Politecnico di Milano
Scuola di Ingegneria Industriale e dell'Informazione
Corso di laurea in Ingegneria Matematica

TITOLO

Relatore: Prof. Laura SANGALLI

Tesi di Laurea di:
Gabriele Mazza Matr. 798794

Indice

Introduzione

Il presente lavoro di tesi illustra un modello statistico spatio-temporal spline regression (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. Sono disponibili le misurazioni per tutta la regione del Veneto, ma per semplicità computazionale e per l'elevato interesse della 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 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

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

della produzione dei rifiuti nella provincia di Venezia, e infine nel Capitolo 6 sono raccolte le conclusioni e i possibili sviluppi futuri.

Capitolo 1

Panoramica sui modelli già esistenti

Capitolo 2

Presentazione modello

Capitolo 3

Applicazione al dominio a forma di C

Capitolo 4

Confronto con gli altri modelli

Il modello STR-PDE rappresenta una generalizzazione del caso puramente spaziale proposto in [6] e, come già 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 covariate. 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 grazie al 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 dei variogrammi sono stimati dal variogramma empirico, e successivamente è possibile calcolare la stima grazie alle funzioni del pacchetto R *spacetime*.

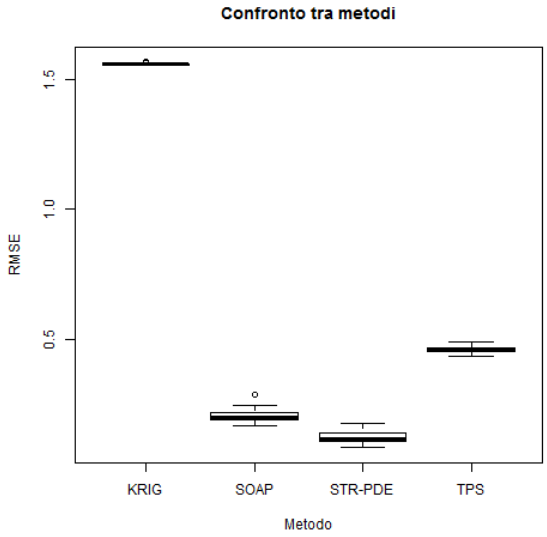

Figura 4.1: Confronto del RMSE, caso senza covariate

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.

4.1 Caso senza covariate

Nel caso senza covariate si hanno i risultati riportati in figura 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 SOAP ha linee di livello meno ordinate che STR-PDE.

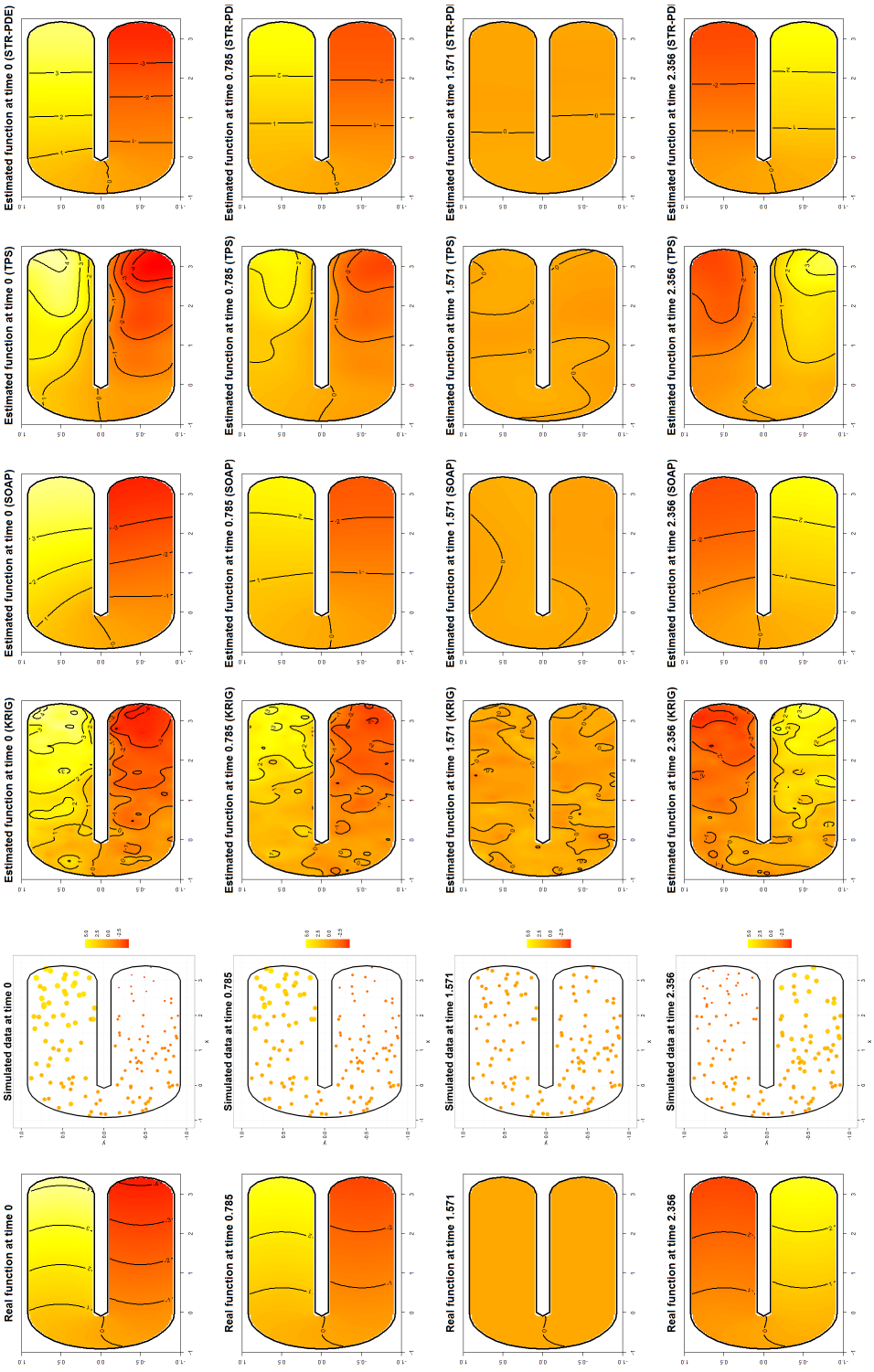

Figura 4.2: Per alcuni istanti di tempo, funzione test $f(\underline{p}, t)$ reale, dati simulati, stime ottenute rispettivamente con kriging spazio-temporale, GAMM con soap film smoothing, GAMM con thin plate splines e dtima con STR-PDE.

4.2 Caso con covariate

Capitolo 5

Produzione di rifiuti 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 di questo codice come base per lo sviluppo di un algoritmo più veloce, attraverso l'integrazione con un linguaggio di programmazione più efficiente (come il C++) 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)