POLITECNICO DI TORINO

Corso di Laurea Magistrale
in Ingegneria Matematica

Tesi di Laurea Magistrale

Analisi della non linearità nei processi di trasporto negli acquiferi

Relatori
prof. Ilaria Butera
prof. Luca Ridolfi
firma dei relatori

. . . . . .
. . . . . .

Candidato
Massimiliano Romana

firma del candidato

. . . . . .

Anno Accademico 2020-2021

A mio papà

Sommario

Per capire a fondo i fenomeni delle acque sotterranee, è utile studiare il tipo di relazione che esiste tra le variabili di flusso e di trasporto. La geostatistica è la parte di statistica che studia i legami lineari tra i dati spaziali ed è una scienza ben consolidata nello studio delle acque sotterranee. Lo scopo di questa tesi è di scegliere e sviluppare uno strumento matematico adatto a rilevare legami non lineari tra dati spaziali ed applicarlo allo studio delle acque sotterranee. A tale scopo, sono stati applicati gli strumenti della teoria dell'informazione, quale l'informazione mutua, che è in grado di rivelare il legame non lineare tra due variabili. Per validare questo approccio si è sviluppato uno studio numerico su dati sintetici attraverso codici di calcolo scritti in Matlab e in Model Muse.

Ringraziamenti

Desidero ringraziare profondamente la Professoressa Ilaria Butera ed il Professore Luca Ridolfi per l'aiuto, la disponibilità e la fiducia che mi hanno mostrato durante il periodo di lavoro per questa tesi.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

Le risorse idriche presenti nel sottosuolo sono di grande importanza, soprattutto perchè fonte ricca e di alta qualità di acqua dolce. Purtroppo, queste risorse sono soggette a sfruttamento e a inquinamento da parte del genere umano. E' normale voler conoscere meglio gli *acquiferi*, cioè comprenderne meglio le sue caratteristiche fisiche, attraverso i suoi parametri, per poter monitorare i processi di inquinamento e dimensionare, e quindi mettere in campo, azioni di bonifica. La difficoltà nel reperire informazioni sulle caratteristiche fisiche degli acquiferi ha portato la comunità scientifica del settore, a orientarsi per un approccio statistico-probabilistico, cioè considerare un acquifero come un evento stocastico e prendere in considerazione le misure di un suo parametro fisico come realizzazione di quello stesso evento casuale. Questo porta ad esempio, a considerare la trasmissività dell'acquifero, come un campo stocastico spazialmente correlato. Di conseguenza, le variabili di flusso e di trasporto sono da considerarsi stocastiche (I. Butera and Ridolfi [2018]).

In qualsiasi processo fisico è di rilevante importanza comprendere se la non linearità del processo stesso gioca un ruolo o meno. In particolare, nei processi di filtrazione dell'acqua nel sottosuolo, tale importanza è legata anche ad alcune classi di modelli di trasporto, come ad esempio quelli geostatistici, che assumono che vi siano legami lineari tra le variabili in gioco, in tempi e in punti dello spazio diversi. Se si mettessero in evidenza tali non linearità, si avrebbe la facoltà di costruire modelli più precisi e quindi più realistici. Da queste considerazioni, nasce l'idea di simulare processi di trasporto, che usano le equazioni complete che descrivono il processo, e a posteriori capire quanto sia presente il contributo della non linearità.

L'obiettivo di questa tesi è di indagare nei processi di trasporto, nello specifico di filtrazione, l'importanza delle non linearità. Per fare questo si è partiti dalla geostatistica e la si è messa a confronto con uno strumento della teoria dell'informazione; in particolare si mostra come si applicare questa teoria in un esperimento numerico di trasporto di un inquinante in un acquifero.

Il resto della tesi è strutturato esssenzialmente in tre capitoli. Nel secondo capitolo si descrive la classica analisi geostatistica, quinidi la statistica di dati spaziali, introducendo nozioni della teoria dell'informazione, in particolare l'informazione mutua. La geostatistica è una scienza ben consolidata, mentre la stima dell'informazione mutua ha bisogno

di una revisione dei metodi per il suo calcolo, quindi vengono passati in rassegna i vari algoritmi. Nel terzo capitolo si validano gli strumenti statistici presentati, su dati generati sinteticamente, e si valutano le performance dei vari metodi di stima dell'informazione mutua presentati nel capitolo precedente. Inoltre viene presentato un metodo per generare campi stocastici spazialmente correlati, come ad esempio un campo di trasmissività. Nel quarto capitolo, si esamina un caso studio: uno studio numerico del flusso in un acquifero e del trasporto di un inquinante di cui si vuole esaminare la non linearità della concentrazione in spazio. L'ultimo capitolo è dedicato alle conclusioni di questo studio: si riassume ciò che è stato svolto e si analizza cosa si può ottenere dalle informazioni raccolte. In appendice, viene riassunta la parte riguardante ai codici di calcolo usati, quindi Modflow per il flusso e MT3DMS per il trasporto dell'inquinante nell'acquifero.

Capitolo 2

Analisi statistica spaziale lineare e non lineare

In questo capitolo, dopo aver descritto le nozioni di base sui processi e campi stocastici, si introduce la geostatistica, che si occupa dell'applicazione della statistica ai dati spaziali, e che in particolare cerca con degli stimatori opportuni di rilevare legami lineari tra i dati. Di particolare interesse sono gli strumenti per il calcolo del variogramma e il covariogramma; strettamente legato a questi strumenti è il concetto di kriging. Dopodiché, si presentano i concetti base della teoria dell'informazione al fine di individuare i legami non lineari nei dati spaziali. In pratica, vengono passati in rassegna i vari metodi di stima della quantità d'interesse, ovvero l'informazione mutua.

2.1 Nozioni di probabilità: processi e campi stocastici

Si definisce variabile stocastica (casuale, aleatoria) una variabile che assume valori dipendentemente da un evento stocastico. Ad esempio un lancio di un dado può essere modellato come un evento stocastico, e il suo risultato è una variabile stocastica. Nel seguito, però, si tratterà soltanto di variabili continue cioè dove l'insieme dei possibili valori ha la potenza del continuo, come ad esempio l'insieme dei reali $\mathbb{R}$.

In generale, una variabile stocastica è completamente caratterizzata dalla sua funzione di densità di probabilità. Quest'ultima non è altro che una funzione non negativa, il cui integrale sul proprio dominio è pari all'unità, che descrive la probabilità che un evento stocastico si possa verificare. La probabilità che un evento A si verifichi data la funzione di densità di probabilità $f_X(x)$, dove X è la variabile stocastica in questione, è pari a:

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx. \tag{2.1}$$

Un esempio di variabile stocastica continua è la variabile distribuita secondo una Gaussiana o normale, la quale ha la funzione di densità di probabilità pari a:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{\sigma^2}}; \tag{2.2}$$

in particolare il supporto di questa variabile è l'insieme $\mathbb{R} = (-\infty, +\infty)$ e μ e σ sono parametri di media e deviazione standard rispettivamente. Un altro esempio di distribuzione continua è la distribuzione Gamma:

$$f_X(x) = \frac{1}{\vartheta^k \Gamma(k)} x^{k-1} e^{-x/\vartheta}; \tag{2.3}$$

in questo caso il supporto è l'insieme $(0, +\infty)$, Γ è la funzione gamma, mentre k e ϑ sono i parametri di forma e scala rispettivamente. Quando si considerano due o più variabili stocastiche e se ne vuole descriverne la distribuzione di probabilità si parla di variabile multivariata. Nel caso di due variabili X, Y si parla di variabile bivariata e la funzione di densità di probabilità f_{XY} è definita su un dominio bidimensionale Ω. Ad esempio la distribuzione Gaussiana bivariata risulta essere:

$$f_{XY}(x, y) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp\left(-\frac{1}{2}(x - \mu_x,\ y - \mu_y)\Sigma^{-1}(x - \mu_x,\ y - \mu_y)^T\right). \tag{2.4}$$

Mentre la distribuzione normale bivariata è unica, non lo è la bivariata le cui marginali sono distribuite secondo una Gamma (Nadarajah and Kotz [2009]); ad esempio si può considerare

$$f_{XY}(x, y) = C(xy)^c \left(\frac{x}{\mu_1} + \frac{y}{\mu_2}\right)^{a-2c} \Gamma\left(2c - a, \frac{x}{\mu_1} + \frac{y}{\mu_2}\right). \tag{2.5}$$

Le distribuzione bivariate e in generale le multivariate sono anch'esse soggette al vincolo che l'integrale sul dominio deve dare l'unità, ma si aggiunge anche che gli integrali marginali devono coincidere con le distribuzione delle singole variabili univariate, dette distribuzioni marginali:

$$f_X(x) = \int_{\Omega_Y} f_{XY}(x, y)dy;\ f_Y(y) = \int_{\Omega_X} f_{XY}(x, y)dx. \tag{2.6}$$

Infine si parla di *indipendenza statistica* tra le due distribuzioni di X e Y se vale

$$f_{XY}(x, y) = f_X(x) f_Y(y). \tag{2.7}$$

Prima di parlare di processi e campi stocastici, si ricordano le nozioni di valore atteso, media, varianza, covarianza, deviazione standard, correlazione, e momento, ovvero gli indici caratteristici di una variabile stocastica.

Per valore atteso della variabile X, s'intende il seguente integrale

$$\mathbb{E}[X] = \int_\Omega x f_X(x)dx, \tag{2.8}$$

dove Ω è il supporto di X. Questo valore viene anche chiamata media di X. La varianza è definita come segue:

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \int_\Omega x^2 f_X(x)dx - \left(\int_\Omega x f_X(x)dx\right)^2 \quad (2.9)$$

mentre la deviazione standard è la radice quadrata della varianza. Nel caso di due o più variabili stocastiche $X_1, X_2, X_3, .., X_i, ..., X_n$, si definisce anche la covarianza o la matrice di covarianza nXn-dimensionale che ha come entrate

$$\mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \int_\Omega (x_i - \mathbb{E}[X_i])(x_j - \mathbb{E}[X_j]) f_{X_1,...,X_n}(x_1, ..., x_n)dx_1 \cdots dx_n. \quad (2.10)$$

La correlazione tra due variabili X, Y risulta essere:

$$\mathrm{Corr}(X, Y) = \frac{\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]}{\sqrt{\mathbb{E}[(X - \mathbb{E}[X])^2]\mathbb{E}[(Y - \mathbb{E}[Y])^2]}}; \quad (2.11)$$

cioè la correlazione è uguale al rapporto tra la covarianza tra le due variabili e il prodotto delle deviazioni standard delle rispettive variabili X, Y. Per la disuguaglianza di Schwarz (vedere A. G. Asuero and Gonz´alez [2006], I. Butera and Ridolfi [2018]) il modulo del numeratore è maggiore del denominatore, quindi il $\mathrm{Corr}(X;Y) \in [-1,1]$. In particolare *la correlazione misura la linearità della relazione tra le due variabili* X e Y: $|\mathrm{Corr}(X;Y)| = 1$ se la relazione tra X e Y è del tipo $Y = aX + b$. Questo è mostrato con chiarezza in Frison [2015]: il termine $\mathrm{Corr}(X;Y)$ viene chiamato anche coefficiente di correlazione lineare di Pearson, in quanto è adimensionale, e se la relazione è del tipo $Y = aX + b$, allora valgono le sequenti espressioni

$$\begin{aligned} \mathbb{E}[Y] &= a\mathbb{E}[X] + b \\ \mathrm{Var}(Y) &= a^2\mathrm{Var}(X) \\ \mathrm{Cov}(X;Y) &= a\mathrm{Var}(X), \end{aligned} \quad (2.12)$$

quindi il coefficiente di correlazione risulta essere

$$\mathrm{Corr}(X;Y) = \frac{\mathrm{Cov}(X;Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{a\mathrm{Var}(X)}{|a|\sqrt{var(X)^2}} = \mathrm{sign}(a) = \pm 1, \quad (2.13)$$

a seconda della positività o negatività del parametro a.

Date N realizzazioni (x_i, y_i) della coppia di variabili stocastiche X, Y, quindi un campione di numerosità N, lo stimatore della correlazione è definito come segue:

$$\rho_{XY} = \frac{\frac{1}{N-1}\sum_i^N (x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_i (x_i - \mu_x)^2 \sum_i (y_i - \mu_y)^2}}, \quad (2.14)$$

dove μ_x, μ_y sono gli stimatori della media del campione x, y rispettivamente. Il momento di ordine k della variabile X è definito nel seguente modo:

$$\mathbb{E}[X^k] = \int_\Omega x^k f_X(x)dx. \quad (2.15)$$

Ritornando agli esempi di prima, si verifica che media e covarianza della variabile stocastica multivariata distribuita secondo una Gaussiana sono rispettivamente μ e Σ. Mentre per una variabile stocastica univariata distribuita secondo una gamma con parametri k e ϑ, la media risulta essere pari a $k\vartheta$ e la varianza pari a $k\vartheta^2$.

Per quanto a riguarda le definizioni rigorose di processo e campo stocastico si rimanda a Carrillo [2018]. Dapprima si considera un insieme di variabili stocastiche indicizzate dal tempo $\{Z_t\}_t \in \mathcal{T}$ dove $\mathcal{T}$ può essere discreto o continuo: questo insieme viene chiamato processo stocastico a tempo discreto o continuo a seconda dell'indicizzazione. In altre parole, X_t è una funzione stocastica del tempo: ad ogni istante di tempo associa una variabile stocastica. Se invece si considera un insieme di punti $\{\boldsymbol{x}\}_i \in \mathbb{R}^d$ (ad esempio $\boldsymbol{x} = (x, y)$), con d = 1,2,3, e si associa ad ogni punto una variabile stocastica $Z(\boldsymbol{x})$, si ottiene un campo stocastico. Come nel caso precedente, il campo stocastico può essere discreto o continuo a seconda dell'indicizzazione dell'insieme $\{\boldsymbol{x}\}_i$. Realisticamente, in geostatistica si considerano un insieme discreto finito di variabili stocastiche, quindi un campo discreto che corrispondono eventualmente a misure puntuali della quantità d'interesse (Cressie [1993]). Se consideriamo un campo stocastico discreto finito $(\{\boldsymbol{x}\}_i^n)$, la sua caratterizzazione è dovuta alla sua funzione di densità di probabilità congiunta o multivariata

$$f_{Z(\boldsymbol{x}_1),...,Z(\boldsymbol{x}_n)}(z_1, ..., z_n). \tag{2.16}$$

Un esempio di processo e di campo stocastico è quello Gaussiano, in cui la distribuzione congiunta è una distribuzione normale multivariata.

2.2 Geostatistica lineare

La geostatistica è il ramo della statistica che si occupa di dati spaziali, cioè a quantità riferite a delle posizioni nello spazio. L'aggettivo *lineare* sta ad indicare che cerca relazioni lineari nell'insieme di dati spaziali (Bez).

2.2.1 Variogramma e covarianza spaziale

Per introdurre i concetti di variogramma e covarianza spaziale, si definiscono prima le varie proprietà di stazionarietà dei campi stocastici.

Un campo stocastico si dice *strettamente stazionario* se vale l'uguaglianza

$$f_{Z(\boldsymbol{x}_1+\boldsymbol{h}),...,Z(\boldsymbol{x}_n+\boldsymbol{h})}(\tilde{z}_1, ..., \tilde{z}_n). = f_{Z(\boldsymbol{x}_1),...,Z(\boldsymbol{x}_n)}(z_1, ..., z_n). \tag{2.17}$$

dove $\boldsymbol{h}$ è la distanza tra i punti $\boldsymbol{x}$ e $\boldsymbol{x}+\boldsymbol{h}$. In altre parole, la funzione di densità di probabilità è invariante per traslazioni nello spazio. Equivalentemente, si parla di stazionarietà stretta se tutti i momenti di ogni ordine sono invarianti per traslazioni. Un altro tipo di stazionarietà, è quella debole: un campo stocastico è *debolmente stazionario* se esiste una funzione $C : (0, +\infty) \to \mathbb{R}$ tale che

$$\mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)) = \tilde{C}(\boldsymbol{h}) = C(h) \ \ \forall h \tag{2.18}$$

dove $h = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ è la distanza tra i due punti, anche chiamata *lag*, e l'ultima uguaglianza è vera se il campo è *isotropo*, e se la media è costante in spazio

$$\mathbb{E}[Z(\boldsymbol{x})] = \mu, \ \forall \boldsymbol{x}. \tag{2.19}$$

In particolare si ha che $C(0) = \mathrm{Var}(Z(\boldsymbol{x})), \forall \boldsymbol{x}$. In altre parole la stazionarietà debole implica che il campo stocastico non ha la media dipendente dallo spazio e la sua funzione di covarianza dipende solo dal distanza tra i punti. Talvolta, questa stazionarietà viene chiamata del *secondo ordine* poichè soltanto i primi due momenti sono indipendenti dallo spazio. La funzione di covarianza viene spesso anche chiamata covariogramma o autocovarianza, ma nel corso di questa tesi, per differenziarla dalla comune covarianza, si aggiunge l'aggettivo *spaziale*. In maniera simile si definisce la correlazione tra due punti distanti h:

$$\mathrm{Corr}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j))] = \frac{C(h)}{C(0)} = \frac{C(h)}{\mathrm{Var}(Z(\boldsymbol{x}))}. \tag{2.20}$$

L'ultimo tipo di stazionarietà che si mostra riguardano gli incrementi del campo stocastico: si dice che il campo stocastico $Z(\boldsymbol{x})$ è *intrinsecamente stazionario* se gli incrementi in spazio hanno media nulla

$$\mathbb{E}[Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_j)] = 0; \boldsymbol{x}_i \neq \boldsymbol{x}_j \tag{2.21}$$

e se la varianza degli incrementi dipende solo dal lag (distanza tra i punti)

$$\mathrm{Var}(Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_j)) = 2\gamma(h) \tag{2.22}$$

dove si è supposta l'isotropia e $\gamma : [0, +\infty) \to [0, +\infty)$ è una funzione del lag, chiamata *semivariogramma* ($2\gamma(h)$ è chiamata variogramma). Da notare che la stazionarietà stretta è più forte della stazionarietà del secondo ordine, e la stazionarietà intrinseca è più debole di quella del secondo ordine. Se il campo stocastico è debolmente stazionario, allora valgono le seguenti relazioni:

$$\gamma(h) = C(0) - C(h), \tag{2.23}$$

$$C(h) = \gamma(h) - \lim_{h \to +\infty} \gamma(h) \tag{2.24}$$

Per dimostrarle si prende come riferimento Lichtenstern [2013]: la varianza delle variazioni della campo stocastico Z risulta essere

$$\begin{aligned} \mathrm{var}(Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h})) &= \mathbb{E}[(Z(\boldsymbol{x}) - Z(\boldsymbol{x}+\boldsymbol{h}))^2] \\ &= \mathbb{E}[(Z(\boldsymbol{x})^2] + \mathbb{E}[(Z(\boldsymbol{x}+\boldsymbol{h})^2] - \mathbb{E}[Z(\boldsymbol{x})Z(\boldsymbol{x}+\boldsymbol{h})] \\ &= \mathrm{Var}(Z(\boldsymbol{x})) + \mu^2 + \mathrm{Var}(Z(\boldsymbol{x}+\boldsymbol{h})) + \mu^2 - 2(C(h) - \mu^2) \\ &= \mathrm{Var}(Z(\boldsymbol{x})) + \mathrm{Var}(Z(\boldsymbol{x}+\boldsymbol{h})) - 2C(h) \\ &= 2\mathrm{Var}(Z(\boldsymbol{x})) - 2C(h) = 2C(0) - 2C(h), \end{aligned} \tag{2.25}$$

quindi $\gamma(h) = \frac{1}{2}\mathrm{Var}(T(\boldsymbol{x}) - T(\boldsymbol{x}+\boldsymbol{h})) = C(0) - C(h)$. Un ultima proprietà importante da enunciare è quella dell'*ergodicità*. Un processo si dice ergodico se la media temporale

su una realizzazione approssima bene la media su tante realizzazioni (Cressie [1993]). In generale si può enunciare il teorema ergodico di Birkhoff che dice che se un processo è ergodico il suo momento di ordine 1 (valore atteso) esiste (ed è costante per ergodicità) allora la media temporale tende al suo valore atteso, detto media d'insieme. La condizione di ergodicità implica la stazionarietà stretta. Dal lavoro di Cressie [1993] si deduce che un esempio un campo stocastico ergodico è un campo stocastico distribuito come una Gaussiana, con media costante in spazio e funzione di correlazione $\rho(h)$ che tende a zero quando $h \to +\infty$.

Si consideri ora una realizzazione del campo stocastico, cioè un campo $Z(\boldsymbol{x})$ noto su un insieme di punti $\{\boldsymbol{x}_i\}_{i=1}^N$. Lo *stimatore del semivariogramma* (anche detto semivariogramma sperimentale, campionario, empirico), risulta essere calcolato mediante la seguente formula:

$$\widehat{\gamma}(h) = \frac{1}{2N_h} \sum_{(i,j)\in\mathcal{I}_h} (Z(\boldsymbol{x}_i) - Z(\boldsymbol{x}_j))^2 \tag{2.26}$$

con $\mathcal{I}_h = \{(i,j) : |h| - \epsilon < |\boldsymbol{x}_i - \boldsymbol{x}_j| < |h| + \epsilon\}$ e $N_h = |\mathcal{I}_h|$. Analogamente si può definire lo *stimatore della covarianza spaziale* (Cressie [1993]) (covarianza spaziale sperimentale, campionaria, empirica) con la seguente formula:

$$\widehat{C}(h) = \frac{1}{N_h} \sum_{(i,j)\in\mathcal{I}_h} \left[Z(\boldsymbol{x}_i) - \sum_{k=1}^{N} Z(\boldsymbol{x}_k) \right] \left[Z(\boldsymbol{x}_j) - \sum_{k=1}^{N} Z(\boldsymbol{x}_k) \right] \tag{2.27}$$

con $\mathcal{I}_h = \{(i,j) : |h| - \epsilon < \|\boldsymbol{x}_i - \boldsymbol{x}_j\| < |h| + \epsilon\}$, $N_h = |\mathcal{I}_h|$. Nella pratica (Lichtenstern [2013]) i due stimatori sono affidabili solo per distanze h pari alla metà del diametro della regione. In particolare, per quanto riguarda gli stimatori si ha che, chiamato $\mu = \frac{1}{N}\sum_{k=1}^{N} T(\boldsymbol{x}_k)$,

$$\widehat{C}(0) - \widehat{C}(h) = -2\mu^2 + \frac{1}{N}\sum_{k}^{N}(Z(\boldsymbol{x}_k))^2 - \frac{1}{N_h}\sum_{I_h} Z(\boldsymbol{x}_i)Z(\boldsymbol{x}_j) + \frac{\mu}{N_h}\sum_{I_h}(Z(\boldsymbol{x}_i) + Z(\boldsymbol{x}_j)) \tag{2.28}$$

mentre

$$\widehat{\gamma}(h) = -\frac{1}{N_h}\sum_{I_h} Z(\boldsymbol{x}_i)Z(\boldsymbol{x}_j) + \frac{1}{2N_h}\sum_{I_h} Z(\boldsymbol{x}_i)^2 + Z(\boldsymbol{x}_j)^2. \tag{2.29}$$

Ne consegue che la relazione $\widehat{\gamma}(h) = \widehat{C}(0) - \widehat{C}(h)$ non sussiste per gli stimatori del semivariogramma e della covarianza spaziale (Cressie [1993]).

Gli stimatori della covarianza e del semivariogramma dipendono dalla particolare realizzazione del campo stocastico ma lo stimatore è consistente (Cressie [1993]) ed in particolare lo si può verificare grazie al teorema di Birkhoff.

Si chiude questa sezione con la presentazione delle leggi che approssimano lo stimatore del semivariogramma sopraesposto. Le leggi più impiegate per modellare la stima del semivariogramma sono le sequenti:

• modello potenza:

$$\widehat{\gamma}(h) = \omega h^{\alpha} \tag{2.30}$$

- modello esponenziale:

$$\widehat{\gamma}(h) = \omega(1 - \exp(-\frac{h}{\alpha}) \tag{2.31}$$

- modello gaussiano:

$$\widehat{\gamma}(h) = \omega(1 - \exp(-\frac{h^2}{\alpha^2}). \tag{2.32}$$

2.2.2 Kriging

Nell'affrontare un problema realistico di geostatistica si ha a che fare con poche misure di una realizzazione del campo stocastico d'interesse. Serve quindi elaborare una strategia per ottenere un insieme di dati più numeroso con una tecnica di estrapolazione: la via più comunemente usata è il metodo del kriging che si serve dei modelli sopraesposti del semivariogramma ed eventualmente della covarianza spaziale. La seguente trattazione è interamente basata sul lavoro di Lichtenstern [2013].

Il kriging è un metodo di stima che consiste nella determinazione di un modello statistico a partire dai dati disponibili, tramite l'uso del variogramma, seguita dalla costruzione di un previsore ottimo dato dalla combinazione lineare dei dati. In alternativa alla regressione lineare, il kriging tiene conto dei volumi di osservazioni e la dipendenza stocastica tra i dati. Si distinguono tre tipi di kriging in relazione alla stazionarietà o meno del campo stocastico considerato:: *semplice* (se stazionario del second'ordine a media nota), *ordinario* (se intrinsecamente stazionario a media incognita) e *universale* (se non stazionario, con media espressa mediante un coefficiente di drift). In questa tesi si esporranno soltanto i primi due metodi di kriging. Il metodo funziona al meglio dentro il cono convesso determinato dai punti del bordo. In particolare il metodo è un interpolatore esatto, cioè nei punti già noti, lo stimatore è riproduce le misure note.

Kriging semplice

Questa tipologia di kriging suppone che la media sia nota e costante pari a $\mu = \mathbb{E}[Z(\boldsymbol{x})]$ per ogni punto $\boldsymbol{x}$. Si suppone anche la stazionarietà del secondo ordine, quindi covarianza espressa in funzione della distanza tra i punti. Lo stimatore di un valore del campo stocastico in un punto non di misura è espresso come segue:

$$\widehat{Z}(\boldsymbol{x}_0) = \mu + \sum_{i=1}^{N} \lambda_0^i (Z(\boldsymbol{x}_i) - \mu) \tag{2.33}$$

dove λ_0^i sono coefficienti da trovare per rendere valida la stima. Si verifica che questo stimatore non è distorto, cioè il suo valore atteso coincide con la media nota:

$$\mathbb{E}[\widehat{Z}(\boldsymbol{x}_0)] = \mu + \sum_{i=1}^{N} \lambda_0^i (\mathbb{E}[Z(\boldsymbol{x}_i)] - \mu) = \mu \tag{2.34}$$

Si impone ora che la varianza dell'errore di stima sia minima:

$$\begin{aligned}
\mathrm{Var}(\widehat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0)) &= \mathbb{E}[(\widehat{Z}(\boldsymbol{x}_0) - Z(\boldsymbol{x}_0))^2] \\
&= \mathbb{E}[(\sum_i \lambda_0^i (Z(\boldsymbol{x}_i) - \mu) + (\mu - Z(\boldsymbol{x}_0)))^2] \\
&= \mathbb{E}[(\sum_i \lambda_0^i (Z(\boldsymbol{x}_i) - \mu))^2] + \mathbb{E}[(\mu - Z(\boldsymbol{x}_0))^2] + \\
&+ 2\mathbb{E}[(\sum_i \lambda_0^i (Z(\boldsymbol{x}_i) - \mu))(\mu - Z(\boldsymbol{x}_0))] \\
&= \sum_j \sum_k \lambda_0^j \lambda_0^k \mathbb{E}[(Z(\boldsymbol{x}_j) - \mu)(Z(\boldsymbol{x}_k) - \mu)] + \mathrm{Cov}(Z(\boldsymbol{x}_0, \boldsymbol{x}_0)) \\
&+ 2\sum_i \lambda_0^i \mathbb{E}[(Z(\boldsymbol{x}_i) - \mu)(\mu - Z(\boldsymbol{x}_0))] \\
&= \sum_j \sum_k \lambda_0^j \lambda_0^k \mathrm{Cov}(Z(\boldsymbol{x}_j), Z(\boldsymbol{x}_k)) + \mathrm{Cov}(Z(\boldsymbol{x}_0), Z(\boldsymbol{x}_0)) \\
&- 2\sum_i \lambda_0^i \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_0)).
\end{aligned} \tag{2.35}$$

Adesso si impone la condizioni di minimo

$$\begin{aligned}
\frac{\partial}{\partial \lambda_0^l} \mathbb{E}[(\widehat{T}(\boldsymbol{x}_0) - T(\boldsymbol{x}_0))^2] &= 2\sum_m \lambda_0^m \mathrm{Cov}(T(\boldsymbol{x}_m), T(\boldsymbol{x}_l)) - 2\mathrm{Cov}(T(\boldsymbol{x}_l), T(\boldsymbol{x}_0)) \\
&= 0
\end{aligned} \tag{2.36}$$

che implica

$$\sum_m \lambda_0^m \mathrm{Cov}(T(\boldsymbol{x}_m), T(\boldsymbol{x}_l)) = \mathrm{Cov}(T(\boldsymbol{x}_l), T(\boldsymbol{x}_0)) \tag{2.37}$$

e scritto in forma matriciale diventa

$$\begin{pmatrix} \ddots & \cdots & \cdots \\ \vdots & \mathrm{Cov}(T(\boldsymbol{x}_m), T(\boldsymbol{x}_l)) & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ \lambda_0^m \\ \vdots \end{pmatrix} = \begin{pmatrix} \vdots \\ \mathrm{Cov}(T(\boldsymbol{x}_0), T(\boldsymbol{x}_l)) \\ \vdots \end{pmatrix}.$$

Kriging ordinario

Con il metodo del kriging ordinario si suppone la stazionarietà intrinseca e si suppone di conoscerne il semivariogramma in forma isotropa del campo stocastico $Z(\boldsymbol{x})$ e con media costante e non nota. Si supponga di voler conoscere la variabile stocastica $Z(\boldsymbol{x}_0)$ con $\boldsymbol{x}_0 \notin \{\boldsymbol{x}_i\}_{i=1}^N$. E' possibile determinare una stima della variabile stocastica di $Z(\boldsymbol{x}_0)$, indicata con $\widehat{Z}(\boldsymbol{x}_0)$), mediante una combinazione lineare pesata di un set di N variabili stocastiche:

$$\widehat{Z}(\boldsymbol{x}_0)) = \sum_{i=1}^N \lambda_0^i Z(\boldsymbol{x}_i) \tag{2.38}$$

La condizione che viene imposta qui è che la varianza sia minima, quindi si cerca l'insieme di $\{\lambda_0^i\}_{i=1}^N$, che sommano a uno $\sum_i^N \lambda_0^i = 1$, tali che $\mathbb{E}[(Z(\boldsymbol{x}_0) - \widehat{Z}(\boldsymbol{x}_0))^2]$ sia minima, avendo supposto che la media sia costante e nota. Quindi si ha che:

$$\{\lambda_0^i\}_{i=1}^N = \underset{\{\lambda_0^i\}_{i=1}^N \in \mathbb{R}^N}{\arg\min}\ \mathbb{E}[(T(\boldsymbol{x}_0) - \widehat{T}(\boldsymbol{x}_0))^2]. \tag{2.39}$$

Si studia ora il termine da minimizzare:

$$\begin{aligned}
\mathbb{E}[(Z(\boldsymbol{x}_0) - \widehat{Z}(\boldsymbol{x}_0))^2] &= \mathbb{E}[(Z(\boldsymbol{x}_0) - \sum_{i=1}^N \lambda_0^i Z(\boldsymbol{x}_i))^2] \\
&= \mathbb{E}[Z(\boldsymbol{x}_0)^2] + \mathbb{E}[\sum_{i=1}^N \lambda_0^i Z(\boldsymbol{x}_i))^2] - 2\mathbb{E}[\sum_{i=1}^N \lambda_0^i Z(\boldsymbol{x}_i) Z(\boldsymbol{x}_0)] \\
&= -\sum_j^N \sum_k^N \lambda_0^j \lambda_0^k \frac{1}{2} \mathbb{E}[(Z(\boldsymbol{x}_j) - Z(\boldsymbol{x}_k))^2] + 2\sum_i^N \lambda_0^j \frac{1}{2} \mathbb{E}[(Z(\boldsymbol{x}_j) - Z(\boldsymbol{x}_0))^2] \\
&= \sum_j^N \sum_k^N \lambda_0^j \lambda_0^k \gamma(|\boldsymbol{x}_j - \boldsymbol{x}_k|) + 2\sum_i^N \lambda_0^j \gamma(|\boldsymbol{x}_j - \boldsymbol{x}_0|)
\end{aligned} \tag{2.40}$$

e minimizzando usando i moltiplicatori di Lagrange (Lichtenstern [2013]) si ottiene il seguente sistema matriciale:

$$\begin{pmatrix} \ddots & \cdots & \cdots & 1 \\ \vdots & \gamma(|\boldsymbol{x}_m - \boldsymbol{x}_l|) & \cdots & 1 \\ \vdots & \vdots & \ddots & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} \vdots \\ \lambda_0^m \\ \vdots \\ \nu \end{pmatrix} = \begin{pmatrix} \vdots \\ \gamma(|\boldsymbol{x}_0 - \boldsymbol{x}_l|) \\ \vdots \\ 1 \end{pmatrix}. \tag{2.41}$$

dove ν è il moltiplicatore di Lagrange.

2.3 Teoria dell'informazione in geostatistica

Lo scopo di questa sezione è di introduzione alla teoria dell'informazione atta a definire il concetto di informazione mutua, della sua interpretazione di misura di non linearità e della sua implementazione nella geostatistica.

La teoria dell'informazione affonda le sue radici nel pioneristico lavoro di Claude Shannon del 1948 (A Mathematical Theory of Communication, Shannon [1948]) dove occupandosi di teoria della comunicazione introdusse i concetti di entropia e informazione mutua.

Dato che si è interessati a variabili stocastiche e campi stocastici a valori continui, la seguente trattazione verrà svolta usando le definizioni di entropia e informazione mutua su distribuzioni continue.

2.3.1 Entropia e informazione mutua

L'*entropia* di una variabile stocastica è una funzione scalare della sua distribuzione di probabilità che caratterizza l'impredicibilità delle realizzazioni. Data la variabile stocastica X a valori in $\Omega \subseteq \mathbb{R}$ e la sua funzione di densità di probabilità f_X, la sua espressione è la seguente:

$$H(X) = -\int_{\Omega} f_X(x) \log f_X(x) dx \tag{2.42}$$

dove il logaritmo è in base esponenziale e di conseguenza l'entropia è misurata in *nats* (natural units). Se cambia la base del logaritmo, cambia anche l'unità di misura: ad esempio in base 2 si ha come unità misura i *bits* (binary units) più comunemente usata in teoria dell'informazione. In letteratura (Thomas M. Cover [1991]) questa quantità $H(X)$ viene chiamata anche entropia differenziale. L'entropia $H(X)$ è sempre non negativa: infatti la funzione di densità di probabilità è a valori nell'intervallo [0,1), quindi l'argomento dell'integrale è sempre negativo (avendo considerato che $0 \log 0 = 0$ per estensione per continuità della funzione $t \log t$ in $t = 0$) e con un segno meno davanti si ha la non negatività.

L'*entropia congiunta* è l'entropia di una distribuzione di probabilità congiunta (multivariata) oppure di una variabile stocastica a più valori. Data una coppia di variabili X, Y e la loro distribuzione di probabilità congiunta f_{XY}, l'entropia congiunta ha la seguente espressione:

$$H(X;Y) = -\int_{\Omega_X} \int_{\Omega_Y} f_{XY}(x,y) \log f_{XY}(x,y) dxdy. \tag{2.43}$$

L'*informazione mutua*, (oppure chiamata *guadagno di informazione*), è una quantità che misura quanta informazione è comunicata in media in una variabile stocastica da un'altra. In altre parole, misura l'incertezza di una variabile rimossa dalla conoscenza dell'altra variabile. In altri termini ancora, è la misura di informazione ottenuta su una variabile conoscendo l'altra variabile. Date due variabili X e Y, loro rispettive funzioni di densità di probabilità f_X, f_Y rispettivamente, e la congiunta f_{XY}, si hanno due definizioni equivalenti (Thomas M. Cover [1991]):

$$I(X;Y) = H(X) + H(Y) - H(X;Y); \tag{2.44}$$

$$I(X;Y) = \int_{\Omega_X} \int_{\Omega_Y} f_{XY}(x,y) \log \left(\frac{f_{XY}(x,y)}{f_X(x) f_Y(y)} \right) dxdy. \tag{2.45}$$

Le due variabili stocastiche saranno statisticamente indipendenti se l'informazione mutua tra loro è nulla: infatti $I(X;Y) = 0$ se e solo se l'argomento del logaritmo è pari a 1, cioè se $f_{XY} = f_X f_Y$, cioè l'indipendenza statistica di X e Y. Questa quantità è simmetrica, cioè $I(X;Y) = I(Y;X)$ ed è sempre non negativa, cioè $I(X;Y) \geq 0$ e vale che $I(X;Y) = +\infty$ se esiste una qualsiasi relazione g tale che $Y = g(X)$ (Brillinger [2004]). L'informazione mutua è invariante per riparametrizzazioni della funzione di densità di probabilità. Ad esempio la lognormale è una riparametrizzazione della normale (Gaussiana) quindi l'informazione mutua rimane invariata rispetto a quella calcolata sulle Gaussiane.

L'informazione mutua, come l'indice di correlazione di Pearson, ha un'importante proprietà: *misura la non linearità tra le due variabili di cui la si calcola*. In altre parole,

l'informazione mutua misura la dipendenza totale della variabile Y dalla variabile X (I. Butera and Ridolfi [2018],Vu [2018]), cioè se esiste una relazione $\mathcal{F}$ tra le variabili tale che $Y = \mathcal{F}(X)$, di qualsiasi tipo. Questo avviene perchè l'informazione mutua misura la distanza dall'indipendenza statistica, ovvero $f_X f_Y = f_{XY}$. In Robin A.A. Ince and Schyns [2017] è riportata una comparazione tra l'indice di correlazione e una versione normalizzata dell'informazione mutua su una serie di "nuvole" di insieme di misure bivariate in modo da verificare la presenza o meno di una qualche funzione che associ una variabile con l'altra. Prima di fare degli esempi concreti, si introduce anche l'*entropia condizionale* (Thomas M. Cover [1991]): si tratta dell'entropia della variabile X conoscendo il valore della variabile Y,

$$H(X|Y) = -\int_{\Omega} f_{XY}(x,y) \log f_{X|Y}(x|y) dx. \tag{2.46}$$

La relazione tra le varie entropie e l'informazione mutua è la seguente:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X;Y) - H(X|Y) - H(Y|X), \end{aligned} \tag{2.47}$$

mentre graficamente si può vedere la relazione tra le quantità esibite in figura 2.1.

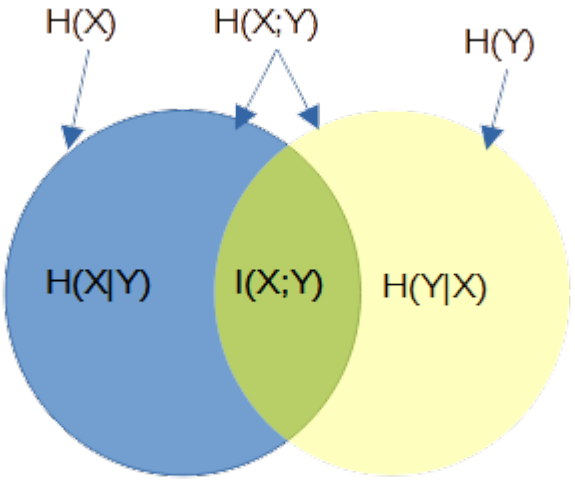

Figura 2.1: Diagramma di Eulero-Venn per l'entropia, entropia congiunta, per l'entropia condizionale e l'informazione mutua.

Per quanto enunciato fino a qui, l'entropia e l'informazione mutua, riguardano le distribuzioni di probabilità (funzioni di densità di probabilità). Per gli esempi partiamo dalla distribuzione Gaussiana o normale già mostrata sopra. L'entropia di una Gaussiana con media μ e deviazione standard σ è pari a

$$H(X_{Gauss}) = \log(\sqrt{2\pi\sigma^2 e}), \tag{2.48}$$

mentre l'informazione mutua di un Gaussiana bivariata risulta essere

$$I(X_{Gauss}; Y_{Gauss}) = -\frac{1}{2}\log(\det(\Sigma)), \tag{2.49}$$

dove Σ è la matrice di covarianza tra le due due variabili X_{Gauss}, Y_{Gauss}. In particolare, se la correlazione tra le due variabili è $\mathrm{Corr}(X_{Gauss}, Y_{Gauss})$ allora l'espressione dell'informazione mutua diventa

$$I(X_{Gauss}; Y_{Gauss}) = -\frac{1}{2}\log(1 - \mathrm{Corr}(X_{Gauss}, Y_{Gauss})). \tag{2.50}$$

Da qui, si ha una *relazione tra una quantità che misura il grado del legame lineare e un'altra quantità che misura il grado del legame totale o non lineare tra le variabili in questione*. Da questa relazione nasce l'idea (Thomas M. Cover [1991]) di normalizzare questo indice:

$$R(X;Y) = \sqrt{1 - \exp(-2I(X;Y))}, \tag{2.51}$$

e che permette di verificare l'equazione

$$R(X_{Gauss}; Y_{Gauss}) = |\mathrm{Corr}(X_{Gauss}, Y_{Gauss})| = |\rho_{X_{Gauss}Y_{Gauss}}|. \tag{2.52}$$

Un altro esempio è la distribuzione Gamma: sia X_{Gamma} distribuita secondo una Gamma di parametri k, ϑ, allora l'entropia risulta essere

$$H(X_{Gamma}) = k + \log\vartheta + \log\Gamma(k) + (1-k)\psi(k) \tag{2.53}$$

dove $\psi(t) = \frac{d}{dt}\log\Gamma(t)$ è la funzione *digamma*. Per quanto riguarda l'espressione dell'informazione mutua, non è possibile ricavarla in forma analitica in funzione dei parametri della distribuzione congiunta, ma con l'ausilio di software per il calcolo analitico di integrali si può verificare che per i paramtetri $a = 1, b = 1, \mu_1 = \mu_2 = 1$ nella distribuzione 2.5, si ha che l'informazione mutua è pari a

$$I(X_{Gamma}; Y_{Gamma}) = 0.2023. \tag{2.54}$$

Per introdurre la teoria dell'informazione in geostatistica, si riporta che l'informazione mutua è già stata applicata su serie temporali in Angeliki Papana [2009], C. Granero-Belinchon [2019], e in via teorica è stata riportata l'idea di implementarla su dati spaziali in Li and Deutsch [2010] senza però vedere dei risultati. Mentre in R. T. Wicks and Dendy [2008] si è applicata lo studio dell'informazione mutua per il rilevamento di correlazione spaziale delle fluttuazioni del vento solare e della dipendenza del ciclo solare.

In questa tesi si cerca di applicare la teoria dell'informazione alla geostatistica, sotto le stesse ipotesi su cui si lavora con la correlazione spaziale (quindi la covarianza spaziale), ovvero la stazionarietà del secondo ordine. In questo modo, data un campo stocastico Z si vuole valutare la quantità

$$I(h) = I(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)), \forall i, j, \tag{2.55}$$

dove $h = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|$, che in questa tesi si propone di denotare con il nome di *informazione mutua spaziale*. Ad esempio, prendendo in considerazione un campo stocastico Gaussiano

Z_{Gauss}, questa quantità risulta essere legata, sfruttando la relazione 2.50, alla correlazione spaziale nel seguente modo:

$$I(h) = I(Z_{gauss}(\boldsymbol{x}_i), Z_{gauss}(\boldsymbol{x}_j)) = \frac{1}{2}\log(1-\rho(h)). \tag{2.56}$$

Proprio come per due variabili stocastiche Gaussiane sussiste la relazione 2.52, anche per un campo stocastico Gaussiano vale

$$R(h) = |\rho(h)|, \tag{2.57}$$

dove $R(h) = \sqrt{1-\exp(-2I(h))}$.

Equivalentemente si potrebbe verificare la relazione

$$P(h) = I(h), \tag{2.58}$$

dove $P(h) = -\frac{1}{2}\log(1-\rho^2(h))$ è il coefficiente di correlazione trasformato nel dominio dell'informazione mutua.

2.3.2 Stima dell'informazione mutua

In una situazione realistica di un problema, si conosce soltanto una realizzazione del campo stocastico che rappresenta la quantità fisica interessata, quindi non si è a conoscenza della funzione di densità di probabilità del campo, ed essendo disponibile un solo campione di dati bisogna stimare l'informazione mutua. Si è già visto come stimare la covarianza spaziale, e quindi la correlazione spaziale, perciò ora si passano in rassegna i vari metodi di stima dell'informazione mutua, a partire da un campione di dati, presenti in letteratura.

Il metodo degli istogrammi

Si consideri ora un campione di misure bivariate $(x_i, y_i)_{i=1}^N$ che rappresentano realizzazioni della coppia di variabili stocastiche (X, Y). Seguendo quanto descritto in Alexander Kraskov and Grassberger [2004] e Paninski [2003], lo stimatore dell'informazione mutua può essere visto come somma delle stime delle singole entropie:

$$\widehat{I}(X;Y) = \widehat{H}(X) + \widehat{H}(Y) + \widehat{H}(X;Y) \tag{2.59}$$

dove

$$\begin{aligned}
\widehat{H}(X) &= -\sum_{i}^{n_b} \widehat{f}_X(i)\log(\widehat{f}_X(i)) \\
\widehat{H}(Y) &= -\sum_{j}^{n_b} \widehat{f}_Y(j)\log(\widehat{f}_Y(j)) \\
\widehat{H}(X;Y) &= -\sum_{i}^{n_b}\sum_{j}^{n_b} \widehat{f}_{XY}(i,j)\log(\widehat{f}_{XY}(i,j))
\end{aligned} \tag{2.60}$$

con n_b uguale al numero di bins usati per la discretizzazione dell'integrale. Le stime delle distribuzioni di probabilità discretizzate sono fatte mediante il conteggio tipo istogramma, quindi

$$\begin{aligned} \widehat{f}_X(i) &= \frac{n_X(i)}{N} \\ \widehat{f}_Y(j) &= \frac{n_Y(j)}{N} \\ \widehat{f}_{XY}(i,j) &= \frac{n_{XY}(i,j)}{N^2}, \end{aligned} \tag{2.61}$$

dove $n_X(i)$ è il numero di punto x_i che cadono all'interno i-esimo bin, e così via per $n_Y(j)$ e $n_{XY}(i,j)$. In Meyer [2008] è stata proposta una versione migliorata rende lo stimatore non distorto (unbiased). In generale non c'è una strategia per scegliere il parametro libero n_b, cioè il numero di bins per gli istogrammi, però si nota che al crescere dei numero di bins, la stima dell'informazione mutua cresce.

Il metodo del Kernel Density Estimator (KDE)

Partendo da quanto fatto in Y. Moon [1995] e dall'idea che l'informazione mutua può essere vista come

$$I(X;Y) = \mathbb{E}\left[\log \frac{f_{XY}}{f_X f_Y}\right] \tag{2.62}$$

e che la sua approssimazione naturale è quello dello stimatore media su N campioni, il Kernel Density Estimator viene definito nel seguente modo:

$$\widehat{I}^{KDE}(X;Y) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\widehat{f}_{XY}(x_i,y_i)}{\widehat{f}_X(x_i)\widehat{f}_Y(y_i)}\right), \tag{2.63}$$

dove le funzioni di densità di probabilità sono determinate dalle seguenti espressioni

$$\begin{aligned} \widehat{f}_{XY}(x,y) &= \frac{1}{N}\sum_{i=1}^{N} K(u) \\ K(u) &= \frac{1}{h^2\sqrt{(2\pi)^2\det(\Sigma_{XY})}}\exp\left(-\frac{u}{2}\right) \\ u &= \frac{([x,y]-[x_i,y_i])\Sigma_{XY}^{-1}([x,y]-[x_i,y_i])^T}{h^2} \end{aligned} \tag{2.64}$$

$$\begin{aligned} \widehat{f}_X(x) &= \frac{1}{N}\sum_{i=1}^{N} K(u) \\ K(u) &= \frac{1}{h^2\sqrt{(2\pi)^2\sigma_X}}\exp\left(-\frac{u}{2}\right) \\ u &= \frac{(x-x_i)\sigma_X^{-1}((x-x_i)}{h^2} \end{aligned} \tag{2.65}$$

$$\begin{aligned} \widehat{f}_Y(y) &= \frac{1}{N}\sum_{i=1}^{N} K(u) \\ K(u) &= \frac{1}{h^2\sqrt{(2\pi)^2\sigma_Y}}\exp\left(-\frac{u}{2}\right) \\ u &= \frac{(y-y_i)\sigma_X^{-1}((y-y_i)}{h^2}. \end{aligned} \tag{2.66}$$

La funzione K viene detta funzione *kernel* e in questo caso è stata scelta del tipo Gaussiano. Il parametro libero di questo stimatore è h, in generale non c'è una scelta ottimale di quest'ultimo tranne che nel caso che X, Y siano distribuite Gaussianamente: allora si ha

$$h_{ott} = .\left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{-\frac{1}{d+4}} \tag{2.67}$$

con $d = 2$.

Il metodo del k-nearest neighbours (KNN) di Kozachenko-Leonenko (KL)

Ora si presenta il primo metodo di stima basato sul metodo k-nearest neighbor (i primi k punti più vicini). Seguendo quanto mostrato in Kozachenko LF [1987] e rivisto in Alexander Kraskov and Grassberger [2004], si ottiene uno stimatore dell'informazione mutua come somma delle stime dell'entropie:

$$\widehat{I}^{KL}(X;Y) = \widehat{H}^{KL}(X) + \widehat{H}^{KL}(Y) + \widehat{H}^{KL}(X;Y). \tag{2.68}$$

Lo stimatore dell'entropia marginale $H(X)$ è così definito:

$$\widehat{H}^{KL}(X) = \psi(N) - \psi(k) + \frac{1}{N}\sum_{i=1}^{N}\log\epsilon_i \tag{2.69}$$

dove ϵ_i è due volte la distanza (indotta dalla norma del massimo) tra il k^{th} punto dell'insieme $\{x_j\}_{j=1}^N$ dal punto x_i e dove la funzione ψ è la funzione digamma precedentemente descritta. In modo analogo si definisce la stima dell'entropia marginale di Y:

$$\widehat{H}^{KL}(Y) = \psi(N) - \psi(k) + \frac{1}{N}\sum_{j=1}^{N}\log\epsilon_j. \tag{2.70}$$

L'entropia congiunta viene stimata secondo

$$\widehat{H}^{KL}(X;Y) = \psi(N) - \psi(k) + \frac{2}{N}\sum_{m=1}^{N}\log\epsilon_m \tag{2.71}$$

dove ϵ_m è due volte la distanza (indotta dalla norma del massimo) tra il k^{th} punto dell'insieme $\{z_i\}_{i=1}^N = \{x_i, y_i\}_{i=1}^N$ dal punto z_m. Il parametro libero di questo stimatore è k: non sembra esserci una scelta ottimale per questo stimatore, ma in base alla letteratura si vede che si comporta bene per valori bassi di k. Di questo stimatore non è garantita la positività, ma questo può essere vista come una qualità: infatti se l'informazione mutua stimata da valori negativi attorno allo zero, allora si può considerare l'informazione mutua nulla, altrimenti sarebbe impossibile misurare la nullità di questa quantità.

Il metodo del k-nearest neighbours (KNN) di Kraskov-Stogbauer-Grassberger (KSG)

Con un ragionamento simile a quello con cui è stato costruito lo stimatore di Kozachenko-Leonenko, si sono costruiti due stimatori dagli autori in Alexander Kraskov and Grassberger [2004]. Il primo stimatore è descritto dalla seguente formula:

$$\widehat{I}^{KSG}(X;Y) = \psi(k) + \psi(N) - \langle \psi(n_x + 1) + \psi(n_y + 1) \rangle \tag{2.72}$$

dove $\langle \cdot \rangle$ rappresenta l'operatore media che agisce su n_x che rappresenta il numero di punti x_j la quale distanza da x_i è strettamente minore della distanza $\epsilon_i/2$ che a sua volta è la distanza tra z_i dal k-esimo punto.

Per la scelta del parametro non c'è una strategia, ma gli stessi autori suggeriscono $k = 2,3,4,5$, ma in generale lo stimatore sembra poco sensibile alla scelta del parametro come si vedrà nel prossimo capitolo. Anche in questo stimatore non vi è garantita la non negatività, ma può essere una qualità infatti se lo stimatore desse sempre valori positivi allora sarebbe impossibile misurare l'informazione mutua nulla. In Weihao Gao [2016] suggeriscono un nuovo metodo basato sul $\widehat{I}^{KSG}$ con una correzione per la non distorsione $\widehat{I}^{BIKSG}$. In Alexander Kraskov and Grassberger [2004] dicono che la fonte di errore nel calcolo dell'informazione mutua con il KNN risiede nella non uniformità locale della densità di probabilità, quindi è stata presentata una versione che corregge questo problema in Shuyang Gao [2015].

2.4 Legame non lineare tra due variabili stocastiche

Si è già visto che per un campo stocastico Gaussiano, non esiste contributo non lineare nello spazio. D'altra parte per una coppia di variabili distribuite secondo una Gamma e correlate, si è già mostrata l'informazione mutua per determinati parametri: ebbene, dato che la correlazione di una bivariata Gamma distribuita secondo l'espressione 2.5 è pari a

$$\mathrm{Corr}(X_{Gamma}; Y_{Gamma}) = \frac{a}{a+b+1} = \frac{1}{3} \approx 0.3333, \tag{2.73}$$

avendo scelto i parametri $a = b = 1$, e dato che il coefficiente R, cioè l'informazione mutua trasformata nel dominio della correlazione, è pari

$$R(X_{Gamma}; Y_{Gamma}) = \sqrt{1 - \exp(-2(0.2023)} = 0.5769, \tag{2.74}$$

allora si verifica che

$$R(X_{Gamma}; Y_{Gamma}) > |\mathrm{Corr}(X_{Gamma}; Y_{Gamma})| \tag{2.75}$$

che sta ad indicare che esiste un legame non lineare tra le due variabili stocastiche X e Y. Per valutare il contributo del legame non lineare in spazio di un campo stocastico, a rigore di logica, basterebbe costruire un campo stocastico distribuito secondo una Gamma, ad esempio seguendo quanto mostrato in J. Liou [2011].

Capitolo 3

Studio su dati sintetici

In questo capitolo si affronta uno studio dei metodi presentati nel capitolo precedente su dati sintetici, cioè generati sinteticamente e non reali, in modo da testare le performatività e sensitività dei vari metodi sia per quanto riguarda la geostatistica lineare che quella non lineare partendo dallo studio sull'informazione mutua.

Prima di mostrare i metodi di calcolo delle quantità sopraesposte, si introducono i metodi per generare le variabili stocastiche e i campi stocastici.

3.1 Genearazione di campi stocastici Gaussiani con correlazione prescritta

I metodi per la generazione di campi stocastici sono ben descritti nella recente review di Yang Liu1 [2019].

3.1.1 Il metodo di decomposizione di Choleski

Il metodo di decomposizione di Choleski è uno dei metodi più semplici da implementare per la generazione di campi stocastici. Si suppone di voler generare una realizzazione del campo stocastico Z in N punti, ad esempio di una griglia bidimensionale, $\{x_i, y_j\}_{ij}$. Un campo stocastico debolmente stazionario, Gaussiano ad esempio, è definito dalla matrice di covarianza R

$$R_{ij} = \mathrm{Cov}(Z(\boldsymbol{x}_i), Z(\boldsymbol{x}_j)) = C(|\boldsymbol{x}_i - \boldsymbol{x}_j|), \tag{3.1}$$

dove la covarianza spaziale deve essere impostata; ad esempio può avere l'espressione esponenziale

$$C(h) = \sigma^2 \exp\left(-\frac{h}{\lambda}\right), \tag{3.2}$$

dove σ^2 è la varianza del campo stocastico e λ la sua *lunghezza di correlazione*. La matrice R, definita positiva, deve essere dunque decomposta con il metodo Choleski

$$R = LU \tag{3.3}$$

dove $L = U^T$ è una matrice triangolare. Il prossimo passo è quello di generare N realizzazioni di una variabile stocastica Gaussiana standard θ, cioè con media nulla e varianza unitaria. In particolare, si possono denotare le N realizzazioni come N variabili stocastiche scorrelate tale $\mathbb{E}[\theta_i] = 0$ e $\mathbb{E}[\theta_i\theta_j] = \delta_{ij}$. Quindi gli N valori del campo stocastico sono descritti dalla moltiplicazione matrice-vettore

$$z = L\theta. \tag{3.4}$$

Questa realizzazione del campo stocastico ha media nulla, infatti la media è un operatore lineare, e matrice di covarianza pari a R, infatti vale

$$\mathrm{Cov}(z) = \mathrm{Cov}(L\theta) = \mathbb{E}[L\theta\theta^T L^T] = L\mathbb{E}[\theta\theta^T]L^T = LIL^T = LU = R. \tag{3.5}$$

Il costo computazionale del metodo è $O(N^3)$ per la decomposizione LU, mentre il costo della moltiplicazione matrice-vettore è $O(N^2)$. Quindi per grandi N il metodo richiede una quantità notevole di memoria (per lo storage della matrice di covarianza R) e di tempo di calcolo.

3.2 Studio dei metodi della geostatistica lineare

Per il calcolo del semivariogramma, della covarianza spaziale e della correlazione spaziale si parte esponendo l'**algoritmo 1**, dati un set di N misure $\{x_i, y_i, Z_i\}_{i=1}^{N}$.

Algorithm 1 Semivariogramma, covarianza spaziale e correlazione spaziale

Input: x, y, Z, n_{classi}
Output: $h, \gamma, C, \rho, n^{coppie}$
$\mu \leftarrow \sum_i^n Z_i/N$
$d_{ij} \leftarrow \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$
$h_k \leftarrow k d_{\max}/n_{classi}$
$\epsilon \leftarrow d_{\max}/(n_{classi})$
(k, i, j) tali che $h_k - \epsilon < d_{ij} < h_k + \epsilon$
$n_k^{coppie} \leftarrow \#(i, j)$ tali che $h_k - \epsilon < d_{ij} < h_k + \epsilon$
$g_{ij,k} \leftarrow (Z_i - Z_j)^2$ con (i, j) tali che $h_k - \epsilon < d_{ij} < h_k + \epsilon$
$c_{ij,k} \leftarrow (Z_i - \mu)(Z_j - \mu)$ con (i, j) tali che $h_k - \epsilon < d_{ij} < h_k + \epsilon$
$\gamma_k \leftarrow \sum_{i,j} g_{ij,k}/2n_k^{coppie}$
$C_k \leftarrow \sum_{i,j} c_{ij,k}/n_k^{coppie}$
$\rho_k \leftarrow C_k/C_0$

Il set up di queste simulazioni è il seguente: si è usato il metodo di Choleski per generare una realizzazione del campo stocastico Z, con una covarianza di tipo esponenziale con parametro $\lambda = 1$, ovvero lunghezza di correlazione unitaria, e varianza pari a $\sigma^2 = 5$. Il dominio del campo è $\Omega = [0{,}50]X[0{,}50]$ dove il numero di punti della realizzazione è pari a $N = 100X100 = 10^4$. I risultati della stima del semivariogramma, della covarianza spaziale e della correlazione spaziale sono presenti in figura 3.1.

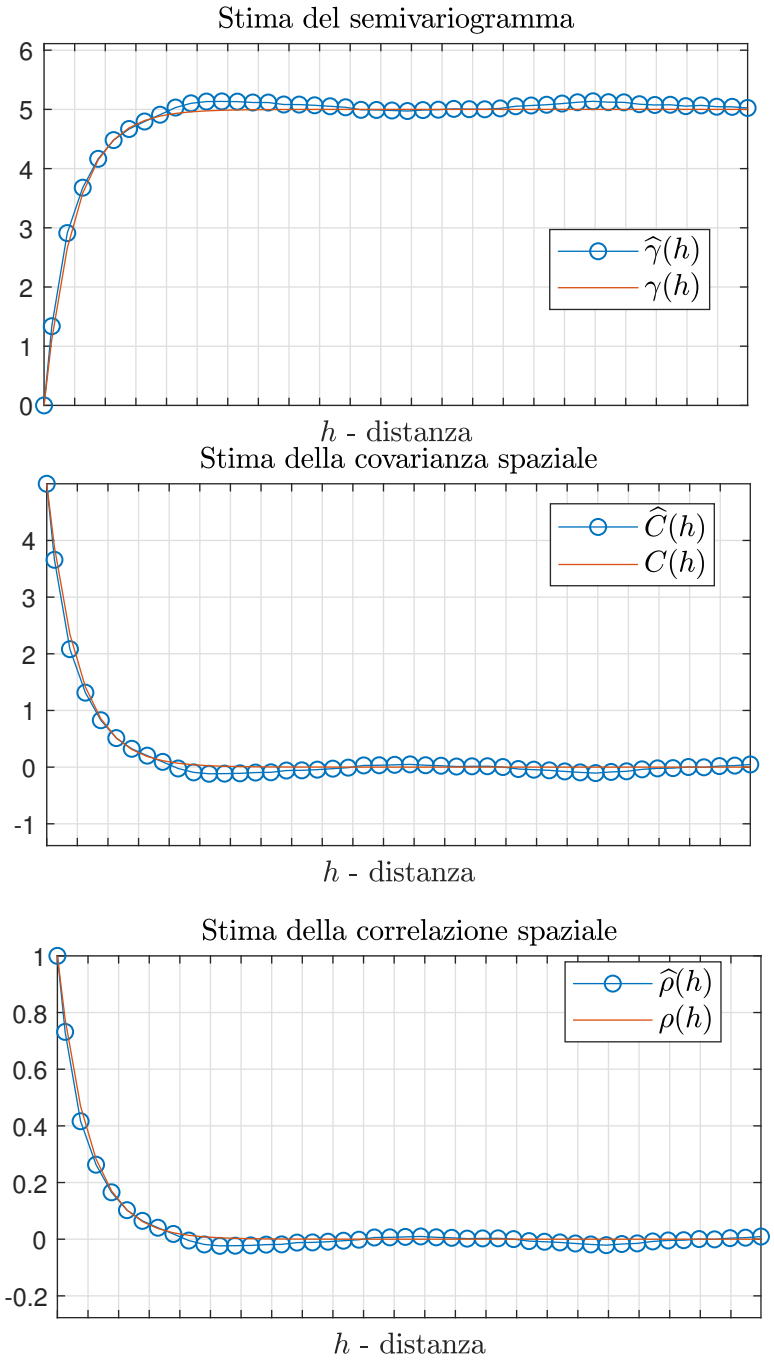

Figura 3.1: In tutte e tre le figure, le linee verticali segnano le distanze di una lunghezza di correlazione. La curva con i pallini blu rappresenta la stima della quantità, mentre la curva rossa rappresenta la quantità teorica.

3.3 Analisi dei metodi di stima dell'informazione mutua

Nelle prossime sottosezioni si valuterà l'efficacia, in termini di accuratezza, al variare del parametro libero della numerosità dei vettori a cui vengono applicati questi metodi. Il set up di queste simulazioni è simile per tutti i metodi: si tratta di valutare l'informazione mutua $\widehat{I}(X_{Gauss}, Y_{Gauss})$ dove X_{Gauss}, Y_{Gauss} sono due vettori distribuiti secondo una Gaussiana bivariata con correlazione impostata ρ. In particolare si calcola anche l'indice R perchè si conosce il suo valore teorico, che corrisponde alla correlazione impostata ρ.

3.3.1 Il metodo degli istogrammi (HIST)

Il metodo degli istogrammi è applicato per stimare l'indice R per vari valori di correlazione, in particolare per tre valori significativi: alta correlazione $\rho = 0.9$, media correlazione $\rho = 0.5$ e bassa o nulla correlazione $\rho = 0.01$. In figura 3.2 sono rappresentate le curve di stima in funzione della numerosità dei vettori per vari valori di numero di bins. Si vede che il metodo è accurato per $n_{bins} \approx 15$ e per altissime numerosità. *Il parametro ottimale che si rileva è* $n_{bins} = 15$ *per alte numerosità.*

3.3.2 Il metodo del Kernel Density Estimator (KDE)

Il metodo del Kernel Density Estimator è applicato per stimare l'indice R per vari valori di correlazione, in particolare per tre valori significativi: alta correlazione $\rho = 0.9$, media correlazione $\rho = 0.5$ e bassa o nulla correlazione $\rho = 0.01$. In figura 3.3 sono rappresentate le curve di stima in funzione della numerosità dei vettori. Si vede che il metodo non è accurato per bassi valori di correlazione. Inoltre, il metodo è computazionalmente costoso per alti valori di numerosità.

3.3.3 Il metodo di Kozachenko-Leonenko (KL)

Il metodo di Kozachenko-Leonenko è applicato per stimare l'indice R per vari valori di correlazione, in particolare per tre valori significativi: alta correlazione $\rho = 0.9$, media correlazione $\rho = 0.5$ e bassa o nulla correlazione $\rho = 0.01$. In figura 3.4 sono rappresentate le curve di stima in funzione della numerosità dei vettori per vari valori di k. Si vede che il metodo non è accurato per bassi valori di correlazione. Il metodo è computazionalmente efficiente per alti valori di numerosità. L'accuratezza è poco sensibile rispetto al parametro k.

3.3.4 Il metodo di Kraskov-Stogbauer-Grassberger (KSG)

Il metodo di Kozachenko-Leonenko è applicato per stimare l'indice R per vari valori di correlazione, in particolare per tre valori significativi: alta correlazione $\rho = 0.9$, media correlazione $\rho = 0.5$ e bassa o nulla correlazione $\rho = 0.01$. In figura 3.5 sono rappresentate le curve di stima in funzione della numerosità dei vettori per vari valori di k. Si vede che il metodo è abbastanza accurato per bassi valori di correlazione. Il metodo è non

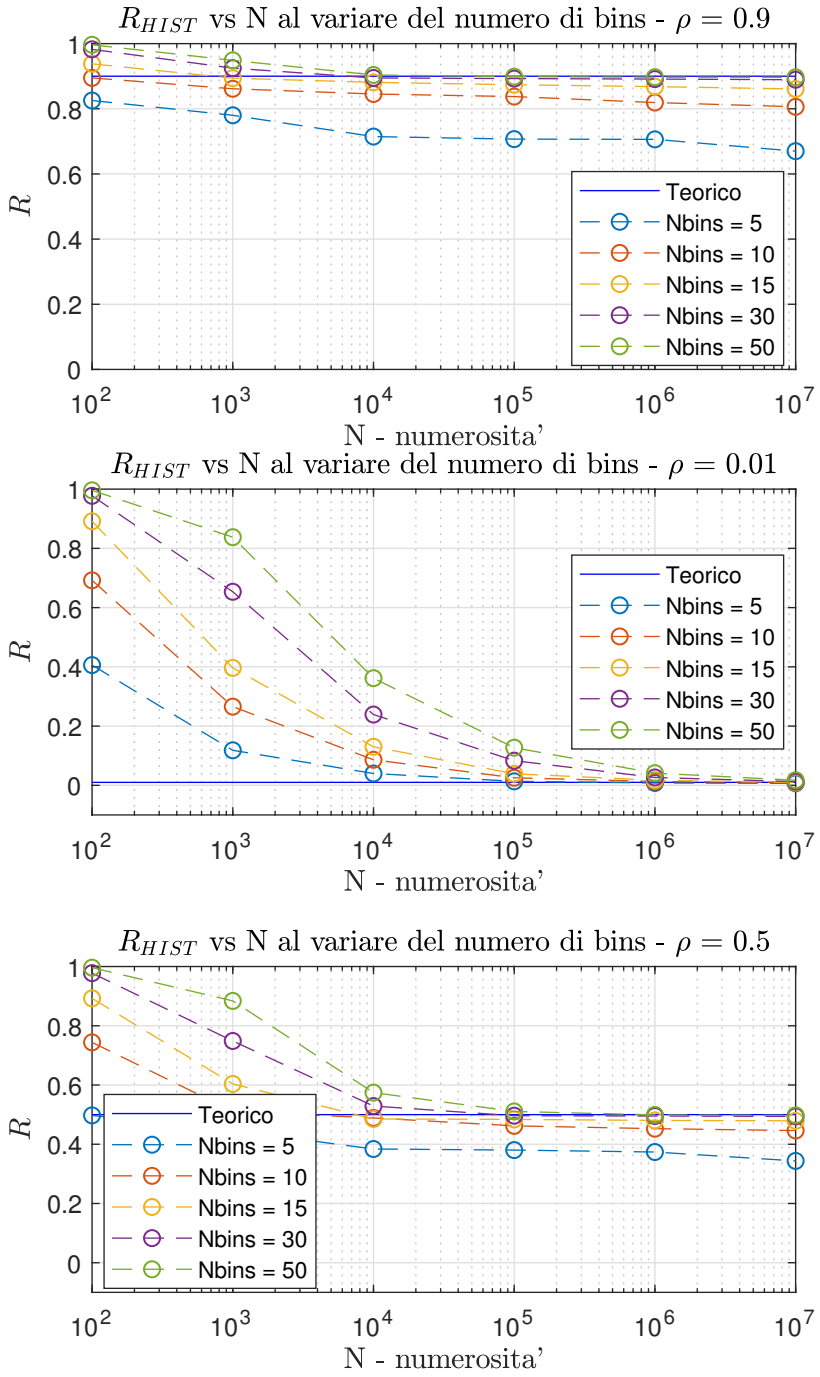

Figura 3.2: Indice R calcolato con il metodo degli istogrammi al variare della numerosità dei due vettori su cui è calcolato e al variare del parametro libero, cioè il numero di bins.

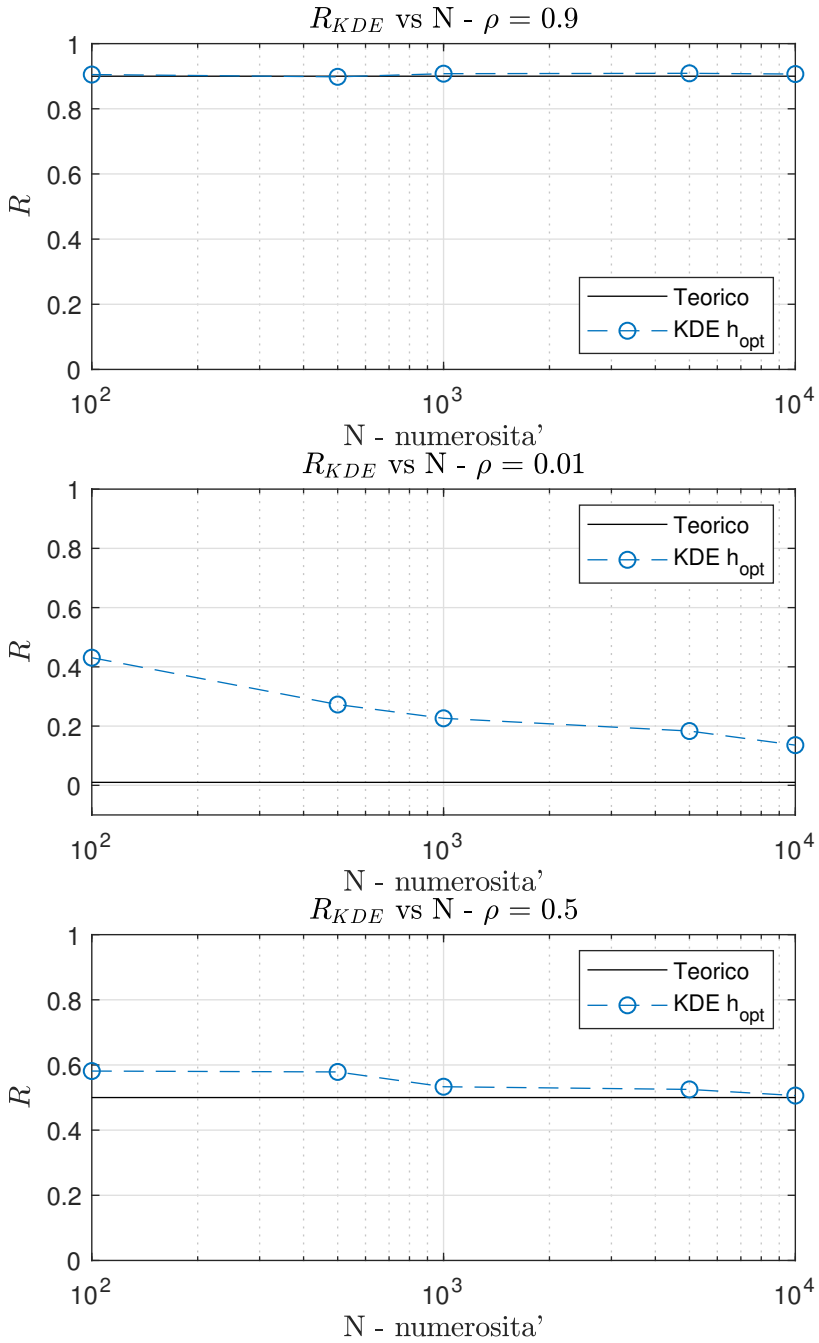

Figura 3.3: Indice R calcolato con il metodo del Kernel Density Estimator al variare della numerosità dei due vettori su cui è calcolato e usando il parametro ottimale visto nel capitolo 2.

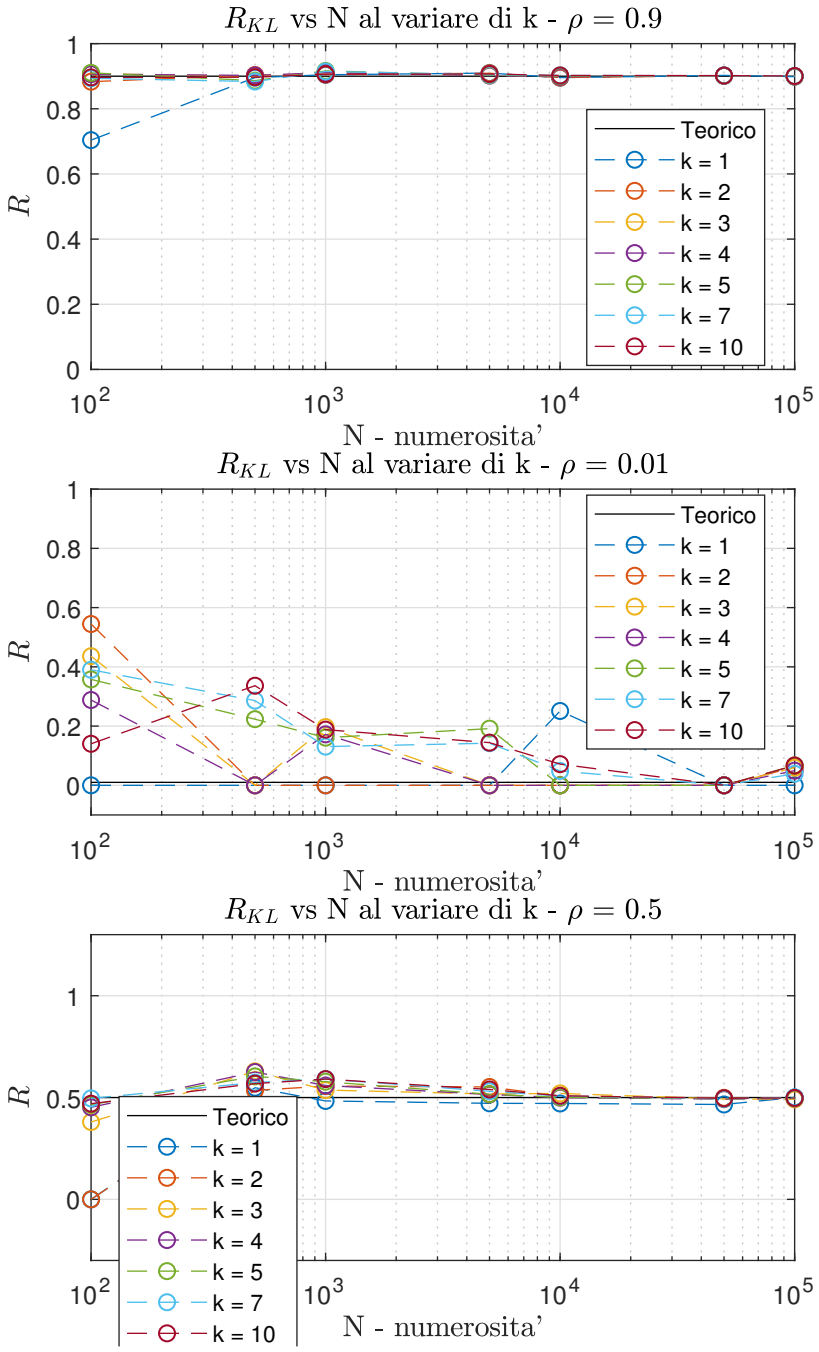

Figura 3.4: Indice R calcolato con il metodo di Kozachenko-Leonenko al variare della numerosità dei due vettori su cui è calcolato e al variare del parametro libero, cioè il numero k di nearest neighbor.

computazionalmente efficiente per alti valori di numerosità. L'accuratezza è poco sensibile rispetto al parametro k.

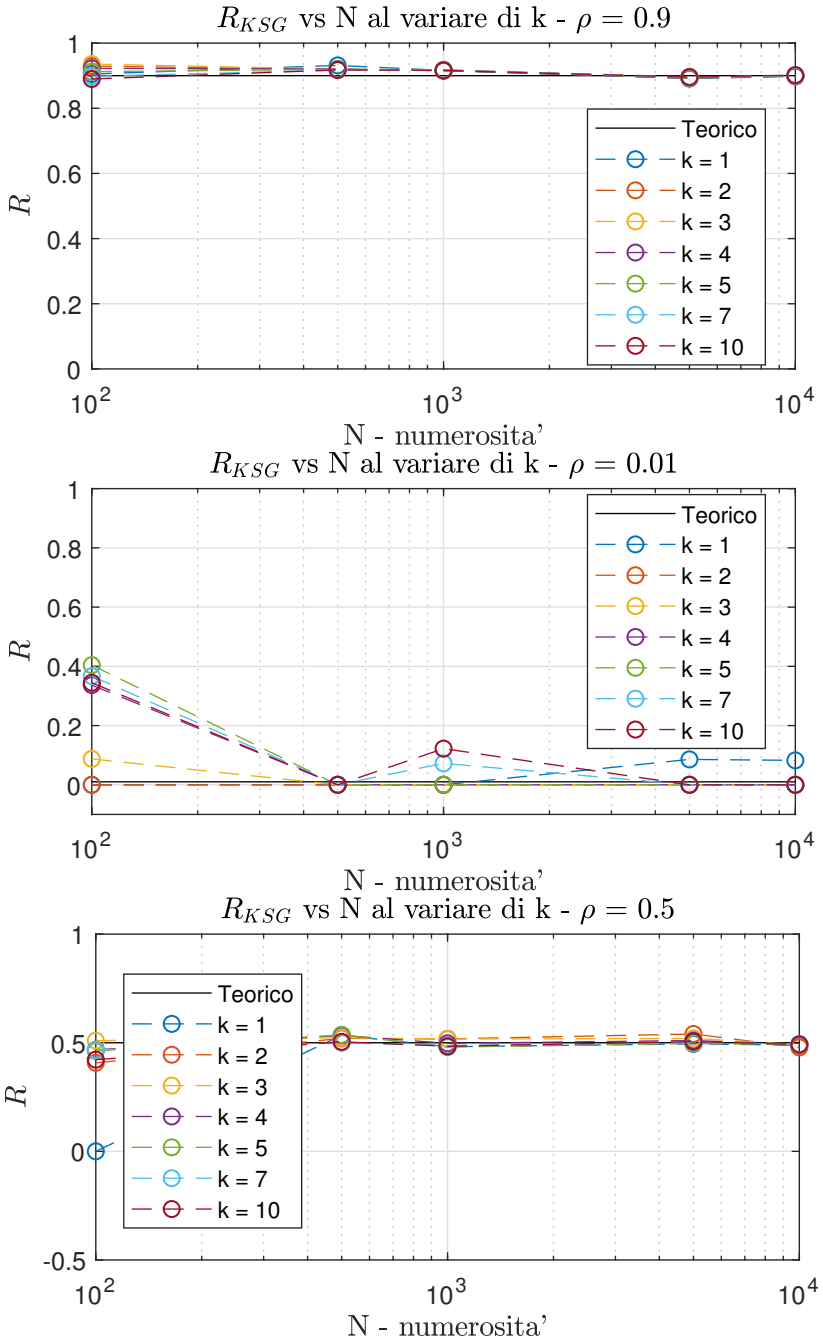

Figura 3.5: Indice R calcolato con il metodo di Kraskov-Stogbauer-Grassberger al variare della numerosità dei due vettori su cui è calcolato e al variare del parametro libero, cioè il numero k di nearest neighbor.

3.3.5 Comparativa fra i vari metodi

Dati i limiti di applicabilità dei metodi in funzione della numerosità, si è scelto di testarli su un campione di 5000 numeri e compararli su una curva che mette in relazione R e ρ. Si vede quindi che il KDE e il metodo degli istogrammmi funzionano male su vettori poco

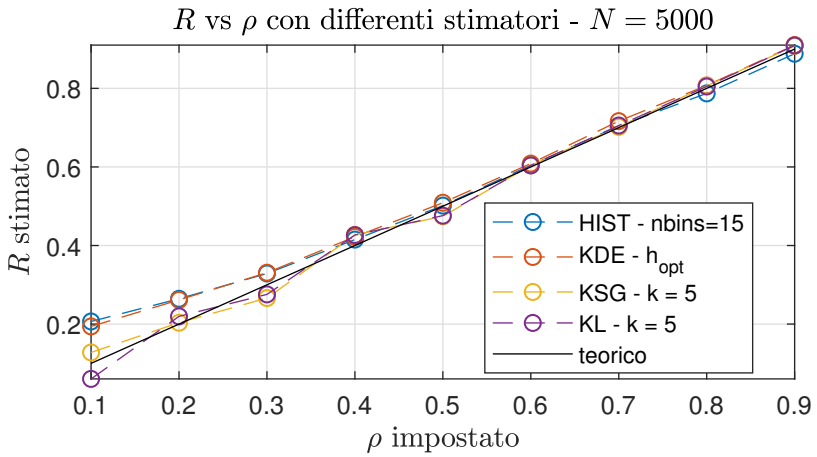

Figura 3.6: Grafico che confronta l'indice R con la correlazione impostata per i vari metodi.

numerosi, mentre il KL e il KSG si avvicinano al valore di riferimento.

3.3.6 Velocità di calcolo dei vari metodi per la stima della mutual information

Infine si valuta la performance di velocità degli algoritmi per il calcolo dell'informazione mutua. I calcoli sono stati eseguiti con una coppia di vettori di lunghezza $1 \cdot 10^4$.

Metodo	Secondi
Istogrammi (HIST)	0.1157
Kozachenko-Leonenko (KL)	1.2509
Kraskov (KSG)	32.0840
Kernel Density Estimator (KDE)	100.5829

Tabella 3.1: Velocità di calcolo per i vari metodi.

3.3.7 Stima dell'informazione mutua spaziale

La stima dell'informazione mutua spaziale è svolta secondo l'**algoritmo 2** sottostante. L'indice R è calcolato per tutti e quattro i metodi a disposizione. Nella pratica, le routines KSGmi e KDEmi sono applicate alle prime 1000 coppie di punti per questioni di efficienza computazionale, mentre le routines HISTmi e KLmi sono applicate a tutte le coppie di punti in quanto sono in grado di gestire numerosità anche molto elevato.

Algorithm 2 Informazione mutua e indice R

Input: x, y, Z, n_{classi}
Output: $h, \widehat{I}^{HIST}, \widehat{I}^{KDE}, \widehat{I}^{KL}, \widehat{I}^{KSG}, \widehat{R}^{HIST}, \widehat{R}^{KDE}, \widehat{R}^{KL}, \widehat{R}^{KSG}, n^{coppie}$
$d_{ij} \leftarrow \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$
$h_k \leftarrow k d_{\max}/n_{classi}$
$\epsilon \leftarrow d_{\max}/(n_{classi})$
(k, i, j) tali che $h_k - \epsilon < d_{ij} < h_k + \epsilon$
$n_k^{coppie} \leftarrow \#(i, j)$ tali che $h_k - \epsilon < d_{ij} < h_k + \epsilon$
$A_{i,k} \leftarrow Z_i, B_{j,k} \leftarrow Z_j$ con (i, j) tali che $h_k - \epsilon < d_{ij} < h_k + \epsilon$
$\widehat{I}_k^{HIST} \leftarrow \text{HISTmi}(A_k, B_k)$
$\widehat{I}_k^{KDE} \leftarrow \text{KDEmi}(A_k, B_k)$
$\widehat{I}_k^{KL} \leftarrow \text{KLmi}(A_k, B_k)$
$\widehat{I}_k^{KSG} \leftarrow \text{KSGmi}(A_k, B_k)$
$\widehat{R}_k^{HIST} \leftarrow \sqrt{1 - \exp(-2\widehat{I}_k^{HIST})}$
$\widehat{R}_k^{KDE} \leftarrow \sqrt{1 - \exp(-2\widehat{I}_k^{KDE})}$
$\widehat{R}_k^{KL} \leftarrow \sqrt{1 - \exp(-2\widehat{I}_k^{KL})}$
$\widehat{R}_k^{KSG} \leftarrow \sqrt{1 - \exp(-2\widehat{I}_k^{KSG})}$

Il set up di questa simulazione è lo stesso della sezione precedente dove si è calcolato il semivariogramma e la correlazione spaziale. In figura 3.7 si comparano i valori dell'indice R e dell'informazione mutua con i valori teorici e con i valori di correlazione e indice P rispettivamente. Questi risultati mostrano che il metodo degli istogrammi (HIST) funziona bene, cioè è il più accurato tra i metodi avendo scelto il parametro ottimale per Gaussiane, mentre gli altri metodi presentano un offset per quanto riguarda l'indice R. Questo fenomeno è dovuto al fatto che i metodi KSG e KDE presentano un piccolo bias nella stima dell'informazione mutua e quando viene trasformato non linearmente nell'indice R questi errori vengono amplificati come si può intuire dalla figura 3.8 per valori bassi di I. Per questo motivo si grafica anche il valore di informazione mutua comparandola con la correlazione spaziale $P(h)$ trasformata nel dominio di $I(h)$. In quest'ultimo plot gli errori sono meno importanti rispetto alla scala di informazione mutua massima. Per quanto riguarda il metodo KL, si nota dai calcoli che non riesce a stimare le singole entropie nel

suo algoritmo quindi la routine restituisce il valore nullo, che però non è rappresentativo della stima.

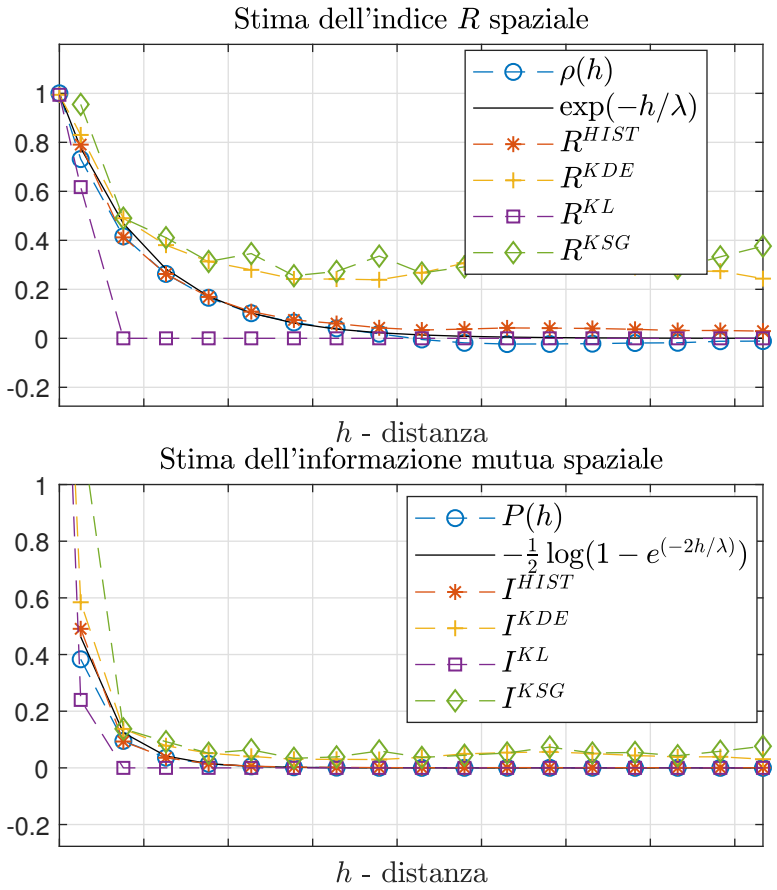

Figura 3.7: Sopra: grafico dell'indice R spaziale confrontata con la funzione teorica e la correlazione spaziale calcolata per vari metodi. Sotto: grafico dell'informazione mutua spaziale confrontata con la funzione teorica e la correlazione spaziale trasformata nel domnio di I, calcolata per vari metodi.

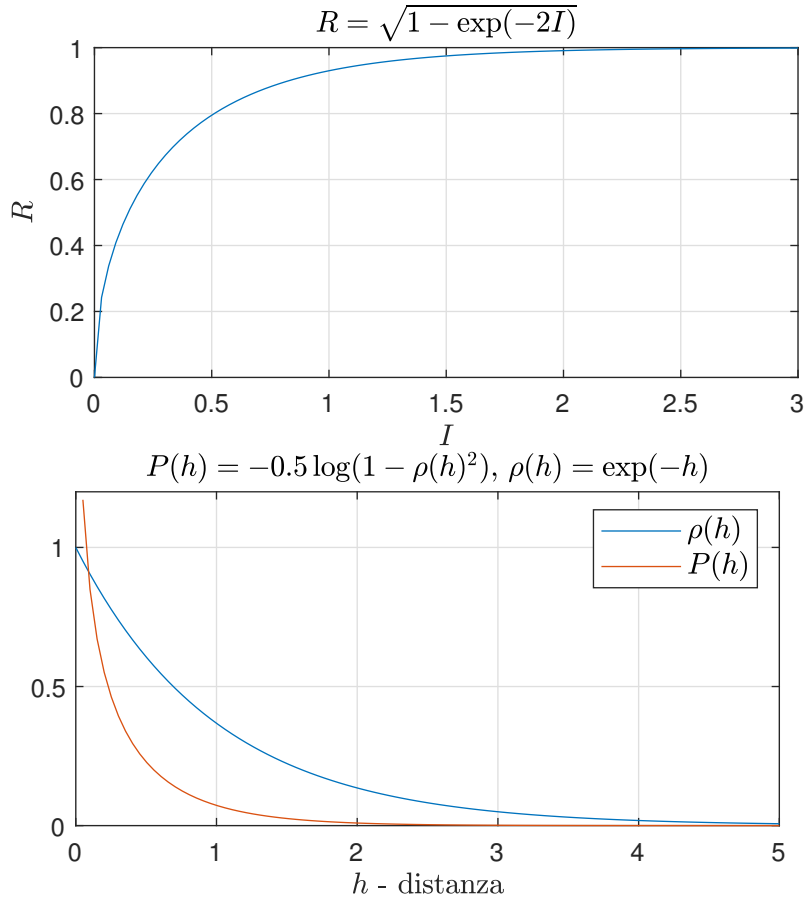

Figura 3.8: Sopra: trasformazione non lineare dell"informazione mutua nell'indice R. Sotto: confronto tra la correlazione spaziale e la correlazione spaziale trasformata nel dominio dell'informazione mutua.

3.3.8 Stima dell'informazione mutua su una coppia di variabili non Gaussiane

Si è già visto nella precedente sezione che i vari metodi di stima, eccetto il KL, si comportano bene su un campo stocastico Gaussiano. In un caso reale, si può avere a disposizione un insieme di misure che si distribuiscono secondo una funzione di densità di probabilità non Gaussiana, come ad esempio la distribuzione Gamma (J. Liou [2011]). Per questo motivo è opportuno scegliere un metodo di stima dell'informazione mutua che stimi correttamente anche in caso di distribuzione non Gaussiane. Inoltre, sinora, si è vista soltanto una *distribuzione "lineare"*, cioè che soddisfa $R(h) = \rho(h)$ e che quindi il cui campo stocastico non ha legami non lineari nello spazio. E' opportuno quindi valutare lo strumento di stima dell'informazione mutua su distribuzioni non lineari, cioè per cui $R > \rho$. Grazie all'esempio fatto nel precedente capitolo, in particolare nella sezione **2.4**, si può testare che gli algoritmi per verificare che stimino correttamente le quantità I e R. Seguendo

quanto svolto in Nadarajah and Kotz [2009], si riesce a generare un campione di dati bivariati distribuiti secondo una Gamma bivariata (2.5) per i quali si conosce il valore dell'informazione mutua e i risultati su dei vettori di lunghezza 5000, sono riportati in tabella 3.2.

Metodo	$\widehat{I}$	$\widehat{R}$
Istogrammi (HIST) $n_{bins} = 15$	0.0913	0.4086
Kozachenko-Leonenko (KL) $k = 5$	0.2056	0.5806
Kraskov (KSG) $k = 5$	0.2107	0.5864
Kernel Density Estimator (KDE) h_{ott}	0.1636	0.5282
Valori teorici	0.2023	0.5769

Tabella 3.2: Accuratezza degli algoritmi per il calcolo dell'informazione mutua per una Gamma bivariata.

Da questa tabella si vede che gli algoritmi KL e KSG sono i migliori in accuratezza. Da qui si conclude che per un generico campo stocastico (non Gaussiano ed eventualmente non lineare), *il miglior algoritmo a disposizione è il KSG*. Un'analisi più approfondita richiederebbe il calcolo dell'informazione mutua su un campo stocastico non gaussiano anziché soltanto su una coppia di vettori.

Capitolo 4

Studio delle non linearità applicato alle tematiche ambientali

Nei precedenti capitoli si sono introdotti e validati gli strumenti di correlazione e di informazione mutua, o meglio l'indice R, ed in questo capitolo, si sceglie di applicarli ad un caso studio numerico in ambito ambientale. Lo scopo è quello di rilevare la presenza di legami non lineari tra le variabili fisiche in gioco.

4.1 Caso studio: trasporto di un inquinante in un acquifero

Si è scelto di studiare numericamente il trasporto di un soluto inquinante in una porzione di un acquifero fittizio. Questo acquifero è spesso 100 metri e di forma quadrata con lato pari a 10000 metri. Dal lato ovest al lato est è stata imposta una differenza di carico di 1000 metri. Il flusso è stazionario e va da ovest verso est ma non omogeneo poichè la trasmissività è un campo eterogeneo isotropo. In particolare la trasmissività è generata mediante il metodo di Choleski e utilizzando una funzione di correlazione esponenziale come mostrato nel precedente capitolo. La lunghezza di correlazione di questo campo di trasmissività è di 500 metri ed il campo è distribuito come una log-normale di varianza $\sigma_Y^2 = 2$, dove $Y = \log T$ e T è la trasmissività. Per quanto riguarda la discretizzazione del dominio, si è optato per una suddivisione in 120X120 celle, in modo da avere 6 celle per lunghezza di correlazione.

La dispersività longitudinale è di 10 metri. Con questa impostazione si esegue Modflow per ottenere il campo dei carichi idraulici dopo 20 anni come mostrato in figura 4.1.

Dopo aver ottenuto il campo dei carichi, si esegue MT3DMS per il trasporto di un soluto inquinante: nella cella indicata in nero nella figura 4.1 si immette il soluto per 20 anni con rate 1 grammo per metro cubo per unità di tempo. Quello che si osserva è che il soluto si disperde da ovest verso est in maniera irregolare tanto quanto è irregolare il

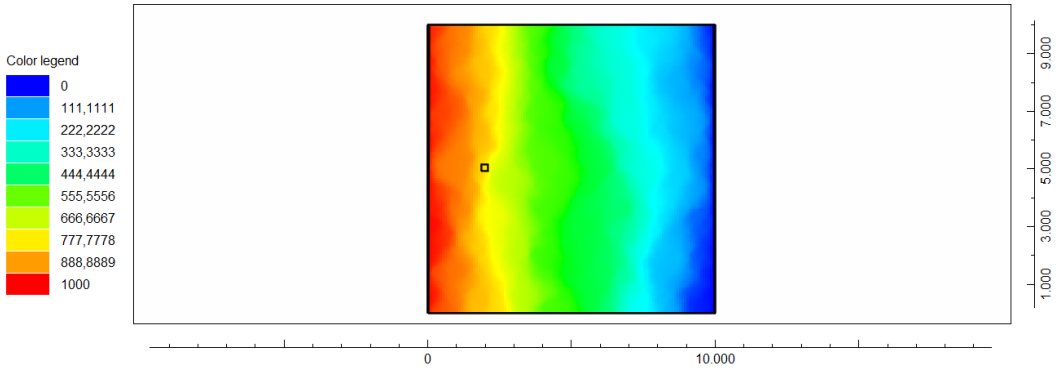

Figura 4.1: Campo dei carichi dopo 20 anni di simulazione.

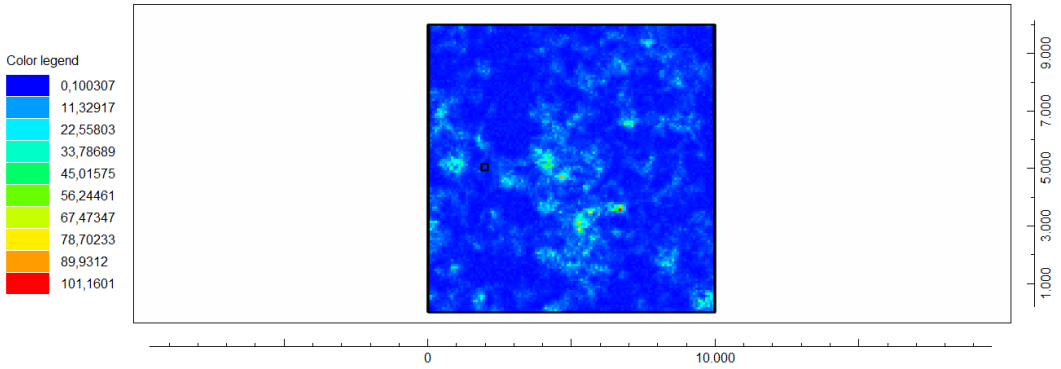

Figura 4.2: Campo di trasmissività spazialmente correlato: una delle 500 realizzazioni.

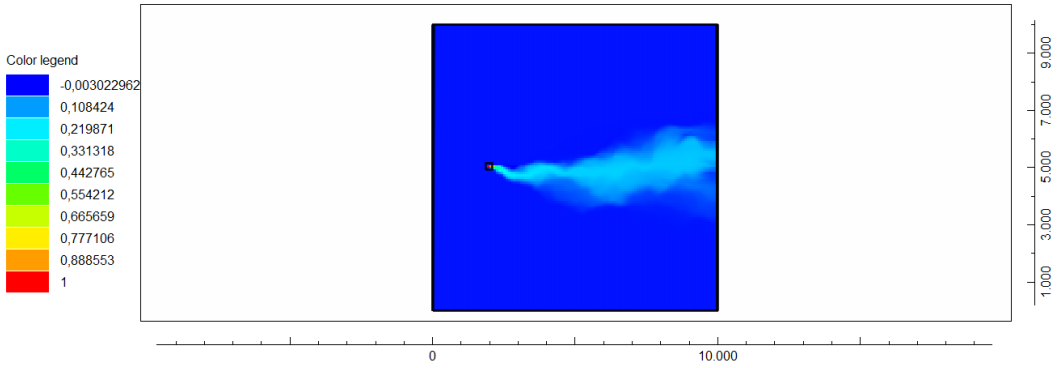

Figura 4.3: Concentrazione del soluto inquinante a 20 anni: una delle 500 realizzazioni.

campo di trasmissività. Per monitorare la concentrazione del soluto, si sono scelte 180 celle (figura 4.4) disposte a valle della sorgente dell'inquinante. Ma nell'analisi verranno considerate solo le 5 righe centrali come mostrate in figura 4.5.

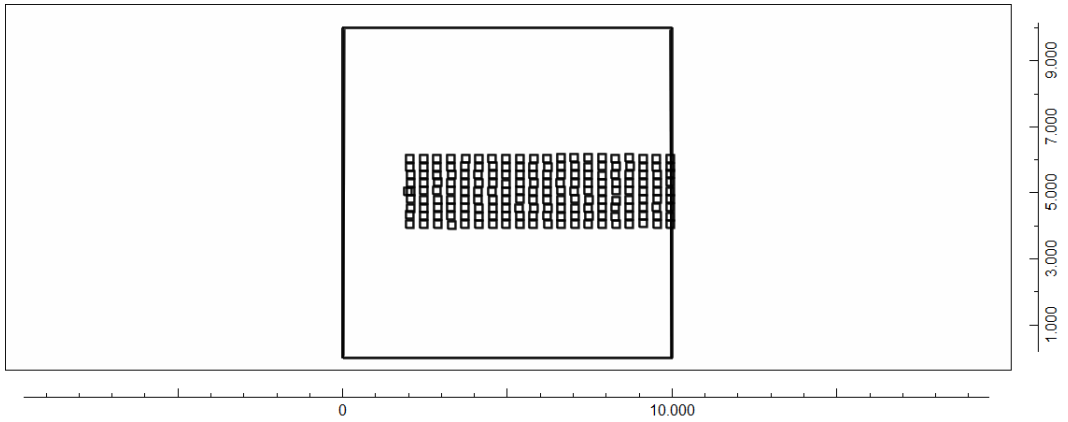

Figura 4.4: Disposizione delle celle di osservazione della concentrazione.

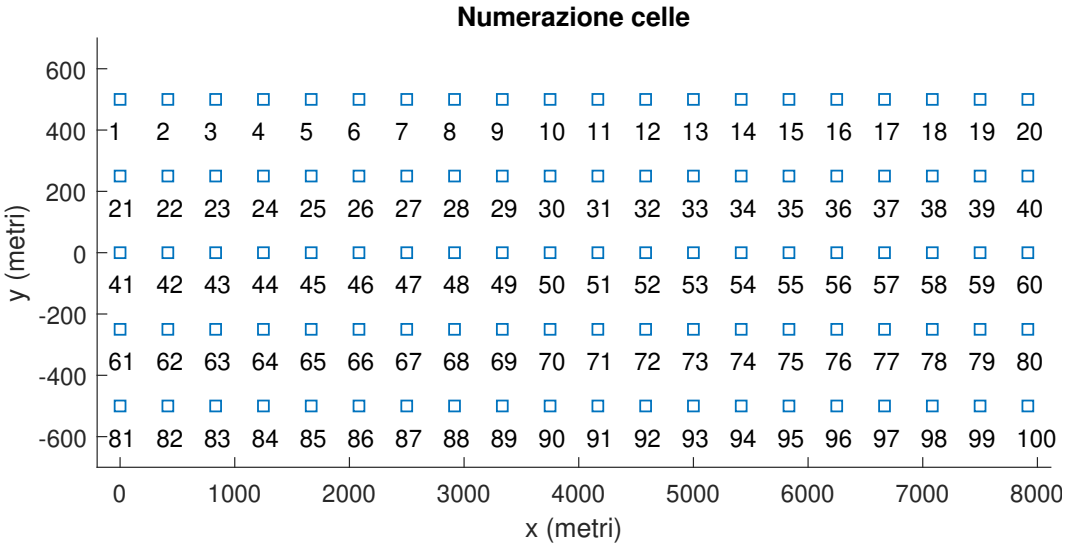

Figura 4.5: Ordinamento delle 5 righe centrali delle celle di osservazione della concentrazione.

4.2 Analisi dei risultati

4.2.1 Campo delle concentrazioni mediato sulle realizzazioni

Dato che si sono ottenute 500 realizzazioni del campo delle concentrazioni del soluto, si può graficare il campo delle concentrazioni mediato sulle realizzazione per ogni istante di tempo preso in considerazione (figura 4.6).

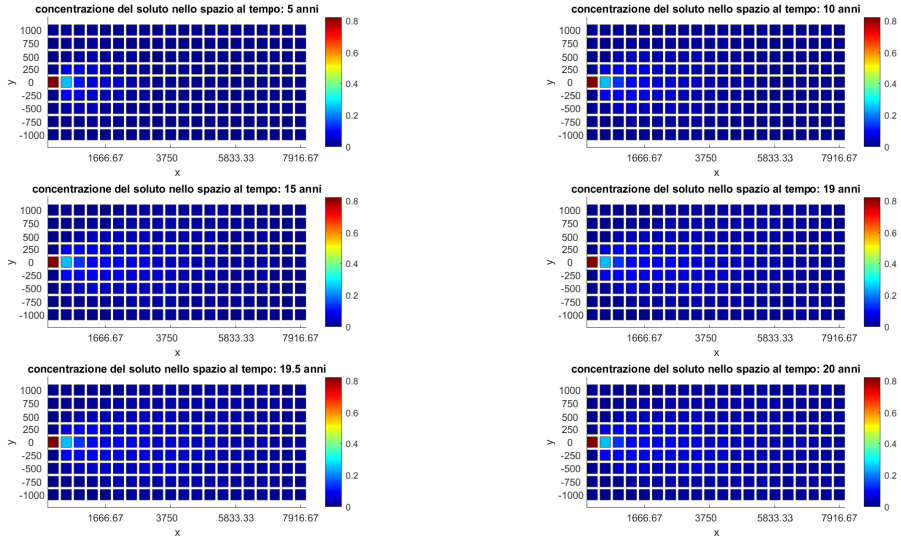

Figura 4.6: Isoconcentrazioni in vari istanti di tempo.

4.2.2 Analisi della correlazione e dell'informazione mutua

In questa sezione si mostrano i risultati dell'applicazione degli strumenti statistici mostrati nei capitoli 2 e 3, per rilevare le non linearità nel campo delle concentrazioni. Dato che il campo delle concentrazioni non è isotropo, non si può applicare lo strumento della correlazione spaziale e dell'informazione mutua spaziale come presentata nel capitolo 2. Invece si decide di prendere in considerazione le 100 celle, che rappresentano 100 posizioni nello spazio, di osservazione della concentrazione e cercarne di rilevare il tipo di legame fra loro; quindi non prendendo tutte le coppie di punti equidistanti tra loro, ma prendendo in considerazioni tutte le coppie possibili tra le 100 celle. Questo lavoro viene svolto per più istanti di tempo, in particolare per i tempi $t = 5$ anni, $t = 10$ anni, $t = 15$ anni, $t = 19$ anni, $t = 19.5$ anni e $t = 20$ anni come si può vedere in figura 4.7. E' stata usata una mappa di colori arcobaleno per meglio evidenziare i valori assunti dalla correlazione, da R e dalla differenza $R - |\rho|$, mentre per i valori negativi di quest'ultima differenza, che dovrebbe essere sempre positiva, si è scelto il colore bianco in modo da contraddistiguerli dai valori positivi.

4.2.3 Dettaglio delle non linearità nelle 20 celle centrali

Adesso, si prende in esame le sole 20 celle centrali, ovvero le celle con numerazione che va da 41 a 60, cioè le celle a valle della sorgente alla sua stessa altezza (figura 4.8). Per un maggiore dettaglio delle non linearità, si decide di graficare la relazione o legame, che c'è tra la cella numero 41 e tutte le altre celle dalla 41esima alla 60esima (figure 4.9, 4.10). Si può notare qui che all'aumentare della distanza, già a partire da 5 anni di simulazione, si presentano delle non linearità, ovvero la differenza tra la correlazione in valore assoluto e l'indice R.

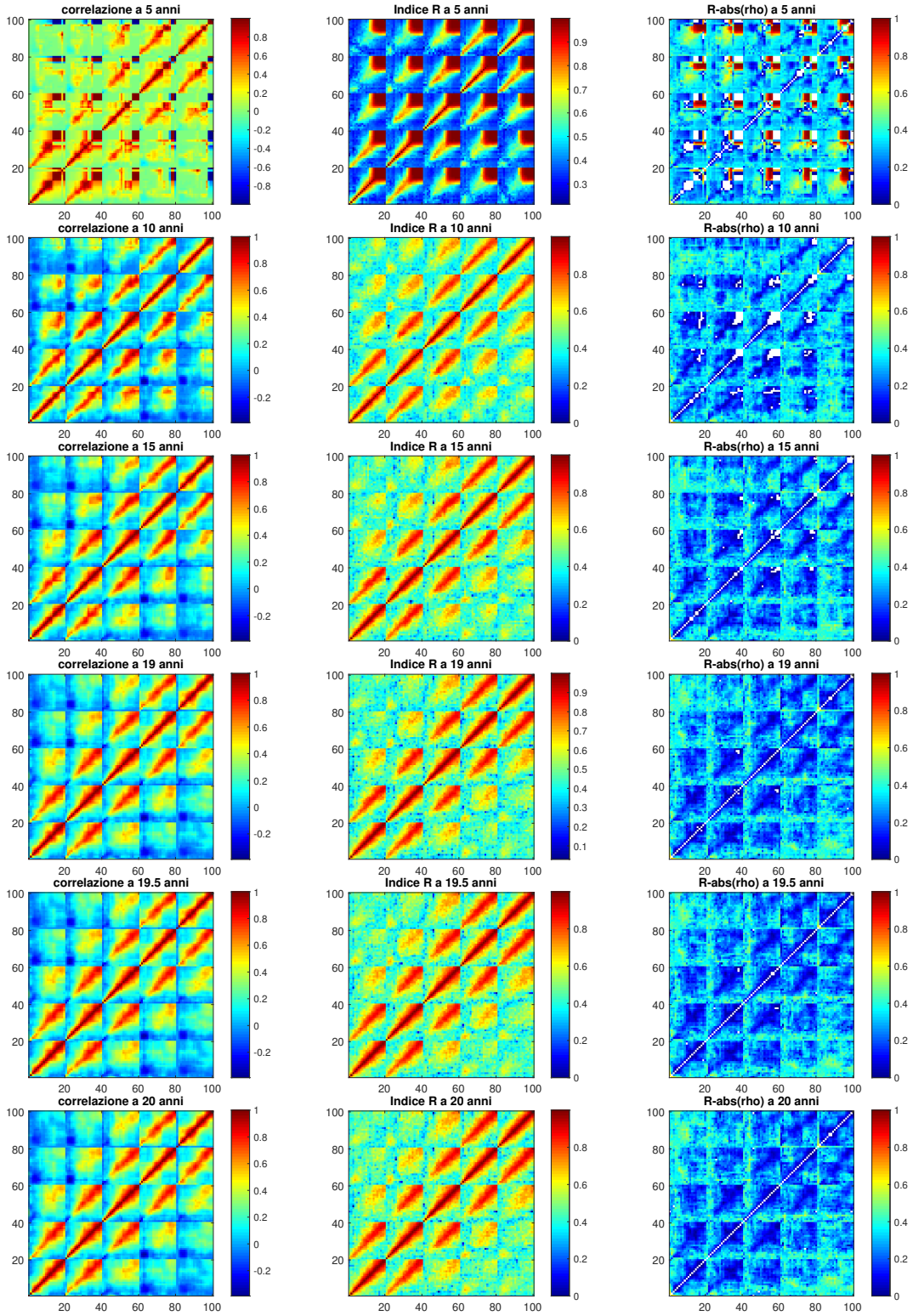

Figura 4.7: Correlazione, indice R e differenza $R - |\rho|$ per i vari istanti di tempo.

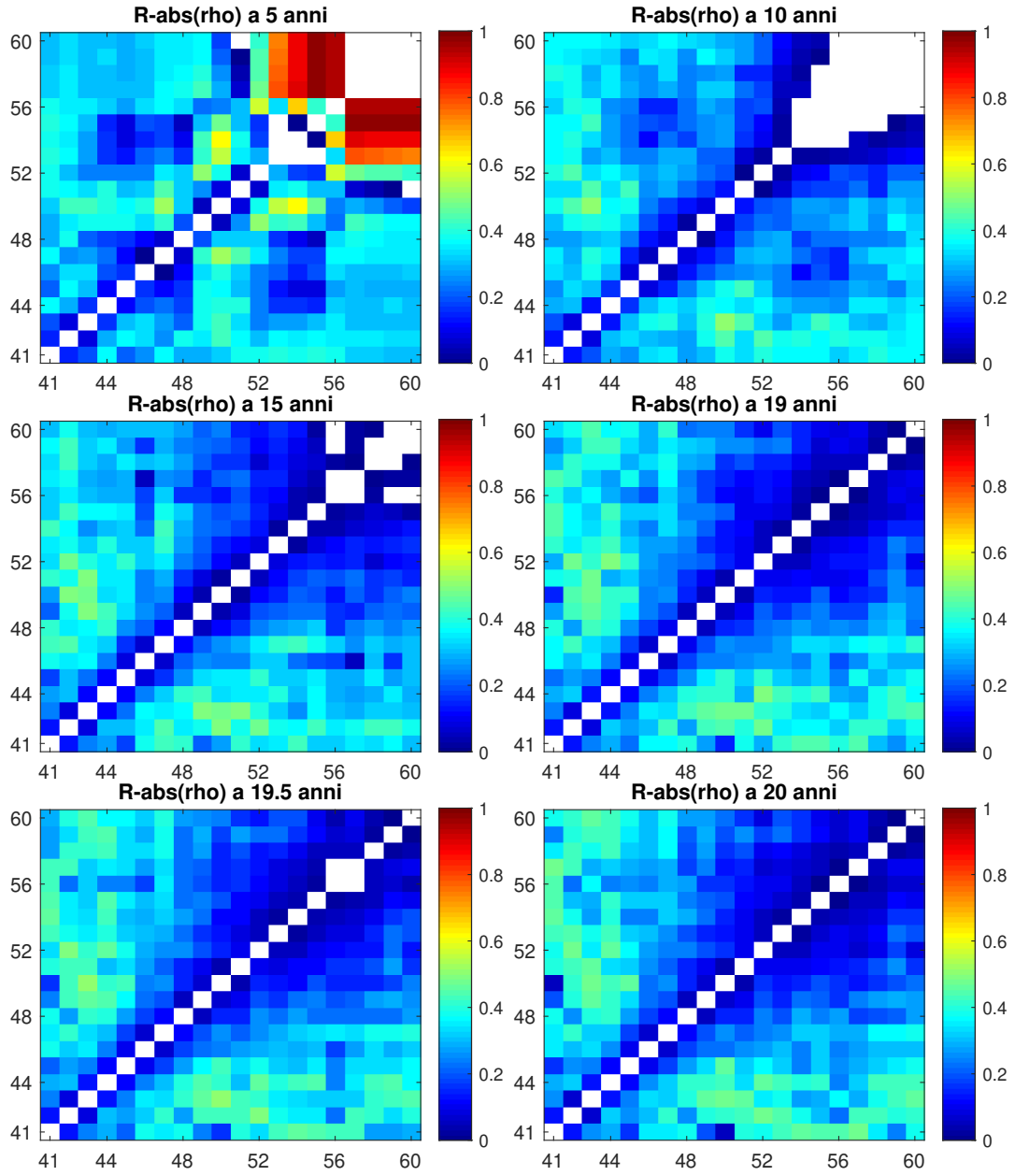

Figura 4.8: Dettaglio della mappa di non linearità delle 20 celle centrali.

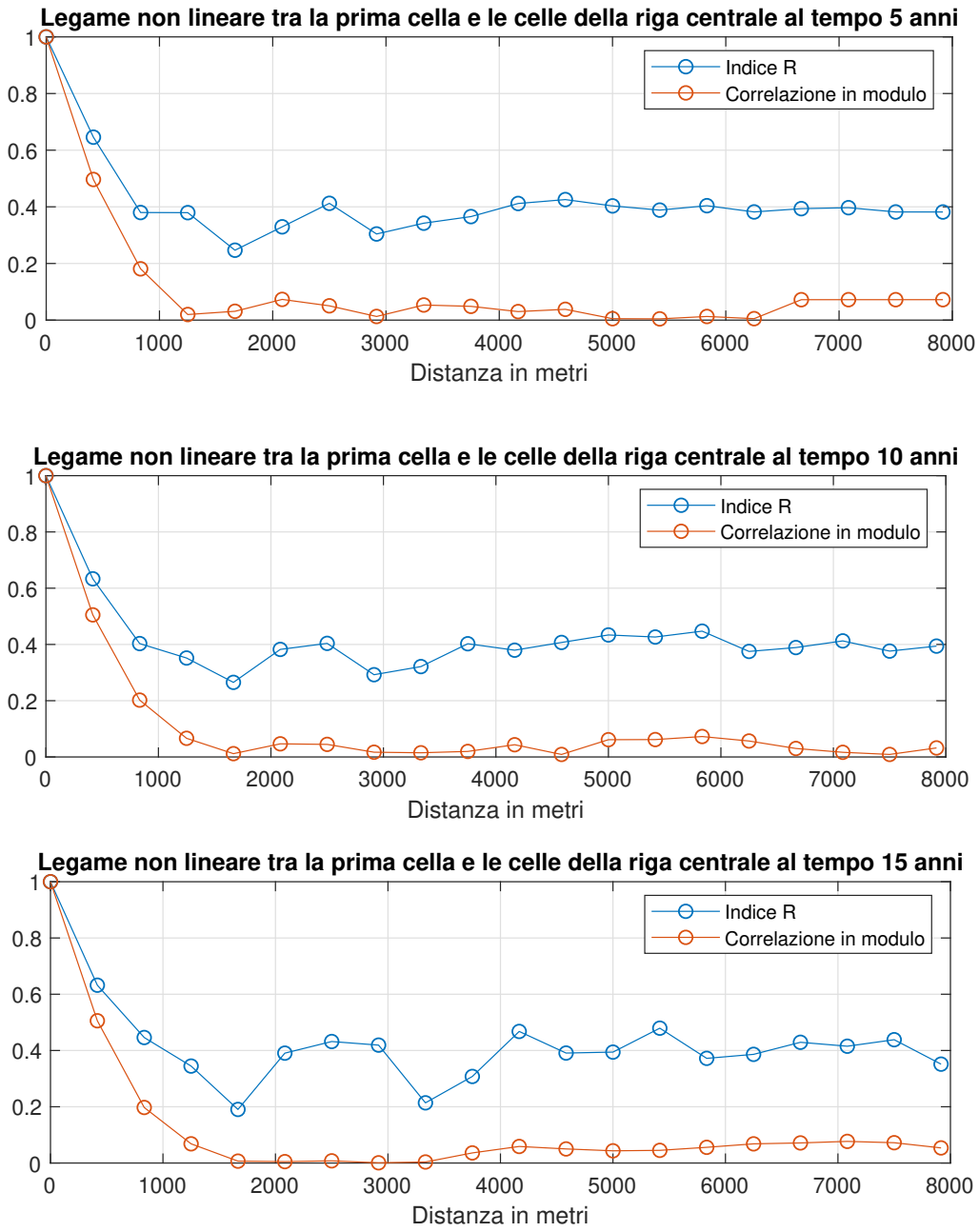

Figura 4.9: Indice R e valore assoluto della correlazione evidenziano non linearità nello spazio per diversi tempi di osservazione.

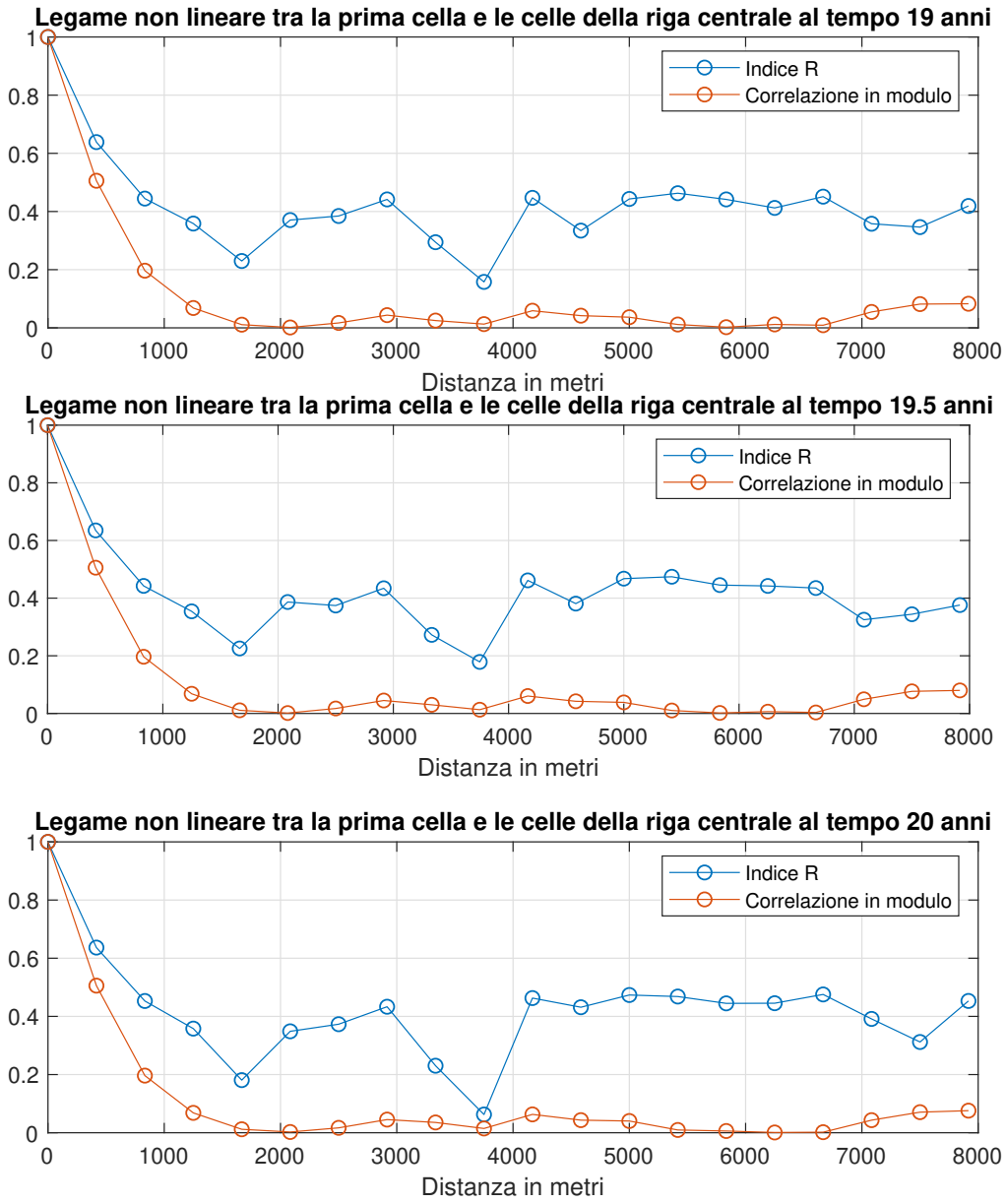

Figura 4.10: Indice R e valore assoluto della correlazione evidenziano non linearità nello spazio per diversi tempi di osservazione.

Capitolo 5

Conclusioni

In questo lavoro di tesi si è cercato di evidenziare i legami non lineari che esistono, spazialmente, nel campo delle concentrazioni di un soluto che si disperde in un acquifero.

Si è partiti presentando gli strumenti classici della geostatistica, come ad esempio la correlazione spaziale, ed uno strumento preso dalla teoria dell'informazione, come l'informazione mutua e la sua trasformazione, l'indice R. Se per la correlazione esiste già un ottimo metodo di stima a partire da un campione di dati, per l'informazione mutua esistono vari metodi tra cui quello degli *istogrammi*, il *Kernel Density Estimator*, il *K-L* ed il *KSG*. Dopo un'analisi delle performance di quest'ultimi, e quindi dopo aver fatto una validazione, si è eletto il metodo di Kraskov a candidato ideale per rilevare legami non lineari.

Dopodichè, è stato condotto uno studio numerico usando un approccio statistico, quindi sfruttando un numero elevato (500) di realizzazioni di acquiferi statisticamente equivalenti. Le analisi mostrano che l'indice R è maggiore della correlazione ρ in valore assoluto. Questo è stato messo in evidenza grazie agli strumenti statistici presentati nel secondo capitolo, ovvero la correlazione e l'informazione mutua, sotto la forma dell'indice di dipendenza non lineare R. La loro differenza, $R - |\rho|$, è stata graficata per alcune coppie di punti di osservazione della concentrazione del soluto inquinante. In particolare, è stata messa in evidenza la relazione che sussiste tra la concentrazione nel punto subito a valle della sorgente e tutti le altre concentrazioni della stessa riga. Questo grafico dice che già a partire da 5 anni di simulazione, il contributo non lineare medio $R-|\rho|$ si attesta intorno a 0.4. Questo risultato non può essere interpretato come risposta positiva alla questione dell'esistenza di non linearità. Piuttosto, come evidenziato nel capitolo 3, bisogna indagare ancora sulla precisione degli algoritmi per il calcolo dell'informazione mutua, poichè anche con valori di quest'ultima prossimi allo zero, il valore di R può essere alto e attestarsi su 0.4. Quindi, si rimanda a studi futuri, una valutazione più accurata sull'esistenza di non linearità in questo tipo di simulazione.

Negli altri possibili futuri sviluppi, c'è sicuramente la possibilità di quantificare la non linearità variando i parametri dell'acquifero come ad esempio la lunghezza di correlazione per la trasmissività oppure il valore di dispersività.

Appendice A

Appendice: Modflow, MT3DMS e Model Muse

I codici di calcolo utilizzati per le simulazioni numeriche sono Modflow e MT3DMS e si tratta di software sviluppati dall'United States Geological Survey. L'interfaccia grafica con cui si è utilizzato questi due software è Model Muse.

A.1 Modflow

Modflow è un codice di calcolo che implementa il metodo delle differenze finite e che risolve le equazioni dei flussi di acque sotterranee.

L'equazione delle acque sotterranee che viene risolta in caso di dominio bidimensionale, stazionarietà del moto, trasmissività eterogenea, isotropa e in assenza di sorgente è la seguente:

$$\frac{\partial}{\partial x}T\frac{\partial h}{\partial x} + \frac{\partial}{\partial y}T\frac{\partial h}{\partial y} = 0, \tag{A.1}$$

con $h = h(x, y)$ carico idraulico e $T = T(x, y)$ trasmissività. L'equazione viene discretizzata secondo le differenze finite e imponendo le condizioni al bordo dell'acquifero; quindi viene risolto un sistema matriciale per trovare il campo dei carichi.

In particolare, viene usata la versione 2005 di Modflow, con il pacchetto BCF6, ovvero il Block-Centered Flow package per il flusso, che consente di lavorare inserendo la trasmissività per ogni cella della discretizzazione. Per le condizioni al bordo viene scelto il pacchetto CHD, ovvero il Time-Variant Specified Head package, per imporre la condizione di carico costante nel bordo est e nel bordo ovest.

Scegliendo di non applicare alcuna condizione al bordo nord e al bordo sud, il programma, automaticamente impone la condizione di "impervious boundary".

A.2 MT3DMS

Il software MT3DMS (mass transport tridimensional for mobile species) è un codice di calcolo che risolve le equazioni di trasporto di massa negli acquiferi. In questo studio si sono utilizzato i pacchetti Basic Transport, Advection, Dispersion, Sink and Source Mixing e Generalized Conjugate Gradient Solver.

L'equazione del trasporto del soluto inquinante, assumendo che la porosità è costante, è la seguente:

$$\frac{\partial c}{\partial t} = -\nabla \cdot (c\boldsymbol{v} - \boldsymbol{D}\nabla c) + s \tag{A.2}$$

dove $c = c(x, y, t)$ è la concentrazione del soluto, $\boldsymbol{v} = \boldsymbol{v}(x, y)$ è il campo di moto ottenibile attraverso il campo dei carichi, s è la sorgente del soluto, e $\boldsymbol{D}$ è il tensore che tiene conto dela diffusione molecolare e della dispersione meccanica.

A.3 Model Muse

Model Muse è un interfaccia grafica per i modelli sopradescritti. L'interfaccia si presenta come in figura A.1.

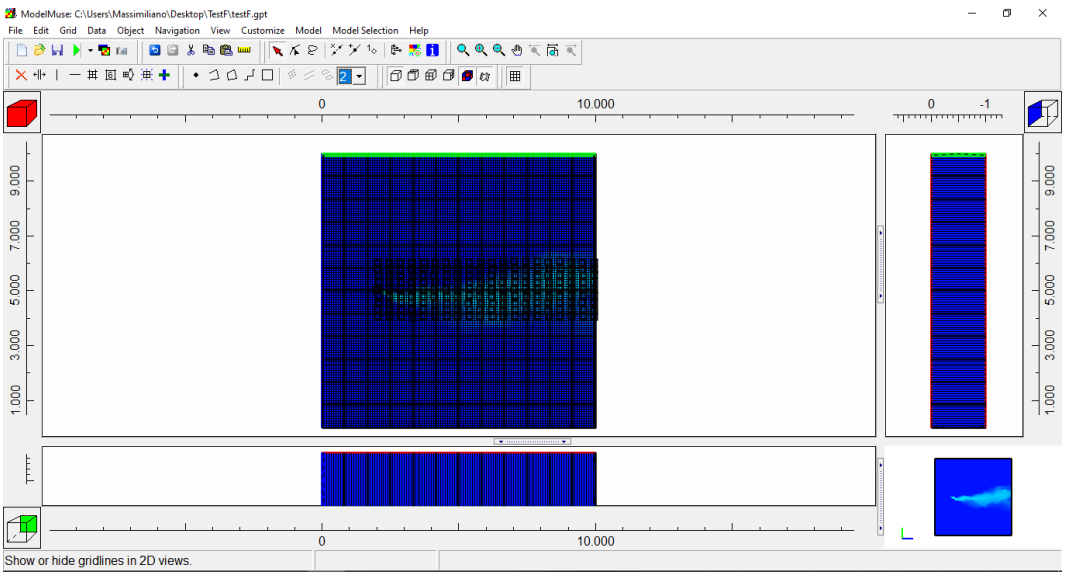

Figura A.1: Schermata di Model Muse.

Bibliografia

A. Sayago A. G. Asuero and A. G. Gonz´alez. The correlation coefficient: An overview. *Critical Reviews in Analytical Chemistry, 36*, pages 41–59, 2006.

Harald St¨ogbauer Alexander Kraskov and Peter Grassberger. Estimating mutual information. *PRE*, 2004.

Dimitris Kugiumtzis Angeliki Papana. Evaluation of mutual information estimators for time series. *https://arxiv.org/abs/0904.4753v1*, 2009.

Nicolas Bez. Linear and applied geostatistics. Doctoral. France. 2019. cel-01998500.

David R. Brillinger. Some data analyses using mutual information. *Brazilian Journal of Probability and Statistics 18*, 2004.

P. Abry N.B. Garnier C. Granero-Belinchon, S.G. Roux. Probing high order dependencies with information theory. *secondarXiv: 1812.05325v1*, 2019.

Christian Caamaño Carrillo. Modeling and estimation of some non gaussian random fields. *Tesi, Instituto de Estadística Universidad de Valparaíso*, 2018.

Noel A. C. Cressie. *Statistics for spatial data revised version.* John Wiley & Sons, INC., 1993.

Davide Frison. Analisi della nozione di transfer entropy e applicazioni. *Tesi, UNIVERSITÀ DEGLI STUDI DI PADOVA*, 2015.

L. Vallivero I. Butera and L. Ridolfi. Mutual information analysis to approach nonlinearity in groundwater stochastic fields. *Stochastic Environmental Research and Risk Assessment*, 32:2933–2942, 2018.

J. Chiang K. Cheng J. Liou, Y. Su. Gamma random field simulation by a covariance matrix transformation method. *Stochastic Environment Resource Risk Assessment*, 2011.

Leonenko NN Kozachenko LF. Sample estimate of the entropy of a random vector. *Problems of information transmission*, 1987.

Yupeng Li and Clayton V. Deutsch. Mutual information and its application in spatial statistics. *Paper 122, CCG Annual Report 12*, 2010.

Andreas Lichtenstern. *Kriging methods in spatial statistics*. Bachelor Thesis, Technische Universitat Munchen, Department of Mathematics, 2013.

Patrick Emmanuel Meyer. Information-theoretic variable selection and network inference from microarray data. *Thèse de l'Université Libre de Bruxelles Bruxelles, Belgique*, 2008.

S. Nadarajah and S. Kotz. Mutual information analysis to approach nonlinearity in groundwater stochastic fields. *Rocky mountain Journal of Mathematics*, 39, 2009.

Liam Paninski. Estimation of entropy andmutual information. *Neural Computation 15, 1191–1253*, 2003.

S. C. Chapman R. T. Wicks and R. O. Dendy. Spatial correlation of solar wind fluctuations and solar cycle dependence. *arXiv:0711.4814v2*, 2008.

Christoph Kayser Guillaume A. Rousselet Joachim Gross Robin A.A. Ince, Bruno L. Giordano and Philippe G. Schyns. A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. *Human Brain Mapping 38:1541–1573*, 2017.

Claude Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 1948.

Aram Galstyan Shuyang Gao, Greg Ver Steeg. Efficient estimation of mutual information for strongly dependent variables. *arXiv:1411.2003v3*, 2015.

Joy A. Thomas Thomas M. Cover. *Elements of Information Theory*. John Wiley Sons, Inc., 1991.

A.K.; Konapala G. Vu, T.M.; Mishra. Information entropy suggests stronger nonlinear associations between hydro-meteorological variables and enso. *Entropy*, pages 20,38, 2018.

Pramod Viswanathz Weihao Gao, Sewoong Ohy. Demystifying fixed k-nearest neighbor information estimators. *arXiv:1604.03006v2*, 2016.

U. Lall Y. Moon, B. Rajagopalan. Estimation of mutual information using kernel density estimators. *PHYSICAL REVIEW E 76, 026209*, 1995.

Shuyu Sun Bo Yu Yang Liu1, Jingfa Li. Advances in gaussian random field generation: a review. *Computational Geosciences 23:1011–1047*, 2019.