Sommario

Il presente lavoro di tesi illustra il modello statistico *Spatio-Temporal Regression model with PDE penalization* (STR-PDE) per l'analisi funzionale di dati distribuiti in un dominio spaziale e in un intervallo temporale, estendendo il caso puramente spaziale proposto in [6]. Il modello ipotizza che i dati possano essere rappresentati dalla somma di una funzione spazio-temporale e di un eventuale termine di covariate. I risultati analitici hanno portato alla creazione di un codice R, con cui il modello ha potuto essere testato nel caso noto del dominio a forma di C descritto in [5] e confrontato con alcune tecniche già esistenti. L'applicazione studiata riguarda l'analisi della produzione di rifiuti urbani pro capite nella provincia di Venezia tra il 1997 e il 2011, con un'attenzione particolare agli effetti legati al turismo

Abstract

The present work of thesis describes the statistical model *Spatio-Temporal Regression model with PDE penalization* (STR-PDE) for the functional analysis of data distributed over a spatial domain and a temporal interval, extending the case purely spatial proposed in [6]. The model assumes that the data could be represented by the sum of a function of space and time and a possible term of covariates. The analytical results led to the creation of a R code, with which the model could be tested in the known case of the C-shaped domain described in [5] and compared with some existing techniques. The application studied concerns the analysis of urban waste production per capita in Venice province between 1997 and 2011, with a particular attention to the effects related to tourism.

Ringraziamenti

da fare

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Il presente lavoro di tesi illustra il modello statistico *Spatio-Temporal Regression model with PDE penalization* (STR-PDE) per l'analisi funzionale di dati distribuiti in spazio e tempo. Quanto fatto può essere considerato un'estensione dei modelli proposti in [6] che studiano la possibilità di costruire una stima funzionale per dati distribuiti su un dominio spaziale attraverso l'approssimazione in basi di elementi finiti. Il modello STR-PDE, invece, sviluppa una tecnica analoga permettendo la variazione temporale alla stima funzionale precedente. Di conseguenza può essere considerato un buon strumento per lo studio di fenomeni varianti in spazio e in tempo. Dalla modellizzazione matematica è stato sviluppato un algoritmo e il codice R per il calcolo della soluzione numerica della stima.

Il lavoro è motivato dalla ricerca di un buon metodo di analisi di un dataset contenente le misurazioni della produzione dei rifiuti urbani pro capite nei comuni della provincia di Venezia tra il 1997 e il 2011. I dati sono stati raccolti ed elaborati dall'Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto (Arpav) e sono disponibili sul sito di Open Data Veneto[1] per la consultazione e il trattamento. Le misurazioni contenute nel dataset in realtà riguardano tutto il Veneto, ma per semplicità computazionale e per l'elevato interesse riguardo la laguna veneta sarà analizzata solo la provincia di Venezia. Il modello STR-PDE permette di stimare l'andamento della produzione dei rifiuti su tutta la regione e ad ogni istante di tempo nell'intervallo considerato, garantendo una chiara visualizzazione del fenomeno.

Il lavoro di tesi sarà strutturato come segue. Nel Capitolo 1 è riportato un excursus sui metodi simili già esistenti in letteratura. Nel Capitolo 2 è presentata la costruzione del modello matematico STR-PDE. Nel Capitolo

[1] http://dati.veneto.it/dataset/produzione-annua-di-rifiuti-urbani-totale-e-pro capite-1997-2011

3 si hanno i primi risultati, derivanti dall'applicazione del modello e del codice R al caso del dominio a forma di C descritto in [5] e [7], per il quale è possibile valutare la bontà delle stime ottenute grazie alla perfetta conoscenza del fenomeno reale in ogni punto e in ogni istante. Nel capitolo 4 il modello STR-PDE è paragonato ad altri metodi già esistenti per il confronto delle stime ottenute. Nel Capitolo 5 si ha l'applicazione allo studio della produzione dei rifiuti nella provincia di Venezia e infine, nel Capitolo 6, sono raccolte le conclusioni e i possibili sviluppi futuri riguardanti miglioramenti del lavoro.

CAPITOLO 1

Panoramica sui modelli già esistenti

Come già accennato nell'introduzione, il lavoro si propone di essere un'estensione al caso tempo-variante di quanto fatto in [6]. In questa pubblicazione si ipotizza che i dati siano distribuiti su un dominio limitato e che possano essere descritti da una funzione con l'aggiunta di rumore:

$$z_i = f(\underline{p}_i) + \varepsilon_i$$

dove $\underline{p}_i$ è il vettore delle coordinate. Estendere ciò al caso tempo-variante significa ipotizzare che la funzione dipenda anche dal tempo e che i dati abbiano anche un'informazione legata al tempo a cui si riferiscono all'interno di un certo intervallo. Inoltre, esattamente come in [6], alla funzione stimata sarà possibile garantire anche una buona regolarità.

L'approccio seguito sarà differente sotto alcuni aspetti. Per poter stimare la funzione solo spaziale in [6] era posto un funzionale di penalizzazione da minimizzare con la somma di un termine di scarti quadratici tra dati e valori stimati dal modello e di un integrale di opportune derivate della funzione (utile ad avere una stima più o meno liscia). Il problema di minimo era ridotto ad un problema variazionale che, per poter essere risolto computazionalmente, necessitava della riduzione in combinazione lineare di opportune funzioni di base per la funzione da stimare. In sostanza si passava da una formulazione complessa ad una più semplice (solitamente risolvibile tramite un sistema lineare) ammettendo che la funzione possa appartenere ad uno spazio finito-dimensionale solo dopo aver garantito una forma variazionale corrispondente al problema di minimo iniziale. Nell'approccio seguito in questo lavoro, come si potrà vedere, non sarà così. Tuttavia l'articolo è stato da ispirazione per molte cose: la modellizzazione elementi finiti come basi in spazio, la scelta di penalizzare il laplaciano, l'uso di alcune matrici (come R_0 e R_1 che si potranno

vedere nel corso della spiegazione del modello nel Capitolo 2) derivano dalla volontà di estendere il caso puramente spaziale.

In [1] e [4] sono disponibili metodi per l'analisi di dati distribuiti in spazio e tempo. In questi casi si ha un approccio che ipotizza da subito che la funzione possa essere estesa come sviluppo di funzioni di base tramite una metodologia più generale. Si hanno più modi di rappresentare la funzione (nel Capitolo 4 e saranno usati come confronto con il modello STR-PDE).

CAPITOLO 2

Presentazione del modello STR-PDE

CAPITOLO 3

Applicazione al dominio a forma di C

CAPITOLO 4

Confronto con altri metodi

Il modello STR-PDE rappresenta una generalizzazione del caso puramente spaziale proposto in [6] e, come è già stato evidenziato nel Capitolo 1, non è l'unico modello disponibile per l'analisi di dati distribuiti sia in spazio che in tempo. Pertanto è necessario che sia confrontato con le altre principali metodologie presenti in letteratura, al fine di poter dire se e quanto il modello proposto possa rappresentare un miglioramento in questo campo.

L'articolo [1] propone l'analisi di dati di questo tipo attraverso modelli misti additivi generalizzati (GAMM) di interazione spazio-tempo. Questo metodo è generalizzato, quindi può essere usato per spiegare anche funzioni del valore atteso della risposta. Nel nostro caso, per avvicinarci al caso STR-PDE, si ipotizza che la risposta sia pari alla somma di una funzione e di un eventuale termine con covariata. Alla funzione è associato lo smoothing secondo il prodotto tensoriale dei termini marginali in spazio e tempo con le loro penalizzazioni. Quindi la costruzione dei GAMM è molto simile a quella analizzata in STR-PDE e, mediante il codice implementato nel pacchetto R *mgcv*, è possibile scegliere tra più tipi di modelli. In particolare ne saranno studiati due, i più simili al modello STR-PDE:

- TPS, in cui sono poste marginalmente *cubic regression splines* in tempo e *thin plate splines* in spazio;
- SOAP, che considera *cubic regression splines* in tempo e *soap film smoothing* in spazio.

Un altro metodo da confrontare è sicuramente il kriging (KRIG) spazio-temporale. Le stime sono ottenute fissando un variogramma separabile e

marginalmente esponenziale in spazio e tempo. I parametri del variogramma sono stimati dall'empirico e, successivamente, è possibile calcolare la stima grazie alle funzioni del pacchetto R *spacetime*. Il confronto con kriging è stato possibile solo nel caso senza covariata.

I quattro modelli sono confrontati sull'esempio del dominio a forma di C proposto precedentemente, poichè garantisce di poter conoscere in ogni punto spaziale e ad ogni istante temporale il valore esatto della funzione. La triangolazione e i dati sono gli stessi che sono stati usati nel Capitolo 3. In aggiunta è stata costruita una griglia spazio-temporale di punti per la validazione: sono stati presi 80 punti equispaziati in $(-1, +3.5)$ per l'ascissa, 40 punti in equispaziati $(-1, +1)$ per l'ordinata e 20 istanti in $(0, 2\pi)$ per il tempo. Ovviamente la validazione è stata studiata soltanto sui punti che ricadevano all'interno del dominio a forma di C.

I modelli sono stati confrontati attraverso il Root Mean Square Error (RMSE) prodotto sui punti di validazione. Quindi se se V è l'insieme dei punti della griglia interni al dominio, e Mod rappresenta la stima ottenuta dal modello, si avrà:

$$\mathrm{RMSE}_V(\mathrm{Mod}) = \sqrt{\frac{\sum_{(\underline{p}_i, t_i) \in V} (\mathrm{Mod}(\underline{p}_i, t_i) - g(\underline{p}_i) cos(t_i))^2}{\mathrm{card}(V)}}$$

Il procedimento è stato iterato 50 volte, per poter escludere possibili andamenti particolari dovuti alla generazione del rumore.

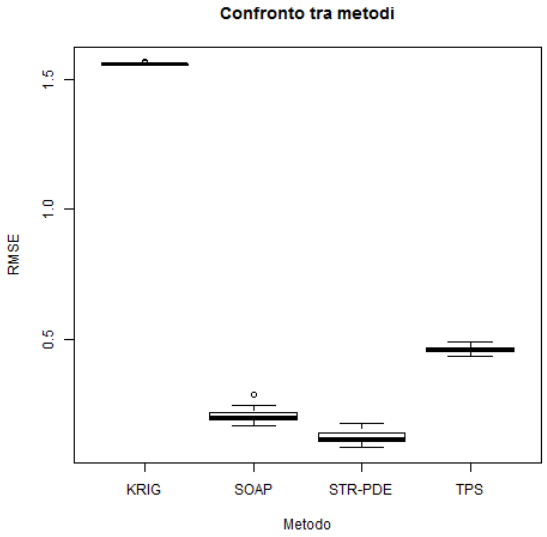


Figura 4.1: Confronto tra i metodi, caso senza covariata

4.1 Caso senza covariata

Nel caso senza covariata si hanno i risultati riportati in fig. 4.1, in cui sono stati tracciati i boxplot dei valori di RMSE raccolti nelle 50 iterazioni per ogni metodo. Subito si nota che l'errore commesso è minore nel caso di STR-PDE, e quindi la stima ottenuta con il modello proposto è la migliore.

Tutto ciò è confermato dai grafici presenti in fig. 4.2. Dai boxplot si nota che l'errore commesso è più alto nei casi di KRIG e TPS, e infatti le stime sono molto distanti dalla funzione reale. Invece SOAP e STR-PDE commettono errori minori, ma tra i due il migliore è STR-PDE, che ha linee di livello più ordinate rispetto a SOAP.

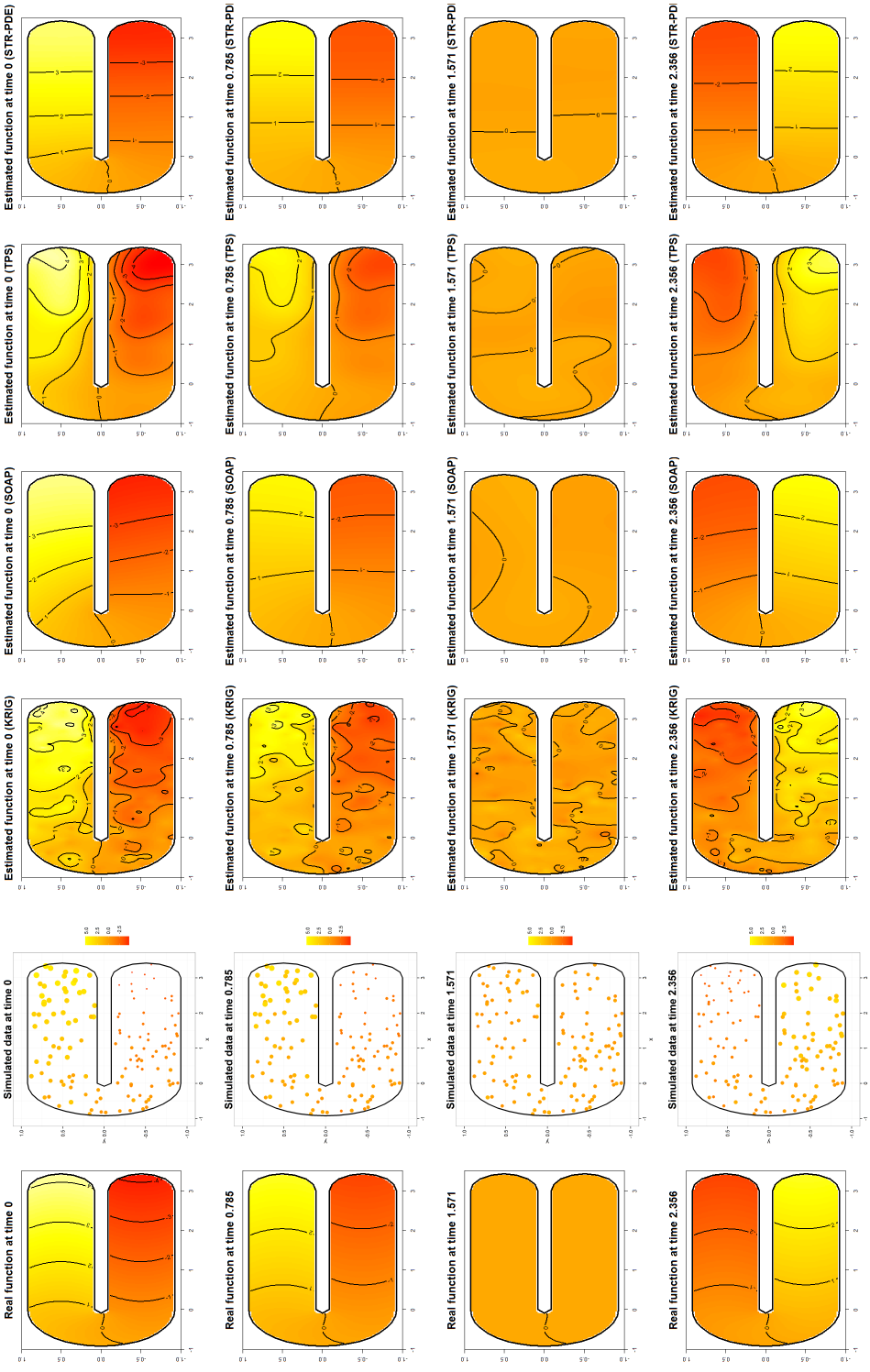
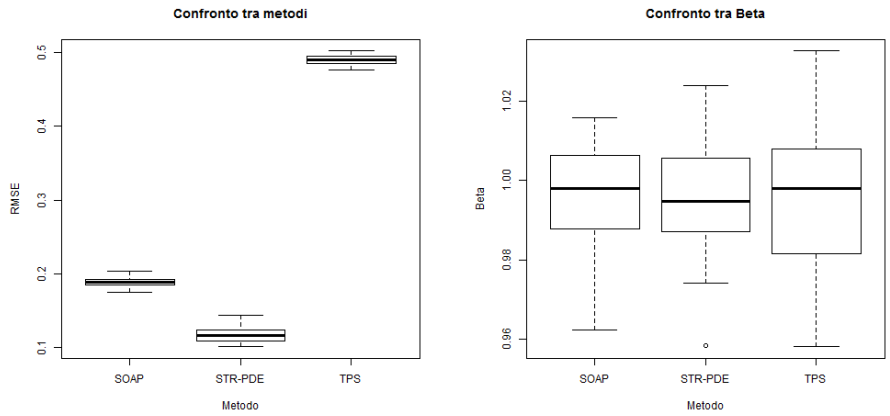


Figura 4.2: Per alcuni istanti di tempo, funzione test $f(p,t)$ reale, dati simulati, stime ottenute rispettivamente con kriging spazio-temporale, GAMM con soap film smoothing, GAMM con thin plate splines e stima con STR-PDE nel caso senza covariata.



(a) RMSE (b) Stime $\hat{\beta}$

Figura 4.3: Confronto tra i metodi, caso con covariata

4.2 Caso con covariata

La stessa analisi, senza il caso del kriging spazio-temporale, è stata eseguita nel caso con covariata. La covariata è stata generata in tutti i punti esattamente come fatto nel Capitolo 3. Nella calcolo del RMSE, poichè non è opportuno generare nuovamente valori per la covariata nei punti di validazione, è stato considerato solo il termine dipendente dalla funzione $f(\underline{p}, t)$. I boxplot riportati in fig. 4.3(a) possono quindi essere considerati come valutazione della bontà della stima della parte funzionale del modello. Per la parte spiegata dalla covariata sono stati tracciati i boxplot in fig. 4.3(b), con le stime di $\hat{\beta}$ calcolate dai metodi.

Le conclusioni sono perfettamente analoghe al caso precedente. La stima di β non presenta differenze, ma nella parte funzionale il caso STR-PDE è nuovamente il migliore.

Analogamente al caso senza covariata, dai plot della funzione stimata ad alcuni istanti di tempo fissati in fig. 4.4 si possono trarre le stesse conclusioni. Il modello STR-PDE è quello che più si avvicina alla funzione reale.

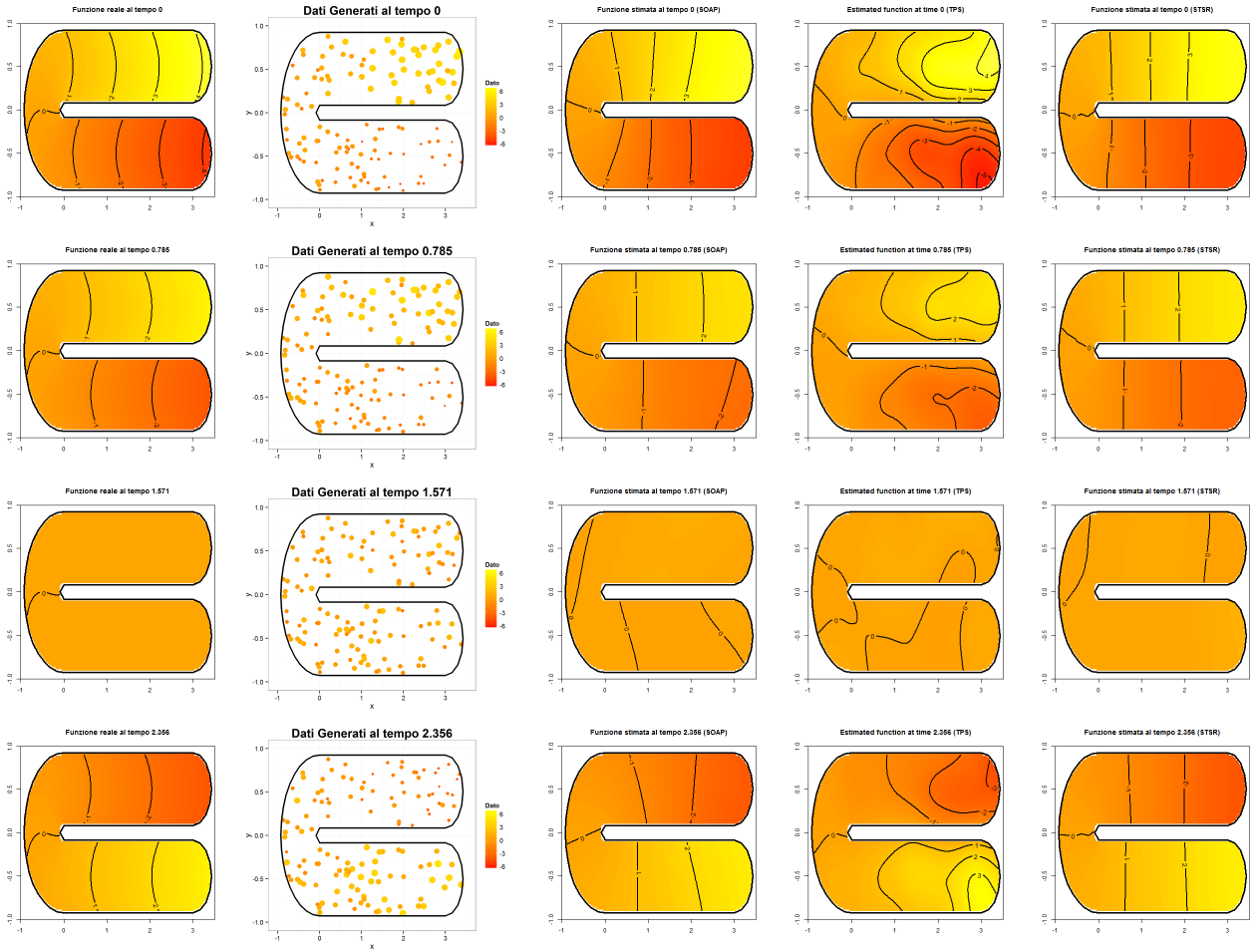


Figura 4.4: Per alcuni istanti di tempo, funzione test $f(\underline{p}, t)$ reale, dati simulati, stime ottenute rispettivamente con GAMM con soap film smoothing, GAMM con thin plate splines e stima con STR-PDE nel caso con covariata.

CAPITOLO 5

Analisi della produzione di rifiuti urbani nella provincia di Venezia

CAPITOLO 6

Conclusioni e sviluppi futuri

In questo lavoro di tesi è stato analizzato nel dettaglio il modello STR-PDE nell'ambito della stima funzionale per dati varianti all'interno di un dominio spaziale e di un intervallo temporale. Il modello, che si propone di essere un'estensione del caso puramente spaziale già analizzato in letteratura, è stato sviluppato in codice R. Dal confronto con gli altri metodi e da quanto ricavato con le stime, soprattutto sul dominio a forma di C in cui è possibile conoscere il valore reale della funzione, si può concludere che i risultati prodotti sono molto buoni.

Diversa è la conclusione per le prestazioni computazionali del codice. Per semplicità computazionale le basi degli elementi finiti sono state scelte lineari e la produzione dei rifiuti è stata analizzata solamente nella provincia di Venezia, pur avendo a disposizione i dati di tutto il Veneto. Inoltre, durante l'esecuzione del codice, si è potuto notare che alcune funzioni come la minimizzazione di GCV($\underline{\lambda}$) o il calcolo dei valori stimati ad un istante di tempo fissato (usati ad esempio per conoscere il profilo della funzione ad un certo anno) sono molto lente. Ovviamente per analisi di dataset di grosse dimensioni deve essere messa in conto una spesa di tempo elevata, ma R certamente non ha aiutato. Infatti, è noto che R non sia un linguaggio di programmazione fortemente efficiente, e questo ha caratterizzato la lentezza di esecuzione. Il più chiaro sviluppo futuro può essere l'uso del codice come base per la creazione di un algoritmo più veloce, attraverso l'integrazione con un linguaggio di programmazione più efficiente (come il C++) o della parallelizzazione nei colli di bottiglia più evidenti.

Dopo che sarà stata sviluppata l'integrazione del codice, sarà possibile garantire una analisi più agile anche per dataset di dimensioni più elevate

o per elementi finiti di ordine maggiore. In questo modo si avrà a disposizione uno strumento di analisi statistica buono non solo dal punto di vista dei risultati, ma anche in termini di efficienza computazionale.

Bibliografia

[1] Nicole H. Augustin, Verena M. Trenkel, Simon N. Wood, Pascal Lorance, *Space-time modelling of blue ling for fisheries stock management*, Environmetrics, 24, 109–119, (2013)

[2] Laura Azzimonti, Laura M. Sangalli, Piercesare Secchi, Maurizio Domanin, Fabio Nobile, *Blood flow velocity field estimation via spatial regression with PDE penalization*, Journal of the American Statistical Association, (2015)

[3] Peter Craven, Grace Wahba, *Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation*, Numerische Mathematik, 31, 377–403, (1979)

[4] Giampiero Marra, David L. Miller, Luca Zanin, *Modelling the spatiotemporal distribution of the incidence of resident foreign population*, Statistica Neerlandica, 66, 133–160, (2012)

[5] Timothy O. Ramsay, *Spline smoothing over difficult regions*, Journal of the Royal Statistical Society: Series B, 64, 307–319, (2002)

[6] Laura M. Sangalli, James O. Ramsay, Timothy O. Ramsay, *Spatial spline regression models*, Journal of the Royal Statistical Society: Series B, 75, 681–703, (2013)

[7] Simon N. Wood, Mark W. Bravington, Sharon L. Hedley, *Soap film smoothing*, Journal of the Royal Statistical Society: Series B, 70, 931–955, (2008)