

Regressione con regolarizzazioni differenziali per dati spazio-temporali, con applicazione all'analisi della produzione di rifiuti urbani nella provincia di Venezia

Gabriele Mazza

29 Aprile 2015



Introduzione

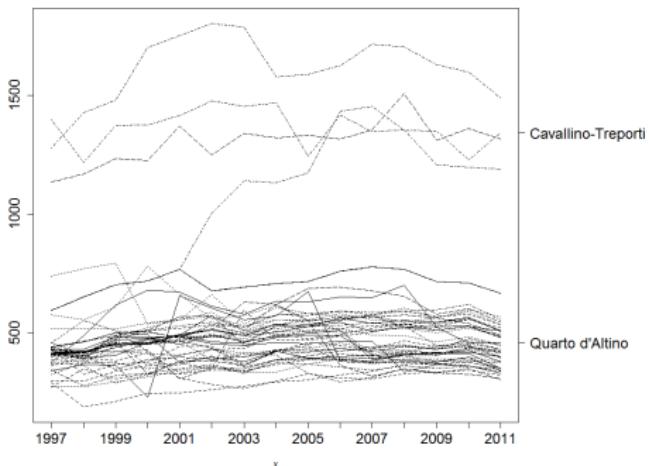
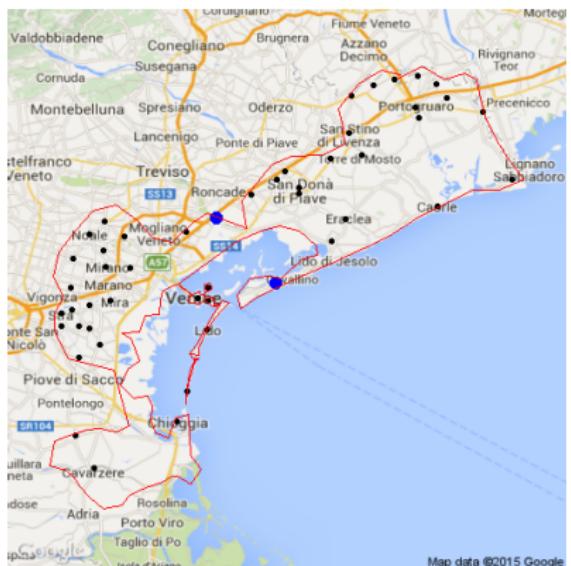
In questo lavoro di tesi è costruito e analizzato il modello di *Regessione Spazio-Temporale con Penalizzazioni Differenziali* (ST-PDE) per dati distribuiti in spazio e tempo:

$$z = f(\mathbf{p}, t)$$

con $\mathbf{p} \in \Omega$, $t \in [T_1, T_2]$.

Grande attenzione sarà dedicata al dominio spaziale.

L'applicazione scelta riguarda l'analisi della produzione di rifiuti urbani nella provincia di Venezia tra il 1997 e il 2011



Presentazione modello ST-PDE

Definisco:

- $\{\boldsymbol{p}_i = (x_i, y_i); i = 1, \dots, n\} \subset \Omega \subset \mathbb{R}^2$
- $\{t_j; j = 1, \dots, m\} \subset [T_1, T_2] \subset \mathbb{R}$
- z_{ij} osservazioni in (\boldsymbol{p}_i, t_j)

Modello:

$$z_{ij} = f(\boldsymbol{p}_i, t_j) + \varepsilon_{ij} \quad i = 1, \dots, n \quad j = 1, \dots, m$$

ε_{ij} rumore iid di media nulla e varianza σ^2

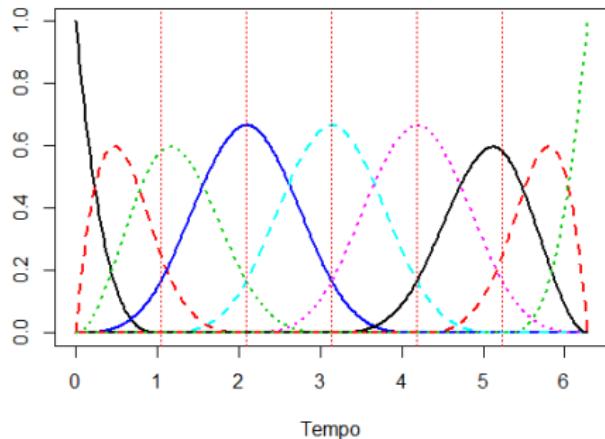
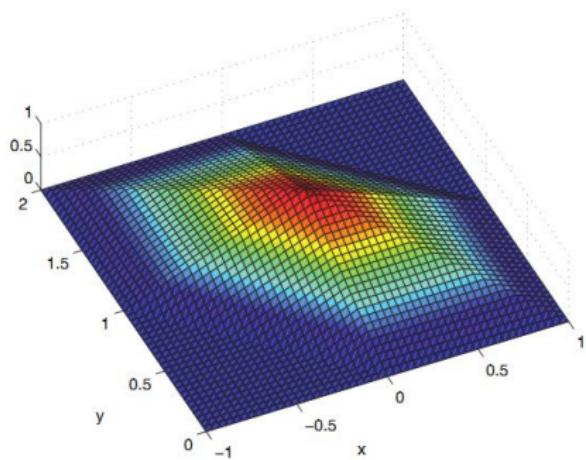
Funzioni di base in spazio e tempo:

$$\{\varphi_k(t); k = 1, \dots, M\}$$

$$\{\psi_l(p); l = 1, \dots, N\}$$

basi temporali definite in $[T_1, T_2]$

basi spaziali definite in Ω



La funzione è espressa tramite le funzioni di base:

$$f(\boldsymbol{p}, t) = \sum_{k=1}^M a_k(\boldsymbol{p}) \varphi_k(t) = \sum_{l=1}^N b_l(t) \psi_l(\boldsymbol{p}) = \sum_{l=1}^N \sum_{k=1}^M c_{lk} \psi_l(\boldsymbol{p}) \varphi_k(t)$$

La soluzione si ricaverà minimizzando il funzionale di penalizzazione:

$$\begin{aligned} J_{\lambda}(f(\boldsymbol{p}, t)) &= \sum_{i=1}^n \sum_{j=1}^m (z_{ij} - f(\boldsymbol{p}_i, t_j))^2 + \\ &+ \lambda_s \sum_{k=1}^M \int_{\Omega} (\Delta(a_k(\boldsymbol{p})))^2 d\boldsymbol{p} + \lambda_T \sum_{l=1}^N \int_{T_1}^{T_2} \left(\frac{\partial^2 b_l(t)}{\partial t^2} \right)^2 dt . \end{aligned}$$

Dati i vettori:

$$\mathbf{c} = \begin{bmatrix} c_{11} \\ \vdots \\ c_{1M} \\ c_{21} \\ \vdots \\ c_{NM} \end{bmatrix} \quad \boldsymbol{\psi} = \begin{bmatrix} \psi_1 \\ \psi_2 \\ \vdots \\ \psi_N \end{bmatrix} \quad \boldsymbol{\psi}_x = \begin{bmatrix} \partial\psi_1/\partial x \\ \partial\psi_2/\partial x \\ \vdots \\ \partial\psi_N/\partial x \end{bmatrix} \quad \boldsymbol{\psi}_y = \begin{bmatrix} \partial\psi_1/\partial y \\ \partial\psi_2/\partial y \\ \vdots \\ \partial\psi_N/\partial y \end{bmatrix} \quad \mathbf{z} = \begin{bmatrix} z_{11} \\ \vdots \\ z_{1m} \\ z_{21} \\ \vdots \\ z_{nm} \end{bmatrix}$$

e le matrici:

- $P_S = R_1 R_0^{-1} R_1 \quad R_0 = \int_{\Omega} \boldsymbol{\psi} \boldsymbol{\psi}^T d\mathbf{p}, R_1 = \int_{\Omega} (\boldsymbol{\psi}_x \boldsymbol{\psi}_x^T + \boldsymbol{\psi}_y \boldsymbol{\psi}_y^T) d\mathbf{p}$
- $P_T|_{k_1, k_2} = \int_{T_1}^{T_2} \varphi''_{k_1}(t) \varphi''_{k_2}(t)$
- $P = \lambda_S (P_S \otimes I_M) + \lambda_T (I_N \otimes P_T)$
- $B = \Psi \otimes \Phi \quad \Psi|_{i,I} = \psi_I(\mathbf{p}_i), \Phi|_{j,k} = \varphi_k(t_j)$

Allora:

$$J_{\lambda}(\mathbf{c}) = (\mathbf{z} - B\mathbf{c})^T (\mathbf{z} - B\mathbf{c}) + \mathbf{c}^T P \mathbf{c} \quad \Rightarrow \quad \hat{\mathbf{c}} = (B^T B + P)^{-1} B^T \mathbf{z}$$

Si inseriscono nel modello p possibili covariate:

$$z_{ij} = \mathbf{w}_{ij}^T \boldsymbol{\beta} + f(\mathbf{p}_i, t_j) + \varepsilon_{ij} \quad i = 1, \dots, n \quad j = 1, \dots, m$$

Data la matrice disegno W , che si ottiene accostando per colonna i vettori di covariate, allora:

$$J_\lambda(\mathbf{c}) = (\mathbf{z} - W\boldsymbol{\beta} - B\mathbf{c})^T (\mathbf{z} - W\boldsymbol{\beta} - B\mathbf{c}) + \mathbf{c}^T S \mathbf{c}$$

Derivando:

$$\begin{cases} W^T W \hat{\boldsymbol{\beta}} = W^T (\mathbf{z} - B\hat{\mathbf{c}}) \\ (B^T B + P)\hat{\mathbf{c}} = B^T (\mathbf{z} - W\hat{\boldsymbol{\beta}}) \end{cases} \Rightarrow \begin{cases} \hat{\boldsymbol{\beta}} = (W^T W)^{-1} W^T (\mathbf{z} - B\hat{\mathbf{c}}) \\ \hat{\mathbf{c}} = A Q \mathbf{z} \end{cases}$$

con

$$Q = [I - W(W^T W)^{-1} W^T] \quad A = [B^T Q B + P]^{-1} B^T$$

Studi di simulazione

Le simulazioni sono state eseguite simulando da $f(\mathbf{p}, t) = g(\mathbf{p})\cos(t)$: del rumore:

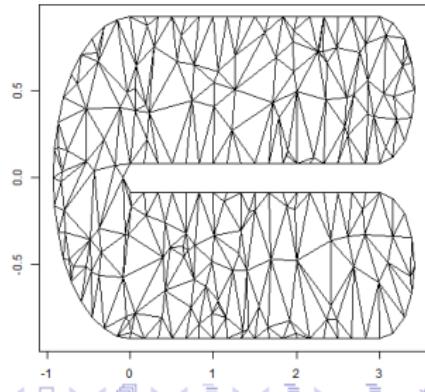
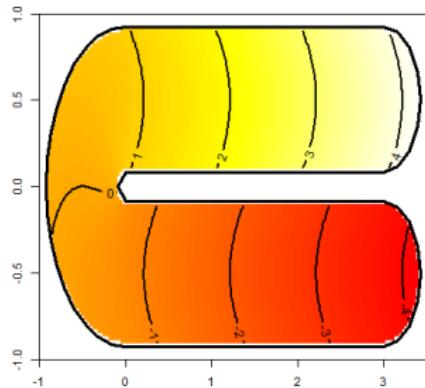
$$z_{ij} = f(\mathbf{p}_i, t_j) + \beta w_{ij} + \varepsilon_{ij} \quad \forall i, \forall j$$

dove:

- $\beta = 1$ (eventuale)
- $w_{ij} \stackrel{\text{iid}}{\sim} N(0, 1) \quad \forall i, \forall j$
- $\varepsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, 0.5^2) \quad \forall i, \forall j$

I parametri di smoothing sono calcolati tramite GCV:

- senza covariata $\lambda = (10^{-0.375}, 10^{-3.25})$
- con covariata $\lambda = (10^{-0.5}, 10^{-3.25})$



Il modello stima:

$$\hat{\beta} \approx 1.001$$

IC approssimato:

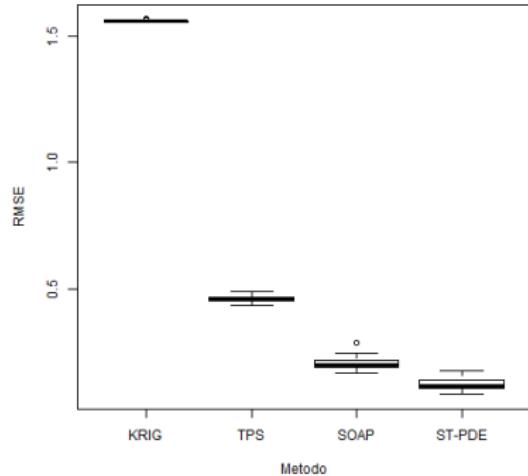
$$\beta \in [0.9809; 1.0225]$$

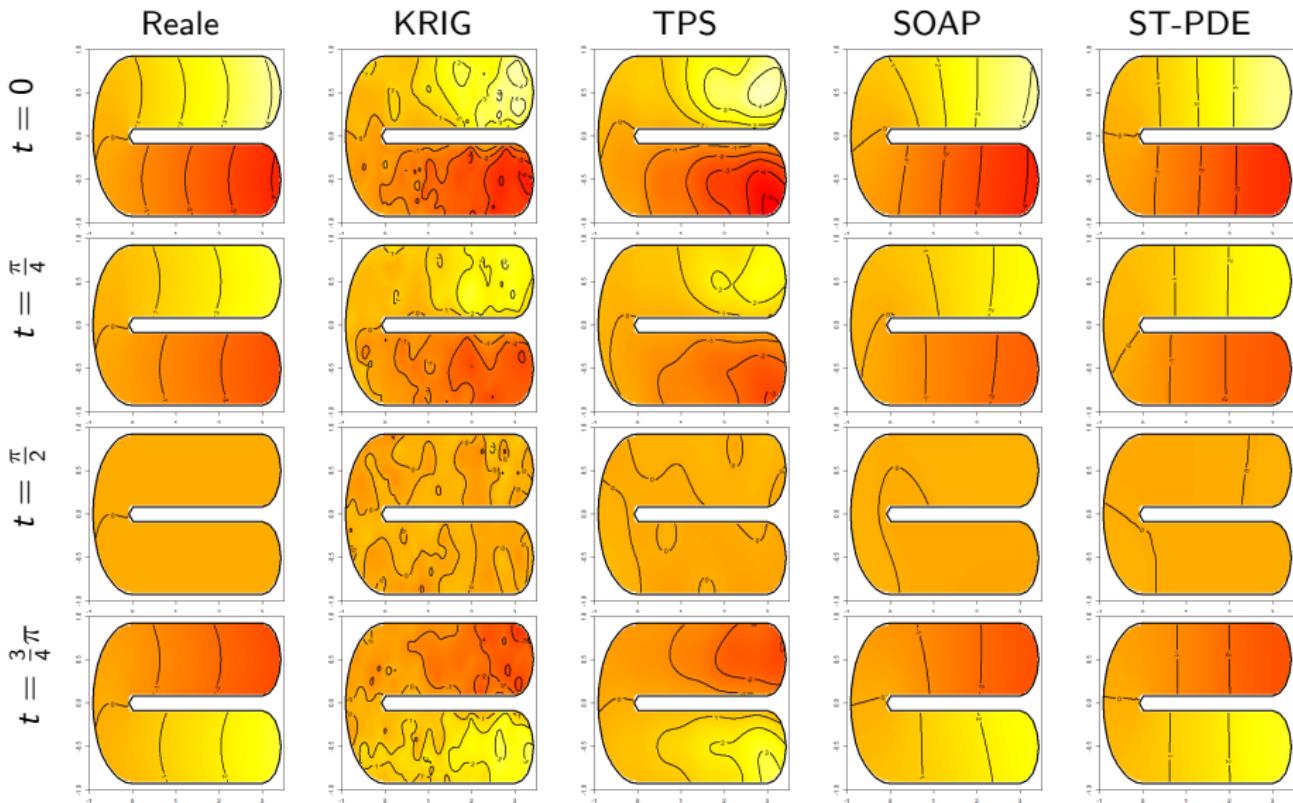
Confronto con altri metodi

L'algoritmo è stato confrontato con altre tecniche già esistenti:

- KRIG: modello basato su kriging spazio-temporale
- TPS: basi in spazio *Thin Plate Splines*, in tempo *Smoothing Splines*
- SOAP: basi in spazio *Soap Film Smoothing*, in tempo *Smoothing Splines*

Solo SOAP e ST-PDE possono tener conto del dominio spaziale!



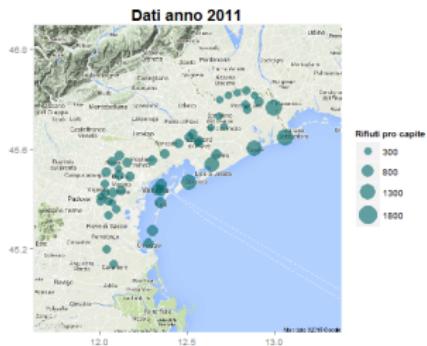


Confronto con altri metodi

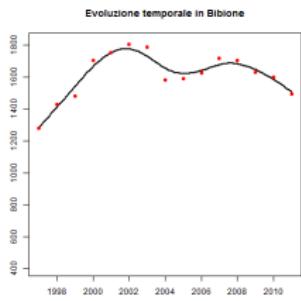
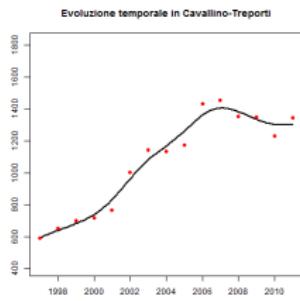
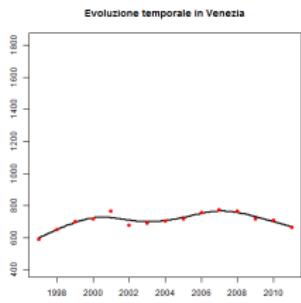
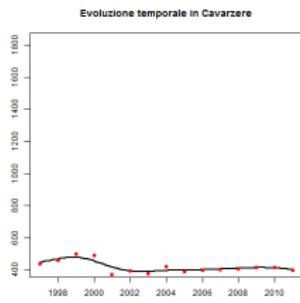
I dati sono stati localizzati in un unico punto, nel paese di riferimento del comune.

Come covariata, per tener conto dell'effetto del turismo, si usa il numero di posti letto pro capite in strutture ricettive.

Sono stati usati i valori pro capite, sia per il dato che per la covariata (possibilità di replicare i dati poichè densità).



Risultati dell'applicazione del modello senza la covariata:



Risultati dell'applicazione del modello con la covariata:

$$\hat{\beta} \approx 30.5563 \quad \beta \in [14.3158; 46.7767]$$

