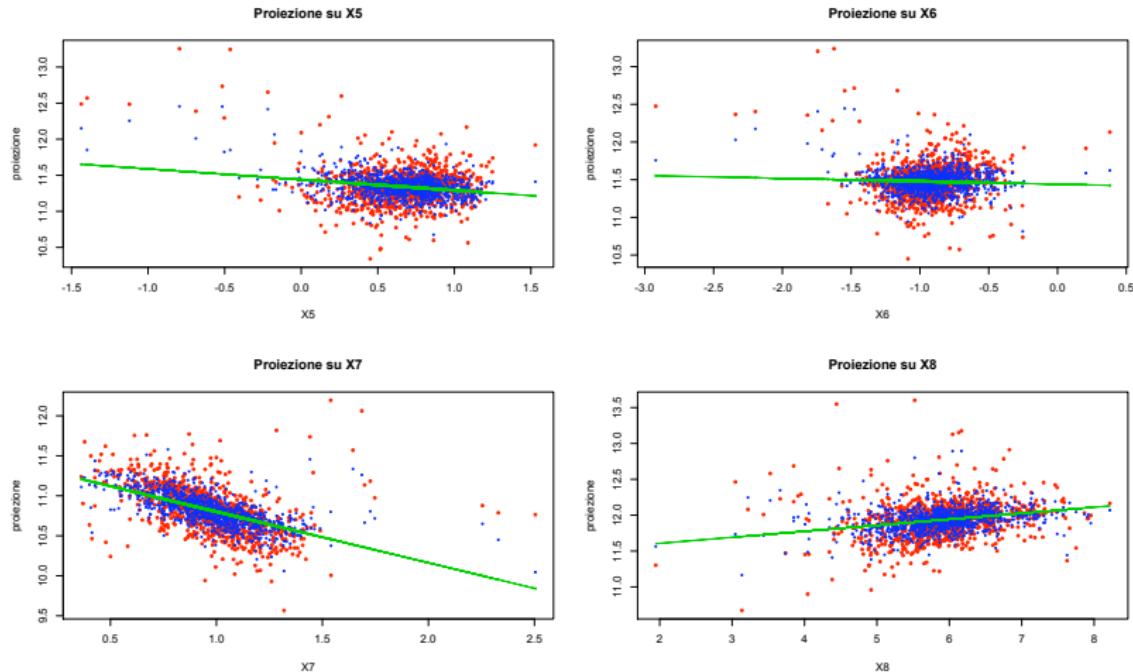
Rappresentazione dei risultati



Rappresentazione dei risultati



Rappresentazione dei risultati



Previsione

Nuovo campione Z' di $n' = 100$ individui

Previsione

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

\mathbf{Y}' non è noto, pertanto con SAR possiamo stimare solo $\check{\mathbf{Y}}' = \check{\boldsymbol{\beta}} \mathbf{Z}'$

Previsione

Nuovo campione Z' di $n' = 100$ individui

Previsione

$$\begin{array}{ll} \text{OLS} & \hat{\mathbf{Y}}' = \hat{\beta} \mathbf{Z}' \\ \text{SAR} & \check{\mathbf{Y}}' = \check{\beta} \mathbf{Z}' + \check{\lambda} \mathbf{W}(\mathbf{Y}' - \check{\beta} \mathbf{Z}') \end{array}$$

\mathbf{Y}' non è noto, pertanto con SAR possiamo stimare solo $\check{\mathbf{Y}}' = \check{\beta} \mathbf{Z}'$

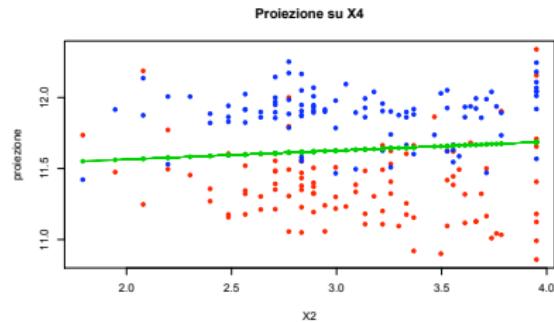
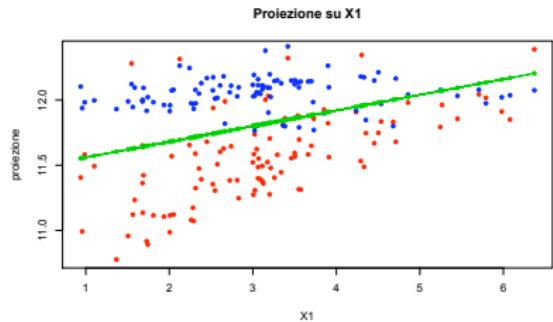
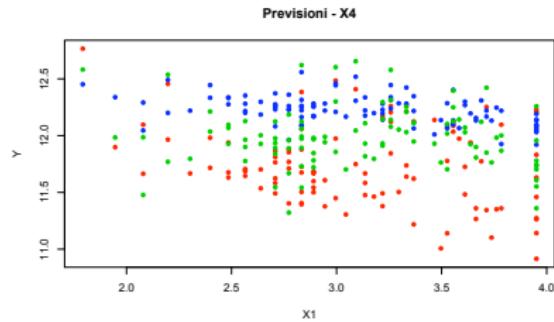
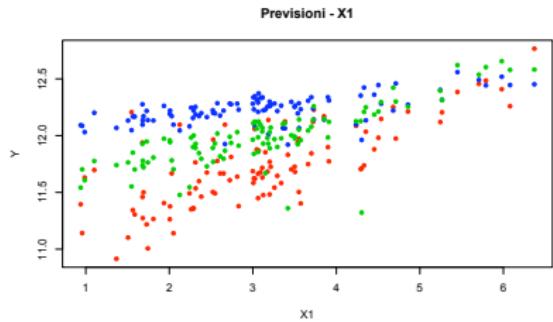
Idea

Sfruttiamo l'eventuale correlazione tra le osservazioni del nuovo campione Z' e quelle del vecchio Z

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

Ogni individuo del nuovo campione dipende dalle m osservazioni più vicine del vecchio campione

Rappresentazione delle previsioni

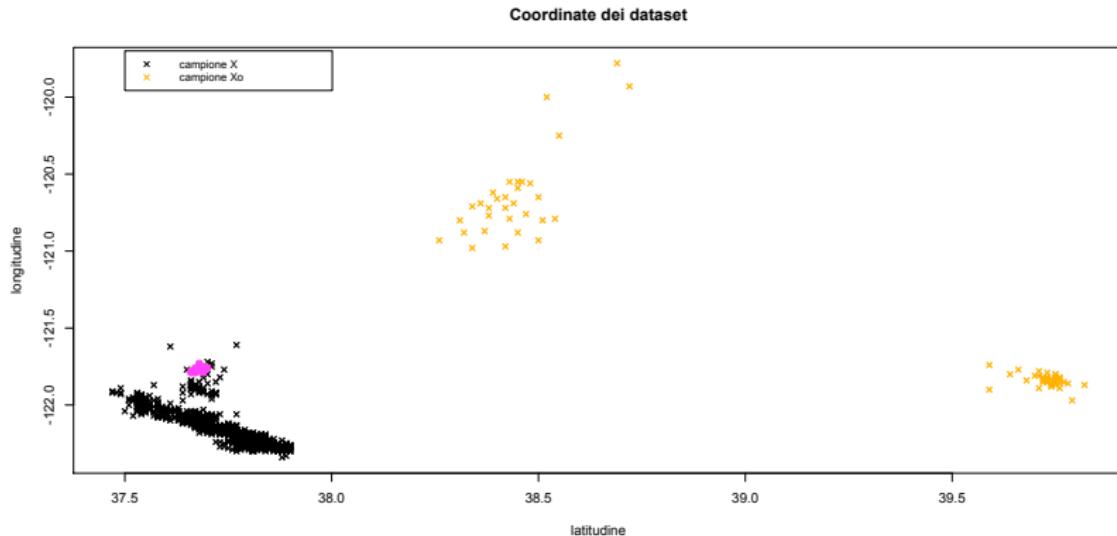


Risultati delle previsioni

SAR fallisce

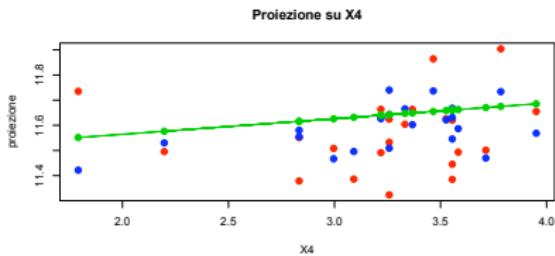
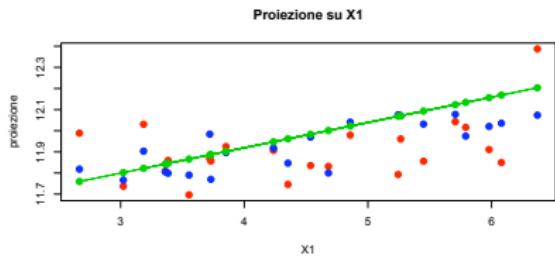
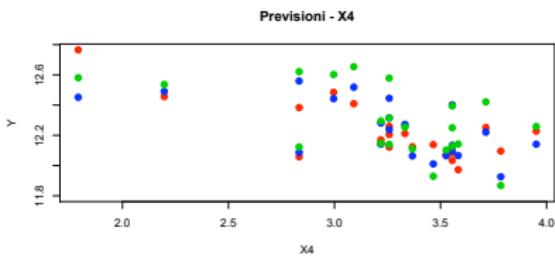
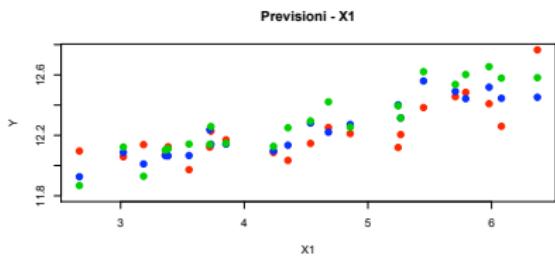
SAR fa previsioni peggiori rispetto a OLS

Vediamo perché...



Nuovo campione

Solo 22 individui del nuovo campione sono vicini spazialmente al vecchio.
Risultato in accordo con il modello SAR.



Conclusioni

Osservazione 1

SAR, cogliendo le correlazioni spaziali tra le osservazioni, è più adeguata di OLS nella rappresentazione del campione

Osservazione 2

SAR ottiene ottimi risultati di previsione solo per individui vicini al campione

Osservazione 3

SAR ha costi computazionali elevati