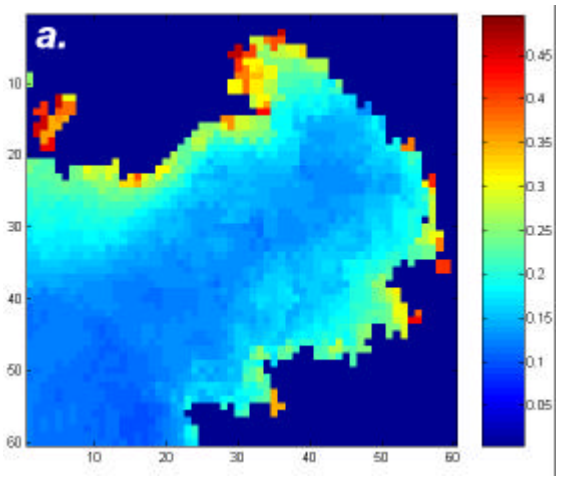
INTERPOLAZIONE DI DATI ATTRAVERSO METODO EOF:

APPLICAZIONE A IMMAGINI MODIS NEL GOLFO DI TRIESTE



ŽIVKO JUŽNIC-ZONTA, ELENA MAURI E PIERRE-MARIE POULAIN

Per comunicato approvato da: ……

Dr. Renzo Mosetti

Direttore, Dipartimento di Oceanografia

Indice

1. Introduzione

Il vantaggio del rilevamento satellitare nelle indagini oceanografiche rispetto ad altri metodi, consiste in una copertura di rilevamento estremamente grande nello spazio e nel tempo. I parametri rilevati sono molteplici: la temperatura superficiale del mare SST (?Sea Surface Temperature?), la clorofilla-a, il coefficiente di attenuazione diffusa a 490 nm, (*K490*) e molti altri. Lavorando nello spettro del visibile si presenta però il problema della mancanza dei dati a causa della presenza di nubi e particolato vario nell'atmosfera. Inoltre possono presentarsi mal funzionamenti del sensore che provocano rumore nei dati e la riduzione della copertura di rilevamento. Per tali circostanze si ottengono immagini con dati mancanti.

Nell'interpolazione di dati mancanti si sono sviluppati in passato diversi metodi: interpolazioni "spline" e metodi inversi, come l'interpolazione ottimale. Lo svantaggio di questi è che nella ricostruzione del set di dati, hanno bisogno a priori dell'errore statistico dei dati, generalmente poco noto.

Come metodo alternativo si presenta un metodo ?self-consistent?, libero da questo tipo di vincolo di informazioni a priori (Rixen & Beckers, 2003). Il metodo permette di calcolare il numero di EOF (?Empirical Orthogonal Functions?) ottimale, tramite il test di cross-validazione, con il quale riempire il set di dati e inoltre fornisce un parametro di stima dell'errore.

L'obiettivo di questa lavoro è di studiare in maniera sistematica l'interpolazione dei dati attraverso il metodo EOF e di applicare questa metodologia a dati di attenuazione diffusa di lunghezza d'onda 490 nm, *K490*, rilevati dal sensore satellitare MODIS, per il Golfo di Trieste.

Questo rapporto è parte di una tesi di laurea del triennio della facoltà di ingegneria per l'ambiente e il territorio. Nella prima parte di questo rapporto è trattata la teoria sulle funzioni empiriche ortogonali (capitolo 2), per la quale vengono presentati due metodi: della matrice di covarianza (?scatter matrix method?) e della decomposizione ai valori singolari SVD (?Singular Value Decomposition?). La parte dedicata alla descrizione del set di dati *K490*, riferiti al Golfo di Trieste, è trattata nel capitolo 3. Il capitolo 4, tratta il metodo d'interpolazione attraverso le EOF, sviluppato con il linguaggio di programmazione MatLab, e i rispettivi risultati.

2. Funzioni empiriche ortogonali

2.1. Teoria generale

Il problema di comprimere un set di dati molto ampio, contenente campionamenti spazio-temporali, è presente in ogni campo della scienza moderna. Un metodo di compressione potrebbe essere utilizzare una combinazione lineare di funzioni ortogonali di ?previsione?, oppure modi spaziali, e funzioni che tengono conto delle variazioni temporali dei dati. Tale combinazione restituisce il segnale osservato. Il set di dati potrebbe essere distribuito su una griglia (regolare o irregolare) di stazioni $x_i(t), y_i(t)$ su un piano orizzontale o su una sezione di profondità, con un piano verticale $x_i(t), z_i(t)$. Questo metodo, più generalmente chiamato col nome di *analisi alle componenti principali* o *PCA* (Principal Component Analysis), in oceanografia è meglio conosciuto con il nome di *analisi alle funzioni empiriche ortogonali* o *analisi con le EOF* (Empirical Orthogonal Functions).

La maggior parte della varianza temporale della serie di dati distribuiti su una griglia spaziale è rappresentata nelle poche prime funzioni ortogonali (o modi), il quale andamento potrebbe essere correlato alla dinamica del fenomeno fisico oggetto di studio. Le funzioni ortogonali sono semplicemente un metodo per frammentare la varianza di una serie spazio-temporale di dati nelle sue componenti principali, ortogonali fra loro. Questo tipo di analisi può essere svolta anche nel dominio delle frequenze, che però non sarà oggetto di questa rapporto tecnico. Qui di seguito viene enunciata una definizione più rigorosa del metodo EOF (Emery & Thomson, 1997).

Si consideri una serie di N immagini al tempo $t = t_j (1 \le j \le N)$, ognuna composta da M grandezze scalari $y_m(t)$ (pixel) registrati nelle corrispondenti località di coordinate $\vec{x_m} (1 \le m \le M)$. Quindi la serie di dati $y_m(t)$ ad ogni punto spaziale di campionamento $\vec{x_m} = (x_m, y_m)$ può venir scritta come somma di M funzioni ortogonali spaziali $f_i(\vec{x_m}) = f_{im}$, tali che

$$y(\overrightarrow{x_m},t) = y_m(t) = \sum_{i=1}^{M}[a_i(t) f_{im}],$$

dove $a_i(t)$ è l'ampiezza della *i*-esima funzione ortogonale spaziale (*i*-esimo modo) al tempo $t = t_n (1 \le n \le N)$. L'equazione indica che la variazione del dato $y_m(t)$ è data dalla combinazione lineare di M modi spaziali delle rispettive funzioni ortogonali, le cui ampiezze temporali sono date dai pesi $a_i(t)$ $(1 \le i \le M)$. Quindi avremo tante funzioni ortogonali quanti sono i siti di campionamento (*M*). Si noti che potremmo esprimere la serie spazio-temporale di dati tramite la combinazione lineare di. *N* funzioni temporali pesate per *N* ampiezze spaziali.

Poiché è necessario che la base di funzioni sia ortogonale, si deve soddisfare la condizione

$$\sum_{m=1}^{M}[f_{im} f_{jm}] = d_{ij}\ ,\ \text{dove } d_{ij} \text{ è la delta di Kronecker } d_{ij} = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases}.$$

Le EOF sono determinate in modo unico fra le tante scelte possibili che soddisfano la condizione di ortogonalità, se le ampiezze $a_i(t)$ sono prive di correlazione per i dati campionati. Ciò equivale a dire che la covarianza media nel tempo delle ampiezze soddisfa

$$\overline{a_i(t) a_j(t)} = l_i d_{ij}\ ,\ \text{dove } l_i = \overline{a_i(t)^2} = \frac{1}{N}\sum_{n=1}^{N}[a_i(t_n)^2],$$

con l_i la varianza nell'*i*-esimo modo ortogonale.

Se a questo punto si prende la matrice di covarianza $\overline{y_m(t) y_k(t)}$, troviamo che

$$\overline{y_m(t) y_k(t)} = \sum_{i=1}^{M}\sum_{j=1}^{M}[a_i(t) a_j(t) f_{im} f_{jk}] = \sum_{i=1}^{M}[l_i f_{im} f_{ik}].$$

Moltiplicando ambo i membri per f_{ik}, sommando per tutti i k e usando la condizione di ortogonalità, si ottiene

$$\sum_{k=1}^{M}\overline{y_m(t) y_k(t)} f_{ik} = l_i f_{im}\ ,$$

per l'*i*-esimo modo all'*m*-esima località (*i,m=1,...,M*). Quest'ultima equazione è semplicemente la forma canonica del *problema degli autovalori*. In questo caso, le EOF, f_{im}, sono gli autovettori *e i* $l_i = \overline{a_i(t)^2}$ sono gli autovalori della matrice di covarianza $R_{mk} = \overline{y_m(t) y_k(t)}$. Se si assume che la media $\overline{y_m(t)}$ è stata rimossa da ogni serie temporale di dati, possiamo riscrivere il problema come

$$Cf - lIf = 0\,,$$

dove *C* è la matrice di covarianza del set di dati e *I* è la matrice unitaria. Il problema agli autovalori consiste nella diagonalizzazione della matrice, che come conseguenza provoca la necessità di trovare gli assi ortogonali di dimensioni spaziali *M*, con elementi fuori diagonale nulli. Quando ciò accade, i diversi modi sono ortogonali fra di loro. Inoltre poiché la matrice *C* è simmetrica con valori reali, gli autovalori saranno anche reali. Risolvendo il *polinomio caratteristico*

$$\det(Cf - lIf) = 0\,,$$

otteniamo che $l_1 > l_2 > ... > l_M$: ciò implica che la varianza associata ad ogni modo statistico è ordinata in ordine decrescente in accordo con il suo corrispondente autovettore. Se sommiamo tutti i contributi di varianza degli. *M* modi otteniamo

$$\sum_{m=1}^{M} \frac{1}{N} \sum_{n=1}^{N} y_m(t_n)^2 = \sum_{j=1}^{M} l_j\,.$$

Ciò significa che la varianza del set di dati uguaglia la somma degli autovalori. Va inoltre ricordato che a questo punto si può determinare lampiezza temporale di ogni modo statistico, che viene espressa come

$$a_i(t) = \sum_{m=1}^{M} y_m(t) f_{im}\,.$$

La bontà del metodo EOF consiste nel fatto che è un'efficiente metodo di compressione dati: col termine efficiente si intende che nessun'altra espansione approssimata del set di dati, in termini di $K < M$ funzioni

$$\hat{y}_m(t) = \sum_{m=1}^{K} a_i(t)\hat{f}_{im}$$

può produrre un errore quadratico medio

$$\sum_{m=1}^{K} \overline{[y_m(t) - \hat{y}_m(t)]^2}$$

inferiore a quello che ottenuto quando $\hat{f}_i$ è un EOF (Emery & Thomson, 1997).

2.2. Applicazione delle EOF nell'analisi dei dati

I pregi del metodo EOF nell'analisi dei dati vengono qui di seguito brevemente riportati. Attraverso l'utilizzo di un numero limitato di modi, quelli con il contenuto di varianza maggiore, si può comprimere il set di dati ed eliminare le scale di variabilità minore, che con gran probabilità presentano una densità di rumore elevato. Le ampiezze temporali ottenute nei primi modi potrebbero riflettere proprietà fisiche. In quest'ottica di applicazione, il metodo EOF assomiglia a quella di Fourier usato per filtrare o analizzare il segnale, depurandolo dalle frequenze (scale di variazione) non desiderate. Una misura della bontà di compressione del segnale è contenuta nell'autovalore riferito al rispettivo modo. Come visto prima la varianza totale del segnale è data dalla sommatoria degli autovalori $\sum_{i=1}^{M} l_i$. Il rapporto

$$f_i = l_i \Big/ \sum_{i=1}^{M} l_i$$

è quindi una misura relativa del contenuto della varianza del *i*-esimo modo, rispetto a quella totale del segnale e viene utilizzato per scegliere con quale numero di modi approssimare il set di dati. Generalmente la compressione avviene conservando il 95% della varianza del sistema iniziale.

Un metodo alternativo per trovare le EOF, rispetto a quello che utilizza la matrice di covarianza (scatter matrix method), è il *metodo di decomposizione ai valori singoli SVD* (Singular Value Decomposition), che a confronto con il primo fornisce le EOF a un costo computazionale molto minore.

Il metodo SVD è basato sul concetto dell'algebra lineare che ogni matrice D di dimensione $M \times N$, dove $M \geq N$, può essere decomposta in un prodotto matriciale di tre matrici: una $M \times N$ ortogonale, U, una $M \times N$ matrice diagonale, S, con elementi positivi o zero e una matrice trasposta ortogonale V^T di dimensione $N \times N$. In notazione matriciale abbiamo

$$D = U \cdot S \cdot V^T$$

dove gli scalari $s_1, s_2, ..., s_N$ della matrice S sono chiamati *valori singolari di D*. I vettori colonna della matrice V^T sono i *vettori singolari destri di D* (EOF temporali), mentre quelli di U sono i *vettori singolari sinistri di D* (EOF spaziali). Nel caso che $M > N$ la matrice S è composta da una matrice $N \times N$ superiore diagonale e da una matrice $(M - N) \times N$ nulla inferiore. Le colonne della matrice U sono gli autovettori, mentre quelli della S sono riferiti agli autovalori.

Poi che

$$DD^T = USV^T VSU^T = US^2U^T$$

per trovere gli autovalori bisogna soddisfare la relazione

$$DD^T f = l f \Rightarrow U^T US^2 U^T f = l U^T f \Rightarrow S^2 \Phi = l \Phi \text{, dove } \Phi = U^T f$$

e quindi

$$\begin{cases} l_i = s_i^2, i = 1, ..., N \\ l_i = 0, i = N+1, ..., M \end{cases},$$

si ottiene una relazione che esprime gli autovalori della matrice di covarianza e i valori singolari della matrice S, ottenuta tramite decomposizione SVD.

Per ottenere le ampiezze temporali abbiamo bisogno di una matrice A, tale che

$$D = UA^T$$

$$A = VS^T$$

Le soluzioni sono identiche a quelle ottenute con la matrice di covarianza C del ?scatter matrix method?.

Il metodo SVD sopra presentato assume che la matrice D è perfettamente piena, priva di dati mancanti. Le immagini satellitari sono invece quasi sempre caratterizzate da dati mancanti, i quali possono essere stimati tramite interpolazione, preferibilmente tale che tenga conto della funzione di correlazione e del rapporto segnale/rumore nei dati. Le EOF possono venir adattate a tale scopo, poiché se è conosciuto il. dominio spaziale e le ampiezze temporali per un determinato campo di dati, la combinazione lineare di esse fornirà dei valori al campo, anche nei punti dove i dati sono mancanti. Partendo dalla considerazione che le scale grandi non dovrebbero essere influenzate da variazioni locali di alcuni punti, si calcola le EOF di una matrice nella quale è stata introdotta una prima approssimazione grezza dei dati mancanti. Tramite tali EOF spaziali e ampiezze temporali potremmo ottenere, per combinazione lineare dei primi k modi, una prima stima dei dati mancanti. Tale procedimento andrebbe iterato fino a convergenza. Il problema che sorge è del numero di modi da usarsi nella ricostruzione del segnale con punti incogniti: esso viene risolto tramite un test di cross-validazione dei dati interpolati con un determinato numero di modi.

Dalla teoria della perturbazione per la decomposizione ai valori singoli SVD, dallo studio del trasferimento di varianza tra diverse EOF e del cambiamento della struttura durante le iterazioni, presentata nelle appendici di Beckers e Rixen (2003), si dimostra che man mano i valori stimati vanno raffinandosi, la varianza nei primi modi principali aumenta, mentre diminuisce quella degli altri modi meno importanti.

Qui di seguito vengono presentati i passi fondamentali del metodo di interpolazione dei dati con le EOF (Beckers & Rixen, 2003), che fa uso della decomposizione SVD:

- Si calcola la media della matrice dei dati iniziali D, la si sottrae dalla matrice di dati D e si inizializza. i dati mancanti col valore zero.

- Si compiono due passi fino a ?convergenza?:
  - si calcola la decomposizione della matrice dei dati *D* per ottenere una prima stima delle EOF spaziali e ampiezze temporali;
  - i punti mancanti vengono rimpiazzati dai valori ottenuti con il primo modo di EOF:

$$D_{i,j} = y(\vec{x}_i, t_j) = \sum_{p=1}^{k} a_p(t_j) f_{pi} \text{ , dove } k = 1$$

avendo ottenuto così una stima migliorata nei punti mancanti, si ripetono i due punti precedenti con i nuovi valori ottenuti fino a ?convergenza?.

- Il passo successivo è rifare il ciclo di ?convergenza? usando sia il primo che il secondo modo. L'uso di una quantità crescente di modi consente di ottenere dati interpolati con 1,2,...,*k*. EOF. Da questo momento in poi si userà l'espressione *numero di modi k* per indicare che nell'interpolazione dei dati sono stati usati 1,2,...,*k*. modi.

- Il numero di modi ottimale da usare viene ottenuto con il test di cross-validazione: si inserisce casualmente alcuni valori zero al posto dei dati. Il numero dei dati sostituiti è un compromesso tra la robustezza statistica e la stabilità dell'informazione principale contenuta nei primi modi. Il numero di modi ottimale viene scelto tale da massimizzare la funzione di cross-correlazione che è funzione del numero di modi impiegati per l'interpolazione.

- Una volta che il numero di modi ottimale è noto, tutta la procedura viene rifatta usando anche i dati messi da parte per la cross-validazione, ottenendo i valori finali dei dati mancanti.

3. Dati del coefficiente di attenuazione diffusa a 490nm

3.1. Il sensore MODIS

Il sensore MODIS (Moderate Resolution Imaging Spectroradiometer) è installato sui satelliti Terra (EOS AM), lanciato il 18 Dicembre 1999 e Aqua (EOS PM), lanciato il 4 Maggio 2002. L'orbita del primo passa da nord a sud, passando l'equatore alle 10:30 (ora locale), mente Aqua ha il verso da sud a nord, passante per l'equatore alle 13:30 (ora locale). Il tipo di orbita per entrambi i satelliti è sincronizzata col periodo solare, circolare, passante vicino ai poli ed ha un'altezza di 705 km. La modalità di scansione è ortogonale alla direzione dell'orbita, di tipo ?cross track scanner?, con una linea di scansione di 2330 km. Il sensore consente di registrare la radianza in 36 bande spettrali, con un intervallo di lunghezza d'onda che va da 0.4 µm a 14.4 µm. Due bande LAC sono date alla risoluzione nominale di 250m, cinque bande LAC a 500m. e le rimanenti 29 che fanno parte del formato GAC a 1km. La copertura terrestre è raggiunta per ogni un giorno, fino a due giorni. Il sensore è progettato per un funzionamento di 6 anni.

3.2. Generalità sul set di dati

Il data set utilizzato consiste di 103 immagini settimanali dal 1 luglio 2002 al 19 giugno 2004. La singola immagine riferita all'Adriatico contiene 1052x1024 pixel e la una risoluzione di un chilometro, formato GAC (Global Area Coverage) del sensore MODIS (satellite Aqua). Ogni pixel rappresenta il valore del coefficiente di attenuazione diffusa *K490* ($k(l = 490nm, z)$, misurato della lunghezza d'onda di 490 nm e mediato lungo la profondità z sotto la superficie del mare) che è un parametro della torbidità dell'acqua. Esso viene quindi influenzato dal fitoplancton e dal particolato non vivente presente in acqua. L'irradiazione spettrale discendente nella colonna d'acqua $E_w^-(l, z)$ è governata dalla legge di Beer

$$E_w^-(l, z) = E_{incidente}^-(l, 0)e^{-k(l,z)z}$$

dove $E^{-}_{incidente}(l,0)$ è l'irradiazione incidente solare immediatamente sotto la superficie marina (Clark, 2004). L'algoritmo usato dal sensore MODIS per calcolare il coefficiente di attenuazione diffusa *K490* si serve delle radianze emesse dalla superficie marina $L^{+}_{w}(l)$ alle lunghezze d'onda di 488nm e 547nm

$$K490 = 0{,}016 + 0{,}15645\left[\frac{L^{+}_{w}(488)}{L^{+}_{w}(547)}\right]^{-1.5401}$$

dove 0,016 è l'attenuazione diffusa a 490nm in acqua pura e gli altri due coefficienti sono ottenuti da regressioni lineari tra radianze e dati di attenuazione a 490 nm in situ.

Tali dati sono messi a disposizione dalla NOAA (National Oceanic and Atmospheric Administration) sulla pagina web http://seawifs.gsfc.nasa.gov/cgi/browse.pl?typ=GAC . Esse sono *immagini giornaliere* registrate attorno alle ore 12:00 (ora locale di Trieste) in formato GAC corrette per l'effetto atmosferico con l'algoritmo atmosferico (livello 2 di elaborazione) . I dati trasferiti sono stati estratti per l'area d'interesse applicando la proiezione di mercatore. con il programma WIM Automatic Module (WAM) nel formato *hdf*.

A causa dell'elevata percentuale di copertura nuvolosa le immagini giornaliere vengono mediate (sette giornate per una immagine settimanale) in *composite settimanali*, che presentano una nuvolosità notevolmente ridotta (figura 1). Tale procedimento di media, condotto con il programma WIM, fa perdere l'informazione a scale piccole, ma conserva quella a scale più grandi. Accanto alle immagini settimanali sono prodotte anche delle *immagini di conteggio* che indicano il numero di pixel usati nella costruzione della composita (figura 1).

3.3. Matrice dei dati iniziali

I dati sono stati ordinati in un cubo matriciale (x,y,z) in cui le colonne x rappresentano la latitudine, le righe y la longitudine e la profondità z il tempo. Per il Golfo di Trieste tale cubo presenta una dimensione 60x60x103. Esso viene successivamente trasformato in una matrice 3600x103, dove l'immagine è rappresentata dal vettore colonna. Per assicurarsi di eliminare i valori presenti sulla terra si è utilizzata una *immagine maschera* 3600x1 che presenta valori NaN sulla terra e valore uno sul mare. Alcuni valori NaN prossimi alla costa sono in realtà punti

del mare o vice versa a causa della limitata risoluzione dell'immagine presente in tale zona limitrofa.

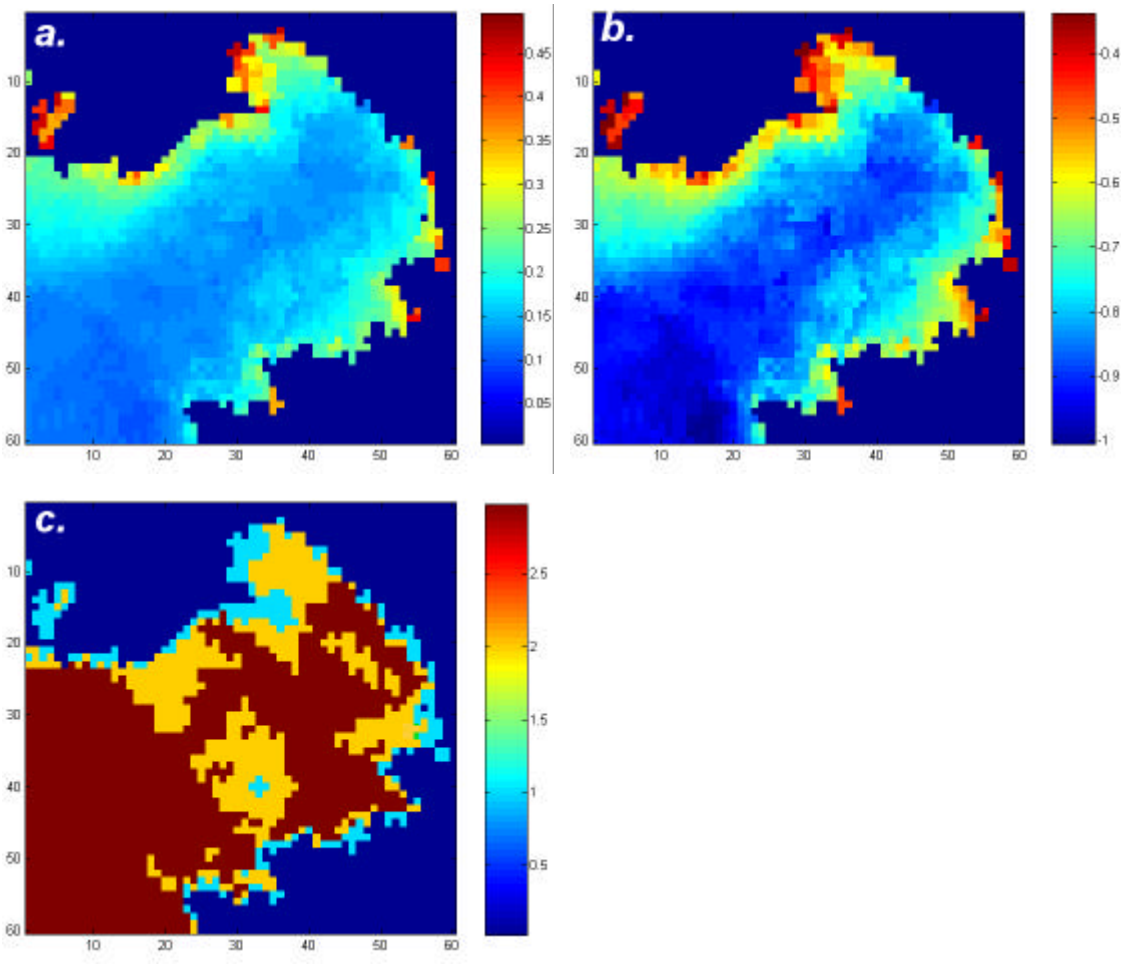


Figura 1. ***(a)*** *Immagine del K490 della prima settimana di settembre 2002 con scala lineare.* ***(b)*** *Composita settimanale che rappresenta la prima settimana di set-tembre 2002 dove i valori del K490 sono stati rappresentati in scala logori tmicaper mettere in evidenza il gradiente dei valori piccoli di K490.* ***(c)*** *Immagine di conteggio della prima settimana di settembre 2002, dove in leggenda è segnato il numero di pixel usati nella composizione dell'immagine mediata. Si fa notare che hanno senso solo i numeri interi.*

Sulla *matrice di dati iniziali* $y(\overrightarrow{x_m}, t_n)$ di dimensione 2267x103, di soli dati *K490* senza valori della terra, ottenuta tramite l'immagine maschera, vengono applicate alcune semplici indagini statistiche per capire alcune caratteristiche principali della distribuzione spazio-temporale dei dati. La matrice così strutturata sarà utilizzata per tutti i calcoli futuri.

3.4. Matrice dei dati *K490*

Gli istogrammi delle settimane prime quattro settimane del data set sono riportati in figura 2 e presentano una distribuzione assimetrica, andamento evidente anche nella maggior parte delle altre settimane. Va notato inoltre che sono pochi i valori che superano 0,25 m^{-1}.

La copertura nuvolosa del set di dati iniziale è del 26%. Si ricorda che i dati mancanti, cioè corrispondenti a pixel nuvolosi sono stati posti uguale a zero.

In figura 3a è rappresentata la media dei dati di ogni immagine iniziale. Si osservano dei picchi di valori estremamente bassi, alcuni zero: questi appartengono alle settimane con nuvolosità molto elevata. In figura 3c si riporta la nuvolosità percentuale per ogni immagine. Essa mette in evidenzia l'apparente elevata correlazione tra media bassa e nuvolosità alta. Si osserva inoltre che le settimane del periodo estivo sono meno coperte che quelle del periodo invernale. La figura 3b riporta la deviazione standard di tale media per darci un idea della variabilita' all'interno di ciascuna composita e capire la significatività della media riportata in figura 3a.

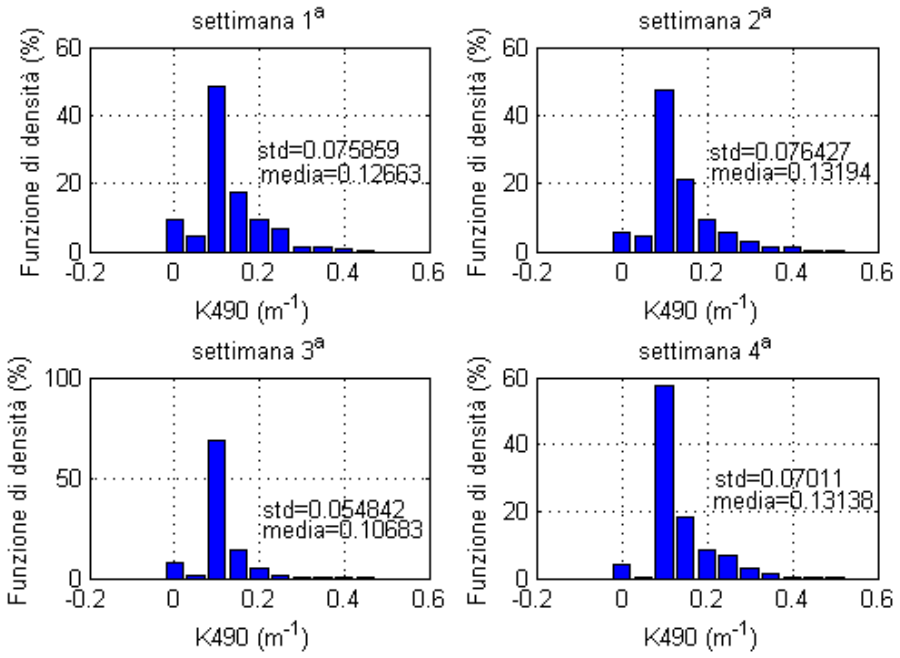


Figura 2. *Istogrammi delle prime quattro settimane che presentano una distribuzione assimetrica, questa e' una caratteristica anche nella maggior parte delle altre settimane. S ipuo' notare una bassa percentuale di valori superiori di* $0{,}25\,m^{-1}$.

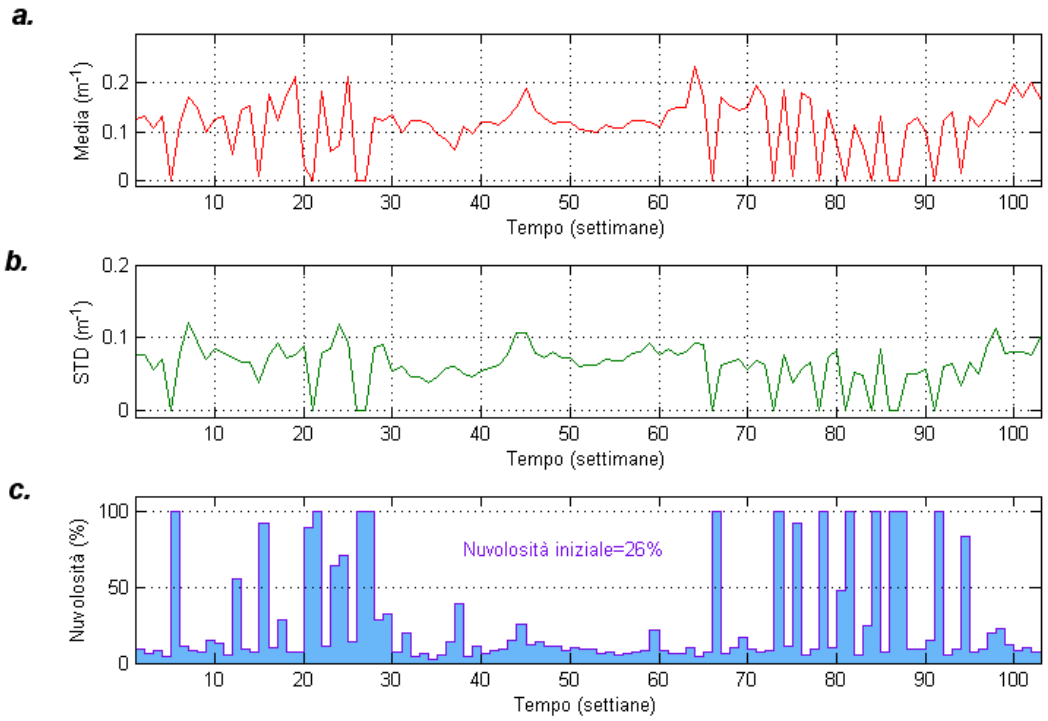


Figura 3. ***(a)*** *Media,* ***(b)*** *deviazione standard (STD) e* ***(c)*** *percentuale di nuvolosità del K490, per ciascuna settimana della matrice di dati K490. I valori di media zero corrispondono alle settimane con copertura del 100%.*

La qualità dei dati *K490* rilevati dal sensore MODIS dipendono dal valore che assumono: i *K490* compresi nell'intervallo $0{,}20\,m^{-1} \le K490 < 0{,}25\,m^{-1}$ hanno un'incertezza del 20% (Clark, 2004). Sopra tale intervallo l'incertezza aumenta. Si è deciso quindi di prendere in considerazione solo dati con valori minori di $0.25\,m^{-1}$. Tutti i valori superiori a questa soglia vengono posti uguale a zero. Tale scelta provoca un aumento apparente della nuvolosità della matrice del 7%.

I dati a questo punto sono stati ulteriormente ridotti togliendo le immagini con copertura nuvolosa maggiore al 95%. Si è ritenuto infatti, che tali immagini non contengono informazioni sufficienti ed inoltre non sono affidabili a causa del non perfetto mascheramento delle nuvole. La matrice così ottenuta viene chiamata di seguito *matrice dei dati ridotti* o *matrice ridotta.* Va

ricordato inoltre che tale scelta impedisce in futuro la stima attraverso il metodo di interpolazione con le EOF di tali settimane. La percentuale di dati da interpolare a questo punto è di 23%, avendo ridotto di tredici settimane il data set iniziale.

Per facilitare la lettura dei grafici che riportano il tempo in settimane progressive a partire dalla prima settimana di luglio 2002, la figura 4 trasforma tali settimane in periodi mensili.

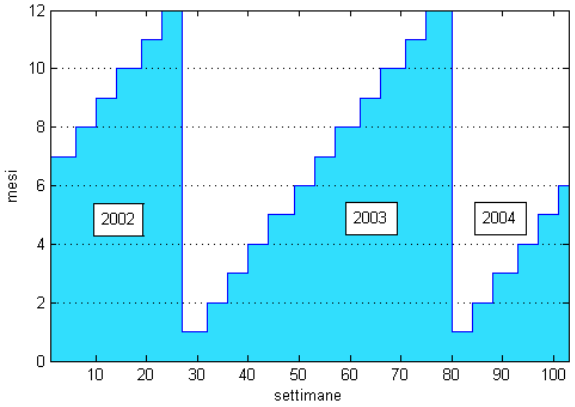


Figura 4. *Trasformazione da settimane in mesi dell'anno per il periodo del set di dati.*

In figura 5 è riportato l'istogramma dei dati con intervalli numerici di *K490* molto piccoli per valutare la risoluzione del data set. Il valore di risoluzione digitale riscontrato è di 0,0002, quindi siamo sicuri che il millesimo ha ancora significato fisico, mentre si può trascurare il decimillesimo. Tale analisi è stata condotta per valutare successivamente il grado di risoluzione numerica da ottenere nei valori stimati dal metodo di interpolazione, sempre se ciò sarà possibile.

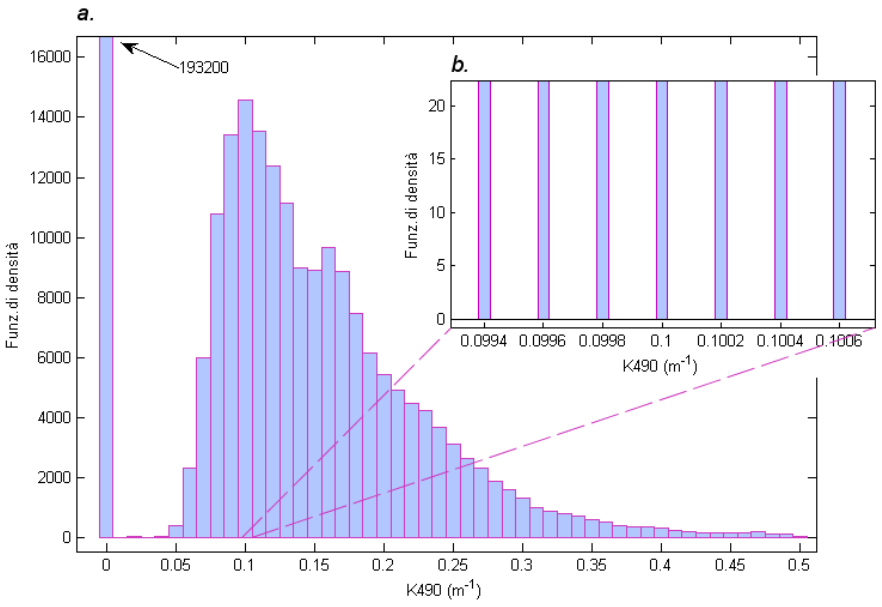


Figura 5. (a) *Istogramma della matrice di dati ridotti (escludendo valori maggiori e uguali a 0,25* m^{-1} *e le settimane con copertura nuvolosa maggiore del 95%).* ***(b)*** *Lo zoom evidenzia la risoluzione digitale di 0,0002.*

Per completare il quadro statistico che concerne la matrice dei dati ridotti, utilizzata nel metodo EOF, si osserva dalla figura 6 che la densità di dati mancanti per ogni pixel nell'arco di tempo considerato è predominante lungo la costa del golfo. Tale condizione suggerisce che i dati interpolati in questa zona saranno inaffidabili a causa della scarsa disponibilità di informazione sulla varianza a disposizione nel metodo.

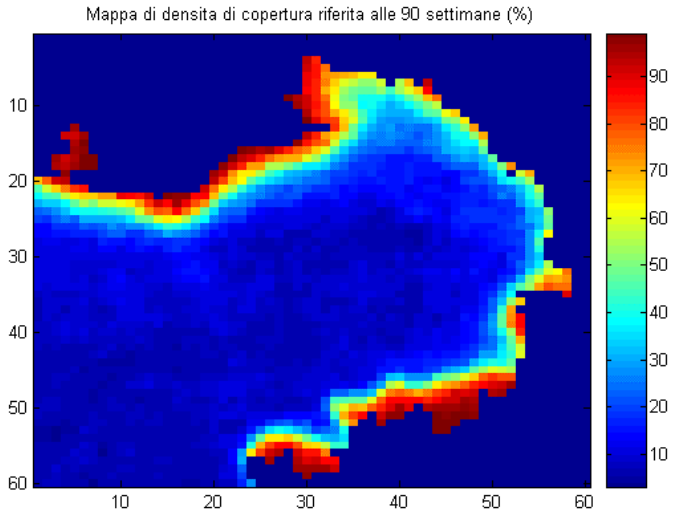
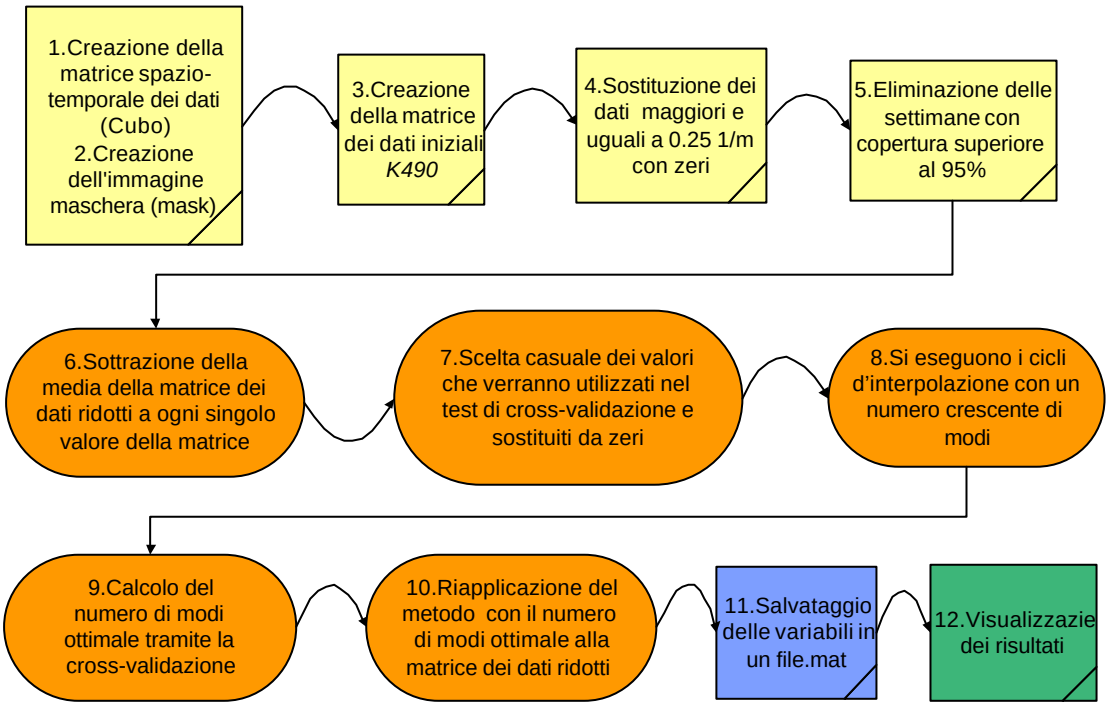


Figura 6. *Mappa della densità di dati mancanti in per-centuale per ogni pixel, riferita alla matrice dei dati ridotti.*

4. Metodo d'interpolazione attraverso EOF applicato al Golfo di Trieste

4.1. Diagramma di flusso del metodo d'interpolazione

L'interpolazione di dati mancanti attraverso metodo EOF è stato descritta per primo da Beckers & Rixen (2003), (vedi paragrafo 2.2). Il metodo sviluppato di seguito è stato addattato alle esigenze di questo lavoro e segue il seguente schema logico:



Il numero di modi ottimale si individua, ripetendo l'interpolazione con un numero di modi crescente, tramite il test di cross-validazione. Noto ciò si può interpolare definitivamente la matrice dei dati ridotti. Infine le variabili vengono salvate in un file *.mat* e graficate.

4.2. Risultati del metodo d'interpolazione con le EOF

4.2.1. Definizione della soglia di troncamento del processo iterativo

L'*errore del ciclo* è rappresentato in figura 7. Esso è definito come il valore assoluto del massimo, della differenza al passo *j*-esimo e di quella al passo *(j-1)*-esimo (con j=1,...,numero di cicli) della matrice dei dati ridotti ed interpolati usando un dato numero di modi k, D_k^j

$$abs(\max(D_k^j - D_k^{j-1}))$$

Si osserva che nei primissimi cicli l'errore tende a diminuire molto velocemente per tutti i numeri di modi d'interpolazione, mentre per cicli avanzati esso decresce molto lentamente. I cicli fatti in certi numeri di modi presentano inoltre un comportamento anomalo, poiché la loro funzione d'errore presenta delle derivate localmente positive.

In seguito si giustificherà perché si ferma il processo iterativo con un errore di troncamento del ciclo dello 0,01, con un evidente risparmio di tempo macchina non influendo sulla precisione dei dati interpolati. Anche se la risoluzione numerica del data set sia 0,0002 m^{-1}, si dimostrerà che le variazioni sotto circa un centesimo rappresentano solo rumore.

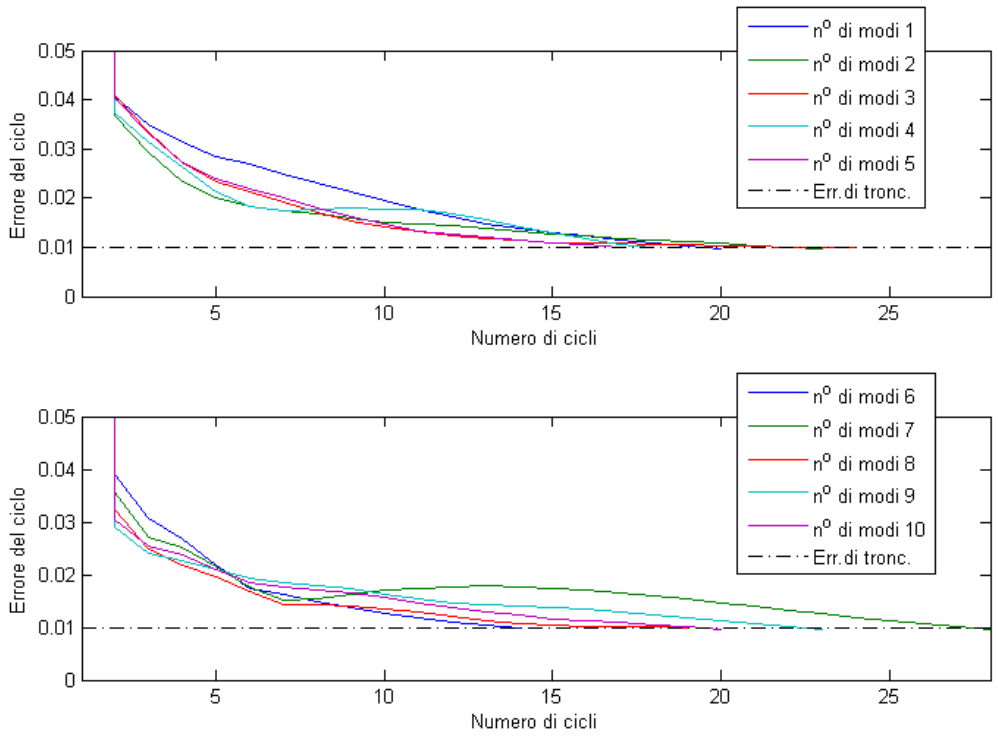


Figura 7. *?Convergenza? del processo iterativo dell'interpolazione dei dati mancanti con un numero di modi da uno a dieci. L'errore di troncamento è di un centesimo.*

4.2.2. Test di cross-validazione

La determinazione del numero di modi ottimale si fa utilizzando il *coefficiente di cross-correlazione*, definito come

$$r_j = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - m_x)(y_{i,j} - m_y^j)}{s_x s_y^j}$$

dove x_i è l'*i*-esimo dato prelevato casualmente dalla matrice dei dati ridotti con i=1,...,n, dove n è il numero di tali dati, $y_{i,j}$ è l'i-esimo dato interpolato tramite il *j*-esimo numero di modi con j=1,..., numero di modi (nel nostro caso $k=10$); m_x è la media dei dati x_i, m_y^j è la media dei dati $y_{i,j}$ ricavato con il rispettivo numero di modi e s_x, s_y^j sono le deviazione standard rispettive.

In figura 8 si nota che usando otto modi il coefficiente di cross-correlazione è massimo. Si può giustificare la decisione di fermarsi ai primi dieci modi, perché l'andamento del coefficiente di cross-correlazione tende a decrescere dopo il decimo modo. Ciò è valido per cinque diversi casi di estrazione casuale di dati per la cross-validazione (vedi figura 9). Si può notare inoltre che i numeri ottimali di modi nei cinque casi non sono uguali, ma comunque molto simili e tutti superiori allo 0,92. Questo valore quindi, potrebbe venir usato come criterio di minima correlazione, accettando un numero di modi tali da avere un coefficiente maggiore o uguale a tale valore limite.

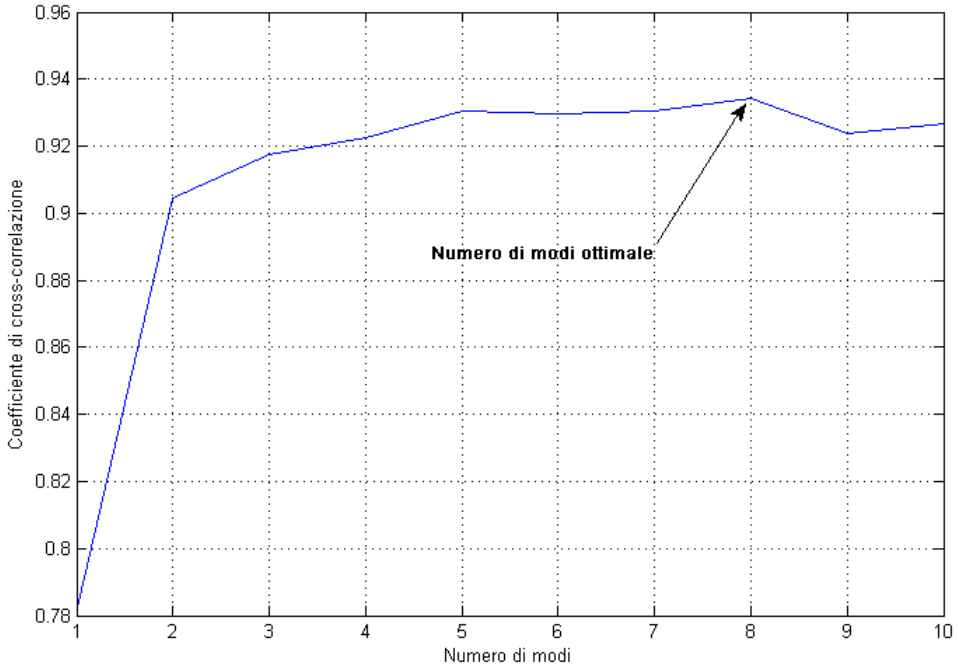


Figura 8. *Coefficiente di cross-correlazione in funzione della quantità dei modi utilizzati per l'interpolazione. Il numero di modio ottimale è otto e corrisponde a* $r_8 = 0{,}93434$.

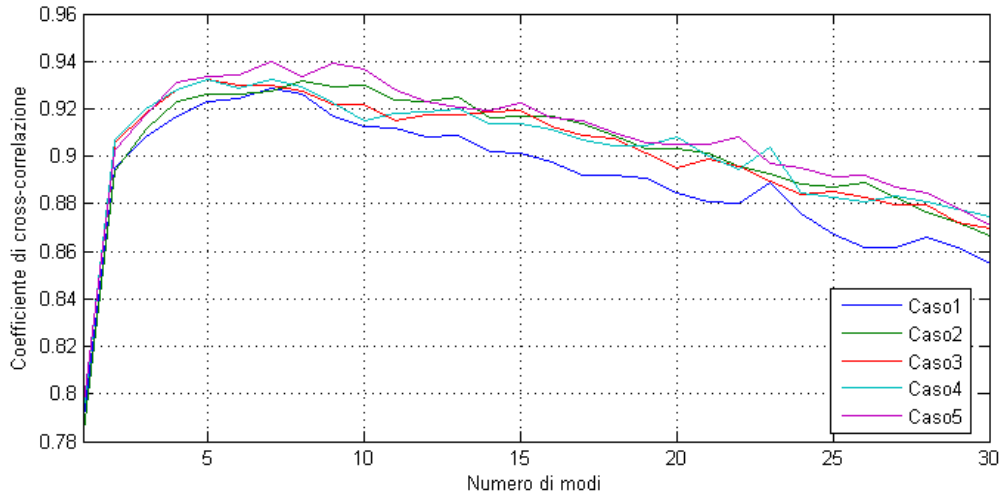


Figura 9. *Coefficiente di cross-correlazione in funzione della quantità di modi utilizzati in cinque casi, in cui si sono estratti casualmente dati per la cross-validazione.*

In figura 10 sono rappresentate le distribuzioni della varianza relativa percentuale in funzione dei modi EOF (i primi dieci), riferita alla matrice ridotta (colore blu) e a quella con dati ridotti interpolati ottimamente usando otto modi (colore rosso). Si nota che la varianza relativa della matrice interpolata è maggiore per il primo modo ed è minore per gli altri rispetto alla varianza della matrice dei dati ridotti.

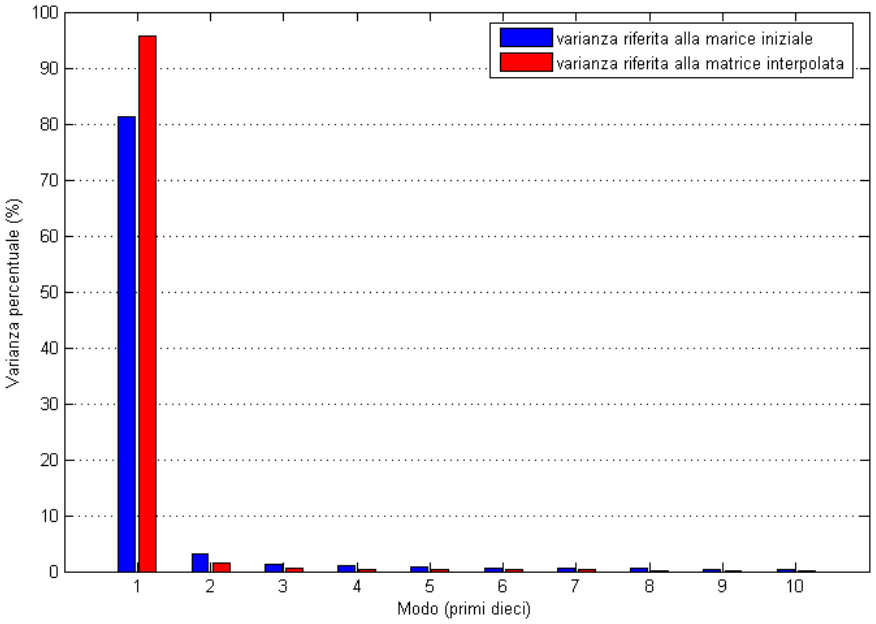


Figura 10: *Distribuzione della varianza relativa percentuale in funzione dei primi dieci modi, relativa alla matrice ridotta (colore blu) e alla matrice ridotta interpolata ottimamente (colore rosso).*

Il coefficiente di cross-correlazione garantisce solo che le due variabili in esame covariano, ma possono differire di un termine costante e un fattore moltiplicativo. E' quindi necessario esaminare la distribuzione delle differenza tra punto interpolato e il valore reale. Quest'ultimo viene riportato in figura 11 che rappresenta uno ?scatterplot? di dati interpolati (nel caso peggiore, medio e ottimale) con i dati di cross-validazione. La linea tratteggiata in nero rappresenta il massimo grado di correlazione, mentre quelle parallele verdi indicano la deviazione standard della differenza SD. La figura 11c mostra che i valori interpolati e i valori reali sono uguali più o meno una differenza standard SD di circa 0,017 m^{-1}. Si può osservare in tale figura che alcuni dati interpolati possono eccedere il valore 0,25 m^{-1}, soglia usata per tagliare i dati incerti. Questi compongono in totale il 2,2% della matrice interpolata e sono probabilmente presenti in maggior numero nelle settimane che corrispondono alla fioritura del mare, cioè in primavera e in autunno (figura 12). Essi sono probabilmente prodotti poiché il metodo ?sente? la tendenza verso tali valori alti.

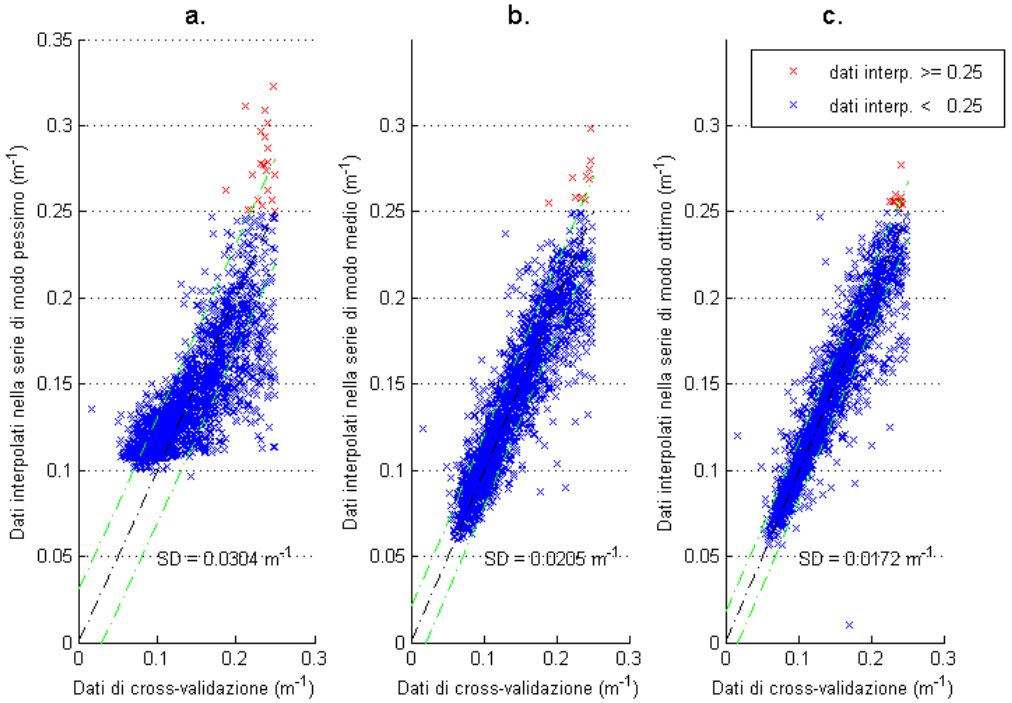


Figura 11. *?Scatterplot? che rappresentano la correlazione tra i dati presi casualmente per il test di cross-validazione e quelli ricavati dall'interpolazione usando un numero di modi che ha un coefficiente di cross-correlazione* ***(a)*** *peggiore (nº di modi 1),* ***(b)*** *medio (nº di modi 3) e* ***(c)*** *massimo (nº di modi 8). La retta tratteggiata y=x indica il massimo grado di cross-correlazione, mentre la verde y=x ± SD .*

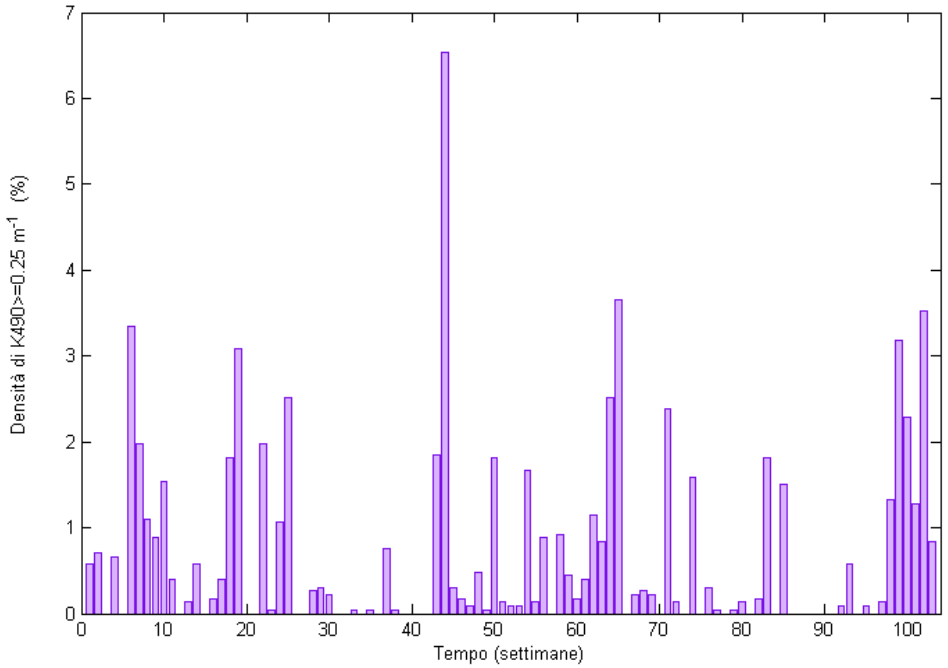


Figura 12. *Densità percentuale di dati interpolati con il numero di modi ottimale che eccedono il valore 0,25 m^{-1} nelle 103 settimane.*

La figura 13 evidenzia che il coefficiente di cross-correlazione potrebbe essere il parametro giusto per decidere quale numero di modi va usato nell'interpolazione finale. Si osserva in maniera ancora più netta di quanto negli ?scatterplot? precedenti che, la tendenza del metodo è di sovrastimare i dati interpolati (evidente nell'interpolazione tramite il primo modo), anche se tale sovrastima non esiste quasi più nel caso si usino 8 modi. L'errore relativo in percentuale riferito al numero di modi *k* usato nell'interpolazione è definito come

$$\frac{x_i^{cross-valid} - y_{i,k}^{interp}}{x_i^{cross-valid}} 100$$

dove $x_i^{cross-valid}$ è l'*i*-esimo dato usato per la cross-validazione, con i=1,...numero di dati usati per la cross-validazione e $y_{i,k}^{interp}$ è l'*i*-esimo dato interpolato con il numero di modi uguale a *k*. Come si osserva dalla figura 14, che riporta l'errore relativo riferito al numero di modi peggiore, medio e ottimale, l'interpolazione tramite il numero di modi ottimale, cioè *k=8*, dà un errore relativo dell'-1% ed una deviazione standard di 18%.

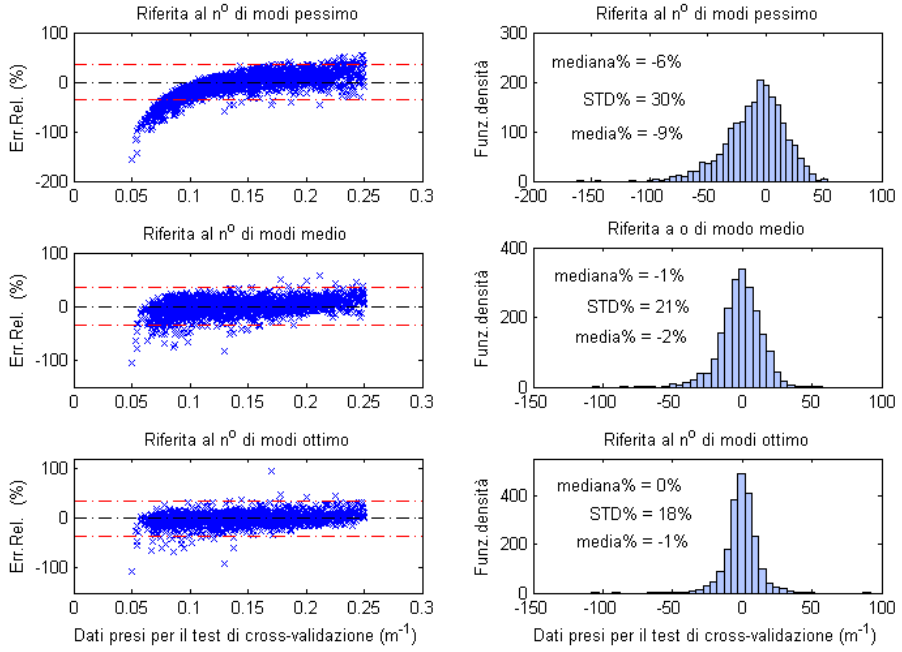


Figura 13. *Grafici a sinistra - indicano l'errore relativo percentuale riferito ai dati usati nel test di cross-validazione, confrontati con quelli interpolati. Grafici a destra - istogrammi degli stessi errori relativi. Dall'alto in basso vengono riportati i grafici per l'interpolazione tramite il numero di modi peggiore (n° di modi 1), medio (n° di modi 3) e ottimale (n° di modi 8). Le linee tratteggiate in rosso indicano una errore relativo del ±35%.*

Come premesso è necessario giustificare il fatto di aver scelto un errore di troncamento dei cicli iterativi di 0,01. Si dimostra tramite una verifica, che consiste nel mantenere costanti i dati presi per la cross-validazione e variando l'errore di troncamento, che per errori maggiori di 0,01 i coefficienti di cross-correlazione non migliorano per i rispettivi numeri di modi, mentre al di sotto di tale errore il peggioramento è notevole, come si osserva nella figura 14. Il valore del centesimo è per cui scelto come compromesso tra una maggiore precisione di interpolazione e una riduzione del tempo di calcolo. In totale il tempo impiegato per completare l'interpolazione ammonta a 3,42 minuti, usando una macchina Windows XP, 256 RAM e AMD 2400+. Nelle figure 15 e 16 sono riportati i tempi impiegati nei cicli di iterazione con la variazione dell'errore di troncamento: si osserva che il tempo aumenta esponenzialmente al diminuire dell'errore di troncamento.

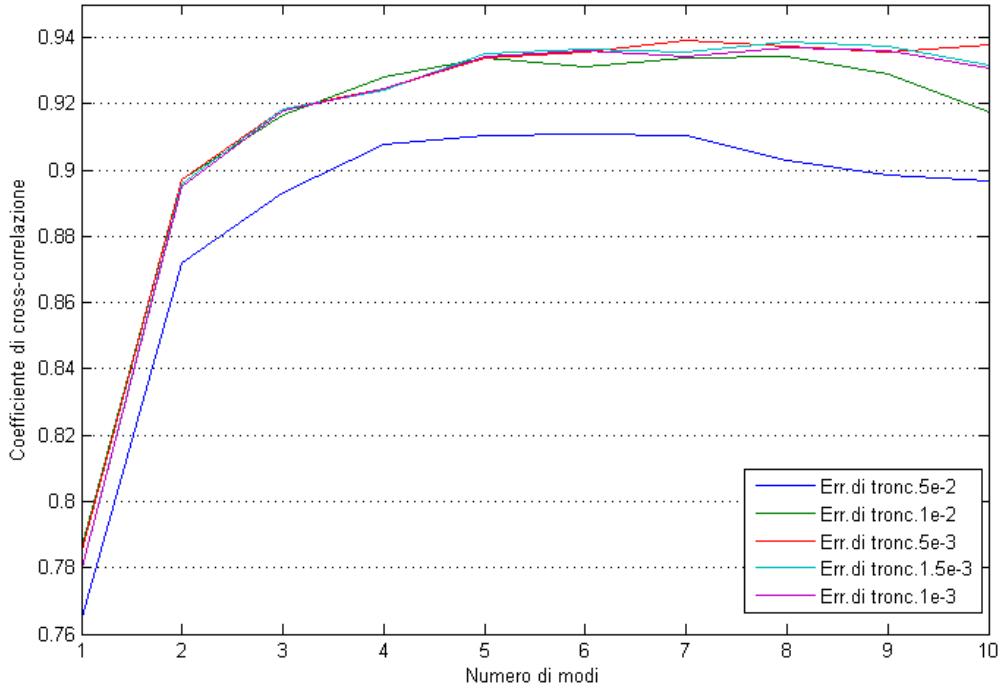


Figura 14. *Coefficiente di cross-correlazione al variare dell'errore di troncamento dei cicli iterativi e del numero di modi.*

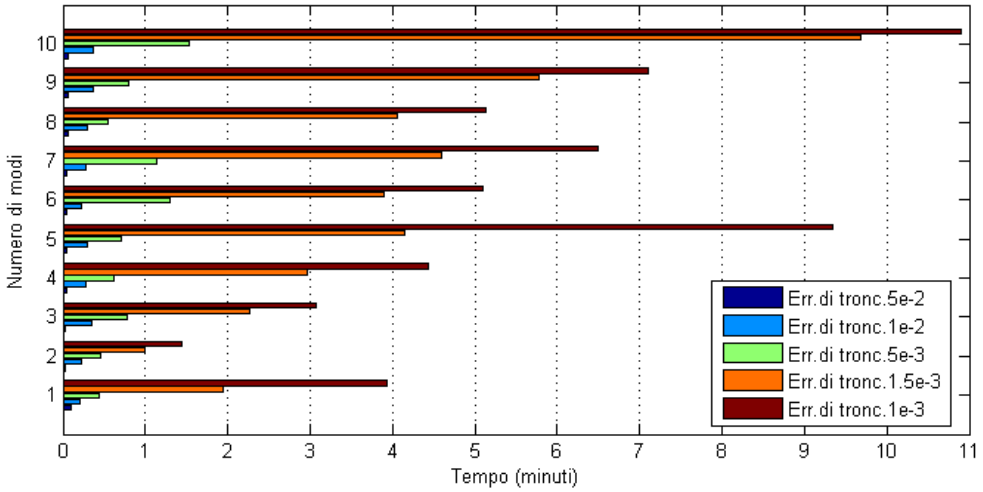


Figura 15. *Tempi macchina in minuti (AMD 2400+, RAM 256, Windows XP) per completare la ?convergenza? in funzione del numero di modi e dell'errore di troncamento.*

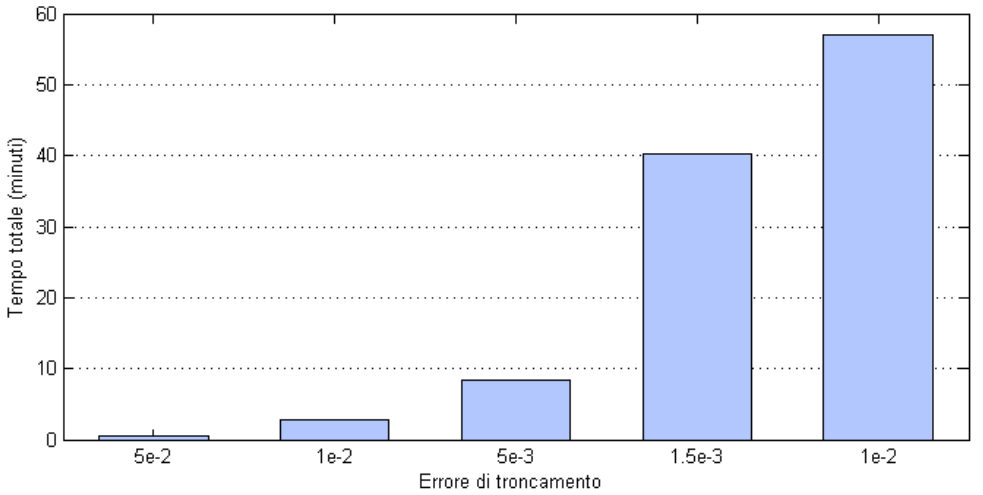


Figura 16. *Tempo totale, ossia la somma dei tempi impiegati per completare i cicli di ?convergenze? al variare dei rispettivi errori di troncamento.*

Il seguente test, i cui risultati sono riassunti in figura 17, consiste nell'eseguire i cicli d'interpolazione, del punto 8 del diagramma di flusso, per trovare i coefficienti di cross-correlazione ottimali, per quindici volte (per avere una base di punti statisticamente valida) e quindi con quindici set di dati per la cross-validazione diversi. Viene inoltre variata la quantità di tali dati di cross-validazione (1% ,3%, 5%, 10% e 50% della matrice dei dati ridotti). Le rette di regressione hanno pendenza casualmente positiva o negativa ed evidenziano il fatto che più si

aumenta la percentuale dei dati di cross-validazione e più i coefficienti di cross-correlazione tendono ad addensarsi lungo le rette di regressione. Tale fatto implica, che il test di cross-validazione diventa sempre più attendibile, ma al contrario si vede ridurre la media dei coefficienti di cross-correlazione ottimali. Ciò é dovuta al fatto che una parte dell'informazione del segnale viene tolta per formare il data set di dati di cross-validazione.

Dalla figura 18 si osserva che il numero di modi ottimale non è costante al variare del numero di casualità, ma comunque i coefficienti di cross-correlazione relativi sono molto simili.

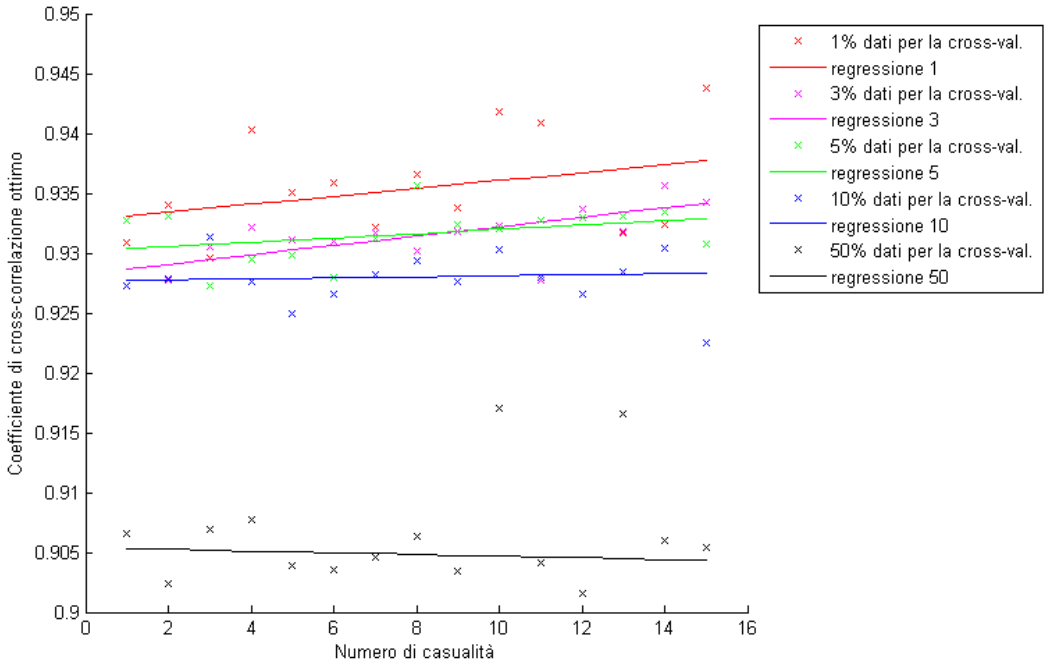


Figura 17. *Coefficiente di cross-correlazione ottimale in funzione della quantità di dati usati nel test di cross-validazione per 15 casi in cui i punti usati per la cross-validazione sono stati estratti casualmente (1% a 50% di punti estratti).*

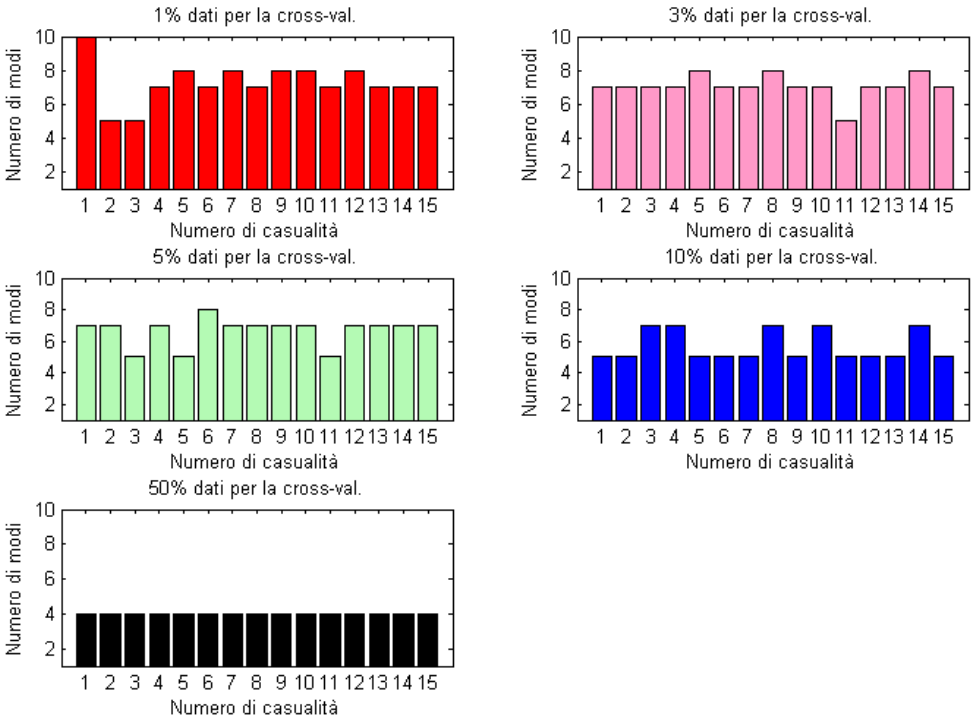


Figura 18. *Numero di modi ottimale in funzione del numero di casualità.*

	Densità di dati usati per la cross-validazione				
	1%	3%	5%	10%	50%
Media	0,93545	0,93124	0,93162	0,92778	0,90641
Mediana	0,93406	0,93169	0,93235	0,92784	0,90536
STD	0,004348	0,00238	0,002235	0,002212	0,004575

Tabella 1. *La media, la mediana e la deviazione standard STD dei coefficienti di cross-correlazione ottimali, riferiti ai cinque casi di densità percentuale di dati usati nella cross-validazione. L'elevato valore della deviazione standard nel caso 50% è dovuto alla presenza di due presunte anomalie.*

In tabella 1 sono riportate la media, la mediana e la deviazione standard dei coefficienti di cross-correlazione ottimali della figura 17. Da tale tabella si può concludere che il numero ottimale di dati per la cross-validazione sono il 3%, 5% e 10% della matrice dei dati ridotti. Il caso del 50% viene scartato poiché presenta una media (oppure una mediana) troppo bassa, così come il caso dell'1% che non è significativo a causa dall'elevata deviazione standard. L'alta deviazione standard del caso 50% è dovuta a due valori del coefficiente di cross-correlazione

molto diversi dal comportamento generale e dell'ancor troppo basso numero di casualità. Probabilmente con un numero di casualità più elevato la deviazione standard sarebbe con buona probabilità più piccola. Si noti inoltre che togliendo il 50% dei dati alla matrice ridotta si interpola una matrice con densità di dati mancanti del 73%. Nonostante l'elevata nuvolosità apparente la qualità dell'interpolazione è ancora buona, con un coefficiente di cross-correlazione circa del 0,90.

Si dimostra che l'ordine delle immagini nella matrice dei dati, non influisce sui risultati dell'interpolazione del metodo EOF. Infatti i modi spaziali e le ampiezze temporali ottenute tramite una matrice con le immagini ordinate temporalmente o mescolate sono identiche. Ciò significa che il metodo interpola in base all'informazione spaziale e che è sensibile alla varianza temporale in maniera indiretta. In pratica si potrebbe riempire un'immagine che contiene un solo pixel, ma un errore sul valore del pixel potrebbe generare un'immagine completamente fuori dall'andamento. Va inoltre notato, che le immagini con elevata nuvolosità hanno una affidabilità minore del dato a causa della condizioni atmosferiche e quindi da analizzarsi con più attenzione.

Una limitazione del test di cross-validazione potrebbe essere che se i dati presi per la cross-validazione sono nelle immediate vicinanze a dati buoni, la stima sarà buona. Se invece sono presi vicino a dati mancanti è difficile dire quale sia l'affidabilità del dato interpolato. Da ciò si ipotizza che le immagini, a parità di nuvolosità, saranno interpolate peggio, se la nuvolosità si presenterà concentrata, formando vaste aree di buco, e meglio quanto più la nuvolosità presenta carattere disperso in maniera casuale. Non sarà oggetto di questo lavoro l'ulteriore indagine in questo senso.

A questo punto si è variata la nuvolosità limite che può essere presente nel set di dati. Per ogni percentuale di nuvolosità si è ripetuto sei volte l'interpolazione, di modo da ottenere un numero di casualità statisticamente significativo. I risultati ottenuti sono riportati in figura 19 e vengono espressi dal parametro di media, mediana e deviazione standard in percentuale riferito all'errore relativo dei dati di cross-correlazione. Il procedimento si è ripetuto per il caso d'interpolazione col numero di modi peggiore (colore blu), medio (colore rosso) e ottimale (colore rosa). Nella figura 20 viene fornita l'informazione di quante immagini settimanali è composta la matrice dei dati, variando il limite di nuvolosità. Si noti che la scala delle ascisse nella figura 19 non è lineare.

Le rette di regressione indicano che nel numero di modi medio e ottimale, la media, la mediana e la deviazione standard hanno un andamento quasi costante. Sembrerebbe quindi valida l'ipotisi sopra citata della non totale adeguatezza del test di cross-validazione nella stima della bontà dell'interpolazione. Se esso descrivesse la bontà dell'interpolazione su ?scala globale?, gli errori relativi percentuali dovrebbero diminuire in termini di valore assoluto, poiché si interpolano immagini sempre più piene.

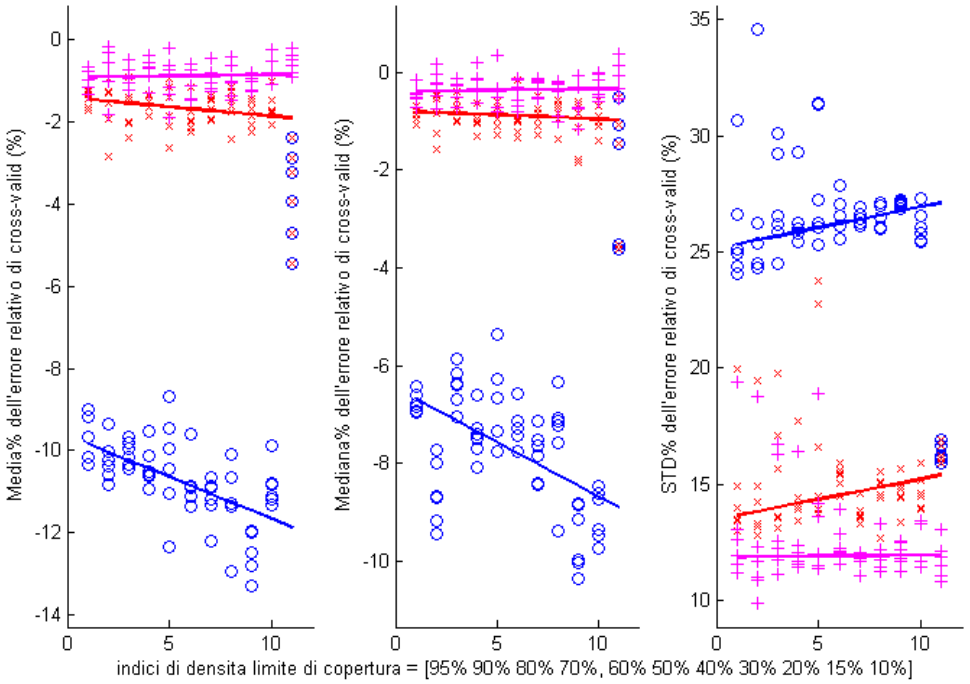


Figura 19. *Da destra verso sinistra - media, mediana e deviazione standard (in percentuale) riferiti all'errore relativo percentuale dei dati di cross-correlazione in funzione degli indici del vettore nuvolosità (in percentuale). Con diversi colori sono segnati le interpolazioni che fanno uso del numero di modi peggiore (colore blu), medio (colore rosso) e ottimale (colore rosa).*

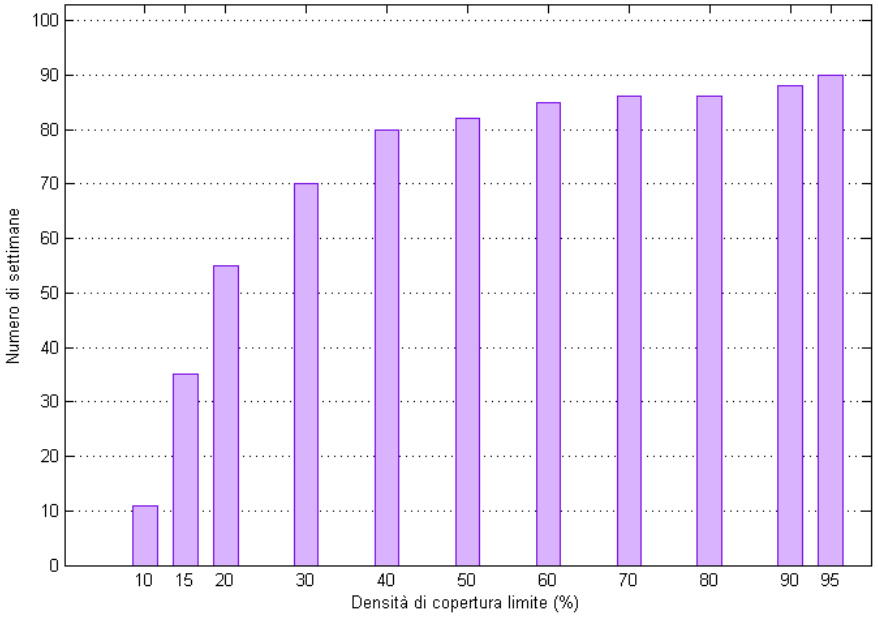


Figura 20. *Numero di immagini settimanali presenti nella matrice di dati ridotti dopo l'eliminazione delle immagini con nuvolosità in percentuale maggiore a quella limite.*

4.2.2. Risultati di interpolazione dei dati mancanti sulle immagini

Avendo provato la validità in termini generali del metodo d'interpolazione con le EOF, vengono presentati qui di seguito i primi otto modi (EOF spaziali) e le rispettive ampiezze temporali. Il metodo è stato applicato sulla matrice dei dati ridotti e i risultati sono relativi all'interpolazione ottimale in figure 21. Si osserva che passando dai primi EOF spaziali a quelli ottenute con più modi EOF il grado di smussamento aumenta e certe strutture vengono traslate. Inoltre si osserva che le immagini delle EOF spaziali e i diagrammi delle ampiezze temporali sono sempre più affette da rumore mano a mano che si aumenta il numero di modi. Tale effetto si può spiegare con il fatto che più fenomeni ad alta frequenza sono contenuti.

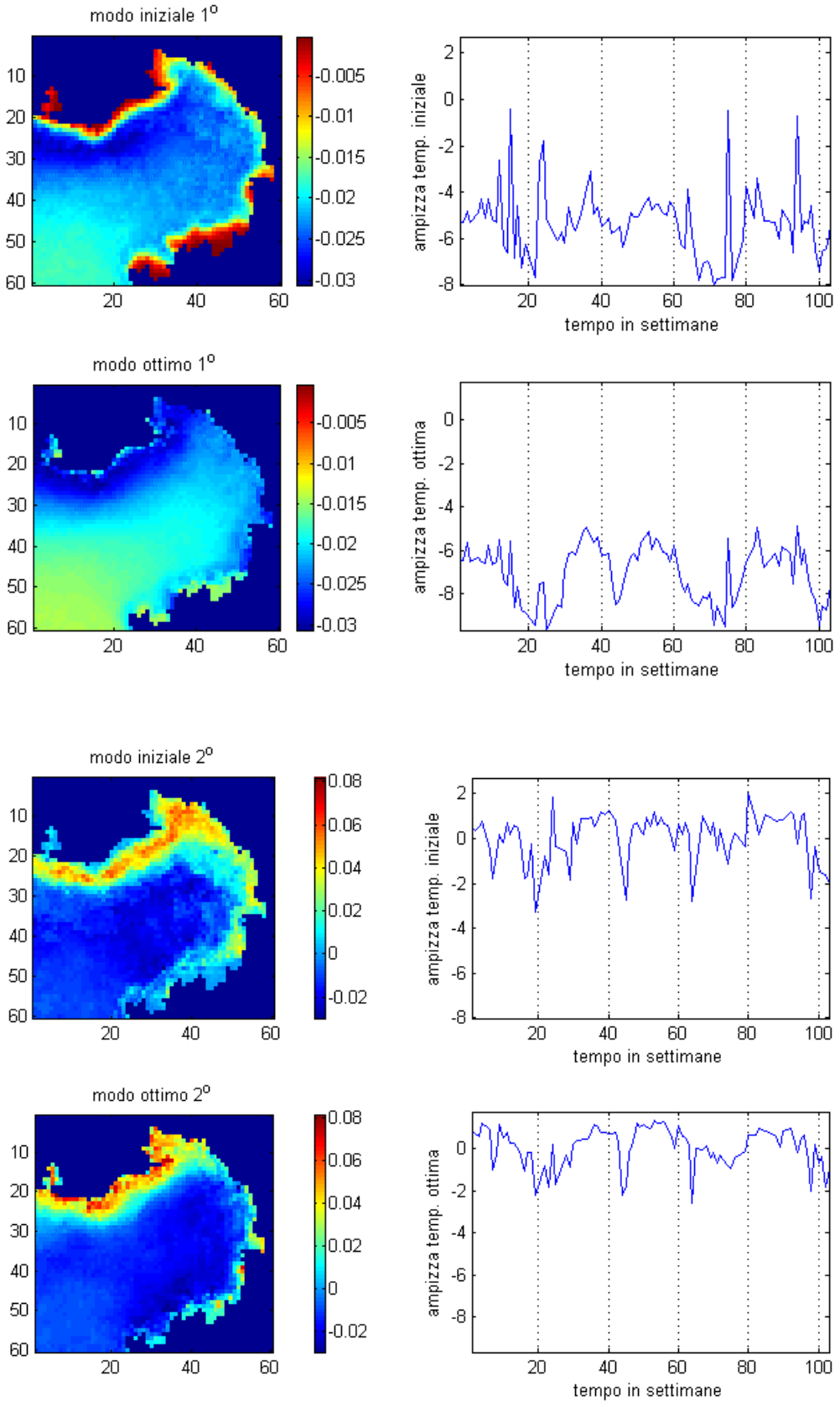

modo iniziale 1°
modo ottimo 1°
modo iniziale 2°
modo ottimo 2°
ampizza temp. iniziale
ampizza temp. ottima
tempo in settimane

Figura 21.

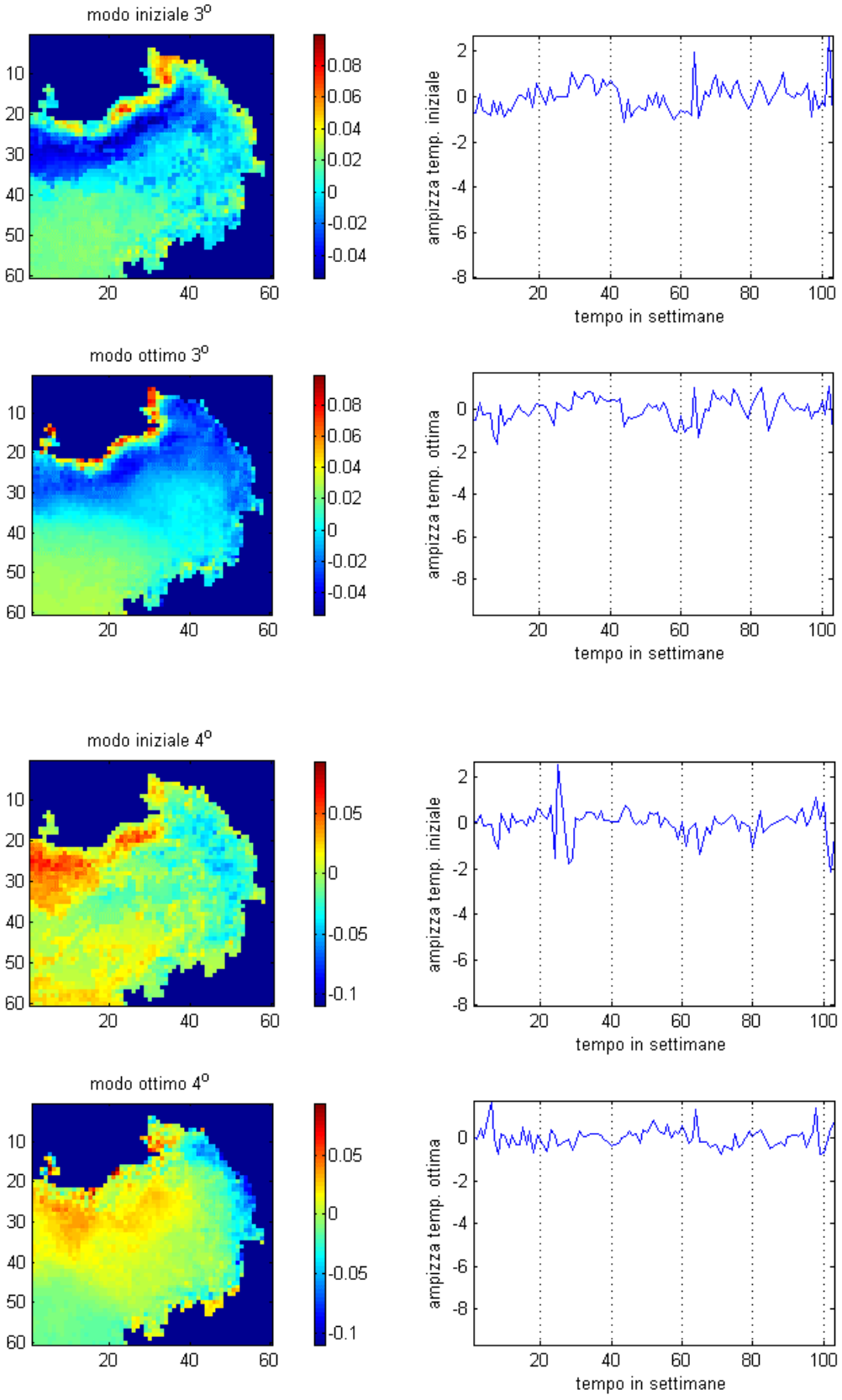

modo iniziale 3°
modo ottimo 3°
modo iniziale 4°
modo ottimo 4°
ampizza temp. iniziale
ampizza temp. ottima
tempo in settimane

Figura 21.

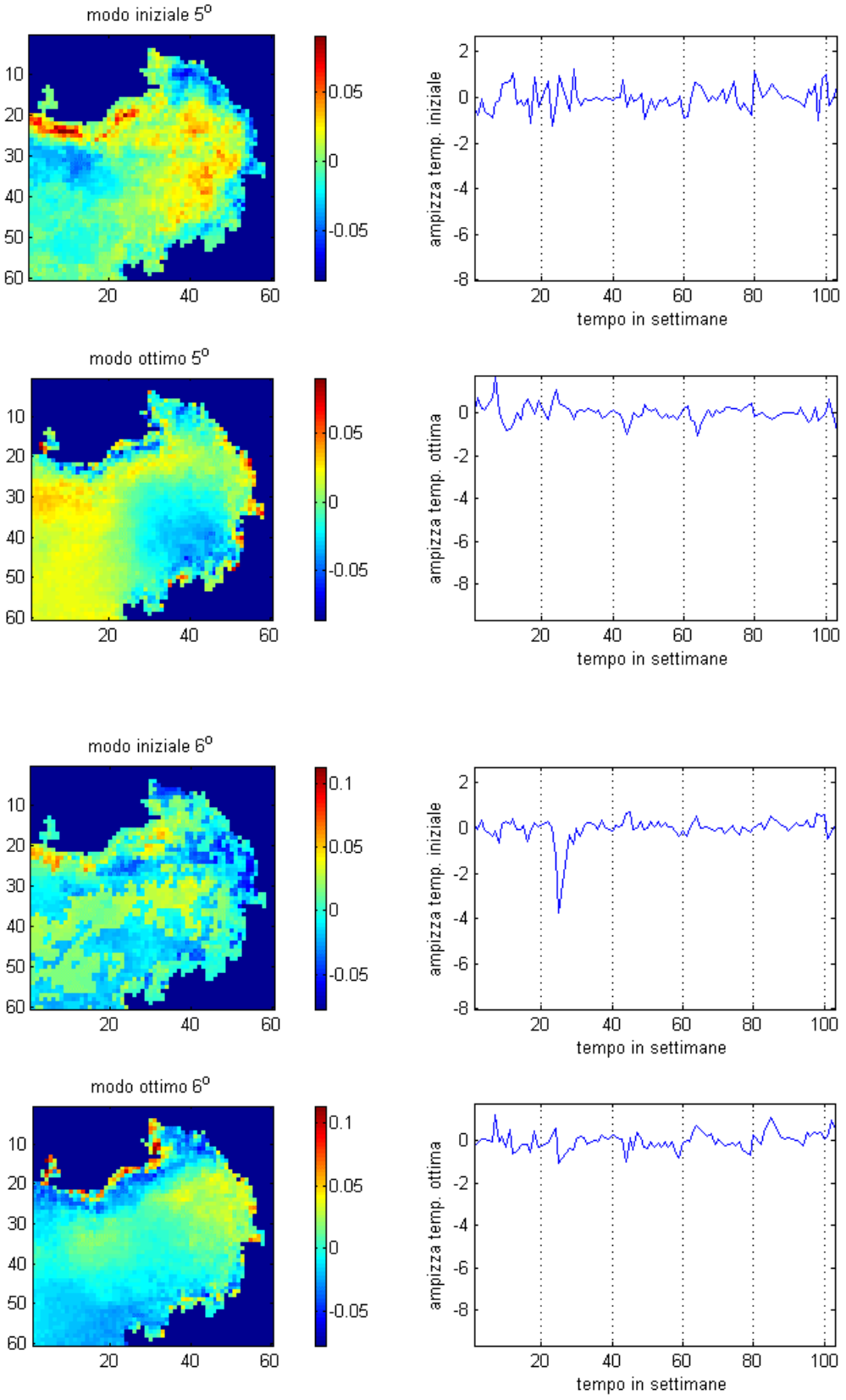

modo iniziale 5°
ampizza temp. iniziale
tempo in settimane
modo ottimo 5°
ampizza temp. ottima
tempo in settimane
modo iniziale 6°
ampizza temp. iniziale
tempo in settimane
modo ottimo 6°
ampizza temp. ottima
tempo in settimane

Figura 21.

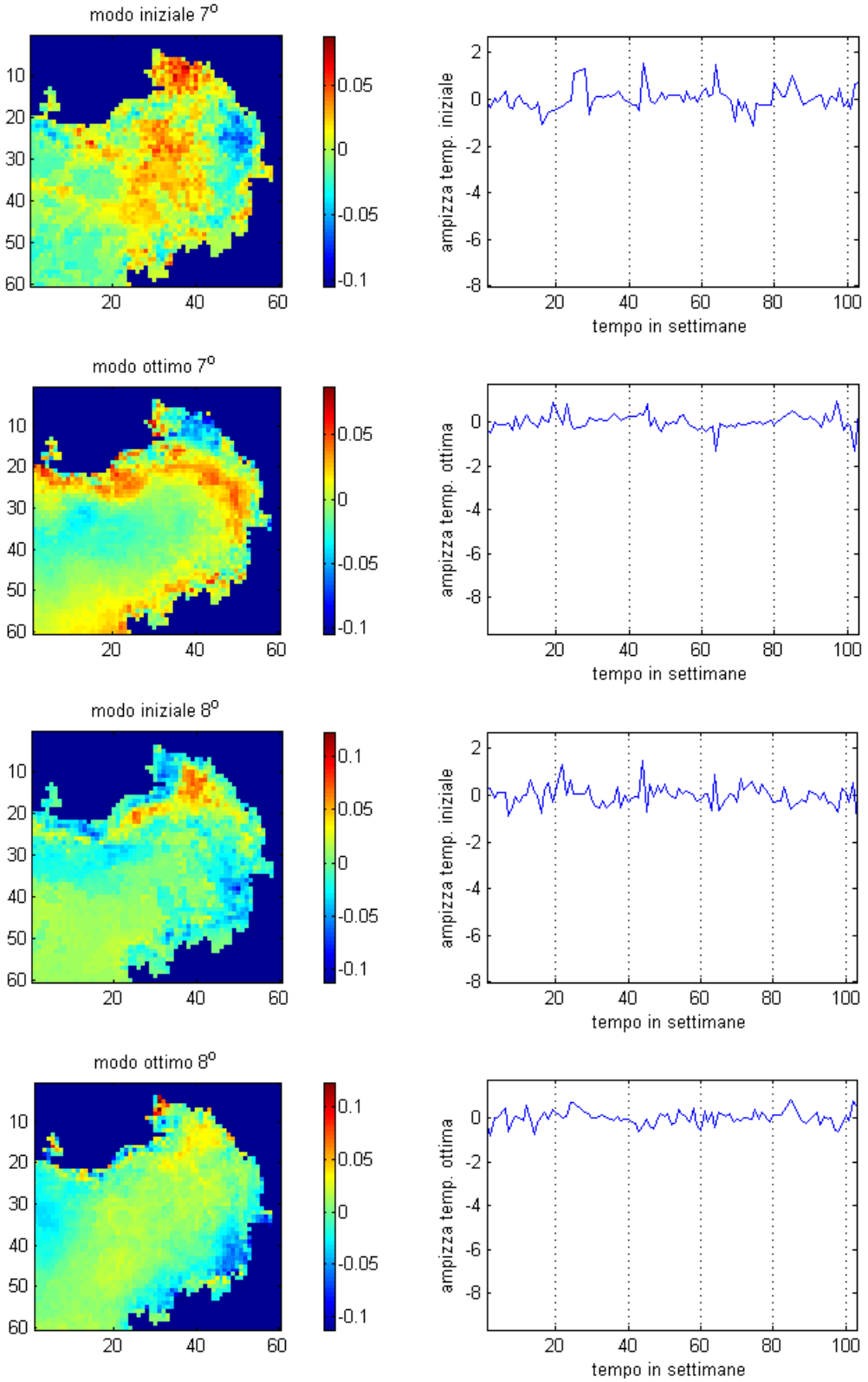


Figura 21. *A destra sono rappresenta le immagini dei modi spaziali. A sinistra quelle delle ampiezze temporali riferite al modo rispettivo. Entrambe sono date sia per la matrice dei dati ridotti e interpolati ottimamente, per le prime otto EOF.*

La figura 22 riporta l'errore relativo (in percentuale) utilizzando i primi quattro modi spaziali EOF della matrice di dati interpolati ove la nuvolosità limite è del 95%, verso gli stessi risultati con il limite di nuvolosità del 20%. Si osserva che, il primo modo può essere ritenuto simile mentre per i rimanenti la similitudine peggiora, anche se le immagini settimanali (quelle comuni) ottenute sono praticamente uguali. Lo stesso fatto si riscontra in figura 23, che mette a confronto le ampiezze temporali dei primi quattro modi per i due casi, con la differenza che la forma delle funzioni ampiezza temporale rimangono simili per le settimane comuni. Le differenze maggiori sono provocate dall'interpolazione lineare applicata al caso con nuvolosità del 20%, applicata a settimane mancanti.

Dalla combinazione lineare delle immagini dei modi (EOF spaziali) U, con le ampiezze temporali A e aggiungendo la media totale della matrice ridotta, si ottiene la matrice dei dati interpolati ottimamente $D = UA^T + media_{tot}$. Alcune immagini prodotte da tale combinazione lineare riportate in figura 24 con le rispettive immagini iniziali.

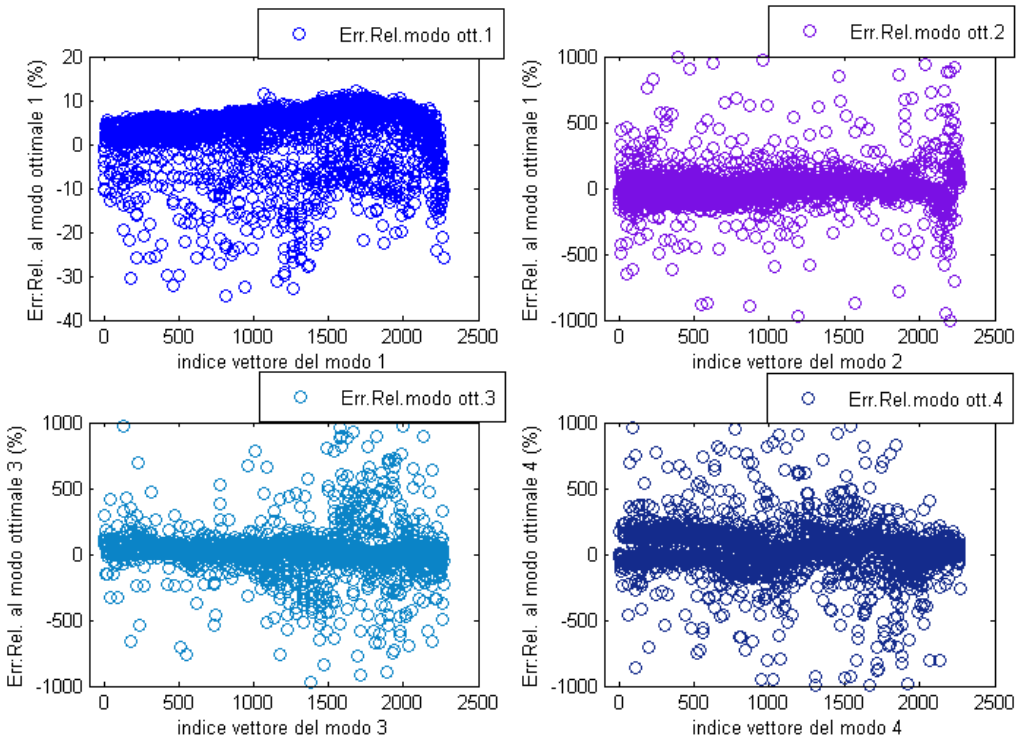


Figura 22. *Errore relativo (in percentuale) utilizzando i primi quattro modi spaziali EOF della matrice di dati interpolati ove la nuvolosità limite è del 95%, verso gli stessi*

risultati con il limite di nuvolosità del 20%. Per il secondo, il terzo e il quarto modo si è ristretto l'intervallo dell'ordinata per evidenziare la distribuzione dell'errore relativo.

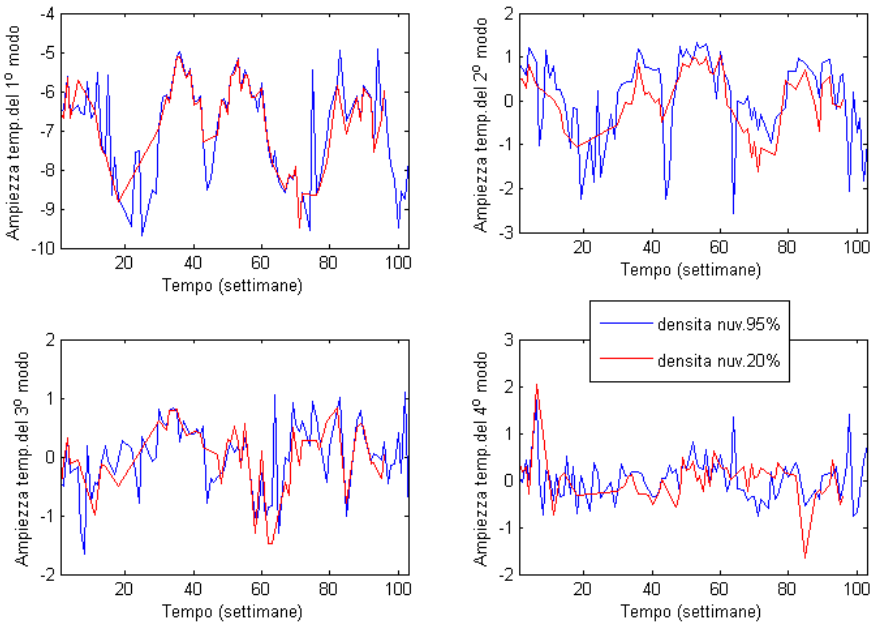


Figura 23. *Ampiezza temporale dei primi quattro modi spaziali EOF della matrice di dati interpolati ove la nuvolosità limite è del 95% e gli stessi risultati con il limite di nuvolosità del 20%. Le ampiezze delle settimane mancanti sono interpolate linearmente.*

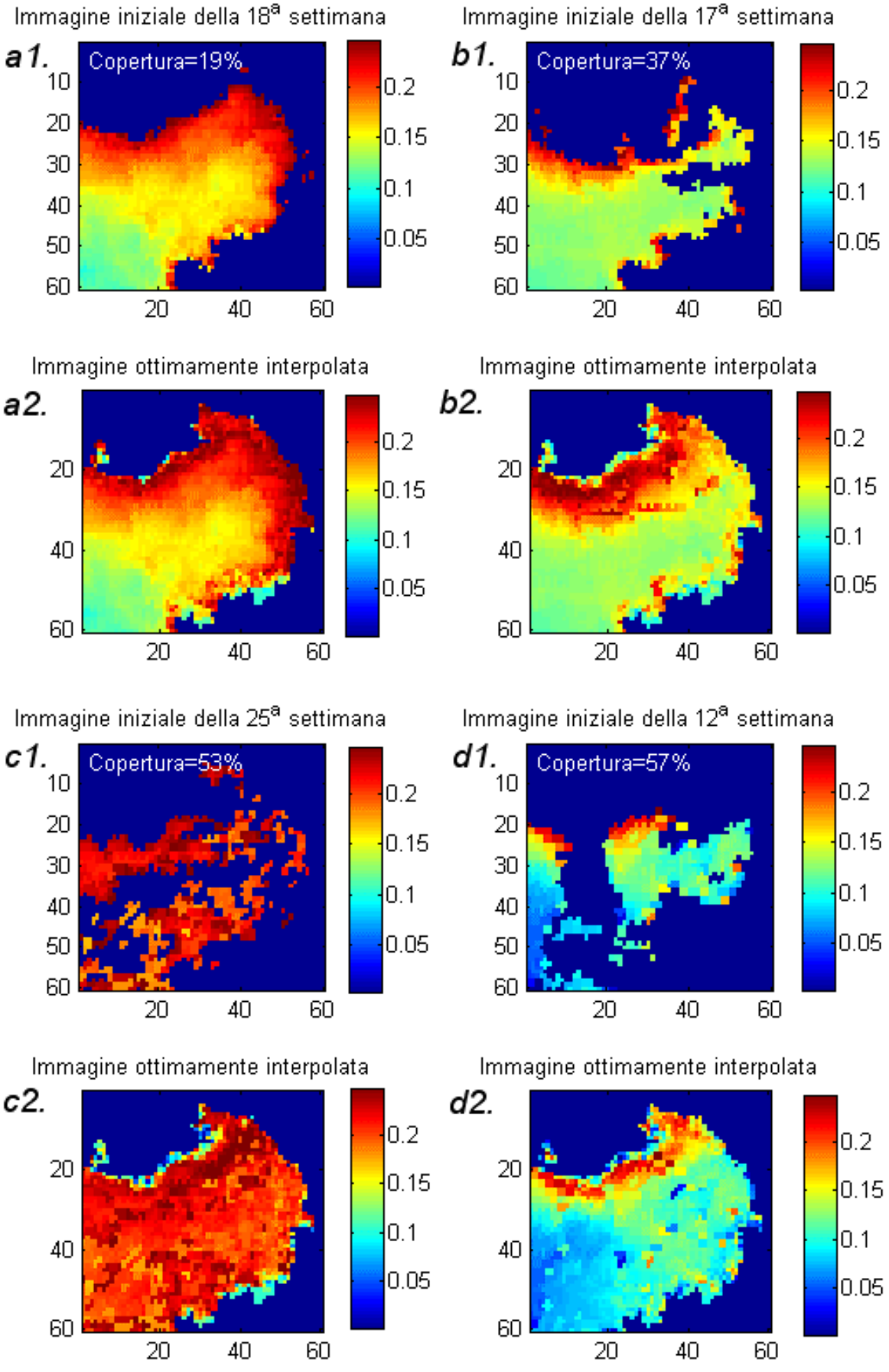


Immagine iniziale della 18a settimana
a1.
Copertura=19%
Immagine iniziale della 17a settimana
b1.
Copertura=37%
Immagine ottimamente interpolata
a2.
Immagine ottimamente interpolata
b2.
Immagine iniziale della 25a settimana
c1.
Copertura=53%
Immagine iniziale della 12a settimana
d1.
Copertura=57%
Immagine ottimamente interpolata
c2.
Immagine ottimamente interpolata
d2.

Figura 24.

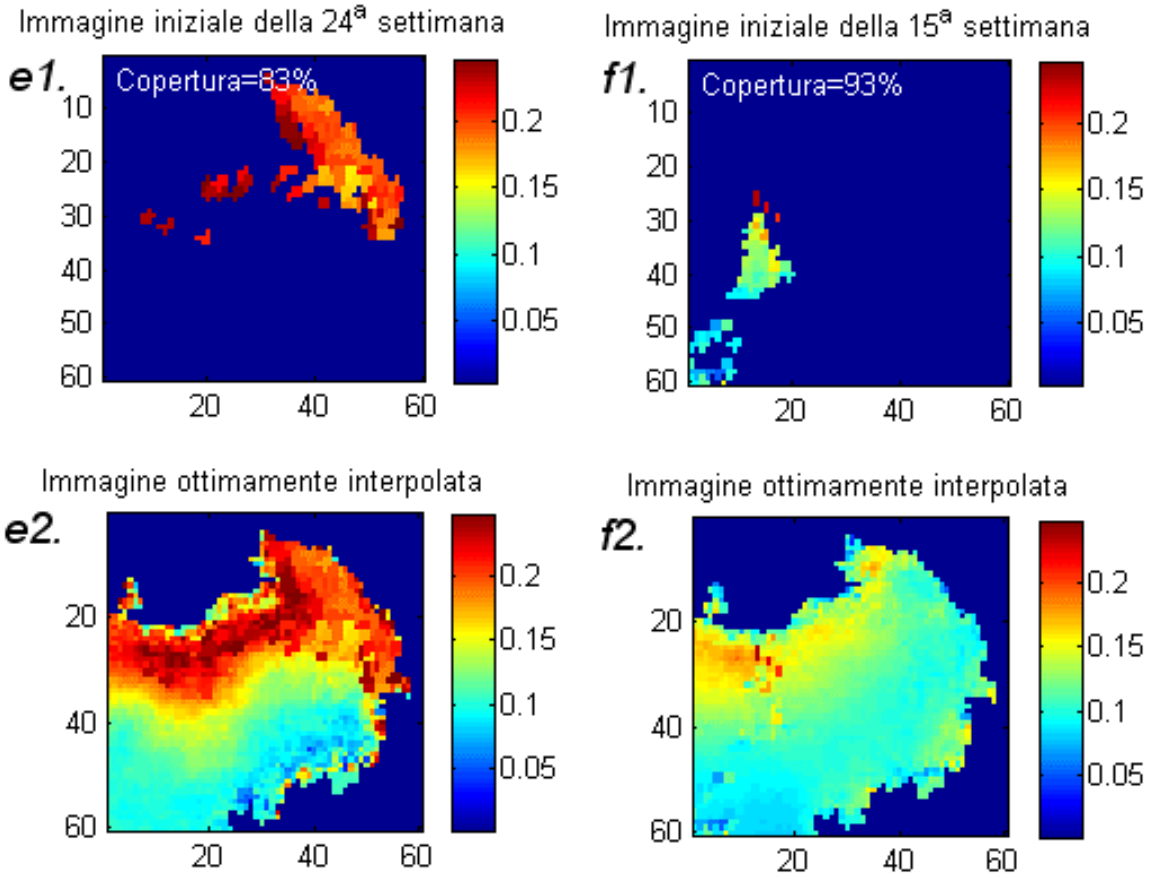


Figura 24. *Alcune immagini settimanali interpolate tramite il numero di modi ottimale, segnate dal numero 2 e mostrate con le rispettive immagini originali, segnate dal numero 1. I valori delle immagini interpolate maggiori o uguali a 0,25 sono stati considerati uguali al valore 0,2499.*

Come si nota in tutte le immagini interpolate, che il metodo presenta grande difficoltà nello stimare i valori vicino alla costa, infatti pronunciata è la variabilità in questa zona. Secondo il metodo, tali valori sono relativamente bassi, il che non rispecchia la realtà dei fatti: ci si aspetterebbe valori alti per la presenza di foci di fiumi, dell'infrangersi delle onde, del fondale basso e altri fenomeni che aumentano la torbidità dell'acqua. Tale fatto si potrebbe spiegare con l'osservazione già presentata dalla figura 6, cioè che si hanno pochi dati lungo la costa e che quindi non viene ?registrata? alcuna struttura spaziale in tale zona da parte del metodo.

Una difficoltà di stima dei valori (poco smussato) si propone anche al margine delle nuvole che coprono il mare: nella figura 24b1 il confine della nuvola che si trova all'incirca a 30 pixel

di ?longitudine?, provoca nell'immagine interpolata una visibile discontinuità dei dati. In questo caso si può attribuire la causa all'elevata imprecisione del rilevamento satellitare, dovuta a un'inefficiente mascheramento in tale zona. Esempio di tale imprecisione ne è parzialmente la figura 24c1, che comunque sembra essere soggetta a meno discontinuità di dato della precedente immagine.

I dati evidentemente anomali della figura 24d1, meglio visibili nella 24d2, sono dovuti al fatto che i dati veri, anche se sbagliati, sono ancora presenti. È importante quindi togliere tali dati già all'inizio per evitare che questi influiscano sull'interpolazione.

L'idea che si propone è di mascherare o smussare le zone di discontinuità e i punti di rumore tramite la riduzione della varianza del segnale, in modo da tagliare le varianze delle scale piccole, contenenti rumore, senza però alterare le scale medie e grandi. Il risultato di tale operazione di riduzione della varianza è presentato dai filmati 1, 2, 3, 4 che presentano un taglio rispettivo del 4,3% (usato 1 modo), 2,1% (usati 3 modi), 1,1% (usati 6 modi) e 0,2% (usati 36 modi) della varianza del segnale. In tali filmati le immagini settimanali interpolate e di varianza ridotta (in alto a sinistra) sono accompagnate dalle immagini di errore relativo percentuale tra l'immagine al 100% di varianza e la rispettiva ridotta in varianza (in alto a destra), per avere uno sguardo sul gradiente di variazione di questa. L'immagine settimanale di varianza 100% è rappresentata in basso a sinistra e l'immagine iniziale è in basso a destra.

Nella figura 25 viene riportato un ?frame? riferito all'immagine della prima settimana di dicembre 2002.

Nella scelta della varianza ottima da usare si è fatto ausilio dei filmati 5, 6 e 7, i quali riportano le immagini settimanali interpolate decisamente problematiche, per le quali viene variata la quantità di varianza contenuta. Il risultato di tale analisi oggettiva indica che la varianza da conservare in modo da ottimizzare il compromesso è dell'99.79%. Si tratta di utilizzare 36 modi nella ricostruzione dell'immagine interpolata. La figura 26 rappresenta vari ?frames? del filmato 5. Si osserva che la fascia di dati anomali (figura 24b1) si estingue progressivamente con l'aumentare della varianza contenuta.

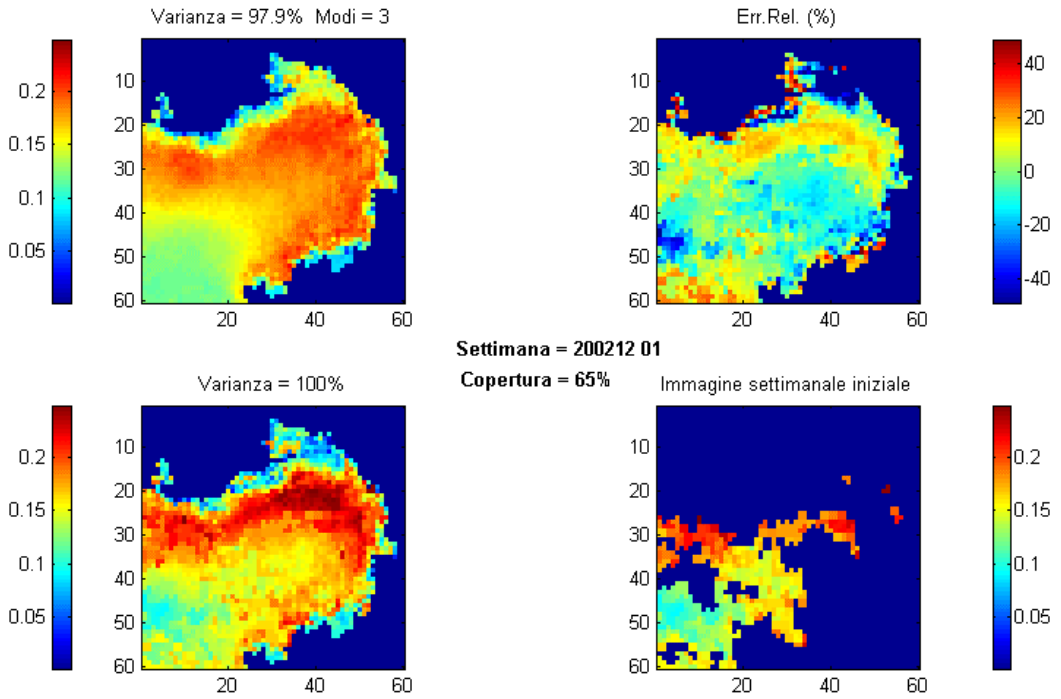


Figura 25. *Immagine ?frame? della prima settimana interpolata di dicembre 2002 ridotta della sua varianza del 2.1%, utilizzando 3 modi (in alto a. sinistra), la rispettiva immagine al 100% della varianza (in basso a sinistra), l'immagine dell'errore relativo in percentuale tra l'immagine a varianza totale e quella ridotta (in alto a destra) e l'immagine iniziale (in basso a destra).*

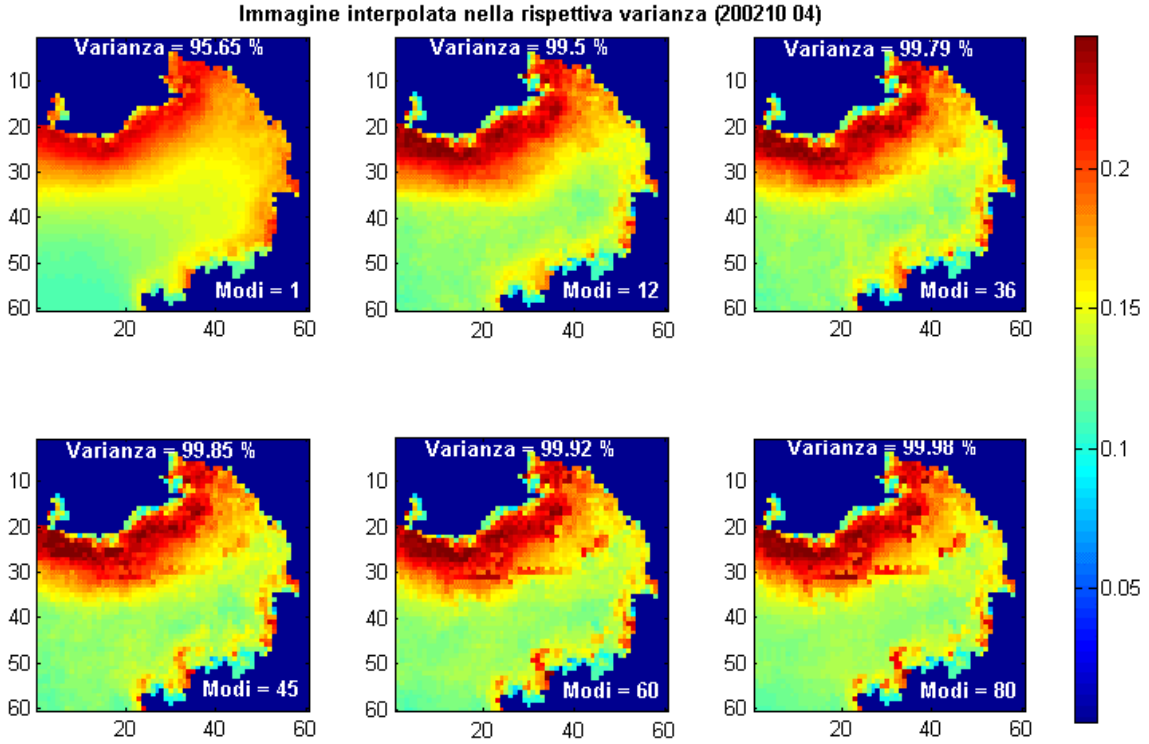


Figura 26. *Immagine interpolata della quarta settimana di ottobre 2002 (17^a settimana) e ricostruita con quantità crescenti di modi EOF.*

I filmati 8 e 9 consentono di osservare le immagini settimanali interpolate di varianza 99,79% e 100% accanto alle loro immagini di conteggio, per poter stimare soggettivamente l'attendibilità dei soli dati iniziali. La figura 27 riporta il caso delle prime sei settimane interpolate con varianza del 99,79% con le rispettive immagini di conteggio.

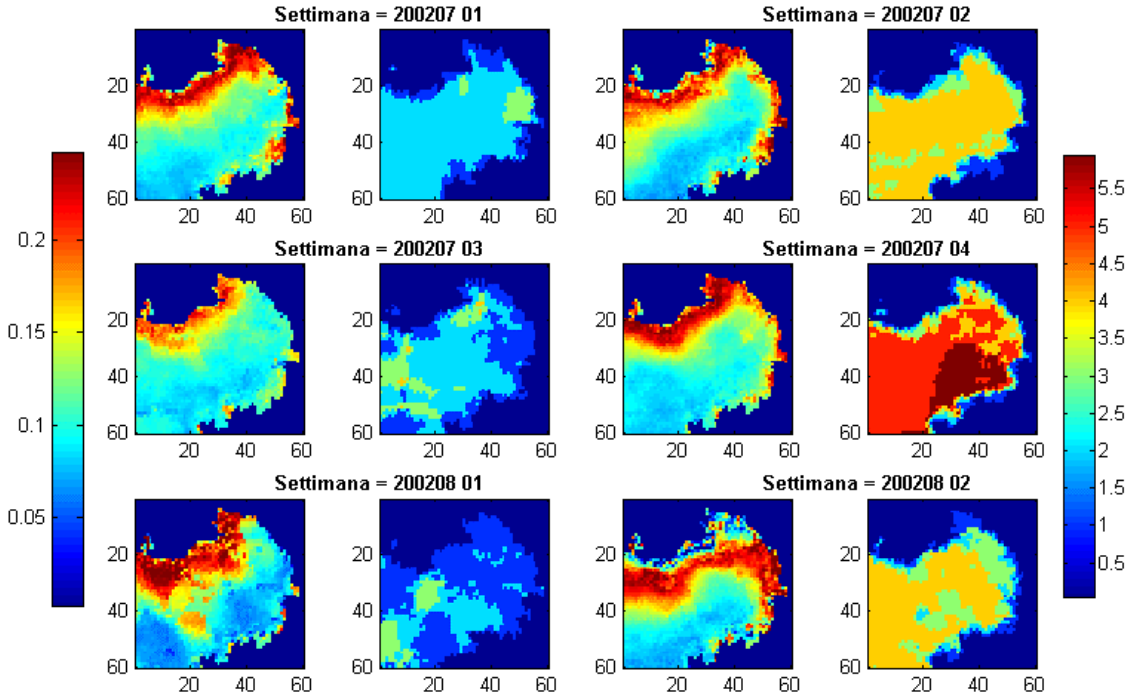


Figura 27. *Prime sei immagini della matrice interpolata espresse col 99.79% della varianza con le rispettive immagini di conteggio. La scala di colore riferita alle immagini di conteggio è significativa solo per numeri interi (barra a destra).*

Infine si osserva chiaramente che il metodo è molto efficace nell'interpolare immagini con un'elevata nuvolosità, come nel caso delle figure 24e1 e 24f1, con però le ipotizzate incertezze sopra espresse riguardo il test di cross-validazione. Nei filmati è evidente un comportamento dubbio della quindicesima settimana interpolata (seconda sett. dell'Ottobre 2002) e settantasettesima (prima sett. di Dicembre 2003). Queste due presentano valori molto bassi per il periodo in cui si trovano: infatti le settimane antecedenti e precedenti hanno valori medi molto alti di *K490* e indicano una fioritura del mare. I valori improvvisamente bassi potrebbero essere dovuti a fattori fisici improbabili in questo periodo dell'anno come a un'improvvisa calma di

vento, oppure un calo della portata dell'Isonzo. Bisognerebbe quindi svolgere un'analisi di correlazione tra i valori del *K490*, dell'intensità del vento e della portata dell'Isonzo per i due periodi. In tal modo si potrebbe valutare se le due immagini sono valide e se i valori interpolati con il metodo attraverso EOF hanno senso, poiché causati da fenomeni fisici correlati.

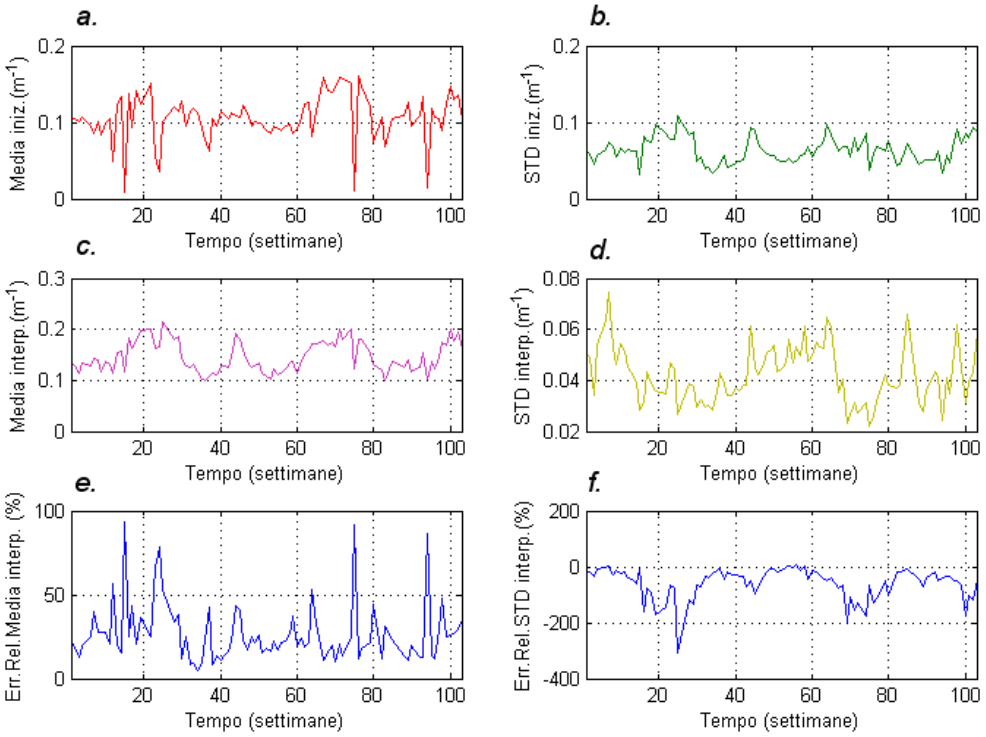


Figura 28. ***(a)*** *Media settimanale della matrice ridotta,* M_{iniz}*; si è interpolato linearmente dove le settimane mancano di dati.* ***(b)*** *Deviazione standard settimanale della matrice iniziale,* STD_{iniz}*.* ***(c)*** *Media settimanale della matrice ottimamente interpolata,* M_{interp}*; si è interpolato linearmente le medie delle settimane mancanti.* ***(d)*** *Deviazione standard settimanale della matrice ottimamente interpolata,* STD_{interp}*.* ***(e)*** *Errore relativo (in. percentuale) tra la media settimanale della matrice interpolata e quella iniziale,* $\frac{M_{interp} - M_{iniz}}{M_{interp}} 100$*.* ***(f)*** *Errore relativo (in percentuale) tra la deviazione standard settimanale della matrice interpolata e quella iniziale,* $\frac{STD_{interp} - STD_{iniz}}{STD_{iniz}} 100$*.*

Nella figura 28a si osservano dei valori molto bassi di media settimanale riferiti alla matrice ridotta, provocati dalle settimane con grande densità di dati mancanti, che però vengono ragionevolmente tagliati nella figura 28c, situazione con densità di dati mancanti del 0%, poiché

riferita alla media settimanale della matrice interpolata. Nelle settimane con bassa nuvolosità (minore del 25%), si osserva che l'errore relativo percentuale riferito alla media della matrice dei dati interpolati non supera il 25% (figura 28e). Ciò è ragionevole, poiché i dati interpolati aggiunti sono relativamente pochi e quindi influenzano di poco la media settimanale iniziale. Dalla figura 28c si interpreta sicuramente meglio i fenomeni fisici a grande scala che non dalla figura 28a e dalla figura 3a (media settimanale riferita alla matrice ridotta), grazie ad un andamento più regolare. I bassi valori della deviazione standard (di figura 28d) rendono più attendibile l'interpolazione delle medie dalla figura 28c.

5. Conclusioni

Il metodo di interpolazione EOF dei dati mancanti in immagini satellitari MODIS è stato applicato con successo nel Golfo di Trieste. Questo metodo auto-consistente e libero da parametri predefiniti presenta il seguenti vantaggi:

- Permette di creare un banca dati con immagini complete, fatta eccezione per le immagini che presentano 100% di copertura nuvolosa. Tali immagini mancanti possono però essere ricostruite tramite l'interpolazione delle ampiezze temporali e la successiva ricostruzione tramite i modi spaziali comuni per tutte le immagini. Con questo metodo è anche possibile ogni pixels calcolarsi l'errore percentuale.
-
- Consente la decomposizione EOF del segnale e quindi l'interpretazione della sua variabilità spazio/temporale. Tale metodo permette inoltre di filtrare la variabilità ad alta frequenza. escludendo il possibile rumore sperimentale.
-

Gli svantaggi del metodo evidenziati sono la necessità di applicare il metodo su immagini interpolate perfettamente mascherate. Valori molto alti o bassi portano all'ottenimento di immagini con strutture spaziali, ma con valori erronei. Si suggerisce quindi, una scrupolosa analisi dell'immagini prima dell'applicazione del metodo e un mascheramento delle nuvole piuttosto conservativo.

6. Bibliografia

Beckers, J.-M.. & M. Rixen (2003): *EOF Calculations and Data Filling from Incomplete Oceanographic Datasets. Journal of Atmospheric and Oceanic Technology*, 20, 1839-1856.

Clark, D.K. (2004): MODIS, *Terra - Diffuse Attenuation Coefficient at 490 nm (K490), Data Quality Summary*. http://modis-ocean.gsfc.nasa.gov/qa/terra/dataqualsum/ k490_qualsum.pdf.

Emery, W.J. & R.E. Thomson (1997): Data *analysis methods in physical oceanography*. Pergamon, pp. 319-336.