POLITECNICO DI MILANO

Facoltà di Ingegneria dei Sistemi

Corso di Studi in INGEGNERIA MATEMATICA

Tesi di Laurea Magistrale

APPLICAZIONE DI MODELLI DI REGRESSIONE SPAZIALE A DATI OCEANOGRAFICI

Relatore: Dr. Luca Bonaventura

Correlatore: Dr.ssa Laura M. Sangalli

Candidato:

Matteo Parnigoni

matr. 751671

Anno Accademico 2012-2013

Alla mia famiglia.

Indice

Ringraziamenti

Vorrei ringraziare il Dottor Luca Bonaventura per la sua grande disponibilità e pazienza nel seguirmi in questi mesi. Lo ringrazio per avermi dato la possibilità di affrontare questi argomenti di mio grande interesse seppur non tutti in linea con il mio indirizzo universitario.

Vorrei ringraziare anche la Dottoressa Laura Sangalli per la sua disponibilità durante questo periodo, per tutto il materiale fornitomi e per le sue preziose consulenze teoriche e pratiche, davvero molto utili per lo svolgimento del progetto della tesi.

Un grande ringraziamento lo devo a quei compagni di studi, a quegli amici e a quei compagni di lunghi viaggi in treno che in questi anni hanno condiviso con me momenti belli, momenti divertenti e anche momenti più difficili, sostenendoci a vicenda. Sono stati tutti davvero importanti!
In particolare grazie a Florence, Marta, Matteo, Marco e Marina che, in modi diversi, mi hanno saputo aiutare maggiormente, condividendo appunti, dandomi consigli e soprattutto dandomi forza nei periodi più difficili.

Un ringraziamento anche a zio Ivo per la sua disponibilità ad accompagnarmi in stazione e per i tanti piccoli, ma fondamentali, aiuti che mi ha dato in questi anni.

Il ringraziamento più speciale e più sentito lo devo ai miei genitori e a mio fratello Paolo, senza i quali non sarei potuto arrivare fino a questo traguardo. Il loro supporto, la loro pazienza e la loro fiducia sono stati davvero fondamentali per me in questi anni.

GRAZIE!

Matteo

Capitolo 1

Introduzione

In molti ambiti delle scienze applicate vi è l'esigenza di ricostruire o stimare dei campi scalari a partire da un numero limitato di osservazioni distribuite irregolarmente nello spazio. Ad esempio, in ambito geologico-minerario si può avere l'interesse di ottenere una ricostruzione di caratteristiche del sottosuolo per stimare la presenza e la distribuzione di un certo minerale; per motivi simili, l'industria estrattiva petrolifera necessita spesso di stimare la distribuzione di idrocarburi nel sottosuolo avendo a disposizione un numero limitato di osservazioni, perché la trivellazione del sottosuolo è molto costosa. In ambito ambientale, è forte l'interesse di tecniche che possano ricostruire la distribuzione di inquinanti atmosferici (un esempio in [Crippa, 2007]). Necessità simili possono essere trovate anche in meteorologia, in climatologia, nell'idrologia e nell'oceanografia. Molte sono le tecniche che negli anni sono state proposte per risolvere il problema di interpolazione spaziale: tra le più note si possono menzionare le tecniche di *Kriging* ([Cressie, 1991], [Wackernagel, 1995], [Kitanidis, 1992]) tradizionalmente note come tecniche di interpolazione geostatistica, introdotte dall'ingegnere D. Krige negli anni '50, le tecniche *thin-plate splines* ([Wahba, 1990]) e *soap-film smoothing* ([Wood et al., 2008]). Recentemente è stato proposto in [Ramsay, 2002] e [Sangalli et al., 2013] un modello di regressione spaziale con spline (che sarà anche chiamato modello SSR) che ha il vantaggio di poter ricostruire campi scalari anche su domini complessi, con forti concavità o tali da non essere semplicemente connessi. L'idea di avere un modello con queste proprietà è senza dubbio molto attraente per molti ambiti scientifici, perché spesso il dominio di interesse ha una forma molto complessa. In questo elaborato si vuole presentare una prima applicazione del metodo proposto in [Sangalli et al., 2013] a dati ambientali, in particolare a dati oceanografici. Più in dettaglio si considererà l'area del Golfo del Messico, perché combina una forma complessa delle coste con dei fenomeni oceanografici caratteristici, come la presenza della Corrente del Golfo. Si vuole inoltre confrontare i campi stimati con questa tecnica con quelli stimati con le tecniche di Kriging, mostrando i punti di forza e i punti deboli di ciascuna delle due tecniche. Nel capitolo 2, verranno descritte dal punto di vista teorico le tecniche tradizionali di Kriging, tra le più note nell'ambito dell'interpolazione spaziale e che saranno utilizzate come termine di confronto per il modello SSR. Dopo una presentazione dell'idea generale del Kriging, verrà introdotto il concetto di variogramma, verrà

spiegato come sceglierlo e verrà presentata la tecnica del Kriging ordinario. In seguito verrà presentato il modello SSR, sempre dal punto di vista teorico, mostrandone le proprietà e spiegando come stimare i parametri del modello nella versione non parametrica attraverso la tecnica degli elementi finiti.
Nel capitolo 3 saranno presentati i dataset di dati oceanografici utilizzati per le applicazioni dei capitoli successivi. Si mostrerà perché non sia possibile nel caso in specifico ricostruire un campo di temperatura della superficie oceanica basandosi su dati rilevati da boe e validarlo con i dati rilevati da satellite.
Infine nel capitolo 4 verranno mostrati alcuni risultati ottenuti utilizzando i dati di satellite: verranno confrontate le stime ottenute col modello SSR, nella sua versione non parametrica, con le stime delle tecniche di Kriging; verranno mostrate le caratteristiche che dovrebbe avere un dataset per consentire al modello SSR di fornire stime migliori e verrà mostrato come cambiano i risultati quando alcune di queste caratteristiche non sono soddisfatte.

Capitolo 2

Metodi di stima per dati spazialmente distribuiti

In questo capitolo vengono presentate le tecniche di interpolazione spaziale e di stima di campi scalari. Inizialmente verrà introdotta l'idea generale dei concetti teorici di base delle tecniche attualmente più utilizzate, le tecniche di Kriging. In seguito, dal paragrafo 2.7 verrà presentato il modello SSR dal punto di vista teorico e verrà spiegato come stimare i suoi parametri.

2.1 Idea generale delle tecniche di Kriging

Uno dei metodi più validi e più utilizzati per l'interpolazione spaziale sono senza dubbio le tecniche note come *Kriging*, dal nome di uno dei suoi ideatori, l'ingegnere minerario Danie Krige. Questi metodi sono nati dalla necessità di poter prevedere la distribuzione delle specie minerali nel sottosuolo avendo a disposizione solo alcuni campionamenti del terreno. Queste tecniche sono state poi estese ad altri ambiti che necessitassero la stima spaziale di una certa variabile incognita avendo a disposizione solamente alcuni campionamenti, come la ricostruzione di campi di temperatura o di concentrazioni di inquinanti atmosferici. L'obiettivo, quindi è quello di stimare un campo stocastico sulla base di alcune misurazioni puntuali.
Fondamento di questi metodi è il concetto di *variogramma* che, sotto opportune ipotesi, consente di poter individuare univocamente la struttura di covarianza del campo stocastico compatibile con la variabilità dei dati misurati. Se si conosce la struttura della covarianza, è possibile ricostruire il campo stocastico e valutarlo su una griglia di punti.
Nei prossimi paragrafi vengono presentate le definizioni che stanno alla base delle tecniche di Kriging; in seguito vengono presentate le metodologie di stima del variogramma e di validazione del modello. Per approfondire l'argomento, si consiglia di consultare [Cressie, 1991], [Kitanidis, 1992] e [Wackernagel, 1995], riferimenti sui quali sono basati i prossimi paragrafi.

2.2 Definizioni basilari sui campi stocastici

Il primo concetto da definire è quello di *campo stocastico*, l'oggetto che si vuole stimare con le tecniche di Kriging; indicando con Ω lo spazio degli eventi, $\mathfrak{F}$ una σ-algebra e $\mathbb{P}$ la relativa misura di probabilità:

Definizione 1 *Sia $(\Omega, \mathfrak{F}, \mathbb{P})$ uno spazio di probabilità; si definisce* ***campo stocastico*** *una funzione $Z = Z(\omega, \mathbf{x})$ che associa un valore reale ad ogni coppia $(\omega, \mathbf{x})$, dove $\omega \in \Omega$ è un evento (che da qui in poi verrà sottointeso) e $\mathbf{x} \in \mathbb{R}^d$.*

Si tratta quindi di una funzione sia a variabili reali che a variabili appartenenti ad uno spazio di probabilità. Il comportamento di un campo stocastico può essere allora completamente determinato sapendo calcolare la probabilità

$$\mathbb{P}\left[Z(\mathbf{x}_1) \in (a_1, b_1), \ldots, Z(\mathbf{x}_n) \in (a_n, b_n)\right]$$

dove $(a_i, b_i) \in \mathbb{R}$ è un intervallo arbitrario e n è il numero di punti $\mathbf{x}_i$ a disposizione. Nel caso di campo stocastico con distribuzioni di probabilità continue, è sufficiente determinare per ogni punto $\mathbf{x}_i$, $i = 1, \ldots, n$, la densità di probabilità $f_Z(\mathbf{u}) = f_{(Z(\mathbf{x}_1),\ldots,Z(\mathbf{x}_n))}(u_1, \ldots, u_n)$. Si possono allora definire media e varianza del campo, come:

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}\left[Z(\mathbf{x})\right] = \int_{-\infty}^{+\infty} u f_{Z(\mathbf{x})}(u) du \\ Var\left[Z(\mathbf{x})\right] &= \sigma_Z^2(\mathbf{x}) = \int_{-\infty}^{+\infty} (u - m(\mathbf{x}))^2 f_{Z(\mathbf{x}}(u) du \end{aligned} \tag{2.1}$$

Se esistono media e varianza, si può ricavare la covarianza del campo:

$$Cov\left[Z(\mathbf{x}), Z(\mathbf{y})\right] = \mathbb{E}\left[(Z(\mathbf{x}) - m(\mathbf{x}))(Z(\mathbf{y}) - m(\mathbf{y}))\right] \tag{2.2}$$

Inoltre, l'esistenza di media e varianza consente di definire il variogramma, la funzione che sta alla base delle tecniche di Kriging, e che descrive la variabilità del campo:

Definizione 2 *Il* ***variogramma*** *di un campo stocastico è definito come:*

$$Var\left[Z(\mathbf{x}) - Z(\mathbf{y})\right] \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d$$

Si indica inoltre con la funzione $\gamma(\mathbf{x}, \mathbf{y})$ il ***semivariogramma*** *del campo,*

$$\gamma(\mathbf{x}, \mathbf{y}) = \frac{1}{2} Var\left[Z(\mathbf{x}) - Z(\mathbf{y})\right] \qquad \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \tag{2.3}$$

Conoscendo il variogramma di un campo stocastico, si ha a disposizione tutta l'informazione necessaria per poter descrivere la sua variabilità in funzione della distanza tra due punti nello spazio.
Per evitare che sia necessaria una grande quantità di dati per la stima del variogramma, è necessario introdurre alcune restrizioni sulla natura del campo stocastico.

Definizione 3 *Un campo stocastico si dice* ***stazionario*** *se per ogni vettore* $\mathbf{h} \in \mathbb{R}^d$*:*

$$\begin{aligned}&\mathbb{P}\left[Z(\mathbf{x_1}+\mathbf{h}) \in (a_1,b_1), \ldots, Z(\mathbf{x_n}+\mathbf{h}) \in (a_n,b_n)\right]\\&= \mathbb{P}\left[Z(\mathbf{x}_1) \in (a_1,b_1), \ldots, Z(\mathbf{x}_n) \in (a_n,b_n)\right]\end{aligned}$$

In altre parole, se un campo è stazionario la sua distribuzione di probabilità è invariante per traslazioni. Da questo segue che i momenti dei singoli punti $\mathbf{x}$ sono costanti: $\mathbb{E}\left[Z(\mathbf{x})^k\right] = m_k \ \ \forall k \in \mathbb{N}$. Di conseguenza, se un campo è stazionario:

- esiste la media ed è invariante rispetto alla posizione $\mathbf{x}$:

$$\mathbb{E}\left[Z(\mathbf{x})\right] = m$$

- esistono la varianza e la funzione di covarianza; inoltre la covarianza non dipende dalla posizione di $\mathbf{x}$ e di $\mathbf{y}$, ma dalla differenza tra i punti $\mathbf{x}$ e $\mathbf{y}$:

$$Cov\left[Z(\mathbf{x}), Z(\mathbf{y})\right] = Cov\left[Z(\mathbf{x}-\mathbf{y})\right]$$

Per quanto riguarda il semivariogramma, vale la seguente definizione:

Definizione 4 *Un campo stocastico è detto* ***intrinsecamente stazionario*** *se il semivariogramma dipende solo dalla differenza tra i due punti* $\mathbf{x}$ *e* $\mathbf{y}$*:*

$$\gamma(\mathbf{x},\mathbf{y}) = \gamma(\mathbf{x}-\mathbf{y}) \tag{2.4}$$

Un campo stazionario è anche intrinsecamente stazionario; infatti, applicando le opportune semplificazioni sulle medie dovute all'invarianza per traslazioni, si ottiene:

$$\begin{aligned}\gamma(\mathbf{x},\mathbf{y}) &= \tfrac{1}{2}Var\left[Z(\mathbf{x})-Z(\mathbf{y})\right] =\\&= \tfrac{1}{2}\mathbb{E}\left[(Z(\mathbf{x})-Z(\mathbf{y}))^2\right] - \tfrac{1}{2}\left\{\mathbb{E}\left[(Z(\mathbf{x})-Z(\mathbf{y}))\right]\right\}^2\\&= \tfrac{1}{2}\left\{\mathbb{E}\left[(Z(\mathbf{x}-\mathbf{y}))^2\right]\right\} = \tfrac{1}{2}Var\left[Z(\mathbf{x}-\mathbf{y})\right]\\&= \gamma(\mathbf{x}-\mathbf{y})\end{aligned} \tag{2.5}$$

La funzione del semivariogramma può anche essere riscritta nel seguente modo:

$$\gamma(\mathbf{x},\mathbf{y}) = Var\left[Z(\mathbf{x})-Z(\mathbf{y})\right] = Var\left[Z(\mathbf{x})\right]+Var\left[Z(\mathbf{y})\right]-2Cov\left[Z(\mathbf{x}),Z(\mathbf{y})\right] \tag{2.6}$$

Nel caso di campo stocastico intrinsecamente stazionario, che ha momento secondo costante, allora, la (2.6) diventa:

$$\gamma(\mathbf{x},\mathbf{y}) = Var\left[Z(\mathbf{x})-Z(\mathbf{y})\right] = Var\left[Z(\mathbf{0})\right] + Var\left[Z(\mathbf{0})\right] - 2Cov\left[Z(\mathbf{x}),Z(\mathbf{y})\right]$$

Questa formulazione mette in risalto il legame tra il semivariogramma e la funzione di covarianza: di fatto, il variogramma descrive come è strutturata la covarianza e quindi la correlazione che c'è tra i diversi punti distribuiti nello spazio.

2.3 Teoremi di caratterizzazione e di rappresentazione

Nel paragrafo 2.6 si tratta il problema della stima del variogramma che, in generale, è affrontato mediante l'utilizzo dei dati a disposizione. Bisogna però tener presente che il variogramma appartiene ad una speciale classe di funzioni in quanto, ad esempio, non può assumere valori negativi essendo per definizione una funzione legata alla varianza, che è positiva. Risulta pertanto necessario identificare la classe di funzioni alla quale appartiene il variogramma e caratterizzarla. Anzitutto si introducono le funzioni condizionatamente definite negative:

Definizione 5 *Una funzione $\phi(\mathbf{x}, \mathbf{y})$ si dice* ***condizionatamente definita negativa*** *se dato $\mathbf{x}_i$, con $i = 1, \ldots, n$, $n \geq 2$, e dato un insieme di numeri reali α_i tali che*

$$\sum_{i=1}^{n} \alpha_i = 0$$

allora

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \phi(\mathbf{x}_i, \mathbf{x}_j) \leq 0$$

La classe delle funzioni condizionatamente definite negative può essere accostata al semivariogramma tramite il teorema:

Teorema 1 *Il semivariogramma di un campo stocastico intrinsecamente stazionario è una funzione condizionatamente definita negativa.*

Questo teorema, di cui si omette la dimostrazione, caratterizza, insieme al teorema seguente, la classe di funzioni a cui appartiene il semivariogramma:

Teorema 2 *Sia $\gamma(\cdot)$ una funzione continua su $\mathbb{R}^d$ tale che $\gamma(\mathbf{0}) = 0$. Le seguenti affermazioni sono equivalenti:*

- *$\gamma(\cdot)$ è condizionatamente definita negativa;*
- *per ogni costante $c > 0$, $e^{-c\gamma(\cdot)}$ è definita positiva;*
- *esistono una forma quadratica $Q(\cdot) \geq 0$ e una una misura positiva $G(\cdot)$, simmetrica, continua all'origine e che soddisfa $\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} (1 + ||\boldsymbol{\omega}||^2) G(d\boldsymbol{\omega}) < +\infty$ tale che:*

$$\gamma(\mathbf{h}) = Q(\mathbf{h}) + \int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \frac{1 - \cos(\boldsymbol{\omega}^T \mathbf{h})}{||\boldsymbol{\omega}||^2} G(d\boldsymbol{\omega})$$

Da questo teorema si ricava pertanto il seguente teorema di rappresentazione:

Teorema 3 ***(di Schoenberg-Yaglom)*** *Una funzione $\phi(\mathbf{x}, \mathbf{y})$ continua, condizionatamente definita negativa e tale che $\phi(\mathbf{x}, \mathbf{x}) = 0$ è il variogramma di un campo stocastico intrinsecamente stazionario.*

Una conseguenza di questo teorema è che la combinazione lineare $\gamma = \sum_{i=1}^{n} \alpha_i \gamma_i$ di un insieme di semivariogrammi γ_i e di coefficienti positivi α_i, con $i = 1, \ldots, n$, è il semivariogramma di un processo intrinsecamente stazionario.

2.4 Variogrammi teorici

Come detto nei paragrafi precedenti, il semivariogramma descrive la variabilità tra due punti $\mathbf{x}$ e $\mathbf{y}$. Si considera allora l'andamento del semivariogramma in funzione della distanza euclidea h tra questi due punti. I modelli teorici si distinguono tra loro sopratutto per il loro comportamento vicino all'orgine e all'infinito.
All'aumentare di h si possono verificare due situazioni:

- il semivariogramma raggiunge un valore soglia detto *sill*, alla distanza *range*;
- il semivariogramma aumenta indefinitamente.

Vicino all'origine, invece, può assumere:

- un comportamento lineare; in tal caso, il campo stocastico non è differenziabile nell'origine;
- un comportamento parabolico; in questo caso, invece, il campo stocastico è differenzabile nell'origine;
- il cosiddetto effetto pepita (o *nugget effect*), che consente al semivariogramma di assumere valore $\lim_{h\to 0}\gamma(h) = c_0$, con $c_0 \neq 0$; questo effetto, può sembrare in contrasto con la definizione di semivariogramma, che impone teoricamente che $\gamma(0) = 0$, ma permette di modellizzare quelle discontinuità vicino all'origine che possono verificarsi trattando dei dati reali, dovute soprattutto ad errori di misurazione e a piccole variazioni su microscala.

Queste caratteristiche sono riscontrabili nei principali modelli teorici di semivariogramma; tra questi si analizzanno quelli disponibili nel pacchetto *gstat* del software R: il modello sferico, il modello esponenziale, il modello gaussiano e il modello Matérn.

2.4.1 Modello sferico

Il modello sferico è descritto dalla seguente funzione:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_1\left(\frac{3h}{2r} - \frac{h^3}{2r^3}\right) & 0 < h \leq r \\ c_0 + c_1 & h \geq r \end{cases} \tag{2.7}$$

dove c_0, c_1, $r \geq 0$. In particolare, c_0 è il parametro che rappresenta la possibile discontinuità nell'orgine del modello (nugget), c_1 è il valore del sill nel caso in cui non ci siano discontinuità nell'origine e r è la distanza (range) a cui il semivariogramma raggiunge il valore di sill $c_0 + c_1$. Il modello sferico prevede quindi il raggiungimento di un valore soglia ed ha un andamento lineare vicino all'origine.

2.4.2 Modello esponenziale

Il modello esponenziale è definito dalla seguente funzione:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_1 \left(1 - e^{-\frac{h}{r}}\right) & h \neq 0 \end{cases} \tag{2.8}$$

dove c_0, c_1, $r \geq 0$ e hanno lo stesso significato del modello sferico, anche se il sill viene raggiunto solo asintoticamente; se $h = 3r$, il modello avrà raggiunto circa il 95% del sill.
Il modello esponenziale, come il modello sferico, prevede quindi il raggiungimento di un valore soglia e presenta un andamento lineare vicino all'origine.

2.4.3 Modello gaussiano

Il modello gaussiano è descritto dalla funzione:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_1 \left(1 - e^{-\frac{h^2}{r^2}}\right) & h \neq 0 \end{cases} \tag{2.9}$$

La funzione è molto simile a quella del modello esponenziale, con la differenza che la variabile h e il parametro r sono elevati al quadrato. Questa differenza rende il modello gaussiano parabolico vicino all'origine, mantenendo un valore asintotico di sill all'infinito. La distanza in cui viene raggiunto il 95% del sill è pari a $\sqrt{3}r$.

2.4.4 Modello Matérn

Il modello Matérn è definito tramite la relazione:

$$\gamma(h) = \begin{cases} 0 & h = 0 \\ c_0 + c_1 \left(1 - \frac{1}{2^{\nu-1}\Gamma(\nu)} \left(\frac{h}{r}\right)^{\nu} K_{\nu}\left(\frac{h}{r}\right)\right) & h \neq 0 \end{cases} \tag{2.10}$$

dove ν è un parametro di smoothing, $K_{\nu}(t) = \frac{\Gamma(r)}{2}\left(\frac{h}{2}\right)^{-\nu}$ è la *funzione di Bessel* e

$$\Gamma(r) = \int_0^{+\infty} t^{r-1} e^{-t} dt$$

è la funzione gamma, calcolabile facilmente come $\Gamma(r) = (r-1)!$ per $r \in \mathbb{N}$. Il modello Matérn generalizza i modelli esponenziale e gaussiano, che ne sono casi particolari: impostando il valore $\nu = \frac{1}{2}$ si ottiene il modello esponenziale, mentre per $\nu \to \infty$ si trova il modello gaussiano.

2.5 Kriging ordinario

Le tecniche di Kriging sono nate con l'obiettivo di predire il valore di una variabile in un sito di cui non si hanno informazioni, conoscendo invece i valori della variabile stessa in altri n siti spazialmente distribuiti. Esistono diverse

tipologie di Kriging: le principali sono note come *Kriging ordinario* e *Kriging universale*. In questo elaborato verrà presentata solo la tecnica del Kriging ordinario, essendo l'unica utilizzata come confronto con il modello SSR.
Un campo stocastico $Z(\mathbf{x})$, intrinsecamente stazionario con media costante m non nota, può essere modellizzato come:

$$Z(\mathbf{x}) = m + Z_R(\mathbf{x})$$

dove con $Z_R(\mathbf{x})$ si è indicato un campo stocastico a media nulla. Il semivariogramma è assunto noto e la sua stima sarà discussa nel prossimo paragrafo. Poiché si è assunta l'ipotesi di campo intrinsecamente stazionario, dalla formula (2.5) vale:

$$\gamma(\mathbf{x}, \mathbf{y}) = \frac{1}{2}\mathbb{E}\left[(Z(\mathbf{x}) - Z(\mathbf{y}))^2\right] = \frac{1}{2}\mathbb{E}\left[(Z_R(\mathbf{x}) - Z_R(\mathbf{y}))^2\right] \tag{2.11}$$

Si può allora definire il Kriging ordinario:

Definizione 6 *La stima del Kriging ordinario in un punto $\mathbf{x}_0$ nota $Z(\mathbf{x}_i)$, con $i = 1, \ldots, n$, è definita dallo stimatore lineare non distorto e con minimo errore quadratico medio:*

$$\hat{Z}(\mathbf{x}_0) = \sum_{i=1}^{n} \lambda_i Z(\mathbf{x}_i) \tag{2.12}$$

L'assunzione di stimatore non distorto è garantita ponendo il vincolo $\sum_{i=1}^{n} \lambda_i = 1$, infatti

$$\mathbb{E}\left[\hat{Z}(\mathbf{x}_0)\right] = \mathbb{E}\left[\sum_{i=1}^{n} \lambda_i Z(\mathbf{x}_i)\right] = \sum_{i=1}^{n} \lambda_i \mathbb{E}\left[Z(\mathbf{x}_i)\right] = m \sum_{i=1}^{n} \lambda_i$$

e la condizione di non distorsione $\mathbb{E}\left[\hat{Z}(\mathbf{x}_0)\right] = m$ è verificata solo se $\sum_{i=1}^{n} \lambda_i = 1$. Per garantire che sia a errore quadratico medio minimo bisogna minimizzare il suo MSE (*Mean Squared Error*): $\mathbb{E}\left[\left(Z(\mathbf{x}_0) - \hat{Z}(\mathbf{x}_0)\right)^2\right]$. Con la tecnica dei moltiplicatori di Lagrange, allora, si riduce il problema nella minimizzazione della funzione non vincolata $\phi(\lambda_1, \ldots, \lambda_n, \mu)$:

$$\phi(\lambda_1, \ldots, \lambda_n, \mu) = \mathbb{E}\left[\left(Z(\mathbf{x}_0) - \sum_{i=1}^{n} \lambda_i Z(\mathbf{x}_i)\right)^2\right] - 2\mu\left(\sum_{i=1}^{n} \lambda_i - 1\right) \tag{2.13}$$

Con alcune elaborazioni matematiche, il problema può essere riscritto come:

$$\begin{aligned}&\phi(\lambda_1, \ldots, \lambda_n, \mu) = \\ &= \mathbb{E}\left[\sum_{i=1}^{n} \lambda_i \left(Z(\mathbf{x}_0) - Z(\mathbf{x}_i)\right)^2\right] - \mathbb{E}\left[\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \left(Z(\mathbf{x}_i) - Z(\mathbf{x}_j)\right)^2\right] - 2\mu\left(\sum_{i=1}^{n} \lambda_i - 1\right)\end{aligned}$$

che diventa quindi:

$$\begin{aligned}
&\phi(\lambda_1, \ldots, \lambda_n, \mu) = \\
&= \sum_{i=1}^{n} \lambda_i \gamma(\mathbf{x}_0, \mathbf{x}_i) + \sum_{i=1}^{n} \lambda_i \left[\gamma(\mathbf{x}_0, \mathbf{x}_i) - \sum_{j=1}^{n} \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) \right] - 2\mu \left(\sum_{i=1}^{n} \lambda_i - 1 \right) \\
&= -\sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma(\mathbf{x}_i, \mathbf{x}_j) + 2 \sum_{i=1}^{n} \lambda_i \gamma(\mathbf{x}_0, \mathbf{x}_i) - 2\mu \left(\sum_{i=1}^{n} \lambda_i - 1 \right)
\end{aligned}$$

Imponendo che i gradienti della funzione $\phi(\lambda_1, \ldots, \lambda_n, \mu)$ siano nulli, si ottiene il sistema lineare:

$$\boldsymbol{\Gamma}_0 \boldsymbol{\lambda}_0 = \boldsymbol{\gamma}_0 \tag{2.14}$$

dove la matrice $\boldsymbol{\Gamma}_0$ di dimensione $(n+1) \times (n+1)$ è:

$$\boldsymbol{\Gamma}_0 = \begin{bmatrix}
\gamma(\mathbf{x}_1, \mathbf{x}_1) & \gamma(\mathbf{x}_1, \mathbf{x}_2) & \ldots & \gamma(\mathbf{x}_1, \mathbf{x}_n) & 1 \\
\gamma(\mathbf{x}_2, \mathbf{x}_1) & \gamma(\mathbf{x}_2, \mathbf{x}_2) & \ldots & \gamma(\mathbf{x}_2, \mathbf{x}_n) & 1 \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
\gamma(\mathbf{x}_n, \mathbf{x}_1) & \gamma(\mathbf{x}_n, \mathbf{x}_2) & \ldots & \gamma(\mathbf{x}_n, \mathbf{x}_n) & 1 \\
1 & 1 & \ldots & 1 & 0
\end{bmatrix}$$

e i vettori $\boldsymbol{\lambda}_0$ e $\boldsymbol{\gamma}_0$ sono:

$$\boldsymbol{\lambda}_0 = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \ldots \\ \lambda_n \\ \mu \end{bmatrix} \quad \boldsymbol{\gamma}_0 = \begin{bmatrix} \gamma(\mathbf{x}_0, \mathbf{x}_1) \\ \gamma(\mathbf{x}_0, \mathbf{x}_2) \\ \ldots \\ \gamma(\mathbf{x}_0, \mathbf{x}_n) \\ 1 \end{bmatrix}$$

Ne consegue che per trovare i coefficienti del Kriging ordinario ottimali, presenti nel vettore $\boldsymbol{\lambda}_0$, è sufficiente risolvere il sistema lineare:

$$\boldsymbol{\lambda}_0 = \boldsymbol{\Gamma}_0^{-1} \boldsymbol{\gamma}_0 \tag{2.15}$$

Il Kriging ordinario, però, presuppone la conoscenza del variogramma, che in genere invece è ignoto. È quindi necessaria una procedura di stima del variogramma in base ai dati a disposizione.

2.6 Stima del variogramma

Quando il variogramma è incognito, è necessario stimare un variogramma empirico, dal quale, con una procedura di fitting, sarà possibile ottenere il variogramma teorico. Si introduce allora un insieme finito di valori positivi h_k, con $k = 1, \ldots, K$; si suppone inoltre che questi valori siano ordinati in modo che $h_k < h_{k+1}$. L'idea è di costruire delle *classi di distanza* disgiunte che coprano lo spazio fino ad una distanza prefissata. Sia δ_k, $k = 1, \ldots, K$, un insieme di valori che rappresenta l'ampiezza desiderata delle K classi di distanza, queste allora si possono definire come:

$$\mathcal{N}(h_k) = \left\{ (\mathbf{x}_i, \mathbf{x}_j) : h_k - \frac{\delta_k}{2} \leq ||\mathbf{x}_i - \mathbf{x}_j|| \leq h_k + \frac{\delta_k}{2} \right\} \tag{2.16}$$

dove i punti $\mathbf{x}_i$, $i = 1, \ldots, n$, sono i punti disponibili. In altre parole, la classe $\mathcal{N}(h_k)$ contiene tutte le coppie di punti che si trovano distanti tra loro per un valore compreso nell'intervallo $[h_k - \frac{\delta_k}{2}, h_k + \frac{\delta_k}{2}]$. Si indica con $|\mathcal{N}(h_k)|$ la cardinalità della classe $\mathcal{N}(h_k)$, ovvero il numero di coppie della classe stessa; empiricamente è auspicabile che ogni classe abbia una numerosità maggiore di 30 coppie per poter garantire significatività alla stima del variogramma.
Gli stimatori principali del variogramma sono due: lo stimatore di Matheron, detto anche stimatore dei momenti, e lo stimatore di Cressie, conosciuto anche come stimatore robusto.

2.6.1 Stimatori di Matheron e di Cressie

Il primo stimatore che è stato proposto è lo stimatore di Matheron, che prende il nome dal suo ideatore e che è stato presentato per la prima volta nel 1962. È detto anche stimatore dei momenti, perché è basato sul metodo dei momenti, ed è formulato come segue:

$$\hat{\gamma}^M(h_k) = \frac{1}{2|\mathcal{N}(h_k)|} \sum_{\mathcal{N}(h_k)} (Z(\mathbf{x}_i) - Z(\mathbf{x}_j))^2 \tag{2.17}$$

Si tratta di uno stimatore ampiamente utilizzato in letteratura, anche se ha il difetto di essere sensibile alla presenza di outliers. Per questo motivo, è stato studiato uno stimatore robusto, lo stimatore di Cressie.
Questo secondo stimatore è stato introdotto da Cressie e Hawkins nel 1980. La sua definizione è la seguente:

$$\hat{\gamma}^C(h_k) = \frac{1}{2\left(0,457 + \frac{0,494}{|\mathcal{N}(h_k)|}\right)} \left(\frac{1}{|\mathcal{N}(h_k)|} \sum_{\mathcal{N}(h_k)} |Z(\mathbf{x}_i) - Z(\mathbf{x}_j)|^{\frac{1}{2}} \right)^4 \tag{2.18}$$

Questo tipo di formulazione consente allo stimatore di essere molto meno sensibile alla presenza di outliers. Proprio per questo motivo si tratta di uno degli stimatori più utilizzati per la stima del variogramma.

2.6.2 Fitting del variogramma col metodo dei minimi quadrati

Si assume che il vero variogramma del campo stocastico appartenga ad una famiglia parametrica di variogrammi validi, come quelli mostrati nel paragrafo 2.4. Per poter scegliere il variogramma teorico che meglio riesca ad approssimare quello vero si stima il variogramma empirico e in seguito lo si confronta con i variogrammi teorici per poter determinare quale di questi riesca ad approssimarlo meglio. Si denoti con $\hat{\gamma}^{\#}(h)$ uno degli stimatori definiti nel paragrafo 2.6.1 e con $\gamma(h, \boldsymbol{\theta})$ un variogramma teorico, dipendente dalla distanza h e dal vettore $\boldsymbol{\theta}$ contenente i parametri del variogramma (ad esempio, per il modello sferico: $\boldsymbol{\theta} = (c_0, c_1, r)^T$). Per stimare il vettore $\boldsymbol{\theta}$ si può utilizzare il cosiddetto *metodo*

ordinario dei minimi quadrati, che consiste nel trovare il valore di $\boldsymbol{\theta}$ che minimizza il funzionale:

$$\sum_{k=1}^{K} \left[\hat{\gamma}^{\#}(h_k) - \gamma(h_k, \boldsymbol{\theta})\right]^2 \tag{2.19}$$

Il problema di questo metodo è che non tiene conto della distribuzione dello specifico stimatore $\hat{\gamma}^{\#}(h)$ che si sta usando, ma svolge solo un fitting puramente geometrico. Considerando il vettore $2\boldsymbol{\gamma}^{\#} = \left(2\gamma^{\#}(h_1), \ldots, 2\gamma^{\#}(h_K)\right)^T$ e indicando con $\mathbf{V}$ la sua matrice di covarianza, $\mathbf{V} = Var\left[2\boldsymbol{\gamma}^{\#}\right]$, si determina il parametro $\boldsymbol{\theta}$ col *metodo dei minimi quadrati generalizzati*, che consiste nel trovare il valore ottimo di $\boldsymbol{\theta}$ tale che sia minimizzato il funzionale

$$\left(2\boldsymbol{\gamma}^{\#} - 2\boldsymbol{\gamma}(\boldsymbol{\theta})\right)^T \mathbf{V}^{-1} \left(2\boldsymbol{\gamma}^{\#} - 2\boldsymbol{\gamma}(\boldsymbol{\theta})\right) \tag{2.20}$$

avendo indicato con $2\boldsymbol{\gamma}(\boldsymbol{\theta}) = (2\gamma(h_1, \boldsymbol{\theta}), \ldots, 2\gamma(h_K, \boldsymbol{\theta}))^T$.
Il metodo dei minimi quadrati generalizzati non richiede particolari assunzioni sulla distribuzione dei dati, poiché necessita solo dei momenti del second'ordine; tuttavia il calcolo di $\mathbf{V}$ e conseguentemente la minimizzazione del funzionale (2.20) possono risultare difficoltosi dal punto di vista computazionale. Esistono però delle tecniche di approssimazione che consentono di poter applicare il metodo dei minimi quadrati generalizzati in modo efficiente e mantenendone la struttura e l'idea generale.
Nei prossimi paragrafi verrà presentato il modello SSR, un'alternativa alle tecniche di Kriging per la determinazione di campi scalari.

2.7 Modello di regressione spaziale con funzione spline

Si consideri un dominio regolare $\Omega \subset \mathbb{R}^2$ e un insieme $\{\mathbf{p}_i = (x_i, y_i); i = 1, \ldots, n)$ di n punti distribuiti in esso. In ogni sito $\mathbf{p}_i$ è osservato il valore della variabile z_i. L'obiettivo della procedura che si vuole introdurre è quello di stimare un campo spaziale definito su tutto il dominio di cui si suppone che le z_i siano i valori. Supponendo di avere a disposizione altre q variabili osservabili nei punti $\mathbf{p}_i$, sia $\mathbf{W}$ la matrice che le raccoglie, detta matrice delle covariate (o matrice disegno):

$$\mathbf{W} = \begin{bmatrix} w_{11} & \ldots & w_{1q} \\ \vdots & & \vdots \\ w_{n1} & \ldots & w_{nq} \end{bmatrix}$$

e sia $\mathbf{w}_i^T$ l'i-esima riga della matrice $\mathbf{W}$. Il modello di regressione spaziale con funzione spline, che nel resto dell'elaborato sarà anche chiamato modello SSR (*Spatial Spline Regression*), per questi dati si può quindi scrivere come:

$$z_i = \mathbf{w}_i^T \boldsymbol{\beta} + f(\mathbf{p}_i) + \epsilon_i \qquad i = 1, \ldots, n \tag{2.21}$$

dove $\epsilon_i \sim N(0, \sigma^2)$ sono i residui, indipendentemente distribuiti tra loro, $\boldsymbol{\beta} \in \mathbb{R}^q$ è il vettore dei coefficienti di regressione e la funzione f è a valori reali e due volte differenziabile. È auspicabile che il campo da stimare con questo modello abbia un errore di stima minimo e che sia sufficientemente regolare. Pertanto, per la determinazione delle incognite $\boldsymbol{\beta}$ e f, si minimizza il funzionale $J_\lambda(\boldsymbol{\beta}, f)$ così definito:

$$J_\lambda(\boldsymbol{\beta}, f) = \sum_{i=1}^{n} \left(z_i - \mathbf{w}_i^T \boldsymbol{\beta} - f(\mathbf{p}_i)\right)^2 + \lambda \int_\Omega (\Delta f)^2 d\Omega \tag{2.22}$$

In $J_\lambda(\boldsymbol{\beta}, f)$ il primo termine è l'errore quadratico medio del modello. Il secondo termine, invece, è composto da un nuovo parametro, λ, e dall'integrale sul dominio Ω del laplaciano di f al quadrato,

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}$$

scelto perché, rispetto ad altre misure di curvatura locale di f, è invariante in caso di trasformazioni euclidee delle coordinate spaziali e quindi è in grado di garantire che la regolarità della superficie stimata non dipenda dal sistema di coordinate. La variabile λ è introdotta proprio per determinare l'importanza della regolarità del campo stimato: infatti, se λ assume un valore molto alto la funzione f è obbligata ad essere molto liscia, mentre se ha un valore basso consente ad f di avere maggiore curvatura locale.
Poiché nei capitoli successivi non si prevede l'uso di covariate, per la mancanza di un numero sufficiente di dati, il modello 2.21 preso in considerazione è quello costituito solo dalla parte non parametrica,

$$z_i = f(\mathbf{p}_i) + \epsilon_i \qquad i = 1, \ldots, n \tag{2.23}$$

Il funzionale da minimizzare diventa allora

$$J_\lambda(f) = \sum_{i=1}^{n} (z_i - f(\mathbf{p}_i))^2 + \lambda \int_\Omega (\Delta f)^2 d\Omega \tag{2.24}$$

Il problema di minimizzazione del funzionale $J_\lambda(f)$ non è banale, ma può essere affrontato con la tecnica nota come *metodo degli elementi finiti*, una delle più efficaci per la risoluzione numerica delle equazioni differenziali alle derivate parziali. Attraverso questo metodo, infatti, è possibile approssimare il funzionale $J_\lambda(f)$ con un funzionale i cui argomenti siano finito-dimensionali, rendendo più semplice la ricerca del minimo. L'appendice A è dedicata ad una trattazione basilare di questa tecnica in modo da consentire una migliore comprensione del prossimo paragrafo, nel quale verrà mostrato come stimare il parametro f del modello SSR (2.23).

2.8 Determinazione dei parametri del modello

In questo paragrafo si analizza il problema di minimizzazione (2.24) seguendo passaggi analoghi a quelli che hanno portato alla risoluzione del problema di Poisson (A.2) in appendice A.
Si considera il vettore $\mathbf{z} = (z_1, \ldots, z_n)^T$ che raggruppa i valori della variabile z e si introduce il vettore delle valutazioni di f negli n siti di coordinate spaziali $\mathbf{p}$, così definito: $\mathbf{f}_n = (f(\mathbf{p}_1), \ldots, f(\mathbf{p}_n))^T$. Si possono cercare le stime di $f \in H_0^2(\Omega)$, confidando in seguito di poter rilassare queste condizioni. Il problema di stima è caratterizzato dalla seguente proposizione

Proposizione 1 *Si consideri $f \in H_0^2(\Omega)$. Lo stimatore $\hat{f}$ che minimizza $J(f)$ del problema (2.24) è $\hat{f}$ tale che sia soddisfatta la relazione:*

$$\mathbf{u}_n^T \hat{\mathbf{f}}_n + \lambda \int_\Omega (\Delta u)(\Delta \hat{f}) = \mathbf{u}_n^T \mathbf{z} \quad \forall u \in H_0^2(\Omega) \tag{2.25}$$

Sotto queste ipotesi, inoltre, lo stimatore $\hat{f}$ è univocamente determinato.

Per la dimostrazione si rimanda a [Sangalli et al., 2013], dove è possibile trovare una versione più ampia della proposizione, che comprende il caso di modello SSR semi-parametrico. Per risolvere il problema (2.25) si fa ricorso al metodo degli elementi finiti. Come per il problema di Poisson nel paragrafo A.2, è necessario riformulare il problema in forma debole. Si introducono perciò la funzione ausiliaria $g \in L^2(\Omega)$ e la funzione test $v \in L^2(\Omega)$. Il problema (2.25) viene riscritto come:

$$\begin{aligned} \mathbf{u}_n^T \hat{\mathbf{f}}_n + \lambda \int_\Omega g(\Delta u) &= \mathbf{u}_n^T \mathbf{z} \qquad \forall u \in H_0^2(\Omega) \\ \int_\Omega gv - \int_\Omega (\Delta \hat{f}) v &= 0 \qquad \forall v \in L^2(\Omega) \end{aligned}$$

Se infatti $\hat{f}$ e g risolvono questo problema $\forall\, u, v$, allora $\hat{f}$ risolve il problema (2.25). Se si utilizza il teorema della divergenza come nel paragrafo A.2 per la

riformulazione di (A.4) in (A.7), si ottiene la seguente formulazione:

$$\begin{aligned}\mathbf{u}_n^T\hat{\mathbf{f}}_n - \lambda\int_\Omega(\nabla u\cdot\nabla g) &= \mathbf{u}_n^T\mathbf{z}\\ \int_\Omega vg - \int_\Omega(\nabla v\cdot\nabla\hat{f}) &= 0\end{aligned}$$

Con questa nuova formulazione è sufficiente imporre che le funzioni $\hat{f}$ e u appartengano allo spazio $(H_0^1(\Omega)\cap C^0(\Omega))$ e che la funzione ausiliaria g e la funzione test v appartengano allo spazio $H^1(\Omega)$, condizione rilassata del tutto simile a quella richiesta nel problema di Poisson. Il problema (2.25) viene quindi riformulato nella formulazione debole come:

trovare $\hat{f}, g \in H_0^1(\Omega)\cap C^0(\Omega)$ tali che $\forall u, v \in H^1(\Omega)$

$$\begin{aligned}\mathbf{u}_n^T\hat{\mathbf{f}}_n - \lambda\int_\Omega(\nabla u\cdot\nabla g) &= \mathbf{u}_n^T\mathbf{z}\\ \int_\Omega vg - \int_\Omega(\nabla v\cdot\nabla\hat{f}) &= 0\end{aligned} \tag{2.26}$$

La forma debole del problema ben si presta all'applicazione del metodo degli elementi finiti. Come per il problema di Poisson, si cerca di discretizzare il dominio e di formulare un problema approssimato risolvibile con un sistema lineare. Si suddivide il dominio Ω in una partizione τ_h di triangoli ottenendo la triangolazione del dominio in elementi K, tali che $\overline{\Omega} = \bigcup_{K\in\tau_h} K$. Per l'approssimazione di $H_0^1(\Omega)$ si considerano gli spazi degli elementi finiti $\mathring{\mathrm{X}}_h^1$ e $\mathring{\mathrm{X}}_h^2$; si richiamano le definizioni degli spazi $\mathring{\mathrm{X}}_h^r$, presenti in appendice A: se P_r è lo spazio dei polinomi di grado r,

$$X_h^r = \{v_h \in C^0(\overline{\Omega}) : v_h \in P_r, \forall K \in \tau_h\}$$

composto da funzioni globalmente continue, polinomiali di grado r in ogni elemento della triangolazione τ_h e

$$\mathring{\mathrm{X}}_h^r = \{v_h \in X_h^r : v_h|_{\partial\Omega} = 0\}$$

Se si sceglie $\mathring{\mathrm{X}}_h^1$, servirà conoscere il valore della funzione f in 3 nodi per ogni elemento (che saranno in particolare i vertici dei triangoli); se invece si sceglie $\mathring{\mathrm{X}}_h^2$ allora serviranno 6 nodi (i vertici dei triangoli e i punti medi dei lati). Indicando con $\boldsymbol{\xi}_j$, $j = 1,\ldots,k$, il vettore dei nodi di un elemento e con $\boldsymbol{\varphi} = (\varphi_1,\ldots,\varphi_k) \in \mathring{\mathrm{X}}_h^1$ l'insieme delle funzioni tali che

$$\varphi_i(\boldsymbol{\xi}_j) = \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \quad i,j = 1,\ldots,k$$

la funzione $f(x,y)$ può essere completamente definita in ogni punto secondo la formula:

$$f(x,y) = \sum_{i=1}^{k} f(\boldsymbol{\xi}_i)\varphi_i(x,y) = \mathbf{f}^T\boldsymbol{\varphi}(x,y)$$

Di seguito vengono definiti gli strumenti che serviranno per un corollario che risolve il problema della stima di $\hat{f}$. Sia $\Omega_\tau = \bigcup_{K \in \tau_h} K$ il dominio partizionato dagli elementi K; definendo i vettori $\boldsymbol{\varphi}_x = (\frac{\partial \varphi_1}{\partial x}, \ldots, \frac{\partial \varphi_k}{\partial x})^T$ e $\boldsymbol{\varphi}_y = (\frac{\partial \varphi_1}{\partial y}, \ldots, \frac{\partial \varphi_k}{\partial y})^T$, si ricavano le matrici $\mathbf{R}_0$ ed $\mathbf{R}_1$:

$$\mathbf{R}_0 = \int_{\Omega_\tau} (\boldsymbol{\varphi}\boldsymbol{\varphi}^T)$$
$$\mathbf{R}_1 = \int_{\Omega_\tau} (\boldsymbol{\varphi}_x\boldsymbol{\varphi}_x^T + \boldsymbol{\varphi}_y\boldsymbol{\varphi}_y^T)$$

Sia $\mathbf{0}$ il vettore nullo e $\mathbf{O}_{m\times l}$ la matrice $m \times l$ identicamente nulla. Si definisce la matrice a blocchi $\mathbf{L}$:

$$\mathbf{L} = \left[\begin{array}{c|c} \mathbf{I}_n & \mathbf{O}_{n\times(k-n)} \\ \hline \mathbf{O}_{(k-n)\times n} & \mathbf{O}_{(k-n)\times(k-n)} \end{array}\right]$$

Vale il seguente corollario:

Corollario 1 *Lo stimatore $\hat{f} \in \mathring{X}_h^r$ che risolve il problema (2.24) discretizzato è $\hat{f}$ identificato dai coefficienti del vettore $\hat{\mathbf{f}}$ che risolve:*

$$\begin{bmatrix} -\mathbf{L} & \lambda\mathbf{R}_1 \\ \lambda\mathbf{R}_1 & \lambda\mathbf{R}_0 \end{bmatrix} \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} = \begin{bmatrix} -\mathbf{I_n z} \\ \mathbf{0} \end{bmatrix} \tag{2.27}$$

Inoltre lo stimatore $\hat{f}$ è univocamente determinato.

La dimostrazione del corollario nella versione più completa che comprende il caso di modello semi-parametrico si può trovare in [Sangalli et al., 2013]. Come per l'equazione di Poisson, il problema è stato approssimato e richiede ora la risoluzione di due sistemi lineari. Il sistema (2.27) ha il vantaggio di essere sparso, perché il prodotto tra le basi $\boldsymbol{\varphi}$ e il prodotto delle derivate di $\boldsymbol{\varphi}$ ha parecchi elementi nulli e questo caratterizza le matrici $\mathbf{R}_0$ e $\mathbf{R}_1$: questo aspetto rende il sistema (2.27) risolvibile abbastanza velocemente, potendo fare affidamento sulle tecniche numeriche adatte alla risoluzione di problemi sparsi.

2.9 Proprietà degli stimatori

Per usare gli stimatori proposti nel corollario 1, è opportuno analizzare le loro proprietà statistiche. Inoltre devono ancora essere stimati il parametro λ e la varianza dei residui σ^2. Sia $\mathbf{B}$ la matrice $2k \times 2k$ definita nel sistema lineare (2.27):

$$\mathbf{B} = \begin{bmatrix} -\mathbf{L} & \lambda\mathbf{R}_1 \\ \lambda\mathbf{R}_1 & \lambda\mathbf{R}_0 \end{bmatrix}$$

e sia $\mathbf{A} = -\mathbf{B}^{-1}$. Indicando con $\mathbf{A}_n$ la matrice quadrata formata dalle prime n righe e n colonne di $\mathbf{A}$ e con $\mathbf{A}_{kn}$ la matrice $k \times n$ formata dalle prime k righe e n colonne di $\mathbf{A}$, tenendo presente che per definizione $\hat{\mathbf{f}}$ è un vettore di dimensione $k > n$, il sistema lineare si può scrivere come:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix} = -\mathbf{A} \begin{bmatrix} -\mathbf{I}_n\mathbf{z} \\ \mathbf{0} \end{bmatrix}$$

quindi

$$\hat{\mathbf{f}}_n = -\mathbf{A}_n(-\mathbf{I}_n\mathbf{z}) = \mathbf{A}_n\mathbf{z} \tag{2.28}$$

e analogamente si ricava che

$$\hat{\mathbf{f}} = \mathbf{A}_{kn}\mathbf{z} = \mathbf{A}_{kn}\mathbf{A}_n^{-1}\hat{\mathbf{f}}_n \tag{2.29}$$

Questa equazione evidenzia il fatto che la soluzione agli elementi finiti per $\hat{\mathbf{f}}$, quindi per tutti i k nodi in cui è valutata la funzione f, è completamente determinata dalla soluzione $\hat{\mathbf{f}}_n$, relativa invece ai soli nodi degli n dati. Grazie a questa proprietà è possibile utilizzare gli strumenti della statistica classica. Si calcolano allora la media e la varianza dello stimatore $\hat{\mathbf{f}}_n$. Sapendo che:

$$\begin{aligned}\mathbb{E}[\mathbf{z}] &= \mathbf{f}_n \\ Var[\mathbf{z}] &= \sigma^2\mathbf{I}_n\end{aligned}$$

e poiché $\mathbf{A}_n$ è simmetrica, si ottengono:

$$\begin{aligned}\mathbb{E}\left[\hat{\mathbf{f}}_n\right] &= \mathbf{A}_n\mathbb{E}[\mathbf{z}] \\ &= \mathbf{A}_n\mathbf{f}_n \\ Var\left[\hat{\mathbf{f}}_n\right] &= \mathbf{A}_n\mathbf{A}_n^T Var[\mathbf{z}] \\ &= \sigma^2\mathbf{A}_n\mathbf{A}_n^T\end{aligned} \tag{2.30}$$

Si possono fare alcune considerazioni a partire dal vettore $\hat{\mathbf{z}}$, vettore dei valori stimati dal modello in corrispondenza degli n punti, che è dato dalla seguente espressione:

$$\hat{\mathbf{z}} = \hat{\mathbf{f}}_n = \mathbf{A}_n\mathbf{z}$$

È evidente come la matrice $\mathbf{A}_n$ abbia il ruolo di matrice di regolarizzazione (o di *smoothing*): questa espressione, infatti, sottolinea come la stima dei valori $\hat{\mathbf{z}}$ del modello dipenda dai valori reali osservati $\mathbf{z}$ filtrati attraverso l'operatore lineare $\mathbf{A}_n$, indipendente da $\mathbf{z}$.
Avendo $\hat{\mathbf{z}}$, si può calcolare la stima della varianza σ^2 dei residui:

$$\hat{\sigma}^2 = \frac{1}{n - tr(\mathbf{A}_n)}(\mathbf{z} - \hat{\mathbf{z}})^T(\mathbf{z} - \hat{\mathbf{z}})$$

essendo $tr(\mathbf{A}_n) = \sum_{i=1}^{n} a_{ii}$ la traccia della matrice $\mathbf{A}_n$, dove con a_{ii} si sono indicati gli elementi sulla diagonale di $\mathbf{A}_n$; dato che lo stimatore è lineare, infatti, $tr(\mathbf{A}_n)$ è una misura dei gradi di libertà del modello.
La scelta del parametro λ può essere effettuata tramite diversi criteri di selezione del modello: in questo elaborato, come in [Sangalli et al., 2013], si sceglie la tecnica del *Generalized-Cross-Validation* (GCV), molto usata in statistica per la scelta di un parametro a partire dai dati a disposizione. L'idea di base di questo criterio è la stessa del *Leave-One-Out Cross-Validation*, dove si considera il valore

stimato dal modello a partire da $n-1$ dati e si cerca di valutare quanto bene riesca a predire l'n-esimo dato. Nella pratica col criterio GCV si cerca $\hat{\lambda}$ tale

$$\hat{\lambda} = \arg\min_{\lambda} \frac{1}{n[1 - tr(\mathbf{A}_n(\lambda))/n]^2} [\mathbf{z} - \hat{\mathbf{z}}(\lambda)]^T [\mathbf{z} - \hat{\mathbf{z}}(\lambda)]$$

Nel caso in cui si voglia predire dal modello la variabile z_{n+1} in un nuovo sito $\mathbf{p}_{n+1}$, si ricava l'espressione

$$\hat{z}_{n+1} = \hat{f}(\mathbf{p}_{n+1}) = \hat{\mathbf{f}}^T \boldsymbol{\varphi}(\mathbf{p}_{n+1})$$

Le potenzialità del modello vanno oltre le proprietà appena descritte. In molti problemi di interpolazione spaziale, infatti, si conoscono le condizioni al contorno del dominio, spesso perché imposte da leggi fisiche, e si vorrebbe poter inserire queste informazioni nel modello, perché la superficie stimata abbia le caratteristiche desiderate. Nel caso si abbiano informazioni sulla soluzione al bordo, si imposterà un problema di Dirichlet, mentre nel caso si abbiano informazioni sulla derivata normale al bordo si tratterà con un problema di Neumann, utilizzando il metodo degli elementi finiti. Si supponga di essere nel primo caso, e di conoscere $f = f_{\partial\Omega}$ sul bordo $\partial\Omega$, con $f_{\partial\Omega}$ sufficientemente regolare. Per questi casi è possibile ricavare una proposizione analoga alla proposizione 1 e il suo corrispondente corollario; nel caso in cui si voglia approfondire questo aspetto, si rimanda a [Sangalli et al., 2013].

Capitolo 3

Descrizione e analisi dei dataset oceanografici

In questo capitolo vengono presentati i dataset oceanografici utilizzati per le applicazioni. Durante il reperimento dei dati si è potuto notare come vi fossero dati riferiti alla stessa variabile ma ottenuti con misurazioni diverse, di fatto consentendo di avere due dataset differenti ma relativi alla stessa variabile; questo ha suggerito di calcolare le stime con il modello SSR utilizzando uno dei due dataset e di validarle con l'altro dataset. Purtroppo, come si mostrato nel corso di questo capitolo, le caratteristiche dei dataset non hanno consentito di seguire questa proceura.
Vengono in seguito descritti i dataset a disposizione e le analisi su di essi eseguite.

3.1 Provenienza e descrizione dei dati

In questo elaborato, lo studio e la validazione del modello SSR in ambito ambientale viene svolto con dati oceanografici. L'oceano, infatti, è fisicamente delimitato dalle coste e quindi è racchiuso in un dominio di forma complessa; con dati oceanografici è possibile studiare il comportamento e l'efficacia del modello sfruttando al meglio le sue proprietà, sapendo anche che la forma del dominio potrebbe essere causa di errore per altre tecniche. Proprio per questo motivo non sono stati considerati dati di tipo atmosferico, seppur disponibili in maggior quantità, perché sono grandezze che non subiscono grandi cambiamenti a causa di vincoli di dominio: si pensi, ad esempio, alla pressione atmosferica, che è una variabile distribuita nell'atmosfera e che quindi non risente della presenza di particolari vincoli spaziali. Tra i dati oceanografici, in particolare, ci si concentra sullo studio della temperatura della superficie dell'oceano.
Il NOAA (*National Oceanic and Atmospheric Administration*) è un'agenzia scientifica americana appartenente al Dipartimento del Commercio degli Stati Uniti e si occupa di ricerche in ambito meteorologico, oceanografico e di gestione del pescato. Dal sito del NOAA (http://www.noaa.gov/) sono presenti diversi database contenenti rilevazioni meteorologiche; essendo un ente molto importante, utilizza diversi sistemi di monitoraggio e rilevazione di dati: ad esempio è possibile trovare rilevazioni oceanografiche ottenute con boe oceaniche,

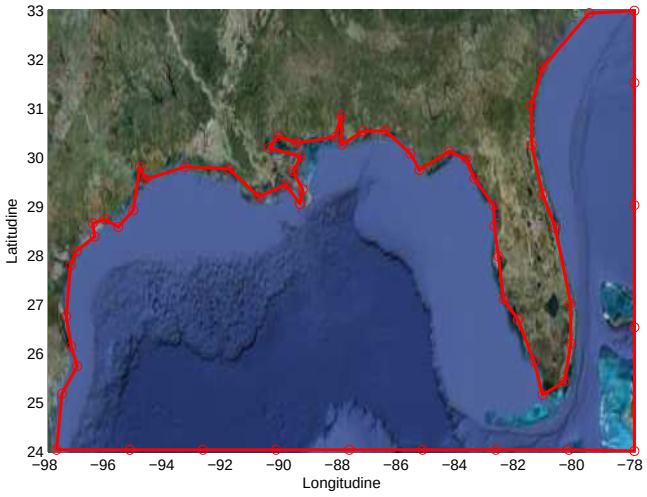


Figura 3.1: Costa del Golfo del Messico ricostruita con Matlab e posizione dei nodi con cui è costruito il bordo (con i cerchi)[1]. Le coordinate dei nodi provengono del database dell'agenzia NOAA.

con sensori installati su navi e con satelliti. Questa varietà di rilevazioni è un'ottima risorsa per lo studio del modello e per la sua validazione: si considerano pertanto le rilevazioni effettuate da boe fisse e le rilevazioni ottenute da satellite. Si cercano dati nella regione del Golfo del Messico: questa zona è caratterizzata da una conformazione costiera particolare e complessa e dalla presenza della corrente calda nota come Corrente del Golfo, che qui nasce e migra verso le regioni settentrionali dell'Europa. Si tratta di un'area strategica dal punto di vista economico e meteorologico, per via dei giacimenti petroliferi sottomarini e per il fatto che proprio lì si forma la maggior parte degli uragani che investono ogni anno gli Stati Uniti; per questi motivi è una regione ben monitorata, che vanta la presenza di un buon numero di boe fisse. Ci si concentra quindi sulla superficie oceanica compresa tra 98°W e 78°W di longitudine e tra 24°N e 33°N di latitudine.
Per caratterizzare il dominio si fa ricorso ad un dataset contenente una serie di coppie di valori di latitudine e longitudine che individuano dei punti su un piano; unendo questi punti, si può ottenere una ricostruzione delle coste. Dataset di questo tipo sono disponibili sul sito del *NOAA* nella sezione dedicata al NGDC (*National Geophysical Data Center*, http://www.ngdc.noaa.gov/mgg_coastline/) in diverse scale di definizione e in diversi formati tra cui il formato compatibile con Matlab. Si considera la scala di definizione più bassa (1:5000000), si individuano i punti relativi all'area geografica desiderata e si eliminano alcuni di questi punti per poter ridurre la complessità del bordo continentale, così da renderlo compatibile con la numerosità dei dati che si avranno a disposizione; si ottiene quindi il bordo del dominio mostrato in figura 3.1, individuato da 63 coppie di latitudine e longitudine.
Dal sito del NGCD (*National Geophysical Data Center*) è possibile trovare anche

[1]Immagine sullo sfondo proveniente da https://maps.google.it/

Figura 3.2: Tre esempi di boe fisse: a sinistra una boa ancorata, al centro un rilevatore meteorologico nei pressi della costa e a destra una struttura estrattiva

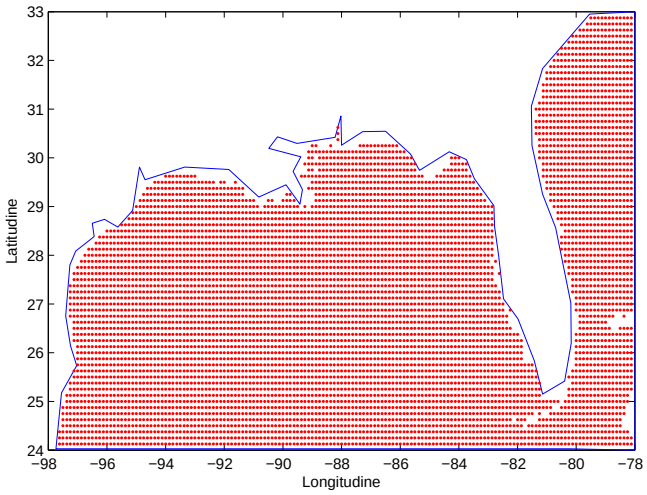


Figura 3.3: Griglia dei punti misurati dal satellite alle ore 22:00 GMT del 15 aprile 2013 nel Golfo del Messico

rilevazioni della temperatura della superficie oceanica del continente Nord Americano rilevate dai satelliti in orbita per conto del *NOAA* ed elaborate ogni 48 ore [2]. Impostando la data di lunedì 15 aprile 2013 si ottiene un dataset rilevato alle ore 22:00 GMT. Dalla figura 3.3 si può notare come la griglia di rilevazioni sia molto fitta, anche se non molto dettagliata nei pressi dell'area costiera.
Per quanto riguarda i dati di boa, dal sito del NDBC (*National Data Buoy Center*, http://www.ndbc.noaa.gov/), dipartimento del *NOAA*, è possibile accedere ad un archivio di dati rilevati da boe che sono distribuite tra tutti i mari e gli oceani. Bisogna però precisare cosa si intende per boe fisse (*moored buoys*), dal momento che questo termine generico sottointende diverse tipologie di rilevatori: le boe ancorate, i rilevatori meteorologici situati nella prossimità della costa e le strutture private estrattive, site nell'oceano, e dotate di sensori meteorologici (figura 3.2)[3].
Poiché i dati di satellite sono rilevati in un orario preciso, le 22:00 GMT, le rilevazioni di boa vengono selezionate tra le boe attive alle ore 22:00 GMT del 15

[2]Dati reperibili da: http://www.class.ngdc.noaa.gov/saa/products/search?datatype_family=SST14NA
[3]Immagini provenienti dal sito http://www.ndbc.noaa.gov/

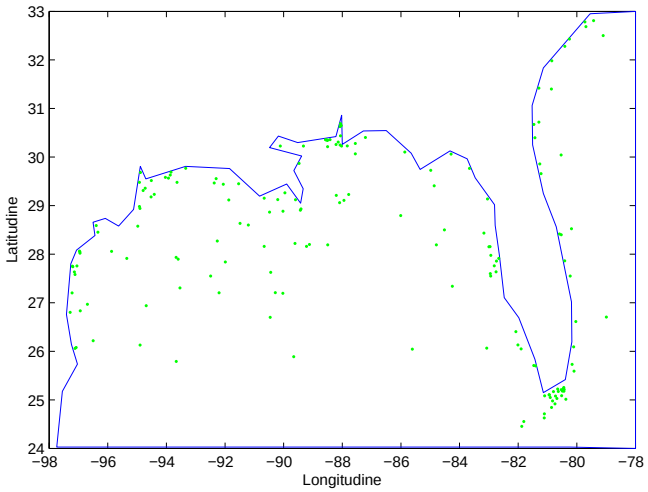


Figura 3.4: Distribuzione delle boe attive alle ore 22:00 GMT del 15 aprile 2013 nel Golfo del Messico

aprile 2013; le rilevazioni di boa, però, non sono registrate sempre con la stessa cadenza temporale e allo stesso istante e pertanto, qualora non fosse possibile ottenere il dato delle ore 22:00 esatte, si considera la rilevazione temporalmente più vicina (ore 21:55 o 21:50) confidando sul fatto che in un arco di tempo così breve non vi siano cambiamenti atmosferici rilevanti. Dalla figura 3.4 è possibile osservare la distribuzione delle 171 boe funzionanti; la loro distribuzione è abbastanza uniforme anche se più concentrata vicino alle coste. Da ogni boa si cercano informazioni su temperatura superficiale dell'acqua, che è la variabile di riferimento, temperatura dell'aria e pressione atmosferica, che sono invece le variabili che si potrebbero scegliere come covariate per il modello; infatti, se temperatura dell'aria e dell'acqua si influenzano reciprocamente, la pressione atmosferica è un buon indicatore delle condizioni meteorologiche, che anch'esse possono influenzare la temperatura dell'acqua. Poiché non tutte le boe hanno la stessa strumentazione a bordo, ogni variabile è rilevata da un numero di boe strettamente inferiore rispetto al totale di quelle disponibili. In particolare, la temperatura superficiale dell'acqua è rilevata da 92 boe, la temperatura atmosferica da 135 boe e la pressione atmosferica da 100 boe; le boe che rilevano sia la temperatura dell'acqua che dell'aria sono 63, quelle che rilevano sia la temperatura dell'acqua che la pressione sono 53 e le boe che rilevano tutte tre le grandezze sono 52.

3.2 Analisi dei dataset

Avendo a disposizione dei dataset provenienti da fonti diverse, si vorrebbe usare i dati di boa per la stima del campo di temperatura col modello SSR e i dati di satellite per validare la stima ottenuta; si rende quindi necessario verificare che le rilevazioni siano compatibili tra loro. Si considerano pertanto i 92 dati di boa, distribuiti come in figura 3.5. Per avere un buon indicatore della differenza tra le

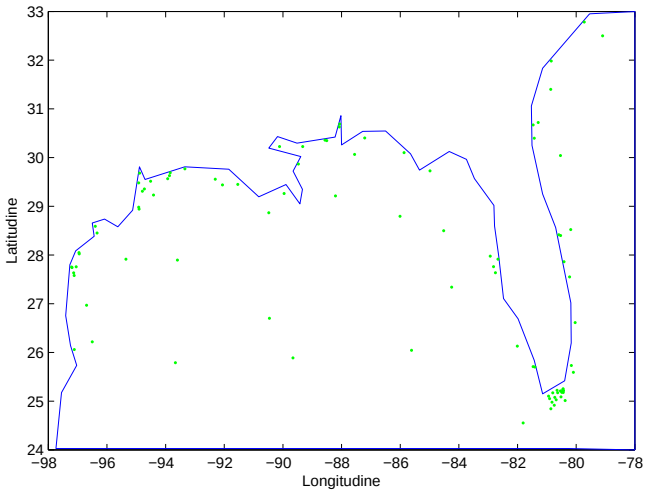


Figura 3.5: Distribuzione delle 92 boe rilevanti la temperatura della superficie oceanica

rilevazioni delle boe e le rilevazioni da satellite, si calcola la radice della loro differenza quadratica media (RDQM),

$$RDQM = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(z_i - z_i^{sat})^2}$$

dove z_i è l'i-esimo dato di boa e dove n è la numerosità dei dati di boa (in questo caso 92). Poiché la maggior parte delle coordinate di longitudine e latitudine in cui sono ormeggiate le boe non coincidono esattamente con le coordinate dei rilevamenti da satellite, ogni boa sarà confrontata con la rilevazione da satellite più vicina (nel senso della distanza euclidea). Il risultato è sorprendentemente elevato, $RDQM = 2,5723°$C, il che obbliga ad una analisi più dettagliata del dataset delle boe. Si vuole infatti capire se questa differenza sia uniforme tra tutte le boe o se sia concentrata maggiormente in alcune aree. Pertanto si suddivide il dataset delle boe in 5 parti così da individuare delle aree oceaniche specifiche: come mostra la figura 3.6, si considera la fascia costiera occidentale del Golfo del Messico (che per comodità si chiamerà area A) e quella settentrionale (area B), l'area centrale di mare aperto (area C), l'area sud-occidentale della Florida (area D) e l'area atlantica (area E). Calcolando la RDQM per ogni area, si ottengono i seguenti risultati:

$$\begin{aligned} RDQM_{AreaA} &= 3,1032°C \\ RDQM_{AreaB} &= 1,4756°C \\ RDQM_{AreaC} &= 1,0223°C \\ RDQM_{AreaD} &= 3,5779°C \\ RDQM_{AreaE} &= 1,1691°C \end{aligned}$$

Dai risultati si nota una certa disomogeneità della differenza tra i due dataset; una spiegazione possibile si può avere osservando la collocazione delle boe per

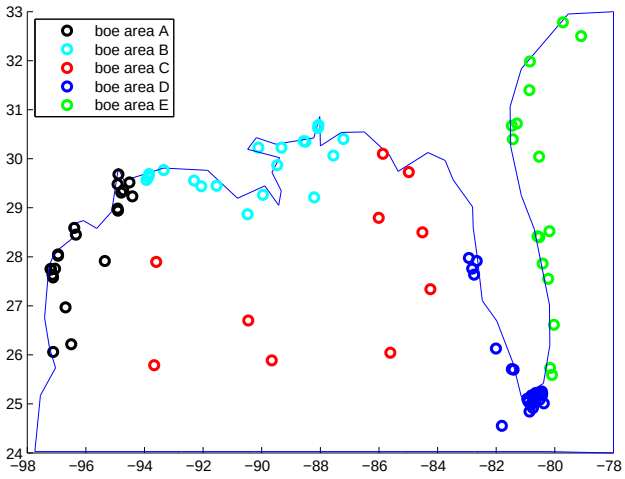


Figura 3.6: Suddivisione delle boe in 5 aree: l'area costiera occidentale del golfo (A), l'area costiera settentrionale del golfo (B), l'area centrale del Golfo del Messico (C), l'area sud-occidentale della Florida (D) e l'area atlantica (E)

ogni area: nell'area A e nell'area B le boe sono collocate prevalentemente nei pressi di baie e sono molto vicine alle coste; in queste aree il satellite non rende disponibili le sue misurazioni: è possibile che in queste aree siano presenti dei fattori che possono cambiare la temperatura dell'acqua, come la vicinanza di nuclei urbani o l'affluenza di fiumi. L'area D è invece caratterizzata dalla presenza di un gran numero di boe ormeggiate in una zona con fondale molto basso e questo può spiegare la presenza di rilevazioni di temperatura dell'acqua più elevate. Da notare inoltre che, essendoci alcune isolette in quell'area, il satellite non rileva la temperatura dell'acqua dove invece sono presenti le boe, proprio a ridosso della punta della Florida. Le aree C ed E sono invece quelle che presentano la discrepanza minore tra dati di boa e dati di satellite. Questi risultati suggeriscono quindi che l'idea di utilizzare il dataset delle boe per la stima del campo di temperatura e quello di satellite per poter validare la stima non sia percorribile con questi dati.
Merita una particolare attenzione anche l'analisi del dataset di satellite. In figura 3.7 è possibile vedere le curve di livello che descrivono il campo di temperatura misurato da satellite: tra queste linee di livello è possibile vedere molte irregolarità che che sono un indice del fatto che la superficie stessa del campo non è regolare a causa della presenza di rumore, che può avere diverse cause come, ad esempio, il passaggio di nubi o di fenomeni meteorologici locali nel momento della misurazione dei dati o come la sensibilità dello strumento che ha svolto la misurazione nel satellite. Queste irregolarità non possono però essere colte dai metodi che si utilizzano in questo elaborato perché sono pensati per stimare campi regolari. L'utilizzo di un campo non regolare avrebbe come conseguenza un aumento dell'errore di stima, provocato proprio dal fatto che la presenza di rumore sia impredicibile (si veda il paragrafo 4.8); risulta perciò una caratteristica non desiderabile per la validazione delle stime. Si rende allora necessario considerare l'idea di eliminare il rumore presente, regolarizzando il campo.

La figura 3.8 è ottenuta filtrando il dataset di satellite tramite la funzione Matlab `ndnanfilter.m`, disponibile all'indirizzo http://www.mathworks.com/matlabcentral/fileexchange/20417-ndnanfilter-m. Questa funzione legge in input la matrice che descrive un'immagine, in questo caso la matrice che contiene i valori rilevati da satellite, ordinati in base alla loro posizione spaziale, e tramite l'operazione di convoluzione elimina discontinuità e rumore. La convoluzione viene eseguita su ogni sottoinsieme di forma rettangolare, detto finestra, dalle dimensioni esplicitate in input e formato da punti della griglia adiacenti tra loro. Nel caso della figura 3.8, si è scelta una finestra quadrata di ampiezza 4 punti. Non vi è il rischio che le aree continentali, che sono descritte nella matrice in input con il valore NaN, condizionino la convoluzione perché la funzione `ndnanfilter.m` può evitare che ciò avvenga. Nel seguito dell'elaborato verranno utilizzati entrambi i dataset, che verranno chiamati "dataset originale" e "dataset regolarizzato".

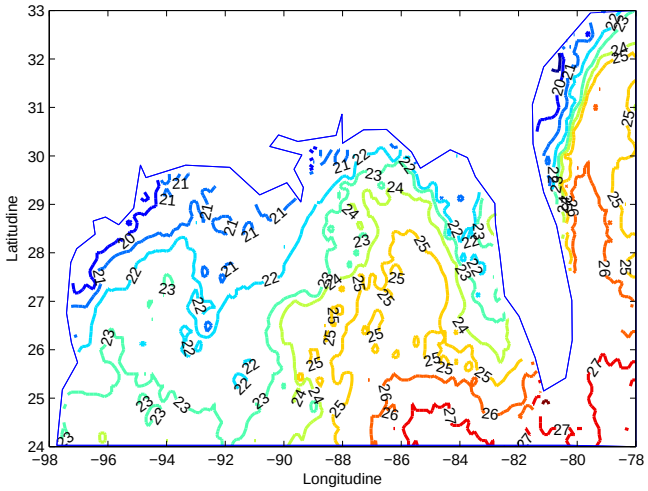


Figura 3.7: Curve di livello che descrivono il campo di temperatura misurato da satellite.

3.3 Software utilizzato

Nelle applicazioni dei prossimi paragrafi e del prossimo capitolo verranno mostrati i risultati di stime ottenute col modello SSR e con le tecniche di Kriging. Per il modello SSR si è usata un'implementazione su Matlab sviluppata da J.O. Ramsay e L.M. Sangalli a partire dalla libreria Matlab relativa alla *functional data analysis* (FDA) di cui l'autore è J.O. Ramsay. La FDA è un insieme di tecniche statistiche che sono nate con l'obiettivo di analizzare dati rappresentabili per mezzo di opportune curve o superfici. È possibile approfondire questo argomento consultando il volume [Ramsay and Silverman, 2005] o tenendo come riferimento l'indirizzo http://www.psych.mcgill.ca/misc/fda/downloads/FDAfuns/. A questo indirizzo, inoltre, è disponibile parte del software sviluppato da J.O. Ramsay.

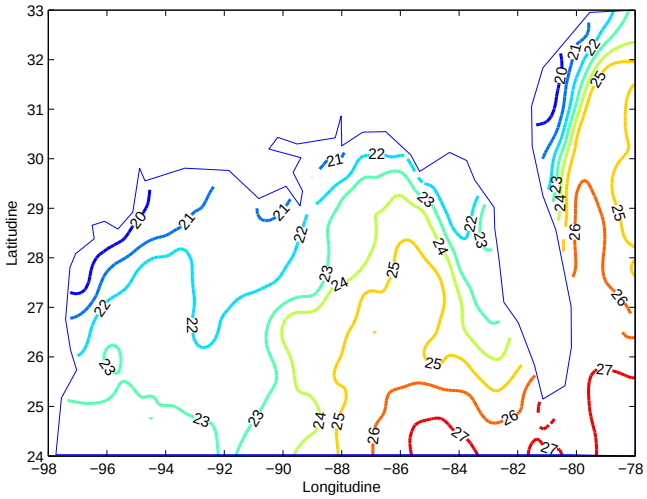


Figura 3.8: Curve di livello che descrivono il campo di temperatura misurato da satellite filtrato dal rumore con la funzione Matlab `ndnanfilter.m` con finestra quadrata ampia 4 punti.

Per la triangolazione del dominio, problema che non è affatto banale, sono suggerite molte tecniche in letteratura che cercano di risolverlo e di cui esistono specifici pacchetti Matlab. In questo elaborato si fa riferimento all'*algoritmo di Delaunay*, utilizzato tramite il comando Matlab *DelaunayTri()*.
Per quanto riguarda il Kriging, invece, si è utilizzato il software R e in particolare il pacchetto *gstat*. Per poter comparare i risultati con quelli del modello SSR si sono poi importate le stime effettuate con R su Matlab. Sebbene non sia stata utilizzata per le simulazioni di questo elaborato, si segnala l'esistenza anche di una libreria in R per il modello SSR.

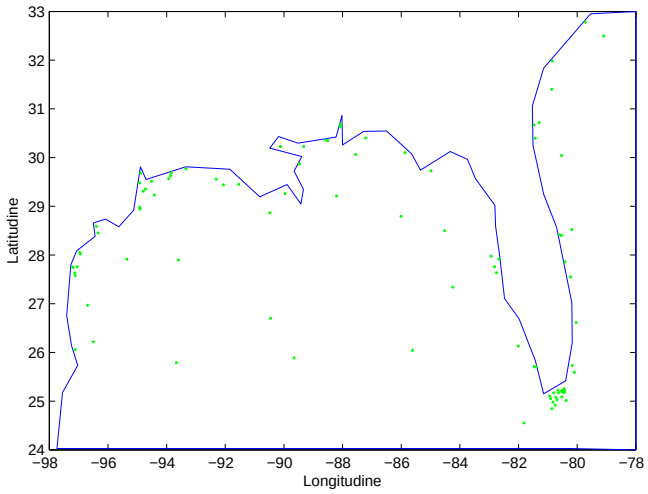
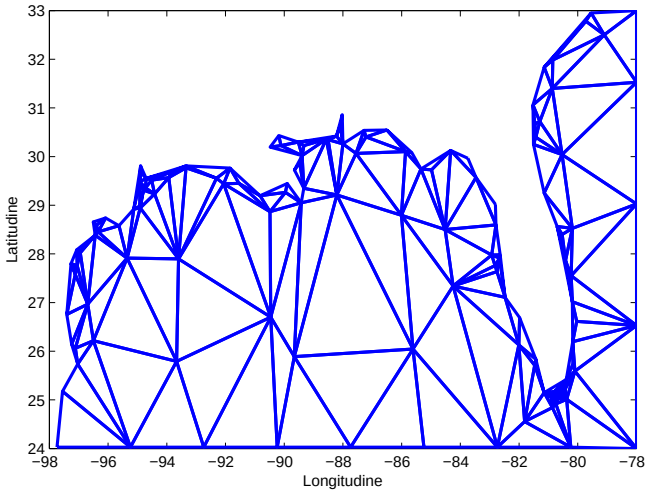


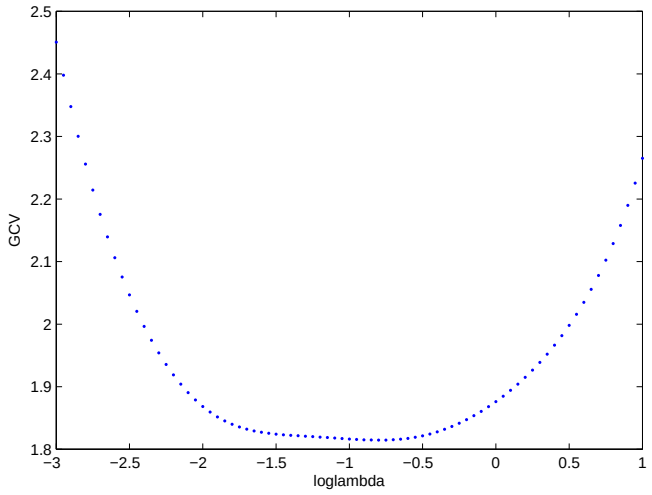
Figura 3.9: Distribuzione delle 92 boe che rilevano la temperatura della superficie oceanica

3.4 Stime del campo di temperatura dal dataset delle boe

Si considera ora il dataset delle 92 boe che misurano la temperatura della superficie oceanica. Dalla figura 3.9 si può notare che la loro collocazione non è delle migliori, essendoci una buona concentrazione di boe in alcune aree costiere e una concentrazione molto minore in altre aree; la triangolazione risente di questa caratteristica del dataset: in figura 3.10(a) si nota la presenza di triangoli piccoli e lunghi alternati a triangoli grandi ed equilibrati nelle dimensioni. La triangolazione, quindi, non è isotropa. Si esegue il modello SSR cercando funzioni nello spazio X_h^2, ovvero utilizzando polinomi di secondo grado su ogni elemento. La stima di $\log \hat{\lambda}$ col metodo GCV è $-0,8$ (figura 3.10(b)), vicino a 0, che renderà il campo abbastanza regolare consentendo però dei piccoli gradienti. La stima della deviazione standard dei residui ϵ del modello è $\hat{\sigma} = 1,1376°C$.
In figura 3.11 è possibile vedere le linee di livello del campo di temperatura stimato col modello SSR a partire dai dati di boa e il campo di temperatura misurato da satellite e regolarizzato. Il confronto tra queste due immagini conferma che il dataset delle boe e quello del satellite non sono compatibili tra di loro: infatti, se dal punto di vista macroscopico vi si può scorgere una certa somiglianza (si osservi che nella stima, ad esempio, le aree più calde sono concentrate a sud nei pressi della Florida, inoltre è presente la Corrente del Golfo), guardando più in dettaglio le linee di livello si nota come localmente i due campi non sono tra loro confrontabili. La differenza tra i due campi, in figura 3.12 conferma quanto appena affermato. In totale, la RDQM tra i due campi è $1,6486$°C.
Anche i risultati con le tecniche di Kriging hanno le stesse caratteristiche: le stime del campo stocastico a partire dai dati di boa, in figura 3.17, mostrano una sostanziale somiglianza alla stima ottenuta con il modello SSR, e forniscono



(a)



(b)

Figura 3.10: Triangolazione con dataset dei dati di boa e calcolo del valore di $\log \lambda$ ottimale col metodo GCV.

un'ultima conferma che in questo caso i dati di satellite non possono essere utilizzati per validare le stime coi dati di boa. Per ottenere le stime con il Kriging nelle immagini di figura 3.17 si è scelto di utilizzare lo stimatore di Cressie per la stima del variogramma empirico con rispettivamente variogramma teorico sferico, in figura 3.13, e variogramma Matérn, in figura 3.14. La RDQM tra il campo con stimatore sferico e quello misurato da satellite è $1,6649°C$, quello con stimatore Matérn è $1,6476°C$, valori simili a quello del SSR.
Non essendo possibile utilizzare contemporaneamente il dataset di satellite e quello delle boe, nel prossimo capitolo si farà riferimento al dataset più completo, quello di satellite, per compiere confronti tra le stime ottenute col modello SSR e quelle ottenute con le tecniche di Kriging.

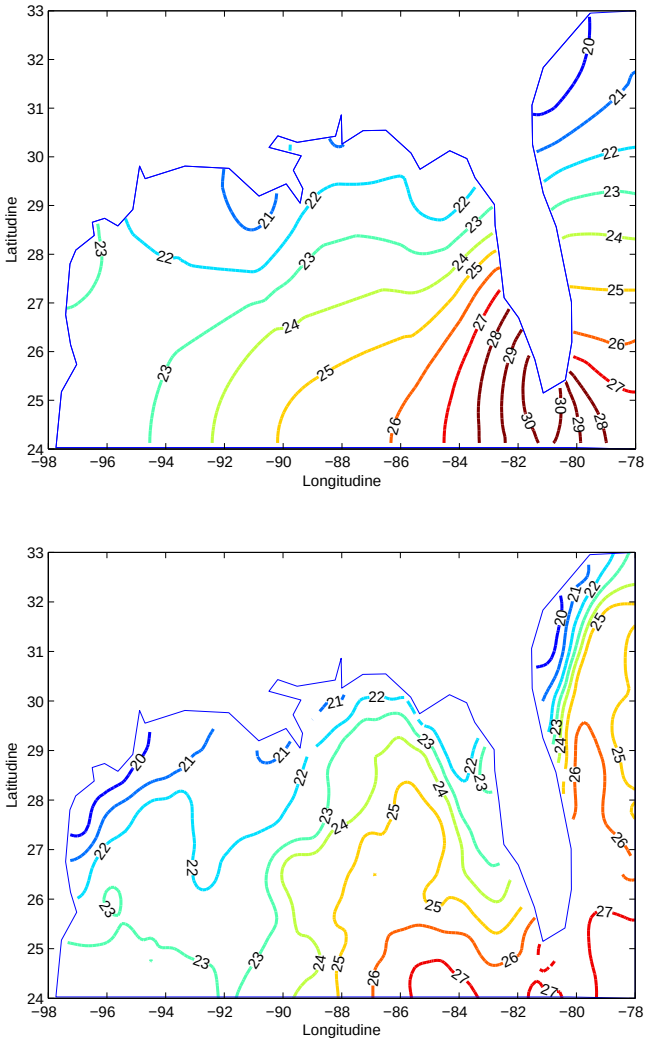


Figura 3.11: In alto le linee di livello del campo di temperatura stimato dal modello SSR col dataset delle boe; in basso le linee di livello del campo di temperatura misurato da satellite.

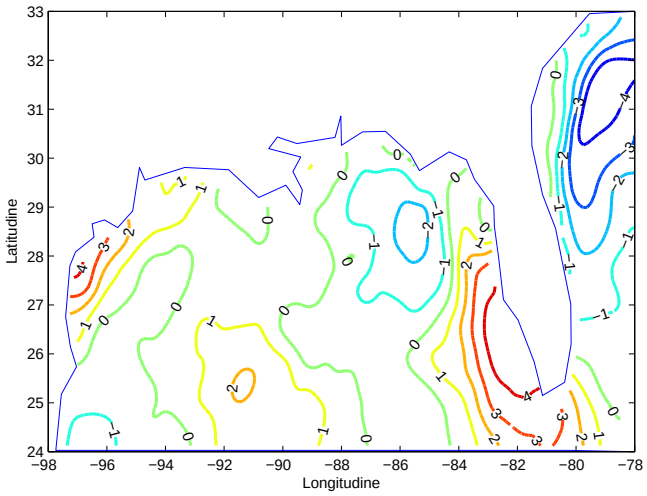


Figura 3.12: Curve di livello della distanza tra il campo di temperatura stimato dal modello SSR con dataset delle boe e quello misurato da satellite

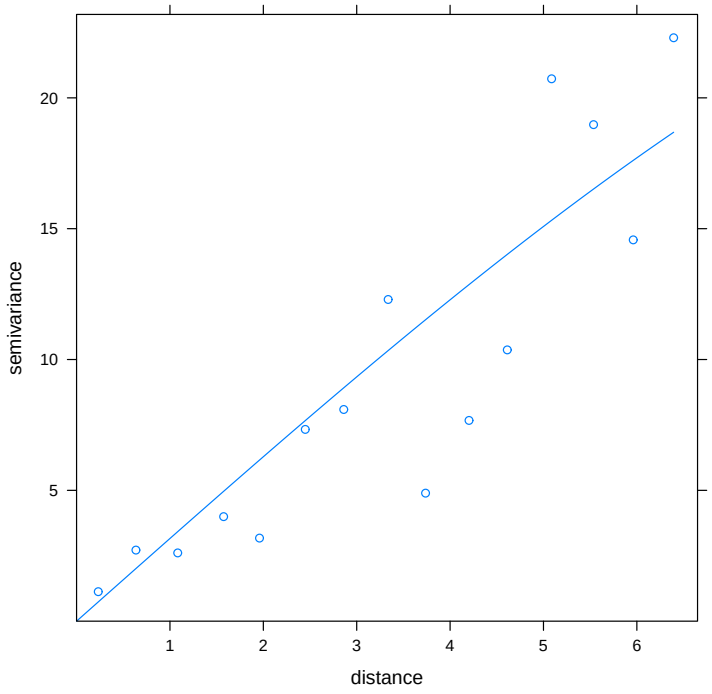


Figura 3.13: Variogramma teorico della famiglia dei modelli sferici (in linea continua) e variogramma empirico stimato con stimatore di Cressie (col simbolo o) per i dati di boa. Sull'asse delle ascisse, 1 unità $\approx$ 111 km.

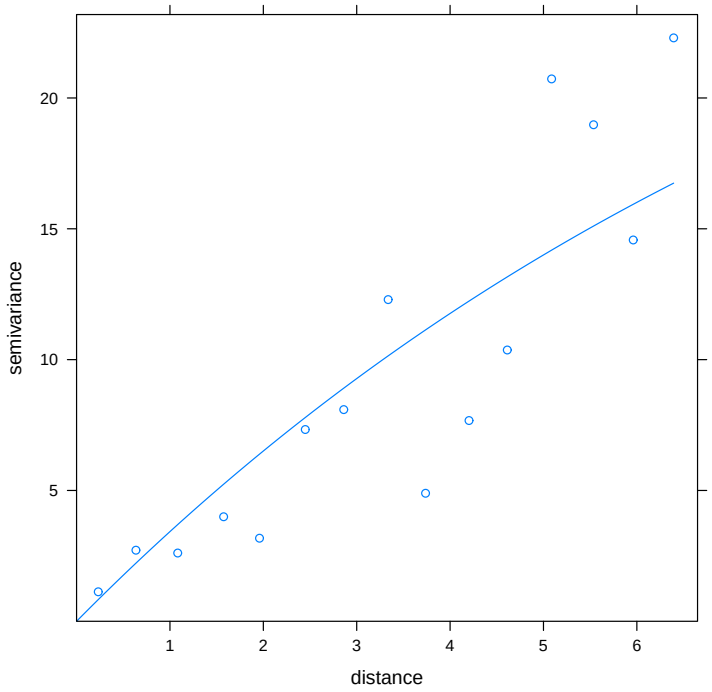


Figura 3.14: Variogramma teorico della famiglia dei modelli Matérn (in linea continua) e variogramma empirico stimato con stimatore di Cressie (col simbolo o) per i dati di boa. Sull'asse delle ascisse, 1 unità $\approx$ 111 km.

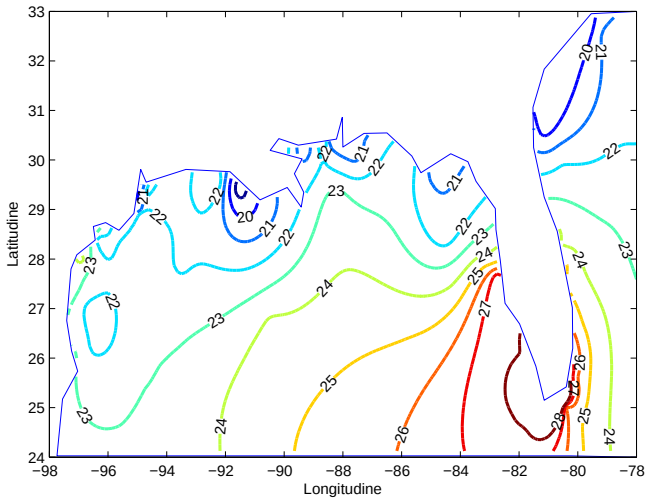


Figura 3.15: Linee di livello del campo di temperatura ottenuto con la tecnica Kriging con variogramma sferico a partire dai dati di boa.

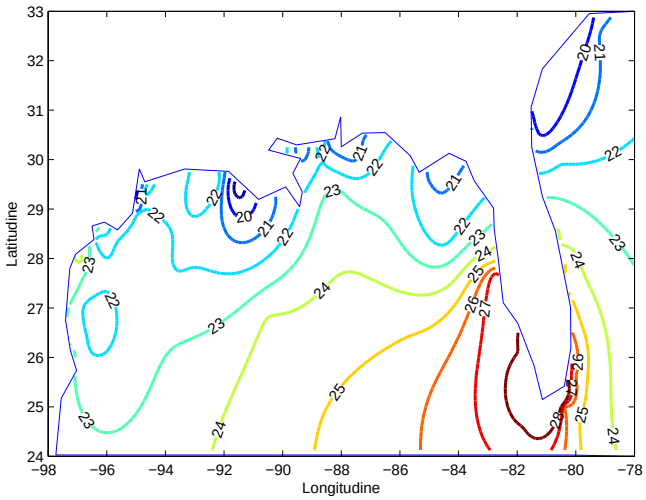


Figura 3.16: Linee di livello del campo di temperatura ottenuto con la tecnica Kriging con variogramma Matérn a partire dai dati di boa.

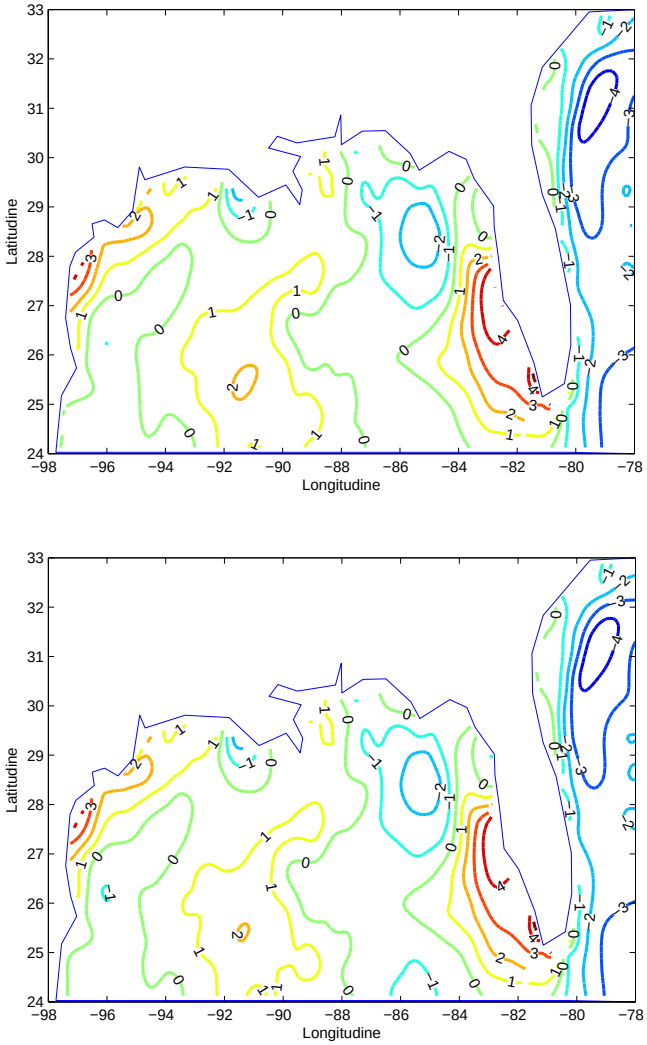


Figura 3.17: In alto, differenza tra il campo stimato con la tecnica di Kriging con variogramma sferico e il dataset di satellite. In basso, differenza tra il campo stimato con la tecnica di Kriging con variogramma Matérn e il dataset di satellite.

Capitolo 4

Confronto del modello SSR con le tecniche di Kriging

Il modello SSR e le tecniche di Kriging hanno l'obiettivo di determinare il valore di una variabile in ogni punto di un dominio spaziale conoscendo solo un numero limitato di campioni. Poiché il dataset di satellite e quello delle boe non sono compatibili tra loro, si considera solo il dataset del satellite perché è formato da un numero significativo di dati, 6953, e tutte le stime vengono fatte utilizzando questi dati. Il dataset del satellite diventa quindi il dataset di riferimento e si suppone che rilevi l'esatta temperatura della superficie oceanica alle ore 22.00 GMT. In questo capitolo si vogliono confrontare le stime ottenute con il modello SSR e quelle ottenute con le tecniche di Kriging. Nei primi tre paragrafi viene illustrata la procedura seguita per confrontare le stime ottenute con il modello SSR e con le tecniche di Kriging, vengono descritte le analisi statistiche eseguite sulle stime ottenute e viene mostrato come sono stati scelti i variogrammi per il Kriging. Nei paragrafi successivi vengono presentati i risultati ottenuti e confrontando le due tecniche.

4.1 Procedimento seguito

Per poter confrontare le stime delle tecniche SSR e Kriging si è seguita la seguente procedura: dopo aver indicizzato i dati di satellite, si è estratto un campione casuale di 100 indici; i dati corrispondenti a tali indici formano il *training set*. Se ne sono scelti solamente 100 perché è un numero confrontabile con i dati di boa a disposizione e perchè, essendo solo una piccola porzione dei dati di satellite a disposizione, anche se vengono tolti dal dataset del satellite non viene modificata in modo apprezzabile la numerosità del *testing set*. Si sono poi eseguiti il modello SSR e le tecniche Kriging ottenendo le rispettive stime del campo di temperatura. Questa operazione è stata ripetuta per 200 volte, così da poter disporre di 200 *training set* differenti, spazialmente distribuiti in modo uniforme, e di 200 stime differenti per ogni tecnica, che è un numero sufficiente perché le statistiche possano essere ritenute significative. Dato che si sta svolgendo un'operazione di stima, si è aggiunto ai *training set* un rumore gaussiano $\mathcal{N}(0; 0, 1)$, un rumore molto lieve se si considera che l'alterazione

massima, in modulo, è di $1,0864°C$; in questo modo i 100 punti estratti possono essere visti come delle boe e il rumore aggiunto come il loro errore di misurazione. Gli obiettivi di questa applicazione sono molteplici: anzitutto si vuole analizzare statisticamente come si comportano le due tecniche quando si hanno a disposizione delle boe ben distribuite; si vuole capire se le due tecniche forniscano risultati tra loro comparabili; si è interessati a capire se ci siano dei punti nel dominio instrinsecamente più problematici per l'una e per l'altra tecnica; infine, si vuole stabilire quali siano le aree con maggiore variabilità di stima.
Per poter implementare un algoritmo che svolga in automatico la stessa procedura per tutti i 200 campioni è necessario introdurre alcune limitazioni sia per il modello SSR che per il Kriging. Nel primo caso si pongono dei limiti per la scelta di λ: può capitare, infatti, che il metodo GCV tenda a stimare valori troppo bassi o troppo alti di $\log\hat{\lambda}$, a seconda della distribuzione spaziale dei dati; questo criterio, però, non è un criterio di scelta assoluto, ma solo un criterio indicativo: si limita la scelta di $\log\hat{\lambda}$ nell'intervallo $[-2;0]$ in modo che $\hat{\lambda}$ possa essere compreso nell'intervallo $I = [0,01;1]$. Valori troppo bassi di $\log\hat{\lambda}$, infatti, farebbero perdere di senso la minimizzazione del laplaciano, di fatto annullando l'operazione di regolarizzazione. Invece gli estremi di I sono tali da consentire sia di poter penalizzare il laplaciano della funzione f, sia di non penalizzarlo eccessivamente.
Per quanto riguarda le tecniche di Kriging, la semplificazione introdotta riguarda la scelta del variogramma teorico: non per tutti i 200 campioni, infatti, può essere ragionevole la scelta di un variogramma appartenente alla stessa famiglia. Non potendo però soffermarsi al controllo di tutti i 200 variogrammi, si suppone che la scelta di un buon variogramma teorico effettuata per alcuni campioni sia idonea per la maggioranza dei campioni. Sebbene questa ipotesi sia forte, si confida sul fatto che le scelte del variogramma sarebbero comunque soggettive e sul fatto che, nella peggiore delle ipotesi, eventuali variogrammi non idonei restituiscano nelle analisi statistiche degli outliers, che quindi non saranno tenuti in considerazione.
Per il modello SSR, un aspetto importante per una buona stima è l'isotropia della triangolazione, che dipende sia dalla distribuzione dei dati nel dominio che dai nodi che compongono il bordo del dominio stesso. Si pone perciò come vincolo sulla scelta casuale dei 100 dati che siano distanti tra loro e dai nodi del bordo almeno di 0,5 unità (1 unità=1 grado di longitudine, ≈ 111 km). Per quanto riguarda il bordo, quello mostrato nel paragrafo precedente è costruito in base alla posizione dei dati di boa, che risultano in molti casi collocati all'interno di piccole baie nei pressi della costa. Il satellite, invece, non ha effettuato misurazioni in queste piccole aree. Per evitare che durante la triangolazione si formino triangoli molto piccoli i cui vertici sono solo nodi del bordo, si ridefinisce il bordo come in figura 4.1, formato anch'esso da 63 punti. Nei paragrafi successivi si utilizzerà anche un bordo formato da 55 punti, molto più somigliante al bordo utilizzato nel capitolo 3, per verificare il comportamento del modello SSR con gli stessi dati ma con una triangolazione meno isotropa; questo bordo meno accurato è visibile in figura 4.2. La ridefinizione del bordo comporta una piccola riduzione della numerosità del dataset di satellite, formato ora da 6944 punti, perché sono state rimosse 9 misurazioni effettuate in un'area esterna al nuovo dominio. In figura 4.3 è mostrata la posizione delle misurazioni di satellite

che vengono utilizzate nelle applicazioni di questo capitolo.

4.2 Descrizione delle analisi statistiche

Le stime ottenute con le tecniche SSR e Kriging vengono analizzate dal punto di vista statistico in modo da capire quale metodo abbia prodotto una stima globalmente più accurata e quali siano le aree dove la stima risulta più difficoltosa. Per valutare l'errore di stima che viene globalmente commesso su tutto il dominio si calcola il RMSE (*Root Mean Squared Error*) per ogni campione j, con $j = 1, \ldots, 200$, così definito:

$$RMSE(j) = \sqrt{\frac{1}{K}\sum_{k=1}^{K}(\hat{f}_j(\mathbf{p}_k) - f(\mathbf{p}_k))^2}$$

dove $K = 6944$ è il numero totale di punti della griglia, $\hat{f}_j(\mathbf{p}_k)$ è la stima del campo ottenuta dal j-esimo campione e calcolata nel punto di coordinate $\mathbf{p}_k$, relativo al k-esimo nodo della griglia, e $f(\mathbf{p}_k)$ è il corrispettivo punto del dataset di satellite. Effettuando questa operazione per tutti i 200 campi stimati, si ottiene un nuovo dataset formato da 200 RMSE, che può essere confrontato statisticamente con i 200 RMSE ottenuti da un'altra tecnica di stima tramite l'utilizzo di boxplot.
Per capire invece quali siano le aree del dominio dove la stima risulta più difficoltosa si considerano su ogni nodo della griglia le 200 simulazioni e si calcola il corrispettivo RMSE, così definito:

$$RMSE(k) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{f}_j(\mathbf{p}_k) - f(\mathbf{p}_k))^2}$$

dove $n = 200$ e $k = 1, \ldots, 6944$. Si hanno così a disposizione 6944 valori di RMSE, ognuno relativo ad un nodo della griglia di satellite, ed è quindi possibile visualizzare graficamente per ogni tecnica quali siano le aree dove la stima è mediamente più affine al dataset di satellite e quali dove mediamente ci sono maggiori differenze.
Sebbene non sia un indice statistico, viene controllato anche l'indice di condizionamento della matrice $\mathbf{A}$ del modello SSR, per controllare il comportamento del modello a seconda del bordo utilizzato e della rispettiva triangolazione.

4.3 Scelta del modello teorico per il Kriging

Dovendo utilizzare la tecnica del Kriging per 200 campioni, ognuno di questi richiederebbe una scelta oculata del variogramma teorico, ma una procedura di questo tipo sarebbe troppo lunga in termini di tempo. Si sceglie quindi il variogramma analizzando i primi 4 campioni e dalla loro analisi si sceglie la famiglia di modelli che in seguito verrà utilizzata per tutti i 200 campioni. In

particolare, la funzione *vgm()* del pacchetto *gstat*, che modellizza il variogramma teorico, consente di scegliere tra il modello sferico, il modello esponenziale, il modello gaussiano e il modello Matérn, descritti nella sezione 2.4. Inoltre è possibile utilizzare lo stimatore di Cressie e lo stimatore di Matheron. Si sceglie quindi di utilizzare due tecniche di Kriging, per verificare che i confronti non dipendano solo dalla scelta particolarmente fortunata del variogramma. Nelle figure 4.4 e 4.5 sono presentati i variogrammi empirici e teorici per i primi 4 campioni: nel primo caso lo stimatore scelto è quello di Matheron, a cui è stato associato il variogramma sferico; nel secondo caso, invece, si è scelto lo stimatore di Cressie, a cui è stato associato il variogramma di Matérn.
La scelta dell'ampiezza delle classi di distanza h_k è stata lasciata al software R. In tabella 4.1 è mostrata la numerosità e l'ampiezza delle classi di distanza scelte di default da R: l'ampiezza media è di $0,5$ unità circa, dove un'unità è un grado di longitudine (circa 111 km); la numerosità delle classi è sempre superiore a 30 coppie.

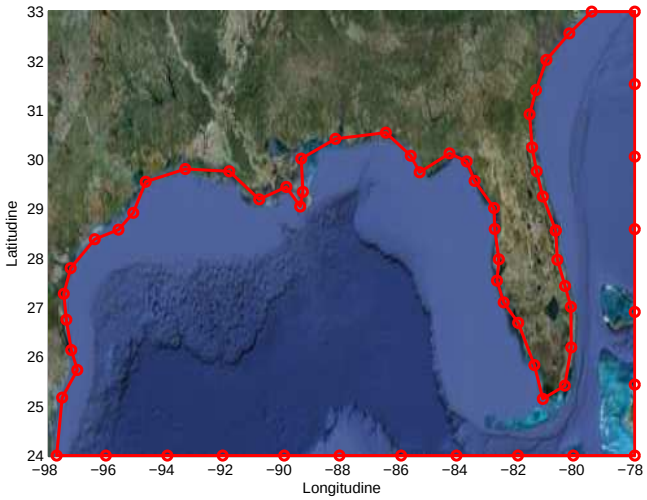


Figura 4.1: Costa del Golfo del Messico ridefinita in modo più accurato e posizione dei 63 nodi con cui è costruito il bordo (con i cerchi).

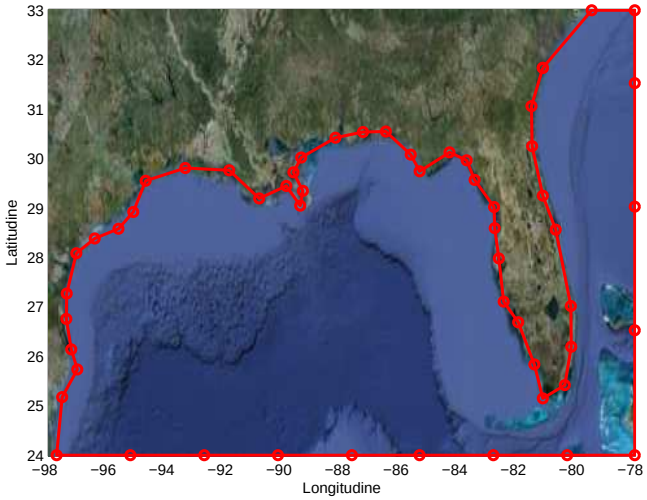


Figura 4.2: Costa del Golfo del Messico ridefinita in modo meno accurato e posizione dei 55 nodi con cui è costruito il bordo (con i cerchi).

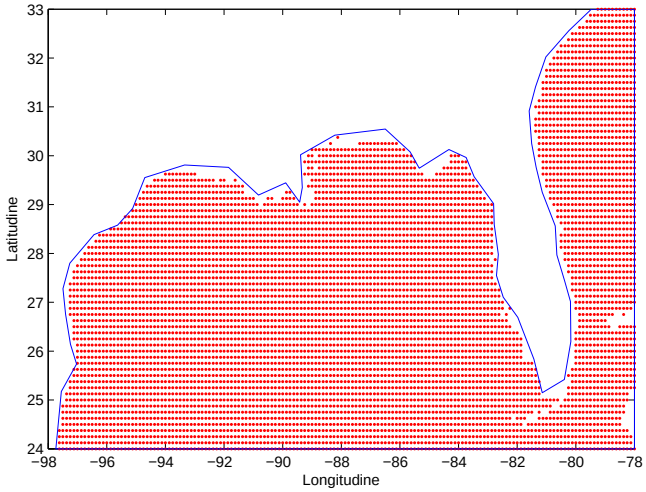


Figura 4.3: Griglia dei punti misurati dal satellite alle ore 22:00 GMT del 15 aprile 2013 nel Golfo del Messico con il nuovo bordo.

Classe	Campione 1		Campione 2		Campione 3		Campione 4	
	num	distanza	num	distanza	num	distanza	num	distanza
1	081	0,6710628	109	0,7171433	098	0,7106926	100	0,7153239
2	131	1,1071533	142	1,1756052	128	1,1237650	127	1,1523795
3	170	1,5523131	159	1,5985290	184	1,5554969	166	1,5810050
4	181	1,9931535	195	2,0497120	211	2,0168373	188	2,0485590
5	205	2,4352723	225	2,4944751	215	2,4387762	200	2,4767932
6	185	2,8910441	235	2,9363267	223	2,8761872	226	2,9585173
7	220	3,3348279	231	3,3839755	241	3,3351377	198	3,3905793
8	195	3,7617514	257	3,8465651	238	3,7870033	262	3,8328489
9	218	4,1941930	248	4,2993477	219	4,2271719	227	4,2956468
10	222	4,6603518	235	4,7505248	238	4,6724004	214	4,7354808
11	228	5,0897886	211	5,1941341	202	5,1235496	214	5,1948019
12	205	5,5352205	203	5,6558754	203	5,5614633	234	5,6602506
13	207	5,9693024	202	6,1076288	185	6,0080249	200	6,1113778
14	210	6,4148388	200	6,5325289	184	6,4361469	211	6,5675916

Tabella 4.1: Riassunto della numerosità (num) e della distanza delle classi sui primi 4 campioni di 100 misurazioni di satellite scelte casualmente.

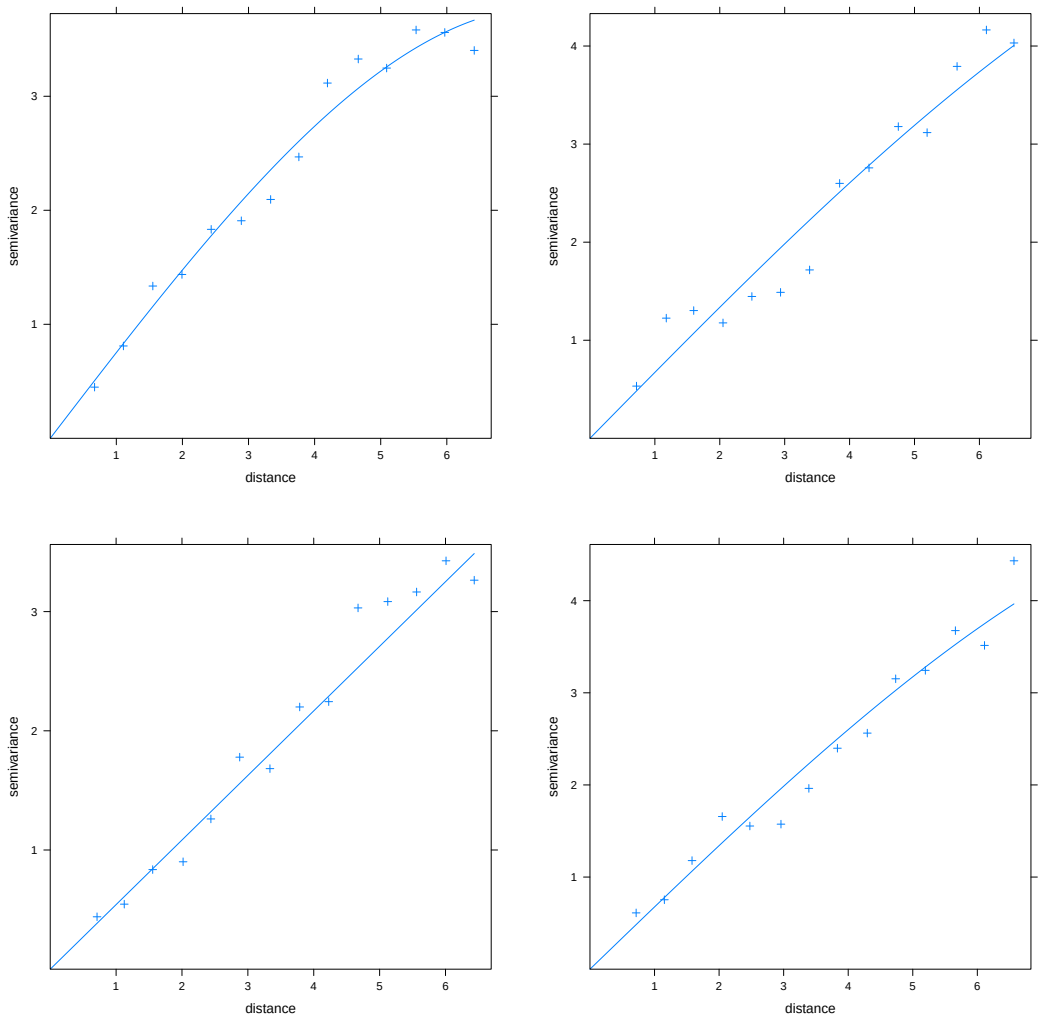


Figura 4.4: Variogrammi teorici della famiglia dei modelli sferici (in linea continua) e variogrammi empirici stimati con stimatore di Matheron (col simbolo +) per i primi 4 campioni di 100 misurazioni di satellite scelte casualmente. Sull'asse delle ascisse, 1 unità $\approx$ 111 km.

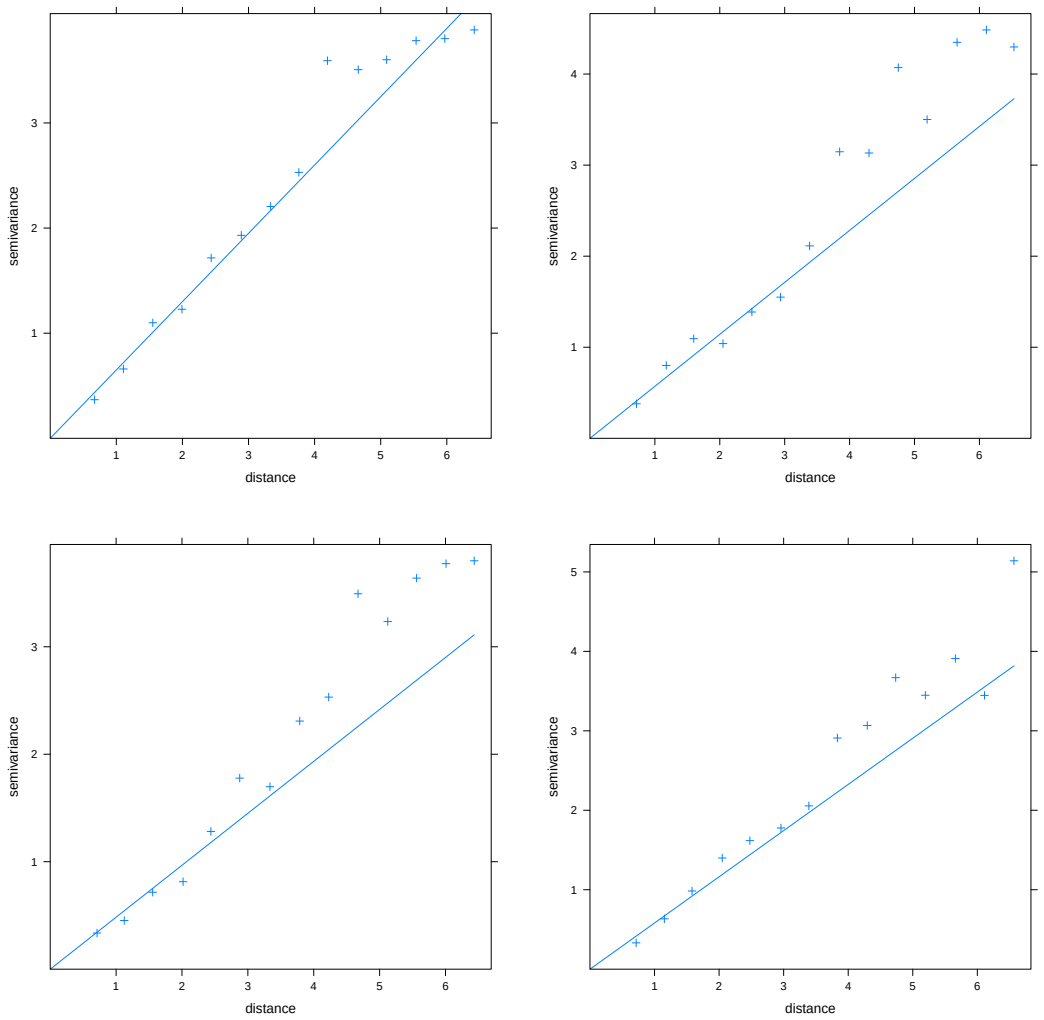


Figura 4.5: Variogrammi teorici della famiglia dei modelli Matérn (in linea continua) e variogrammi empirici stimati con stimatore di Cressie (col simbolo +) per i primi 4 campioni di 100 misurazioni di satellite scelte casualmente. Sull'asse delle ascisse, 1 unità $\approx$ 111 km.

4.4 Confronto tra il modello SSR e le tecniche di Kriging

Si considerano i 200 campioni ottenuti a partire dal dataset di satellite regolarizzato, campionati in modo tale che tra di loro ci sia una distanza euclidea sul dominio di almeno 0,5 unità (1 unità=1 grado di longitudine, ≈ 111 km), e si eseguono il modello SSR e la tecnica di Kriging con variogramma sferico e con variogramma di Matérn; se ne confrontano quindi i risultati. In figura 4.6 è possibile vedere come sono le triangolazioni ottenute per i primi 4 campioni con il bordo più accurato e in figura 4.7 quelle ottenute col bordo meno accurato.
In figura 4.8 sono mostrati i boxplot ottenuti dai 200 RMSE calcolati dopo aver stimato il campo di temperatura con ogni tecnica. I boxplot relativi alle stime col metodo SSR sono più bassi rispetto a quelli relativi alle stime col Kriging; la tabella 4.2 indica che l'errore quadratico medio del modello SSR è inferiore rispetto a quello delle tecniche SSR e che l'uso di due bordi in parte diversi non ha inciso sensibilmente sulle simulazioni. Per entrambi i bordi l'indice di condizionamento sulla matrice $\mathbf{A}$ per tutte le 200 simulazioni è di un valore dell'ordine di 10^4.

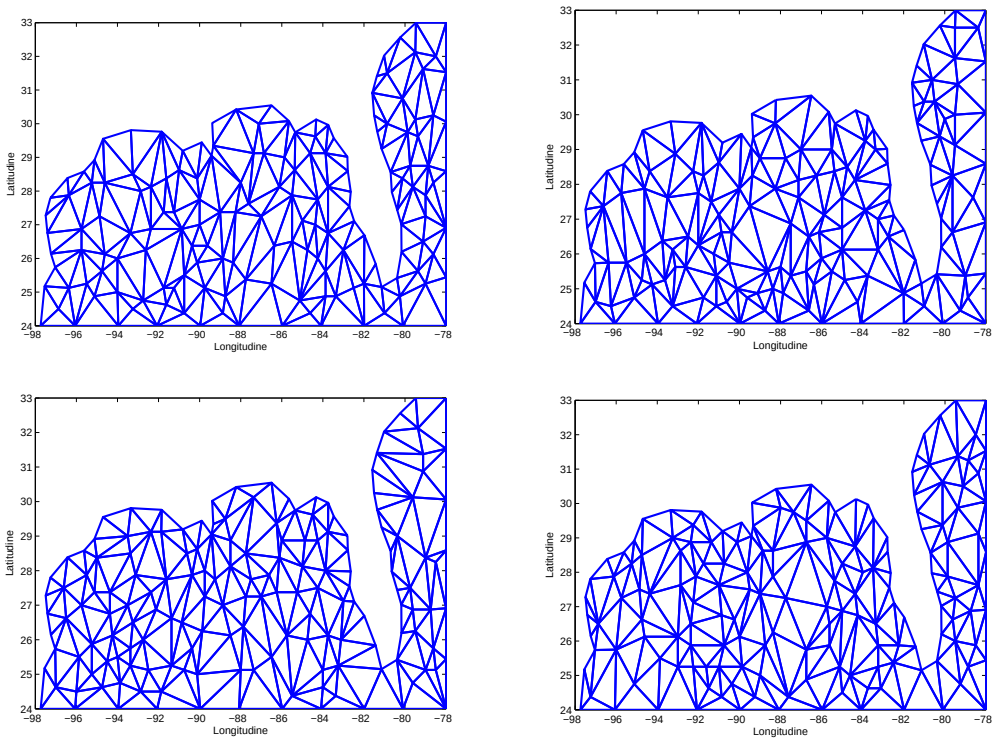


Figura 4.6: Triangolazione di 100 dati casuali dei primi 4 campioni con bordo accurato.

Si passa allora all'analisi delle caratteristiche locali delle stime. In figura 4.9 è mostrato com'è distribuito spazialmente il RMSE per ogni tecnica di stima utilizzata; sia per il modello SSR che per il Kriging l'errore maggiore è commesso nei pressi delle coste, soprattutto nell'area a nord della Florida, sia all'interno del Golfo che nella zona atlantica. Questo risultato indica che la regione più delicata

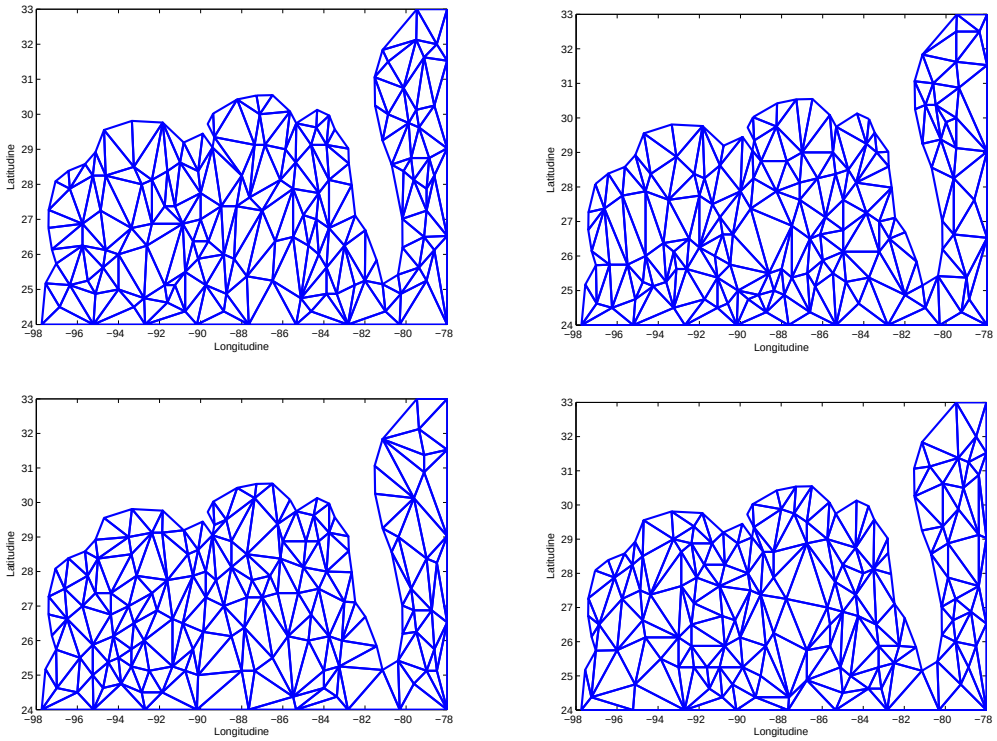


Figura 4.7: Triangolazione di 100 dati casuali dei primi 4 campioni con bordo meno accurato.

Tecnica	1° quartile	Mediana	3° quartile
Modello SSR Bordo A	0,3990	0,4313	0,4794
Modello SSR Bordo B	0,3971	0,4298	0,4792
Kriging-Var. Sferico	0,4275	0,4636	0,5022
Kriging-Var. Matérn	0,4253	0,4624	0,5020

Tabella 4.2: Quartili del **RMSE** valutato sulla stima del campo di temperatura con 200 campioni noto il campo misurato da satellite.

da stimare è proprio questa area, che non è lambita dalla Corrente del Golfo sebbene non si trovi troppo lontana da essa.
Per avere un esempio di una stima, si considera il primo campione: in figura 4.10 è mostrato il campo di temperatura stimato con la due tecniche; l'aspetto più evidente è che le stime con Kriging non siano influenzate dalla presenza della Florida: si potrebbe infatti prolungare le linee di livello ad est della Florida e congiungerle idealmente con quelle a ovest. Questo non avviene per quanto riguarda le stime con il modello SSR: risulterebbe molto difficile pensare di congiungere idealmente le linee di livello a est con quelle a ovest della penisola. La presenza di RMSE più elevato nelle stime con SSR localizzato in quest'area può essere dovuto proprio a questa caratteristica particolare, presente anche del dataset di satellite. Nelle immagini in figura 4.11 si può vedere la differenza tra i campi stimati della figura 4.10 e il campo di temperatura misurato da satellite e regolarizzato.

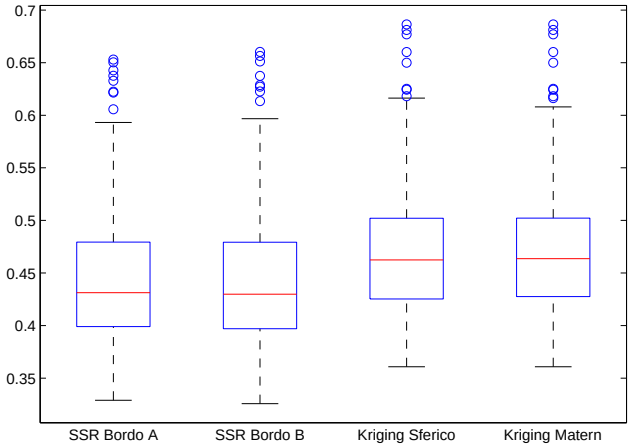


Figura 4.8: Boxplot dei 200 RMSE calcolati dopo aver stimato il campo di temperatura della superficie oceanica con il modello SSR con bordo più accurato (bordo A) e con quello meno accurato (bordo B), con Kriging con variogramma sferico e con Kriging con variogramma Matérn.

A partire dai risultati di questo paragrafo è interessante considerare come cambino i risultati e i confronti nel caso in cui si abbia una situazione diversa: in particolare, quando si hanno a disposizione un numero maggiore o un numero minore di campioni, quando si implementa il modello SSR cercando soluzioni agli elementi finiti di tipo P1, nello spazio $\mathring{\mathrm{X}}_h^1$, quando viene eliminato il vincolo di distanza euclidea nel dominio tra i punti campionati e quando non si considera il dataset di satellite regolarizzato, ma quello non regolarizzato.

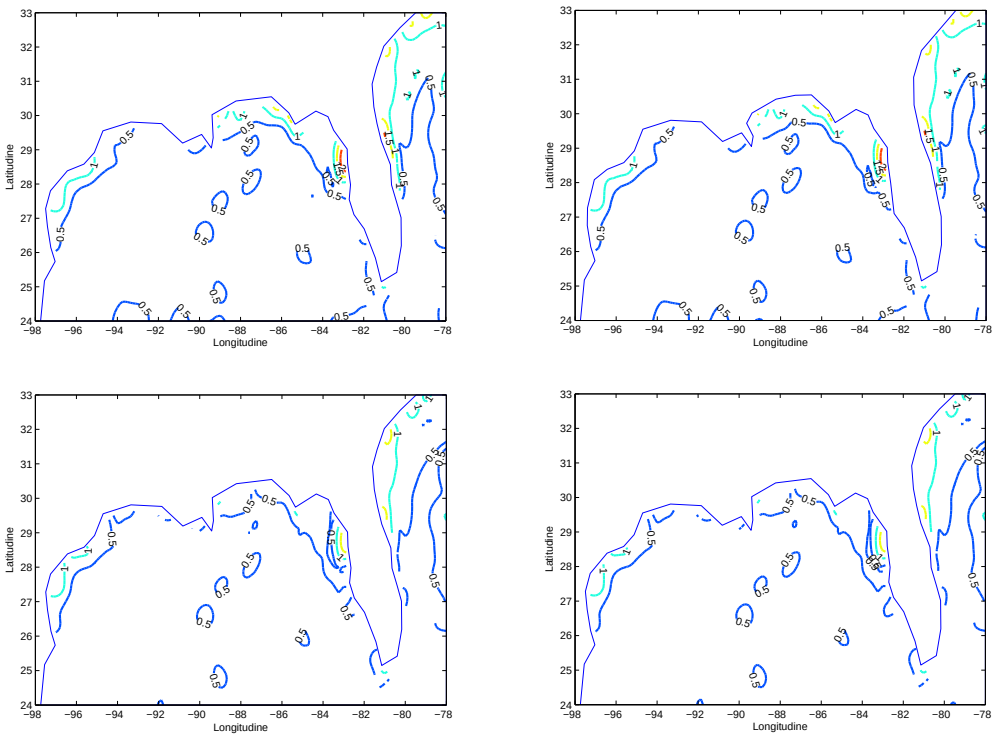


Figura 4.9: Rappresentazione della distribuzione sul dominio del **RMSE** calcolato in ogni nodo della griglia a partire dalle 200 simulazioni con 100 campioni. In particolare, in alto a sinistra, RMSE calcolato a partire dalla stima effettuata con il modello SSR con bordo accurato, in alto a destra con il modello SSR con bordo meno accurato; in basso a sinistra RMSE calcolato a partire dalla stima effettuata con Kriging con variogramma sferico e a destra con variogramma Matérn.

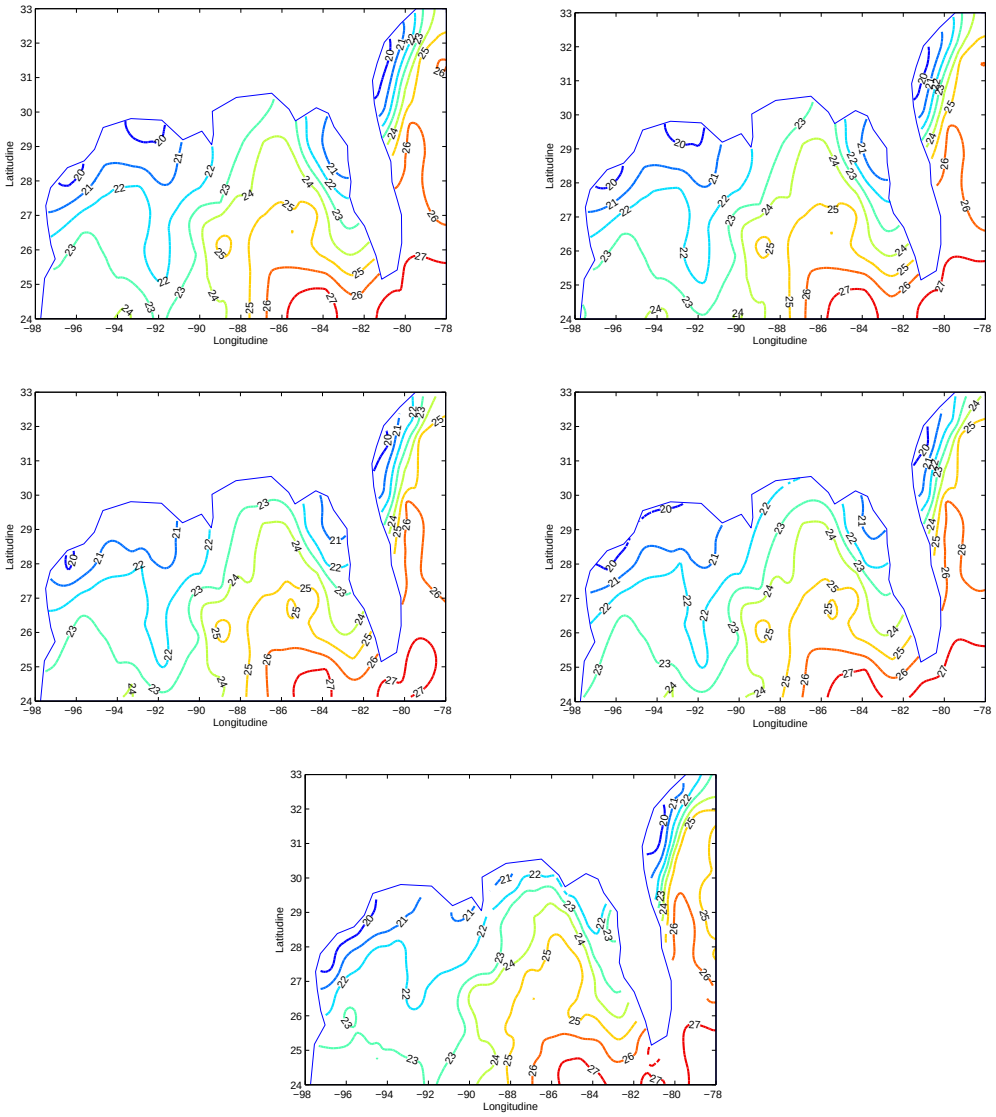


Figura 4.10: In alto, campo stimato col primo campione casuale di 100 dati con la tecnica SSR con bordo accurato (a sinistra) e meno accurato (a destra). In mezzo, campo stimato con la tecnica Kriging con variogramma teorico sferico (a sinistra) e di Matérn (a destra). In basso, campo di temperatura misurato dal satellite e regolarizzato.

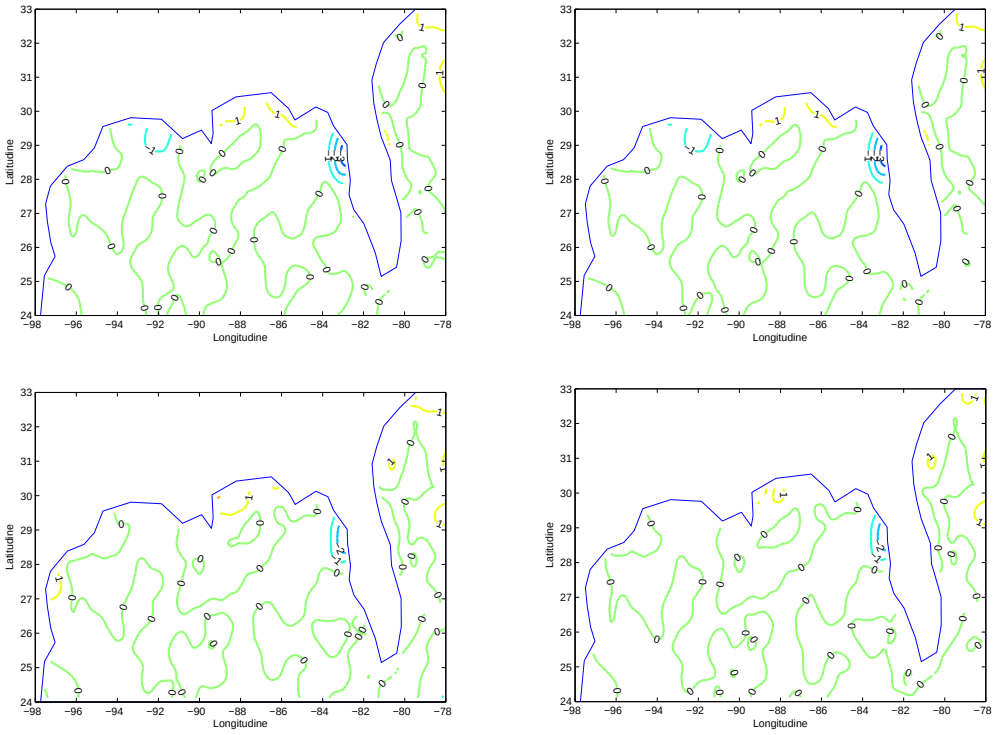


Figura 4.11: Differenza tra i campi stimati col primo campione casuale di 100 dati dal modello SSR con bordo accurato e meno accurato e con le tecniche di Kriging con variogramma teorico sferico e di Matérn e il campo di temperatura regolarizzato misurato dal satellite.

4.5 Confronti con dataset più grandi e più piccoli

In questo paragrafo vengono mostrati i risultati ottenuti seguendo lo stesso procedimento del paragrafo 4.4, ma utilizzando campioni di 200 dati e di 50 dati. L'obiettivo è di verificare come cambino i confronti tra le due tecniche aumentando l'informazione portata dai dati o diminuendola. Le caratteristiche dei 200 dataset sono le stesse del paragrafo 4.4: campioni distanziati almeno di $0,5$ unità tra loro e rispetto ai nodi che formano il bordo, valori dei campioni basati sul dataset regolarizzato e perturbati dallo stesso rumore gaussiano del paragrafo 4.4. L'unica differenza è la numerosità dei campioni e di conseguenza anche la loro posizione.

Campioni di 200 dati

Si considerano prima di tutto campioni di 200 dati. Le triangolazioni che si ottengono con i due bordi sono visibili in figura 4.12; è evidente che con più dati ci siano triangolazioni isotrope, grazie anche all'imposizione che tra due dati ci sia almeno una distanza di $0,5$ unità, che evita la formazione di triangoli troppo piccoli. Confrontando i due bordi, pur avendo triangolazioni molto simili, si può notare che nel bordo meno accurato ci siano alcuni triangoli di forma allungata. Questo aspetto, però, incide molto poco sulle simulazioni. L'indice di condizionamento sulla matrice $\mathbf{A}$ è di un valore dell'ordine di 10^4 per tutte le stime.
I boxplot in figura 4.13 mostrano due caratteristiche principali: la prima è che il divario tra l'errore quadratico medio delle stime con SSR e quello con Kriging è aumentato a favore della prima tecnica; il secondo aspetto è evidente dalla tabella 4.3: i boxplot delle stime con le tecniche SSR con bordi differenti hanno gli stessi quartili del RMSE, così come avviene per le stime con le tecniche di Kriging con i due differenti variogrammi teorici; si può presumere che il numero elevato di dati abbia portato ad una convergenza delle stime.

Tecnica	1° quartile	Mediana	3° quartile
Modello SSR Bordo A	0,2713	0,2832	0,2955
Modello SSR Bordo B	0,2715	0,2827	0,2946
Kriging-Var. Sferico	0,3188	0,3309	0,3459
Kriging-Var. Matérn	0,3190	0,3309	0,3459

Tabella 4.3: Quartili del **RMSE** valutato sulla stima del campo di temperatura con 200 campioni noto il campo misurato da satellite

Considerando le caratteristiche delle stime di punto di vista locale, in figura 4.14 è mostrato il RMSE distribuito spazialmente nel dominio per ogni tecnica di stima utilizzata; si può notare che, oltre che essere diminuito per tutte le tecniche, rimane principalmente concentrato nella fascia costiera settentrionale della Florida, proprio dove la temperatura subisce un aumento repentino a causa del fatto che a pochi kilometri in mare aperto vi è passaggio della Corrente del Golfo. In figura 4.15 sono confrontate le linee di livello dei campi di temperatura stimati con le tecniche SSR e Kriging a partire dal primo campione di 200 elementi; come

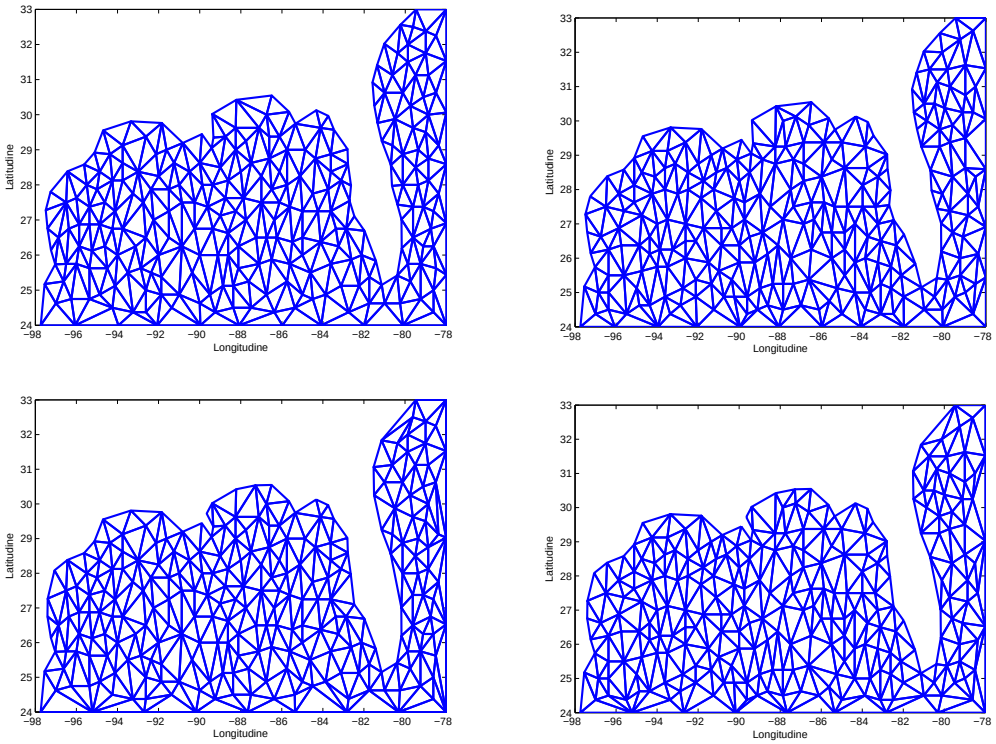


Figura 4.12: In alto, triangolazioni dei i primi due dataset formati da 200 campioni e con bordo più accurato; in basso, triangolazioni dei primi due dataset formati da 200 campioni e con bordo meno accurato.

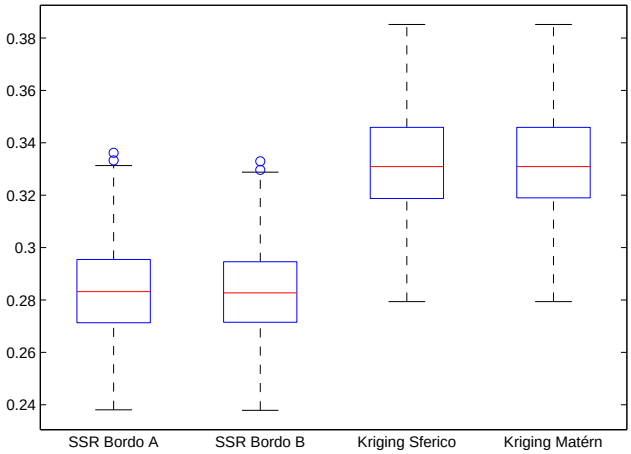


Figura 4.13: Boxplot dei 200 RMSE calcolati dopo aver stimato il campo di temperatura della superficie oceanica con 200 dati con il modello SSR con bordo più accurato (bordo A) e con quello meno accurato (bordo B), con Kriging con variogramma sferico e con Kriging con variogramma Matérn.

termine di confronto sono presenti anche le linee di livello del dataset dei dati misurati da satellite regolarizzato. L'uso di maggiore informazione (200 dati) ha prodotto delle stime migliori per tutte le tecniche, come confermano anche le immagini in figura 4.16. Si prova allora a verificare come cambiano le stime riducendo il numero di dati.

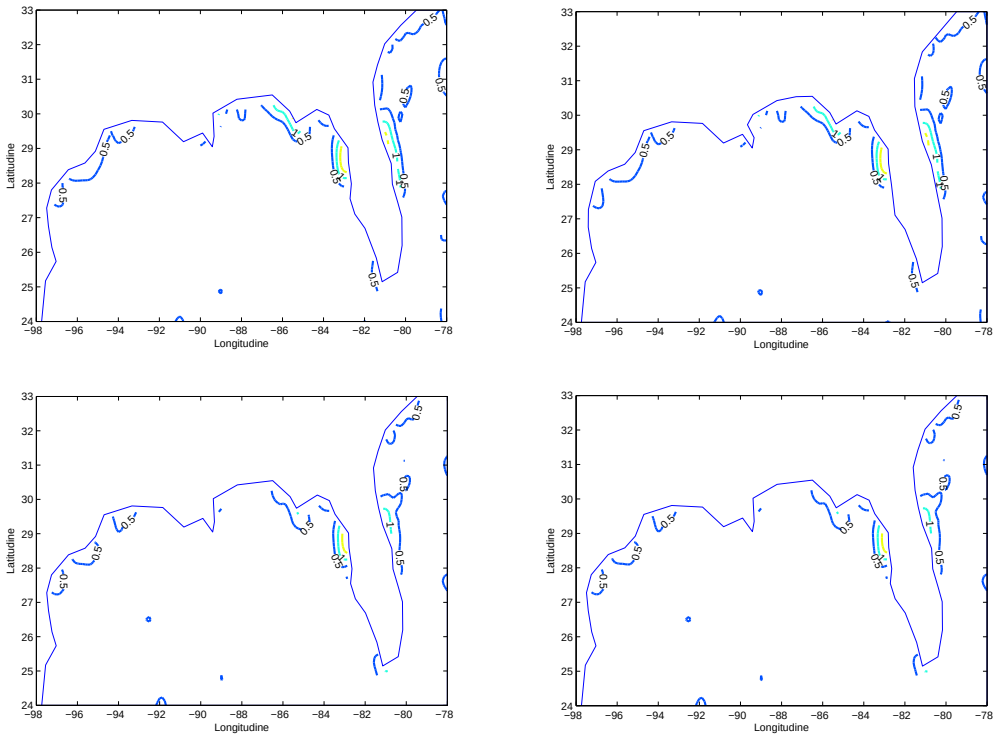


Figura 4.14: Rappresentazione della distribuzione sul dominio del **RMSE** calcolato in ogni nodo della griglia a partire dalle 200 simulazioni con 200 campioni. In particolare, in alto a sinistra, RMSE calcolato a partire dalla stima effettuata con il modello SSR con bordo accurato, in alto a destra con il modello SSR con bordo meno accurato; in basso a sinistra RMSE calcolato a partire dalla stima effettuata con Kriging con variogramma sferico e a destra con variogramma Matérn.

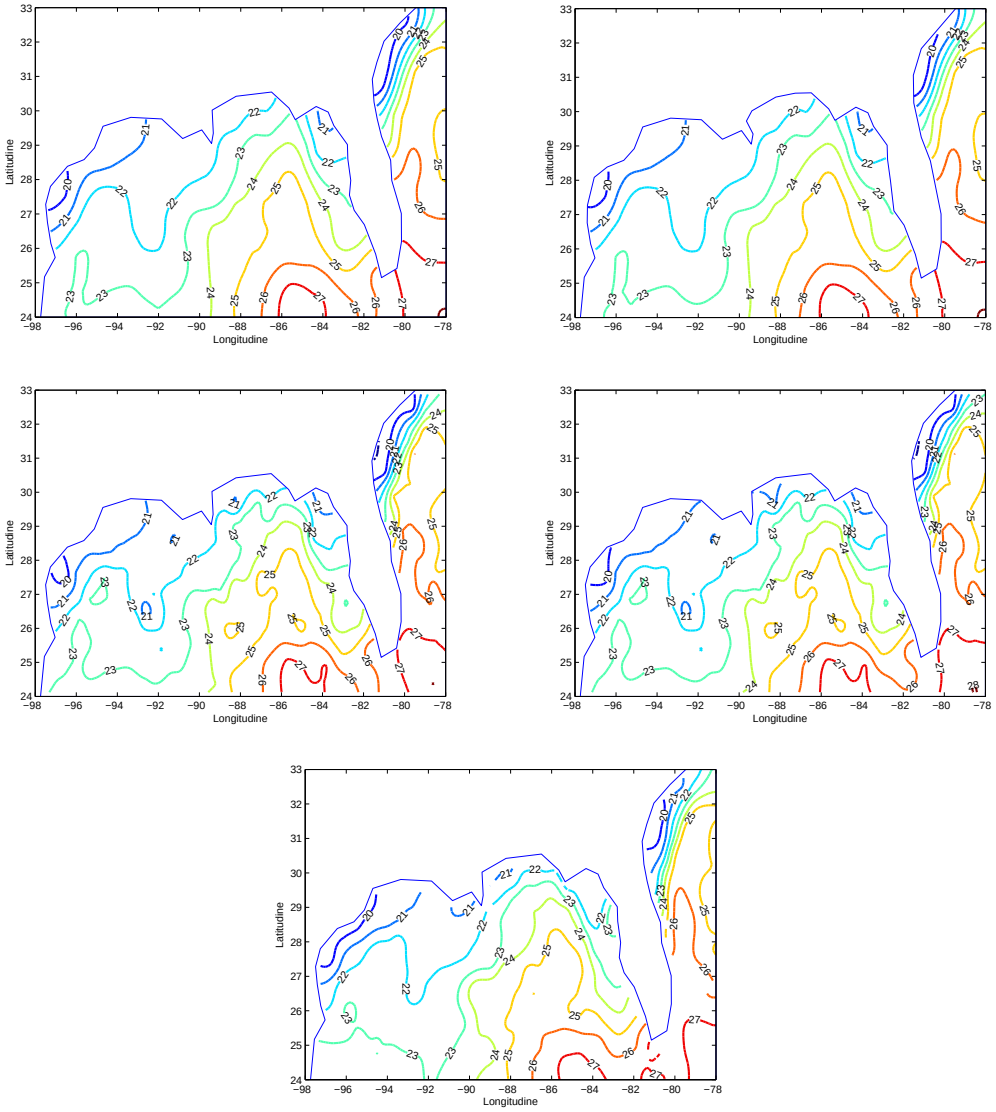


Figura 4.15: In alto, campo stimato col primo campione casuale di 200 dati con la tecnica SSR con bordo accurato (a sinistra) e meno accurato (a destra). In mezzo, campo stimato con la tecnica Kriging con variogramma teorico sferico (a sinistra) e di Matérn (a destra). In basso, campo di temperatura misurato dal satellite e regolarizzato.

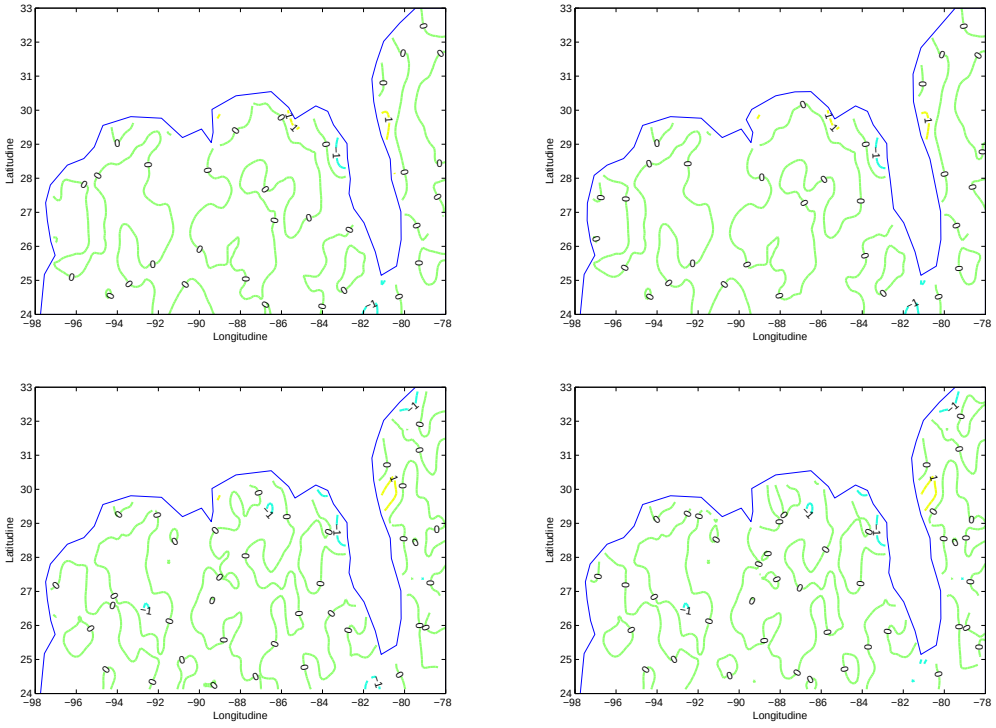


Figura 4.16: Differenza tra i campi stimati col primo campione casuale di 200 dati dal modello SSR con bordo accurato e meno accurato e con le tecniche di Kriging con variogramma teorico sferico e di Matérn e il campo di temperatura regolarizzato misurato dal satellite.

Campioni di 50 dati

Riducendo i campioni a 50 elementi, le triangolazioni perdono in isotropia e compaiono anche triangoli formati solo da nodi di bordo, come si vede in figura 4.17. L'indice di condizionamento della matrice **A** resta comunque dell'ordine di 10^4.
I boxplot relativi al RMSE sono quasi allineati e i valori di mediana sono molto simili tra loro, differenziandosi di qualche millesimo di grado (si veda la tabella 4.4); dalla figura 4.18 si può anche notare come la variabilità dell'errore quadratico medio delle stime con Kriging sia minore rispetto a quella dell'errore delle stime con SSR; un riscontro quantitativo di quanto appena affermato lo si può avere calcolando la differenza interquartile a partire dalla tabella tab:quartili50. La riduzione della numerosità dei campioni comporta anche dei

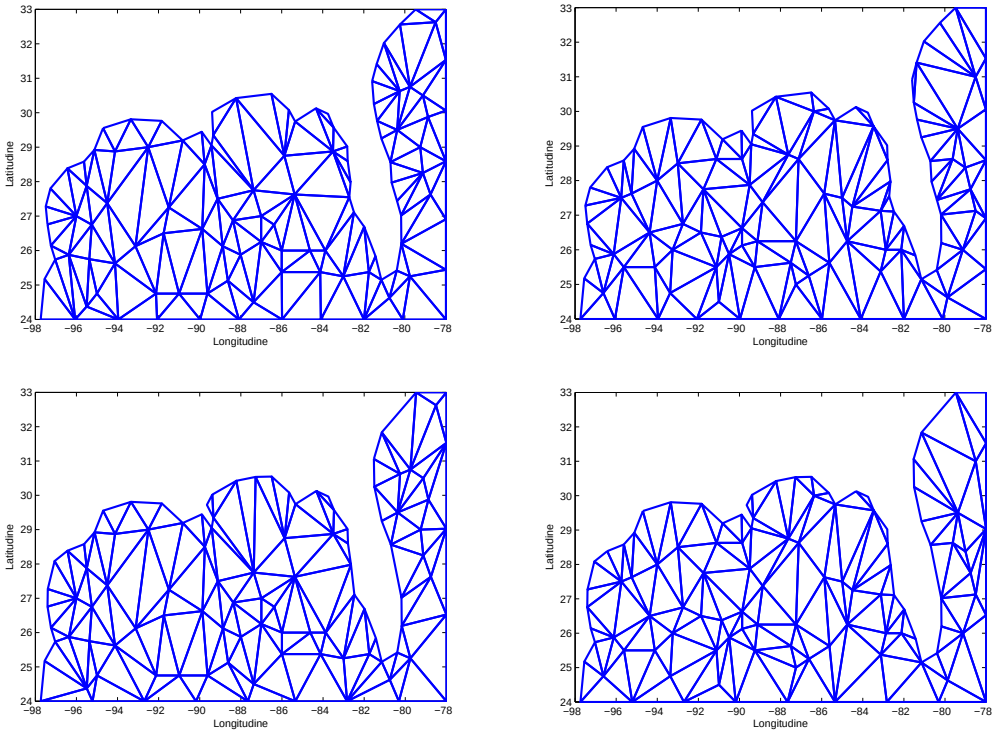


Figura 4.17: In alto, triangolazioni dei i primi due dataset formati da 50 campioni e con bordo più accurato; in basso, triangolazioni dei primi due dataset formati da 200 campioni e con bordo meno accurato.

Tecnica	1° quartile	Mediana	3° quartile
Modello SSR Bordo A	0,6029	0,6911	0,7855
Modello SSR Bordo B	0,5984	0,6898	0,7882
Kriging-Var. Sferico	0,6172	0,6907	0,7575
Kriging-Var. Matérn	0,6089	0,6781	0,7552

Tabella 4.4: Quartili del **RMSE** valutato sulla stima del campo di temperatura con 50 campioni noto il campo misurato da satellite

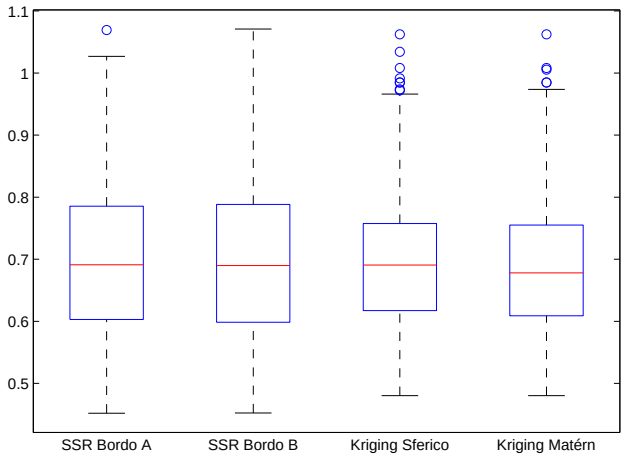


Figura 4.18: Boxplot dei 200 RMSE calcolati dopo aver stimato il campo di temperatura della superficie oceanica con 50 dati con il modello SSR con bordo più accurato (bordo A) e con quello meno accurato (bordo B), con Kriging con variogramma sferico e con Kriging con variogramma Matérn.

cambiamenti dal punto di vista locale, specie in alcune aree: si rileva infatti un notevole aumento del RMSE per entrambe le tecniche, che presentano valori abbastanza elevati soprattutto lungo le coste, in particolare la costa atlantica, come si vede in figura 4.19. Anche in altre aree, però, le tecniche di stima commettono mediamente più errore, come era lecito attendersi.
In figura 4.20 sono confrontate le linee di livello dei campi di temperatura stimati con le tecniche SSR e Kriging a partire dal primo campione di 50 elementi; anche in questo caso, come termine di confronto sono presenti le linee di livello del dataset regolarizzato dei dati misurati da satellite. Come ci si poteva attendere, le stime sono molto più approssimative per entrambe le tecniche, soprattutto nell'area atlantica. La figura 4.21 conferma quanto appena affermato.
I confronti di questo paragrafo mostrano che aumentando il numero di dati si ottengono stime generalmente più accurate con il modello SSR rispetto alle tecniche di Kriging, ma con un numero di dati troppo basso i vantaggi del modello SSR si riducono e le stime delle due tecniche tendono a commettere un errore equivalente.

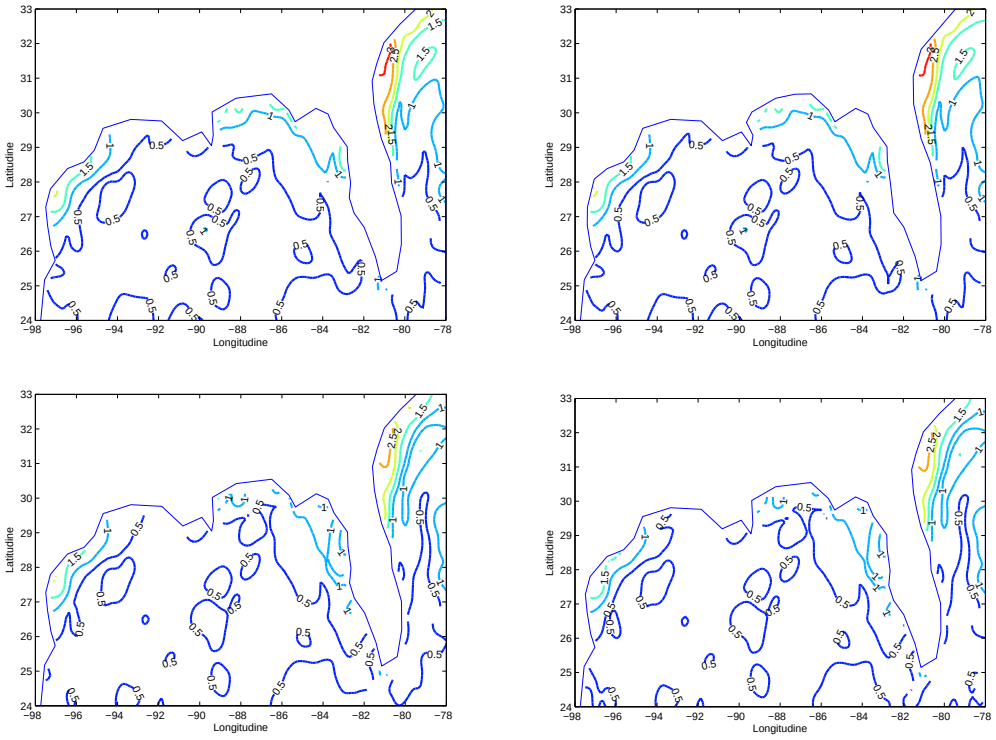


Figura 4.19: Rappresentazione della distribuzione sul dominio del **RMSE** calcolato in ogni nodo della griglia a partire dalle 200 simulazioni con 50 campioni. In particolare, in alto a sinistra, RMSE calcolato a partire dalla stima effettuata con il modello SSR con bordo accurato, in alto a destra con il modello SSR con bordo meno accurato; in basso a sinistra RMSE calcolato a partire dalla stima effettuata con Kriging con variogramma sferico e a destra con variogramma Matérn.

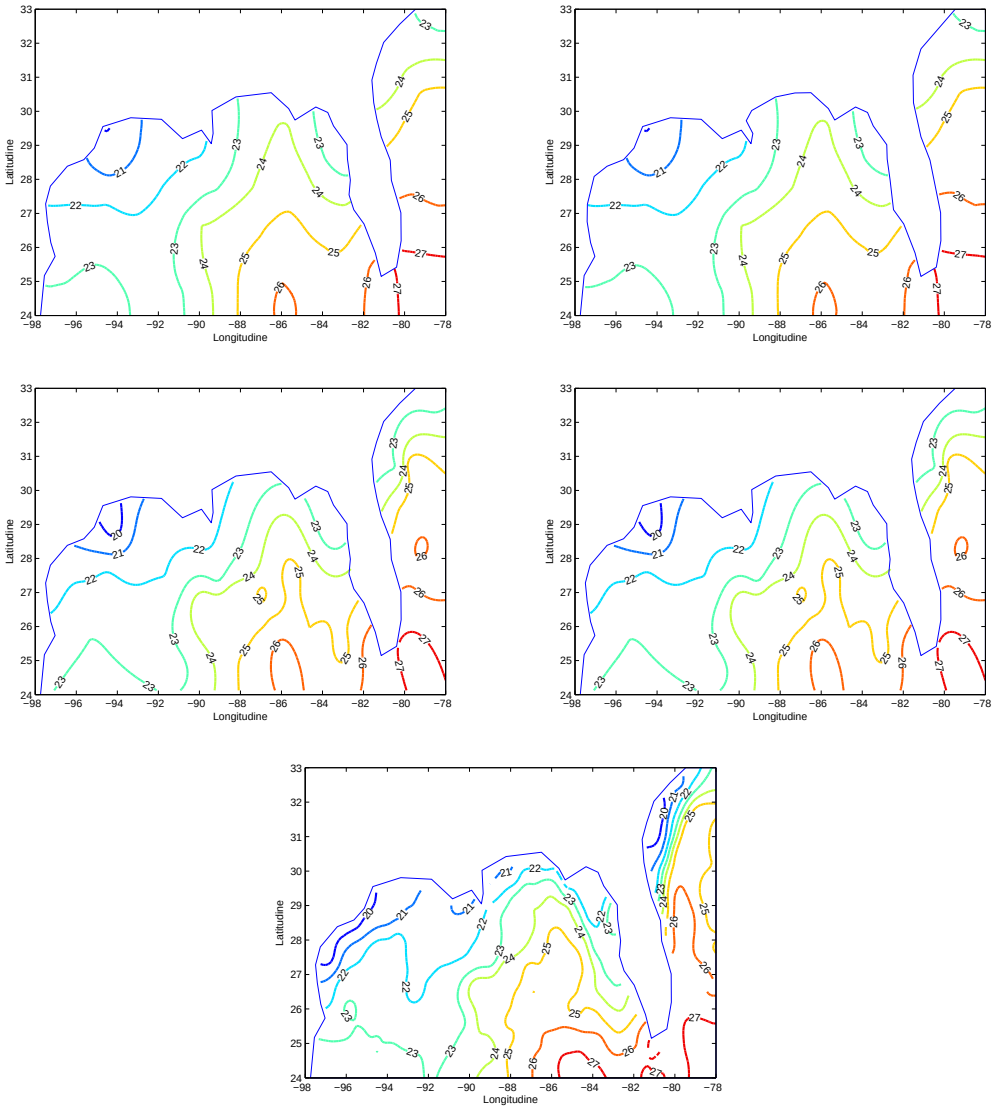


Figura 4.20: In alto, campo stimato col primo campione casuale di 50 dati con la tecnica SSR con bordo accurato (a sinistra) e meno accurato (a destra). In mezzo, campo stimato con la tecnica Kriging con variogramma teorico sferico (a sinistra) e di Matérn (a destra). In basso, campo di temperatura misurato dal satellite e regolarizzato.

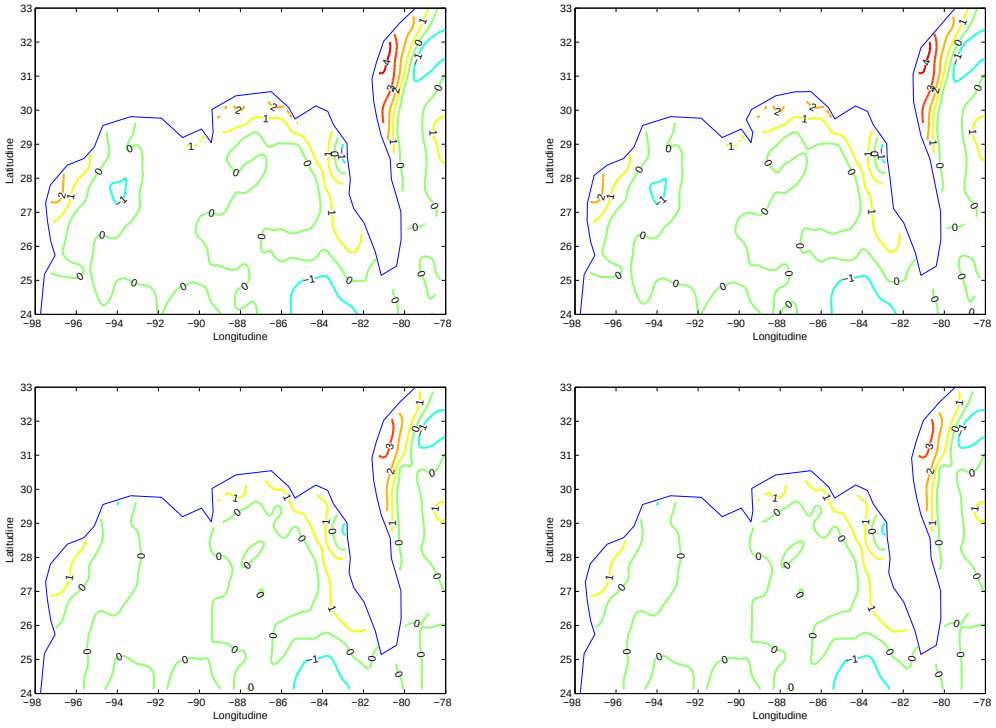


Figura 4.21: Differenza tra i campi stimati col primo campione casuale di 50 dati dal modello SSR con bordo accurato e meno accurato e con le tecniche di Kriging con variogramma teorico sferico e di Matérn e il campo di temperatura regolarizzato misurato dal satellite.

4.6 Risultati con modello SSR con soluzioni in $\mathring{\mathrm{X}}_h^1$

Un ulteriore aspetto che si vuole analizzare è il comportamento del modello SSR nel caso in cui si cerchino soluzioni in $\mathring{\mathrm{X}}_h^1$ (o anche soluzioni di tipo P1): i risultati nel paragrafo 4.4, infatti, sono ottenuti cercando soluzioni con gli elementi finiti nello spazio $\mathring{\mathrm{X}}_h^2$ (soluzioni di tipo P2).
Dal boxplot del RMSE in figura 4.22 e dalla rispettiva tabella dei quartili 4.5 si nota che le differenze sono minime, dell'ordine di qualche millesimo e che sono a vantaggio delle soluzioni P2. Anche nelle analisi locali, la distribuzione del RMSE è molto simile per le differenti stime, come si può vedere in figura 4.23.
In figura 4.24 sono confrontate le linee di livello dei campi di temperatura stimati sia con soluzione agli elementi finiti P1 che P2 a partire dal primo campione di 100 elementi; in figura 4.24 sono presenti anche le linee di livello del dataset regolarizzato dei dati misurati da satellite, per consentire un migliore confronto tra le stime e il campo da stimare. Nel caso di soluzioni di tipo P1, le stime sono molto meno regolari, ma ricalcano abbastanza l'andamento delle stime di tipo P2.
Questi risultati mostrano come, per problemi la cui complessità richieda molto sforzo computazionale, la scelta di utilizzare il modello SSR cercando le soluzioni con gli elementi finiti in $\mathring{\mathrm{X}}_h^1$ possa essere una valida alternativa, consentendo di avere una buona accuratezza con una minore complessità di calcolo. È però importante rimarcare che le buone stime ottenute anche nel caso di soluzioni P1 sono in parte dovute alla triangolazione isotropa (la stessa delle figure 4.6 e 4.7) dovuta al particolare vincolo imposto sul campionamento dei dati, che richiede che ci sia una distanza di almeno $0,5$ unità tra due punti campionati.

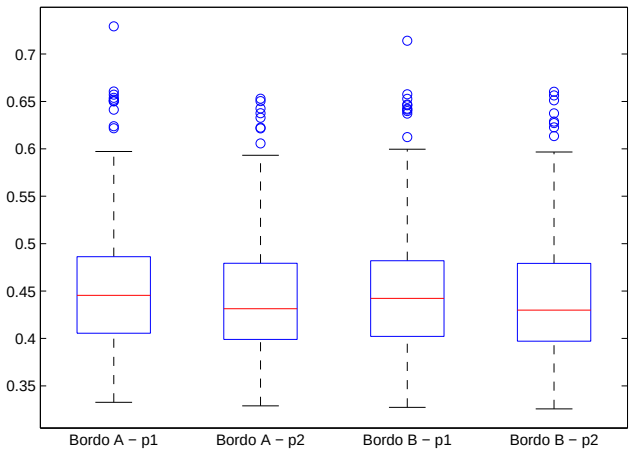


Figura 4.22: Confronto tra i boxplot dei 200 RMSE calcolati dopo aver stimato il campo di temperatura della superficie oceanica con il modello SSR con bordo accurato (bordo A) e bordo meno accurato (bordo B) cercando la soluzione degli elementi finiti di tipo P1 e di tipo P2.

Tecnica	1° quartile	Mediana	3° quartile
Modello SSR Bordo A P1	0,4055	0,4454	0,4862
Modello SSR Bordo A P2	0,3990	0,4313	0,4794
Modello SSR Bordo B P1	0,4021	0,4423	0,4819
Modello SSR Bordo B P2	0,3971	0,4298	0,4792

Tabella 4.5: Quartili del **RMSE** valutato sulla stima del campo di temperatura con 200 campioni noto il campo misurato da satellite.

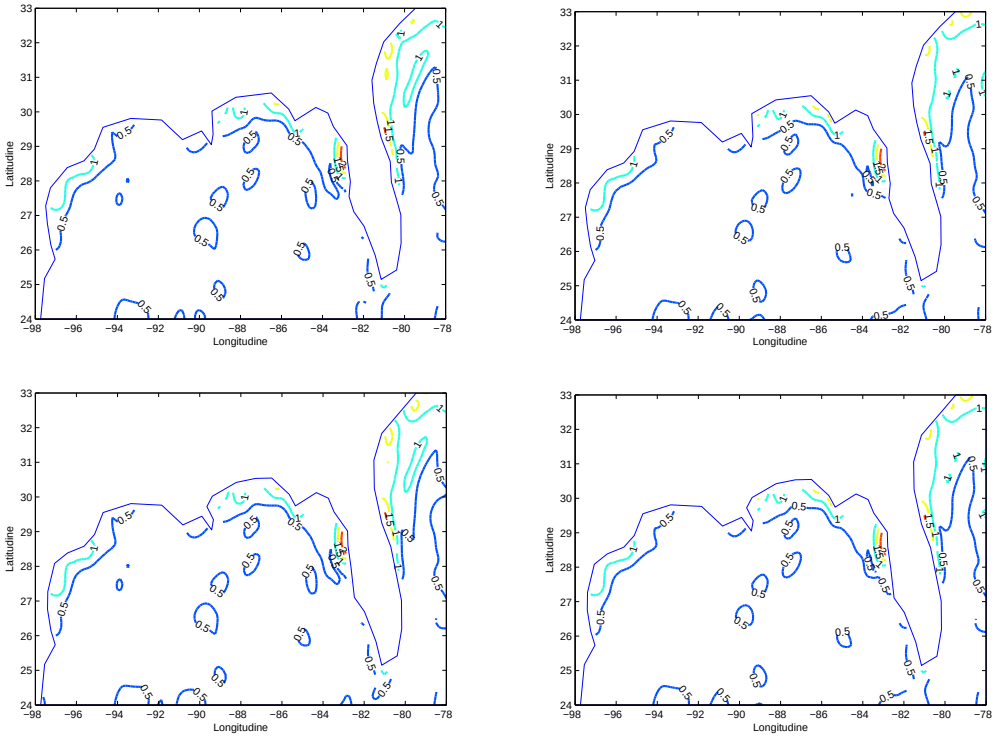


Figura 4.23: Rappresentazione della distribuzione sul dominio del **RMSE** calcolato in ogni nodo della griglia a partire dalle 200 simulazioni con 100 campioni. In particolare, in alto a sinistra, RMSE calcolato a partire dalla stima di tipo P1 effettuata con il modello SSR con bordo accurato e in alto a destra con la stima di tipo P2; in basso a sinistra RMSE calcolato a partire dalla stima di tipo P1 effettuata con il modello SSR con bordo meno accurato e a destra con la stima di tipo P2.

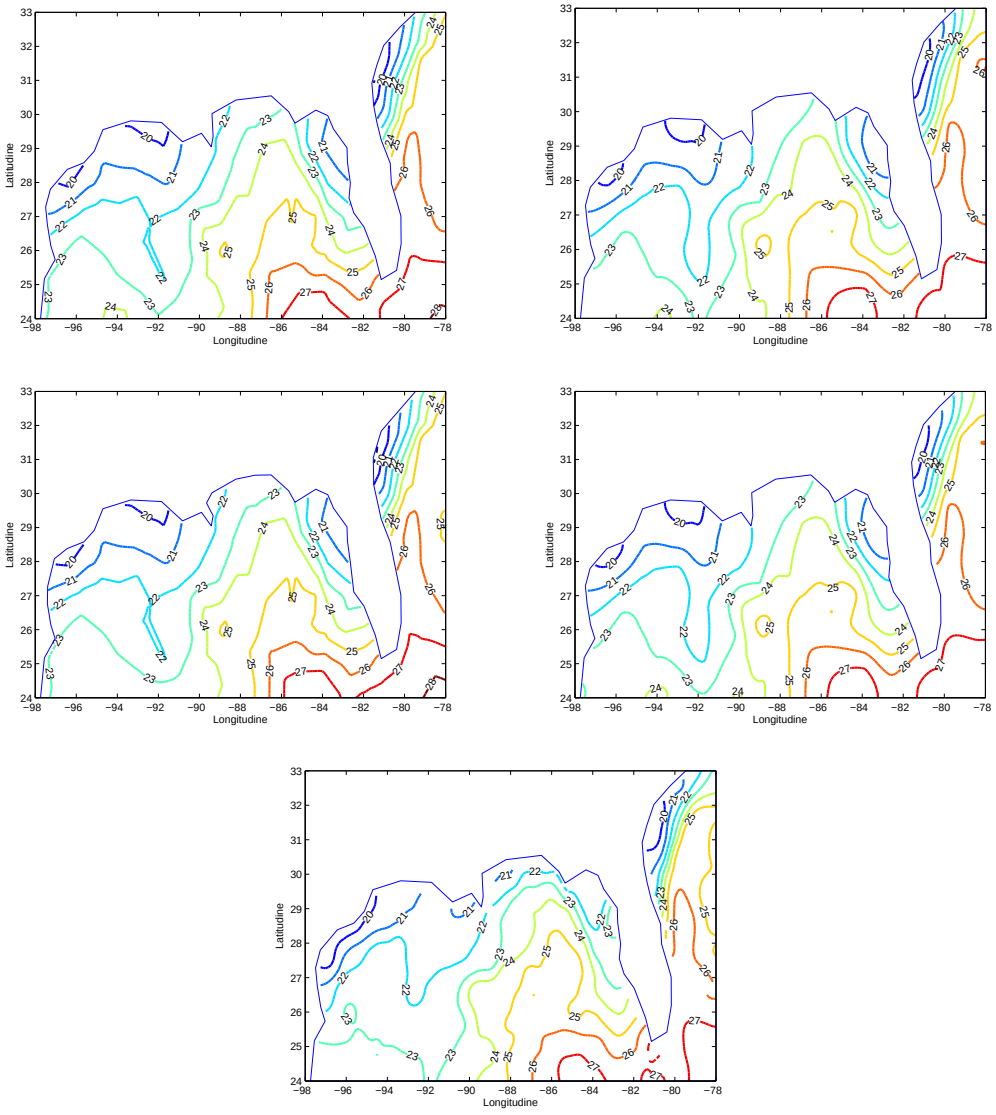


Figura 4.24: In alto, campo stimato col primo campione casuale di 100 dati con la tecnica SSR con bordo accurato con soluzione agli elementi finiti P1 (a sinistra) e P2 (a destra). In mezzo, campo stimato con la tecnica SSR con bordo meno accurato con soluzione agli elementi finiti P1 (a sinistra) e P2 (a destra). In basso, campo di temperatura misurato dal satellite e regolarizzato.

4.7 Confronti con dataset senza vincoli di distanza

I risultati dei paragrafi 4.4, 4.5 e 4.6 sono ottenuti imponendo un vincolo forte sul campionamento dei dati per i 200 dataset: si è infatti richiesto che i punti dovessero essere ad una distanza di almeno $0,5$ unità. In questo paragrafo viene mostrato cosa cambia quando questa richiesta non è soddisfatta. Vengono allora scelti casualmente 200 campioni di 100 dati e si esegue la stessa procedura degli altri paragrafi, confrontando i risultati ottenuti sia con il modello SSR che con le tecniche di Kriging.
Nelle figure 4.26 e 4.27 si possono vedere le triangolazioni dei primi 4 campioni effettuate con i due differenti bordi: la presenza di triangoli molto acuti indica che la triangolazione è piuttosto anisotropa; nonostante questo gli indici di condizionamento della matrice $\mathbf{A}$ sono tutti dell'ordine di 10^4.
L'aspetto più interessante viene dal boxplot del RMSE in figura 4.28: il modello SSR ha i boxplot spostati su valori più elevati rispetto al Kriging e questo indica che tendenzialmente le sue stime sono meno accurate. La tabella 4.6, se confrontata con la tabella 4.2, indica che, se l'eliminazione del vincolo di distanza ha causato una perdita di precisione nella stima con la tecnica SSR, la stessa cosa non è avvenuta con il Kriging; in altre parole, il Kriging non ha risentito della differente disposizione dei punti, ma con lo stesso numero di punti (100 per dataset) ha prodotto stime che commettono circa lo stesso errore (la mediana è sempre circa $0,46$).
L'analisi delle stime dal punto di vista locale presenta un RMSE distribuito nel dominio in modo simile rispetto al paragrafo 4.4 sia per il modello SSR che per le tecniche di Kriging, come si vede in figura 4.29. Si può notare, però, come sia più diffuso nel dominio, essendo comparse delle linee di livello all'interno del Golfo del Messico.
In figura 4.30 è possibile vedere le linee di livello dei campi di temperatura stimati con le tecniche SSR e Kriging a partire dal primo campione di 100 elementi e confrontarle con le linee di livello del dataset regolarizzato dei dati misurati da satellite.

Tecnica	1° quartile	Mediana	3° quartile
Modello SSR Bordo A	0,4587	0,4961	0,5493
Modello SSR Bordo B	0,4555	0,4953	0,5454
Kriging-Var. Sferico	0,4295	0,4676	0,5080
Kriging-Var. Matérn	0,4287	0,4636	0,5059

Tabella 4.6: Quartili del **RMSE** valutato sulla stima del campo di temperatura con 200 campioni noto il campo misurato da satellite

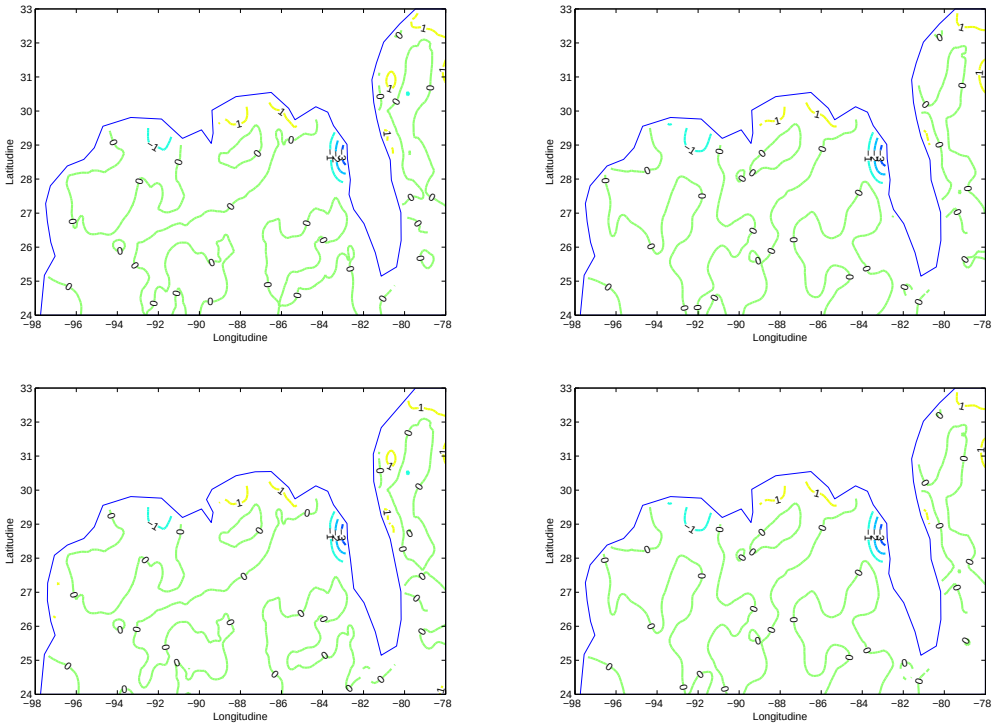


Figura 4.25: Differenza tra i campi di temperatura stimati di figura 4.24 e il campo di temperatura regolarizzato misurato dal satellite.

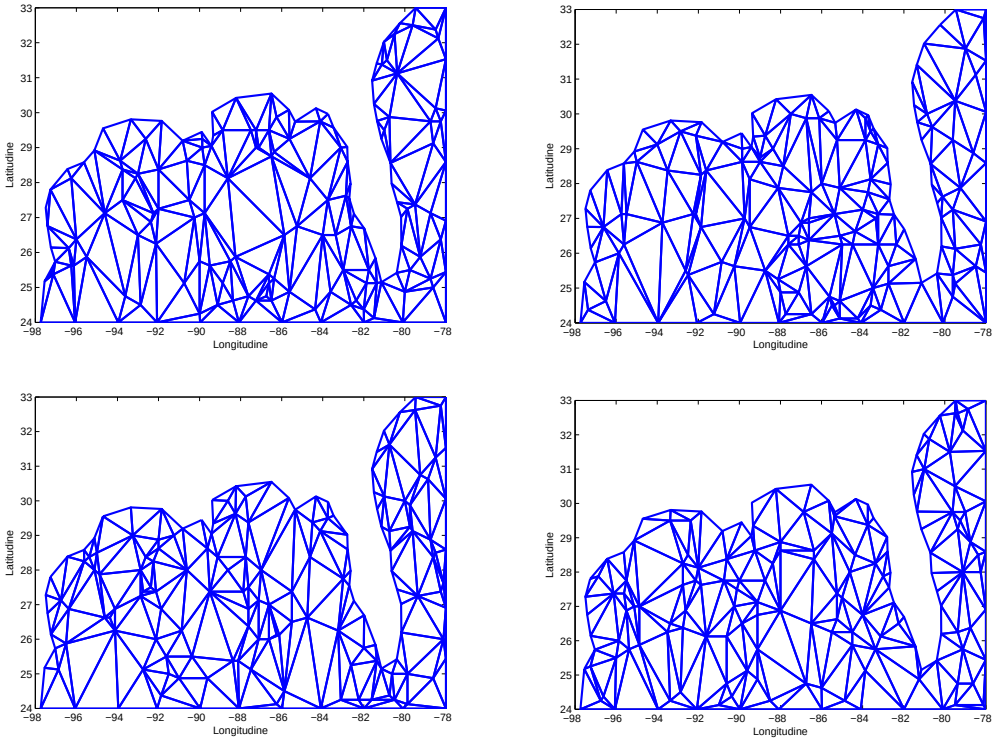


Figura 4.26: Triangolazione di 100 dati casuali dei primi 4 campioni con bordo accurato.

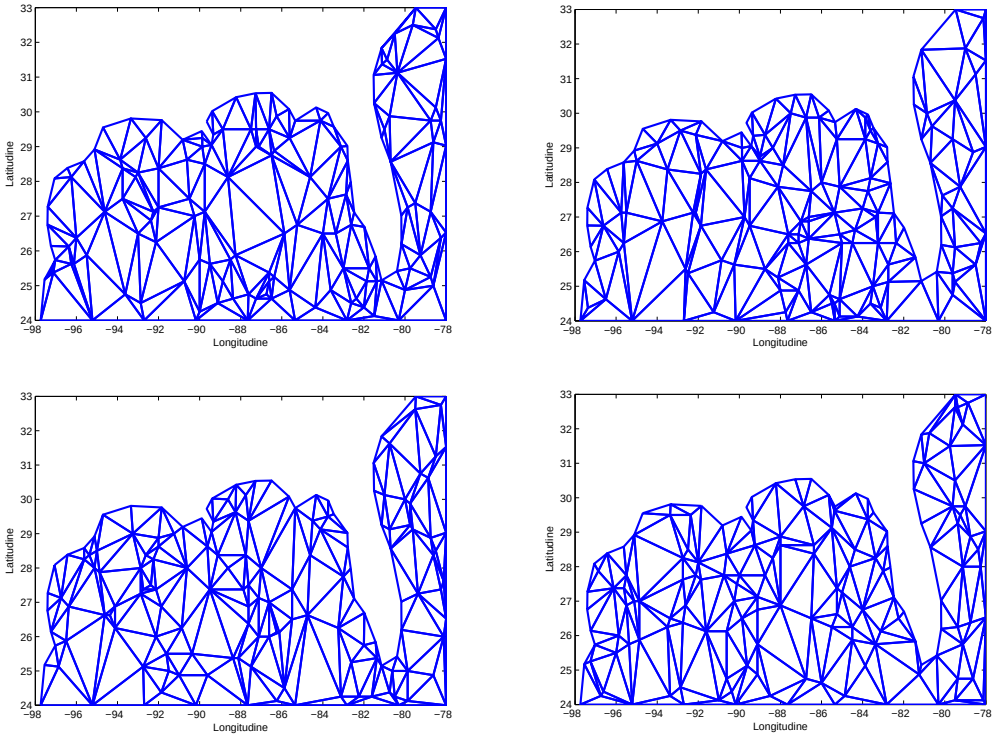


Figura 4.27: Triangolazione di 100 dati casuali dei primi 4 campioni con bordo meno accurato.

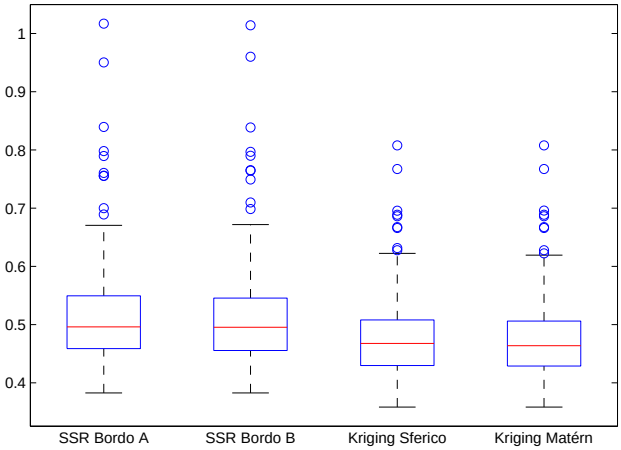


Figura 4.28: Boxplot dei 200 RMSE calcolati dopo aver stimato il campo di temperatura della superficie oceanica con il modello SSR con bordo più accurato (bordo A) e con quello meno accurato (bordo B), con Kriging con variogramma sferico e con Kriging con variogramma Matérn.

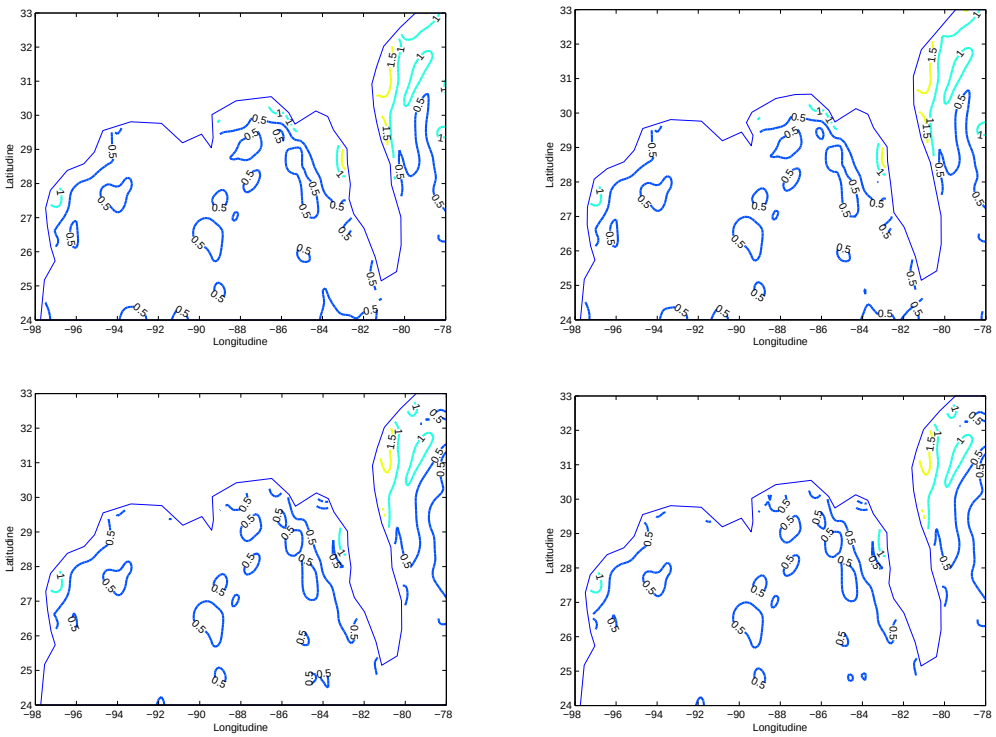


Figura 4.29: Rappresentazione della distribuzione sul dominio del **RMSE** calcolato in ogni nodo della griglia a partire dalle 200 simulazioni con 100 campioni senza vincoli di distanza. In particolare, in alto a sinistra, RMSE calcolato a partire dalla stima effettuata con il modello SSR con bordo accurato, in alto a destra con il modello SSR con bordo meno accurato; in basso a sinistra RMSE calcolato a partire dalla stima effettuata con Kriging con variogramma sferico e a destra con variogramma Matérn.

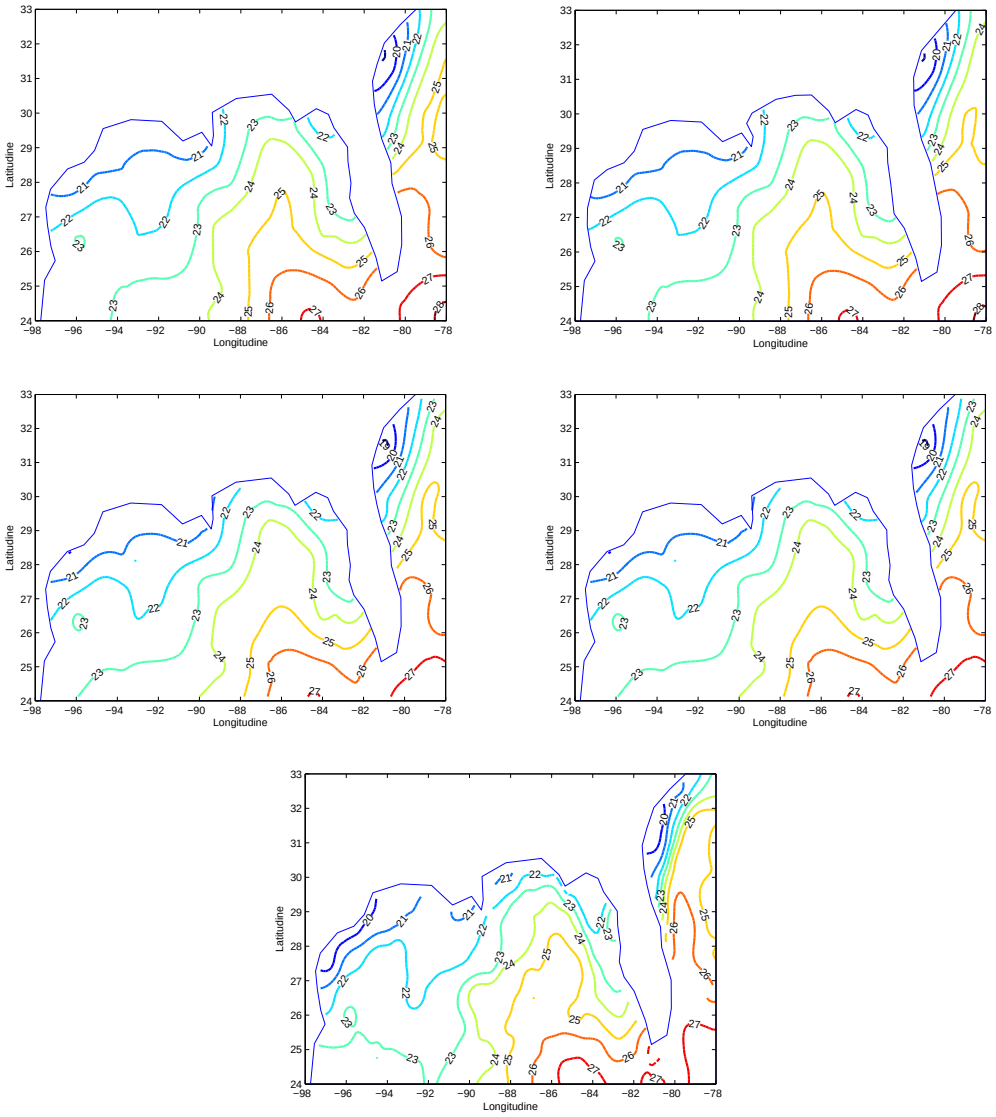


Figura 4.30: In alto, campo stimato col primo campione casuale di 100 dati senza una distanza minima tra loro con la tecnica SSR con bordo accurato (a sinistra) e meno accurato (a destra). In mezzo, campo stimato con la tecnica Kriging con variogramma teorico sferico (a sinistra) e di Matérn (a destra). In basso, campo di temperatura misurato dal satellite e regolarizzato.

4.8 Confronti con dataset non regolarizzato

Fino a questo punto si è sempre utilizzato il dataset di satellite regolarizzato. Ci si può chiedere, allora, quali variazioni possano avere i risultati se si considerasse il dataset di satellite originale, non filtrato dal rumore. Si segue la stessa procedura del paragrafo 4.4 e si utilizzano i campioni posti sugli stessi nodi della griglia di satellite e perturbati dallo stesso rumore del paragrafo 4.4. Il RMSE, di cui sono visionabili i boxplot in figura 4.32, viene calcolato sottraendo dai valori stimati i valori del campo di temperatura non regolarizzato. I risultati sono solo traslati rispetto a quelli del paragrafo 4.4, ma valgono le stesse considerazioni fatte in quel paragrafo.
Per quanto riguarda l'analisi delle stime dal punto di vista locale, la figura 4.33 presenta la comparsa di parecchie "macchie" all'interno del dominio. Quello che può sembrare errore di stima è, però, rumore presente nel dataset di satellite e che nessuna tecnica è in grado di simulare.
Questo aspetto risulta più chiaro visualizzando le linee di di livello dei campi di temperatura stimati con le tecniche SSR e Kriging a partire dal primo campione di 100 elementi, in figura 4.34 e confrontandole con le linee di livello del dataset non regolarizzato dei dati misurati da satellite. Dalla figura 4.35 si può notare come sia presente dell'errore impredicibile, dovuto proprio al dataset di satellite. Le tecniche sono infatti pensate per stimare dei campi regolari; non sarebbe quindi molto sensato confrontarle con un dataset non regolarizzato: sarebbe difficile, infatti, poter definire se l'errore commesso da una tecnica sia un errore di stima o un errore dovuto alle caratteristiche del dataset con cui è confrontato.

Tecnica	1° quartile	Mediana	3° quartile
Modello SSR Bordo A	0,5330	0,5621	0,6113
Modello SSR Bordo B	0,5307	0,5637	0,6152
Kriging-Var. Sferico	0,5745	0,6000	0,6455
Kriging-Var. Matérn	0,5705	0,5968	0,6454

Tabella 4.7: Quartili del **RMSE** valutato sulla stima del campo di temperatura con 200 campioni noto il campo misurato da satellite

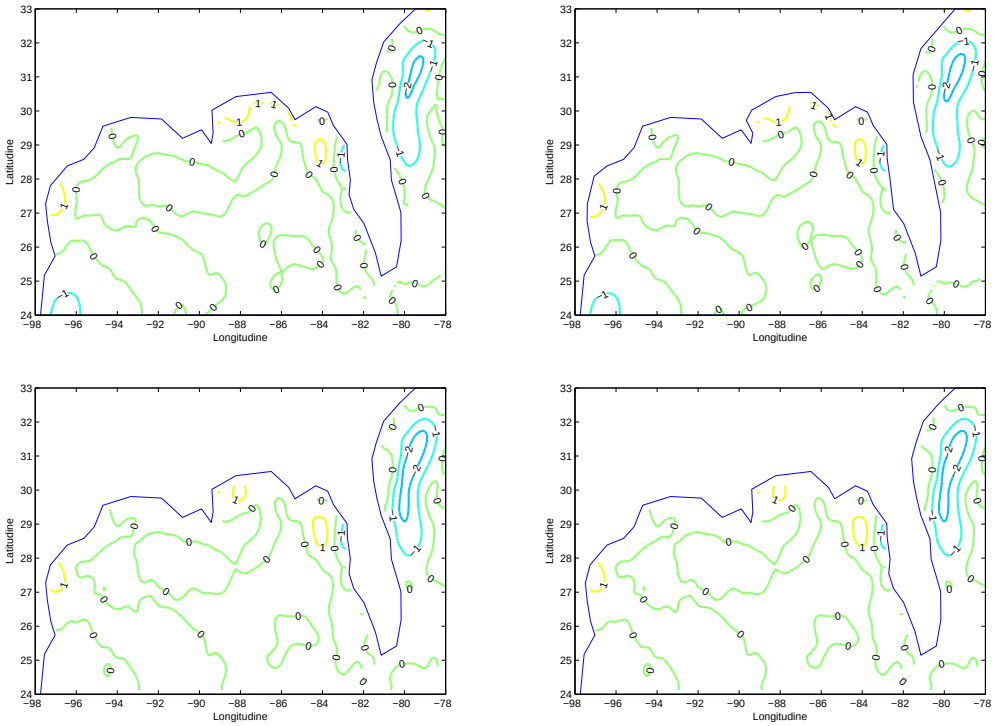


Figura 4.31: Differenza tra i campi di temperatura di figura 4.30 e il campo misurato da satellite e regolarizzato.

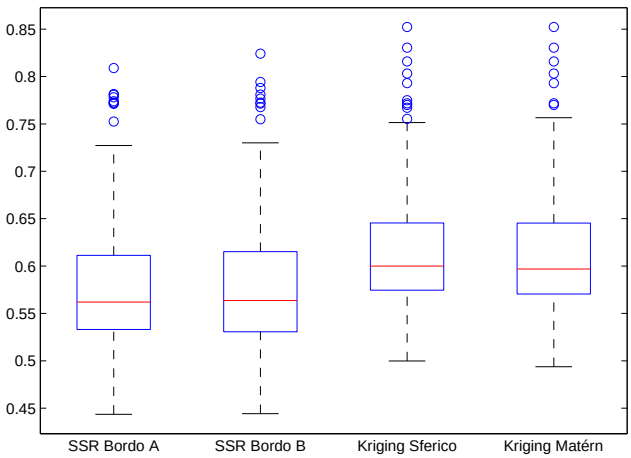


Figura 4.32: Boxplot dei 200 RMSE calcolati dopo aver stimato il campo di temperatura della superficie oceanica con il modello SSR con bordo più accurato (bordo A) e con quello meno accurato (bordo B), con Kriging con variogramma sferico e con Kriging con variogramma Matérn.

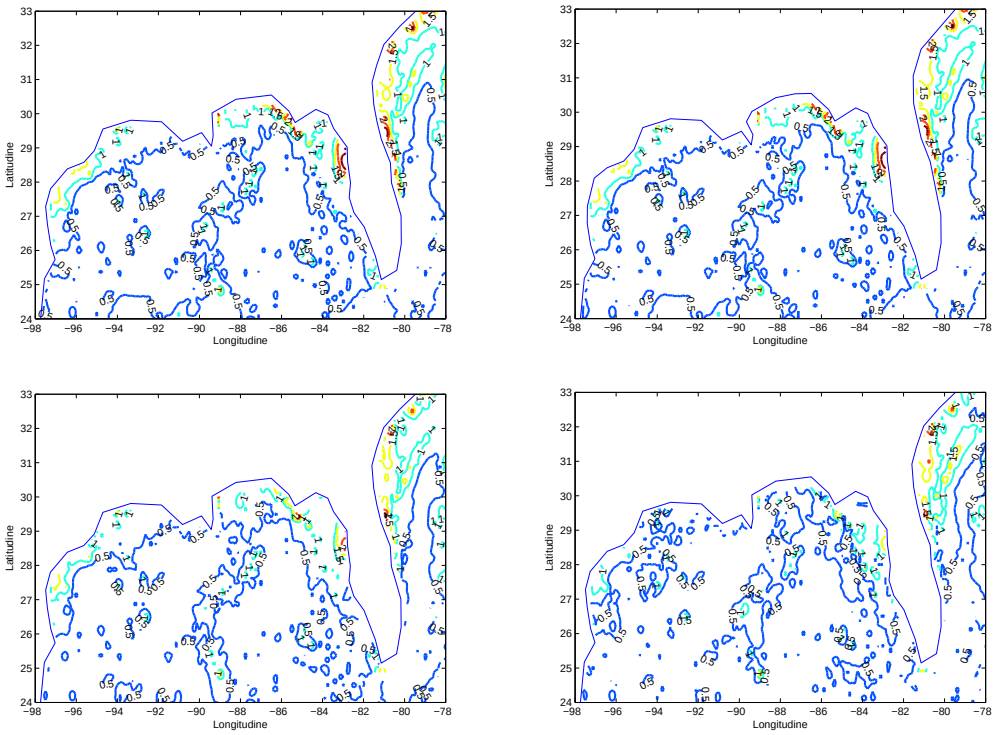


Figura 4.33: Rappresentazione della distribuzione sul dominio del **RMSE** calcolato in ogni nodo della griglia a partire dalle 200 simulazioni con 100 campioni e con dataset non regolarizzato. In particolare, in alto a sinistra, RMSE calcolato a partire dalla stima effettuata con il modello SSR con bordo accurato, in alto a destra con il modello SSR con bordo meno accurato; in basso a sinistra RMSE calcolato a partire dalla stima effettuata con Kriging con variogramma sferico e a destra con variogramma Matérn.

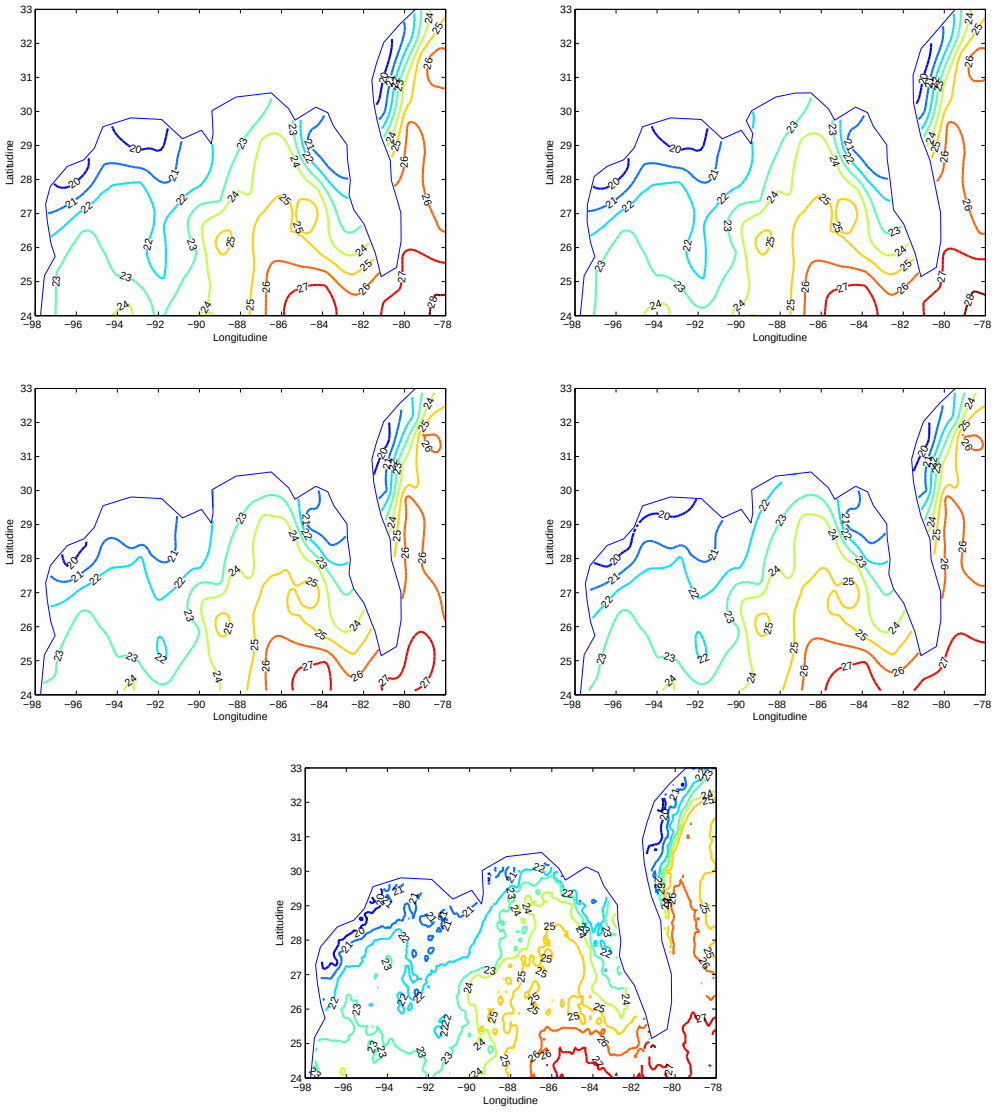


Figura 4.34: In alto, campo stimato col primo campione casuale di 100 dati con la tecnica SSR con bordo accurato (a sinistra) e meno accurato (a destra). In mezzo, campo stimato con la tecnica Kriging con variogramma teorico sferico (a sinistra) e di Matérn (a destra). In basso, campo di temperatura misurato dal satellite e non regolarizzato.

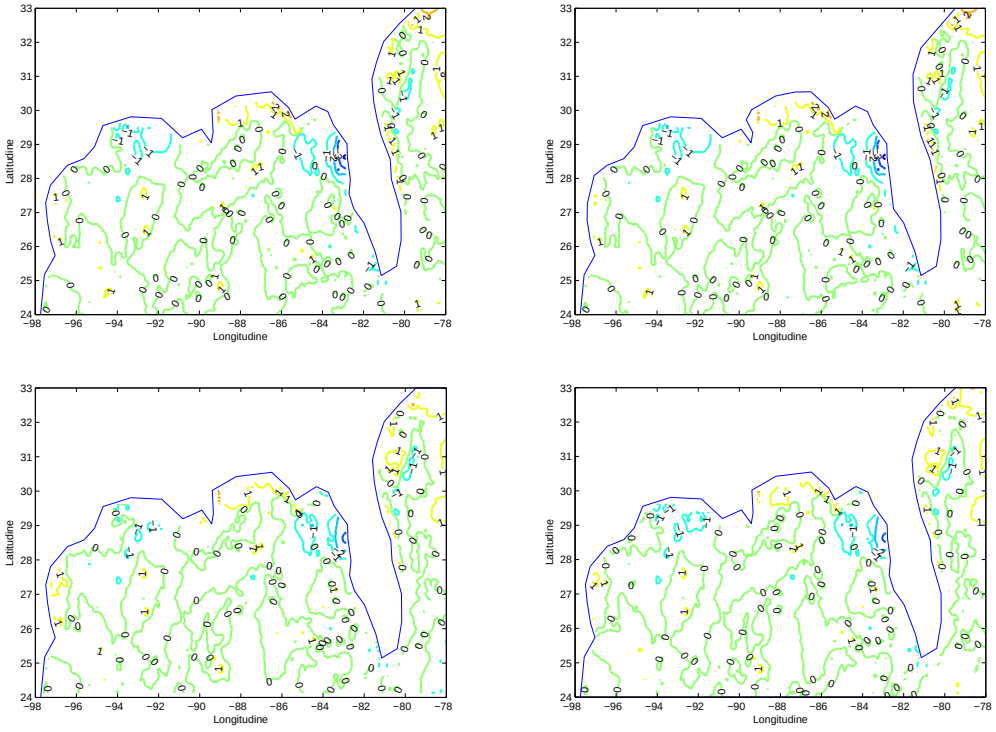


Figura 4.35: Differenza tra i campi stimati col primo campione casuale di 100 dati dal modello SSR con bordo accurato e meno accurato e con le tecniche di Kriging con variogramma teorico sferico e di Matérn e il campo di temperatura non regolarizzato misurato dal satellite.

4.9 Confronto con dati di satellite posizionati in prossimità delle boe

Si prova a sostituire il dataset delle 92 boe con le 92 rilevazioni da satellite più prossime alle boe. In questo modo si può confrontare il dataset del satellite regolarizzato con le stime ottenute da un dataset con la distribuzione delle boe. Anche in questo caso, ai dati selezionati è aggiunto un rumore gaussiano $\mathcal{N}(0; 0, 1)$. In figura 4.36 è presente il nuovo dataset. La triangolazione (in figura 4.37) è sostanzialmente simile a quella ottenuta col dataset delle boe nel paragrafo 3.4: sia col bordo accurato che con quello meno accurato la triangolazione è anisotropa. Dalla figura 4.38 si può vedere che il metodo GCV stima per $\log \hat{\lambda}$ il valore ottimale di $-1, 7$. L'indice di condizionamento è dell'ordine di 10^5, superiore di un ordine rispetto ai paragrafi precedenti. Con il bordo accurato si ottiene una stima della deviazione standard dell'errore ϵ del modello $\hat{\sigma} = 0, 3293°\text{C}$ mentre, con il bordo meno accurato, $\hat{\sigma} = 0, 3333°\text{C}$.

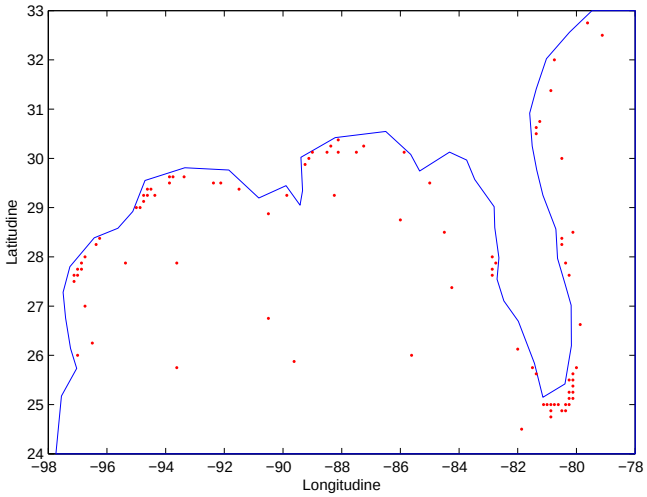


Figura 4.36: In rosso la posizione dei 92 dati di satellite più prossimi ai dati di boa

In figura 4.39 sono confrontati il campo di temperatura misurato dal satellite e regolarizzato e quelli stimati col modello SSR, sia con bordo accurato che meno accurato, e con le tecniche di Kriging con variogramma teorico sferico e di Matérn. Tutte queste stime risentono della posizione non ottimale dei dati: dalla tabella 4.8 si può notare che il valore del RMSE è molto superiore rispetto ai RMSE ottenuti nelle 200 simulazioni del paragrafo 4.4. La collocazione dei dati e di conseguenza anche la triangolazione non ottimale, penalizzano maggiormente la tecnica SSR.

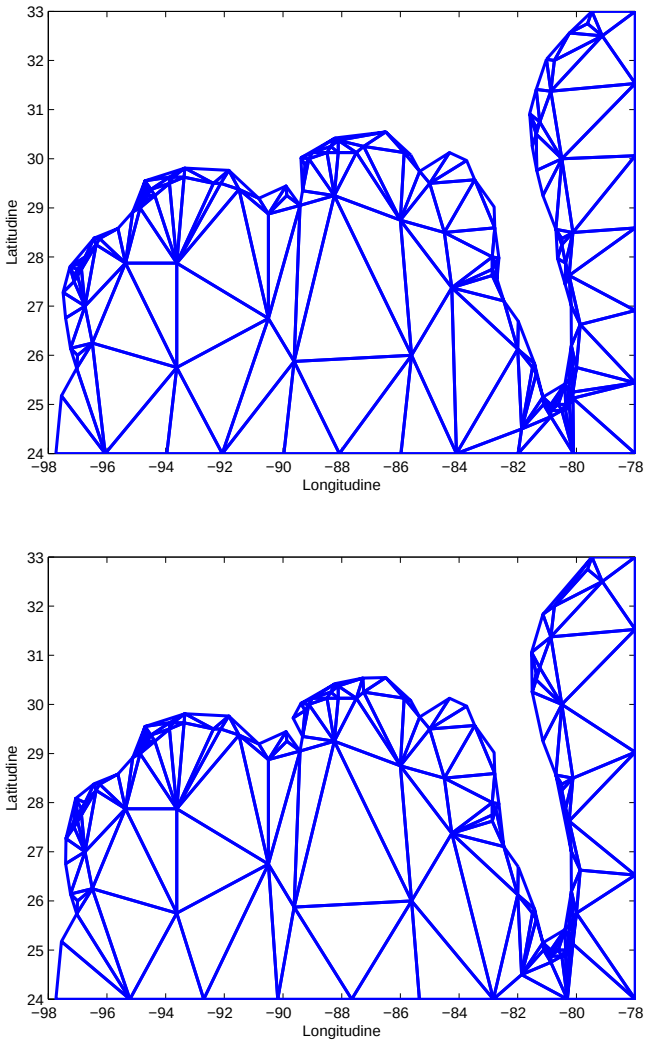


Figura 4.37: Triangolazione con dataset dei 92 dati di satellite posizionati in prossimità delle boe con bordo accurato e bordo meno accurato.

Tecnica	RMSE
Modello SSR Bordo A	0,7561
Modello SSR Bordo B	0,7635
Kriging-Var. Sferico	0,7263
Kriging-Var. Matérn	0,7259

Tabella 4.8: Tabella contenente il valore dell'RMSE tra le quattro stime effettuate e il campo misurato dal satellite.

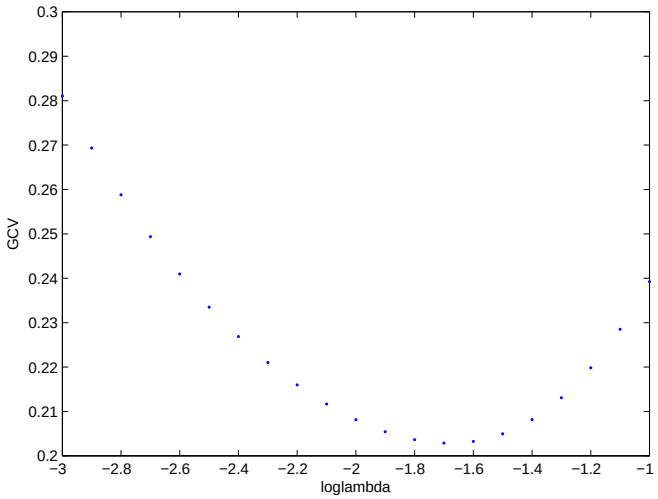


Figura 4.38: Calcolo del valore di $\log \lambda$ ottimale col metodo GCV.

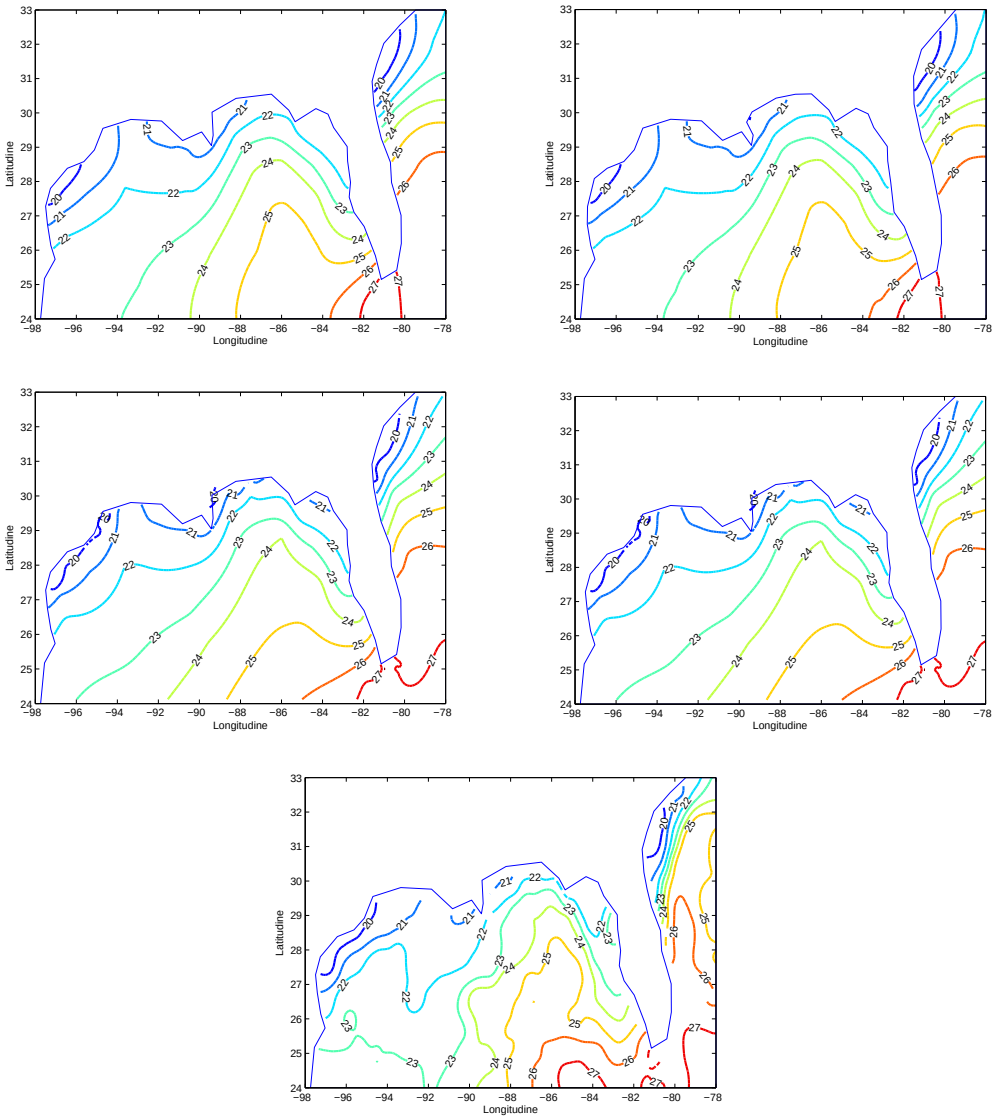


Figura 4.39: In alto, campo stimato con la tecnica SSR con bordo accurato (a sinistra) e meno accurato (a destra). In mezzo, campo stimato con la tecnica Kriging con variogramma teorico sferico (a sinistra) e di Matérn (a destra). In basso, campo di temperatura misurato dal satellite e regolarizzato.

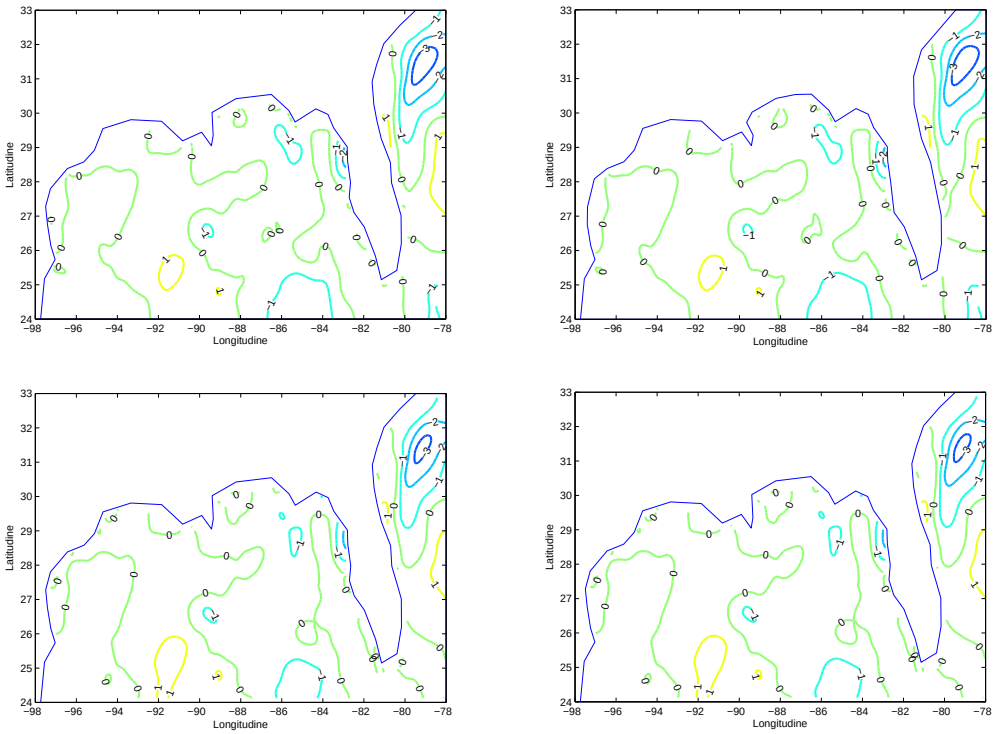


Figura 4.40: Differenza tra i campi stimati dal modello SSR con bordo accurato e meno accurato e con le tecniche di Kriging con variogramma teorico sferico e di Matérn e il campo di temperatura regolarizzato misurato dal satellite.

Capitolo 5

Conclusioni

In molti ambiti delle scienze applicate vi è l'esigenza di ottenere dei campi che descrivano spazialmente una variabile di interesse avendo a disposizione solo un numero limitato di campioni distribuiti nello spazio. Nel corso degli ultimi 50 anni sono state proposte molte tecniche per la risoluzione di questo problema, come le tecniche di *Kriging* ([Cressie, 1991], [Wackernagel, 1995], [Kitanidis, 1992]), le tecniche *thin-plate splines* ([Wahba, 1990]) e di *soap-film smoothing* ([Wood et al., 2008]). Una proprietà auspicabile è che le tecniche tengano conto della forma del dominio per poter ottenere dei campi coerenti con la fisica dei problemi. Recentemente è stato proposto in [Ramsay, 2002] e [Sangalli et al., 2013] un modello di regressione spaziale con spline (SSR) in grado di conciliare questa esigenza con una buona stima del campo della variabile.
In questo elaborato sono state anzitutto presentate la tecnica tradizionale del Kriging e il modello SSR, entrambe analizzate dal punto di vista teorico. La differenza principale tra queste due tecniche è che la prima è basata sul concetto di variogramma e ha l'obiettivo di stimare un campo stocastico sulla base della correlazione tra i dati a disposizione, a prescindere dalla forma del dominio. Questo comporta che, nel caso di domini dalla forma complessa, alcune ipotesi generalmente assunte per effettuare la stima con le tecniche di Kriging, come la stazionarietà del campo stocastico, non siano sempre rispettate. Il modello SSR, invece, ha una natura semiparametrica e concilia una parte di regressione con una parte non parametrica determinata dalla stima di una funzione in grado di attuare un'operazione di smoothing in tutto il dominio.
Poiché l'oceano è vincolato dalle coste e quindi è racchiuso in un dominio di forma complessa, si è proceduto ad una prima applicazione del modello SSR in ambito ambientale su dati oceanografici. In particolare ci si è concentrati su dati di temperatura della superficie oceanica nell'area del Golfo del Messico, caratterizzata dalla presenza della Corrente del Golfo e dalla forma particolare del dominio, dovuta alla principalmente alla Florida, penisola che divide le acque interne del golfo da quelle dell'oceano Atlantico. I dataset disponibili sono stati due: un primo formato da 92 dati rilevati da boe fisse e distribuiti in maniera non ottimale e un secondo rilevato da satellite e formato da 6953 dati, distribuiti secondo una griglia regolare. Entrambi i dataset sono stati rilevati alle ore 22:00 GMT del giorno 15 aprile 2013. A causa della scarsa disponibilità di dati da utilizzare come covariate del modello (pressione e temperatura atmosferica), non

è stato possibile utilizzare il modello SSR nella sua versione completa, ma ci si è dovuti concentrare solo sulla sua parte non parametrica. Si è notato, inoltre, che il dataset di satellite presentasse del rumore che avrebbe potuto aggiungere dell'errore di stima in quanto impredicibile; si è allora rimosso il rumore utilizzando un filtro basato sulla convoluzione.
Se l'idea iniziale era di stimare dei campi di temperatura utilizzando i dati di boa e validare le stime con i dati di satellite, la differenza tra questi due dataset, dovuta soprattutto al fatto che i dati di boa sono molto più concentrati in aree costiere, dove invece mancano rilevazioni di satellite, ha costretto a rinunciare a questo proposito. Si è seguito allora un procedimento diverso per validare il modello SSR e per confrontarlo con la tecnica di Kriging ordinario: estraendo casualmente 200 campioni di 100 dati dal dataset di satellite, dopo averli perturbati con un rumore gaussiano, si è proceduto alla stima del campo di temperatura con le tecniche SSR e Kriging. In questo modo è stato possibile analizzare il RMSE (*Root Mean Squared Error*) sia globalmente che localmente, potendo vedere quali siano le aree dove la stima risulta più difficoltosa. Per ottenere una triangolazione non troppo anisotropa si è imposto che i dati estratti avessero una distanza tra di loro e rispetto ai nodi del bordo di almeno $0,5$ unità (corrispondenti a circa 55 km).
Le stime ottenute con il modello SSR sono state più accurate dal punto di vista globale, presentando valori di RMSE generalmente inferiori. Svolgendo un'analisi locale, si è notato come nelle aree prossime alla costa settentrionale della Florida fossero presenti valori di RMSE mediamente più elevati per tutte le tecniche, in particolare per il modello SSR, indicando che si tratta dell'area più complicata da stimare. Risultati simili si sono ottenuti aumentando la numerosità dei 200 dataset a 200 campioni mentre, riducendo la numerosità a 50 campioni, le due tecniche hanno portato a stime con pari accuratezza. Dal momento che si è utilizzato il modello SSR cercando soluzioni agli elementi finiti di tipo P2, si sono confrontati i risultati con quelli ottenuti con soluzioni P1, mostrandone la somiglianza. Si sono poi confrontati i risultati delle due tecniche rimuovendo il vincolo che i dati estratti dovessero essere posti ad una distanza maggiore di una certa soglia: in questo caso le stime con SSR sono state globalmente meno accurate in termini di RMSE, a causa di triangolazioni meno isotrope, mentre le stime con Kriging hanno mantenuto inalterata la loro accuratezza. Utilizzando dati provenienti dal dataset di satellite non regolarizzato è aumentato il RMSE sia globalmente che localmente; in quest'ultimo caso, inoltre, le linee di livello della distribuzione del RMSE hanno mostrato come questo aumento fosse dovuto proprio alla presenza di rumore nel dataset di satellite, giustificando la scelta fatta nei casi precedenti di filtrarlo.
Si è infine provato a utilizzare i 92 dati di satellite più prossimi spazialmente ai dati di boe verificando che la cattiva distribuzione delle boe porta ad un aumento del RMSE per entrambe le tecniche.
Da questi risultati si può concludere che in questo caso specifico una disposizione sufficientemente distanziata tra i dati nel dominio consente al modello SSR di fornire stime più accurate. Se invece i dati non sono distribuiti in modo da garantire una buona triangolazione, si perdono i vantaggi del modello SSR a vantaggio delle tecniche di Kriging. Inoltre, osservando le stime ottenute con le

tecniche di Kriging, si nota che queste tecniche non hanno tenuto conto della presenza di un dominio complesso: si potrebbe infatti prolungare idealmente le linee a ovest della Florida congiungendole con quelle a est; questo non accade invece con le stime ottenute dal modello SSR. Il fatto che il dataset di satellite utilizzato in questo elaborato presenti delle linee di livello tali per cui idealmente si possano congiungere quelle a est con quelle a ovest della Florida è una possibile spiegazione per il fatto che localmente, in quest'area, le stime con il modello SSR presentino un RMSE più elevato.
Durante le prove di validazione, si sono evidenziate le criticità del criterio GCV (*Generalized Cross Validation*) per la scelta del parametro λ per il modello SSR: l'introduzione di altri criteri di selezione del modello, come AIC e BIC, può essere una possibile prospettiva futura. Inoltre, avendo mostrato la sensibilità del modello SSR alla triangolazione, si potrebbero ripetere le simulazioni utilizzando tecniche differenti per la triangolazione, come quella proposta in [Azzimonti et al., 2013], non vincolata alla posizione dei dati. Restano anche da confrontare le stime del modello SSR nella sua versione completa con le stime delle tecniche di Kriging: questo sarebbe possibile se si avessero a disposizione dati ambientali più completi, che presentassero un numero sufficiente di dati per le covariate.

Appendice A

Metodo degli elementi finiti

Vengono qui presentate alcune definizioni di base e l'idea matematica generale da cui deriva la tecnica degli elementi finiti. Per approfondire l'argomento si consiglia la consultazione di [Salsa, 2004], [Quarteroni, 2008] e [Gockenbach, 2006], riferimenti sui quali è basato questo stesso paragrafo.

A.1 Equazioni a derivate parziali

Il metodo degli elementi finiti si applica principalmente alla risoluzione numerica delle equazioni differenziali a derivate parziali; è quindi utile definire anzitutto cosa si intende per equazione alle derivate parziali:

Definizione 7 *Sia u una funzione in n variabili $(x_1, \ldots, x_n)$. Si dice* ***equazione alle derivate parziali*** *(EDP) l'equazione*

$$F\left(x_1, \ldots, x_n, u, \frac{\partial u}{\partial x_1}, \ldots, \frac{\partial u}{\partial x_n}, \ldots, \frac{\partial^{p_1+\cdots+p_n} u}{\partial^{p_1} x_1 \ldots \partial^{p_n} x_n}\right) = 0 \qquad \text{(A.1)}$$

con $p_1, \ldots, p_n \in \mathbb{N}$

Si dice che l'equazione (A.1) è di *ordine* q se $q = p_1 + \cdots + p_n$, ordine massimo delle derivate parziali. Se l'equazione (A.1) dipende linearmente da u e dalle sue derivate, si dice *lineare*, altrimenti si dice *non lineare*; se l'equazione è non lineare, ma è lineare rispetto alle derivate di ordine massimo, si dice equazione *quasi-lineare* e in tal caso, se i coefficienti delle derivate di ordine massimo sono indipendenti da u, si dice equazione *semi-lineare*.
Vi è inoltre una ulteriore classificazione che viene data alle EDP: se ci si limita al caso di EDP del secondo ordine lineare a coefficienti costanti (ma la classificazione è valida per le equazioni di tutti gli ordini), l'equazione (A.1) diventa

$$A\frac{\partial^2 u}{\partial x_1^2} + B\frac{\partial^2 u}{\partial x_1 \partial x_2} + C\frac{\partial^2 u}{\partial x_2^2} + D\frac{\partial u}{\partial x_1} + E\frac{\partial u}{\partial x_2} + Fu = G$$

con G funzione assegnata e $A, B, C, D, E, F \in \mathbb{R}$. In base al valore del discriminante $\Delta = B^2 - 4AC$, l'equazione sarà classificata come:

- *ellittica*, se $\Delta < 0$;

- *parabolica*, se $\Delta = 0$;
- *iperbolica*, se $\Delta > 0$;

Esempio 1 *L'equazione di Poisson è largamente utilizzata per descrivere fenomeni fisici di diffusione, come la diffusione del calore, oltre che essere usata ampiamente in elettrostatica e rappresenta un ottimo esempio di EDP per poter chiarire i concetti appena introdotti e per applicare il metodo degli elementi finiti. Sia $\Omega \subset \mathbb{R}^2$ un dominio limitato e connesso, $\mathbf{x} = (x_1, x_2)$ la coppia di variabili spaziali e $f = f(\boldsymbol{x})$ una funzione assegnata. L'equazione di Poisson è così formulata:*

$$-\Delta u = f \tag{A.2}$$

Dalle definizioni date in precedenza, si può quindi dire che l'equazione di Poisson è un'equazione lineare, del second'ordine ed ellittica.

Assegnando delle opportune condizioni al contorno, l'equazione di Poisson (A.2) ammette un'unica soluzione. Se si forniscono informazioni relative al comportamento della soluzione u sulla frontiera $\partial\Omega$ del dominio, ad esempio assegnando una funzione g sul bordo e quindi imponendo $u = g$ su $\partial\Omega$, si parla di *problema di Dirichlet*; se invece si hanno informazioni sulla *derivata normale* di u e si assegna come condizione al contorno la funzione $h = \nabla u \cdot \boldsymbol{\nu}$, con $\boldsymbol{\nu}(\mathbf{x}) = (\nu_1(\mathbf{x}), \nu_2(\mathbf{x}))^T$ normale uscente a Ω, allora si parla di *problema di Neumann*.

A.2 Formulazione debole

L'equazione (A.2) è scritta nella cosiddetta *formulazione forte* e la soluzione u deve essere $C^2(\Omega)$. Si vorrebbe però rilassare questa condizione e consentire che le soluzioni siano funzioni meno regolari. È possibile perciò riformulare il problema nella *formulazione debole* o *formulazione variazionale*, che, come si vedrà, ha il vantaggio di poter essere facilmente approssimata col metodo degli elementi finiti. Di seguito, quindi, si mostra come ottenere la formulazione debole del problema scelto come esempio per questo paragrafo, il problema di Poisson (A.2); si considerano in particolare le condizioni al contorno di Dirichlet omogenee:

$$\begin{cases} -\Delta u = f & \text{in } \Omega \\ u = 0 & \text{su } \partial\Omega \end{cases} \tag{A.3}$$

Si moltiplica la (A.3) per una funzione v arbitraria e si integra il tutto sul dominio Ω, ottenendo:

$$-\int_\Omega \Delta u v \, d\Omega = \int_\Omega f v \, d\Omega \tag{A.4}$$

Si consideri ora il *teorema della divergenza*: data una funzione vettoriale sufficientemente regolare $\mathbf{a}(\mathbf{x}) = (a_1(\mathbf{x}), a_2(\mathbf{x}))^T$, allora

$$\int_\Omega div(\mathbf{a}) d\Omega = \int_{\partial\Omega} \mathbf{a} \cdot \mathbf{n} \, d\gamma \tag{A.5}$$

Applicando il teorema della divergenza (A.5) su $\mathbf{a} = (\varphi\psi, 0)^T$ e $\mathbf{a} = (0, \varphi\psi)^T$, dove φ e ψ sono due funzioni, si ottengono le relazioni

$$\int_\Omega \frac{\partial \varphi}{\partial x_i}\psi d\Omega = -\int_{\partial\Omega} \varphi \frac{\partial \psi}{\partial x_i} d\Omega + \int_{\partial\Omega} \varphi\psi n_i \, d\gamma \qquad i = 1, 2 \tag{A.6}$$

Tenendo conto che $\Delta u = div\nabla u = \sum_{i=1}^{2} \frac{\partial}{\partial x_i}\left(\frac{\partial u}{\partial x_i}\right)$, allora si applica la (A.6) e il teorema di Fubini alla prima parte di (A.3) ottenendo:

$$\begin{aligned}
-\int_\Omega \Delta u v \, d\Omega &= -\sum_{i=1}^{2} \int_\Omega \frac{\partial}{\partial x_i}\left(\frac{\partial u}{\partial x_i}\right) v \, d\Omega \\
&\overset{(A.6)}{=} \sum_{i=1}^{2} \int_\Omega \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i} d\Omega - \sum_{i=1}^{2} \int_{\partial\Omega} \frac{\partial u}{\partial x_i} v n_i d\gamma \\
&\overset{(Fubini)}{=} \int_\Omega \sum_{i=1}^{2} \frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_i} d\Omega - \int_{\partial\Omega} \left(\sum_{i=1}^{2} \frac{\partial u}{\partial x_i} n_i\right) v \, d\gamma \\
&= \int_\Omega \nabla u \cdot \nabla v \, d\Omega - \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, d\gamma
\end{aligned}$$

A questo punto è sufficiente considerare funzioni test v nulle al bordo per poter eliminare l'integrale al contorno. Il problema (A.3) diventa:

$$\int_\Omega \nabla u \cdot \nabla v \, d\Omega = \int_\Omega f v \, d\Omega \tag{A.7}$$

Prima di proseguire, bisogna fare alcune considerazioni sullo spazio in cui sono definite u e v: nell'equazione (A.7) compaiono i gradienti delle due funzioni e questo porterebbe a pensare che sia sufficiente scegliere delle funzioni $C^1(\Omega)$. In realtà $C^1(\Omega)$ è uno spazio di funzioni troppo regolari, perché si possono trovare esempi in cui la soluzione u non è derivabile con continuità. Si cerca uno spazio più idoneo che consenta anche l'utilizzo di distribuzioni.

Definizione 8 *Si indica con $\mathcal{D}(\Omega)$ lo spazio delle funzioni infinitamente derivabili e a supporto compatto su Ω:*

$$\mathcal{D}(\Omega) = \{f \in C^\infty(\Omega) : \exists H \subset \Omega, \ compatto: \ supp\, f \subset H\}$$

Si consideri una trasformazione lineare $T : \mathcal{D}(\Omega) \to \mathbb{R}$ e un elemento $\varphi \in \mathcal{D}(\Omega)$, denotando con $\langle T, \varphi\rangle$ il valore assunto da T su φ: la trasformazione lineare T è continua se $\lim_{k\to\infty} \langle T, \varphi_k\rangle = \langle T, \varphi\rangle$, con $(\varphi_k)_{k=1}^{\infty}$ successione di $\mathcal{D}(\Omega)$ convergente a φ.

Definizione 9 *Si chiama **distribuzione** su Ω una qualunque trasformazione $T : \mathcal{D}(\Omega) \to \mathbb{R}$ lineare e continua. Si indica con $\mathcal{D}'(\Omega)$ lo spazio delle distribuzioni su Ω, e viene anche detto duale di $\mathcal{D}(\Omega)$.*

Anche per le distribuzioni esiste il concetto di derivazione:

Definizione 10 *Sia $T \in \mathcal{D}'(\Omega)$ con $\Omega \subset \mathbb{R}^n$, le sue derivate (nel senso delle distribuzioni) sono definite dalla relazione:*

$$\left\langle \frac{\partial T}{\partial x_i}, \varphi \right\rangle = - \left\langle T, \frac{\partial \varphi}{\partial x_i} \right\rangle \qquad \forall \varphi \in \mathcal{D}(\Omega), \;\; i = 1, \ldots, n$$

Dato il multi-indice $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$, le derivate successive sono definite come:

$$\langle D^\alpha T, \varphi \rangle = (-1)^{|\alpha|} \langle T, D^\alpha \varphi \rangle \qquad \forall \varphi \in \mathcal{D}(\Omega)$$

Si hanno allora gli strumenti per poter introdurre gli *spazi di Sobolev*:

Definizione 11 *Si chiama **spazio di Sobolev** di ordine k su Ω lo spazio formato da quelle funzioni di $L^2(\Omega)$ aventi tutte le derivate distribuzionali fino all'ordine k appartenti ad $L^2(\Omega)$:*

$$H^k(\Omega) = \{ f \in L^2(\Omega) : D^\alpha f \in L^2(\Omega), \; \forall \boldsymbol{\alpha} : |\boldsymbol{\alpha}| \leq k \}$$

Gli spazi di Sobolev $H^k(\Omega)$ sono anche *spazi di Hilbert* rispetto al prodotto scalare

$$(f, g)_k = \sum_{|\boldsymbol{\alpha}| \leq k} \int_\Omega (D^\alpha f)(D^\alpha g) d\Omega.$$

Per il problema di Dirichlet (A.3), avendo solo la necessità di operare una sola derivazione, è sufficiente considerare lo spazio $H^1(\Omega)$. Nel caso specifico del problema di Poisson con condizioni di Dirichlet omogenee, si cercano però funzioni $u, v \in H^1_0(\Omega)$, con $H^1_0(\Omega) = \{ v \in H^1(\Omega) : v = 0 \;\; su \;\; \Omega \}$.
Il problema (A.3) può essere quindi scritto in forma debole:

$$\text{trovare } u \in H^1_0(\Omega) : \quad \int_\Omega \nabla u \cdot \nabla v \; d\Omega = \int_\Omega f v \; d\Omega \quad \forall v \in H^1_0(\Omega) \tag{A.8}$$

Si introduce ora la nozione di *forma*:

Definizione 12 *Dato uno spazio funzionale normato V si dice **forma** un'applicazione $a : V \times V \to \mathbb{R}$, che associa ad ogni coppia di elementi di V un numero reale.*

Una forma si dice:

- *bilineare* se è lineare rispetto ai suoi argomenti:
  $a(\lambda u + \mu w, v) = \lambda a(u, v) + \mu a(w, v), \; \forall \lambda, \mu \in \mathbb{R}, \; \forall u, v, w \in V$
- *continua* se $\exists M > 0$: $|a(u, v)| \leq M \|u\|_V \|v\|_V \; \forall v \in V$
- *coerciva* se $\exists \alpha > 0$: $a(u, v) \geq \alpha \|v\|^2_V \; \forall v \in V$

Si riscrive pertanto il problema (A.8) inserendo la forma bilineare, continua e coerciva a e il funzionale lineare e limitato F così definiti:

$$a : H^1_0(\Omega) \times H^1_0(\Omega) \to \mathbb{R}, \quad a(u, v) = \int_\Omega \nabla u \cdot \nabla v \; d\Omega$$

$$F : H^1_0(\Omega) \to \mathbb{R}, \quad F(v) = \int_\Omega f v \; d\Omega$$

ottenendo una nuova formulazione del problema:

$$\text{trovare } u \in H_0^1(\Omega): \quad a(u,v) = F(v) \quad \forall v \in H_0^1(\Omega) \tag{A.9}$$

Si è nelle ipotesi del *lemma di Lax-Milgram*[1] e quindi si può dire che il problema ammette un'unica soluzione.

A.3 Metodo di Galerkin-elementi finiti

Dopo aver scritto il problema in forma debole (A.9), si può provare ad applicare un metodo per la sua risoluzione numerica. Il primo passo è di cercare di approssimare il problema discretizzando lo spazio e i funzionali. Si consideri una famiglia di spazi $V_h \subset H_0^1(\Omega)$ dipendente da un parametro h e di dimensione $N_h < \infty \ \forall h > 0$. Il problema (A.9) si può riscrivere in forma approssimata come:

$$\text{trovare } u_h \in V_h: \quad a(u_h, v_h) = F(v_h) \quad \forall v_h \in V_h \tag{A.10}$$

Il problema (A.10) viene anche detto *problema di Galerkin*. Si sceglie come funzione test v_h una funzione φ_j della base $\{\varphi_j,\ j = 1, \ldots, N_h\}$ di V_h , in questo modo è possibile scrivere $u_h = \sum_{j=1}^{N_h} u_j \varphi_j$, perché anche $u_h \in V_h$, e il problema di Galerkin (A.10) diventa:

$$\sum_{j=1}^{N_h} u_j a(\varphi_j, \varphi_i) = F(\varphi_i) \qquad i = 1, \ldots, N_h \tag{A.11}$$

Denotando con $\mathbf{A}$ la matrice i cui elementi sono $a_{ij} = a(\varphi_j, \varphi_i)$, detta matrice di *rigidezza* o di *stiffness*, e con $\mathbf{f} = (F(\varphi_1), \ldots, F(\varphi_{N_h}))$, il problema (A.11) si può riscrivere come sistema lineare:

$$\mathbf{Au} = \mathbf{f} \tag{A.12}$$

L'approssimazione della funzione u richiede anche un'approssimazione dello spazio $H_0^1(\Omega)$ in cui è definita. Si suddivide perciò il dominio Ω in una partizione τ_h di poligoni tali che $\overline{\Omega} = \bigcup_{K \in \tau_h} K$, avendo indicato con $\overline{\Omega}$ la chiusura di Ω; ogni sottoinsieme K è detto *elemento*. Quando i poligoni sono dei triangoli, si parla di *triangolazione* e un esempio lo si trova in figura A.1. Una qualità desiderabile per ogni elemento $K \in \tau_h$ è che valga la proprietà di *isotropia*: è auspicabile che non ci siano triangoli deformati o allungati, ma che le proporzioni di ogni triangolo siano tali che, inscrivendoci una circonferenza, il rapporto tra il diametro del triangolo (la distanza massima tra due punti del triangolo) e quello della

[1] Il lemma di Lax-Milgram afferma che dato uno spazio di Hilbert V, una forma $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ bilineare, continua e coerciva, un funzionale $F(\cdot) : V \to \mathbb{R}$ lineare e limitato, allora esiste unica la soluzione del problema

$$\text{trovare } u \in V: \quad a(u,v) = F(v) \quad \forall v \in V$$

circonferenza sia sempre inferiore ad una costante $\delta > 0$. Con griglie anisotrope, infatti, le matrici potrebbero essere malcondizionate, amplificando quindi l'errore numerico rendendo la soluzione poco accurata.

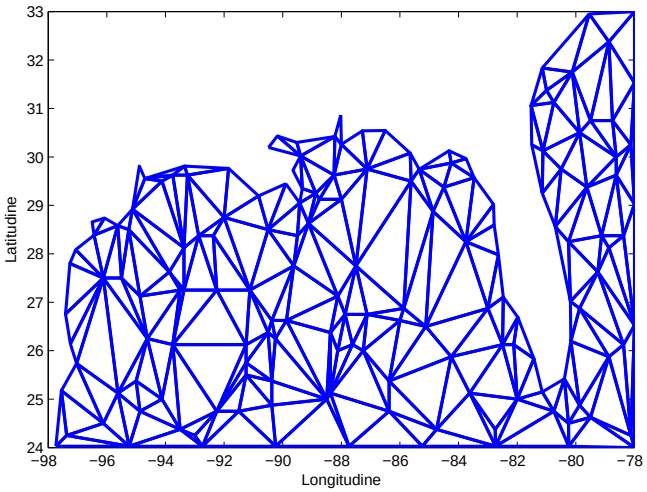


Figura A.1: Esempio di triangolazione

Si consideri ora lo spazio $P_r = \{p(x_1, x_2) = \sum_{i,j \geq 0}$
i+j$\leq r a_{ij} x_1^i x_2^j$, $a_{ij} \in \mathbb{R}\}$ dei polinomi di grado r. Si può verificare che la sua dimensione è

$$dim\, P_r = l = \frac{(r+1)(r+2)}{2}$$

pertanto è sufficiente conoscere l valori della funzione v_h su ogni elemento $K \in \tau_h$ perché v_h sia ben definita.
Si può allora definire lo spazio degli elementi finiti

$$X_h^r = \{v_h \in C^0(\overline{\Omega}) : v_h \in P_r, \forall K \in \tau_h\}$$

composto da funzioni globalmente continue, polinomiali, di grado r in ogni elemento della triangolazione τ_h. In X_h^r è contenuto lo spazio

$$\mathring{X}_h^r = \{v_h \in X_h^r : v_h|_{\partial\Omega} = 0\}$$

Gli spazi trovati X_h^r e $\mathring{X}_h^r$ sono idonei per approssimare $H^1(\Omega)$ e $H_0^1(\Omega)$ in virtù di una proprietà che assicura che se $v \in C^0(\overline{\Omega})$ e allo stesso tempo
$v \in H_1(K) \quad \forall K \in \tau_h$, allora $v \in H_1(\Omega)$ (per ulteriori dettagli si consulti [Quarteroni, 2008], pag. 54).
Per completare l'esempio 1, si può applicare ora il metodo degli elementi finiti al problema di Poisson con condizioni al contorno di Dirichlet omogenee (A.3). Si pone $V_h = \mathring{X}_h^r$ e si definisce $\boldsymbol{\xi}_i$ il vettore dei nodi dell'elemento i della triangolazione τ_h in cui è stato partizionato il dominio Ω, con $i = 1, \ldots, k$, dove k è numero di elementi K. Una base dello spazio V_h è l'insieme delle funzioni $\varphi \in V_h$ tali che

$$\varphi_j(\boldsymbol{\xi}_i) = \delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \quad i, j = 1, \ldots, k$$

Questo significa che se $r = 1$, i nodi sono 3 e corrispondono ai vertici dei triangoli, φ_j è lineare su ogni triangolo e vale 1 nel nodo $\boldsymbol{\xi}_i$ e 0 su tutti gli altri. Se $r = 2$ il discorso è analogo a quanto detto per $r = 1$ con la sola differenza che i nodi sono 6.

Bibliografia

[Azzimonti et al., 2013] Azzimonti, L., Sangalli, L., Secchi, P., Domanin, M., and Nobile, F. (2013). Blood flow velocity field estimation via spatial regression with PDE penalization. Technical Report 19/2013, MOX, Dipartimento di Matematica, Politecnico di Milano.

[Cressie, 1991] Cressie, N. (1991). *Statistics for Spatial Data*. Wiley.

[Crippa, 2007] Crippa, P. (2007). Tecniche di interpolazione geostatistica delle concentrazioni di PM10: applicazione e validazione nel caso del bacino padano. Master's thesis, Politecnico di Milano, Milano.

[Gockenbach, 2006] Gockenbach, M. S. (2006). *Understanding and Implementing the Finite Element Method*. PA: Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

[Kitanidis, 1992] Kitanidis, G. (1992). *Geostatistics*, page 153165. McGraw Hill.

[Quarteroni, 2008] Quarteroni, A. (2008). *Modellistica Numerica per Problemi Differenziali*. Springer-Verlag Italia, Milano.

[Ramsay and Silverman, 2005] Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer.

[Ramsay, 2002] Ramsay, T. O. (2002). Spline Smoothing over Difficult Regions. *Journal of the Royal Statistical Society Ser. B, Statistical Methodology*, 64:307–319.

[Salsa, 2004] Salsa, S. (2004). *Equazioni a derivate parziali - Metodi, modelli e applicazioni*. Springer-Verlag Italia, Milano.

[Sangalli et al., 2013] Sangalli, L. M., Ramsay, J. O., and Ramsay, T. O. (2013). Spatial Spline Regression Models. *Journal of the Royal Statistical Society Ser. B, Statistical Methodology*, 75:1–23.

[Wackernagel, 1995] Wackernagel, H. (1995). *Multivariate Geostatistics*. Springer Verlag, Berlin.

[Wahba, 1990] Wahba, G. (1990). *Spline Models for Observational Data*. PA: SIAM, Philadelphia.

[Wood et al., 2008] Wood, S. N., Bravington, M. V., and Hedley, S. L. (2008). Soap Film Smoothing. *J. R. Stat. Soc. B Stat. Methodol.*, 70:931–955.