POLITECNICO DI MILANO

Scuola di Ingegneria dei Sistemi

Corso di Laurea in Ingegneria Matematica

Geostatistica per Elementi di uno Spazio di Hilbert: Teoria e Applicazioni ai Dati Funzionali

Tesi di Laurea Magistrale

Relatore: Prof. Piercesare Secchi

Correlatore: Dott. Matilde Dalla Rosa

Tesi di Laurea di: :

Alessandra Menafoglio

Matricola 748300

Anno Accademico 2010-2011

Do not let the day end without having grown a bit, without being happy,
without having risen your dreams.
Do not let overcome by disappointment.
Do not let anyone you remove the right to express yourself,
which is almost a duty.
Do not forsake the yearning to make your life something special. [...]
Never stop dreaming, because in a dream, man is free. [...]
Enjoy the panic that leads you have life ahead. Vivel intensely, without mediocrity.

Walt Whitman, *Do not let*

Sommario

Il presente lavoro di tesi si colloca nell'ambito della statistica per dati funzionali georeferenziati. L'obiettivo del lavoro è l'estensione di alcune metodologie geostatistiche, note nel caso di dati vettoriali, alla situazione in cui i dati siano funzionali e spazialmente distribuiti in modo non stazionario, considerando ciascuna osservazione come un punto di uno spazio infinito-dimensionale, assunto di Hilbert, secondo l'approccio tipico della Analisi di Dati Funzionali.

Il problema è affrontato, dal punto di vista teorico, proponendo un'estensione dei metodi di Ordinary e Universal Kriging a coefficienti costanti al caso di elementi di uno spazio di Hilbert, in un quadro coerente di definizioni e ipotesi. L'approccio usato è di tipo astratto, basato su nuove definizioni globali di covariogramma e variogramma, di cui è individuato il legame con la definizione operatoriale di cross-covarianza.

Le proposte algoritmiche del lavoro si focalizzano sulla stima della funzione variogramma e sulla previsione dei dati non osservati. È proposta l'applicazione di metodi bootstrap semiparametrici per l'approssimazione della distribuzione dello stimatore variografico empirico, l'adattamento a un modello valido e la costruzione di stime intervallari, in ipotesi distribuzionali lasche. La bontà dell'approssimazione ottenuta è analizzata con uno studio di simulazione.

È quindi sviluppata una metodologia per la previsione di dati funzionali non stazionari, coerente con i nuovi risultati teorici e consistente in tre fasi: la selezione della forma funzionale per la modellazione del *drift*, il disaccoppiamento del *drift* dal residuo, l'Universal Kriging. La validazione degli algoritmi sviluppati è condotta attraverso l'applicazione a dati sintetici e reali.

La metodologia proposta è applicata al dataset *Canada's Maritime Provinces Temperatures*, che raccoglie le temperature medie giornaliere registrate in 35 stazioni meteorologiche delle province Marittime del Canada. L'analisi geostatistica è condotta adottando una distanza non euclidea sul dominio spaziale e modellando un termine di *drift*. I risultati ottenuti sono confrontati con le mappe di temperatura del Servizio Meteorologico del Canada e con le analisi dello stesso dataset già presenti in letteratura.

Abstract

This work concerns the statistical field of spatially distributed functional data. The aim is the extension of some geostatistical methodologies, already known in vectorial data cases, to non-stationary functional random fields. This purpose is reached with a Functional Data Analysis approach, considering each statistical unit as a point of an infinite-dimensional space, assumed to be Hilbert.

An extension of Ordinary and Universal Kriging methods, based on constant coefficients, to elements of a Hilbert space is proposed, in coherent frame of definitions and hypotheses. The theoretical study is carried out through an abstract approach, based on global definition of covariogram and variogram; moreover, a relation with operatorial definition of cross-covariance is derived.

The algorithms developed in this work are focused on variogram estimation and spatial prediction of functional variables at unsampled location. Strong distributional hypotheses on the functional process are avoided applying semiparametric bootstrap methods for assessing the uncertainty of the variogram empirical estimator. The obtained approximation is used to construct confidence intervals and fit a variogram valid model. The behavior of the proposed algorithms is evaluated through a simulation study on stationary functional data.

Moreover, a methodology for the prediction of non-stationary spatial dependent functional data, consistently with the established theoretical results, is proposed and then developed in three steps: model selection for the drift term, dichotomization of the original process into a deterministic term (the drift) and a residual stochastic process, Universal Kriging prediction. Algorithms validation is carried out through their application to simulated and real data.

The proposed procedures are applied to Canada's Maritime Provinces Temperatures dataset, that collects daily mean temperatures data, observed in 35 meteorological stations located in Canada's Maritimes Provinces. Geostatistical analysis is performed using a non-Euclidean metric on the spatial domain and modeling the deterministic variability through a drift term. The results are finally compared with the reference temperature maps of the Meteorological Service of Canada and with the analysis of the same dataset already presented in literature.

Ringraziamenti

Vorrei rivolgere il mio primo sentito ringraziamento al Prof. Piercesare Secchi, per essere stato per me non soltanto un relatore, ma un vero Maestro, supportandomi in ogni fase di questo lavoro con nuove idee e vivo entusiasmo. Questo lavoro è stato possibile perché ha creduto in me e nelle mie capacità più di quanto abbia fatto io stessa, trasmettendomi con carisma la sua passione per la ricerca, la sua curiosità rispetto alla conoscenza, il rigore formale e il senso applicativo, spronandomi sempre a migliorare me stessa e a superare i miei limiti, valorizzando le mie inclinazioni.

Un grazie di cuore a Matilde, che mi ha sostenuta giorno per giorno, dedicandomi il suo tempo, le sue energie e il suo bagaglio di conoscenze. In questi mesi ha condiviso con me sia l'entusiasmo per i traguardi raggiunti, sia le delusioni per i momentanei fallimenti, aiutandomi, come un'amica, a non scoraggiarmi nei momenti più duri del lavoro: una disponibile tutor, una 'geniale' correlatrice, un preparato ingegnere matematico, ma, soprattutto, una grande persona.

Ringrazio il Dott. Paolo Ruffo, per il costante interesse, la disponibilità e la gentilezza dimostrati nei miei confronti e per la fiducia riposta in me e nel mio lavoro, e tutto il personale di Eni S.p.A., per avermi accolta con simpatia e affiancata con competenza durante il periodo di stage.

Vorrei inoltre ringraziare il Prof. Marco Fuhrman, che in questi anni è stato per me un punto di riferimento importante, sempre disponibile a consigliarmi nei momenti cruciali della mia carriera universitaria e a supportarmi nelle scelte per il mio futuro professionale. Un grazie al MOX-stat, in particolare a Valeria Vitelli e a Simone Vantini, per il supporto di questi mesi.

Un semplice ringraziamento non sarebbe sufficiente a esprimere la mia gratitudine verso la mia famiglia, che mi è stata accanto in questi anni, nei momenti felici e nelle difficoltà, sostenendomi con forza nel momento più buio della mia vita e aiutandomi a riscoprire me stessa e il mio carattere, dando luce ai miei pensieri: a Mamma, Papà e Paolo ed ai miei meravigliosi nonni devo la parte più preziosa di quello che sono ora.

Ringrazio i miei stupendi amici, quelli di una vita e quelli nuovi, che hanno colorato in modo vivace il mio percorso di vita e di studio. Un grazie a Sheyla, che con il suo sorriso solare ha rallegrato tanti momenti in famiglia. Un grazie a Paoly e Mary, per il nostro trio inseparabile. Un grazie alla mia cara amica Anna, perché, in tanti anni, ha sempre capito i miei pensieri meglio di me, aiutandomi a trovare le parole per dar loro voce. Un grazie a tutte le mie compagne di viaggio e in particolare a Chiara, al mio fianco fin dal primo giorno di università, a Lisa, con la quale ho condiviso le soddisfazioni e i timori per i progetti universitari e di vita, ricevendo sempre un appoggio disinteressato e completo, a Jemmina, Fede, Roberta e Cri, per i tanti momenti passati insieme e i pranzi gustati con 'estrema calma' alla dolce sosta: senza di voi questi anni non sarebbero stati così belli.

E infine un grazie a Saverio, perché illumina ogni giorno della mia vita, rendendolo speciale; perché l'essere l'uno accanto all'altra non è una limite, ma la vera libertà; perché gli ho affidato tutto il mio cuore e lui se n'è preso cura. "*Io e te, come le stelle*".

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Il presente lavoro di tesi si colloca nell'ambito della statistica per dati funzionali georeferenziati. Questo campo di ricerca, di recente sviluppo, sorge dal connubio tra alcune metodologie statistiche classiche, applicate a dati spazialmente correlati, e l'approccio dell'Analisi di Dati Funzionali (FDA). Questo lavoro nasce in un ambito di ricerca non solo molto innovativo da un punto di vista scientifico, ma anche di notevole interesse industriale: il progetto è stato infatti proposto e sviluppato in collaborazione con Eni S.p.A., con la finalità di generare nuove e originali metodologie geostatistiche per l'analisi di dati funzionali, da applicare a un problema di geofisica durante un periodo di stage presso l'azienda.

La geostatistica è una branca della statistica applicata che, dalla sua nascita nel 1962 con il lavoro di Matheron (1962), ha ottenuto attenzione sempre crescente in letteratura grazie alla sua estrema utilità nel trattamento di dati distribuiti nello spazio e nel tempo: l'obiettivo della geostatistica è infatti quello di fornire metodologie statistiche che consentano la descrizione quantitativa di variabili naturali distribuite nello spazio, o nello spazio e nel tempo. L'ipotesi chiave su cui si basa la geostatistica è l'esistenza di un processo stocastico sottostante al fenomeno, dalle cui osservazioni inferire dapprima sulla struttura di variabilità spaziale del campo, quindi sul dato non osservato.

In questo contesto, i metodi di kriging consentono di costruire, sulla base di un modello di variabilità spaziale, il previsore BLUP (Best Linear Unbiased Predictor) della variabile di interesse nei siti in cui non è disponibile l'osservazione diretta. Si distinguono tre tipi di kriging in relazione alla stazionarietà o meno del processo considerato (Cressie, 1993): *Simple Kriging* (SK), *Ordinary Kriging* (OK) o *Universal Kriging* (UK); i primi sono finalizzati alla previsione nel caso stazionario, l'ultimo trova applicazione nell'analisi di fenomeni non stazionari.

L'obiettivo della tesi è l'estensione di queste metodologie al caso in cui i dati spazialmente distribuiti siano curve, considerando ciascuna osservazione $\{\chi(t), t \in \mathcal{T}\}$ come un punto in uno spazio infinito dimensionale H, assunto di Hilbert. L'approccio adottato è dunque quello tipico della FDA, nata negli anni '90 del secolo scorso in risposta alle nuove esigenze nel trattamento di osservazioni che si presentino come curve, o siano registrate su griglie fitte tanto da permetterne la rappresentazione funzionale.

L'estensione delle metodologie geostatistiche all'analisi di dati funzionali sorge come naturale risposta alla necessità di analisi ed elaborazione di dati provenienti da problemi in

ambito industriale e, in generale, nelle scienze ambientali, trovando applicazione nei campi dove già la geostatistica detiene un ruolo cruciale per fornire una descrizione quantitativa delle variabili in oggetto. Inoltre, in questi contesti, frequentemente si presenta la necessità di analizzare curve che presentino un comportamento spazialmente non stazionario; basti pensare alle applicazioni geofisiche, laddove i fenomeni naturali in analisi sono molto complessi e raramente sono uniformi sull'intero dominio spaziale. Per questo motivo una parte consistente del lavoro di tesi è dedicata all'analisi di processi funzionali non stazionari.

Il problema proposto è affrontato in primo luogo dal punto di vista teorico. Infatti, la teoria relativa ai processi stocastici funzionali e, in particolare, lo sviluppo di tecniche di kriging e cokriging sotto ipotesi di stazionarietà, è stata sviluppata in tempi molto recenti ed è tuttora in via di consolidamento (Goulard e Voltz, 1993), (Monestiez e Nerini, 2008), (Giraldo *e altri*, 2010a). D'altra parte, sebbene la previsione di dati funzionali georeferenziati spazialmente non stazionari sia molto rilevante ai fini applicativi, in letteratura non è ancora presente una metodologia di kriging non stazionario che la consenta.

Il primo obiettivo del lavoro di tesi è dunque quello di fornire una formulazione di sintesi dei metodi di Ordinary e Universal Kriging a coefficienti costanti all'interno di un quadro coerente ed elegante di definizioni e ipotesi, quali la stazionarietà e l'isotropia. Questo scopo è perseguito in questo lavoro attraverso l'introduzione di definizioni globali di covarianza spaziale e variogramma, individuandone inoltre il legame con le definizioni operatoriali presenti in letteratura (Hörmann e Kokoszka, 2011). Sono quindi riformulati i metodi di kriging stazionario e non stazionario, estendendo la formulazione di Ordinary Kriging esistente (Giraldo *e altri*, 2008a) a processi stocastici su un qualunque spazio di Hilbert, e proponendo un metodo di Universal Kriging per elementi spazialmente non stazionari di uno spazio di Hilbert. In particolare, la formulazione del metodo di Universal Kriging è basata sull'assunzione che il processo non stazionario ammetta una decomposizione in un termine di variabilità deterministica, il *drift*, descritto da un modello lineare con coefficienti funzionali spazialmente costanti, e un termine di variabilità stocastica, il residuo, ipotizzato spazialmente stazionario.

Dal punto di vista applicativo, il lavoro risponde a due esigenze: in primo luogo la stima del variogramma empirico nel contesto dei dati funzionali (*trace-variogram*); in secondo luogo la previsione di dati funzionali spazialmente non stazionari sulla base dei risultati teorici stabiliti. Per quanto riguarda il *trace-variogram*, se da un lato è di rilievo teorico l'estensione formale di tale concetto, dall'altro è fondamentale individuarne una procedura di stima efficace, per mezzo della quale valutare la struttura di covarianza del campo aleatorio e verificarne le assunzioni di stazionarietà e isotropia. Sono qui proposti stimatori sperimentali puntuali, da adattare a un modello parametrico valido con un'opportuna procedura di minimizzazione, affiancandovi una stima intervallare. In particolare, è proposta l'applicazione di tecniche di ricampionamento e, nello specifico, di metodi bootstrap semiparametrici, per l'approssimazione della distribuzione dello stimatore empirico, essenziale per la costruzione degli intervalli di confidenza associati, superando in questo modo le difficoltà legate alla formulazione e alla verifica di ipotesi distribuzionali nel contesto funzionale.

Per quanto concerne la previsione nel caso in cui il processo sia non stazionario, la metodologia qui proposta è sviluppata in tre fasi: la selezione della forma funzionale per la modellazione del *drift*, il disaccoppiamento della variabilità deterministica dalla variabilità stocastica e l'Universal Kriging. Infatti, è dapprima messo a punto un metodo a criterio previsivo per l'individuazione di una forma di *drift* ottimale tra una collezione di modelli lineari candidati; è quindi presentato un algoritmo iterativo per la stima ai minimi quadrati generalizzati del termine di *drift*, modellato con la forma funzionale selezionata nella fase precedente e, infine, grazie alla stima variografica sul residuo, calcolato per differenza, è ottenuta la previsione ottima con il metodo di Universal Kriging.

Le metodologie proposte in questo lavoro sono inoltre analizzate e messe alla prova attraverso l'applicazione a dati sintetici, per i quali sono noti il modello di variogramma e il *drift* di generazione: i casi test presentati sono volti a far emergere i vantaggi, gli svantaggi e le peculiarità delle tecniche sviluppate, sottolineandone le proprietà di robustezza rispetto a vari fattori, quali la scelta iniziale del modello parametrico di variogramma, e suggerendo ulteriori prospettive di ricerca.

Le metodologie sviluppate e sottoposte a studio di simulazione sono infine applicate a un dataset reale, analizzato nei principali lavori presentati in letteratura sul tema del kriging per dati funzionali. L'oggetto dell'analisi è il dataset *Canada's Maritime Provinces Temperatures*, che contiene le informazioni relative alle temperature medie giornaliere registrate tra gli anni 1960 e 1994, in 35 stazioni meteorologiche dislocate nelle province Marittime del Canada. Lo studio di questo dataset consente di confrontare le proposte metodologiche di questo lavoro con le tecniche alternative ad oggi disponibili. L'adozione di una metrica non euclidea per il dominio spaziale e l'introduzione di un termine di *drift* sono tuttavia elementi di discontinuità rispetto agli studi precedenti, il cui impatto sarà valutato in termini variografici e previsivi.

La trattazione è sviluppata in sei capitoli, la cui struttura è sintetizzata di seguito.

- **Capitolo 1**: *Geostatistica per Processi Reali.* In questo capitolo è presentata una breve introduzione alla geostatistica per dati monovariati, illustrandone le definizioni e i risultati principali, seguendo la trattazione presente in (Chilès e Delfiner, 1999) e (Cressie, 1993). L'attenzione è rivolta in particolare all'introduzione dei concetti di variogramma e alle relative ipotesi di stazionarietà e isotropia; infine sono formulati i metodi di kriging stazionario e non stazionario.

- **Capitolo 2**: *Introduzione all'Analisi di Dati Funzionali.* È qui introdotta l'Analisi di Dati Funzionali, soffermandosi dapprima sull'approccio funzionale, quindi illustrando le potenzialità e le difficoltà connesse al trattamento di dati infinito-dimensionali. Infine è presentata una breve panoramica sui modelli lineari, con particolare riferimento al lavoro di Ramsay e Dalzell (1991).

- **Capitolo 3**: *Kriging per Elementi di uno Spazio di Hilbert: una Formulazione di Sintesi.* Questo capitolo raccoglie i risultati teorici originali sviluppati nella tesi. In particolare, sono definite le funzioni covariogramma e variogramma come misure di dipendenza spaziale

di tipo globale, valide per processi funzionali su spazi di Hilbert; sono quindi stabilite le condizioni di stazionarietà e isotropia, che consentono di introdurre gli stimatori del variogramma. Sono qui proposte le formulazioni di sintesi dei metodi di Ordinary e Universal Kriging, a coefficienti costanti, individuando, coerentemente con le definizioni introdotte, il previsore BLUP, sia nel caso stazionario, sia nel caso non stazionario. È inoltre introdotto lo stimatore ottimo ai minimi quadrati del termine di *drift*, descritto da un modello lineare, e ricavata una formula di decomposizione della varianza per il caso funzionale. I riferimenti bibliografici essenziali per il capitolo sono (Hörmann e Kokoszka, 2011), (Giraldo *e altri*, 2010c) e (Ramsay e Dalzell, 1991).

- **Capitolo 4**: *Metodi Bootstrap Semiparametrici per la Stima della Distribuzione del Trace-Variogram.* Il capitolo è dedicato all'introduzione della metodologia di stima del variogramma proposta in questo lavoro. Dopo una breve introduzione ai metodi bootstrap non parametrici e semiparametrici (Efron, 1979), (Olea e Pardo-Igúzquiza, 2011), è proposto un algoritmo iterativo di tipo MC-bootstrap per l'approssimazione della distribuzione dello stimatore empirico del variogramma. L'approssimazione ottenuta è quindi sfruttata per l'adattamento di un modello valido di variogramma e per la costruzione di stimatori intervallari per il *trace-variogram*. È qui illustrato il primo studio di simulazione del lavoro di tesi, volto alla valutazione del comportamento del metodo in termini di convergenza, bontà dell'approssimazione e probabilità di copertura delle stime intervallari, indagando la sensibilità della procedura proposta all'ipotesi di gaussianità del campo.

- **Capitolo 5**: *Previsione per Dati Funzionali Georeferenziati Non Stazionari: l'Universal Kriging.* Questo capitolo è interamente dedicato alla presentazione e all'analisi degli algoritmi sviluppati per la previsione di dati funzionali georeferenziati, spazialmente non stazionari. Sono dapprima introdotti gli algoritmi per il disaccoppiamento del *drift* dal residuo e per la selezione del *drift* ottimo con criterio previsivo, quindi ne è studiato il comportamento attraverso l'applicazione a dati sintetici. È valutata in particolare l'influenza del modello parametrico di variogramma sulle prestazioni previsive e il comportamento del metodo al decrescere della numerosità del campione osservato.

- **Capitolo 6**: *Un Caso Studio: Analisi delle Temperature delle Province Marittime del Canada.* È qui illustrata l'analisi svolta sul dataset di temperature medie giornaliere registrate in 35 località della regione Marittima del Canada. All'introduzione dei dati, seguono alcune riflessioni sulle possibili metriche da adottare per il dominio spaziale, individuando nella metrica indotta dalla distanza geodetica la scelta più appropriata. Dopo l'analisi variografica, sono applicati gli algoritmi proposti in questo lavoro, fornendo un'interpretazione climatica dei risultati ottenuti. All'analisi geostatistica segue la validazione dei risultati, dapprima per cross-validazione, quindi per confronto con le mappe di riferimento del Servizio Meteorologico del Canada (`http://atlas.nrcan.gc.ca/`). Infine, l'analisi geostatistica svolta con le metodologie proposte in questo lavoro è confrontata con le analisi del medesimo dataset presenti in letteratura, sottolineando il ruolo cruciale rivestito dalla

metrica sul dominio spaziale e dall'introduzione di un termine di *drift* nel modello per il fenomeno.

- **Appendice A**: *Notazioni*. In questa appendice sono riportate le convenzioni di scrittura adottate nel lavoro di tesi, con riferimento agli stili e ai simboli introdotti. Questa parte è pensata in particolare come supporto alla lettura e alla comprensione del Capitolo 3.

- **Appendice B**: *Codici*. Sono qui riportati i codici in linguaggio `R` (R Development Core Team, 2009), implementati per l'applicazione degli algoritmi sviluppati nel lavoro di tesi. L'implementazione fa uso di due pacchetti aggiuntivi di `R`, denominati `geoR` (Jr e Diggle, 2001) e `fda` (Ramsay *e altri*, 2010).

Capitolo 1

Geostatistica per Processi Reali

La geostatistica è una branca della statistica applicata che studia le tecniche e i metodi per l'inferenza a partire da campioni che abbiano una struttura di covarianza spaziale.

L'interesse della geostatistica si focalizza sullo studio delle caratteristiche statistiche di un campo aleatorio

$$\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\} \tag{1.1}$$

dove D indica un sottoinsieme di $\mathbb{R}^d$, $d = 2, 3, ...$, di cui è nota una realizzazione $(Z_{\boldsymbol{s}_1}, ..., Z_{\boldsymbol{s}_n})$, corrispondente al campionamento del processo in n località $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$.

A partire dai dati disponibili, l'obiettivo delle analisi è in generale l'individuazione di opportuni modelli per la descrizione della variabilità spaziale del fenomeno, la stima dei relativi parametri e quindi la previsione del dato laddove non sia disponibile l'osservazione.

In questo capitolo saranno introdotte alcune delle principali definizioni e dei risultati più significativi nell'ambito dell'analisi geostatistica di processi a valori reali. Le notazioni che saranno introdotte, coerenti con quelle adottate nel seguito, così come le definizioni i risultati citati, sono tratti in particolare da (Armstrong, 1998), (Chilès e Delfiner, 1999), (Cressie, 1993), (Porcu, 2004).

1.1 Processi Stocastici e Stazionarietà

Nell'ambito della geostatistica monovariata, si considera un processo stocastico a valori reali caratterizzato da una variabilità di larga scala, detta *drift*, e di piccola scala, caratterizzata da una struttura di dipendenza. Il modello per il campo aleatorio è quindi della forma

$$Z(\boldsymbol{s}) = m(\boldsymbol{s}) + \delta(\boldsymbol{s}), \quad \boldsymbol{s} \in D. \tag{1.2}$$

Il *drift* m è supposto essere un termine deterministico, eventualmente dipendente da fattori esterni (*external drift*), mentre δ è un termine stocastico che pertanto definisce le caratteristiche probabilistiche del campo aleatorio.

Per ogni insieme di localizzazioni spaziali $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$, il vettore aleatorio $\boldsymbol{Z}(\boldsymbol{s}) = (Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_n))^t$ è definito, dal punto di vista distribuzionale, dalla funzione di ripartizione congiunta,

detta legge finito-dimensionale:

$$F(z_1, ..., z_n) = \mathbb{P}(Z(\boldsymbol{s}_1) \leq z_1, ..., Z(\boldsymbol{s}_n) \leq z_n), \quad z_1, ..., z_n \in \mathbb{R}.$$

Qualora siano soddisfatte le condizioni di omogeneità di Kolmogorov (Cressie, 1993), la famiglia F di distribuzioni al variare di $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$, caratterizza la legge dei processo Z.

Al fine di fare inferenza, è essenziale fissare alcune ipotesi tale la famiglia e, in particolare, determinarne una classe di appartenenza, affinché possano essere sfruttati i risultati noti e le proprietà asintotiche della classe stessa. Un'assunzione fondamentale è la stazionarietà, che può essere declinata in diverse definizioni più o meno stringenti, a seconda delle quali diversi metodi statistici possono essere applicati.

Definizione 1.1. $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ *si dice processo stocastico* stazionario in senso stretto *se* $\boldsymbol{Z}(\boldsymbol{s}) = (Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_n))^t$ *e* $\boldsymbol{Z}(\boldsymbol{s} + \boldsymbol{h}) = (Z(\boldsymbol{s}_1 + \boldsymbol{h}), ..., Z(\boldsymbol{s}_n + \boldsymbol{h}))^t$ *hanno la medesima distribuzione congiunta, per ogni* $\boldsymbol{h} \in \mathbb{R}^d$*, per ogni successione* $\{\boldsymbol{s}_k\}_{k=1}^n$ *in* $D \subseteq \mathbb{R}^d$*, per ogni* n*.*

Dal momento che questa assunzione è molto forte e difficilmente verificabile, è possibile indebolirla richiedendo che il processo sia *debolmente stazionario*. In questo caso l'ipotesi di lavoro risulta la seguente.

Definizione 1.2. $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ *si dice processo stocastico* debolmente stazionario *se*

$$\mathbb{E}[Z(\boldsymbol{s})] = m, \quad \forall\, \boldsymbol{s} \in D \subseteq \mathbb{R}^d$$
$$\mathrm{Cov}(Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)) = \mathbb{E}[(Z(\boldsymbol{s}_i) - m)(Z(\boldsymbol{s}_j) - m)] = C(\boldsymbol{h}), \quad \forall\, \boldsymbol{s}_i, \boldsymbol{s}_j \in D \subseteq \mathbb{R}^d,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j.$$

C è detta *funzione di covarianza* o *covariogramma* ed è una funzione definita positiva, ovvero tale che:

$$\sum_i \sum_j \lambda_i \lambda_j C(\boldsymbol{s}_i - \boldsymbol{s}_j) \geq 0, \quad \forall\, \lambda_i, \lambda_j \in \mathbb{R};\ \boldsymbol{s}_i, \boldsymbol{s}_j \in D.$$

Questa condizione assicura che le combinazioni lineari con pesi $\lambda_1, ..., \lambda_n$ delle variabili aleatorie $Z(\boldsymbol{s}_1), ..., Z(\boldsymbol{s}_n)$ abbiano varianza positiva, per ogni insieme di pesi $\lambda_1, ..., \lambda_n$. Inoltre, vale che:

$$C(\boldsymbol{h}) = C(-\boldsymbol{h})$$
$$|C(\boldsymbol{h})| \leq C(\boldsymbol{0})$$

Si dimostra (Cressie, 1993) che la stazionarietà in senso stretto implica quella debole, mentre in generale non vale il viceversa. L'implicazione contraria è verificata nel caso di processo gaussiano, ovvero le cui leggi finito dimensionali siano gaussiane: in tal caso il processo è completamente caratterizzato dalla definizione dei momenti del primo e del second'ordine e, pertanto, le definizioni 1.1 e 1.2 coincidono.

Un'ultima definizione di stazionarietà, di interesse per il lavoro di tesi, è la seguente.

Definizione 1.3. $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ *è un campo aleatorio* intrinsecamente stazionario *se è caratterizzato dalle seguenti ipotesi:*

$$\mathbb{E}[Z(\boldsymbol{s})] = m, \quad \forall \boldsymbol{s} \in D$$
$$\mathrm{Var}(Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)) = \mathbb{E}[(Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j))^2] = 2\gamma(\boldsymbol{h}), \quad \forall \boldsymbol{s}_i, \boldsymbol{s}_j \in D,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j.$$

In questo caso $\gamma(h)$ è detto *semivariogramma* ed è una funzione condizionatamente definita negativa, cioè tale che

$$\sum_i \sum_j \lambda_i \lambda_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) \leq 0, \quad \forall\ \boldsymbol{s}_i, \boldsymbol{s}_j, \in D, \forall\ \lambda_i, \lambda_j \quad \text{t.c.} \quad \sum_i \lambda_i = 0. \tag{1.3}$$

Tale condizione assicura la positività della varianza di una qualsiasi combinazione lineare ammissibile di $Z(s_1), ..., Z(s_n)$, cioè di una combinazione lineare per la quale sia definito la varianza che, per ipotesi di stazionarietà, è garantita esistere solo per incrementi del processo. Anche in questo caso si verifica che il semivariogramma è una funzione pari; inoltre esso è non negativo e nullo nell'origine:

$$\gamma(\boldsymbol{h}) = \gamma(-\boldsymbol{h});$$
$$\gamma(\boldsymbol{h}) \geq 0;$$
$$\gamma(0) = 0.$$

Un processo debolmente stazionario è anche intrinsecamente stazionario, il viceversa è vero con la seguente ipotesi aggiuntiva (Cressie, 1993).

Definizione 1.4. *Un processo di dice* ergodico *qualora sia debolmente stazionario e*

$$\lim_{||\boldsymbol{h}|| \to +\infty} C(\boldsymbol{h}) = 0.$$

In tali ipotesi si ha allora che

$$\lim_{||\boldsymbol{s}_i - \boldsymbol{s}_j|| \to +\infty} \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) = C(\boldsymbol{0}),$$

e

$$\gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) = C(\boldsymbol{0}) - C(\boldsymbol{s}_i - \boldsymbol{s}_j). \tag{1.4}$$

Nella valutazione delle proprietà asintotiche di un processo aleatorio e degli stimatori ad esso associati, è possibile considerare almeno tre tipi di distribuzione dei punti campionati (Cressie, 1993): *infill domain sampling*, *increasing domain sampling* e *nearly infill sampling*.

Il primo caso si verifica quando la regione D_N, all'interno della quale le osservazioni sono campionate nei punti $\{\boldsymbol{s}_{i,N}; 1 \leq i \leq N\}$, rimane limitata per $N \to \infty$; in questo tipo di disegno sperimentale, i risultati asintotici classici, come la legge dei grandi numeri, non sono generalmente applicabili.

La seconda situazione possibile, *increasing domain sampling*, si verifica quando il diametro della regione D_N tende all'infinito come conseguenza della richiesta che i punti campionati in posizione $\{\boldsymbol{s}_{i,N}\}$ siano separati da una distanza almeno pari a d_{min}, per ogni i e N.

Infine, si ha un campionamento di tipo *nearly infill* qualora il numero di siti esplorati tenda all'infinito in ciascuna sottoregione del dominio D_N, illimitato per $N \to \infty$.

In questo contesto si inserisce un'altra nozione di ergodicità, la microergodicità, legata alle caratteristiche asintotiche nel caso di disegno di esperimento di tipo *infill domain*. Tale proprietà è definita da Matheron (1978) come segue.

Definizione 1.5. *Sia $D \subseteq \mathbb{R}^d$ il dominio spaziale di riferimento e sia $\mathcal{P} = \{P_\vartheta | \vartheta \in \Theta\}$ una classe di misure di probabilità su D. Data una funzione h definita su Θ, $h(\vartheta)$ si dice* microergodico *se il suo valore può essere ottenuto correttamente, a partire da una singola realizzazione del campo aleatorio* (1.1) *con probabilità 1.*

L'ipotesi di microergodicità è importante in quanto è una condizione necessaria, sebbene non sufficiente, per la consistenza di uno stimatore $h(\vartheta)$, qualora la sua stima si basi su uno stimatore $\widehat{\boldsymbol{\vartheta}}_N$ del parametro $\boldsymbol{\vartheta}$ consistente; inoltre, come sottolineato da Chilès e Delfiner (1999), i parametri microergodici sono tipicamente parametri di significato fisico.

In alcune applicazioni accade tuttavia che il processo considerato non risulti nemmeno intrinsecamente stazionario: questo può accadere ad esempio in presenza di una media non costante. In questo caso, la teoria geostatistica esistente in ambito finito dimensionale consente di procedere in due modi (Chilès e Delfiner, 1999): il primo metodo consiste nell'ipotizzare un *drift* e quindi procedere dicotomizzando la variabile in esame come somma del *drift* stesso e di un residuo stazionario; il secondo approccio, che non sarà approfondito nel lavoro di tesi, prevede invece l'introduzione di una teoria più generale, ossia la teoria delle funzioni aleatorie intrinseche di ordine k (IRF-k), di cui le funzioni intrinsecamente stazionarie sono un caso particolare (IRF-0).

Nel seguito del lavoro di tesi, l'ipotesi di non stazionarietà sarà pensata in riferimento a un modello della forma (1.2), caratterizzato pertanto da una componente $m(s)$ deterministica variabile spazialmente e da una componente stocastica stazionaria (al second'ordine o intrinsecamente).

1.2 Il Variogramma: dalla Stima Empirica ai Modelli Validi

Si consideri un processo stocastico $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$ stazionario in senso debole o intrinseco, caratterizzato da una funzione di media m costante nel dominio spaziale D e da un semivariogramma $\gamma(\boldsymbol{h})$.

Dal momento che è possibile mostrare esempi di processi intrinsecamente stazionari, ma non debolmente stazionari, per i quali il covariogramma non è definito, è preferibile analizzare

la struttura di covarianza del processo tramite il variogramma, essendo una nozione più generale.

In questa sezione, saranno quindi introdotte, senza pretesa di completezza, le caratteristiche teoriche e i metodi di stima della funzione variogramma, le cui generalizzazioni al caso funzionale saranno oggetto di studio nei Capitoli 3 e 4.

1.2.1 Proprietà del Variogramma

Il variogramma descrive le proprietà probabilistiche del second'ordine di un processo stazionario, in senso debole o intrinseco, ed è per questo di grande interesse nell'analisi della realizzazione di un campo aleatorio. Esistono alcune caratteristiche della funzione variogramma che sono particolarmente significative nella determinazione delle proprietà fisiche del campo stesso; esse saranno brevemente descritte in questa sezione, per ulteriori approfondimenti ci si riferisca ad esempio a (Chilès e Delfiner, 1999) e (Cressie, 1993).

La funzione $2\gamma(\boldsymbol{h})$ è simmetrica e nulla nell'origine, tuttavia è possibile che il comportamento di tale funzione in prossimità dell'origine sia discontinuo, presentando un limite non nullo per $\|\boldsymbol{h}\|$ tendente a 0:

$$\lim_{\|\boldsymbol{h}\|\to 0} \gamma(\boldsymbol{h}) = \tau^2 \neq 0 = \gamma(\mathbf{0});$$

in questo caso τ^2 è detto effetto *nugget*. Quest'ultimo modellizza e quantifica la discontinuità del processo introdotta dalla variazione di 'microscala': sebbene dal punto di vista matematico non sia possibile che un processo L^2-continuo sia caratterizzato da una tale discontinuità, in questa ipotesi l'effetto *nugget* è generalmente indicazione della presenza di un errore di misura nel dato.

Come enfatizzato da Cressie (1993), è di fatto impossibile determinare il valore esatto del variogramma per distanze molto piccole: avendo a disposizione una realizzazione $z_{\boldsymbol{s}_1}, ..., z_{\boldsymbol{s}_n}$ non sono infatti disponibili informazioni per distanze inferiori a $\min\{\|\boldsymbol{s}_i - \boldsymbol{s}_j\|,\, 1 \leq i, j \leq n\}$. Tuttavia, l'effetto *nugget* può essere determinato per estrapolazione o, in alternativa, fissato a 0 se sono disponibili informazioni sul processo di misura e sulla variazione di microscala del campo aleatorio.

Una seconda caratteristica importante del variogramma è il *sill*. Questo, qualora esista, è definito come:

$$\tau^2 + \sigma^2 := \lim_{\|\boldsymbol{h}\|\to\infty} \gamma(\boldsymbol{h}),$$

dove τ^2 è l'effetto *nugget* e σ^2 è detto *partial sill*. L'esistenza di un *sill* finito indica che il processo in esame è debolmente stazionario, caratterizzato da una varianza $C(\mathbf{0}) = \tau^2 + \sigma^2$.

Si definisca infine $\boldsymbol{R}$ come quel valore di $\boldsymbol{h}$ tale che:

$$\gamma(\boldsymbol{R}) = \tau^2 + \sigma^2,$$

ovvero il valore del vettore incremento $\boldsymbol{h}$ tale che il variogramma raggiunge il *sill*. La lunghezza $\|\boldsymbol{R}\|$ di tale vettore, detta *range*, quantifica il raggio di influenza del processo, nel senso

che, a direzione fissata $\boldsymbol{h}/\|\boldsymbol{h}\|$, si può considerare nulla la correlazione tra due osservazioni il cui vettore congiungente $\boldsymbol{h}$ abbia lunghezza superiore a $\|\boldsymbol{R}\|$. Il *range* può non esistere o non essere finito: può infatti accadere che il *sill* non esista (indice di non stazionarietà o stazionarietà intrinseca) o sia raggiunto asintoticamente.

Un'ulteriore proprietà del variogramma è l'isotropia che, se verificata, consente di ridurre notevolmente la complicazione dell'analisi. Essa è infatti definita come segue.

Definizione 1.6. *Un processo stocastico stazionario si dice* isotropo *se il variogramma che lo caratterizza è isotropo, ovvero è tale che:*

$$\mathrm{Var}(Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)) = 2\gamma(h), \quad h = \|\boldsymbol{h}\| = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|, \forall \boldsymbol{s}_i, \boldsymbol{s}_j \in D.$$

In caso contrario, il processo è detto anisotropo.

L'ipotesi isotropia è dunque verificata qualora ci sia uniformità nella struttura di covarianza in tutte le direzioni di $\mathbb{R}^d$. In caso contrario, esistono due tipi principali di anisotropia: l'*anisotropia geometrica* e l'*anisotropia zonale*. La prima si verifica qualora il variogramma presenti o il *range* o la pendenza nell'origine variabile al variare delle direzione di osservazione: in questo caso si opera un cambiamento di coordinate per correggere l'anisotropia stessa. L'anisotropia di tipo zonale si verifica invece quando il *sill* del variogramma non è lo stesso in tutte le direzioni: in tal caso il variogramma deve essere definito in termini di vettore di separazione $\boldsymbol{h}$.

Dal momento che la geostatistica per dati funzionali è un ambito di ricerca ancora molto aperto e con basi teoriche tuttora in via di consolidamento, in questo lavoro è stato scelto di focalizzare le analisi sul caso di processi isotropi. Per questo motivo nel seguito, salvo contrario avviso, sarà fatto riferimento a modelli di variogramma (e covariogramma) isotropi.

1.2.2 Modelli Validi di Variogramma

Si consideri ora un processo stocastico debolmente stazionario e isotropo (1.1) e sia $\gamma(h)$ il suo semivariogramma. Tra le proprietà che deve soddisfare la funzione $\gamma(h)$, si è citata la definitezza negativa condizionata, espressa dalla (1.3). Una funzione $\gamma(h)$ che soddisfi tale ipotesi è detta modello di semivariogramma *valido*.

Esistono alcune tecniche per costruire modelli di (semi)variogramma validi (ad esempio (Cressie, 1993) e (Armstrong e Diamond, 1984)), tuttavia sono generalmente usati modelli parametrici, per i quali sono noti sia le proprietà analitiche che il significato fisico dei relativi parametri.

Gli esempi più noti di modelli di variogramma validi sono i seguenti.

– *Puro Nugget*:

$$\gamma(h) = \begin{cases} \tau^2, & h > 0 \\ 0, & h = 0 \end{cases}, \tag{1.5}$$

con $\tau \in \mathbb{R}$. Il processo è un rumore bianco di varianza τ^2.

– *Modello Lineare*:

$$\gamma(h) = \begin{cases} a^2 h, & h > 0 \\ 0, & h = 0 \end{cases}, \tag{1.6}$$

con $a \in \mathbb{R}$. In questo caso il processo è intrinsecamente stazionario ma non debolmente stazionario (non è definita la funzione di covarianza).

– *Modello Esponenziale*:

$$\gamma(h) = \begin{cases} \sigma^2(1 - e^{-h/a}), & h > 0 \\ 0, & h = 0 \end{cases}, \tag{1.7}$$

dove $a, \sigma \in \mathbb{R}$. In questo caso il *sill* è σ^2, il *range* non è definito: è possibile definire il *practical range*, ovvero il valore di h tale che è raggiunto il 95% del *sill*, che è pari a $3a$.

– *Modello Sferico*:

$$\gamma(h) = \begin{cases} \sigma^2\{\frac{3}{2}\frac{h}{a} + \frac{1}{2}(\frac{h}{a})^3\}, & h > a \\ \sigma^2, & h \leq a \\ 0, & h = 0 \end{cases}, \tag{1.8}$$

con $a, \sigma \in \mathbb{R}$. I parametri del modello hanno ora un significato fisico importante: a è il *range*, σ^2 il *sill*.

A partire da modelli di variogramma validi, è possibile ottenerne altri in modo semplice grazie alla seguente Proposizione (Armstrong, 1998).

Proposizione 1.7. *i) Sia $\gamma(\cdot)$ un variogramma valido e $c \in \mathbb{R}$ una costante non negativa, allora $c\gamma(\cdot)$ è un variogramma valido.*

ii) Siano $\gamma_1(\cdot)$ e $\gamma_2(\cdot)$ variogrammi validi, allora la loro somma è ancora un variogramma valido.

Una caratteristica del variogramma che influenza significativamente le peculiarità del processo in termini di regolarità è il comportamento nell'origine. Si distinguono infatti tre situazioni:

- Discontinuità nell'origine. Il processo stocastico presenta un comportamento altamente irregolare e non risulta continuo in media quadratica. Tale discontinuità è presente in caso di effetto *nugget*, ovvero qualora siano presenti errori di misura e variabilità di microscala.

- Lineare nell'origine. Tale comportamento è tipico dei processi continui ma non differenziabili. I modelli lineare, sferico ed esponenziale sono esempi di questo caso.

- Parabolico nell'origine. Il campo aleatorio associato risulta tipicamente molto regolare e presenta spesso un termine di *drift*.

1.2.3 Stima del Variogramma: lo Stimatore Empirico e l'Adattamento a un Modello Valido

Una volta definite formalmente la funzione variogramma $\gamma(h)$ e le relative proprietà, dal punto di vista applicativo è necessario poter stimare in modo adeguato tale funzione, a partire dai dati disponibili.

Il processo di stima del variogramma avviene generalmente in due fasi: dapprima si stima il *semivariogramma empirico*, ovvero uno stimatore sperimentale del semivariogramma che in generale non gode delle proprietà della funzione $\gamma(h)$; in un secondo momento, nota la stima empirica, si procede a determinare i parametri di un modello parametrico opportunamente scelto, che siano ottimi secondo un criterio stabilito (massima verosimiglianza, minimi quadrati, etc.).

Lo Stimatore Empirico

Una stima empirica del semivariogramma può essere ottenuta attraverso il metodo dei momenti e risulta:

$$\widehat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{(i,j)\in N(h)} [z(\boldsymbol{s}_i) - z(\boldsymbol{s}_j)]^2, \tag{1.9}$$

dove $N(h) = \{(i,j) : ||\boldsymbol{s}_i - \boldsymbol{s}_j|| = h\}$ e $|N(h)|$ ne indica la cardinalità.

Dal momento che solitamente non sono disponibili coppie di osservazioni per ogni distanza h, si procede usualmente nell'analisi e visualizzazione degli stimatori *semivariogram cloud* e *binned semivariogram*.

La prima è la nuvola di punti corrispondente ai valori osservati di $\frac{1}{2}[Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)]^2$ per ogni coppia di variabili $Z(\boldsymbol{s}_i), Z(\boldsymbol{s}_j)$, $i, j = 1, ..., n$: tale nuvola fornisce in particolare le informazioni relative al numero di coppie disponibili per la stima e alla disposizione nel piano degli incrementi al quadrato osservati (ad esempio, una distribuzione di forma triangolare più densa nella parte bassa del piano è tipica di un variogramma stazionario).

Lo stimatore *binned semivariogram* $\widehat{\gamma}(\boldsymbol{h})$ è invece spesso usato nelle applicazioni poiché mostra le caratteristiche del variogramma in modo semplice e immediato. Esso è un vettore aleatorio $(\widehat{\gamma}(h_1), ..., \widehat{\gamma}(h_K))$ definito per un numero limitato di classi di distanze ed è calcolato come media dei valori osservati in ciascuna classe:

$$\widehat{\gamma}(h_k) = \frac{1}{2|N(h_k)|} \sum_{(i,j)\in N(h_k)} [Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)]^2, \quad k = 1, ..., K, \tag{1.10}$$

dove $N(h_k)$ indica l'insieme delle coppie appartenenti alla k-esima classe di distanze e $|N(h_k)|$ ne indica la cardinalità.

In questo caso, è fondamentale la definizione dell'ampiezza delle classi e, di conseguenza, del numero di lag K. Esiste infatti un *trade-off* tra l'*overfitting* e l'*oversmoothing*: se da un lato occorre che la cardinalità delle classi sia sufficiente per evitare che lo stimatore sia eccessivamente perturbato dalla componente di rumore, dall'altro lato il numero di lag K

deve essere sufficientemente elevato per cogliere le peculiarità del variogramma sottostante (ad esempio il comportamento nell'origine, il *sill* e il *range*).

Adattamento a un Modello Valido

Una volta disponibile una stima del *binned semivariogram* per K lag opportunamente scelti, occorre determinare i parametri di un modello valido, al fine di garantire che la stima ottenuta risulti condizionatamente definita negativa.

Tali parametri possono essere ottenuti attraverso la massimizzazione di un criterio di ottimo opportunamente scelto. Nel caso sia nota la verosimiglianza del modello (1.1), è possibile determinare i parametri ottimi con un approccio di tipo massima verosimiglianza (ML, REML (Cressie, 1993), (Pardo-Igúzquiza e Dowd, 2003)), tuttavia, in assenza di ipotesi distribuzionali, i metodi più usati sono i minimi quadrati. Nel Capitolo 4 sarà illustrato con più precisione il metodo di stima dei minimi quadrati ordinari e generalizzati, approfondendo in particolare le tecniche numeriche necessarie ad ottenere una stima del variogramma basata sulle proprietà statistiche dello stimatore $\widehat{\gamma}(h)$.

1.3 La Previsione per Dati Spaziali: il Kriging

Una tematica incontrata comunemente nell'ambito della geostatistica è la previsione: date n osservazioni, $Z(s_1), ..., Z(s_n)$, relative a un numero limitato di siti $s_1, ..., s_n$ del dominio spaziale D, è sovente di interesse per le applicazioni la stima del dato in un punto non osservato, o su una griglia equispaziata di punti, al fine di studiare il fenomeno in oggetto sull'intero dominio D.

Esistono due approcci matematici differenti a tale problema di stima (Chilès e Delfiner, 1999): l'interpolazione e il kriging.

Nel primo caso l'attenzione è concentrata sulla funzione interpolante, per la quale è formulato preliminarmente un modello, in modo esplicito (ad esempio un modello polinomiale) o in modo implicito (ad esempio imponendo la condizione di minima curvatura); i parametri di tale modello sono quindi determinati in modo da ottimizzare un criterio di adattamento ai punti campionati, che può essere di tipo deterministico (adattamento esatto, i.e. interpolazione esatta) o di tipo statistico (minimi quadrati).

Il punto di vista del kriging è invece rivolto alla determinazione di un modello statistico per il fenomeno in esame a partire dai dati disponibili, identificando dapprima la struttura di covarianza spaziale del campo aleatorio e implementando in un secondo momento un previsore ottimo come combinazione lineare dei dati.

In questo paragrafo sarà approfondita la descrizione delle tecniche di kriging, che saranno sviluppate nel seguito con particolare riferimento a campi aleatori funzionali. Dal punto di vista formale, si consideri il processo stocastico (1.1), rappresentato dalla dicotomia (1.2). Le tecniche di kriging prevedono che la stima puntuale di $Z(s_0)$, a partire dai dati osservati in posizione $s_1, ..., s_n$, sia ottenuta attraverso la determinazione del previsore BLUP (Best Linear

Unbiased Predictor), cioè di un previsore che sia funzione lineare dei dati a disposizione,

$$Z^*(s_0) = m(s_0) + \sum_{i=1}^{n} \lambda_i^* [Z(s_i) - m(s_i)],$$

attraverso dei pesi ottimi $\boldsymbol{\lambda}^*$ ricavati minimizzando l'errore quadratico medio sotto il vincolo di non distorsione:

$$\boldsymbol{\lambda}^* = \underset{\{\boldsymbol{\lambda} \in \mathbb{R}^n \mid \mathbb{E}[Z^*(s_0)] = [Z(s_0)]\}}{\operatorname{argmin}} \operatorname{Var} \left[\sum_{i=1}^{n} \lambda_i Z(s_i) - Z(s_0) \right].$$

Si distinguono tre tipi di kriging in relazione alla stazionarietà o meno del processo considerato: *Simple Kriging* (SK), *Ordinary Kriging* (OK) e *Universal Kriging* (UK).

Il kriging di tipo semplice e quello di tipo ordinario sono usati in ambito stazionario, nel caso in cui la media $m(s) = m$, $s \in D$ sia rispettivamente nota o incognita. L'Universal Kriging è invece applicato qualora il processo sia non stazionario nella forma (1.2), caratterizzato da un *drift*:

$$m(s) = \sum_{l=0}^{L} a_l f_l(s),$$

con a_l coefficienti costanti e $f_l(s)$ funzioni note variabili spazialmente (o regressori della risposta $Z(s)$).

I coefficienti ottimi di kriging possono essere ottenuti in tutti i casi attraverso la soluzione di un sistema lineare che incorpori opportunamente le condizioni di non distorsione. Dal momento che il Simple Kriging non necessita di vincoli di questo tipo, essendo nota la media, i pesi ottimi derivanti dalla minimizzazione dell'errore quadratico medio di previsione sono ottenibili risolvendo il sistema:

$$\begin{pmatrix} \gamma(0) & \gamma(s_1 - s_2) & \cdots & \gamma(s_1 - s_n) \\ \gamma(s_2 - s_1) & \gamma(0) & \cdots & \gamma(s_2 - s_n) \\ \vdots & \vdots & \ddots & \vdots \\ \gamma(s_n - s_1) & \gamma(s_n - s_2) & \cdots & \gamma(0) \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} \gamma(s_0 - s_1) \\ \gamma(s_0 - s_2) \\ \vdots \\ \gamma(s_0 - s_n) \end{pmatrix} \tag{1.11}$$

I pesi relativi all'Ordinary Kriging sono invece determinabili risolvendo il sistema lineare seguente:

$$\begin{pmatrix} \gamma(0) & \gamma(s_1 - s_2) & \cdots & \gamma(s_1 - s_n) & 1 \\ \gamma(s_2 - s_1) & \gamma(0) & \cdots & \gamma(s_2 - s_n) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(s_n - s_1) & \gamma(s_n - s_2) & \cdots & \gamma(0) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(s_0 - s_1) \\ \gamma(s_0 - s_2) \\ \vdots \\ \gamma(s_0 - s_n) \\ 1 \end{pmatrix} \tag{1.12}$$

ottenuto attraverso il metodo dei moltiplicatori di Lagrange: l'ultima equazione corrisponde infatti alla condizione di non distorsione:

$$\sum_{i=1}^{n} \lambda_i = 1$$

e μ ne è il relativo moltiplicatore. La scrittura a blocchi del precedente sistema risulta:

$$\begin{pmatrix} \gamma_{ij} & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_i \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma_{0,i} \\ 1 \end{pmatrix}$$

L'Universal Kriging prevede infine la soluzione del sistema lineare:

$$\begin{pmatrix} \gamma(0) & \cdots & \gamma(s_1 - s_n) & 1 & f_1(s_1) & \cdots & f_L(s_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(s_n - s_1) & \cdots & \gamma(0) & 1 & f_1(s_n) & \cdots & f_L(s_n) \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ f_1(s_1) & \cdots & f_1(s_n) & 0 & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_L(s_1) & \cdots & f_L(s_n) & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu_0 \\ \mu_1 \\ \vdots \\ \mu_L \end{pmatrix} = \begin{pmatrix} \gamma(s_0 - s_1) \\ \vdots \\ \gamma(s_0 - s_n) \\ 1 \\ f_1(s_0) \\ \vdots \\ f_L(s_0) \end{pmatrix}, \tag{1.13}$$

ottenuto anche in questo caso tramite il metodo dei moltiplicatori di Lagrange per imporre la condizione di non distorsione:

$$\sum_{i=1}^{n} \lambda_i f_l(s_i) = f_l(s_0), \quad \forall\, l = 1, ..., L.$$

La scrittura a blocchi del sistema precedente risulta:

$$\begin{pmatrix} \gamma_{ij} & 1 & f_i^l \\ 1 & 0 & 0 \\ f_i^l & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_i \\ \mu_0 \\ \mu_l \end{pmatrix} = \begin{pmatrix} \gamma_{0,i} \\ 1 \\ f_0^l \end{pmatrix}$$

Un aspetto fondamentale nel kriging è la possibilità di determinare in forma chiusa un'espressione della variabilità associata alla stima puntuale fornita dal metodo. Questa varianza, nota con il nome di *varianza di kriging*, è ottenibile nei tre casi precedenti rispettivamente come

$$\sigma^2_{SK} = \sum_{i=1}^{n} \lambda_i \gamma(s_0 - s_i), \quad s_0 \in D \tag{1.14}$$

$$\sigma^2_{OK} = \sum_{i=1}^{n} \lambda_i \gamma(s_0 - s_i) + \mu, \quad s_0 \in D \tag{1.15}$$

$$\sigma^2_{UK} = \sum_{i=1}^{n} \lambda_i \gamma(s_0 - s_i) + \sum_{l=0}^{L} \mu_l f_l(s_0), \quad s_0 \in D;\ f_0(s) = 1,\ \forall s \in D. \tag{1.16}$$

Capitolo 2

Introduzione all'Analisi di Dati Funzionali

L'Analisi di Dati Funzionali (FDA) è una branca della statistica che nasce e si sviluppa per rispondere alle crescenti esigenze, emerse in molte applicazioni, rispetto al trattamento di dati che corrispondono, per ciascuna unità statistica, all'osservazione di un fenomeno lungo una variabile continua, ad esempio il tempo, una coordinata spaziale o una variabile di altra natura. Questa sezione si propone di introdurre i concetti principali della FDA, focalizzandosi in particolare sugli aspetti che verranno usati o estesi in questo lavoro.

I dati funzionali risultano essere tipicamente generati da un campionamento lungo una griglia fitta a piacere di una funzione, che può essere ad esempio una curva o un'immagine. Sebbene la raccolta di informazioni avvenga lungo una griglia discreta, dal punto di vista teorico l'analisi di dati funzionali si basa sulla considerazione dell'intera funzione come unica entità e non come collezione di punti lungo una griglia: il dato funzionale è dunque pensato come punto in un opportuno spazio infinito-dimensionale.

Pertanto, sebbene la FDA possa essere pensata come naturale sviluppo dell'analisi di dati multivariati (MDA), conseguente all'aumentare della dimensionalità del dato, essa sorge in molti casi dall'astrazione dell'analisi monovariata, nella considerazione di punti in uno spazio funzionale, anziché di punti nell'insieme dei reali.

L'analisi di dati funzionali non costituisce soltanto un insieme di metodi atti all'analisi di dati complessi, ma rappresenta un cambiamento del punto di vista dell'analisi: la considerazione di un approccio di tipo astratto nella trattazione di segnali di varia natura apporta numerosi vantaggi in relazione alla possibilità di caratterizzare il dato in base a grandezze non solo di tipo vettoriale, ma anche di tipo differenziale, sfruttando tutte le informazioni disponibili derivanti dal processo di misura. D'altra parte, le potenzialità di un'ottica funzionale e le difficoltà tecniche ad essa connesse sono due facce della stessa medaglia.

Dal punto di vista applicativo, esistono essenzialmente tre tipi di motivazioni che spingono ad adottare questo tipo di paradigma (Ramsay e Dalzell, 1991):

1. Le osservazioni funzionali sono sempre più spesso presenti in diversi contesti applicativi;
2. Alcuni problemi sono di natura intrinsecamente funzionale;
3. Considerare la regolarità naturale dei dati può portare a risultati significativi che non si otterrebbero qualora gli stessi dati fossero trattati con le tecniche multivariate consuete.

Il lavoro di tesi si colloca in quest'ambito, con l'obiettivo di rispondere alle problematiche e alle finalità dell'analisi geostatistica in un contesto più generale, coniugando l'approccio esposto nel Capitolo 1 con le tecniche e le definizioni tipiche dell'ambito funzionale. Per questo motivo, in questo capitolo saranno dapprima introdotte le definizioni di base e le notazioni adottate nel lavoro; in un secondo momento, sarà fatto breve accenno ad alcune procedure preliminari all'analisi di dati funzionali, quali la proiezione su una base e la registrazione dei dati; saranno quindi esposte alcune riflessioni legate alla dimensionalità dei dati e all'importanza della scelta del giusto spazio nel quale immergere i dati stessi. Infine, sarà presentata una breve panoramica sui modelli lineari per dati funzionali, fornendo gli strumenti teorici essenziali ai fini della presentazione delle estensioni al caso spazialmente correlato, proposte nel Capitolo 3.

2.1 I Dati Funzionali

Si supponga di poter osservare una variabile di interesse $X(\cdot)$ in un intervallo $\mathcal{T} = [t_{min}, t_{max}]$ di istanti successivi t_j. Grazie agli strumenti di misura moderni, in molti casi è possibile ottenere misurazioni del fenomeno lungo una griglia sempre più fitta, ottenendo dati relativi ad istanti più ravvicinati; pertanto, sebbene il dato possa essere considerato come osservazione della famiglia aleatoria $\{X(t_j)\}_{j=1,...,J}$, è spesso naturale pensare le osservazioni come campionamento dalla famiglia continua $\boldsymbol{\chi} = \{X(t); t \in \mathcal{T}\}$. Nel seguito del lavoro si penserà alla famiglia aleatoria $\{X(t_j)\}_{j=1,...,J}$ come al campionamento di una funzione $\boldsymbol{\chi}$ definita da $\mathcal{T}$, sottoinsieme compatto di $\mathbb{R}$, a $\mathbb{R}$ (i.e. $X(\cdot) \in \mathbb{R}$). Tuttavia, la FDA può essere usata in ambiti nei quali i dati raccolti non corrispondono a curve, ma, ad esempio, a una superficie, un vettore di curve o un oggetto infinito-dimensionale più complesso. In quest'ottica, si collocano le seguenti definizioni, presenti in Ferraty e Vieu (2006).

Definizione 2.1. *Un ente aleatorio $\boldsymbol{\chi}$ è detto* variabile funzionale *se assume valori in uno spazio infinito dimensionale (o spazio funzionale). Un'osservazione χ di $\boldsymbol{\chi}$ è detta* dato funzionale.

Definizione 2.2. *Un* dataset *(o* campione*)* funzionale $\chi_1, ..., \chi_n$ *è un'osservazione di n variabili funzionali $\boldsymbol{\chi}_1, ..., \boldsymbol{\chi}_n$, identicamente distribuite rispetto a $\boldsymbol{\chi}$.*

La Definizione 2.1 è valida in contesti molto generali. Per inquadrare formalmente il contesto particolare che sarà considerato in questo lavoro, si denoti con $(\Omega, \mathfrak{F}, \mathbb{P})$ uno spazio di probabilità e con H uno spazio vettoriale, dotato di una σ-algebra $\mathfrak{H}$. Nel seguito, si

supporrà che H sia uno spazio di Hilbert, dotato del prodotto interno $\langle \cdot, \cdot \rangle_H$ e della norma da esso indotta $\| \cdot \|_H$ (indicati, qualora non si creino ambiguità, con $\langle \cdot, \cdot \rangle$ e $\| \cdot \|$). Allora, con l'espressione *variabile aleatoria funzionale*, si farà riferimento a una funzione misurabile $\boldsymbol{\chi}$:

$$\boldsymbol{\chi} : \Omega \to H,$$

mentre con *dato funzionale* sarà indicata la funzione:

$$\chi := \boldsymbol{\chi}(\omega) : \mathcal{T} \to \mathbb{R},$$

per $\omega \in \Omega$ fissato, ovvero una realizzazione di $\boldsymbol{\chi}$.

Nel contesto stabilito dalle precedenti assunzioni, una variabile aleatoria funzionale può essere definita come la realizzazione di una funzione aleatoria, identificando il dato funzionale con una traiettoria del processo stesso (Tarpey e Kinateder, 2003). L'immersione dell'analisi di dati funzionali nel contesto della teoria dei processi stocastici è molto utile dal punto di vista teorico, fornendo la possibilità di usarne i metodi e i risultati, tuttavia è implicita l'assunzione che i dati funzionali possiedano una natura matematica differente, caratterizzata da proprietà di regolarità. Non è però scopo di questo lavoro addentrarsi nelle questioni filosofiche legate a queste considerazioni, per il cui approfondimento si rimanda ad esempio a (Ferraty e Vieu, 2006), (Tarpey e Kinateder, 2003), (Ramsay e Dalzell, 1991).

Lavorando in uno spazio infinito-dimensionale è inoltre necessario estendere in modo opportuno le definizioni dell'analisi multivariata, come ad esempio le definizioni di media, varianza e covarianza: questo può essere fatto sia in modo astratto, sia adottando la terminologia e le definizioni dal contesto dei processi stocastici sfruttando la stretta relazione con tale teoria. Sulla base delle notazioni precedentemente introdotte, nel corso del lavoro saranno adottate le seguenti definizioni.

Definizione 2.3. *Sia $(\Omega, \mathfrak{F}, \mathbb{P})$ uno spazio di probabilità e sia H uno spazio vettoriale, dotato di una σ-algebra $\mathfrak{H}$. Si assuma inoltre che $\boldsymbol{\chi}$ sia integrabile rispetto alla misura $\mathbb{P}$. Allora la* media *di $\boldsymbol{\chi}$ è definita come:*

$$m := \mathbb{E}[\boldsymbol{\chi}] = \int_\Omega \boldsymbol{\chi}(\omega) \mathbb{P}(d\omega).$$

In particolare, $m \in H$ è una funzione tale che:

$$m(t) = \mathbb{E}[\boldsymbol{\chi}(t)], \quad t \in \mathcal{T}.$$

In analogia a quanto fatto per il valore atteso, è possibile definire la mediana e la moda funzionale, come mostrato ad esempio in (Ferraty e Vieu, 2006). Per quanto riguarda la misura di variabilità del dato esistono due definizioni utili da citare. Infatti, il parallelo con la teoria dei processi stocastici conduce alla formulazione della seguente definizione.

Definizione 2.4. *Con le notazioni della Definizione 2.3, si assuma che $\mathbb{E}[\boldsymbol{\chi}^2(t)] < \infty$ per ogni $t \in \mathcal{T}$. Allora, si definisce la* funzione varianza *come:*

$$\sigma(t) = \mathbb{E}[(\boldsymbol{\chi}(t) - m(t))^2], \quad t \in \mathcal{T}.$$

Inoltre si definisce la funzione autocovarianza Σ *come:*

$$\begin{aligned}\Sigma(t,z) &= \mathrm{Cov}(\boldsymbol{\chi}(t),\boldsymbol{\chi}(z)) = \\ &= \mathbb{E}[(\boldsymbol{\chi}(t)-m(t))\cdot(\boldsymbol{\chi}(z)-m(z))], \quad t,z\in\mathcal{T}.\end{aligned}$$

È altresì possibile fornire una definizione di operatore di covarianza, valida per un ente aleatorio di un qualunque spazio di Hilbert.

Definizione 2.5. *Sia* $(\Omega,\mathfrak{F},P)$ *uno spazio di probabilità e sia* $\boldsymbol{\chi}$ *una variabile aleatoria funzionale a valori in uno spazio di Hilbert* H*, dotato di prodotto scalare* $\langle\cdot,\cdot\rangle$ *e norma indotta* $||\cdot||$*, tale che* $\mathbb{E}[||\boldsymbol{\chi}||^4]<\infty$*. Si definisce operatore di covarianza* C *di* $\boldsymbol{\chi}$*, l'operatore che ad ogni* $x\in H$ *associa* $C(x)\in H$*:*

$$C(x)=\mathbb{E}[\langle\boldsymbol{\chi}-\mathbb{E}[\chi],x\rangle(\boldsymbol{\chi}-\mathbb{E}[\chi])].$$

La Definizione 2.5 è strettamente connessa alla Definizione 2.4, infatti:

$$C(x;t)=\langle\Sigma(t,\cdot),x\rangle, \quad t\in\mathcal{T}.$$

Le usuali statistiche usate nell'analisi monovariata si applicano in modo analogo al caso funzionale. In particolare, è possibile fornire una stima delle funzioni di media, varianza e autocovarianza, così come dell'operatore di covarianza, usando gli stimatori campionari corrispondenti. Nel contesto che sarà trattato in questo lavoro, vale infatti la seguente definizione.

Definizione 2.6. *Dato un campione casuale* $\chi_1,...,\chi_n$ *si definiscono i seguenti stimatori campionari:*

- Funzione Media Campionaria*:* $\overline{\chi}(t)=\frac{1}{n}\sum_{i=1}^{n}\chi_i(t), \quad t\in\mathcal{T}$*;*
- Funzione Varianza Campionaria*:* $s^2(t)=\frac{1}{n-1}\sum_{i=1}^{n}(\chi_i(t)-\overline{\chi}(t))^2, \quad t\in\mathcal{T}$*;*
- Funzione Autocovarianza Campionaria*:* $S(s,t)=\frac{1}{n-1}\sum_{i=1}^{n}(\chi_i(s)-\overline{\chi}(s))(\chi_i(t)-\overline{\chi}(t)), \quad t\in\mathcal{T}$*;*
- Operatore Covarianza Campionario*:* $\overline{C}(x)=\frac{1}{n-1}\sum_{i=1}^{n}\langle\chi_i,x\rangle\chi_i, \quad t\in\mathcal{T}$*.*

Dalla stima dell'operatore di covarianza è possibile ottenere la stima delle componenti principali funzionali, che può portare a una riduzione dimensionale (Ramsay e Silverman, 2005). Si noti tuttavia come queste misure di variabilità dipendano in modo sostanziale dalla metrica dello spazio H: la scelta dello spazio H, ovvero del prodotto scalare $\langle\cdot,\cdot\rangle$, nel quale svolgere le analisi costituisce quindi uno dei punti cruciali preliminari all'analisi stessa; a questo tema è dedicata la sezione seguente.

2.2 Scelta della Metrica, Curse of Dimensionality e Dimensionalità Pratica dei Dati Funzionali

L'approccio funzionale all'analisi di oggetti infinito-dimensionali comporta l'insorgere di alcune difficoltà tecniche rispetto a un punto di vista vettoriale. In particolare, come sottolineato da Ferraty e Vieu (2006), la sparsità dei dati aumenta inevitabilmente a fronte della crescita nella dimensionalità dei dati stessi, che, nel caso in esame, appartengono per la loro stessa natura a uno spazio infinito-dimensionale. D'altra parte, la sparsità dei dati è direttamente connessa alla misura con la quale si valuta la vicinanza tra i dati stessi, ovvero la metrica (o semi-metrica) scelta per lo spazio H.

Per questo motivo, la scelta della metrica più adatta all'analisi rappresenta un punto cruciale della FDA: se in uno spazio finito-dimensionale, come $\mathbb{R}^n$, tutte le metriche sono equivalenti, appena la dimensione dello spazio diventa infinita, l'equivalenza tra le norme decade e la scelta preliminare di una metrica rispetto ad un'altra può creare una sostanziale differenza.

Infatti, da un lato la norma è strettamente connessa alle richieste di regolarità delle funzioni in esame, identificando lo spazio di appartenenza dei dati e l'ambiente delle analisi, dall'altro lato in base alla metrica scelta è possibile lo studio e la visualizzazione delle caratteristiche salienti del campione. In particolare, la nozione di distanza influenza notevolmente l'identificazione della struttura di dipendenza degli enti aleatori in esame, quindi la stima delle componenti principali e la conseguente riduzione dimensionale: ad esempio, come sarà ancora più evidente nel corso del Capitolo 3, l'adozione della norma di L^2 comporta la considerazione dei soli valori puntuali della curva, mentre una norma di uno spazio di Sobolev consente di valutare la variabilità del dato anche attraverso le sue caratteristiche differenziali.

Si noti fin da ora che l'equivalenza matematica delle norme in $\mathbb{R}^n$ non corrisponde all'equivalenza statistica delle norme stesse (Johnson e Wichern, 2007): infatti, la necessità di tenere in considerazione la deformazione della geometria indotta dalla struttura di covarianza del campione è di importanza fondamentale nelle analisi a prescindere dalla dimensionalità del dato e sarà oggetto di discussione nel seguito (*cfr.* Definizione 3.15 in Sezione 3.5).

L'immersione del dato in uno spazio di dimensione infinita, unitamente alla scelta di una metrica adeguata, consente altresì di oltrepassare il limite dettato dal *curse of dimensionality*. Con questa espressione si fa riferimento al problema già citato della sparsità dei dati al crescere della dimensione dei dati stessi, la cui incidenza può essere valutata attraverso il numero $N(p)$ di unità nel campione appartenenti a un sottoinsieme (fissato) dello spazio campionario $\mathbb{R}^p$, al crescere della dimensione $p \in \mathbb{N}$ di quest'ultimo. Nel caso funzionale, l'analogo di tale quantità è il numero N_d di unità appartenenti a un sottoinsieme (fissato) di H, dove (H, d) indica lo spazio metrico (o semi-metrico) al quale appartiene ciascuna delle n osservazioni disponibili. Il confronto tra l'approccio della FDA e quello di tipo vettoriale, può essere condotto comparando il numero N_d di osservazioni interne a una palla di raggio r centrata in 0, ottenuto considerando l'intera curva come punto dello spazio infinito-

dimensionale H:

$$N_d = \sum_{i=1}^{n} I_{\{\frac{d(\chi_i,0)}{\max_i d(\chi_i,0)}<0.01\}}(\chi_i),$$

dove I è la funzione indicatrice, con il numero $N(p)$:

$$N(p) = \sum_{i=1}^{n} I_{\{\frac{\delta_p(\boldsymbol{x}_{i,p},0)}{\max_i \delta_p(\boldsymbol{x}_{i,p},0)}<0.01\}}(\boldsymbol{x}_{i,p}),$$

ottenuto considerandone la versione discretizzata $\{\boldsymbol{x}_{i,p} = (\chi_i(t_1), \chi_i(t_2), ..., \chi_i(t_p))\}_{i=1,...,n}$, $p = 1, 2, ..., P$, ovvero la matrice delle osservazioni:

$$\mathbb{X}_p = \begin{pmatrix} \chi_1(t_1) & \chi_1(t_2) & \cdots & \chi_1(t_p) \\ \chi_2(t_1) & \chi_2(t_2) & \cdots & \chi_2(t_p) \\ \vdots & \vdots & \vdots & \vdots \\ \chi_n(t_1) & \chi_n(t_2) & \cdots & \chi_n(t_p) \end{pmatrix}$$

che può essere considerata come un dataset multivariato costituito da n unità statistiche, ciascuna in $\mathbb{R}^p$, e valutando la distanza tra punti di $\mathbb{R}^p$ attraverso la distanza euclidea $\delta(\cdot, \cdot)$.

L'analisi della quantità $N(p)$ al crescere di p, ovvero al crescere della risoluzione di osservazione della variabile discretizzata, evidenzia come il *curse of dimensionality* sia un problema non trascurabile appena le curve osservate diventino poco regolari e autocorrelate, tendendo a 0 al crescere di p qualora l'autocorrelazione sia molto bassa (Ferraty e Vieu, 2006).

Il valore di N_d è invece dipendente soltanto dalla metrica d adottata e non risente della discretizzazione del dato: la scelta di una metrica funzionale consente quindi di superare il limite legato all'alta dimensionalità dei dati $\boldsymbol{x_{i,p}}$, fornendo una misura di sparsità indipendente dalla griglia di valutazione dei dati stessi.

In questo senso, il vantaggio di un approccio astratto è evidenziato tanto più il dato risulta irregolare e poco autocorrelato, essendo possibile, attraverso un'opportuna definizione della metrica, o semi-metrica, d, evitare alcune delle difficoltà discendenti dalla discretizzazione delle curve osservate, inevitabili in caso di approccio vettoriale.

D'altro canto, l'irregolarità del dato si riflette sulla *dimensionalità* dello stesso, intesa come la quantità di informazioni necessarie a fornirne una descrizione esaustiva (Ramsay e Silverman, 2005). Ad esempio, in assenza di errore di misura, ogni minimo o massimo locale della funzione può essere approssimato localmente da un polinomio quadratico, che è determinato attraverso tre coefficienti. La dimensionalità 'pratica' della funzione può quindi essere pensata come il numero totale dei coefficienti necessari a descrivere localmente ciascuna caratteristica saliente della funzione.

I dati funzionali sono potenzialmente di dimensionalità infinita: infatti, la completa descrizione di un dato χ potrebbe richiedere la conoscenza del valore $\chi(t)$ per ogni valore dell'ascissa t in $\mathcal{T}$. Questo accadrebbe qualora la funzione avesse un comportamento particolarmente erratico, come il moto Browniano, per il quale la conoscenza di un valore $\chi(t)$ non fornisce informazioni sul valore di $\chi(t + \tau)$, per ogni $\tau > 0$. Di norma tuttavia, grazie

alla richiesta implicita di regolarità citata in precedenza, la dimensionalità pratica dei dati funzionali trattati dalla FDA risulta finita e, in particolare, una descrizione esaustiva del dato può essere ottenuta da un numero finito di osservazioni puntuali.

2.3 Dai Dati Raccolti ai Dati Funzionali

Dal punto di vista applicativo, un dataset funzionale può derivare dalla collezione di dati in due forme:

1. Il campione di dati è costituito dalle espressioni analitiche dei dati stessi: in questo caso è possibile accedere alle proprietà analitiche e differenziali attraverso il calcolo esplicito;

2. Il campione è costituito da una collezione di dati discreti, corrispondenti a misurazioni della funzione per valori consecutivi del relativo argomento.

Nell'ultimo caso, un primo momento critico è costituito dalla rappresentazione del dato in termini funzionali. Qualora la griglia di acquisizione sia sufficientemente fitta e il processo di raccolta dei dati privo di errore, questo obiettivo può essere conseguito attraverso metodi di approssimazione numerica quali l'interpolazione; d'altra parte, se si ritiene che il processo di misura sia stato affetto da errore, è possibile filtrare tale componente di rumore ricorrendo a opportune tecniche di *smoothing*, tanto più sofisticate quanto più i dati siano sparsi o registrati lungo griglie irregolari. In particolare, in caso di presenza di errore di misura, si rende necessaria l'introduzione di una base di funzioni (B-Spline, Fourier, Wavelet, etc.) che permetta di sopperire alla mancanza di regolarità tipica di osservazioni discrete consentendo quindi di ricavare derivate fino all'ordine desiderato. Ovviamente anche la scelta della base risulta cruciale e per questo costituisce una delle problematiche più importanti da affrontare nel trattamento dei dati funzionali. In letteratura esistono numerose tecniche per questo scopo, in particolare si può fare riferimento a (Ramsay e Silverman, 2005), nel quale è presentato un quadro riassuntivo preciso a tal riguardo.

Un secondo passo preliminare all'analisi di dati funzionali è la *registrazione* dei dati (Ramsay e Silverman, 2005). Si consideri a titolo di esempio il campione di curve in Figura 2.1 (Ramsay *e altri*, 1995), relativo a 20 misurazioni della forza esercitata su uno strumento di misura durante un breve pizzico tra pollice e indice, raccolte con lo scopo di studiare la neurofisiologia della muscolatura tra le prime due dita.

Come si può notare, l'inizio della pressione è collocata arbitrariamente nel tempo rispetto all'inizio dell'esperimento, così come il picco della forza applicata allo strumento è posizionato in un istante temporale diverso per ogni osservazione. Un confronto corretto tra le curve può avvenire soltanto se le osservazioni sono state preliminarmente allineate e rappresentate nella giusta scala, temporale o spaziale, rispetto alla quale misurare il fenomeno e quindi confrontare i soggetti (Figura 2.1 a destra). Dal punto di vista formale, il processo di registrazione dei dati consiste quindi nell'identificazione di una trasformazione $r_i^{-1}(\cdot)$ per

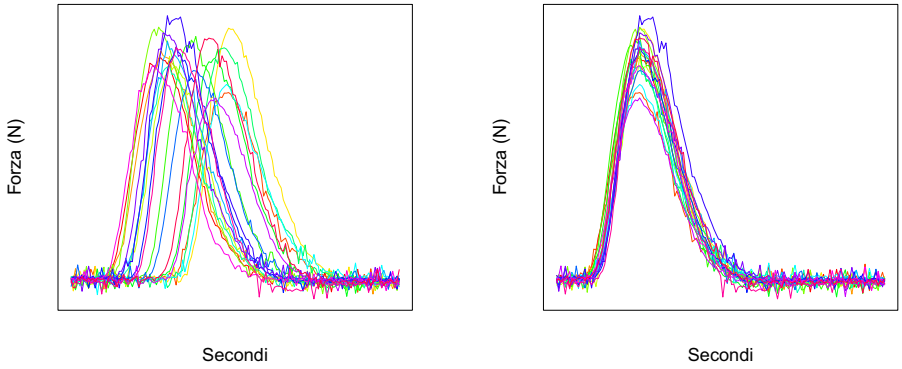

Figura 2.1: Registrazione di dati Funzionali. A sinistra: curve non registrate. A destra: curve registrate

ogni osservazione x_i che disaccoppi la variabilità di fase, ovvero il disallineamento delle curve lungo l'ascissa, dalla variabilità di ampiezza, osservabile nella differente ordinata raggiunta durante il picco di pressione.

Nel seguito del lavoro si assumerà che il dataset a disposizione sia già in forma funzionale e registrato, senza soffermarsi sui metodi e sulle procedure che abbiano condotto dai dati originali alla loro forma analitica finale.

2.4 Modelli Lineari nella FDA

L'obiettivo di questa Sottosezione è la presentazione di una breve panoramica sui modelli lineari funzionali, seguendo in particolare (Ferraty e Vieu, 2006), (Ramsay e Dalzell, 1991) e (Ramsay e Silverman, 2005).

In Tabella 2.1 sono schematizzate le situazioni che si possono presentare qualora si voglia formulare un modello lineare per dati funzionali, indicando i modelli con la notazione vettoriale e per componenti.

In alto a sinistra, è presentato un modello lineare multivariato (Johnson e Wichern, 2007), dal quale si può ricavare un modello a regressori funzionali (in alto a destra), generalizzando opportunamente la matrice disegno $\mathbb{Z}$ e adottando coefficienti $\boldsymbol{\beta}$ funzionali:

$$x_i = \sum_{l=1}^{L} \langle z_{il}, \beta_l \rangle + \varepsilon_i(t), \quad t \in \mathcal{T},\ i = 1, ..., n,$$

o un modello a risposta funzionale (in basso a sinistra), considerando coefficienti $\boldsymbol{\beta}$ funzionali e regressori multivariati:

$$x_i(t) = \sum_{l=1}^{L} z_{il}\beta_l(t) + \varepsilon_i(t), \quad t \in \mathcal{T},\ i = 1, ..., n. \tag{2.1}$$

	REGRESSORI MULTIVARIATI	REGRESSORI FUNZIONALI
RISP. MULTIVARIATA	$\boldsymbol{x} = \mathbb{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ $x_i = \sum_{l=1}^{L} z_{il}\beta_l + \varepsilon_i$ $\boldsymbol{x}, \boldsymbol{\varepsilon} \in \mathbb{R}^n$ $\boldsymbol{\beta} \in \mathbb{R}^L$ $\mathbb{Z} \in \mathbb{R}^{n,L}$	$\boldsymbol{x} = \mathbb{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ $x_i = \sum_{l=1}^{L} \langle z_{il}, \beta_l \rangle + \varepsilon_i(t)$ $\boldsymbol{x}, \boldsymbol{\varepsilon} \in \mathbb{R}^n$ $\boldsymbol{\beta} \in G \subseteq H^L$ $\mathbb{Z} \in \mathcal{L}(G, \mathbb{R}^n)$
RISP. FUNZIONALE	$\boldsymbol{x}(t) = \mathbb{Z}\boldsymbol{\beta}(t) + \boldsymbol{\varepsilon}(t)$ $x_i(t) = \sum_{l=1}^{L} z_{il}\beta_l(t) + \varepsilon_i(t)$ $\boldsymbol{x}, \boldsymbol{\varepsilon} \in H^n$ $\boldsymbol{\beta} \in G \subseteq H^L$ $\mathbb{Z} \in \mathbb{R}^{n,L}$	$\boldsymbol{x}(t) = (\mathbb{Z}\boldsymbol{\beta})(t) + \boldsymbol{\varepsilon}(t)$ $x_i(t) = \sum_{l=1}^{L} \langle z_{il}(t)\beta_l(t) \rangle + \varepsilon_i(t)$ $x_i(t) = \sum_{l=1}^{L} \langle z_{il}, \beta_l(t, \cdot) \rangle + \varepsilon_i(t)$ $\boldsymbol{x}, \boldsymbol{\varepsilon} \in H^n$ $\boldsymbol{\beta} \in G \subseteq H^L$ $\mathbb{Z} \in \mathcal{L}(G, H^n)$

Tabella 2.1: Schema dei modelli lineari funzionali.

Un modello lineare di tipo (2.1) è contemplato ad esempio in (Ramsay e Silverman, 2005) per modellare dati meteorologici simili a quelli trattati nel Capitolo 6, in cui l'obiettivo è prevedere la curva della temperatura annuale, utilizzando come regressori le zone geografiche, quindi dati multivariati.

I modelli lineari più generali sono costituiti dai modelli a regressori e risposta funzionali (in basso a destra). In particolare, qualora si ritenga che la dipendenza della risposta funzionale dai regressori sia presente soltanto per valori coincidenti dell'ascissa t, si può adottare il modello lineare:

$$\boldsymbol{x}_i(t) = \sum_{l=1}^{L} z_{il}(t)\beta_l(t) + \varepsilon_i(t), \quad t \in \mathcal{T},\ i = 1, ..., N;$$

denominato *cuncurrent model* (Ramsay e Silverman, 2005).

La scelta di un modello 'totale' (*total model*, (Ramsay e Silverman, 2005)):

$$x_i(t) = \sum_{l=1}^{L} \langle z_{il}, \beta_l(t, \cdot) \rangle + \varepsilon_i(t), \quad t \in \mathcal{T},\ i = 1, ..., N \tag{2.2}$$

che in $H = L^2$ risulta:

$$x_i(t) = \sum_{l=1}^{L} \int_{\mathcal{T}} z_l(\tau)\beta_l(\tau, t)d\tau + \varepsilon_i(t), \quad t \in \mathcal{T},\ i = 1, ..., N$$

è invece preferibile qualora si ritenga che l'influenza dei regressori avvenga trasversalmente rispetto all'ascissa del dato, consentendo in questo modo una modellazione più accurata del fenomeno.

Si precisa che la distinzione tra modelli lineari proposta in Tabella 2.1 è motivata prevalentemente da ragioni applicative legate all'efficienza computazionale. Infatti, dal punto di vista teorico, i modelli presentati rappresentano un caso particolare del modello totale (2.2) dal momento che i primi possono essere ricavati da quest'ultimo osservando che una variabile aleatoria reale può essere considerata come una variabile aleatoria funzionale di tipo costante. In particolare, i modelli presentati possono essere trattati con i medesimi procedimenti di stima, grazie al formalismo illustrato di seguito, tratto dal lavoro di Ramsay e Dalzell (1991).

Un dataset funzionale $x_1, ..., x_n$ è la realizzazione di un vettore aleatorio $(\boldsymbol{x}_1, ..., \boldsymbol{x}_n)$, e, come tale, assume valori in $H^n = H \times H \times ... \times H$, se la generica variabile $\boldsymbol{x}_i$, $i = 1, ..., n$ è a valori in H (di Hilbert).

Un modello lineare per x_i, $i = 1, ..., n$, induce una relazione lineare tra lo spazio H e qualche altro spazio G, ovvero $x_i \approx Z_i\boldsymbol{\beta}$, dove l'operatore lineare continuo $Z_i : G \to H$ è noto, $i = 1, ..., n$ e $\boldsymbol{\beta} \in G$. Non è necessario assumere che G sia uno spazio di Hilbert, mentre si assume che Z_i sia iniettivo e la sua immagine $Z_i(G)$ chiusa.

Considerando contemporaneamente tutti gli indici $i = 1, ..., n$, gli operatori Z_i definiscono un operatore $\mathbb{Z} : G \to H^n$, tale che $\mathbb{Z} = (Z_1, ..., Z_n)^T$ e $\boldsymbol{x} \approx \mathbb{Z}\boldsymbol{\beta}$, quindi l'operatore $\mathbb{Z}$ corrisponde alla matrice disegno dell'analisi multivariata.

Fissato un prodotto scalare in norma H^n, quindi una norma $\|\cdot\|$, la stima ai minimi quadrati di $\boldsymbol{\beta} \in G$ corrisponde all'elemento $\widehat{\beta} \in G$ che minimizza $\|\boldsymbol{x} - \mathbb{Z}\boldsymbol{\beta}\|^2$ e l'approssimazione $\widehat{\boldsymbol{x}} = \mathbb{Z}\boldsymbol{\beta}$ è la proiezione di $\boldsymbol{x}$ sull'immagine (chiusa) $\mathbb{Z}(G)$ di G in H^n sotto la mappa $\mathbb{Z}$. Il fatto che $\mathbb{Z}$ sia iniettiva ed abbia un'immagine chiusa assicura che l'elemento minimizzante $\boldsymbol{\beta}$ esista e sia unico.

Dal punto di vista applicativo, è tuttavia opportuno trattare separatamente i quattro casi distinti in Tabella 2.1, potendo in questo modo implementare algoritmi efficienti che sfruttino le caratteristiche di ciascun caso particolare, come avviene all'interno del pacchetto `fda` del software R.

Inoltre, la determinazione della soluzione ai minimi quadrati per i modelli a regressori multivariati può essere ricavata esplicitamente dai dati attraverso opportune matrici di proiezione, che saranno introdotte nel Capitolo 3; invece, per quanto concerne il caso di regressori funzionali, che non sarà trattato in questo lavoro, il calcolo dei parametri ottimi $\boldsymbol{\beta}$ è generalmente trattato a partire dalle proiezioni dei dati su un'opportuna base funzionale, come dettagliato ad esempio in (Ramsay e Silverman, 2005).

Nel seguito del lavoro di tesi, sarà rivolta particolare attenzione al procedimento di stima relativo al modello lineare di forma (2.1), al fine di sviluppare una metodologia che consenta di interpolare un campo funzionale con un metodo di Universal Kriging per elementi di uno spazio di Hilbert. In particolare, la procedura di stima di tale modello sarà generalizzata nel Capitolo 3 e usata nei capitoli successivi per la modellazione del *drift* di un processo stocastico funzionale.

Capitolo 3

Kriging per Elementi di uno Spazio di Hilbert: una Formulazione di Sintesi

In un numero crescente di applicazioni provenienti da svariati ambiti della ricerca industriale, i dati osservati sono curve spazialmente distribuite, caratterizzate cioè dalla presenza di una struttura di dipendenza spaziale. In questo caso le tecniche sviluppate nell'ambito della Functional Data Analysis risultano talvolta inappropriate, in quanto non incorporano nei modelli e nelle procedure di stima la struttura di dipendenza esistente tra le funzioni osservate.

Per questo motivo, è sorta la necessità di estendere al caso funzionale alcune tecniche statistiche in uso in ambito finito-dimensionale per il trattamento di dati spazialmente correlati, come i metodi di regressione spaziale e la geostatistica.

Nella prima direzione è sviluppato, ad esempio, il lavoro di Yamanishi e Tanaka (2003), dedicato alla formulazione di *geographically weighted regression models*, coniugando i relativi metodi finito-dimensionali (Brunsdon *e altri*, 1996), con le tecniche di stima dei modelli lineari funzionali (*cfr.* Capitolo 2, Sezione 2.4)

L'estensione delle tecniche geostatistiche all'ambito funzionale ha recentemente ottenuto particolare attenzione nel contesto della previsione spaziale per dati funzionali. La ricerca in quest'ambito si è infatti concentrata nella direzione delle tecniche di kriging e cokriging, a partire dal lavoro pionieristico di Goulard e Voltz (1993), nel quale sono proposti tre approcci alla previsione di curve, di cui due parametrici multivariati basati su metodi di cokriging e uno di kriging funzionale non parametrico.

L'approccio funzionale non parametrico, proposto ma non indagato in (Goulard e Voltz, 1993), è stato adottato e sviluppato in tempi recenti nell'ottica della FDA in (Giraldo *e altri*, 2008a), (Giraldo *e altri*, 2008b), (Giraldo, 2009), (Giraldo *e altri*, 2010c), (Delicado *e altri*, 2010), (Giraldo *e altri*, 2010a) e, contemporaneamente, in (Monestiez e Nerini, 2008). In questi lavori, sono proposti stimatori lineari dai dati, a coefficienti costanti e non costanti,

per la previsione di curve dello spazio L^2, nella considerazione dei dati funzionali come punti in uno spazio infinito-dimensionale (*cfr.* Capitolo 6, Sottosezione 6.4.1).

L'obiettivo del presente capitolo è fornire una formulazione di sintesi del metodo di Ordinary Kriging a coefficienti costanti presente in (Giraldo *e altri*, 2008a), all'interno di un quadro coerente di definizioni e ipotesi, quali la stazionarietà e l'isotropia, estendendo la formulazione esistente a processi stocastici spazialmente non stazionari su spazi di Hilbert. Questo scopo sarà perseguito attraverso un approccio astratto in linea con quello presente in (Hörmann e Kokoszka, 2011).

Per questo motivo, dopo aver introdotto la definizione di media e una nuova definizione di covarianza spaziale di tipo 'globale' per processi stocastici funzionali, saranno individuate opportune definizioni di stazionarietà e isotropia, grazie alle quali sarà possibile procedere alla riformulazione dei metodi di Ordinary Kriging (Giraldo *e altri*, 2008a) e alla formulazione del metodo di Universal Kriging, tuttora non presente in letteratura. Inoltre, saranno proposti opportuni stimatori per il variogramma di un processo aleatorio funzionale stazionario e, per il caso non stazionario, sarà individuata un procedimento di stima per il termine di *drift*, sulla base di un modello lineare a residuo spazialmente correlato.

3.1 Funzioni di Media e Covarianza per Processi Stocastici Funzionali

Si consideri un processo stocastico funzionale:

$$\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}, \tag{3.1}$$

dove $\chi_{s_i} : \mathcal{T} \to \mathbb{R}$, sia una variabile funzionale per ogni $s_i \in D$, con $\mathcal{T}$ sottoinsieme compatto di $\mathbb{R}$. Siano $s_1, ..., s_n$ n localizzazioni nel dominio spaziale D, in corrispondenza delle quali è osservata una realizzazione del processo stocastico (3.1), denotata con $\chi_{s_1}, ..., \chi_{s_n}$.

Al fine di riformulare in modo astratto l'Ordinary Kriging, è necessario introdurre delle nuove nozioni di media e covarianza del processo, attraverso le quali ridefinire le assunzioni di stazionarietà e isotropia.

Definizione 3.1. *Sia $\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}$ un processo stocastico funzionale, tale che $\boldsymbol{\chi}_s$ è un elemento aleatorio di uno spazio di Hilbert H, dotato del prodotto interno $\langle \cdot, \cdot \rangle$ e della norma da questo indotta $\| \cdot \|$. Si definisce* media spaziale *del processo:*

$$m_s = \mathbb{E}[\boldsymbol{\chi}_s], \tag{3.2}$$

dove la media per $s \in D$ fissato è da intendersi come in Definizione 2.3. Nelle stesse ipotesi, si definiscono, qualora esistano, la funzione di varianza spaziale *del processo:*

$$\sigma^2(s) = \mathrm{Var}(\boldsymbol{\chi}_s) = \mathbb{E}[\|\boldsymbol{\chi}_s\|^2] - \|\mathbb{E}[\boldsymbol{\chi}_s]\|^2 \tag{3.3}$$

e la funzione di covarianza spaziale *o* covariogramma *del processo:*

$$C(s_i, s_j) = \mathrm{Cov}(\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j}) = \mathbb{E}[\langle \boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j} \rangle] - \langle \mathbb{E}[\boldsymbol{\chi}_{s_i}], \mathbb{E}[\boldsymbol{\chi}_{s_j}] \rangle. \tag{3.4}$$

Quando opportuno, per evitare confusione, si indicherà a pedice lo spazio di riferimento per il prodotto interno $\langle \cdot, \cdot \rangle$, ovvero si userà la notazione $\langle \cdot, \cdot \rangle_H$ al posto della più semplice $\langle \cdot, \cdot \rangle$. Lo stesso dicasi per la norma $\| \cdot \|$ indotta da $\langle \cdot, \cdot \rangle_H$.

La funzione di covarianza spaziale, qualora sia ben definita, gode delle proprietà usuali della covarianza: in particolare, essa definisce un prodotto scalare nello spazio degli elementi aleatori definiti su $H(\lambda \otimes \mathbb{P}) = (\mathcal{T} \times \Omega, \mathfrak{B}(\mathcal{T}) \otimes \mathfrak{F}, \lambda \otimes \mathbb{P})$, costituito dalle classi di equivalenza di funzioni quadrato integrabili rispetto alla misura prodotto $\lambda \otimes \mathbb{P}$, ovvero delle funzioni $\boldsymbol{\mathcal{X}}$ tali che:

$$\mathbb{E}[\|\boldsymbol{\mathcal{X}}\|_H^2] < \infty.$$

Si consideri infatti su questo spazio il prodotto scalare definito come:

$$\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle_{H(\lambda \otimes \mathbb{P})} := \mathrm{Cov}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}}) = \mathbb{E}[\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle_H] - \langle \mathbb{E}[\boldsymbol{\mathcal{X}}], \mathbb{E}[\boldsymbol{\mathcal{Y}}] \rangle_H.$$

Allora si verifica facilmente che $\langle \cdot, \cdot \rangle_{H(\lambda \otimes \mathbb{P})}$ è una forma che gode delle seguenti proprietà:

$$\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}} \rangle_{H(\lambda \otimes \mathbb{P})} = \langle \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{X}} \rangle_{H(\lambda \otimes \mathbb{P})} \tag{3.5}$$

$$\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \rangle_{H(\lambda \otimes \mathbb{P})} = \mathbb{E}[\|\boldsymbol{\mathcal{X}}\|_H^2] - \|\mathbb{E}[\boldsymbol{\mathcal{X}}]\|_H^2 \geq 0 \tag{3.6}$$

$$\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{X}} \rangle_{H(\lambda \otimes \mathbb{P})} = 0 \quad \Rightarrow \quad \boldsymbol{\mathcal{X}} = 0, \quad [\lambda \otimes \mathbb{P}] - q.o. \tag{3.7}$$

$$\langle a\boldsymbol{\mathcal{X}} + b\boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Z}} \rangle_{H(\lambda \otimes \mathbb{P})} = a\langle \boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Z}} \rangle_{H(\lambda \otimes \mathbb{P})} + b\langle \boldsymbol{\mathcal{Y}}, \boldsymbol{\mathcal{Z}} \rangle_{H(\lambda \otimes \mathbb{P})} \tag{3.8}$$

Le proprietà di simmetria (3.5), annullamento (3.7) e linearità (3.8) (e la conseguente bilinearità) discendono direttamente dalle medesime proprietà del prodotto scalare $\langle \cdot, \cdot \rangle_H$; la (3.6) segue invece dalla disuguaglianza di Jensen (valida in qualsiasi spazio topologico).

Dalla Definizione 3.1, discende la seguente.

Definizione 3.2. *Nelle ipotesi della Definizione 3.1, si definisce* semivariogramma *del processo* (3.1) *la funzione:*

$$\gamma(s_i, s_j) = \frac{1}{2} \mathrm{Var}(\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}). \tag{3.9}$$

Dalla definizione appena fornita, si ricava che, qualora il covariogramma e il variogramma siano definiti, la relazione tra essi è data dalla seguente estensione della (1.4):

$$\gamma(s_i, s_j) = C(0) - C(s_i, s_j). \tag{3.10}$$

Le funzioni di varianza e covarianza spaziale definite dalla (3.3) e (3.4), così come il semivariogramma definito dalla (3.9), sono misure di variabilità globali del campo aleatorio, poichè dipendono soltanto dalla localizzazione spaziale delle variabili. È possibile considerare altre definizioni di funzione di varianza e covarianza, fornendone una definizione puntuale (Giraldo, 2009), tuttavia la definizione proposta in questo lavoro consente di riformulare in modo naturale ed elegante i metodi di kriging ordinario e universale, in ipotesi di stazionarietà e isotropia il più possibile lasche.

Esempio 3.3 (Funzione di covarianza spaziale e variogramma in L^2)**.** *Si consideri come spazio H, lo spazio $L^2 = L^2(\mathcal{T})$ delle classi di equivalenza di funzioni a quadrato integrabile su $\mathcal{T} = [0, T]$:*

$$L^2(\mathcal{T}) = \{f : \mathcal{T} \to \mathbb{R},\ t.c.\ \int_{\mathcal{T}} |f(t)|^2 dt < \infty\}.$$

Si doti L^2 del prodotto scalare e della norma da esso indotta:

$$< f, g > = \int_{\mathcal{T}} f(t) \cdot g(t) dt$$
$$\|f\| = \sqrt{\int_{\mathcal{T}} |f(t)|^2 dt},$$

dove $f, g \in L^2$.

Si noti che, poiché i rappresentanti di una stessa classe di equivalenza sono di fatto indistinguibili, con abuso di linguaggio si parlerà nel seguito di funzioni di L^2, indicando un rappresentante della classe di equivalenza in oggetto.

Si assuma inoltre che siano definite per il processo la funzione di media $m_s, s \in D$, la funzione di covarianza $C(s_i, s_j)$, $s_i, s_j \in D$ e la funzione di covarianza puntuale: $C(s_i, s_j; t) = \mathrm{Cov}(\boldsymbol{\chi}_{s_i}(t), \boldsymbol{\chi}_{s_j}(t))$ (Giraldo e altri*, 2008a).*

Allora, con semplici calcoli, si ottengono le seguenti relazioni:

$$\mathrm{Cov}(\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j}) = \mathbb{E}\left[\int_{\mathcal{T}} \boldsymbol{\chi}_{s_i} \cdot \boldsymbol{\chi}_{s_j} dt\right] - \int_{\mathcal{T}} \mathbb{E}[\boldsymbol{\chi}_{s_i}](t) \cdot \mathbb{E}[\boldsymbol{\chi}_{s_j}](t) dt = \tag{3.11}$$
$$= \int_{\mathcal{T}} \mathbb{E}[\boldsymbol{\chi}_{s_i}(t) \cdot \boldsymbol{\chi}_{s_j}(t)] dt - \int_{\mathcal{T}} m_{s_i}(t) \cdot m_{s_j}(t) dt = \tag{3.12}$$
$$= \int_{\mathcal{T}} \{\mathbb{E}[\boldsymbol{\chi}_{s_i}(t) \cdot \boldsymbol{\chi}_{s_j}(t)] - m_{s_i}(t) \cdot m_{s_j}(t)\} dt =$$
$$= \int_{\mathcal{T}} \{\mathbb{E}[\boldsymbol{\chi}_{s_i}(t) \cdot \boldsymbol{\chi}_{s_j}(t)] - \mathbb{E}[\boldsymbol{\chi}_{s_i}(t)] \cdot \mathbb{E}[\boldsymbol{\chi}_{s_j}(t)]\} dt =$$
$$= \int_{\mathcal{T}} \mathrm{Cov}(\boldsymbol{\chi}_{s_i}(t), \boldsymbol{\chi}_{s_j}(t)) dt$$
$$= \int_{\mathcal{T}} C(s_i, s_j; t) dt,$$

dove nella (3.11) *la media è intesa come nella Definizione* (3.1)*, mentre dalla* (3.12) *media e covarianza sono definite puntualmente con la definizione usuale di $\mathbb{R}$. Il passaggio da* (3.11) *a* (3.12) *è giustificato dal teorema di Fubini-Tonelli.*

In modo analogo, per la varianza si ricava (sempre usando Fubini):

$$\mathrm{Var}(\boldsymbol{\chi}_{s_i}) = \mathbb{E}\left[\int_{\mathcal{T}} |\boldsymbol{\chi}_{s_i}|^2 dt\right] - \int_{\mathcal{T}} |\mathbb{E}[\boldsymbol{\chi}_{s_i}](t)|^2 dt =$$
$$= \int_{\mathcal{T}} \left\{\mathbb{E}[|\boldsymbol{\chi}_{s_i}(t)|^2] - |m_{s_i}(t)|^2\right\} dt =$$
$$= \int_{\mathcal{T}} \mathrm{Var}(\boldsymbol{\chi}_{s_i}(t)) dt.$$

Infine per il variogramma:

$$\begin{aligned}\mathrm{Var}(\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}) &= \mathbb{E}[\|\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}\|^2] - \|\mathbb{E}[\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}]\|^2 = \\ &= \mathbb{E}[\|\boldsymbol{\chi}_{s_i}\|^2] + \mathbb{E}[\|\boldsymbol{\chi}_{s_j}\|^2] - 2\mathbb{E}[<\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j}>] - 2\|m_{s_i} - m_{s_j}\|^2 = \\ &= \mathrm{Var}(\boldsymbol{\chi}_{s_i}) + \mathrm{Var}(\boldsymbol{\chi}_{s_j}) - 2\,\mathrm{Cov}(\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j}),\end{aligned}$$

quindi il variogramma risulta:

$$\begin{aligned}\gamma_{s_i,s_j} &= C_0 - C(s_i, s_j) = \\ &= \frac{1}{2}\left\{\mathbb{E}\left[\int_{\mathcal{T}} \boldsymbol{\chi}_{s_i}(t)^2 dt\right] + \mathbb{E}\left[\int_{\mathcal{T}} \boldsymbol{\chi}_{s_j}(t)^2 dt\right] - 2\mathbb{E}\left[\int_{\mathcal{T}} \boldsymbol{\chi}_{s_i}(t) \cdot \boldsymbol{\chi}_{s_j}(t) dt\right]\right\} + \\ &\quad - \frac{1}{2}\left\{\int_{\mathcal{T}} \mathbb{E}[\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}](t)^2 dt\right\} = \\ &= \frac{1}{2}\left\{\mathbb{E}\left[\int_{\mathcal{T}} (\boldsymbol{\chi}_{s_i}(t) - \boldsymbol{\chi}_{s_j}(t))^2 dt\right] - \int_{\mathcal{T}} \mathbb{E}[\boldsymbol{\chi}_{s_i}(t) - \boldsymbol{\chi}_{s_j}(t)]^2 dt\right\} = \\ &= \frac{1}{2}\left\{\int_{\mathcal{T}} \{\mathbb{E}[(\boldsymbol{\chi}_{s_i}(t) - \boldsymbol{\chi}_{s_j}(t))^2] - \mathbb{E}[\boldsymbol{\chi}_{s_i}(t) - \boldsymbol{\chi}_{s_j}(t)]^2\} dt\right\} = \\ &= \frac{1}{2}\left\{\int_{\mathcal{T}} \mathrm{Var}(\boldsymbol{\chi}_{s_i}(t) - \boldsymbol{\chi}_{s_j}(t)) dt\right\} = \\ &= \int_{\mathcal{T}} \gamma(s_i, s_j; t) dt.\end{aligned}$$

Pertanto, in L^2, il variogramma spaziale è la versione integrata del variogramma definito puntualmente nell'ipotesi di stazionarietà dei processi reali $\{\boldsymbol{\chi}_s(t) : s \in D \subseteq \mathbb{R}^d\}$, per ogni $t \in \mathcal{T}$.

Esempio 3.4 (Funzione di covarianza spaziale e variogramma in $\mathcal{H}^k$)**.** *Si consideri un processo stocastico funzionale a valori nello spazio di Sobolev $\mathcal{H}^k, k \geq 1$, ossia il sottoinsieme di L^2 costituito dalle classi di equivalenza di funzioni con derivate in senso debole appartenenti a L^2 fino a un certo ordine k.*

$$\mathcal{H}^k(\mathcal{T}) = \{f : \mathcal{T} \to \mathbb{R},\ t.c.\ D^\alpha f \in L^2, \forall \alpha \leq k,\ \alpha \in \mathbb{N}\}.$$

Si doti $\mathcal{H}^k$ del prodotto scalare e della norma definiti come:

$$\begin{aligned}< f, g >_{\mathcal{H}^k} &= \sum_{\alpha=1}^{k} \langle D^\alpha f, D^\alpha g \rangle_{L^2} \\ \|f\|_{\mathcal{H}^k} &= \sqrt{\sum_{\alpha=1}^{k} \|D^\alpha f\|_{L^2}^2} = \sqrt{\sum_{\alpha=1}^{k} \int_{\mathcal{T}} |D^\alpha f(t)|^2 dt},\end{aligned}$$

dove $f, g \in \mathcal{H}^k$. Si assuma che siano definite per il processo le funzioni media $m_s, s \in D$, covarianza spaziale $C(s_i, s_j)$, $s_i, s_j \in D$ e le funzioni covarianza puntuale: $C^\alpha(s_i, s_j; t) = \mathrm{Cov}(D^\alpha \boldsymbol{\chi}_{s_i}(t), D^\alpha \boldsymbol{\chi}_{s_j}(t))$, $\alpha = 0, ..., k$.

Analogamente a quanto ricavato nell'Esempio 3.3, sfruttando il teorema di Fubini-Tonelli, si ottiene allora il seguente risultato:

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j})_{\mathcal{H}^k} &= \mathbb{E}[\langle \boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j} \rangle_{\mathcal{H}^k}] - \langle m_{s_i}, m_{s_j} \rangle_{\mathcal{H}^k} = \\
&= \mathbb{E}\left[\sum_{\alpha=1}^{k} \langle D^\alpha \boldsymbol{\chi}_{s_i}, D^\alpha \boldsymbol{\chi}_{s_j} \rangle_{L^2} \right] - \sum_{\alpha=1}^{k} \langle D^\alpha m_{s_i}, D^\alpha m_{s_j} \rangle_{L^2} = \\
&= \sum_{\alpha=1}^{k} \left\{ \mathbb{E}[\langle D^\alpha \boldsymbol{\chi}_{s_i}, D^\alpha \boldsymbol{\chi}_{s_j} \rangle_{L^2}] - \langle D^\alpha m_{s_i}, D^\alpha m_{s_j} \rangle_{L^2} \right\} = \\
&= \sum_{\alpha=1}^{k} \mathrm{Cov}(D^\alpha \boldsymbol{\chi}_{s_i}, D^\alpha \boldsymbol{\chi}_{s_j})_{L^2} = \\
&= \sum_{\alpha=1}^{k} \int_{\mathcal{T}} C^\alpha(s_i, s_j; t) dt.
\end{aligned}$$

Si noti che i passaggi precedenti sono ben definiti in quanto la funzione di media m_s, $s \in D$ è della stessa natura funzionale di $\boldsymbol{\chi}_s$, quindi, in particolare, possiede derivare deboli in L^2 fino all'ordine k.

Applicando gli stessi ragionamenti alla varianza, si ottiene:

$$\begin{aligned}
\mathrm{Var}(\chi_{s_i})_{\mathcal{H}^k} &= \sum_{\alpha=1}^{k} \mathrm{Var}(D^\alpha \boldsymbol{\chi}_{s_i})_{L^2} = \\
&= \sum_{\alpha=1}^{k} \int_{\mathcal{T}} \mathrm{Var}(D^\alpha \chi_{s_i}(t)) dt.
\end{aligned}$$

Infine per il variogramma:

$$\begin{aligned}
\gamma(s_i, s_j) &= \frac{1}{2} \mathrm{Var}(\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j})_{\mathcal{H}^k} = \\
&= \frac{1}{2} \sum_{\alpha=1}^{k} \mathrm{Var}(D^\alpha \boldsymbol{\chi}_{s_i} - D^\alpha \boldsymbol{\chi}_{s_i})_{L^2} = \\
&= \sum_{\alpha=1}^{k} \int_{\mathcal{T}} \gamma^\alpha(s_i, s_j; t) dt,
\end{aligned}$$

con la convenzione: $\gamma^\alpha(s_i, s_j; t) = \mathrm{Var}(D^\alpha \boldsymbol{\chi}_{s_i}(t) - D^\alpha \boldsymbol{\chi}_{s_j}(t))$: $\gamma^\alpha(s_i, s_j)$ risultano ben definite per ogni $\alpha = 0, 1, ..., k$ grazie all'ipotesi di esistenza delle funzioni $C^\alpha(s_i, s_j; \cdot)$.

Dal punto di vista teorico, è interessante sottolineare la relazione esistente tra la definizione di covarianza espressa dalla (3.4) e l'operatore covarianza definito come segue (Hörmann e Kokoszka, 2011).

Definizione 3.5. *Sia $\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}$ un processo stocastico a valori in uno spazio di Hilbert H, dotato di prodotto scalare $\langle \cdot, \cdot \rangle$ e norma indotta $\| \cdot \|$, tale che $\mathbb{E}[\|\boldsymbol{\chi}_s\|^2] < \infty$; sia*

m_s *la sua media. Si definisce* operatore di cross-covarianza C_{s_1,s_2} *del processo, l'operatore che ad ogni* $x \in H$ *associa* $C_{s_1,s_2}(x) \in H$*:*

$$C_{s_1,s_2}(x) = \mathbb{E}[\langle \boldsymbol{\chi}_{s_1} - m_{s_1}, x\rangle(\boldsymbol{\chi}_{s_2} - m_{s_2})], \quad s_1, s_2 \in D. \tag{3.13}$$

Si definisce inoltre operatore di covarianza*:*

$$C_{s,s}(x) = \mathbb{E}[\langle \boldsymbol{\chi}_s - m_s, x\rangle(\boldsymbol{\chi}_s - m_s)], \quad s \in D. \tag{3.14}$$

Sia ora $\{e_j, j \in \mathbb{N}\}$ una base ortonormale dello spazio H, che esiste in quanto H è assunto essere di Hilbert, allora, grazie all'identità di Parseval,

$$\begin{aligned} \|\boldsymbol{\chi}_s - m_s\|^2 &= \sum_{j=1}^{\infty} |\langle \boldsymbol{\chi}_s, e_j\rangle|^2 \\ \langle \boldsymbol{\chi}_{s_1} - m_{s_1}, \boldsymbol{\chi}_{s_2} - m_{s_2}\rangle &= \sum_{j=1}^{\infty} \langle \boldsymbol{\chi}_{s_1} - m_{s_1}, e_j\rangle\langle \boldsymbol{\chi}_{s_2} - m_{s_2}, e_j\rangle. \end{aligned}$$

Si noti inoltre che, per la linearità della valore atteso e del prodotto scalare in H, valgono le seguenti uguaglianze:

$$\begin{aligned} \langle C_{s_1,s_2}(e_j), e_j\rangle &= \langle \mathbb{E}[\langle \boldsymbol{\chi}_{s_1} - m_{s_1}, e_j\rangle \boldsymbol{\chi}_{s_2} - m_{s_2}], e_j\rangle = \\ &= \mathbb{E}[\langle \boldsymbol{\chi}_{s_1} - m_{s_1}, e_j\rangle\langle \boldsymbol{\chi}_{s_2} - m_{s_2}, e_j\rangle]. \end{aligned}$$

La relazione tra l'operatore di cross-covarianza e la covarianza spaziale definita in precedenza si ottiene quindi con i seguenti calcoli:

$$\begin{aligned} C(s_1, s_2) = \mathbb{E}[\langle \boldsymbol{\chi}_{s_1} - m_{s_1}, \boldsymbol{\chi}_{s_2} - m_{s_1}\rangle] &= \mathbb{E}\left[\sum_{j=1}^{\infty} \langle \boldsymbol{\chi}_{s_1} - m_{s_1}, e_j\rangle\langle \boldsymbol{\chi}_{s_2} - m_{s_2}, e_j\rangle\right] = \\ &= \sum_{j=1}^{\infty} \mathbb{E}[\langle \boldsymbol{\chi}_{s_1} - m_{s_1}, e_j\rangle\langle \boldsymbol{\chi}_{s_2} - m_{s_2}, e_j\rangle] = \\ &= \sum_{j=1}^{\infty} \langle C_{s_1,s_2}(e_j), e_j\rangle. \end{aligned} \tag{3.15}$$

Il primo passaggio è giustificato dalla citata identità di Parseval, mentre la terza uguaglianza è valida grazie al teorema della convergenza dominata per serie in quanto $\langle \boldsymbol{\chi}_{s_1} - m_{s_1}, e_j\rangle\langle \boldsymbol{\chi}_{s_2} - m_{s_2}, e_j\rangle$ ha valore atteso finito se l'operatore di covarianza è ben definito e:

$$\begin{aligned} \sum_{j=1}^{\infty} \mathbb{E}[|\langle \boldsymbol{\chi}_{s_1} - m_{s_1}, e_j\rangle\langle \boldsymbol{\chi}_{s_2} - m_{s_2}, e_j\rangle|] &\leq \sum_{j=1}^{\infty} \mathbb{E}[\max\{|\langle \boldsymbol{\chi}_{s_1} - m_{s_1}, e_j\rangle|^2, |\langle \boldsymbol{\chi}_{s_2} - m_{s_2}, e_j\rangle|^2\}] \leq \\ &\leq \sum_{j=1}^{\infty} \mathbb{E}[|\langle \boldsymbol{\chi}_{s_1} - m_{s_1}, e_j\rangle|^2 + |\langle \boldsymbol{\chi}_{s_2} - m_{s_2}, e_j\rangle|^2] = \\ &= \mathbb{E}[\|\boldsymbol{\chi}_{s_1} - m_{s_1}\|]^2 + \mathbb{E}[\|\boldsymbol{\chi}_{s_2} - m_{s_2}\|]^2 < \infty. \end{aligned}$$

La (3.15) indica che la covarianza spaziale $C(s_1, s_2)$ definita dalla (3.4), è in realtà la traccia dell'operatore cross-covarianza C_{s_1,s_2}, la cui stima è sufficiente per determinare la previsione lineare ottima con il metodo di kriging, come sarà mostrato nella Sezione 3.4. Per questo motivo, nel seguito del lavoro saranno indicati con i nomi di *trace-covariogram* e *trace-variogram* rispettivamente le funzioni $C(s_i, s_j)$ e $2\gamma(s_i, s_j)$ definiti dalle (3.4) e (3.9). Il termine *trace-variogram* è tratto dal lavoro di Giraldo *e altri* (2008a), nel quale con tale espressione è indicata la quantità:

$$\int_{\mathcal{T}} \gamma(s_i, s_j; t)dt, \tag{3.16}$$

ottenuta nella formulazione dell'Ordinary Kriging, sulla base di definizioni puntuali di varianza e covarianza del processo. Tuttavia, come ricavato in precedenza nell'Esempio 3.3, il semivariogramma spaziale è pari a (3.16) solo nel caso particolare di L^2, pertanto, in questo lavoro, con tale terminologia si farà riferimento al concetto più generale di semivariogramma espresso dalla Definizione 3.2, valido in qualsiasi spazio di Hilbert H.

3.2 Stazionarietà e Isotropia

Una volta definita la funzione media e un'opportuna misura di variabilità spaziale, è fondamentale stabilire delle ipotesi sotto le quali queste quantità siano ben definite. Per questo motivo saranno ora fornite le definizioni di stazionarietà e isotropia, estendendo in modo naturale le nozioni introdotte nel Capitolo 1 relativamente ai processi aleatori a valori in spazi finito-dimensionali.

Nel corso del lavoro sarà sempre pensato, anche in ambito infinito-dimensionale, che la variabile aleatoria $\boldsymbol{\chi}_s$ possa essere disaccoppiata in una componente deterministica, la media locale chiamata *drift*, e una componente stocastica stazionaria, nel senso che sarà ora precisato. Tale dicotomia sarà indicata con la seguente notazione:

$$\boldsymbol{\chi}_s = m_s + \boldsymbol{\delta}_s. \tag{3.17}$$

Nella formulazione delle ipotesi distribuzionali del processo $\boldsymbol{\chi}_s$, ovvero del residuo $\boldsymbol{\delta}_s$, analogamente al caso finito-dimensionale possono essere stabilite tre diverse nozioni di stazionarietà: forte, debole (o al second'ordine) e intrinseca.

Definizione 3.6. $\{\boldsymbol{\chi}_s, s \in D\}$ *si dice processo stocastico funzionale* stazionario in senso stretto *se* $(\boldsymbol{\chi}_{s_1}, ..., \boldsymbol{\chi}_{s_n})^t$ *e* $(\boldsymbol{\chi}_{s_1+h}, ..., \boldsymbol{\chi}_{s_n+h})^t$ *hanno la medesima distribuzione congiunta, per ogni* h*, per ogni successione* $\{s_k\}_{k=1}^n$ *in* $D \subseteq \mathbb{R}^d$*, per ogni* n*.*

Definizione 3.7. $\{\boldsymbol{\chi}_s, s \in D\}$ *si dice processo stocastico funzionale* debolmente stazionario *se:*

$$\mathbb{E}[\boldsymbol{\chi}_s] = m, \quad \forall\, s \in D \subseteq \mathbb{R}^d$$
$$\mathrm{Cov}(\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j}) = \mathbb{E}[\langle \boldsymbol{\chi}_{s_i} - m, \boldsymbol{\chi}_{s_i} - m \rangle] = C(\boldsymbol{h}), \quad \forall\, s_i, s_j \in D \subseteq \mathbb{R}^d,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j.$$

Definizione 3.8. $\{\boldsymbol{\chi}_s, s \in D\}$ *è un campo aleatorio* intrinsecamente stazionario *se è caratterizzato dalle seguenti ipotesi:*

$$\mathbb{E}[\boldsymbol{\chi}_s] = m, \quad \forall \boldsymbol{s} \in D$$
$$\mathrm{Var}(\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}) = \mathbb{E}[\|\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}\|^2] = 2\gamma(\boldsymbol{h}), \quad \forall s_i, s_j \in D,\ \boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j.$$

Anche in questo contesto, il covariogramma e il variogramma sono caratterizzati da alcune proprietà.

Proprietà 3.9 (Proprietà del covariogramma)**.** *Sia $C(\boldsymbol{h})$ la funzione di covarianza spaziale del processo* (3.1)*, allora:*

(i) $C(\boldsymbol{h})$ *è una funzione definita positiva:*

$$\sum_i \sum_j \lambda_i \lambda_j C(s_i - s_j) \geq 0, \quad \forall\, \lambda_i, \lambda_j, s_i, s_j \in D.$$

(ii) $C(\boldsymbol{h}) = C(-\boldsymbol{h})$

(iii) $|C(\boldsymbol{h})| \leq C(\boldsymbol{0})$

Proprietà 3.10 (Proprietà del variogramma)**.** *Sia $\gamma(\boldsymbol{h})$ il semivariogramma del processo* (3.1)*, allora:*

(i) $\gamma(\boldsymbol{h})$ *è una funzione condizionatamente definita negativa:*

$$\sum_i \sum_j \lambda_i \lambda_j \gamma(s_i - s_j) \leq 0, \quad \forall\, s_i, s_j, \in D, \forall\, \lambda_i, \lambda_j \quad t.c. \quad \sum_i \lambda_i = 0$$

(ii) $\gamma(\boldsymbol{h}) = \gamma(-\boldsymbol{h})$*;*

(iii) $\gamma(\boldsymbol{h}) \geq 0$*;*

(iv) $\gamma(\boldsymbol{0}) = 0$.

Le Proprietà 3.9 (i) e 3.10 (i) garantisono che le combinazioni lineari (ammissibili) di variabili $\boldsymbol{\chi}_s$ abbiano varianza positiva. Le altre proprietà sono invece proprietà strutturali delle funzioni covariogramma e variogramma, che seguono immediatamente dalla relativa definizione. Anche per il caso funzionale sarà quindi necessaria l'adozione di opportuni modelli parametrici validi, per la cui stima si rimanda al Capitolo 4.

La definizione di isotropia si estende in modo naturale a partire da quanto illustrato nel Capitolo 1.

Definizione 3.11. *Un processo stocastico funzionale stazionario si dice* isotropo *se il variogramma che lo caratterizza è isotropo, ovvero è tale che:*

$$\mathrm{Var}(\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}) = 2\gamma(h), \quad h = \|s_i - s_j\|, \forall s_i, s_j \in D.$$

In caso contrario, il processo è detto anisotropo.

Si noti che le definizioni appena fornite dipendono fortemente dalla metrica sullo spazio H indotta dal prodotto scalare definito sullo spazio stesso. Si considerino gli Esempi 3.3 e 3.4: svolgendo le analisi in L^2 l'ipotesi di stazionarietà è limitata alla sola funzione, mentre in H^k l'assunzione di stazionarietà contempla anche le prime k derivate deboli. Pertanto, anche in questo caso, la scelta preliminare dello spazio nel quale condurre l'analisi risulta cruciale e la selezione dello spazio opportuno deve avvenire in base alle caratteristiche di regolarità della funzione e alle finalità dell'analisi.

Si consideri ora un processo stocastico funzionale a valori nello spazio di Hilbert H di media m_s; sia $\{e_j, j \geq 1\}$ una base ortonormale fissata di H. In queste ipotesi, qualsiasi campo aleatorio (3.1) scritto in forma (3.17) ammette il seguente sviluppo:

$$\boldsymbol{\chi}_s = m_s + \sum_{j=1}^{\infty} \xi_j(s) e_j, \tag{3.18}$$

dove $\xi_j(s)$ sono variabili aleatorie a media nulla.

La scrittura (3.18) è significativa poiché assicura l'esistenza di processi spaziali funzionali stazionari nei sensi specificati, essendo possibile fornirne una costruzione diretta attraverso tale definizione (Hörmann e Kokoszka, 2011). Ad esempio, qualora la media del processo sia costante rispetto alla variabile spaziale s, $m_s = m$, dal momento che i coefficienti aleatori $\xi_j(s)$ sono le proiezioni sulla base ortonormale della variabile aleatoria funzionale $\boldsymbol{\chi}_s$:

$$\xi_j(s) = \langle \boldsymbol{\chi}_s - m, e_j \rangle,$$

il processo funzionale (3.1) è stazionario in senso stretto se e solo se ciascun processo scalare ξ_j lo è. Inoltre, dall'identità di Parseval,

$$\|\boldsymbol{\chi}_s - m\|^2 = \sum_{j=1}^{\infty} \xi_j^2(s),$$

segue la seguente equivalenza:

$$\mathbb{E}[\|\boldsymbol{\chi}_s\|^2] < \infty \quad \Leftrightarrow \quad \sum_{j=1}^{\infty} \mathbb{E}[\xi_j^2(s)] < \infty.$$

Pertanto, in linea di principio, tutte le proprietà di $\boldsymbol{\chi}_s$, inclusa la struttura di dipendenza spaziale, possono essere formulate equivalentemente come proprietà della famiglia di campi scalari ξ_j.

Si noti infine che le Definizioni 3.6, 3.7 e 3.8 corrispondono a ipotesi di stazionarietà di tipo "globale", in quanto discendono da definizioni di variabilità globali. In particolare, l'esistenza del covariogramma (variogramma) in H è una condizione necessaria per l'esistenza dei covariogrammi puntuali; analogamente, la stazionarietà intesa come nelle definizioni precedenti è una condizione necessaria per la stazionarietà definita puntualmente (*cfr.* Esempi 3.3 e 3.4).

3.3 Lo Stimatore del Variogramma

Si consideri il processo stocastico funzionale (3.1):

$$\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}$$

a valori in H, di cui si suppone nota una realizzazione $\chi_{s_1}, ..., \chi_{s_n}$, osservata in $s_1, ..., s_n \in D$. Si assuma inoltre che il processo sia debolmente stazionario e isotropo, caratterizzato da una media m, un covariogramma $C(h)$ e un variogramma a questo associato $2\gamma(h)$.

La struttura di covarianza influenza notevolmente il comportamento del processo (3.1) (*cfr.* Sezione 1.2), per questo motivo risulta di grande importanza ottenerne una stima quanto più accurata.

La stima di tale struttura può essere condotta analizzando sia il covariogramma, che il variogramma; tuttavia, è preferita la stima di quest'ultimo poiché è ben definito in condizioni di stazionarietà meno stringenti.

La metodologia di stima del variogramma finora adottata in ambito funzionale si sviluppa in due momenti, dapprima con una stima empirica, quindi con l'adattamento di un modello valido (Giraldo *e altri*, 2008a), (Giraldo *e altri*, 2008b), (Giraldo *e altri*, 2010c), (Delicado *e altri*, 2010), in analogia con quanto illustrato nel Capitolo 1 riguardo al dato monodimensionale.

Al fine di ottenere una stima della struttura di covarianza del campo aleatorio espressa attraverso il covariogramma o il variogramma, occorre dunque definire opportuni stimatori empirici, la cui costruzione può essere ottenuta con il metodo dei momenti, come proposto, per $H = L^2$, da Giraldo *e altri* (2008a). In particolare, considerando le funzioni varianza e covarianza spaziali (Definizione 3.1), si possono definire gli stimatori campionari come segue.

Definizione 3.12. *Sia* $\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}$ *un processo stocastico stazionario e isotropo, di media* m*, covariogramma* $C(h)$ *e varianza* $\sigma^2 = C(0)$*. Si definisce* varianza spaziale campionaria*:*

$$\widehat{C}(0) = \frac{1}{n}\sum_{k=1}^{n} \|\boldsymbol{\chi}_{s_k} - \overline{\boldsymbol{\chi}}_n\|^2 \tag{3.19}$$

dove:

$$\overline{\boldsymbol{\chi}}_n = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{\chi}_{s_k}, \tag{3.20}$$

è la media campionaria. Si definisce inoltre covariogramma (empirico) campionario

$$\widehat{C}(h) = \frac{1}{|N(h)|}\sum_{(i,j)\in N(h)} \langle \boldsymbol{\chi}_{s_i} - \overline{\boldsymbol{\chi}}_n, \boldsymbol{\chi}_{s_j} - \overline{\boldsymbol{\chi}}_n \rangle, \tag{3.21}$$

dove $N(h)$ *indica l'insieme delle coppie* $\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j}$ *separate da una distanza pari a* h*:*

$$N(h) = \{(i,j) : ||s_i - s_j|| = h\}$$

e $|N(h)|$ *ne indica la cardinalità.*

Con la medesima costruzione, lo stimatore empirico del variogramma risulta l'analogo infinito-dimensionale dell'espressione (1.9), definito come segue.

Definizione 3.13. *Sia* $\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}$ *un processo stocastico stazionario e isotropo, di media* m *e variogramma* $2\gamma(h)$. *Si definisce* variogramma (empirico) campionario

$$\widehat{\gamma}(h) = \frac{1}{|2N(h)|} \sum_{(i,j)\in N(h)} \|\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}\|^2 \tag{3.22}$$

dove $N(h) = \{(i,j) : ||s_i - s_j|| = h\}$ *e* $|N(h)|$ *ne indica la cardinalità.*

Dal punto di vista teorico, non sono presenti in letteratura studi relativi alle proprietà asintotiche degli stimatori appena definiti. Tuttavia, esistono alcune relazioni tra questi ultimi e gli stimatori campionari degli operatori covarianza e cross-covarianza (Definizione 3.5) dei quali sono state recentemente indagate le proprietà di consistenza (Hörmann e Kokoszka, 2011). Tali relazioni possono essere ricavate seguendo le stesse argomentazioni della Sezione 3.2.

In particolare, si definisca l'operatore covarianza campionaria come (Hörmann e Kokoszka, 2011):

$$\widehat{C}_N^0 = \frac{1}{N} \sum_{k=1}^{N} (\boldsymbol{\chi}_{s_k} - \overline{\boldsymbol{\chi}}_N) \otimes (\boldsymbol{\chi}_{s_k} - \overline{\boldsymbol{\chi}}_N) \tag{3.23}$$

dove per ogni $\mathcal{X}, \mathcal{Y}$ variabili aleatorie a valori in H e $x \in H$ non aleatorio:

$$(\mathcal{X} \otimes \mathcal{Y})(x) = \langle \mathcal{X}, x \rangle \, \mathcal{Y}.$$

Si definisca inoltre l'operatore cross-covarianza campionaria come:

$$\widehat{C}_N^h = \frac{1}{|N(h)|} \sum_{(i,j)\in N(h)} (\boldsymbol{\chi}_{s_i} - \overline{\boldsymbol{\chi}}_N) \otimes (\boldsymbol{\chi}_{s_j} - \overline{\boldsymbol{\chi}}_N). \tag{3.24}$$

Seguendo il procedimento introdotto nel capitolo precedente, si ottiene che gli stimatori definiti dalle (3.19) e (3.21) corrispondono rispettivamente alla traccia degli operatori definiti dalle (3.23) e (3.24). Infatti, sia $\{e_j, j \in \mathbb{N}\}$ una base ortonormale di H, allora per il primo:

$$\widehat{C}(0) = \sum_{j=1}^{\infty} \langle \widehat{C}_N^0(e_j), e_j \rangle,$$

mentre per il covariogramma campionario:

$$\widehat{C}(h) = \sum_{j=1}^{\infty} \langle \widehat{C}_N^h(e_j), e_j \rangle.$$

Nel lavoro di Hörmann e Kokoszka (2011) è dimostrata la consistenza degli stimatori (3.20) e (3.23) sotto alcune ipotesi sul disegno sperimentale, che possono essere sintetizzate nella richiesta che il campionamento sia di tipo *increasing domain* o *nearly infill*, e sulle

proprietà probabilistiche del second'ordine del processo: da questi risultati discende quindi la consistenza dello stimatore varianza spaziale campionaria definito dalla (3.19). Non è invece stata studiata la consistenza dello stimatore (3.24), pertanto risultati di consistenza per lo stimatore (3.21) e, conseguentemente, per (3.22) non sono tuttora disponibili.

Dal punto di vista applicativo non è possibile determinare una stima del variogramma usando direttamente l'espressione (3.22), in quanto di norma non sono disponibili coppie di campioni per ogni distanza h: per questo motivo, in analogia al caso finito dimensionale, la stima del variogramma può essere ottenuta attraverso il *binned-variogram* e la *variogram cloud*, definibili in ambito funzionale come segue.

Definizione 3.14. *Dato un processo stocastico stazionario e isotropo* (3.1)*, con variogramma* $\gamma(h)$*, e fissati* $0 = h_1 < ... < h_{K-1}$*, si definisce stimatore* binned (semi-)variogram *il vettore aleatorio* $\widehat{\gamma}(\boldsymbol{h}) = (\widehat{\gamma}(h_1), ..., \widehat{\gamma}(h_K))$*, dove*

$$\widehat{\gamma}(h_k) = \frac{1}{2|N(h_k)|} \sum_{(i,j)\in N(h_k)} \|\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}\|^2,$$

per $N(h_k) = \{(i,j) : ||s_i - s_j|| \in [h_k - \tau, h_k + \tau], s_i, s_j \in D, i, j = 1, ..., n, i \neq j\}$*,* $\tau = \frac{1}{2}(h_k - h_{k-1})$*,* $k = 2, ..., K$*,* $N(h_1) = \{(i,j) : ||s_i - s_j|| \in [0, \tau], s_i, s_j \in D, i, j = 1, ..., n, i \neq j\}$*. La quantità* $h_k - h_{k-1}$*,* $k = 2, ..., K$*, è supposta costante ed è denominata* lag*.*

Nelle stesse ipotesi, si definisce variogram cloud *la nuvola di punti corrispondente ai valori osservati di* $\frac{1}{2}\|\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}\|^2$ *per ogni coppia di variabili* $\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j}$*,* $i, j = 1, ..., n$*,* $i \neq j$*.*

Gli stimatori empirici del variogramma introdotti dalla Definizione 3.14 consentono di indagare le proprietà di stazionarietà del processo considerato attraverso la valutazione del comportamento in prossimità dell'origine e del tipo di andamento asintotico. Una volta verificata l'assunzione di stabilità, è tuttavia necessario stimare i parametri di un modello valido di variogramma opportunamente scelto (sferico, esponenziale, etc.), al fine di ottenere un variogramma che goda delle Proprietà 3.10.

A questo scopo, è essenziale identificare un criterio di ottimo per la procedura di adattamento che tenga in considerazione le caratteristiche distribuzionali dello stimatore empirico e che sia al contempo sufficientemente generale da essere adottato in ipotesi di lavoro non troppo restrittive.

In assenza di ipotesi distribuzionali, le stime ML e REML (Cressie, 1993) non sono applicabili: un metodo che si applica in modo semplice a questo contesto è invece il metodo dei minimi quadrati ordinari, che determina il vettore di parametri $\boldsymbol{\vartheta}$ minimizzando la cifra di merito:

$$\sum_{k=1}^{K} (\widehat{\gamma}(h_k) - \gamma(h_k; \boldsymbol{\vartheta}))^2.$$

Tuttavia, il metodo dei minimi quadrati ordinari attribuisce peso uniforme a tutte le osservazioni, trascurando la struttura di covarianza delle stesse. Basando la procedura di

minimizzazione su una distanza 'statistica', si invece è in grado di quantificare correttamente la deformazione della geometria dello spazio operata dalla struttura di covarianza esistente.

Una metrica che tenga opportunamente in considerazione la struttura di dipendenza dell'oggetto considerato è la distanza di Mahalanobis (Mahalanobis, 1936), definita come segue.

Definizione 3.15 (Distanza di Mahalanobis in $\mathbb{R}^K$)**.** *Si consideri un vettore aleatorio* $\mathbb{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_K)$*, e sia* Ψ *la sua matrice di covarianza, con* $\operatorname{Det}\Psi > 0$*. La* distanza di Mahalanobis $d_{\Psi^{-1}}$ *è la distanza indotta dalla forma quadratica semidefinita positiva* Ψ^{-1}*, ovvero*

$$d^2_{\Psi^{-1}}(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x} - \boldsymbol{y})^T \Psi^{-1} (\boldsymbol{x} - \boldsymbol{y}), \quad \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^K.$$

Dal momento che la matrice di covarianza Ψ è generalmente incognita, si definisce la versione campionaria di tale distanza, $d_{\widehat{\Psi}^{-1}}$, adottando la metrica indotta dallo stimatore campionario $\widehat{\Psi}$ di Ψ.

Un'opportuna misura di adattamento nell'individuazione del criterio di ottimo per il problema in esame è dunque la distanza di Mahalanobis in $\mathbb{R}^K$, tra la stima empirica $\widehat{\gamma}(\boldsymbol{h})$ e il modello parametrico di semivariogramma $\gamma(\boldsymbol{h}; \boldsymbol{\vartheta})$, ossia il funzionale:

$$d_{\widehat{\Psi}^{-1}}(\widehat{\gamma}(\boldsymbol{h}), \gamma(\boldsymbol{h}; \boldsymbol{\vartheta})) = (\widehat{\gamma}(\boldsymbol{h}) - \gamma(\boldsymbol{h}; \boldsymbol{\vartheta}))^T \widehat{\Psi}^{-1} (\widehat{\gamma}(\boldsymbol{h}) - \gamma(\boldsymbol{h}; \boldsymbol{\vartheta})), \tag{3.25}$$

dove Ψ è la matrice di covarianza dello stimatore $\widehat{\gamma}(\boldsymbol{h})$, $\Psi_{ij} = \operatorname{Cov}(\gamma(h_i), \gamma(h_j))$, $i, j = 1, ..., K$, e $\widehat{\Psi}$ ne è una sua stima.

La possibilità di introdurre il metodo dei minimi quadrati generalizzati è tuttavia subordinata alla determinazione di una stima per Ψ. Quest'ultima sarà ottenuta nel seguito attraverso l'uso di metodi di ricampionamento, nello specifico di tecniche bootstrap, che consentiranno di valutare l'incertezza legata allo stimatore empirico in ipotesi distribuzionali lasche. A queste metodologie sarà dedicato il Capitolo 4.

3.4 I Metodi di Kriging

3.4.1 Ordinary Kriging

Si consideri un processo stocastico funzionale $\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}$, tale che $\chi_{s_i} : \mathcal{T} \to \mathbb{R}$ sia una elemento aleatorio per ogni $s_i \in D$, con $\boldsymbol{\chi}_s \in H$ di Hilbert. Siano $s_1, ..., s_n$ punti arbitrariamente scelti in D, presso i quali sia osservata una realizzazione $\chi_{s_1}, ..., \chi_{s_n}$ del processo funzionale.

Si assuma inoltre che il processo stocastico sia stazionario al second'ordine e isotropo, caratterizzato dalle seguenti funzioni di media e varianza:

$$\begin{aligned} \mathbb{E}[\boldsymbol{\chi}_s] &= m, \quad m \in H,\, s \in D \\ \operatorname{Var}(\boldsymbol{\chi}_s) &= \sigma^2, \quad \sigma^2 \in \mathbb{R},\, s \in D, \end{aligned}$$

e con covariogramma e semivariogramma:

$$\begin{aligned} \mathrm{Cov}(\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j}) &= C(h), \quad h = \|s_i - s_j\|, \forall\, s_i, s_j \in D \\ \frac{1}{2}\mathrm{Var}(\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}) &= \gamma(h), \quad h = \|s_i - s_j\|, \forall\, s_i, s_j \in D. \end{aligned}$$

Si vuole ora riformulare il previsore di kriging per dati funzionali, ovvero si vuole determinare il previsore lineare dai dati, non distorto e ottimo nel senso della minimizzazione della varianza dell'errore di previsione (stimatore BLUP).

Il previsore BLUP $\boldsymbol{\chi}^*_{s_0}$ è quindi della forma:

$$\boldsymbol{\chi}^*_{s_0} = \sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i},$$

dove i pesi $\lambda_1, ..., \lambda_n$ sono selezionati in modo da minimizzare la varianza dell'errore di previsione $\boldsymbol{\varepsilon}_{s_0} = \boldsymbol{\chi}^*_{s_0} - \boldsymbol{\chi}_{s_0}$, definita in accordo con la (3.3):

$$\mathrm{Var}(\boldsymbol{\varepsilon}_{s_0}) = \mathrm{Var}(\boldsymbol{\chi}^*_{s_0} - \boldsymbol{\chi}_{s_0}), \tag{3.26}$$

sotto il vincolo di non distorsione:

$$\mathbb{E}[\boldsymbol{\chi}^*_{s_0}] = \mathbb{E}[\boldsymbol{\chi}_{s_0}] = m. \tag{3.27}$$

Si noti in particolare che, dal momento che uno spazio di Hilbert è in particolare uno spazio vettoriale, H possiede la proprietà di chiusura rispetto alle combinazioni lineari di suoi punti: per ogni insieme $x_1, ..., x_n$ tale che $x_i \in H$, $i = 1, ..., n$, la funzione definita come $x^* = \sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i}$, è ancora un elemento di H, per ogni $\lambda_1, ..., \lambda_n \in \mathbb{R}$. Un previsore lineare è quindi una variabile aleatoria funzionale della stessa natura delle variabili aleatorie funzionali osservate, appartenendo allo spazio generato dai dati:

$$\boldsymbol{\chi}^*_{s_0} \in span\{\boldsymbol{\chi}_{s_1}, ..., \boldsymbol{\chi}_{s_n}\},$$

dove

$$span\{\boldsymbol{\chi}_{s_1}, ..., \boldsymbol{\chi}_{s_n}\} = \left\{ \boldsymbol{\chi} = \sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i},\ \lambda_i \in \mathbb{R} \right\}.$$

Per questo motivo, la quantità (3.26) risulta ben definita, così come l'uguaglianza (3.27), essendo m e $\mathbb{E}[\chi^*_{s_0}]$ nel medesimo spazio.

La condizione di non distorsione si traduce nella seguente semplice limitazione sui pesi $\lambda_1, ..., \lambda_n$

$$\sum_{i=1}^{n} \lambda_i = 1, \tag{3.28}$$

attraverso i seguenti passaggi:

$$\begin{aligned}
0 = \mathbb{E}[\boldsymbol{\varepsilon}_{s_0}] = \mathbb{E}\left[\sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_0}\right] = \\
= \sum_{i=1}^{n} \lambda_i \mathbb{E}[\boldsymbol{\chi}_{s_i}] - \mathbb{E}[\boldsymbol{\chi}_{s_0}] = \\
= \sum_{i=1}^{n} \lambda_i m - m = \\
= \left(\sum_{i=1}^{n} \lambda_i - 1\right) m.
\end{aligned}$$

I pesi ottimi di kriging possono quindi essere ottenuti risolvendo il problema di minimizzazione vincolato:

$$\min_{\lambda_1,\dots,\lambda_n} \operatorname{Var}(\boldsymbol{\chi}^*_{s_0} - \boldsymbol{\chi}_{s_0}), \quad \text{t.c} \quad \sum_{i=1}^{n} \lambda_i = 1,$$

riconducibile ad un problema di ottimizzazione non vincolato attraverso il metodo dei moltiplicatori di Lagrange.

Si noti dapprima che, grazie all'ipotesi di stazionarietà e alla condizione (3.28)

$$\begin{aligned}
\operatorname{Var}(\boldsymbol{\varepsilon}_{s_0}) = \mathbb{E}[\|\boldsymbol{\varepsilon}\|^2] - \|\mathbb{E}[\boldsymbol{\varepsilon}]\|^2 = \\
= \mathbb{E}\left[\left\|\sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_0}\right\|^2\right] - \left\|\mathbb{E}\left[\sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_0}\right]\right\|^2 = \\
= \mathbb{E}\left[\langle \sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_0}, \sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_0} \rangle\right] = \\
= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \mathbb{E}[\langle \boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j} \rangle] + \mathbb{E}[\langle \boldsymbol{\chi}_{s_0}, \boldsymbol{\chi}_{s_0} \rangle] - 2\sum_{i=1}^{n} \lambda_i \mathbb{E}[\langle \boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_0} \rangle] = \\
= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \operatorname{Cov}(\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j}) + m^2 + \operatorname{Var}(\boldsymbol{\chi}_{s_0}) + m^2 - 2\sum_{i=1}^{n} \lambda_i \operatorname{Cov}(\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_0}) - 2m^2 = \\
= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(s_i, s_j) + C(0) - 2\sum_{i=1}^{n} \lambda_i C(s_i, s_0).
\end{aligned}$$

Introducendo ora il moltiplicatore di Lagrange μ corrispondente alla condizione di non distorsione, il problema di ottimizzazione si riconduce alla determinazione dei pesi ottimi

$\lambda_1^*, ..., \lambda_n^*$ che rendano minimo il funzionale

$$\Phi = \mathrm{Var}(\boldsymbol{\varepsilon}_{s_0}) + 2\mu\left(\sum_{i=1}^{n}\lambda_i - 1\right) = \\ = \sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j C(s_i, s_j) + C(0) - 2\sum_{i=1}^{n}\lambda_i C(s_i, s_0) + 2\mu\left(\sum_{i=1}^{n}\lambda_i - 1\right).$$

Il minimo del funzionale Φ può essere individuato semplicemente imponendo l'annullamento delle derivate prime parziali, ottenendo i pesi ottimi dal seguente sistema:

$$\begin{pmatrix} C(0) & C(s_1,s_2) & \cdots & C(s_1,s_n) & 1 \\ C(s_2,s_1) & C(0) & \cdots & C(s_2,s_n) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ C(s_n,s_1) & C(s_n,s_2) & \cdots & C(0) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} C(s_0,s_1) \\ C(s_0,s_2) \\ \vdots \\ C(s_0,s_n) \\ 1 \end{pmatrix}, \tag{3.29}$$

ovvero, nella scrittura a blocchi:

$$\begin{pmatrix} C(s_i,s_j) & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \lambda_j \\ \mu \end{pmatrix} = \begin{pmatrix} C(s_i,s_0) \\ 1 \end{pmatrix}.$$

Tale sistema è identico al sistema (1.12) ed ammette la scrittura equivalente in termini di variogramma:

$$\begin{pmatrix} \gamma(0) & \gamma(s_1,s_2) & \cdots & \gamma(s_1,s_n) & 1 \\ \gamma(s_2,s_1) & \gamma(0) & \cdots & \gamma(s_2,s_n) & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma(s_n,s_1) & \gamma(s_n,s_2) & \cdots & \gamma(0) & 1 \\ 1 & 1 & \cdots & 1 & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(s_0,s_1) \\ \gamma(s_0,s_2) \\ \vdots \\ \gamma(s_0,s_n) \\ 1 \end{pmatrix}, \tag{3.30}$$

ovvero, nella scrittura a blocchi:

$$\begin{pmatrix} \gamma(s_i,s_j) & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \lambda_j \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma(s_i,s_0) \\ 1 \end{pmatrix}.$$

I sistemi (3.29) e (3.30) possiedono un'unica soluzione se e solo se la matrice

$$\begin{pmatrix} C(s_i,s_j) & 1 \\ 1 & 0 \end{pmatrix}$$

è non singolare. Questo è verificato se la sottomatrice $\Sigma = (C(s_i, s_j))$ è strettamente definita positiva, condizione che è assicurata dall'uso di una funzione di covarianza strettamente definita positiva e dall'eliminazione dei dati duplicati.

In tal caso, la stima ottenuta è caratterizzata dalla seguente varianza:

$$\sigma^2_{s_0,OK} = C(0) - \sum_{i=1}^{n}\lambda_i C(s_i,s_0) - \mu = \sum_{i=1}^{n}\lambda_i\gamma(s_i,s_0) - \mu. \tag{3.31}$$

Essa può essere considerata come una misura di incertezza globale legata alla stima di kriging in s_0 e sarà indicata nel seguito con il nome di *(Ordinary) Kriging variance*, per l'evidente analogia con il caso finito-dimensionale.

3.4.2 Universal Kriging

Si consideri ora un processo stocastico funzionale non stazionario $\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}$ a valori in uno spazio di Hilbert H. Si assuma inoltre che il processo ammetta una scrittura della forma (3.17):

$$\boldsymbol{\chi}_s = m_s + \boldsymbol{\delta}_s.$$

Come di consueto, si indichi con m_s il *drift* (deterministico) del processo, e con $\boldsymbol{\delta}_s$ il residuo debolmente stazionario e isotropo, supposto essere a media nulla, con covariogramma $C(h)$ e semivariogramma $\gamma(h)$:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\chi}_s] &= m_s, \quad \forall s \in D, \\
\mathbb{E}[\boldsymbol{\delta}_s] &= 0, \quad \forall s \in D, \\
\mathrm{Cov}(\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j}) &= \mathrm{Cov}(\boldsymbol{\delta}_{s_i}, \boldsymbol{\delta}_{s_j}) = C(\|s_i - s_j\|), \quad \forall\, s_i, s_j \in D, \\
\mathrm{Var}(\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}) &= \mathrm{Var}(\boldsymbol{\delta}_{s_i} - \boldsymbol{\delta}_{s_j}) = 2\gamma(\|s_i - s_j\|), \quad \forall\, s_i, s_j \in D.
\end{aligned}$$

Il primo passo necessario per l'estensione al caso funzionale dell'Universal Kriging è l'identificazione di una forma funzionale adeguata per il termine di *drift* m_s. In analogia alla teoria dei modelli lineari funzionali, illustrati nel Capitolo 2, si possono infatti considerare almeno quattro tipi differenti di *drift* (Ramsay e Silverman, 2005):

(i) *Modello identicamente costante in t, spazialmente variabile attraverso funzioni f identicamente costanti in t*:

$$m_s(t) = \sum_{l=0}^{L} a_l f_l(s), \quad s \in D,\, t \in \mathcal{T}, \tag{3.32}$$

dove $a_l \in \mathbb{R}$, $f_0(s) = 1$ per ogni $s \in D$.

(ii) *Modello funzionale in t, spazialmente variabile attraverso funzioni f identicamente costanti in t*:

$$m_s(t) = \sum_{l=0}^{L} a_l(t) f_l(s), \quad s \in D,\, t \in \mathcal{T}, \tag{3.33}$$

dove a_l è una funzione nello spazio di Hilbert H, $f_0(s) = 1$ per ogni $s \in D$.

(iii) *Modello funzionale in t, spazialmente variabile attraverso funzioni f non costanti in t, di tipo 'cuncurrent'*:

$$m_s(t) = \sum_{l=0}^{L} a_l(t) f_l(t, s), \quad s \in D,\, t \in \mathcal{T}, \tag{3.34}$$

dove a_l è una funzione nello spazio di Hilbert H, $f_0(t, s) = 1$ per ogni $s \in D$, $t \in \mathcal{T}$.

(iv) *Modello funzionale in t, spazialmente variabile attraverso funzioni f non costanti in t, di tipo 'totale'*:

$$m_s(t) = \sum_{l=0}^{L} \langle a_l(t,\cdot), f_l(\cdot,s)\rangle_H, \quad s \in D,\, t \in \mathcal{T}, \tag{3.35}$$

dove $f(\cdot,s) \in H$ per ogni $s \in D$, $f_0(t,s) = 1$ per ogni $s \in D$, $t \in \mathcal{T}$ e $a_l(\cdot,\cdot)$ è una funzione tale che $a_l(t,\cdot) \in H$ per ogni $t \in \mathcal{T}$, $\{\langle a_l(t,\cdot), f_l(\cdot,s)\rangle_H, t \in \mathcal{T}\} \in H$ per ogni $s \in D$.

Il primo modello di *drift* prevede che la media m_s si mantenga costante lungo l'ascissa del dato funzionale, ipotesi che nelle applicazioni si rivela molto spesso troppo restrittiva. I modelli (ii)-(iii)-(iv) consentono invece che la media del processo dipenda dalla variabile t attraverso dei coefficienti funzionali a_l, con complessità via via crescente.

Inoltre, la differenza fondamentale tra i modelli (i)-(ii) e i modelli (iii)-(iv) è il fatto che nel terzo e quarto caso si consente che la dipendenza della funzione m_s dalla variabile dipendente t avvenga anche attraverso le funzioni f_l, supposte note a priori: questo comporta una maggiore flessibilità nella modellazione del *drift*, accompagnata dalla crescita della difficoltà tecnica associata alla stima. Si noti che in tutti i casi la variazione spaziale della media si assume che sia spiegata interamente dalle funzioni f_l che, a differenza dei coefficienti a_l, sono le uniche a rientrare nel sistema di Universal Kriging.

Nel seguito del lavoro di tesi si focalizzerà l'attenzione sui modelli a f_l identicamente costanti in t e, in particolare, sul modello (ii): quest'ultimo consente infatti di estendere in modo elegante il sistema di kriging (1.13), inserendosi in modo naturale nell'ambiente definito nelle Sezioni 3.1 e 3.2. Si sottolinea inoltre che, sebbene i modelli (iii)-(iv) siano potenzialmente molto interessanti in presenza di un *external drift* funzionale, il modello (ii) consente una stima accurata del *drift* in tutti i casi di *external drift* scalare e, in generale, di *drift* separabile.

Si assuma dunque che il *drift* m_s sia della forma (3.33) (o della forma (3.32)) e si consideri la famiglia di previsori lineari della variabile $\boldsymbol{\chi}_{s_0}$:

$$\boldsymbol{\chi}^*_{s_0} = \sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i}.$$

Per ottenere il previsore ottimo di kriging si procederà imponendo dapprima la condizione di non distorsione, quindi la minimalità rispetto alla varianza dell'errore di previsione risolvendo un problema di ottimizzazione vincolata.

La prima condizione è tradotta in questo caso nella seguente:

$$\sum_{i=1}^{n} \lambda_i f_l(s_i) = f_l(s_0), \quad \forall\, l = 0, ..., L; \tag{3.36}$$

infatti, indicando con $\boldsymbol{\varepsilon}_{s_0}$ l'errore commesso nella previsione di $\boldsymbol{\chi}_{s_0}$ con il previsore $\boldsymbol{\chi}^*_{s_0}$:

$$\boldsymbol{\varepsilon}_{s_0} = \boldsymbol{\chi}^*_{s_0} - \boldsymbol{\chi}_{s_0},$$

si ricava:

$$
\begin{aligned}
0 = \mathbb{E}[\boldsymbol{\varepsilon}_{s_0}] = \mathbb{E}\left[\sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_0}\right] = \\
= \sum_{i=1}^{n} \lambda_i \mathbb{E}[\boldsymbol{\chi}_{s_i}] - \mathbb{E}[\boldsymbol{\chi}_{s_0}] = \\
= \sum_{i=1}^{n} \lambda_i \mathbb{E}[m_{s_i} + \boldsymbol{\delta}_{s_i}] - \mathbb{E}[m_{s_0} + \boldsymbol{\delta}_{s_0}] = \\
= \sum_{i=1}^{n} \lambda_i \sum_{l=0}^{L} a_l f_l(s_i) - \sum_{l=0}^{L} a_l f_l(s_0) = \\
= \sum_{l=0}^{L} a_l \left(\sum_{i=1}^{n} \lambda_i f_l(s_i) - f_l(s_0)\right).
\end{aligned}
$$

Si osservi ora che:

$$
\begin{aligned}
\mathrm{Var}(\boldsymbol{\varepsilon}_{s_0}) = \mathrm{Var}(\boldsymbol{\chi}^*_{s_0} - \boldsymbol{\chi}_{s_0}) = \\
= \mathrm{Var}\left(\sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_0}\right) = \\
= \mathrm{Var}\left(\sum_{i=1}^{n} \lambda_i (m_{s_i} + \boldsymbol{\delta}_{s_i}) - (m_{s_0} + \boldsymbol{\delta}_{s_0})\right) = \\
= \mathrm{Var}\left(\sum_{i=1}^{n} \lambda_i \boldsymbol{\delta}_{s_i} - \boldsymbol{\delta}_{s_0}\right) = \\
= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(s_i, s_j) + C(0) - 2\sum_{i=1}^{n} \lambda_i C(s_i, s_0).
\end{aligned}
$$

Quindi, introducendo $L+1$ moltiplicatori di Lagrange, corrispondenti alle $L+1$ condizioni (3.36), il problema di ottimizzazione vincolato si riconduce alla minimizzazione del funzionale:

$$
\begin{aligned}
\Phi = \mathrm{Var}(\boldsymbol{\varepsilon}_{s_0}) + 2\sum_{l=0}^{L} \mu_l \left(\sum_{i=1}^{n} \lambda_i f_l(s_i) - f_l(s_0)\right) = \\
= \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j C(s_i, s_j) + C(0) - 2\sum_{i=1}^{n} \lambda_i C(s_i, s_0) + 2\sum_{l=0}^{L} \mu_l \left(\sum_{i=1}^{n} \lambda_i f_l(s_i) - f_l(s_0)\right).
\end{aligned}
$$

L'ottimo globale può essere infine individuato annullando le derivate prime parziali in λ_i, $i = 1, ..., n$ e μ_l, $l = 0, ..., L$, ovvero risolvendo il seguente sistema lineare:

$$\begin{pmatrix} C(0) & \cdots & C(s_1,s_n) & 1 & f_1(s_1) & \cdots & f_L(s_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ C(s_n,s_1) & \cdots & C(0) & 1 & f_1(s_n) & \cdots & f_L(s_n) \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ f_1(s_1) & \cdots & f_1(s_n) & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_L(s_1) & \cdots & f_L(s_n) & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu_0 \\ \mu_1 \\ \vdots \\ \mu_L \end{pmatrix} = \begin{pmatrix} C(s_0,s_1) \\ \vdots \\ C(s_0,s_n) \\ 1 \\ f_1(s_0) \\ \vdots \\ f_L(s_0) \end{pmatrix}, \tag{3.37}$$

che nella scrittura a blocchi risulta:

$$\begin{pmatrix} C(s_i,s_j) & 1 & f_l(s_i) \\ 1 & 0 & 0 \\ f_l(s_j) & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_j \\ \mu_0 \\ \mu_l \end{pmatrix} = \begin{pmatrix} C(s_i,s_0) \\ 1 \\ f_l(s_0) \end{pmatrix}.$$

Sfruttando la relazione (1.4), è possibile esprimere il sistema (3.37) in termini di variogramma come segue:

$$\begin{pmatrix} \gamma(0) & \cdots & \gamma(s_1,s_n) & 1 & f_1(s_1) & \cdots & f_l(s_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \gamma(s_n,s_1) & \cdots & \gamma(0) & 1 & f_1(s_n) & \cdots & f_l(s_n) \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ f_1(s_1) & \cdots & f_1(s_n) & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ f_L(s_1) & \cdots & f_L(s_n) & 0 & 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu_0 \\ \mu_1 \\ \vdots \\ \mu_L \end{pmatrix} = \begin{pmatrix} \gamma(s_0,s_1) \\ \vdots \\ \gamma(s_0,s_n) \\ 1 \\ f_1(s_0) \\ \vdots \\ f_L(s_0) \end{pmatrix}, \tag{3.38}$$

ovvero, nella scrittura a blocchi,

$$\begin{pmatrix} \gamma(s_i,s_j) & 1 & f_l(s_i) \\ 1 & 0 & 0 \\ f_l(s_j) & 0 & 0 \end{pmatrix} \begin{pmatrix} \lambda_j \\ \mu_0 \\ \mu_l \end{pmatrix} = \begin{pmatrix} \gamma(s_i,s_0) \\ 1 \\ f_l(s_0) \end{pmatrix}.$$

La soluzione dei sistemi (3.37) e (3.38) esiste ed è unica se e solo se la matrice

$$\begin{pmatrix} C(s_i,s_j) & 1 & f_l(s_i) \\ 1 & 0 & 0 \\ f_l(s_j) & 0 & 0 \end{pmatrix}$$

è non singolare. Questo è verificato se la sottomatrice $\Sigma = (C(s_i,s_j))$ è strettamente definita positiva e la sottomatrice $\mathbb{F}_{\boldsymbol{s}}$, definita come

$$\mathbb{F}_{\boldsymbol{s}} = \begin{pmatrix} 1 & f_1(s_1) & \cdots & f_L(s_1) \\ 1 & f_1(s_2) & \cdots & f_L(s_2) \\ \vdots & \vdots & \cdots & \vdots \\ 1 & f_1(s_n) & \cdots & f_L(s_n) \end{pmatrix}$$

è di rango pari a $L+1$. La prima condizione è assicurata dall'uso di una funzione di covarianza strettamente definita positiva e dall'eliminazione dei dati duplicati, la seconda è invece una condizione standard sul disegno sperimentale, che è incontrata ad esempio nella teoria della stima ai minimi quadrati nella richiesta che $\mathbb{F}_{\boldsymbol{s}}^T\mathbb{F}_{\boldsymbol{s}}$ sia non singolare.

In tali ipotesi, l'*(Universal) Kriging variance* risulta:

$$\sigma^2_{s_0,UK} = C(0) - \sum_{i=1}^{n} \lambda_i C(s_i, s_0) - \mu = \sum_{i=1}^{n} \lambda_i \gamma(s_i, s_0) - \mu, \tag{3.39}$$

che corrisponde all'incertezza legata alla previsione di kriging in s_0.

3.5 Lo Stimatore del Drift

Si è visto nella sezione precedente che, qualora il processo stocastico (3.1) sia caratterizzato da un *drift* deterministico della forma (3.33), l'insieme di pesi $\lambda_1, ..., \lambda_n$ e i moltiplicatori di Lagrange $\mu_0, ..., \mu_L$ relativi alla condizione di non distorsione (3.36) sono ottenibili risolvendo un sistema di $L + n + 1$ equazioni in $L + n + 1$ incognite che, con assunzioni non troppo restrittive, possiede un'unica soluzione.

In tale sistema non compaiono i coefficienti funzionali a_l, ma solo le funzioni note $f_l(s)$, tuttavia questo non deve indurre a pensare che la stima di tali termini non sia necessaria. Infatti, lo stimatore sperimentale (3.22) fornisce una stima distorta del variogramma $\gamma(h)$ qualora sia calcolato a partire dalle osservazioni di un processo di media non costante. Dal momento che la variabilità spaziale deterministica del termine di *drift* risulta spesso preponderante sulla variabilità stocastica del residuo stazionario, che invece determina le caratteristiche probabilistiche del second'ordine del processo e, in particolare, il variogramma, tale distorsione è nella pratica molto influente, conducendo a una stima empirica non in linea con le caratteristiche attese da un variogramma stazionario, risultando quindi non fruibile nelle applicazioni (ad esempio per la manifestazione di un andamento superquadratico o per l'assenza dell'asintoto orizzontale, ovvero del *sill*). Il calcolo del residuo, che è dunque essenziale per la stima del variogramma, necessita tuttavia della stima preliminare del termine di *drift*, ovvero dei coefficienti a_l.

Infatti, a partire dal dataset funzionale $\chi_{s_1}, ..., \chi_{s_n}$ e considerando le espressioni (3.17) e (3.33):

$$\begin{aligned} \boldsymbol{\chi}_s &= m_s + \boldsymbol{\delta}_s, \quad s \in D, \\ m_s &= \sum_{l=0}^{L} a_l f_l(s), \quad s \in D, \end{aligned}$$

il modello per il fenomeno è della forma:

$$\boldsymbol{\chi}_s = \sum_{l=0}^{L} a_l f_l(s) + \boldsymbol{\delta}_s, \quad s \in D. \tag{3.40}$$

La procedura di stima del *drift* si inserisce dunque nella teoria dei modelli lineari: le funzioni $f_l(s)$ possono essere pensate come regressori (scalari nel caso in esame, funzionali nei modelli di *drift* (3.34), (3.35)) aventi come risposta funzionale $\boldsymbol{\chi}_s$.

A differenza dei modelli lineari trattati nell'ambito dell'analisi di dati funzionali (*cfr.* Capitolo 2, Sezione 2.4), in questo caso i residui $\boldsymbol{\delta}_{s_1}, ..., \boldsymbol{\delta}_{s_n}$ non sono indipendenti e identicamente distribuiti: al contrario, essi sono caratterizzati da una struttura di covarianza spaziale, molto rilevante ai fini dell'analisi geostatistica, espressa attraverso la matrice $\Sigma = (C(\|s_i - s_j\|))$.

Inoltre, nell'ambito geostatistico, il residuo del modello lineare che descrive il *drift*, $\boldsymbol{\delta}_s$, riveste un ruolo particolarmente importante, dal momento che determina le caratteristiche distribuzionali del processo $\boldsymbol{\chi}_s$: il residuo $\boldsymbol{\delta}_s$ rappresenta quindi una componente di variabilità molto interessante, la cui stima è essenziale ai fini modellistici e previsivi.

In assenza di ipotesi distribuzionali, il metodo di stima preferito per un modello lineare di forma 3.40 è generalmente il metodo dei minimi quadrati, che, in ambito funzionale, si basa sulla minimizzazione della varianza dell'errore di stima, nel senso che sarà ora indicato.

Formalmente, il dataset funzionale $\chi_{s_1}, ..., \chi_{s_n}$ è la realizzazione del vettore aleatorio $(\boldsymbol{\chi}_{s_1}, ..., \boldsymbol{\chi}_{s_n})$, e, come tale, assume valori in H^n, se la generica variabile $\boldsymbol{\chi}_{s_i}$, $i = 1, ..., n$ è a valori in H (di Hilbert). Per inquadrare il problema di stima, si doti lo spazio H^n del prodotto scalare:

$$\langle \mathcal{X}, \mathcal{Y} \rangle_{H^n} = \sum_{i=1}^{n} \langle \mathcal{X}_i, \mathcal{Y}_i \rangle_H, \tag{3.41}$$

e della norma indotta:

$$\|\mathcal{X}\|^2_{H^n} = \sum_{i=1}^{n} \|\mathcal{X}_i\|^2_H, \tag{3.42}$$

con $\mathcal{X} = (\mathcal{X}_1, ..., \mathcal{X}_n)$, $\mathcal{Y} = (\mathcal{Y}_1, ..., \mathcal{Y}_n)$, $\mathcal{X}_i, \mathcal{Y}_i \in H$, per ogni $i = 1, ..., n$.

Il metodo dei minimi quadrati ordinari (OLS) prevede che la stima $\widehat{m}_s$ del *drift* m_s sia determinata attraverso la minimizzazione del funzionale:

$$\mathbb{E}[\|\boldsymbol{\chi}_s - \widehat{\boldsymbol{m}}_s\|^2_H].$$

Tuttavia, avendo a disposizione un dataset funzionale di dimensione n, la quantità precedente è incognita ed è dunque sostituita dalla sua versione campionaria, minimizzando infine:

$$\sum_{i=1}^{n} \|\boldsymbol{\chi}_{s_i} - \widehat{\boldsymbol{m}}_{s_i}\|^2_H = \|\boldsymbol{\chi_s} - \widehat{\boldsymbol{m}}_{\boldsymbol{s}}\|^2_{H^n},$$

dove $\boldsymbol{\chi_s} = (\boldsymbol{\chi}_{s_1}, ..., \boldsymbol{\chi}_{s_n})$, $\widehat{\boldsymbol{m}}_{\boldsymbol{s}} = (\widehat{\boldsymbol{m}}_{s_1}, ..., \widehat{\boldsymbol{m}}_{s_n})$ ed è tralasciata la costante di proporzionalità $1/n$.

Una stima basata sui minimi quadrati ordinari (OLS), si rivela tuttavia inadeguata nel caso in cui i residui del modello non siano indipendenti. Infatti, affinchè nella procedura di minimizzazione sia opportunamente considerata la geometria dello spazio campionario, la distanza corretta da considerare nel criterio di ottimalità è la distanza di Mahalanobis $d_{\Sigma^{-1}}$

(Definizione 3.15), la cui definizione può essere estesa in ambito infinito-dimensionale come segue.

Definizione 3.16 (Distanza di Mahalanobis in H)**.** *Sia* $\boldsymbol{\mathcal{X}} = (\boldsymbol{\mathcal{X}}_1, ..., \boldsymbol{\mathcal{X}}_n)$ *un vettore aleatorio di* H^n *(ossia un vettore di elementi aleatori di* H*), dotato di prodotto interno* (3.41) *e norma* (3.42)*. Sia* Σ *la matrice di covarianza di* $\boldsymbol{\mathcal{X}}$*,* $\Sigma_{ij} = \mathrm{Cov}(\boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{X}}_j)$ *(Def. 3.1), e si supponga che esista la sua inversa* Σ^{-1}*. La* distanza di Mahalanobis $d_{\Sigma^{-1}}$ *è la distanza indotta dalla forma quadratica definita positiva* Σ^{-1}*, nel senso:*

$$d_{\Sigma^{-1}}(\mathcal{X}, \mathcal{Y}) = \|\mathcal{X} - \mathcal{Y}\|_{\Sigma^{-1} - H^n} = \|\Sigma^{-1/2}(\mathcal{X} - \mathcal{Y})\|_{H^n}, \quad \forall\, \mathcal{X}, \mathcal{Y} \in H^n,$$

dove $\Sigma = \Sigma^{1/2}\Sigma^{T/2}$.

In generale, in ambito finito-dimensionale, la metrica indotta dalla distanza di Mahalanobis su $\mathbb{R}^n$ è da preferire alla metrica euclidea per la procedura di minimizzazione, poiché $d_{\Sigma^{-1}}$ rappresenta la distanza statistica tra i punti di $\mathbb{R}^n$, corrispondente alla deformazione della metrica euclidea apportata dalla struttura di covarianza del campione, quantificata attraverso la matrice Σ.

Per il medesimo motivo, una corretta procedura di stima dei parametri di un modello lineare funzionale di tipo (3.40), ovvero caratterizzato da residui spazialmente correlati, non deve ignorare tale struttura di covarianza. È quindi importante dal punto di vista metodologico adottare un approccio ai minimi quadrati generalizzati (GLS), minimizzando la distanza di Mahalanobis tra il vettore di valori osservati $\boldsymbol{\chi_s}$ e il vettore di valori predetti dal modello $\widehat{\boldsymbol{m}}_{\boldsymbol{s}}$, corrispondente al funzionale:

$$\|\boldsymbol{\chi_s} - \widehat{\boldsymbol{m}}_{\boldsymbol{s}}\|^2_{\Sigma^{-1} - H^n}. \tag{3.43}$$

Al fine di ottimizzare il funzionale (3.43) è interessante notare che, adottando la notazione vettoriale:

$$(\mathbb{A}g)_i = \sum_{j=1}^{n} \mathbb{A}_{ij} g_j, \quad \mathbb{A} = (\mathbb{A}_{ij}) \in \mathbb{R}^{n,n}, g \in H^n,$$

risulta:

$$\begin{aligned} \|\boldsymbol{\chi_s} - \widehat{\boldsymbol{m}}_{\boldsymbol{s}}\|_{\Sigma^{-1} - H^n} &= \|\Sigma^{-1/2}(\boldsymbol{\chi_s} - \widehat{\boldsymbol{m}}_{\boldsymbol{s}})\|_{H^n} = \\ &= \|\Sigma^{-1/2}\boldsymbol{\chi_s} - \Sigma^{-1/2}\mathbb{F}_{\boldsymbol{s}} a_{\boldsymbol{l}}\|_{H^n} = \\ &= \|\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} - \widetilde{\mathbb{F}}_{\boldsymbol{s}} a_{\boldsymbol{l}}\|_{H^n}, \end{aligned}$$

dove:

$$\begin{aligned} \mathbb{F}_{\boldsymbol{s}} &= \begin{pmatrix} 1 & f_1(s_1) & \cdots & f_L(s_1) \\ 1 & f_1(s_2) & \cdots & f_L(s_2) \\ \vdots & \vdots & \cdots & \vdots \\ 1 & f_1(s_n) & \cdots & f_L(s_n) \end{pmatrix} \\ a_{\boldsymbol{l}} &= \begin{pmatrix} 1 & a_1 & \cdots & a_L \end{pmatrix}^T \\ \widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} &= \Sigma^{-1/2}\boldsymbol{\chi_s} \\ \widetilde{\mathbb{F}}_{\boldsymbol{s}} &= \Sigma^{-1/2}\mathbb{F}_s. \end{aligned}$$

Si noti in particolare che il vettore $\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}}$ risulta scorrelato, infatti:

$$\Sigma := \mathrm{Cov}(\boldsymbol{\chi}_{\boldsymbol{s}}) = \mathrm{Cov}(\Sigma^{1/2}\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}}) = \Sigma^{1/2}\,\mathrm{Cov}(\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}})\Sigma^{T/2}$$

che è pari a $\Sigma^{1/2}\Sigma^{T/2}$ se e solo se $\mathrm{Cov}(\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}})$ è la matrice identità in $\mathbb{R}^n$. Pertanto, il metodo di ottimizzazione dei minimi quadrati generalizzati non è altro che il metodo dei minimi quadrati ordinari applicato a un dataset scorrelato, ottenuto dai dati originali attraverso un'opportuna trasformazione lineare.

Si assuma nota la matrice di covarianza Σ del campione e si denoti quindi con $\widetilde{V}$ il sottospazio (chiuso) di H^n, generato dalle composizioni lineari a coefficienti in H delle colonne di $\widetilde{\mathbb{F}}_{\boldsymbol{s}}$, e con $\widetilde{V}^{\perp}$ il suo complemento ortogonale:

$$\widetilde{V} = \{\widetilde{\boldsymbol{v}} \in H^n : \widetilde{\boldsymbol{v}} = \widetilde{\mathbb{F}}_{\boldsymbol{s}}\widetilde{a}_{\boldsymbol{l}}, \quad \widetilde{a}_{\boldsymbol{l}} \in H^L\}, \tag{3.44}$$

$$\widetilde{V}^{\perp} = \{\widetilde{\boldsymbol{w}} \in H^n : \langle\widetilde{\boldsymbol{w}}, \widetilde{\boldsymbol{v}}\rangle_{H^n} = 0, \quad \forall\widetilde{\boldsymbol{v}} \in \widetilde{V}\}, \tag{3.45}$$

allora lo stimatore ottimo $\widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}}$ del *drift* $\widetilde{m}_{\boldsymbol{s}}$ del campione scorrelato (i.e. la sua media), che individua il minimo del funzionale (3.43), esiste, è unico e corrisponde alla proiezione del campione scorrelato $\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}}$ su $\widetilde{V}$, mentre il vettore di residui $\widetilde{\boldsymbol{\delta}}_{\boldsymbol{s}} = \widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} - \widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}}$ ne è la proiezione sul complemento ortogonale $\widetilde{V}^{\perp}$ (Ramsay e Dalzell, 1991).

Inoltre, grazie alla matrice $\widetilde{\mathbb{H}}$ di proiezione ortogonale su $\widetilde{V}$:

$$\widetilde{\mathbb{H}} = \widetilde{\mathbb{F}}_{\boldsymbol{s}}(\widetilde{\mathbb{F}}_{\boldsymbol{s}}^T\widetilde{\mathbb{F}}_{\boldsymbol{s}})^{-1}\widetilde{\mathbb{F}}_{\boldsymbol{s}}^T,$$

gli stimatori ottimi ai minimi quadrati generalizzati $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}$ e $\widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}}$, rispettivamente del vettore di coefficienti $a_{\boldsymbol{l}}$ e del *drift* $\widetilde{m}_{\boldsymbol{s}}$, risultano lineari e per essi è possibile determinare un'espressione esplicita. Infatti, tali stimatori possono essere ottenuti come:

$$\begin{aligned} \widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS} &= (\widetilde{\mathbb{F}}_{\boldsymbol{s}}^T\widetilde{\mathbb{F}}_{\boldsymbol{s}})^{-1}\widetilde{\mathbb{F}}_{\boldsymbol{s}}^T\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}}; \\ \widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}} &= \widetilde{\mathbb{H}}\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} = \widetilde{\mathbb{F}}_{\boldsymbol{s}}(\widetilde{\mathbb{F}}_{\boldsymbol{s}}^T\widetilde{\mathbb{F}}_{\boldsymbol{s}})^{-1}\widetilde{\mathbb{F}}_{\boldsymbol{s}}^T\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}}. \end{aligned} \tag{3.46}$$

Applicando la trasformazione inversa, lo stimatore ottimo del *drift* relativo al processo $\chi_{\boldsymbol{s}}$ può essere ottenuto come:

$$\widehat{\boldsymbol{m}}_{\boldsymbol{s}} = \mathbb{H}\boldsymbol{\chi}_{\boldsymbol{s}} = \mathbb{F}_{\boldsymbol{s}}(\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}}\Sigma^{-1}\boldsymbol{\chi}_{\boldsymbol{s}}, \tag{3.47}$$

dove:

$$\mathbb{H} := \mathbb{F}_{\boldsymbol{s}}(\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}}\Sigma^{-1},$$

mentre lo stimatore $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}$ può essere riscritto come:

$$\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS} = (\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}}\Sigma^{-1}\boldsymbol{\chi}_{\boldsymbol{s}}. \tag{3.48}$$

Con argomenti di linearità, per gli stimatori precedenti possono essere calcolate le funzioni di media e di covarianza spaziale, secondo la Definizione 3.1.

In particolare, per quanto concerne il valor medio e la matrice di covarianza Λ dello stimatore $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}$, $\Lambda = (\mathrm{Cov}(\widehat{\boldsymbol{a}}_i^{GLS}, \widehat{\boldsymbol{a}}_j^{GLS}))$, si ricava:

$$\mathbb{E}[\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}] = a_{\boldsymbol{l}} \tag{3.49}$$

$$\Lambda = (\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1}. \tag{3.50}$$

Infatti:

$$\begin{aligned}\mathbb{E}[\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}] &= \mathbb{E}[(\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} \mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \boldsymbol{\chi}_{\boldsymbol{s}}] = \\ &= (\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} \mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{E}[\boldsymbol{\chi}_{\boldsymbol{s}}] = \\ &= (\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} \mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}} \boldsymbol{a}_{\boldsymbol{l}} = a_{\boldsymbol{l}};\end{aligned}$$

e

$$\begin{aligned}\Lambda := \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}) &= \mathrm{Cov}((\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} \mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \boldsymbol{\chi}_{\boldsymbol{s}}) = \\ &= (\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} \mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \,\mathrm{Cov}(\boldsymbol{\chi}_{\boldsymbol{s}}) \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}} (\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} = \\ &= (\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} \mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \Sigma \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}} (\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} = \\ &= (\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1}.\end{aligned}$$

Relativamente al valor medio e all'incertezza globale legata alla stima del *drift*, si ricava invece:

$$\mathbb{E}[\widehat{\boldsymbol{\chi}}_{\boldsymbol{s}}] = \mathbb{E}[\mathbb{F}_{\boldsymbol{s}} \widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}] = \mathbb{F}_{\boldsymbol{s}} a_{\boldsymbol{l}} = m_{\boldsymbol{s}} \tag{3.51}$$

$$\mathrm{Cov}(\widehat{\boldsymbol{m}}_{\boldsymbol{s}}) = \mathrm{Cov}(\mathbb{F}_{\boldsymbol{s}} \widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}) = \tag{3.52}$$

$$= \mathbb{F}_{\boldsymbol{s}}^T \Lambda \mathbb{F}_{\boldsymbol{s}} = \mathbb{F}_{\boldsymbol{s}}^T (\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} \mathbb{F}_{\boldsymbol{s}}. \tag{3.53}$$

Inoltre, gli stimatori $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}$ e $\widehat{\boldsymbol{\chi}}_{\boldsymbol{s}}$ sono BLUE (*Best Linear Unbiased Estimators*): la non distorsione è verificata per le uguaglianze (3.49) e (3.51), mentre l'ottimalità segue dai medesimi procedimenti reperibili nell'ambito della teoria della stima finito-dimensionale (Spagnolini, 2010).

Infatti, sia $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}$ un generico stimatore lineare dai dati dei coefficienti $a_{\boldsymbol{l}}$ del *drift*:

$$\widehat{\boldsymbol{a}}_{\boldsymbol{l}} = \mathbb{A} \boldsymbol{\chi}_{\boldsymbol{s}} + \boldsymbol{b}, \tag{3.54}$$

con $\mathbb{A} \in \mathbb{R}^{l,n}$, $\boldsymbol{b} \in H^L$, allora la condizione di non distorsione si traduce nei vincoli:

$$\mathbb{A} \mathbb{F}_{\boldsymbol{s}} = \mathbb{I}_n \tag{3.55}$$

$$\boldsymbol{b} = \boldsymbol{0}, \quad q.o. \tag{3.56}$$

indicando con $\mathbb{I}_n$ la matrice identità in $\mathbb{R}^n$, con $\boldsymbol{0} \in H^L$ il vettore di L funzioni identicamente nulle. L'ottimalità dello stimatore $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{BLUE}$ si traduce nella richiesta che per ogni stimatore lineare $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}$, la matrice:

$$\mathrm{Cov}(\widehat{\boldsymbol{a}}_{\boldsymbol{l}}) - \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{BLUE}),$$

risulti semidefinita positiva, che è equivalente alla condizione:

$$\boldsymbol{x}^T(\mathrm{Cov}(\widehat{\boldsymbol{a}}_{\boldsymbol{l}}) - \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{BLUE}))\boldsymbol{x} \geq 0, \quad \forall \boldsymbol{x} \in \mathbb{R}^n.$$

La matrice di covarianza del generico stimatore lineare (3.54), sotto il vincoli di non distorsione (3.56), risulta:

$$\mathrm{Cov}(\widehat{\boldsymbol{a}}_{\boldsymbol{l}}) = \mathbb{A}\Sigma\mathbb{A}^T;$$

e, grazie alla disuguaglianza (Shumway e Dean, 1968):

$$\boldsymbol{\alpha}\mathbb{C}^{-1}\boldsymbol{\alpha}^T \geq \boldsymbol{\alpha}\mathbb{D}(\mathbb{D}^T\mathbb{C}\mathbb{D})^{-1}\mathbb{D}^T\boldsymbol{\alpha}^T,$$

valida per $\mathbb{C} \in \mathbb{R}^{n,n}$ semidefinita positiva, $\mathbb{D} \in \mathbb{R}^{n,L}$ e $\boldsymbol{\alpha} \in \mathbb{R}^n$, si può ottenere un limite inferiore per $\boldsymbol{x}^T\,\mathrm{Cov}(\widehat{\boldsymbol{a}}_{\boldsymbol{l}})\boldsymbol{x}$, ponendo $\mathbb{C} = \Sigma^{-1}$, $\alpha = \boldsymbol{x}^T\mathbb{A}$ e $\mathbb{D} = \mathbb{F}_{\boldsymbol{s}}$:

$$\boldsymbol{x}^T\mathbb{A}\Sigma\mathbb{A}^T\boldsymbol{x} \geq \boldsymbol{x}^T\mathbb{A}\mathbb{F}_{\boldsymbol{s}}(\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}}^T\mathbb{A}^T\boldsymbol{x}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n$$

che è raggiunto per:

$$\mathbb{A}^{BLUE} = (\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}.$$

Lo stimatore lineare ottimo risulta quindi:

$$\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{BLUE} = (\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\boldsymbol{\chi}_{\boldsymbol{s}} = \widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS},$$

che in particolare minimizza gli scarti quadratici medi MSE_l simultaneamente per ogni $l = 1, ..., L$:

$$MSE_l = \mathbb{E}[\|\widehat{\boldsymbol{a}}_l - \boldsymbol{a}_l\|^2] = (\mathbb{A}\Sigma\mathbb{A}^T)_{ll},$$

sotto i vincoli di non distorsione (3.55) e (3.56). Come conseguenza, per linearità, lo stimatore $\widehat{\boldsymbol{m}}_{\boldsymbol{s}}$ è uno stimatore BLUE del *drift*.

Grazie all'ortogonalità tra gli stimatori del *drift* e del residuo e alla struttura geometrica dello spazio di Hilbert H, è inoltre possibile fornire una decomposizione della matrice di covarianza Σ di $\boldsymbol{\chi}_{\boldsymbol{s}}$, che segue dall'identità:

$$\|\boldsymbol{\chi}_{\boldsymbol{s}}\|^2_{\Sigma^{-1}-H} = \|\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}}\|^2_H$$

e dalle scomposizioni equivalenti:

$$\begin{aligned} \|\boldsymbol{\chi}_{\boldsymbol{s}}\|^2_{\Sigma^{-1}-H^n} &= \|\widehat{\boldsymbol{m}}_{\boldsymbol{s}}\|^2_{\Sigma^{-1}-H^n} + \|\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}\|^2_{\Sigma^{-1}-H^n} \\ \|\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}}\|^2_{H^n} &= \|\widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}}\|^2_{H^n} + \|\widehat{\widetilde{\boldsymbol{\delta}}}_{\boldsymbol{s}}\|^2_{H^n}, \end{aligned}$$

dove $\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}$ è il residuo del modello stimato con i minimi quadrati generalizzati (o, equivalentemente, lo stimatore GLS del residuo) e $\widehat{\widetilde{\boldsymbol{\delta}}}_{\boldsymbol{s}} = \Sigma^{-1/2}\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}$ è il residuo del modello stimato con i minimi quadrati ordinari dai dati scorrelati:

$$\begin{aligned} \widehat{\boldsymbol{\delta}}_{\boldsymbol{s}} &= \boldsymbol{\chi}_{\boldsymbol{s}} - \widehat{\boldsymbol{m}}_{\boldsymbol{s}} \\ \widehat{\widetilde{\boldsymbol{\delta}}}_{\boldsymbol{s}} &= \widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} - \widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}} = \Sigma^{-1/2}\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}. \end{aligned}$$

Infatti, indicando con Σ^{GLS} la matrice di covarianza dello stimatore $\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}$ e adottando la seguente notazione matriciale:

$$\begin{aligned} gg^T &= (\langle g_i, g_j \rangle), \quad g = (g_1, ..., g_n) \in H^n \\ \mathbb{E}[\mathbb{A}] &= (\mathbb{E}[\mathbb{A}_{ij}]), \quad \mathbb{A} = (\mathbb{A}_{ij}) \in \mathbb{R}^n, \end{aligned}$$

risulta:

$$\begin{aligned} \Sigma := \mathrm{Cov}(\boldsymbol{\chi}_{\boldsymbol{s}}) &= \mathbb{E}[(\boldsymbol{\chi}_{\boldsymbol{s}} - m_{\boldsymbol{s}})(\boldsymbol{\chi}_{\boldsymbol{s}} - m_{\boldsymbol{s}})^T] = \\ &= \mathbb{E}[\Sigma^{1/2}(\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} - \widetilde{m}_{\boldsymbol{s}})(\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} - \widetilde{m}_{\boldsymbol{s}})^T \Sigma^{T/2}] = \\ &= \mathbb{E}[\Sigma^{1/2}(\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} \pm \widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}} - \widetilde{m}_{\boldsymbol{s}})(\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} \pm \widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}} - \widetilde{m}_{\boldsymbol{s}})^T \Sigma^{T/2}] = \\ &= \Sigma^{1/2}\mathbb{E}[(\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} - \widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}})(\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} - \widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}})^T + (\widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}} - m_{\boldsymbol{s}})(\widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}} - m_{\boldsymbol{s}})^T]\Sigma^T = \\ &= \mathbb{E}[\Sigma^{1/2}(\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} - \widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}})(\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} - \widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}})^T \Sigma^{T/2}] + \mathbb{E}[\Sigma^{1/2}(\widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}} - m_{\boldsymbol{s}})(\widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}} - m_{\boldsymbol{s}})^T \Sigma^{T/2}] = \\ &= \mathbb{E}[(\boldsymbol{\chi}_{\boldsymbol{s}} - \widehat{\boldsymbol{m}}_{\boldsymbol{s}})(\boldsymbol{\chi}_{\boldsymbol{s}} - \widehat{\boldsymbol{m}}_{\boldsymbol{s}})^T] + \mathbb{E}[(\widehat{\boldsymbol{m}}_{\boldsymbol{s}} - m_{\boldsymbol{s}})(\widehat{\boldsymbol{m}}_{\boldsymbol{s}} - m_{\boldsymbol{s}})^T] = \\ &= \mathbb{E}[\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}^T] + \mathbb{E}[(\widehat{\boldsymbol{m}}_{\boldsymbol{s}} - m_{\boldsymbol{s}})(\widehat{\boldsymbol{m}}_{\boldsymbol{s}} - m_{\boldsymbol{s}})^T] = \\ &= \Sigma^{GLS} + \mathrm{Cov}(\widehat{\boldsymbol{m}}_{\boldsymbol{s}}). \end{aligned}$$

Dai precedenti calcoli si può quindi concludere che la matrice di covarianza Σ risulta essere la somma della matrice di covarianza dello stimatore del vettore di *drift* e della matrice di covarianza Σ^{GLS} dello stimatore del residuo:

$$\Sigma = \Sigma^{GLS} + \mathbb{F}_{\boldsymbol{s}}^T \Lambda \mathbb{F}_{\boldsymbol{s}}. \tag{3.57}$$

Pertanto, la matrice di covarianza Σ^{GLS} di $\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}$, pur essendo lo stimatore naturale della matrice di covarianza Σ, fornisce una stima distorta della struttura di dipendenza dello spazio, sottostimandola sistematicamente di un fattore:

$$\mathbb{B} = \mathbb{F}_{\boldsymbol{s}}^T \Lambda \mathbb{F}_{\boldsymbol{s}}. \tag{3.58}$$

D'altra parte, la distorsione $\mathbb{B}$ della stima di Σ individuata tramite Σ^{GLS} corrisponde alla distorsione 'minima' ottenibile da un procedimento di stima ai minimi quadrati (nel senso che per ogni distorsione $\mathbb{B}'$ conseguente alla scelta di uno stimatore lineare dai dati $\widehat{\boldsymbol{a}}'_{\boldsymbol{l}}$ la matrice $\mathbb{B}' - \mathbb{B}$ è semidefinita positiva), in quanto gli stimatori $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}$ e $\widehat{\boldsymbol{\chi}}_{\boldsymbol{s}}$ risultano essere stimatori BLUE.

Pertanto, dal momento che la distorsione dello stimatore Σ^{GLS} corrisponde esattamente alla matrice di covarianza dello stimatore $\widehat{\boldsymbol{m}}_{\boldsymbol{s}}$, la precisione della stima ai minimi quadrati del *drift* assume una duplice importanza, determinando da un lato l'accuratezza della stima della componente deterministica, dall'altro la distorsione nella stima della struttura di dipendenza spaziale relativa alla componente stocastica, essenziale ai fini previsivi.

Queste osservazioni costituiscono un'ulteriore conferma dell'importanza dell'adozione del metodo dei minimi quadrati generalizzati. Infatti, qualora il criterio di ottimalità fosse basato

sui minimi quadrati ordinari, con gli stessi passaggi riportati in precedenza si otterrebbero gli stimatori lineari OLS:

$$\begin{aligned} \widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{OLS} &= (\mathbb{F}_{\boldsymbol{s}}^T\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}}^T\boldsymbol{\chi}_{\boldsymbol{s}} \\ \widehat{\boldsymbol{m}}_{\boldsymbol{s}} &= \mathbb{F}_{\boldsymbol{s}}\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{OLS} = \mathbb{F}_{\boldsymbol{s}}(\mathbb{F}_{\boldsymbol{s}}^T\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}}^T\boldsymbol{\chi}_{\boldsymbol{s}}, \end{aligned} \tag{3.59}$$

caratterizzati dai seguenti vettori media e matrici di covarianza:

$$\begin{aligned} \mathbb{E}[\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{OLS}] &= a_{\boldsymbol{l}} \\ \mathbb{E}[\widehat{\boldsymbol{m}}_{\boldsymbol{s}}] &= m_{\boldsymbol{s}} \\ \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{OLS}) &= (\mathbb{F}_{\boldsymbol{s}}^T\mathbb{F}_{\boldsymbol{s}})^{-1}(\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\mathbb{F}_{\boldsymbol{s}})(\mathbb{F}_{\boldsymbol{s}}^T\mathbb{F}_{\boldsymbol{s}})^{-1} \\ \mathrm{Cov}(\widehat{\boldsymbol{m}}_{\boldsymbol{s}}) &= \mathbb{F}_{\boldsymbol{s}}\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{OLS} = \mathbb{F}_{\boldsymbol{s}}(\mathbb{F}_{\boldsymbol{s}}^T\mathbb{F}_{\boldsymbol{s}})^{-1}(\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\mathbb{F}_{\boldsymbol{s}})(\mathbb{F}_{\boldsymbol{s}}^T\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}}^T. \end{aligned}$$

Inoltre, in questo caso, dalla scomposizione:

$$\|\boldsymbol{\chi}_{\boldsymbol{s}}\|_{H^n}^2 = \|\widehat{\boldsymbol{m}}_{\boldsymbol{s}}\|_{H^n}^2 + \|\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}\|_{H^n}^2,$$

valida per l'ortogonalità degli stimatori $\widehat{\boldsymbol{m}}_{\boldsymbol{s}}$ e $\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}$, seguirebbe la decomposizione della varianza:

$$\Sigma = \Sigma^{OLS} + \mathbb{F}_{\boldsymbol{s}}^T\Lambda^{OLS}\mathbb{F}_{\boldsymbol{s}}, \tag{3.60}$$

dove $\Lambda^{OLS} = \mathrm{Cov}(\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{OLS})$.

In particolare, dall'ottimalità degli stimatori lineari GLS in quanto BLUE, la matrice:

$$\Delta_{\mathbb{B}} := \mathbb{F}_{\boldsymbol{s}}^T\Lambda^{OLS}\mathbb{F}_{\boldsymbol{s}} - \mathbb{F}_{\boldsymbol{s}}^T\Lambda\mathbb{F}_{\boldsymbol{s}}$$

risulta definita positiva. Questo si riflette direttamente sulla distorsione dello stimatore $\widehat{\Sigma}$, che, nel passaggio da una stima OLS a una stima GLS, si riduce esattamente di un fattore $\Delta_{\mathbb{B}}$.

L'entità della riduzione è chiaramente dipendente dalla struttura di dipendenza spaziale, dal momento che la distorsione $\mathbb{B}$ ne è funzione attraverso la matrice Σ:

$$\begin{aligned} \mathbb{B} &= \mathbb{F}_{\boldsymbol{s}}^T(\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}} \\ \Delta_{\mathbb{B}} &= \mathbb{F}_{\boldsymbol{s}}(\mathbb{F}_{\boldsymbol{s}}^T\mathbb{F}_{\boldsymbol{s}})^{-1}(\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\mathbb{F}_{\boldsymbol{s}})(\mathbb{F}_{\boldsymbol{s}}^T\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}}^T - \mathbb{F}_{\boldsymbol{s}}^T(\mathbb{F}_{\boldsymbol{s}}^T\Sigma^{-1}\mathbb{F}_{\boldsymbol{s}})^{-1}\mathbb{F}_{\boldsymbol{s}} \end{aligned}$$

e al limite si annichilisce quando $\Sigma = \sigma^2\mathbb{I}_n$: in questo caso, infatti, il campione risulta scorrelato e le stime OLS e GLS coincidono, poiché la minimizzazione dei funzionali:

$$\|\boldsymbol{\chi}_{\boldsymbol{s}} - \widehat{\boldsymbol{m}}_{\boldsymbol{s}}\|_{\Sigma^{-1}-H^n}^2 = \|\boldsymbol{\chi}_{\boldsymbol{s}} - \widehat{\boldsymbol{m}}_{\boldsymbol{s}}\|_{\sigma^{-2}\mathbb{I}_n-H^n}^2$$

e

$$\|\boldsymbol{\chi}_{\boldsymbol{s}} - \widehat{\boldsymbol{m}}_{\boldsymbol{s}}\|_{H^n}^2$$

produce la medesima soluzione (la verifica è immediata sostituendo $\Sigma = \sigma^2\mathbb{I}_n$ nelle espressioni (3.48) e (3.59)).

Dal punto di vista variografico, il residuo di un processo scorrelato è caratterizzato da una struttura di puro *nugget*, dal momento che, in tal caso, la discrepanza tra le osservazioni dovuta alla componente stocastica è indipendente dalla distanza tra i siti di rilevamento ed è mediamente pari alla varianza σ^2 del processo.

La stima e l'analisi del variogramma del residuo consentono dunque di individuare la procedura di stima più adeguata al trattamento statistico delle osservazioni, come sarà evidente dagli algoritmi sviluppati nel Capitolo 5.

In particolare, la struttura di covarianza del processo, generalmente incognita nelle applicazioni, sarà stimata nel seguito per mezzo della stima variografica del residuo $\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}$, con i metodi proposti nel Capitolo 4. La stima $\widehat{\Sigma}$ della matrice Σ sarà quindi determinata attraverso il modello di variogramma stimato per $\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}}$, dunque attraverso uno stimatore di $\widehat{\Sigma}^{GLS}$ di Σ^{GLS} e gli stimatori saranno calcolati attraverso la stima $\widehat{\Sigma}$.

La procedura di stima proposta sarà basata su un metodo iterativo, che dopo una stima iniziale del *drift*, ricavi i parametri del modello di variogramma, quindi aggiorni i coefficienti del modello lineare con il metodo dei minimi quadrati generalizzati. Tale procedura sarà oggetto di studio nel Capitolo 5, nel quale saranno sviluppati gli opportuni algoritmi, con applicazione a dataset sintetici e reali.

Capitolo 4

Metodi Bootstrap Semiparametrici per la Stima della Distribuzione del Trace-Variogram

La stima della struttura di covarianza di un processo stocastico funzionale riveste un ruolo fondamentale nell'analisi geostatistica: infatti, da questa possono essere verificate le ipotesi di stazionarietà e isotropia del campo stesso, consentendo inoltre di applicare i metodi di kriging per la previsione delle curve non osservate a partire dai dati a disposizione.

L'analisi della struttura di dipendenza spaziale può essere condotta attraverso lo studio del *trace-variogram*, dapprima determinandone una stima empirica (Capitolo 3, Sezione 3.3), quindi adattandovi un modello valido.

In questo contesto, la considerazione della struttura di covarianza dello stimatore empirico consente da un lato di associare alla stima puntuale opportuni intervalli di confidenza, dall'altro di individuare, per la procedura di adattamento, un criterio di ottimo che tenga in considerazione la deformazione della geometria operata dalla struttura di dipendenza dello stimatore.

Non essendo disponibile, nel caso funzionale, un'espressione analitica per la matrice di covarianza dello stimatore sperimentale, nel presente capitolo è proposta una procedura di ricampionamento di tipo bootstrap, al fine di stimare l'intera distribuzione dello stimatore stesso.

In particolare, dopo una breve introduzione ai metodi bootstrap in ambito finito- dimensionale (Sezione 4.1), per il caso funzionale georeferenziato sarà proposta una metodologia semiparametrica basata su un algoritmo iterativo (Sezione 4.2), del quale sarà studiato il comportamento attraverso uno studio di simulazione (Sezione 4.3).

4.1 Introduzione ai Metodi Bootstrap

I metodi di ricampionamento si basano sull'uso dei dati osservati per generare campioni aggiuntivi in modo casuale. Tali tecniche, di cui fanno parte i metodi bootstrap, hanno ottenuto molta attenzione in letteratura, poiché costituiscono uno strumento importante per indagare le caratteristiche distribuzionali di una variabile aleatoria, evitando l'introduzione di ipotesi forti sul modello statistico in oggetto.

Dal punto di vista storico, i metodi di ricampionamento furono introdotti negli anni '30 da Fisher e Pitman, con i metodi di permutazione, ed esplorati successivamente con i metodi jackknife e delta a partire dal lavoro di Quenouille (1949). L'origine del bootstrap è invece più recente, essendo stato introdotto nella sua forma non parametrica nel lavoro di Efron (1979), nel quale è presentato come elemento unificatore di alcune tecniche precedentemente adottate, come i già citati metodi jackknife e delta.

L'obiettivo dei metodi bootstrap è quello di fornire uno strumento per la stima della variabilità di uno stimatore attraverso lo studio delle caratteristiche della distribuzione empirica del campione, che è nota, evitando in questo modo la difficoltà di formulare ipotesi parametriche. Infatti, a partire dalle informazioni derivanti dal campione, la procedura di ricampionamento, cioè di campionamento dalla distribuzione empirica, consente di accedere alle proprietà distribuzionali dello stimatore per via analitica o con l'ausilio di metodi di tipo Monte Carlo.

Più precisamente, si consideri un campione casuale $\boldsymbol{X} = (X_1, ..., X_n)$ di dimensione n da un modello statistico $P_\vartheta \in \mathcal{P} = \{P_\vartheta, \vartheta \in \Theta\}$, la cui realizzazione è denotata con $\boldsymbol{x} = (x_1, ..., x_n)$, dove il modello $\mathcal{P}$ può indicare sia un modello parametrico (se $\Theta \subseteq \mathbb{R}^k$) che non parametrico. Sia inoltre $\widehat{\vartheta}(\boldsymbol{X})$ uno stimatore di $\vartheta \in \Theta$ e si denoti con $P_{\widehat{\vartheta}}$ il relativo stimatore di P_ϑ.

Qualora non sia possibile determinare in modo esplicito la distribuzione dello stimatore $\widehat{\vartheta}$ e, in particolare, la variabilità ad esso associata, l'idea di base dei metodi bootstrap è quella di ricavarne una stima attraverso la distribuzione dello stesso stimatore ottenuto da un differente campione $\boldsymbol{X}^\star$, costruito campionando dalla distribuzione empirica del dataset di partenza: la stima della distribuzione di $\widehat{\vartheta}(\boldsymbol{X})$ è dunque ottenuta dalla valutazione della distribuzione di $\widehat{\vartheta}(\boldsymbol{X}^\star)$.

Se $\mathcal{P}$ è un modello parametrico, tale approccio conduce al bootstrap parametrico; nel caso in cui $\mathcal{P}$ sia invece un modello non parametrico, il metodo prende il nome di bootstrap non parametrico.

Nelle sezioni seguenti saranno introdotti i metodi bootstrap ed il loro fondamento teorico, con particolare riferimento al bootstrap non parametrico e alla sua applicazione nell'ambito spaziale.

4.1.1 Il Bootstrap Non Parametrico

Siano $X_1, ..., X_n$ variabili aleatorie indipendenti e identicamente distribuite a valori reali, di cui $x_1, ..., x_n$ siano una generica realizzazione, e si denoti con F la funzione di ripartizione di $X_i,\ i = 1, ..., n$.

Sia inoltre $T(\boldsymbol{X}, F)$ un funzionale di $\boldsymbol{X} = (X_1, ..., X_n)$ e F: esso definisce dunque una variabile aleatoria, determinata dal campione e dalla sua funzione di ripartizione. La stima di $T(\boldsymbol{X}, F)$ è fornita da $T_n = T(\boldsymbol{x}, F_n)$, dove $\boldsymbol{x} = x_1, ..., x_n$ e F_n è la distribuzione empirica:

$$F_n = \frac{1}{n} \sum_{i=1}^{n} I_{\{X_i \leq x\}}.$$

Più in generale, se $T(\boldsymbol{X}, P)$ è un funzionale di P e $\boldsymbol{X} = (X_1, ..., X_n)$ sono i.i.d da P, allora uno stimatore naturale di $T(\boldsymbol{X}, P)$ è $T(\boldsymbol{X}, P_n)$, dove P_n è la misura empirica:

$$P_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}.$$

La stima bootstrap non parametrica di $T(\boldsymbol{X}, F)$ (o $T(\boldsymbol{X}, P)$) è ottenuta attraverso il *principio di sostituzione* (Efron e Tibshirani, 1993), che prescrive che la stima avvenga attraverso la funzione di ripartizione empirica F_n (o misura empirica P_n) qualora F (o P) sia incognito.

Il fondamento teorico del metodo risiede nel fatto che la distribuzione della statistica $T^\star$, che è definita come:

$$T^\star = T(\boldsymbol{X}^\star, F_n),$$

ed è teoricamente calcolabile in forma analitica nota la realizzazione $\boldsymbol{x}$, è pari alla distribuzione di $T(\boldsymbol{X}, F)$ se $F = F_n$; tale risultato è la consistenza di Fisher applicata al problema di stima in esame.

Più precisamente, la consistenza di Fisher è definita come segue (Fisher, 1922).

Definizione 4.1. *Sia $\boldsymbol{X}$ un campione casuale tale che X_i abbia distribuzione F_ϑ per ogni $i = 1, 2, ..., n$, con $\vartheta \in \Theta$ parametro incognito. Se uno stimatore di ϑ ammette una rappresentazione come funzionale della distribuzione empirica F_n:*

$$\widehat{\vartheta} = T(F_n),$$

allora tale stimatore è consistente *se:*

$$T(F_\vartheta) = \vartheta.$$

Equivalentemente, per la legge dei grandi numeri, $\widehat{\vartheta} = T(F_n)$ si dice consistente se

$$T(\lim_{n \to \infty} F_n) = \vartheta.$$

Si noti che l'ipotesi che la statistica T sia rappresentabile come funzione di F_n non è particolarmente stringente: infatti, nell'ipotesi di scambiabilità di $X_1, ..., X_n$, dato uno stimatore T definito in termini di X_i, è possibile ricondursi a uno stimatore T' definito come funzionale di F_n mediando T su tutte le permutazioni dei dati; lo stimatore risultante ha lo stesso valore atteso di T e varianza non superiore.

Oltre al concetto di consistenza della statistica $T(\boldsymbol{X}, F)$, i metodi di ricampionamento trovano una base teorica nella legge dei grandi numeri e nel Teorema di Glivenko-Cantelli. Infatti, dalla legge dei grandi numeri discende la convergenza puntuale della funzione di ripartizione empirica a F_ϑ, convergenza che è anche uniforme grazie al Teorema di Glivenko-Cantelli, riportato di seguito.

Teorema 4.2 (Teorema di Glivenko-Cantelli)**.** *Sia $\mathcal{P}$ l'insieme delle distribuzioni di probabilità continue su uno spazio misurabile $(E, \mathcal{E})$ e sia D_n definito come:*

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| = ||F_n - F||_\infty. \tag{4.1}$$

Allora

$$\lim_{n \to \infty} D_n = 0 \quad q.c. \quad \forall\, F \in \mathcal{P},$$

cioè $F_n \to F$ uniformemente con probabilità 1.

La bontà dell'approssimazione della distribuzione di T attraverso la distribuzione di $T^\star$ dipende dalla forma del funzionale T. Sia ad esempio $\vartheta(F)$ un parametro di interesse (come la media o la deviazione standard di F) e sia $t(X)$ un suo stimatore. Poiché nel primo caso la dipendenza del funzionale T da F è solo indiretta attraverso $\boldsymbol{X}$, se $T(\boldsymbol{X}, F) = t(X)$ ci si aspetta che il metodo bootstrap sia meno efficace rispetto al caso in cui:

$$T(\boldsymbol{X}, F) = \frac{t(X) - \mathbb{E}_F[t(X)]}{\sqrt{\mathrm{Var}_F(t)}}.$$

I metodi bootstrap sono pertanto utili per l'analisi della accuratezza e dell'incertezza degli stimatori e in genere non sono usati per fornire delle migliori stime puntuali.

Per illustrare il metodo bootstrap sarà ora introdotto un semplice esempio, sviluppabile per via analitica.

Esempio 4.3 (Stima bootstrap della distorsione dello stimatore della media per una Bernoulli)**.** *Sia $\boldsymbol{X} = (X_1, ..., X_n)$ un campione casuale di variabili aleatorie reali con distribuzione Bernoulli di parametro $\vartheta(F) = \mathbb{P}_F\{X = 1\}$. Si consideri il funzionale definito da:*

$$T(\boldsymbol{X}, F) = \overline{X} - \vartheta(F)$$

dove $\overline{X}$ indica la media campionaria del campione $\boldsymbol{X}$.

A partire da n osservazioni $\boldsymbol{X} = \boldsymbol{x}$, occorre dunque costruire la funzione di ripartizione empirica F_n che risulta della forma:

$$F_n(x) = \begin{cases} 0, & x \leq 0 \\ 1 - \overline{x}, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases} \tag{4.2}$$

Fissata F_n, il campione bootstrap $\boldsymbol{X}^\star = (X_1^\star, X_2^\star, ..., X_n^\star)$ ha quindi componenti pari a 1 con probabilità $\overline{x}$, 0 con probabilità $1-\overline{x}$. Grazie alle proprietà della distribuzione Binomiale si ha perciò che la variabile aleatoria:

$$T_n^\star = T(\boldsymbol{X}^\star, F_n) = \overline{X}^\star - \overline{x}$$

ha media e varianza pari a:

$$\mathbb{E}_\star[\overline{X}^\star - \overline{x}] = 0$$
$$\mathrm{Var}_\star(\overline{X}^\star - \overline{x}) = \frac{\overline{x}(1-\overline{x})}{n}.$$

Ovvero i risultati ottenuti attraverso l'analisi della distribuzione bootstrap sono in accordo con i risultati ottenibili dalla statistica asintotica: è noto infatti che $\overline{X}$ è uno stimatore non distorto per ϑ con varianza $\vartheta(1-\vartheta)/n$ che è approssimativamente pari a $\overline{x}(1-\overline{x})/n$.

4.1.2 Il Bootstrap e le Approssimazioni Monte Carlo

Dal momento che generalmente non sono disponibili espressioni in forma chiusa per gli stimatori bootstrap, la loro valutazione avviene spesso in modo indiretto. Infatti, sebbene sia possibile, in linea di principio, enumerare tutti i possibili campioni di dimensione n da F_n ottenuti con reimmissione, essendo la distribuzione empirica F_n discreta, il numero di tali campioni è dell'ordine di n^n e cresce dunque molto velocemente al crescere della dimensione del campione ($\binom{2n-1}{n}$ se si considerano i campioni bootstrap distinti). D'altra parte, un'approssimazione di tipo Monte-Carlo della distribuzione bootstrap è semplice da ottenere: è infatti sufficiente generare B campioni indipendenti di dimensione n estratti con reimmissione da F_n:

$$\boldsymbol{X}_j = (X_{j,1}^\star, ..., X_{j,n}^\star), \quad j = 1, ..., B,$$

e quindi considerare la distribuzione empirica del j-esimo campione:

$$F_{j,n}^\star(x) = \frac{1}{n}\sum_{i=1}^{n} I_{\{X_{j,i}^\star \leq x\}}, \tag{4.3}$$

ed infine valutare la stima di interesse attraverso:

$$T_{j,n}^\star = T(\boldsymbol{X}_j, F_{j,n}^\star), \quad j = 1, ..., B. \tag{4.4}$$

L'algoritmo bootstrap con stima Monte Carlo è riportato di seguito.

Algoritmo 4.4. *Data una realizzazione $\boldsymbol{x} = (x_1, ..., x_n)$ di un campione casuale $\boldsymbol{X}$:*

1. *Costruire la distribuzione empirica F_n attribuendo massa $1/n$ ad ogni punto $x_1, ..., x_n$.*
2. *Fissato F_n, estrarre un campione casuale di dimensione n da F_n:*

$$X_i^\star = x_i^\star, \quad x_i^\star \sim F_n, \quad i = 1, 2, ..., n.$$

Il campione $\boldsymbol{X}^\star$ è detto campione bootstrap, di cui $\boldsymbol{x}^\star = (x_1^\star, ..., x_n^\star)$ è una realizzazione.

3. *Ripetere il passo 2. B volte ottenendo i campioni $\boldsymbol{X}_j = (X^\star_{j,1}, ..., X^\star_{j,n}), \quad j = 1, ..., B$, calcolare $F^\star_{j,n}$ e $T^\star_{j,n}$ con la (4.3) e (4.4).*

4. *Approssimare la distribuzione bootstrap di $T^\star = T(\boldsymbol{X}^\star, F_n)$ con la distribuzione empirica di:*

$$T^\star_{j,n} = T(\boldsymbol{X}_j, F^\star_{j,n}), \quad j = 1, ..., B.$$

5. *Approssimare la distribuzione campionaria di $T(\boldsymbol{X}, F)$ con la distribuzione bootstrap approssimata tramite metodo Monte Carlo di:*

$$T^\star = T(\boldsymbol{X}^\star, F_n)$$

ovvero la distribuzione di $T^\star$ indotta dal passo 2., mantenendo fissato F_n al suo valore osservato.

Gli istogrammi delle stime $T^\star_{j,n}$ così ottenute costituiscono quindi un'approssimazione della distribuzione bootstrap; analogamente, gli stimatori puntuali delle quantità caratteristiche della distribuzione Monte Carlo di $T^\star$ sono approssimazioni Monte Carlo della stima bootstrap delle rispettive quantità per la distribuzione di T_n.

Un parametro importante per tenere sotto controllo la varianza dell'approssimazione Monte Carlo è il numero di ripetizioni B. Infatti, date B repliche ottenute per ricampionamento, si denoti con σ^2_B la varianza dello stimatore di $\vartheta(F)$ ottenuto con un singolo campione bootstrap: dal momento che l'approssimazione Monte Carlo della stima bootstrap è calcolata come media di B stime indipendenti, la varianza dell'approssimazione Monte Carlo è pari a $B^{-1}\sigma^2_B$ ed è quindi decrescente al crescere di B.

La crescita del numero di ripetizioni comporta tuttavia un aumento dell'onere computazionale; pertanto, la scelta del valore di B appropriato per la stima Monte Carlo richiede particolare attenzione: occorre infatti da un lato fornire un'approssimazione della distribuzione bootstrap con varianza Monte Carlo sufficientemente contenuta, mantenendo d'altro canto un onere computazionale non eccessivo e commisurato all'accuratezza della stima.

Una regola solitamente accettata nell'ambito di dataset reali è la considerazione di un numero di ripetizioni attorno a $B = 100, 200$ per la stima bootstrap dell'errore standard e della distorsione, mentre almeno $B = 1000$ per la costruzione di bande di confidenza. Queste regole sono per lo più empiriche, basate su studi di simulazione e su applicazioni in molti ambiti. Tuttavia, nel caso in cui il dataset sia multivariato il numero di ripetizioni necessarie per ottenere una stima sufficientemente accurata deve necessariamente essere più alto del caso reale monovariato. In questo caso, infatti, la crescita della dimensione della variabile aleatoria in oggetto porta al problema noto come "*curse of dimensionality*" (*cfr.* Capitolo 2).

Si noti infine che le argomentazioni precedenti sono tipiche delle approssimazioni Monte Carlo, delle quali la tecnica MC-bootstrap costituisce un'applicazione particolare: infatti, quest'ultima ha per obiettivo accedere numericamente alle caratteristiche distribuzionali

	30	40	50	100	150	250	500	1000
$P_n(X=0)$	0.700	0.700	0.760	0.780	0.760	0.740	0.748	0.761
$P_n(X=1)$	0.300	0.300	0.240	0.220	0.240	0.260	0.252	0.239

Tabella 4.1: Densità di probabilità empirica dei campioni simulati al variare di n={30; 40; 50; 100; 150; 250; 500; 1000}.

dello stimatore bootstrap, scopo che viene perseguito attraverso la particolare scelta della distribuzione dalla quale campionare, ossia la distribuzione empirica.

Specifiche argomentazioni per la valutazione dell'accuratezza della stima MC-bootstrap nascono invece nella teoria dei processi empirici, per i quali tecnicismi, che esulano dallo scopo di questo lavoro, si rimanda ad esempio a (van der Vaart e Wellner, 1996), (Kosorok, 2008) e (Wellner, 2011). Come conclusione della sezione sarà ora mostrato il comportamento dell'Algoritmo 4.4 appena introdotto attraverso la sua applicazione ad un semplice esempio, continuazione dell'Esempio 4.3.

Esempio 4.5 (Esempio di stima bootstrap con approssimazione Monte Carlo)**.** *Sia* $\boldsymbol{X} = (X_1, X_2, ..., X_n)$ *un campione i.i.d. con distribuzione Bernoulli di parametro* $\vartheta = 0.25$ *e si denoti con* $\boldsymbol{x} = (x_1, ..., x_n)$ *una sua realizzazione, caratterizzata da una funzione di ripartizione empirica* F_n*.*

Per indagare il comportamento dell'approssimazione bootstrap al variare della dimensione campionaria n è stata effettuata la seguente simulazione. Al variare di n nell'insieme $\{30; 40; 50; 100; 150; 250; 500; 1000\}$*, è stato applicato l'Algoritmo 4.4 a partire da un campione* $\boldsymbol{x}_n$*, fissando* $B = 1000$ *e valutando per ogni n la distribuzione bootstrap dello stimatore media campionaria* $\overline{X}$ *e della distorsione della media campionaria* $T(\boldsymbol{X}, F) = \overline{X} - \vartheta(F)$*. In Tabella 4.1 sono riportate le distribuzioni empiriche dai campioni* $\boldsymbol{x}_n$ *generati.*

In accordo con i risultati di statistica asintotica (Casella e Berger, 2002), i grafici in Figura 4.1 evidenziano che la media campionaria si assesta per n crescente verso il valore del parametro incognito ϑ*, con varianza associata evanescente, mostrando una distribuzione sempre più concentrata attorno al valor medio.*

In Figura 4.2 sono riportati i violin-plot degli stimatori $\overline{X}$ *e* $T(\boldsymbol{X}, F)$ *al variare della dimensione campionaria n: come si può osservare, la distribuzione bootstrap, coerentemente con la distribuzione asintotica degli stimatori, si approssima alla distribuzione gaussiana per n crescente, concentrandosi attorno al valor medio.*

Il secondo obiettivo di questo esempio è lo studio della bontà dell'approssimazione Monte Carlo al variare del numero di ripetizioni B nell'insieme $\mathcal{B}$={50; 100; 200; 350; 500; 1000; 2500; 5000}*.*

Analogamente a quanto fatto in precedenza, è stato quindi applicato l'Algoritmo 4.4 al crescere di B, ottenendo le distribuzioni MC-bootstrap, rappresentate in Figura 4.3. Per $n = 30$ *è possibile adottare l'approssimazione gaussiana per lo stimatore* $\overline{X}$ *e, in accordo con*

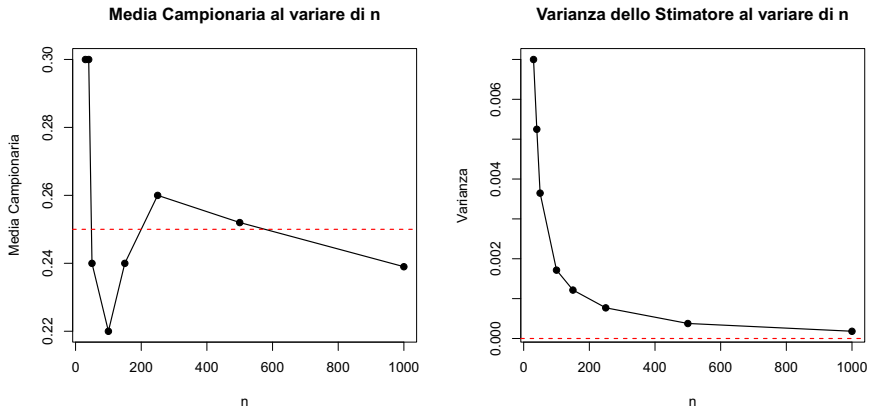

Figura 4.1: Media campionaria $\overline{x}$ e varianza dello stimatore $\overline{x}$, al variare della numerosità campionaria n in $\{30; 40; 50; 100; 150; 250; 500; 1000\}$.

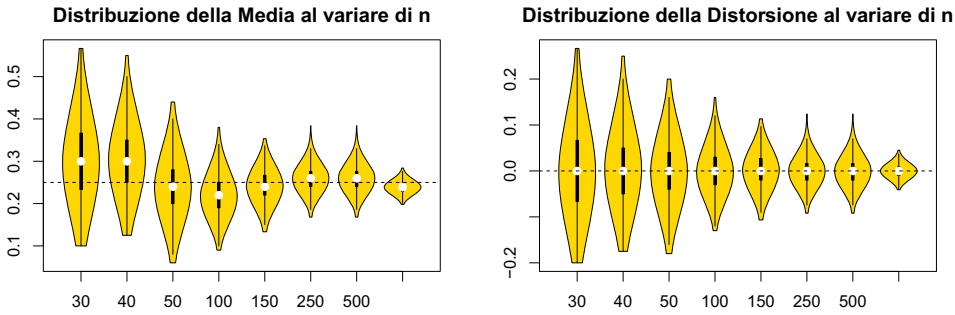

Figura 4.2: Violin-plot della distribuzione bootstrap della media campionaria $\overline{X}$ e della sua distorsione al variare della numerosità campionaria n in $\{30; 40; 50; 100; 150; 250; 500; 1000\}$, approssimata con il metodo Monte Carlo per $B = 1000$. Al centro dei violin-plot è rappresentato il box-plot delle osservazioni, i bordi corrispondono alla densità di probabilità empirica stimata dal campione.

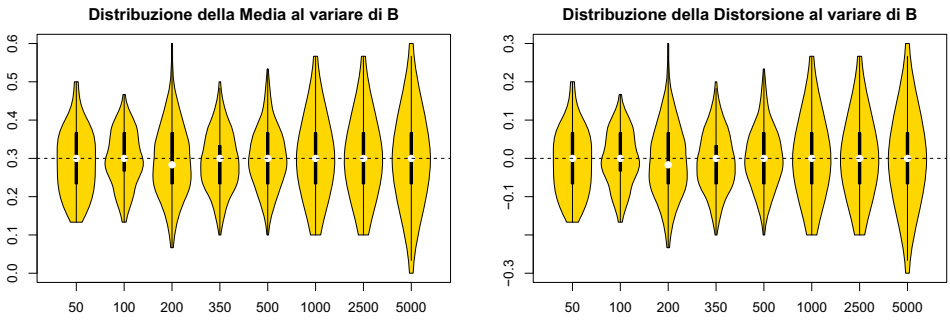

Figura 4.3: Violin-plot della distribuzione MC-bootstrap dello stimatore media campionaria $\overline{X}$ e della sua distorsione al variare del numero di iterazioni bootstrap B in $\{50; 100; 200; 350; 500; 1000; 2500; 5000\}$, per $n = 30$.

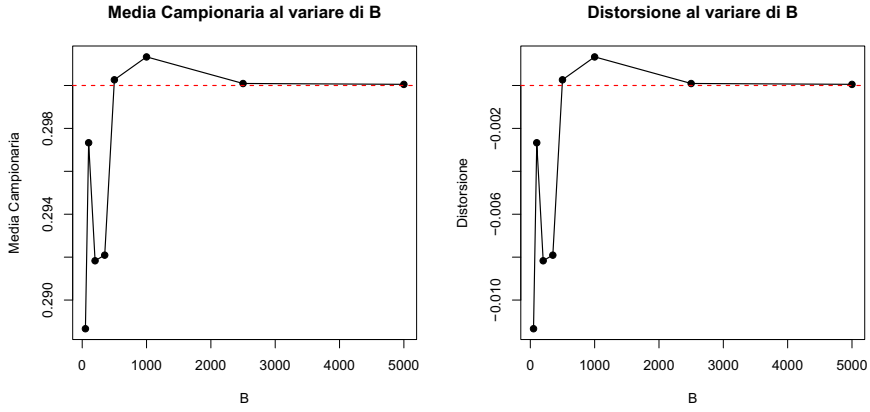

Figura 4.4: Stime Monte Carlo di media campionaria $\overline{x}$ e distorsione $T(\boldsymbol{x}, F)$, per $n = 30$ e B={50; 100; 200; 350; 500; 1000; 2500; 5000}.

questa, l'immagine mostra che la distribuzione approssimata assume la forma tipica della distribuzione gaussiana già per $B = 500$*, non evidenziando particolari variazioni al crescere della numero di iterazioni Monte Carlo oltre* $B = 1000$*.*

In Figura 4.4 sono invece riportati i grafici delle stime di $\overline{X}$ *e* $T(\boldsymbol{X}, F) = \overline{X} - \vartheta(F)$ *ottenute con approssimazione Monte Carlo, al variare del numero di ripetizioni* B*. Si può notare che, per* B *crescenti, tali valori si assestano sulle rispettive stime campionarie calcolate dal campione originale, indicando che l'approssimazione è accurata con uno scarto inferiore all'1% già per* $B = 100$*.*

L'Esempio 4.5 permette di comprendere quanto asserito in precedenza circa l'utilità dei metodi bootstrap. Nel caso in esame, il modello parametrico che ha generato i dati è completamente specificato; inoltre, per il parametro ϑ è noto lo stimatore di massima verosimiglianza

e le sue caratteristiche distribuzionali e asintotiche. Lo stimatore bootstrap non può quindi avere un comportamento migliore di questo: infatti, esso corrisponde esattamente alla media campionaria e la stima Monte Carlo ne fornisce un'approssimazione. Applicare la procedura bootstrap a questo caso sembra un onere computazionale inutile, tuttavia tale metodo, come asserito in precedenza, non è pensato per fornire una stima puntuale migliore, ma per valutare l'accuratezza degli stimatori in contesti molto più generali. L'esempio qui mostrato è pensato per illustrare il funzionamento del metodo in un caso nel quale tutte le variabili aleatorie siano sotto controllo, tuttavia è bene ricordare che i metodi di ricampionamento risultano davvero utili qualora sia difficile formulare ipotesi distribuzionali e parametriche, come nel caso funzionale che sarà studiato in seguito.

4.1.3 Bootstrap per Dati Spaziali

Sebbene il teorema di Glivenko-Cantelli ammetta una generalizzazione al caso in cui $X_1, ..., X_n$ siano variabili aleatorie di un processo stocastico a tempo discreto $\{X_t, t \geq 1\}$ strettamente stazionario ed ergodico (Tucker, 1959), nel caso di dati georeferenziati i metodi di ricampionamento non sono più applicabili nella loro forma standard: al contrario, essi devono essere adattati affinché si tenga in considerazione la correlazione esistente tra le osservazioni.

Infatti, Sebastién (2004) propose l'uso del bootstrap non parametrico standard affiancato al metodo dei percentili (Chernick, 1999) al fine di costruire intervalli di confidenza per il variogramma; tuttavia, come mostrato in Tang *e altri* (2006), l'applicazione *naive* di una procedura di ricampionamento a dati con una struttura di dipendenza spaziale fornisce una probabilità di copertura anche notevolmente diversa dalla nominale, a meno che la correlazione non sia trascurabile.

Per questi motivi sono stati proposti diversi adattamenti dei metodi di ricampionamento al caso di dati georeferenziati, adottando opportune modifiche che tengano conto della distribuzione spaziale del campione in esame.

Esistono essenzialmente tre classi di modifiche apportabili ai metodi standard per tenere in considerazione la dipendenza spaziale caratteristica del campione nell'implementazione delle procedure di ricampionamento:

- effettuare un ricampionamento a blocchi;
- costruire un modello parametrico per il processo stocastico e simulare un campione da tale modello con la stima dei parametri ottenuta dal campione osservato;
- costruire un modello di covariogramma e usare opportunamente quest'ultimo per ricampionare da un dataset scorrelato.

La prima classe di tecniche dà luogo ai metodi block-bootstrap, la seconda al bootstrap parametrico, la terza al bootstrap semiparametrico.

Il primo a proporre l'uso di tecniche di ricampionamento a blocchi nell'ambito spaziale fu Hall (1985), seguito da Künsch (1989) che propose l'applicazione del block-bootstrap per lo

studio della media campionaria delle osservazioni da un processo debolmente stazionario. Di questi metodi sono state studiate le proprietà asintotiche con particolare riferimento a processi stocastici reali, nel caso di disegno sperimentale *increasing-domain* (Lahiri, 1992),(Lahiri, 1993) e a dominio fisso (Loh e Stein, 2008).

Contemporaneamente a Hall, Solow (1985) propose un approccio semiparametrico per dati provenienti da serie temporali monodimensionali, metodo successivamente modificato ed esteso a dati bidimensionali in (Tang *e altri*, 2006), con l'introduzione di un metodo bootstrap parametrico.

Le procedure a blocchi e semiparametriche sono state riviste, confrontate e affiancate a studi di simulazione in lavori recenti dedicati alla valutazione dell'incertezza legata alla stima del variogramma empirico (Olea e Pardo-Igúzquiza, 2011) e (Clark e Alligham, 2011).

Infatti, le espressioni analitiche della matrice di covarianza dello stimatore empirico (1.9) (Marchant e Lark (2004), Pardo-Igúzquiza e Dowd (2001)) sono fondate sull'ipotesi che il processo che genera le osservazioni sia gaussiano, ipotesi che nella pratica è difficile, se non impossibile, da verificare; per questo motivo, i metodi di ricampionamento e, in particolare il bootstrap semiparametrico, sono stati sfruttati da Olea e Pardo-Igúzquiza (2011) per superare le difficoltà legate alla stima di un modello parametrico basata sul metodo GLS.

Dal punto di vista formale, si consideri un processo stocastico a valori reali, stazionario al second'ordine e isotropo (Definizioni 1.2 e 1.6):

$$\{Z(s), s \in D \subseteq \mathbb{R}^d\}, \tag{4.5}$$

e si indichino con $z_{s_1}, ..., z_{s_n}$ le realizzazioni di $Z_{s_1}, ..., Z_{s_n}$ in posizione $s_1, ..., s_n$ nel dominio spaziale D.

Denotato con $C(h)$ il covariogramma del campo aleatorio (4.5), con $\gamma(h)$ il corrispondente semivariogramma, sia $\Sigma = (\Sigma_{ij})$ la matrice di covarianza del vettore aleatorio $\boldsymbol{Z} = (Z_{s_1}, ..., Z_{s_n})$, indotta da $C(h)$:

$$\Sigma_{ij} = \mathrm{Cov}(Z_{s_i}, Z_{s_j}) = C(||s_i - s_j||).$$

Il metodo bootstrap semiparametrico si basa sulla considerazione del vettore aleatorio ausiliario:

$$\widetilde{\boldsymbol{Z}} = \Sigma^{-1/2}\boldsymbol{Z}$$

le cui componenti sono supposte essere i.i.d., ipotesi che è certamente verificata nel caso in cui il processo stocastico sia gaussiano (i.e. tutte le leggi finito dimensionali siano gaussiane). Si noti che le componenti del vettore $\widetilde{\boldsymbol{Z}}$ risultano almeno scorrelate per definizione, infatti:

$$\Sigma := \mathrm{Cov}(\boldsymbol{Z}) = \Sigma^{1/2}\mathrm{Cov}(\widetilde{\boldsymbol{Z}})\Sigma^{T/2}$$

che è pari a $\Sigma^{1/2}\Sigma^{T/2}$ se e solo se $\mathrm{Cov}(\widetilde{\boldsymbol{Z}})$ è la matrice identità in $\mathbb{R}^n$.

La procedura bootstrap è quindi applicata al campione $\widetilde{\boldsymbol{Z}}$ le cui componenti sono indipendenti per ipotesi (o almeno scorrelate per l'osservazione precedente), seguendo l'Algoritmo 4.4.

Una volta ottenuto un campione bootstrap $\widetilde{\boldsymbol{Z}}^{\star}$ campionando dalla distribuzione empirica, le sue componenti vengono ricorrelato con la trasformazione inversa, ottenendo il corrispondente campione bootstrap con dipendenza spaziale $\boldsymbol{Z}^{\star}$, sul quale calcolare le statistiche di interesse.

Per garantire il soddisfacimento dell'ipotesi di indipendenza delle componenti del campione $\widetilde{\boldsymbol{Z}}$, nel lavoro di Olea e Pardo-Igúzquiza (2011) è proposta l'introduzione di un ulteriore passo preliminare per rendere il campione approssimativamente gaussiano attraverso una funzione non lineare opportunamente scelta (Deutsch e Journel, 1998). Tale passo preliminare rende tuttavia inapplicabile la procedura al caso multivariato e funzionale, non esistendo un modo efficiente ed efficace per la stima di una trasformazione non lineare dei dati che mappi lo spazio campionario in uno spazio gaussiano. Non è stato altresì effettuato uno studio di simulazione nel caso monovariato, né multivariato, della robustezza del metodo nel caso di campione non gaussiano, studio che sarà condotto nel seguito per il caso funzionale.

4.2 Bootstrap Semiparametrico per Dati Funzionali Georeferenziati

4.2.1 L'Algoritmo

Si consideri un processo stocastico funzionale stazionario e isotropo:

$$\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}, \tag{4.6}$$

e si indichi con $C(h)$ il suo covariogramma, con $\gamma(h)$ il suo variogramma e con Σ la matrice di covarianza del vettore $\boldsymbol{\chi_s} = (\boldsymbol{\chi}_{s_1}, ..., \boldsymbol{\chi}_{s_n})$.

Obiettivo di questa sezione è l'introduzione di un algoritmo MC-bootstrap per dati funzionali georeferenziati, con applicazione alla costruzione di intervalli di confidenza per lo stimatore campionario di $\gamma(h)$ e all'adattamento di un modello valido di variogramma con il criterio dei minimi quadrati generalizzati.

Nel contesto dei dati funzionali indipendenti e identicamente distribuiti, risultati teorici per il bootstrap sono stati stabiliti all'interno della teoria dei processi empirici (Giné e Zinn (1990), Politis e Romano (1994), van der Vaart e Wellner (1996)) e recentemente studiati in contesti più applicativi in particolare da Castro *e altri* (2005), Cuevas *e altri* (2006) e Ferraty *e altri* (2010).

Nell'ambito dei dati funzionali georeferenziati, ad oggi non sono presenti in letteratura riferimenti all'uso di tecniche di ricampionamento. La metodologia proposta nel seguito è dunque una prima estensione al caso funzionale del bootstrap semiparametrico, inserito nella cornice teorica introdotta nel Capitolo 3; nello specifico, sarà proposto un procedimento iterativo ispirato in particolare ai lavori di Pardo-Igúzquiza e Dowd (2001) e Olea e Pardo-Igúzquiza (2011).

Sia dunque $\widehat{\gamma}(h)$ il semivariogramma empirico definito dalla (3.22) e si denoti con $\widehat{\gamma}(\boldsymbol{h}) = (\widehat{\gamma}(h_1), ..., \widehat{\gamma}(h_H))$ lo stimatore *binned variogram* (3.14). In analogia con il caso reale, si con-

sideri il dataset funzionale ausiliario $\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} \in H^n$, ottenuto premoltiplicando il vettore aleatorio $\boldsymbol{\chi}_{\boldsymbol{s}}$ per la matrice $\Sigma^{-1/2}$:

$$\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} = \Sigma^{-1/2}\boldsymbol{\chi}_{\boldsymbol{s}}, \tag{4.7}$$

dove, con questa scrittura, si intende che $\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} = (\widetilde{\boldsymbol{\chi}}_{s_1}, ..., \widetilde{\boldsymbol{\chi}}_{s_n})$ è un vettore la cui generica componente $\widetilde{\boldsymbol{\chi}}_{s_i}$, $i = 1, ..., n$, è pari a:

$$\widetilde{\boldsymbol{\chi}}_{s_i} = \sum_{j=1}^{n} \Sigma_{ij}^{-1/2} \boldsymbol{\chi}_{s_j}.$$

Come osservato in precedenza (Capitolo 3, Sezione 3.5), le componenti del campione funzionale (4.7) risultano per costruzione scorrelate: per questo motivo, tale campione è il candidato naturale al quale applicare la procedura di ricampionamento.

Dal punto di vista applicativo, la matrice di covarianza Σ nella (4.7) deve essere sostituita da una sua stima $\widehat{\Sigma}$, che tuttavia non può essere ricavata dallo stimatore $\widehat{\gamma}(\boldsymbol{h})$ senza determinarne preliminarmente i parametri di un opportuno modello valido $\gamma(h; \boldsymbol{\vartheta})$.

In assenza di ipotesi distribuzionali sul campo aleatorio (3.1), il vettore di parametri $\boldsymbol{\vartheta}$ può essere ottenuto attraverso il metodo dei minimi quadrati ordinari (OLS), sulla base del quale calcolare una stima $\widehat{\Sigma}$ della matrice di covarianza Σ. Dal momento che $\gamma(h; \boldsymbol{\vartheta})$ è un modello valido, la stima $\widehat{\Sigma}$ è semidefinita positiva e può quindi essere espressa, grazie alla decomposizione di Cholesky (Quarteroni *e altri*, 2008), come:

$$\widehat{\Sigma} = LL^T,$$

dove $L \in \mathbb{R}^{n \times n}$ è una matrice triangolare inferiore che individua una stima della matrice $\Sigma^{1/2}$, calcolabile in modo efficiente. Il campione ausiliario sul quale svolgere la procedura bootstrap è quindi definito come:

$$\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} = L^{-1}\boldsymbol{\chi}_{\boldsymbol{s}}. \tag{4.8}$$

Una volta determinato il campione scorrelato $\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}}$ ed ottenuto un campione bootstrap $\widetilde{\boldsymbol{\chi}}^{\star}_{\boldsymbol{s}}$ per ricampionamento, il dataset funzionale bootstrap georeferenziato è ottenuto attraverso la trasformazione inversa:

$$\boldsymbol{\chi}^{\star}_{\boldsymbol{s}} = L\widetilde{\boldsymbol{\chi}}^{\star}_{\boldsymbol{s}},$$

calcolando, infine, su quest'ultimo le statistiche di interesse. Il procedimento illustrato è sintetizzato nell'Algoritmo 4.6 riportato di seguito.

Algoritmo 4.6. *Data una realizzazione $\chi_{\boldsymbol{s}} = (\chi_{s_1}, ..., \chi_{s_n})$ di una successione di variabili aleatorie da un processo stocastico funzionale stazionario e isotropo $\{\boldsymbol{\chi}_s, s \in D\}$, $D \subset \mathbb{R}^k$:*

1. *Stimare tramite OLS i parametri di un modello parametrico $\gamma(h; \boldsymbol{\vartheta})$ per il variogramma e calcolare stima $\widehat{\Sigma}$ della matrice di covarianza Σ di $\boldsymbol{\chi}_{\boldsymbol{s}}$.*

2. *Calcolare la decomposizione di Cholesky di $\widehat{\Sigma}$:*

$$\widehat{\Sigma} = LL^T$$

e definire:

$$\widetilde{\chi}_{\boldsymbol{s}} = L^{-1}\chi_{\boldsymbol{s}}.$$

3. *Costruire la distribuzione empirica F_n attribuendo massa $1/n$ a ciascun dato funzionale $\widetilde{\chi}_{s_1}, ..., \widetilde{\chi}_{s_n}$.*

4. *Fissato F_n, estrarre un campione i.i.d. di dimensione n da F:*

$$\widetilde{\boldsymbol{\chi}}^{\star}_{s_i} = \widetilde{\chi}^{\star}_{s_i}, \quad \widetilde{\chi}^{\star}_{s_i} \sim F_n, \quad i = 1, 2, ..., n.$$

Il campione $\widetilde{\boldsymbol{\chi}}^{\star}_{\boldsymbol{s}}$ è detto campione bootstrap, $\widetilde{\chi}^{\star}_{\boldsymbol{s}} = (\widetilde{\chi}^{\star}_{s_1}, ..., \widetilde{\chi}^{\star}_{s_n})$ ne è una realizzazione.

5. *Antitrasformare il campione bootstrap per ottenere un campione bootstrap le cui componenti siano spazialmente dipendenti:*

$$\boldsymbol{\chi}^{\star}_{\boldsymbol{s}} = L\widetilde{\boldsymbol{\chi}}^{\star}_{\boldsymbol{s}}.$$

6. *Approssimare la distribuzione campionaria di $T(\boldsymbol{\chi}_{\boldsymbol{s}}, F)$ con la distribuzione bootstrap di:*

$$T^{\star} = T(\boldsymbol{\chi}^{\star}_{\boldsymbol{s}}, F_n)$$

ovvero la distribuzione di $T^{\star}$ indotta dal passo 4., mantenendo fissato F_n al suo valore osservato.

4.2.2 Il Metodo MC-Bootstrap per la Stima del Variogramma

Una volta definita la procedura di stima bootstrap semiparametrica, è necessario adattare il metodo al fine di applicarlo al problema in oggetto.

La distribuzione di interesse nel caso specifico in esame è la legge dello stimatore empirico $\widehat{\gamma}(\boldsymbol{h})$: da questa è infatti possibile costruire gli intervalli di confidenza per lo stimatore campionario e individuare la stima GLS dei parametri $\boldsymbol{\vartheta}$ di un modello valido $\gamma(h; \boldsymbol{\vartheta})$.

Si fissi dunque la statistica T come:

$$T(\boldsymbol{\chi}_{\boldsymbol{s}}, F) = \widehat{\gamma}(\boldsymbol{h})$$

e si considerino i seguenti problemi:

1. Individuare K intervalli di confidenza di livello $1 - \alpha$

$$IC^{k}_{1-\alpha}(\gamma(h_k)) = [\gamma^{min}(h_k), \gamma^{max}(h_k)], \quad k = 1, ..., K. \tag{4.9}$$

2. Determinare i parametri ottimi $\boldsymbol{\vartheta}$ secondo il criterio dei minimi quadrati generalizzati, minimizzando cioè il funzionale:

$$d_{\widehat{\Psi}^{-1}}(\widehat{\gamma}(\boldsymbol{h}), \gamma(\boldsymbol{h}; \boldsymbol{\vartheta})) = (\widehat{\gamma}(\boldsymbol{h}) - \gamma(\boldsymbol{h}; \boldsymbol{\vartheta}))^{T}\widehat{\Psi}^{-1}(\widehat{\gamma}(\boldsymbol{h}) - \gamma(\boldsymbol{h}; \boldsymbol{\vartheta})), \tag{4.10}$$

dove $\Psi = \mathrm{Cov}(\widehat{\gamma}(\boldsymbol{h}))$, $\widehat{\Psi}$ ne è una stima e $\boldsymbol{h} = (h_1, ..., h_K)$.

Dall'approssimazione della distribuzione dello stimatore *binned variogram*, ottenuta seguendo l'Algoritmo 4.6, gli intervalli di confidenza (4.9) possono essere calcolati con il metodo dei percentili (Chernick, 1999), ovvero valutando due opportuni quantili γ_{β_1}, γ_{β_2} della distribuzione di $\gamma(h_k)$, $k = 1, ..., K$, al cui interno siano registrate il $100\alpha\%$ delle realizzazioni bootstrap:

$$IC^{*k}_{1-\alpha}(\gamma(h_k)) = [\gamma_{\alpha/2}(h_k), \gamma_{1-\alpha/2}(h_k)], \quad k = 1, ..., K.$$

Il problema di adattamento 2. può invece essere affrontato e risolto valutando la stima bootstrap $\widehat{\Psi}^\star$ della covarianza dello stimatore empirico, ovvero:

$$\widehat{\Psi}^\star = \mathrm{Cov}_\star(\widehat{\gamma}^\star(\boldsymbol{h})),$$

dove $\widehat{\gamma}^\star(\boldsymbol{h})$ è il *binned variogram* calcolato sul campione bootstrap.

Si noti tuttavia che l'approssimazione della distribuzione di $\widehat{\gamma}(\boldsymbol{h})$, potrebbe dipendere dalla stima iniziale del modello di variogramma, peraltro basata su un metodo OLS. Una possibile soluzione a questo inconveniente è l'introduzione di un procedimento di stima iterativo, con aggiornamenti successivi della stima della matrice Σ: sulla base di una prima stima OLS del modello di variogramma, si procede nella stima della matrice Σ secondo l'Algoritmo 4.6; si usa quindi la stima ottenuta per determinare una nuova stima dei parametri del modello tramite GLS, proseguendo fino a convergenza. Il procedimento appena illustrato è sintetizzato nell'Algoritmo 4.7.

Algoritmo 4.7. *Data una realizzazione $\chi_{\boldsymbol{s}} = (\chi_{s_1}, ..., \chi_{s_n})$ di una successione di variabili aleatorie da un processo stocastico funzionale stazionario e isotropo $\{\boldsymbol{\chi}_s, s \in D\}$, $D \subset \mathbb{R}^k$:*

1. *Applicare l'Algoritmo 4.6 a $T(\boldsymbol{Z}, F) = \widehat{\gamma}(\boldsymbol{h})$*
2. *A partire dalla stima $T^\star$, stimare i parametri di un modello parametrico $\gamma(h; \boldsymbol{\vartheta})$ con il metodo GLS; calcolare il corrispondente modello di covariogramma $C(h; \boldsymbol{\vartheta})$ e la matrice di covarianza $\widehat{\Sigma}$ delle osservazioni.*
3. *Applicare l'Algoritmo 4.6 partendo dalla stima di $\widehat{\Sigma}$ ottenuta al passo 2.*
4. *Ripetere 2. e 3. fino a convergenza.*

Gli Algoritmi 4.6 e 4.7 consentono dunque di fornire una stima della variabilità dello stimatore empirico del variogramma, sfruttando le informazioni derivanti dal campione. Tuttavia, come si è anticipato in precedenza, non è presente in letteratura la dimostrazione della convergenza di tali procedure applicate all'ambito funzionale georeferenziato. In aggiunta, i metodi introdotti prevedono la determinazione preliminare di alcuni parametri caratteristici, dai quali dipende la bontà delle approssimazioni calcolate.

Nella prossima sezione, sarà indagato attraverso uno studio di simulazione il comportamento degli algoritmi introdotti, con particolare riferimento alla loro convergenza, alla differenza tra la stima GLS e la stima OLS e alla probabilità di copertura degli intervalli di confidenza costruiti. Infine, ne saranno valutati i risultati al variare dei parametri

caratteristici, in particolare del numero di lag K in cui calcolare lo stimatore *binned variogram*, del numero di iterazioni dell'approssimazione Monte Carlo B (che saranno denominate nel seguito *iterazioni bootstrap*) relativa all'Algoritmo 4.6 e del numero di iterazioni N_{max} dell'Algoritmo 4.7, che saranno indicate con il nome di *iterazioni esterne*.

4.3 Applicazione a Dati Simulati

L'Algoritmo 4.7 formulato nella Sezione 4.2 fornisce un metodo per determinare la stima GLS di un modello valido di variogramma sulla base di una stima del *binned-variogram* $\widehat{\gamma}(h)$, accompagnando quest'ultimo da opportuni intervalli di confidenza.

In questa sezione sarà illustrato il comportamento degli algoritmi proposti attraverso uno studio di simulazione, svolto sui dataset funzionali sintetici che saranno introdotti di seguito.

4.3.1 I Dataset Funzionali Sintetici

4.3.1.1 Campione Gaussiano

Sarà ora descritto il principale dataset sintetico sul quale sono state testate le metodologie sviluppate nel lavoro di tesi. La scelta della particolare forma funzionale (esponenziale) è stata dettata dal tipo di dati provenienti dal contesto industriale nel quale ha avuto origine il lavoro di tesi: si è infatti cercato di generare un campione sintetico verosimile, fissando con questo criterio la dimensione del dominio spaziale D, il dominio $\mathcal{T}$ dei dati funzionali e le relative unità di misura.

Si consideri dunque il seguente processo stocastico funzionale:

$$\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}, \tag{4.11}$$

di cui la variabile aleatoria funzionale $\boldsymbol{\chi}_s$, $s \in D$ sia della forma:

$$\boldsymbol{\chi}_s(z) = \Phi_0(s) \exp\{-z/\beta\}, \quad z \in \mathcal{T} = [0, 10000], \tag{4.12}$$

dove l'ascissa z indica la profondità, espressa in metri, il parametro $\beta \in \mathbb{R}$ è costante, mentre $\Phi_0(s)$ è un processo stocastico gaussiano stazionario e isotropo a valori reali:

$$\{\Phi_0(s) : s \in D \subseteq \mathbb{R}^d\}, \tag{4.13}$$

di media m_Φ e variogramma $\gamma_\Phi(h)$.

Si fissi come spazio H di Hilbert lo spazio L^2 delle funzioni quadrato integrabili, dotato del prodotto interno e della norma indotta:

$$< f, g > = \int_{\mathcal{T}} f(t) \cdot g(t) dt \tag{4.14}$$

$$\|f\| = \sqrt{\int_{\mathcal{T}} |f(t)|^2 dt},$$

per $f, g \in L^2$.

Grazie allo sfondo teorico del Capitolo 3, noti la media e la struttura di covarianza del campo aleatorio (4.13), le funzioni media m e il variogramma $2\gamma(h)$ del processo (4.11) possono essere ricavati rispettivamente come:

$$m_s(z) = m_\Phi \exp\{-z/\beta\} \tag{4.15}$$

e

$$\gamma(h) = \gamma_\Phi(h) \| \exp\{- \cdot /\beta\} \|^2. \tag{4.16}$$

Per lo studio di simulazione, i parametri sono stati fissati come segue:

1. $\beta = 5500$;
2. $m_\Phi = 0$;
3. $\gamma_\Phi(h)$ pari a un modello esponenziale di *practical range* $3a = 75$ km, *sill* $\sigma^2 = 80$ e *nugget* $\tau^2 = 0$:
$$\gamma(h) = \begin{cases} 80(1 - e^{-3h/75}), & h > 0 \\ 0, & h = 0, \end{cases} \tag{4.17}$$
con h espresso in chilometri.

Una volta fissati i parametri, il campo è stato simulato 100 volte su una fitta griglia rettangolare ($D = [0, 200] \times [0, 300]$ km con passo 2 km); in corrispondenza di 100 localizzazioni $s_1, ..., s_{100}$ campionate in modo uniforme dalla griglia stessa, sono state quindi estratte 100 realizzazioni $(\chi^i_{s_1}, ..., \chi^i_{s_{100}})$ del campione $(\boldsymbol{\chi}^i_{s_1}, ..., \boldsymbol{\chi}^i_{s_{100}})$, $i = 1, ..., 100$, usate nelle analisi come segue.

La prima realizzazione, mostrata in Figura 4.5, è stata considerata come l'unica osservazione $\chi_{s_1}, ..., \chi_{s_{100}}$ disponibile del campo aleatorio e a questa sono stati applicati gli Algoritmi 4.6 e 4.7, valutando in particolare il variogramma sperimentale e le simulazioni bootstrap. Gli altri campioni sono invece stati usati per la valutazione della probabilità di copertura degli intervalli di confidenza costruiti tramite approssimazione MC-bootstrap.

La generazione dei parametri Φ_0 con il metodo illustrato, ottenuta con una simulazione non condizionata (Chilès e Delfiner, 1999), è stata svolta con il software geostatistico ISATIS®.

Si noti in particolare che, grazie alla normalità del campo $\Phi_0(s)$, il campo $\boldsymbol{\chi}_s$ risulta essere un processo gaussiano: per questo motivo, la scorrelatezza del campione $\widetilde{\chi}_{s_1}, ..., \widetilde{\chi}_{s_n}$ conseguente alla trasformazione (4.7) è equivalente all'indipendenza delle sue componenti.

Si precisa infine che la rappresentazione delle curve relative ai dataset presentati per lo studio di simulazione nel presente capitolo e nel Capitolo 5, avverrà visualizzando le ascisse lungo il semi-asse verticale negativo e le ordinate lungo l'asse orizzontale; il motivo di tale scelta, comune in ambito geofisico, è il significato della variabile indipendente, che corrisponde nello specifico alla profondità z alla quale è osservata la variabile dipendente $\chi_s(z)$.

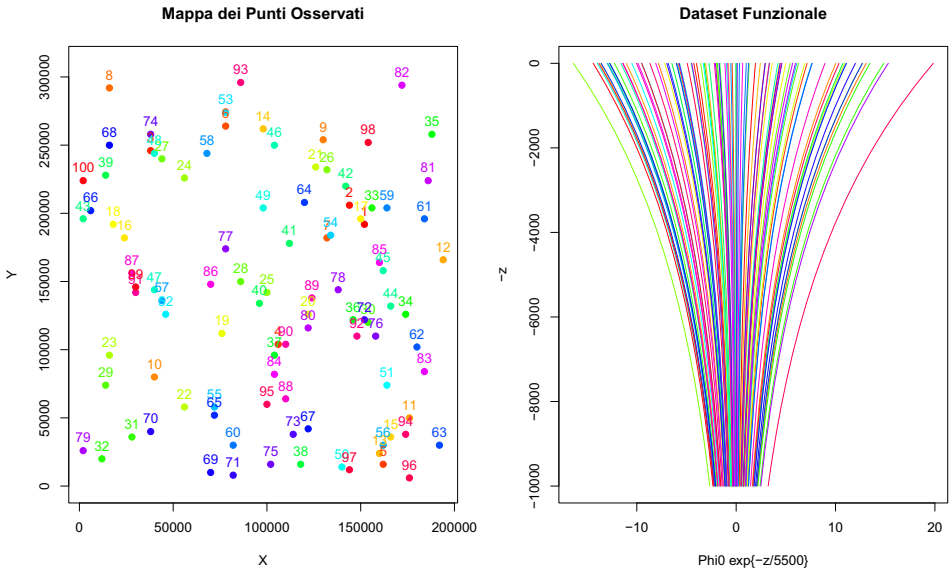

Figura 4.5: Dataset funzionale sintetico gaussiano, spazialmente stazionario. A sinistra: Mappa dei punti campionati $s_1, ..., s_{100}$. A destra: dataset funzionale sintetico $\chi_{\boldsymbol{s}} = (\chi_{s_1}, ..., \chi_{s_{100}})$. Per il pannello di destra, le ascisse sono rappresentate lungo il semi-asse verticale negativo in virtù del loro significato fisico (profondità) e l'asse delle ordinate lungo l'asse orizzontale.

4.3.1.2 Campione Non Gaussiano

Nella formulazione degli Algoritmi 4.6 e 4.7 è stato scelto di omettere le trasformazioni iniziali per rendere il campione gaussiano (indicate ad esempio negli algoritmi proposti, in ambito finito dimensionale, da Olea e Pardo-Igúzquiza (2011)). A supporto di questa scelta, motivata anche dalla difficoltà di definire una trasformazione di un campo funzionale generico in un campo gaussiano, sarà svolta un'analisi del comportamento degli algoritmi sviluppati su un secondo dataset funzionale, non gaussiano, ottenuto come segue.

Il dataset non gaussiano $\boldsymbol{\chi_s}$ è stato costruito generando opportunamente sette coefficienti da comporre con una base $\{B_j\}$ di B-spline cubiche ($m = 4$), fissando $n_k = 3$ nodi interni $z_k = \{2500; 5000; 7500\}$ m:

$$\boldsymbol{\chi}_s(z) = \sum_{j=1}^{m+n_k} \boldsymbol{\Phi}_j(s) B_j(z), \quad z \in \mathcal{T},\ s \in D. \tag{4.18}$$

Più precisamente, la generazione dei campi aleatori per i coefficienti Φ_j, $j = 1, ..., 7$, è avvenuta in due momenti:

1. generazione di sette campi gaussiani stazionari $\widetilde{\boldsymbol{\Phi}}_j$ con una simulazione non condizionata (Chilès e Delfiner, 1999), avvenuta con l'ausilio del software geostatistico ISATIS®;

2. trasformazione delle realizzazioni ottenute con opportune funzioni non lineari al fine di ottenere un campione non gaussiano, $\boldsymbol{\Phi}_j(s) = f(\widetilde{\boldsymbol{\Phi}}_j(s))$.

Le strutture variografiche di generazione dei campi gaussiani sono state di tipo sferico ed esponenziale, secondo i parametri di *sill*, *range* e *nugget* riportati in Tabella 4.2.

Una volta stabiliti i parametri dei variogrammi di generazione, il procedimento di simulazione è stato analogo al caso descritto nella sezione precedente: il campo è stato simulato 100 volte su una griglia rettangolare $D = [0, 200] \times [0, 300]$ km, con passo 2 km, campionando da essa in modo uniforme 100 localizzazioni $s_1, ..., s_{100}$ ed estraendo infine 100 realizzazioni $(\chi^i_{s_1}, ..., \chi^i_{s_{100}})$ del campione $(\boldsymbol{\chi}^i_{s_1}, ..., \boldsymbol{\chi}^i_{s_{100}})$, $i = 1, ..., 100$, presso i siti $s_1, ..., s_{100}$ individuati.

Di questi campioni, la prima realizzazione, $(\chi^1_{s_1}, ..., \chi^1_{s_{100}})$, mostrata in Figura 4.5, è stata usata come caso test per lo studio del comportamento degli Algoritmi 4.6 e 4.7, mentre le restanti realizzazioni sono state usate per accedere al variogramma di generazione attraverso una sua approssimazione Monte Carlo e per la valutazione delle probabilità di copertura degli intervalli di confidenza bootstrap.

Nella Sottosezione seguente saranno introdotte le misure di convergenza adottate nell'analisi; nelle Sottosezioni 4.3.3 e 4.3.4 saranno quindi illustrati i principali risultati ottenuti per simulazione rispettivamente per il campione gaussiano e il secondo campione, confrontando le stime determinate applicando l'Algoritmo 4.7 con il variogramma di riferimento e con le stime calcolate con il metodo OLS. Infine, saranno mostrati gli intervalli di confidenza puntuali per lo stimatore $\widehat{\gamma}(\boldsymbol{h})$ e ne sarà fornita una stima della probabilità di copertura.

	Struttura	*Sill*	(*Practical*) *Range*
Φ_1	Esponenziale	80	75 km
Φ_2	Sferica	80	75 km
Φ_3	Esponenziale	80	150 km
Φ_4	Sferica	80	150 km
Φ_5	Sferica	40	75 km
	Esponenziale	40	75 km
Φ_6	Sferica	40	75 km
	Esponenziale	40	75 km
Φ_7	Sferica	60	150 km
	Esponenziale	20	75 km

Tabella 4.2: Modelli di variogramma di generazione per i coefficienti Φ_j, $j = 1, ..., 7$, usati per la costruzione del processo stocastico funzionale stazionario non gaussiano. I coefficienti associati a più righe (Φ_5, Φ_6, Φ_7) sono stati generati dalla somma delle strutture variografiche indicate. L'effetto *nugget* è stato fissato a 0 per tutti i modelli variografici considerati.

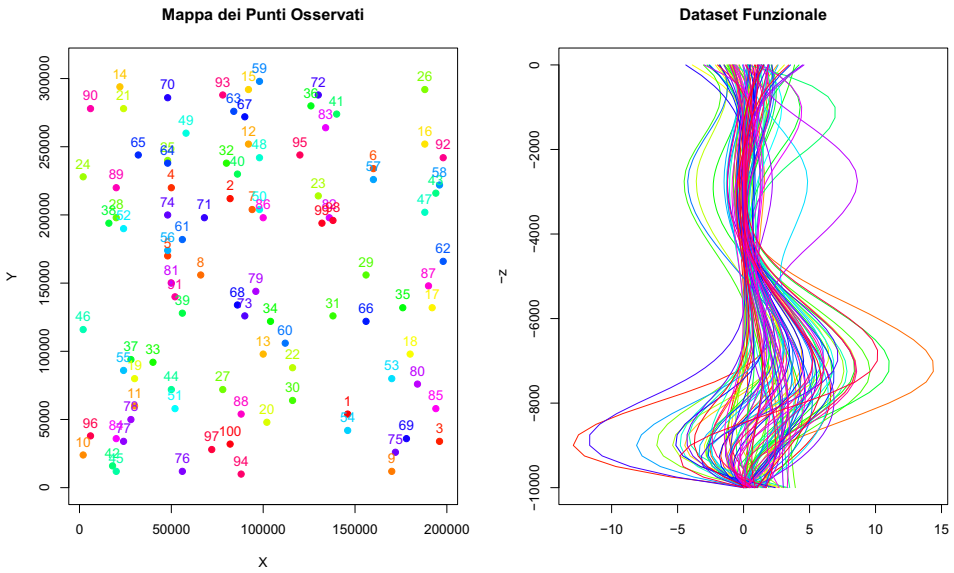

Figura 4.6: Dataset funzionale sintetico non gaussiano, spazialmente stazionario. A sinistra: Mappa dei punti campionati $s_1, ..., s_{100}$. A destra: dataset funzionale sintetico $\chi_{\boldsymbol{s}} = (\chi_{s_1}, ..., \chi_{s_{100}})$. Per il pannello di destra, le ascisse sono rappresentate lungo il semi-asse verticale negativo in virtù del loro significato fisico (profondità) e l'asse delle ordinate lungo l'asse orizzontale.

4.3.2 Misure di Convergenza

A partire dai dataset funzionali $\chi_{s_1}, ..., \chi_{s_n}$ generati come illustrato in Sottosezione 4.3.1 un primo obiettivo applicativo del lavoro di tesi è stata la valutazione tramite simulazione del comportamento dell'Algoritmo 4.7 in presenza di un campo esponenziale gaussiano (4.12) e di un campo esponenziale non gaussiano (4.18).

A questo scopo, sono state adottate opportune misure di distanza tra variogrammi, affiancandole alla visualizzazione grafica.

Infatti, in primo luogo è stata valutata la convergenza del metodo è attraverso la distanza assoluta in L^2 tra i modelli di variogramma stimati a iterazioni successive, cioè determinati da stime successive dei parametri $\boldsymbol{\vartheta}_N^{GLS}$. Tale distanza è inoltre stata normalizzata rispetto alle norme L^2, L^∞ (pari al *sill* nei modelli parametrici considerati) e alla varianza campionaria $\widehat{V}(\chi_s)$:

$$\begin{aligned}
d_{L^2}^{A,N} &= \|\gamma(\cdot, \boldsymbol{\vartheta}_N^{GLS}) - \gamma(\cdot, \boldsymbol{\vartheta}_{N-1}^{GLS})\|_{L^2}; \\
d_{L^2,L^2}^{R,N} &= \frac{\|\gamma(\cdot, \boldsymbol{\vartheta}_N^{GLS}) - \gamma(\cdot, \boldsymbol{\vartheta}_{N-1}^{GLS})\|_{L^2}}{\|\gamma(\cdot, \boldsymbol{\vartheta}_N^{GLS})\|_{L^2}}; \\
d_{L^2,L^\infty}^{R,N} &= \frac{\|\gamma(\cdot, \boldsymbol{\vartheta}_N^{GLS}) - \gamma(\cdot, \boldsymbol{\vartheta}_{N-1}^{GLS})\|_{L^2}}{\|\gamma(\cdot, \boldsymbol{\vartheta}_N^{GLS})\|_{L^\infty}}; \\
d_{L^2,var.}^{R,N} &= \frac{\|\gamma(\cdot, \boldsymbol{\vartheta}_N^{GLS}) - \gamma(\cdot, \boldsymbol{\vartheta}_{N-1}^{GLS})\|_{L^2}}{\widehat{V}(\chi_s)}.
\end{aligned}$$

Verificata la convergenza, l'approssimazione MC-bootstrap è stata quindi confrontata con la stima OLS e il variogramma di riferimento con la valutazione in questo caso delle seguenti misure:

$$\begin{aligned}
d_{L^2,L^2}^{OLS,N} &= \frac{\|\gamma(\cdot, \boldsymbol{\vartheta}_N^{GLS}) - \gamma(\cdot, \boldsymbol{\vartheta}^{OLS})\|_{L^2}}{\|\gamma(\cdot, \boldsymbol{\vartheta}_N^{GLS})\|_{L^2}}; \\
d_{L^2,L^2}^{ref,N} &= \frac{\|\gamma(\cdot, \boldsymbol{\vartheta}_N^{GLS}) - \gamma(\cdot, \boldsymbol{\vartheta}^{ref})\|_{L^2}}{\|\gamma(\cdot, \boldsymbol{\vartheta}_N^{GLS})\|_{L^2}};
\end{aligned}$$

dove $\boldsymbol{\vartheta}^{ref}$ è il vettore di parametri del modello variografico di generazione.

In particolare, l'attenzione è stata concentrata sulla determinazione dei parametri B (iterazioni bootstrap) e N_{max} (iterazioni esterne), considerando da un lato le approssimazioni MC-bootstrap ottenute per $B = \{1000; 2000; 3000; ...; 10000\}$ e $N_{max} = \{1; 2; 3; ...; 10\}$, dall'altro l'onere computazionale necessario ad ottenerle.

4.3.3 Stima MC-Bootstrap su Campione Gaussiano

Si consideri ora il campione gaussiano $\boldsymbol{\chi_s} = (\boldsymbol{\chi}_{s_1}, ..., \boldsymbol{\chi}_{s_n})$, dove:

$$\boldsymbol{\chi}_s(z) = \Phi_0(s) \exp\{-z/5500\}, \quad z \in [0, 10000], s \in D,$$

caratterizzato da una media $m(s)$ e un variogramma $\gamma(h)$, determinati dal campo Φ_0 (4.13), generato come descritto nella Sottosezione 4.3.1. Il variogramma di riferimento è, in questo caso:

$$\gamma(h) = \begin{cases} 80(1 - e^{-3h/75}) \| \exp\{- \cdot /5500\}\|^2, & h > 0 \\ 0, & h = 0, \end{cases} \tag{4.19}$$

con $\| \exp\{- \cdot /5500\}\|^2 \simeq 4607.237$ e dunque caratterizzato da un *sill* pari a $80 \cdot \| \exp\{- \cdot /5500\}\|^2 \simeq 184289.5$. Si fissi inoltre il modello esponenziale come modello parametrico di variogramma $\gamma(h, \boldsymbol{\vartheta})$ da adattare alla stima empirica, calcolata con un lag $(h_{k+1} - h_k)$ pari a 10 km e con $K = 9$.

Saranno ora mostrati i risultati di simulazione ottenuti attraverso l'Algoritmo 4.7, grazie all'implementazione riportata in Appendice B.3; l'ambiente di simulazione è stato `R` 2.13.1.

Tutte le simulazioni sono state svolte centrando le variabili osservate rispetto alla media campionaria, ovvero considerando il dataset $\chi'_{s_1}, ..., \chi'_{s_{100}}$, con:

$$\chi'_{s_j} = \chi_{s_i} - \overline{\chi}_{100},$$

dove $\overline{\chi}_{100}$ indica la media campionaria calcolata sul dataset $\chi_{s_1}, ..., \chi_{s_{100}}$. Si noti che il dataset centrato è caratterizzato dalla stessa struttura di covarianza del dataset originale; tuttavia, questa scelta è stata motivata dal fatto che gli algoritmi sono risultati più stabili considerando $\chi'_{s_1}, ..., \chi'_{s_{100}}$ rispetto al dataset originale.

Per fissare le idee, si consideri la Figura 4.7: nella parte sinistra, sono mostrati i variogrammi sperimentali calcolati a partire dai campioni bootstrap $\chi^\star_{\boldsymbol{s}}$ generati fissando $B = 1000$ e $N_{max} = 10$, sovrapposte al variogramma sperimentale calcolato dal campione originale e alla stima Monte Carlo del variogramma. Quest'ultimo è stato ottenuto come media dei *trace-variogram* stimati da ciascuna realizzazione $\chi^i_{s_1}, ..., \chi^i_{s_{100}}$, $i = 1, ..., 100$, e, come mostrato nel pannello di destra della Figura 4.7, risulta essere aderente al modello di variogramma di riferimento.

Dal fascio di stime sperimentali $\widehat{\gamma}^\star_b(\boldsymbol{h})$, $b = 1, ..., B$, la stima $\widehat{\Psi}$ della matrice di covarianza di $\widehat{\gamma}(\boldsymbol{h})$ è stata ottenuta attraverso lo stimatore empirico calcolato su $\widehat{\gamma}^\star_b(\boldsymbol{h})$, $b = 1, ..., B$, ovvero come:

$$\widehat{\Psi}^\star = \frac{1}{B-1} \sum_{b=1}^{B} [(\widehat{\gamma}^\star_b(\boldsymbol{h}) - \overline{\widehat{\gamma}^\star}_B(\boldsymbol{h}))(\widehat{\gamma}^\star_b(\boldsymbol{h}) - \overline{\widehat{\gamma}^\star}_B(\boldsymbol{h}))^T],$$

dove $\overline{\widehat{\gamma}^\star_B}(\boldsymbol{h})$ è la media campionaria del variogramma sperimentale calcolata sui campioni bootstrap come:

$$\overline{\widehat{\gamma}^\star}_B(\boldsymbol{h}) = \frac{1}{B} \sum_{b=1}^{B} \widehat{\gamma}^\star_b(\boldsymbol{h}).$$

4.3.3.1 Determinazione dei parametri B e N_{max}

Con il procedimento appena dettagliato, l'Algoritmo 4.7 è stato applicato al variare di $B \in \mathcal{B} = \{1000, 2000, 3000, ..., 10000\}$ e $N_{max} \in \{1, 2, ..., 10\}$.

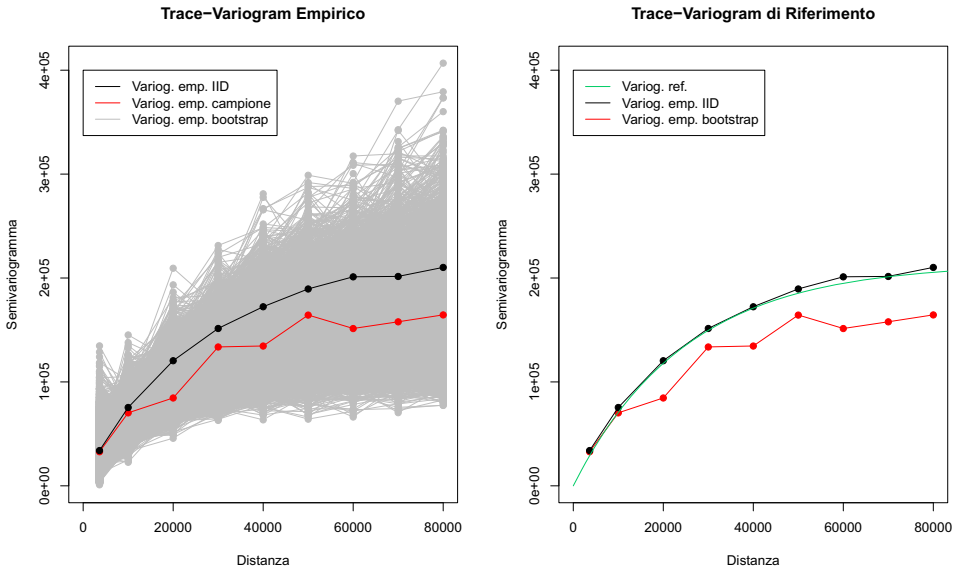

Figura 4.7: A sinistra: *Trace-variogram* dei campioni bootstrap (in grigio), *trace-variogram* del campione $\chi^1_{s_1}, ..., \chi^1_{s_{100}}$ (in rosso) e *trace-variogram* calcolato con il metodo Monte Carlo grazie alle realizzazioni $\chi^i_{s_1}, ..., \chi^i_{s_{100}}$, $i = 1, ..., 100$ (in nero). A destra: *Trace-variogram* del campione $\chi^1_{s_1}, ..., \chi^1_{s_{100}}$ (in rosso), *trace-variogram* calcolato con il metodo Monte Carlo (in nero) e variogramma di riferimento (in verde). In entrambi i pannelli h è espresso in metri.

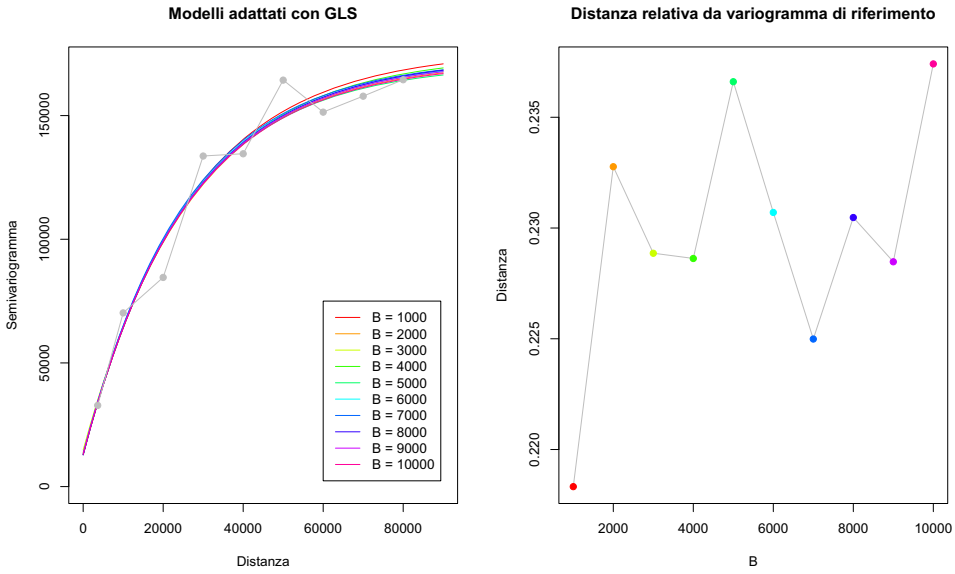

Figura 4.8: A sinistra: Modelli di variogramma $\gamma(h; \boldsymbol{\vartheta}_B^{GLS})$ addattati con il metodo GLS al variare di $B \in \mathcal{B}$, con h espresso in metri. A destra: Distanza tra il modello di variograma di riferimento e il modello $\gamma(h; \boldsymbol{\vartheta}_B^{GLS})$ stimato, per $B \in \mathcal{B}$.

In Figura 4.8, sono mostrati i risultati di convergenza ottenuti al variare del numero B di iterazioni bootstrap. In particolare, nel pannello di sinistra sono raffigurati i modelli di variogramma $\gamma(h; \boldsymbol{\vartheta}_B^{GLS})$, $B \in \mathcal{B}$, mentre nella parte destra è mostrata la distanza $d_{L^2,L^2}^{ref,N}$, $N = 1, ..., N_{max}$, tra il variogramma di riferimento e il variogramma adattato con il metodo proposto, rappresentato in funzione del numero di iterazioni.

Come si può notare, i modelli $\gamma(h; \boldsymbol{\vartheta}_B^{GLS})$ così stimati non mostrano differenze sostanziali al variare del numero di iterazioni bootstrap.

In aggiunta, la distanza $d_{L^2,L^2}^{ref,N}$ non evidenzia un andamento monotono al crescere del parametro B, mantenendosi attorno a 0.23: questo valore, che a prima vista può sembrare alto, è dovuto principalmente allo scostamento esistente tra il variogramma di riferimento e il variogramma sperimentale dal campione $\chi_{s_1}, ..., \chi_{s_{100}}$, rispetto al quale è valutato l'adattamento ottimo.

La valutazione dell'approssimazione MC-bootstrap al variare di B può essere condotta a partire dalla Figura 4.9, nella quale è raffigurato lo scostamento tra la stima MC-bootstrap $\widehat{\Psi}^{\star}$ e la stima Monte Carlo $\widehat{\Psi}^{MC}$, normalizzato rispetto al raggio spettrale di $\widehat{\Psi}^{MC}$:

$$d^{\star,MC} = \frac{\widehat{\Psi}^{\star} - \widehat{\Psi}^{MC}}{\rho(\widehat{\Psi}^{MC})}. \tag{4.20}$$

Dal grafico si può vedere che lo scostamento $d^{\star,MC}$ si mantiene entro 0.13, assumendo valori di un ordine di grandezza inferiore in corrispondenza dei primi lag, e non evidenzia significative

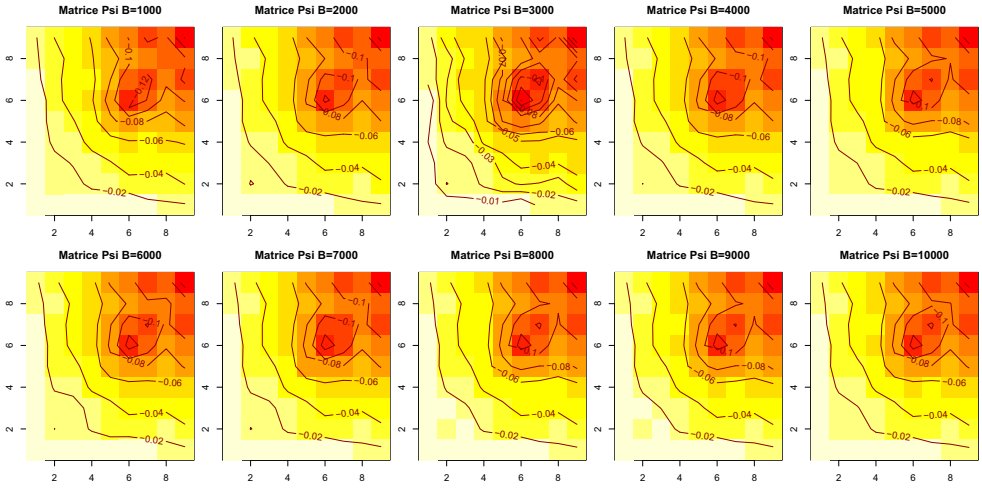

Figura 4.9: Andamento dello scostamento tra le stime $\widehat{\Psi}^{\star}$ e $\widehat{\Psi}^{MC}$, valutato attraverso la misura $d^{\star,MC}$.

differenze al variare di $B \in \mathcal{B}$.

Sebbene lo scostamento relativo rimanga limitato per ogni $B \in \mathcal{B}$, la differenza assoluta tra le due matrici $\widehat{\Psi}^{\star} - \widehat{\Psi}^{MC}$ non è trascurabile, essendo in qualche caso dell'ordine di grandezza delle componenti di $\widehat{\Psi}$. Una possibile spiegazione di questo potrebbe risiedere nel fatto che l'alta dimensionalità dei dati potrebbe ripercuotersi sulla velocità di convergenza della distribuzione empirica F_n alla distribuzione vera F, rendendo l'approssimazione bootstrap meno affidabile.

Dalle simulazioni è risultato inoltre che il numero di iterazioni Monte Carlo influenza il comportamento del metodo rispetto al numero di iterazioni esterne. Nel pannello a sinistra della Figura 4.10 è infatti rappresentato l'andamento delle distanze tra i variogrammi stimati a iterazioni successive, $d^{R,N}_{L^2,L^2}$, $N = 1, ..., N_{max}$, al crescere del numero di iterazioni esterne ($N_{max} \in \{1, 2, ..., 10\}$) e al variare delle iterazioni bootstrap ($B \in \mathcal{B}$). Come si può notare, l'ampiezza di tali oscillazioni diminuisce al crescere del valore di B, mantenendosi al di sotto del 2% a partire dalla seconda iterazione esterna per $B > 2000$ e in tutti i casi al di sotto del 5%.

Risultati in accordo con questi sono mostrati nel pannello di destra della Figura 4.10, dove è rappresentato l'andamento della distanza $d^{OLS,N}_{L^2,L^2}$, $N = 1, ..., N_{max}$ al variare dei parametri B e N_{max}. I valori di $d^{OLS,N}_{L^2,L^2}$, $N = 1, ..., N_{max}$ si assestano infatti attorno a 0.012 entro la seconda iterazione, mentre le oscillazioni attorno al valor medio si riducono al crescere di $B \in \mathcal{B}$. Tali oscillazioni, che sono sempre presenti nelle simulazioni effettuate, sono probabilmente dovute all'approssimazione Monte-Carlo della distribuzione bootstrap; tuttavia esse non sembrano influenzare in modo sostanziale i parametri del modello di variogramma stimati, rappresentati in Figura 4.11.

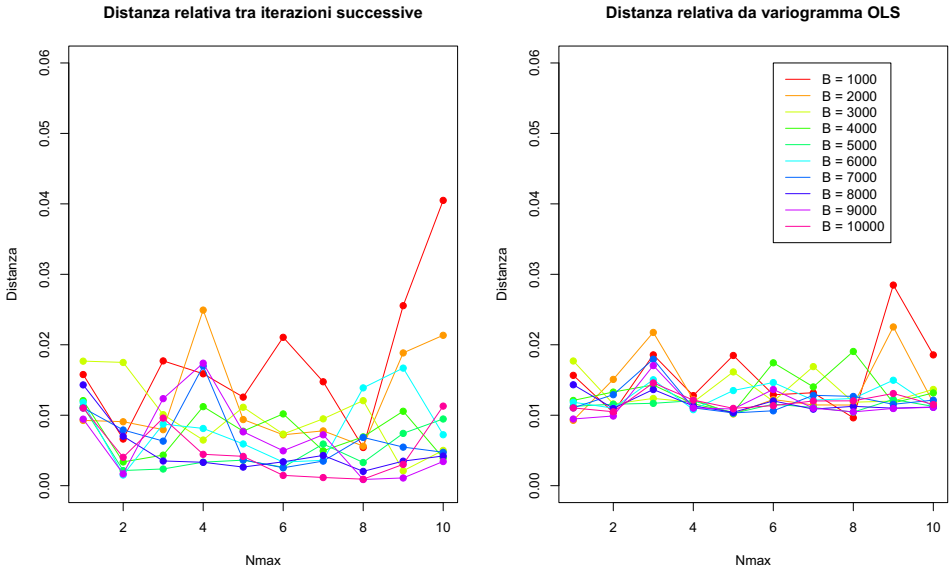

Figura 4.10: Convergenza del metodo MC-bootstrap per campione gaussiano. A sinistra: Rappresentazione della distanza $d^{R,N}_{L^2,L^2}$, $N = 1, ..., N_{max}$, al variare di $N_{max} \in \{1, 2, ..., 10\}$ e $B \in \mathcal{B}$. A destra: Andamento della distanza $d^{OLS,N}_{L^2,L^2}$, $N = \{1; ...; N_{max}\}$, al variare di $N_{max} \in \{1, 2, ..., 10\}$ e $B \in \mathcal{B}$. In entrambi i pannelli, in ascissa è rappresentato il numero di iterazioni esterne, in ordinata il corrisponente valore di distanza.

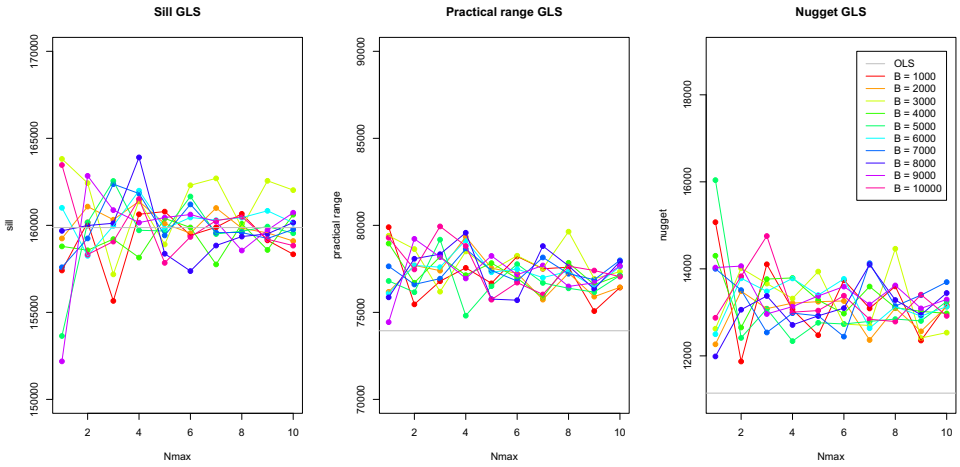

Figura 4.11: Parametri del modello di variogramma stimati con metodo MC-bootstrap per campione gaussiano. I parametri ϑ^{GLS}_B, $B \in \mathcal{B}$ (linee colorate), sono confrontati con la stima OLS (in grigio). In ascissa è rappresentato il numero di iterazioni esterne, in ordinata il valore del parametro; i colori distinguono il numero di iterazioni Monte Carlo.

Alla luce delle considerazioni precedenti, è stato deciso di fissare il numero di iterazioni esterne come $N_{max} = 2$, in quanto tale valore è risultato sufficiente per la perdita della dipendenza dalla stima OLS iniziale. Per quanto riguarda il parametro B, esso è invece stato fissato a $B = 1000$ o $B = 5000$ nel caso in cui fossero desiderate rispettivamente efficienza computazionale o maggiore precisione della stima.

4.3.3.2 Probabilità di Copertura degli Intervalli di Confidenza Calcolati con il Metodo dei Percentili

Uno dei vantaggi dell'uso dei metodi bootstrap è la possibilità di avere una stima dell'intera distribuzione della statistica T di interesse che, nel caso in esame, corrisponde allo stimatore $\widehat{\gamma}(\boldsymbol{h})$. Tale distribuzione, usata finora per la stima GLS di un modello di variogramma, sarà ora impiegata nella costruzione di intervalli di confidenza con il metodo dei percentili, introdotto in precedenza.

Si consideri ancora una volta il vettore aleatorio di stime empiriche:

$$\widehat{\gamma}(\boldsymbol{h}) = (\widehat{\gamma}(h_1), ..., \widehat{\gamma}(h_K));$$

si denoti con $\mathcal{L}(\widehat{\gamma}(\boldsymbol{h}))$ la legge congiunta di $\widehat{\gamma}(\boldsymbol{h})$ e con $\mathcal{L}(\widehat{\gamma}(h_k))$ la legge marginale della componente l dello stimatore sperimentale, per $k = 1, ..., K$.

Si assuma ora che la legge marginale $\mathcal{L}(\widehat{\gamma}(h_k))$, ammetta una densità $f_{\widehat{\gamma}(h_k)}$ rispetto alla misura di Lebesgue su $\mathbb{R}$, per ogni $k = 1, ..., K$.

Le densità $f_{\widehat{\gamma}(h_k)}$, $k = 1, ..., K$, possono essere stimate, in accordo con l'Algoritmo 4.6, tramite l'istogramma delle statistiche ottenute dai campioni bootstrap e i percentili, utili ai fini della costruzione degli intervalli di confidenza, possono essere ottenuti da questo con i relativi stimatori campionari. In particolare, fissato il livello di confidenza a $1-\alpha = 0.90$, gli intervalli di confidenza saranno costruiti tramite i quantili simmetrici $\gamma_{0.05}(h_k)$, $\gamma_{0.95}(h_k)$:

$$IC^{\star k}_{0.90}(\gamma(h_k)) = [\gamma_{0.05}(h_k), \gamma_{0.95}(h_k)], \quad k = 1, ..., K.$$

In Figura 4.12, sono rappresentate le distribuzioni marginali ottenute applicando l'Algoritmo 4.7, per $N_{max} = 2$ e $B = 5000$, fissando il lag, come in precedenza, a 10 km e $K = 9$. I quantili 0.05 e 0.95, indicati con le linee rosse verticali, sono stati usati per la costruzione degli intervalli di confidenza riportati nel pannello in alto a sinistra della Figura 4.12.

Gli stessi intervalli di confidenza sono rappresentati anche in Figura 4.13, nel pannello di sinistra, sovrapposti al modello di variogramma (4.19). Dal grafico si nota che l'ampiezza degli intervalli cresce al crescere della distanza tra le osservazioni, indicando una maggiore incertezza legata alla stima empirica per lag grandi, dove si registra il massimo scostamento della *binned trace-variogram* dal modello di riferimento: la crescita dell'ampiezza degli intervalli di confidenza consente dunque alla stima intervallare di contenere il variogramma di riferimento per tutti i lag di osservazione.

Nel pannello di destra della Figura 4.13 sono invece riportati i risultati ottenuti con il medesimo procedimento, fissando come modello parametrico il modello sferico. Come si

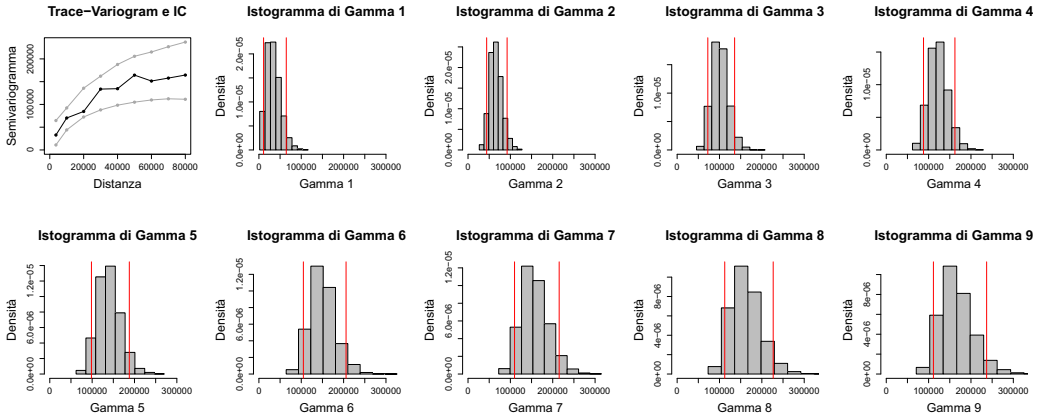

Figura 4.12: Intervalli di confidenza puntuali per $\widehat{\gamma}(h_k)$, $k = 1, ..., 9$ (in alto a sinistra) stimati con il metodo MC-bootstrap, a partire dalla stima delle distribuzioni marginali $f_{\widehat{\gamma}(h_k)}$, $k = 1, ..., 9$ (istogrammi riportati nei rimanenti pannelli).

nota dal confronto dei due pannelli, la forma della banda di confidenza varia al variare del modello parametrico fissato: questo accade perché il modello di variogramma influenza il modo in cui le componenti del campione sono dapprima scorrelate, quindi ricorrelate a seguito del campionamento (attraverso la matrice $\widehat{\Sigma}$) e, di conseguenza, influisce sulla stima della distribuzione dello stimatore empirico. Si noti tuttavia che, malgrado la forma del modello di variogramma non corrisponda, nel secondo caso, alla struttura del modello di riferimento, la banda di confidenza calcolata contiene ugualmente il variogramma (4.19).

Da una prima analisi, gli intervalli di confidenza costruiti con il metodo MC-bootstrap sembrano dunque fornire risultati interessanti. Al fine di usare tali intervalli nelle analisi successive, è stato deciso di stimare, tramite simulazione, la probabilità di copertura di tali intervalli, a fronte di un livello nominale del 90%.

La stima della probabilità di copertura degli stimatori intervallari $IC^{\star k}_{0.90}(\gamma(h_k))$, è stata condotta come segue.

Per ogni realizzazione $\chi^i_{\boldsymbol{s}} = (\chi^i_{s_1}, ..., \chi^i_{s_{100}})$, $i = 1, ..., 30$, sono stati costruiti gli intervalli di confidenza puntuali, $IC^{\star k}_{0.90}(\gamma(h_k); \chi^i_{\boldsymbol{s}})$, $k = 1, ..., K$, con il metodo precedentemente illustrato. Per ciascuno di essi è stata verificata la copertura, ponendo:

$$c^i_k = \begin{cases} 0, & \gamma(h_k) \notin IC^{\star k}_{0.90}(\gamma(h_k); \chi^i_{\boldsymbol{s}}) \\ 1, & \gamma(h_k) \in IC^{\star k}_{0.90}(\gamma(h_k); \chi^i_{\boldsymbol{s}}), \end{cases} \tag{4.21}$$

per $i = 1, ..., 30$, $k = 1, ..., K$. Per ogni valore di k fissato, è stato quindi calcolato il valor medio $\overline{c}_k$ di c^i_k, ed è infine stato considerato $\overline{c}_k$, $k = 1, ..., K$, come una stima della probabilità di copertura del relativo stimatore intervallare.

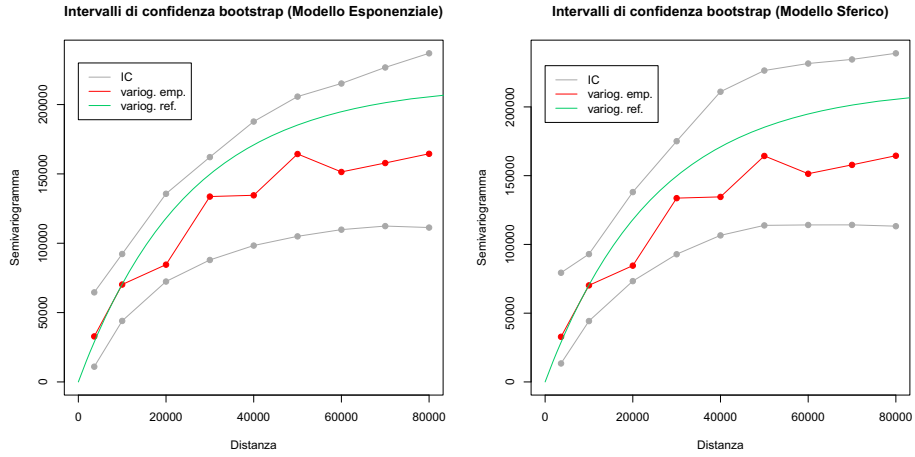

Figura 4.13: Intervalli di confidenza stimati per il variogramma empirico scegliendo il modello esponenziale (a sinistra) e il modello sferico (a destra). In verde è riportato il variogramma che ha generato i dati; h è espresso in metri.

Il procedimento descritto è stato ripetuto per quattro scelte del lag, ovvero $\{8; 10; 13; 17\}$ km, fissando come distanza massima 100 km (ovvero $K = \{13; 10; 8; 6\}$). I risultati ottenuti sono riportati in Figura 4.14 e in Tabella 4.3.

Dalla Tabella 4.3, così come dall'osservazione della Figura 4.14, si nota che gli stimatori intervallari si dimostrano in tutti i casi conservativi evidenziando un livello di confidenza stimato mediamente pari a 0.93, a fronte di un livello nominale di 0.9.

In Tabella 4.4 sono invece riportate le stime della probabilità di copertura degli intervalli di confidenza simultanei, ricavati a partire dagli intervalli puntuali $IC^{\star k}_{0.90}$, $k = 1, ..., K$ al variare del lag in $\{8; 10; 13; 17\}$ km. Si noti in particolare che con l'espressione 'intervalli di confidenza simultanei' si indicano, in questo caso, gli intervalli simultanei per le componenti $\widehat{\gamma}(h_k)$ dello stimatore *binned trace-variogram* $\widehat{\gamma}(\boldsymbol{h})$ e non la banda di confidenza per l'intera curva $\gamma(h)$.

Le stime delle suddette probabilità di copertura sono state ottenute calcolando il numero medio di realizzazioni $\chi^i_{\boldsymbol{s}}$ per le quali il variogramma di generazione fosse contenuto nella banda costruita per ogni lag, ovvero ponendo:

$$c^i_{(sim)} = \begin{cases} 0, & \exists\, k : \gamma(h_k) \notin IC^{\star k}_{0.90}(\gamma(h_k); \chi^i_{\boldsymbol{s}}) \\ 1, & \gamma(h_k) \in IC^{\star k}_{0.90}(\gamma(h_k); \chi^i_{\boldsymbol{s}}), \quad \forall k \in \{1, 2, ..., K\}, \end{cases} \tag{4.22}$$

e considerando come stima della probabilità di copertura simultanea:

$$\overline{c}_{(sim)} = \frac{1}{30} \sum_{i=1}^{30} c^i_{(sim)}.$$

	$\overline{c}_{\cdot k}$ lag=8 km	$\overline{c}_{\cdot k}$ lag=10 km	$\overline{c}_{\cdot k}$ lag=13 km	$\overline{c}_{\cdot k}$ lag=17 km
$IC^{\star 1}_{0.90}$	1.000	0.967	1.000	1.000
$IC^{\star 2}_{0.90}$	1.000	0.967	1.000	0.900
$IC^{\star 3}_{0.90}$	0.967	0.900	0.933	0.933
$IC^{\star 4}_{0.90}$	0.933	0.933	0.933	0.933
$IC^{\star 5}_{0.90}$	0.933	0.933	0.933	0.933
$IC^{\star 6}_{0.90}$	0.933	0.933	0.933	0.967
$IC^{\star 7}_{0.90}$	0.933	0.933	0.933	-
$IC^{\star 8}_{0.90}$	0.933	0.933	0.933	-
$IC^{\star 9}_{0.90}$	0.933	0.967	-	-
$IC^{\star 10}_{0.90}$	0.933	0.967	-	-
$IC^{\star 11}_{0.90}$	0.933	-	-	-
$IC^{\star 12}_{0.90}$	0.967	-	-	-
$IC^{\star 13}_{0.90}$	0.967	-	-	-

Tabella 4.3: Probabilità di copertura degli intervalli di confidenza puntuali stimata per simulazione, al variare del lag.

	$\overline{c}_{\cdot (sim)}$ lag=8 km	$\overline{c}_{\cdot (sim)}$ lag=10 km	$\overline{c}_{\cdot (sim)}$ lag=13 km	$\overline{c}_{\cdot (sim)}$ lag=17 km
$IC^{\star (sim)}_{0.90}$	0.933	0.867	0.933	0.867

Tabella 4.4: Probabilità di copertura degli intervalli di confidenza simultanei stimata per simulazione, al variare del lag.

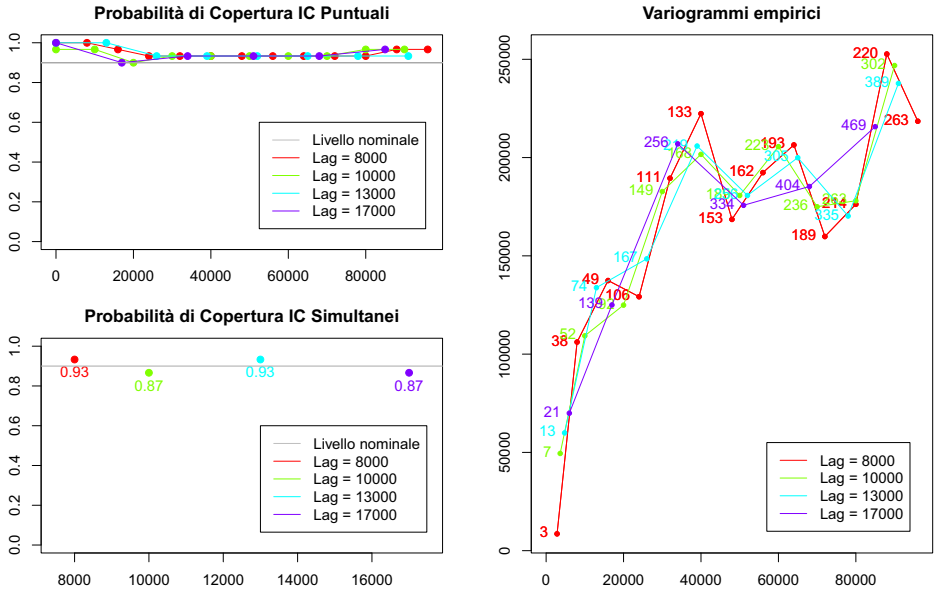

Figura 4.14: A sinistra: Stima della probabilità di copertura degli intervalli di confidenza puntuali (in alto) e simultanei (in basso), costruiti con il metodo dei percentili. In ascissa è riportata la distanza h_k, $k = 1, ..., K$ espressa in metri, in ordinata è invece riportata la stima $\overline{c}_k$ ($\overline{c}_{(sim)}$) della probabilità di copertura del corrispondente stimatore intervallare $IC^{\star k}_{0.90}$ ($IC^{\star (sim)}_{0.90}$). La linea grigia corrisponde al livello di confidenza nominale ($1-\alpha = 0.9$). A destra: Stimatori empirici, calcolati al variare del lag in $\{8000; 10000; 13000; 17000\}$ m; per ogni valore della distanza h_k (in metri), è indicato il numero $|N(h_k)|$ di coppie appartenenti alla classe $N(h_k)$, per $k = 1, ..., K$. In tutti i pannelli i colori distinguono il valore fissato per il lag.

Non essendo nota a priori la probabilità di copertura degli intervalli di confidenza bootstrap $IC^{\star k}_{1-\alpha}$ ottenuti con il metodo dei percentili, si è deciso di costruire la banda di confidenza simultanea di livello 0.90 a partire da $IC^{\star k}_{0.90}$, $k = 1, ..., K$, senza applicare correzioni di Bonferroni.

Le stime della probabilità di copertura simultanea ottenute con il metodo introdotto, riportate in Tabella 4.3, non si discostano in modo significativo dal livello di confidenza nominale del 90%, assestandosi su 0.87 o 0.93 a seconda del lag. Gli stimatori intervallari simultanei, a differenza degli intervalli di confidenza puntuali, non evidenziano dunque un comportamento conservativo. Si precisa, infine, che le probabilità di copertura stimate sono dipendenti dal numero di lag considerati ($K = \{13; 10; 8; 6\}$), e non sono dunque valide come stime dalla probabilità di copertura per l'intera curva.

4.3.4 Stima MC-Bootstrap su Campione Non Gaussiano

4.3.4.1 Metodo MC-Bootstrap per l'Adattamento di un Modello Valido

Al fine di verificare l'influenza della distribuzione dei dati sul comportamento degli Algoritmi 4.6 e 4.7, è stata ripetuta l'analisi svolta nella Sottosezione 4.3.3 a partire dalla prima realizzazione $\chi_{\boldsymbol{s}} = \chi^1_{\boldsymbol{s}}$ del processo non gaussiano (Figura 4.6), generato con il metodo illustrato nella Sottosottosezione 4.3.1.2.

In particolare, la valutazione della metodologia proposta per l'adattamento di un modello valido di variogramma alla stima empirica è avvenuta in termini di velocità di convergenza dell'Algoritmo 4.7 e di qualità della stima fornita.

L'Algoritmo 4.7 è stato quindi applicato fissando il numero di iterazioni bootstrap a $B = 5000$, il lag a 10 km e il parametro $K = 9$, e analizzando il comportamento del metodo al variare delle iterazioni esterne. La convergenza del metodo è stata quindi studiata attraverso le distanze $d^{R,N}_{L^2,L^2}$ e $d^{OLS,N}_{L^2,L^2}$ definite nella Sottosezione 4.3.2; per la valutazione delle approssimazioni MC-bootstrap è invece stata considerata la stima Monte Carlo ottenuta grazie alle realizzazioni $\chi^i_{s_1}, ..., \chi^i_{s_100}$, $i = 1, ..., 100$.

Infatti, dal momento che la costruzione del dataset non gaussiano ha richiesto l'introduzione di trasformazioni non lineari nei coefficienti $\widetilde{\Phi}(s)$, per il campione in esame non è nota l'espressione esplicita del variogramma di generazione. Per questo motivo, è stato considerato come variogramma di riferimento l'approssimazione Monte Carlo della stima sperimentale, calcolata come media degli stimatori empirici ottenuti dalle 100 realizzazioni $\chi^i_{\boldsymbol{s}}$, $i = 1, ..., 100$ (Figura 4.15, a destra), stima che si è dimostrata essere aderente al variogramma di generazione nello studio di Sezione 4.3.3 (Figura 4.7).

Per quanto concerne le misure di convergenza, nel pannello di sinistra di Figura 4.16 sono raffigurate le distanze relative tra i modelli stimati a iterazioni successive, $d^{R,N}_{L^2,L^2}$, e le distanze relativa dal variogramma stimato alla prima iterazione (stima OLS), $d^{OLS,N}_{L^2,L^2}$, al variare del numero di iterazioni esterne N_{max}. Come si nota, a differenza del comportamento del metodo nel caso gaussiano, la stima MC-bootstrap non presenta in questo caso oscillazioni

Figura 4.15: A sinistra: *Trace-variogram* dei campioni bootstrap (in grigio), *trace-variogram* del campione $\chi^1_{s_1}, ..., \chi^1_{s_{100}}$ (in rosso) e *trace-variogram* calcolato con il metodo Monte Carlo da $\chi^i_{s_1}, ..., \chi^i_{s_{100}}$, $i = 1, ..., 100$ (in nero). A destra: *Trace-variogram* del campione $\chi^1_{s_1}, ..., \chi^1_{s_{100}}$ (in rosso) e *trace-variogram* calcolato con il metodo Monte Carlo (in verde). In entrambi i pannelli h è espresso in metri.

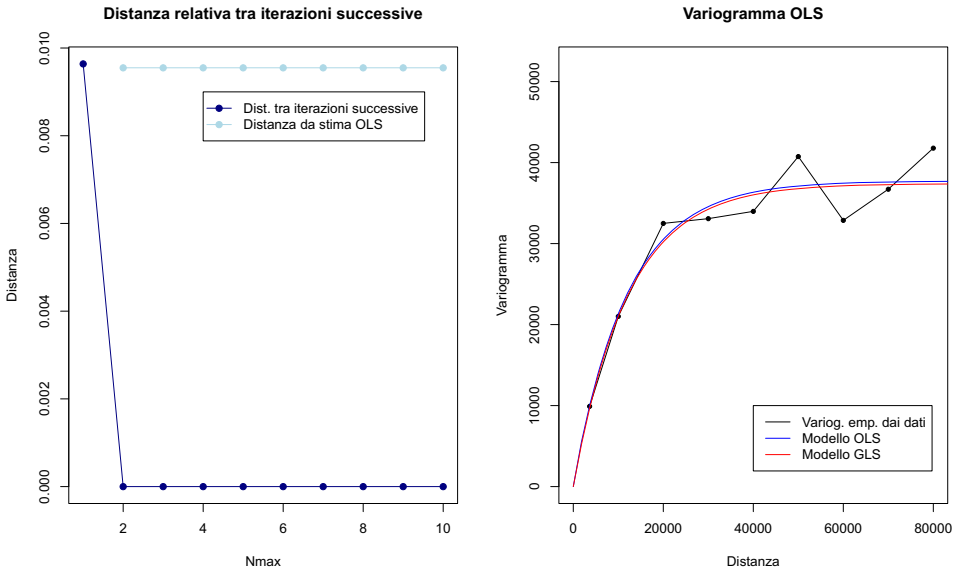

Figura 4.16: A sinistra: Rappresentazione della distanza relativa tra iterazioni successive $d^{R,N}_{L^2,L^2}$, $N = 1, ..., N_{max}$, al variare di $N_{max} \in \{1, 2, ..., 10\}$ (in blu) e della distanza dalla stima iniziale $d^{OLS,N}_{L^2,L^2}$, $N = 1, ..., N_{max}$, al variare di $N_{max} \in \{1, 2, ..., 10\}$ (in azzurro). In ascissa è rappresentato il numero di iterazioni esterne, in ordinata il corrispondente valore di distanza. A destra: Stima empirica del variogramma sovrapposta al modello stimato con OLS e modello stimato con MC-bootstrap al termine della procedura di stima.

significative a partire dalla seconda iterazione, assestandosi su una stima a distanza relativa 0.1 dalla stima OLS.

Risultati in accordo con i precedenti si rilevano dalla Figura 4.17; infatti, dai grafici si nota che la stima dei parametri non evidenzia oscillazioni significative al crescere di N_{max}.

Dalle considerazioni precedenti si può dunque concludere che l'Algoritmo 4.7 applicato al campione non gaussiano si dimostra convergente entro due iterazioni esterne.

Nel pannello di destra di Figura 4.16 sono invece mostrati, a confronto, la stima OLS del modello parametrico di variogramma e la corrispondente stima GLS ottenuta con il metodo MC-bootstrap.

Da tale immagine, così come dalla Figura 4.17, si nota che la principale differenza ottenuta dall'applicazione dei due metodi risiede nella stima dell'effetto *nugget*: infatti, se i parametri *sill* e *range* stimati coincidono fino alla seconda cifra decimale per i metodi OLS e GLS, la differenza della stima del *nugget* è invece di circa 300 unità. Dal momento che il comportamento del modello di variogramma in prossimità dell'origine e, in particolare, l'effetto *nugget*, influenzano notevolmente le previsioni di kriging, l'accuratezza della stima di tale parametro è molto importante ai fini predittivi. Relativamente a questo parametro,

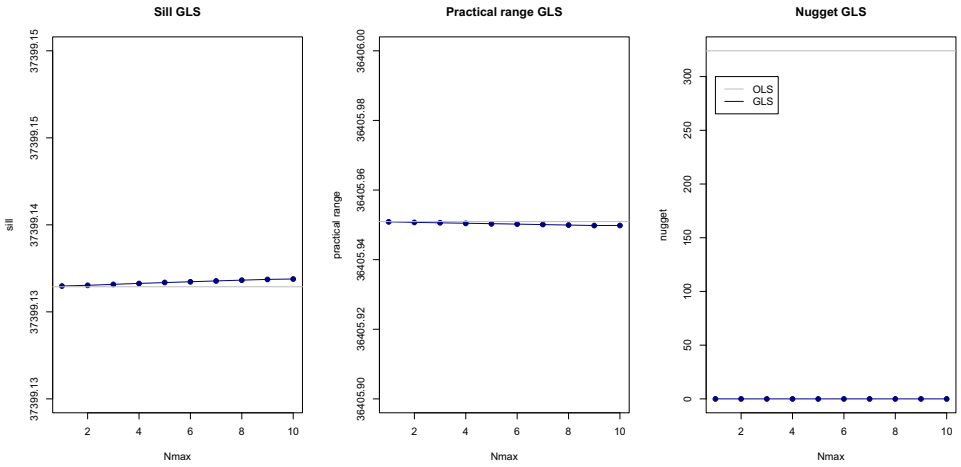

Figura 4.17: Parametri del modello di variogramma stimati con metodo MC-bootstrap per campione non gaussiano. I parametri $\boldsymbol{\vartheta}^{GLS}$, (linea blu), sono confrontati con la stima OLS (in grigio). In ascissa è rappresentato il numero di iterazioni esterne, in ordinata il valore del parametro.

la stima GLS si accosta maggiormente al variogramma di generazione, dal momento che tutti i parametri $\widetilde{\Phi}_j(s)$, $j = 1, ..., 7$, sono stati generati da un modello privo di *nugget* e l'applicazione di trasformazioni non lineari non introduce discontinuità nel processo.

Una valutazione della bontà dell'approssimazione ottenuta con il metodo MC-bootstrap può essere fornita considerando la distanza $d^{\star,MC}$ definita dalla (4.20), la cui rappresentazione è riportata in Figura 4.18. Dal grafico si può notare che lo scarto relativo registrato si mantiene entro il valore di 0.2, denotando, in accordo con il caso gaussiano, una stima molto precisa in corrispondenza dei lag più bassi.

L'approssimazione MC-bootstrap risulta pertanto accettabile anche nel caso non gaussiano fornendo risultati coerenti con i risultati ottenuti dalle analisi illustrate nella Sottosezione 4.3.3

4.3.4.2 Probabilità di Copertura degli Intervalli di Confidenza Calcolati con il Metodo dei Percentili

Una volta analizzato il comportamento dell'Algoritmo 4.7 in termini di convergenza e qualità dell'approssimazione MC-bootstrap in presenza di un campo non gaussiano, le distribuzioni del variogramma empirico ottenute con la metodologia proposta sono state usate per la costruzione di intervalli di confidenza per lo stimatore sperimentale.

Gli stimatori intervallari sono stati quindi determinati con il metodo dei percentili, adottando il medesimo procedimento usato nelle analisi illustrate in Sottosottosezione 4.3.3.2.

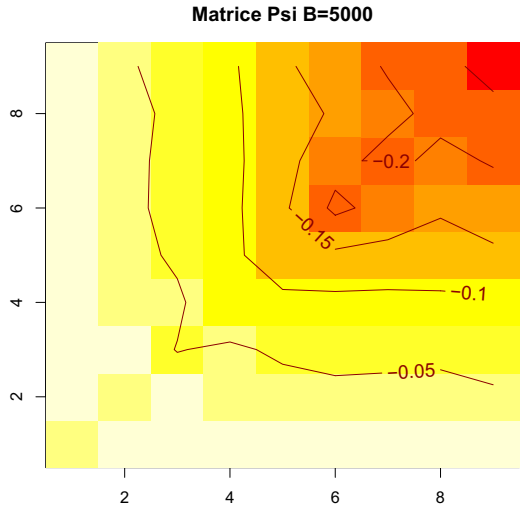

Figura 4.18: A sinistra: Scostamento tra le stime $\widehat{\Psi}^{\star}$ e $\widehat{\Psi}^{MC}$, valutato attraverso la misura $d^{\star,MC}$.

Le distribuzioni marginali ottenute applicando l'Algoritmo 4.7, per $N_{max} = 2$, $B = 5000$ e fissando il lag a 10 km con $K = 9$ sono rappresentate in Figura 4.19. In particolare, la costruzione degli intervalli di confidenza riportati nel pannello in alto a sinistra della Figura 4.19 è stata basata sui quantili 0.05 e 0.95, indicati con le linee rosse verticali.

Il livello di confidenza degli intervalli MC-bootstrap puntuali e simultanei stimati è quindi stato valutato attraverso l'uso delle realizzazioni $\chi^i_{\boldsymbol{s}}$, $i = 1, ..., 30$: per ogni campione $\chi^i_{\boldsymbol{s}}$ è stato verificata la copertura puntuale e simultanea degli intervalli stimati $IC^{\star k,i}_{0.90}$ e $IC^{\star (sim),i}_{0.90}$, determinando i valori delle statistiche c^i_k e $c^i_{(sim)}$, definiti dalle (4.21) e (4.22).

I valori medi $\overline{c}_k$ e $\overline{c}_{(sim)}$ delle statistiche c^i_k e $c^i_{(sim)}$, che costituiscono la stima empirica del livello di copertura degli intervalli costruiti, sono riportati in Tabella 4.5 e 4.6.

Dall'osservazione dei valori riportati in Tabella 4.5, così come dalla visualizzazione dei grafici in Figura 4.20 si nota che, a differenza del caso gaussiano, gli intervalli di confidenza puntuali costruiti dai campioni non gaussiani non si dimostrano conservativi: la stima della probabilità di copertura è infatti molto vicina al livello nominale del 90%.

Inoltre, gli intervalli di confidenza simultanei evidenziano una probabilità di copertura mediamente inferiore del 20% circa rispetto al livello nominale, suggerendo che l'uso di tali intervalli dovrebbe essere accompagnato da una opportuna correzione di Bonferroni.

4.4 Conclusioni e Sviluppi Futuri

In questo Capitolo è stata proposta una metodologia di ricampionamento finalizzata alla stima della distribuzione dello stimatore sperimentale del variogramma per un processo

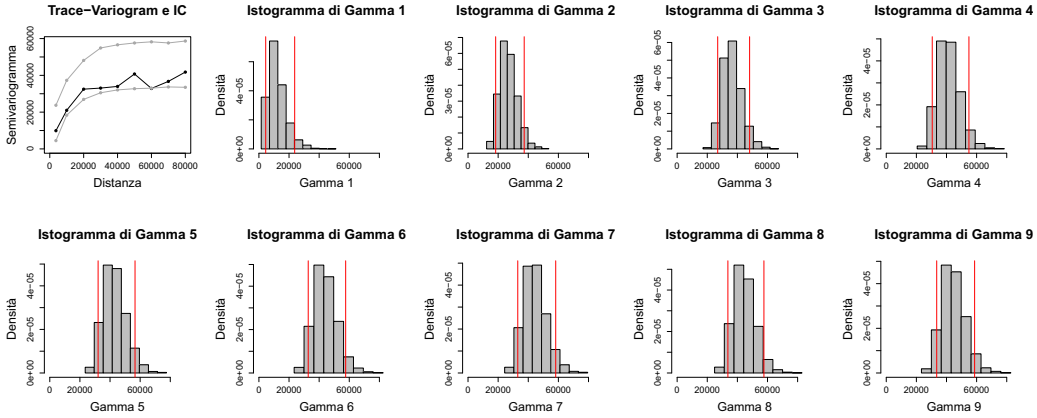

Figura 4.19: Intervalli di confidenza puntuali per $\widehat{\gamma}(h_k)$, $k = 1, ..., 9$ (in alto a sinistra) stimati con il metodo MC-bootstrap, a partire dalla stima delle distribuzioni marginali $f_{\widehat{\gamma}(h_k)}$, $k = 1, ..., 9$ (istogrammi riportati nei rimanenti pannelli).

	$\overline{c}_k$ lag=8 km	$\overline{c}_k$ lag=10 km	$\overline{c}_k$ lag=13 km	$\overline{c}_k$ lag=17 km
$IC^{\star 1}_{0.90}$	1.000	0.833	0.933	0.933
$IC^{\star 2}_{0.90}$	1.000	0.900	0.867	0.900
$IC^{\star 3}_{0.90}$	0.867	0.900	0.833	0.867
$IC^{\star 4}_{0.90}$	0.900	0.867	0.867	0.867
$IC^{\star 5}_{0.90}$	0.900	0.933	0.867	0.800
$IC^{\star 6}_{0.90}$	0.900	0.867	0.867	0.833
$IC^{\star 7}_{0.90}$	0.900	0.900	0.867	-
$IC^{\star 8}_{0.90}$	0.867	0.867	0.867	-
$IC^{\star 9}_{0.90}$	0.867	0.867	-	-
$IC^{\star 10}_{0.90}$	0.900	0.833	-	-
$IC^{\star 11}_{0.90}$	0.867	-	-	-
$IC^{\star 12}_{0.90}$	0.900	-	-	-
$IC^{\star 13}_{0.90}$	0.900	-	-	-

Tabella 4.5: Probabilità di copertura degli intervalli di confidenza puntuali stimata per simulazione, al variare del lag.

	$\overline{c}_{(sim)}$ lag=8 km	$\overline{c}_{(sim)}$ lag=10 km	$\overline{c}_{(sim)}$ lag=13 km	$\overline{c}_{(sim)}$ lag=17 km
$IC^{\star(sim)}_{0.90}$	0.733	0.633	0.733	0.767

Tabella 4.6: Probabilità di copertura degli intervalli di confidenza simultanei stimata per simulazione, al variare del lag.

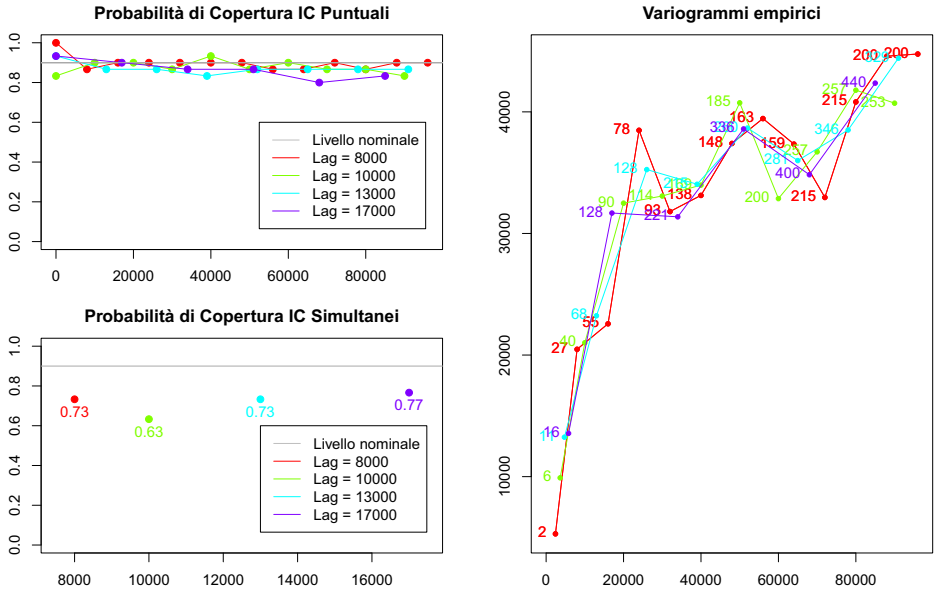

Figura 4.20: A sinistra: Stima della probabilità di copertura degli intervalli di confidenza puntuali (in alto) e simultanei (in basso), costruiti con il metodo dei percentili. In ascissa è riportata la distanza h_k, $k = 1, ..., K$ espressa in metri, in ordinata è invece riportata la stima $\overline{c}_k$ ($\overline{c}_{(sim)}$) della probabilità di copertura del corrispondente stimatore intervallare $IC^{\star k}_{0.90}$ ($IC^{\star (sim)}_{0.90}$). La linea grigia corrisponde al livello di confidenza nominale ($1-\alpha = 0.9$). A destra: Stimatori empirici, calcolati al variare del lag in $\{8000; 10000; 13000; 17000\}$ m; per ogni valore della distanza h_k (in metri), è indicato il numero $|N(h_k)|$ di coppie appartenenti alla classe $N(h_k)$, per $k = 1, ..., K$. In tutti i pannelli i colori distinguono il valore fissato per il lag.

stocastico funzionale stazionario e isotropo. Gli Algoritmi 4.6 e 4.7 costituiscono infatti un'estensione al caso funzionale dei metodi bootstrap semiparametrici proposti in letteratura in riferimento a campi aleatori a valori reali.

Un vantaggio della metodologia sviluppata è la possibilità di accedere alle caratteristiche probabilistiche degli stimatori costruiti evitando l'introduzione di ipotesi distribuzionali per il processo stocastico: i metodi bootstrap semiparametrici necessitano infatti soltanto di una stima preliminare del modello di variabilità spaziale.

La distribuzione dello stimatore sperimentale del variogramma approssimata con il metodo MC-bootstrap è stata usata nel capitolo in due direzioni: l'adattamento di un modello valido $\gamma(h;\boldsymbol{\vartheta})$ di variogramma attraverso una stima GLS dei parametri $\boldsymbol{\vartheta}$ e la costruzione di opportuni intervalli di confidenza per lo stimatore empirico.

Gli Algoritmi 4.6 e 4.7 sviluppati sono quindi stati validati attraverso uno studio di simulazione, applicando la procedura a un dataset funzionale sintetico gaussiano (Sottosezione 4.3.3) e a uno non gaussiano (Sottosezione 4.3.4). In particolare, gli obiettivi delle analisi sono state la valutazione della convergenza del procedimento iterativo, della qualità dell'approssimazione fornita e del livello di copertura delle stime intervallari individuate.

Dallo studio effettuato sul campione gaussiano, è possibile affermare che le approssimazioni MC-bootstrap risultano accettabili già per $B = 1000$ e $N_{max} = 2$, e piuttosto accurate per $B = 5000$, presentando oscillazioni limitate nelle stime al variare di B e N_{max}. L'analisi del campione non gaussiano ha mostrato un comportamento coerente con le prestazioni della metodologia applicata al campione gaussiano, evidenziando che l'Algoritmo 4.7 si dimostra robusto rispetto all'ipotesi di gaussianità del campo.

In riferimento alle stime intervallari, le approssimazioni MC-bootstrap hanno fornito intervalli di confidenza puntuali di livello stimato non lontano dal livello nominale, dimostrandosi leggermente conservativi solo nel caso gaussiano. Gli intervalli di confidenza simultanei sono invece risultati più sensibili all'ipotesi di gaussianità del campo, evidenziando la necessità di introdurre correzioni di Bonferroni qualora il processo stocastico non risulti gaussiano.

Inoltre, nelle analisi svolte, il metodo è risultato robusto rispetto alla scelta del modello di variogramma e del lag caratteristico dello stimatore empirico.

In conclusione, l'inferenza sulla stima del variogramma necessiterebbe di una maggiore indagine teorica: la dimostrazione delle proprietà di consistenza dello stimatore empirico del variogramma funzionale, così come l'analisi della distribuzione asintotica, sono campi di ricerca tuttora quasi inesplorati.

Inoltre, dal punto di vista algoritmico, sarebbe sicuramente di interesse lo studio teorico delle proprietà di convergenza della procedura proposta, in termini di velocità di convergenza e qualità della stima. Infine, particolare attenzione andrebbe posta sullo studio teorico dell'influenza delle ipotesi di gaussianità del processo e della numerosità del dataset sulla qualità delle approssimazioni MC-bootstrap fornite dalla metodologia proposta.

Capitolo 5

Previsione per Dati Funzionali Georeferenziati Non Stazionari: l'Universal Kriging

Nell'ambito dell'analisi geostatistica, un momento molto rilevante ai fini applicativi è costituito dalla previsione del fenomeno dove non sia disponibile l'osservazione. Nel Capitolo 3, sono stati formalizzati a questo scopo i metodi di Ordinary e Universal Kriging per elementi di uno spazio di Hilbert, che consentono di ottenere la previsione voluta rispettivamente in ipotesi di stazionarietà e di non stazionarietà.

Dal punto di vista applicativo, non è tuttavia possibile applicare direttamente i metodi di kriging individuati, in quanto generalmente non è disponibile una stima della struttura di dipendenza spaziale e quest'ultima, nel caso non stazionario, può essere calcolata soltanto dopo una stima preliminare del termine di *drift*.

Per questo motivo, in questo capitolo è proposta una metodologia per l'analisi di dati funzionali georeferenziati spazialmente non stazionari, integrata alla procedura di stima variografica sviluppata nel Capitolo 4 e articolata in tre fasi: la selezione della forma funzionale per la modellazione del *drift*, il disaccoppiamento della variabilità deterministica dalla variabilità stocastica e l'Universal Kriging.

In particolare, dopo l'introduzione degli algoritmi necessari allo svolgimento delle prime due fasi, illustrati nelle Sezioni 5.1 e 5.2, la metodologia proposta sarà analizzata nel suo comportamento attraverso l'applicazione a dati sintetici, che saranno introdotti e studiati nelle Sezioni 5.3 e 5.4.

5.1 Stima del *Drift* e del Modello di Variogramma: l'Algoritmo

Si consideri un processo stocastico funzionale non stazionario $\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}$ a valori in uno spazio di Hilbert H e si assuma che tale processo possa essere rappresentato secondo la dicotomia (3.17), ossia come:

$$\boldsymbol{\chi}_s = m_s + \boldsymbol{\delta}_s, \quad s \in D,$$

dove m_s indica il *drift* (deterministico) del processo e $\boldsymbol{\delta}_s$ il residuo debolmente stazionario e isotropo (Definizioni 3.7 e 3.11), ipotizzato a media nulla, con covariogramma $C(h)$ e semivariogramma $\gamma(h)$:

$$\begin{aligned}
\mathbb{E}[\boldsymbol{\chi}_s] &= m_s, \quad \forall s \in D, \\
\mathbb{E}[\boldsymbol{\delta}_s] &= 0, \quad \forall s \in D, \\
\mathrm{Cov}(\boldsymbol{\chi}_{s_i}, \boldsymbol{\chi}_{s_j}) &= \mathrm{Cov}(\boldsymbol{\delta}_{s_i}, \boldsymbol{\delta}_{s_j}) = C(\|s_i - s_j\|), \quad \forall\, s_i, s_j \in D, \\
\mathrm{Var}(\boldsymbol{\chi}_{s_i} - \boldsymbol{\chi}_{s_j}) &= \mathrm{Var}(\boldsymbol{\delta}_{s_i} - \boldsymbol{\delta}_{s_j}) = 2\gamma(\|s_i - s_j\|), \quad \forall\, s_i, s_j \in D.
\end{aligned}$$

Al fine di applicare il metodo di Universal Kriging attraverso la soluzione del sistema lineare (3.38) è essenziale disporre di una stima del variogramma $\gamma(h)$ (*cfr.* Sezione 3.4 del Capitolo 3). D'altra parte, quest'ultimo è calcolabile con i metodi del Capitolo 4 a partire da una realizzazione del residuo $\boldsymbol{\delta}_s$, la cui stima è determinabile per differenza dopo una stima preliminare del termine di *drift*.

Nel contesto del Capitolo 3 e con le notazioni ivi introdotte, tale stima sarà individuata nel seguito secondo il criterio dei minimi quadrati generalizzati, ovvero minimizzando il funzionale (3.43):

$$\|\boldsymbol{\chi_s} - \widehat{\boldsymbol{m}}_{\boldsymbol{s}}\|^2_{\Sigma^{-1} - H^n},$$

dove $\boldsymbol{\chi_s} = (\boldsymbol{\chi}_{s_1}, ..., \boldsymbol{\chi}_{s_n})$ è il dataset funzionale, mentre $\widehat{\boldsymbol{m}}_{\boldsymbol{s}}$ individua una stima del *drift*, che è supposto della forma:

$$m_{\boldsymbol{s}} = \mathbb{F}_{\boldsymbol{s}} a_{\boldsymbol{l}} = \sum_{l=0}^{L} a_l f_l(s), \quad s \in D,$$

con $f_l(\cdot)$ funzioni della sola variabile spaziale $s \in D$ e $a_l \in H$, per ogni $l = 1, ..., L$.

Gli stimatori lineari dei coefficienti $a_{\boldsymbol{l}}$, $\widehat{\boldsymbol{a}}^{GLS}_{\boldsymbol{l}}$, e del vettore di *drift* $m_{\boldsymbol{s}}$, $\widehat{\boldsymbol{m}}_{\boldsymbol{s}}$, ammettono le espressioni esplicite rispettivamente (3.48) e (3.47), ovvero:

$$\begin{aligned}
\widehat{\boldsymbol{a}}^{GLS}_{\boldsymbol{l}} &= (\mathbb{F}^T_{\boldsymbol{s}} \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} \mathbb{F}_{\boldsymbol{s}} \Sigma^{-1} \boldsymbol{\chi_s}, \\
\widehat{\boldsymbol{m}}_{\boldsymbol{s}} &= \mathbb{H} \boldsymbol{\chi_s} = \mathbb{F}_{\boldsymbol{s}} (\mathbb{F}^T_{\boldsymbol{s}} \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} \mathbb{F}_{\boldsymbol{s}} \Sigma^{-1} \boldsymbol{\chi_s};
\end{aligned}$$

per linearità, le espressioni del vettore media e della matrice di covarianza possono essere determinate come (*cfr.* Sezione 3.5):

$$\begin{aligned}
\mathbb{E}[\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}] &= a_{\boldsymbol{l}}; \\
\mathbb{E}[\widehat{\boldsymbol{\chi}}_{\boldsymbol{s}}] &= \mathbb{F}_{\boldsymbol{s}} a_{\boldsymbol{l}} = m_{\boldsymbol{s}}; \\
\mathrm{Cov}(\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}) &= (\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1}; \\
\mathrm{Cov}(\widehat{\boldsymbol{\chi}}_{\boldsymbol{s}}) &= \mathbb{F}_{\boldsymbol{s}}^T \Lambda \mathbb{F}_{\boldsymbol{s}} = \mathbb{F}_{\boldsymbol{s}}^T (\mathbb{F}_{\boldsymbol{s}}^T \Sigma^{-1} \mathbb{F}_{\boldsymbol{s}})^{-1} \mathbb{F}_{\boldsymbol{s}},
\end{aligned}$$

concludendo in particolare che gli stimatori introdotti risultano BLUE.

Tuttavia gli stimatori (3.48) e (3.47), dipendono in modo sostanziale dalla matrice di covarianza dei dati Σ, che in generale è incognita ed è sostituita da una sua stima $\widehat{\Sigma}$. Quest'ultima può essere determinata soltanto a partire da una stima del residuo, che, a sua volta, dipende dalla stima del *drift*, individuata con il metodo GLS attraverso la matrice $\widehat{\Sigma}$ stessa. Al fine di usare il criterio di ottimo dei minimi quadrati generalizzati, la cui importanza metodologica è stata più volte sottolineata, è quindi necessario ricorrere ad un algoritmo di tipo iterativo, basato su una stima iniziale del *drift* ottenuta con il criterio OLS.

Più precisamente, l'algoritmo proposto è il seguente:

Algoritmo 5.1. *Data una realizzazione* $\chi_{\boldsymbol{s}} = (\chi_{s_1}, ..., \chi_{s_n})$ *del processo stocastico non stazionario* $\{\boldsymbol{\chi}_s : s \in D \subseteq \mathbb{R}^d\}$, $D \subseteq \mathbb{R}^d$*:*

1. *Stimare i coefficienti* $a_{\boldsymbol{l}}$ *con OLS attraverso la* (3.59)*, porre* $\widehat{a}_{\boldsymbol{l}} := \widehat{a}_{\boldsymbol{l}}^{OLS}$*;*

2. *Ricavare la stima del* drift $\widehat{m}_{\boldsymbol{s}} = (\widehat{m}_{s_1}, ..., \widehat{m}_{s_n})$ *come:*

$$\widehat{m}_{\boldsymbol{s}} = \mathbb{F}_{\boldsymbol{s}} \widehat{a}_{\boldsymbol{l}}$$

   e il residuo $\widehat{\delta}_{\boldsymbol{s}} = (\widehat{\delta}_{s_1}, ..., \widehat{\delta}_{s_n})$ *per differenza:*

$$\widehat{\delta}_{\boldsymbol{s}} = \chi_{\boldsymbol{s}} - \widehat{m}_{\boldsymbol{s}};$$

3. *Stimare il trace-variogram* $2\gamma(h)$ *di* $\boldsymbol{\delta}_s$ *dal vettore* $\widehat{\delta}_{\boldsymbol{s}}$*, dapprima con lo stimatore empirico, quindi adattando un modello valido di variogramma (con metodo MC-bootstrap).*

4. *Ricavare la stima* $\widehat{\Sigma}$ *(e calcolarne la decomposizione di Cholesky* $\widehat{\Sigma} = LL^T$*).*

5. *Stimare i coefficienti* $a_{\boldsymbol{l}}$ *con* $\widehat{a}_{\boldsymbol{l}}^{GLS}$ *ottenuto da* $\chi_{\boldsymbol{s}}$ *per mezzo della* (3.48) *(o da* $\widetilde{\chi}_{\boldsymbol{s}} = L^{-1}\chi_{\boldsymbol{s}}$ *per mezzo della* (3.46)*); porre* $\widehat{a}_{\boldsymbol{l}} := \widehat{a}_{\boldsymbol{l}}^{GLS}$*.*

6. *Ripetere 2.-5. fino a convergenza.*

Dal punto di vista teorico, il calcolo del vettore di coefficienti $\widehat{a}_{\boldsymbol{l}}^{GLS}$ può essere svolto indifferentemente sul campione originale, $\chi_{\boldsymbol{s}}$, o sul campione scorrelato $\widetilde{\chi}_{\boldsymbol{s}}$, ottenuto premoltiplicando $\chi_{\boldsymbol{s}}$ per l'inverso del fattore di Cholesky L^{-1}.

Tuttavia, dal punto di vista implementativo, nel secondo caso possono essere usati i codici (ottimizzati) contenuti all'interno del pacchetto `fda` del software R, eventualmente imponendo una base funzionale per i coefficienti $a_{\boldsymbol{l}}$, come sarà fatto nella prima parte della Sezione 5.3. Si noti che l'imposizione di una base per i coefficienti si traduce nella scelta di un sottospazio $\widetilde{V}_0 \subseteq \widetilde{V}$, definito dalla (3.44), sul quale proiettare il campione $\chi_{\boldsymbol{s}}$: questo caso, dettagliato in (Ramsay e Silverman, 2005), non è stato sviluppato esplicitamente nel Capitolo 3, in quanto si è ritenuta di maggior generalità e rilievo applicativi l'assunzione di un forma (3.33) per il termine di *drift*. La ragionevolezza di questa assunzione sarà considerata nell'analisi di robustezza svolta in Sottosezione 5.3.2.5.

Il comportamento dell'algoritmo introdotto sarà studiato nel seguito del capitolo con applicazione a dataset sintetici e reali. In particolare, attraverso uno studio di simulazione, saranno indagate la velocità di convergenza dell'algoritmo, quindi come fissare il numero di iterazioni N_{max}^{GLS}, la qualità delle stime fornite e la robustezza del metodo rispetto alla misclassificazione del modello di variogramma e della base funzionale per i coefficienti a_l, $l = 0, ..., L$.

5.2 La Scelta del *Drift*: l'Algoritmo di Ordinamento dei *Drift* a Criterio Previsivo

I metodi introdotti nella sezione precedente e, in particolare, l'Algoritmo 5.1, si basano sull'ipotesi che le funzioni f_l siano note, per ogni $l = 0, ..., L$ ($f_0 = 1$). Tuttavia, in assenza di *external drift*, è possibile sia che tali funzioni non siano fissate a priori, sia che esistano diverse collezioni di regressori plausibili per il *drift*.

Nel primo caso, è una pratica comune considerare per f_l, $l = 0, ..., L$, alcune forme polinomiali, il cui grado sia sufficientemente alto per descrivere esaustivamente il *drift* e rendere il residuo stazionario, mantenendo allo stesso tempo ridotta, per quanto possibile, la complessità del modello.

Per questo motivo, in entrambe le situazioni, si rende necessaria l'introduzione di un ordinamento di tali collezioni di funzioni secondo un criterio prestabilito, che, per il lavoro di tesi, sarà il criterio previsivo. L'algoritmo proposto in questa sezione è un'estensione al contesto funzionale di quanto attualmente implementato all'interno del software ISATIS®.

Dal punto di vista formale, si considerino N_f collezioni di funzioni $f_{\boldsymbol{l}}^k = \{f_0^k, ..., f_L^k\}$, corrispondenti a N_f possibili *drift*:

$$m_s^k = \sum_{l=0}^{L} a_l f_l^k(s), \quad s \in D,\, k = 1, ..., N_f,$$

e si denoti con $\{(1), ..., (N_f)\}$ una permutazione di $\{1, ..., N_f\}$. L'obiettivo di questa sezione è l'introduzione di un algoritmo che consenta di determinare la permutazione degli indici $\{1, ..., N_f\}$ corrispondente all'ordinamento delle collezioni $f_{\boldsymbol{l}}^k = \{f_0^k, ..., f_L^k\}$ secondo il criterio

previsivo, quantificato attraverso il funzionale scarto quadratico medio:

$$MSE_k = \mathbb{E}[\|\boldsymbol{\chi}_s - \boldsymbol{\chi}_s^{*k}\|^2], \quad k = 1, ..., N_f,$$

dove $\boldsymbol{\chi}_s^{*k}$ indica il valore previsto in posizione s, adottando il modello di *drift* m_s^k, $k = 1, ..., N_f$.

Per rispondere alla necessità che la valutazione del funzionale MSE_k sia efficiente, la metodologia proposta prevede l'adozione di una tecnica di tipo cross-validazione a partire da una previsione spaziale semplificata, basata su un modello di variogramma di tipo puro *nugget*, rimandando quindi il calcolo di una stima accurata all'analisi conseguente alla scelta della forma del *drift*.

Più precisamente, per ogni collezione $f_{\boldsymbol{l}}^k = \{f_0^k, ..., f_L^k\}$, il metodo di cross-validazione (*leave-one-out*) consente di determinare una stima campionaria del termine MSE_k, $k = 1, ..., N_f$, come media di n stime $\chi_{s_i}^{*k}$ di χ_{s_i}, $i = 1, ..., n$, ottenute dai dataset $\chi_{\boldsymbol{s}^{-i}} = (\chi_{s_j})_{j \neq i}$ come illustrato di seguito.

Una volta costruito il dataset ridotto $\chi_{\boldsymbol{s}^{-i}}$ rimuovendo il dato χ_{s_i} dal dataset originale, la stima $\chi_{s_i}^{*k}$ è costruita considerando lo stimatore di kriging:

$$\chi_{s_i}^{*k} = \sum_{j \neq i} \lambda_j^{*k} \chi_{s_j}, \quad i = 1, ..., n;\ k = 1, ..., N_f,$$

dove i pesi λ_j^{*k} sono determinati risolvendo il sistema (3.38), assumendo la struttura di *drift* m_s^k e un modello di variogramma puro *nugget*, ossia ponendo $\gamma(h) = C(0) I_{\{0\}}$, dove I rappresenta la funzione indicatrice.

Si noti che la stima ottenuta da un modello di puro *nugget* corrisponde alla media dei dati:

$$\chi_{s_i}^{*k} = \frac{1}{n-1} \sum_{j \neq i} \chi_{s_j}, \quad i = 1, ..., n;\ k = 1, ..., N_f,$$

in quanto i pesi ottimi risultano, in tal caso:

$$\lambda_j^{*k} = \frac{1}{n-1}, \quad j \neq i.$$

A partire dai valori predetti $\chi_{s_i}^{*k}$, una stima dei valori MSE_k può essere espressa attraverso lo stimatore campionario:

$$MSE_k = \frac{1}{n} \sum_{i=1}^{n} \|\chi_{s_i} - \chi_{s_i}^{*k}\|^2.$$

Infine, la permutazione degli indici $\{1, ..., N_f\}$ è ottenuta dall'ordinamento dei valori di MSE_k o, equivalentemente, dall'ordinamento di SSE_k, definito come:

$$SSE_k = \sum_{i=1}^{n} \|\chi_{s_i} - \chi_{s_i}^{*k}\|^2 = \|\chi_{\boldsymbol{s}} - \chi_{\boldsymbol{s}}^{*k}\|_{H^n}^2,$$

ponendo $\chi_{\boldsymbol{s}}^{*k} = (\chi_{s_1}^{*k}, ..., \chi_{s_n}^{*k})$.

Il metodo descritto è sintetizzato nell'algoritmo riportato di seguito.

Algoritmo 5.2. *Data una realizzazione $\chi_{s_1}, ..., \chi_{s_n}$ del processo stocastico non stazionario $\{\boldsymbol{\chi}_s, s \in D\}$ e N_f collezioni di funzioni $f_{\boldsymbol{l}}^k = \{f_0^k, ..., f_L^k\}$ (candidate per il* drift*):*

1. *Fissare una collezione $f_{\boldsymbol{l}}^k$, $k = 1, ..., N_f$;*
2. *Per ogni $i = 1, ..., n$ fissato:*
   a. *Selezionare il dato χ_i e rimuoverlo;*
   b. *Prevedere χ_i dal campione $\chi_{\boldsymbol{s}^{-i}}$ con un modello puro* nugget*:*
   $$\chi_{s_i}^{*k} = \sum_{j \neq i} \lambda_j^* \chi_{s_j},$$
   con λ_j^* soluzione del sistema di kriging (3.38) posto $\gamma(h) = C(0)I_{\{0\}}$, $f_l = f_l^k$, $l = 1, ..., L$.
3. *Calcolare la somma degli scarti quadratici tra i dati osservati χ_{s_i} e i valori predetti $\chi_{s_i}^{*k}$:*
   $$SSE_k = \sum_{i=1}^{n} \|\chi_{s_i} - \chi_{s_i}^{*k}\|^2,$$
   o lo scarto quadratico medio:
   $$MSE_k = \frac{1}{n} \sum_{i=1}^{n} \|\chi_{s_i} - \chi_{s_i}^{*k}\|^2;$$
4. *Ripetere 1.-2.-3. per ogni collezione $f_{\boldsymbol{l}}^k$, $k = 1, ..., N_f$;*
5. *Ordinare $\{SSE_1, ..., SSE_{N_f}\}$ (o $\{MSE_1, ..., MSE_{N_f}\}$) in modo crescente, individuando una permutazione $\{(1), ..., (N_f)\}$ di $\{1, ..., N_f\}$:*
   $$(SSE_{(1)}, ..., SSE_{(N_f)}).$$
6. *Ordinare le collezioni $\{f_{\boldsymbol{l}}^k\}_{k=1,...,N_f}$, secondo l'ordinamento dettato dalla permutazione $\{(1), ..., (N_f)\}$:*
   $$\{f_{\boldsymbol{l}}^{(k)}\}_{(k)=1,...,N_f}$$
   e selezionare il drift *ottimo:*
   $$m_s^{opt} = \sum_{l=0}^{L} a_l f_l^{opt}(s), \quad s \in D,$$
   con $f_{\boldsymbol{l}}^{opt} = f_{\boldsymbol{l}}^{(1)}$.

L'Algoritmo 5.2, la cui implementazione è riportata in Appendice B.3, sarà nel seguito affiancato all'analisi del variogramma risultante dall'applicazione di un'iterazione dell'Algoritmo 5.1, per verificare che il modello di *drift* ottimo secondo il criterio introdotto dia luogo a un residuo stazionario.

k	L	f_l^k	$m_s^k \quad [s=(x,y)]$
1	1	$\{1, x\}$	$a_0 + a_1 x$
2	1	$\{1, y\}$	$a_0 + a_1 y$
3	2	$\{1, x, y\}$	$a_0 + a_1 x + a_2 y$
4	1	$\{1, x^2\}$	$a_0 + a_1 x^2$
5	1	$\{1, y^2\}$	$a_0 + a_1 y^2$
6	1	$\{1, xy\}$	$a_0 + a_1 xy$
7	2	$\{1, x^2, y^2\}$	$a_0 + a_1 x^2 + a_2 y^2$
8	3	$\{1, x^2, y^2, xy\}$	$a_0 + a_1 x^2 + a_2 y^2 + a_3 xy$

Tabella 5.1: Forme funzionali per il *drift* testate dall'Algoritmo 5.2 in assenza di informazioni a priori.

Le forme funzionali scelte per il *drift* in assenza di informazioni a priori e testate in questo lavoro attraverso l'Algoritmo 5.2 sono riportate in Tabella 5.1.

Gli Algoritmi 5.1 e 5.2 saranno l'oggetto dello studio di simulazione illustrato nelle Sezioni 5.3 e 5.4 e saranno in un secondo momento applicati a dati reali (Capitolo 6).

5.3 Applicazione a Dati Sintetici con Modello di *Drift* a Coefficienti Costanti

L'Algoritmo 5.1 introdotto nella Sezione 5.1 consente di disaccoppiare la variabilità spaziale deterministica, rappresentata dal termine di *drift*, dalla variabilità spaziale di carattere aleatorio, ovvero il residuo $\boldsymbol{\delta}_s$ del processo, individuando una stima GLS della prima componente e, di conseguenza, una stima del residuo stazionario.

Con l'ausilio dell'Algoritmo 5.2 è possibile determinare il *drift* ottimo in senso previsivo tra un insieme di N_f *drift* candidati, selezionando tra $\{f_l^1, ..., f_l^{N_f}\}$ la collezione di funzioni f_l^{opt} più adatta a descrivere la media spaziale m_s del processo funzionale.

Questi algoritmi saranno ora applicati a dati sintetici non stazionari, costruiti a partire dal dataset gaussiano stazionario oggetto di studio nella Sottosezione 4.3.3 del Capitolo 4.

5.3.1 I Dati

Il primo dataset $\chi_{s_1}, ..., \chi_{s_n}$ sul quale è stato valutato il comportamento degli Algoritmi 5.1 e 5.2 è stato costruito come somma di un residuo stazionario a media nulla e di un termine di *drift* a coefficienti costanti, attenendosi alla dicotomia (3.17):

$$\boldsymbol{\chi}_s = m_s + \boldsymbol{\delta}_s.$$

In particolare, per il residuo $\boldsymbol{\delta}_s$ è stato considerato il medesimo campo aleatorio funzionale gaussiano usato nello studio di simulazione del Capitolo 4, corrispondente alla forma

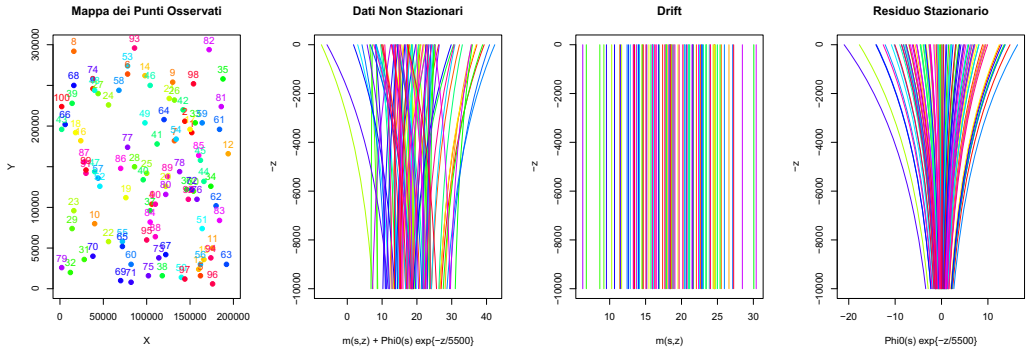

Figura 5.1: Da sinistra a destra: Mappa dei punti campionati $s_1, ..., s_{100}$; primo dataset funzionale sintetico non stazionario $\chi_{\boldsymbol{s}} = (\chi_{s_1}, ..., \chi_{s_{100}})$; *drift* $m_{\boldsymbol{s}} = (m_{s_1}, ..., m_{s_{100}})$; residuo stazionario $\delta_{\boldsymbol{s}} = (\delta_{s_1}, ..., \delta_{s_{100}})$.

funzionale (4.12), riportata di seguito:

$$\boldsymbol{\delta}_s(z) = \Phi_0(s) \exp\{-z/\beta\}, \quad z \in \mathcal{T} = [0, 10000],$$

dove $\Phi_0(s)$ è un campo gaussiano a valori reali, a media nulla e variogramma esponenziale:

$$\gamma(h) = \begin{cases} 80(1 - e^{-3h/75}), & h > 0 \\ 0, & h = 0, \end{cases}$$

indicando le distanze in chilometri.

Il campione $\delta_{\boldsymbol{s}} = (\delta_{s_1}, ..., \delta_{s_n})$ usato per la costruzione del dataset non stazionario al quale sono stati applicati in questa sezione gli Algoritmi 5.1 e 5.2 è il medesimo generato per lo studio di simulazione del Capitolo 4 e riportato in Figura 4.5.

Per il termine di *drift* è invece stata considerata la forma:

$$m_s(z) = a_0 + a_1 x + a_2 y, \quad s = (x, y) \in D, \, z \in \mathcal{T},$$

dove a_0, a_1, a_2 sono coefficienti reali pari a $a_0 = 5.0$, $a_1 = 7.5 \cdot 10^{-3}$, $a_2 = 4.25 \cdot 10^{-3}$, considerando come unità di misura di x e y i chilometri.

Il dataset risultante è rappresentato nelle Figure 5.1, 5.2, 5.3 e 5.4. In particolare, le Figure 5.2, 5.3 e 5.4 mostrano i grafici contour dell'intera griglia di simulazione, dalla quale è stato estratto il campione di numerosità $n = 100$, $\chi_{s_1}, ..., \chi_{s_{100}}$, per $z = \{0; 2500; 7500; 10000\}$ m; tali mappe saranno confrontate nel seguito con le mappe ricostruite con il metodo di Universal Kriging a partire dal campione $\chi_{\boldsymbol{s}}$.

Coerentemente con lo studio del Capitolo 4, lo spazio di Hilbert considerato per le analisi successive è lo spazio L^2 delle funzioni quadrato integrabili, dotato del prodotto interno e della norma usuali, espressi dalla (4.14).

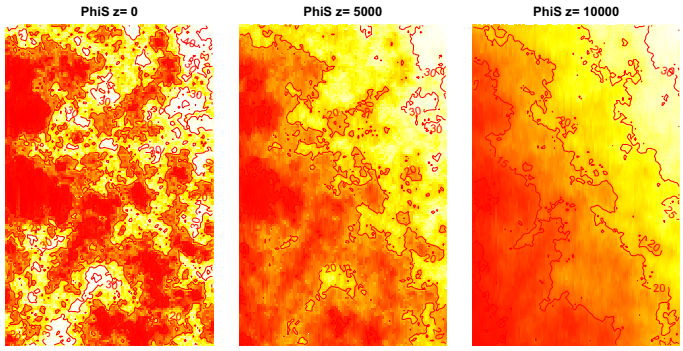

Figura 5.2: Contour-plot della realizzazione del processo $\boldsymbol{\chi}_s$ sull'intera griglia di simulazione, per $z = \{0; 5000; 10000\}$ m.

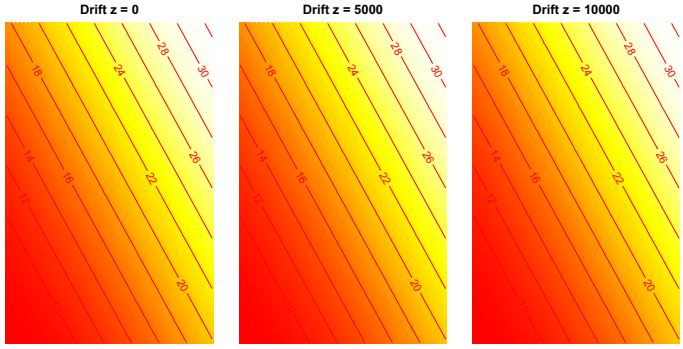

Figura 5.3: Contour-plot del termine di *drift* m_s sull'intera griglia di simulazione, per $z = \{0; 5000; 10000\}$ m.

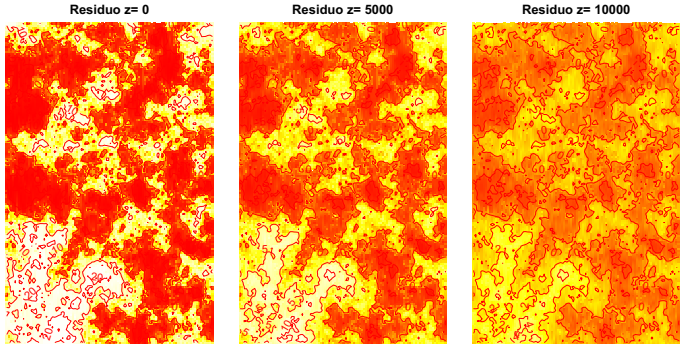

Figura 5.4: Contour-plot della realizzazione del residuo processo $\boldsymbol{\delta}_s$ sull'intera griglia di simulazione, per $z = \{0; 5000; 10000\}$ m.

(k)	k	f_l^k	SSE_k
(1)	3	$\{1, x, y\}$	151217.4
(2)	8	$\{1, x^2, y^2, xy\}$	159290.2
(3)	6	$\{1, xy\}$	171264.6
(4)	7	$\{1, x^2, y^2\}$	188620.4
(5)	2	$\{1, y\}$	311713.1
(6)	4	$\{1, x^2\}$	338550.7
(7)	1	$\{1, x\}$	342074.3
(8)	5	$\{1, y^2\}$	343117.9

Tabella 5.2: Ordinamento dei *drift* per il dataset sintetico $\chi_{\boldsymbol{s}}$.

5.3.2 I Risultati di Simulazione

In questa sezione saranno mostrati i risultati ottenuti applicando gli Algoritmi 5.1 e 5.2 ai dati sintetici introdotti nella Sottosezione 5.3.1.

In particolare, l'analisi sarà svolta in due momenti: con l'ausilio dell'Algoritmo 5.2, sarà dapprima selezionato il *drift* ottimale tra le forme riportate in Tabella 5.1 (Sezione 5.2), quindi sarà applicato l'Algoritmo 5.1 a partire dalla struttura di *drift* individuata.

Il comportamento del metodo sarà valutato in termini previsivi attraverso un'analisi di cross-validazione, quindi rispetto alle proprietà di convergenza. Infine, sarà valutata la robustezza della metodologia introdotta relativamente alla misclassificazione del modello di variogramma e della base per i coefficienti del *drift* a_l, $l = 0, ..., L$.

5.3.2.1 Ordinamento dei *Drift*

Si consideri dunque il dataset non stazionario sintetico $\chi_{s_1}, ..., \chi_{s_n}$ introdotto nella Sottosezione 5.3.1.

Al fine di ricostruire un campo di funzioni attraverso il metodo di Universal Kriging è necessario in primo luogo selezionare la forma di *drift* più adatta alla descrizione della variabilità spaziale deterministica del fenomeno.

Per fare questo, l'Algoritmo 5.2 è stato applicato al dataset funzionale, ottenendo l'ordinamento dei *drift* riportato in Tabella 5.2.

La struttura selezionata risulta pertanto:

$$m_s(z) = a_0 + a_1 x + a_2 y, \quad s = (x, y) \in D,\ z \in \mathcal{T},$$

ovvero il medesimo modello di *drift* di generazione.

Nel grafico riportato in Figura 5.5 sono rappresentati i variogrammi ottenuti dalla prima iterazione dell'Algoritmo 5.1. Come si può notare, i variogrammi corrispondenti ai *drift* 3, 8 e 6, individuati dall'Algoritmo 5.2 come i modelli migliori, corrispondono ai variogrammi empirici dalla forma più stazionaria: essi mostrano infatti una chiara concavità verso il basso

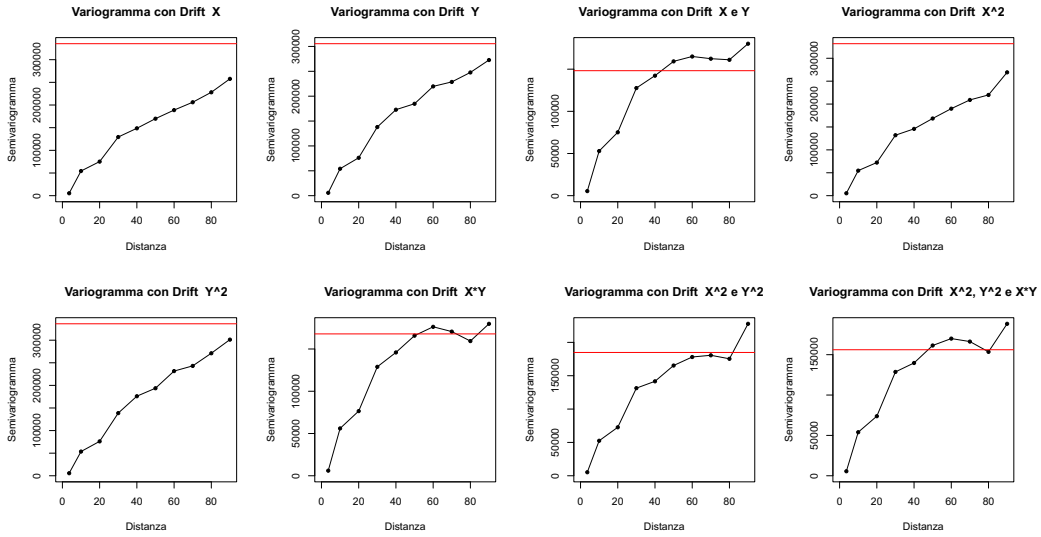

Figura 5.5: Variogrammi dei residui $\delta^k_{\boldsymbol{s}}$, al variare delle forme di *drift* in Tabella 5.1, ottenuti da una prima iterazione dell'Algoritmo 5.1. Si presti attenzione nel confronto tra i grafici, in quanto, per consentire la valutazione delle forme delle stime variografiche, la scala adottata per l'asse delle ordinate non è comune tra i pannelli.

in prossimità dell'origine, assestandosi verso un *sill* non lontano dalla varianza campionaria, già attorno alla distanza di 60 km.

Dalla forma dei variogrammi sperimentali corrispondenti ai modelli di *drift* 2, 4, 1 e 5, così come dal confronto dei corrispondenti SSE_k, si osserva invece che, in accordo con la forma del modello di riferimento, i polinomi nella sola variabile x o y non riescono a catturare in maniera sufficiente la variabilità spaziale deterministica del fenomeno.

5.3.2.2 Disaccoppiamento del *Drift* e Universal Kriging

Una volta selezionato il *drift* secondo l'ordinamento fornito dall'Algoritmo 5.2 e verificato che il residuo prodotto risulti stazionario, il disaccoppiamento della variabilità spaziale deterministica dal residuo stazionario è stato ottenuto applicando l'Algoritmo 5.1.

Sarà ora dettagliato lo studio effettuato fissando un modello di variogramma esponenziale e una base costante per i coefficienti. Saranno quindi illustrati i risultati di simulazione ottenuti attraverso l'Algoritmo 5.1 grazie all'implementazione riportata in Appendice B.3, nell'ambiente di simulazione di `R` 2.13.1.

L'Algoritmo 5.1 è stato applicato fissando il numero di iterazioni a $N^{GLS}_{max} = 5$, che, come sarà mostrato nella Sottosezione 5.3.2.4 è risultato sufficiente a stabilizzare la stima del termine di *drift*; per ciascuna iterazione sono stati registrati:

i. i parametri $\widehat{a}^{GLS}$ stimati;

ii. il campo di *drift* m_s stimato con il metodo GLS;

iii. la stima empirica e parametrica del variogramma del residuo funzionale, ottenuto con i metodi del Capitolo 4;

iv. il residuo funzionale δ_s^* stimato con il metodo di Ordinary Kriging a partire dal modello di variogramma stimato;

v. il campo funzionale $\chi_{\boldsymbol{s}}^*$ stimato con il metodo di Universal Kriging a partire dal campione $\chi_{\boldsymbol{s}}$ e dal modello di variogramma stimato.

In particolare, le stime $\chi_{\boldsymbol{s}}^*$ e δ_s^* dei punti iv.-v., così come la stima del campo di *drift* al punto ii., sono state valutate su una griglia equispaziata ampia $[0, 100] \times [0, 150]$ km, con passo rispettivamente 2 e 3 km.

In Figura 5.6 sono riportati i coefficienti $\widehat{a}_0^{GLS}, \widehat{a}_1^{GLS}, \widehat{a}_2^{GLS}$ stimati con il metodo proposto, a confronto con la stima OLS e i parametri di generazione.

L'incertezza legata alla stima è stata quantificata attraverso la media integrale della varianza globale, ovvero attraverso la quantità:

$$\varsigma_l^2 = \frac{1}{|\mathcal{T}|}\Lambda_{ll}, \tag{5.1}$$

dove Λ è la matrice di covarianza spaziale dello stimatore $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}$. Infatti, l'espressione (5.1) corrisponde alla traccia dell'operatore covarianza (Definizione 3.5) di a_l, normalizzata rispetto alla misura del dominio; tuttavia, la (5.1) è equivalente in L^2 alla media integrale della funzione varianza $\varsigma_l^2(t)$ (Definizione 2.4), grazie ai calcoli sviluppati nell'Esempio 3.3:

$$\begin{aligned}\varsigma_l^2 &= \frac{1}{|\mathcal{T}|}\Lambda_{ll} = \\ &= \frac{1}{|\mathcal{T}|}(\mathbb{E}[\|\widehat{\boldsymbol{a}}_l^{GLS}\|^2] - \|\mathbb{E}[\widehat{\boldsymbol{a}}_l^{GLS}]\|^2) = \\ &= \frac{1}{|\mathcal{T}|}\int_{\mathcal{T}} \varsigma_l^2(t)dt.\end{aligned}$$

Per questo motivo, al fine di fornire un'indicazione della variabilità legata alla stima, in Figura 5.6 alle stime puntuali sono state accompagnate le curve corrispondenti a $\widehat{a}_l^{GLS} \pm 2\widehat{\varsigma}_l$ (linee tratteggiate), dove $\widehat{\varsigma}_l$ indica lo stimatore campionario di ς_l, ottenuto dalla stima campionaria $\widehat{\Lambda}$ di Λ come:

$$\widehat{\varsigma}_l^2 = \frac{1}{|\mathcal{T}|}\widehat{\Lambda}_{ll}, \quad l = 0, ..., L.$$

Tuttavia, è difficile individuare un'interpretazione per i parametri $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}$, non potendo riconoscere direttamente in questi le peculiarità del termine di *drift*; a questo scopo, i grafici contour risultano più intuitivi, potendo rilevare da essi le dipendenze spaziali.

In Figura 5.7, 5.8 e 5.9 sono quindi confrontate le mappe relative alla realizzazione di riferimento con le mappe ricostruite alla prima (stima OLS) e alla quinta iterazione (stima

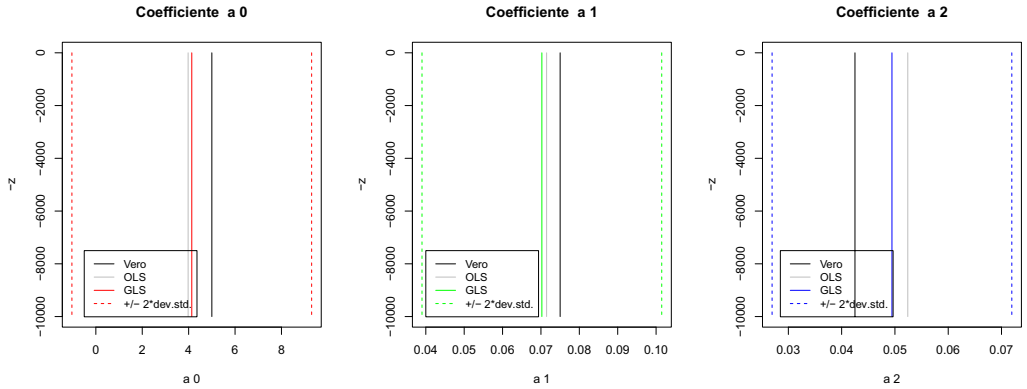

Figura 5.6: Coefficienti $\widehat{a}_l^{GLS}$, $l = 0, 1, 2$, stimati con il metodo GLS (linea continua colorata) a confronto con la stima OLS (linea continua grigia) e con i valori veri (linea continua nera). Le linee tratteggiate colorate forniscono un'indicazione della variabilità puntuale, attraverso la valutazione di $\widehat{a}_l^{GLS} \pm 2\widehat{\varsigma}_l$.

GLS) dell'Algoritmo 5.1, rispettivamente per il processo $\boldsymbol{\chi}_s$, per il *drift* m_s e per i residuo $\boldsymbol{\delta}_s$.

In particolare, dall'osservazione delle stime OLS e GLS nelle Figure 5.7, 5.9 (seconda e terza riga) e dal confronto di queste ultime con le mappe di riferimento (prima riga), è molto evidente l'effetto regolarizzante dell'interpolazione di kriging, già presente nel caso finito-dimensionale.

Questo effetto si riflette soprattutto in una previsione locale talvolta imprecisa per valori dell'ascissa z prossimi allo zero (seconda colonna di Figura 5.7 e 5.9), dove solo parte della ricostruzione è soddisfacente.

Questo risultato è tuttavia giustificato dal tipo di metodo che si sta usando, che si basa sulla minimizzazione di una variabilità globale del fenomeno e che individua la stima con uno stimatore lineare a coefficienti costanti: la presenza di variabilità non costante lungo l'ascissa può essere dunque un elemento problematico nella procedura stima, che in questi casi può risultare poco flessibile.

Il problema riscontrato è da imputarsi quasi completamente all'interpolazione di kriging in quanto, come si evince dal confronto della prima e della terza riga in Figura 5.8, la stima del termine di *drift* è invece piuttosto precisa.

Infine, il residuo, calcolato per differenza, risulta stazionario: questo è evidente dall'osservazione della Figura 5.10 dove il variogramma sperimentale del residuo finale, sovrapposto al modello di variogramma adattato con il metodo MC-bootstrap, è affiancato ai relativi intervalli di confidenza bootstrap (*cfr.* Capitolo 4, Sezione 4.2).

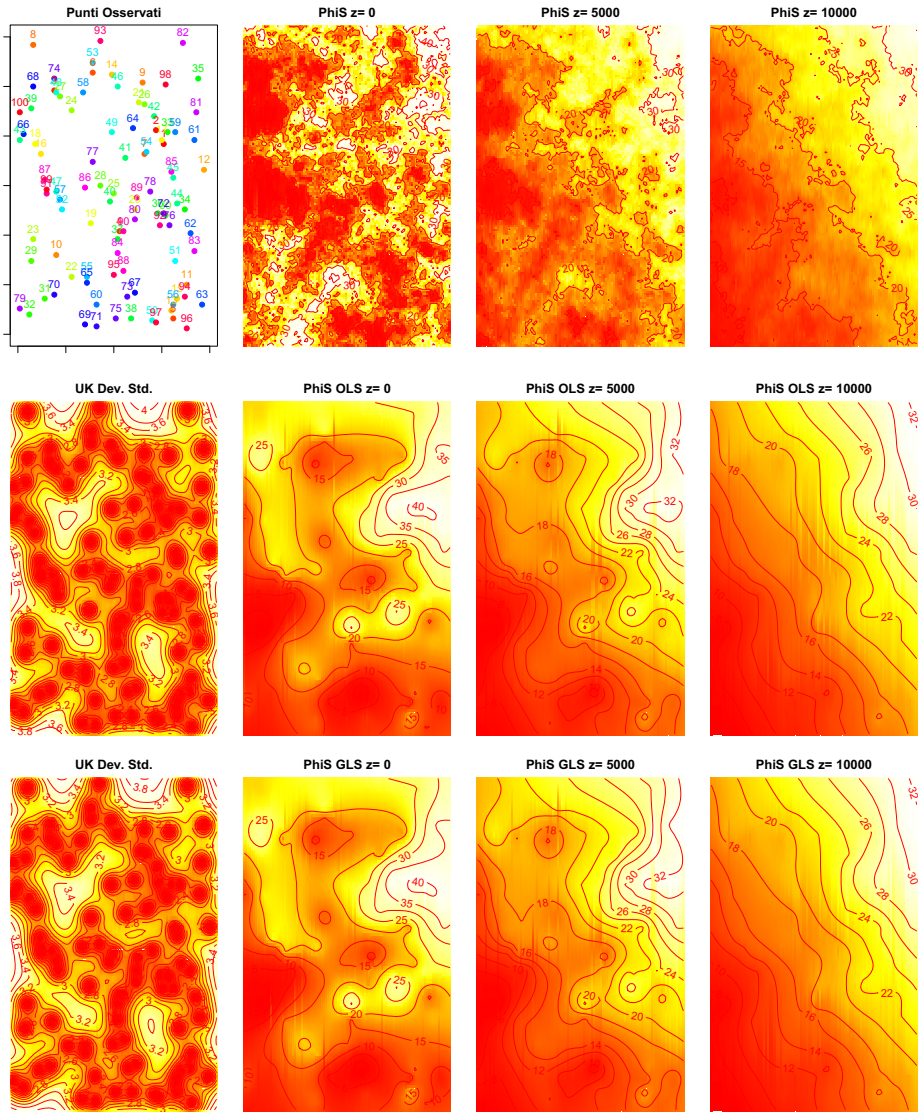

Figura 5.7: Confronto tra la realizzazione di riferimento (prima riga) del processo $\boldsymbol{\chi}_s$, l'interpolazione di Universal Kriging alla prima iterazione dell'Algoritmo 5.1 (stima OLS, seconda riga) e la stima di Universal Kriging dopo 5 iterazioni dell'Algoritmo 5.1 (stima GLS, terza riga). Per la realizzazione di riferimento: punti campionati $s_1, ..., s_{100}$ e contour-plot della realizzazione per $z = \{0; 5000; 10000\}$ m. Per le stime OLS e GLS: contour-plot della deviazione standard normalizzata, $\sqrt{\widehat{\sigma}^2_{UK}/10000}$ (a sinistra), e mappe di Universal Kriging per $z = \{0; 5000; 10000\}$ m.

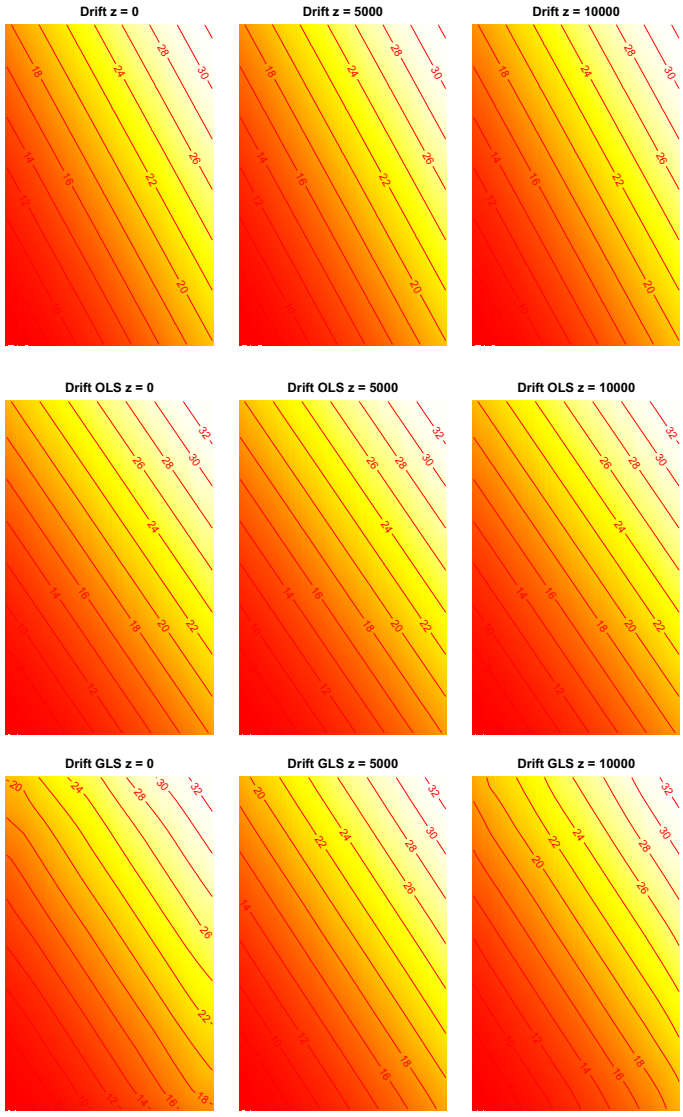

Figura 5.8: Confronto tra il *drift* generato (prima riga), la stima OLS del *drift* (seconda riga) e la stima GLS del *drift* ottenuta dopo 5 iterazioni dell'Algoritmo 5.1 (terza riga), per $z = \{0; 5000; 10000\}$ m.

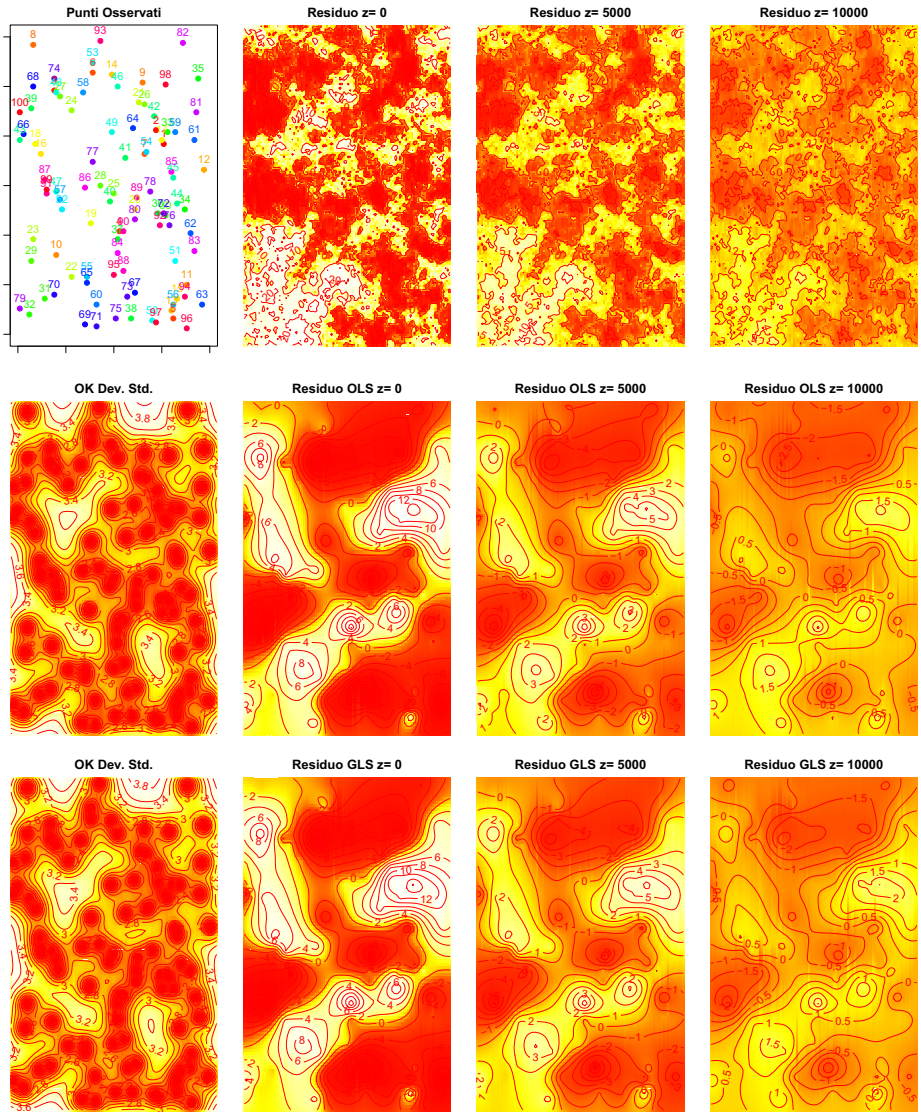

Figura 5.9: Confronto tra la realizzazione di riferimento (prima riga) del residuo $\boldsymbol{\delta}_s$, l'interpolazione di Ordinary Kriging alla prima iterazione dell'Algoritmo 5.1 (stima OLS, seconda riga) e la stima di Ordinary Kriging dopo 5 iterazioni dell'Algoritmo 5.1 (stima GLS, terza riga). Per la realizzazione di riferimento: punti campionati $s_1, ..., s_{100}$ e contour-plot della realizzazione per $z = \{0; 5000; 10000\}$ m. Per le stime OLS e GLS: contour-plot della deviazione standard normalizzata, $\sqrt{\widehat{\sigma}^2_{OK}/10000}$ (a sinistra), e mappe di Ordinary Kriging per $z = \{0; 5000; 10000\}$ m.

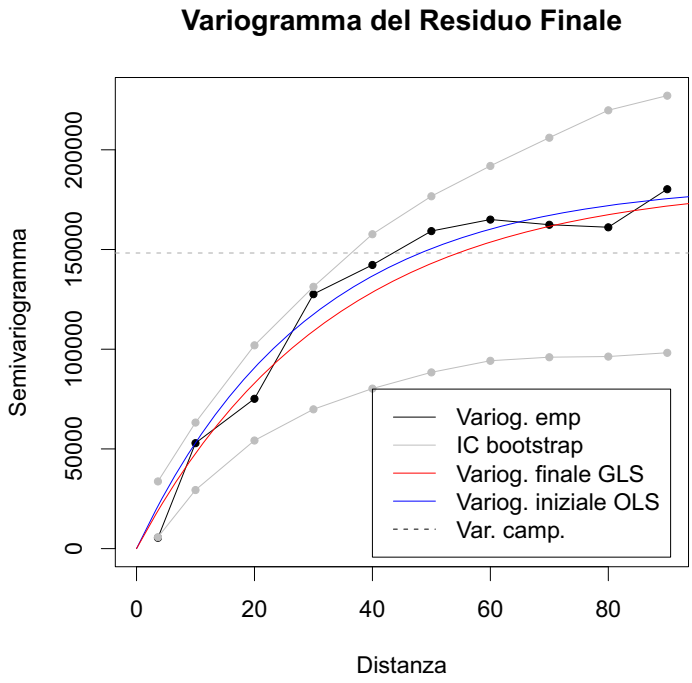

Figura 5.10: Variogramma finale del residuo stimato applicando l'Algoritmo 5.1; in particolare sono rappresentati lo stimatore empirico (linea nera), gli intervalli di confidenza bootstrap (linee grigie), il modello finale adattato con GLS-bootstrap (linea rossa), il modello iniziale adattato con OLS (linea blu) e la varianza campionaria del residuo finale.

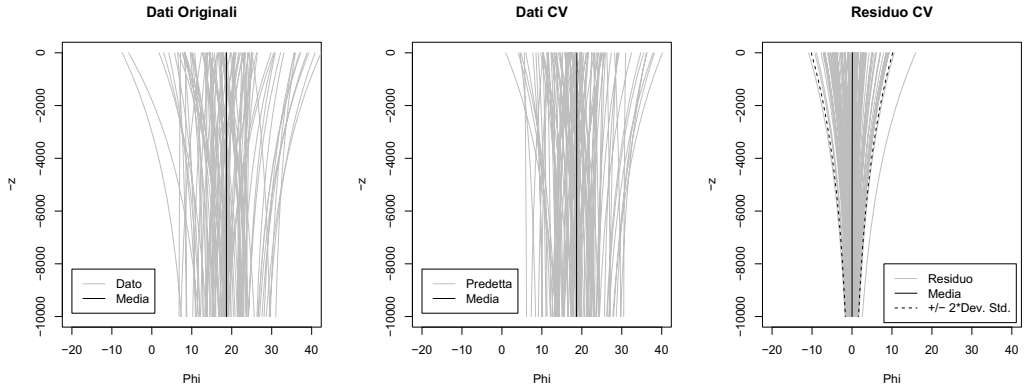

Figura 5.11: Analisi di Cross-Validazione. A sinistra: dati originali (in grigio) e relativa media campionaria (in nero). Al centro: dati predetti con Universal Kriging (in grigio) e loro media (in nero). A destra: differenza r_i^{CV}, $i = 1, ..., n$, tra i dati originali e i dati predetti (in grigio), loro media $\widehat{m}_r$ (in nero) e banda di confidenza puntuale $\widehat{m}_r(z) \pm 2\widehat{\sigma}_r(z)$ (in nero tratteggiato), dove $\widehat{\sigma}_r(z)$ è la deviazione standard stimata puntualmente dal residuo di cross-validazione.

5.3.2.3 Risultati di Cross-Validazione

L'analisi di cross-validazione *leave-one-out* svolta sul campione ha mostrato risultati in accordo con le osservazioni precedenti.

In particolare, dalla visualizzazione delle curve predette con il metodo di Universal Kriging, riportate in Figura 5.11, è evidente la problematicità dell'interpolazione per valori di z prossimi allo zero, mentre la precisione della previsione è crescente al crescere di z.

Lo scarto tra i valori osservati e i valori predetti è stato quantificato attraverso la statistica SSE_i:

$$SSE_i = \|\chi_{s_i} - \chi^*_{s_i}\|^2, \quad i = 1, ..., n.$$

I risultati dell'analisi di cross-validazione sono mostrati in Tabella 5.3. Nell'ultima riga della Tabella 5.3 è riportata la stima campionaria del valor medio della norma quadratica di χ_s:

$$\mathbb{E}_n[\|\chi_{\boldsymbol{s}}\|^2] = \frac{1}{n}\sum_{i=1}^{n} \|\chi_{s_i}\|^2.$$

Le statistiche della distribuzione empirica di SSE_i possono essere infatti confrontate con questa quantità, ottenendo un SSE relativo, $SSE^{rel.} = \frac{SSE}{\mathbb{E}_n[\|\chi_s\|^2]}$, le cui statistiche sono indicate nella terza colonna della Tabella 5.3.

Da queste ultime statistiche, si può notare che lo scarto relativo è mediamente inferiore al 2%, pertanto il comportamento dell'Algoritmo è nel complesso molto soddisfacente.

	SSE	$SSE^{(rel.)}$
Minimo	0.21	$5.38 \cdot 10^{-8}$
Mediana	29980.0	0.008
Media	68920.0	0.018
Massimo	682400.0	0.173
Dev. Std.	97713.2	0.025
Somma	6891570.0	1.750
$\mathbb{E}_n[\|\|\chi_s\|\|^2]$	3937577.1	

Tabella 5.3: Analisi di Cross-Validazione.. Statistiche della distribuzione dello scarto quadratico SSE e dello scarto quadratico relativo $SSE^{(rel.)}$

5.3.2.4 Convergenza dell'Algoritmo

Al fine di studiare la convergenza dell'Algoritmo 5.1 sono state considerate le differenze medie tra le stime del *drift* a iterazioni consecutive sui punti della griglia considerata, registrando ad ogni iterazione, per ogni z fissato:

$$d^{GLS}(\widehat{m}(z)^{N+1}, \widehat{m}(z)^N) = \frac{1}{N_g} \sum_{g=1}^{N_g} [\widehat{m}_{s_g}(z)^{N+1}, \widehat{m}_{s_g}(z)^N], \quad z \in \mathcal{T}, \tag{5.2}$$

dove $N_g = 10201$ indica il numero di punti sulla griglia, mentre $\widehat{m}_{s_g}(z)^N$ indica il *drift* ricostruito all'iterazione N in posizione s_g.

In Figura 5.12 è mostrato l'andamento della distanza $d^{GLS}(\widehat{m}^{N+1}, \widehat{m}^N)$ nelle prime 5 iterazioni dell'algoritmo, per 5 valori di z fissati ($z = \{0; 2500; 5000; 7500; 10000\}$ m) e mediando sui valori di $z \in [0, 10000]$ m, cioè considerando:

$$\overline{d}^{GLS}(\widehat{m}^{N+1}, \widehat{m}^N) = \frac{1}{|\mathcal{T}|} \int_{\mathcal{T}} d^{GLS}(\widehat{m}(z)^{N+1}, \widehat{m}(z)^N) dz. \tag{5.3}$$

Dall'immagine risulta chiaro come l'algoritmo si stabilizzi in modo molto veloce, mantenendo un'oscillazione al di sotto dello 0.5%, localmente e globalmente, già entro la quinta iterazione. Si noti inoltre che, a partire dalla seconda iterazione, le oscillazioni più ampie si verificano dove è maggiore la variabilità del fenomeno, ossia per valori di z prossimi allo 0. Questo è in accordo con quanto notato in precedenza riguardo alla maggiore difficoltà di interpolazione in presenza di alta variabilità del fenomeno, presente in $z = 0$ m a causa della forma esponenziale del termine di residuo.

5.3.2.5 Analisi di Robustezza

Durante lo studio del primo dataset non stazionario è stata sottoposta a verifica la robustezza del metodo rispetto a due fattori: il modello parametrico di variogramma, $\gamma(h, \boldsymbol{\vartheta})$, e la base per i coefficienti a_0, a_1, a_2 del *drift* individuato attraverso l'Algoritmo 5.2.

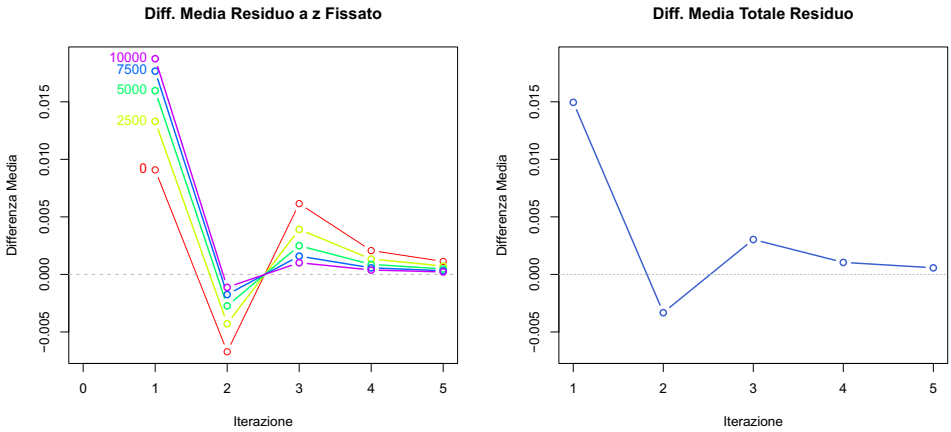

Figura 5.12: Analisi di convergenza. A sinistra: Andamento della distanza $d^{GLS}(\widehat{m}(z)^{N+1}, \widehat{m}(z)^{N})$, per $N = 1, ..., 5$, $z = \{0; 2500; 5000; 7500; 10000\}$ m. A destra: Andamento della distanza media $\overline{d}^{GLS}(\widehat{m}^{N+1}, \widehat{m}^{N})$.

Dal momento che a priori non è noto se i coefficienti del termine di *drift* siano costanti o funzionali, uno degli obiettivi dello studio è stata la valutazione del comportamento del metodo qualora ci si ponga nella situazione più generale di coefficienti funzionali, a fronte di un modello generativo a coefficienti costanti.

L'Algoritmo 5.1 è stato applicato quindi nei seguenti casi:

1. Modello di variogramma sferico e base costante per i coefficienti a_l, $l = 0, 1, 2$ (misclassificazione del modello di variogramma);

2. Modello di variogramma sferico e base di B-spline cubiche per i coefficienti a_l, $l = 0, 1, 2$, con nodi interni in $z_k = \{2500, 5000, 7500\}$ m (misclassificazione del modello di variogramma e della base dei coefficienti).

Il comportamento dell'Algoritmo 5.1 è stato valutato considerando le mappe di kriging generate e i risultati di cross-validazione.

In Figura 5.13, 5.14 e 5.15 sono confrontati i contour-plot relativi alla realizzazione di riferimento con i grafici ottenuti nei tre casi in analisi. Si può notare che in tutti i casi è evidente l'effetto regolarizzante del metodo kriging; tuttavia, non si registrano differenze significative nelle mappe ricostruite nei tre casi.

Infine, in Figura 5.16 sono riportati i variogrammi stimati dal residuo finale nelle tre situazioni analizzate. Dai grafici si può notare che in tutti i casi il residuo finale è caratterizzato da una varianza attorno al valore di 150000 e non esistono differenze sostanziali nel comportamento dei tre variogrammi, chiaramente stazionari.

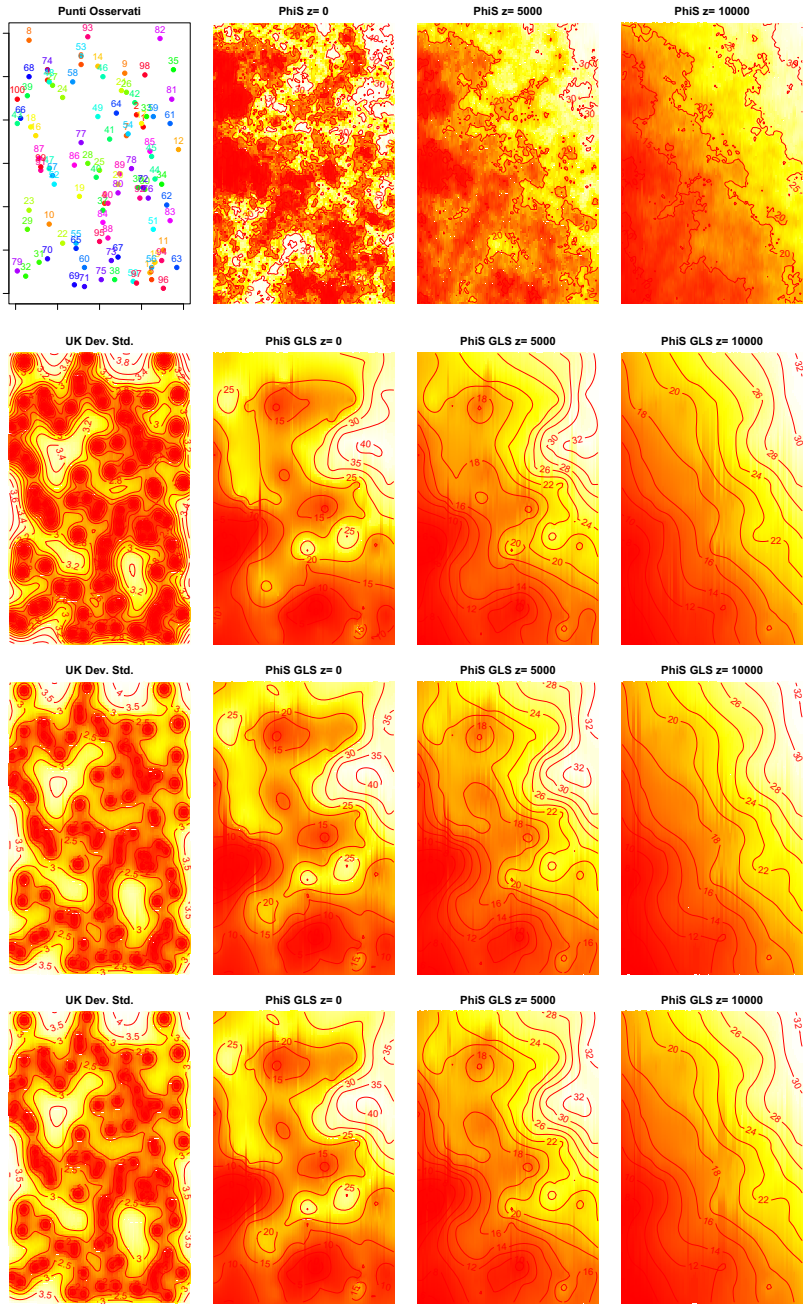

Figura 5.13: Analisi di Robustezza. Confronto tra la realizzazione di riferimento del processo $\boldsymbol{\chi}_s$ (prima riga) e l'interpolazione di Universal Kriging dopo 5 iterazioni dell'Algoritmo 5.1 ottenuta con modello di variogramma esponenziale e base costante per i coefficienti a_l (seconda riga), modello sferico e base costante (terza riga), modello sferico e base funzionale (quarta riga). Per la realizzazione di riferimento: punti campionati $s_1, ..., s_{100}$ e contour-plot della realizzazione per $z = \{0; 5000; 10000\}$ m. Per le stime: contour-plot della deviazione standard normalizzata, $\sqrt{\widehat{\sigma}^2_{UK}/10000}$ (a sinistra), e mappe di Universal Kriging per $z = \{0; 5000; 10000\}$ m.

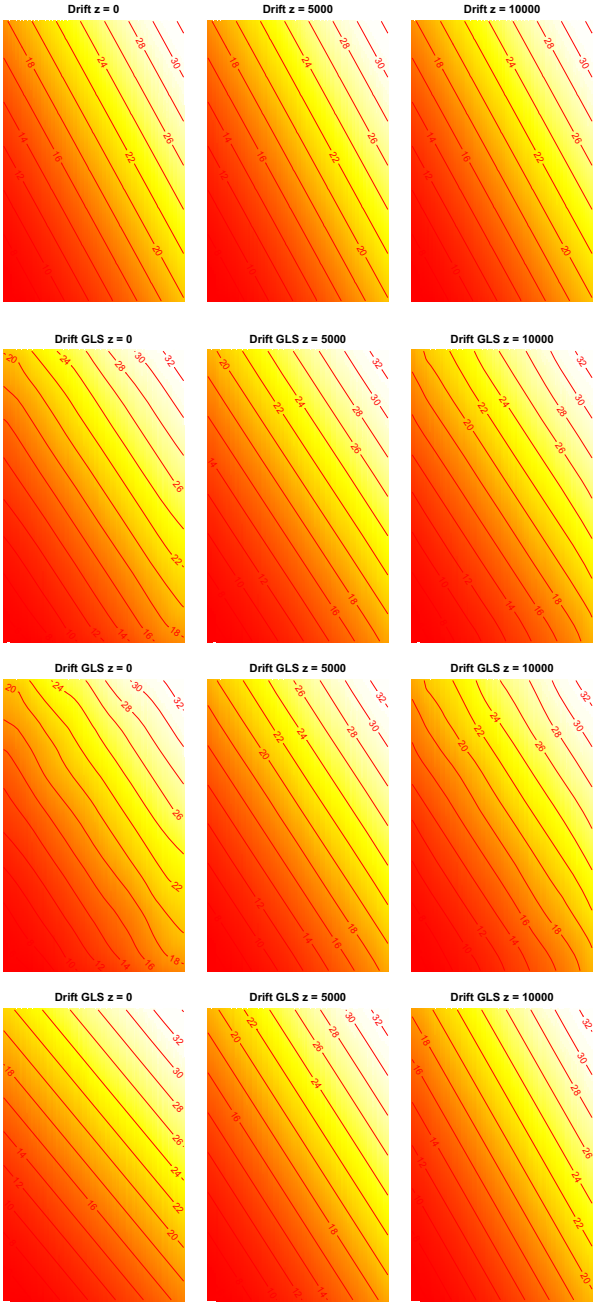

Figura 5.14: Analisi di Robustezza. Confronto tra il *drift* generato (prima riga), la stima GLS del *drift* ottenuta con modello di variogramma esponenziale e base costante per i coefficienti a_l (seconda riga), modello sferico e base costante (terza riga), modello sferico e base funzionale (quarta riga), per $z = \{0; 5000; 10000\}$ m.

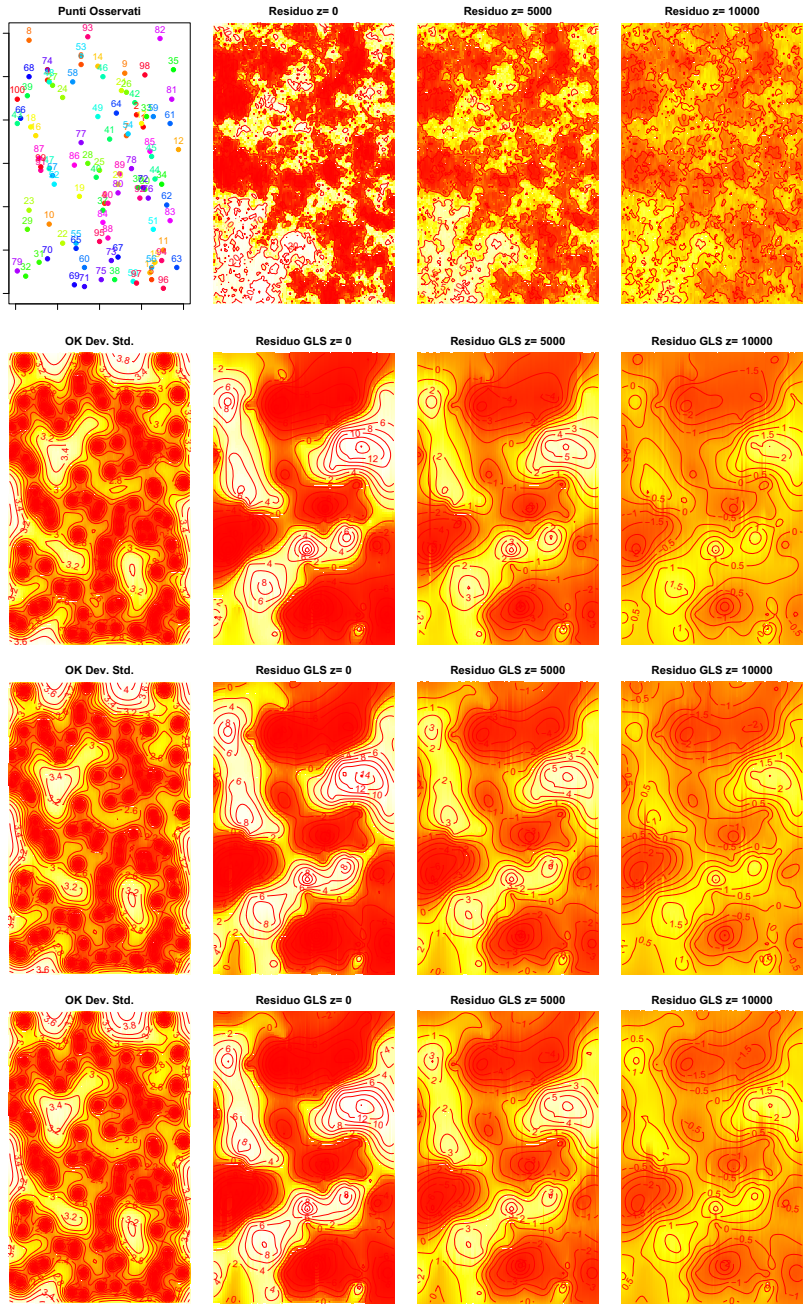

Figura 5.15: Analisi di Robustezza. Confronto tra la realizzazione di riferimento del residuo $\boldsymbol{\delta}_s$ (prima riga) e l'interpolazione di Ordinary Kriging dopo 5 iterazioni dell'Algoritmo 5.1 ottenuta con modello di variogramma esponenziale e base costante per i coefficienti a_l (seconda riga), modello sferico e base costante (terza riga), modello sferico e base funzionale (quarta riga). Per la realizzazione di riferimento: punti campionati $s_1, ..., s_{100}$ e contour-plot della realizzazione per $z = \{0; 5000; 10000\}$ m. Per le stime: contour-plot della deviazione standard normalizzata, $\sqrt{\widehat{\sigma}^2_{OK}/10000}$ (a sinistra), e mappe di Universal Kriging per $z = \{0; 5000; 10000\}$ m.

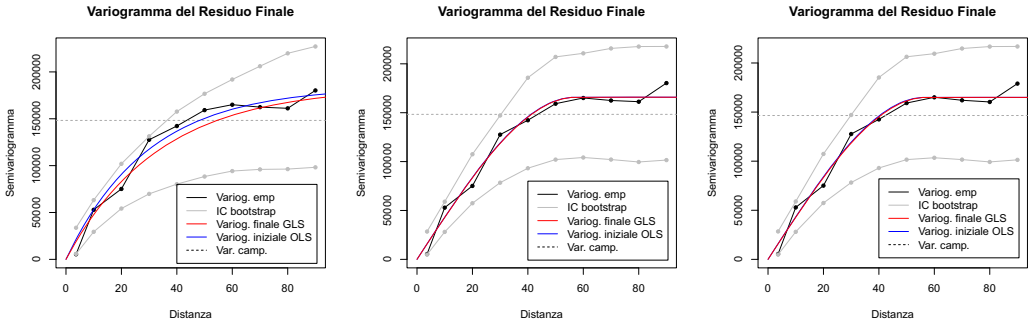

Figura 5.16: Analisi di Robustezza. Confronto dei variogrammi finali del residuo stimati applicando l'Algoritmo 5.1, nei casi di modello di variogramma esponenziale e base costante per i coefficienti $a_{\boldsymbol{l}}$ (a sinistra), modello sferico e base costante (al centro), modello sferico e base funzionale (a destra). In particolare sono rappresentati lo stimatore empirico (linea nera), gli intervalli di confidenza bootstrap (linee grigie), il modello finale adattato con GLS-bootstrap (linea rossa), il modello iniziale adattato con OLS (linea blu) e la varianza campionaria del residuo finale.

I risultati di cross-validazione si sono dimostrati in accordo con le osservazioni precedenti, come evidenziato dalla distribuzione dello scarto quadratico SSE_i, e dello scarto relativo $SSE_i^{(rel)}$, le cui statistiche sono riportate in Tabella 5.4.

Dalla Tabella 5.4 si può notare che la distribuzione del SSE_i nel secondo e terzo caso non è sostanzialmente differente rispetto al primo caso, presentando statistiche molto simili. Non si registra dunque un'influenza significativa, sul potere previsivo del metodo, del cambiamento del modello di variogramma, né, a modello fissato, della diversa scelta della base per i coefficienti $a_{\boldsymbol{l}}$.

Come si evince dai grafici in Figura 5.17, in tutti i casi il residuo di cross-validazione raggiunge valori più alti per valori di z prossimi a zero, laddove l'incertezza è massima: questa è un'ulteriore indicazione dell'effetto regolarizzante del metodo di kriging.

Dalle precedenti analisi si può dunque concludere che la scelta iniziale del modello parametrico di variogramma $\gamma(h;\boldsymbol{\vartheta})$ non è particolarmente influente sul potere previsivo del metodo di Universal Kriging associato all'Algoritmo 5.1, fornendo buoni risultati in termini di stazionarietà del residuo (Figura 5.16).

Occorre tuttavia precisare che la robustezza del metodo è da intendersi rispetto alla forma del modello di variogramma, sferico o esponenziale, non rispetto alle stime dei parametri $\boldsymbol{\vartheta}$, che si riflettono sulle caratteristiche strutturali del modello stimato, ovvero sulle stime di *sill*, *range* e *nugget*. Queste ultime, al contrario, sono molto influenti sul potere previsivo del metodo di Universal Kriging; tuttavia, al termine della procedura definita dall'Algoritmo 5.1, esse si assestano in tutti i casi presentati su valori molto simili tra loro ed questo è il motivo principale per il quale il metodo risulta robusto rispetto alla forma del variogramma.

Infine, il metodo è risultato robusto rispetto al tipo di base fissata per i coefficienti del

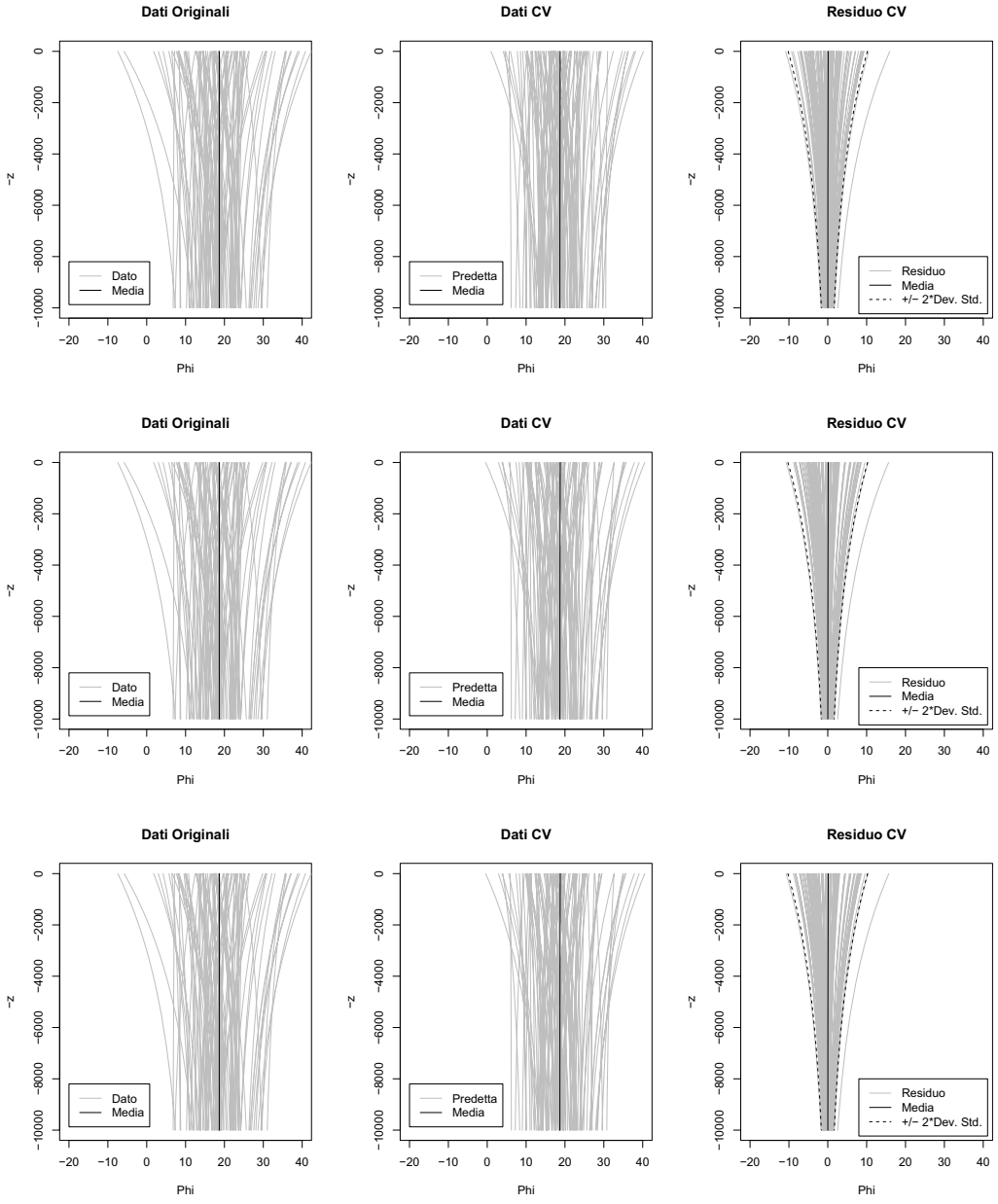

Figura 5.17: Analisi di robustezza. A confronto dati predetti con cross-validazione per Universal Kriging con modello di variogramma esponenziale e base costante per i coefficienti a_l (prima riga), modello sferico e base costante (seconda riga), modello sferico e base funzionale (terza riga). A sinistra: dati originali (in grigio) e relativa media campionaria (in nero). Al centro: dati predetti con Universal Kriging (in grigio) e loro media (in nero). A destra: differenza r_i^{CV}, $i = 1, ..., n$, tra i dati originali e i dati predetti (in grigio), loro media $\widehat{m}_r$ (in nero) e banda di confidenza puntuale $\widehat{m}_r(z) \pm 2\widehat{\sigma}_r(z)$ (in nero tratteggiato), dove $\widehat{\sigma}_r(z)$ è la deviazione standard stimata puntualmente dal residuo di cross-validazione.

	Modello Esponenziale Base Costante		Modello Sferico Base Costante		Modello Sferico Base Funzionale	
	SSE	$SSE^{(rel.)}$	SSE	$SSE^{(rel.)}$	SSE	$SSE^{(rel.)}$
Minimo	0.21	$5.38{\cdot}10^{-8}$	8.346	$2.12{\cdot}10^{-6}$	7.24	$1.84{\cdot}10^{-6}$
Mediana	29980.0	0.008	29710.0	0.0075	29780.0	0.0076
Media	68920.0	0.018	69650.0	0.0177	69720.0	0.0177
Massimo	682400.0	0.173	656100.0	0.166	661300.0	0.168
Dev. Std.	97713.2	0.025	93273.7	0.024	93614.7	0.024
Somma	6891570.0	1.750	6964502.1	1.768	6971918.7	1.771
$\mathbb{E}_n[\|\chi_s\|^2]$			3937577.1			

Tabella 5.4: Analisi di Robustezza. Statistiche della distribuzione dello scarto quadratico SSE e dello scarto quadratico relativo $SSE^{(rel.)}$ per le tre situazioni considerate: modello esponenziale e base costante, modello sferico e base costante, modello sferico e base funzionale.

drift, non evidenziando differenze marcate nel potere previsivo al variare della base scelta.

5.4 Applicazione a Dati Sintetici con Modello di *Drift* a Coefficienti Funzionali

Gli Algoritmi 5.1 e 5.2 introdotti nelle Sezioni 5.1 e 5.2 hanno fornito risultati soddisfacenti dal punto di vista della stima, della previsione e della robustezza nell'applicazione a dati non stazionari con *drift* a coefficienti costanti.

L'obiettivo della presente sezione è l'applicazione della metodologia proposta a dati non stazionari, caratterizzati da una media non costante lungo l'ascissa z, valutando in particolare il comportamento del metodo in termini di precisione della stima, potere previsivo e robustezza rispetto alla misclassificazione del modello parametrico di variogramma e alla numerosità del campione.

La struttura della sezione è analoga alla struttura adottata per la Sezione 5.3, con la presentazione iniziale dei dati seguita dall'illustrazione dei risultati di simulazione.

5.4.1 I Dati

Il secondo dataset funzionale georeferenziato usato per le simulazioni è stato generato a partire dal medesimo residuo usato per la costruzione del primo campione, sommandovi un *drift* la cui variabilità dipenda anche dall'ascissa z.

Il *drift* considerato per la seconda parte dello studio di simulazione è della forma:

$$m_s(z) = a_0(z) + a_1(z)x + a_2(z)y, \quad s = (x, y) \in D,\, z \in \mathcal{T},$$

dove a_0, a_1, a_2 sono coefficienti funzionali appartenenti allo spazio $H = L^2$.

	β_1	β_2	β_3	β_4	β_5	β_6	β_7
a_0	0.71431	0.71427	0.71423	0.71418	0.71440	0.71433	0.71425
a_1	$8.2{\cdot}10^{-5}$	$-1.6{\cdot}10^{-5}$	$11.7{\cdot}10^{-5}$	$6.2{\cdot}10^{-5}$	$18.7{\cdot}10^{-5}$	$14.34{\cdot}10^{-5}$	$18.9{\cdot}10^{-5}$
a_2	$6.6{\cdot}10^{-5}$	$3.5{\cdot}10^{-5}$	$19.1{\cdot}10^{-5}$	$16.2{\cdot}10^{-5}$	$-2.7{\cdot}10^{-5}$	$12.6{\cdot}10^{-5}$	$-6.1{\cdot}10^{-5}$

Tabella 5.5: Coefficienti β_j, $j = 1, ..., 7$, generati per la costruzione del *drift* funzionale.

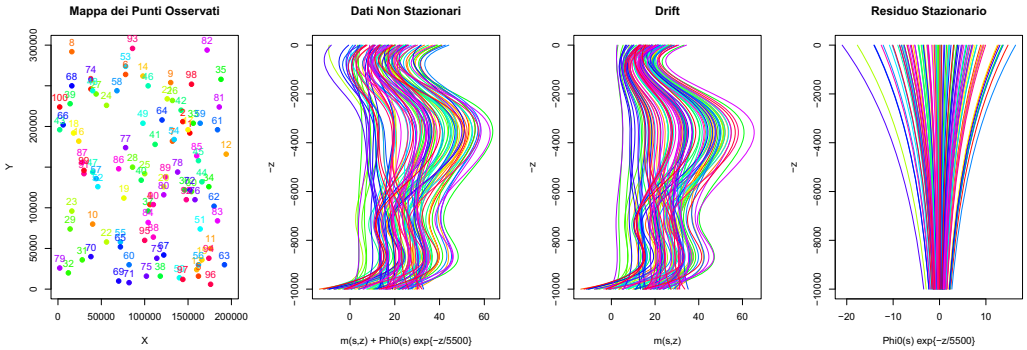

Figura 5.18: Da sinistra a destra: Mappa dei punti campionati $s_1, ..., s_{100}$; secondo dataset funzionale sintetico non stazionario $\chi_{\boldsymbol{s}} = (\chi_{s_1}, ..., \chi_{s_{100}})$; *drift* $m_{\boldsymbol{s}} = (m_{s_1}, ..., m_{s_{100}})$; residuo stazionario $\delta_{\boldsymbol{s}} = (\delta_{s_1}, ..., \delta_{s_{100}})$.

Tali coefficienti sono stati costruiti su una base $\{B_j\}$ di B-spline cubiche $(m = 4)$, fissando $n_k = 3$ nodi interni $z_k = (2500, 5000, 7500)$ m:

$$a_l(z) = \sum_{j=1}^{m+n_k} \beta_j B_j(z),$$

dove si indicano con β_j i coefficienti dell'espansione, per $j = 1, ..., m + n_k$. I coefficienti delle spline cubiche generati per la simulazione sono riportati in Tabella 5.5.

I grafici relativi ai dataset generati sono rappresentati in Figura 5.18, 5.19 e 5.20. Il grafico contour del residuo è il medesimo della Figura 5.4 ed è pertanto omesso.

Infine, con lo scopo di testare la robustezza degli Algoritmi 5.1 e 5.2 al variare della numerosità campionaria del dataset, sono stati considerati dei sottocampioni di $(\chi_{s_1}, ..., \chi_{s_{100}})$, selezionando in particolare i primi 30 dati e i primi 10 dati del campione. I dataset ridotti sono riportati in Figura 5.21 e 5.22.

5.4.2 I Risultati di Simulazione

Analogamente all'analisi svolta nella Sezione 5.3, saranno ora applicati in successione gli Algoritmi 5.2 e 5.1; in un secondo momento sarà valutata la convergenza del metodo, svolta un'a-

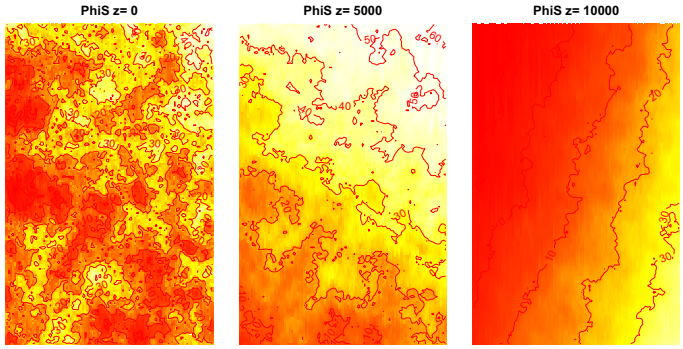

Figura 5.19: Contour-plot della realizzazione del processo $\boldsymbol{\chi}_s$ sull'intera griglia di simulazione, per $z = \{0; 5000; 10000\}$ m.

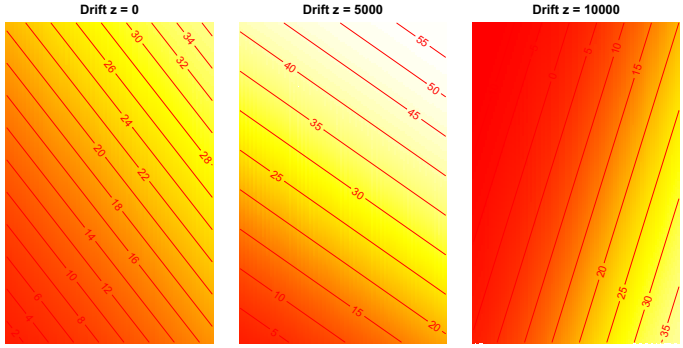

Figura 5.20: Contour-plot del termine di *drift* m_s sull'intera griglia di simulazione, per $z = \{0; 5000; 10000\}$ m.

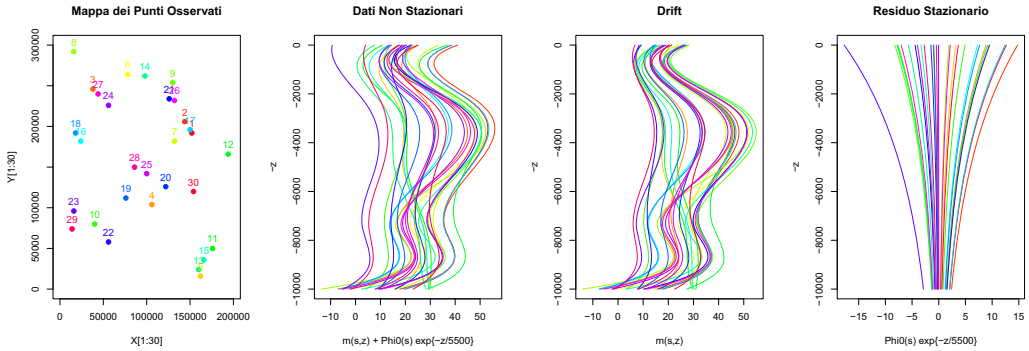

Figura 5.21: Da sinistra a destra: Mappa dei punti campionati $s_1, ..., s_{30}$; dataset funzionale sintetico non stazionario $\chi_{\boldsymbol{s}} = (\chi_{s_1}, ..., \chi_{s_{30}})$; *drift* $m_{\boldsymbol{s}} = (m_{s_1}, ..., m_{s_{30}})$; residuo stazionario $\delta_{\boldsymbol{s}} = (\delta_{s_1}, ..., \delta_{s_{30}})$.

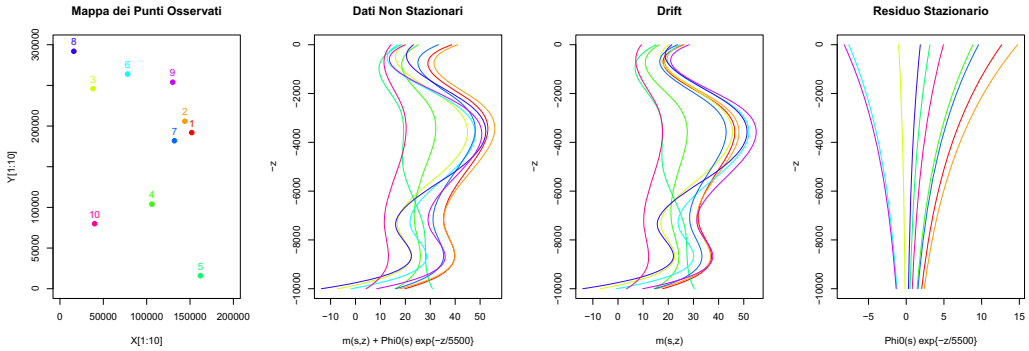

Figura 5.22: Da sinistra a destra: Mappa dei punti campionati $s_1, ..., s_{10}$; dataset funzionale sintetico non stazionario $\chi_{\boldsymbol{s}} = (\chi_{s_1}, ..., \chi_{s_{10}})$; *drift* $m_{\boldsymbol{s}} = (m_{s_1}, ..., m_{s_{10}})$; residuo stazionario $\delta_{\boldsymbol{s}} = (\delta_{s_1}, ..., \delta_{s_{10}})$.

(k)	k	$f_{\boldsymbol{l}}^k$	SSE_k
(1)	3	$\{1, x, y\}$	149352.2
(2)	8	$\{1, x^2, y^2, xy\}$	219273.4
(3)	7	$\{1, x^2, y^2\}$	292113.3
(4)	6	$\{1, xy\}$	535381.8
(5)	2	$\{1, y\}$	547783.8
(6)	5	$\{1, y^2\}$	710451.1
(7)	1	$\{1, x\}$	1063733.0
(8)	4	$\{1, x^2\}$	1081310.1

Tabella 5.6: Ordinamento dei *drift* con criterio previsivo ottenuto dall'applicazione dell'Algoritmo 5.2.

nalisi di cross-validazione ed infine una verifica di robustezza rispetto alla misclassificazione del modello di variogramma e alla numerosità del campione.

5.4.2.1 Ordinamento dei *Drift*

A partire dalle otto forme funzionali riportate in Tabella 5.1, è stato selezionato il *drift* ottimale con il criterio previsivo, attraverso l'applicazione dell'Algoritmo 5.2. L'ordinamento ottenuto è riportato in Tabella 5.6.

Anche in questo caso, il modello di *drift* selezionato corrisponde al *drift* di generazione, ovvero alla forma funzionale 3:

$$m_s(z) = a_0(z) + a_1(z)x + a_2(z)y, \quad s = (x, y) \in D,\ z \in \mathcal{T}. \tag{5.4}$$

Al fine di verificare la stazionarietà del residuo derivante dalla scelta del *drift* (5.4), è stato applicata un'iterazione dell'Algoritmo 5.1, valutando in particolare i variogrammi

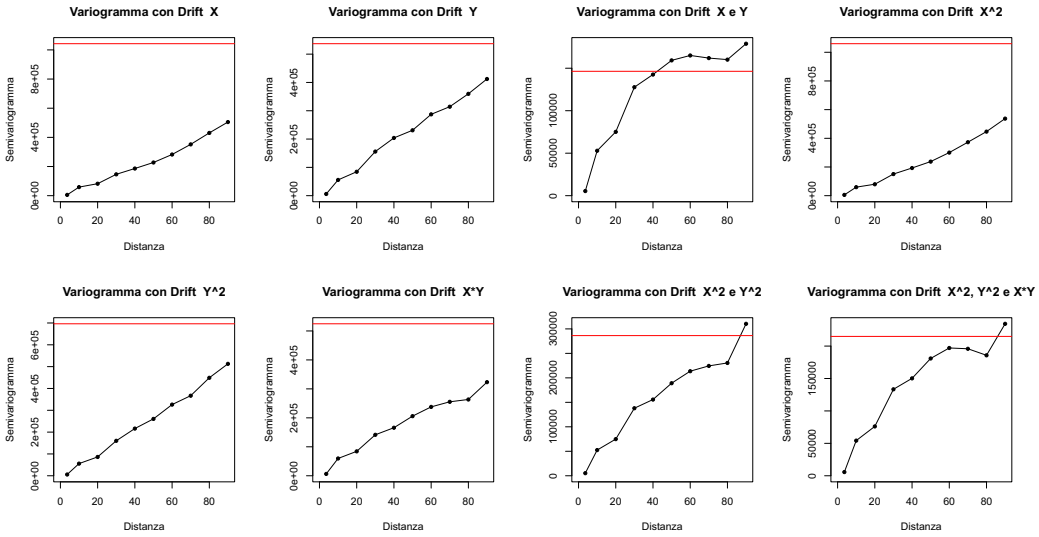

Figura 5.23: Variogrammi dei residui $\delta_{\boldsymbol{s}}^k$, al variare delle forme di *drift* in Tabella 5.1, ottenuti da una prima iterazione dell'Algoritmo 5.1. Nel confronto tra i pannelli, si presti attenzione alle diverse scale per l'asse delle ordinate.

sperimentali dei residui calcolati per ciascuna forma del *drift*.

In Figura 5.23 sono riportate tali stime empiriche: come si può osservare, il variogramma corrispondente al modello di *drift* selezionato è caratterizzato dalla forma più stazionaria, mentre i variogrammi empirici relativi ai modelli 5, 1 e 4, che sono posti nelle ultime posizioni dell'ordinamento proposto dall'Algoritmo 5.2, non hanno ancora le caratteristiche di stazionarietà attese per il residuo (in particolare si noti la crescita superquadratica delle stime relative ai modelli 1 e 4).

5.4.2.2 Disaccoppiamento del *Drift* e Universal Kriging

Una volta selezionato il *drift* ottimale con l'ausilio dell'Algoritmo 5.2, è stato applicato l'Algoritmo 5.1 con procedimento analogo a quello descritto nella Sottosezione 5.3.2.2.

Nelle Figure 5.24 sono riportati i coefficienti di generazione a confronto con i coefficienti stimati alla prima e all'ultima iterazione del procedimento: dall'immagine si può notare che il procedimento fornisce una stima molto buona per il secondo e terzo coefficiente del termine di *drift*, mentre la stima del primo coefficiente è leggermente meno accurata. Quest'ultima è tuttavia legata a una variabilità maggiore, valutata anche in questo caso attraverso $\widehat{\varsigma}_l^2$, $l = 0, 1, 2$, come riportato nel grafico.

In Figura 5.25, 5.26 e 5.27 sono invece mostrati i grafici contour ottenuti per simulazione alla prima e ultima iterazione dell'Algoritmo 5.1 a confronto con le mappe di riferimento.

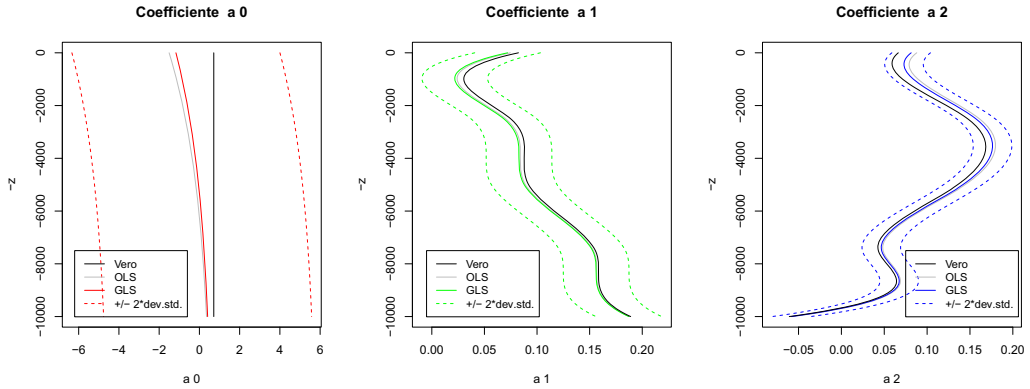

Figura 5.24: Coefficienti del *drift* $\widehat{a}_l^{GLS}$, $l = 0, 1, 2$, stimati con il metodo GLS (linea continua colorata) a confronto con la stima OLS (linea continua grigia) e con i valori veri (linea continua nera). Le linee tratteggiate colorate forniscono un'indicazione della variabilità puntuale, attraverso la valutazione di $\widehat{a}_l^{GLS} \pm 2\widehat{\varsigma}_l$.

Anche in questo caso è evidente l'effetto regolarizzante dell'interpolazione di Ordinary e Universal Kriging, particolarmente significativo per valori di z prossimi allo zero. La stima finale del termine di *drift* è tuttavia risultata molto aderente al *drift* originale, in accordo con le buone stime dei coefficienti fornite dal metodo. Non si registrano invece differenze sostanziali tra le stime OLS e le stima GLS.

Infine, la verifica della stazionarietà del residuo δ_s è avvenuta attraverso la visualizzazione del variogramma sperimentale riportato in Figura 5.28, caratterizzato da una concavità verso il basso in corrispondenza dell'origine e dalla presenza di un *sill* non lontano dalla varianza campionaria.

5.4.2.3 Risultati di Cross-Validazione

Il comportamento previsivo del procedimento di Universal Kriging è stato valutato attraverso un'analisi di cross-validazione di tipo *leave-one-out*.

In Tabella 5.7 sono riportate le statistiche relative allo scarto quadratico SSE_i e allo scarto quadratico relativo $SSE_i^{(rel.)}$, risultanti dalla simulazione.

Anche in questo caso la previsione è nel complesso molto buona, presentando uno scarto medio relativo inferiore all'1%.

Dai grafici in Figura 5.29 si può infine notare che, anche nel caso di *drift* variabile lungo l'ascissa z, il residuo di cross-validazione è superiore laddove la variabilità del fenomeno è più alta, ovvero per valori di z prossimi allo zero.

Si noti in particolare che la forma del residuo di cross-validazione relativo al dataset in esame è molto simile alla forma dello stesso residuo calcolato sul dataset trattato nella Sezione 5.3 (*cfr.* Figura 5.11). Tale somiglianza indica che la difficoltà previsiva del metodo

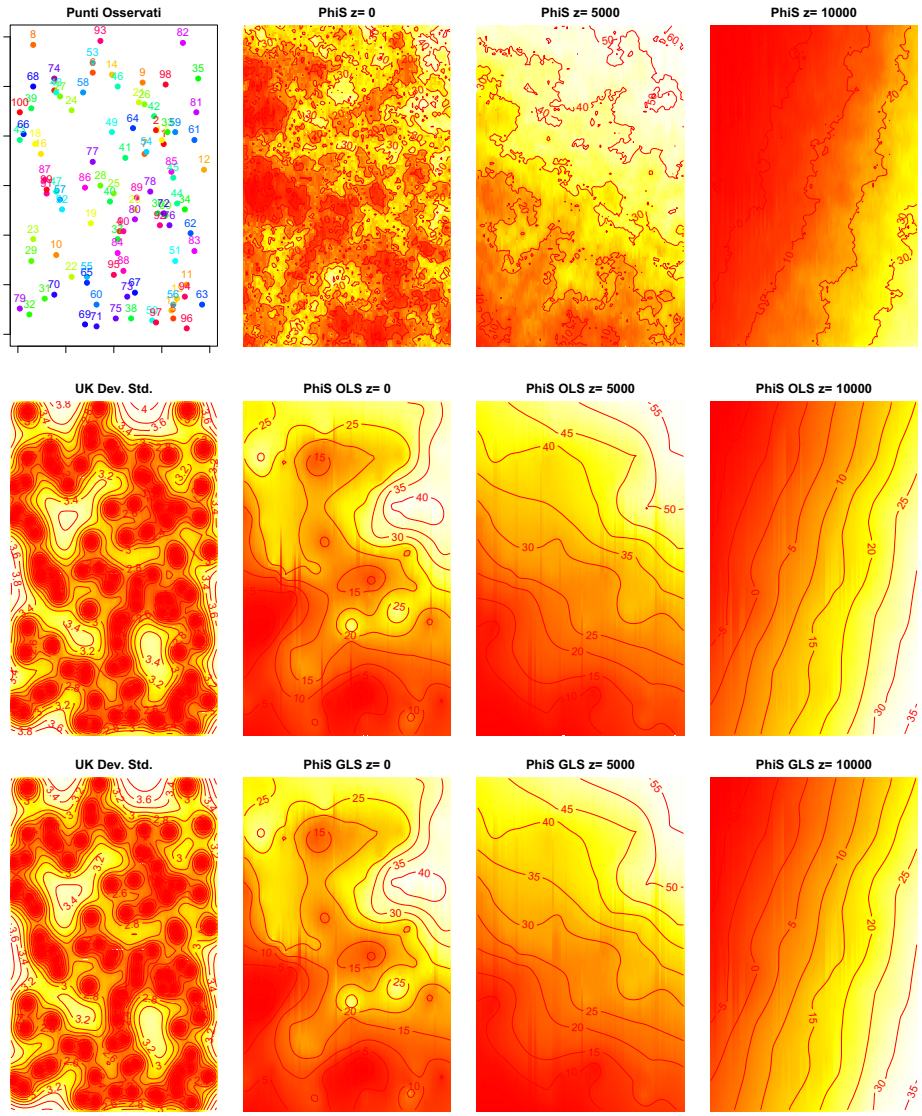

Figura 5.25: Confronto tra la realizzazione di riferimento (prima riga) del processo $\boldsymbol{\chi}_s$, l'interpolazione di Universal Kriging alla prima iterazione dell'Algoritmo 5.1 (stima OLS, seconda riga) e la stima di Universal Kriging dopo 5 iterazioni dell'Algoritmo 5.1 (stima GLS, terza riga). Per la realizzazione di riferimento: punti campionati $s_1, ..., s_{100}$ e contour-plot della realizzazione per $z = \{0; 5000; 10000\}$ m. Per le stime OLS e GLS: contour-plot della deviazione standard normalizzata, $\sqrt{\widehat{\sigma}^2_{UK}/10000}$ (a sinistra), e mappe di Universal Kriging per $z = \{0; 5000; 10000\}$ m.

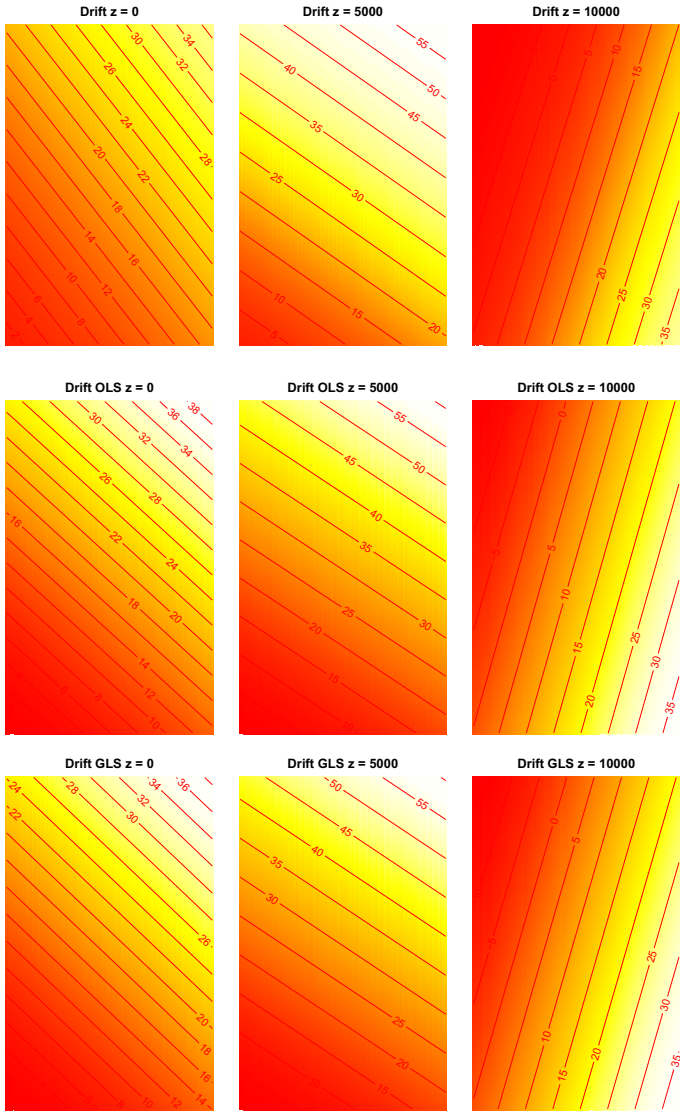

Figura 5.26: Confronto tra il *drift* di riferimento (prima riga), la stima OLS del *drift* (seconda riga) e la stima GLS del *drift* ottenuta dopo 5 iterazioni dell'Algoritmo 5.1 (terza riga), per $z = \{0; 5000; 10000\}$ m.

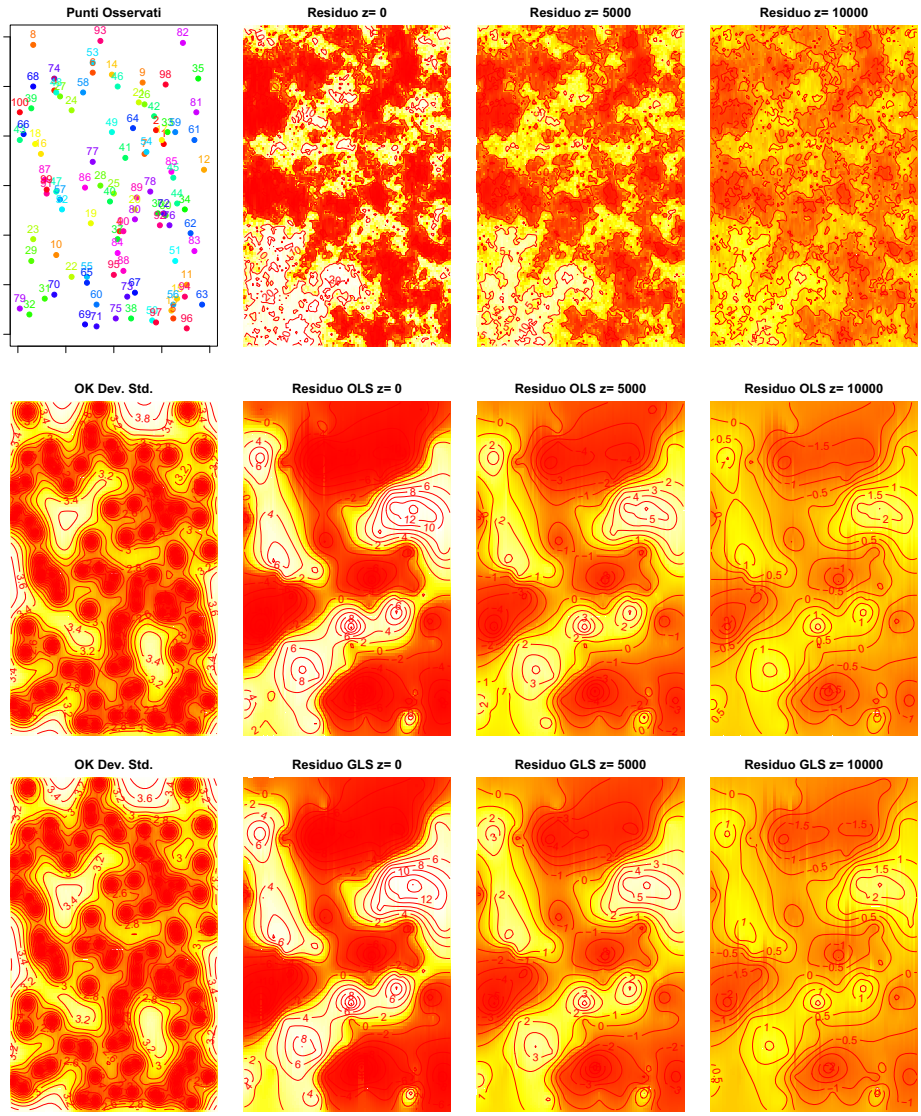

Figura 5.27: Confronto tra la realizzazione di riferimento (prima riga) del residuo $\boldsymbol{\delta}_s$, l'interpolazione di Ordinary Kriging alla prima iterazione dell'Algoritmo 5.1 (stima OLS, seconda riga) e la stima di Ordinary Kriging dopo 5 iterazioni dell'Algoritmo 5.1 (stima GLS, terza riga). Per la realizzazione di riferimento: punti campionati $s_1, ..., s_{100}$ e contour-plot della realizzazione per $z = \{0; 5000; 10000\}$ m. Per le stime OLS e GLS: contour-plot della deviazione standard normalizzata, $\sqrt{\widehat{\sigma}^2_{OK}/10000}$ (a sinistra), e mappe di Ordinary Kriging per $z = \{0; 5000; 10000\}$ m.

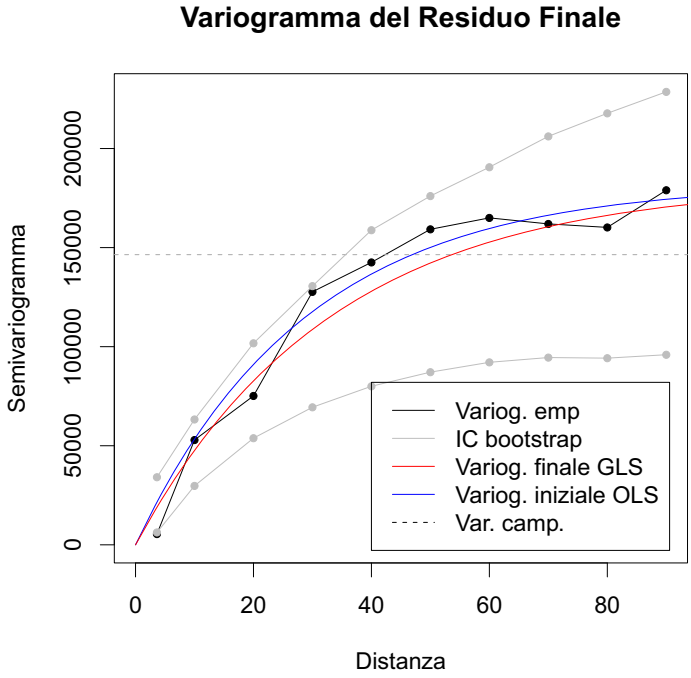

Figura 5.28: Variogramma finale del residuo stimato applicando l'Algoritmo 5.1; in particolare sono rappresentati lo stimatore empirico (linea nera), gli intervalli di confidenza bootstrap (linee grigie), il modello finale adattato con GLS-bootstrap (linea rossa), il modello iniziale adattato con OLS (linea blu) e la varianza campionaria del residuo finale.

	SSE	$SSE^{(rel.)}$
Minimo	0.67	$8.73 \cdot 10^{-8}$
Mediana	30050.0	$3.927 \cdot 10^{-3}$
Media	68880.0	$9.001 \cdot 10^{-3}$
Massimo	680400.0	$8.891 \cdot 10^{-2}$
Dev. Std.	97556.5	0.01275
Somma	6887798.2	0.90001
$\mathbb{E}_n[\|\|\chi_s\|\|^2]$	7652444.0	

Tabella 5.7: Analisi di cross-validazione. Statistiche della distribuzione dello scarto quadratico SSE e dello scarto quadratico relativo $SSE^{(rel.)}$.

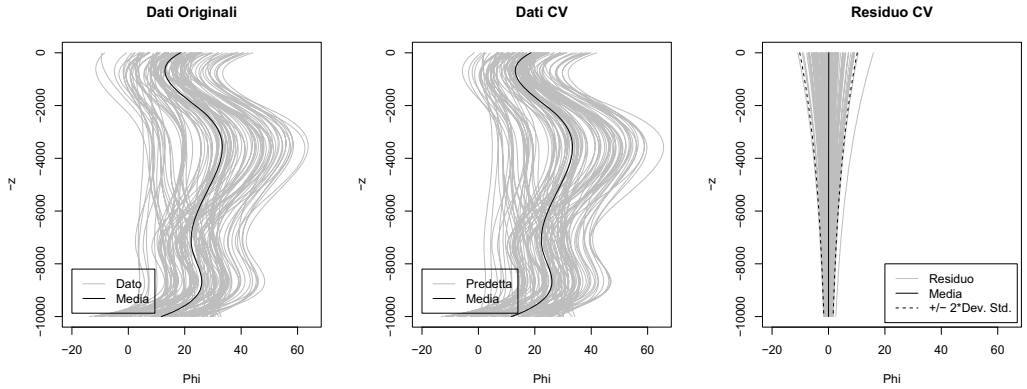

Figura 5.29: Analisi di Cross-Validazione. A sinistra: dati originali (in grigio) e relativa media campionaria (in nero). Al centro: dati predetti con Universal Kriging (in grigio) e loro media (in nero). A destra: differenza r_i^{CV}, $i = 1, ..., n$, tra i dati originali e i dati predetti (in grigio), loro media $\widehat{m}_r$ (in nero) e banda di confidenza puntuale $\widehat{m}_r(z) \pm 2\widehat{\sigma}_r(z)$ (in nero tratteggiato), dove $\widehat{\sigma}_r(z)$ è la deviazione standard stimata puntualmente dal residuo di cross-validazione.

è riscontrata in particolar modo nell'estrapolazione del termine di residuo, caratterizzato da variabilità non costante lungo l'ascissa z.

5.4.2.4 Convergenza dell'Algoritmo

La convergenza dell'Algoritmo 5.1 è stata valutata localmente e globalmente attraverso le misure $d^{GLS}(\widehat{m}(z)^{N+1}, \widehat{m}(z)^N)$ e $\overline{d}^{GLS}(\widehat{m}^{N+1}, \widehat{m}^N)$, introdotte nella Sottosezione 5.3.2.4 dalla (5.2) e dalla (5.3).

Come si può notare dalla Figura 5.30, anche in questo caso la stima del *drift* fornita dal metodo si stabilizza molto velocemente, indicazione della convergenza dell'Algoritmo 5.1 entro le prime cinque iterazioni.

Le maggiori oscillazioni tra iterazioni consecutive sono registrate per valori di z prossimi allo zero, come conseguenza della maggiore variabilità del fenomeno.

5.4.2.5 Analisi di Robustezza

Analisi di Robustezza rispetto alla Numerosità del Campione Nel corso delle analisi precedenti, il comportamento degli Algoritmi 5.1 e 5.2 è stato studiato a partire da dataset di numerosità 100; tuttavia, dal punto di vista applicativo, accade spesso che la numerosità del campione sia invece molto scarsa. Per numerosità del campione si intende il numero n di siti, $s_1, ..., s_n$, presso i quali è registrato il dato funzionale: si vuole pertanto condurre un'analisi rispetto alla risoluzione spaziale del campionamento, non rispetto alla qualità dell'acquisizione dei dati lungo l'ascissa z della variabile funzionale.

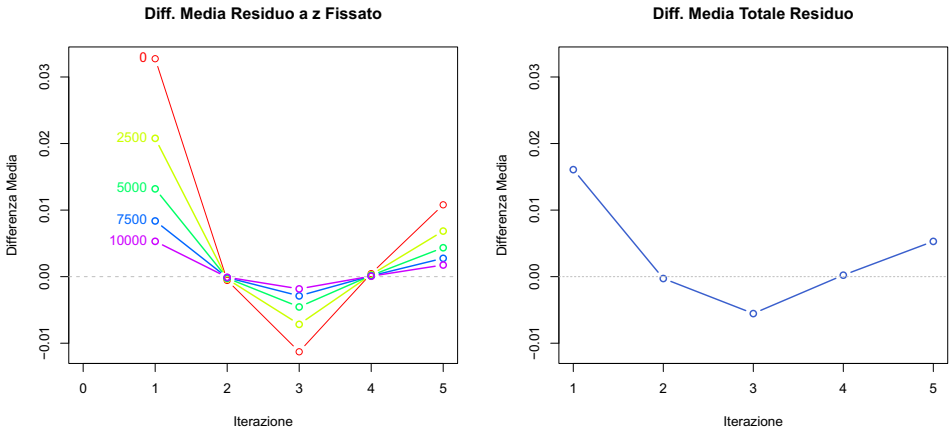

Figura 5.30: A sinistra: Andamento della distanza $d^{GLS}(\widehat{m}(z)^{N+1}, \widehat{m}(z)^{N})$, per $N = 1, ..., 5$, $z = \{0; 2500; 5000; 7500; 10000\}$. A destra: Andamento della distanza media $\overline{d}^{GLS}(\widehat{m}^{N+1}, \widehat{m}^{N})$.

Per questo motivo, è risultato di interesse valutare la capacità previsiva del metodo in due situazioni più critiche, ovvero con dataset di numerosità rispettivamente pari a $n = 30$ e $n = 10$ osservazioni (Figure 5.21 e 5.22).

Coerentemente con le analisi svolte sul dataset completo, ai dataset ridotti è stato applicato il metodo di Universal Kriging associato alla stima del *drift* ottenuta attraverso l'Algoritmo 5.1, selezionando preliminarmente il modello di *drift* tramite l'Algoritmo 5.2.

Sul dataset di numerosità $n = 10$, è stato scelto di non applicare la stima MC-bootstrap del modello di variogramma a causa della scarsa precisione delle stime bootstrap per numerosità del campione molto esigue.

Attraverso l'Algoritmo 5.2, in entrambi i casi è stato individuato come modello di *drift* ottimale:

$$m_s(z) = a_0(z) + a_1(z)x + a_2(z)y, \quad s = (x, y) \in D, z \in \mathcal{T}, \tag{5.5}$$

corrispondente al modello di generazione.

Una volta selezionato il modello (5.5), sono stati stimati i coefficienti a_0, a_1, a_2 del *drift* per mezzo di cinque iterazioni dell'Algoritmo 5.1, ottenendo le stime riportate in Figura 5.31. Dalle immagini si evince che, quantificando la variabilità locale di $\widehat{a}_l^{GLS}$ tramite la stima $\widehat{\varsigma}_l^2$ della media integrale definita dalla (5.1), $l = 0, 1, 2$, le stime ottenute risultano caratterizzate da un'incertezza crescente alla diminuzione della numerosità campionaria.

Le mappe di Ordinary e Universal Kriging, riportate in Figura 5.32 e 5.33 per il campione di numerosità $n = 30$ e in Figura 5.34 e 5.35 per il campione di numerosità $n = 10$, evidenziano una ricostruzione nel complesso accettabile; tuttavia, dalle immagini mostrate,

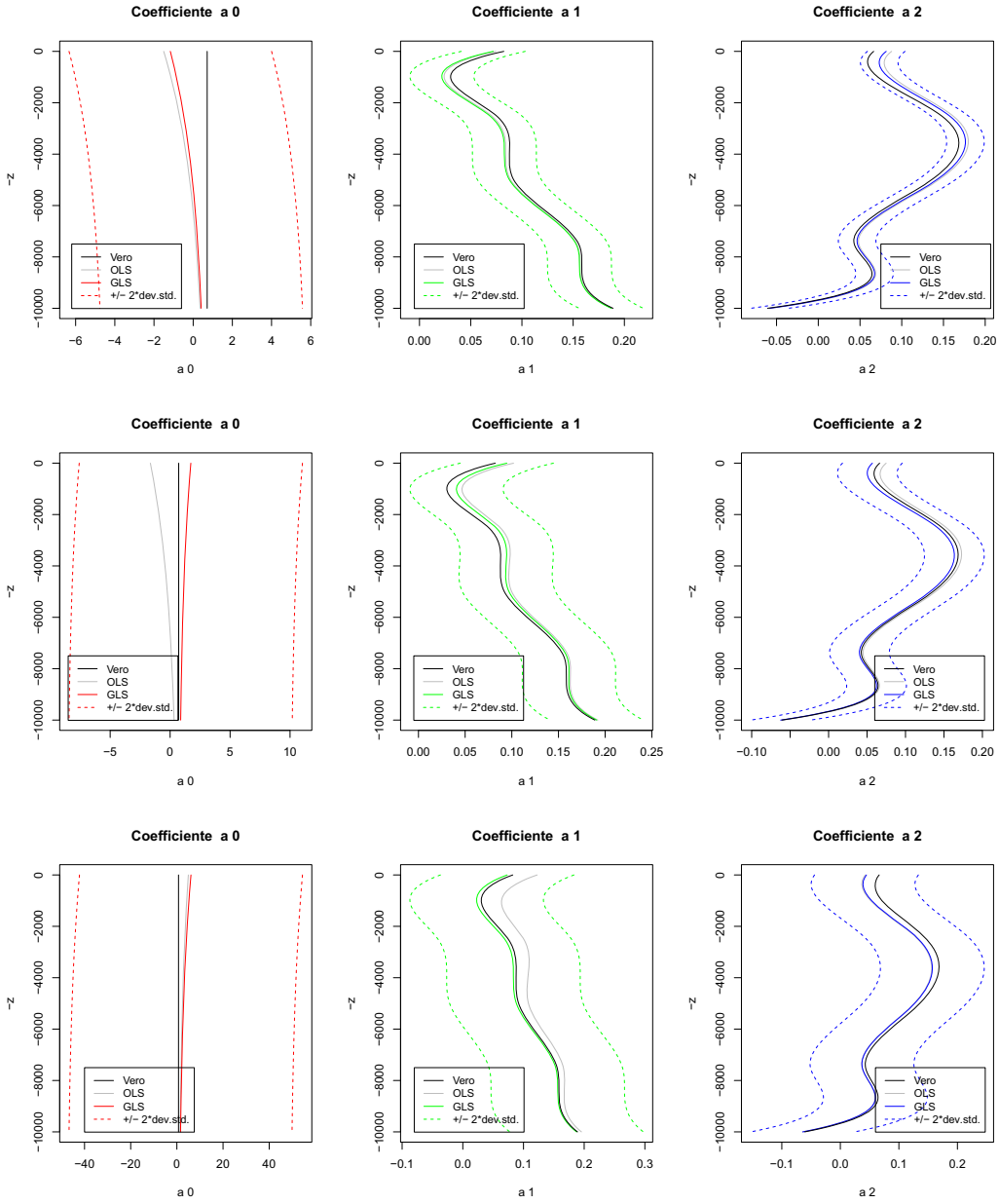

Figura 5.31: Analisi di Robustezza rispetto alla numerosità campionaria, per $n = 100$ (prima riga), $n = 30$ (seconda riga), $n = 10$ (terza riga). Nei grafici sono rappresentati i coefficienti $\widehat{a}_l^{GLS}$, $l = 0, 1, 2$, stimati con il metodo GLS (linea continua colorata) a confronto con la stima OLS (linea continua grigia) e con i valori veri (linea continua nera). Le linee tratteggiate colorate forniscono un'indicazione della variabilità puntuale, attraverso la valutazione di $\widehat{a}_l^{GLS} \pm 2\widehat{\varsigma}_l^2$.

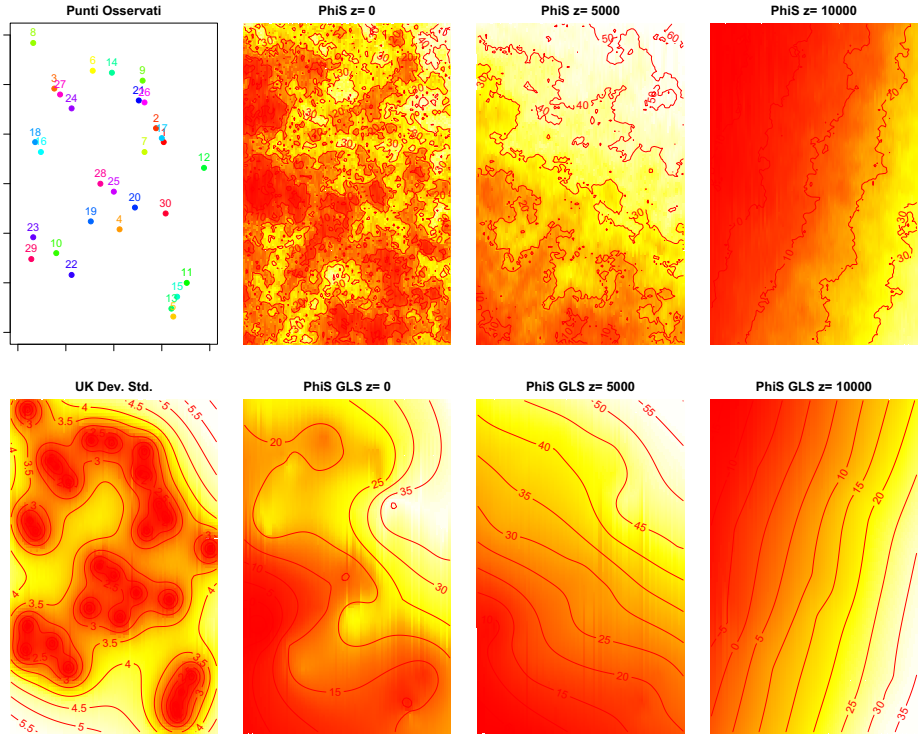

Figura 5.32: Analisi di Robustezza rispetto alla numerosità campionaria, $n = 30$. Confronto tra la realizzazione di riferimento (prima riga) del processo $\boldsymbol{\chi}_s$, l'interpolazione di Universal Kriging dopo 5 iterazioni dell'Algoritmo 5.1 (stima GLS, seconda riga) per il dataset di numerosità $n = 30$. Per la realizzazione di riferimento: punti campionati $s_1, ..., s_{30}$ e contour-plot della realizzazione per $z = \{0; 5000; 10000\}$ m. Per le stime GLS: contour-plot della deviazione standard normalizzata, $\sqrt{\widehat{\sigma}^2_{UK}/10000}$ (a sinistra), e mappe di Universal Kriging per $z = \{0; 5000; 10000\}$ m.

è evidente l'effetto regolarizzante del metodo, più significativo al decrescere della numerosità campionaria.

Dal punto di vista previsivo, il comportamento del metodo è stato valutato attraverso un'analisi di cross-validazione *leave-one-out*. In Tabella 5.8 sono riportate, a confronto, le statistiche relative allo scarto quadratico SSE e allo scarto quadratico relativo $SSE^{(rel.)}$, per $n = 100, 30, 10$.

Dalla Tabella 5.8 si nota che, al decrescere della numerosità campionaria, il potere previsivo del metodo subisce una diminuzione; tuttavia le stime si mantengono ragionevoli e, infatti, i valori dello scarto quadratico relativo sono accettabili.

Analisi di Robustezza rispetto al Modello di Variogramma A partire dal dataset in esame e dai dataset ridotti di numerosità $n = 30$ e $n = 10$, è stata svolta un'analisi di

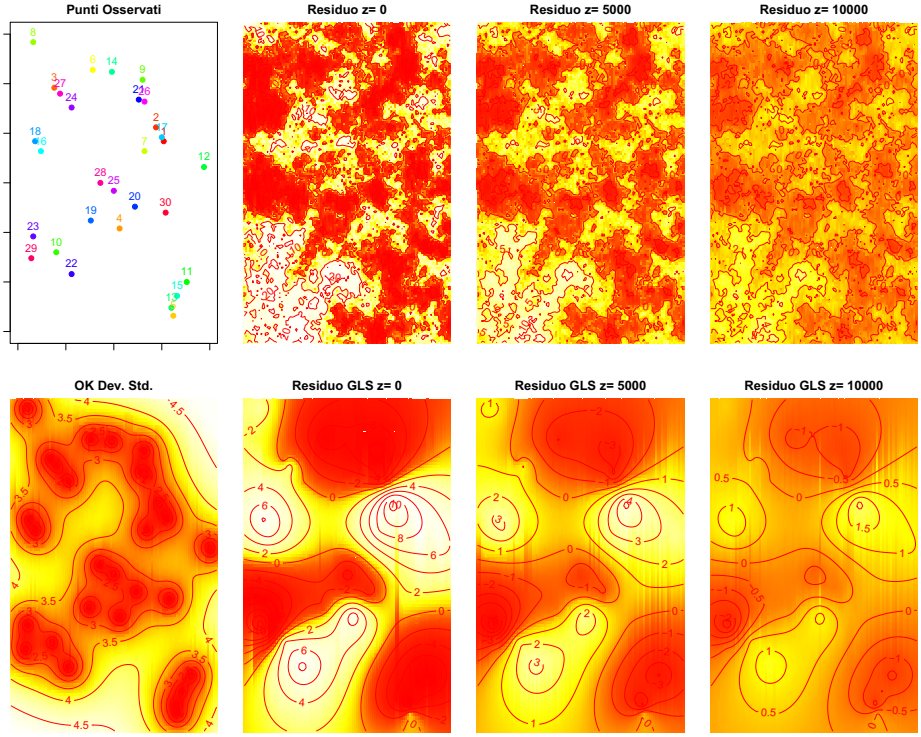

Figura 5.33: Analisi di Robustezza rispetto alla numerosità campionaria, $n = 30$. Confronto tra la realizzazione di riferimento (prima riga) del residuo $\boldsymbol{\delta}_s$ e l'interpolazione di Ordinary Kriging dopo 5 iterazioni dell'Algoritmo 5.1 (stima GLS, seconda riga). Per la realizzazione di riferimento: punti campionati $s_1, ..., s_{30}$ e contour-plot della realizzazione per $z = \{0; 5000; 10000\}$ m. Per le stime GLS: contour-plot della deviazione standard normalizzata, $\sqrt{\widehat{\sigma}^2_{OK}/10000}$ (a sinistra), e mappe di Ordinary Kriging per $z = \{0; 5000; 10000\}$ m.

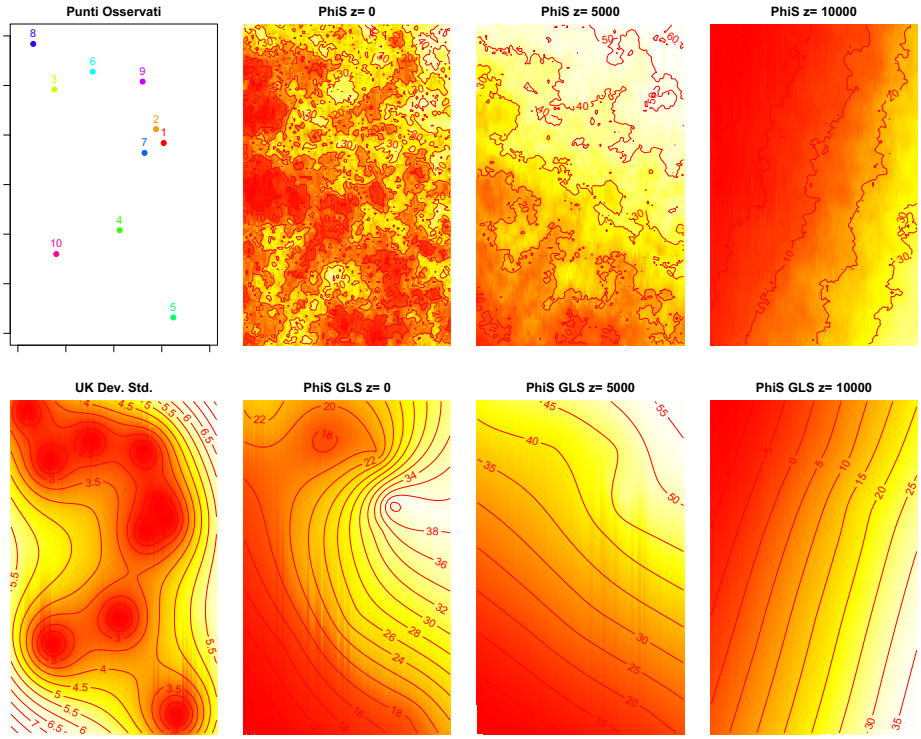

Figura 5.34: Analisi di Robustezza rispetto alla numerosità campionaria, $n = 10$. Confronto tra la realizzazione di riferimento (prima riga) del processo $\boldsymbol{\chi}_s$ e l'interpolazione di Universal Kriging dopo 5 iterazioni dell'Algoritmo 5.1 (stima GLS, seconda riga) per il dataset di numerosità $n = 10$. Per la realizzazione di riferimento: punti campionati $s_1, ..., s_{10}$ e contour-plot della realizzazione per $z = \{0; 5000; 10000\}$. Per le stime GLS: contour-plot della deviazione standard normalizzata, $\sqrt{\widehat{\sigma}^2_{UK}/10000}$ (a sinistra), e mappe di Universal Kriging per $z = \{0; 5000; 10000\}$ m.

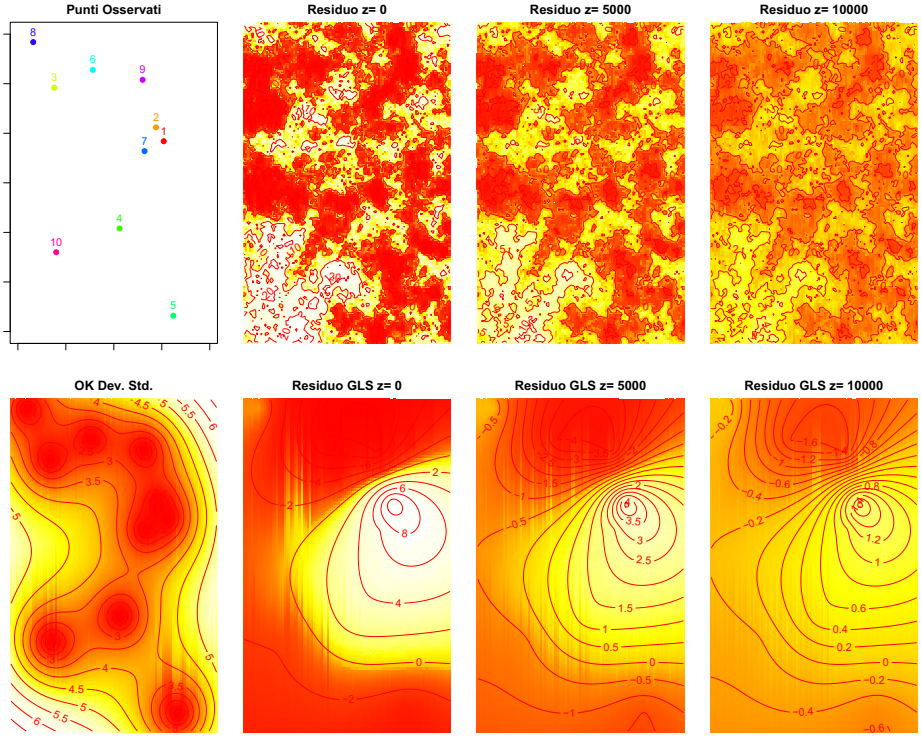

Figura 5.35: Analisi di Robustezza rispetto alla numerosità campionaria, $n = 10$. Confronto tra la realizzazione di riferimento (prima riga) del residuo $\boldsymbol{\delta}_s$ e l'interpolazione di Ordinary Kriging dopo 5 iterazioni dell'Algoritmo 5.1 (stima GLS, seconda riga). Per la realizzazione di riferimento: punti campionati $s_1, ..., s_{10}$ e contour-plot della realizzazione per $z = \{0; 5000; 10000\}$ m. Per le stime GLS: contour-plot della deviazione standard normalizzata, $\sqrt{\widehat{\sigma}^2_{OK}/10000}$ (a sinistra), e mappe di Ordinary Kriging per $z = \{0; 5000; 10000\}$ m.

	Modello Esponenziale					
	100 Dati		30 Dati		10 Dati	
	SSE	$SSE^{(rel.)}$	SSE	$SSE^{(rel.)}$	SSE	$SSE^{(rel.)}$
Minimo	0.67	$8.73{\cdot}10^{-8}$	1408.0	$1.68{\cdot}10^{-4}$	29.99	$2.87{\cdot}10^{-6}$
Mediana	30050.0	$3.93{\cdot}10^{-3}$	38940.0	$4.64{\cdot}10^{-3}$	42500.0	$4.07{\cdot}10^{-3}$
Media	68880.0	$9.00{\cdot}10^{-3}$	92040.0	$1.11{\cdot}10^{-2}$	121100.0	$1.16{\cdot}10^{-2}$
Massimo	680400.0	$8.89{\cdot}10^{-2}$	880000.0	0.1048	432900.0	4.15e-2
Dev. Std.	97556.5	0.0127	165317.6	0.0197	152572.3	0.0146
Somma	6887798.2	0.9000	2761055.0	0.3289	1210679.2	0.1160
$\mathbb{E}_n[\|\chi_s\|^2]$	7652444.0		8393745.0		10439214.7	

Tabella 5.8: Analisi di Robustezza. Statistiche della distribuzione dello scarto quadratico SSE e dello scarto quadratico relativo $SSE^{(rel.)}$ al variare della numerosità campionaria, per $n = 100, 30, 10$ osservazioni.

	Modello Esponenziale		Modello Sferico	
	SSE	$SSE^{(rel.)}$	SSE	$SSE^{(rel.)}$
Minimo	0.67	$8.73{\cdot}10^{-8}$	1.97	$2.57{\cdot}10^{-7}$
Mediana	30050.0	$3.93{\cdot}10^{-3}$	28920.0	$3.77{\cdot}10^{-3}$
Media	68880.0	$9.00{\cdot}10^{-3}$	70150.0	$9.17{\cdot}10^{-3}$
Massimo	680400.0	$8.89{\cdot}10^{-2}$	685200.0	$8.95{\cdot}10^{-2}$
Dev. Std.	97556.5	0.0128	95334.4	0.0125
Somma	6887798.2	0.9000	7015451.7	0.9168
$\mathbb{E}_n[\|\chi_s\|^2]$	7652444.0			

Tabella 5.9: Analisi di Robustezza sul dataset completo. Statistiche della distribuzione dello scarto quadratico SSE e dello scarto quadratico relativo $SSE^{(rel.)}$ al variare del modello parametrico di variogramma, per $n = 100$ osservazioni.

robustezza rispetto alla scelta modello parametrico di variogramma. La valutazione delle proprietà del metodo è stata svolta con criterio previsivo, confrontando in particolare i risultati di cross-validazione nel caso di adozione di un modello di variogramma esponenziale (corrispondente al modello di generazione) o sferico.

Le statistiche caratterizzanti la distribuzione empirica dello scarto quadratico SSE e dello scarto quadratico relativo $SSE_i^{(rel)}$, ottenute dall'analisi del dataset di numerosità $n = 100$, sono riportate in Tabella 5.9.

Dalle statistiche si evince che, dal punto di vista previsivo, il metodo non mostra evidenti variazioni nella distribuzione dello scarto quadratico conseguenti alla differente scelta della struttura di variogramma. Inoltre, in entrambi i casi lo scarto medio è inferiore all'1%, dimostrando un'ottima capacità previsiva.

	Modello Sferico					
	100 Dati		30 Dati		10 Dati	
	SSE	$SSE^{(rel.)}$	SSE	$SSE^{(rel.)}$	SSE	$SSE^{(rel.)}$
Minimo	1.97	$2.57{\cdot}10^{-7}$	441.5	$5.26{\cdot}10^{-5}$	387.9	$3.72{\cdot}10^{-5}$
Mediana	28920.0	$3.77{\cdot}10^{-3}$	33890.0	$4.04{\cdot}10^{-3}$	66850.0	$6.40{\cdot}10^{-3}$
Media	70150.0	$9.17{\cdot}10^{-3}$	87880.0	$1.05{\cdot}10^{-2}$	112600.0	$1.08{\cdot}10^{-2}$
Massimo	685200.0	$8.95{\cdot}10^{-2}$	781300.0	0.0931	446400.0	$4.54{\cdot}10^{-2}$
Dev. Std.	95334.4	0.0125	148574.8	0.0177	156914.7	0.0150
Somma	7015451.7	0.9168	2636397.6	0.3141	1125548.8	0.1078
$\mathbb{E}_n[\|\chi_s\|^2]$	7652444.0		8393745.0		10439214.7	

Tabella 5.10: Analisi di Robustezza. Statistiche della distribuzione dello scarto quadratico SSE e dello scarto quadratico relativo $SSE^{(rel.)}$ al variare della numerosità campionaria, per $n = 100, 30, 10$ osservazioni.

Le osservazioni precedenti sono in accordo con quanto mostrato dalla Figura 5.36, dove i dati predetti per cross-validazione sono confrontati con i dati osservati.

La verifica del comportamento previsivo del metodo al variare del modello di variogramma è ancora più importante qualora la numerosità dei dati decresca. Infatti, al decrescere del numero di osservazioni, l'influenza del modello di variogramma sul metodo di Kriging (Ordinary o Universal) diventa più significativa, passando da un'interpolazione di tipo *data driven* a una di tipo *model driven*.

Nella Tabella 5.10 sono dunque riportate le statistiche degli scarti SSE e $SSE^{(rel)}$, ottenuti con la scelta di un modello sferico a fronte di un modello di generazione esponenziale, al variare di $n = 100, 30, 10$ osservazioni.

Dal confronto delle Tabella 5.8 e 5.10, relative rispettivamente al modello esponenziale e al modello sferico, si nota che non esistono significative differenze, dal punto di vista previsivo, al variare del modello parametrico di variogramma scelto.

Inoltre, dal confronto dei risultati di cross-validazione a modello sferico fissato si può dedurre che anche in questo caso la diminuzione del potere previsivo al decrescere della numerosità campionaria si mantiene limitata.

5.5 Conclusioni e Sviluppi Futuri

A partire dal contesto teorico sviluppato nel Capitolo 3, nel presente capitolo è stata proposta una metodologia per la selezione e la stima della variabilità spaziale deterministica del campo, espressa attraverso il *drift* (Algoritmi 5.1 e 5.2). Quest'ultima consente di determinare una stima della struttura di covarianza spaziale dal residuo stazionario del processo e di fornire una stima interpolatoria di Universal Kriging del campo funzionale di generazione sulla base del dataset funzionale osservato.

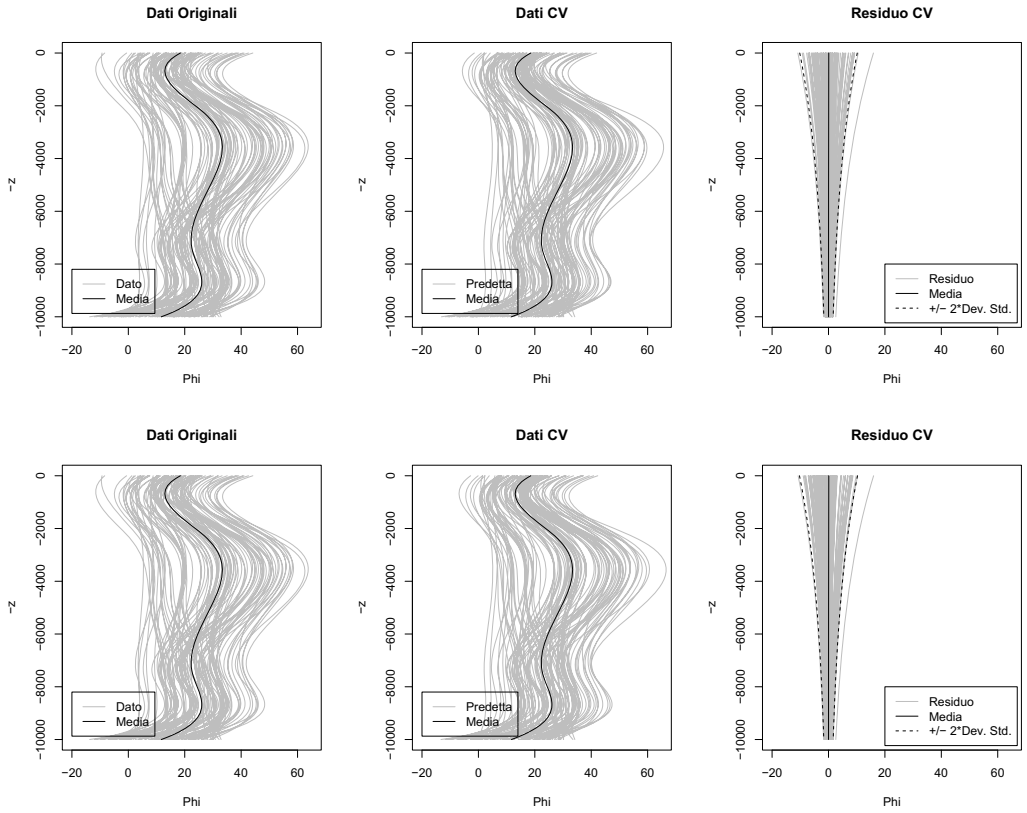

Figura 5.36: Analisi di robustezza sul dataset completo. A confronto dati predetti con cross-validazione per Universal Kriging con modello di variogramma esponenziale (prima riga) e modello sferico (seconda riga). A sinistra: dati originali (in grigio) e relativa media campionaria (in nero). Al centro: dati predetti con Universal Kriging (in grigio) e loro media (in nero). A destra: differenza r_i^{CV}, $i = 1, ..., n$, tra i dati originali e i dati predetti (in grigio), loro media $\widehat{m}_r$ (in nero) e banda di confidenza puntuale $\widehat{m}_r(z) \pm 2\widehat{\sigma}_r(z)$ (in nero tratteggiato), dove $\widehat{\sigma}_r(z)$ è la deviazione standard stimata puntualmente dal residuo di cross-validazione.

La metodologia proposta è stata quindi sottoposta ad uno studio di simulazione, durante il quale è stato valutato il comportamento degli algoritmi sviluppati in termini di convergenza, potere previsivo e robustezza.

Dal punto di vista algoritmico, la convergenza del metodo è stata verificata per simulazione, adottando come criterio generale di arresto il numero massimo di iterazioni, fissato a $N_{max}^{(GLS)} = 5$ in quanto risultato sufficiente a stabilizzare la stima iterativa del *drift* (Sottosottosezioni 5.3.2.4 e 5.4.2.4): maggiore attenzione potrebbe essere posta sull'individuazione di ulteriori criteri d'arresto e, soprattutto, sulla formulazione di risultati teorici riguardo la velocità di convergenza dell'Algoritmo 5.1.

In tutti i casi sottoposti a verifica, l'Algoritmo 5.2 ha selezionato la forma di *drift* corrispondente alla struttura di generazione, dimostrando robustezza sia rispetto al modello parametrico di variogramma, sia rispetto alla risoluzione spaziale del campionamento (Sottosottosezioni 5.3.2.5 e 5.4.2.5).

Inoltre, dalle analisi svolte è emerso che anche l'Algoritmo 5.1 si dimostra robusto rispetto al modello parametrico di variogramma scelto per l'analisi, fornendo in tutti i casi stime finali delle caratteristiche salienti del modello di variabilità spaziale quali *sill*, *range* e *nugget* (Sottosottosezione 5.4.2.5) molto simili tra loro.

Grazie alle conclusioni ottenute nella Sottosottosezione 5.3.2.5, si può ritenere che l'applicazione dell'Algoritmo 5.1 possa essere condotta adottando un modello di *drift* generale di forma (3.33), ovvero:

$$m_s(t) = \sum_{l=0}^{L} a_l(t) f_l(s), \quad s \in D, t \in \mathcal{T},$$

anche in presenza di un *drift* di forma (3.32), ossia costante lungo l'ascissa del dato funzionale. L'adozione di una struttura di *drift* più generale della struttura di generazione consente infatti di svolgere le analisi senza variazioni significative sul potere previsivo del metodo.

La diminuzione della numerosità campionaria, intesa nel senso della risoluzione spaziale (Sottosottosezione 5.4.2.5), non altera l'affidabilità del metodo in termini previsivi, conducendo a stime di precisione inferiore rispetto al caso di dataset di alta numerosità, ma comunque nel complesso accettabili.

Non è stata tuttavia considerata l'influenza della rappresentazione del dato funzionale, né della qualità dell'acquisizione dei dati, sui risultati finali forniti dal metodo: tale robustezza, tralasciata in questa sede poiché al di là delle finalità dello studio, dovrebbe invece essere verificata in caso di applicazione del metodo a dati caratterizzati da un cattivo campionamento.

La precisione della stima è risultata essere influenzata sia dalla numerosità campionaria, sia dalla variabilità locale del fenomeno: infatti, a causa dell'effetto regolarizzante del metodo di Universal Kriging, dalle analisi svolte è emerso che la previsione si dimostra più accurata dove il fenomeno è meno variabile, mentre può risultare imprecisa in presenza di un'alta

variabilità spaziale del fenomeno stesso, localizzata in corrispondenza di alcuni intervalli di valori dell'ascissa del dato funzionale.

Una possibile soluzione al problema registrato potrebbe essere l'introduzione di coefficienti funzionali $\lambda(t)$ all'interno della formulazione dell'Universal Kriging, in modo simile a quanto proposto, limitatamente all'ambito stazionario, nei lavori di Giraldo *e altri* (2008b), Delicado *e altri* (2010) e Giraldo *e altri* (2010a).

Dal punto di vista teorico, un ulteriore sviluppo della procedura di stima potrebbe consistere nell'inserimento di una correzione alla distorsione dello stimatore $\widehat{\Sigma}$, quantificata nel Capitolo 3 attraverso la matrice $\mathbb{B}$, definita dalla (3.58): una metodologia bootstrap semiparametrica analoga a quella introdotta nel Capitolo 4, potrebbe essere utile a tal scopo, fornendo uno strumento per l'approssimazione della distribuzione dello stimatore $\widehat{\Sigma}$. Inoltre, la possibilità di determinare degli intervalli di confidenza bootstrap sul *trace-variogram* sperimentale (Capitolo 4) potrebbe essere sfruttata per la costruzione di bande di confidenza per le previsioni di kriging, il cui calcolo incorpori anche l'incertezza e la distorsione dello stimatore del variogramma.

Maggiore attenzione potrebbe inoltre essere rivolta alla definizione delle forme funzionali scelte per il *drift*, con una possibile estensione delle formulazioni fornite nel Capitolo 3 al caso di modello lineare di variabilità deterministica a regressori funzionali, consentendo in questo modo di modellare in modo più accurato la media spaziale del processo, eventualmente con l'introduzione di un *external drift*.

L'importanza applicativa di questa evoluzione potrebbe essere molto rilevante, fornendo la possibilità di integrare la tecnica di Universal Kriging con il processo modellistico del fenomeno in oggetto. In tal caso, potrebbero tuttavia presentarsi dei vincoli sulla forma funzionale attesa per il dato ricostruito, che dovrebbero essere incorporati nella procedura di minimizzazione descritta nella Sezione 3.4 del Capitolo 3: quest'ambito di ricerca è tuttora molto aperto e meriterebbe sicuramente ulteriore attenzione teorica.

Capitolo 6

Un Caso Studio: Analisi delle Temperature delle Province Marittime del Canada

6.1 I Dati

I dati considerati per lo studio di questo capitolo consistono nelle temperature giornaliere registrate in 35 stazioni meteorologiche dislocate nella regione marittima del Canada. Tale regione è costituita dalle 'Maritimes provinces' (Figura 6.1), ossia da tre delle quattro province della costa orientale del Canada: Nova Scotia, New Brunswich e Prince Edward Island.

Le province Marittime, che coprono circa l'1% della superficie del Canada, insieme alle province Newfoundland e Labrador, costituiscono la regione atlantica del Paese.

Dal punto di vista geografico (Figura 6.1), la regione considerata è per lo più collinare e ricoperta da foreste, mentre i picchi più alti sono situati nelle zone montuose attorno al Mount Carleton (820 m), nella parte nord-centrale della provincia di New Brunswich, dove si registrano le temperature invernali più rigide.

Tuttavia, la caratteristica distintiva della regione è la lunga zona costiera, che si sviluppa verso l'Atlantico nella parte orientale delle provincie di Prince Edward Island e Nova Scotia, mentre si affaccia sulla Deep Bay nella parte occidentale di quest'ultima e nella provincia di New Brunswich.

L'esposizione della regione verso il mare ne influenza notevolmente il clima, specialmente per la Corrente del Golfo proveniente da est. Per questo motivo, il clima di tali province è temperato, contraddistinto da inverni miti ed estati fresche: masse di aria fredda provenienti dalle regioni nord-occidentali si alternano infatti con arie calde-umide marittime sud-occidentali (Stanley, 2002).

In questo contesto geografico e climatico, il dataset considerato contiene le informazioni relative alle temperature medie giornaliere registrate tra gli anni 1960 e 1994 (combinando

Figura 6.1: Immagine da satellite del Canada (a sinistra) e delle province Marittime del Canada (a destra). Fonte: Google map.

i dati raccolti il 29 febbraio con quelli relativi al 28 febbraio), analizzati in (Giraldo, 2009), (Delicado *e altri*, 2010), (Giraldo *e altri*, 2010a). I dati relativi ad ogni stazione meteorologica sono raccolti nel database del Servizio Meteorologico del Canada (Giraldo, 2009); il dataset completo usato per l'analisi è disponibile in rete all'indirizzo `http://www.docentes.unal.edu.co/rgiraldoh/docs/` ed è mostrato nel pannello di destra della Figura 6.2.

6.2 Scelta dello Spazio Funzionale e dello Spazio Metrico

A partire dal dataset funzionale di temperature introdotto nella Sezione 6.1, il primo momento dell'analisi è consistito nella scelta dello spazio funzionale per i dati e della metrica più adatta per il calcolo delle distanze tra i siti osservati.

In continuità con le analisi svolte nel lavoro di Giraldo (2009), è stato fissato come spazio di Hilbert H lo spazio L^2 delle funzioni quadrato integrabili, dotato del prodotto interno e della norma da esso indotta:

$$< f, g > = \int_{\mathcal{T}} f(t) \cdot g(t) dt$$
$$\|f\| = \sqrt{\int_{\mathcal{T}} |f(t)|^2 dt},$$

per $f, g \in L^2$.

I dati puntuali sono quindi stati proiettati su una base di 65 funzioni di Fourier, scelta in (Giraldo, 2009) con un approccio di cross-validazione non parametrica funzionale. Il dataset funzionale così ottenuto è mostrato nel pannello di destra della Figura 6.2.

Nel seguito sarà dunque denotato con:

$$\{\boldsymbol{\chi}_s, s \in D\}$$

il processo stocastico funzionale a valori in L^2 che descrive la temperatura media giornaliera nelle province Marittime del Canada, mentre con D sarà indicato il dominio spaziale di

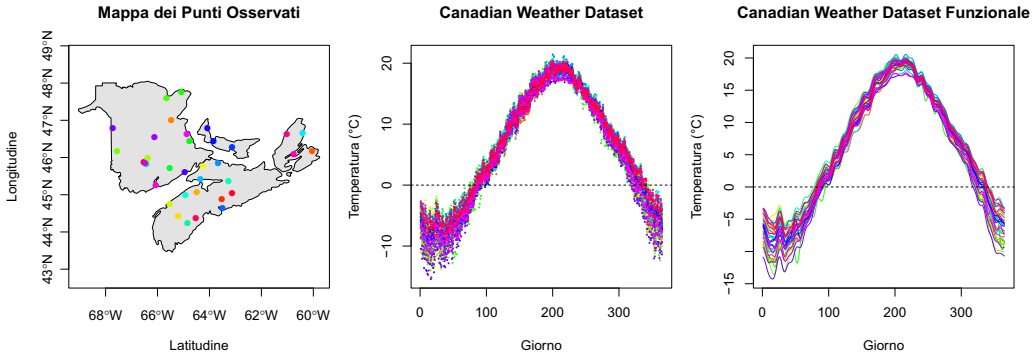

Figura 6.2: Dataset *Canada's Maritime Provinces Temperatures*. A sinistra: mappa delle 35 stazioni meteorologiche presso le quali sono stati registrati i dati. Al centro: temperature medie giornaliere osservate in un anno tra il 1960 e 1994 (valori puntuali). A destra: dataset funzionale corrispondente alle temperature medie giornaliere registrate nelle stazioni meteorologiche di osservazione.

osservazione, pari a $D = [43°N, 48°N] \times [-69°E, -60°E]$, individuando le coordinate di ciascun punto osservato attraverso la coppia latitudine[Nord]-longitudine[Est], $s = (\zeta, \varphi)$, espressi in gradi (Figura 6.2).

Per quanto riguarda la metrica per il dominio D, si sono individuate tre possibili classi di distanze:

i. Distanza euclidea calcolata sulla superficie sferica a partire da opportune trasformazioni delle coordinate geografiche:

$$d_e(\widetilde{s}_1, \widetilde{s}_2) = \sqrt{(\widetilde{\zeta}_1 - \widetilde{\zeta}_2)^2 + (\widetilde{\varphi}_1 - \widetilde{\varphi}_2)^2}, \quad \widetilde{s}_i = (\widetilde{\zeta}_i, \widetilde{\varphi}_i) = (\widetilde{f}(\zeta_i), \widetilde{f}(\varphi_i)) \in D, i = 1, 2. \tag{6.1}$$

ii. Distanza euclidea calcolata direttamente delle coordinate geografiche:

$$d_e(s_1, s_2) = \sqrt{(\zeta_1 - \zeta_2)^2 + (\varphi_1 - \varphi_2)^2}, \quad s_i = (\zeta_i, \varphi_i) \in D, i = 1, 2. \tag{6.2}$$

iii. Distanza geodetica, calcolata, grazie all'approssimazione sferica della Terra, attraverso la formula esplicita seguente:

$$d_g(s_1, s_2) = 2R_m \arcsin\left(\sqrt{\sin^2\left(\frac{\zeta_1 - \zeta_2}{2}\right) + \cos(\zeta_1)\cos(\zeta_2)\sin^2\left(\frac{\varphi_1 - \varphi_2}{2}\right)}\right), \tag{6.3}$$

dove $s_i = (\zeta_i, \varphi_i), i = 1, 2$ e $R_m \simeq 6371$ km indica il raggio medio terrestre. Fissando in tal modo il parametro R_m, l'unità di misura della distanza fornita dalla (6.3) è chilometri.

Esistono diversi metodi per il calcolo della distanza euclidea i., che si basano su opportune proiezioni delle coordinate di partenza (Banjerjee, 2005). La distanza geodetica è però in

generale preferibile alla distanza euclidea i. per superfici sufficientemente estese, in quanto la seconda può portare in questi casi a deformazioni importanti nella metrica stessa.

Inoltre, come sottolineato da Banjerjee (2005), l'esistenza di una proiezione su una mappa piana che preservi le distanze tra i punti è preclusa dal Teorema Egregio di Gauss in geometria differenziale (Guggenheimer, 1977): infatti, mentre esistono proiezioni che preservano le aree e le forme, le distanze risultano sempre distorte.

Sebbene nel caso in esame l'area in considerazione non sia molto estesa, coprendo distanze inferiori a 445 km in direzione N-S e 640 km in direzione W-E, e dunque tali approssimazioni siano accettabili, è stato deciso di adottare la distanza geodetica in quanto si è ritenuto preferibile lavorare con le coordinate di partenza calcolando le distanze attraverso l'espressione (6.3).

Inoltre, si è ritenuto formalmente importante adottare una distanza che tenesse in considerazione la geometria del dominio spaziale, consentendo al contempo di ottenere risultati nel sistema metrico internazionale, cosa che la distanza euclidea ii. non consente di fare, generando, al contrario, ambiguità sull'unità di misura della distanza calcolata.

D'altra parte, la distanza definita dalla (6.2), ottenuta in unità di grado, può essere convertita in chilometri convertendo dapprima in radianti, quindi moltiplicando per R_m: in questo modo tuttavia, le distorsioni della metrica sono significative già per distanze inferiori a 300 km (Banjerjee, 2005). Infatti, come si può osservare dalla Figura 6.3, la differenza tra l'adozione delle metriche citate è sensibile anche nella considerazione di uno dominio spaziale relativamente limitato come quello delle province Marittime del Canada.

Per questi motivi, la distanza geodetica è stata preferita alle distanze euclidee i. e ii. ed è stata adottata per le analisi geostatistiche che saranno illustrate nel seguito del capitolo.

L'introduzione della metrica indotta dalle geodetiche è dunque un punto di discontinuità rispetto alle analisi precedenti presenti in letteratura, che ha implicazioni importanti in termini di valutazione della stazionarietà e isotropia del campo, come sarà mostrato nella Sezione 6.4.

6.3 Analisi Geostatistica

6.3.1 Analisi di Stazionarietà

Una volta ottenuto il dataset funzionale rappresentato in Figura 6.2 e fissata la metrica come illustrato nella Sezione 6.2, si è proceduto con la valutazione della stazionarietà del campo aleatorio $\boldsymbol{\chi}_s$.

A partire dai dati, è stato stimato il variogramma omnidirezionale tramite gli stimatori empirici *variogram cloud* e *binned-variogram* (Definizione 3.14), ottenendo le stime riportate in Figura 6.4.

Dalle stime sperimentali ottenute vi è evidenza di non stazionarietà del processo. Infatti, la *variogram cloud* presenta una forma parabolica incompatibile con la forma triangolare attesa da un variogramma di generazione stazionario: dalle differenze, in norma, tra le osser-

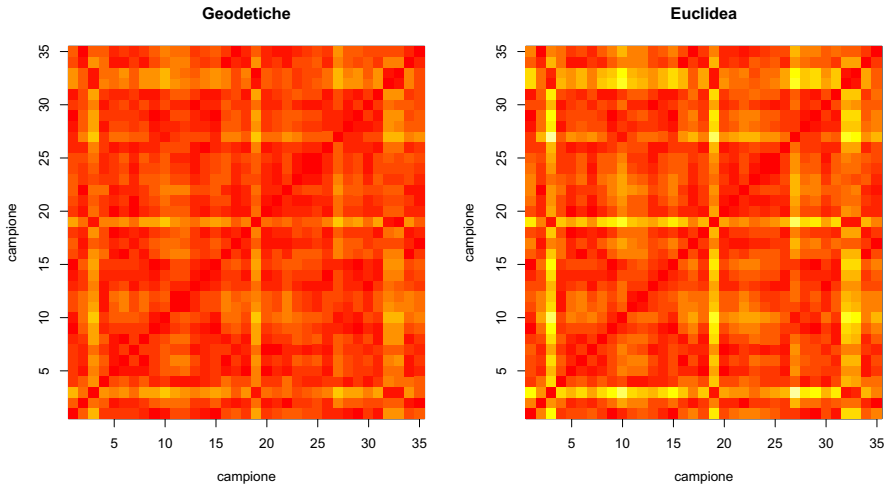

Figura 6.3: Confronto tra distanze geodetiche ed euclidee. Matrici di distanze reciproche tra i 35 siti del dataset calcolate tramite la distanza geodetica d_g (a sinistra) e la distanza euclidea d_e convertita in km. I valori delle distanze sono rappresentate attraverso colori dal rosso (valori bassi) al giallo (valori alti); in posizione (i, j) del grafico è rappresentata la distanza $d_g(s_i, s_j)$ o $d_e(s_i, s_j)$. Il *range* di distanze della prima matrice è [0, 591.16] km, della seconda [0, 855.74] km.

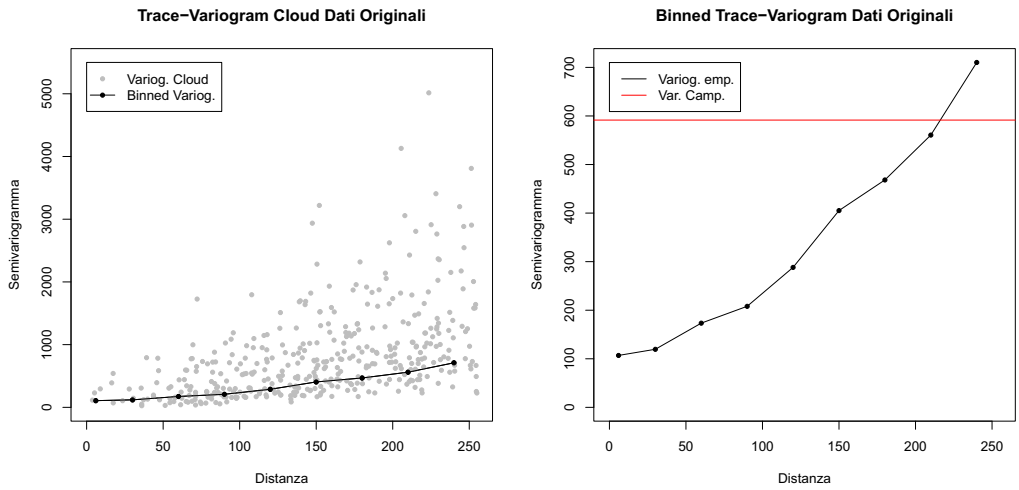

Figura 6.4: Variogramma dai Dati. A sinistra: *Trace-Variogram Cloud*. A destra: *Binned Trace-Variogram*.

(k)	k	f_l^k	SSE_k
(1)	8	$\{1, x^2, y^2, xy\}$	157.13
(2)	3	$\{1, x, y\}$	160.91
(3)	7	$\{1, x^2, y^2\}$	161.11
(4)	6	$\{1, xy\}$	229.17
(5)	5	$\{1, y^2\}$	381.75
(6)	2	$\{1, y\}$	384.85
(7)	1	$\{1, x\}$	410.64
(8)	4	$\{1, x^2\}$	412.51

Tabella 6.1: Ordinamento delle forme funzionali di *drift* per il dataset *Canada's Maritime Provinces Temperatures*.

vazioni si deduce quindi che la variabilità del fenomeno aumenta in modo superquadratico al crescere della distanza e non si riconosce un *range* di distanze oltre il quale la discrepanza media registrata tra le osservazioni si accosta alla varianza campionaria stimata dai dati.

Conclusioni analoghe si traggono dall'osservazione della stima *binned-variogram*, che nell'origine presenta un'evidente concavità verso l'alto, mentre asintoticamente non sembra stabilizzarsi in corrispondenza di un asintoto orizzontale.

Pertanto, le stime sperimentali indicano che la dissimilarità tra le osservazioni aumenta, al crescere della distanza, in modo non stazionario.

6.3.2 Ordinamento dei *Drift* con Criterio Previsivo

Dal momento che dall'analisi variografica il dataset funzionale è risultato spazialmente non stazionario, è stato applicato l'Algoritmo 5.2 a partire dalle otto collezioni di funzioni f_l^k riportate nella Tabella 5.1 di Sezione 5.2.

L'ordinamento dei *drift* ottenuto con criterio previsivo attraverso l'Algoritmo 5.2 è riportato nella Tabella 6.1.

Il modello di *drift* selezionato è dunque il modello quadratico:

$$m(s,t) = a_0(t) + a_1(t)x^2 + a_2(t)y^2 + a_3(t)xy, \quad s = (x,y),\ t \in \mathcal{T} = [0, 365], \tag{6.4}$$

identificando le coordinate (x, y) con le coordinate geografiche latitudine e longitudine, (ζ, φ).

Sebbene anche i modelli più parsimoniosi 3 e 7 diano dei buoni risultati in termini previsivi, è stato deciso di proseguire l'analisi adottando il modello 8, in quanto il rischio di *overfitting* dei dati è ridotto dall'uso di un criterio previsivo derivante da una stima di cross-validazione.

Infatti, l'adozione di una struttura di *drift* eccessivamente complessa rispetto alla complessità della variabilità deterministica del fenomeno non consentirebbe di filtrare il rumore presente nei dati, cogliendo anche parte della variabilità stocastica dei dati stessi. D'altra parte, l'eccessivo adattamento del modello alle osservazioni a disposizione (i.e. *overfit-*

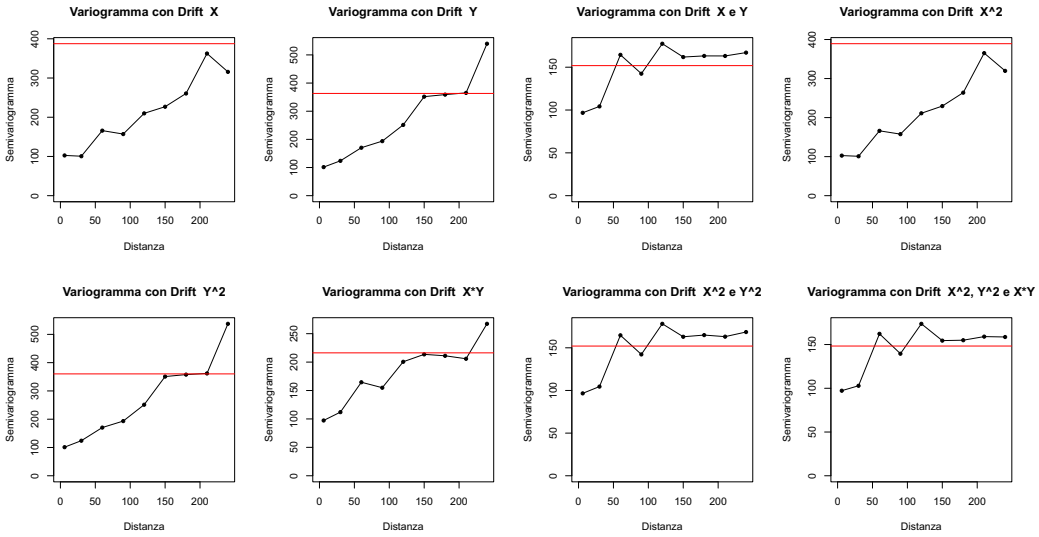

Figura 6.5: Variogrammi dei residui ottenuti dopo un'iterazione dell'Algoritmo 5.1, adottando ciascuno degli 8 modelli di *drift* della Tabella 5.1. Si presti attenzione alla scala sull'asse delle ordinate, in quanto, per consentire l'analisi della forma delle stime variografiche, è stato necessario usare scale diverse per ciascun pannello.

ting), riduce in modo consistente le prestazioni previsive, secondo le quali è determinato l'ordinamento individuato dall'Algoritmo 5.2.

In Figura 6.5 sono riportati i variogrammi sperimentali calcolati dai residui dopo un'iterazione dell'Algoritmo 5.1, adottando ciascuno degli 8 modelli di *drift* della Tabella 5.1. Come si può osservare, il residuo relativo alla struttura di *drift* selezionata dal metodo è caratterizzato da una varianza campionaria inferiore rispetto alle rimanenti strutture, indicando che la forma funzionale selezionata modella la variabilità deterministica del fenomeno nel modo più esaustivo.

Dai grafici si può notare che la stazionarietà del residuo è subordinata all'introduzione di entrambe le coordinate geografiche, latitudine e longitudine: infatti, i variogrammi derivanti dall'adozione delle forme di *drift* 1, 2, 4 e 5 non rispettano ancora i requisiti si stazionarietà richiesti per il residuo. Dal punto di vista climatico, questo potrebbe indicare che, sebbene la regione in oggetto sia limitata, la media spaziale della temperatura giornaliera non possa essere considerata costante né in direzione N-S, né in direzione W-E.

Nella Sottosezione 6.3.3 sarà quindi applicato l'Algoritmo 5.1 a partire dal modello di *drift* (6.4), al fine di disaccoppiare la variabilità deterministica del fenomeno dalla componente aleatoria; successivamente, sarà stimato l'intero campo di temperatura con il metodo di Universal Kriging, sulla base del modello di variogramma stimato (Sottosezione 6.3.4).

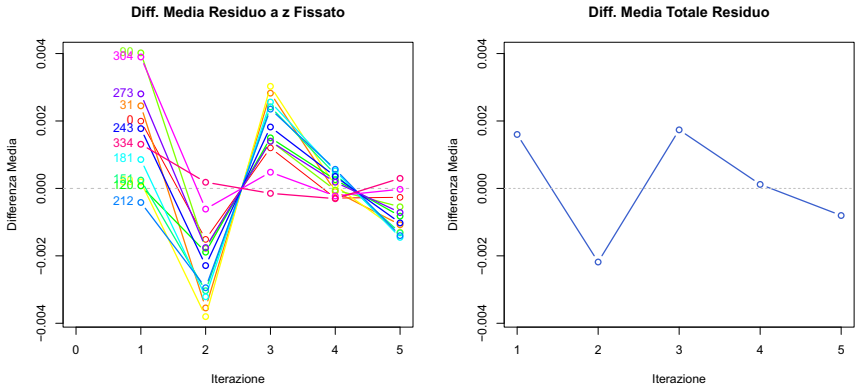

Figura 6.6: Analisi di Convergenza. A sinistra: Andamento della distanza $d^{GLS}(\widehat{m}(z)^{N+1}, \widehat{m}(z)^{N})$, per $N = 1, ..., 5$, il primo giorno di ogni mese. A destra: Andamento della distanza media $\overline{d}^{GLS}(\widehat{m}^{N+1}, \widehat{m}^{N})$.

6.3.3 Stima del *Drift*: Disaccoppiamento della Variabilità Deterministica dal Residuo

La stima empirica del variogramma calcolata dal residuo corrispondente al *drift* ottimale individuato nella Sottosezione 6.3.2, mostra l'esistenza di una struttura di covarianza spaziale più complessa della struttura di puro *nugget* (pannello in basso a destra di Figura 6.5). I residui del modello risultano quindi correlati ed è ragionevole procedere nell'applicazione dell'Algoritmo 5.1 per la determinazione di una stima GLS dei coefficienti a_l, $l = 0, 1, 2, 3$.

Il modello parametrico di variogramma adottato per le analisi illustrate nel seguito è il modello esponenziale, che ben si adatta alle peculiarità del variogramma empirico stimato. Si noti che, adottando metriche non euclidee, non vi è in generale garanzia che i modelli di variogramma comunemente usati rimangano validi (Curriero, 2006). Tuttavia, si dimostra che, nel caso della geometria sferica, i modelli di variogramma esponenziale e sferico risultano modelli validi (Huang *e altri*, 2011) e ne è quindi giustificato l'utilizzo.

Una volta fissato il modello di variogramma, è stato applicato l'Algoritmo 5.1, verificandone in primo luogo la convergenza. L'Algoritmo 5.1 si è dimostrato convergente entro le prime cinque iterazioni, dopo le quali lo scostamento medio tra iterazioni consecutive, $\overline{d}^{GLS}(\widehat{m}^{N+1}, \widehat{m}^{N})$, è risultato inferiore allo 0.1% (Figura 6.6); per questo motivo, i risultati mostrati nel seguito si riferiscono alle stime ottenute dopo cinque iterazioni dell'Algoritmo stesso.

In Figura 6.7 sono quindi riportate le stime dei coefficienti a_l, $l = 0, 1, 2, 3$, ottenute al termine della procedura di stima. Dal confronto dei quattro pannelli si osserva che la stima del primo coefficiente, associata a una variabilità più bassa rispetto alle altre, è quasi indistinguibile rispetto alla stima OLS.

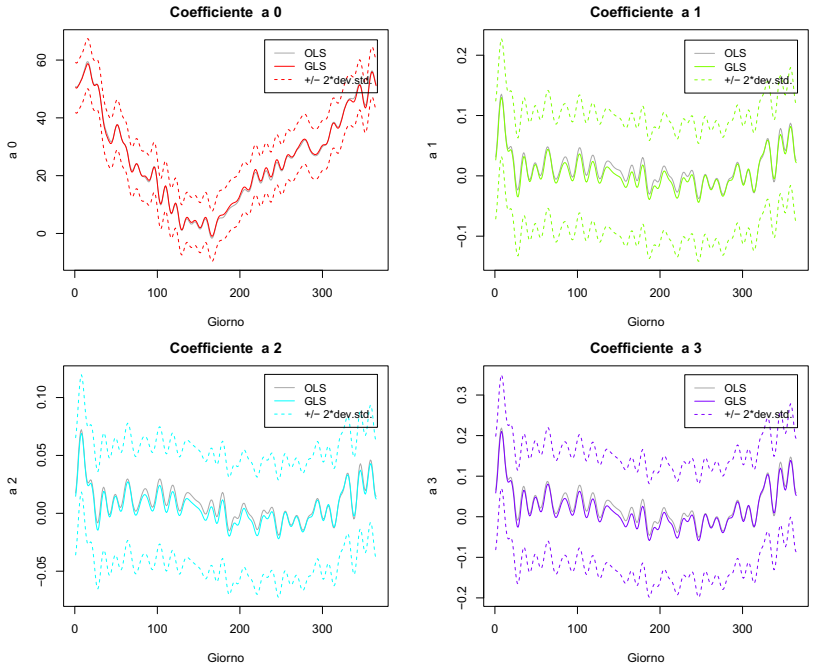

Figura 6.7: Stima dei coefficienti del modello lineare di drift, $\widehat{a}_l^{GLS}$, $l = 0, 1, 2, 3$. La linea grigia indica la stima OLS, la linea colorata la stima GLS; le linee colorate tratteggiate forniscono un'indicazione della variabilità puntuale, attraverso la valutazione di $\widehat{a}_l^{GLS} \pm 2\widehat{\varsigma}_l$.

Si registrano invece differenze più significative tra le stime OLS e GLS per i coefficienti a_1, a_2, a_3, associati a una variabilità più alta, con scostamenti più evidenti nella parte centrale dell'anno.

Le stime sperimentali del variogramma, calcolate dal residuo del modello di *drift* (6.4) stimato attraverso i coefficienti $\widehat{a}_l^{GLS}$, $l = 0, 1, 2, 3$, sono riportate in Figura 6.8. L'incertezza legata allo stimatore empirico, calcolata con il metodo MC-bootstrap fissando il numero di iterazioni esterne a $N_{max} = 2$ e il numero di iterazioni bootstrap a $B = 5000$, è piuttosto significativa, in particolar modo in prossimità dell'origine, come indicato dall'ampiezza dell'intervallo di confidenza bootstrap. Malgrado questo, sia lo stimatore *binned variogram*, sia lo stimatore *variogram cloud* mostrano un comportamento chiaramente stazionario. In particolare, nella *variogram cloud* si riconosce una forma triangolare, evidenziando, la presenza di un effetto *nugget* piuttosto accentuato (si noti la forma nell'origine).

Anche nel modello parametrico adattato con i minimi quadrati generalizzati (linea rossa) è presente un forte effetto *nugget*, che spiega una parte consistente della variabilità spaziale. Inoltre il *sill* è raggiunto molto velocemente rispetto alle dimensioni dell'area di interesse, con un *practical range* di circa 122 km.

Come si nota dal grafico, non vi è una forte differenza tra la stima OLS iniziale del

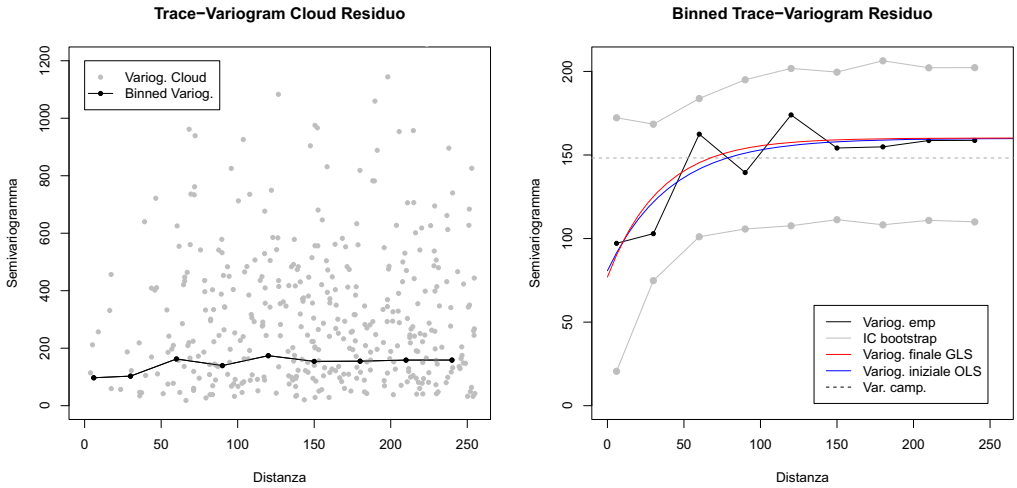

Figura 6.8: Variogramma dal residuo finale. A sinistra: *Trace-Variogram Cloud*. A destra: *Binned Trace-Variogram*, sovrapposto alla stima del modello parametrico MC-bootstrap finale, agli intervalli di confidenza puntuali MC-bootstrap di livello 0.90% e alla stima OLS iniziale.

variogramma e la stima GLS finale, indicando che la struttura di covarianza del residuo alla prima iterazione non è sostanzialmente differente dalla struttura di dipendenza spaziale del residuo finale.

Questo è dovuto probabilmente alla presenza di un effetto *nugget* molto accentuato rispetto al *partial sill* relativo alla struttura variografica esponenziale. Infatti, grazie alle decomposizioni della varianza (3.60) e (3.57), l'aumento (o la diminuzione) della variabilità del residuo è controbilanciato dal cambiamento di segno opposto della variabilità dello stimatore $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}^{GLS}$.

Più precisamente, dal momento che un processo caratterizzato da un effetto *nugget* può essere pensato come somma di un processo continuo nell'origine e di un contributo di rumore bianco (Cressie, 1993), qualora l'effetto *nugget* sia molto accentuato, la componente di rumore bianco diventa molto influente sul processo di partenza. Il rumore bianco è un processo scorrelato e, in presenza di tale proprietà, la stima OLS e la stima GLS coincidono.

In altri termini, qualora il campo aleatorio sia un rumore bianco e dunque la struttura di covarianza spaziale sia di tipo puro *nugget*, un qualsiasi campione $\chi_{\boldsymbol{s}}$ risulta scorrelato, con matrice di covarianza pari all'identità in $\mathbb{R}^n$, $\Sigma = \mathbb{I}_n$: in questo caso, nessuna deformazione interviene nello spazio campionario ad opera della struttura di covarianza del campione e, dunque, la metrica in $\Sigma^{-1} - H^n$ corrisponde esattamente alla metrica in H^n. Pertanto, all'aumentare dell'effetto *nugget*, la diminuzione della varianza dello stimatore $\widehat{\boldsymbol{a}}_{\boldsymbol{l}}$ nel passaggio da una stima OLS a una stima GLS, così come la differenza tra la stima iniziale e finale della varianza del residuo, diventano meno evidenti.

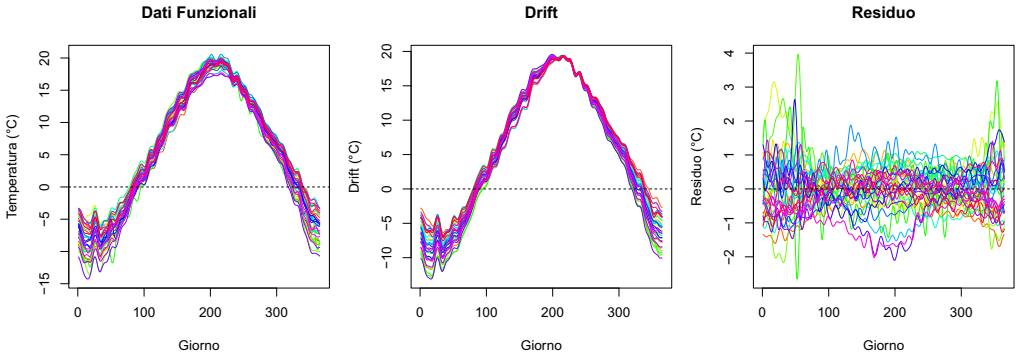

Figura 6.9: Decomposizione del processo come somma di *drift* e residuo, dopo cinque iterazioni dell'Algoritmo 5.1. A sinistra: dati funzionali. Al centro: *drift* stimato. A destra: residuo stimato.

Infine, in Figura 6.9 sono rappresentati i dati (pannello di sinistra), affiancati alla loro decomposizione come somma del termine di *drift* (al centro) e del residuo stazionario (a destra). Come si può notare, la media spaziale esprime le peculiarità comuni alle funzioni di temperatura osservate, quali la curvatura, la periodicità e l'andamento nel tempo; il termine di residuo descrive invece la componente stocastica, cogliendo la variabilità spaziale residua del fenomeno.

6.3.4 Previsione di Universal Kriging e Interpretazione dei Risultati

A partire dal dominio spaziale D, è stata costruita una griglia equispaziata composta da 101 nodi in direzione N-S (tolleranza $\sim 0.05°$) e 101 nodi in direzione W-E (tolleranza $\sim 0.09°$).

Grazie all'interpolazione di Universal Kriging è stato stimato il campo funzionale sulla suddetta griglia; una parte di tali curve è riportata nel pannello di sinistra di Figura 6.10, con le relative componenti di *drift* e di residuo.

Nelle Figure 6.11, 6.12 e 6.13 sono invece riportate le mappe contour dei campi di temperatura, *drift* e residuo, a tempo fissato: nelle immagini sono raffigurate le stime relative al primo giorno di ogni mese. Si precisa che, per rendere apprezzabili le variazioni spaziali del fenomeno, non è stato possibile adottare una scala colori comune; pertanto, occorre prestare attenzione qualora si vogliano confrontare i grafici mostrati.

Le mappe rappresentate in Figura 6.11 confermano il ruolo cruciale che l'esposizione verso il mare riveste nell'andamento delle temperature.

Infatti, dall'osservazione dei pannelli corrispondenti ai mesi di luglio e agosto si può notare che la zona costiera risulta essere più fresca rispetto alla zona interna, sia nella parte orientale della regione, esposta verso l'Oceano Atlantico, sia nella zona esposta verso la Deep Bay. Quest'ultima influenza notevolmente la distribuzione spaziale della temperatura nei mesi estivi, (fattore evidente soprattutto nel pannello corrispondente al primo giorno di

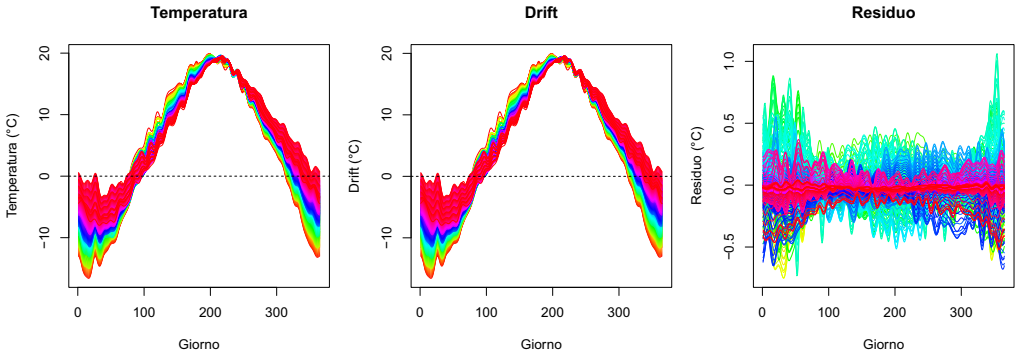

Figura 6.10: Funzioni ricostruite tramite Universal Kriging (a sinistra), *drift* stimato con GLS (al centro) e residuo stimato con Ordinary Kriging (a destra) per 511 dei $N_g = 10201$ punti $s = (\zeta, \varphi)$ della griglia di interpolazione considerata.

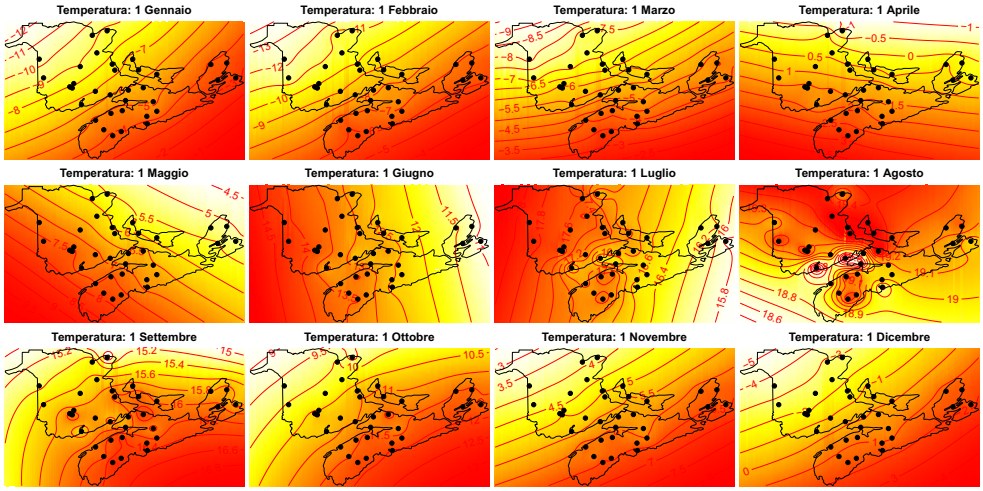

Figura 6.11: Contour-plot dell'interpolazione di Universal Kriging dopo 5 iterazioni dell'Algoritmo 5.1 per il primo giorno di ogni mese. La scala colori indica con il rosso temperature più alte, con il bianco temperature più basse.

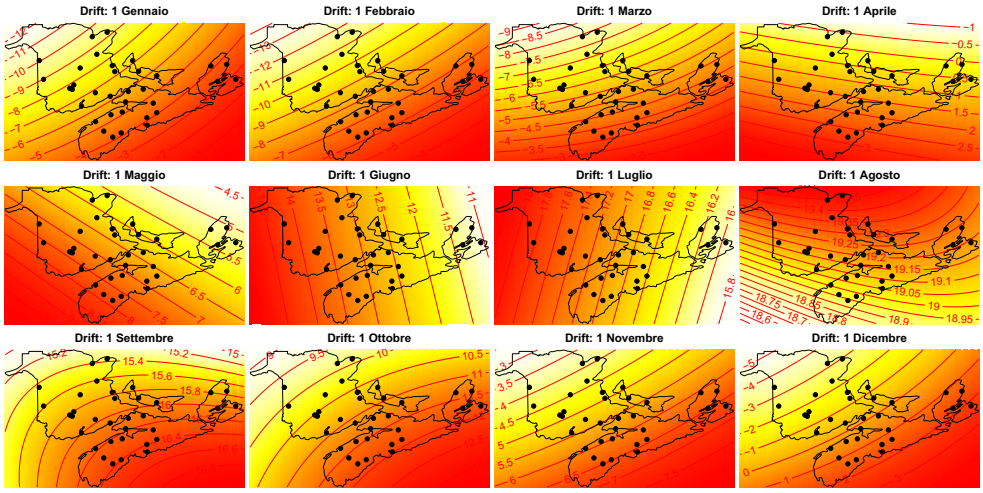

Figura 6.12: Contour-plot del *drift* delle province Marittime del Canada: stima GLS ottenuta dopo 5 iterazioni dell'Algoritmo 5.1 per il primo giorno di ogni mese. La scala colori indica con il rosso temperature più alte, con il bianco temperature più basse.

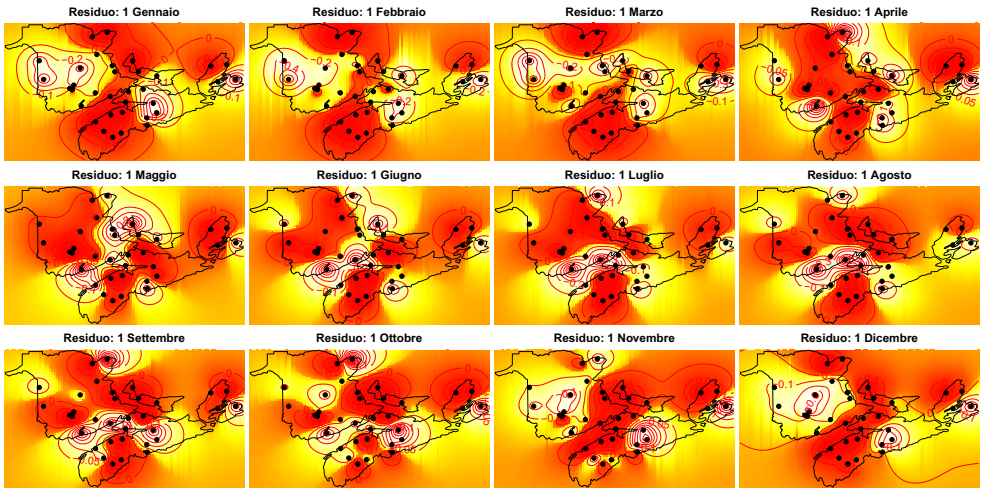

Figura 6.13: Contour-plot del residuo del processo, per il primo giorno di ogni mese, nelle province Marittime del Canada. Le mappe sono state ottenute tramite un'interpolazione di Ordinary Kriging dopo 5 iterazioni dell'Algoritmo 5.1. La scala colori indica con il rosso temperature più alte, con il bianco temperature più basse.

agosto), creando una zona concentrata di bassa temperatura in corrispondenza della baia stessa, circondata da zone a temperatura più alta in corrispondenza delle zone più interne.

Inoltre, in accordo con quanto affermato da Stanley (2002), la zona di Halifax, sulla costa orientale del Canada, mantiene temperature molto fresche durante i mesi estivi, a causa della brezza proveniente dall'oceano: questo si riflette in una zona a temperatura più bassa in corrispondenza della costa atlantica.

Differente è invece l'andamento delle temperature nei mesi autunnali e invernali, durante i quali la componente di esposizione verso il mare assume un ruolo opposto, rendendo le temperature più miti.

L'alternanza del ruolo che l'esposizione verso il mare riveste nella determinazione della temperatura è particolarmente evidente dall'osservazione delle mappe del *drift* in Figura 6.12.

Infatti, i pannelli relativi ai primi mesi dell'anno mostrano come le temperature più basse si verifichino in corrispondenza della zona continentale delle province canadesi, evidenziando un gradiente di temperatura in direzione NW-SE. La stima della temperatura media (i.e. *drift*) è dunque coerente con quanto affermato in (Stanley, 2002): le temperature invernali più rigide si registrano nella zona interna del New Brunswich, causate dalla circolazione di masse di aria fredda provenienti dalle regioni nord-occidentali.

A partire dai mesi primaverili, invece, il gradiente di temperatura inizia a subire una rotazione, fino ad arrivare alla situazione tipica dei mesi estivi, nella quale le temperature medie più fresche sono registrate in corrispondenza della zona marittima, dove si percepisce il refrigerio proveniente dalle brezze oceaniche, mentre l'entroterra è caratterizzato dalle temperature più alte.

Infine, la direzione del gradiente all'inizio di settembre muta nuovamente, conducendo a temperature più miti in corrispondenza delle zone marittime, raccordandosi infine nel corso dei mesi autunnali con la direzione di massima crescita della temperatura individuata nel mese di gennaio.

6.3.5 Confronto con le Mappe di Temperatura di Riferimento del Servizio Meteorologico del Canada

Al fine di verificare la coerenza dei risultati ottenuti attraverso la metodologia proposta con i risultati ufficiali reperibili attraverso il Servizio Meteorologico del Canada, le mappe calcolate con l'interpolazione di Universal Kriging sono state confrontate con le mappe disponibili in rete (`http://atlas.nrcan.gc.ca/`).

Le temperature di riferimento sono temperature medie registrate tra il 1971 e il 2000, calcolate dall'organismo Environment Canada in modo consistente con la metodologia dell'Organizzazione Meteorologica Mondiale. In particolare, i risultati reperibili in rete sono mappe di temperature minime e massime stagionali normali, ovvero calcolate come media aritmetica della temperatura media minima o massima per il periodo specifico (mesi di gennaio, aprile, luglio e ottobre), sulla base delle relative osservazioni giornaliere.

I modelli spaziali per la generazione di tali mappe di temperatura sono stati sviluppati con l'uso di *thin plate spatial smoothing spline*, attraverso l'algoritmo ANUSPLIN (Hutchinson, 2004).

In Figura 6.14, 6.15, 6.16 e 6.17 sono riportate, a confronto, le mappe di temperatura minima (a sinistra) e massima (al centro) di riferimento, a confronto con le mappe di temperatura media stimate tramite Universal Kriging, per i mesi di gennaio, aprile, luglio e ottobre.

Le mappe di Universal Kriging riportate sono state costruite considerando la temperatura media ricostruita nel mese relativo, ovvero come:

$$\overline{\chi}_s^* = \frac{1}{|\mathcal{T}_i|}\int_{\mathcal{T}_i} \chi_s^*(t)dt,$$

dove χ_s^* indica la stima del dato in posizione $s \in D$ ottenuta con Universal Kriging, mentre $\mathcal{T}_i \subset \mathcal{T}$ indica l'intervallo temporale corrispondente al mese considerato, tra gennaio, aprile, luglio e ottobre.

Dal confronto grafico si nota che le interpolazioni ottenute con il metodo proposto nel presente lavoro sono coerenti con le mappe di temperatura di riferimento, fornendo pertanto un'ulteriore validazione del modello stimato.

Si noti in particolare che grazie all'approccio funzionale adottato, le curve osservate sono state trattate come punti di un opportuno spazio funzionale, con lo scopo di stimare l'intera curva nel generico sito non osservato, e non la collezione di singoli punti corrispondenti alla temperatura ad ascissa fissata. Tuttavia, l'analisi delle curve ricostruite limitatamente a sottointervalli $\mathcal{T}_i$ del dominio $\mathcal{T}$, evidenzia come il metodo fornisca risultati ragionevoli anche localmente.

Dal punto di vista climatico, dal confronto delle mappe ricostruire con le mappe di riferimento si ottiene una conferma delle conclusioni tratte nella Sottosezione 6.3.4.

Infatti, nei mesi invernali (Figura 6.14) le temperature più alte si registrano nella parte costiera del Canada, e in particolare nella zona meridionale della Nova Scotia; invece, durante il mese di aprile (Figura 6.15), le temperature medie risultano ancora piuttosto rigide nella zona settentrionale della regione, mentre nella parte meridionale il prolungarsi della luminosità diurna e l'irraggiamento più forte del sole rendono le temperature più miti, con medie giornaliere sopra lo 0°C.

La stagione estiva è caratterizzata da temperature medie oltre i 10°C, evidenziando picchi attorno ai 20°C in alcune zone centrali del New Brunswich, dove in questo periodo si sperimentano le temperature massime annuali grazie all'irraggiamento solare e alla diminuzione delle precipitazioni. Le temperature massime nella zona costiera delle regioni Atlantiche del Canada sono invece maggiormente moderate dalla presenza dell'Oceano (Figura 6.16).

Infine, nei mesi autunnali (Figura 6.17), il prolungarsi delle notti conduce a temperature continentali più basse, mentre le zone costiere presentano ancora temperature miti per la stagione, grazie all'influsso delle correnti calde-umide dal mare.

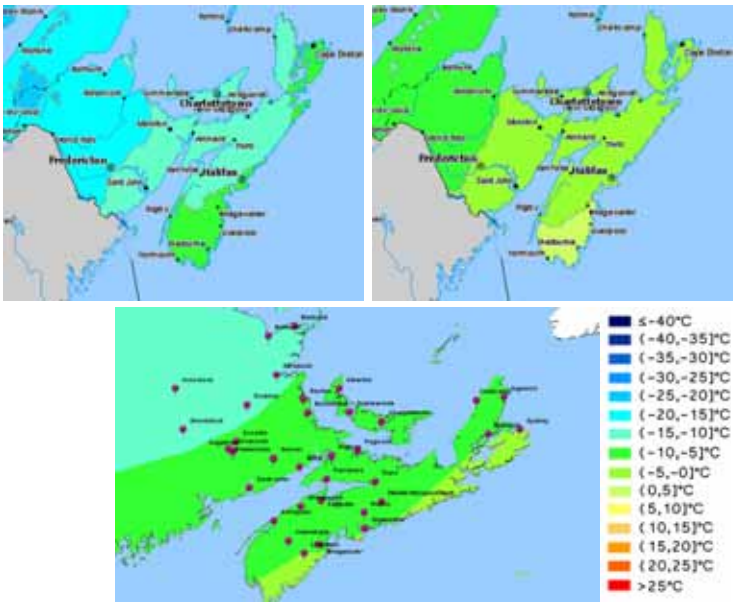

Figura 6.14: Confronto delle mappe di temperatura media giornaliera nel mese di gennaio stimate con Universal Kriging con le mappe di riferimento. In alto a sinistra: mappa di temperatura minima di riferimento; in alto a destra: mappa di temperatura massima di riferimento; in basso: mappa di Universal Kriging.

6.3.6 Analisi di Cross-Validazione

Il comportamento previsivo del metodo di Universal Kriging basato sul modello di variogramma individuato tramite cinque iterazioni dell'Algoritmo 5.1 è stato valutato attraverso un'analisi di cross-validazione *leave-one-out*.

In Figura 6.18 sono rappresentati i dati originali (pannello a sinistra) a confronto con le curve predette per cross-validazione (pannello al centro). Come si può notare, le funzioni predette presentano una varianza inferiore rispetto ai dati, sebbene mostrino le medesime peculiarità delle funzioni originali, come si osserva dalla visualizzazione del residuo di cross-validazione (pannello a destra) che non evidenzia una forte eteroschedasticità rispetto all'ascissa temporale.

Dall'osservazione della Figura 6.18 si osserva che il residuo di cross-validazione si mantiene nel complesso limitato. Inoltre, dalla stima della relativa deviazione standard (pannello di destra) si può notare che l'incertezza legata alla stima di cross-validazione risulta leggermente più alta nei mesi invernali, durante i quali la variabilità dei dati è superiore, mentre la qualità della previsione è massima nei mesi primaverili, associati a una variabilità inferiore del fenomeno.

I residui più significativi sono registrati nei mesi di gennaio, febbraio e dicembre relativamente alle località di Bertrand e Bathurst nella zona nord-orientale della provincia di

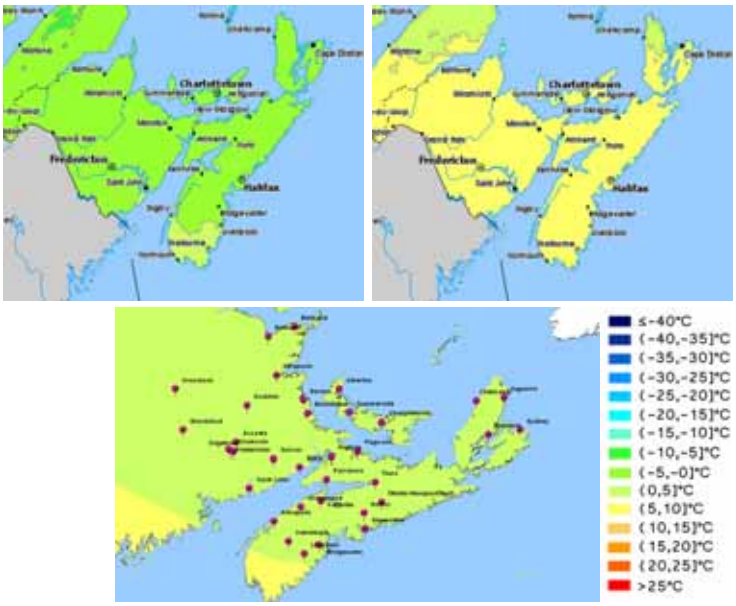

Figura 6.15: Confronto delle mappe di temperatura media giornaliera nel mese di aprile stimate con Universal Kriging con le mappe di riferimento. In alto a sinistra: mappa di temperatura minima di riferimento; in alto a destra: mappa di temperatura massima di riferimento; in basso: mappa di Universal Kriging.

New Brunswich (Figura 6.19). Questo risultato, in accordo con quanto ottenuto in (Giraldo *e altri*, 2010a), è motivato dal fatto che, sebbene la temperatura media giornaliera nelle due località presenti un comportamento molto simile, in alcuni giorni dei mesi invernali (in particolare nei giorni 19 gennaio, 17-25 febbraio, 18, 19, 21 dicembre, ovvero $t = \{19; 49-57; 353; 354; 356\}$ (Giraldo *e altri*, 2010a)), la differenza di temperatura registrata presso tali stazioni meteorologiche è stata superiore a 4°C. Dal momento che questi siti sono molto vicini, l'influenza reciproca nella previsione di cross-validazione è significativa, conducendo a residui di cross-validazione elevati rispetto alla media.

In Figura 6.20 sono invece rappresentati, a confronto, i violin-plot relativi alle temperature medie mensili osservate e di cross-validazione, nei mesi di gennaio, aprile, luglio e ottobre. Tali temperature sono state calcolate come media integrale delle rispettive funzioni di temperatura, ovvero come:

$$\begin{aligned} \overline{\chi}_s &= \frac{1}{|\mathcal{T}_i|}\int_{\mathcal{T}_i} \chi_s(t)dt, \\ \overline{\chi}_s^{*CV} &= \frac{1}{|\mathcal{T}_i|}\int_{\mathcal{T}_i} \chi_s^{*CV}(t)dt, \end{aligned}$$

dove $\mathcal{T}_i$ indica l'intervallo temporale corrispondente al mese considerato tra gennaio, aprile, luglio e ottobre.

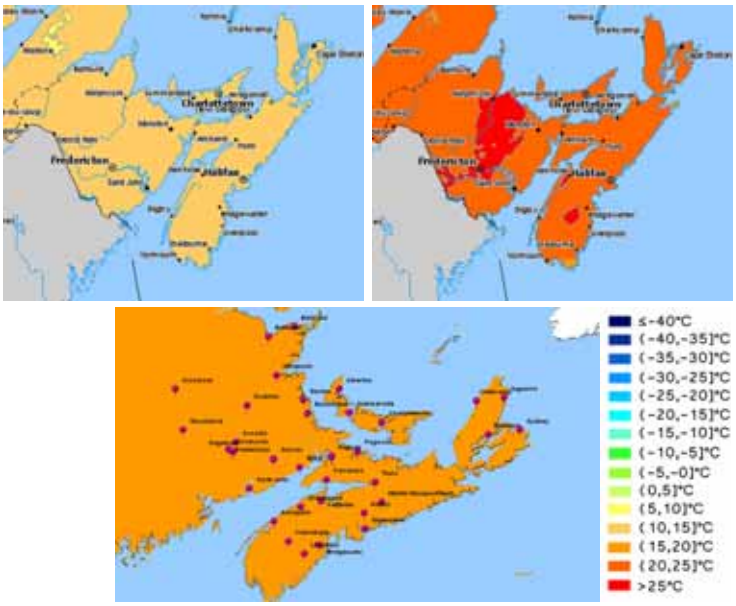

Figura 6.16: Confronto delle mappe di temperatura media giornaliera nel mese di luglio stimate con Universal Kriging con le mappe di riferimento. In alto a sinistra: mappa di temperatura minima di riferimento; in alto a destra: mappa di temperatura massima di riferimento; in basso: mappa di Universal Kriging.

Le parti centrali dei violin-plot in Figura 6.20, corrispondenti ai box-plot dei dati, evidenziano come, dal punto di vista distribuzionale, la previsione di cross-validazione sia fedele ai dati relativamente ai mesi di gennaio, aprile e ottobre, mostrando soltanto una mediana leggermente più alta rispetto al dato durante il mese di aprile.

Il mese di luglio presenta invece una distribuzione predetta decisamente più concentrata rispetto alla distribuzione del dato: infatti, come si osserva dalla Figura 6.18, la previsione di cross-validazione nei mesi estivi è caratterizzata dalla variabilità minima.

Analoghe conclusioni si ricavano dall'osservazione dei bordi dei violin-plot, corrispondenti alle densità di probabilità stimate dai dati e dalle previsioni di cross-validazione, che evidenziano una notevole differenza nella distribuzione del dato rispetto alla stima di cross-validazione soltanto per il mese di luglio.

I risultati relativi al mese estivo non sorprendono qualora si analizzino con attenzione i pannelli relativi ai mesi di luglio e agosto in Figura 6.11. Infatti, tali mesi sono caratterizzati da un comportamento spaziale piuttosto complesso, con particolare riferimento alla parte centrale della regione, corrispondente alla Deep Bay: in questo periodo, l'influenza delle funzioni osservate è dunque molto significativa e la previsione risulta più difficoltosa, anche a causa dell'effetto regolarizzante del metodo di kriging.

Infine, in Tabella 6.2 sono riportate le statistiche relative alla distribuzione dello scarto quadratico SSE e dello scarto quadratico relativo $SSE^{(rel.)}$, stimate dall'analisi di cross-

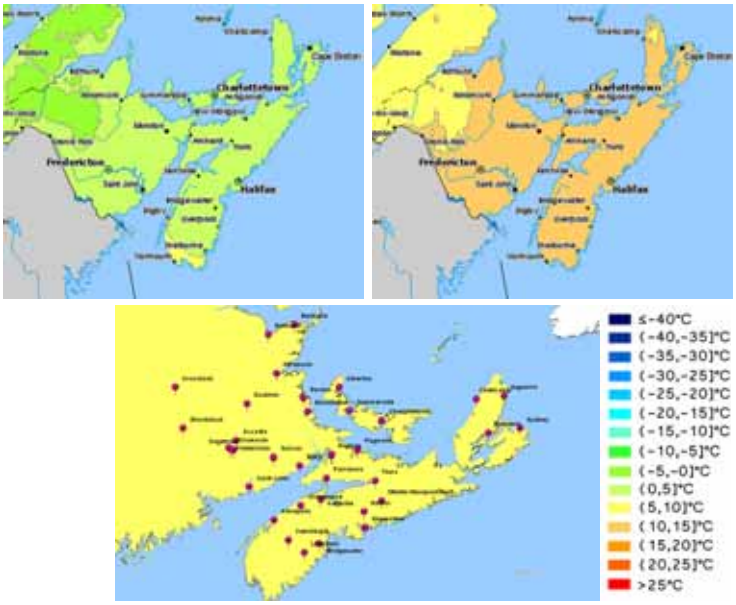

Figura 6.17: Confronto delle mappe di temperatura media giornaliera nel mese di ottobre stimate con Universal Kriging con le mappe di riferimento. In alto a sinistra: mappa di temperatura minima di riferimento; in alto a destra: mappa di temperatura massima di riferimento; in basso: mappa di Universal Kriging.

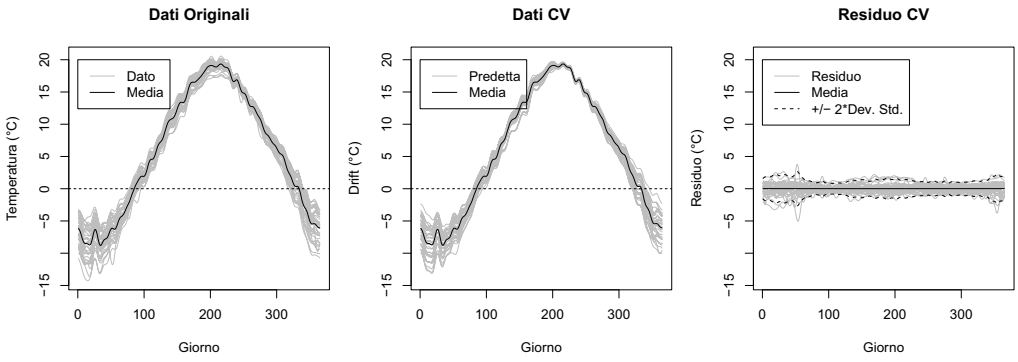

Figura 6.18: Analisi di Cross-Validazione. A sinistra: dati originali (in grigio) e relativa media campionaria (in nero). Al centro: dati predetti con Universal Kriging (in grigio) e loro media (in nero). A destra: differenza r_i^{CV}, $i = 1, ..., n$, tra i dati originali e i dati predetti (in grigio), loro media $\widehat{m}_r$ (in nero) e banda di confidenza puntuale $\widehat{m}_r(z) \pm 2\widehat{\sigma}_r(z)$ (in nero tratteggiato), dove $\widehat{\sigma}_r(z)$ è la deviazione standard stimata puntualmente dal residuo di cross-validazione.

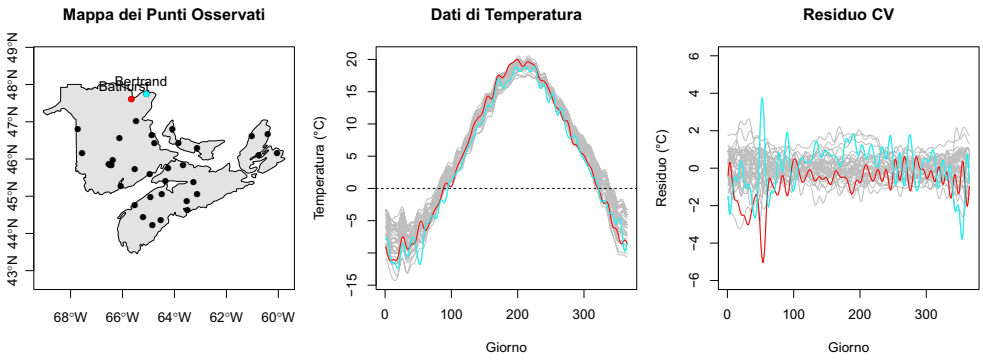

Figura 6.19: Analisi di Cross-Validazione. A sinistra: mappa dei punti osservati con indicazione delle località di Bertrand (in celeste) e Bathurst (in rosso) (New Brunswich). Al centro: dati di temperatura nelle 35 stazioni meteorologiche; le linee colorate corrispondono ai dati di Bertrand (in celeste) e Bathurst (in rosso). A destra: residui di cross-validazione; le linee colorate si riferiscono ai residui relativi a Bertrand (in celeste) e Bathurst (in rosso).

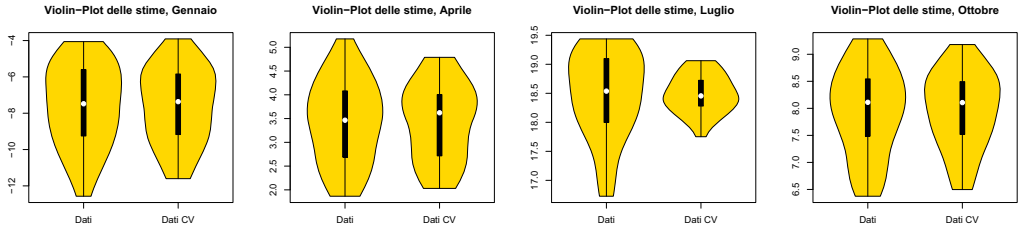

Figura 6.20: Analisi di Cross-Validazione. Violin-plot relativi alle temperature medie nei mesi di gennaio ($\mathcal{T}_1 = [1, 31]$), aprile ($\mathcal{T}_4 = [91, 120]$), luglio ($\mathcal{T}_7 = [182, 212]$) e ottobre ($\mathcal{T}_{10} = [274, 305]$), a confronto con le temperature medie di cross-validazione nei medesimi mesi.

	SSE	$SSE^{(rel.)}$
Minimo	22.14	$5.08 \cdot 10^{-4}$
Mediana	125.2	$2.87 \cdot 10^{-3}$
Media	180.2	$4.13 \cdot 10^{-3}$
Massimo	538.4	$1.23 \cdot 10^{-2}$
Dev. Std.	155.26	$3.56 \cdot 10^{-3}$
Somma	6308.65	0.144
$\mathbb{E}_n[\|\|\chi_s\|\|^2]$	43624.63	

Tabella 6.2: Analisi di Cross-Validazione. Statistiche della distribuzione dello scarto quadratico SSE e dello scarto quadratico relativo $SSE^{(rel.)}$ per il dataset *Canada's Maritime Provinces Temperatures*.

validazione.

Le statistiche riportate in Tabella 6.2 confermano le prestazioni soddisfacenti del metodo dal punto di vista previsivo. I valori di tali statistiche saranno confrontate nella Sezione 6.4 con le statistiche ottenute dalle analisi effettuate con i metodi di kriging funzionali alternativi, presenti in letteratura.

6.4 Confronto con le Analisi Presenti in Letteratura

La Sezione 6.3 è stata dedicata all'analisi geostatistica del dataset *Canada's Maritime Provinces Temperatures* attraverso le metodologie sviluppate nei Capitoli 4 e 5, all'interno della cornice teorica fornita dal Capitolo 3.

L'obiettivo di questa sezione è il confronto dei risultati ottenuti nel lavoro di tesi con i risultati delle analisi dello stesso dataset presentati in letteratura nell'ambito dei metodi di kriging per dati funzionali.

Per questo motivo, dopo una breve introduzione alle tecniche di kriging proposte in letteratura, saranno valutate, a confronto, le prestazioni previsive dei modelli sviluppati attraverso la metodologia proposta in questo lavoro e dei metodi alternativi. In un secondo momento, sarà studiata l'influenza della metrica non euclidea e, quindi, dell'introduzione del termine di *drift* sulla stima del campo e sull'incertezza ad essa associata.

6.4.1 I Modelli Alternativi

Il dataset di temperature studiato nel presente capitolo è stato analizzato in letteratura come applicazione dei seguenti metodi di kriging per dati funzionali:

1. Ordinary Kriging per Dati Funzionali (OKFD) (Giraldo *e altri*, 2008a), (Giraldo, 2009), (Giraldo *e altri*, 2010c);

2. Point-Wise Kriging per Dati Funzionali (PWKFD) (Giraldo *e altri*, 2008b), (Giraldo *e altri*, 2010a);

3. Functional Kriging Total Model (FKTM) (Giraldo, 2009), (Delicado *e altri*, 2010).

Tali metodi si basano sull'ipotesi di stazionarietà e isotropia (puntuale) del processo che ha generato i dati e differiscono tra loro soltanto per la costruzione dello stimatore lineare di kriging.

Più precisamente, il metodo OKFD, si basa sullo stimatore lineare a coefficienti costanti descritto nella Sezione 3.4.1:

$$\boldsymbol{\chi}^*_{s_0} = \sum_{i=1}^{n} \lambda_i \boldsymbol{\chi}_{s_i}, \quad s_0, s_1, ..., s_n \in D.$$

La formulazione di tale metodo, presente ad esempio in (Giraldo *e altri*, 2008a), è stata proposta a partire da definizioni puntuali di stazionarietà e isotropia, limitatamente allo spazio $H = L^2$.

Ancora in $H = L^2$ è stato sviluppato il metodo PWKFD (Giraldo *e altri*, 2008b), che prevede invece la costruzione di uno stimatore lineare a coefficienti funzionali di tipo:

$$\boldsymbol{\chi}^*_{s_0}(t) = \sum_{i=1}^{n} \lambda_i(t) \boldsymbol{\chi}_{s_i}(t), \quad s_0, s_1, ..., s_n \in D, \, t \in \mathcal{T}.$$

Infine, lo stimatore proposto dal metodo FKTM (Giraldo, 2009) è costruito in modo più complesso tramite la seguente combinazione lineare con coefficienti di due variabili:

$$\boldsymbol{\chi}^*_{s_0}(t) = \sum_{i=1}^{n} \int_{\mathcal{T}} \lambda_i(t,z) \boldsymbol{\chi}_{s_i}(z) dz, \quad s_0, s_1, ..., s_n \in D, \, t \in \mathcal{T}.$$

In tutti i casi, la determinazione dei parametri ottimi è stata stabilita sulla base del criterio dei minimi quadrati, con il vincolo di non distorsione, ovvero risolvendo il problema di ottimo vincolato seguente:

$$\min_{\lambda_1,...,\lambda_n} \mathbb{E}[\|\boldsymbol{\chi}^*_{s_0} - \boldsymbol{\chi}_{s_0}\|^2], \quad \text{t.c} \quad \mathbb{E}[\boldsymbol{\chi}^*_{s_0} - \boldsymbol{\chi}_{s_0}] = 0.$$

Per il metodo OKFD, è possibile determinare esplicitamente la soluzione (unica) del problema di minimizzazione vincolata, risolvendo il sistema lineare esibito nella Sottosezione 3.4.1 del Capitolo 3.

Per quanto concerne i metodi PWKFD e FKTM, l'espressione esplicita degli stimatori lineari di kriging è disponibile per dati proiettati su una base funzionale, non necessariamente ortogonale, ovvero per dati della forma:

$$\sum_{j=1}^{M} \beta_j B_j(t),$$

dove β_j indica il coefficiente dell'espansione relativo al j-esimo elemento della base $\{B_j\}$, con $j = 1, ..., M$. In questo caso, la soluzione al problema di ottimo è stata individuata sulla base di un modello lineare di coregionalizzazione (LMC, (Cressie, 1993)) per i vettori di coefficienti $\boldsymbol{\beta} = (\beta_1, ..., \beta_M)$ (e.g. (Giraldo, 2009), (Giraldo *e altri*, 2010a)).

Al confronto dei risultati dell'analisi illustrata in questo lavoro con le analisi presenti in letteratura è opportuno premettere alcune considerazioni circa le definizioni stazionarietà.

Innanzi tutto occorre sottolineare che il metodo di Universal Kriging (inteso associato a una stima preliminare del *drift* con l'Algoritmo 5.1 eventualmente selezionato con l'Algoritmo 5.2) si basa su ipotesi diverse rispetto ai metodi di OKFD, PWKFD e FKTM: infatti, questi ultimi sono fondati sull'assunzione di stazionarietà (puntuale) del processo, mentre il primo si basa sull'ipotesi, più debole, di stazionarietà (globale) del residuo del processo.

Più precisamente, le definizioni di stazionarietà e isotropia fornite nei lavori citati si basano sulle seguenti condizioni puntuali (Giraldo *e altri*, 2010a):

(i) $\mathbb{E}[\boldsymbol{\chi}_s(t)] = m(t), \quad \forall t \in \mathcal{T},\ s \in D;$

(ii) $\mathrm{Cov}(\boldsymbol{\chi}_{s_i}(t), \boldsymbol{\chi}_{s_j}(u)) = C(h; t, u), \quad t, u \in \mathcal{T};\ s_i, s_j \in D;\ h = \|s_i - s_j\|;$

(iii) $\frac{1}{2}\mathrm{Var}(\boldsymbol{\chi}_{s_i}(t) - \boldsymbol{\chi}_{s_j}(u)) = \gamma(h; t, u), \quad t, u \in \mathcal{T};\ s_i, s_j \in D;\ h = \|s_i - s_j\|.$

Nella cornice teorica stabilita da tali definizioni risulta difficile verificare la stazionarietà e isotropia del campo aleatorio esaminato, non potendo determinare una stima del variogramma e dei cross-variogrammi per ogni ascissa nel dominio di osservazione. Questa difficoltà è invece completamente superata dall'introduzione del formalismo del Capitolo 3 e in particolare dalle Definizioni 3.1, 3.2, 3.7 e 3.11, sufficienti alla formulazione di Ordinary e Universal Kriging.

Si noti tuttavia che non è dimostrato che le assunzioni di stazionarietà adottate nel presente lavoro implichino la stazionarietà del campo multivariato dei coefficienti $\boldsymbol{\beta}$, intesa nel senso dei LMC; non è altresì evidente che le ipotesi (i)-(iii) siano sufficienti a garantire la stazionarietà dei coefficienti stessi[1].

Nel caso specifico, dalle analisi svolte nella Sottosezione 6.3.1, è emerso che, adottando la metrica indotta dalla distanza geodetica, il processo di partenza non risulta stazionario secondo la Definizione 3.7 (quindi nemmeno secondo la nozione di stazionarietà derivante dalle (i)-(iii), *cfr.* Esempio 3.3): la scelta della metrica euclidea dalle coordinate geografiche, definita dalla (6.2), potrebbe dunque aver creato distorsioni nella stima variografica, conducendo a un modello giudicato come stazionario, a fronte di un processo stocastico in realtà non stazionario.

All'influenza della metrica euclidea e delle conseguenti ipotesi di stazionarietà sulla previsione di kriging per il metodo OKFD sarà dedicata la Sottosezione 6.4.3.

[1] È invece vero che la stazionarietà al second'ordine riferita all'operatore di covarianza spaziale (3.13) è equivalente alla stazionarietà al second'ordine dei coefficienti dell'espansione su una base ortonormale (cfr. Sottoezione 3.2): in tali ipotesi e rispetto a una siffatta base è sviluppata la stima del modello FKTM in (Monestiez e Nerini, 2008).

	UKFD	OKFD	PWKFD	FKTM
Minimo	114.1	103.7	104.4	104.3
Mediana	227.3	253.4	252.7	252.8
Media	300.8	299.5	299.2	298.9
Massimo	843.9	890.8	902.1	899.8
Dev. Std.	179.7	178.4	175.4	176.1
Somma	10528.1	10483.9	10471.3	10461

Tabella 6.3: Confronto del potere previsivo dell'Universal Kriging per dati funzionali (UKFD) proposto in questo lavoro e dei metodi precedentemente in uso (OKFD, PWKFD, FKTM) attraverso le statistiche della distribuzione dello scarto quadratico SSE rispetto ai dati puntuali osservati.

6.4.2 Confronto dei Risultati di Cross-Validazione

Nella Sottosezione 6.3.6 è stato analizzato il comportamento previsivo del metodo di Universal Kriging associato alla stima del *drift* ottenuta da cinque iterazioni dell'Algoritmo 5.1. In particolare, si è concluso che i risultati di cross-validazione sono soddisfacenti in termini di SSE, evidenziando uno scarto molto limitato rispetto alla norma quadratica della funzione.

I risultati di cross-validazione saranno ora confrontati con i risultati di cross-validazione ottenuti dall'analisi del dataset di temperature con i metodi di kriging per dati funzionali proposti in letteratura.

Si puntualizza che, nei lavori citati, la valutazione della qualità della previsione fornita dai metodi proposti è stata valutata in relazione alle osservazioni puntuali del dataset originale (Figura 6.2, pannello centrale), non rispetto al dataset funzionale ottenuto per proiezione sulla base di Fourier (Figura 6.2, pannello di destra), come invece è stato fatto nel corso di tutte le analisi svolte.

Per questo motivo in Tabella 6.3, coerentemente con le valutazioni fornite in letteratura, sono confrontate le statistiche della distribuzione dello scarto quadratico $SSE^{(p)}$ tra i dati predetti per cross-validazione e i dati puntuali, definendo:

$$SSE_i^{(p)} = \sum_{i=1}^{n}\sum_{j=1}^{M}(\chi_{s_i}(t_j) - \chi_{s_i}^{*CV}(t_j))^2, \quad s_i \in D, t_1, ..., t_M \in \mathcal{T},$$

dove $t_j = j$, $j = 1, ..., 365$, indica il giorno di rilevamento dell'osservazione.

Dall'analisi della distribuzione dello scarto SSE attraverso le statistiche in Tabella 6.3 si può osservare che, nel passaggio da un approccio di tipo stazionario basato sulla metrica euclidea a un approccio non stazionario supportato da una metrica non euclidea, la riduzione dello scarto di cross-validazione è di circa il 10% relativamente alla mediana e di circa il 5% per il massimo, mentre non si registrano differenze significative rispetto alle altre statistiche.

Tale osservazione è confermata dai violin-plot mostrati nel pannello di sinistra di Figura 6.21, nel quale si nota che la distribuzione dello scarto quadratico relativo al metodo UKFD

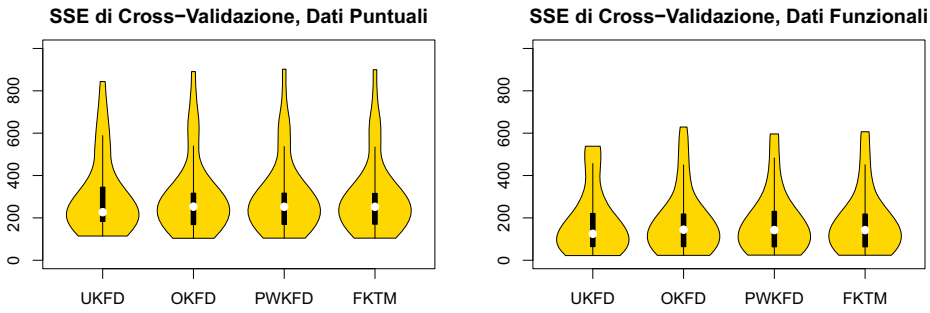

Figura 6.21: Analisi di Cross-Validazione. Violin-plot relativi alle temperature medie nei mesi di gennaio ($\mathcal{T}_1 = [1, 31]$), aprile ($\mathcal{T}_4 = [91, 120]$), luglio ($\mathcal{T}_7 = [182, 212]$) e ottobre ($\mathcal{T}_{10} = [274, 305]$), a confronto con le temperature medie di cross-validazione nei medesimi mesi.

risulta più asimmetrica e concentrata su valori attorno alla mediana, presentando una coda destra meno pesante rispetto ai metodi alternativi.

In continuità con le valutazioni svolte nel lavoro di tesi, in Tabella 6.4 sono riportate le statistiche dello scarto quadratico valutato rispetto al dataset funzionale. Si precisa che le statistiche relative ai metodi di kriging funzionale OKFD, PWKFD e FKTM sono state calcolate grazie ai codici disponibili in rete (`http://www.docentes.unal.edu.co/rgiraldoh/docs/`) e all'interno del pacchetto `geofd` del software `R` (Giraldo *e altri*, 2010b), e non sono state pubblicate nei lavori citati.

Tuttavia, si è ritenuto interessante valutare il comportamento del metodo anche attraverso lo scarto quadratico SSE rispetto al dataset funzionale con il quale sono costruiti gli stimatori, dal momento che il dato predetto per cross-validazione, in caso di stima esatta, coinciderebbe con il dato funzionale, non con la collezione di dati puntuali registrati inizialmente.

Per questo motivo, in Tabella 6.4 sono riportate le statistiche di cross-validazione ottenute con i quattro metodi. Come si può notare dalla seconda riga della Tabella, la mediana dello scarto SSE presenta in questo caso una riduzione del 13% circa per il metodo di UKFD rispetto ai metodi OKFD, PWKFD e FKTM (del 10% circa rispetto al massimo). Analogamente al caso precedente, dai violin-plot nel pannello di destra di Figura 6.21 si osserva che la distribuzione dello scarto SSE relativo al metodo proposto risulta più concentrata rispetto alla distribuzione della scarto relativa alle analisi svolte con i precedenti metodi.

Si noti peraltro che la metodologia di kriging usata per la presente analisi si basa su uno stimatore lineare a coefficienti costanti, indicando che la variazione più significativa della distribuzione dello scarto si registra nel cambiamento della metrica e, dunque, del tipo di ipotesi si stazionarietà, piuttosto che nella variazione della forma funzionale proposta per lo stimatore, come si nota rispettivamente dal confronto della prima colonna con le altre tre e dal confronto delle ultime tre colonne di Tabella 6.3.

	UKFD	OKFD	PWKFD	FKTM
Minimo	22.14	23.13	24.23	23.68
Mediana	125.2	144.7	142.8	142.6
Media	180.2	179.2	178.9	178.6
Massimo	538.4	629.0	596.3	606.7
Dev. Std.	155.3	153.8	150.1	150.9
Somma	6308.7	6273.7	6260.3	6250.6

Tabella 6.4: Confronto del potere previsivo dell'Universal Kriging per dati funzionali (UKFD) proposto in questo lavoro e dei metodi precedentemente in uso (OKFD, PWKFD, FKTM) attraverso le statistiche della distribuzione dello scarto quadratico SSE rispetto ai dati proiettati sulla base di Fourier.

Per questo motivo, nella sottosezione seguente saranno analizzati, a confronto, i risultati ottenuti per Ordinary Kriging sulla base della metrica euclidea e i risultati di Universal Kriging discendenti dall'adozione di una metrica non euclidea, essendo possibile inserire tali metodologie nel contesto teorico introdotto nel Capitolo 3.

6.4.3 Confronto delle Previsioni Fornite da Ordinary e da Universal Kriging

Sebbene dal punto di vista previsivo le variazioni nei valori delle statistiche riportate in Tabella 6.3 e 6.4 siano apprezzabili solo in termini di mediana e di massimo, dal punto di vista concettuale è importante rimarcare che l'analisi geostatistica illustrata nel presente capitolo evidenzia come la scelta iniziale della metrica sia fondamentale nell'analisi di stazionarietà del processo, portando a risultati in direzioni completamente opposte variando il tipo di distanza adottata.

L'obiettivo della presente sottosezione è il confronto delle mappe di kriging ottenute con i metodi OKFD e UKFD, al fine di valutare l'influenza della metrica per D e del modello di variogramma conseguentemente stimato, sulla previsione finale di kriging e sull'incertezza ad essa associata. Si precisa che le stime dei variogrammi e la previsione del campo attraverso il metodo OKFD sono state ottenute applicando i codici disponibili in rete (riportati in Appendice B.3) e non sono pubblicate nei lavori citati.

In Figura 6.22 sono mostrati gli stimatori empirici *variogram cloud* e *binned variogram* dai dati, calcolate in base alle metriche euclidea (a sinistra) e non euclidea (a destra). Come si può notare dai grafici, la forma della stima *binned variogram* risente della metrica adottata, evidenziando, nel caso euclideo, un andamento irregolare. La presenza di tali irregolarità nella forma del variogramma, indice di una leggera anisotropia (Figura 6.23), potrebbe indicare in questo caso che la metrica con la quale si misurano le distanze tra i punti introduce delle deformazioni nella stima variografica.

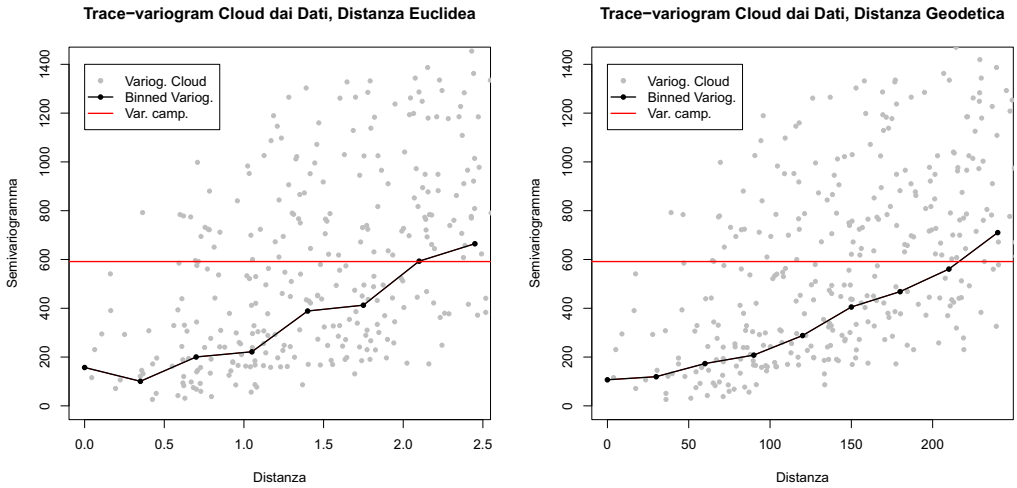

Figura 6.22: *Trace-Variogram Cloud* dai dati adottando la metrica euclidea (a sinistra) o la metrica indotta dalla distanza geodetica (a destra).

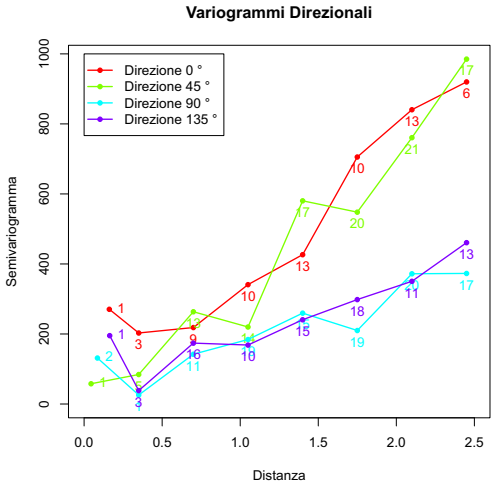

Figura 6.23: *Binned trace-variogram* direzionali calcolati con la metrica euclidea. Per ogni valore della distanza h_k, è indicato il numero di coppie $|N(h_k)|$ appartenenti alla classe $N(h_k)$, per $k = 1, ..., K$. In tutti i pannelli i colori distinguono la direzione considerata.

Per quanto concerne l'adattamento di un modello valido $\gamma(h;\boldsymbol{\vartheta})$ alla stima empirica, la metodologia di stima adottata nel lavoro di tesi è differente dalla metodologia proposta nelle analisi del dataset di temperatura come applicazione del metodo OKFD presentate in letteratura (Giraldo *e altri*, 2008a), (Giraldo, 2009), (Giraldo *e altri*, 2010c).

Infatti, nell'ultimo caso, la procedura di stima ha previsto l'adattamento di un modello sferico allo stimatore empirico *variogram cloud*, con il metodo dei minimi quadrati ordinari. In questo lavoro, invece, è stato adottato il modello esponenziale, a partire dallo stimatore *binned variogram*, adattato con il metodo dei minimi quadrati generalizzati (attraverso l'approssimazione MC-bootstrap).

Inoltre, in tutte le analisi svolte nel lavoro di tesi, la stima empirica *binned variogram* è stata troncata a una distanza di circa la metà dell'ampiezza minima del dominio, al fine di ottenere una stima empirica affidabile per ciascuna classe di distanze: la distanza massima di osservazione, nel caso piano, non dovrebbe eccedere il raggio del cerchio inscritto nel dominio spaziale D (Armstrong, 1998). In linea con tale principio, nel caso di geometria non euclidea, la distanza massima è stata valutata considerando circa la metà della distanza geodetica minima tra le coordinate geografiche degli estremi del dominio ($\sim$250 km). Non è invece stato inserito alcun troncamento nella procedura di stima adottata per l'analisi con il metodo OKFD (il troncamento secondo il criterio proposto sarebbe stato a 2.5° circa).

I modelli validi di variogramma adattati con i metodi illustrati e usati nella stima di kriging sono riportati in Figura 6.24. Si noti che il modello usato nella stima di OKFD (a sinistra) è stimato dai dati, mentre il modello usato per la stima di UKFD (a destra) è calcolato dai residui: questo comporta che la varianza campionaria indicata con la linea rossa sia differente nei due casi, in quanto relativa a quantità diverse.

Dall'osservazione dei grafici in Figura 6.24 si nota che il modello di variogramma adottato per la stima di Ordinary Kriging è caratterizzato da valori decisamente più alti rispetto al modello di variogramma adottato nella stima di Universal Kriging: nel secondo caso, infatti, una parte consistente della variabilità spaziale è attribuita al termine deterministico di *drift*.

Oltre a un *sill* notevolmente più alto, il primo modello stimato è caratterizzato da un *range* superiore alla massima distanza tra i dati, indicazione che, secondo il modello di dipendenza spaziale, le osservazioni del processo si mantengono correlate fino a distanze superiori a 8° (pari a circa 640 km, (6.2)); al contrario, il *practical range* relativo al secondo modello indica che la dipendenza spaziale della componente stocastica è considerata evanescente entro 150 km, oltre i quali la variabilità del fenomeno è da imputarsi al termine di *drift* deterministico.

Le differenze tra i modelli di dipendenza spaziale stimati nei due casi si ripercuotono notevolmente sulle stime di Ordinary e Universal Kriging, le cui mappe mensili sono riportate rispettivamente in Figura 6.25 e 6.26 (la scala colori è comune, a mese fissato, per le mappe di OKFD e UKFD).

In particolare di osserva che la stima di OKFD risulta sensibilmente *data driven*, presentando linee di livello dall'andamento irregolare, soprattutto nei mesi estivi, per i quali la previsione è più difficoltosa anche per il metodo di UKFD. Per quest'ultimo è invece più

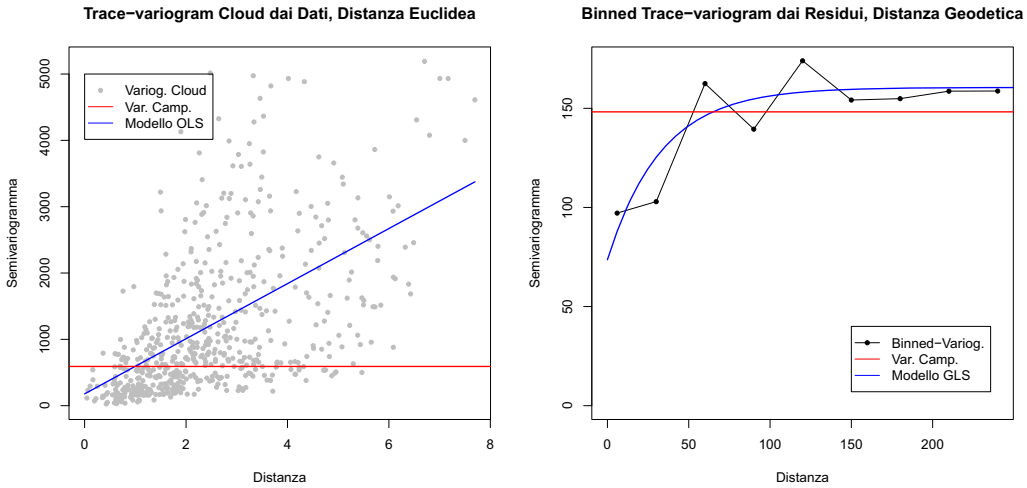

Figura 6.24: Modelli di variogramma stimati per le analisi del dataset di temperature con il metodo di OKFD (a sinistra) e UKFD (a destra). Per il metodo OKFD lo stimatore empirico (*cloud*) è calcolato dai dati, così come la varianza campionaria. Per il metodo UKFD lo stimatore sperimentale (*binned*) e la varianza campionaria sono calcolati dai residui.

evidente l'influenza del termine di *drift*, che determina una stima spazialmente più liscia, come si nota dall'andamento delle linee di livello.

Inoltre, il metodo di OKFD risente notevolmente dell'effetto regolarizzante del kriging, effetto che invece è parzialmente mitigato, nella stima di UKFD, dalla presenza del *drift*. Questo fattore è particolarmente evidente durante i mesi invernali nelle zone periferiche della regione considerata (soprattutto nelle parti N-W e S-E) nelle quali le stime fornite dal metodo di Universal Kriging raggiungono valori di oltre 1.5°C più estremi delle previsioni di Ordinary Kriging.

Le differenze individuate nelle mappe di kriging evidenziano come il modello di variogramma adottato $\gamma(h;\widehat{\boldsymbol{\vartheta}})$, combinato con la presenza (o assenza) del termine di *drift*, risulti molto influente sulla stima spaziale. Infatti, a causa del valore contenuto del *range* stimato, nelle zone periferiche del dominio spaziale D, lontane dai siti di osservazione, la stima di Universal Kriging si assesta sul valore medio, individuato dal *drift*; d'altra parte, l'influenza del termine di *drift* è evidente anche nella parte centrale della regione stessa, a causa della bassa variabilità stocastica residua del fenomeno, indicata dal parametro di *sill*.

Si noti infine che l'effetto del modello di variogramma si ripercuote anche sulla varianza di kriging: infatti, come evidenziato dalla Figura 6.27, la deviazione standard normalizzata è decisamente inferiore per la stima di Universal Kriging, indicando che l'introduzione del termine di *drift* nel modello riduce significativamente l'incertezza legata alla previsione.

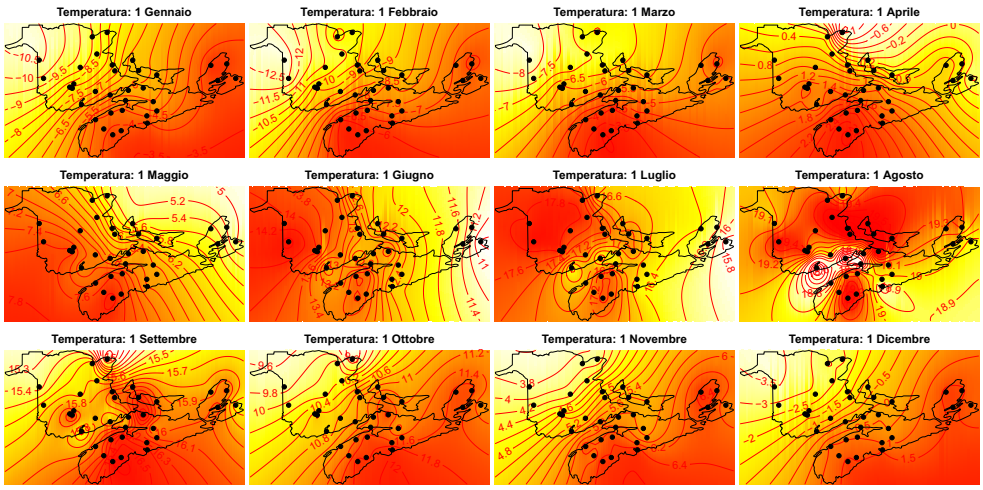

Figura 6.25: Contour-plot dell'interpolazione di Ordinary Kriging basata sulla metrica euclidea per il primo giorno di ogni mese. La scala colori indica con il rosso temperature più alte, con il bianco temperature più basse.

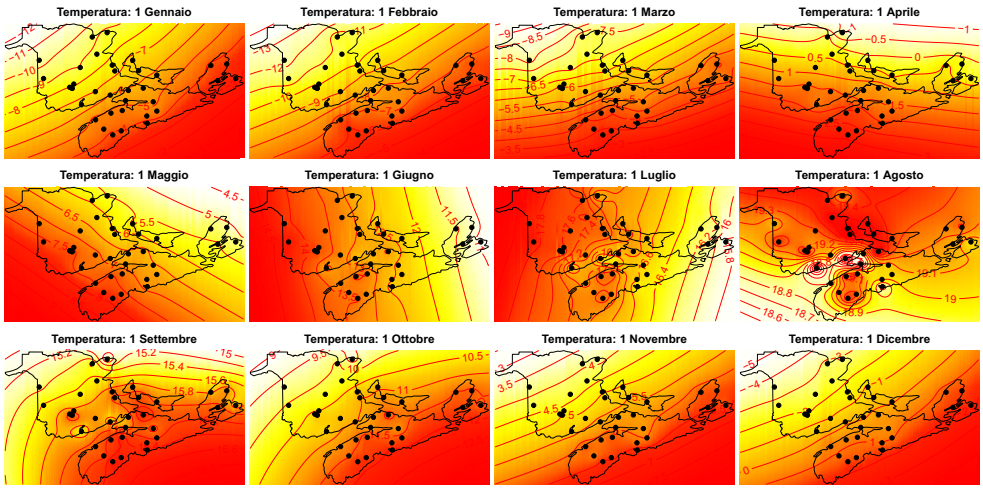

Figura 6.26: Contour-plot dell'interpolazione di Universal Kriging dopo 5 iterazioni dell'Algoritmo 5.1 per il primo giorno di ogni mese. La scala colori indica con il rosso temperature più alte, con il bianco temperature più basse.

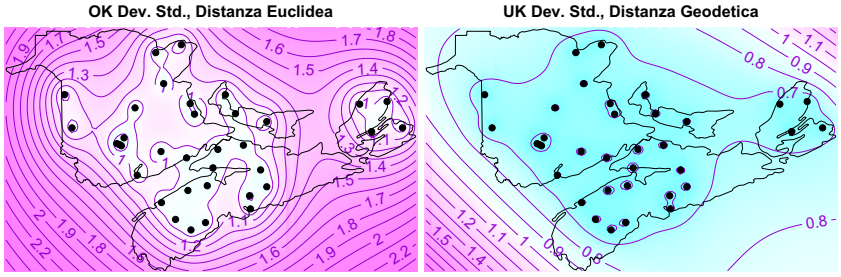

Figura 6.27: Confronto delle deviazioni standard di Ordinary Kriging (a sinistra) e Universal Kriging associato alla stima del *drift* con 5 iterazioni dell'Algoritmo 5.1 (a destra). La scala colori indica con il viola valori alti, con il celeste valori bassi, ed è la medesima per i due pannelli.

6.5 Conclusioni e Sviluppi Futuri

In questo capitolo, la metodologia proposta nel lavoro di tesi è stata applicata a un dataset funzionale reale, corrispondente alle temperature medie giornaliere registrate in 35 siti nella zona marittima del Canada.

Dopo aver individuato lo spazio funzionale opportuno per lo svolgimento dell'analisi e la metrica non euclidea per il calcolo delle distanze internamente al dominio spaziale (Sezione 6.2), l'analisi geostatistica si è svolta in quattro momenti (Sezione 6.3): l'analisi variografica iniziale con la verifica dell'ipotesi di stazionarietà, l'individuazione della struttura del termine di *drift* e la sua stima, l'interpolazione del campo di temperatura e la validazione del modello individuato.

Dal primo momento dell'analisi è infatti emersa la non stazionarietà del processo (Sottosezione 6.3.1), che ha reso necessaria l'introduzione nel modello di un termine di variabilità spaziale deterministica, la cui forma è stata selezionata attraverso l'Algoritmo 5.2 (Sottosezione 6.3.2). La stima del termine di *drift* è stata quindi determinata con la metodologia illustrata nel Capitolo 5, applicando cinque iterazioni dell'Algoritmo 5.1 ed il campo di temperature è stato interpolato attraverso il metodo di Universal Kriging.

Le stime individuate si sono dimostrate in linea sia con le mappe di riferimento del Servizio Meteorologico del Canada, sia con le descrizioni climatiche in (Stanley, 2002), evidenziando una variazione ciclica del gradiente di temperatura, in accordo con l'alternanza stagionale delle correnti d'aria provenienti dall'entroterra e dall'Oceano Atlantico (Sottosezioni 6.3.4 e 6.3.5).

Dal punto di vista previsivo, il modello stimato è risultato soddisfacente, presentando statistiche della distribuzione dello scarto quadratico relativo molto ridotte (Sezione 6.3.6).

Dal confronto dell'analisi illustrata in questo lavoro con le analisi presentate in letteratura (Sottosezione 6.4) è emerso che, dal punto di vista previsivo, la variazione delle prestazioni dei metodi di kriging derivanti dall'uso di una metrica non euclidea e dall'introduzione del

termine di *drift*, sono apprezzabili limitatamente alle statistiche mediana e massimo dello scarto quadratico di cross-validazione.

Si ritiene tuttavia che l'adozione di una metrica non euclidea, formalmente preferita alla metrica euclidea per le analisi svolte nel lavoro di tesi, abbia aperto possibilità più ampie rispetto alla metrica euclidea, denunciando la non stazionarietà del processo e suggerendo a ulteriori prospettive di sviluppo.

Infatti, il riconoscimento dell'esistenza di una variabilità deterministica nel processo, la cui modellazione è avvenuta in termini di coordinate geografiche, ha consentito di mitigare l'effetto regolarizzante della stima di kriging, diminuendo sensibilmente la varianza di kriging associata alla stima.

Infine, l'introduzione di una forma funzionale più complessa per il *drift*, eventualmente incorporando regressori funzionali, potrebbe condurre a miglioramenti ancora più significativi in termini previsivi, traducendosi in una modellazione più accurata la variabilità spaziale deterministica e stocastica del fenomeno.

Conclusione

Questo lavoro è stato dedicato allo sviluppo di metodologie geostatistiche volte all'analisi di dati funzionali spazialmente distribuiti, con l'obiettivo di estendere, attraverso l'approccio dell'Analisi di Dati Funzionali, alcune tecniche geostatistiche in uso nel caso di dati vettoriali.

Il problema è stato trattato in primo luogo da un punto di vista teorico, dapprima costruendo un impianto formale di definizioni e ipotesi, quindi stabilendo nuove formulazioni dei metodi di Ordinary e Universal Kriging a coefficienti costanti, valide per elementi di un qualsiasi spazio di Hilbert. Lo sforzo teorico è stato rivolto alla formulazione delle ipotesi essenziali per lo sviluppo dei metodi proposti e all'individuazione dei legami delle nuove definizioni globali di covariogramma e variogramma con la teoria esistente, determinando in particolare, sulla base delle opportune assunzioni di stazionarietà introdotte, le espressioni esplicite per i previsori di Ordinary e Universal Kriging e per gli stimatori del *drift*.

Lo studio teorico di questo lavoro si è concentrato sul caso di previsori lineari ottimi (BLUP) a coefficienti costanti, affiancati, nel caso non stazionario, a stimatori lineari ottimi (BLUE) per il *drift*, determinati a partire da un modello lineare per il termine di variabilità deterministica, i cui regressori sono ipotizzati solo spazialmente variabili. Lo sviluppo naturale di questo lavoro di tesi è la considerazione di stimatori più generali di quelli qui introdotti, ammettendo coefficienti funzionali per gli stimatori di kriging, come proposto nell'ambito stazionario in (Giraldo *e altri*, 2010a), e regressori funzionali per la modellazione del *drift*. Questa prospettiva avrebbe notevoli implicazioni anche dal punto di vista applicativo, consentendo una modellazione ancora più accurata dei fenomeni oggetti di studio.

La seconda parte di questo lavoro è stata invece dedicata allo sviluppo di nuovi algoritmi che, coerentemente con i risultati teorici stabiliti, permettessero il trattamento di dati funzionali spazialmente distribuiti. In particolare, si è ritenuto interessante focalizzare l'attenzione in due direzioni: la stima della funzione variogramma e la previsione per campi funzionali spazialmente non stazionari.

La novità più rilevante nella direzione della stima variografica è la proposta dell'uso di metodi bootstrap semiparametrici per dati funzionali spazialmente distribuiti, applicata in questo lavoro all'approssimazione della distribuzione dello stimatore empirico (Capitolo 4). In particolare, la metodologia di ricampionamento introdotta ha permesso di quantificare l'incertezza legata alla stima della struttura di dipendenza del campo aleatorio in analisi, consentendo da un lato l'adattamento della stima empirica a un modello valido attraverso

un'opportuna procedura di minimizzazione, dall'altro la costruzione di intervalli di confidenza per lo stimatore sperimentale, la cui probabilità di copertura, stimata per simulazione, si è dimostrata in linea con il livello di confidenza nominale.

Più in generale, la disponibilità di un metodo MC-bootstrap per dati funzionali spazialmente distribuiti apre ampie prospettive di ricerca, proponendosi come uno strumento molto versatile, il cui rilievo statistico è amplificato dalla possibilità di ottenere un'approssimazione della distribuzione di una qualunque statistica dai dati, evitando la difficoltà di formulare ipotesi distribuzionali.

In questo senso, la metodologia bootstrap semiparametrica proposta attraverso l'Algoritmo 4.6, costituisce un'ulteriore prospettiva di sviluppo anche per la procedura di analisi e previsione per dati funzionali georeferenziati spazialmente non stazionari, introdotta nel Capitolo 5: ad esempio, un'approssimazione della distribuzione dello stimatore del *trace-variogram* potrebbe essere sfruttata per la costruzione di intervalli di confidenza per la previsione di kriging, incorporando opportunamente l'incertezza e la distorsione legata alla stima della struttura di dipendenza del campo.

L'impianto algoritmico studiato nel Capitolo 5, che è stato ottenuto dall'integrazione del metodo MC-bootstrap per stima variografica, proposto nel Capitolo 4, con il metodo di stima ai minimi quadrati generalizzati del termine di *drift*, la cui base teorica è stata stabilita nel Capitolo 3, è risultato essere di particolare rilievo applicativo, essendo sufficientemente generale per descrivere in modo esaustivo la variabilità di un processo funzionale non stazionario e per fornirne previsioni.

In particolare, una delle potenzialità evidenziate dalla metodologia sviluppata nel lavoro di tesi è il fatto di fornire una procedura di analisi geostatistica che può essere usata per un qualsiasi processo funzionale la cui variabilità spaziale possa essere spiegata attraverso un termine deterministico di *drift* e un residuo debolmente stazionario: ad esempio, il caso stazionario rientra in questa procedura pur di considerare una media spazialmente costante, che si traduce nella riduzione della matrice disegno $\mathbb{F}_s$ alla sola colonna $\mathbf{1} = (1, ..., 1)^T \in \mathbb{R}^n$; allo stesso modo, il caso di dataset spazialmente non correlato discende dal caso generale qui studiato, particolarizzando il modello variografico a un modello puro *nugget*.

Inoltre, sebbene il campo di ricerca in cui si inserisce questo lavoro sia ancora molto aperto e i risultati qui presentati offrano molteplici spunti di riflessione teorica e di sviluppo applicativo, il comportamento delle metodologie proposte, studiato attraverso l'analisi di dati sintetici e reali, si è dimostrato già molto soddisfacente nella considerazione di coefficienti costanti per i previsori di kriging e di regressori variabili solo spazialmente per il termine di *drift*.

Infatti, la possibilità di introdurre un termine di *drift* nel modello per il fenomeno, che contraddistingue questa nuova metodologia rispetto ai metodi di kriging stazionario per dati funzionali già presenti in letteratura, è sicuramente un vantaggio dal punto di vista applicativo, fornendo uno strumento molto più flessibile, adatto ad una più accurata modellazione del fenomeno. I risultati ottenuti nel Capitolo 6, dedicato al caso studio relativo al dataset

di temperature nelle province Marittime del Canada, suffragano la precedente affermazione: dalle analisi è infatti emerso che la modellazione della variabilità deterministica del fenomeno permette di ottenere dei risultati di interessante interpretazione climatica e coerenti con le mappe di riferimento del Servizio Meteorologico del Canada, consentendo al contempo di abbattere la varianza della previsione di kriging.

Un ulteriore miglioramento nell'accuratezza della stima, che costituisce un possibile sviluppo dell'analisi del dataset di temperature, potrebbe essere ottenuto dall'ambientazione delle analisi stesse in uno spazio funzionale differente. Infatti, sebbene le analisi riportate in questo lavoro siano state condotte nello spazio L^2, i risultati teorici stabiliti nella prima parte del lavoro si mantengono validi in un qualsiasi spazio di Hilbert: in particolare, ambientando l'analisi in un opportuno spazio di Sobolev, si consentirebbe l'inclusione, nella modellazione della struttura di dipendenza spaziale, delle caratteristiche differenziali delle curve osservate, sfruttando anche in questo modo la natura funzionale dei dati.

In conclusione, le metodologie sviluppate in questo lavoro, sostenute dal contesto teorico qui formalizzato, sono risultate adeguate per una prima applicazione in ambito industriale: nel periodo di stage svolto presso Eni S.p.A., gli algoritmi proposti sono stati applicati a un problema di natura geofisica ottenendo risultati promettenti, non illustrati in questa trattazione per motivi di confidenzialità.

Appendice A

Notazioni

La presente Appendice è pensata per fornire un punto di riferimento per la notazione adottata nel lavoro di tesi e in particolare supportare la lettura del Capitolo 3. Di seguito sono quindi riportate, in modo schematico, le convenzioni di scrittura adottate, con particolare riferimento al significato dei principali simboli e degli stili introdotti.

A.1 Convenzioni per Processi Monovariati e Vettori di $\mathbb{R}^n$

- $D \subseteq \mathbb{R}^d$: dominio spaziale;
- $\boldsymbol{s}$, $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$; qualora non crei ambiguità, s, $s_1, ..., s_n$: siti nel dominio spaziale;
- $\{Z(s), s \in D\}$ o Z(s): processo stocastico monovariato;
  $Z^*(s_0)$: previsore di kriging (con $\boldsymbol{\lambda}^*$ vettore di pesi ottimi);
- $m(s)$: *drift* del processo $Z(s)$;
- $f_l(s)$: l-esimo regressore nel modello lineare per il *drift* valutato in s, $l = 1, ..., L$;
- a_l: coefficiente del modello lineare per il *drift* associato al regressore f_l, $l = 1, ..., L$.
- $\delta(s)$: residuo del processo $Z(s)$;
- $C(\boldsymbol{h})$: covariogramma stazionario, $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$, $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$;
  $C(h)$: covariogramma stazionario e isotropo, $h = \|\boldsymbol{h}\| = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$, $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$;
- $\gamma(\boldsymbol{h})$: semivariogramma stazionario, $\boldsymbol{h} = \boldsymbol{s}_i - \boldsymbol{s}_j$, $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$;
  $\gamma(h)$: semivariogramma stazionario e isotropo, $h = \|\boldsymbol{h}\| = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$, $\boldsymbol{s}_i, \boldsymbol{s}_j \in D$;
  $\gamma(h; \boldsymbol{\vartheta})$: modello valido di semivariogramma stazionario e isotropo.

Inoltre, in generale, salvo diversa indicazione:

- Il grassetto indica vettori di $\mathbb{R}^n$, il corsivo semplice numeri reali (e.g. $\boldsymbol{h} \in \mathbb{R}^d$, $h \in \mathbb{R}$);
- Il maiuscolo indica variabili aleatorie (e.g. Z);
- Il minuscolo indica realizzazioni (e.g. z);
- La localizzazione della variabile è indicata tra parentesi (e.g. $Z(s)$);
- L'asterisco ad apice $\cdot^*$ indica previsori di kriging (o pesi di kriging) (e.g. $Z^*(s_0)$);
- La stella ad apice $\cdot^\star$ indica campioni bootstrap o statistiche bootstrap (e.g. $Z^\star$, $T^\star$);
- In $\|s_i - s_j\|$, $\|\cdot\|$ indica la norma euclidea.

A.2 Convenzioni per Processi Funzionali e Vettori di H^n

Relativamente alla teoria dei processi stocastici su spazi di Hilbert, sono adottati i seguenti simboli:

- $\{\boldsymbol{\chi}_s, s \in D\}$ o $\boldsymbol{\chi}_s$: processo stocastico funzionale;
  χ_s: realizzazione del processo $\boldsymbol{\chi}_s$ in $s \in D$;
  $\boldsymbol{\chi_s} = (\boldsymbol{\chi}_{s_1}, ..., \boldsymbol{\chi}_{s_n})$: vettore di elementi aleatori del processo $\boldsymbol{\chi}_s$, localizzati in $s_1, ..., s_n$;
  $\chi_{\boldsymbol{s}} = (\chi_{s_1}, ..., \chi_{s_n})$: vettore di osservazioni;
  $\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}} = (\widetilde{\chi}_{s_1}, ..., \widetilde{\chi}_{s_n})$: campione scorrelato (da intendere: le cui componenti sono scorrelate);
  $\widetilde{\chi}_{\boldsymbol{s}}$: realizzazione di $\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}}$;
  $\boldsymbol{\chi}^*_{s_0}$: previsore di kriging in s_0;
  $\chi^*_{s_0}$: previsione di kriging in s_0;

- m_s: *drift* del processo $\boldsymbol{\chi}_s$ (deterministico);
  $m_{\boldsymbol{s}} = (m_{s_1}, ..., m_{s_n})$: *drift* in posizione $s_1, ..., s_n$;
  $\widehat{\boldsymbol{m}}_{\boldsymbol{s}} = (\widehat{\boldsymbol{m}}_{s_1}, ..., \widehat{\boldsymbol{m}}_{s_n})$: stimatore del *drift* $m_{\boldsymbol{s}}$ in $s_1, ..., s_n$;
  $\widehat{m}_{\boldsymbol{s}} = (\widehat{m}_{s_1}, ..., \widehat{m}_{s_n})$: stima del *drift* in $s_1, ..., s_n$;
  $\widetilde{m}_s$, $\widetilde{m}_{\boldsymbol{s}}$, $\widehat{\widetilde{\boldsymbol{m}}}_{\boldsymbol{s}}$, $\widehat{\widetilde{m}}_{\boldsymbol{s}}$: analoghe alle quantità precedenti, ma riferite al campione scorrelato $\widetilde{\chi}_{\boldsymbol{s}}$;

- $f_{\boldsymbol{l}}(s)$: regressore nel modello lineare per il *drift* valutato in $s \in D$, $l = 1, ..., L$;
  $f_{\boldsymbol{l}}(s) = (f_1(s), ..., f_L(s))$: vettore di regressori in $s \in D$;
  $\mathbb{F}_{\boldsymbol{s}}$: matrice disegno per il modello lineare del *drift*;

- a_l: coefficiente del modello lineare per il *drift* associato al regressore f_l, $l = 1, ..., L$;
  $a_{\boldsymbol{l}} = (a_1, ..., a_L)$: vettore di coefficienti (funzionali) del modello lineare per il *drift*;
  $\widehat{\boldsymbol{a}}_{\boldsymbol{l}} = (\widehat{\boldsymbol{a}}_1, ..., \widehat{\boldsymbol{a}}_L)$: generico stimatore di $a_{\boldsymbol{l}}$ (vettore di elementi aleatori di H);
  $\widehat{\boldsymbol{a}}^{GLS}_{\boldsymbol{l}} = (\widehat{\boldsymbol{a}}^{GLS}_1, ..., \widehat{\boldsymbol{a}}^{GLS}_L)$: stimatore GLS di $a_{\boldsymbol{l}}$ (vettore di elementi aleatori di H);
  $\widehat{\boldsymbol{a}}^{BLUE}_{\boldsymbol{l}} = (\widehat{\boldsymbol{a}}^{BLUE}_1, ..., \widehat{\boldsymbol{a}}^{BLUE}_L)$: stimatore BLUE di $a_{\boldsymbol{l}}$ (vettore di elementi aleatori di H);

- $\{\boldsymbol{\delta}_s, s \in D\}$ o $\boldsymbol{\delta}_s$: (campo aleatorio del) residuo del processo $\boldsymbol{\chi}_s$;
  $\boldsymbol{\delta_s} = (\boldsymbol{\delta}_{s_1}, ..., \boldsymbol{\delta}_{s_n})$: vettore di residui (aleatori) in posizione $s_1, ..., s_n$;
  $\delta_{\boldsymbol{s}} = (\delta_{s_1}, ..., \delta_{s_n})$ realizzazione del vettore $\boldsymbol{\delta_s}$;
  $\boldsymbol{\delta}^*_{s_0}$: previsore di kriging in s_0;
  $\delta^*_{s_0}$: previsione di kriging in s_0;
  $\widehat{\boldsymbol{\delta}}_{\boldsymbol{s}} = (\widehat{\boldsymbol{\delta}}_{s_1}, ..., \widehat{\boldsymbol{\delta}}_{s_n})$ previsore del residuo $\boldsymbol{\delta_s}$ in $s_1, ..., s_n$;
  $\widehat{\delta}_{\boldsymbol{s}} = (\widehat{\delta}_{s_1}, ..., \widehat{\delta}_{s_n})$ previsione del residuo in $s_1, ..., s_n$;

- $\boldsymbol{\varepsilon}_{s_0}$: errore di previsione in s_0 commesso con il previsore di kriging $\boldsymbol{\chi}^*_{s_0}$ (elemento aleatorio);
  ε_{s_0}: realizzazione di $\boldsymbol{\varepsilon}_{s_0}$;

- $C(s_i, s_j)$: covariogramma;
  $C(\boldsymbol{h})$: covariogramma stazionario;
  $C(h)$: covariogramma stazionario e isotropo;

- $C^\alpha(s_i, s_j; t)$: covariogramma puntuale di $D^\alpha \boldsymbol{\chi}_s(t)$;
  $C^\alpha(h; t)$: covariogramma puntuale stazionario e isotropo di $D^\alpha \boldsymbol{\chi}_s(t)$;

- C_{s_i, s_j}: operatore di cross-covarianza tra $\boldsymbol{\chi}_{s_i}$ e $\boldsymbol{\chi}_{s_j}$;
  C_h: operatore di cross-covarianza, in ipotesi di stazionarietà e isotropia, tra $\boldsymbol{\chi}_{s_i}$ e $\boldsymbol{\chi}_{s_j}$, $h = \|s_i - s_j\|$;

- $\gamma(s_i, s_j)$: semivariogramma;
  $\gamma(\boldsymbol{h})$: semivariogramma stazionario;
  $\gamma(h)$: semivariogramma stazionario e isotropo;
  $\gamma(h; \boldsymbol{\vartheta})$: modello valido di semivariogramma stazionario e isotropo;
  $\widehat{\gamma}(h)$: stimatore del variogramma;
  $\widehat{\gamma}(\boldsymbol{h}) = \widehat{\gamma}(h_1, ..., h_K)$: stimatore *binned variogram*;

- $\gamma^\alpha(s_i, s_j; t)$: semivariogramma puntuale di $D^\alpha \boldsymbol{\chi}_s(t)$;
  $\gamma^\alpha(h; t)$: semivariogramma puntuale stazionario e isotropo di $D^\alpha \boldsymbol{\chi}_s(t)$;

- Σ: matrice di covarianza di $\boldsymbol{\chi_s}$;

- Ψ: matrice di covarianza di $\widehat{\gamma}(\boldsymbol{h})$;

- Λ: matrice di covarianza di $\widehat{\boldsymbol{a}}^{GLS}$;

- Λ^{OLS}: matrice di covarianza di $\widehat{\boldsymbol{a}}^{OLS}$;

- $\mathbb{B}$: distorsione dello stimatore $\widehat{\Sigma} = \widehat{\Sigma}^{GLS}$ di Σ.

Inoltre, in generale, salvo diversa indicazione:

- Il grassetto indica un elemento aleatorio o un processo aleatorio (e.g. $\boldsymbol{\chi}_s$);

- Il pedice in grassetto indica un vettore (e.g. $\chi_{\boldsymbol{s}}$);

- Le lettere calligrafiche indicano insiemi (e.g. $\mathcal{T}$ dominio di χ_s);
- Le lettere in stile 'blackboard' indicano matrici (e.g. $\mathbb{B}$);
- Le lettere greche maiuscole indicano matrici di covarianza (e.g. Σ);
- La localizzazione della variabile è indicata come pedice, l'ascissa della variabile tra parentesi (e.g. $\chi_s(t)$);
- Il simbolo $\widehat{\cdot}$ indica che la quantità in esame è uno stimatore o una stima (e.g. $\widehat{\boldsymbol{a}}$, $\widehat{a}$);
- Il simbolo $\widetilde{\cdot}$ indica che la quantità in oggetto è un campione scorrelato o è riferito a un campione scorrelato (e.g. $\widetilde{\boldsymbol{\chi}}_{\boldsymbol{s}}$ o $\widetilde{m}_s$);
- L'asterisco ad apice $\cdot^*$ indica previsori di kriging (o pesi di kriging) (e.g. $\chi^*_{s_0}$, $\boldsymbol{\lambda}^*$);
- La stella ad apice $\cdot^\star$ indica campioni bootstrap o statistiche bootstrap (e.g. $\boldsymbol{\chi}^\star$, $T^\star$);
- In $\|s_i - s_j\|$, $\|\cdot\|$ indica la norma euclidea;
- In $\|\chi_{s_i} - \chi_{s_j}\|$, $\|\cdot\|$ indica la norma su H.

Appendice B

Codici

L'applicazione delle metodologie proposte in questo lavoro di tesi è stata basata sull'implementazione degli algoritmi sviluppati nell'ambiente `R` (R Development Core Team, 2009).

`R` è un insieme integrato di software per la manipolazione di dati e per la gestione e la produzione di grafici. `R` è un sistema coerente e completo, all'interno del quale sono implementate tutte le principali tecniche statistiche attualmente proposte in letteratura: tra i software concepiti per l'analisi statistica, è quello maggiormente usato dai ricercatori in questo campo.

Inoltre, `R` è un software open source, scaricabile da Internet sul sito *The R Project for Statistical Computing* all'indirizzo `http://www.r-project.org`, sul quale possono anche essere reperiti i pacchetti aggiuntivi che consentono di espanderne le funzionalità.

In questa appendice sono riportati i codici, in linguaggio `R`, delle principali funzioni usate per le analisi illustrate nei Capitoli 4, 5 e 6. In particolare, nella prima parte dell'appendice sono riportate le implementazioni usate per la stima del variogramma e per l'applicazione degli Algoritmi 4.6, 4.7 sviluppati in merito al metodo bootstrap; nella seconda parte sono invece presentati i codici riguardanti gli Algoritmi 5.1 e 5.2; infine, nella Sezione B.3, è illustrato il codice usato per l'analisi del dataset di temperature, presentata nel Capitolo 6.

I codici riportati non sono ottimizzati e non sono pensati per un uso industriale; al contrario, essi sono intesi come indicativi della procedura numerica svolta nelle analisi condotte. Migliori prestazioni dal punto di vista dell'efficienza algoritmica e dei tempi di calcolo sarebbero ottenute dall'implementazione delle procedure proposte in linguaggio C o C++.

B.1 Stima del Variogramma

In questa sezione sono presentate le funzioni principali usate nell'ambito della stima variografica; la scelta delle funzioni riportate risponde al criterio di originalità dei codici: non sono qui illustrate le funzioni parzialmente o totalmente già implementate in `R`. Le funzioni sfruttate per le analisi fanno uso del pacchetto aggiuntivo `geoR` (Jr e Diggle, 2001) e sono le seguenti:

- `trapzc` (non riportata) calcola l'integrale numerico di una funzione data con la formula dei trapezi;

- `dist.geod` (non riportata) calcola la distanza geodetica tra due punti la cui localizzazione sia espressa in coordinate geografiche espresse in gradi;

- `Tracevariogram.emp.fn`: calcola la stima *binned trace-variogram*, prevedendo la possibilità di scelta della distanza da adottare per il dominio spaziale (euclidea o geodetica) e consentendo il calcolo variogrammi direzionali. Si basa sulle funzioni `dist.geod` e `trapzc`.

- `variofit.fn` (non riportata): modificazione della funzione `variofit` del pacchetto `geoR`, che consente l'adattamento di un modello valido di variogramma con il metodo GLS. Riceve in input il risultato della funzione `Tracevariogram.emp.fn` e la matrice di pesi per la minimizzazione, indicata nel testo con Ψ^{-1}.

- `discr.pesi.uniformi` (non riportata): estrae un campione i.i.d. di dimensione n nell'insieme $\{1, ..., n\}$.

- `bootstrap.fn`: svolge la procedura MC-bootstrap semiparametrica, secondo il procedimento dettagliato nell'Algoritmo 4.6, fissato il numero di iterazioni bootstrap B. Fa uso delle funzioni `Tracevariogram.emp.fn`, `variofit.fn` e `discr.pesi.uniformi`.

- `bootstrap.fn.Nmax`: consente l'applicazione dell'Algoritmo 4.7, a partire da valori fissati del numero di iterazioni bootstrap B e di iterazioni esterne N_{max}. Si basa sulle funzioni `Tracevariogram.emp.fn`, `variofit.fn`, `discr.pesi.uniformi`,`bootstrap.fn`.

Tracevariogram.emp.fn

INPUT:

```
X, Y = coordinate (se geografiche: X=lat, Y=long)
phiS = matrice di dimensione (n,N_samples),
       contenente i valori discretizzati delle funzioni
N_samples = numero di campioni
z_step = passo di discretizzazione ascissa
N_lag = numero di lag
lmax = distanza massima di osservazione
class_vec = classi per il calcolo della stima binned variogram
punto_lag = distanza h_k
directions = 'omnidirectional' o direzioni di calcolo del variogramma
tol = tolleranza sulla direzione di calcolo
      (usato solo se directions!='omnidirectional')
h = matrice di distanze tra i siti di osservazione
dist = 'euclidean' o 'geodetic': distanza sul dominio spaziale
```

OUTPUT:

```
trace_variogram = oggetto della classe 'variog' contenente la stima
         empirica e le informazioni di input alla funzione variofit.fn.
```

```
Tracevariogram.emp.fn=function(X, Y, phiS, N_samples, z_step, lag=lmax/N_lag,
                N_lag=ceiling(lmax/lag), lmax, class_vec=NULL, punto_lag=NULL,
                directions='omnidirectional',tol=pi/8, h=NULL, ang=NULL,
                dist='euclidean')
{   ## Corpo della funzione disponibile su richiesta all'autrice ##
    ##          alessandra.menafoglio@mail.polimi.it             ##
}
```

bootstrap.fn

Input:

```
X, Y = coordinate
h = matrice di distanze tra i siti di osservazione
phiS = matrice di dimensione (n,N_samples),
       contenente i valori discretizzati delle funzioni
N_samples = numero di campioni
N_iter = numero di iterazioni bootstrap
N_lag = numero di lag
lmax = distanza massima di osservazione
model = modello parametrico di variogramma
param = parametri del modello parametrico preliminarmente stimato
z_step = passo di discretizzazione ascissa
n = numero di ascisse sulle quali è valutata la funzione
class_vec = classi per il calcolo della stima binned variogram
punto_lag = distanza h_k
```

Output:

```
V = matrice di dimensione (N_iter,N_lag) contenente le stime bootstrap del
    binned variogram

bootstrap.fn=function(X, Y, h=NULL, phiS, N_samples, N_iter, N_lag, model,
                param=c(sill,range,nugget=0), n=max(z)/z_step, z_step, lmax,
                class_vec=NULL, punto_lag=NULL)
{   ## Corpo della funzione disponibile su richiesta all'autrice ##
    ##          alessandra.menafoglio@mail.polimi.it             ##
}
```

bootstrap.fn.Nmax

Input:

```
X, Y = coordinate
h = matrice di distanze tra i siti di osservazione
phiS = matrice di dimensione (n,N_samples),
       contenente i valori discretizzati delle funzioni
N_samples = numero di campioni
N_iter = numero di iterazioni bootstrap (N_iter=5000 di default)
Nmax = numero di iterazioni esterne (Nmax=2 di default)
N_lag = numero di lag
lmax = distanza massima di osservazione
model = modello parametrico di variogramma
param = parametri del modello parametrico preliminarmente stimato (OLS)
fix.nugget = booleano, se fix.nugget=TRUE il nugget è imposto nella stima
```

```
z_step = passo di discretizzazione ascissa
n = numero di ascisse sulla quale è valutata la funzione
class_vec = classi per il calcolo della stima binned variogram
punto_lag = distanza h_k
```

Output:

```
list.return = lista contente:
       - V = matrice di dimensione (N_iter,N_lag) contenente le stime bootstrap
             del binned variogram all'ultima iterazione
       - T.fit.b = modello di variogramma adattato all'ultima iterazione esterna
                   (output della funzione variofit.fn)
       - gamma.trace.IC = matrice di dimensioni (2,N_lag) le cui righe corrispon-
                          dono ai percentili 0.5% e 0.95%
       - param.GLS = parametri del modello di variogramma, stimati all'ultima
                     iterazione esterna
       - param.OLS = parametri iniziali del modello di variogramma


bootstrap.fn.Nmax=function(X, Y, phiS, h, N_samples, N_lag,  lmax, n, z_step, model,
                           param=c(sill_ols,range_ols,nugget_ols), fix.nugget=FALSE,
                           Nmax=2 , N_iter=5000)
{   ## Corpo della funzione disponibile su richiesta all'autrice ##
    ##          alessandra.menafoglio@mail.polimi.it             ##
}
```

Le funzioni riportate saranno inserite in un codice finale, attraverso il quale comprendere meglio quale sia stata la successione logica delle funzioni implementate durante le analisi illustrate nel testo.

B.2 Metodi di Previsione per Dati Funzionali Georeferenziati

Nella presente sezione sono riportati i codici, opportunamente commentati, delle funzioni necessarie per l'applicazione delle metodologie proposte nel Capitolo 5. Tali codici fanno uso dei pacchetti aggiuntivi `geoR` (Jr e Diggle, 2001) e `fda` (Ramsay *e altri*, 2010).

Le implementazioni riguardanti l'ordinamento delle forme funzionali per il *drift* e la costruzione dei relativi stimatori ai minimi quadrati, riportate di seguito e nella Sezione B.3, si basano sull'assunzione che i dati siano proiettati su una base funzionale, facendo uso dei codici ottimizzati interni al pacchetto `fda`. Tuttavia, qualora fosse disponibile la forma funzionale esplicita per i dati funzionali e non si volesse imporre una base costante per i coefficienti, si è lavorato direttamente con le espressioni esplicite ricavate per gli stimatori nella Sezione 3.5.

Di seguito sono sintetizzati i nomi delle funzioni usate con le rispettive funzionalità:

- `rank.drift`: svolge la procedura di selezione del modello di *drift* tra quelli riportati in Tabella 5.1. Fa uso della funzione `Tracevariogram.emp.fn`, riportata in precedenza, e della funzione `fRegress` interna al pacchetto `fda`;

- `rank.drift.constantbasis` (non riportata): svolge la medesima procedura della funzione `rank.drift`, ma imponendo una base costante per i coefficienti del *drift* a_l;

- `ord.kriging`: determina la previsione di Ordinary Kriging su una griglia fornita in input, prevedendo la possibilità di usare la distanza euclidea o geodetica per il dominio spaziale e di fissare un *neighborhood* (non sfruttata nelle analisi illustrate nel testo);

- `univ.kriging`: determina la previsione di Universal Kriging su una griglia fornita in input, con regressori f_l fissati, prevedendo la possibilità di usare la distanza euclidea o geodetica per il dominio spaziale e di fissare un *neighborhood* (non sfruttata nelle analisi illustrate nel testo).

rank.drift

INPUT:

```
X, Y = coordinate (se geografiche: X=lat, Y=long)
phiS.fd = oggetto della classe fd, corrispondente alla proiezione delle
          funzioni su una base funzionale
lag = ampiezza del lag
lmax = distanza massima di osservazione
which.drift = sottoinsieme di {1,2,...,8} contenente i modelli di drift da ordinare
```

OUTPUT:

```
grafico dei variogrammi dei residui, per ogni modello di drift;
stampa a video dell'ordinamento individuato dall'algoritmo;
file rank.txt contenente l'ordinamento ottenuto e gli scarti stimati SSE(k);
list.return = lista contenente:
        - drift = lista i cui elementi corrispondono al drift stimato alla prima
                  iterazione dell'algoritmo di stima GLS, per ogni modello di drift
                  in which.drift
        - deltaS = lista i cui elementi corrispondono al residuo stimato alla prima
                  iterazione dell'Algoritmo di stima GLS, per ogni modello di drift
                  in which.drift
        - trace_variogram = lista contente i variogrammi sperimentali dei residui
                  contenuti in deltaS
        - fit.a.GLS = lista contenente l'output della funzione fRegress, per ogni
                  modello di drift in which.drift
        - diff.drift = vettore di lunghezza length(which.drift) contenente i valori
                  stimati per SSE(k)
        - rank = ordinamento dei drift con criterio previsivo
```

```
rank.drift.constbasis=function(phiS.fd,X,Y,lag,lmax,which.drift=seq(1,8))
{   ## Corpo della funzione disponibile su richiesta all'autrice ##
    ##          alessandra.menafoglio@mail.polimi.it             ##
}
```

ord.kriging

INPUT:

```
X, Y = coordinate (se geografiche: X=lat, Y=long)
phiS = matrice di dimensione (n,N_samples),
       contenente i valori discretizzati delle funzioni
xlim, ylim = limiti della griglia spaziale sulla quale prevedere il campo
tol.x, tol.y = passo della griglia spaziale
model = modello parametrico di variogramma
model.param = parametri del modello parametrico
dist = 'euclidean' o 'geodetic', distanza da usare sul dominio spaziale
nb = 'unique' se non si vuole usare un neighborhood, o valore numerico
       indicante l'ampiezza del neighborhood
```

OUTPUT:

```
list.return = lista contenente
        - gridOutput = griglia spaziale sulla quale sono predetti i punti
        - phiS.grid = matrice le cui colonne contengono i valori discretizzati
                delle funzioni predette
        - lambda.opt = pesi ottimi stimati come soluzione del sistema di kriging
        - s2.OK = varianza di kriging delle funzioni predette

ord.kriging=function(X,Y,phiS,xlim,ylim,tol.x=(xlim[2]-xlim[1])/100,
            tol.y=(ylim[2]-ylim[1])/100,model,model.param=c(sill,range,nugget),
            dist='euclidean',nb='unique')
{   ## Corpo della funzione disponibile su richiesta all'autrice ##
    ##          alessandra.menafoglio@mail.polimi.it             ##
}
```

univ.kriging

INPUT:

```
X, Y = coordinate (se geografiche: X=lat, Y=long)
phiS = matrice di dimensione (n,N_samples),
       contenente i valori discretizzati delle funzioni
Fs = matrice disegno
Fs = matrice con la valutazione dei regressori sulla griglia da prevedere
xlim, ylim = limiti della griglia spaziale sulla quale prevedere il campo
tol.x, tol.y = passo della griglia spaziale
model = modello parametrico di variogramma
model.param = parametri del modello parametrico
dist = 'euclidean' o 'geodetic', distanza da usare sul dominio spaziale
nb = 'unique' se non si vuole usare un neighborhood, o valore numerico
     indicante l'ampiezza del neighborhood
```

OUTPUT:

```
list.return = lista contenente
        - gridOutput = griglia spaziale sulla quale sono predetti i punti
        - phiS.grid = matrice le cui colonne contengono i valori discretizzati
                delle funzioni predette
        - lambda.opt = pesi ottimi stimati come soluzione del sistema di kriging
        - s2.UK = varianza di kriging delle funzioni predette
```

```
univ.kriging=function(X,Y,phiS,Fs,Fs.grid,xlim,ylim,tol.x=(xlim[2]-xlim[1])/100,
            tol.y=(ylim[2]-ylim[1])/100,model,model.param=c(sill,range,nugget),
            dist='euclidean',nb='unique')
{   ## Corpo della funzione disponibile su richiesta all'autrice ##
    ##          alessandra.menafoglio@mail.polimi.it             ##
}
```

B.3 Analisi del Dataset *Canada's Maritime Provinces Temperatures*

In questa sezione è riportato il codice usato nell'analisi del dataset *Canada's Maritime Provinces Temperatures*, i cui risultati sono stati discussi nel Capitolo 6.

Il codice riportato si basa sulle funzioni descritte nelle Sezioni e e fa uso dei pacchetti `geoR` e `fda`.

main.Canada.Maritimes.R

```
################################################
#               CARICAMENTO INPUT              #
################################################

rm(list=ls(all=TRUE))

# Caricamento pacchetti aggiuntivi
library(geoR)
library(fda)

# Caricamento funzioni aggiuntive
source('FunzioniAusiliarie.R')

# Caricamento dati
coord=matrix(scan('coordata.txt',0, dec='.'), 35, 2, byrow=TRUE)
temp=matrix(scan('dailtemp.txt',0, dec=','), 365, 35,byrow=TRUE)
day=matrix(scan('day.txt',0), 365, 35, byrow=TRUE)

place=c('Fredericton','Halifax','Sydney','Miramichi','Kentville','Keminkujik',
'Nappan', 'Annapolis', 'Accadia', 'Woodstock', 'Bathurst', 'Bertrand',
'Buctouche', 'Sussex', 'Gagetown', 'Liverpool', 'Truro', 'Greenwood',
'Ingonish', 'Parrsboro', 'Pugwash', 'Shearwater', 'Charlottetown', 'Summerside',
'Alberton', 'Alma', 'Aroostook', 'Doakton', 'Oromocto', 'Rexton',
'Saint John', 'Baddeck', 'Cheticamp', 'Bridgewater', 'Middle Musquodoboit')
dimnames(temp) <- list(NULL,place)

# Inizializzazioni
Y=coord[,1]
X=coord[,2]
N_samples=length(X)
n=dim(temp)[1]

################################################
```

```
#          COSTRUZIONE DELLE FUNZIONI            #
#            SULLA BASE DI FOURIER               #
#################################################

# Parametri
nbasis=65
argvals=seq(1,n,by=1)
z=argvals
range=c(1,n)
period=n
# Costruzione della base
basis=create.fourier.basis(range, nbasis, period)
# Proiezione dati
datafd=data2fd(temp,argvals,basis)
coef=datafd$coef
z_step=1

# Valutazione su griglia temporale
temp=eval.fd(seq(1,365,by=z_step), datafd)

# Rappresentazione dei dati
plot(Y,X,xlab='Longitudine.W',ylab='Latitudine.N')
matplot(temp, type='l', lty=1, ylab='Temperatura (gradi C)',
        xlab='Giorno',col=rainbow(35), main='Canadian Weather Dataset')
abline(h=0,lty=2)
dev.off()

# Valutazione delle distanze massime per la stima variografica
dist.geod(c(44.0,-68),c(44.0,-60))/2 #max.dist longitudine
dist.geod(c(44.0,-68),c(48.0,-68))/2 #max.dist latitudine
dist.geod(c(44.0,-68),c(48.0,-60))/2 #max.dist 45°

#################################################
#               ANALISI INIZIALI:               #
#           STATISTICHE DAL CAMPIONE;           #
#           SCELTA DELLA METRICA SU D           #
#################################################

phiS.fd=datafd
phiS=temp

## Statistiche dal dataset funzionale ##
# Media
mean_phiS=(apply(phiS,1,'mean'))
# Varianza
var=rep(0,N_samples)
for(i in 1:N_samples)
    var[i]= trapzc( z_step,((phiS[,i]-mean_phiS)^2) )

## Scelta della metrica ##
# Calcolo della matrice di distanze tra i siti di osservazione:
# distanza geodetica
```

```
h=matrix(0,ncol=N_samples,nrow=N_samples)
for (i in 1:(N_samples-1))
{   for(j in (i+1):N_samples)
    {   h[i,j]=dist.geod(x=c(X[i],Y[i]),y=c(X[j],Y[j]))
        h[j,i]=h[i,j]
}    }


# distanza euclidea (in gradi)
h.eu=as.matrix(dist(cbind(X,Y),method='euclidian'))

# distanza euclidea (in km)
h.eu.km=h.eu/180*pi*6371

# Grafici di confronto delle matrici
par(mfrow=c(1,2),cex=1.25)
image(1:N_samples,1:N_samples,h, main='Geodetiche',xlab='campione',ylab='campione',
      breaks=seq(0,950,by=50),col=heat.colors(19))
image(1:N_samples,1:N_samples,h.eu.km, main='Euclidea',xlab='campione',ylab='campione',
      breaks=seq(0,950,by=50),col=heat.colors(19))
dev.off()


##################################################
#            ANALISI GEOSTATISTICA:              #
#           ANALISI DI STAZIONARIETA'            #
##################################################

# Parametri dello stimatore
lag=30
lmax=250
N_lag=ceiling(lmax/lag)
lmax=lag*N_lag

# Calcolo della stima binned variogram
trace_variogram=Tracevariogram.emp.fn(X=X, Y=Y, phiS=phiS, N_samples=N_samples,
                z_step=z_step, N_lag=N_lag, lmax=lmax, h=h, lag=lag)
gamma_trace=trace_variogram$v
punto_lag=trace_variogram$u
N_h=trace_variogram$n
N_class=N_lag

# Grafico del variogramma
par(cex=1.25)
plot(punto_lag,gamma_trace,pch=20,xlab='Distanza',ylab='Semivariogramma',
     ylim=c(0,max(gamma_trace)),main='Variogramma Dati Originali',
     xlim=c(0,max(punto_lag)))
lines(punto_lag,gamma_trace,pch=20,xlab='Distanza',ylab='Semivariogramma')
abline(h=mean(var),col='red',lty=1)
legend(145,125,c('Variog. emp.','Var. Camp.'),col=c('black','red'),lty=c(1,1))
dev.off()
```

```
# Poiché si evidenzia non stazionarietà si procede con la selezione del drift ottimo


#####################################################
#             ANALISI GEOSTATISTICA:                #
#              SELEZIONE DEL DRIFT                  #
#####################################################

# Inizializzazione parametri
const=rep(1,N_samples)

# Parametri per il variogramma
model='exponential'
fix.nugget=FALSE

rank=rank.drift(phiS.fd=phiS.fd,X=X,Y=Y,lag=lag,lmax=lmax,which.drift=seq(1,8))

# Grafici output
barplot(sort(rank$diff.drift,dec=FALSE), main='Rank Drift')
text(seq(1,9.5,length=8),rep(2,8),order(rank$diff.drift))
names.drift[order(rank$diff.drift)]

# Selezione del drift
which.drift=rank$rank[1]

#####################################################
#             ANALISI GEOSTATISTICA:                #
#           STIMA ITERATIVA DEL DRIFT               #
#####################################################

## Codice omesso, disponibile su richiesta all'autrice ##
##       alessandra.menafoglio@mail.polimi.it          ##

#####################################################
#             ANALISI GEOSTATISTICA:                #
#         CALCOLO DEL VARIOGRAMMA FINALE            #
#####################################################

Nmax=2   # numero iteraz. esterne bootstrap
N_iter=5000 # numero iteraz. MC

# Stima binned variogram
trace_variogram=Tracevariogram.emp.fn(X=X, Y=Y, phiS=deltaS, N_samples=N_samples,
                z_step=z_step, N_lag=N_lag, lmax=lmax, h=h)
gamma_trace=trace_variogram$v

# Stima modello valido con OLS
tracefit_ols=suppressWarnings(variofit(vario=trace_variogram, cov.model=model,
                wei='equal',messages=FALSE,fix.nugget=fix.nugget))
param=c(as.list(tracefit_ols)$cov.pars,as.list(tracefit_ols)$nugget)

# Stima modello valido con MC-bootstrap
```

```
fit.boot[[l]]=bootstrap.fn.Nmax(X=X, Y=Y, phiS=deltaS, h=h, N_samples=N_samples,
                N_lag=N_lag,  lmax=lmax, n=n, z_step=z_step, model=model,
                param=param,fix.nugget=fix.nugget, Nmax=Nmax, N_iter=N_iter)

sill.f=fit.boot[[l]]$param.GLS[1]
range.f=fit.boot[[l]]$param.GLS[2]
nugget.f=fit.boot[[l]]$param.GLS[3]

var=rep(0,N_samples)
for(i in 1:N_samples)
    var[i]= trapzc( z_step,(deltaS[,i]^2) )

## Codice per la produzione dei grafici omesso ##

################################################
#            ANALISI GEOSTATISTICA:            #
#         ORDINARY E UNIVERSAL KRIGING         #
################################################

# Previsione del residuo
fit.OK=ord.kriging(X=X,Y=Y,phiS=deltaS,xlim=xlim,ylim=ylim,tol.x=tol.x,tol.y=tol.y,
                 model=model,model.param=c(sill.f,range.f,nugget.f),dist='geodetic')
delta.grid.OK=fit.OK$phiS.grid
s2.OK=fit.OK$s2.OK

# Previsione della temperatura
fit.UK=univ.kriging(X=X,Y=Y,phiS=phiS,Fs=Fs.matrix,Fs.grid=Fs.pred,xlim=xlim,
                 ylim=ylim,tol.x=tol.x,tol.y=tol.y,model=model,
                 model.param=c(sill[l],range[l],nugget[l]),dist='geodetic')
phiS.grid=fit.UK$phiS.grid
s2.UK=fit.UK$s2.UK

################################################
#         ANALISI DI CROSS-VALIDAZIONE         #
################################################
H=h
phiS.cv=matrix(0,nrow=n,ncol=N_samples)
s2.UK.cv=rep(0,N_samples)
for(i in 1: N_samples)
{   ## Codice omesso, disponibile su richiesta all'autrice ##
    ##       alessandra.menafoglio@mail.polimi.it          ##
}
h=H

# Calcolo dello scarto SSE
SSE=rep(0,N_samples)
for(i in 1:N_samples)
    SSE[i]=trapzc(1,(phiS.cv[,i]-phiS[,i])^2)

# Calcolo dello scarto relativo SSE(rel)
norma.phi=rep(0,N_samples)
for(i in 1:N_samples)
```

```
    norma.phi[i]=sqrt(trapzc(1,phiS[,i]^2))
SSE.rel=SSE/(norma.phi^2)

# Produzione output
sse.table=c(summary(SSE)[c(1,3,4,6)],sd(SSE),sum(SSE),mean(norma.phi^2))
names(sse.table)=c('min','med','mean','max','sd','sum','E[norm^2]')

sse.rel.table=c(summary(SSE.rel)[c(1,3,4,6)],sd(SSE.rel),sum(SSE.rel))
names(sse.rel.table)=c('min','med','mean','max','sd','sum')

write.table(sse.table,'SSE.txt')
write.table(sse.rel.table,'SSErel.txt')

################################################
#              CONFRONTO CON OKFD              #
################################################

# Definizione parametri variogramma
h.eu=as.matrix(dist(cbind(X,Y),'euclidean'))
lag.eu=0.35
lmax.eu=2.5
N_lag.eu=ceiling(lmax.eu/lag.eu)
lmax.eu=lag.eu*N_lag.eu
model='spherical'

# Stima binned variogram omnidirezionale
trace_variogram.eu=Tracevariogram.emp.fn(X=X, Y=Y, phiS=phiS, N_samples=N_samples,
                   z_step=z_step, N_lag=N_lag.eu, lmax=lmax.eu, h=h.eu,lag=lag.eu)
gamma_trace.eu=trace_variogram.eu$v
punto_lag.eu=trace_variogram.eu$u
N_h.eu=trace_variogram.eu$n
N_class.eu=N_lag.eu

# Stima binned variogram direzionale

directions=c(0,pi/4,pi/2,3*pi/4)
trace_variogram_anis=Tracevariogram.emp.fn(X=X, Y=Y, phiS=phiS, N_samples=N_samples,
                   z_step=z_step, N_lag=N_lag.eu, lmax=lmax.eu, h=h.eu,
                   directions=directions,tol=pi/8)
gamma_trace=punto_lag=N_h=list()
N_class=rep(0,length(directions))
for(k in 1:length(directions))
{   gamma_trace[[k]]=trace_variogram_anis[[k]]$v
    punto_lag[[k]]=trace_variogram_anis[[k]]$u
    N_h[[k]]=trace_variogram_anis[[k]]$n
    N_class[k]=N_lag
}

# Grafici del variogramma direzionale
k=1
col=rainbow(length(directions))
plot(punto_lag[[k]],gamma_trace[[k]],xlab='Distanza',ylab='Semivariogramma',pch=20,
```

```
           lwd=2,ylim=c(0,max(gamma_trace[[2]])),xlim=c(0,max(lmax.eu)),col=col[k])
title(main='Variogrammi Direzionali')
lines(punto_lag[[k]],gamma_trace[[k]],col=col[k],lwd=2)
text(punto_lag[[k]],gamma_trace[[k]], N_h[[k]],pos=c(4,rep(1,N_lag-1)),col=col[k])
for(k in 2:length(directions))
{   points(punto_lag[[k]],gamma_trace[[k]],col=col[k],pch=20,lwd=2)
    lines(punto_lag[[k]],gamma_trace[[k]],col=col[k],lwd=2)
    text(punto_lag[[k]],gamma_trace[[k]], N_h[[k]],pos=c(4,rep(1,N_lag-1)),col=col[k])
}
legend(0,1000,paste('Direzione',directions/pi*180,'°'),col=col,
        lty=rep(1,length(directions)),lwd=rep(2,length(directions)),
        pch=rep(20,length(directions)))

######## Codice disponibile in rete ########
########      non modificato        ########
L2norm<-matrix(0,nrow=N_samples,ncol=N_samples)
M<-fourierpen(basis,Lfdobj=0)
for (i in 1:(N_samples-1))
{   coef.i<-coef[,i]
    for (j in (i+1):N_samples)
    {   coef.j<-coef[,j]
        L2norm[i,j]<-t(coef.i-coef.j)%*%M%*%(coef.i-coef.j)
    L2norm[j,i]<-L2norm[i,j]
}   }

# Euclidian distance among sites
Eu.d <-as.matrix(dist(coord,method='euclidian'))

# fitting a theorethical model

sigma2.0<-quantile(emp.trace.vari$v,0.75)
phi.0<-quantile(Eu.d,.75)
if(is.null(nugget.fix))
{   fix.nugget<-FALSE
   nugget<-0
   }else{
   fix.nugget=TRUE
   nugget<-nugget.fix
}
if (is.null(max.dist.variogram))
max.dist.variogram<-max(emp.trace.vari$u)
trace.vari<-variofit(emp.trace.vari,ini.cov.pars=c(sigma2.0,phi.0),
                     max.dist=max.dist.variogram,fix.nugget=fix.nugget,
                     nugget=nugget,cov.model='spherical',message=FALSE)

########            Fine            ########
######## Codice disponibile in rete ########

# Parametri del modello valido stimato
sill.eu=as.list(trace.vari)$cov.pars[1]
range.eu=as.list(trace.vari)$cov.pars[2]
nugget.eu=as.list(trace.vari)$nugget
```

```
# Ordinary Kriging
fit.OK.eu=ord.kriging(X=X,Y=Y,phiS=phiS,xlim=xlim,ylim=ylim,tol.x=tol.x,
          tol.y=tol.y,model=model,model.param=c(sill.eu,range.eu,nugget.eu))
phiS.grid.eu=fit.OK.eu$phiS.grid
s2.OK.eu=fit.OK.eu$s2.OK

## Codice per la produzione dei grafici omesso ##
```

Bibliografia

Armstrong M. (1998). *Basic linear geostatistics.* Springer, New York.

Armstrong M.; Diamond P. (1984). Testing variogram for positive-definiteness. *Mathematical geology*, **16**(4), 407–421.

Banjerjee S. (2005). On geodetic distance computations in spatial modeling. *Biometrics*, **61**, 617–625.

Brunsdon C.; Fotheringham S.; Charlton M. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, **28**(4), 281–298.

Casella G.; Berger R. L. (2002). *Statistical Inference.* Duxbury, seconda edizione.

Castro B. F.; Guillas S.; Manteiga W. G. (2005). Functional samples and bootstrap for predicting sulfur dioxide levels. *Technometrics*, **47**(2), 212–222.

Chernick M. R. (1999). *Bootstrap methods.* Wiley series in probability and statistics. John Wiley & Sons, New York.

Chilès J. P.; Delfiner P. (1999). *Geostatistics: Modeling Spatial Uncertainty.* John Wiley & Sons, New York.

Clark R. G.; Alligham S. (2011). Robust resampling confidence intervals for empirical variograms. *Mathematical Geosciences*, **43**(2), 243–259.

Cressie N. (1993). *Statistics for Spatial data.* John Wiley & Sons, New York.

Cuevas A.; Febrero M.; Fraiman R. (2006). On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis*, **51**(2), 1063–1074.

Curriero F. (2006). On the use of non-euclidean distance measures in geostatistics. *Mathematical Geology*, **38**, 907–926.

Delicado P.; Giraldo R.; Comas C.; Mateu J. (2010). Statistics for spatial functional data. *Environmetrics*, **21**(3-4), 224–239.

Deutsch C. V.; Journel A. G. (1998). *GSLIB: Geostatistical software library, user's guide*. Oxford University Press, New York, seconda edizione.

Efron B. (1979). Bootstraps methods: another look at the jackknife. *Ann. Statist.*, **7**, 1–26.

Efron B.; Tibshirani R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall/CRC.

Ferraty F.; Vieu P. (2006). *Nonparametric functional data analysis : theory and practice*. Springer, New York.

Ferraty F.; Keilegom I. V.; Vieu F. (2010). On the validity of the bootstrap in non-parametric functional regression. *Scandinavian Journal of Statistics*, **37**(2), 286–306.

Fisher R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society*, **A**, 309–368.

Giné E.; Zinn J. (1990). Bootstrapping general empirical measures. *Ann. Probab.*, **18**, 851–869.

Giraldo R. (2009). *Geostatistical Analysis of Functional Data*. Tesi di Dottorato di Ricerca, Universitat Politècnica da Catalunya, Barcellona.

Giraldo R.; Delicado P.; Mateu J. (2008a). Geostatistics for functional data: An ordinary kriging approach. Relazione tecnica. Universitat Politècnica de Catalunya, http://hdl.handle.net/2117/1099.

Giraldo R.; Delicado P.; Mateu J. (2008b). *Point-wise kriging for spatial prediction of functional data*, capitolo 22, pp. 135–142. Functional and operatorial statistics. Proceedings of the first international workshop on functional and operatorial statistics. Springer, Toulouse, France.

Giraldo R.; Delicado P.; Mateu J. (2010a). Continuous time-varying kriging for spatial prediction of functional data: An environmental application. *Journal of Agricultural, Biological, and Environmental Statistics*, **15**(1), 66–82.

Giraldo R.; Delicado P.; Mateu J. (2010b). *geofd: a package for prediction for functional data*. Tesi di Dottorato di Ricerca, Universitat Politecnica de Catalunya.

Giraldo R.; Delicado P.; Mateu J. (2010c). Geostatistics for functional data: An ordinary kriging approach. *Environmental and Ecological Statistics*. In Stampa.

Goulard M.; Voltz M. (1993). Geostatistical interpolation of curves: A case study in soil science In *Geostatistics Tróia '92*. A cura di Soares A., volume 2, pp. 805–816. Dordrecht: Kluwer Academic.

Guggenheimer H. W. (1977). *Differential Geometry*. Dover Pubblications, New York.

Hall P. (1985). Resampling a coverage pattern. *Stochastic Process. Appl*, **20**, 231–246.

Hörmann S.; Kokoszka P. (2011). Consistency of the mean and the principal components of spatially distributed functional data.

Huang C.; Zhang H.; Robeson S. M. (2011). On the validity of commonly used covariance and variogram functions on the sphere. *Mathematical Geosciences*, **43**(6), 721–733.

Hutchinson M. F. (2004). Anusplin version 4.3. *Centre for Resource and Environmental Studies*. The Australian National University, Canberra, Australia.

Johnson R. A.; Wichern D. W. (2007). *Applied Multivariate Statistical Analysis*. Prentice Hall, sesta edizione.

Jr P. J. R.; Diggle P. J. (2001). geoR: a package for geostatistical analysis. *R-NEWS*, **1**(2), 14–18. ISSN 1609-3631.

Kosorok M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Series in Statistics. Springer, New York.

Künsch H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, **17**, 1217–1241.

Lahiri S. N. (1992). *Edgeworth correction by 'moving block' bootstrap for stationary and nonstationary data,* in *Exploring the Limits of Bootstrap*. John Wiley & Sons, New York.

Lahiri S. N. (1993). On the moving block bootstrap under long range dependence. *Statist. Probab. Lett.*, **18**, 405–413.

Loh J. M.; Stein M. L. (2008). Spatial bootstrap with increasing observations in a fixed domain. *Statistica Sinica*, **18**, 667–688.

Mahalanobis P. C. (1936). On the generalised distance in statistics. *Proceedings National Institute of Science,* India, **2**(1), 49–55.

Marchant B. P.; Lark R. M. (2004). Estimating variogram uncertainty. *Mathematical Geology*, **36**(8), 867–898.

Matheron G. (1962). *Traité de géostatistique appliqué, Tome I; Tome II: Le krigeage*. I: Mémoires du Bureau de Recherches Géologiques et Miniéres, Numero 14 (1962), Edizioni Technip, Parigi; II: Mémoires du Bureau de Recherches Géologiques et Miniéres, Numero 24 (1963), Edizioni B.R.G.M., Parigi.

Matheron G. (1978). *Estimer et choisir*. Numero 7. Cahiers du centre de morphologie mathématique de fontainbleau. Ecole des mines de Paris.

Monestiez P.; Nerini D. (2008). A cokriging method for spatial functional data with applications in oceanology. In *Functional and operatorial statistics*. A cura di Dabo-Niang S., Ferraty F. Springer.

Olea R.; Pardo-Igúzquiza E. (2011). Generalized bootstrap method for assessment of uncertainty in semivariogram inference. *Mathematical Geosciences*, **43**(2), 203–228.

Pardo-Igúzquiza E.; Dowd P. (2001). Variance-covariance matrix of the experimental variogram: assessing variogram uncertainty. *Mathematical Geology*, **33**(4), 397–419.

Pardo-Igúzquiza E.; Dowd P. (2003). Assessment of the uncertainty of spatial covariance parameters of soil properties and its use in applications. *Soil Science*, **168**(11), 769–782.

Politis D. N.; Romano J. P. (1994). Limit theorems for weakly dependent hilbert space valued random variables with application to the stationarity bootstrap. *Statist. Sinica*, **4**, 461–476.

Porcu E. (2004). *Geostatistica Spazio-temporale: Nuove Classi di Covarianza, Variogramma e Densità Spettrali*. Tesi di Dottorato di Ricerca, Università di Milano Bicocca, Milano.

Quarteroni A.; Sacco S.; Saleri F. (2008). *Matematica Numerica*. Springer Verlag, Italia.

Quenouille M. H. (1949). Approximate tests of correlation in time series. *J. R. Statist. Soc.*, **11**, 353–360.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ramsay J.; Silverman B. (2005). *Functional data analysis*. Springer, New York, seconda edizione.

Ramsay J. O.; Dalzell C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society*, **53**(3), 539–572.

Ramsay J. O.; Wang X.; Flanagan R. (1995). A functional data analysis of the pinch force of the human fingers. *Applied Statistics*, **44**(3), 17–30.

Ramsay J. O.; Wickham H.; Graves S.; Hooker G. (2010). *fda: Functional Data Analysis*. R package version 2.2.5.

Sebastién Z. (2004). Uncertainty investigation of the semivariogram by means of the bootstrap method. *Acta Geol Hung*, **47**(1), 83–91.

Shumway R. H.; Dean W. (1968). Best linear unbiased estimation for multivariate stationary processes. *Technometrics*, **10**(3), 523–534.

Solow A. R. (1985). Bootstrapping correlated data. *Mathematical Geosciences*, **17**(7), 769–775.

Spagnolini U. (2010). Elaborazione statistica dei segnali per le telecomunicazioni. Appunti di Lezione.

Stanley D. (2002). *Canada's Maritime provinces*. Marybirnong: Lonely Planet Pubblications, prima edizione.

Tang L.; Schucany W.; Woodward W.; Gunst R. (2006). A parametric spatial bootstrap. Relazione tecnica. Southern Methodist University, Dallas, Texas.

Tarpey T.; Kinateder K. K. J. (2003). Clustering functional data. *Journal of Classification*, **20**, 93–114.

Tucker H. G. (1959). A generalization of Glivenko-Cantelli theorem. *The Annals of Mathematical Statistics*, **30**(3), 828–830.

van der Vaart A.; Wellner J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, New York.

Wellner J. A. (2011). Advanced theory of statistical inference. Lecture notes. University of Washington, Department of Statistics, Seattle.

Yamanishi Y.; Tanaka Y. (2003). Geographically weighted functional multiple regression analysis: a numerical investigation. *Journal of Japanese Society of Computational Statistics*, (15), 307–317.