ALMA MATER STUDIORUM - UNIVERSITÁ DI BOLOGNA

FACOLTÀ DI INGEGNERIA

Corso di Laurea Magistrale in Ingegneria per l'Ambiente e il Territorio

DICAM

Tesi di Laurea Magistrale
In
Modellistica Idrologica

TECNICHE DI INTERPOLAZIONE GEOSTATISTICA PER LA STIMA DELLA PIENA DI PROGETTO IN BACINI NON STRUMENTATI

Candidato:
Alessio Pugliese

Relatore:
Prof. Ing. Attilio Castellarin

Correlatori:
Prof. Ing. Alberto Montanari
Stacey A. Archfield, Ph.D.

Anno Accademico 2011-2012
Sessione I

Non mi fido molto delle statistiche,
perché un uomo con la testa nel forno acceso
e i piedi nel congelatore statisticamente
ha una temperatura media.
Charles Bukowski

Indice

Elenco delle figure

Elenco delle tabelle

Sommario

La presente dissertazione analizza e mette a confronto due tecniche di *kriging*, che recentemente sono state proposte nella letteratura scientifica, affrontando il problema della stima della piena di progetto in bacini idrografici non strumentati. La prima tecnica, che prende il nome di Canonical Kriging (CK), considera la variabile regionalizzata a supporto puntuale, definita su uno spazio bidimensionale, ottenuto dall'analisi della correlazione canonica dei descrittori geomorfoclimatici con le variabili idrometriche; la seconda tecnica, chiamata Toplogical Kriging (TK), lavora su supporto areale considerando la dimensione e la mutua posizione dei bacini, tenendo conto della loro struttura annidata. Attraverso una procedura di cross-validazione *leave-one-out* sono state messe a confronto le prestazioni del CK con il TK su quantili di piena al colmo con tempo di ritorno 10, 50, 100, 500 anni di 61 bacini strumentati degli USA sud-orientali. Sia il CK che il TK risultano essere procedure efficaci per la stima della piena di progetto in bacini non strumentati, tuttavia, nell'area di studio il TK supera significativamente il CK (con riferimento ai valori cross validati della Q_{100} l'efficienza di Nash & Sutcliffe e l'errore medio relativo valgono rispettivamente 0.877 e 0.176 per il TK e 0.826 e 0.299 per il CK). L'analisi ha inoltre mostrato che l'accoppiamento delle tecniche CK e TK migliora leggermente le prestazioni di entrambe le metodologie applicate singolarmente, in particolare si è osservata una riduzione dell'errore medio relativo e della radice dell'errore quadratico medio, specialmente per i bacini medio-piccoli.

Abstract

Concerning the problem of design-flood prediction in ungauged basins, the Dissertation analyses and compares two *kriging* techniques that have recently been proposed in the scientific literature. The first technique, named Canonical Kriging (CK), considers the regionalized variable with point support defined on a synthetic two-dimensional space, obtained from canonical correlation analysis of geomorfoclimatic descriptors and hydrometric variables; the second technique, called Topological Kriging (TK), works on a non-point area based support, considering basins' dimension and location, taking their nested structure into account. By means of a leave-one-out cross validation procedure, the performance of CK and TK was compared on 10-, 50-, 100- and 500-year floods for 61 streamgauges in the southeast United States. Both CK and TK are effective procedures for design-flood prediction in ungauged sites; however, TK significantly outperforms CK for the study area (Nash-Sutcliffe and mean-absolute- error Q_{100} cross-validated estimates are equal to 0.877 and 0.176 for TK and 0.826 and 0.299 for CK). The analysis also shows that coupling TK and CK slightly improves the performance of each methodology applied as a standalone technique, in particular reducing both BIAS and root-mean-square-error for small-to-medium-catchments.

Introduzione

Nei problemi di difesa idraulica del territorio e di progettazione di opere idrauliche si pone come necessaria la valutazione della portata di piena temibile con assegnato livello di rischio (o tempo di ritorno), in una determinata sezione fluviale. Tuttavia la ricerca di un valore di portata di progetto è ostacolata dalla mancanza di informazioni idrologiche su scala locale, circostanza che introduce l'annoso problema legato alla stima, argomento verso il quale la comunità scientifica ha riposto un'attenzione crescente, come testimonia il progetto di ricerca internazionale PUB (*Prediction in Ungauged Basin*) promosso dall'International Association of Hydrological Sciences (Sivapalan et al.,2003).

La regionalizzazione statistica del regime di frequenza degli estremi idrologici è ad oggi tra le tecniche comunemente usate per la stima dei deflussi fluviali in siti per i quali non sono disponibili o reperibili dati. Secondo l'approccio tradizionale, che fa riferimento al metodo della "piena indice", alla base di tali tecniche vi è la ricerca di regioni *omogenee*, ovvero di raggruppamenti di bacini sulla base di criteri di vicinanza geografica o di similitudine idrologica. Tuttavia la fase di classificazione e raggruppamento dei bacini rappresenta una problematica tuttora aperta e dibattuta dalla comunità scientifica.

Recentemente, nell'ambito della regionalizzazione idrologica, si stanno affermando tecniche geostatistiche di interpolazione basate sul *Kriging*. Tale tecnica, nata in ambito minerario, permette la stima di una variabile regionalizzata in punti in cui essa non è nota a priori, a partire da un campione di dati distribuiti sul dominio di interpolazione.

L'obiettivo posto nella presente dissertazione è quello di accoppiare due diverse tecniche di krigaggio, note in letteratura come Canonical Kriging e Topologocal Kriging: il primo approccio, identificato con l'acronimo PSBI (*Physiografical Space Based Interpolation*), considera la variabile idrometrica su un supporto puntuale, definita su un piano cartesiano sintetico (v. Chokmani e Ouarda 2004; Castiglioni et al. 2011), a sua volta ottenuto mediante l'analisi della correlazione canonica dei descrittori geomorfoclimatici e delle variabili idrometriche di piena al colmo; il secondo, lavora su supporto non puntuale nello spazio geografico, tenendo conto delle dimensioni, della mutua posizione e dell'eventuale struttura annidata dei bacini (v. Skøien, Merz e Blöschl 2006). Per simulare le condizioni non strumentate si può fare ricorso alla tecnica della cross-validazione *jack-*

knife, anche detta *leave-one-out*, la quale permette di confrontare, in corrispondenza della stessa stazione di misura, i valori empirici con i valori simulati.

Recenti studi (v. Castiglioni et al. 2011) hanno dimostrato che le due tecniche sembrano avere caratteristiche di complementarietà: il Canonical Kriging lavora bene per bacini simili, ma lontani tra loro, come ad esempio i bacini di testata o i piccoli bacini montani; il Topological Kriging al contrario ottiene migliori performance per le sezioni fluviali a valle delle sezioni strumentate, ovvero per bacini di grandi dimensioni. Alla base del lavoro, dunque, c'è l'intenzione di considerare i residui generati dai due stimatori in fase di cross-validazione come una variabile regionalizzata, interpolabile, questa, mediante l'utilizzo, a seconda dei casi, di una delle seguenti tecniche: il Top-Kriging per la stima dei residui generati dal Canonical Kriging; il Canonical Kriging per i residui generati dal Top-Kriging. In questo modo, per ciascuna delle due linee d'intervento, si hanno a disposizione una superficie di interpolazione della variabile idrometrica e una del residuo, per cui la portata "modificata", scopo ultimo dell'analisi, è ottenuta semplicemente sommando le due stime.

L'analisi, condotta su 61 bacini degli U.S.A. sud-orientali compresi tra i territori della Georgia, dell' Alabama e della Florida, si pone due obiettivi principali: in primo luogo il confronto, in termini prestazionali, delle due principali tecniche di interpolazione (Canonical Kriging e Top-Kriging); in secondo luogo, di verificare l'applicabilità della sovrapposizione dei due approcci, modellando i residui in uscita dalla prima stima.

Il lavoro di tesi si articola come segue.

Nel *primo capitolo* vengono richiamati alcuni concetti e formalismi della Geostatistica, propedeutici alla trattazione.

Nel *secondo capitolo* sono illustrate due tecniche di analisi statistica multivariata, mettendo in luce le analogie e le differenze tra l'analisi delle componenti principali e l'analisi della correlazione canonica, utilizzata, quest'ultima, nella tecnica Canonical Kriging.

Nel *terzo capitolo* vengono esposte e descritte, con dovizia di particolari, le due tecniche di kriging considerate in questo lavoro: Canonical Kriging e Topological Kriging.

Nel *quarto capitolo* viene presentata l'area di studio su cui si è concentrata l'analisi e la base dati a disposizione.

Nel *quinto capitolo* si articola la struttura dell'indagine sulle due linee d'intervento, basate sul kriging, che hanno portato alla stima dei quantili di piena al colmo in bacini non strumentati e le modalità con cui sono stati interpolati i residui.

Nel *sesto capitolo* vengono riportati tutti i risultati ottenuti sia in termini di diagrammi di dispersione, sia di indici statistici prestazionali, tabellati in modo da evidenziare miglioramenti e peggioramenti.

Capitolo 1

Cenni di Geostatistica

1.1 Introduzione

La *Geostatistica* studia i fenomeni naturali che si sviluppano su base spaziale a partire dalle informazioni derivanti da un loro campionamento. In particolare studia la variabilità spaziale dei parametri che descrivono i suddetti fenomeni, estraendone le regole in un quadro modellistico di riferimento e usandole per effettuare le operazioni volte a dare soluzione a specifiche problematiche riguardanti la caratterizzazione e la stima dei fenomeni stessi (v. Raspa e Bruno 1994a).

Si consideri un fenomeno che ha caratteristiche di variabilità spaziale indicando con $z(\vec{x})$ la variabile nel punto di coordinate planimetriche $\vec{x} = (u, v)$. Si possono dare le seguenti definizioni:

Variabile Regionalizzata

Si intende la funzione $z(\vec{x})$ il cui valore dipende dalla localizzazione e che si presenta strutturata spazialmente.

Campo

È il dominio nel quale la variabile regionalizzata è suscettibile di assumere determinati valori e all'interno del quale se ne studia la variabilità.

Supporto

È l'entità geometrica sulla quale la variabile regionalizzata è definita o anche misurata; essa è caratterizzata dalle sue dimensioni e dalla sua forma. Quando le dimensioni sono molto piccole rispetto al campo il supporto può considerarsi *puntuale*.

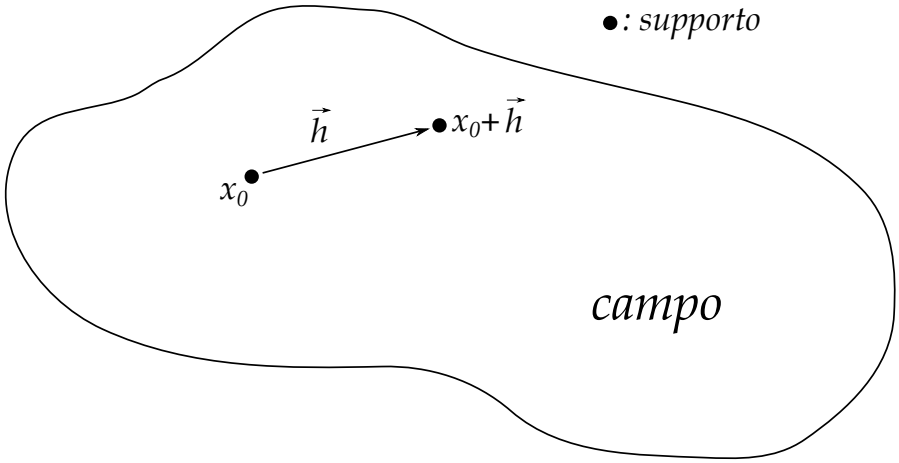

Figura 1.1: Punti distanti $|\vec{h}|$ nel campo

1.2 Approccio probabilistico

Secondo la definizione data di *Variabile Regionalizzata* la funzione $z(\vec{x})$ che descrive il fenomeno in esame è chiaramente una funzione deterministica.

Consideriamo ora un particolare punto $\vec{x}_0$ del campo, in esso si può identificare una variabile aleatoria (VA) continua $Z(x_0)$, cioè una variabile che assume valori numerici appartenenti ad un certo intervallo secondo una legge di densità di probabilità $f_0(Z)$. In queste condizioni il valore deterministico $z(\vec{x}_0)$ può essere considerato come una realizzazione della VA $Z(x_0)$. Così come in x_0 può essere definita una VA in ogni altro punto generico x del campo; allora l'insieme di tutte le VA definite nel campo costituisce una *Funzione Aleatoria* (FA).

La FA $Z(x)$ sarà caratterizzata dall'insieme di tutte le funzioni di distribuzione multivariabili che si possono definire nel campo per qualsiasi intero k e per qualsiasi configurazione dei k punti $x_1, x_2, \ldots, x_k$:

$$F_{Z_1,\ldots,Z_k}(z_1,\ldots,z_k) = \text{prob}\{Z(x_1) < z_1, \ldots, Z(x_k) < x_k\} \tag{1.1}$$

Questo insieme di funzioni, data la natura spaziale del fenomeno che si vuole modellizzare, costituisce la legge spaziale della FA $Z(x)$; nel caso di indipendenza a due a due delle variabili $Z_1, \ldots, Z_k$, la legge spaziale $Z(x)$ si riduce all'insieme di funzioni di distribuzione monovariabile:

$$F_{Z(x)}(z) = \text{prob}\{Z(x) < z\} \qquad \forall x \in S \tag{1.2}$$

dove per S è inteso il dominio (*campo*).

Il vantaggio dell'approccio probabilista è quello di poter far uso di *modelli* di cui è necessario stimare i parametri a partire dai dati di campionatura del fenomeno. Per ciò

che concerne le stime (o *predizioni* in termini statistici) sarà sufficiente la conoscenza dei primi due *momenti* della funzione aleatoria e della funzione più comune in ambito geostatistico, ovvero la funzione *variogramma*:

Momento Primo

Sia S il campo di indagine, in accordo con l'interpretazione probabilista, in ogni punto $x \in S$ è definita una VA $Z(x)$ tale che il suo momento primo è dato da[1]:

$$m(x) = \mathrm{E}[Z(x)] \tag{1.3}$$

Momento Secondo

Si considerino due punti x_1 e x_2 di S, la *covarianza* in quei punti tra le variabili aleatorie $Z(x_1)$ e $Z(x_2)$ è data da:

$$\mathrm{Cov}(x_1, x_2) = \mathrm{E}[(Z(x_1) - m(x_1))(Z(x_2) - m(x_2))] = \mathrm{E}[Z(x_1)Z(x_2)] - m(x_1)m(x_2) \tag{1.4}$$

se i due punti coincidono si ottiene la *varianza*[2]

$$\mathrm{Var}(x) = \mathrm{Cov}(x, x) = \mathrm{E}[Z^2(x)] - \big(\mathrm{E}[Z(x)]\big)^2 \tag{1.5}$$

Variogramma

Sia S il dominio di variabilità della FA $Z(x)$ e siano x_0 e $x_0 + \vec{h}$ una coppia di punti di S e distanti $|\vec{h}|$ (v. fig. 1.1). La differenza tra $Z(x_0)$ e $Z(x_0 + \vec{h})$ definisce una nuova variabile aleatoria detta *incremento*:

$$Z(x_0 + \vec{h}) - Z(x_0)$$

Si definisce *variogramma* la sua semivarianza:

$$\gamma(x_0, \vec{h}) = \frac{1}{2}\,\mathrm{Var}\{Z(x_0 + \vec{h}) - Z(x_0)\} \tag{1.6}$$

[1] si ricorda che per definizione il valore atteso di una variabile casuale continua con funzione densità di probabilità $f(z)$ è dato da

$$\mathrm{E}[Z] = \int_{-\infty}^{+\infty} z f(z)\, dz$$

[2] per definizione la *varianza* di Z è definita come il valore atteso al quadrato della variabile aleatoria centrata $Y = Z - \mathrm{E}[Z]$:

$$\mathrm{Var}(Z) = \mathrm{E}[Y^2] = \mathrm{E}[(Z - \mathrm{E}[Z])^2]$$

Con questi strumenti statistici è possibile caratterizzare il modello di FA preso in considerazione per lo studio del fenomeno in esame.

Molto spesso però è utile poter restringere il dominio della FA ad una piccola porzione del campo, riducendo così lo studio ad un problema locale in modo da poter controllare meglio le statistiche associate e le stime conseguenti: questa porzione di campo viene chiamata *vicinaggio*. L'idea di restringere il campo di applicazione del modello deriva dall'ipotesi di poter traslare il vicinaggio in modo da coprire tutto il dominio e per questo motivo viene definito anche *vicinaggio mobile*. L'artificio del vicinaggio mobile risulta molto utile nella distinzione dei modelli descrittori del fenomeno naturale che si vuole studiare.

1.3 Modelli stazionari

I modelli stazionari si basano sulla proprietà di stazionarietà *strictu sensu* della funzione aleatoria, ovvero sull'invarianza per traslazione della legge spaziale del processo aleatorio. Più precisamente, preso un qualsiasi insieme di k punti $x_1, \ldots, x_k$ del campo S e un qualsiasi vettore $\vec{h}$, i due vettori aleatori $\{Z(x_1), \ldots, Z(x_k)\}$ e $\{Z(x_1+\vec{h}), \ldots, Z(x_k+\vec{h})\}$ hanno la stessa funzione di distribuzione k-variabile di probabilità.

Piuttosto che fare riferimento alla legge di distribuzione, a fini pratici, risulta molto più comodo associare direttamente la stazionarietà ai momenti primo e secondo dei quali si contraddistingue una determinata legge di distribuzione.

Modelli stazionari di oridine 2

Un modello di FA si dice stazionario di ordine 2 quando sono verificate entrambe le due seguenti condizioni:

1. il momento primo esiste ed è invariante rispetto ad x;

2. la covarianza, o momento secondo, esiste e non dipende dalla posizione assoluta dei punti, ma dalla loro reciproca distanza.

Con la prima condizione si assume che il momento primo (la *media*) è costante su tutto il dominio, ovvero:

$$m(x) = \mathrm{E}[Z(x)] = cost. \qquad \forall x \in S. \tag{1.7}$$

Nella seconda condizione, detta $\vec{h}$ la distanza tra i punti x_1 e x_2 tale che $x_2 = x_1 + \vec{h}$, si ammette che $\mathrm{Cov}(x_1, x_2)$ è una funzione di $\vec{h}$:

$$\mathrm{Cov}(x_1, x_2) = C(x_1, x_1 + \vec{h}) = C(\vec{h}) \tag{1.8}$$

si nota che per $|\vec{h}| \to 0$ la covarianza decade nella varianza che è quindi anch'essa invariante per traslazione:

$$\lim_{|\vec{h}|\to 0} C(x_1, x_1 + \vec{h}) = \mathrm{Var}(x_1) = C(0) \tag{1.9}$$

la funzione $C(\vec{h})$ viene chiamata *funzione covarianza* ed esprime la correlazione tra le variabili $Z(x)$ e $Z(x + \vec{h})$ in funzione delle mutua distanza tra i punti del campo. In base alle assunzioni fatte è possibile trovare una formulazione compatta per la funzione variogramma; infatti, tenendo conto dell'invarianza per traslazione del momento primo $\mathrm{E}[Z(x + \vec{h})] = \mathrm{E}[Z(x)]$, per la (1.6) si ha:

$$\gamma(x, \vec{h}) = \frac{1}{2}\mathrm{E}[\big(Z(x + \vec{h}) - Z(x)\big)^2] \tag{1.10}$$

sviluppando in (1.10) il quadrato del binomio al secondo membro e sfruttando la linearità dell'operatore E (valore atteso), si ha:

$$\gamma(x, \vec{h}) = \frac{1}{2}\big(\mathrm{Var}\{Z(x + \vec{h})\} + \mathrm{Var}\{Z(x)\} - 2\,\mathrm{Cov}\{Z(x + \vec{h}), Z(x)\}\big) \tag{1.11}$$

in virtù dell'ipotesi di invarianza per traslazione del momento secondo $\mathrm{Var}\{Z(x + \vec{h})\} = \mathrm{Var}\{Z(x)\}$, per cui la (1.11) diventa:

$$\gamma(x, \vec{h}) = \frac{1}{2}[2\,\mathrm{Var}\{Z(x)\} - 2C(\vec{h})] \tag{1.12}$$

infine, tenendo conto della (1.9), si giunge alla formulazione compatta:

$$\gamma(\vec{h}) = C(0) - C(\vec{h}). \tag{1.13}$$

Questa formulazione dimostra che la funzione variogramma, sotto le ipotesi di stazionarietà, è strettamente legata alla funzione covarianza e si può affermare che anche il variogramma è invariante per traslazione.

Poiché, come è ragionevole pensare, la correlazione tra le variabili $Z(x)$ e $Z(x+\vec{h})$ tende ad indebolirsi con l'aumentare della mutua distanza $|\vec{h}|$ tra i punti, si ha che la funzione $C(\vec{h})$ tende a decrescere con h, fino a potersi annullare se le due variabili diventano indipendenti. La funzione variogramma risulta pertanto limitata superiormente da un *sill*

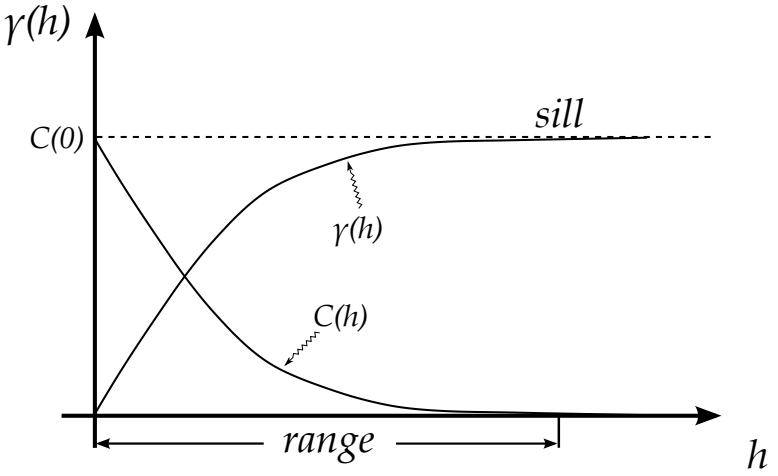

Figura 1.2: Andamento delle funzioni covarianza $C(h)$ e variogramma $\gamma(h)$ in funzione della distanza h

(o *soglia*) e tale limite è la varianza $C(0)$ (v. Fig. 1.2).

1.4 Modelli non stazionari

Le FA non stazionarie sono quelle che soddisfano anche solo una delle due seguenti condizioni:

1. la media $\mathrm{E}[Z(x)] = m(x)$ non è costante nel campo;
2. la funzione covarianza non è invariante per traslazione.

Dato che gran parte dei fenomeni naturali descritti da una variabile regionalizzata ha caratteristiche spiccatamente non stazionarie, si cerca di ricondurre l'analisi alle più favorevoli condizioni di stazionarietà attraverso due approcci: il primo considera la *deriva* della FA, il secondo lavora sugli *accrescimenti*.

Modelli con deriva

La FA presenta un *trend*, vale a dire una variazione sistematica della variabile più o meno accentuata. Tale comportamento può essere modellato considerando la media della variabile calcolata su vicinaggi mobili all'interno dell'area di indagine.

La FA $Z(x)$ viene decomposta in due componenti, una deterministica rappresentata dalla sua media $m(x) = \mathrm{E}[Z(x)]$ e detta anche *deriva*, l'altra stocastica definita *residuo* $Y(x) = Z(x) - m(x)$. La media viene modellizzata da una funzione polinomiale:

$$m(\mathbf{x}) = \sum_{k=0}^{K} a_k f_k(\mathbf{x}) = a_0 + a_1 x + a_2 y + a_3 x^2 + a_4 y^2 + a_5 xy + \dots \qquad (1.14)$$

dove a_k sono coefficienti da stimare e $f_k(\mathbf{x})$ sono monomi di grado crescente delle coordinate spaziali dei punti. La stazionarietà è garantita dal fatto che la media dei residui risulta costante poiché nulla.

Modello intrinseco di ordine k

Piuttosto che considerare in ciascun punto la variabile $Z(x)$ è conveniente riferirsi all'accrescimento $Z(x+\vec{h}) - Z(x)$, ovvero considerare, in maniera più generale, una *combinazione lineare autorizzata*[3] di ordine k della FA $Z(x)$:[4]

$$Z(x) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{1.15}$$

dove λ_i sono coefficienti da stimare.

Affinché tale combinazione possa definirsi stazionaria è necessario che venga soddisfatto il vincolo:

$$\sum_{i=1}^{n} \lambda_i f_l(x_i) = 0 \qquad con \quad l = 0, \ldots, k \tag{1.16}$$

con $f^l(x_i)$ monomi di grado l delle coordinate spaziali degli n punti. Si può riconoscere che l'accrescimento stazionario $Z(x+\vec{h}) - Z(x)$ è una combinazione lineare autorizzata di ordine 0 a coefficienti rispettivamente 1 e -1.

1.5 Variogramma sperimentale e teorico

La stima della funzione variogramma viene effettuata sulla base dei dati provenienti dal campionamento del fenomeno oggetto di studio o da altri dati puntuali indiretti. A partire dall'espressione (1.10), scelta una direzione[5] principale, il calcolo del variogramma *sperimentale* viene effettuato sugli n punti del campo dei quali si hanno informazioni attraverso uno stimatore:

$$\gamma^*(h) = \frac{1}{2n} \sum_{i=1}^{n} [Z(x_i + h) - Z(x_i)]^2 \tag{1.17}$$

[3] per *autorizzata* si intende che la combinazione lineare ammette varianza finita.

[4] anche detta FAI-k: Funzione Aleatoria Intrinseca di ordine k

[5] le *variabili regionalizzate* possono presentare differente variabilità spaziale in funzione della direzione presa per il calcolo del variogramma sperimentale, ciò è dovuto alla presenza di *anisotropie* tipiche dei fenomeni naturali.

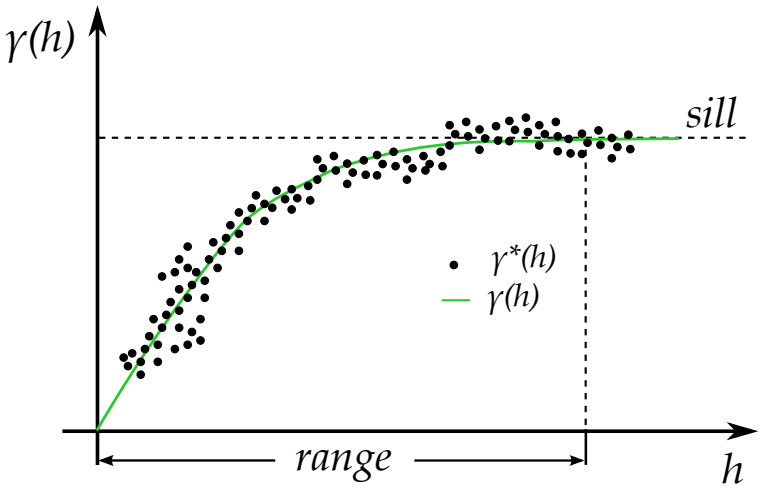

Figura 1.3: Variogramma sperimentale con valore di soglia fittato dal variogramma teorico

L'andamento del variogramma sperimentale in funzione di h esprime la variabilità del fenomeno considerato e ne rivela il comportamento stocastico, suggerendo di fatto il modello di FA da adottare.

Il variogramma sperimentale può essere analizzato ponendo l'attenzione in due zone principali del grafico:

vicino l'origine

Per $h \to 0$ è possibile distinguere tre diversi comportamenti riferibili alla regolarità della *VR*:

1. *parabolico*. Elevata regolarità e continuità della *VR*;
2. *lineare*. Continuità, ma non regolarità della *VR*;
3. *discontinuo*. *VR* con andamento irregolare. La variabilità tra due punti vicini è molto elevata, si ha il cosiddetto *effetto pepita* (o *nugget*) nell'origine.

per h crescenti

All'aumentare di h il variogramma aumenta di valore ed evolve secondo due forme:

1. raggiunge un valore di soglia (*sill*). FA stazionaria;
2. aumenta indefinitamente. FA non stazionaria.

Data l'evidente scarsa praticità dei variogrammi sperimentali si utilizzano funzioni analitiche che vengono assunte comunemente per descriverne il comportamento. Tali funzioni devono rispondere a determinate proprietà matematiche integrando le principali caratteristiche dei variogrammi:

1. è positiva $\gamma(h) \geq 0$

2. per $h = 0$ si ha $\gamma(0) = 0$

3. è una funzione pari $\gamma(h) = \gamma(-h)$

4. quando la FA è stazionaria $\gamma(h) = C(0) - C(h)$

5. cresce all'infinito meno rapidamente rispetto ad h^2:

$$\lim_{h \to 0} \frac{\gamma(h)}{h^2} = 0 \qquad quando \quad h \to 0 \tag{1.18}$$

6. deve essere tale da dar luogo a *combinazioni lineari autorizzate*.

La letteratura geostatistica propone alcune funzioni matematiche (v. Fig. 1.4) adatte a descrivere il comportamento del variogramma sperimentale; una volta scelta la funzione opportuna, tramite il *fitting* dei parametri (solitamente con tecniche ai minimi quadrati) si ottiene il *variogramma teorico*; il risultato è una funzione continua e derivabile $\forall h \in [0, +\infty)$.

Modello Pepitico

$$\gamma(h) = c(1 - \delta(h)) \qquad dove \quad \delta(h) = \begin{cases} 1 & in \quad r = 0 \\ 0 & \forall r \neq 0 \end{cases} \tag{1.19}$$

$\delta(h)$ è la funzione impulso di *Dirac*, il parametro c esprime il *sill* del variogramma. Questo modello esprime una discontinuità nell'origine.

Modello Sferico

$$\gamma(h) = \begin{cases} c\left(\frac{3}{2}\frac{h}{r} - \frac{1}{2}\frac{h^3}{r^3}\right) & per \quad 0 \leq h \leq r \\ c & per \quad h > r \end{cases} \tag{1.20}$$

r e c sono paraemtri del modello ed esprimono rispettivamente *range* e *sill*. Il comportamento vicino l'origine è lineare[6].

Modello Esponenziale

$$\gamma(h) = c\left(1 - e^{-\frac{h}{r}}\right) \qquad per \quad h \geq 0 \tag{1.21}$$

In questo caso il valore c di *sill* è raggiunto asintoticamente, pertanto il *range* risulta infinito. Nella pratica il valore di *range* viene considerato alla distanza per la quale viene raggiunto il 95% del *sill*, cioè pari a circa $3r$.

[6] vicino l'origine si può ritenere il termine cubico $\frac{h^3}{r^2}$ trascurabile.

Modello Gaussiano

$$\gamma(h) = c\left(1 - e^{-\frac{h^2}{r^2}}\right) \qquad per \quad h \geq 0 \tag{1.22}$$

come per il modello esponenziale il *range* viene calcolato considerando la distanza alla quale viene raggiunto il 95% del *sill*, ovvero a circa $\sqrt{3}r$.

Modelli Potenza

Sono modelli che non prevedono la presenza di un valore di soglia, ma crescono indefinitamente e vengono utilizzati per FA non stazioanrie.

$$\gamma(h) = ch^{\beta} \tag{1.23}$$

dove $c > 0$ e $\beta \in (0, 2)$. Con $\beta = 1$ si ottiene un modello lineare, spesso utilizzato nella pratica per FA intrinseche di ordine 0 (v. Castiglioni 2009).

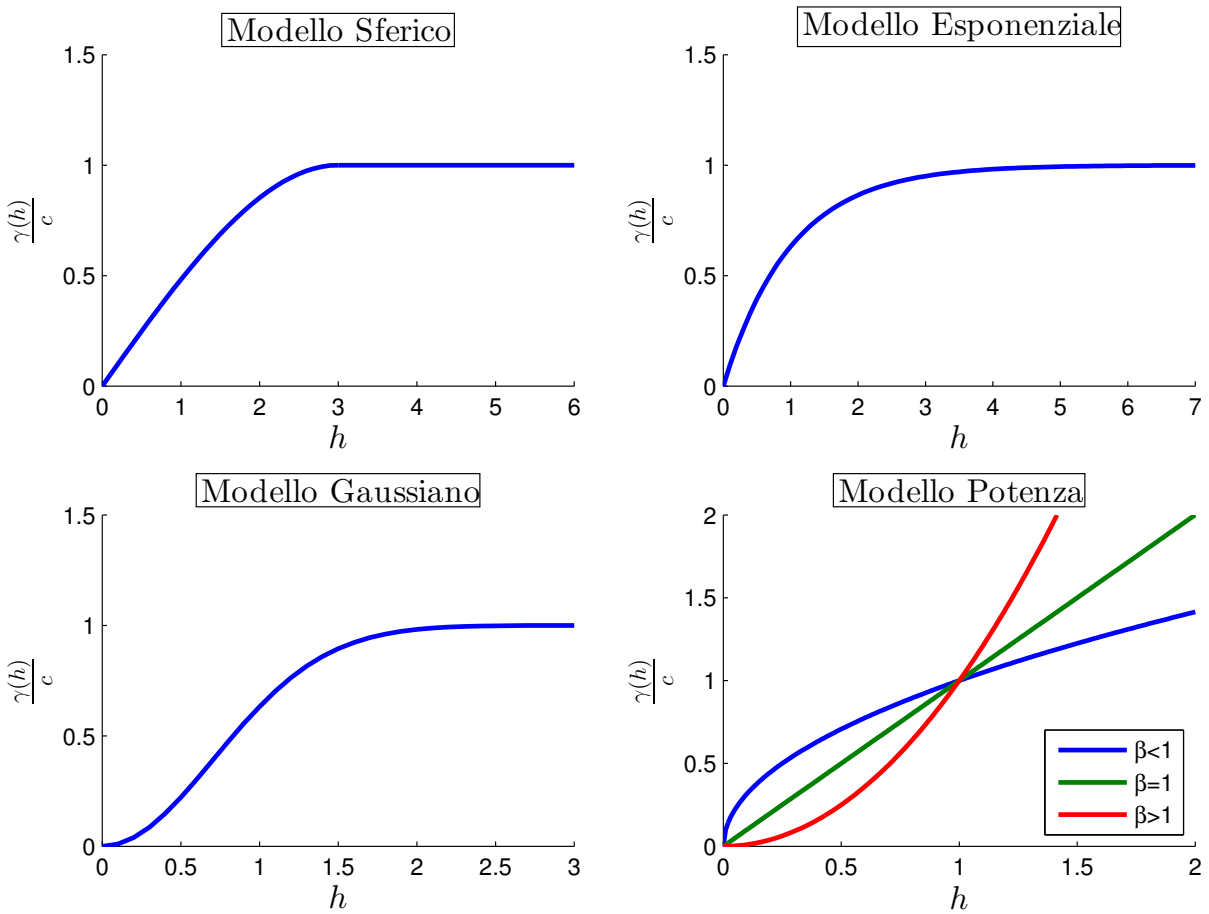

Figura 1.4: Grafici qualitativi dei variogrammi teorici elementari più utilizzati in geostatistica

Le funzioni appena citate fanno riferimento ad un comportamento ideale nell'origine del variogramma sperimentale; se, al contrario, questo dovesse presentare un effetto pepitico o *nugget* (discontinuità nell'origine) è essenziale fittare un modello con un parametro aggiuntivo che esprime lo *shift* verso l'alto del variogramma nell'origine:

$$g(h) = nug + \gamma(h) \tag{1.24}$$

dove per nug si intende una costante positiva e γ uno dei modelli sopra proposti.

Ci si può trovare nella condizione in cui nessuno dei precedenti modelli interpreti al meglio il variogramma sperimentale, purtuttavia si può riconoscere nel suo andamento una serie di strutture note, dette anche a *strutture annidate* riconducibili a variogrammi teorici elementari. Sfruttando il principio della *sovrapposizione degli effetti* si può dimostrare che, se $Z(x)$ è composta dalla somma di m variabili indipendenti e stazionarie[7] $Z_i(x)$ con $i = 1, \ldots, m$ aventi ciscuno variogramma $\gamma_i(h)$, anche il suo variogramma $\gamma(h)$ può essere espresso come:

$$\gamma(h) = \sum_{i=1}^{m} \gamma_i(h) \tag{1.25}$$

1.6 Tecniche di stima

Nei problemi geostatistici è di fondamentale importanza riuscire a prevedere il comportamento della variabile regionalizzata in punti dello spazio in cui non è nota. Tale procedura viene detta *stima* e viene normalmente utilizzata per la produzione di carte tematiche.

L'operazione di stima ha carattere locale, in quanto solitamente non riguarda le caratteristiche generali (o globali), ma studia la variabilità alla piccola scala, in un dominio ristretto detto *vicinaggio di stima*. Gli stimatori più adatti e più usati per questo tipo di operazione sono quelli *lineari*, cioè sono combinazioni lineari della variabile nei punti noti del campo situati nelle vicinanze del punto da stimare.

Consideriamo n punti del campo x_i con $i = 1, \ldots, n$, disposti casualmente, nei quali risulta nota la variabile regionalizzata $z(x_i)$; sia $Z(x)$ la FA stazionaria assunta per descrivere in senso probabilistico il fenomeno di studio e siano $C(h)$ e $\gamma(h)$ rispettivamente le funzioni covarianza e variogramma. È possibile esprimere la stima della in un punto

[7] dette anche *componenti spaziali* di $Z(x)$.

qualsiasi x_0 del campo come:[8]

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{1.26}$$

dove i λ_i sono costanti incognite chiamate *ponderatori*.

Alla stima (1.26) è associato l'*errore di stima* $\varepsilon(x_0)$, risultante dalla differenza tra la variabile *stimata* e quella *esatta*:

$$\varepsilon(x_0) = Z(x_0) - Z^*(x_0) = Z(x_0) - \sum_{i=1}^{n} \lambda_i Z(x_i) \tag{1.27}$$

Si dice che lo stimatore è *corretto* se risulta:[9]

$$\mathrm{E}[\varepsilon(x_0)] = \mathrm{E}\Big[Z(x_0) - \sum_{i=1}^{n} \lambda_i Z(x_i)\Big] = 0 \tag{1.28}$$

ovvero

$$\mathrm{E}[Z(x_0)] - \sum_{i=1}^{n} \lambda_i \mathrm{E}[Z(x_i)] = 0 \tag{1.29}$$

per l'ipotesi fatta di stazionarietà $\mathrm{E}[Z(x)] = m = cost. \quad \forall x \in S$ la (1.29) si riduce a:

$$m\Big[1 - \sum_{i=1}^{n} \lambda_i\Big] = 0$$

ottenendo infine l'importante condizione sui ponderatori:

$$\sum_{i=1}^{n} \lambda_i = 1 \tag{1.30}$$

Non avendo a disposizione la funzione di densità per la VA, non la si può avere nemmeno per l'errore di stima. L'accuratezza della stima nel punto x_0 può, però, essere valutata a partire dalla *varianza di stima* $\sigma_s^2(x_0)$, intimamente legata alla covarianza e alla funzione variogramma della VA.

Riprendendo la definizione di varianza data dalla (1.5) si ha per la varianza di stima:

$$\mathrm{Var}_s(x_0) = \sigma_s^2(x_0) = \mathrm{E}\Big[\Big(Z(x_0) - \sum_{i=1}^{n} \lambda_i Z(x_i)\Big)^2\Big] \tag{1.31}$$

[8] si ricorda che il simbolo '*' sta ad indicare la variabile *stimata*

[9] In letteratura anglosassone viene detto *Unbiased Estimator*

sviluppando il quadrato del binomio a secondo membro si ottiene:

$$\sigma_s^2(x_0) = \mathrm{E}\Big[Z^2(x_0) + \sum_i \sum_j \lambda_i \lambda_j Z(x_i) Z(x_j) - 2 \sum_i \lambda_i Z(x_i) Z(x_0)\Big]$$

tenendo conto delle (1.4),(1.9) e delle proprietà di linearità dell'operatore *valore atteso* si ha:

$$\begin{aligned}\sigma_s^2(x_0) &= C(0) + \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j C(x_i, x_j) - 2 \sum_{i=1}^{n} \lambda_i C(x_i, x_0) \\ &= C(0) + \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \big(C(0) - \gamma(h_{i,j})\big) - 2 \sum_{i=1}^{n} \lambda_i \big(C(0) - \gamma(h_{i,0})\big)\end{aligned}$$

tenendo in considerazione la (1.30) si ottiene infine:[10]

$$\sigma_s^2(x_0) = 2 \sum_{i=1}^{n} \lambda_i \gamma_{i,0} - \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j \gamma_{i,j} \tag{1.32}$$

la (1.32) mette in evidenza che la varianza di stima σ_s^2 non dipende dalla covarianza ma solo dal variogramma, quindi può essere rimossa l'ipotesi di stazionarietà globale introdotta in precedenza.

I *ponderatori* λ_i, che consentono di effettuare la stima della variabile regionalizzata in punti non noti, sono parametri del modello di stima e possono essere individuati con diverse tecniche che fanno riferimento a due approcci principali:

1. *tecniche deterministiche*: non tengono in considerazione le caratteristiche dell'approccio probabilistico e si basano su impostazioni prettamente geometriche;

2. *tecniche geostatistiche*: meglio conosciute come *Kriging*,[11] sono tecniche di regressione che permettono di individuare i ponderatori minimizzando la varianza di stima.

1.6.1 Tecniche deterministiche

Le principali tecniche deterministiche, come già accennato, sono quelle che prescindono dal contesto aleatorio e calcolano i ponderatori in funzione della distribuzione esclusivamente geometrica dei punti noti nel vicinaggio di stima.

[10] le notazioni $\gamma(h_{i,j})$ o più brevemente $\gamma_{i,j}$ equivalgono alla notazione $\gamma(x_i, x_j)$ ed indica la funzione variogramma calcolata tra i punti i e j distanti $h_{i,j}$.

[11] in francese *Krigeage* nome coniato da G. Matheron in onore dell'ingegnere sudafricano D. G. Krige che per primo le formalizzò.

Una prima tecnica, storicamente utilizzata soprattutto in ambito idrologico, è la *poligonazione di Thiessen* (o tassellazione di Dirichelet-Voronoi), che prevede la costruzione di *tasselli* aventi come vertici gli ortocentri dei triangoli formati congiungendo i punti noti. All'interno di ogni tassello o poligono ricadrà una sola misura della variabile regionalizzata e si assume che tale valore sia costante sull'intero dominio V_i individuato dal poligono i-esimo. I ponderatori quindi banalmente discriminano tra l'appartenenza all'uno o all'altro tassello, ciò può essere espresso come:

$$\lambda_i = \begin{cases} 1 & se \quad x_0 \in V_i \\ 0 & altrimenti \end{cases} \tag{1.33}$$

I limiti evidenti di questa tecnica risiedono in primo luogo nell'assumere costante il valore della variabile all'interno di un' area più o meno estesa (ipotesi piuttosto stringente) e nella discontinuità della stima: l'andamento a gradini fa sì che tra i punti di confine, ai margini dei tasselli, si può passare da un valore all'altro della variabile in infinitesimi di distanza.

La seconda tecnica che più realisticamente può interpretare il comportamento della variabile regionalizzata, superando le incongruenze della poligonazione di Thiessen, è quella della *distanza inversa*. I ponderatori da attribuire ai campioni compresi entro il vicinaggio di stima hanno un peso proporzionale all'inverso della distanza euclidea d_i fra il punto noto x_i e il punto da stimare x_0, elevata ad un esponente di ordine $p \in \mathbb{R}^+$, ovvero:

$$\lambda_i = \frac{k}{d_i^p} \qquad con \quad k = \frac{1}{\sum_j \frac{1}{d_j^p}} \quad e \quad i, j = 1, \ldots, n \tag{1.34}$$

dove la potenza p, detta *ordine della distanza* può variare a seconda della variabile regionalizzata da interpolare: scelte diverse della potenza p implicano stime diverse. Si è soliti nella applicazioni idrologiche assumere $p = 2$, ma l'ordine è del tutto arbitrario. Bisogna considerare, però, che per $p \to 0$ i pesi tendono ad assumere valori simili anche per distanze elevate fino al limite in cui la stima si riduce ad una media aritmetica (tutti i ponderatori sono uguali), al contrario se $p \to \infty$ si ottiene il caso limite in cui la stima si riduce ad una poligonazione.

1.6.2 Tecniche geostatistiche (o Kriging)

Le tecniche geostatistiche di stima prevedono la formulazione della FA in un processo aleatorio associato alla variabile regionalizzata attraverso lo studio degli indici statistici come, per esempio, il variogramma. Tra le tecniche di stima principali spiccano quelle del *Kriging*, le quali in letteratura anglosassone vengono definite *B.L.U.E.* (Best Linear

Unbiased Estimator) che sta ad indicare come il Kriging sia in grado di fornire stime non deviate ed esatte.

Le tecniche di "krigaggio" sono diverse a seconda della condizioni in cui opera la FA (se stazionarie o non stazionarie), ma discendono tutte da un logica matematica comune. Come si è visto, la varianza di stima σ_s^2, ricavata nell'equazione (1.32), esprime la qualità e la correttezza della stima, per cui è immediato porsi come obiettivo la ricerca di quei ponderatori che la rendano minima, nel senso di una maggiore precisione. Il problema appena esposto può essere risolto utilizzando il metodo di ottimizzazione dei *moltiplicatori di Lagrange* sotto opportune condizioni di vincolo.[12]

Kriging Ordinario

Si tratta di applicare il metodo dei moltiplicatori di Lagrange all'equazione (1.32) sotto il vincolo di stima corretta (o non deviata) data dalla (1.30). Sia $\mu \in \mathbb{R}$ il parametro lagrangiano incognito, la funzione lagrangiana è data da:

$$\mathcal{L}(\vec{\lambda}, \mu) = \underbrace{2\sum_{i=1}^{n} \lambda_i \gamma_{i,0} - \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j \gamma_{i,j}}_{\sigma_s^2} + \underbrace{2\mu\Big(1 - \sum_{i=1}^{n} \lambda_i\Big)}_{vincolo} \tag{1.35}$$

si tratta, quindi, di risolvere il sistema di minimo vincolato dato da:

$$\nabla \mathcal{L}(\vec{\lambda}, \mu) = 0 \tag{1.36}$$

ovvero

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \lambda_i} = 2\gamma_{i,0} - 2\displaystyle\sum_{j=1}^{n} \lambda_j \gamma_{i,j} - 2\mu = 0 \qquad per \quad i = 1, \ldots, n \\ \dfrac{\partial \mathcal{L}}{\partial \mu} = 2\big(1 - \displaystyle\sum_{i=1}^{n} \lambda_i\big) = 0 \end{cases}$$

si ottiene quindi il sistema lineare di $n+1$ equazioni in $n+1$ incognite:

$$\begin{cases} \displaystyle\sum_{j=1}^{n} \lambda_j \gamma_{i,j} + \mu = \gamma_{i,0} \qquad per \quad i = 1, \ldots, n \\ \displaystyle\sum_{j=1}^{n} \lambda_j = 1 \end{cases} \tag{1.37}$$

[12] le le tecniche geostatistiche di stima introdotte in questo capitolo sono quelle che serviranno nelle applicazioni. Si ricorda che esiste anche il "Kriging semplice", che prescinde dall'ottimizzazione con i moltiplicatori di Lagrange.

che scritto in forma esplicita

$$\begin{aligned}
\lambda_1\gamma_{1,1} + \lambda_2\gamma_{1,2} + \cdots + \lambda_n\gamma_{1,n} + \mu &= \gamma_{1,0} \\
\lambda_1\gamma_{2,1} + \lambda_2\gamma_{2,2} + \cdots + \lambda_n\gamma_{2,n} + \mu &= \gamma_{2,0} \\
&\vdots \\
\lambda_1\gamma_{n,1} + \lambda_2\gamma_{n,2} + \cdots + \lambda_n\gamma_{n,n} + \mu &= \gamma_{n,0} \\
\lambda_1 + \lambda_2 + \cdots + \lambda_n &= 1
\end{aligned} \tag{1.38}$$

il sistema precedente scritto in termini matriciali diventa:

$$\begin{bmatrix}
\gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,n} & 1 \\
\gamma_{2,1} & \gamma_{2,2} & \cdots & \gamma_{2,n} & 1 \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
\gamma_{n,1} & \gamma_{n,2} & \cdots & \gamma_{n,n} & 1 \\
1 & 1 & 1 & 1 & 0
\end{bmatrix}
\begin{bmatrix}
\lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_n \\ \mu
\end{bmatrix}
=
\begin{bmatrix}
\gamma_{1,0} \\ \gamma_{2,0} \\ \vdots \\ \gamma_{n,0} \\ 1
\end{bmatrix} \tag{1.39}$$

La (1.39) mostra che la matrice dei coefficienti $\mathbf{\Gamma}$ e il termine noto $\mathbf{g_0}$ dipendono unicamente dalla funzione variogramma, che, come si è visto, è possibile calcolare (teoricamente[13]) in tutti i punti del campo; per cui, se le posizioni i-esime dei punti sono distinte, il sistema ammette sempre un'unica soluzione. Concludendo è possibile calcolare per ogni punto x_0 una n-pla $\vec{\lambda}$ di ponderatori λ_i risolvendo il sistema lineare (1.39):

$$\boldsymbol{\lambda} = \mathbf{\Gamma}^{-1}\mathbf{g_0}$$

Kriging Universale

Nel caso in cui la variabile regionalizzata presenti un *trend* (o *deriva*) sicuramente la FA associata deve essere interpretata tramite modelli non stazionari. In questo caso nell'applicare il metodo dei moltiplicatori di Lagrange bisogna tener conto delle mutate condizioni di stazionarietà e applicare alla (1.35) un ulteriore vincolo che tenga conto del fenomeno della deriva.

Si assume che le media possa essere descritta da una funzione polinomiale del tipo (2.3a) che calcolata nel punto x_0 di stima vale:

$$m(x_0) = \sum_{k=1}^{K} a_k f_k(x_0) \tag{1.40}$$

[13] s'intende che si ha già a disposizione un *variogramma teorico*.

poiché il kringig è uno stimatore esatto, utilizzando la proprietà (1.29), la media può essere scritta come combinazione degli stessi coefficienti λ_i, ovvero

$$m(x_0) = \sum_{i=1}^{n} \lambda_i \mathrm{E}\big[Z(x_i)\big] = \sum_{i=1}^{n} \lambda_i m(x_i) \tag{1.41}$$

combinando la (1.41) e la (1.40) si ha

$$\sum_{i=1}^{n} \lambda_i \Big(\sum_{k=1}^{K} a_k f_k(x_i) \Big) = \sum_{k=1}^{K} a_k \Big(\sum_{i=1}^{n} \lambda_i f_k(x_i) \Big)$$

da cui risulta la seconda condizione di vincolo

$$\sum_{i=1}^{n} \lambda_i f_k(x_i) = f_k(x_0) \tag{1.42}$$

Sia $\vec{\mu} = (\dots, \mu_k, \dots)$ con $k = 0, \dots, K$ un vettore di moltiplicatori di Lagrange, si può, quindi, formalizzare la funzione lagrangiana come:

$$\mathcal{L}(\vec{\lambda}, \vec{\mu}) = \sigma_s^2 + \underbrace{2\mu_0 \Big(1 - \sum_{i=1}^{n} \lambda_i\Big)}_{vincoloI} + \underbrace{2 \sum_{k=1}^{K} \mu_k \Big(f_k(x_0) - \sum_{i=1}^{n} \lambda_i f_k(x_i) \Big)}_{vincoloII} \tag{1.43}$$

i parametri incogniti $(\lambda_1, \dots, \lambda_n, \mu_0, \dots, \mu_K)$ vengono individuati risolvendo il sistema di minimo vincolato:

$$\nabla \mathcal{L}(\vec{\lambda}, \vec{\mu}) = 0$$

ovvero

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial \lambda_i} = 2\gamma_{i,0} - 2 \displaystyle\sum_{j=1}^{n} \lambda_j \gamma_{i,j} - 2\mu_0 - 2 \sum_{k=1}^{K} \mu_k f_k(x_i) = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \mu_0} = 2\big(1 - \displaystyle\sum_{i=1}^{n} \lambda_i\big) = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \mu_k} = -2\big(f_k(x_0) - \displaystyle\sum_{i=1}^{n} \lambda_i f_k(x_i)\big) = 0 \end{cases} \qquad \begin{array}{r} per \quad i = 1, \dots, n \\ e \quad k = 1, \dots, K \end{array}$$

il sistema precedente scritto in termini matriciali diventa:

$$\begin{bmatrix} \gamma_{1,1} & \dots & \gamma_{1,n} & 1 & f_1(x_1) & \dots & f_K(x_1) \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \gamma_{n,1} & \dots & \gamma_{n,n} & 1 & f_1(x_n) & \dots & f_K(x_n) \\ 1 & \dots & 1 & 0 & \dots\dots\dots & & 0 \\ f_1(x_1) & \dots & f_1(x_n) & 0 & \dots\dots\dots & & 0 \\ \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ f_k(x_1) & \dots & f_k(x_n) & 0 & \dots\dots\dots & & 0 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_n \\ \mu_0 \\ \mu_1 \\ \vdots \\ \mu_k \end{bmatrix} = \begin{bmatrix} \gamma_{1,0} \\ \vdots \\ \gamma_{n,0} \\ 1 \\ f_1 x_0 \\ \vdots \\ f_k(x_0) \end{bmatrix} \tag{1.44}$$

il quale evidenzia una struttura a blocchi del tipo:

$$\begin{bmatrix} \mathbf{\Gamma} & \mathbf{F} \\ \mathbf{F^T} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\mu} \end{bmatrix} = \begin{bmatrix} \mathbf{g_0} \\ \mathbf{f_0} \end{bmatrix} \tag{1.45}$$

Anche in questo caso si osserva come la matrice dei coefficienti e il termine noto siano formati da grandezze note, ovvero il variogramma e le funzioni monomie delle coordinate spaziali, il cui ordine massimo K è del tutto arbitrario. La soluzione, formata dal vettore dei ponderatori λ_i e dai moltiplicatori di Lagrange μ_i, unica se i punti x_i sono distinti, viene individuata invertendo la matrice a blocchi nella (1.45).

Capitolo 2

Tecniche di analisi multivariata

Nelle applicazioni pratiche ci si trova spesso a dover analizzare ed utilizzare uno o più *dataset* di n campioni caratterizzati ciascuno da k variabili del tipo

$$X = \begin{bmatrix} x_{1,1} & \dots & x_{1,k} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,k} \end{bmatrix} \tag{2.1}$$

Se le k variabili fossero incorrelate tra loro ciascuna di esse potrebbe essere esaminata singolarmente utilizzando k modelli monovariabili di regressione semplice. Sfortunatamente, però, è raro che le colonne di una matrice di dati risultino tra loro incorrelate: un insieme di variabili incorrelate è praticamente ottenibile solo tramite una trasformazione delle variabili osservate (v. Pollice 2011).

È utile, specialmente nella pratica geostatistica, ridurre la dimensione del *dataset* in modo da poter riferire la variabilità totale espressa dalle k variabili originarie ad un numero inferiore (o al massimo uguale) a k di variabili *sintetiche*[1], tutte incorrelate tra loro, ottenute tramite opportune trasformazioni: questa operazione viene fatta attraverso l'*analisi delle componenti principali*[2].

Se si hanno a disposizione due dataset degli stessi campioni, dal numero delle variabili k il primo ed m il secondo, del tipo

$$X = \begin{bmatrix} x_{1,1} & \dots & x_{1,k} \\ \vdots & \ddots & \vdots \\ x_{n,1} & \dots & x_{n,k} \end{bmatrix} \qquad Y = \begin{bmatrix} y_{1,1} & \dots & y_{1,m} \\ \vdots & \ddots & \vdots \\ y_{n,1} & \dots & y_{n,m} \end{bmatrix} \tag{2.2}$$

[1] il termine indica che le nuove variabili non hanno significato fisico ed hanno il compito di *sintetizzare* l'informazione proveniente dalle variabili iniziali.

[2] in inglese viene chiamata con l'acronimo *PCA*(Principal Component Analysis).

lo studio delle relazioni di interdipendenza tra i due gruppi di variabili X e Y costituisce l'obiettivo dell'*analisi della correlazione canonica*[3]. Così come nello studio delle componenti principali, l'analisi della correlazione canonica, a partire dalla rappresentazione originaria delle unità statistiche fornita dai due gruppi di variabili, effettua una sintesi tale da rappresentarle attraverso due nuovi gruppi di variabili sintetiche che siano incorrelati al loro interno e massimamente correlati tra loro.

2.1 Analisi delle componenti principali

Sia X un universo campionario k-dimensionale del tipo (2.1) e siano le statistiche campionarie *media* e *varianza* date rispettivamente da

$$\bar{X} = \frac{1}{n} X^{\mathrm{T}} u_n = \left[\bar{X}_1, \ldots, \bar{X}_k\right]^{\mathrm{T}} \tag{2.3a}$$

$$S = \frac{1}{n}\left(X - u_n \bar{X}^{\mathrm{T}}\right)^{\mathrm{T}}\left(X - u_n \bar{X}^{\mathrm{T}}\right) \tag{2.3b}$$

essendo u_n un vettore n-dimensionale di elementi unitari.

Lo scopo dell'analisi è quello di cercare un vettore e_1 k-dimensionale tale che possa essere espresso come combinazione lineare delle unità statistiche di X, ovvero

$$e_1 = X a_1 \tag{2.4}$$

dove a_1 rappresenta un vettore k-dimensionale di termini costanti, e che abbia varianza massima[4]. Essendo $\mathrm{Var}\{e_1\} = a_1^T S a_1$ si osserva che la varianza di e_1 è una funzione crescente delle costanti contenute in a_1. Affinché il problema sia *ben posto* si deve richiedere ad a_1 che sia di norma[5] unitaria, cioè che $||a_1||_2 = \sqrt{a_1^T a_1} = 1$. In definitiva la *prima componente principale* e_1 si trova risolvendo il sistema di massimo vincolato con l'ausilio del metodo dei moltiplicatori di lagrange. La funzione lagrangiana è del tipo

$$\mathcal{L}(a_1, \lambda) = \underbrace{a_1^{\mathrm{T}} S a_1}_{\mathrm{Var}\{e_1\}} + \lambda \underbrace{(1 - a_1^{\mathrm{T}} a_1)}_{vincolo} \tag{2.5}$$

e viene risolto ponendo

$$\nabla \mathcal{L}(a_1, \lambda) = 0 \tag{2.6}$$

[3] in inglese *CCA*(Canonical Correlation Analysis)

[4] questa condizione riflette la considerazione fatta nel preambolo in cui si richiedeva che la maggior parte della varianza venga *spiegata* da un numero inferiore di variabili.

[5] per norma s'intende norma 2.

ovvero

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial a_1} = 2Sa_1 - 2\lambda a_1 = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \lambda} = 1 - a_1^{\mathrm{T}} a_1 = 0 \end{cases} \tag{2.7}$$

dalla prima equazione del sistema (2.7) si evince che il problema iniziale si riduce ad un problema agli autovalori, infatti a_1 rappresenta un autovettore a norma unitaria di S associato all'autovalore λ. Se nella stessa equazione moltiplichiamo ambo i membri per a_1^{T} si ottiene:

$$a_1^{\mathrm{T}} S a_1 = \mathrm{Var}\{e_1\} = \lambda \underbrace{a_1^{\mathrm{T}} a_1}_{=1} = \lambda \tag{2.8}$$

La (2.8) mette in evidenza che la prima componente principale $e_1 = Xa_1$ risulta essere univocamente definita dall'autovettore a_1 associato al più grande autovalore di S indicato con λ_1.

Per la *seconda componente principale*, definita come

$$e_2 = Xa_2 \tag{2.9}$$

si tratta di inserire nella (2.5) un' ulteriore condizione di vincolo in cui venga esplicata la necessità di incorrelazione tra le variabili sintetiche e_1 ed e_2, ossia che $a_2^{\mathrm{T}} S a_1 = 0$. La lagrangiana assume la forma

$$\mathcal{L}(a_2, \lambda, \mu) = \underbrace{a_2^{\mathrm{T}} S a_2}_{\mathrm{Var}\{e_2\}} + \lambda \underbrace{(1 - a_2^{\mathrm{T}} a_2)}_{vincolo \quad I} - \underbrace{\mu a_2^{\mathrm{T}} S a_1}_{vincolo \quad II} \tag{2.10}$$

da cui impnendo $\nabla \mathcal{L} = 0$ si ottiene il sistema

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial a_2} = 2Sa_2 - 2\lambda a_2 - \mu S a_1 = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \lambda} = 1 - a_2^{\mathrm{T}} a_2 = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \mu} = a_2^{\mathrm{T}} S a_1 = 0 \end{cases} \tag{2.11}$$

se nella prima equazione del sistema (2.11) si moltiplicano ambo i membri per a_1^{T} si ha

$$2a_1^{\mathrm{T}} S a_2 - 2\lambda a_1^{\mathrm{T}} a_2 - \mu a_1^{\mathrm{T}} S a_1 = 0 \tag{2.12}$$

ed essendo $Sa_1 = \lambda_1 a_1$, sostituendo nella condizione $a_2^{\mathrm{T}} S a_1 = 0$, si ottiene che $a_2^{\mathrm{T}}\big(\lambda_1 a_1\big) = \lambda_1 a_2^{\mathrm{T}} a_1 = 0$ solo se $a_2^{\mathrm{T}} a_1 = 0$, poiché come già dimostrato λ_1 non può essere nullo. Per cui dalla (2.12) si deduce che $\mu = 0$ e che anche in questo caso la ricerca del vettore dei

coefficienti a_2 della seconda componente principale risulta dal problema agli autovalori

$$Sa_2 = \lambda a_2$$

inoltre, premoltiplicando la prima del sistema (2.11) per a_2^{T}, si ha che

$$a_2^{\mathrm{T}} S a_2 = \mathrm{Var}\{e_2\} = \lambda a_2^{\mathrm{T}} a_2 = \lambda \tag{2.13}$$

l'autovalore che definisce la seconda componente principale, detto λ_2 non può che essere il secondo autovalore più grande di S.

Con questa tecnica, se S è definita positiva, è possibile definire un numero di componenti principali pari al rango k di S. Infatti se $a_1, \ldots, a_k$ sono gli autovettori associati agli autovalori $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_k$ della matrice S, le k componenti principali possibili sono date dalle combinazioni lineari

$$\begin{aligned} e_1 &= Xa_1 \\ e_2 &= Xa_2 \\ &\vdots \\ e_k &= Xa_k \end{aligned}$$

e godono della proprietà

$$\sum_{j=1}^{k} \mathrm{Var}\{e_j\} = \sum_{j=1}^{k} \mathrm{Var}\{X_j\} = \mathrm{tr}(S) \tag{2.14}$$

ovvero che la somma delle varianze delle componenti principali è uguale alla somma delle varianze campionarie delle k colonne di X.

Di fondamentale importanza è l'interpretazione delle componenti principali in base ai coefficienti che le mettono in relazione con le variabili rilevate. Sia $a_j = (a_{1j}, \ldots, a_{kj})^{\mathrm{T}}$ con $j = 1, \ldots, k$ il vettore dei coefficienti associato all'autovalore j-esimo di S, la j-esima componente principale e_j può essere scritta come:

$$e_j = a_{1j}X_1 + \cdots + a_{kj}X_k \tag{2.15}$$

in cui si osserva che ciascuna componente principale è espressione di una combinazione lineare delle k variabili di partenza. Quindi il coefficiente a_{hj} con $h = 1, \ldots, k$ può essere interpretato come il peso della variabile X_h nella determinazione della componente j-esima: quanto maggiore è a_{hj} in valore assoluto, tanto più e_j *interpreta* la variabile X_h.

È possibile dimostrare (v. Pollice 2011) che il coefficiente di correlazione campionario tra e_j e X_h è dato da

$$\rho_{e_j,X_h} = \frac{\text{Cov}(e_j, X_h)}{\sqrt{\text{Var}\{e_j\}\sigma_h^2}} = \frac{\lambda_j a_{hj}}{\sqrt{\lambda_j \sigma_h^2}} = a_{hj}\sqrt{\frac{\lambda_j}{\sigma_h^2}} \tag{2.16}$$

Dalla (2.16) si ottiene la quota di variabilità di X_h *spiegata* dalla j-esima componente principale. È evidente che al diminuire di λ_j anche la correlazione tra X_h e e_j si riduce, fino ad annullarsi se gli autovalori sono molto piccoli in valore assoluto; allora, tenendo conto della proprietà (2.14), si può attribuire a ciascun λ_j il significato di misura della quota di varianza totale spiegata dalla j-esima componente principale. Se gli ultimi $q+1, \ldots, k$ autovalori possono ritenersi trascurabili rispetto ai primi q, l'indice che misura la quota di varianza spiegata dalle prime q componenti principali può essere calcolato come

$$I_q = \frac{\lambda_1 + \cdots + \lambda_q}{\lambda_1 + \cdots + \lambda_k} = \frac{\lambda_1 + \cdots + \lambda_q}{\text{tr}(S)} \tag{2.17}$$

Generalmente viene fissata[6] una soglia I^* che indica la frazione della varianza totale che si vuole venga spiegata dalle prime q componenti principali in modo che q sia il più piccolo intero tale che $I_q > I^*$.

2.2 Analisi della correlazione canonica

Nella stessa ottica che ha portato allo studio delle componenti principali, il metodo proposto per affrontare lo studio delle interdipendenze tra due gruppi di variabili del tipo (2.2) consiste nell'individuare un doppio sistema di variabili latenti che riproducano la correlazione tra i due gruppi di variabili osservate al netto di quella presente al loro interno. In altri termini dalla rappresentazione originaria delle unità statistiche, fornita dai due gruppi di variabili rilevate, si vuole ottenere una sintesi tale da rappresentarle tramite due nuovi gruppi di variabili artificiali che siano incorrelate al loro interno e massimamente correlate tra loro.

Siano X e Y due gruppi di variabili di n campioni statistici del tipo (2.2) di dimensioni rispettivamente $n \times k$ e $n \times m$ e siano

$$\begin{aligned} \bar{X} &= \frac{1}{n} X^{\mathrm{T}} u_n = \left(\bar{X_1}, \ldots, \bar{X_k}\right)^{\mathrm{T}} \\ \bar{Y} &= \frac{1}{n} Y^{\mathrm{T}} u_n = \left(\bar{Y_1}, \ldots, \bar{X_m}\right)^{\mathrm{T}} \end{aligned} \tag{2.18}$$

[6] la soglia è del tutto arbitraria e dipende dagli obiettivi dell'analisi, spesso, specialmente in applicazioni idrologiche, si utilizzano valori soglia dell'ordine del $90 - 95\%$.

le loro medie campionarie e

$$\begin{aligned} S_X &= \frac{1}{n}\big(X - u_n X^{\mathrm{T}}\big)^{\mathrm{T}}\big(X - u_n X^{\mathrm{T}}\big) \\ S_Y &= \frac{1}{n}\big(Y - u_n Y^{\mathrm{T}}\big)^{\mathrm{T}}\big(Y - u_n Y^{\mathrm{T}}\big) \\ S_{XY} &= \frac{1}{n}\big(X - u_n X^{\mathrm{T}}\big)^{\mathrm{T}}\big(Y - u_n Y^{\mathrm{T}}\big) = S_{YX}^{\mathrm{T}} \end{aligned} \tag{2.19}$$

le loro matrici di varianza e covarianza campionarie. Lo scopo è quello di produrre variabili sintetiche U e V espresse come combinazione lineare delle variabili originarie, per cui si ha

$$U = Xa \qquad V = Yb \tag{2.20}$$

ed essendo

$$\begin{aligned} \mathrm{Var}\{U\} &= a^{\mathrm{T}} S_X a \\ \mathrm{Var}\{V\} &= b^{\mathrm{T}} S_Y b \\ \mathrm{Cov}\{U, V\} &= a^{\mathrm{T}} S_{XY} b = b^{\mathrm{T}} S_{YX} a \end{aligned} \tag{2.21}$$

il coefficiente di correlazione tra U e V dato da

$$\rho_{UV} = \frac{\mathrm{Cov}\{U, V\}}{\sqrt{\mathrm{Var}\{U\}\,\mathrm{Var}\{V\}}} = \frac{a^{\mathrm{T}} S_{XY} b}{\sqrt{a^{\mathrm{T}} S_X a b^{\mathrm{T}} S_Y b}} \tag{2.22}$$

deve essere tale da massimizzare la correlazione tra le due variabili sintetiche. Operativamente deve essere massimizzato il numeratore della (2.22) con la condizione di unitarietà delle varianze delle due variabili U e V:

$$\begin{cases} \max\limits_{a}\big\{a^{\mathrm{T}} S_{XY} b\big\} \\ \max\limits_{b}\big\{a^{\mathrm{T}} S_{XY} b\big\} \\ a^{\mathrm{T}} S_X a = 1 \\ b^{\mathrm{T}} S_Y b = 1 \end{cases} \tag{2.23}$$

Come nell'analisi *PCA* la risoluzione del problema viene condotta utilizzando il metodo dei moltiplicatori di Lagrange; detti $\frac{\mu}{2}$ e $\frac{\eta}{2}$ i moltiplicatori, la funzione lagrangiana assume la forma

$$\mathcal{L}(a, b, \mu, \eta) = a^{\mathrm{T}} S_{XY} b + \frac{\mu}{2}(1 - a^{\mathrm{T}} S_X a) + \frac{\eta}{2}(1 - b^{\mathrm{T}} S_Y b) \tag{2.24}$$

quindi si tratta di risolvere il sistema di massimo vincolato

$$\nabla \mathcal{L}(a, b, \mu, \eta) = 0$$

che scritto in forma esplicita diventa

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial a} = S_{XY}b - \mu S_X a = 0 \\ \dfrac{\partial \mathcal{L}}{\partial b} = S_{XY}^{\mathrm{T}} a - \eta S_Y b = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \mu} = 1 - a^{\mathrm{T}} S_X a = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \eta} = 1 - b^{\mathrm{T}} S_Y b = 0 \end{cases} \tag{2.25}$$

moltiplicando per a^{T} la prima e per b^{T} la seconda delle (2.25) e ricordando che $a^{\mathrm{T}} S_{XY} b = b^{\mathrm{T}} S_{YX} a$ si ottene che $\mu = \eta$; il sistema diventa

$$\begin{cases} a = \dfrac{1}{\mu} S_X^{-1} S_{XY} b \\ b = \dfrac{1}{\mu} S_Y^{-1} S_{YX} a \\ a^{\mathrm{T}} S_X a = 1 \\ b^{\mathrm{T}} S_Y b = 1 \end{cases} \tag{2.26}$$

dalle (2.26) sostituendo la prima nella seconda e viceversa si ottiene un doppio problema agli autovalori, infatti

$$\begin{aligned} a\mu^2 &= S_X^{-1} S_{XY} S_Y^{-1} S_{YX} a \\ b\mu^2 &= S_Y^{-1} S_{YX} S_X^{-1} S_{XY} b \end{aligned} \tag{2.27}$$

posto $E_1 = S_X^{-1} S_{XY} S_Y^{-1} S_{YX}$, $E_2 = S_Y^{-1} S_{YX} S_X^{-1} S_{XY}$ e $\mu^2 = \lambda$ un autovalre di E_1 e E_2 il sistema assume la forma classica di un problema agli autovalori

$$\begin{aligned} E_1 a &= \lambda a \\ E_2 b &= \lambda b \end{aligned} \tag{2.28}$$

Le due matrici E_1 ed E_2, per essere invertibili, devono necessariamente essere di rango pieno rispettivamente $\mathrm{rg}(S_X) = k$ e $\mathrm{rg}(S_Y) = m$, mentre $\mathrm{rg}(S_{XY}) = (S_{YX}) = r \leq \min(k, m)$, quindi anche E_1 ed E_2 avranno rango r. Si osserva infine che $\mathrm{Cov}\{U, V\} = a^{\mathrm{T}} S_{XY} b = \sqrt{\lambda}$, per cui, detto λ_1 l'autovalore più grande in modulo del sistema (2.27) al quale sono associati gli autovettori a_1 e a_2, vengono dette *prime componenti canoniche*

$$U_1 = X a_1 \qquad V_1 = Y b_1 \tag{2.29}$$

la correlazione reciproca è espressa da

$$\rho_{U_1V_1} = \mathrm{Cov}\{U_1, V_1\} = \sqrt{\lambda_1} \tag{2.30}$$

Con la stessa metodologia si possono individuare le componenti canoniche successive imponendo che i due gruppi di variabili siano massimamente correlati tra loro con l'aggiunta del vincolo di incorrelazione al loro interno. Siano a_2 e b_2 due vettori di dimensioni rispettivamente k ed m di elementi costanti incogniti; le *seconde componenti canoniche* vengono definite come combinazioni lineari del tipo

$$U_2 = Xa_2 \qquad V_2 = Yb_2 \tag{2.31}$$

dove a_1 e b_2 vengono determinati in modo tale che la correlazione tra U_2 e V_2 deve essere massima e che, sia U_1 con U_2, sia V_1 con V_2 devono necessariamente essere indipendenti tra loro. Ciò si traduce nel sistema di massimo vincolato

$$\begin{cases} \max\limits_{a_2}\{a_2^{\mathrm{T}} S_{XY} b_2\} \\ \max\limits_{b_2}\{a_2^{\mathrm{T}} S_{XY} b_2\} \\ \mathrm{Var}\{U_2\} = a_2^{\mathrm{T}} S_X a_2 = 1 \\ \mathrm{Var}\{V_2\} = b_2^{\mathrm{T}} S_Y b_2 = 1 \\ \mathrm{Cov}\{U_1, U_2\} = a_1^{\mathrm{T}} S_X a_2 = 0 \\ \mathrm{Cov}\{V_1, V_2\} = b_1^{\mathrm{T}} S_Y b_2 = 0 \end{cases} \tag{2.32}$$

per i quattro vincoli vengono definite altrettante costanti moltiplicatori di Lagrange: $\frac{\mu}{2}, \frac{\eta}{2}, \phi, \psi$, pertanto la lagrangiana assume la forma

$$\mathcal{L}(a_2, b_2, \mu, \eta, \phi, \psi) = a_2^{\mathrm{T}} S_{XY} b_2 + \frac{\mu}{2}(1 - a_2^{\mathrm{T}} S_X a_2) + \frac{\eta}{2}(1 - b_2^{\mathrm{T}} S_Y b_2) - \phi a_1^{\mathrm{T}} S_X a_2 - \psi b_1^{\mathrm{T}} S_Y b_2 \tag{2.33}$$

si tratta quindi di risolvere il sistema

$$\nabla \mathcal{L}(a_2, b_2, \mu, \eta, \phi, \psi) = 0 \tag{2.34}$$

che scritto in forma esplicita

$$\begin{cases} \dfrac{\partial \mathcal{L}}{\partial a_2} = S_{XY} b_2 - \mu S_X a_2 - \phi S_X a_1 = 0 \\ \dfrac{\partial \mathcal{L}}{\partial b_2} = S_{YX} a_2 - \eta S_X b_2 - \psi S_Y b_1 = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \mu} = 1 - a_2^{\mathrm{T}} S_X a_2 = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \eta} = 1 - b_2^{\mathrm{T}} S_Y b_2 = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \phi} = a_1^{\mathrm{T}} S_X a_2 = 0 \\ \dfrac{\partial \mathcal{L}}{\partial \psi} = b_1^{\mathrm{T}} S_Y b_2 = 0 \end{cases} \tag{2.35}$$

moltiplicando la prima equazione per a_1^{T} e la seconda per b_1^{T}, tenendo conto dei vincoli, le prime due equazioni diventano

$$\phi = a_1^{\mathrm{T}} S_{XY} b_2 = b_2^{\mathrm{T}} S_{YX} a_1$$
$$\psi = b_1^{\mathrm{T}} S_{YX} a_2 = a_2^{\mathrm{T}} S_{XY} b_1$$

sostituendo nelle precedenti espressioni le (2.26) ricavate per le prime componenti canoniche si ha

$$\phi = \frac{1}{\sqrt{\lambda_1}} b_2^{\mathrm{T}} S_{YX} S_X^{-1} S_{XY} b_1$$
$$\psi = \frac{1}{\sqrt{\lambda_1}} a_2^{\mathrm{T}} S_{XY} S_Y^{-1} S_{YX} a_1$$

ricordando che $S_{YX} S_X^{-1} S_{XY} b_1 = \lambda_1 S_Y b_1$ e $S_{XY} S_Y^{-1} S_{YX} a_1 = \lambda_1 S_X a_1$ e sostituendo nelle precedenti si conclude che $\phi = \psi = 0$: nel calcolo delle seconde componenti principali ci si può quindi svincolare dalla condizione di incorrelazione riconducendo il sistema ad una risoluzione identica a quanto già visto in (2.25). Ponendo, infatti, $\mu^2 = \eta^2 = \lambda$ si ottiene

$$a_2 = \frac{1}{\sqrt{\lambda}} S_X^{-1} S_{XY} b_2$$
$$b_2 = \frac{1}{\sqrt{\lambda}} S_Y^{-1} S_{YX} a_2$$

sotituendo la prima nella seconda e visceversa si ottiene il doppio problema agli autovalori

$$\begin{aligned} E_1 a_2 &= \lambda a_2 \\ E_2 b_2 &= \lambda b_2 \end{aligned} \tag{2.36}$$

dove E_1 ed E_2 sono le stesse matrici prodotte per il calocolo delle prime componenti canoniche e, per quanto dimostrato, $\lambda = \lambda_2$ è il secondo maggiore autovalore delle matrici

E_1 ed E_2 a cui sono associati gli autovettori a_2 e b_2 tali da realizzare le combinazioni lineare (2.31) la cui correlazione è definita da

$$\rho_{U_2,V_2} = \sqrt{\lambda_2} \tag{2.37}$$

mentre risulta nullo il coefficiente di correlazione tra le componenti dello stesso gruppo.

È facile intuire che la stessa procedura può essere applicata in cascata con le successive componenti canoniche, in generale la h-esima coppia di componenti canoniche consiste nelle combinazioni lineri

$$U_h = Xa_h \qquad V_h = Yb_h \tag{2.38}$$

tali da essere massimamente correlate tra loro e tali che ciascuna variabile precedente $U_1, \ldots, U_{h-1}$, $V_1, \ldots, V_{h-1}$ risulta indipendente. Si ricava facilmente che a_h e b_h sono gli autovettori associati all'h-esimo autovalore λ_h delle matrici E_1 ed E_2.

Riassumendo, se $r = \mathrm{rg}(S_{XY}) \leq \min(k, m)$, l'*analisi della correlazione canonica* consiste nel trasformare le k colonne di X e le m colonne di Y in r coppie di vettori

$$(U_1, V_1), \ldots, (U_r, V_r) \tag{2.39}$$

generando le due matrici $U = (U_1, \ldots, U_r)$ e $V = (V_1, \ldots, V_r)$ le cui rispettive varianze sono unitarie e $\mathrm{Cov}\{U, V\} = \mathrm{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_r})$. Nelle applicazioni finalizzate ad analizzare l'interdipendenza tra due gruppi di variabili multidimensionali, l'analisi della correlazione canonica consente di ridurre il numero delle osservazioni attraverso variabili sintetiche, non correlate al loro interno e allo stesso tempo ciascuna variabile di un gruppo è "spiegata" dall'altro gruppo e viceversa; questa proprietà risulta molto efficace nelle applicazioni geostatistiche in ambito idrologico.

Capitolo 3

Tecniche di interpolazione spaziale applicate alla regionalizzazione idrologica

3.1 Analisi regionale di frequenza delle piene

La regionalizzazione del regime di frequenza degli estremi idrologici e meteorologici è ad oggi tra le tecniche comunemente usate per la stima dei deflussi fluviali e delle precipitazioni, in siti dove non sono disponibili o reperibili dati.

Data la difficoltà di descrivere le portate di piena attraverso modelli di tipo deterministico, vengono adottate schematizzazioni delle stesse come variabili aleatorie. Tale approccio permette di determinare la portata che può essere superata con una certa probabilità. In ambito idraulico/idrologico si preferisce sostituire al concetto di probabilità quello di *tempo di ritorno* $T(x)$, a cui è univocamente legato, definito come l'intervallo di tempo che mediamente intercorre tra il verificarsi di due successivi eventi in cui il valore x è uguagliato o superato. Detta $F_X(x)$ la funzione di distribuzione di probabilità cumulata dei massimi annuali di portata al colmo, si definisce tempo di ritorno il rapporto

$$T(x) = \frac{1}{1 - F_X(x)} = \frac{1}{F'_X(x)} \tag{3.1}$$

nel quale $F_X(x)$ prende il nome di probabilità di *non superamento*, mentre $F'_X(x)$ quello di probabilità di *superamento*. La probabilità espressa in termini di tempo di ritorno esprime più efficacemente la frequenza attesa con cui una certa portata viene eguagliata o superata, pertanto, nel dimensionamento dei manufatti, viene messa in luce la frequenza con cui questi possono trovarsi in condizioni critiche. Fissato il tempo di ritorno di progetto,

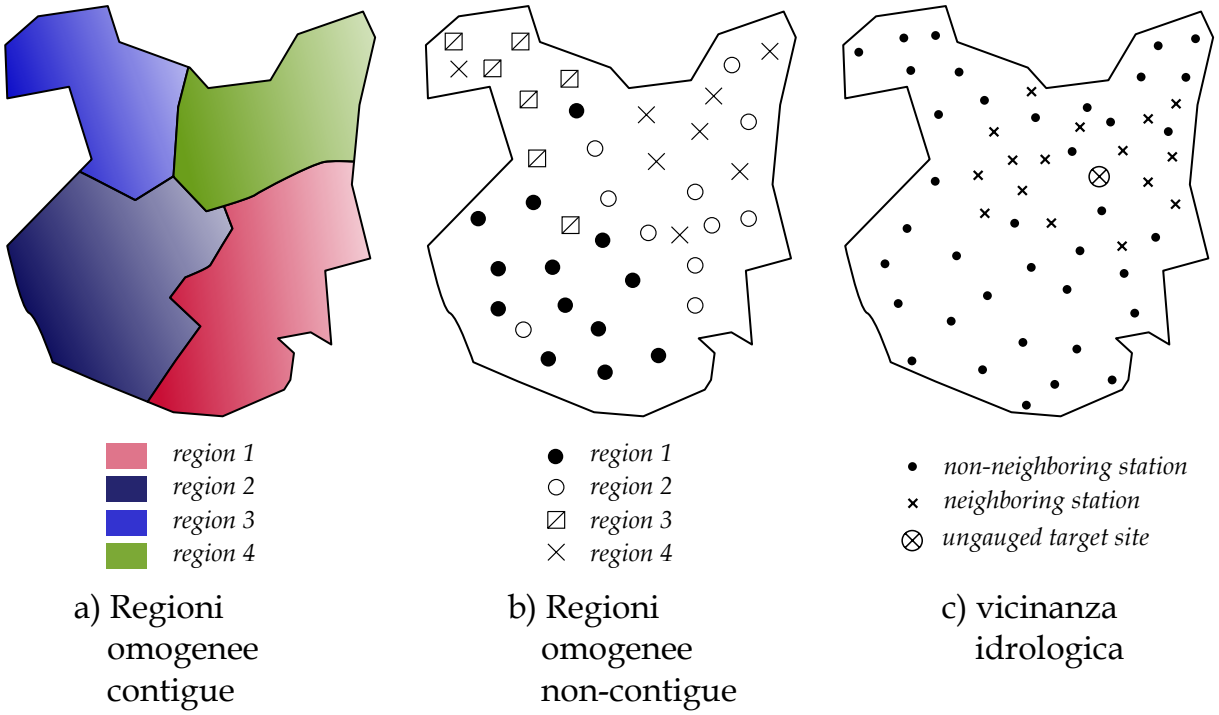

Figura 3.1: Vari approcci per l'individuazione delle regioni omogenee (v. Ouarda et al. 2001)

mediante il legame $x = x(T)$, è possibile valutare la portata associata che transiterà, con una certa probabilità, in una assegnata sezione fluviale (v. Brath 1995).

Il problema della stima di quantili di piena al colmo in sezioni fluviali prive di osservazioni idrometriche, è stato storicamente affrontato nell'ambito dell'analisi *regionale* di frequenza delle piene, e per il quale sono stati proposti diversi approcci da numerosi autori. La letteratura scientifica riporta esempi di applicazione di metodi di regionalizzazione, attraverso i quali i regimi di deflusso ed i volumi idrici disponibili in corrispondenza di sezioni fluviali non strumentate vengono valutati estendendo le informazioni idrometriche note in altre stazioni di misura. Tutte le tecniche regionali proposte in letteratura presentano come prerequisito la ricerca e l'individuazione di regioni *omogenee*, ovvero regioni nelle quali tutti i siti (stazioni di misura) in esse contenuti presentano caratteristiche comuni in termini di distribuzione in frequenza delle portate di piena. I metodi che portano all'identificazione di tali regioni sono drasticamente evoluti nel corso degli anni, passando da criteri *soggettivi*, con i quali le regioni omogenee venivano identificate su criteri esclusivamente geografici, ad *oggettivi*, basati sulla *cluster analysis* che identifica bacini con caratteristiche simili a partire da informazioni geomorfologiche e climatiche. Proprio per quanto riguarda i confini possono essere definite tre tipologie di classificazione di regioni omogenee (v. Fig. 3.1): *geograficamente contigue*, *geograficamente non contigue*, o per *vicinanza idrologica* (v. Ouarda et al. 2001).

Tuttavia l'individuazione di regioni omogenee risulta un problema ancora dibattuto dalla comunità scientifica, poiché non esiste una procedura univoca che porta alle determinazione dei confini delle regioni. Tale limite può essere agevolmente superato utilizzando tecniche d'interpolazione geostatistica, storicamente utilizzate in ambito minerario, e recentemente adattati nel contesto idrologico.

3.2 Tecniche di regionalizzazione geostatistica

Nell'ambito della regionalizzazione del regime di frequenza delle portate, le tecniche di interpolazione geostatistica basate sul *kriging* stanno avendo largo impiego nella valutazione delle variabili idrologiche in bacini non strumentati. Si riconoscono due metodi principali di *krigaggio*: un primo metodo, che prende il nome di Canonical Kriging, utilizza la variabile idrologica su supporto puntuale interpolata su un dominio *non* geografico; il secondo, detto Topological Kriging, adotta una variabile idrologica regolarizzata, ovvero su supporto areale, a scala di bacino. Recenti studi (v. Chokmani e Ouarda 2004; Castiglioni et al. 2011) hanno dimostrato che i due metodi sembrano avere caratteristiche di complementarietà: il Canonical Kriging lavora bene per bacini simili, ma lontani tra loro, come ad esempio i bacini di testata o i piccoli bacini montani; il Topological Kriging al contrario ottiene migliori performance per le sezioni fluviali a valle delle sezioni strumentate, ovvero per bacini di grandi dimensioni.

3.2.1 Canonical Kriging

Secondo l'impostazione proposta da *Chokmani e Ouarda*, piuttosto che interpolare i valori noti su un dominio *reale* di coordinate geografiche georeferenziate, viene definito uno spazio bidimensionale *sintetico*, le cui coordinate sono ottenute da una combinazione lineare di parametri caratteristici della morfologia, del clima, delle precipitazioni appartenenti al dominio geografico nel quale si vuole studiare la variabilità della grandezza presa in esame. Tale dominio viene anche definito *spazio dei descrittori geomorfoclimatici* e un tale approccio nella letteratura anglosassone viene identificato con la sigla PSBI (*Physiographical space-based interpolation*).

La definizione di un dominio *ad hoc* parte dalla considerazione che le due coordinate sintetiche devono formare un spazio ortonormale, ovvero che i punti appartenenti ai due assi devono essere incorrelati tra loro. Supponiamo di avere un dataset di variabili geomorfoclimatiche di un fissato numero di bacini, l'obiettivo è quello di ridurre il numero di variabili facendo sì che nel passaggio dalle vecchie alle nuove variabili la perdita d'informazione contenuta nei dati sia minima, ossia che le nuove variabili *spieghino* gran

parte della varianza delle variabili originarie. Questa operazione può essere agevolmente condotta utilizzando la procedura statistica di analisi multivariata che prende il nome di *analisi delle componenti principali* (PCA, Principal Component Analysis).

Studi di interpolazione della variabile idrometrica su domini sintetici ottenuti da PCA hanno dimostrato buoni rendimenti in termini di indici statistici dei modelli di interpolazione (v. Castiglioni, Castellarin e Montanari 2009; Chokmani e Ouarda 2004), sia se applicati agli indici di magra, sia ai valori di piena al colmo per assegnati tempi di ritorno. Di contro però, nella genesi del dominio, la tecnica PCA non tiene conto dell'influenza che possono avere i dati idrometrici, limite, questo, che può essere superato utilizzando la tecnica statistica multivariata di *analisi della correlazione canonica* (CCA, *Canonical Correlation Analysis*). Come già accennato nel Capitolo 2, a partire da due dataset che presentano caratteristiche di dipendenza reciproca, come ad esempio le variabili geomorfoclimatiche e quelle strettamente idrometriche, la CCA permette di accoppiare i dati in un sistema di variabili che siano massimamente correlate tra loro e incorrelate al loro interno. È possibile con questa procedura estrapolare due variabili artificiali U_1 e U_2 indipendenti tra loro, frutto di una combinazione lineare a coefficienti costanti delle variabili appartenenti al dataset X dei descrittori geomorfclimatici del tipo:

$$\begin{aligned} U_1 &= Xa_1 \\ U_2 &= Xa_2 \end{aligned} \tag{3.2}$$

dove i coefficienti a_1 e a_2 sono autovettori associati ai primi due maggiori autovalori della matrice

$$E_X = S_X^{-1} S_{XY} S_Y^{-1} S_{YX} \tag{3.3}$$

nella quale S rappresenta le matrici di varianza e covarianza campionarie dei due dataset che si vuole correlare. Nella (3.3) è evidente come la presenza della varianza del dataset Y, rappresentato ad esempio da dati idrometrici, influenzi la variabilità totale del fenomeno oggetto di studio; seppur discendente da una combinazione lineare a coefficienti costanti ottenuti da un problema agli autovalori, questa particolarità non è presente nell'analisi delle componenti principali.

Ammettendo che il dataset X sia composto come in (2.2) da k colonne, quindi k variabili geomorfoclimatiche, le (3.2) possono essere scritte (vedi Tabella 3.1) come:

$$\begin{aligned} U_1 &= a_{1,1}A + a_{1,2}LAT + a_{1,3}H_{max} + \cdots + a_{1,j}X_j + \cdots + a_{1,k}X_k \\ U_2 &= a_{2,1}A + a_{2,2}LAT + a_{2,3}H_{max} + \cdots + a_{2,j}X_j + \cdots + a_{2,k}X_k \end{aligned} \tag{3.4}$$

dove i vettori dei coefficienti, come dimostrato nel Capitolo 2, vengono individuati risol-

vendo un problema agli autovalori.

Tabella 3.1: Principali variabili geomorfoclimatiche utilizzate nel metodo PSBI

Variabili geomorfoclimatiche		
Variabile	Descrizione	Unità di misura
A	Area Bacino	(km^2)
LAT	Latitudine (centro)	$(^o)$
LONG	Longitudine (centro)	$(^o)$
L	Lunghezza asta principale	(km)
S	Pendenza Asta principale	$(m \cdot km^{-1})$
P	Perimetro del bacino	(km)
F_f	Fattore di forma	$(-)$
H_m	Quota media del bacino	(m)
H_{min}	Quota minima del bacino	(m)
H_{max}	Quota massima del bacino	(m)
S_m	Pendenza Media del bacino	$(\%)$
F_i	Frazione di bacino impermeabile	$(\%)$
F_{for}	Frazione di bacino a foreste	$(\%)$
D_d	Densità di drenaggio	$(km \cdot km^{-2})$
I_d	Indice di drenaggio del suolo	$(-)$
I_h	Indice di suolo idrologico	$(-)$
MAP	Precipitazione media annua	(mm)
MDP_T	Precipitazione massima gionaliera con assegnato tempo di ritorno	(mm)

In ultimo si evidenzia che le grandezze descrittrici di ciascun bacino, essendo grandezze fisiche, presentano caratteristiche dimensionali diverse; il problema può essere agevolmente superato operando una *standardizzazione* delle variabili in modo da omogeneizzare i dati:

$$x_1 = \frac{X_1 - u_n \bar{X}_1}{\sqrt{S_1^2}}; \ldots; x_k = \frac{X_k - u_2 \bar{X}_k}{\sqrt{S_k^2}} \tag{3.5}$$

Calcolati i vettori dei coefficienti a_1 e a_2, ciascun bacino è quindi univocamente identificato da una coppia (U_1^*, U_2^*) come un punto nel piano sintetico ortonormale U_1, U_2 (v. Fig. 3.2).

Creato lo spazio dei descrittori, a ciascun punto del piano è possibile associare nella terza dimensione un valore idrometrico corrispondente alle variabili del secondo dataset Y a disposizione; ad esempio nel caso oggetto di studio sono stati utilizzati valori di portata al colmo di piena con quattro differenti tempi di ritorno, anche se la dimensione del numero di variabili può essere qualunque[1].

[1] si ricorda che la dimensione dei due dataset in termini di numeri di righe deve essere la stessa, poiché si riferiscono allo stesso numero di bacini.

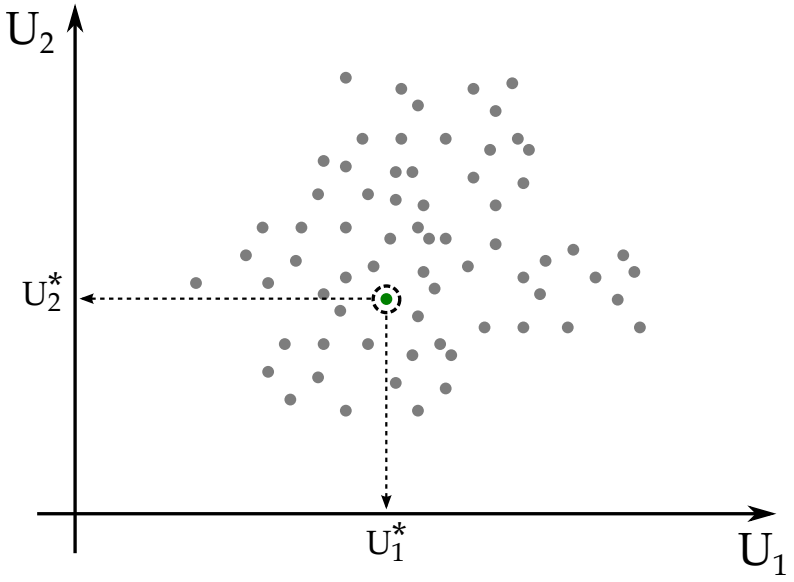

Figura 3.2: Dataset di n bacini nel piano sintetico U_1, U_2.

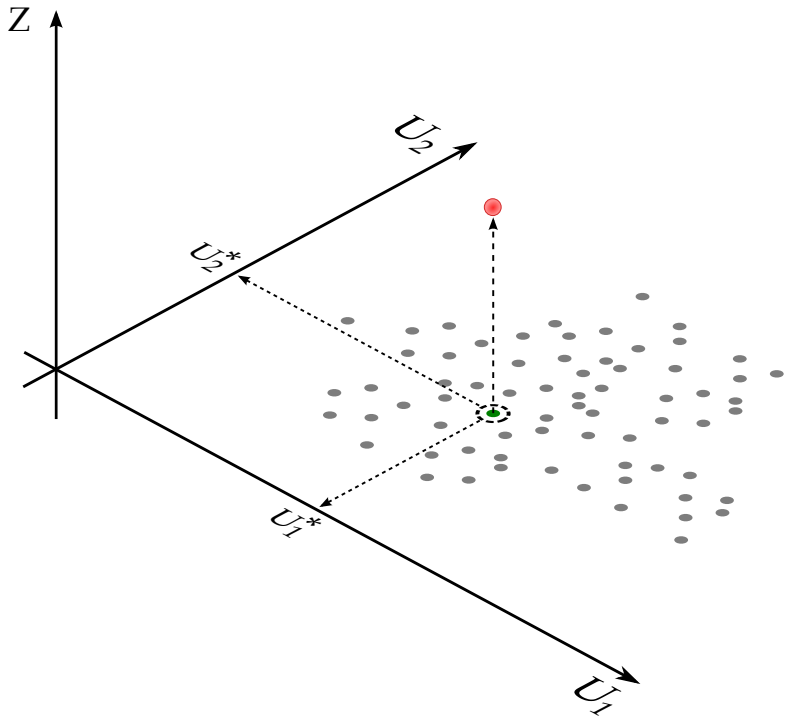

Figura 3.3: Variabile regionalizzata a supporto puntuale nel dominio dei descrittori geomorfoclimatici.

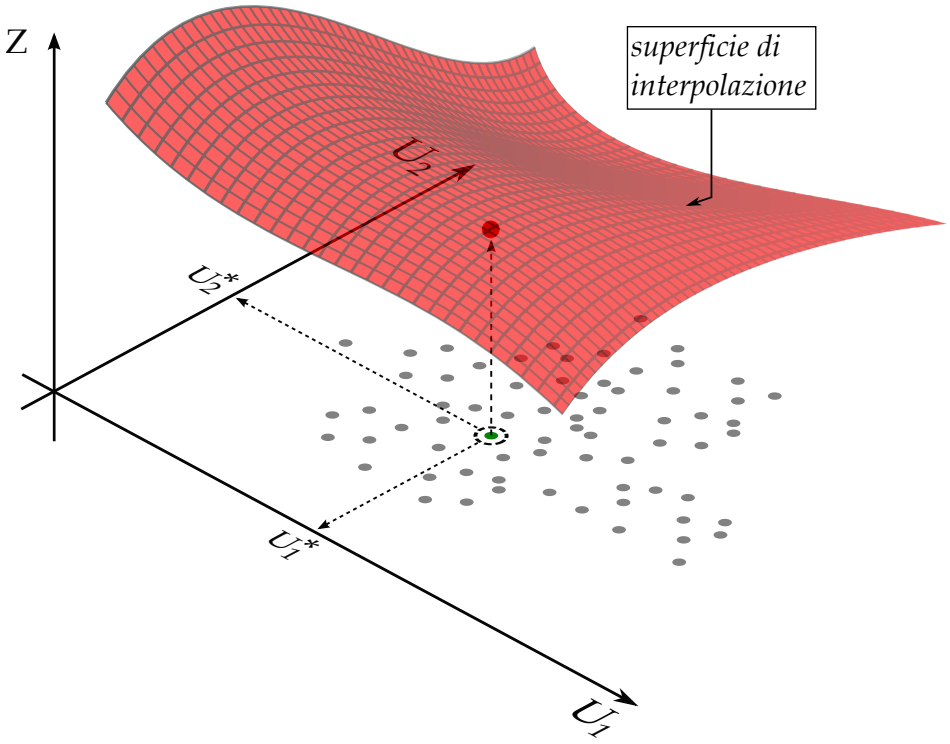

Figura 3.4: Grafico qualitativo di una superficie di interpolazione ottenuta con Canonical Kriging

A questo punto appare chiaro che le variabili idrometriche possono essere considerate a tutti gli effetti variabili regionalizzate (VR) (v. Fig. 3.3) a supporto puntuale consentendo l'interpolazione dei dati campionari tramite una delle tecniche di *kriging* esposte nel Capitolo 1.

Tutte le tecniche di *krigaggio* prevedono la conoscenza quantitativa e qualitativa della dispersione della variabile regionalizzata nel dominio; se per ciascun punto del piano U_1, U_2 è possibile associare un valore della VR, ad esempio il valore di portata al colmo ad assegnato tempo di ritorno, la dispersione può essere interpretata dal *variogramma sperimentale*, che, come già visto, può essere agevolmente calcolato con la (1.17). La presenza o meno di un *sill*, del *nugget* e in generale la regolarità del variogramma sperimentale sono parametri che suggeriscono il modello di funzione aleatoria da adottare, stazionaria o non stazionaria. Tramite una regressione ai minimi quadrati è possibile associare al variogramma sperimentale uno teorico, assimilabile alla funzione $\gamma(h)$, che ha il pregio di essere continua e derivabile quindi utilizzabile ai fini della stima.

Stabilita la natura della funzione aleatoria, il *krigaggio* consiste nell'interpolazione tramite una combinazione lineare a coefficienti costanti, anche detti ponderatori, della variabile regionalizzata

$$Z^*(x_0) = \sum_{i=1}^{n} \lambda_i Z(x_i)$$

dove i λ_i sono costanti determinabili con le due tecniche geostatistiche di *Kriginig Ordinario* risolvendo il sistema (1.39) e di *Kriging Universale* risolvendo il sistema (1.44). La scelta delle due tecniche dipende dal comportamento della funzione aleatoria e può ricadere su una delle due nel caso si manifesti stazionarietà o al contrario la presenza di una deriva. La scelta molto spesso non è banale e richiede un'analisi accurata dei variogrammi, inoltre non è detto che la distinzione tra FA stazionaria e non stazionaria sia riconoscibile a priori, ma potrebbero essere necessari dei tentativi prima di decretare il comportamento della FA.

Il risultato finale del *Canonical Kriging* è una superficie di interpolazione (v. Fig. 3.4) estesa sul dominio fissato dalla prima e seconda componente canonica. In conclusione, conoscendo le caratteristiche geomorfologiche e climatiche (come quelle descritte nella tabella 3.1) di un determinato bacino di cui, però, *non* si hanno a disposizione dati idrometrici, con gli strumenti messi a disposizione dall'analisi della corellazione canonica, si possono calcolare le coordinate sintetiche con le (3.2) e trovare il valore che la superficie di interpolazione assume in quel punto.

3.2.2 Topological Kriging

Il Topological Kriging o *Top-Kriging* è un metodo di interpolazione basato sulle tecniche tradizionali di kriging nel quale però si ammette che la variabile regionalizzata sia a supporto *non puntuale*, ovvero che il supporto abbia un'estensione areale finita. Pertanto la VR può definirsi *variabile regolarizzata* perché il passaggio da supporto puntuale ad areale comporta una maggiore regolarità nella VR (v. Raspa e Bruno 1994a). Il metodo presenta caratteristiche innovative nell'ambito delle tecniche di interpolazione geostatistiche applicate ai deflussi superficiali; recenti studi (v. Skøien, Merz e Blöschl 2006; Skøien et al. 2011) hanno dimostrato, infatti, che il *Top-Kriging* può essere ritenuto uno dei migliori approcci all'interpolazione di variabili idrometriche, proprio perché esiste una correlazione tra deflusso superficiale misurato in una determinata sezione fluviale ed estensione del bacino sotteso da quella sezione.

Supponiamo di avere una FA $Z(x)$, puntuale e stazionaria di media m, covarianza $C(h)$ e variogramma $\gamma(h)$, è possibile effettuare una trasformazione lineare definendo una nuova FA derivata da quella puntuale su un supporto esteso di area A come media integrale

$$\begin{aligned} Z_r(x) &= \frac{1}{A}\int_{A_x} Z(\xi)\, d\xi \\ Z_r(x+h) &= \frac{1}{A}\int_{A_{x+h}} Z(\zeta)\, d\zeta \end{aligned} \tag{3.6}$$

ipotizzando[2] che l'estensione del supporto non vari nel campo. Per la definizione di covarianza data con la (1.4) si ha che la funzione covarianza regolarizzata vale

$$C_r(h) = \mathrm{E}\big[Z_r(x+h)Z_r(x)\big] - m^2$$

e sostituendo le (3.6) nell'espressione della covarianza si ha

$$C_r(h) = \frac{1}{A^2}\int_{A_x}\int_{A_{x+h}} \mathrm{E}\big[Z(\xi)Z(\zeta)\big]\,d\xi d\zeta - m^2 = \frac{1}{A^2}\int_{A_x}\int_{A_{x+h}} C(\xi-\zeta)\,d\xi d\zeta \qquad (3.7)$$

per cui la covarianza regolarizzata può interpretarsi come il valor medio che la covarianza puntuale assume quando gli estremi del segmento $\xi-\zeta$ variano indipendentemente l'uno in A_x, l'altro in A_{x+h}. Essendo la funzione covarianza $C(h)$ dipendente solo dalla distanza tra i punti $\xi-\zeta$ e non dalla particolare posizione, si ha che anche $C_r(h)$ non dipende dal punto di appoggio e quindi anche $Z_r(x)$ può ritenersi stazionaria (v. Raspa e Bruno 1994a).

Per quanto riguarda il variogramma, partendo dall' espressione ricavata nella (1.10), per le FA stazionarie si può scrivere

$$\gamma_r(h) = \frac{1}{2}\mathrm{E}\big[\big(Z_r(x+h) - Z_r(h)\big)^2\big] \qquad (3.8)$$

anche qui, sostituendo le (3.6) e sviluppando il quadrato del binomio, si ha

$$\gamma_r(h) = \frac{1}{2}\left\{\frac{1}{A^2}\iint_{A_x}\mathrm{E}\big[Z(\xi)Z(\zeta)\big]\,d\xi d\zeta + \frac{1}{A^2}\iint_{A_{x+h}}\mathrm{E}\big[Z(\xi)Z(\zeta)\big]\,d\xi d\zeta\right\} - \frac{1}{A^2}\int_{A_x}\int_{A_{x+h}}\mathrm{E}\big[Z(\xi)Z(\zeta)\big]\,d\xi d\zeta$$

ovvero

$$\gamma_r(h) = \frac{1}{2}\left\{\frac{1}{A^2}\iint_{A_x} C(\xi-\zeta)\,d\xi d\zeta + \frac{1}{A^2}\iint_{A_{x+h}} C(\xi-\zeta)\,d\xi d\zeta\right\} - \frac{1}{A^2}\int_{A_x}\int_{A_{x+h}} C(\xi-\zeta)\,d\xi d\zeta \qquad (3.9)$$

[2]ipotesi che potrà in seguito essere rimossa.

ed osservando che $C(\xi-\zeta) = C(0) - \gamma(\xi-\zeta)$ e sostituendo nella precedente si ottiene

$$\gamma_r(h) = \frac{1}{A^2}\int\limits_{A_x}\int\limits_{A_{x+h}} \gamma(\xi-\zeta)\,d\xi d\zeta - \frac{1}{2}\left\{\frac{1}{A^2}\iint\limits_{A_x}\gamma(\xi-\zeta)\,d\xi d\zeta - \frac{1}{A^2}\iint\limits_{A_{x+h}}\gamma(\xi-\zeta)\,d\xi d\zeta\right\} \quad (3.10)$$

nella quale si può notare come al primo termine del secondo membro, rappresentante la varianza tra le due posizioni x e $x+h$, venga sottratta una quantità pari alla varianza interna ai rispettivi supporti, producendone, di fatto, una più bassa di quella della variabile puntuale: questo spiega perché le trasformazioni (3.6) generano un effetto *regolarizzante* sulla variabile regionalizzata.

Dalla (3.10) è banale poter espandere l'espressione del variogramma regolarizzato ricavato sotto l'ipotesi del supporto ad area costante a quella con supporto ad area qualsiasi (v. Skøien et al. 2011):

$$\gamma^r_{i,j} = \frac{1}{A_iA_j}\int\limits_{A_i}\int\limits_{A_j}\gamma(x_i-x_j)\,dx_idx_j - \frac{1}{2}\left\{\frac{1}{A_i^2}\iint\limits_{A_i}\gamma(x_i-x_j)\,dx_idx_j - \frac{1}{A_j^2}\iint\limits_{A_j}\gamma(x_i-x_j)\,dx_idx_j\right\} \quad (3.11)$$

La complessità di dover risolvere analiticamente integrali quadrupli porta, nella pratica, ad implementare uno schema di calcolo numerico a partire da una discretizzazione regolare dei supporti considerati. È importante tener conto che la maglia di discretizzazione delle aree deve essere la stessa, ma può anche capitare, specialmente nelle operazioni di stima di variabili idrometriche, di avere a che fare con bacini di estensioni areali che, abbracciando diversi ordini di grandezza, possono determinare una perdita di accuratezza soprattutto nei bacini più piccoli. Secondo quanto proposto da *Skøien*, viene fissata una griglia di base che per un determinato supporto può essere ridefinita aumentando la risoluzione finché all'interno non si ha un numero minimo accettabile di punti di calcolo (v. fig 3.5). Questa tecnica assicura che i punti usati per i supporti più grandi vengono riutilizzati nella discretizzazione di supporti più piccoli contenuti nei primi, com'è naturale se si tratta di sottobacini contenuti in bacini più grandi.

Lo stesso autore propone per il calcolo della varianza regolarizzata del *nugget* (effetto pepita) tra due supporti di diversa taglia A_1 e A_2 l'espressione

$$C_0(A_1,A_2) = \frac{1}{2}\left(\frac{C_0}{A_1} + \frac{C_0}{A_2} - \frac{2C_0\mathrm{Mis}(A_1\cap A_2)}{A_1A_2}\right) \quad (3.12)$$

dove $\mathrm{Mis}(A_1\cap A_2)$ rappresenta l'intersezione (area condivisa) tra le due aree A_1 e A_2 e possono manifestarsi due casi:

Aree in sovrapposizione

$$\text{Mis}(A_1 \cap A_2) = \min\{A_1, A_2\}$$

Aree disgiunte

$$\text{Mis}(A_1 \cap A_2) = 0$$

L'equazione (3.11) può essere utilizzata in fase di stima nel calcolo dei ponderatori risolvendo il sistema lineare (1.39), introdotto nel Capitolo 1 nell'ambito del Kriging Ordinario, nel quale i vari $\gamma_{i,j}$ vengono sostituiti con $\gamma^r_{i,j}$ ottenuti con la trasformazione di regolarizzazione.[3]

Come illustrato in Figura 3.6 si riporta uno schema tratto da Skøien (2006) che evidenzia l'influenza delle dimensioni delle aree e della struttura annidata dei bacini nella determinazione dei ponderatori λ_i. In tutti e tre gli esempi riportati i bacini hanno baricentro posto nella medesima distanza da quello per il quale si intende stimare la grandezza di interesse. Nella sottofigura a) viene mostrato come ad un bacino avente area maggiore sia assegnato un peso maggiore; nella sottofigura b) si evidenzia come un sottobacino appartenente ad un bacino più grande per il quale si vuole effettuare la stima ha un peso maggiore rispetto ad un bacino di pari estensione posto nelle vicinanze; mentre nella sottofigura c) si mostra l'effetto contrario a b), ovvero che l'influenza di un bacino su un suo sottobacino di cui si vuole effettuare la stima risulta maggiore rispetto ad un bacino di pari estensione posto nelle vicinanze. In enrambi i casi b) e c) rivelano l'effetto della struttura annidata dei bacini.

Concludendo si può affermare che nelle applicazioni geostatistiche ed in particolare in quelle idrologiche è possibile *regolarizzare* una variabile regionalizzata applicando una trasformazione lineare che consideri la non puntualità del supporto per poi procedere all'interpolazione finale che tiene conto dell'estensione areale e della mutua vicinanza, o sovrapposizione, dei bacini di cui si vuole effettuare la stima.

[3]Nelle ipotesi considerate si è fatto riferimento alla stazionarietà della FA che può risultare in alcuni casi stringente in presenza del fenomeno di deriva, problema che può essere risolto adottando modelli *quasi-stazionari*.

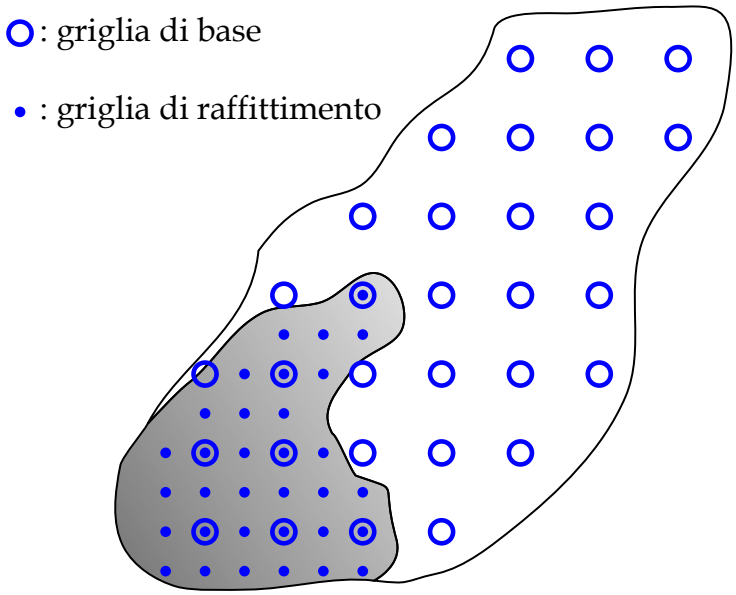

Figura 3.5: Discretizzazione di due aree in sovrapposizione di diversa estensione, i punti piccoli rappresentano lo schema di raffittimento per le aree più piccole

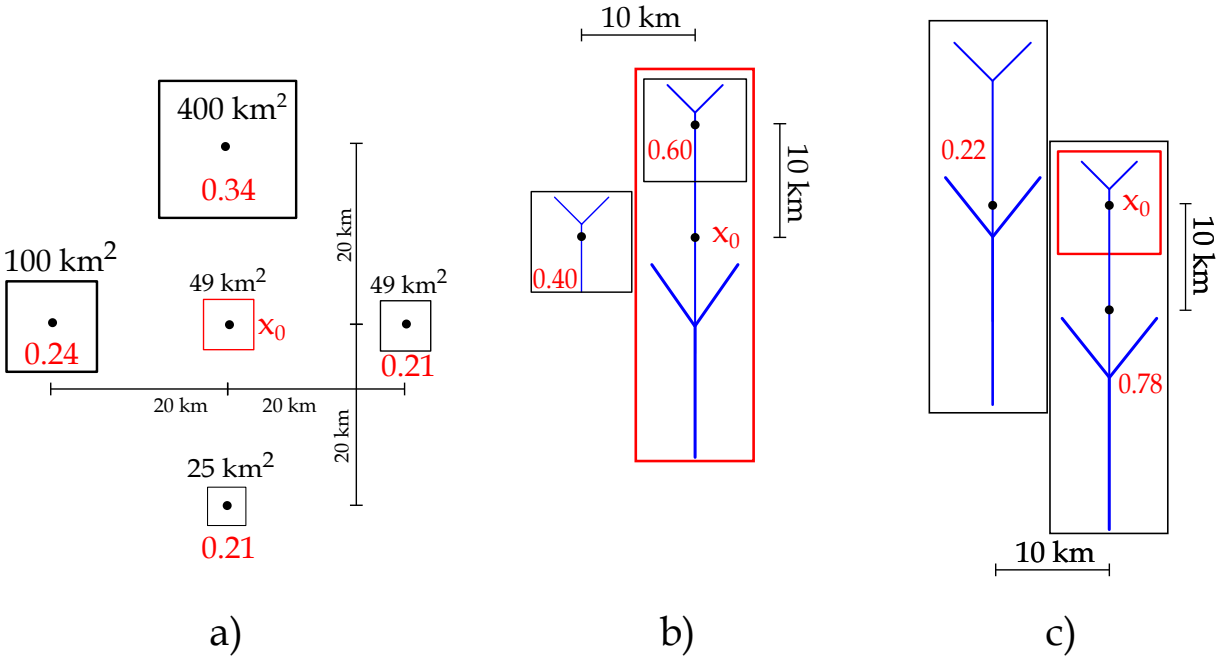

Figura 3.6: a) Effetto della dimensione dei supporti areali sui ponderatori λ_i (in rosso); b) e c) struttura annidata dei bacini

Capitolo 4

Area di studio

Gli studi sull'interpolazione spaziale di quantili di portata mediante l'impiego delle due principali tecniche di *kriging*, Canonical Kriging (CK) e Top-Kriging (TK), utilizzate nella presente dissertazione, sono basati su dati idrometrici, geomorfologici e climatici di una vasta area, composta da 61 bacini degli U.S.A. sud-orientali. Nello specifico l'area di studio si estende sui tre stati contigui di Florida, Georgia e Alabama i cui dati impiegati sono stati forniti dal USGS(*United States Geological Survey*) e sono il risultato di studi condotti su un'area più ampia che si estende dalla Georgia al North Carolina e pubblicati in un *report* del USGS nel 2009 (v. Gotvald, Feaster e Weaver 2009).

4.1 Inquadramento geomorfologico

La maggior parte dei bacini e le stazioni di misura delle corrispondenti sezioni di chiusura presenti nel dataset fornito, interessano prevalentemente il territorio della Georgia; dei restanti, tre sono situati in Alabama e uno in Florida. Da quanto si evince dal report dell'USGS, il territorio può essere suddiviso, seguendo le disposizioni dell'EPA (*Environmental Protection Agency*), in diverse regioni omogenee chiamate *ecoregioni*(v. Fig. 4.1). Queste sono individuate, attraverso un'analisi delle caratteristiche spaziali del territorio, da proprietà biotiche e abiotiche come la geologia, la morfologia, la vegetazione, il clima, i suoli e le precipitazioni. Tuttavia la zona oggetto di studio rappresenta solo una porzione dell'area mostrata in figura 4.1 ed è situata nell'immagine a sud-ovest (area cerchiata in blu in figura).

In questa regione è possibile osservare la presenza di 3 o 4 *ecoregioni* (v. legenda in fig. 4.1) che comprendono a sud zone di pianura costiera e zone pianeggianti del sudest; salendo poi verso nord, con l'aumento progressivo di quota, s'incontrano vaste aree pedemontane fino a toccare le catene montuose del *Blue Ridge*, nei territori di confine tra

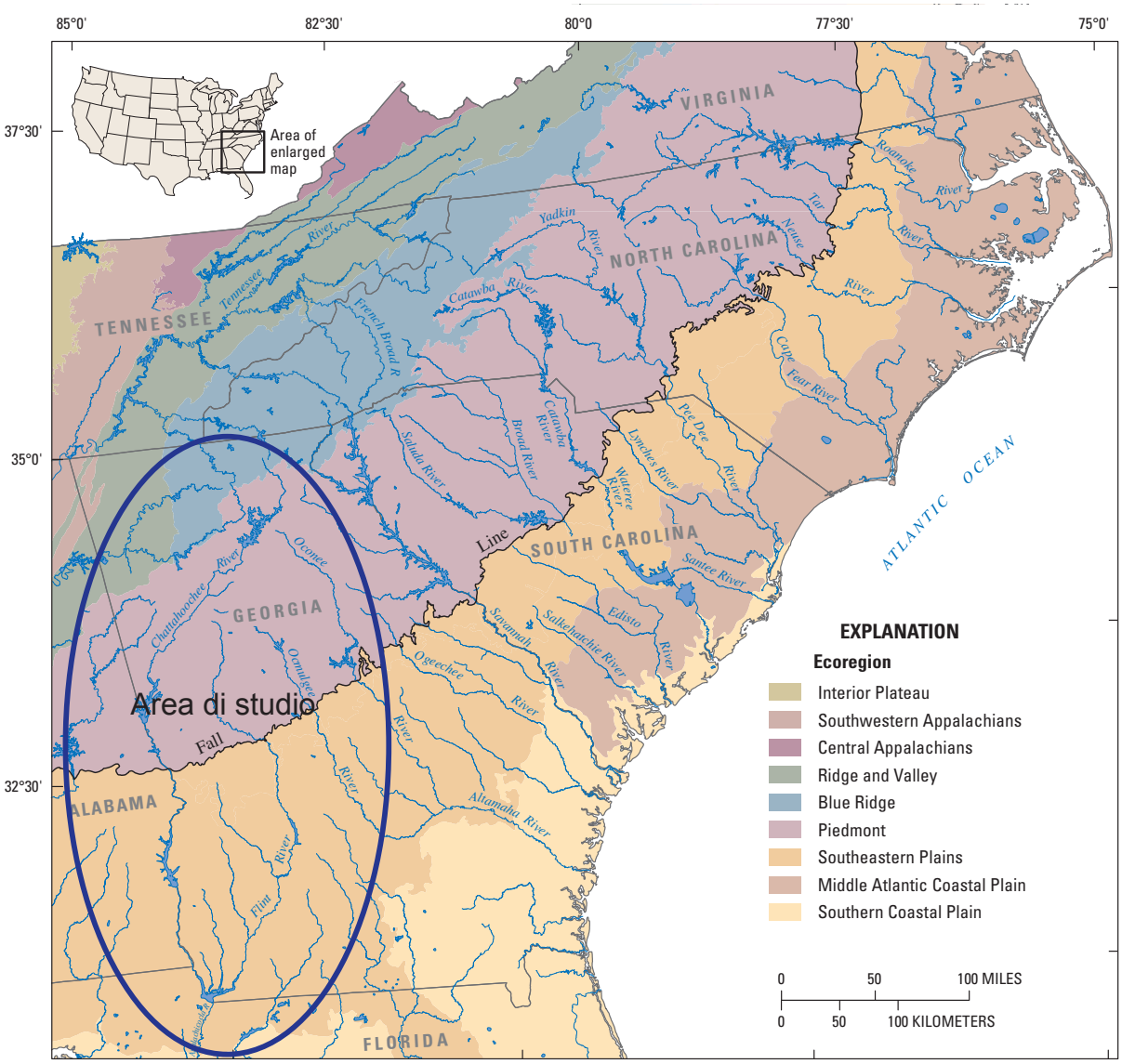

Figura 4.1: Contestualizzazione geografica dell'area di studio. Suddivisione in ecoregioni. USGS *"Magnitude and Frequency of Rural Floods in the Southeastern United States"*.

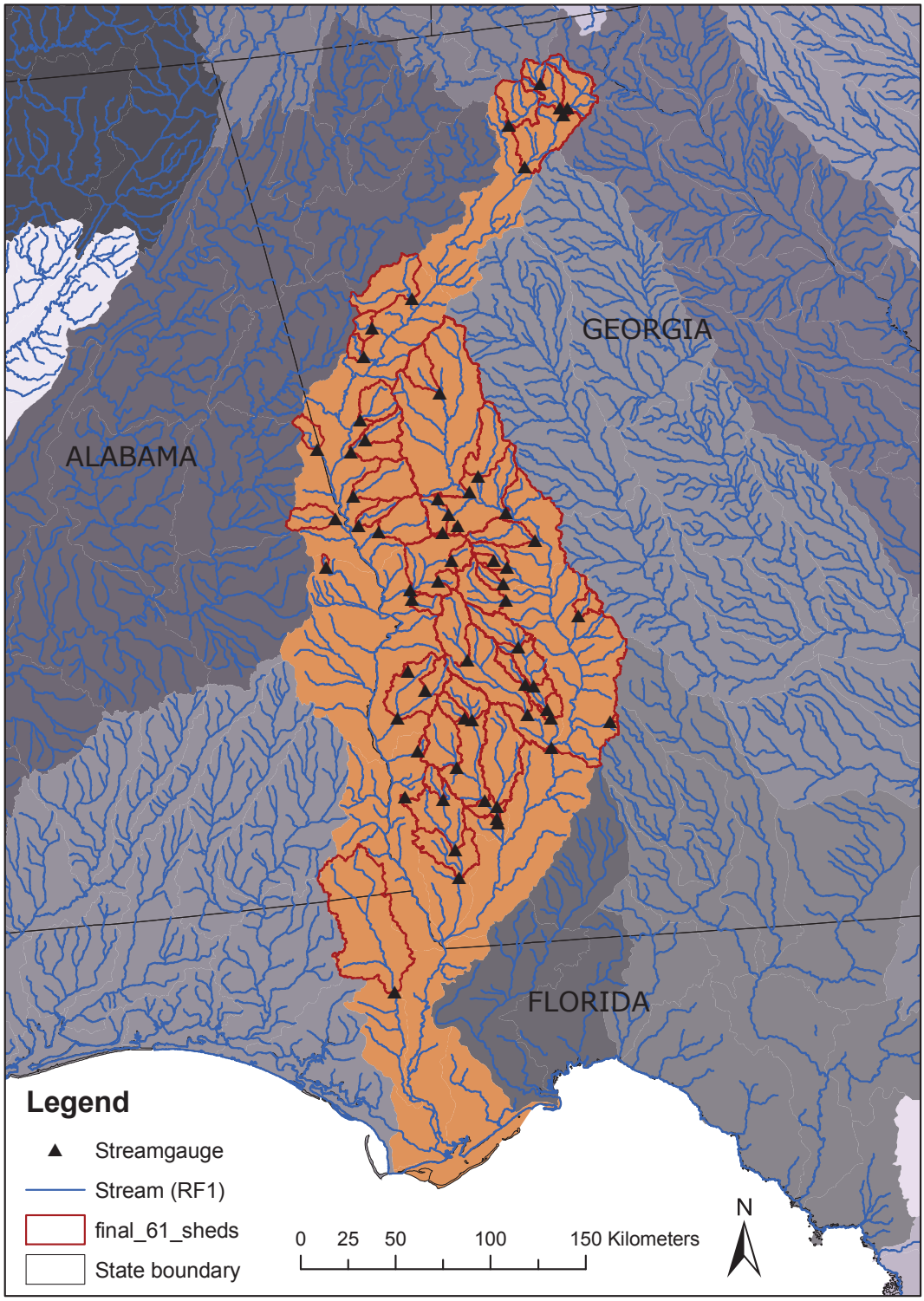

Figura 4.2: Area di studio: bacini idrografici in contorno rosso, reticolo idrografico in blu, sezioni di chiusura rappresentate da triangoli neri.

Georgia, Tennessee e North Carolina. Tuttavia la quasi totalità dell'estensione superficiale dei bacini è compresa nelle due zone *Piedmont* e *Southeastern Plains*: la prima è una zona di transizione tra le catene montuose degli Appalachi e le vaste aree pianeggianti che precedono la costa a sud, caratterizzata da formazioni rocciose ignee e metamorfiche con alternanza di zone pianeggianti e collinari; la seconda è formata da pianure irregolari costituite da un mosaico di terreni agricoli, pascoli, boschi e foreste, e presenta formazioni sedimentarie più giovani come sabbie e argille. Inoltre si può notare come la separazione tra le queste due regioni dalle marcate differenze orografiche, climatiche e geologiche, è evidenziata in figura da una linea detta *Fall Line*. Infine le pianure costiere sono strettamente pianeggianti ed eterogenee e presentano zone paludose lagunari che, in generale, rendono i suoli più umidi rispetto a quelle di pianura del sud-est.

4.2 Dati geomorfologici e climatici

Per ciascuno dei 61 bacini si hanno a disposizione 22 diversi descrittori geomorfologici e climatici, estratti da analisi di dataset digitali mediante software GIS. Secondo quanto descrive il *report* del USGS 2009, per quanto concerne i bacini della Georgia, lo spartiacque è stato ricavato a partire da DEM (*Digital Elevation Model*) con risoluzione a 30 metri di dati contenuti nel NED (*National Elevation Dataset*). Qui si riportano (v. tabelle 4.1, 4.2, 4.3) per ciascuna variabile il valore minimo, il valore massimo, la media, il 25° percentile, la mediana, il 75° percentile estratti dal dataset dei 61 bacini, mentre per una consultazione più ampia e dettagliata si rimanda all'Appendice A in cui sono stati trascritti i dati al completo. Per la descrizione delle variabili si consulti la tabella 4.5.

Tabella 4.1: Variabili geomorfologiche. Parte 1.

	A (km^2)	LAT (^oN)	LONG (^oE)	L (km)	S $(m \cdot km^{-1})$	P (km)	F_f $(-)$
min	1,9	30,885883	-85,332292	2,9	0,3971	5,7	2,48
25^o**percent.**	121,7	31,998449	-84,814377	22,6	0,7476	73,2	4,21
mediana	328,9	32,515723	-84,601870	43,1	1,3102	113,7	5,30
75^o**percent.**	727,7	33,107740	-84,316985	65,2	3,0581	205,5	7,23
max	13752,1	34,761731	-83,535129	410,5	14,2724	1071,6	12,25
media	792,7	32,618562	-84,534311	54,8	2,2721	159,9	5,81

Per quanto riguarda le variabili climatiche si sono considerati esclusivamente parametri descrittori delle precipitazioni come il MAP (precipitazione media annua) e l' MDP_T (precipitazione massima giornaliera con assegnato tempo di ritorno) ottenuti con una media pesata sulle aree dei bacini (v. Gotvald, Feaster e Weaver 2009).

Tabella 4.2: Variabili geomorfologiche. Parte 2.

	H_m (m)	H_{max} (m)	H_{min} (m)	S_m (%)	F_i (%)	F_{for} (%)	I_d $(-)$	I_h $(-)$	D_d (km/km^2)
min	47,3	103,6	9,3	1,0	0,14	3,54	2,4	1,83	0,75
25^o**percent.**	129,5	176,7	82,0	4,0	0,35	38,91	3,01	2,05	1,17
mediana	174,1	252,2	110,3	5,3	0,56	58,99	3,09	2,12	1,28
75^o**percent.**	243,0	411,7	192,3	6,5	1,14	64,63	3,3	2,26	1,38
max	758,2	1351,5	435,3	33,0	7,95	94,63	4,04	2,86	5,76
media	214,3	367,9	144,9	6,4	1,02	51,64	3,18	2,18	1,33

Tabella 4.3: Variabili climatiche.

	MAP (mm)	MDP_2 (mm)	MDP_{10} (mm)	MDP_{25} (mm)	MDP_{50} (mm)	MDP_{100} (mm)
min	1079,5	101,6	152,4	177,8	197,6	203,2
25^o**percent.**	1265,0	101,6	152,6	177,8	201,4	215,7
mediana	1300,2	101,6	157,3	181,8	203,2	221,4
75^o**percent.**	1397,0	107,3	165,1	192,1	211,0	232,1
max	1899,2	126,5	194,2	226,5	256,3	282,9
media	1342,0	105,5	159,9	185,9	207,0	224,6

4.3 Dati idrometrici

I dati idrometrici messi a disposizione dall'USGS provengono da un'analisi in frequenza dei valori al colmo di piena annuali, o anche definiti portate annuali di picco (*peak-flow data*).[1] In tale analisi sono state utilizzate solo le stazioni con un *record* minimo di 10 anni di osservazioni, e sono stati accettati valori di portata che non subivano una sostanziale influenza da parte della regolazione di dighe, della presenza di casse di espansione, di maree o di centri urbani.

Nelle sezioni strumentate l'analisi di frequenza dei valori di piena è stata condotta *fittando* le serie di portate al colmo annuali logaritmiche con un'assegnata distribuzione statistica, che per l'area di studio è stata scelta la *Pearson* tipo III in accordo con le linee guida del "Bulletin 17B of the Hydrology Subcommittee of the Interagency Advisory Committee on Water Data (1982)".

In definitiva una stima delle portate al colmo di piena con un determinato tempo di ritorno oppure, che è lo stesso, le portate al colmo con una percentuale P di probabilità di superamento,[2] vengono valutate inserendo i tre momenti (media, varianza e asimmetria)

[1] Con *portate annuali di picco* s'intende il più alto valore di portata registrato nell'arco di un anno in una determinata stazione di misura.

[2] viene calcolata come $P = \frac{1}{T} \times 100$ dove T è il tempo di ritorno espresso in anni.

della funzione di distribuzione in frequenza nell'equazione:

$$\log Q_P = X + KS, \tag{4.1}$$

dove

Q_P**:** è il valore di portata al colmo con probabilità di superamento P;

X**:** è la media della serie logaritmica delle portate al colmo annuali;

K**:** è un fattore dipendente dal coefficiente di asimmetria e dalla probabilità di superamento assegnata (o dal tempo di ritorno);

S**:** è la varianza della serie logaritmica delle portate al colmo annuali.

In questo modo si sono ottenute per ciascuno dei 61 bacini le portate al colmo per diversi tempi di ritorno: $T_{rit} = 10, 50, 100, 500$ (anni)(v. tab. 4.4). Anche qui, per brevità, si riportano il minimo, il 25° percentile, la mediana, il 75° percentile, il massimo e la media che sintetizzano i caratteri principali dei dati. Per una visione completa dei valori di portata al colmo con assegnato tempo di ritorno si rimanda all'Appendice A.

Tabella 4.4: Variabili idrometriche con assegnato tempo di ritorno.

	$\mathbf{Q_{10}}$ (m^3/sec)	$\mathbf{Q_{50}}$ (m^3/sec)	$\mathbf{Q_{100}}$ (m^3/sec)	$\mathbf{Q_{500}}$ (m^3/sec)
min	3,5	7,5	9,9	13,8
25^o**percentile**	57,5	96,8	118,1	165,1
mediana	168,8	303	368,1	535,2
75^o**percentile**	264,5	481,4	603,1	923,1
max	1673,5	2489,1	2860	3822,8
media	243	397,9	475,7	690,9

Tabella 4.5: Descrizione delle variabili geomorfoclimatiche utilizzate nel metodo PSBI.

Variabili geomorfoclimatiche		
Variabile	Descrizione	Unità di misura
A	Area Bacino	(km^2)
LAT	Latitudine (centro)	$(^o)$
LONG	Longitudine (centro)	$(^o)$
L	Lunghezza asta principale	(km)
S	Pendenza Asta principale	$(m \cdot km^{-1})$
P	Perimetro del bacino	(km)
F_f	Fattore di forma	$(-)$
H_m	Quota media del bacino	(m)
H_{min}	Quota minima del bacino	(m)
H_{max}	Quota massima del bacino	(m)
S_m	Pendenza Media del bacino	$(\%)$
F_i	Frazione di bacino impermeabile	$(\%)$
F_{for}	Frazione di bacino a foreste	$(\%)$
D_d	Densità di drenaggio	$(km \cdot km^{-2})$
I_d	Indice di drenaggio del suolo	$(-)$
I_h	Indice di suolo idrologico	$(-)$
MAP	Precipitazione media annua	(mm)
MDP_T	Precipitazione massima gionaliera con assegnato tempo di ritorno	(mm)

Capitolo 5

Applicazione e accoppiamento delle tecniche CK e TK

Come ampiamente descritto nel capitolo 3 le tecniche di regionalizzazione, basate sull'interpolazione geostatistica delle variabili idrometriche, permettono di stimare valori di portata in siti non strumentati o di cui si conoscono solo dati parziali o non completi. In particolare, Il Canonical Kriging consente di ricavare valori di portata in bacini non strumentati altrimenti incogniti, a partire dalla conoscenza dei soli parametri geomorfologici e climatici (v. Fig. 3.4) di più immediata reperibilità; il Top-Kriging si appoggia alla teoria geostatistica della *regolarizzazione* e tiene in conto della superficie e della collocazione spaziale dei bacini, particolarmente vantaggiosa in quanto i dati sono facilmente ottenibili usando un qualsiasi software GIS (*Geografical Information System*). In questo capitolo verrà mostrato com'è possibile e con quali tecniche accoppiare i due approcci, traendo i vantaggi dell'una e dell'altra tecnica.

5.1 Struttura dell'indagine

L'idea sulla quale si basa il lavoro svolto è quella di orientare lo studio sull'interpolazione geostatistica dei valori di piena al colmo con assegnato tempo di ritorno focalizzando l'attenzione sull'analisi dei *residui* generati dalla stima. Nelle applicazioni idrologiche, e in generale nei modelli fisico-statistici, viene detto *residuo* la differenza in un determinato punto tra la variabile osservata, ovvero il dato proveniente da misure dirette o indirette, e la variabile stimata attraverso un determinato modello a cui appartengono specifici parametri:

$$\varepsilon(x_0) = Z(x_0) - \hat{Z}(x_0) \tag{5.1}$$

dove:

$\hat{Z}(x_0)$**:** è la variabile stimata nel punto x_0

$Z(x_0)$**:** è la variabile osservata nello stesso punto x_0

Riuscire a *modellare* i residui, nel senso stretto di trovare delle relazioni funzionali tra il residuo generato dalla stima e le variabili indipendenti, siano esse del tipo PSBI o geografiche nel Top-Kriging, può essere per molti aspetti vantaggioso ai fini della ricerca di un modello che simuli nella maniera più fedele possibile i dati osservati. L'analisi dei residui prodotti dai modelli geostatistici può essere condotta assumendo che gli stessi possono essere trattati come una variabile regionalizzata alla stregua di variabili idrometriche, quindi sottoponibili ad interpolazione con le tecniche del kriging:

$$\varepsilon \xrightarrow{kriging} \varepsilon^*$$

Ecco che quindi si profilano due possibilità di agire sui residui generati dai due modelli di interpolazione Canonical Kriging e Top-Kriging:

CK corretto via TK

L'interpolazione della variabile viene effettuata con il Canonical Kriging e si ricorre al Top-Kriging per la modellazione dei residui.

TK corretto via CK

L'interpolazione della variabile si effettua con il Top-Kriging e si utilizza il Canonical Kriging per la modellazione dei residui.

Le due linee d'intervento ed i passaggi che hanno portato ai risultati finali sono schematizzate nel diagramma di flusso in figura 5.1.

5.1.1 Analisi preliminari

Alcune indagini preliminari hanno mostrato che l'applicazione di entrambe le tecniche, Canonical kriging e Top-Kriging (esposte in dettaglio nei paragrafi seguenti), risultano raggiungere maggiori prestazioni e risultati significativi se applicate non direttamente ai quantili di portata tal quali, bensì operando sui quantili *specifici* cioè adimensionalizzati rispetto all'area del bacino A. Come efficacemente illustrato dai diagrammi di dispersione in scala bilogaritmica delle figure da 5.2 a 5.5, sono state messe in relazione, per ciascun bacino, l'area sottesa dalla sezione di chiusura con i quantili di portata ad assegnato tempo di ritorno calcolati per la stessa sezione; utilizzando una regressione ai minimi quadrati si è messo in luce che la legge di scala dei quantili di piena Q_T può essere, con buona approssimazione, del tipo

$$Q_T \sim A^{0.65} \tag{5.2}$$

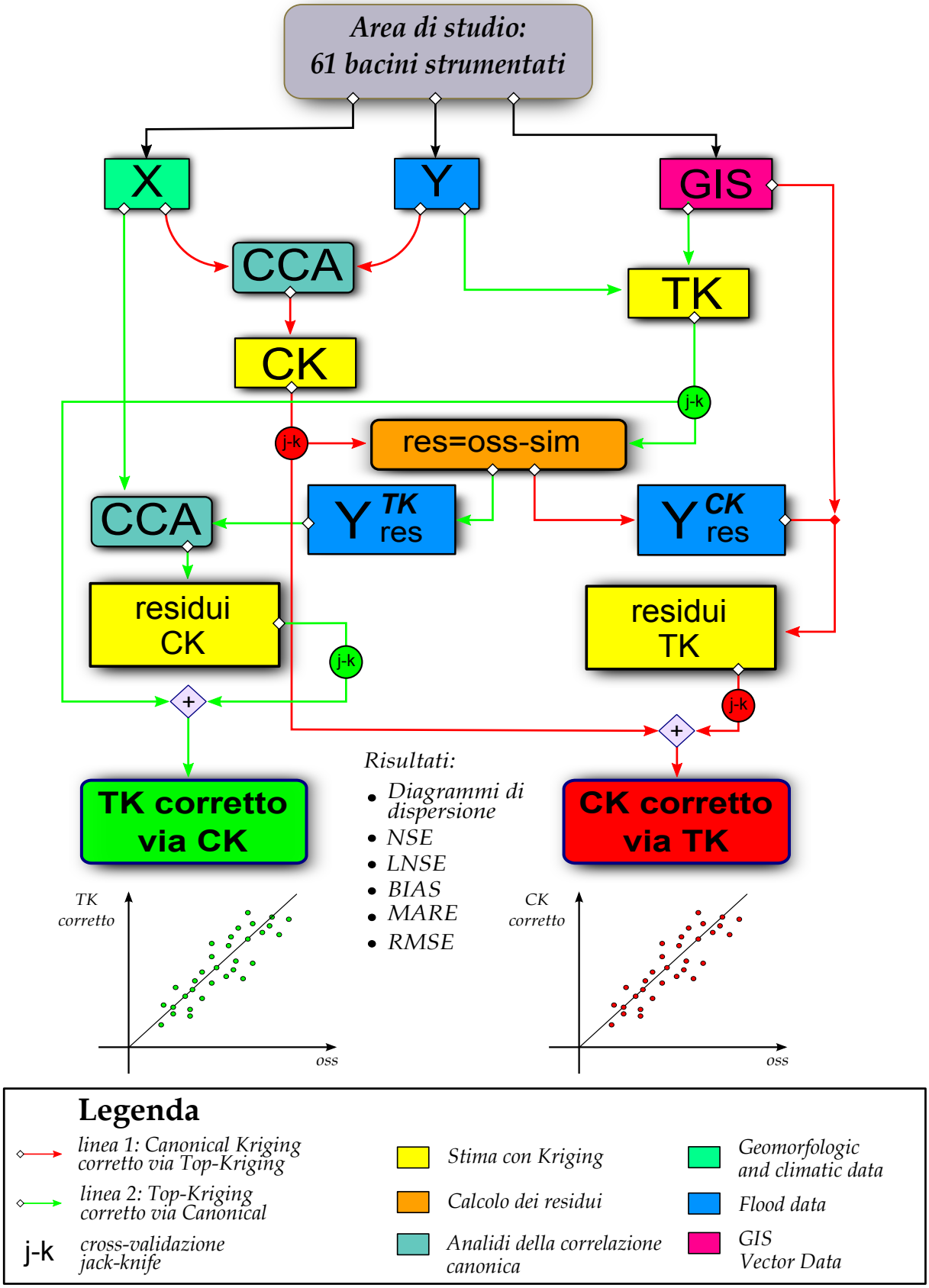

Figura 5.1: Schema a blocchi dell'impianto dell'indagine. Le linee rosse indirizzano l'approccio Canonical Kriging corretto via Top-Kriging: le linee in verde indirizzano l'approccio Top-Kriging corretto via Canonical Kriging.

qualunque sia il tempo di ritorno. Applicando, infatti, la regressione alle singole serie, lo scostamento dei parametri (in uscita dal fittaggio) tra una serie e l'altra, può essere ritenuto trascurabile, per cui si è scelto un valore medio pari a 0.65.

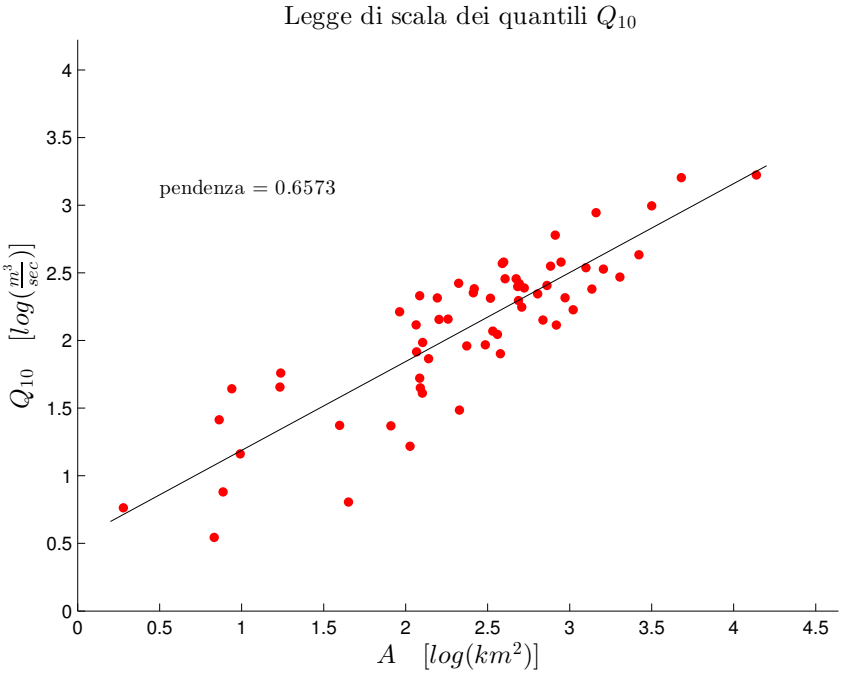

Figura 5.2: Legge di scala dei quantili di portata al colmo di piena Q_{10}. Regressione lineare in scala bilogaritmica.

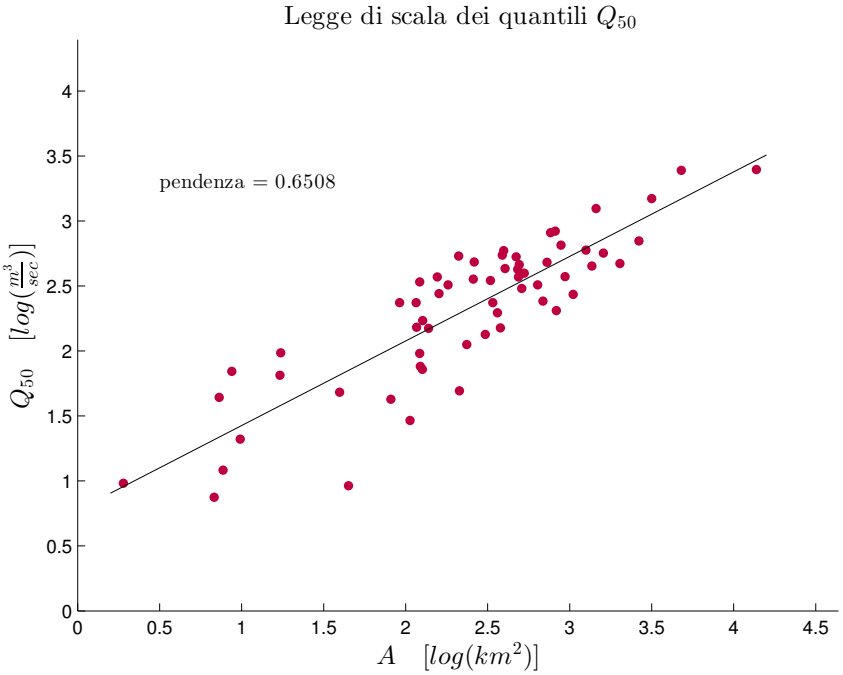

Figura 5.3: Legge di scala dei quantili di portata al colmo di piena Q_{50}. Regressione lineare in scala bilogaritmica.

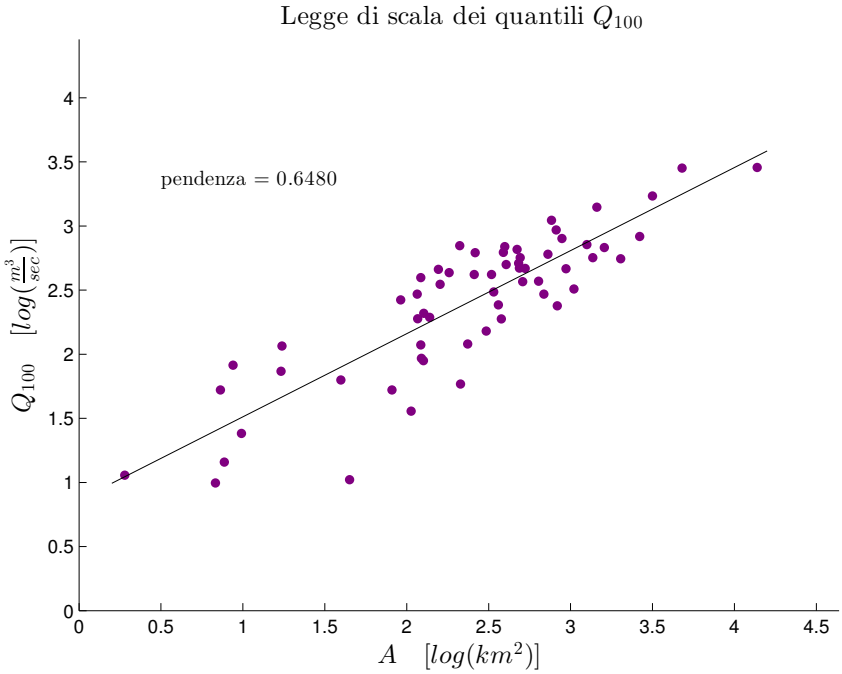

Figura 5.4: Legge di scala dei quantili di portata al colmo di piena Q_{100}. Regressione lineare in scala bilogaritmica.

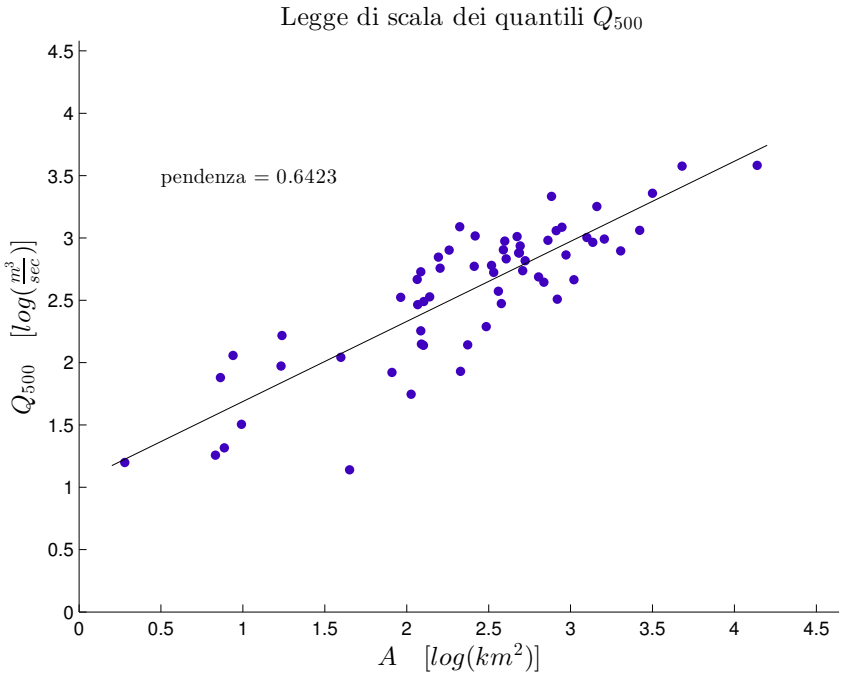

Figura 5.5: Legge di scala dei quantili di portata al colmo di piena Q_{500}. Regressione lineare in scala bilogaritmica.

D'ora in avanti in questo testo si da per assunto che i quantili di portata, utilizzati negli interpolatori geostatistici, sono del tipo[1]

$$Q'_T = \frac{Q_T}{A^{0.65}}. \tag{5.3}$$

5.1.2 Canonical Kriging corretto via Top-Kriging

Partendo dai dati disponibili, ottenuti in collaborazione con l'USGS (*United States Geological Survey*) e consultabili in Appendica A, per 61 bacini (sulle modalità di reperimento e di calcolo dei dati si veda Capitolo 4) sono stati separati due dataset differenti e complementari: nella prima matrice di dati, che chiameremo X, ogni bacino è caratterizzato da un vettore di valori corrispondenti ciascuno ad una variabile geomorfologica (area del bacino, latitudine, longitutdine, lunghezza asta pincipale, ecc...) e climatica (precipitazioni) per un totale di 22 variabili; nella seconda matrice, detta Y, abbiamo che a ciascun bacino appartiene un vettore di valori corrispondenti alle variabili idrometriche, nel nostro caso rappresentate da valori di portata al colmo di piena con diversi tempi di ritorno, ovvero 10, 50, 100, 500 anni. È fondamentale ai fini dell'analisi della correlazione canonica e della comparazione delle tecniche di krigaggio che le due matrici presentino lo stesso numero N di righe, ossia lo stesso numero di bacini; quindi riassumendo le due matrici utilizzate hanno dimensioni:

$$\dim(X) = (61, 22)$$
$$\dim(Y) = (61, 4)$$

Assumendo che il rango r della matrice di covarianza S_{XY} sia pieno, dette k il numero di colonne di X ed m il numero di colonne di Y, per quanto affermato nel Capitolo 2, $r = \mathrm{rg}(S_{XY}) = \min(k, m) = 4$, quindi c'è da aspettarsi che l'analisi delle correlazione canonica produca 4 coppie di variabili sintetiche (U_h, V_h) con $h = 1, \ldots, 4$.

Ottenuto il dataset definitivo si è proceduto con l'effettuare l'analisi della correlazione canonica delle due matrici di dati X ed Y in modo da estrapolare le prime due componenti canoniche indipendenti U_1 ed U_2, così da poterle utilizzare come assi di un piano cartesiano nel quale ciascun bacino è rappresentato da un punto di coordinate (U_1^*, U_2^*) (v. fig 3.2). Si ricorda[2] che le prime due componenti canoniche discendono dalla soluzione di un problema agli autovalori nel quale convergono le varianze $\mathrm{Var}\{X\}$ e $\mathrm{Var}\{Y\}$ dei due dataset, caratteristica distintiva rispetto all'analisi delle componenti principali che non

[1] è bene precisare che le portate specifiche sono state adoperate solo nel caso si fosse utilizzato un interpolatore, mentre nell'operazione di confronto tra i dati cross-validati con quelli empirici si è fatto riferimento alle portate tal quali.

[2] per la teoria sull'analisi della correlazione canonica si veda Capitolo 2.

considera la dipendenza dai dati Y. L'analisi della correlazione canonica è stata prodotta in ambiente MATLAB®mediante uno script ad hoc che utilizza la *function* `canoncorr` automatizzando la ricerca delle componenti canoniche. Si nota in figura 5.6 come i punti dei diagrammi a dispersione delle prime due coppie di componenti principali (U_1, V_1) e (U_2, V_2) posti sulla diagonale, si orientino grossomodo lungo la bisettrice del primo e terzo quadrante, con poca dispersione attorno ad essa. Questo risultato discende proprio dalla ricerca di variabili sintetiche ottenute massimizzando la correlazione tra le variabili originarie dei dataset X e Y. Come illustra la figura 5.7, facendo un test statistico sul livello di significatività dei dati, o *p-value test*, è possibile dimostrare che le prime due componenti canoniche riescono a spiegare più del 95% della varianza totale dei dati.

Estrapolato il riferimento spaziale cartesiano sintetico U_1, U_2 dall'analisi della correlazione canonica, è immediato poter associare a ciascun punto del piano una tra le variabili idrometriche, a disposizione nel dataset, lungo la terza dimensione (v. Fig. 3.3). L'interpolazione dei dati puntuali è stata condotta in ambiente MATLAB®mediante l'utilizzo del software `EasyKrig3.0`.[3], che permette di calcolare il variogramma sperimentale e teorico, assegnando come input la variabile regionalizzata sul dominio sintetico. Risulta quindi un prerequisito confezionare i dati in uscita dall'analisi della correlazione canonica, in modo tale da poter essere caricati adeguatamente dal software. L'operazione fondamentale del software consiste nel calcolare i ponderatori ottenuti risolvendo i sistemi lineari (1.39) nel caso di Kriging Ordinario, oppure (1.44), nel caso di Kriging Universale. La scelta dell'uno o dell'altro metodo è affidata all'utente, ma in generale è dettata dall'andamento del variogramma sperimentale, a seconda della presenza o meno di un *sill* o dal comportamento vicino l'origine (v. Raspa e Bruno 1994a).

Dopo diverse prove preliminari di interpolazione, la scelta globale in termini di prestazioni statistiche sembra ricadere sull'Universal Kriging, modello più robusto in quanto tiene in conto della presenza di un *drift* della media nel campo, utilizzando portate specifiche $Q'_T = \frac{Q_T}{A^{0.65}}$ dove A è l'area del bacino[4]. L'interpolazione geostatistica tramite Kriging Universale si presenta come una superficie continua sopra il piano dei descrittori geomorfoclimatici U_1, U_2 come mostrato nelle figure 5.8 e 5.9 in cui viene rappresentato il dato interpolato per le Q'_{10} e Q'_{100} in uscita dal software `EasyKrig3.0`.

All'interpolazione segue la validazione del modello eseguito con la tecnica *leave-one-out* (si veda il paragrafo 5.3) la quale, permettendo la simulazione delle condizioni non strumentate, dà la possibilità di poter esprimere dei giudizi sulla qualità e sulle prestazioni del modello, analizzando i diagrammi di dispersione e gli indici statistici ottenuti da un confronto tra la variabile stimata e quella osservata nei medesimi punti del piano.

[3]Copiright(c)1998,2001,2004 property of Dezhang Chu and Woods Hole Oceanographic Institution.
[4]sulle motivazione di tale scelta si veda il paragrafo precedente.

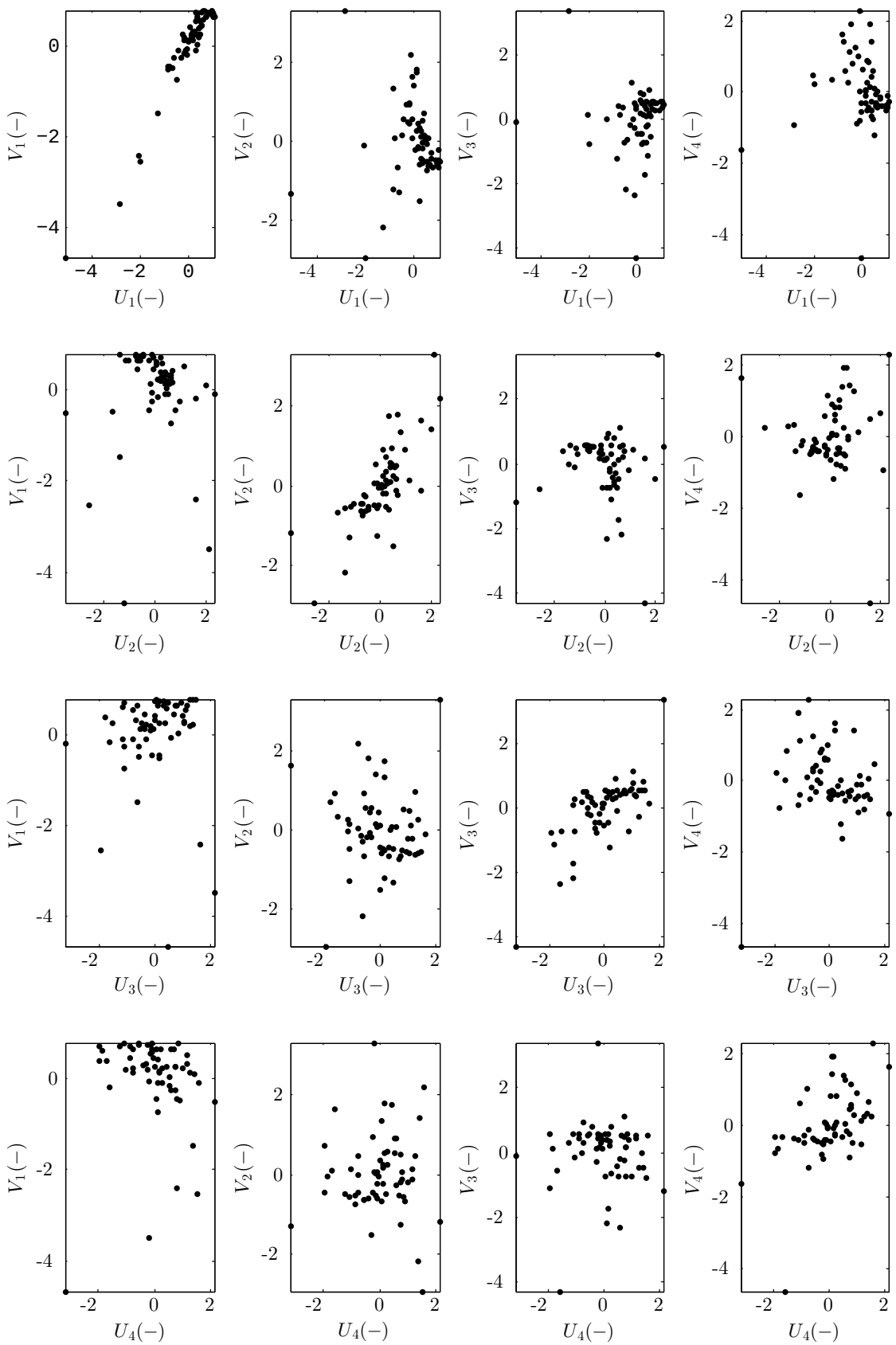

Figura 5.6: Diagrammi di dispersione delle componenti canoniche associate ai dataset X ed Y.

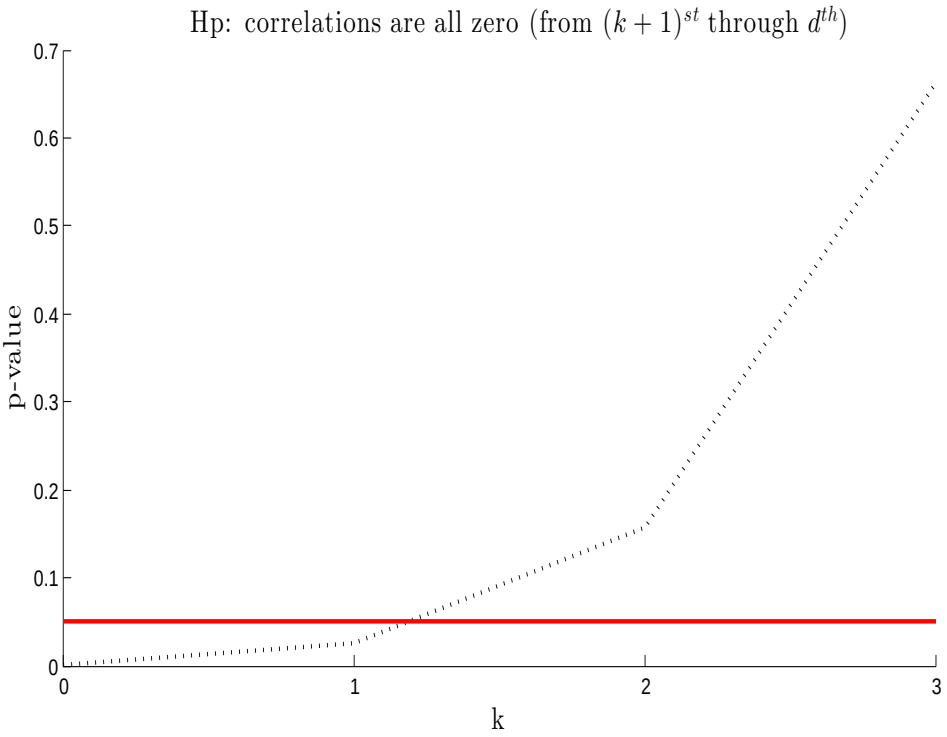

Figura 5.7: Test delle ipotesi. Il *p-value* rappresenta la probabilità che la varianza dei dati *non* sia spiegata dalle componenti canoniche.

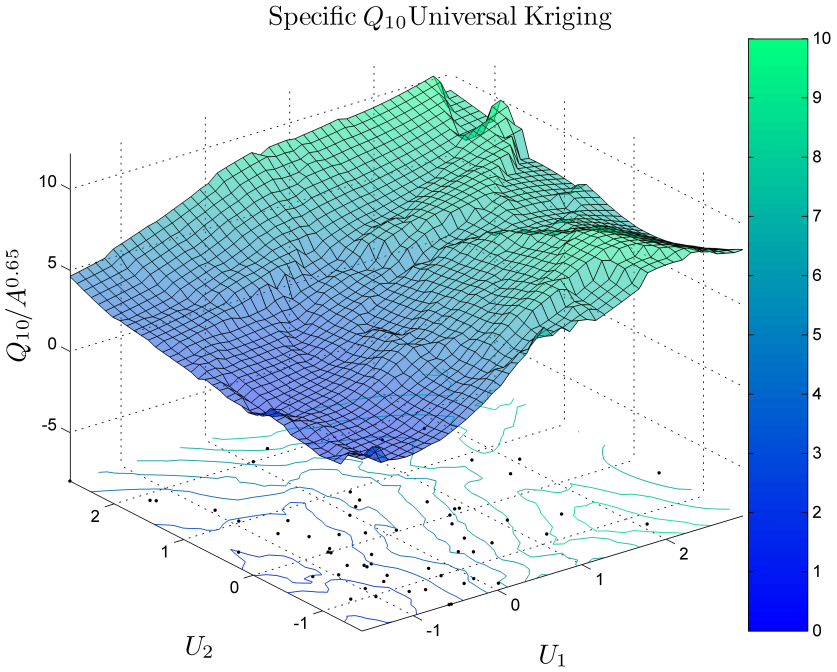

Figura 5.8: Superficie di interpolazione su dominio PSBI per portate specifiche con tempo di ritorno $T = 10$ anni ottenuta tramite Universal Kriging.

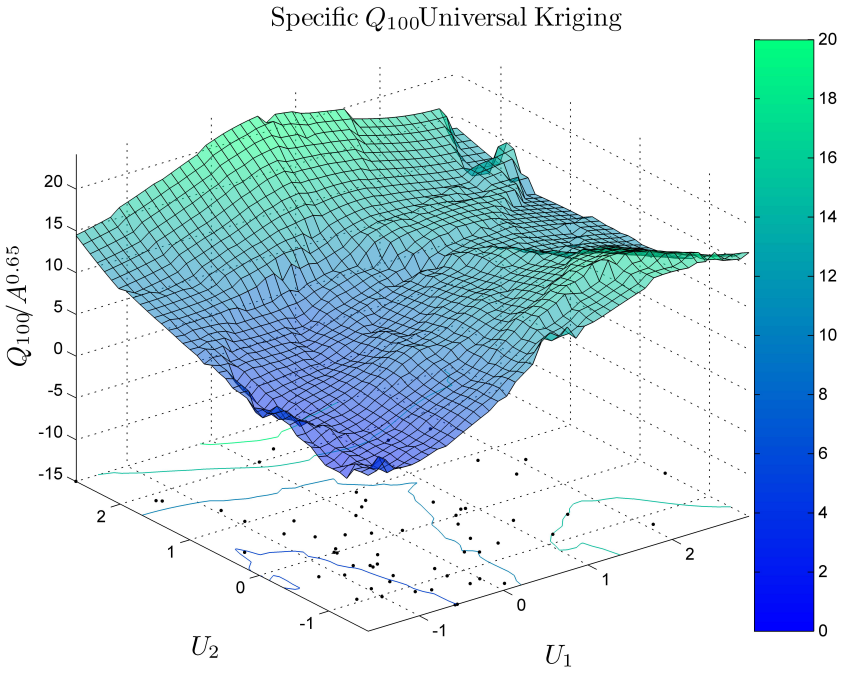

Figura 5.9: Superficie di interpolazione su dominio PSBI per portate specifiche con tempo di ritorno $T = 100$ anni ottenuta tramite Universal Kriging.

A questo punto l'analisi si sposta sui residui. Vengono calcolati gli scarti tra la variabile osservata e ed i valori ottenuti in cross-validazione della variabile stimata con il Canonical Kriging, in modo tale da generare una nuova variabile regionalizzata, il *residuo*, che può assumere questa volta sia valori positivi sia negativi. La nuova variabile viene interpolata utilizzando il Top-Kriging, che, come già detto (si veda Capitolo 3), si basa su un'interpolazione della VR a supporto *non* puntuale o areale, inducendo un effetto di regolarizzazione nel variogramma. L'analisi dei residui viene condotta utilizzando un pacchetto specifico dal nome `rtop` del software freeware ed open source **R**-project, scritto e sviluppato da J. O. Skøien (v. Skøien et al. 2011) e reperibile da CRAN (The Comprehensive **R** Archive Network, `http://cran.r-project.org`). Il software/pacchetto riesce a integrare ed elaborare dati GIS, tipicamente file di estensione `.shp`, con l'informazione idrometrica; una tale commistione consente di interpolare i dati idrometrici tenendo conto dell' estensione areale dei bacini, della loro posizione geografica e della eventuale presenza di strutture annidate. Anche in questo caso il software, basandosi sulla formulazione introdotta con l'equazione (3.11), calcola dapprima i variogrammi sperimentale e teorico, propedeutici alla risoluzione del sistema lineare che porta alla conoscenza dei ponderatori λ_i necessari alla stima, successivamente effettua l'interpolazione dei residui nel dominio e, in ultimo, opera una validazione del modello con la tecnica *leave-one-out*. Un tale approccio consente di affiancare una stima dei residui su tutto il dominio alla stima iniziale della

variabile idrometrica, ottenuta tramite Canonical Kriging, semplicemente sommando le due stime e, di fatto, sovrapponendone gli effetti.

In definitiva, a partire da dati geomorfologici e climatici di un bacino B non strumentato, si procede inizialmente alla valutazione della portata al colmo $Q_{T,sim}$ con assegnato tempo di ritorno T, simulata mediante Canonical Kriging, e successivamente alle stima di un residuo ε^* per quel determinato bacino e per lo stesso tempo di ritorno T; per cui, sommando le due stime, è possibile calcolare un valore di portata Q_T^* "modificata" che tenga in conto dell'errore indotto dal modello.

$$\begin{array}{ccc} B & \xrightarrow{CK} & Q_{T,sim} \\ {\scriptstyle TK}\downarrow & & \downarrow \\ \varepsilon^* & \longrightarrow & Q_T^* \end{array}$$

5.1.3 Top-Kriging corretto via Canonical Kriging

È l'approccio complementare alla combinazione delle tecniche Canonical Kriging e Top-Kriging.

Il primo step consiste nell'interpolazione con la tecnica Top-Kriging dei valori di portata al colmo tramite il pacchetto `rtop` che calcola i variogrammi, esegue l'interpolazione ed effettua la validazione *leave-one-out*; questa operazione viene iterata per ogni variabile idrometrica a disposizione: Q_{10}, Q_{50}, Q_{100}, Q_{500}.[5] L'analisi dei residui viene condotta mediante Canonical Kriging: in questo caso il dataset Y non è più la matrice di dati idrometrici ma una nuova matrice, dimensionalmente identica, in cui le variabili, o le colonne della matrice, sono rappresentate dai residui ottenuti dall'interpolazione delle variabili idrometriche. Anche qui, in analogia a quanto visto finora, i residui in ingresso all'interpolatore vengono rapportati all'area del bacino.

Il secondo step riguarda l'analisi dei residui. Si parte dall'analisi della correlazione canonica, che produce le variabili sintetiche U, V a partire dai dati geomorfoclimatici contenuti in X e dalla nuova matrice dei residui Y_{res}, dalla quale vengono estratte le prime due componenti principali U_1, U_2, che, a loro volta, formeranno il piano cartesiano sintetico su cui effettuare l'interpolazione. A ciascun punto nel piano viene associato nella terza dimensione un valore di residuo, diventando così una nuova variabile regionalizzata. L'operazione di interpolazione del residuo, eseguita anche in questo caso mediante l'utilizzo del software `EasyKrig3.0` in ambiente MATLAB®, è stata iterata per ciascuna

[5] anche in questo caso sono state utilizzate variabili specifiche, rapportate all'area del bacino.

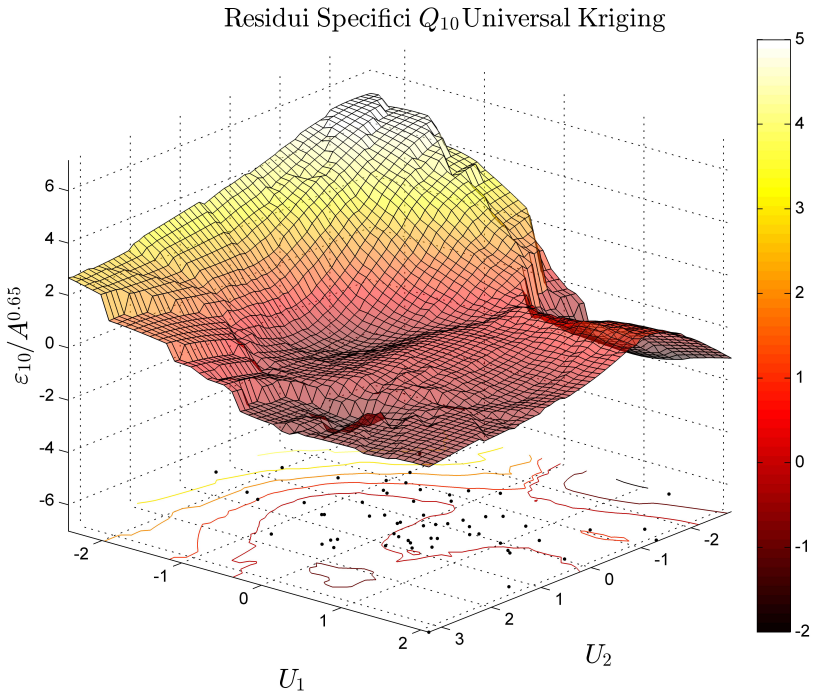

Figura 5.10: Superficie di interpolazione su dominio PSBI per residui specifici con tempo di ritorno $T = 10$ anni ottenuta tramite (Canonical) Universal Kriging.

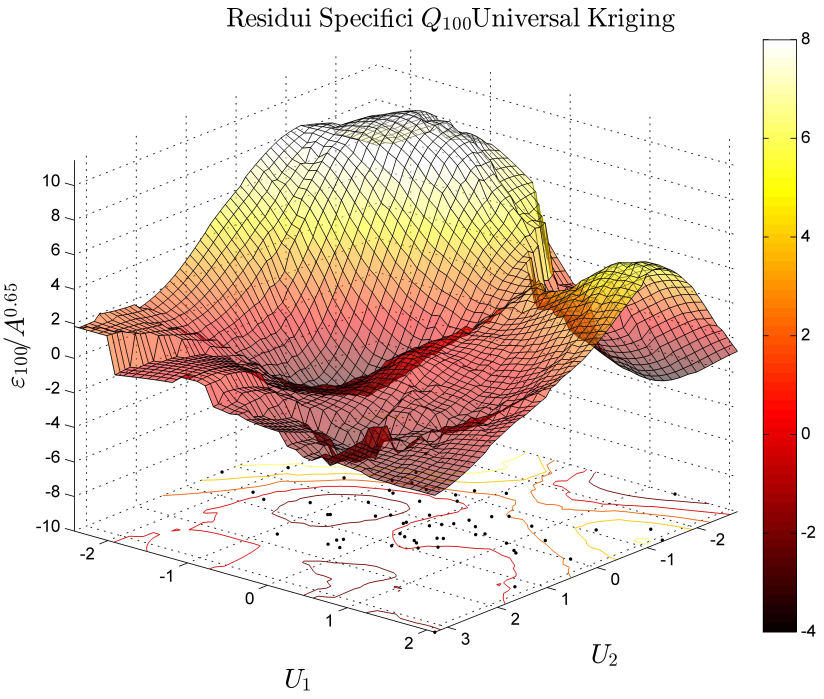

Figura 5.11: Superficie di interpolazione su dominio PSBI per residui specifici con tempo di ritorno $T = 100$ anni ottenuta tramite (Canonical) Universal Kriging.

variabile residuo ε_T con $T = 10, 50, 100, 500$ (anni), corrispondenti alle rispettive portate al colmo con assegnato tempo di ritorno. Nelle figure 5.10 e 5.11 è possibile visualizzare l'andamento dei residui ε_{10} e ε_{100} interpolati mediante Canonical Kriging. Il modello infine viene validato con la tecnica *leave-one-out*.

In uscita si ha quindi una stima dei residui che può essere sommata alla prima stima della variabile idrometrica; anche qui, come prima, a partire da dati geomorfologici e climatici di un bacino B non strumentato, inizialmente si valuta la portata al colmo $Q_{T,sim}$ con assegnato tempo di ritorno T simulata mediante Top-Kriging; successivamente si stima un residuo ε^* per quel determinato bacino e per lo stesso tempo di ritorno; pertanto sommando le due stime è possibile calcolare un valore di portata Q_T^* "modificata" che tenga in conto dell'errore indotto dal modello.

$$\begin{array}{ccc} B & \xrightarrow{CK} & \varepsilon^* \\ {\scriptstyle TK}\downarrow & & \downarrow \\ Q_{T,sim} & \longrightarrow & Q_T^* \end{array}$$

5.2 Principali indici statistici prestazionali

Nell'analizzare i risultati si è fatto largo uso di diversi indici statistici che permettono la comparazione e la quantificazione dell'accuratezza dei vari modelli utilizzati. Si riporta un elenco degli indici adottati:

NSE

Efficienza di Nesh & Sutcliffe

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{N} \left(Z_{i,obs} - \hat{Z}_{i,sim}\right)^2}{(N-1)\,\mathrm{Var}\left\{Z_{obs}\right\}} \tag{5.4}$$

LNSE

Efficienza di Nesh & Sutcliffe in scala logaritmica

$$\mathrm{LNSE} = 1 - \frac{\sum_{i=1}^{N} \left(\log Z_{i,obs} - \log \hat{Z}_{i,sim}\right)^2}{(N-1)\,\mathrm{Var}\left\{\log Z_{obs}\right\}} \tag{5.5}$$

BIAS

Errore medio relativo

$$\text{BIAS} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{Z_{i,obs} - \hat{Z}_{i,sim}}{Z_{i,obs}}\right) \tag{5.6}$$

MARE

Errore medio relativo in valore assoluto

$$\text{MARE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{|Z_{i,obs} - \hat{Z}_{i,sim}|}{Z_{i,obs}}\right) \tag{5.7}$$

RMSE

Radice dell'errore quadratico medio

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{Z_{i,obs} - \hat{Z}_{i,sim}}{Z_{i,obs}}\right)^2} \tag{5.8}$$

dove $Z_{i,obs}$ sta ad indicare il valore i-esimo di variabile idrometrica osservata o misurata, mentre $\hat{Z}_{i,sim}$ è il valore i-esimo di variabile stimata in cross-validazione dal modello.

5.3 Cross-Validazione *jack-knife*

La procedura di cross-validazione *jack-knife* o anche detta *leave-one-out* rappresenta una strumento estremamente versatile, di semplice implementazione e in grado di fornire un adeguato controllo sulle prestazioni degli interpolatori utilizzati. Questa metodologia di validazione dei modelli consente di valutare l'accuratezza della stima, simulando le condizioni non strumentate in ciascuno dei bacini dell'area di studio. La procedura *jack-knife*, utilizzata sia nel Canonical Kriging che nel Top-Kriging, può essere descritta dal seguente algoritmo:

1. dall'insieme di N bacini strumentati si scarta l'i-esimo bacino, avente determinate caratteristiche;

2. viene effettuata l'interpolazione (con CK o TK) utilizzando un dataset ridotto a $N - 1$ bacini;

3. la stima del quantile Q_T del bacino i-esimo è determinabile dalle superfici generate dall'interpolazione;

4. vengono iterati gli step da 1 a 3 per ciascuno degli N bacini appartenenti al dataset;

Con questa procedura si ottiene una serie di valori $Q_{T,sim}$ prodotti dal modello di interpolazione adottato nello stesso punto in cui si hanno a disposizione dati empirici $Q_{T,obs}$: le due serie sono pertanto della stessa lunghezza e possono essere confrontate mediante diagrammi di dispersione e indici statistici visti nel paragrafo 5.2. L'algoritmo così definito risulta nativamente implementato nei software `EasyKrig3.0` in ambiente MATLAB®e `rtop` in ambiente **R**, i quali eseguono la cross-validazione con estrema rapidità.

Capitolo 6

Analisi dei risultati

L'applicazione dei due approcci, Canonical Kriging corretto da Top-kriging e viceversa, descritti con dovizia di particolari nel Capitolo 5, è stata condotta sulle quattro variabili idrometriche di piena al colmo con assegnato tempo di ritorno Q_{10}, Q_{50}, Q_{100}, Q_{500} delle quali si disponevano dati su un campione di 61 bacini idrografici del sud-est degli U.S.A.(per il dataset completo si veda Appendice A). Nelle figure seguenti vengono riportati, per ciascuno dei 61 bacini, in ordinata i valori ottenuti in cross-validazione *leave-one-out*[1] e in ascissa i dati empirici; in ciascuna figura il diagramma di dispersione di sinistra indica la correlazione tra i valori empirici e quelli ottenuti dal primo modello di kriging simulati in cross-validazione $(Q_{T,obs}; Q_{T,sim})$, quello di destra mostra la correlazione tra i valori empirici e i valori ottenuti dal primo krigaggio ai quali viene sommato il valore cross-validato del residuo ottenuto con il modello di kriging complementare $(Q_{T,obs}; Q^*_{T,sim})$. Gli stessi valori vengono riportati anche in scala bi-logaritmica.

Canonical Kriging corretto via Top-Kriging

Da figura 6.1 a figura 6.8 si rappresenta a sinistra la correlazione tra i valori ottenuti da Canonical Kriging applicato in cross-validazione e i valori empirici, a destra, alla correlazione precedente, viene sovrapposta, in rosso, la correlazione con i valori ottenuti da Canonical Kriging corretto via Top-Kriging (sempre in cross-validazione) e i valori empirici.

Si riportano inoltre i principali indici statistici, calcolati mediante le formule da (5.4) a (5.8), associati alle due serie di valori cross-validati in funzione dei valori empirici: nelle tabelle 6.1, 6.2, 6.3 le colonne con intestazione asteriscata si riferiscono agli indici di prestazione calcolati per il modello Canonical Kriging corretto via Top-Kriging; mentre

[1] permette di simulare le condizioni non strumentate in un bacino di cui si dispongono dati idrometrici, si veda l'algoritmo al par. 5.3.

quelle senza asterisco afferiscono al Canonical Kriging puro. La colonna con Δ_{indice} si riferisce agli scarti tra le due colonne appena descritte:

$$\Delta_{\text{indice}} = \text{indice}^* - \text{indice}.$$

Gli scarti mettono in evidenza se c'è stato un miglioramento o un peggioramento a seconda del segno posseduto dallo scarto. Si ricorda che per quanto riguarda l'efficienza di Nash & Sutcliffe, buoni risultati si ottengono quanto più tale indice è vicino ad 1; al contrario tutti gli indici che riguardano l'errore globale evidenziano buone prestazioni del modello quanto più sono vicini a 0. Per cui nel primo caso (NSE) una variazione positiva sta ad indicare un incremento di prestazioni, nel secondo (indici di errore) è una variazione negativa ad indicare migliorie dei risultati. In ogni caso, a scanso di equivoci, nelle tabelle sono evidenziati con colore blu i miglioramenti, con colore magenta i peggioramenti.

Tabella 6.1: Principali indici statistici per la linea Canonical Kriging corretto via Top-Kriging. Parte 1.

T	NSE	NSE*	Δ_{NSE}	LNSE	LNSE*	Δ_{LNSE}
10	0.5865	0.3930	-0.1935	0.8442	0.6889	-0.1553
50	0.4411	0.2435	-0.1976	0.8276	0.8010	-0.0267
100	0.3854	0.1563	-0.2291	0.8259	0.7715	-0.0544
500	0.3261	0.2812	-0.0449	0.8127	0.7883	-0.0244

Tabella 6.2: Principali indici statistici per la linea Canonical Kriging corretto via Top-Kriging. Parte 2.

T	BIAS	BIAS*	Δ_{BIAS}	MARE	MARE*	Δ_{MARE}
10	0.2546	0.1507	-0.1039	0.5156	0.4796	-0.0360
50	0.2833	0.1346	-0.1487	0.5321	0.4885	-0.0436
100	0.2989	0.0967	-0.2022	0.5409	0.5109	-0.0300
500	0.3375	0.1847	-0.1528	0.5731	0.5220	-0.0511

Tabella 6.3: Principali indici statistici per la linea Canonical Kriging corretto via Top-Kriging. Parte 3.

T	RMSE	RMSE*	Δ_{RMSE}
10	0.8448	0.7465	-0.0983
50	0.9084	0.7273	-0.1811
100	0.9510	0.7642	-0.1868
500	1.0602	0.8137	-0.2465

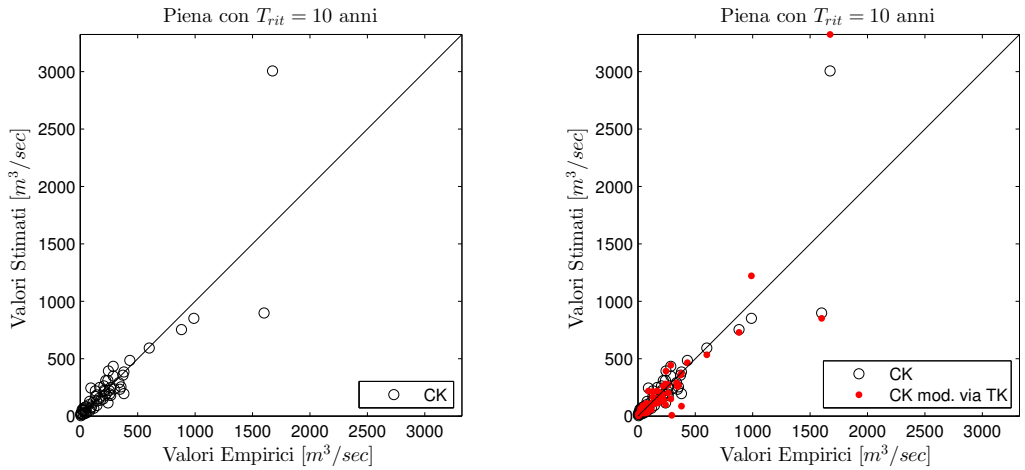

Figura 6.1: Diagrammi di dispersione dei valori di piena Q_{10}: CK e CK corretto via TK.

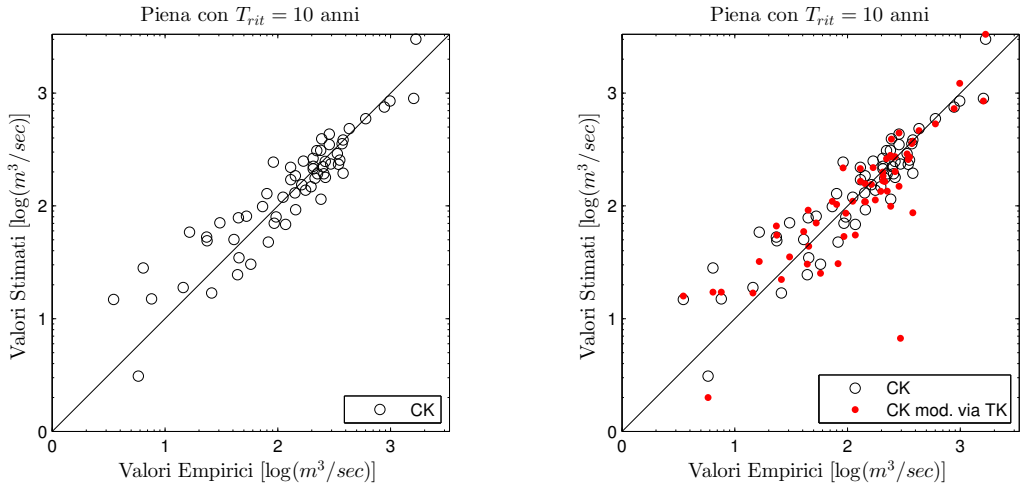

Figura 6.2: Diagrammi di dispersione in scala bilogaritmica dei valori di piena Q_{10}: CK e CK corretto via TK.

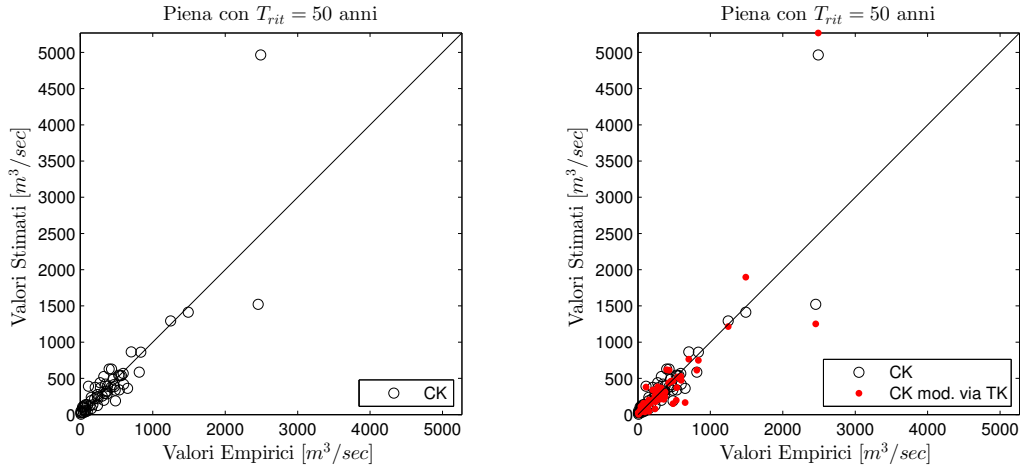

Figura 6.3: Diagrammi di dispersione dei valori di piena Q_{50}: CK e CK corretto via TK.

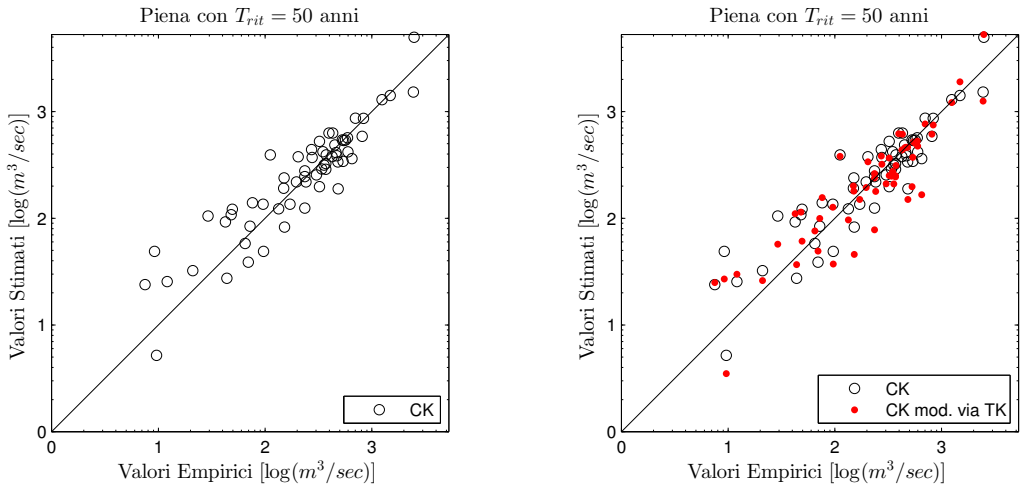

Figura 6.4: Diagrammi di dispersione in scala bilogaritmica dei valori di piena Q_{50}: CK e CK corretto via TK.

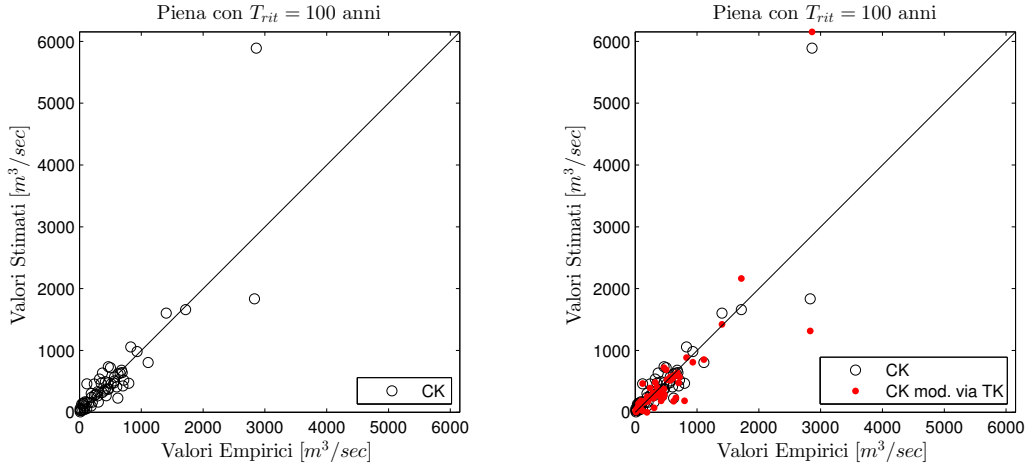

Figura 6.5: Diagrammi di dispersione dei valori di piena Q_{100}: CK e CK corretto via TK.

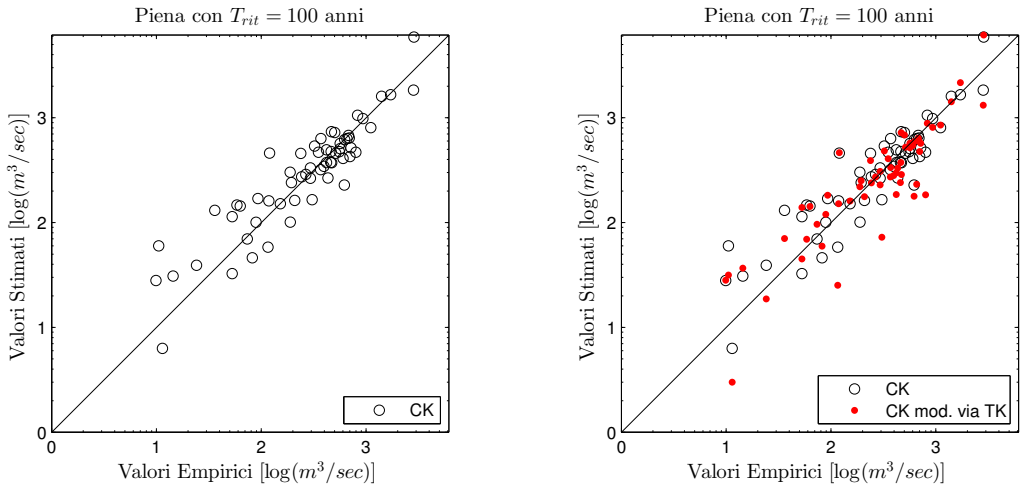

Figura 6.6: Diagrammi di dispersione in scala bilogaritmica dei valori di piena Q_{100}: CK e CK corretto via TK.

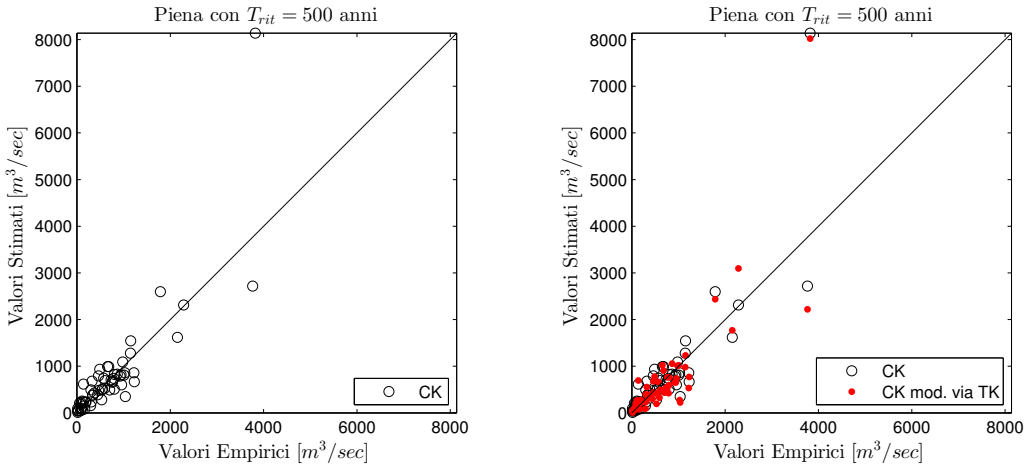

Figura 6.7: Diagrammi di dispersione dei valori di piena Q_{500}: CK e CK corretto via TK.

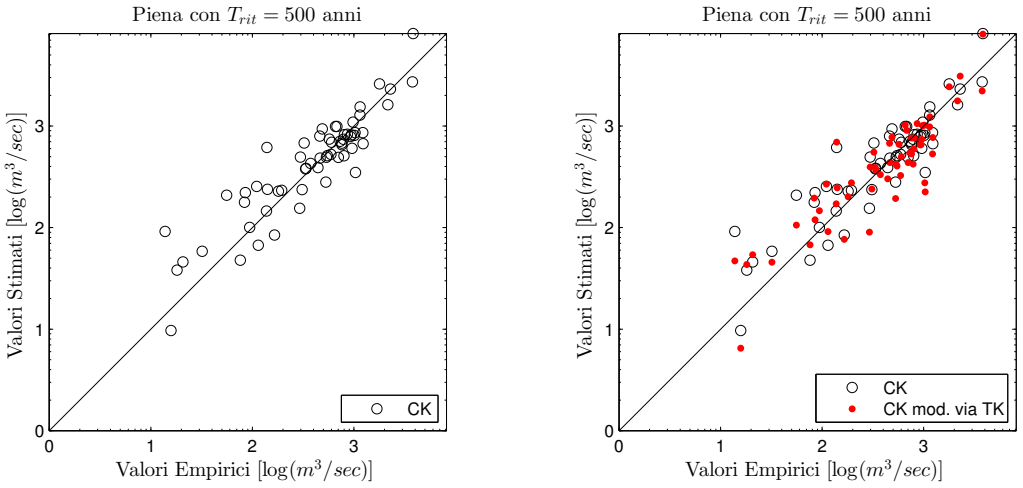

Figura 6.8: Diagrammi di dispersione in scala bilogaritmica dei valori di piena Q_{500}: CK e CK corretto via TK.

Top-Kriging corretto via Canonical Kriging

Da figura 6.9 a figura 6.16 s'illustra a sinistra la correlazione tra i valori ottenuti da Top-Kriging ottenuti in cross-validazione e i valori empirici; a destra, in verde, la correlazione con i valori ottenuti da Top-Kriging corretto via Canonical Kriging, sempre in cross-validazione, e i valori empirici, sovrapposta alla precedente.

Anche qui si riportano gli stessi indici statistici di riferimento associati alle due serie di valori cross-validati in funzione dei valori empirici: nelle tabelle 6.4, 6.5, 6.6 le colonne con intestazione asteriscata si riferiscono agli indici di prestazione calcolati per il modello Top-Kriging corretto via Canonical Kriging; quelle senza asterisco al Top-Kriging puro. Per la colonna Δ_{indice} si veda sopra.

Tabella 6.4: Principali indici statistici per la linea Top-Kriging corretto via Canonical Kriging. Parte 1.

T	NSE	NSE*	Δ_{NSE}	LNSE	LNSE*	Δ_{LNSE}
10	0.9330	0.7413	-0.1917	0.9020	0.9134	+0.0114
50	0.8982	0.5537	-0.3445	0.8766	0.8843	+0.0077
100	0.8774	0.5388	-0.3386	0.8628	0.8808	+0.0180
500	0.8262	0.2393	-0.5869	0.8207	0.7994	-0.0213

Tabella 6.5: Principali indici statistici per la linea Top-Kriging corretto via Canonical Kriging. Parte 2.

T	BIAS	BIAS*	Δ_{BIAS}	MARE	MARE*	Δ_{MARE}
10	0.1350	0.1429	+0.0079	0.3768	0.3465	-0.0303
50	0.1674	0.1642	-0.0032	0.4273	0.4036	-0.0237
100	0.1765	0.1735	-0.0030	0.4505	0.4177	-0.0328
500	0.2113	0.2040	-0.0073	0.5224	0.5169	-0.0055

Tabella 6.6: Principali indici statistici per la linea Top-Kriging corretto via Canonical Kriging. Parte 3.

T	RMSE	RMSE*	Δ_{RMSE}
10	0.5712	0.5032	-0.0680
50	0.6612	0.5998	-0.0614
100	0.6972	0.6202	-0.0770
500	0.8276	0.7614	-0.0662

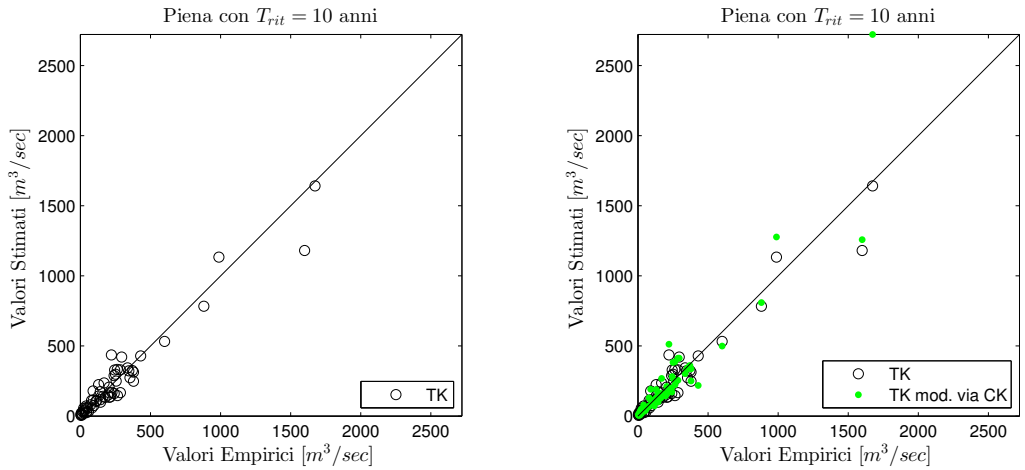

Figura 6.9: Diagrammi di dispersione dei valori di piena Q_{10}: TK e TK corretto via CK.

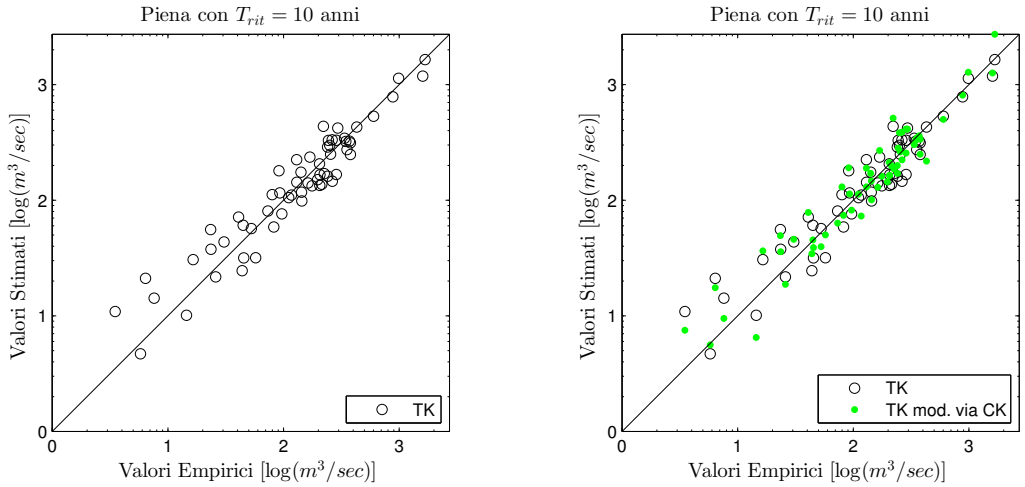

Figura 6.10: Diagrammi di dispersione in scala bilogaritmica dei valori di piena Q_{10}: TK e TK corretto via CK.

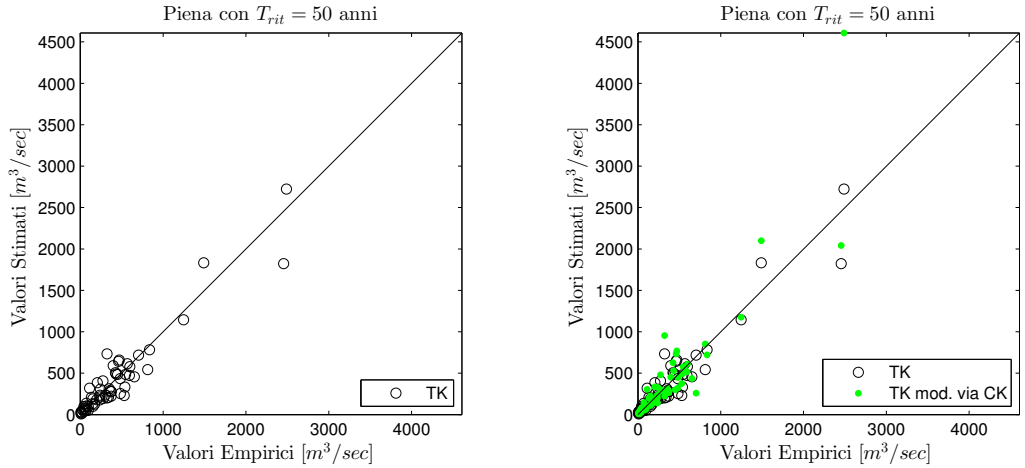

Figura 6.11: Diagrammi di dispersione dei valori di piena Q_{50}: TK e TK corretto via CK.

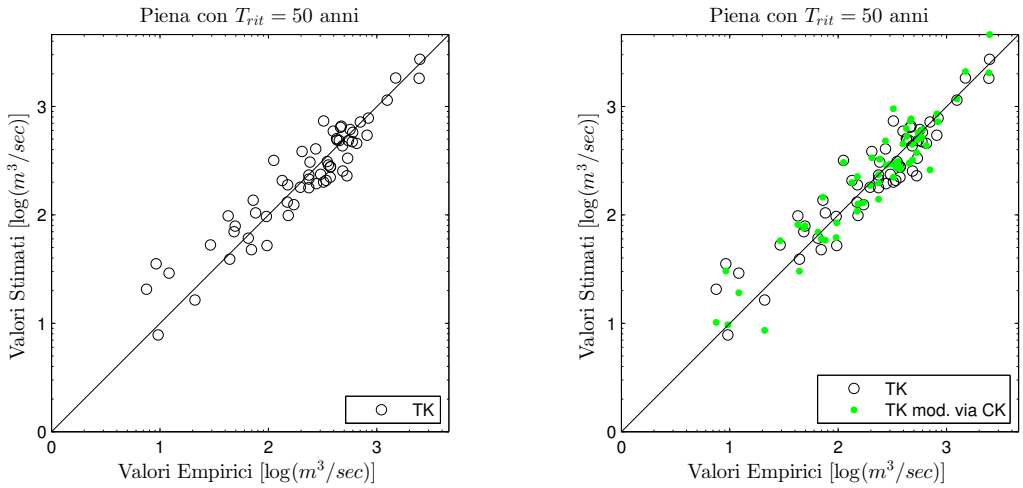

Figura 6.12: Diagrammi di dispersione in scala bilogaritmica dei valori di piena Q_{50}: TK e TK corretto via CK.

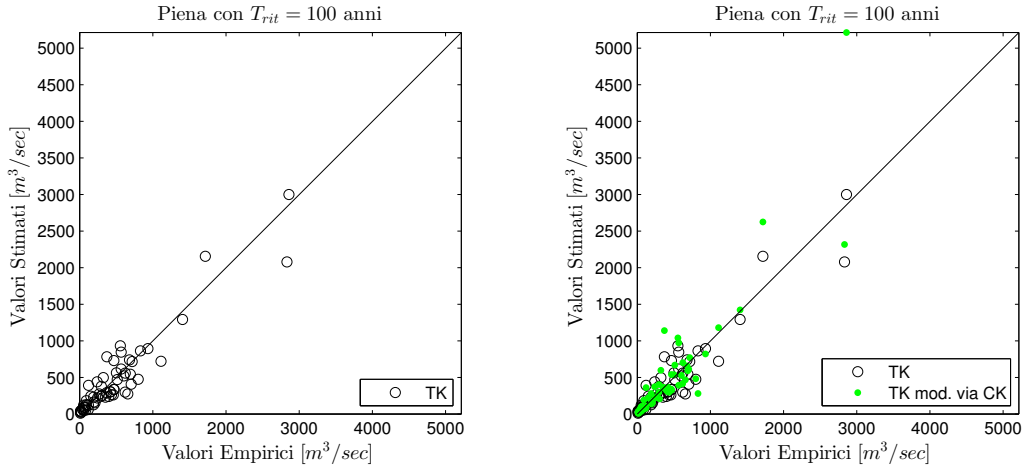

Figura 6.13: Diagrammi di dispersione dei valori di piena Q_{100}: TK e TK corretto via CK.

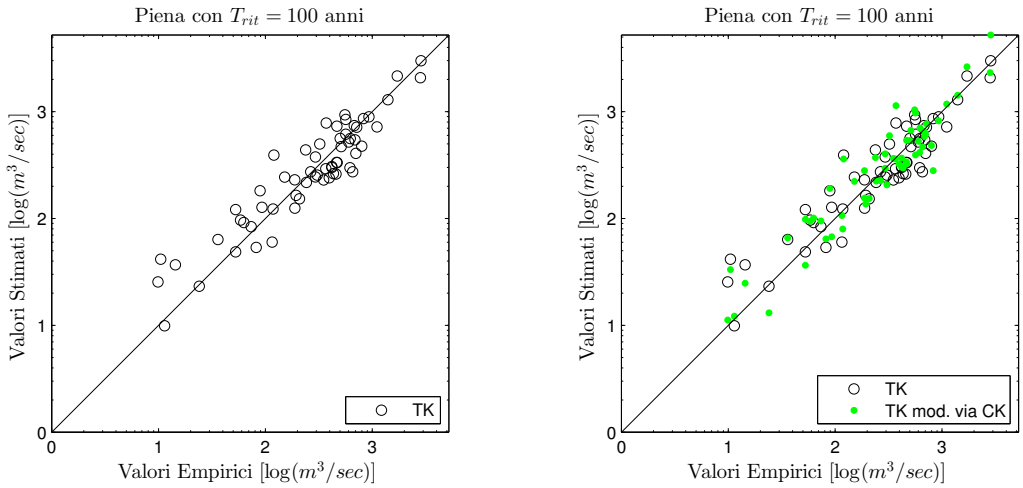

Figura 6.14: Diagrammi di dispersione in scala bilogaritmica dei valori di piena Q_{100}: TK e TK corretto via CK.

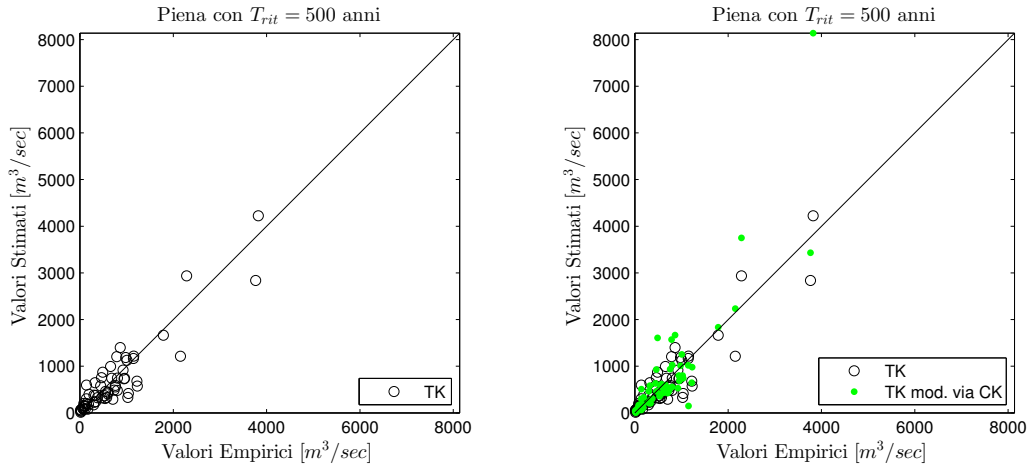

Figura 6.15: Diagrammi di dispersione dei valori di piena Q_{500}: TK e TK corretto via CK.

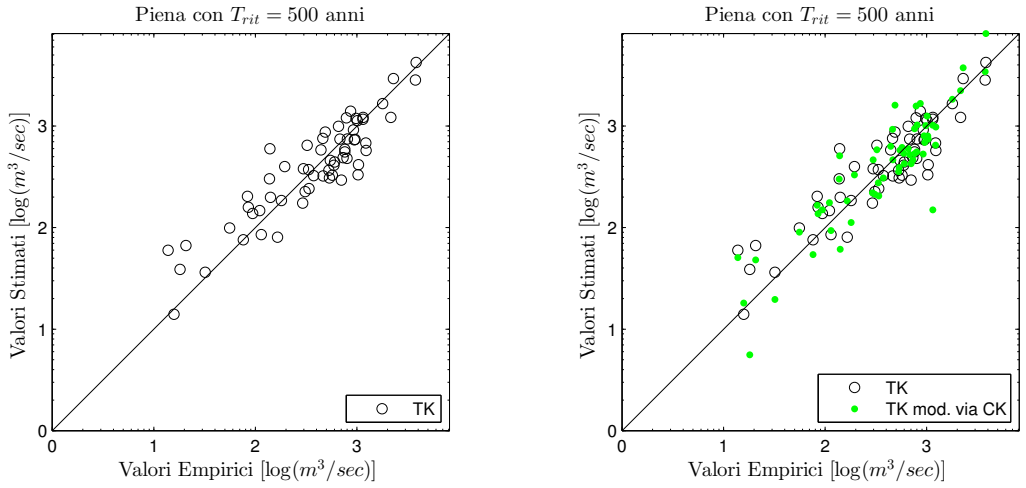

Figura 6.16: Diagrammi di dispersione in scala bilogaritmica dei valori di piena Q_{500}: TK e TK corretto via CK.

Considerazioni finali

Come risulta ben chiaro dalle tabelle riassuntive 6.1, 6.2, 6.3 nel caso di Canonical Kriging corretto dal Top-Kriging (colonne asteriscate) l'efficienza di Nash & Sutcliffe (NSE) cala rispetto all'utilizzo del Canonical Kriging puro, sia nel caso si utilizzi la coppia (valore simulato,valore osservato) in scala 1:1, sia in scala logaritmica, sinonimo di un peggioramento delle capacità predittive del modello. Questo risultato può essere spiegato dal fatto che l'NSE risulta significativamente influenzato dall'errore consistente commesso sui maggiori bacini, già nella prima interpolazione. Al contrario si evidenzia un buon risultato sugli indici rappresentanti l'errore *globale*, introdotto in fase di stima, mostrando tutti una riduzione significativa, specialmente in termini di errore medio relativo (BIAS) e di radice dell'errore quadratico medio (RMSE).

Nella seconda linea, facendo riferimento alle tabelle 6.4, 6.5, 6.6 si nota un peggioramento dell'NSE dei valori in scala 1:1, un miglioramento dello stesso per i valori confrontati in scala logaritmica (tranne che per le portate cinquecentennali), una sostanziale invarianza nell'errore medio relativo ed un miglioramento apprezzabile per errore medio relativo in valore assoluto (MARE) e la radice dell'errore quadratico medio (RMSE). Il Top-Kriging mostra già nella prima interpolazione ottimi risultati, potendo, infatti, apprezzare valori di NSE intorno a 0.9. Ciò mette in evidenza che, al contrario del Canonical Kriging, il Top-Kriging è molto più efficiente sui bacini grandi. In questo contesto risulta molto difficile riuscire ad incrementare significativamente l'efficienza, già di per se alta, modellando i residui con il Canonical Kriging.

In ogni caso, entrambe le interpolazioni effettuate sui residui mostrano una debole struttura spaziale, con la presenza di forti gradienti e difficilmente interpretabile da ciascuna delle due tecniche di interpolazione.

Ciononostante, specialmente per la prima linea Canonical Kriging corretto via Top-Kriging, si sono ottenuti risultati apprezzabili e prospettive promettenti di miglioramento da poter applicare in altri contesti.

Conclusioni

Questo lavoro di tesi è stato indirizzato all'analisi di alcuni modelli per la stima di variabili idrometriche in bacini idrografici non strumentati, ponendo particolari attenzioni ai residui generati in fase di stima. Attraverso tali modelli è possibile effettuare le regionalizzazione dell'informazione idrometrica senza dover ricorrere a quella fase critica di classificazione e raggruppamento dei bacini in regioni omogenee che, storicamente, ha contraddistinto buona parte delle tecniche di regionalizzazione.

L'analisi è stata condotta con riferimento a 61 bacini idrografici distribuiti su una vasta area degli U.S.A. sud-orientali, compresa tra gli stati Georgia, Alabama e Florida, di cui si disponevano sia dati geomorfologici (area del bacino, lunghezza asta principale, pendenza media, ecc...) e climatici (precipitazioni), sia dati idrometrici di portate di piena al colmo per assegnati tempi di ritorno Q_T con $T = 10, 50, 100, 500$.

Lo studio ha esaminato e messo a confronto due tecniche di interpolazione geostatistica basate sul *kriging*, ossia il Canonical Kriging e il Toplogical Kriging: il primo approccio, identificato con l'acronimo PSBI (*Physiografical Space Based Interpolazion*), considera la variabile idrometrica su un supporto puntuale definita su un piano sintetico dei descrittori geomorfologici e climatici, ottenuto mediante l'analisi della correlazione canonica; la seconda lavora su supporto non puntuale nello spazio geografico, tenendo conto delle dimensioni e della mutua posizione dei bacini. Nell'idea di poter sfruttare le potenzialità di entrambe, si sono analizzati i residui generati dai due modelli scambiando le tecniche d'interpolazione, per cui i residui generati dal Canonical Kriging sono stati modellati con il Top-Kriging e viceversa i residui ottenuti dalla stima con Top-Kriging modellati con Canonical Kriging. Per valutare l'affidabilità delle stime in un contesto di assenza di dati idrometrici, ciascuna interpolazione è stata applicata mediante una procedura di ricampionamento *jack-knife* consentendo di simulare in maniera efficace le condizioni non strumentate. La cross-validazione, quindi, crea l'opportunità di poter confrontare per la stessa stazione di misura i valori stimati e i valori empirici mediante l'utilizzo di opportuni indici statistici.

La prima linea d'intervento ha mostrato per ogni valore di portata un generale peggioramento della capacità predittiva, in termini di efficienza di Nash & Sutcliffe (NSE),

Tabella 6.7: Comportamento anomalo di 2 bacini nel dataset

ID	$\mathbf{Q_{100}}$ (m^3/s)	**A** (km^2)	$\frac{\mathbf{Q_{100}}}{\mathbf{A^{0.65}}}$
02347500	2831.7	4791.2	11.5
02352500	2860.0	13752.1	5.8
		Media	8.7
		Mediana	7.9

dell'approccio Canonical Kriging corretto via Top-Kriging; tuttavia si è notato una sensibile e non trascurabile diminuzione dell'errore *globale* sia in termini di errore medio relativo (BIAS), sia di errore medio relativo in valore assoluto (MARE) e sia di radice dell'errore quadratico medio (RMSE). Facendo riferimento alla portata centennale, una spiegazione di questo anomalo comportamento può essere identificata nella presenza nel dataset di due bacini con portate simili ma con portate specifiche ($\frac{Q_{100}}{A^{0.65}}$) molto diverse tra loro: una eccessivamente alta e l'altra eccessivamente bassa rispetto alla media e alla mediana (v. tab. 6.7). In questa situazione il Canonical Kriging sovrastima pesantemente l'una e sottostima significativamente l'altra impattando negativamente sull'efficienza di Nash & Sutcliffe. In tali condizioni l'applicazione del Top-Kriging sui residui non può migliorare i rendimenti perché gli scarti tra variabile osservata e simulata, afferenti ai suddetti bacini, risultano maggiori rispetto alla media di 1 o 2 ordini di grandezza, riflettendosi negativamente sull'efficienza del primo approccio.

La seconda linea d'intervento, caratterizzata da una prima interpolazione basata sul Top-Kriging, ha mostrato buoni risultati già nella prima fase. Sempre con riferimento alla portata centennale, i rendimenti del Top-Kriging in termini di NSE e LNSE risultano molto elevati (86-87%), negando, di fatto, la possibilità di un miglioramento significativo con la modellazione dei residui attraverso il Canonical Kriging. In generale si è riscontrato un lieve incremento dell'efficienza di Nash & Sutcliffe (in scala logaritmica) e una leggera diminuzione degli indici di errore globale.

In conclusione il primo approccio ha messo in evidenza che il Canonical Kriging sembra particolarmente soffrire la possibile presenza di *outliers* (dati anomali) nella base dati, tuttavia l'applicazione del Top-Kriging può migliorare l'errore indotto nella prima stima; al contrario il secondo approccio ha dimostrato che il Top-Kriging risulta un interpolatore molto robusto e performante, non risente della presenza di outliers ed è di facile applicazione.

Appendice A

Dataset completo

Dati idrometrici

Tabella A.1: Portate di piena al colmo con assegnato tempo di ritorno $T = 10, 50, 100, 500$ anni. GA=Georgia; AL=Alabama; FL=Florida.

N°	**ID**	**Stazione di misura**	**Stato**	$\mathbf{Q_{10}}$ $\left(\frac{m^3}{s}\right)$	$\mathbf{Q_{50}}$ $\left(\frac{m^3}{s}\right)$	$\mathbf{Q_{100}}$ $\left(\frac{m^3}{s}\right)$	$\mathbf{Q_{500}}$ $\left(\frac{m^3}{s}\right)$
1	02330450	Chattahoochee River at Helen	GA	130,3	235,6	294,5	464,4
2	02331000	Chattahoochee River near Leaf	GA	371,0	546,5	623,0	804,2
3	02331500	Soque River at St Rt 105 near Demorest	GA	286,0	430,4	501,2	679,6
4	02331600	Chattahoochee River near Cornelia	GA	600,3	835,3	931,6	1144,0
5	02333000	Chattahoochee River near Gainesville	GA	880,7	1245,9	1404,5	1786,8
6	02333500	Chestatee River near Dahlonega	GA	379,4	591,8	690,9	943,0
7	02337000	Sweetwater Creek near Austell	GA	221,2	322,8	371,0	487,0
8	02337400	Dog River near Douglasville	GA	214,1	339,8	396,4	535,2
9	02337500	Snake Creek near Whitesburg	GA	162,8	235,6	265,6	334,1
10	02338660	New River at Ga 100, near Corinth	GA	205,0	348,3	419,1	603,1
11	02338840	Yellowjacket Creek-Hammett Rd, Blw Hogansville	GA	91,2	111,9	120,1	138,8
12	02339000	Yellowjacket Creek near La Grange	GA	286,0	529,5	657,0	1025,1
13	02339225	Wehadkee Creek Below Rock Mills	AL	206,1	371,0	458,7	702,3
14	02340250	Flat Shoal Creek at Ga 18, near West Point	GA	244,7	396,4	467,2	657,0
15	02340500	Mountain Oak Creek near Hamilton	GA	143,0	276,1	351,1	572,0
16	02340750	Osanippa Creek near Fairfax	AL	225,1	356,8	419,1	591,8
17	02341220	Mulberry Creek near Mulberry Grove	GA	263,3	461,6	566,3	863,7
18	02341600	Juniper Creek near Geneva	GA	44,7	76,2	92,9	140,7
19	02341723	Pine Knot Creek at Ga 355, near Juniper	GA	23,4	42,5	52,7	83,5
20	02341800	Upatoi Creek near Columbus	GA	379,4	651,3	798,5	1220,5
21	02341900	Ochillee Creek at Hourglass Road near Cussetta	GA	73,3	149,2	193,7	337,0
22	02342200	Phelps Creek near Opelika	AL	57,5	96,8	115,8	165,1
23	02343200	Pataula Creek near Lumpkin	GA	143,8	322,8	433,2	798,5
24	02343219	Bluff Springs Branch at Ga 27, near Lumpkin	GA	7,6	12,1	14,4	20,7
25	02343225	Pataula Creek near Georgetown	GA	354,0	812,7	1110,0	2152,1
26	02343244	Cemochechobee Creek near Coleman	GA	23,6	48,1	62,9	110,4
27	02343267	Temple Creek at Ga 39, near Blakely	GA	3,5	7,5	9,9	18,1
28	02344700	Line Creek near Senoia	GA	241,3	484,2	620,1	1036,4
29	02346180	Flint River near Thomaston	GA	988,3	1489,5	1716,0	2285,2
30	02346193	Scott Creek near Talbotton	GA	43,9	69,7	82,1	114,1
31	02346195	Lazer Creek at Ga 41, near Talbotton	GA	264,5	538,0	702,3	1229,0
32	02346210	Kimbrough Creek near Talbotton	GA	45,3	65,1	73,6	94,0
33	02346217	Coleoatchee Creek near Manchester	GA	25,9	43,9	52,7	75,9
34	02346500	Potato Creek near Thomaston	GA	250,6	424,8	512,5	758,9

Tabella A.1: continua nella prossima pagina

Tabella A.1: continua dalla pagina precedente

N°	ID	Stazione di misura	Stato	$\mathbf{Q_{10}}$ $\left(\frac{m^3}{s}\right)$	$\mathbf{Q_{50}}$ $\left(\frac{m^3}{s}\right)$	$\mathbf{Q_{100}}$ $\left(\frac{m^3}{s}\right)$	$\mathbf{Q_{500}}$ $\left(\frac{m^3}{s}\right)$
35	02347500	Flint River near Culloden	GA	1599,9	2455,1	2831,7	3766,1
36	02348300	Patsiliga Creek near Reynolds	GA	117,2	235,6	305,8	529,5
37	02348485	Whitewater Creek at Ga 137, near Butler	GA	6,4	9,2	10,5	13,8
38	02349000	Whitewater Cr Below Rambulette Cr Nr Butler	GA	30,6	49,3	58,6	85,2
39	02349030	Cedar Creek at Us 19, near Rupert	GA	16,5	29,2	36,0	55,8
40	02349350	Buck Creek at Us 19, near Ellaville	GA	79,9	150,6	188,6	297,3
41	02349695	Horsehead Creek at Ga 224, near Montezuma	GA	5,8	9,6	11,4	15,8
42	02349900	Turkey Creek at Byromville	GA	82,4	152,3	188,9	291,7
43	02350520	Abrams Creek Tributary near Doles	GA	14,5	21,0	24,1	32,0
44	02350600	Kinchafoonee Creek at Preston	GA	176,4	303,0	368,1	546,5
45	02350900	Kinchafoonee Creek near Dawson	GA	240,1	450,2	566,3	923,1
46	02351500	Muckalee Creek near Americus	GA	111,0	196,8	242,4	373,8
47	02351700	Muckalee Creek near Smithville	GA	141,6	241,8	294,5	441,7
48	02351800	Muckaloochee Creek at Smithville	GA	52,7	95,7	118,1	180,1
49	02351890	Muckalee Creek at Ga 195, near Leesburg	GA	206,7	373,8	464,4	730,6
50	02351900	Muckalee Creek near Leesburg	GA	168,8	272,4	322,8	461,6
51	02352500	Flint River at Albany	GA	1673,5	2489,1	2860,0	3822,8
52	02353100	Ichawaynochaway C (Us Hwy 82) Nr Graves	GA	92,9	133,9	151,8	194,5
53	02353200	Little Ichawaynochaway Creek near Shellman	GA	40,8	72,2	89,2	137,6
54	02353400	Pachitla Creek near Edison	GA	197,1	371,0	470,1	758,9
55	02353500	Ichawaynochaway Creek at Milford	GA	337,0	566,3	679,6	979,8
56	02354500	Chickasawhatchee Creek at Elmodel	GA	130,0	204,4	238,4	322,8
57	02355000	Ichawaynochaway Creek near Newton	GA	430,4	702,3	829,7	1149,7
58	02356100	Spring Creek near Arlington	GA	96,6	171,3	208,7	308,7
59	02356640	Spring Creek at Us 27, at Colquitt	GA	255,4	481,4	603,1	957,1
60	02357000	Spring Creek near Iron City	GA	345,5	597,5	716,4	1008,1
61	02359000	Chipola River Nr Altha	FL	294,5	470,1	555,0	787,2

Dati geomorfologici e climatici

Tabella A.2: Dati geomorfologici. Parte 1. Per la descrizione delle variabili consultare la tabella 4.5.

N°	**A** (km^2)	**LAT** (^oN)	**LONG** (^oE)	**L** (km)	**S** $\left(\frac{m}{km}\right)$	**P** (km)	$\mathbf{F_f}$ $(-)$	$\mathbf{H_m}$ (m)	$\mathbf{H_{max}}$ (m)	$\mathbf{H_{min}}$ (m)	$\mathbf{S_m}$ (%)
1	115,8	34,761 731	−83,758 385	20,5	14,2724	70,6	3,632	758,2	1347,3	435,3	33,0
2	388,5	34,702 945	−83,708 450	47,8	4,4710	142,1	5,876	591,3	1347,3	376,0	22,9
3	404,0	34,662 315	−83,535 129	59,9	3,1354	162,9	8,870	500,0	1341,1	362,4	13,9
4	815,8	34,678 446	−83,620 346	65,8	2,3285	214,6	5,304	540,9	1347,3	344,5	18,1
5	1447,7	34,579 043	−83,678 992	118,0	1,3730	328,9	9,613	475,8	1347,3	321,7	14,4
6	396,2	34,626 200	−83,866 944	47,4	3,7658	145,8	5,670	558,2	1351,5	348,1	20,3
7	637,1	33,846 032	−84,735 927	64,4	0,5900	183,9	6,500	309,5	544,1	260,5	5,5
8	121,7	33,674 438	−84,902 446	18,9	3,3466	73,2	2,935	338,4	413,9	277,1	6,8
9	91,9	33,575 397	−84,967 631	24,1	2,8086	59,6	6,292	331,7	422,2	255,5	8,1
10	328,9	33,306 653	−84,882 674	34,1	1,3102	110,6	3,526	241,9	305,3	195,0	6,4
11	235,7	33,178 844	−84,855 384	33,2	1,2255	109,9	4,686	243,0	293,3	198,0	5,6
12	471,4	33,131 704	−84,885 725	45,5	0,9042	156,2	4,397	238,3	299,0	192,3	5,4
13	155,9	33,203 908	−85,283 836	35,2	3,4996	64,0	7,936	272,2	421,5	200,1	8,9
14	528,3	32,929 370	−84,879 017	62,2	1,0501	154,8	7,331	250,5	411,7	176,0	5,8
15	159,8	32,793 278	−84,958 531	43,1	1,5250	90,6	11,629	233,1	400,7	162,9	6,7
16	258,2	32,803 057	−85,332 292	44,5	0,7727	79,3	7,668	216,6	280,9	173,4	7,1
17	492,1	32,734 237	−84,796 328	57,4	1,2388	140,6	6,699	230,3	422,7	160,4	6,1
18	122,8	32,541 289	−84,508 984	18,1	2,0191	73,9	2,661	179,5	242,6	119,5	5,3
19	81,3	32,442 432	−84,600 102	15,7	4,0089	55,5	3,037	164,8	246,4	105,0	6,0

Tabella A.2: continua nella prossima pagina

Tabella A.2: continua dalla pagina precedente

N°	**A** (km^2)	**LAT** (^oN)	**LONG** (^oE)	**L** (km)	**S** ($\frac{m}{km}$)	**P** (km)	**F_f** ($-$)	**H_m** (m)	**H_{max}** (m)	**H_{min}** (m)	**S_m** (%)
20	885,7	32,520 005	−84,652 725	66,9	0,9920	205,5	5,057	156,9	261,8	82,0	5,8
21	138,0	32,338 164	−84,731 326	25,7	1,9223	82,1	4,796	146,2	235,2	86,7	6,5
22	17,3	32,594 443	−85,277 317	9,0	4,3486	19,2	4,642	195,4	252,6	164,2	6,4
23	181,3	31,984 219	−84,730 742	27,6	2,1582	99,1	4,215	137,8	199,5	88,3	5,9
24	7,7	32,050 769	−84,892 254	5,9	11,1779	17,3	4,519	166,4	201,2	114,2	6,6
25	764,0	31,950 642	−84,842 100	53,6	1,1089	200,3	3,757	130,6	205,4	65,5	6,7
26	39,6	31,676 412	−84,851 595	12,3	2,7603	41,1	3,808	119,0	149,3	78,5	5,7
27	6,8	31,444 179	−84,995 614	4,1	0,7476	16,3	2,482	95,8	103,6	91,4	1,0
28	261,6	33,435 832	−84,613 857	40,8	1,0947	113,3	6,368	272,7	323,1	225,0	4,9
29	3159,6	33,231 184	−84,535 137	151,1	0,4316	409,0	7,230	257,0	419,3	145,9	4,8
30	8,7	32,655 339	−84,621 915	6,7	4,5236	21,2	5,094	207,6	242,8	180,9	4,1
31	210,6	32,746 531	−84,626 565	26,8	3,4432	99,2	3,415	225,6	419,3	165,0	6,6
32	17,1	32,667 726	−84,518 379	8,9	4,6218	26,8	4,616	205,9	259,1	174,2	5,2
33	7,3	32,837 854	−84,604 020	5,0	7,6184	16,1	3,461	283,4	364,0	241,5	8,5
34	481,7	33,052 755	−84,286 964	60,2	0,4987	175,5	7,522	245,9	395,2	193,3	5,1
35	4791,2	33,107 740	−84,480 136	187,9	0,6707	501,0	7,365	241,8	419,3	102,3	5,3
36	339,3	32,622 024	−84,266 991	62,0	0,8082	146,2	11,343	167,0	241,8	100,0	5,0
37	44,8	32,515 723	−84,383 497	13,2	6,2135	43,4	3,915	187,5	255,0	127,2	5,7
38	212,9	32,527 997	−84,350 518	31,2	2,1582	90,6	4,580	179,3	255,0	110,3	5,3
39	106,4	32,425 228	−84,373 629	23,5	3,0581	68,1	5,182	180,8	252,2	118,9	5,2
40	378,1	32,372 287	−84,439 942	51,9	1,4031	156,2	7,115	174,1	252,9	107,2	5,6
41	1,9	32,364 110	−83,933 951	2,9	2,2788	5,7	4,363	132,2	158,0	117,6	4,9
42	116,5	32,240 957	−83,839 012	22,6	0,5538	95,9	4,371	124,3	166,3	94,2	3,6
43	9,8	31,690 266	−83,773 526	6,4	3,9361	20,4	4,132	115,1	139,6	91,2	4,0
44	510,2	32,211 608	−84,601 870	49,5	0,8459	159,3	4,812	167,4	246,3	99,7	5,6
45	1364,8	32,069 424	−84,500 626	109,7	0,6872	293,4	8,816	141,8	246,3	66,9	4,3
46	362,6	32,188 682	−84,356 477	44,8	1,0642	149,3	5,525	152,5	226,8	97,0	4,7
47	686,3	32,107 426	−84,293 665	72,4	0,8380	233,7	7,634	135,6	226,8	80,6	4,2
48	121,7	31,998 449	−84,316 985	29,7	2,0487	89,2	7,234	128,1	176,2	85,2	3,6
49	937,5	32,053 642	−84,276 946	91,2	0,7342	277,6	8,869	126,9	226,8	66,9	3,7
50	1048,9	32,024 930	−84,259 238	98,2	0,7435	303,1	9,199	122,2	226,8	66,6	3,4
51	13 752,1	32,484 864	−84,270 614	410,5	0,3971	1071,6	12,255	165,7	419,3	47,1	4,4
52	305,6	31,878 654	−84,576 264	34,7	1,2919	113,7	3,938	129,5	176,7	84,0	3,9
53	126,4	31,840 964	−84,676 129	22,0	1,8475	75,3	3,839	135,3	176,7	88,3	4,0
54	486,9	31,686 265	−84,743 777	41,4	1,1295	153,6	3,520	112,8	164,5	67,0	3,4
55	1605,7	31,681 752	−84,648 075	95,5	0,6118	297,3	5,684	103,6	176,7	48,7	2,7
56	828,8	31,592 115	−84,428 487	73,6	0,5799	216,6	6,545	74,0	127,9	45,4	1,0
57	2641,6	31,634 891	−84,571 970	116,6	0,5177	345,3	5,147	91,4	176,7	36,5	2,1
58	126,9	31,505 053	−84,814 377	27,3	1,5531	79,5	5,859	85,8	124,9	51,9	1,6
59	727,7	31,371 423	−84,805 592	65,2	0,6171	208,3	5,840	67,9	124,9	38,0	1,5
60	1256,1	31,292 526	−84,793 948	87,0	0,4908	286,8	6,032	61,5	124,9	31,3	1,4
61	2022,7	30,885 883	−85,279 113	109,9	0,4249	312,8	5,971	47,3	116,0	9,3	2,5

Tabella A.3: Dati geomorfologici (parte 2) e climatici (precipitazioni).

N°	**F_i** (%)	**F_{for}** (%)	**I_d** ($-$)	**I_h** ($-$)	**D_d** ($\frac{km}{km^2}$)	**MAP** (mm)	**MDP_2** (mm)	**MDP_{10}** (mm)	**MDP_{25}** (mm)	**MDP_{50}** (mm)	**MDP_{100}** (mm)
1	0,23	94,63	2,86	2,10	1,34	1899,2	126,5	177,7	205,4	227,4	255,1
2	0,46	16,42	3,03	2,12	1,24	1786,2	122,1	173,5	202,5	220,1	250,4
3	1,51	63,02	3,12	2,10	1,38	1675,9	117,3	169,0	196,1	213,9	246,0
4	0,99	72,26	3,08	2,11	1,33	1724,6	119,2	170,6	198,5	216,4	247,7
5	1,74	63,07	3,05	2,07	0,80	1587,5	111,9	162,5	189,2	210,0	232,1
6	0,76	77,33	3,10	2,23	1,36	1726,1	114,5	162,6	191,9	208,3	237,7
7	7,95	44,12	3,05	2,23	1,23	1383,1	101,6	152,4	177,8	201,7	203,2
8	2,19	60,65	3,13	2,06	1,51	1373,4	101,6	152,4	177,8	203,2	203,2
9	1,26	63,91	3,12	2,06	1,30	1361,6	101,6	152,4	177,8	200,4	203,3

Tabella A.3: continua nella prossima pagina

Tabella A.3: continua dalla pagina precedente

N°	$\mathbf{F_i}$ (%)	$\mathbf{F_{for}}$ (%)	$\mathbf{I_d}$ (−)	$\mathbf{I_h}$ (−)	$\mathbf{D_d}$ $\left(\frac{km}{km^2}\right)$	**MAP** (mm)	$\mathbf{MDP_2}$ (mm)	$\mathbf{MDP_{10}}$ (mm)	$\mathbf{MDP_{25}}$ (mm)	$\mathbf{MDP_{50}}$ (mm)	$\mathbf{MDP_{100}}$ (mm)
10	1,51	60,36	3,01	2,01	1,48	1330,3	101,6	152,4	177,8	203,2	206,7
11	1,15	62,02	3,00	2,00	1,42	1397,0	101,6	152,4	177,8	203,2	208,8
12	1,14	60,20	2,96	1,97	1,35	1397,0	101,6	152,4	177,8	203,2	209,9
13	0,59	59,36	3,10	2,09	1,41	1397,0	101,6	152,4	177,8	197,6	212,2
14	0,70	59,00	3,01	2,06	1,48	1299,3	101,6	152,7	177,8	203,2	213,5
15	0,62	71,92	3,07	2,25	1,53	1279,6	101,6	154,5	178,3	203,2	216,6
16	0,71	63,33	3,01	2,26	1,47	1397,0	103,7	157,2	181,2	202,3	219,7
17	0,72	68,92	3,17	2,31	1,60	1397,0	101,6	154,1	178,0	203,2	216,3
18	0,57	31,30	2,80	2,02	1,21	1269,6	101,6	154,0	177,9	199,6	217,5
19	0,31	63,17	2,87	1,99	1,27	1268,2	101,6	156,1	180,0	200,8	219,9
20	0,56	66,97	2,88	2,03	1,95	1269,0	101,6	155,7	179,4	201,2	219,1
21	0,43	76,88	2,93	2,26	1,84	1272,9	102,2	158,7	183,5	202,2	222,8
22	0,53	72,87	3,10	2,00	1,28	1397,0	105,1	159,6	184,4	203,2	222,9
23	0,15	64,63	3,10	2,35	1,28	1281,7	105,8	163,8	190,1	208,5	228,3
24	0,14	68,41	3,06	2,24	1,17	1287,2	106,8	163,9	190,9	209,5	228,7
25	0,19	65,76	3,18	2,42	1,27	1300,2	107,3	165,1	192,1	211,0	231,4
26	0,22	65,78	3,09	2,05	1,15	1397,0	110,5	169,5	197,5	217,6	240,4
27	0,44	14,30	3,40	2,10	1,24	1390,5	114,8	174,2	203,8	225,2	250,8
28	7,03	45,60	3,07	2,04	1,51	1314,9	101,6	152,4	177,8	203,2	203,2
29	4,30	26,37	3,08	2,14	1,22	1286,6	101,6	152,4	177,8	202,8	206,2
30	0,32	62,47	3,05	2,03	1,33	1397,0	101,6	153,8	177,8	201,4	216,8
31	0,43	70,72	3,09	2,14	1,51	1397,0	101,6	152,8	177,8	202,2	215,0
32	0,65	72,36	3,01	2,01	1,54	1397,0	101,6	152,6	177,8	199,7	215,7
33	0,83	26,91	3,00	2,86	1,01	1280,7	101,6	152,4	177,8	202,6	213,2
34	2,30	21,99	3,08	2,13	1,37	1268,8	101,6	152,4	177,8	199,1	207,4
35	3,22	55,68	3,08	2,14	1,29	1281,1	101,6	152,4	177,8	201,5	207,9
36	0,50	55,33	2,40	1,90	1,29	1238,3	101,6	152,4	177,8	197,7	214,1
37	0,34	59,47	2,62	2,02	0,82	1079,5	101,6	153,0	177,8	199,1	216,6
38	0,48	46,49	2,64	2,02	0,92	1252,7	101,6	152,7	177,8	198,8	216,1
39	0,49	64,29	2,80	2,01	1,23	1254,4	101,6	154,2	177,9	199,7	217,9
40	0,34	59,12	2,90	2,07	1,26	1261,8	101,6	155,5	179,4	200,5	219,3
41	0,32	3,54	3,50	2,25	1,05	1190,0	101,6	152,4	177,8	197,9	216,0
42	0,56	26,50	3,67	2,36	1,41	1204,1	101,6	152,4	177,8	198,2	217,2
43	0,32	64,95	3,66	2,57	5,76	1244,6	101,6	157,4	182,7	203,2	225,4
44	0,25	67,86	3,09	2,25	1,38	1273,3	102,3	159,3	184,3	202,7	223,4
45	0,29	55,40	3,32	2,26	1,22	1276,2	102,6	160,4	185,8	203,6	224,5
46	0,71	56,28	3,33	2,24	1,32	1259,4	101,6	157,3	181,8	201,4	221,4
47	1,42	46,05	3,53	2,29	1,30	1079,5	101,6	157,6	182,6	201,7	222,2
48	0,53	41,32	3,72	2,33	1,24	1265,0	101,6	159,4	184,9	202,6	224,0
49	1,17	26,79	3,63	2,34	1,27	1262,2	101,6	158,1	183,3	202,0	222,9
50	1,14	39,94	3,64	2,35	1,22	1262,8	101,6	158,2	183,6	202,1	223,2
51	1,68	48,79	3,30	2,24	1,23	1258,5	101,8	154,8	180,2	201,2	216,2
52	0,35	48,15	3,24	2,10	0,88	1079,5	105,2	164,0	190,4	208,5	228,2
53	0,19	51,36	2,88	1,83	0,75	1079,5	106,8	165,6	192,3	211,0	231,0
54	0,41	47,63	3,24	2,12	1,09	1341,2	109,2	168,5	196,0	215,7	237,7
55	0,41	38,91	3,46	2,21	0,97	1329,4	108,3	167,5	195,0	214,3	235,8
56	0,55	36,30	3,92	2,46	0,88	1324,1	107,1	166,1	193,9	212,7	233,8
57	0,45	37,76	3,62	2,30	0,92	1337,5	108,1	167,2	194,9	214,1	235,6
58	0,35	22,67	3,94	2,46	1,28	1363,2	112,2	171,8	200,4	221,1	245,3
59	0,73	26,67	4,03	2,56	1,08	1397,0	114,0	173,6	203,1	224,1	249,5
60	0,77	25,86	4,04	2,57	0,97	1384,6	114,8	174,7	205,0	225,9	252,0
61	0,79	19,92	3,44	2,21	0,88	1397,0	123,8	194,2	226,5	256,3	282,9

Bibliografia

Brath, A. (1995). «Metodologie di valutazione delle portate di piena». In: *Moderni criteri per la sistemazione degli alvei fluviali*. A cura di U. Maione e A. Brath. Atti del I corso di aggiornamento, Politecnico di Milano,10-14 ottobre 1994. Cosenza: Editoriale BIOS.

Castiglioni, S. (2009). «Modelli per la stima delle risorse idriche superficiali in bacini idrografici non strumentati». Tesi di dott. Alma Mater Studiorum - Università di Bologna.

Castiglioni, S., A. Castellarin e A. Montanari (2008). «Stima delle portate di magra in siti non strumentati mediante tecniche di interpolazione spaziale». In: *XXXI Convegno Nazionale di Idraulica e Costruzioni Idrauliche*.

— (2009). «Prediction of low-flow indices in ungauged basins through physiographical space-based interpolation». In: *Journal of Hydrology* 378, pp. 272 –280. ISSN: 0022-1694. DOI: 10.1016/j.jhydrol.2009.09.032. URL: http://www.sciencedirect.com/science/article/pii/S0022169409006064.

Castiglioni, S. et al. (2011). «Smooth regional estimation of low-flow indices: physiographical space based interpolation and top-kriging». In: *Hydrology and Earth System Sciences* 15.3, pp. 715–727. DOI: 10.5194/hess-15-715-2011. URL: http://www.hydrol-earth-syst-sci.net/15/715/2011/.

Chokmani, K. e T. B. M. J. Ouarda (dic. 2004). «Physiographical space-based kriging for regional flood frequency estimation at ungauged sites». In: *Water Resour. Res.* 40.12, W12514–. ISSN: 0043-1397. URL: http://dx.doi.org/10.1029/2003WR002983.

Gotvald, A. J., T. D. Feaster e J. C. Weaver (2009). *Magnitude and Frequency of Rural Floods in the Southeastern United States*. Rapp. tecn. United States Geological Survey. URL: http://pubs.usgs.gov/sir/2009/5043/.

Morichini, M. (2006). «Applicazione delle tecniche di interpolazione geografica nello spazio dei descrittori idrologici per la stima della piena di progetto». Tesi di laurea mag. Alma Mater Studiorum - Università di Bologna.

Nezhad, M. Kamali et al. (2010). «Regional flood frequency analysis using residual kriging in physiographical space». In: *Hydrological Processes* 24.15, pp. 2045–2055. ISSN: 1099-1085. DOI: 10.1002/hyp.7631. URL: http://dx.doi.org/10.1002/hyp.7631.

Ouarda, Taha B.M.J. et al. (2001). «Regional flood frequency estimation with canonical correlation analysis». In: *Journal of Hydrology* 254.1–4, pp. 157 –173. ISSN: 0022-1694. DOI: 10.1016/S0022-1694(01)00488-7. URL: http://www.sciencedirect.com/science/article/pii/S0022169401004887.

Pollice, A. (2011). «Analisi della correlazione canonica». Dispense del corso Statistica Multivariata - Università di Bari. URL: http://www.dip-statistica.uniba.it/html/docenti/pollice/sm2012/Dispense/disp8.pdf.

Raspa, G. e R. Bruno (1994a). *Dispense di Geostatistica Applicata*. Capitolo 3 - Geostatistica di base. URL: http://w3.uniroma1.it/geostatistica/Geostatistica/Dispense.pdf.

— (1994b). *La pratica della geostatistica lineare: il trattamento dei dati spaziali*. Edizioni Angelo Guerini ed Associati S.r.l., p. 170.

Skøien, J. O., R. Merz e G. Blöschl (2006). «Top-kriging - geostatistics on stream networks». In: *Hydrology and Earth System Sciences* 10.2, pp. 277–287. DOI: 10.5194/hess-10-277-2006. URL: http://www.hydrol-earth-syst-sci.net/10/277/2006/.

Skøien, J.O. et al. (2011). «rtop - an R package for interpolation of data with a variable spatial support - examples from river networks».

Ringraziamenti

Desidero innanzitutto ringraziare il Prof. Ing. Attilio Castellarin per la fiducia accordatami, per la passione che mi ha trasmesso nello studio e nello sviluppo di questo lavoro, e, infine, per avermi dato la possibilità di partecipare per la prima volta in vita mia ad un lavoro scientifico di ricerca internazionale.

Ringrazio poi Satcey A. Archfield, Julie E. Kiang e John O. Skøien per il supporto costante offerto in tutte le fasi che hanno caratterizzato l'evoluzione di questa ricerca.

Un ringraziamento particolare va alla mia famiglia, sopratutto a nonna Maria, mia instancabile motivatrice, e ai miei coinquilini Francesco, Nicola, Nico e Silvia, gli amici più vicini in questo percorso.

Infine, ma che in assenza di formalità sarebbe stato il primo, grazie a Ilaria, il mio vero pensiero felice.