POLITECNICO DI MILANO

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

KRIGING E RBF-NETWORK PER L'INTERPOLAZIONE DI DATI METEOROLOGICI IN LOMBARDIA

Relatrice: Prof. Giovanna Venuti

Correlatore: Stefano Barindelli

Tesi di laurea magistrale di:

Valerio Guglieri

Matr. 884235

Anno accademico 2017/2018

Sommario

1. INTRODUZIONE

1.1. Il progetto LAMPO

L'obiettivo del progetto LAMPO ("Lombardy-based Advanced Meteorological Predictions and Observations") è "sperimentare un sistema innovativo a basso costo per la previsione a brevissimo termine dei temporali e mitigare l'impatto di esondazioni e dissesti idrogeologici attraverso una rete capillare di stazioni GPS/GNSS dislocate sul territorio lombardo"[1]. La fase sperimentale del progetto interessa il bacino del Seveso, un fiume molto noto ai milanesi per i danni associati alle sue violente e improvvise esondazioni.

Il Seveso è un torrente che nasce a Cavallasca (CO) e "sfocia" nel Naviglio Martesana dopo un percorso di 52km, di cui l'ultimo tratto di circa 9km si trova sotto la città di Milano. Il suo bacino è situato interamente in Lombardia. Il fiume è a carattere torrentizio e, anche a causa della cementificazione del suo bacino, risente in modo particolarmente violento dei fenomeni meteorici estremi. Inoltre, a causa della sua lunghezza ridotta, la propagazione delle onde di piena è molto rapida e spesso non consente di mettere in opera con un adeguato preavviso i sistemi di protezione.

L'idea alla base del progetto, quindi, è di spostare la previsione dall'evento di piena a una sua possibile sua causa, ovvero l'evento meteorologico estremo. Il primo passo per l'implementazione di questa idea è mettere in relazione diverse variabili meteorologiche tramite un modello che le correli alla probabilità di accadimento di un evento estremo e, possibilmente, con la sua intensità.

In particolare, si è deciso di ricorrere a un approssimatore universale di tipo rete neurale.

1.2. La tesi

A prescindere dal tipo di modello scelto per l'implementazione, i dati di input vanno opportunamente pre-elaborati ad esempio per rimuovere possibili errori grossolani, per colmare assenze di dato e per rendere uniformi gli intervalli temporali di campionamento.

Il database considerato è composto da serie temporali di temperatura dell'aria, umidità relativa, pressione atmosferica, velocità e direzione del vento, precipitazione. A queste si aggiungono serie temporali di Ritardi troposferici Zenitali o ZTD derivati dalla elaborazione dei dati GNSS raccolti dalla rete di sensori a basso costo del progetto Lampo. Le osservazioni meteorologiche derivano da sensori posti in tutta la Lombardia: di ogni sensore si conoscono il codice identificativo e la posizione.

[1] https://www.arpalombardia.it

Le serie temporali delle variabili in esame hanno lunghezze diverse a seconda dei periodi di funzionamento di ciascun sensore e frequenze di campionamento diverse. Ad esempio, i valori di ZTD sono disponibili con un passo temporale di 30 secondi, mentre la precipitazione è rilevata ogni minuto; le altre serie hanno tendenzialmente un passo temporale di 10 minuti, anche se si possono trovare intervalli di 20, 30 o 60 minuti tra due misure successive. Oltre al pre-processing richiesto dal particolare modello che si deciderà di implementare, queste serie richiedono un lavoro di interpolazione e downsampling per essere rese omogenee. Lo scopo di questo elaborato è, pertanto, portare a termine questi compiti analizzando i dati a disposizione e confrontando due diversi metodi per l'interpolazione dei record mancanti: kriging e Radial Basis Function Network (RBFN).

Nell'analisi di serie temporali di variabili meteorologiche è frequente imbattersi in sequenze di dati disomogenee e discontinue. Le cause di mancanza di dato sono essenzialmente due: il sensore non ha effettuato la misura oppure in post-processing si è deciso di eliminare una misura (ad esempio un outlier) o di aggiungere record "vuoti" per rendere omogenea la serie. I motivi per cui un sensore non effettua la misura sono svariati: cali di tensione che inficiano il comportamento dell'apparecchio, malfunzionamenti generici causati da eventi esterni o interventi di manutenzione del sensore. Questi comportamenti creano, nelle serie temporali, 'buchi' irregolari di lunghezza e frequenza casuale.

Un comportamento diverso si nota se i 'buchi' temporali sono sintetici, ovvero creati artificialmente in post-processing. I motivi per cui si inseriscono record vuoti in questa fase sono due: un analista potrebbe aver rimosso un outlier, e allora i 'buchi' sono tendenzialmente isolati e, per definizione, hanno frequenza bassissima; oppure potrebbe aver adeguato il passo temporale a un'altra serie a più elevata risoluzione. In quest'ultimo caso, i record vuoti hanno lunghezza costante e prevedibile. Si ammetta, ad esempio, di avere un sensore che rileva la temperatura dell'aria ogni 30 minuti: se si volesse creare una serie omogenea a 10 minuti, bisognerebbe aggiungere 2 record vuoti dopo ogni misura.

Queste differenze sono rilevanti nella scelta degli algoritmi da usare per l'interpolazione dei dati. Nel caso di buchi isolati, ad esempio, risulta più conveniente dal punto di vista della velocità computazionale adottare algoritmi di interpolazione lineare, o anche qualcosa di più semplice come la media aritmetica tra i valori successivi e precedenti al record mancante. In alternativa si potrebbe optare per un algoritmo più lento ma più preciso come, ad esempio, un interpolatore spline. Un discorso analogo vale qualora i buchi siano più lunghi ma distribuiti in modo omogeneo e costante lungo tutta la serie temporale. È compito dell'analista distinguere i diversi casi e valutare il trade-off tra velocità di calcolo e accuratezza dell'interpolazione.

Nel caso in cui i record mancanti siano raggruppati in intervalli molto lunghi, i principali algoritmi di interpolazione potrebbero dare risultati molto scarsi, se non addirittura controproducenti in quanto rischiano di inficiare pesantemente la capacità predittiva del modello per cui verranno utilizzati. In tal caso è meglio non riempire l'intervallo vuoto con valori interpolati, lasciando gestire questa mancanza di informazione al modello da implementare successivamente. Questa, peraltro, è una condizione che si verifica spesso quando si lavora con variabili ambientali. Si pensi, ad esempio, al guasto di un sensore posto in montagna, in un luogo difficile da raggiungere: il rischio che non vengano effettuate misurazioni per qualche mese è concreto. L'algoritmo per la predizione degli eventi meteorici estremi dovrà, quindi, essere in grado di funzionare nonostante alcuni valori in input possano essere nulli.

Il primo passo da compiere per la pulizia dei dati è l'analisi dei record mancanti. Nel prossimo capitolo verranno mostrati i risultati di questa operazione, oltre alle statistiche principali che descrivono le diverse serie temporali. In riferimento ai record mancanti, si intende darne una caratterizzazione utile come guida per la scelta degli algoritmi di interpolazione, evidenziando la distribuzione della lunghezza dei buchi e della frequenza con la quale si verificano.

Si effettuerà, poi, l'interpolazione dei dati mancanti, o si valuterà l'opportunità di lasciare inalterati i buchi di dimensioni troppo grandi. Nel capitolo [3] verranno mostrati i presupposti teorici dei due metodi principali scelti per il confronto. Per quanto riguarda le interpolazioni più comuni, come la media aritmetica, non vengono fornite descrizioni matematiche. Infine, nel capitolo [4] verranno mostrati i risultati dell'applicazione di tali algoritmi, si verificherà la bontà delle interpolazioni ottenute e si confronteranno tra lori i diversi metodi.

I due algoritmi scelti per l'interpolazione sono kriging e Radial Basis Function Network (RBFN). Pur partendo da presupposti teorici molto diversi e pur essendo stati sviluppati in ambiti disciplinari completamente differenti, le formulazioni matematiche dei due modelli sono piuttosto simili. Viene qui fornita una rapida spiegazione dei concetti base dei due algoritmi: si rimanda il lettore al capitolo 3 per un'esposizione approfondita e completa dei necessari passaggi matematici.

Il kriging è un metodo di interpolazione stocastico, in cui le variabili oggetto di predizione sono modellizzate come un insieme di variabili casuali, parametrizzate nel tempo (segnali stocastici o, nello spazio, campi stocastici). La predizione della variabile in un tempo t è ottenuta come combinazione lineare delle osservazioni della stessa variabile (in un insieme di tempi $t_1, \dots, t_N$), i cui coefficienti dipendono dalla covarianza tra le osservazioni e tra quella delle osservazioni con il punto di predizione. È stato sviluppato a partire dagli anni '50 e ha trovato un uso molto ampio in geostatistica; è stato contemporaneamente studiato anche in statistica con il nome di "Gaussian

process regression" e applicato a problemi di machine learning. Sviluppato nell'ambito dell'ingegneria mineraria, è basato sul presupposto che variabili temporalmente (o spazialmente) più vicine siano più simili (ovvero più correlate) tra loro che non quelle lontane.

Il grado di similitudine, espresso in termini di covarianza, è stimato empiricamente attraverso il semivariogramma, una funzione che interpola la semi-varianza delle differenze tra i valori osservati. Matematicamente, quindi, la stima in un punto può essere interpretata come la combinazione lineare di trasformazioni dell'input, dove per "input" si intende il valore dell'ascissa (epoca) del punto da stimare e le trasformazioni sono funzioni della distanza, in questo caso temporale, tra l'input e i punti noti. I parametri dipendono anche dalle distanze e dai valori dei punti noti. La stima in un punto è ottenuta sfruttando tutti i punti noti della serie temporale, quindi i tempi di predizione sono O(n).

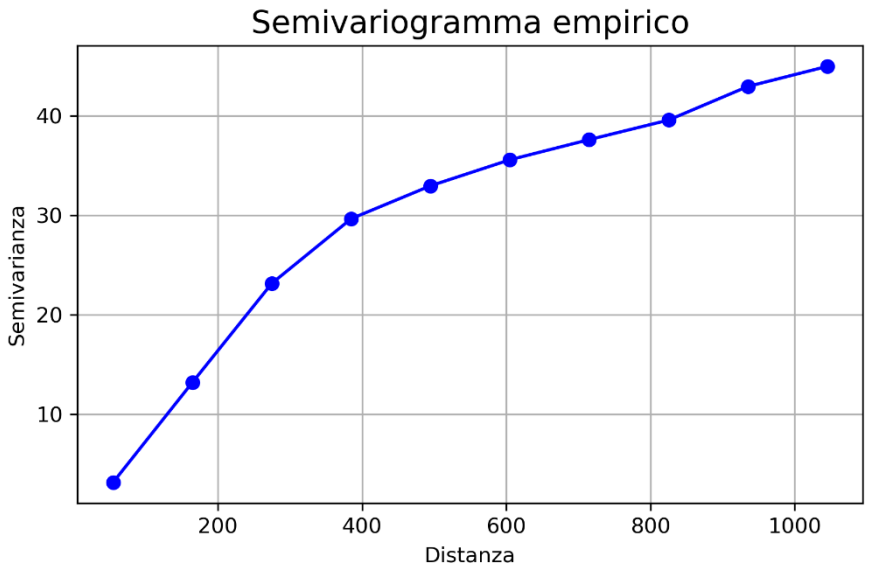

Fig. 1 - Esempio di semivariogramma empirico.

Le RBFN sono modelli nati nell'ambito dello sviluppo di reti neurali artificiali (ANN, dall'inglese "Artificial Neural Network"). Di fatto, esse sono delle ANN costituite da 3 layer: uno per gli input, uno per le funzioni di attivazione e uno per l'output. A differenza delle classiche ANN, che usano come funzioni di attivazione le tangenti iperboliche, qualche funzione sigmoidale o una ReLU, le RBFN usano delle funzioni a base radiale (RBF, dall'inglese "Radial Basis Function"). Queste sono funzioni a valori reali, il cui valore dipende solo dalla distanza tra l'input e un punto fisso del dominio. Esistono diverse famiglie di RBF, tra cui le più famose sono le gaussiane. Nel caso in esame, le RBF sono centrate nei punti noti della serie temporale. Come per le ANN ordinarie, il calcolo dell'output è effettuato come combinazione lineare di trasformazioni dell'input. È importante notare che i

collegamenti tra layer successivi sono “pesati”: i pesi tra il layer degli input e quello delle funzioni di attivazione possono essere imposti pari a 1, riducendo i tempi di training del modello e garantendo che gli input non vengano modificati prima dell’attivazione.

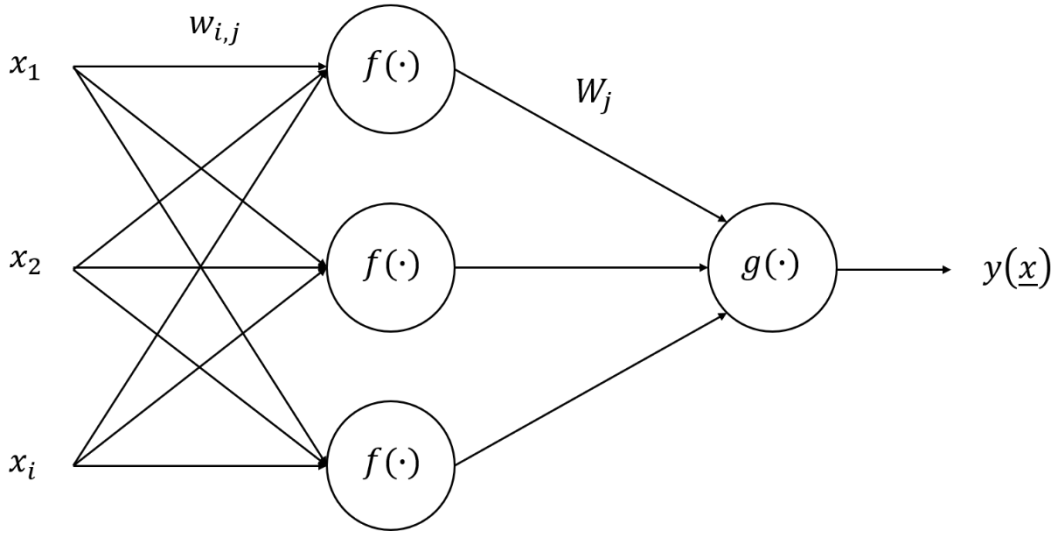

Fig. 2 - Esempio di architettura di una RBFN.

Nell’ambito del machine learning, entrambi gli algoritmi qua nominati rientrano nella categoria delle “macchine a kernel”. I kernel sono funzioni che, oltre ad avere particolari proprietà e vantaggi che verranno elencate nell’apposito capitolo, esprimono la distanza tra due punti. In questo senso, il semivariogramma del kriging e le funzioni a base radiale della RBFN possono essere definiti kernel. La formulazione dei due modelli, pur partendo da considerazioni differenti, porta a un risultato matematicamente molto simile: entrambi forniscono un output che è combinazione lineare di funzioni della distanza tra l’input e i punti noti della serie temporale. Le differenze tra i due metodi sono nella definizione del kernel: se nel kriging bisogna ricorrere all’esperienza dell’analista per calcolare e interpolare il semivariogramma, nelle RBFN le funzioni di attivazione possono essere ottimizzate in fase di training tramite cross-validazione, rendendo quindi il processo automatico.

2. DESCRIZIONE DATI

Il progetto LAMPO è geograficamente collocato in Lombardia e ha tra i suoi partner l'Agenzia Regionale per la Protezione dell'Ambiente (ARPA) della stessa regione. Di conseguenza i dati a disposizione, almeno nelle prime fasi, derivano da sensori posti entro i confini territoriali di questo ente. La rete ARPA della Lombardia è composta da 1830 sensori, raggruppati in stazioni, che coprono tutto il territorio regionale come illustrato in Figura 1. Le variabili meteorologiche osservate dai diversi sensori sono temperatura, pressione atmosferica, umidità relativa, precipitazione, velocità e direzione del vento. Di ogni sensore sono noti un codice identificativo e le coordinate geografiche. I dati di ciascun sensore sono forniti in file di testo contenenti tutte le osservazioni registrate negli anni dal 2001 al 2017. Ogni record del file è composto da tre voci: ID del sensore, data e ora con risoluzione al secondo in cui è stata effettuata la misura e il suo valore. I record vuoti sono contrassegnati dal flag "-999", unico per tutti i sensori e tutte le grandezze. Per la elaborazione dei dati con Python, tale valore è stato sostituito con il flag "NaN" ("Not a Number"), che consente di calcolare le statistiche dei dati omettendo automaticamente tali valori.

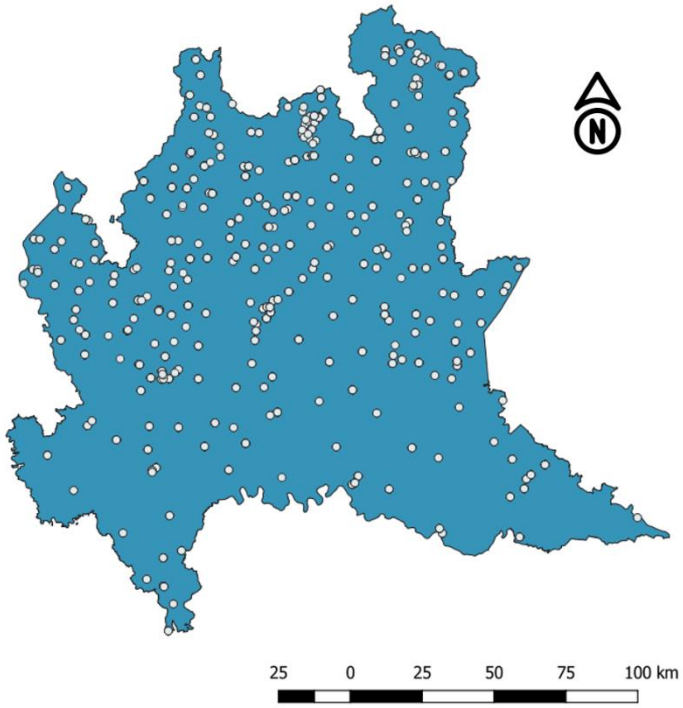

Fig. 3 - Mappa della rete di stazioni di rilevamento ARPA per la misura di variabili meteorologiche.

I dati della rete sono pre-processati da ARPA, che provvede alla rimozione degli outlier. Inoltre, la risoluzione temporale dei dati dei diversi sensori è riportata a quella più elevata introducendo record di no-data ('-999') tra osservazioni a più bassa risoluzione. Il lavoro di ARPA, però, non va oltre queste semplici operazioni e ogni successivo step di pre-elaborazione dei dati è lasciato a chi li utilizzerà. Com'è noto le operazioni di interpolazione introducono degli errori, ed è chi utilizzerà il dato a dover stabilire la soglia di errore accettabile e di conseguenza stabilire l'opportunità e la tipologia di interpolazione da effettuare.

In questo progetto di tesi si analizzeranno nello specifico solo osservazioni di pressione atmosferica e di temperatura.

2.1. Pressione atmosferica

La pressione atmosferica è una delle variabili più utilizzate per descrivere le condizioni meteorologiche di un'area geografica. Essa è, infatti, una delle variabili principali nella caratterizzazione di eventi meteorologici: variazioni di pressione sono infatti responsabili di cambiamenti nella circolazione atmosferica e sono di fondamentale importanza nelle previsioni meteorologiche. La rete di ARPA Lombardia comprende 65 sensori di pressione.

Di seguito è mostrata una panoramica dei principali metadati di tali sensori, in particolare sono analizzati: passi temporali, date di inizio e fine delle osservazioni, numero di record totali e numero di record mancanti. Inoltre, verrà preso come esempio un sensore e verranno mostrate le sue principali statistiche descrittive: media, deviazione standard, quartili, minimo e massimo. L'unità di misura in cui sono espressi tutti i dati è ettopascal [hPa], tranne dove indicato diversamente.

Tutti i sensori analizzati in questo capitolo hanno registrato i valori di pressione atmosferica con un passo temporale di 10 minuti. Ciò si traduce in 144 valori al giorno, 1008 alla settimana e 52.560 all'anno. La lunghezza complessiva delle diverse serie temporali è in realtà molto varia e dipende dal momento in cui ogni sensore è entrato in funzione. L'ultima osservazione per tutte le serie in esame è stata registrata il 1° gennaio 2018 alle 00:00, ovvero è presente solamente un dato per questo anno. Di conseguenza, la lunghezza delle serie temporali dipende solo dall'anno di entrata in funzione dei sensori: questi sono distribuiti in modo disomogeneo tra il 2002 e il 2013, come mostrato in fig.4.

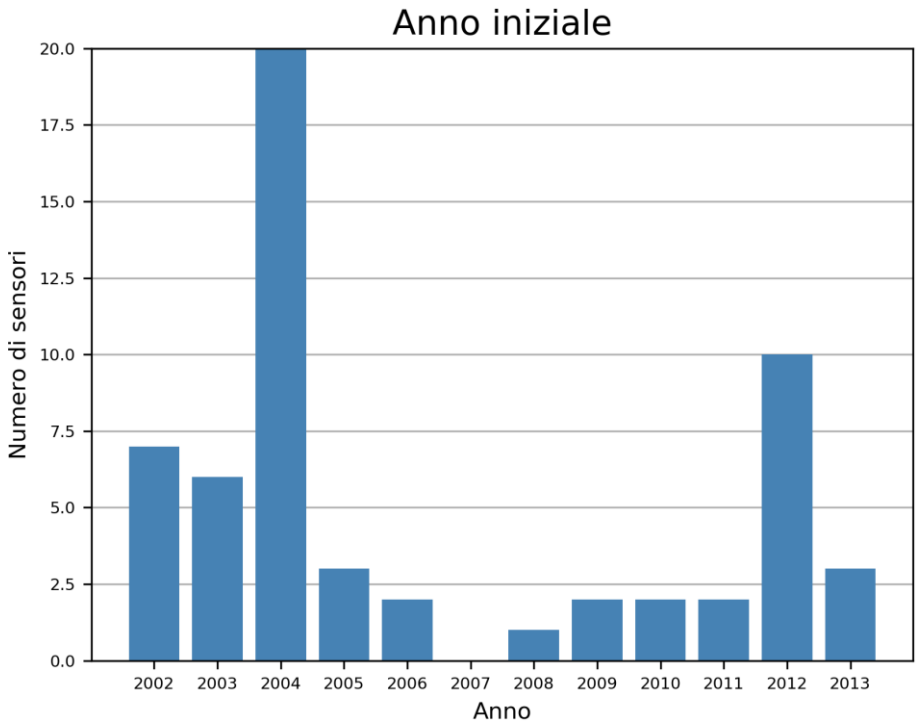

Fig. 4 - Distribuzione degli anni di entrata in funzione dei sensori di pressione.

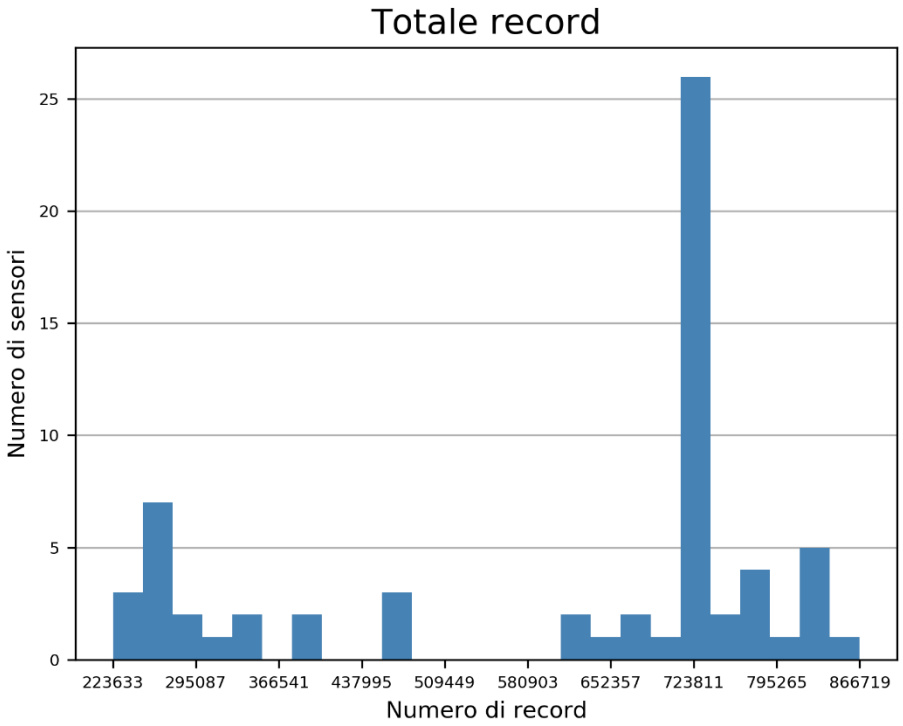

Fig. 5 - Istogramma del numero di record in ogni sensore.

La lunghezza delle serie temporali di pressione varia tra i circa 220.000 record per i sensori più "giovani" e 870.000 record per i sensori più longevi. È interessante notare che nel 2007 non è entrato in funzione nessun sensore, mentre nel 2004 ne sono stati attivati 20: questo comporta che, come si vede in fig.5, la distribuzione delle lunghezze delle serie temporali è unimodale attorno al valore 723.800. Questo numero è molto vicino a 735.840, che sarebbe il numero di record nominali registrati dal 1° gennaio 2004 al 1° gennaio 2018. In realtà non tutti i sensori sono stati attivati il 1° gennaio, quindi tra i due valori di lunghezza indicati sopra vi è una leggera differenza.

Per quanto riguarda i record mancanti, prima di analizzare nel dettaglio i singoli sensori è possibile fare alcune considerazioni generali, che ci serviranno per definire meglio i successivi passi di elaborazione. Anzitutto occorre determinare la quantità di record mancanti in ogni sensore: questo valore è utile a distinguere se una determinata serie di dati potrà essere utilizzata o no. Ad esempio, è improbabile che una serie con più dell'80% di valori mancanti possa essere interpolata con risultati soddisfacenti (circostanza che non si riscontra mai, come mostrato in fig.6). Tuttavia, potrebbero essere presenti delle regolarità o periodicità nei record mancanti: se la percentuale di record mancanti ha un valore particolare come 50% o 66%, è probabile che il dato sia stato acquisito a intervalli di tempo maggiori dei 10 minuti, ad esempio registrando una misura ogni 20 o 30 minuti, e quindi la serie ha un dato ogni 2 o 3 record. In questo caso si può supporre che l'interpolazione darà risultati sufficientemente accurati anche se il numero di osservazioni è pari o inferiore al numero di record mancanti ovvero al numero di punti di predizione. Di conseguenza, per decidere quali serie temporali utilizzare e quali eliminare non si può utilizzare solo la percentuale di record mancanti, ma occorre operare un'analisi più approfondita: lo scarto di una serie comporta una perdita di informazioni, quindi nessuna serie è stata scartata osservando la sola percentuale di record mancanti.

Nel nostro caso abbiamo deciso di creare, per ogni sensore, un vettore contenente il numero di no-data per ogni 'buco', ovvero la lunghezza dei buchi, e di calcolarne la distribuzione. Se la frequenza si concentra su valori di lunghezza pari a uno o due e la percentuale dei dati mancanti è prossima al 50% o al 66% allora si tratta di sensori vecchi con intervalli di campionamento maggiore dei 10 minuti (20 o 30 minuti, rispettivamente). Se invece la percentuale di vuoti è inferiore al 50% può darsi che la serie sia in parte costituita dalle osservazioni di un sensore di vecchia generazione (con intervallo di campionamento inferiore ai 10 minuti) poi sostituito con di nuova generazione con campionamento ai 10 minuti. In tab.1 sono riportate le frequenze, ovvero il numero di sensori, con una distribuzione di buchi concentrata intorno ai valori 1, 2 e 5.

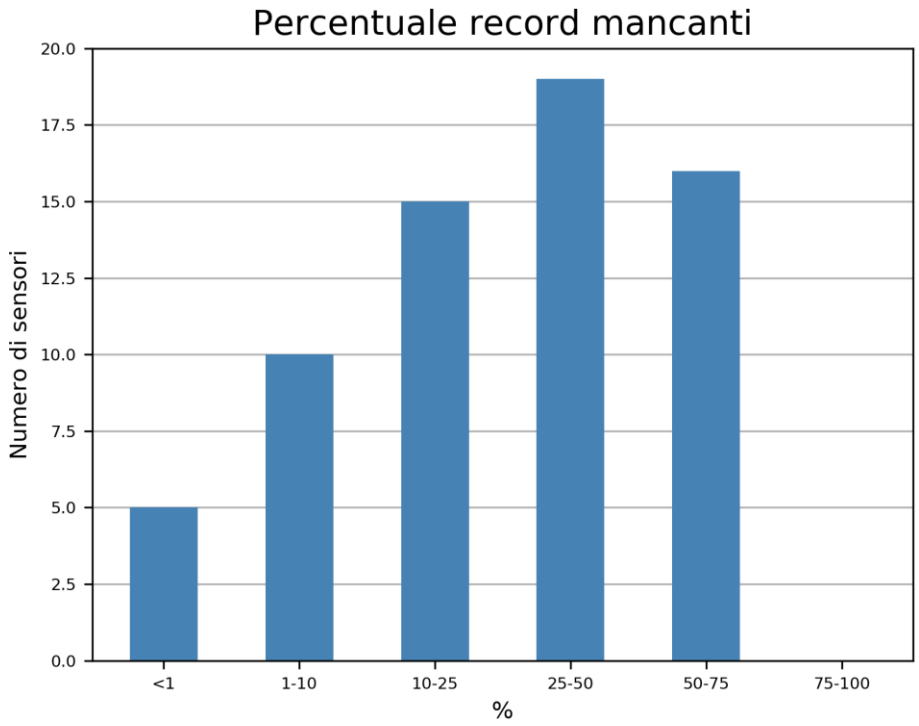

Fig. 6 - Distribuzione delle percentuali di record nulli in ogni sensore.

Mediana	N° sensori
1	13
2	34
5	18

Tab. 1 - Numero di sensori per ogni mediana.

Tutti i sensori con mediana pari a 1 hanno una percentuale di vuoti molto inferiore al 50% (tutti inferiori al 15%), anche se si restringe l'intervallo di valutazione della ampiezza dei buchi a un solo anno o a sei mesi, si può supporre quindi che le rispettive distribuzioni dei vuoti siano casuali, non legate alla diversa modalità di campionamento. I sensori con mediana uguale a 2 sono invece quelli che hanno una misura ogni tre, ovvero un intervallo di campionamento di mezz'ora. I sensori con mediana uguale a 5, hanno un dato ogni 6 ovvero un campionamento orario.

Viene ora mostrata come esempio l'analisi dei dati rilevati da un sensore di pressione. Il sensore in esame ha codice identificativo "9060", si trova a Molteno (LC), con coordinate (5069985N, 523872E). Le statistiche descrittive di base di questa serie temporale sono contenute in tab.2, mentre in tab.3 sono evidenziate alcune informazioni aggiuntive.

Numerosità	561.835
Media	984,66 hPa
Dev. St.	7,59 hPa
Min	943,7 hPa
25%	980,3 hPa
50%	984,8 hPa
75%	989,1 hPa
Max	1016,7 hPa

Tab. 2 - Principali statistiche descrittive della serie temporale.

Anni	14
Totale record	735.840
Numerosità	561.835
Numero NaN	174.005
Percentuale NaN	23,65%

Tab. 3 - Altre informazioni sul sensore '9060'.

Nei grafici seguenti vengono mostrati il primo e l'ultimo mese di osservazioni registrate dal sensore. Le differenze tra i grafici in fig.7 e fig.8 mostrano che probabilmente il sensore è stato sostituito nel corso degli anni con uno più sensibile e con maggiore risoluzione temporale (in grado cioè di registrare la pressione al decimo di ettopascal). Il grafico del primo mese di osservazioni, infatti, mostra evidenti discontinuità sia sulle ascisse che sulle ordinate; esse sono invece assenti nel grafico dell'ultimo mese. Andando a leggere le prime 10 osservazioni della serie e le ultime 10, si possono notare le cause di queste differenze.

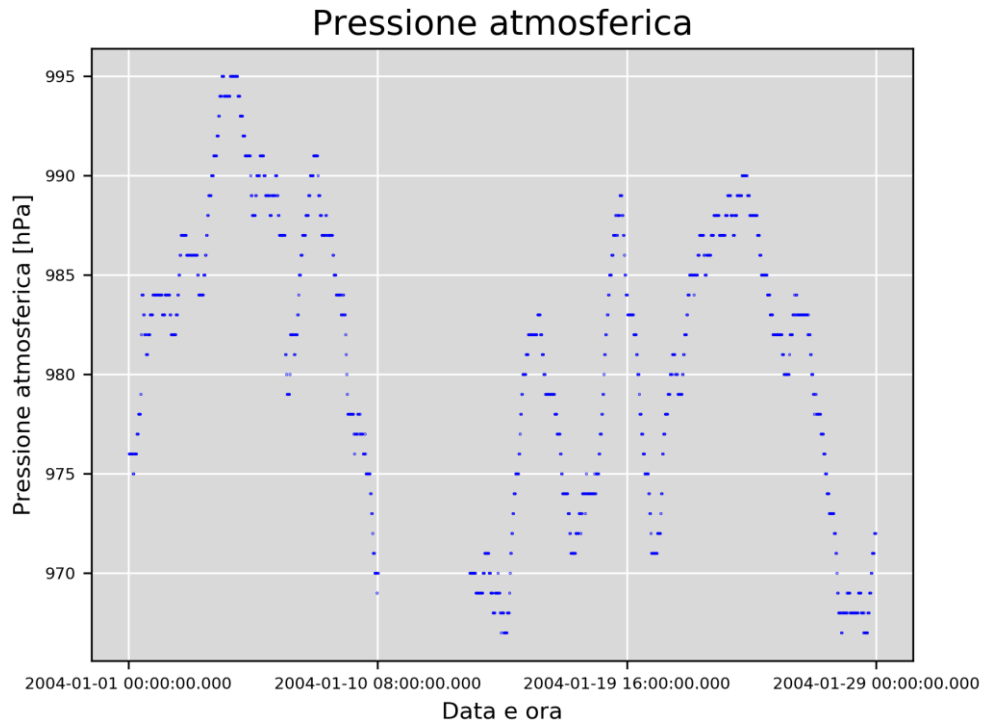

Fig. 7 - Primo mese di osservazioni.

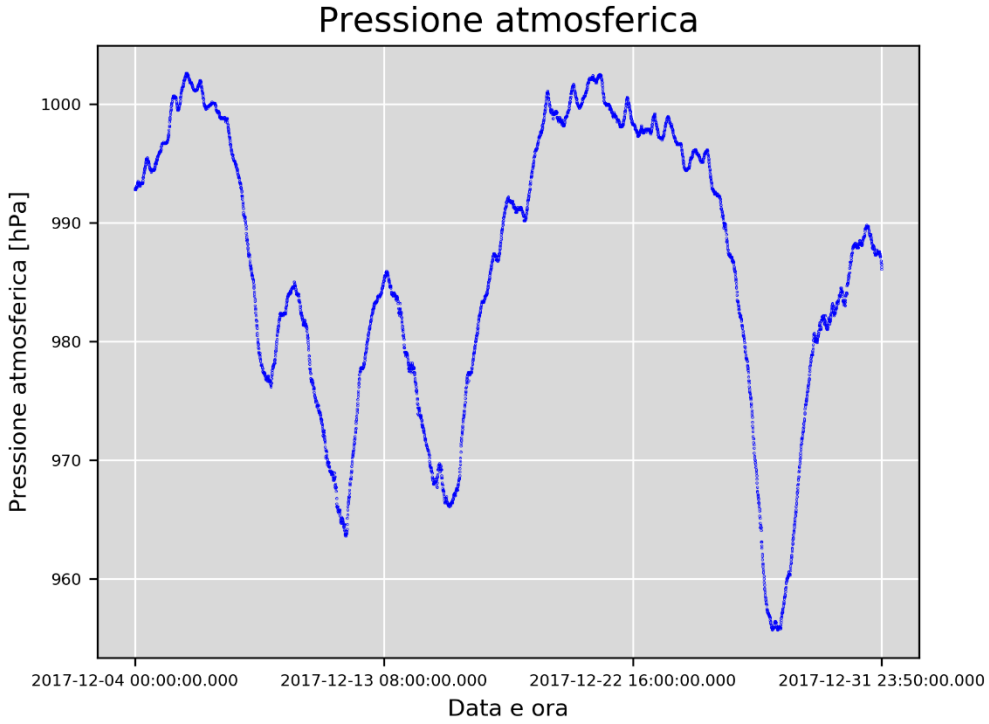

Fig. 8 - Ultimo mese di osservazioni.

Data e ora	Pressione [hPa]
2004-01-01 00:00:00.000	NaN
2004-01-01 00:10:00.000	NaN
2004-01-01 00:20:00.000	NaN
2004-01-01 00:30:00.000	976,0
2004-01-01 00:40:00.000	NaN
2004-01-01 00:50:00.000	NaN
2004-01-01 01:00:00.000	976,0
2004-01-01 01:10:00.000	NaN
2004-01-01 01:20:00.000	NaN
2004-01-01 01:30:00.000	976,0

Tab. 4 - Primi dieci record.

Data e ora	Pressione [hPa]
2017-12-31 22:20:00.000	987,2
2017-12-31 22:30:00.000	987,2
2017-12-31 22:40:00.000	987,1
2017-12-31 22:50:00.000	987,1
2017-12-31 23:00:00.000	986,8
2017-12-31 23:10:00.000	986,8
2017-12-31 23:20:00.000	986,6
2017-12-31 23:30:00.000	986,5
2017-12-31 23:40:00.000	986,3
2017-12-31 23:50:00.000	986,1

Tab. 5 - Ultimi dieci record.

La differenza più evidente è la discretizzazione temporale: in tab.4 si nota che le misure vengono effettuate ogni 30 minuti, allo scoccare dell'ora e della mezz'ora. Di conseguenza, la serie ha un record valido ogni due e questo risulta in una discontinuità molto regolare lungo l'asse delle ascisse. In secondo luogo, è da notare che in tab.5 le misure hanno una precisione che arriva al decimo di ettopascal, mentre in tab.3 la cifra dopo la virgola è sempre zero: questo si traduce graficamente in una discontinuità lungo l'asse delle ordinate, in quanto tutti i punti si trovano su valori interi.

A questo punto si procede con l'analisi dei vuoti. Il primo passo è la creazione di un vettore contenente le grandezze di ogni buco, ovvero la lunghezza di tutti gli intervalli di record nulli consecutivi. Di tale vettore vengono fornite le principali statistiche descrittive in tab.6. Si può notare che nella serie analizzata sono presenti 60.892 intervalli di record nulli; la lunghezza media degli intervalli è 2,86 record mentre il valore della mediana, oltre a quello degli altri due quartili, è pari a 2: questo significa che i dati da interpolare hanno una struttura molto solida. Un'ulteriore prova è costituita dalla distribuzione delle frequenze di queste lunghezze, mostrate in fig.9: una percentuale superiore al 99% dei buchi ha lunghezza pari a 2.

Numerosità	60.892
Media	2,86
Dev. St.	145,62
Min	1
25%	2
50%	2
75%	2
Max	35.423

Tab. 6 - Statistiche descrittive delle lunghezze di buchi della serie in esame.

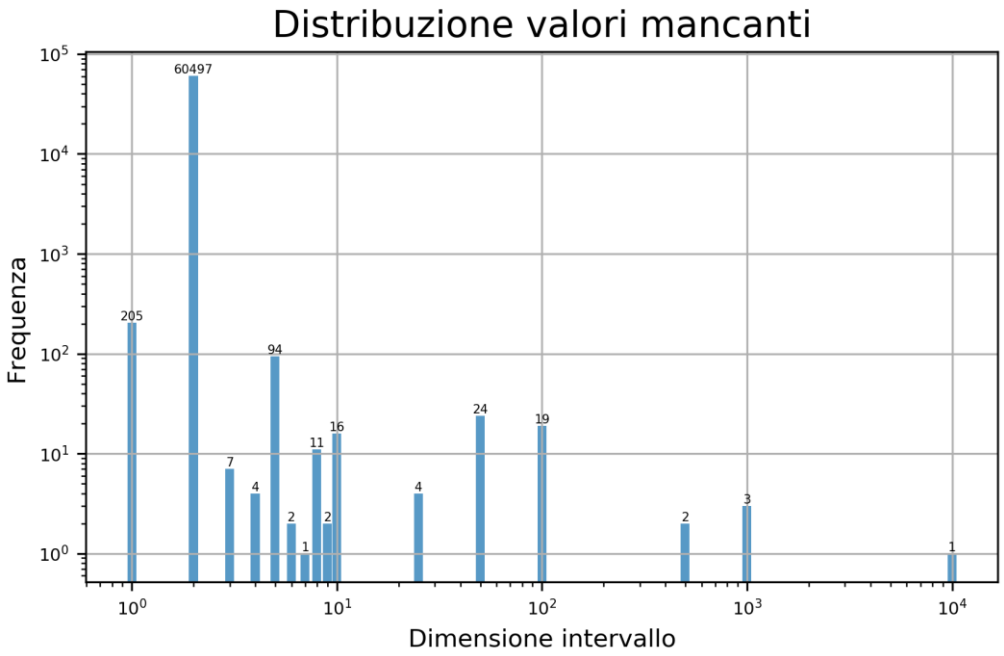

Fig. 9 - Distribuzione delle lunghezze degli intervalli di record nulli; notare che gli assi sono in scala logaritmica.

Il dato relativo al numero di intervalli di lunghezza 2 è preponderante e tende a oscurare ogni altra informazione contenuta nel vettore in analisi: di conseguenza i conti successivi verranno effettuati dopo aver rimosso tale valore. Oltre alle informazioni già ricavate, si può notare che una percentuale pari all'83% degli intervalli considerati ha lunghezza inferiore a 10. Un ulteriore 16%, inoltre, ha lunghezza compresa tra 10 e 1000, pari a circa una settimana di osservazioni, ovvero la lunghezza massima di auto-correlazione per questo sensore[2]. Solamente 4 intervalli risultano di lunghezza superiore a 1000: 3 di questi misurano tra 2000 e 5000 record, mentre l'intervallo di dimensioni maggiori, come evidenziato in tab.5, è lungo 35.423 record, ovvero circa 8 mesi di osservazioni mancanti. Quest'ultimo intervallo da solo contiene circa il 20% dei record nulli in tutta la serie temporale e risulta troppo lungo per essere interpolato con sufficiente accuratezza, quindi verrà lasciato vuoto; oltre a questo, anche gli intervalli compresi tra 2000 e 5000 record non verranno interpolati. Tutti gli altri record nulli verranno, invece, sostituiti da valori predetti.

Questo tipo di analisi è stata svolta per ogni sensore di pressione e verrà ripetuta anche per i sensori di temperatura. Le considerazioni sull'interpolazione dei record mancanti verranno riprese nel capitolo 4, poiché verranno utilizzate per definire la procedura di interpolazione.

[2] si discuterà di ciò in maniera più approfondita nel capitolo 4.

2.2. Temperatura

In questo paragrafo verranno mostrate alcune informazioni sui 162 sensori che compongono la rete di rilevamento di temperatura di ARPA Lombardia. Nello specifico, sono analizzati: passi temporali, date di inizio e fine delle misurazioni, numero di record totali e numero di record mancanti. Inoltre, verranno mostrate le statistiche descrittive complete di un sensore come esempio. L'unità di misura in cui sono espressi tutti i dati è gradi Celsius [°C], tranne dove indicato diversamente.

Come per la pressione atmosferica, il passo temporale delle serie di temperatura e la data dell'ultimo record sono uguali per tutti i sensori. In particolare, il passo temporale è pari a 10 minuti e la data dell'ultima rilevazione è il primo gennaio 2018 alle 00:00. I sensori in esame sono stati attivati tra il 2001 e il 2016: nel grafico in fig.10 si nota come la maggior parte dei sensori sia attiva da prima del 2007. Di conseguenza, il grafico in fig.11 evidenzia come la lunghezza delle serie sia clusterizzata attorno al valore 750.000. I due grafici sono speculari.

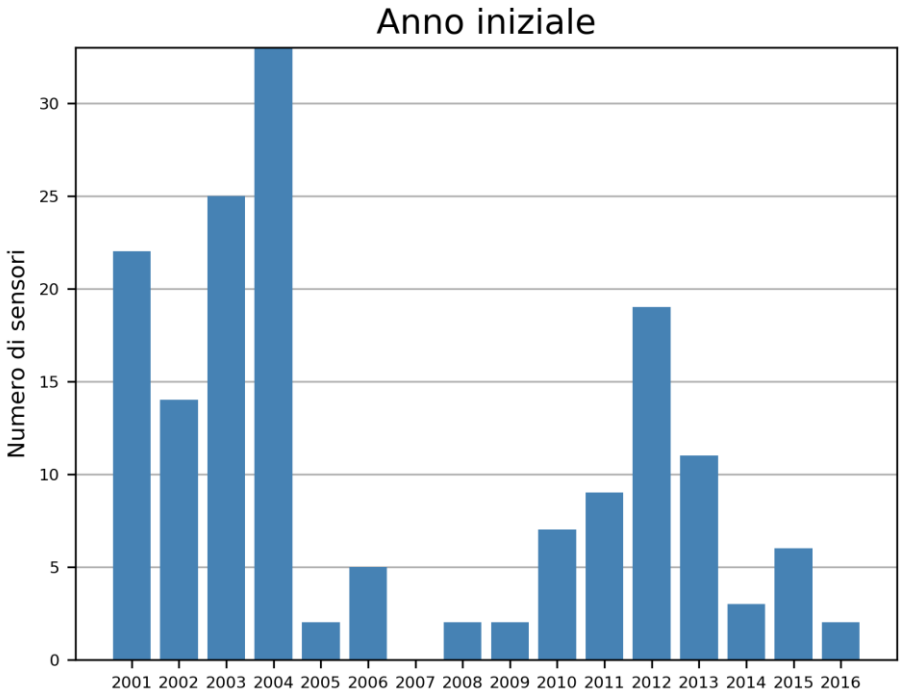

Fig. 10 - Distribuzione degli anni iniziali dei sensori di temperatura.

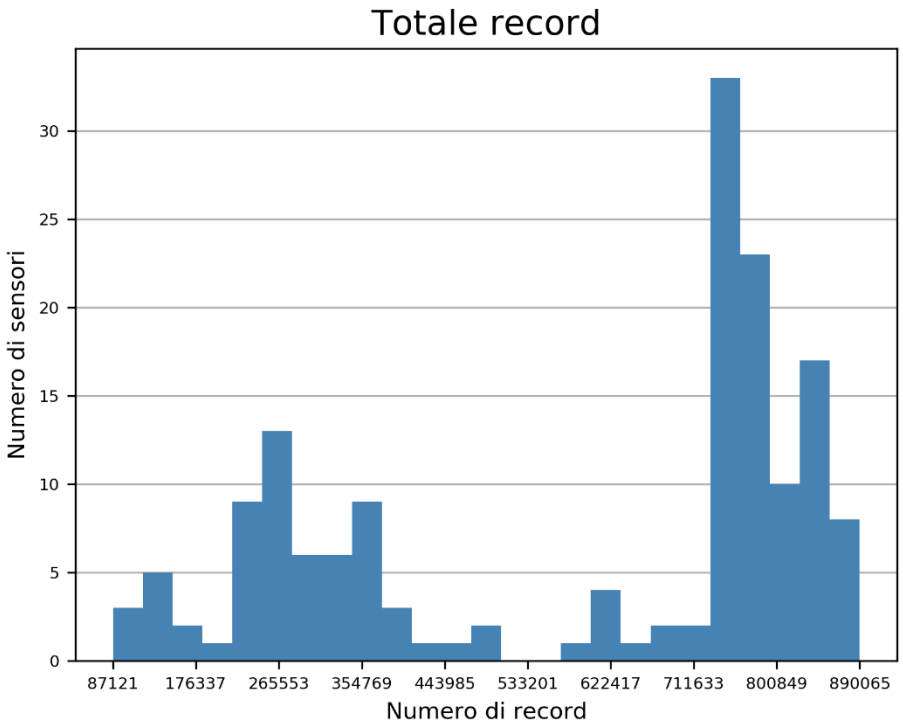

Fig. 11 - Istogramma del numero di record in ogni sensore.

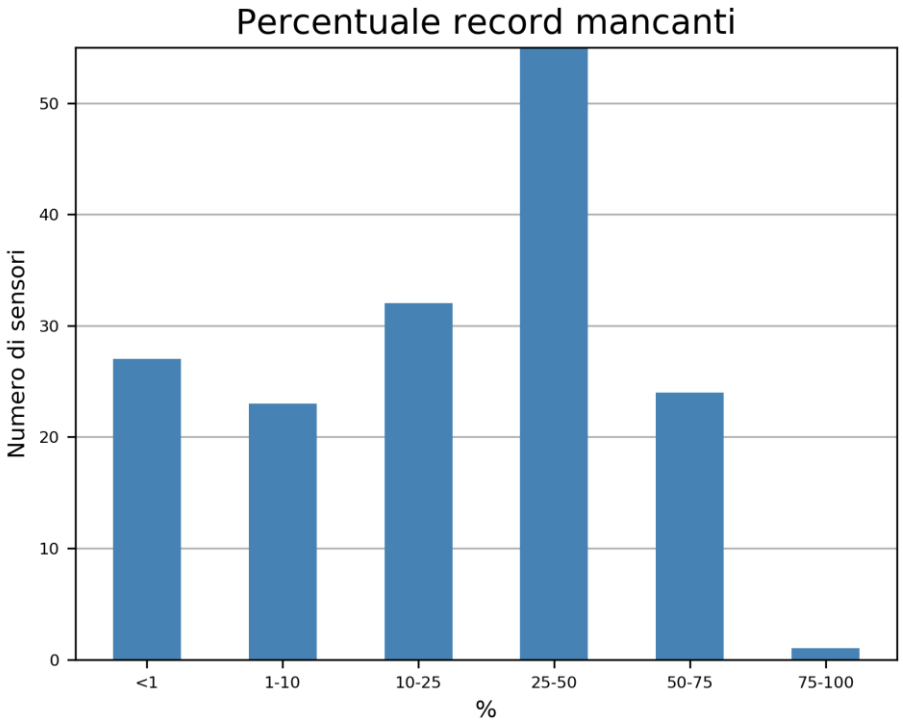

Fig. 12 - Distribuzione delle percentuali di record nulli in ogni sensore.

Per quanto riguarda le percentuali di valori mancanti in ciascuna serie, il grafico in fig.12 mostra alcune analogie con quello in fig.6 del paragrafo precedente. In particolare si nota che entrambi i grafici hanno un andamento unimodale centrato sull'intervallo 25-50% e circa il 30% dei sensori appartiene a questa categoria per entrambe le variabili considerate. Tuttavia, mentre nel caso della pressione la maggior parte dei sensori ha una percentuale di record mancanti tra il 10% e il 75%, nel caso della temperatura si nota una frazione notevole di sensori con meno dell'1% di vuoti: circa il 17% dei sensori rientra in questa categoria.

Anche in questo caso è stata eseguita un'analisi più approfondita dei record mancanti. Come mostrato nel paragrafo precedente, infatti, la sola percentuale di vuoti non fornisce una caratterizzazione adeguata di eventuali pattern presenti nelle serie temporali. Nel caso della temperatura, metà dei sensori ha mediana pari a 2, ovvero una rilevazione ogni mezz'ora; il 30% dei sensori ha mediana pari a 1 e il restante 20% ha un record ogni 6 epoche. Anche nel caso delle serie temporali di temperatura, quindi, i vuoti sono ben strutturati e dipendono dalle caratteristiche dei rilevatori.

Mediana	N° sensori
1	51
2	78
5	33

Tab. 7 - Numero di sensori per ogni mediana.

Viene ora mostrata come esempio l'analisi dei dati rilevati da un sensore di pressione. Il sensore in esame ha codice identificativo "8229", che si trova a Varese, con coordinate (5075452N, 486300E). Le statistiche descrittive di base di questa serie temporale sono contenute in tab.8, mentre in tab.9 sono evidenziate alcune informazioni aggiuntive.

Numerosità	469.869
Media	13,1 °C
Dev. St.	8,25 °C
Min	-12,7 °C
25%	6,5 °C
50%	12,8 °C
75%	19,4 °C
Max	36,2 °C

Tab. 8 - Principali statistiche descrittive della serie temporale.

Anni	15
Totale record	754.128
Numerosità	469.869
Numero NaN	284.259
Percentuale NaN	37,69%

Tab. 9 - Altre informazioni sul sensore '8229'.

I grafici seguenti mostrano due mesi di osservazioni di temperatura: i due periodi considerati sono di 4 settimane ciascuno, quindi 28 giorni, e sono stati scelti all'inizio e alla fine della serie in esame. L'intervallo temporale del primo grafico è tra il 24 settembre 2003 e il 22 ottobre dello stesso anno; nel secondo grafico sono rappresentate le osservazioni tra il 4 e il 31 dicembre 2017 (si tratta delle ultime 4 settimane di osservazioni).

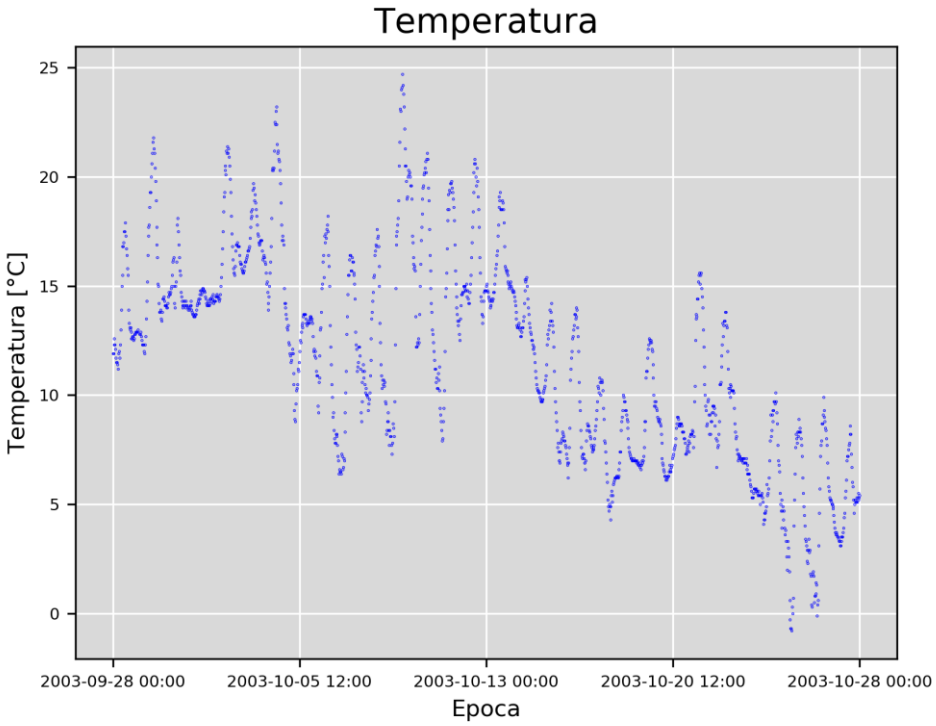

Fig. 13 – Osservazioni tra il 28 settembre e il 28 ottobre 2003.

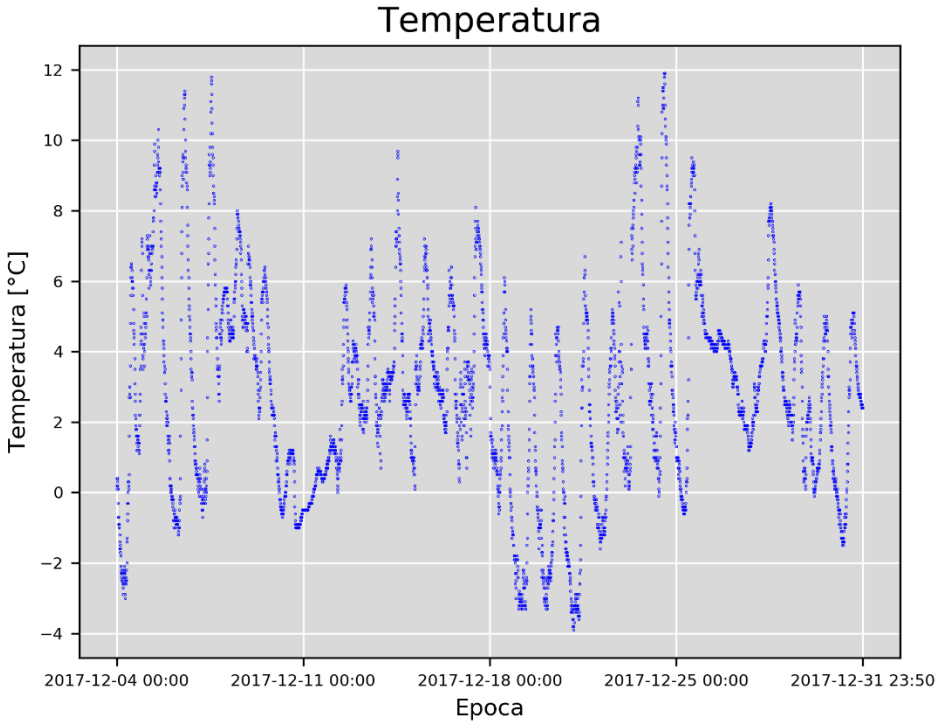

Fig. 14 - Ultimo mese di osservazioni del sensore '8229'.

Osservando i due grafici si nota che la densità dei punti è diversa: il secondo grafico contiene più punti del primo. Come nel caso del sensore di pressione atmosferica analizzato precedentemente, il passo temporale tra la registrazione di due misure successive è cambiata nel tempo. Quando il sensore è stato attivato, tale intervallo era di 30 minuti, come si può vedere dai dati in tab.10; ciò è coerente con i dati in tab.12, che mostrano le principali statistiche dei record nulli: la lunghezza mediana dei buchi della serie è pari a 2. Successivamente il sensore è stato modificato e il passo temporale è diminuito fino a 10 minuti. Si può notare, inoltre, che l'accuratezza del sensore non è cambiata nel corso degli anni: sia nel 2003 che nel 2017 la precisione della misura raggiunge il decimo di grado.

Data e ora	Temperatura [°C]
2003-09-24 00:00:00.000	15,7
2003-09-24 00:10:00.000	NaN
2003-09-24 00:20:00.000	NaN
2003-09-24 00:30:00.000	15,6
2003-09-24 00:40:00.000	NaN
2003-09-24 00:50:00.000	NaN
2003-09-24 01:00:00.000	15,5
2003-09-24 01:10:00.000	NaN
2003-09-24 01:20:00.000	NaN
2003-09-24 01:30:00.000	15,4

Tab. 10 - Prime dieci osservazioni del sensore '8229'.

Data e ora	Temperatura [°C]
2017-12-31 22:20:00.000	2,5
2017-12-31 22:30:00.000	2,5
2017-12-31 22:40:00.000	2,4
2017-12-31 22:50:00.000	2,5
2017-12-31 23:00:00.000	2,5
2017-12-31 23:10:00.000	2,5
2017-12-31 23:20:00.000	2,4
2017-12-31 23:30:00.000	2,4
2017-12-31 23:40:00.000	2,4
2017-12-31 23:50:00.000	2,4

Tab. 11 - Ultime dieci osservazioni.

Per completare l'analisi della serie temporale occorre caratterizzare gli intervalli di record nulli, calcolandone le principali statistiche descrittive.

Numerosità	284.259
Media	2,02
Dev. St.	2,47
Min	1
25%	2
50%	2
75%	2
Max	650

Tab. 12 - Principali statistiche descrittive del vettore di record nulli.

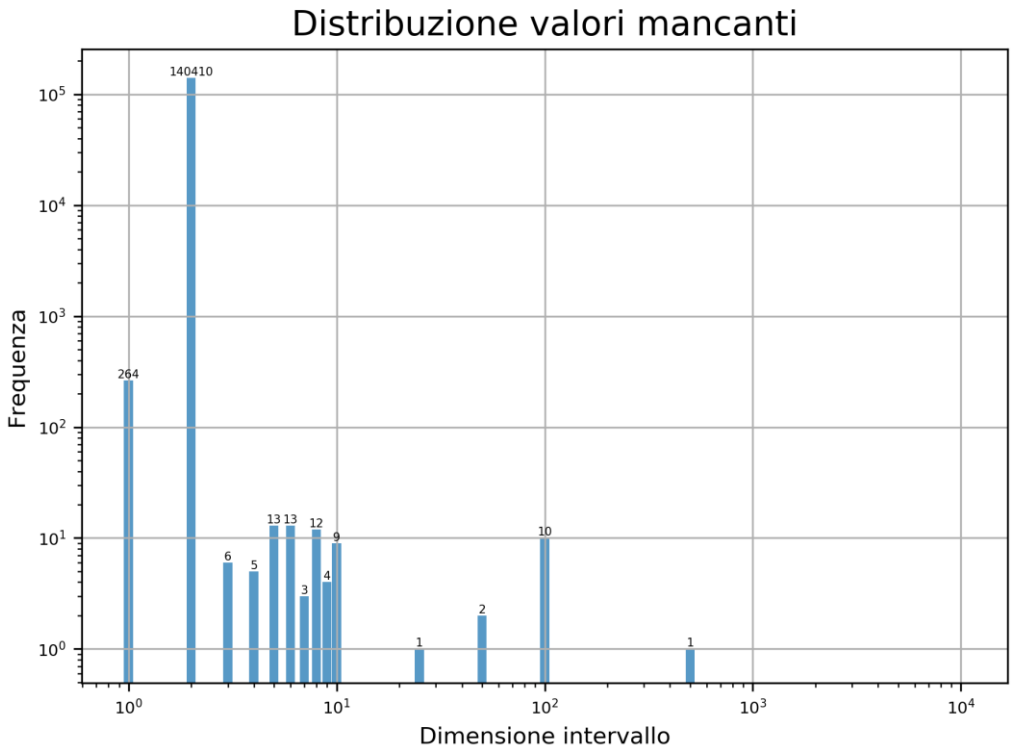

Fig. 15 - Frequenze degli intervalli di record mancanti; assi in scala logaritmica.

Il grafico in fig.15 mostra la distribuzione delle lunghezze degli intervalli vuoti. La seconda barra da sinistra mostra la totale dominanza degli intervalli di lunghezza 2, che da soli coprono più del 99% dei vuoti. Eliminando questo valore e ricalcolando le percentuali si nota che circa il 77% dei vuoti restanti sono isolati, mentre un ulteriore 16% dei buchi rimane comunque sotto la lunghezza di 10 epoche. Solamente 23 intervalli risultano di lunghezza maggiore di 10, rimanendo al di sotto di 1000 istanti temporali, distanza oltre la quale decade l'autocorrelazione del segnale[3].

[3] si discuterà di ciò in maniera più approfondita nel capitolo 4.

3. METODI

3.1. Kriging

Il kriging è una procedura di interpolazione sviluppata in ambito geostatistico a partire dagli anni ’50 del secolo scorso.

Il metodo modellizza le osservazioni come un campione discreto di una realizzazione di un processo o campo stocastico e utilizza tale ipotesi per predire il segnale o il campo in uno o più epoche o posizioni come una combinazione lineare delle osservazioni i cui coefficienti vengono calcolati imponendo la minimizzazione della varianza dell’errore di predizione. I coefficienti così determinati dipendono dalla covarianza tra le osservazioni e tra queste e i punti di predizione. Tale covarianza viene altresì stimata a partire dalle osservazioni utilizzando, ove possibile, l’ipotesi aggiuntiva di stazionarietà o di omogeneità e isotropia rispettivamente del processo o del campo stocastico considerato.

Di seguito vengono riportate in dettaglio le ipotesi e le espressioni necessarie all’implementazione dell’algoritmo. Considereremo il caso specifico di processo stocastico debolmente stazionario, ovvero di una famiglia di variabili casuali indicizzata dal parametro tempo che varia in maniera continua su R, che abbiano tutte stessa media e stessa varianza e covarianza tra coppie di variabili X(*t*) e X(*t’*) dipendente solo dalla differenza in modulo tra le epoche *t* e *t’*: $\tau = |t - t'|$.

Si consideri il problema della predizione di un valore incognito di tale processo stocastico stazionario u(t), a partire da N osservazioni Y di una realizzazione del processo. Si ipotizzi che il processo abbia media costante nell’area di interesse e che le osservazioni siano affette da un rumore v_i.

$$i = 1, 2, \dots, N$$

$$Y_i = u(t_i) + v_i$$

$$E[u(t)] = \bar{u}, \quad \forall t$$

$$E[u(t) * u(t')] - \bar{u}^2 = C(|t - t'|), \quad \forall t, t'$$

Si ipotizzi, inoltre, che v_i sia un rumore bianco, ovvero che valgano le equazioni:

$$E[v_i] = 0 \quad \forall i$$

$$C_{vv} = E\left[\underline{v}\underline{v}'\right] = \sigma_v^2 I_N$$

$$\sigma_v^2 = E[v_i^2] \; \forall i$$

dove $E[\cdot]$ indica l'operatore di valore atteso, σ_v^2 è la varianza dell'errore e C_{vv} è la sua matrice di covarianza. Inoltre, si suppone che il processo e il rumore siano incorrelati tra loro $E[u(t) * v(t')] = 0, \ \forall t, t'$.

Per risolvere il problema occorre definire una funzione che descriva l'auto-correlazione tra coppie di variabili casuali per qualunque distanza temporale τ. Nel caso di processo stocastico stazionario a media non nulla questa può essere derivata dal semivariogramma, definito come:

$$\gamma(t, t') = \frac{1}{2} E[(u(t) - u(t'))^2]$$

Questa funzione può essere ricavata empiricamente dai dati e senza conoscere il valore della media del processo. Il semivariogramma gode di diverse importanti proprietà:

- il valore nell'origine è uguale a 0: $\gamma(t, t) = 0 \quad \forall t$
- è una funzione simmetrica: $\gamma(t, t') = \gamma(t', t) \quad \forall t, t'$
- è condizionalmente definito negativo
- se il processo stocastico è spazialmente incorrelato, il semivariogramma è costante.

Inoltre, se la funzione di covarianza del processo esiste, essa è legata al semivariogramma tramite l'equazione:

$$2\gamma(t, t') = C(t, t) + C(t', t') - 2C(t, t').$$

Se il processo è stazionario, il semivariogramma dipende da $\tau = |t - t'|$ e l'ultima equazione diventa:

$$\gamma(\tau) = C(0) - C(\tau)$$

Da queste proprietà segue che non tutte le funzioni possono essere utilizzate in questo algoritmo. In particolare, una funzione $\gamma(t, t')$ è un semivariogramma se e solo se ogni matrice

$$\Gamma \equiv [\gamma(t_i, t_k)], \qquad i, k = 1, 2, \dots, N$$

è condizionalmente definita negativa, ovvero se $\forall\{t_1, \dots, t_N\},\ \forall N$, vale:

$$\underline{\lambda}^T \Gamma \underline{\lambda} \leq 0,$$

$$\forall \underline{\lambda} \in R^N \neq 0 \text{ per cui vale la condizione}$$

$$\underline{e}^T \underline{\lambda} \equiv 0,$$

dove *e* è il vettore N-dimensionale composto da tutti 1.

È importante notare che se il segnale, stazionario, è combinazione lineare di due processi stazionari indipendenti

$$y(t) = u(t) + x(t)$$

vale la relazione

$$\gamma_y(\tau) = \gamma_u(\tau) + \gamma_x(\tau).$$

Questo comporta che è possibile ottenere un semivariogramma da una combinazione lineare a coefficienti positivi di funzioni ammissibili.

La stima empirica del semivariogramma viene effettuata interpolando alcuni valori disposti su una griglia con k nodi e ampiezza Δ. La formula per il calcolo è:

$$\hat{\gamma}(k\Delta) + \hat{\sigma}_v^2 = \frac{1}{2}\frac{1}{N_{k\Delta}} \sum_{k\Delta - \frac{\Delta}{2} < t_i - t_j < k\Delta + \frac{\Delta}{2}} [y(t_i) - y(t_j)]^2$$

Lo stimatore empirico del variogramma restituisce una stima del variogramma traslata verticalmente per una costante pari alla varianza del noise delle osservazioni. Una stima di tale varianza può essere ricavata imponendo che nell'origine il variogramma sia nullo.

Con la formula appena citata si ricavano i valori del semivariogramma empirico per un numero k di distanze separate da una lunghezza pari a Δ. Questi punti vanno poi interpolati da una funzione ammissibile; alcuni esempi di semivariogramma ammissibile sono:

- esponenziale: $\gamma(\tau) = A(1 - e^{\alpha\tau})$
- normale: $\gamma(\tau) = A(1 - e^{\alpha\tau^2})$
- logaritmico: $\gamma(\tau) = \alpha \log(1 + \frac{\tau}{a})$

- lineare: $\gamma(\tau) = A + \alpha\tau$

- esponenziale con esponente positivo minore di 1: $\gamma(\tau) = A\tau^{\alpha}$, con $0 < \alpha < 1$

- sferico: $\gamma(\tau) = \begin{cases} 1 - \frac{3}{2}\left(\frac{\tau}{\alpha}\right) + \frac{1}{2}\left(\frac{\tau}{\alpha}\right)^3 & 0 \leq \tau \leq \alpha \\ 0 & \tau \geq \alpha \end{cases}$

Per $\tau \to +\infty$ alcune tra queste funzioni giungono a saturazione sono quelle relative a processi stocastici che ammettono funzione di covarianza, altre divergono. Nella caratterizzazione di un semivariogramma sono importanti, oltre al già citato nugget, due elementi: il "sill" è il valore di $\gamma(\tau)$ a saturazione, mentre il "range" indica la distanza alla quale le differenze tra il sill e il semivariogramma diventano trascurabili. È importante notare che risultano semivariogrammi ammissibili anche funzioni che non raggiungono saturazione, come ad esempio quella lieare: di conseguenza, la classe di segnali su cui è applicabile il kriging non è limitata a quelli con covarianza finita.

Dopo aver definito un'opportuna misura dell'auto-correlazione del processo stocastico e la sua derivazione, vediamo più in dettaglio l'algoritmo di predizione. L'obiettivo è trovare un predittore lineare del tipo:

$$\hat{u}(t) = \underline{\lambda}^T \underline{Y} = \underline{\lambda}^T \underline{u} + \underline{\lambda}^T \underline{\nu}$$

Tra tutti gli stimatori lineari di u(t) senza bias, quello che minimizza la varianza dell'errore di predizione si ottiene come soluzione del seguente sistema:

$$\begin{cases} (\Gamma - C_{\nu\nu})\underline{\lambda} + \alpha\underline{e} = \underline{\gamma}(t) \\ \underline{e}^T\underline{\lambda} = 1 \end{cases}$$

dove:

- $\Gamma = \{\gamma(t_i - t_k)\}$ è la matrice con le semivarianze dei punti noti

- $\underline{\gamma} = \{\gamma(t - t_k)\}$ è il vettore con le semivarianze tra i punti noti e i punti in cui si vuole effettuare la predizione

- $\mathcal{E}^2 = E[(u(t) - \hat{u}(t))^2]$ è lo scarto quadratico medio dell'errore di predizione

- $\alpha, \underline{\lambda}$ sono le incognite del sistema.

La stima della varianza dell'errore di predizione vale:

$$\mathcal{E}^2 = \alpha + \underline{\lambda}^T \underline{\gamma}(t).$$

Il sistema precedente può essere trasformato in forma matriciale e riscritto in questo modo:

$$\begin{vmatrix} \Gamma - C_{vv} & \underline{e} \\ \underline{e}^T & 0 \end{vmatrix} \begin{vmatrix} \underline{\lambda} \\ \alpha \end{vmatrix} = \begin{vmatrix} \underline{\gamma}(t) \\ 1 \end{vmatrix}$$

Nel caso in cui si voglia operare un'interpolazione esatta, è sufficiente imporre che la varianza dell'errore di osservazione σ_v^2 sia uguale a zero: la matrice di covarianza dell'errore diventa così nulla e non viene modificato il valore della misura nei punti noti. In questo caso specifico, inoltre, si può dimostrare che se il semivariogramma è della forma,

$$\gamma(\tau) = A * f(\tau)$$

il parametro moltiplicativo A può essere posto uguale a 1 senza perdita di informazione. Considerando, quindi, $C_{vv} = 0$ ed esplicitando il parametro A, il sistema da risolvere diventa:

$$\begin{cases} A\Gamma\underline{\lambda} + \alpha\underline{e} = A\underline{\gamma}(t) \\ \underline{e}^T\underline{\lambda} = 1 \end{cases}$$

dividendo tutto per A, ponendo $\beta = \frac{\alpha}{A}$ e sostituendo, si ottiene:

$$\begin{cases} \Gamma\underline{\lambda} + \beta\underline{e} = \underline{\gamma}(t) \\ \underline{e}^T\underline{\lambda} = 1 \end{cases}$$

Ciò comporta che lo stimatore $\hat{u}(t) = \underline{\lambda}^T\underline{Y}$ è indipendente dal parametro A.

3.2. Radial Basis Function Network

Le RBFN sono modelli sviluppati in ambito statistico per funzionare come approssimatori universali. Esse sono delle particolari reti neurali artificiali che utilizzano delle funzioni a base radiale come funzioni di attivazione. Solitamente queste reti sono costituite da tre layer: uno per gli input, uno per le funzioni di attivazione e uno per l'output, disposti come in figura. Ogni layer è completamente connesso al successivo e ogni collegamento è moltiplicato per un peso.

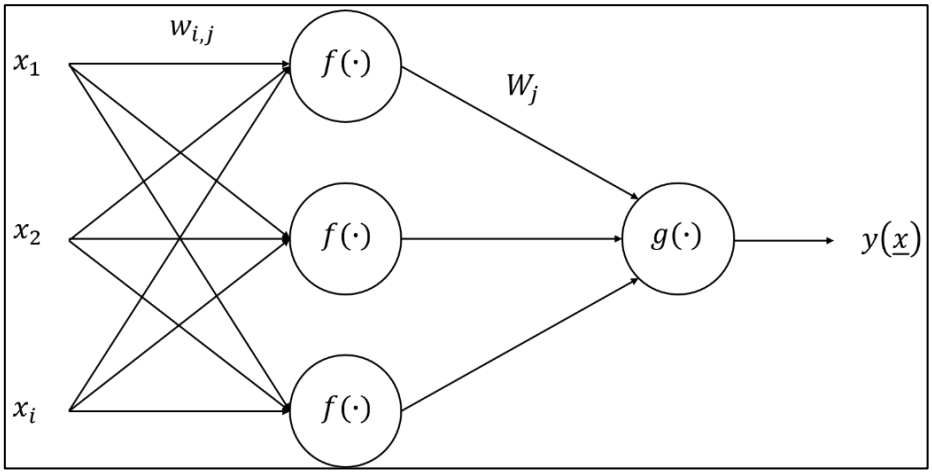

Fig. 16 - Schema generico di una RBF-Net.

Nella figura compaiono tutti gli elementi generici che caratterizzano una RBFN:

- $x_1, \dots, x_i$ sono i valori in input della rete
- $w_{i,j}$ sono i pesi di ciascun arco: rappresentano il peso dell'input i-esimo rispetto al nodo j-esimo
- $f(\cdot)$ sono le funzioni di attivazione, di cui discuteremo in seguito
- W_j sono i pesi del nodo j-esimo rispetto al nodo dell'output
- $g(\cdot)$ è la funzione di aggregazione dei nodi di attivazione
- $y(x)$ è il "*firing power*", ovvero l'output della rete.

Il funzionamento della rete è concettualmente molto semplice: ogni input x_i viene moltiplicato per gli opportuni pesi $w_{i,j}$ ed entra in ciascun nodo $f_j(\cdot)$ di attivazione; l'output di questi nodi viene pesato tramite W_j e aggregato tramite la funzione $g(\cdot)$ nel nodo dell'output, che calcola il valore in

uscita dalla rete. L'aggregazione avviene come una combinazione lineare delle uscite dei nodi di attivazione, quindi l'output dell'intera rete vale:

$$y(\underline{x}) = \sum_j (W_j * f_j \left(\sum_i w_{i,j} * x_i \right))$$

L'algoritmo di training di una RBFN risulta molto più efficiente rispetto alla backpropagation rule utilizzata comunemente per addestrare le reti neurali artificiali. Nelle RBFN, infatti, si impone che $w_{i,j} = 1$, $\forall i, j$, in modo da garantire che gli input non vengano modificati prima di entrare nei nodi di attivazione; questa operazione ha anche un senso "fisico", poiché le funzioni di attivazione hanno come argomento la distanza tra l'input e il centro della funzione. Di conseguenza, gli unici pesi da tarare sono quelli tra i nodi di attivazione e l'output: la relazione tra questi, però, è lineare e si può procedere in modo molto efficiente utilizzando l'algoritmo dei minimi quadrati.

Se $w_{i,j} = 1$, è lecito porre:

$$f_j(\sum_i x_i) = a_j$$

Il calcolo dell'output diventa:

$$y(\underline{x}) = \sum_j (W_j * a_j)$$

ovvero, informa vettoriale:

$$\underline{y}(\underline{x}) = \underline{W}^T \underline{a}$$

Applicando l'algoritmo dei Minimi Quadrati, si ottiene la stima dei pesi come:

$$\underline{\widehat{W}} = (\underline{a}^T \underline{a})^{-1} \underline{a}^T \underline{y}$$

In una RBFN, l'attivazione avviene tramite funzioni a base radiale, ovvero funzioni a valori reali il cui output dipende solo dalla distanza tra l'input e un punto fisso. Queste possono assumere diverse forme, le più note sono:

- Gaussiana: $f(r) = e^{-(ar)^2}$
- Multiquadric: $f(r) = \sqrt{1 + (ar)^2}$
- Quadratica inversa: $f(r) = \frac{1}{1+(ar)^2}$
- Multiquadric inversa: $f(r) = \frac{1}{\sqrt{1+(ar)^2}}$
- Lineare: $f(r) = r$

dove $r = \|x - x_0\|$ è la distanza tra l'input x e il punto fisso x_0, mentre a rappresenta un parametro, solitamente l'inverso del raggio critico, che misura quanto sia ampia la finestra identificata dalla funzione.

I parametri da tarare durante l'addestramento di una RBFN sono di due tipi: i pesi W che agiscono sui collegamenti tra i nodi di attivazione e l'output e i punti fissi delle funzioni di attivazione. A causa di ciò il training della rete avviene in due passaggi: prima si calibrano le singole funzioni a base radiale, poiché senza di esse è impossibile proseguire, poi si calcolano i pesi W. I punti noti della serie temporale vengono utilizzati sia come punti fissi dei diversi nodi di attivazione e successivamente per tarare i pesi dei diversi archi tramite le formule presentate sopra. La RBFN avrà, quindi, un numero di nodi di attivazione pari alla quantità di valori noti della serie da analizzare.

La RBFN usata nel caso in esame ha un'architettura diversa da quella mostrata in fig.1: lo schema ivi mostrato rappresenta il caso generico, tipico di un modello "Multi Input, Single Output" (MISO), in cui l'input è multidimensionale. Nel caso in esame, tuttavia, l'input è rappresentato solamente dall'istante temporale in cui avviene la misura (o in cui si vuole operare la previsione), quindi il modello è del tipo "Single Input, Single Output" (SISO) e la struttura della rete sarà come quella presentata in fig.2. A dispetto del kriging, in cui la predizione avviene simultaneamente per ogni punto ignoto, l'output della RBFN è calcolato per ogni punto separatamente: gli input sono quindi forniti in modo sequenziale.

Di conseguenza, le formule indicate prima possono essere opportunamente modificate sostituendo il vettore $\underline{x}$ con uno scalare e rimuovendo la sommatoria:

$$f_j(x) = a_j$$

$$\underline{y}(x) = \underline{W}^T \underline{a}$$

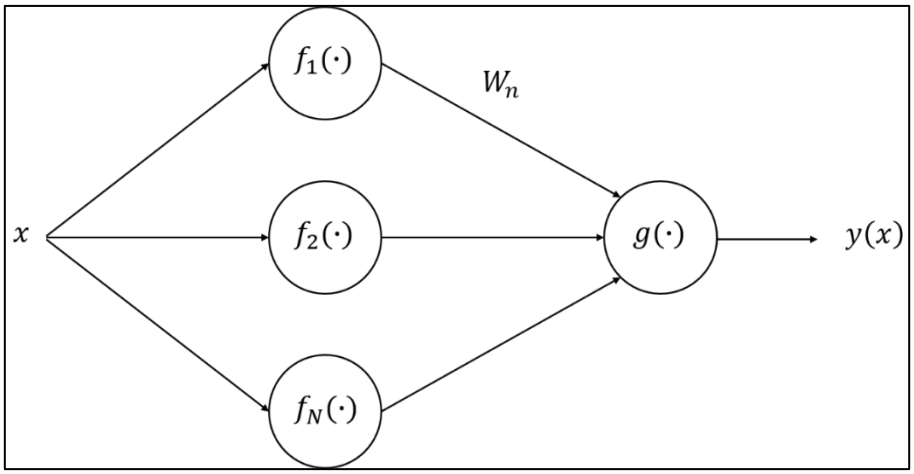

Fig. 17 - Schema di una RBF-Net SISO.

L'algoritmo presentato in questo paragrafo risulta in generale più veloce rispetto a quello del kriging, poiché la scelta delle trasformazioni non-lineari dell'input non richiede la stima del semivariogramma. Questa velocità di calcolo si paga in termini di informazioni che le RBFN non forniscono: nel calcolo dell'output, infatti, non viene fornito il valore della varianza dell'errore. Quest'ultima rappresenta una misura "built-in" di bontà del fitting e nel kriging è calcolata insieme alla stima dei parametri.

I due algoritmi esaminati in questo paragrafo sono stati sviluppati in ambiti disciplinari molto diversi e sono basati su presupposti differenti. Ciò nonostante, i modelli che ne conseguono possono essere interpretati sotto la stessa luce: sia il semivariogramma che le funzioni di attivazione della RBFN servono a valutare la distanza tra l'input e un punto fisso. Inoltre, il calcolo dell'output di entrambi i modelli avviene come combinazione lineare di trasformazioni non-lineari dell'input.

4. RISULTATI

4.1. Pre-processing

Obiettivo di questo lavoro di tesi è la definizione di una procedura di pre-elaborazione di serie storiche di osservazioni di variabili meteorologiche, registrate da sensori diversi, che consenta di ottenere sequenze con una stessa risoluzione temporale, interpolando, ove possibile, i dati mancanti a partire dalle osservazioni disponibili. La scelta del metodo di interpolazione è centrale nella definizione della procedura. Si è deciso di confrontare due metodi uno stocastico, basato sulla minimizzazione della varianza dell'errore di predizione, l'altro puramente empirico: il kriging e le RBF-Network. In questo capitolo si descrivono le modalità di applicazione dei due metodi: le ipotesi fatte, l'implementazione degli algoritmi e i risultati ottenuti per finire con il confronto tra le prestazioni dei metodi utilizzati in termini di tempi di calcolo. Tutte le elaborazioni esposte in questo capitolo sono state effettuate attraverso software sviluppato ad hoc utilizzando il linguaggio Python.

Prima di applicare gli algoritmi di interpolazione ai dati si è dapprima rimossa la componente con periodicità annuale, stimata epoca per epoca come media di tulle le osservazioni disponibili nei diversi anni considerati dal 2010 al 2018. Per fare ciò è stato necessario riportare tutte le serie temporali dei diversi sensori allo stesso numero di epoche, inserendo ove necessario dei no-data (NaN). Più precisamente tutte le serie sono state riportate a un numero di epoche multiplo di 52.560, ovvero il numero di epoche con passo temporale di 10 minuti contenute in un anno canonico. In particolare si è aggiunto all'inizio di ciascuna serie un numero di record nulli tale da coprire l'intervallo tra il 1° gennaio dell'anno di inizio della serie e il primo dato registrato, se necessario. Sempre per garantire che tutti gli anni abbiano lo stesso numero di record, si è eliminato il 29 febbraio dagli anni bisestili e l'unico record del 1° gennaio 2018, presente alla fine di tutte le serie.

Quindi si è calcolata per ciascun sensore la serie annuale di valori di riferimento o climatologia mediando tutte le osservazioni disponibili per una certa epoca nei diversi anni considerati (dal 2010 al 2018). Ad esempio, il valore del 1° gennaio alle ore 00:00:00 della serie di riferimento è pari alla media dei valori del 1° gennaio alle ore 00:00:00 di ciascun anno tra il 2010 e il 2018. Questa operazione, effettuata per epoca, ha restituito un vettore di 52.560 elementi, che è stato sottratto alle serie temporali osservate ottenendo dei residui poi utilizzati per la successiva interpolazione. Si fa notare che i buchi di lunghezza maggiore a 100 record consecutivi sono interpolati ma vengono poi sostituiti con la serie temporale di riferimento a posteriori.

I residui ottenuti sono stati utilizzati per la stima del semivariogramma empirico, necessario per la interpolazione con il kriging, che descrive l'auto-correlazione temporale del segnale. Come già riportato nel paragrafo 3.1, la formula la stima del semivariogramma empirico è:

$$\hat{\gamma}(k\Delta) + \hat{\sigma}_v^2 = \frac{1}{2}\frac{1}{N_{k\Delta}} \sum_{k\Delta - \frac{\Delta}{2} < t_i - t_j < k\Delta + \frac{\Delta}{2}} [y(t_i) - y(t_j)]^2$$

Essa fornisce la stima del semivariogramma per un insieme discreto di distanze grigliate che vanno successivamente interpolati con una funzione ammissibile, che cioè rispetti le proprietà del semivariogramma già enunciate sempre nel paragrafo 3.1.

Per quanto riguarda le osservazioni di pressione atmosferica raccolte dal sensore '9060', il semivariogramma è rappresentato in fig.18. I punti indicano che il semivariogramma debba approssimare una curva monotona crescente, con concavità verso il basso. La semivarianza massima si trova a una distanza di 1008 epoche: quest'ultimo valore corrisponde a una settimana di osservazioni e rappresenta la lunghezza massima di auto-correlazione temporale del segnale.

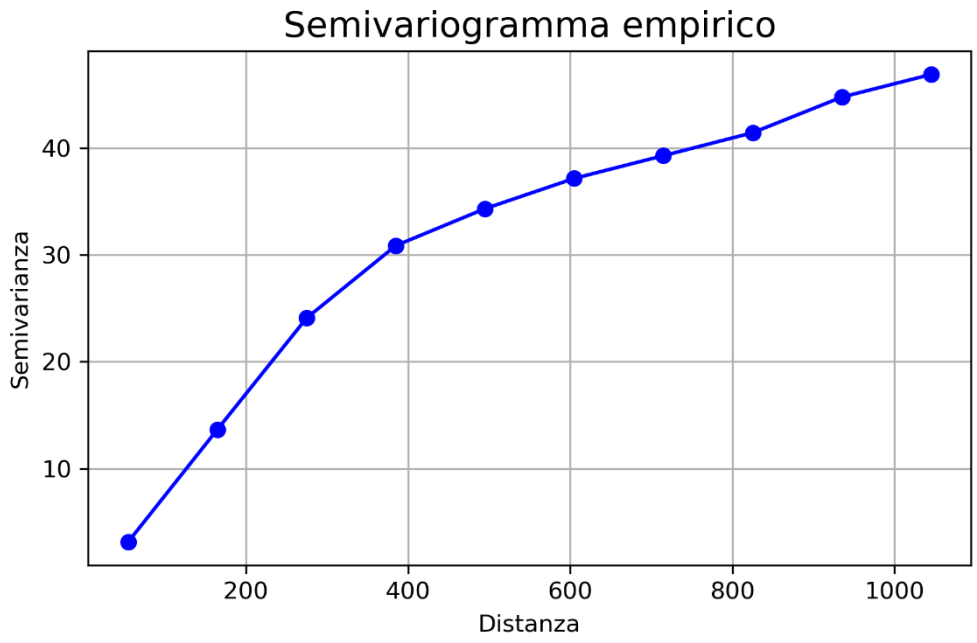

Fig. 18 - Punti del semivariogramma del sensore di pressione atmosferica '9060'.

L'interpolazione del semivariogramma empirico è stata effettuata tramite due funzioni ammissibili: una retta e un'esponenziale. La stima dei parametri di queste due funzioni è stata effettuata ai minimi quadrati (MQ): nel caso della funzione lineare l'algoritmo MQ è stato utilizzato direttamente, mentre per l'esponenziale l'algoritmo MQ è stato applicato sui logaritmi delle variabili d'interesse. I risultati

delle interpolazioni sono illustrati in fig.19, fig.20 e fig.21; nelle didascalie sono indicati i valori di R^2 per valutare la bontà dell'interpolazione e i valori dei parametri stimati.

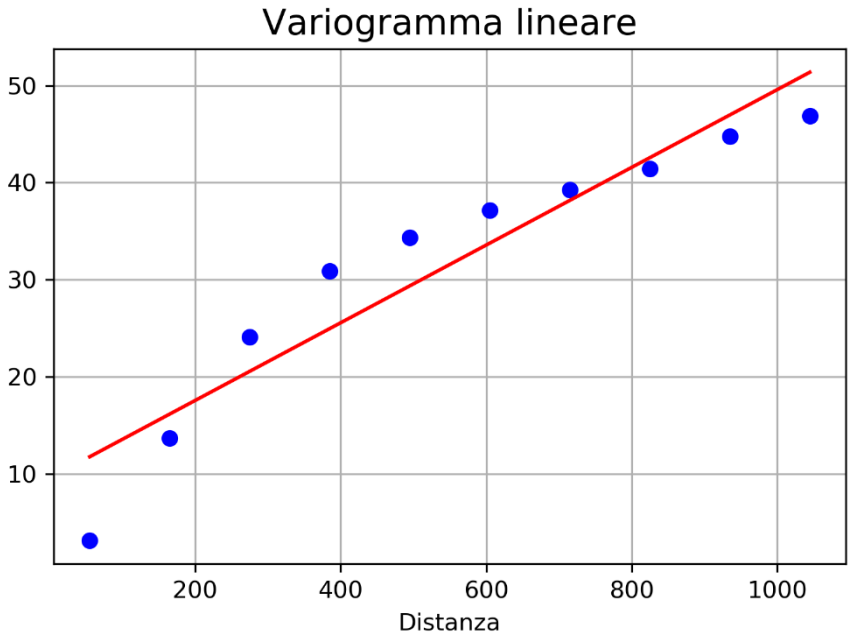

Fig. 19 - Interpolazione tramite funzione lineare.
R^2 = 0,95; coefficiente angolare = 0,04; intercetta = 9,55.

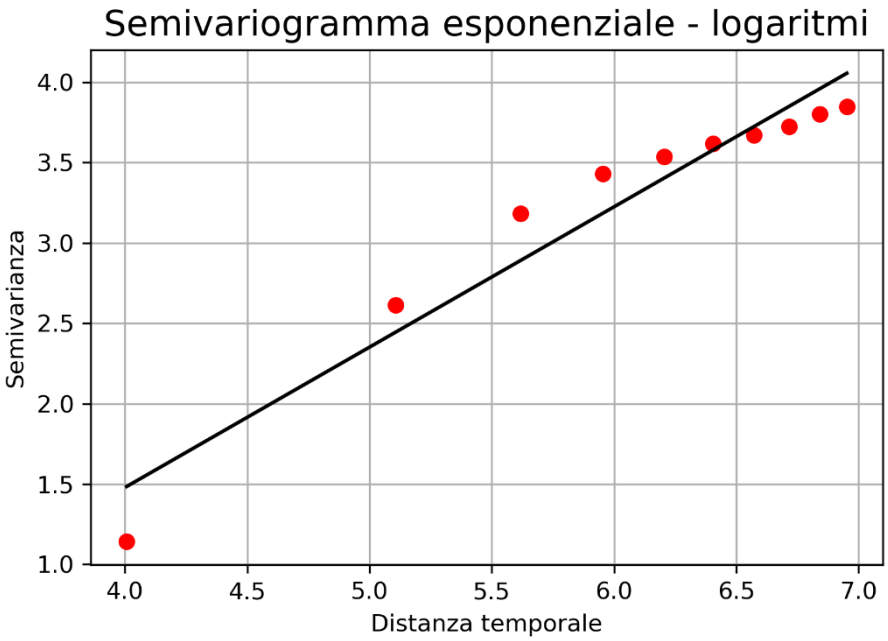

Fig. 20 - Algoritmo MQ applicato sui logaritmi per l'interpolazione esponenziale.
R^2 = 0,97; coefficiente angolare = 0,87; intercetta = -2,01.

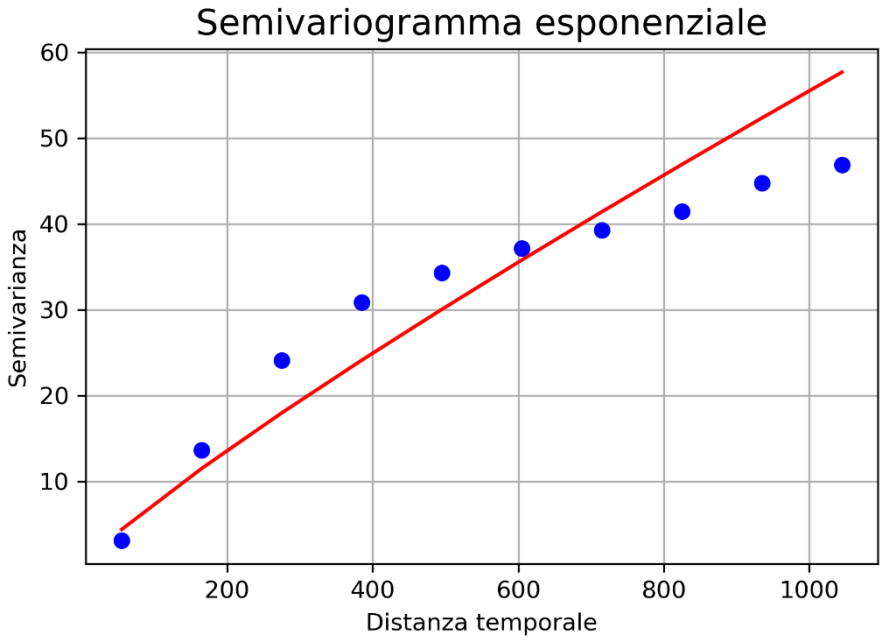

Fig. 21 - Interpolazione tramite semivariogramma esponenziale.
R^2 = 0,96; parametro moltiplicativo = 0,13; esponente = 0,87.

L'interpolazione tramite il modello esponenziale spiega meglio i dati rispetto a quella lineare, soprattutto per quanto riguarda i primi due punti del grafico. Al contrario, la retta approssima meglio i valori sulla coda destra del grafico.

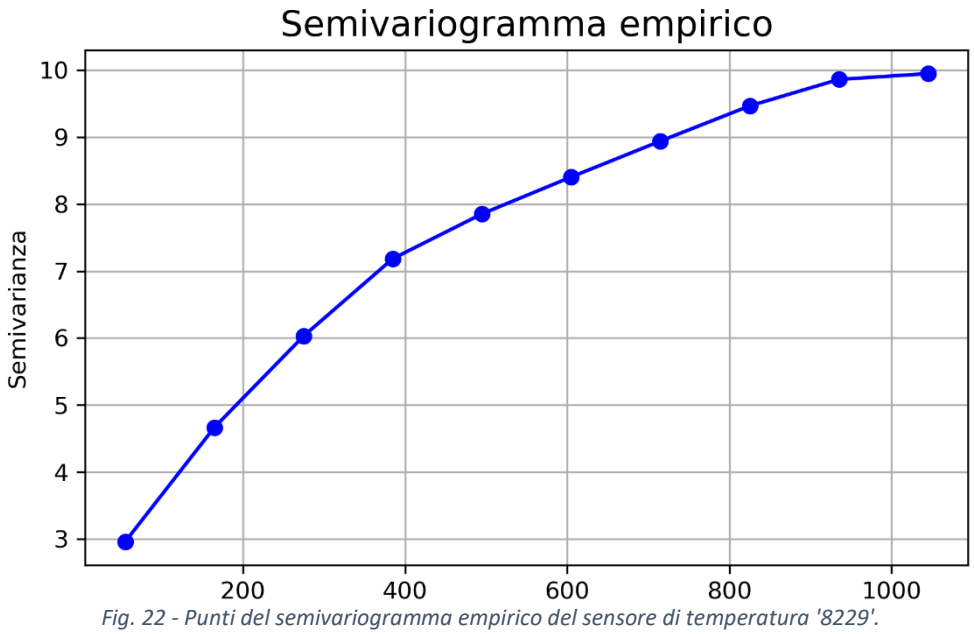

Fig. 22 - Punti del semivariogramma empirico del sensore di temperatura '8229'.

Il semivariogramma delle osservazioni di temperatura registrate dal sensore ‘8229’ è illustrato nel grafico in fig.22. Come si può notare, ha una forma molto simile al suo analogo per la pressione, con due differenze: la scala è, ovviamente, diversa e il valore massimo di semivarianza si trova in ascissa pari all’epoca numero 1080. Ciò significa che queste misure di temperatura hanno un’auto-correlazione massima di 7,5 giorni: dato che le epoche distano 10 minuti, ce ne sono 144 in un giorno, 1008 in una settimana e 1080 in 7 giorni e mezzo.

Come per il sensore analizzato precedentemente, anche questo semivariogramma è stato interpolato con una funzione lineare e una esponenziale, con esponente compreso tra 0 e 1, e anche in questo caso l’interpolazione esponenziale ha una performance leggermente migliore rispetto a quella lineare.

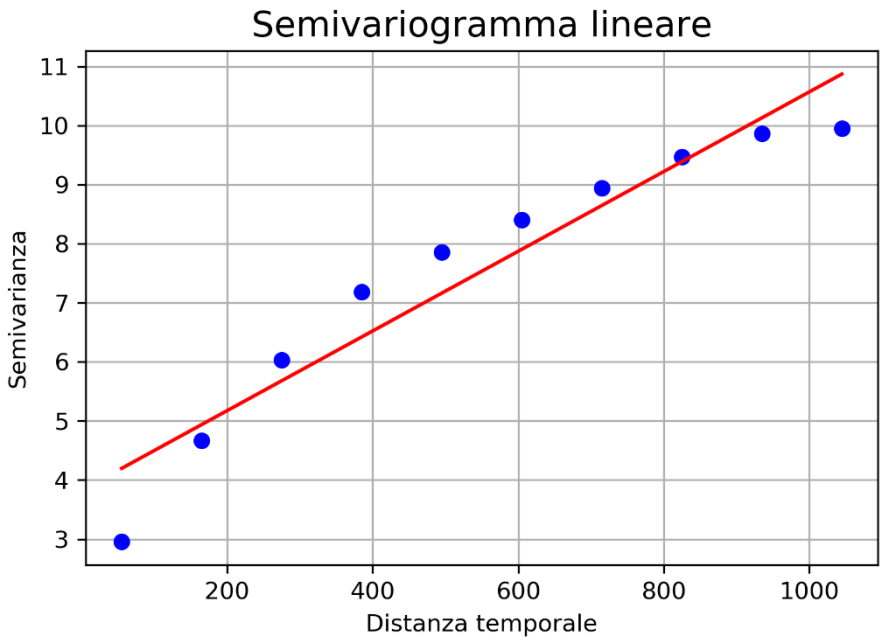

Fig. 23 - Interpolazione tramite funzione lineare.
R^2 = 0,96; coefficiente angolare = 0,0067; intercetta = 3,83.

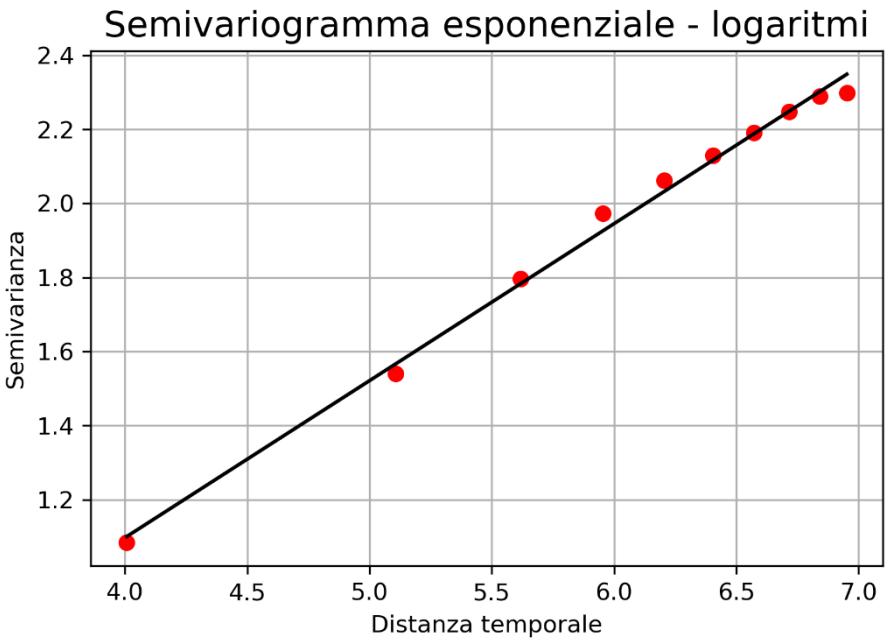

Fig. 24 - Algoritmo MQ applicato sui logaritmi per l'interpolazione esponenziale. R2 = 0,99; coefficiente angolare = 0,42; intercetta = -0,60.

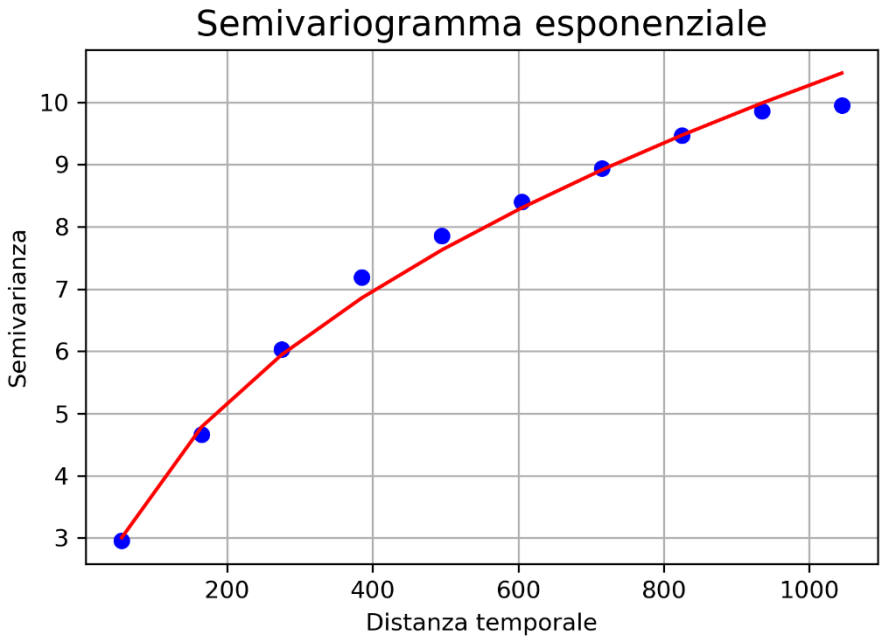

Fig. 25 - Interpolazione tramite semivariogramma esponenziale. R^2 = 0,97; parametro moltiplicativo = 0,55; esponente = 0,42.

Per quanto riguarda il kriging, l'interpolazione è stata effettuata usando entrambi i semivariogrammi appena stimati ma ponendo la varianza del noise uguale a 0 (ordinata della funzione interpolante per distanza temporale nulla), ovvero effettuando una interpolazione esatta. Questa operazione comporta la stima del segnale in un numero di epoche pari a quelle con dato mancante, più alcune scelte tra quelle con dato noto utilizzate per la validazione del metodo. L'operazione è stata poi ripetuta con la sola funzione lineare, ipotizzando che le osservazioni siano la somma di segnale più un rumore bianco e incorrelato con il segnale. In questo caso il numero di punti in cui è stata effettuata la stima è pari al numero complessivo di punti presenti nella serie temporale, in quanto la validazione è effettuata utilizzando i valori noti.

L'algoritmo della RBFN è stato utilizzato, invece, solamente per effettuare interpolazioni esatte. Tuttavia sono stati testati tre diversi kernel: multiquadric, gaussiano e lineare. Le funzioni di questi kernel sono già state illustrate nel paragrafo 3.2.

Verrà ora mostrata una panoramica dei risultati ottenuti: complessivamente si valuterà la bontà del fitting tramite il coefficiente di determinazione (R^2) e i tempi di calcolo richiesti dall'algoritmo. Sono mostrati prima i risultati relativi alla pressione atmosferica, poi quelli relativi alla temperatura. A causa della elevata densità di punti, non è possibile fare un confronto usando grafici di lunghezza mensile, come quelli mostrati nel capitolo 2. Il confronto visivo, pertanto verrà effettuato su intervalli di tempo di una settimana.

4.2. Pressione atmosferica

4.2.1. Risultati kriging

La tabella sottostante raccoglie le principali statistiche della serie temporale interpolata. Nella prima colonna a sinistra compaiono i valori della serie originale, già mostrati nel capitolo 2.

	Serie originale	Kriging con variogramma lineare - senza noise	Kriging con variogramma esponenziale - senza noise	Kriging con variogramma esponenziale – con noise
Numerosità	561.835	735.840	735.840	735.840
Media	984,66 hPa	984,72 hPa	984,72 hPa	984,72 hPa
Dev. St.	7,59 hPa	7,36 hPa	7,36 hPa	7,30 hPa
Min	943,7 hPa	943,7 hPa	943,7 hPa	945,9 hPa
q25%	980,3 hPa	980,7 hPa	980,7 hPa	980,7 hPa
q50%	984,8 hPa	984,9 hPa	984,9 hPa	984,8 hPa
q75%	989,1 hPa	989,0 hPa	989,0 hPa	988,9 hPa
Max	1016,7 hPa	1016,7 hPa	1016,7 hPa	1016,3 hPa

Tab. 13 - Statistiche descrittive delle serie interpolate.

Come si nota esaminando i dati in tabella, non ci sono differenze sostanziali tra le diverse serie di dati interpolati. Questo è un buon indicatore della bontà dell'interpolazione. I grafici in fig.26, fig.27 e fig.2 mostrano i valori dei dati nella prima settimana di osservazioni, interpolati in modo esatto tramite kriging con semivariogramma lineare; in rosso sono evidenziati i valori del dato reale utilizzato per il calcolo del coefficiente di determinazione.

Mentre tra i primi due grafici non si notano differenze sostanziali tra valori osservati e valori predetti, nel terzo grafico i valori osservati differiscono da quelli predetti, per via della ipotesi di osservazioni con noise. Nel caso di interpolazioni esatte, infatti, si nota che il segnale segue molto bene l'andamento della serie originale, rispettandone, a tratti, la discretizzazione. Quando viene effettuata un'interpolazione non esatta, invece, il segnale non è vincolato al passaggio per i punti noti della serie, quindi il risultato è una curva più liscia. È interessante notare che in quest'ultimo caso l'errore di predizione è più ampio rispetto ai grafici precedenti e non assume mai valori pari a zero. Esso, inoltre, non aumenta in modo drastico agli estremi della serie come invece succede negli altri due grafici.

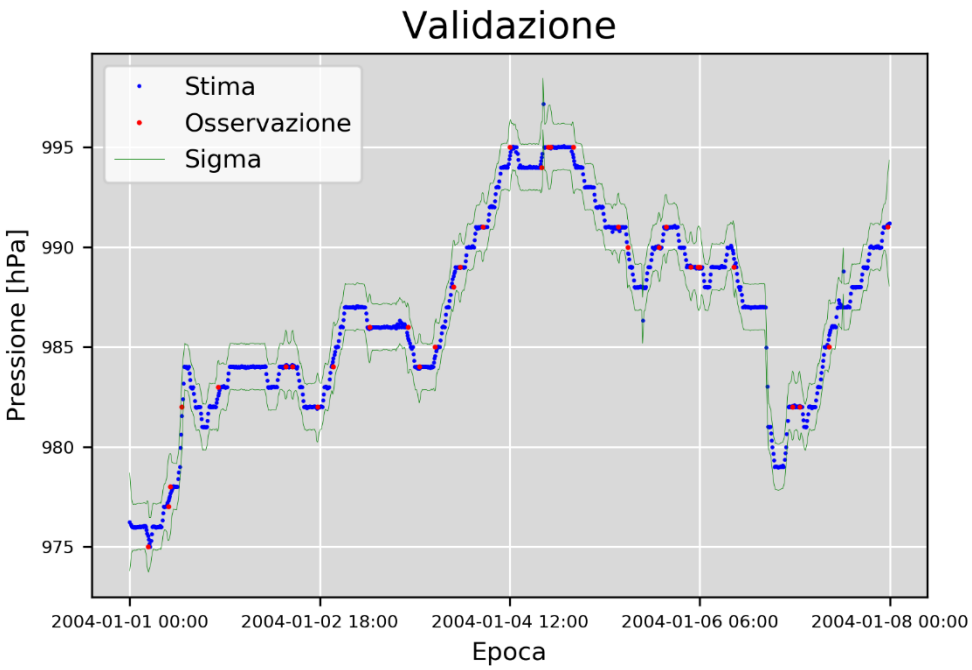

Fig. 26 - Osservazioni interpolate con kriging esatto, semivariogramma lineare; prima settimana di osservazioni.

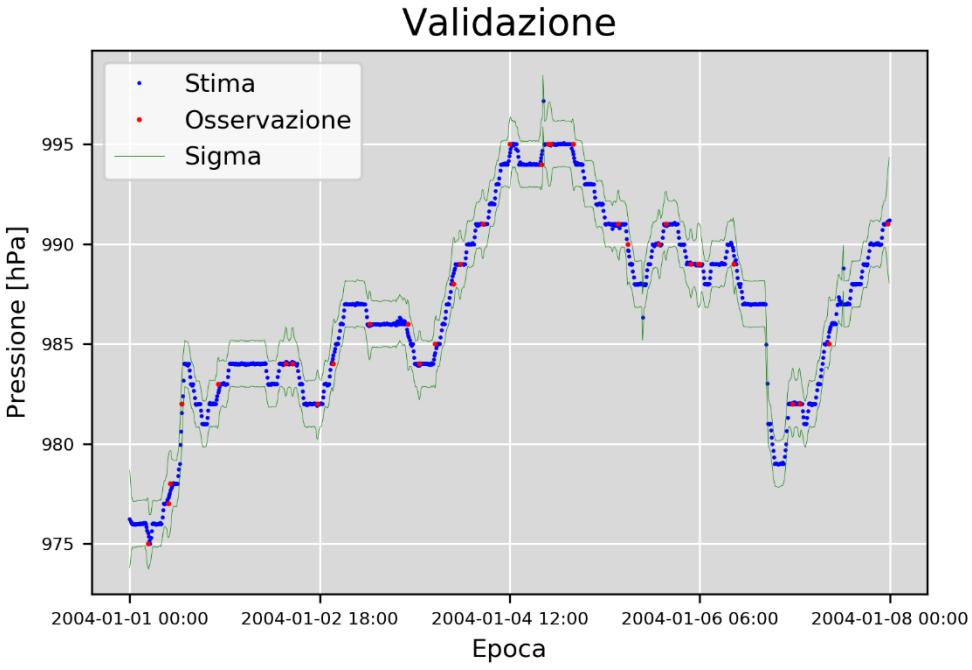

Fig. 27 - Osservazioni interpolate con kriging esatto, semivariogramma esponenziale.

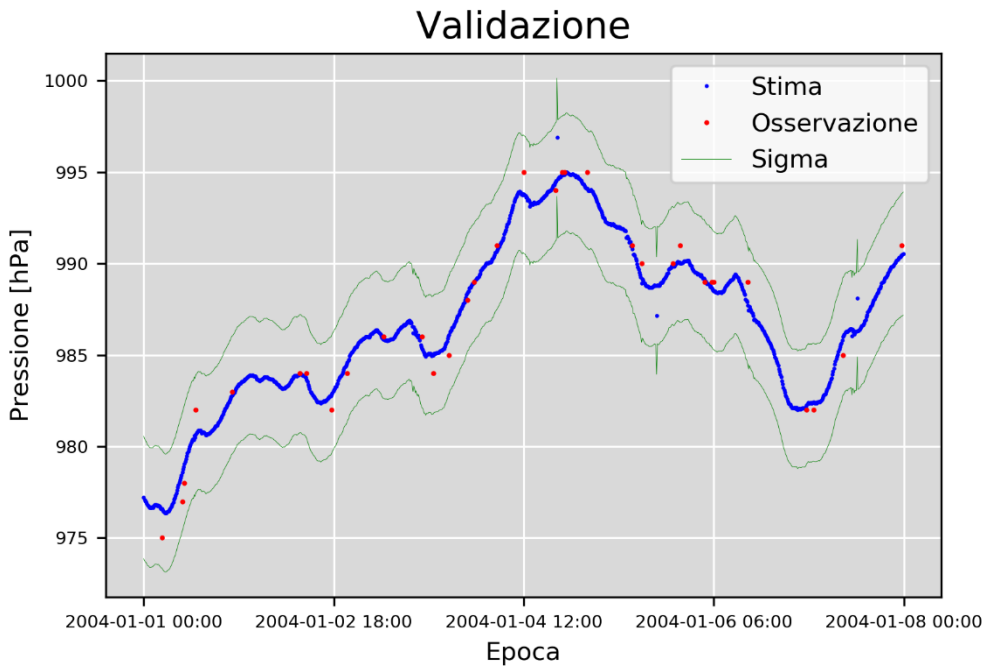

Fig. 28 - Osservazioni interpolate con kriging non esatto e variogramma lineare.

Per valutare la bontà dei fitting si riportano anche gli scatterplot tra valori osservati e predetti nei punti di validazione, fig.29, fig.30 e fig.31: questi corrispondono ai punti rossi presenti nei grafici precedenti. La prima caratteristica che si nota, pertanto, è che i punti nel grafico in fig.31 sono molto più numerosi rispetto alle altre immagini. La seconda caratteristica è che, com'era intuibile guardando i grafici presentati sopra, l'utilizzo di un semivariogramma lineare o esponenziale non cambia di molto la predizione, dato che i relativi grafici sono pressoché identici.

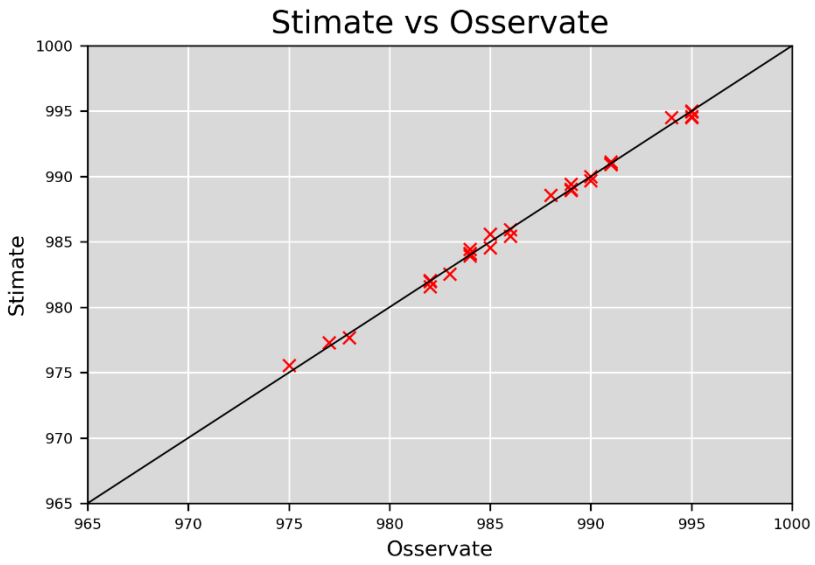

Fig. 30 - Osservazioni contro stima, interpolazione esatta con semivariogramma lineare.

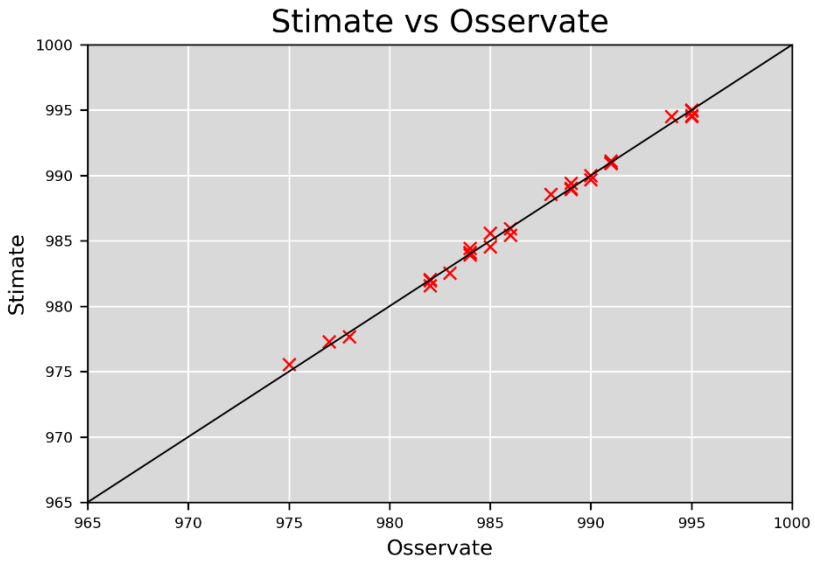

Fig. 29 - Osservazioni contro stima, interpolazione esatta con semivariogramma esponenziale.

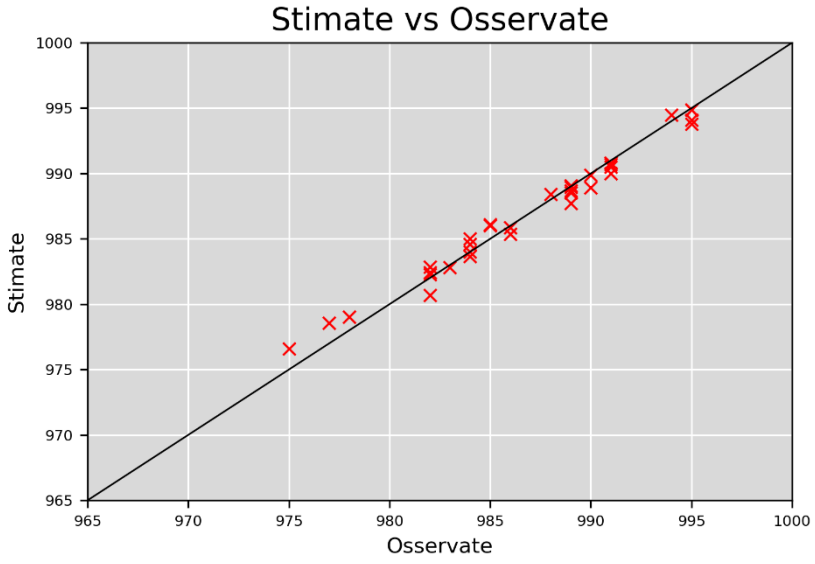

Fig. 31 - Osservazioni contro stime, interpolazione non esatta, semivariogramma lineare.

Per quanto riguarda l'ultima settimana di osservazioni interpolate si possono fare considerazioni del tutto analoghe: l'utilizzo della funzione esponenziale piuttosto che quella lineare nell'interpolazione esatta non influisce sul risultato finale. D'altra parte, le statistiche mostrate nella tabella precedente sono piuttosto esplicative del trend generale delle serie interpolate con questi metodi. Per quanto riguarda l'interpolazione non esatta effettuata con semivariogramma lineare, invece, si può notare come, analogamente a quanto accaduto nella prima settimana di osservazioni, la curva sia più smussata.

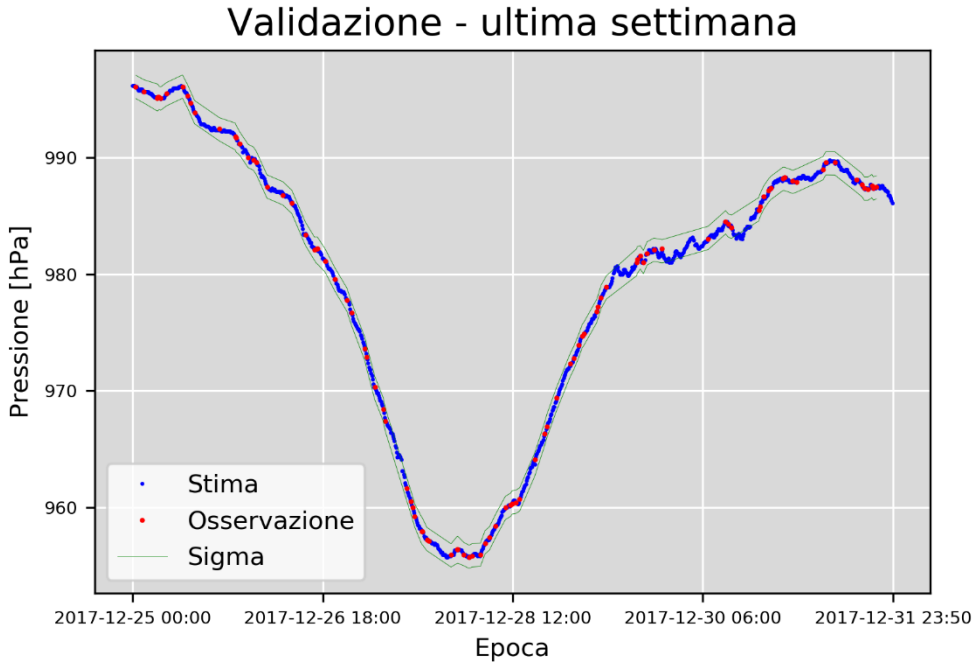

Fig. 32 - Osservazioni interpolate con kriging esatto, semivariogramma lineare; ultima settimana di osservazioni.

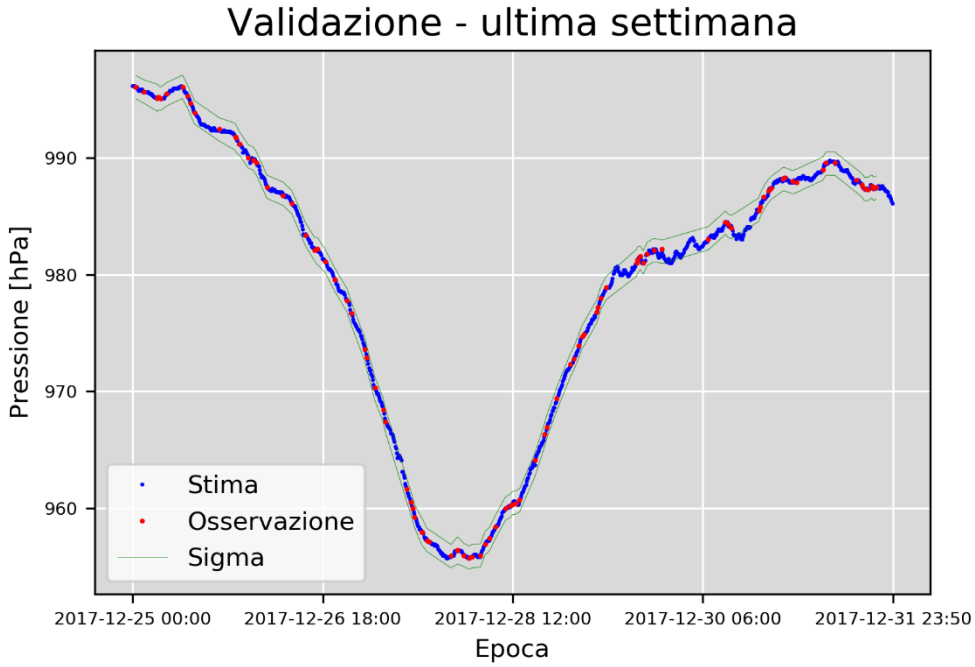

Fig. 33 - Osservazioni interpolate con kriging esatto, semivariogramma esponenziale.

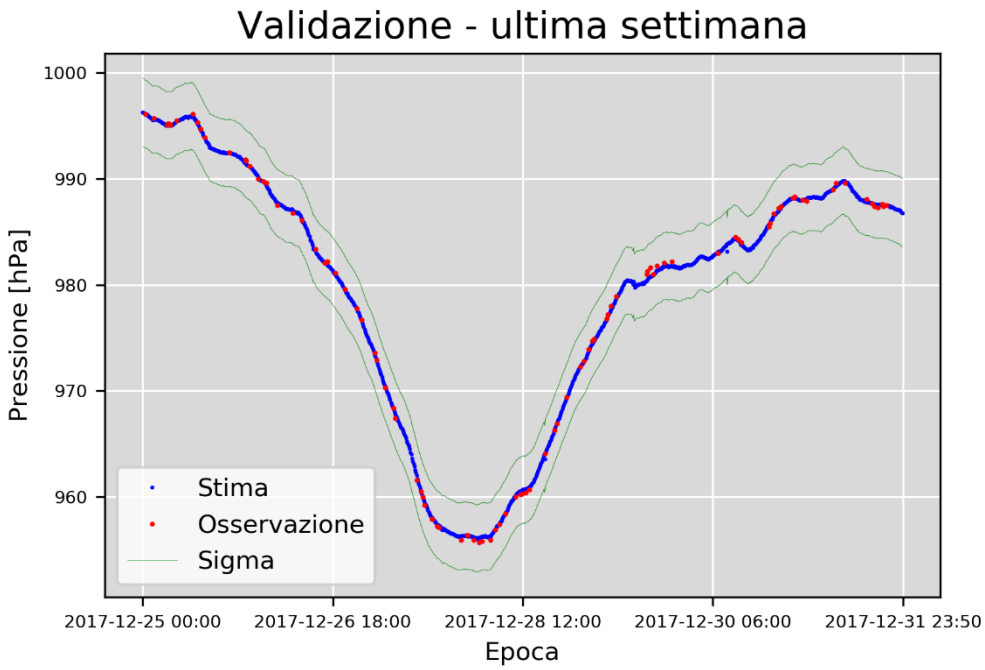

Fig. 34 - Osservazioni interpolate con kriging non esatto, semivariogramma lineare.

La bontà del fitting dell'interpolazione dell'ultima settimana di osservazioni rispecchia i risultati già mostrati nel caso della prima settimana, come è d'altronde intuibile dai grafici appena illustrati. L'interpolazione esatta effettuata con entrambe le funzioni lineare ed esponenziale è molto buona, quella non esatta è leggermente peggiore. In questo secondo lasso temporale, però, l'accuratezza è nettamente maggiore: ciò si evince dal grafico in fig.35, che mostra come la nuvola di punti sia decisamente più concentrata sulla diagonale rispetto al suo analogo nel periodo precedente.

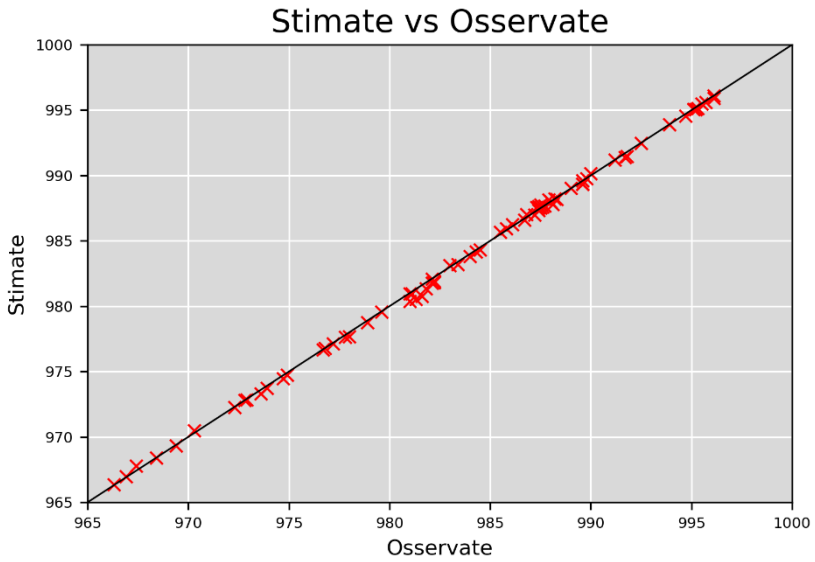

Fig. 35 - Osservazioni contro stime.

Dopo aver mostrato gli andamenti esemplificativi delle serie interpolate, è comunque necessario analizzare le distribuzioni globali per valutare la bontà del fitting dei diversi metodi. In aggiunta alle statistiche mostrate in tab.13, sono stati calcolati tutti i percentili per verificare la similitudine tra le distribuzioni delle serie interpolate e quella originale. In tab.14 sono mostrati i percentili "limite", ovvero quelli pari a 1% e a 99%.

Quantile	Serie originale	Kr. lineare	Kr. esponenziale	Kr. non esatto
1%	964,6 hPa	965,1 hPa	965,1 hPa	965,3 hPa
99%	1003,0 hPa	1002,8 hPa	1002,8 hPa	1002,6 hPa

Tab. 14 - 1° e 99-esimo quantile per tutte le serie esaminate.

Successivamente sono stati plottati i quantili delle tre serie interpolate contro quelli della serie originale, in modo da verificarne l'aderenza. Oltre a ciò sono mostrati anche i grafici con le funzioni densità di probabilità.

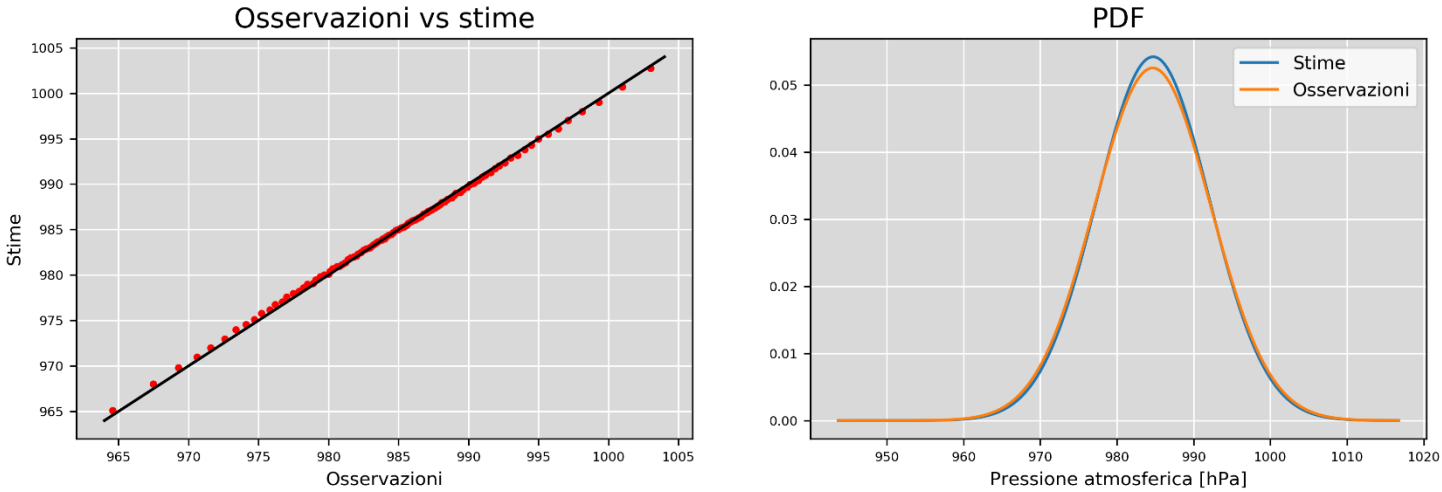

Fig. 36 - Kriging esatto con semivariogramma lineare.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

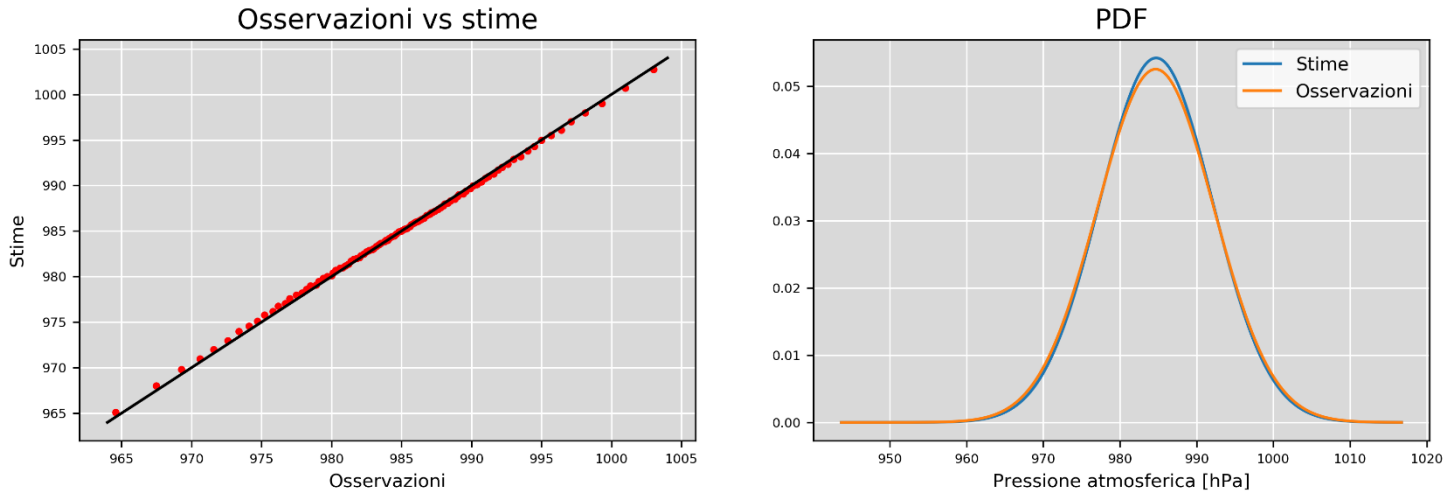

Fig. 37 - Kriging esatto con semivariogramma esponenziale.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

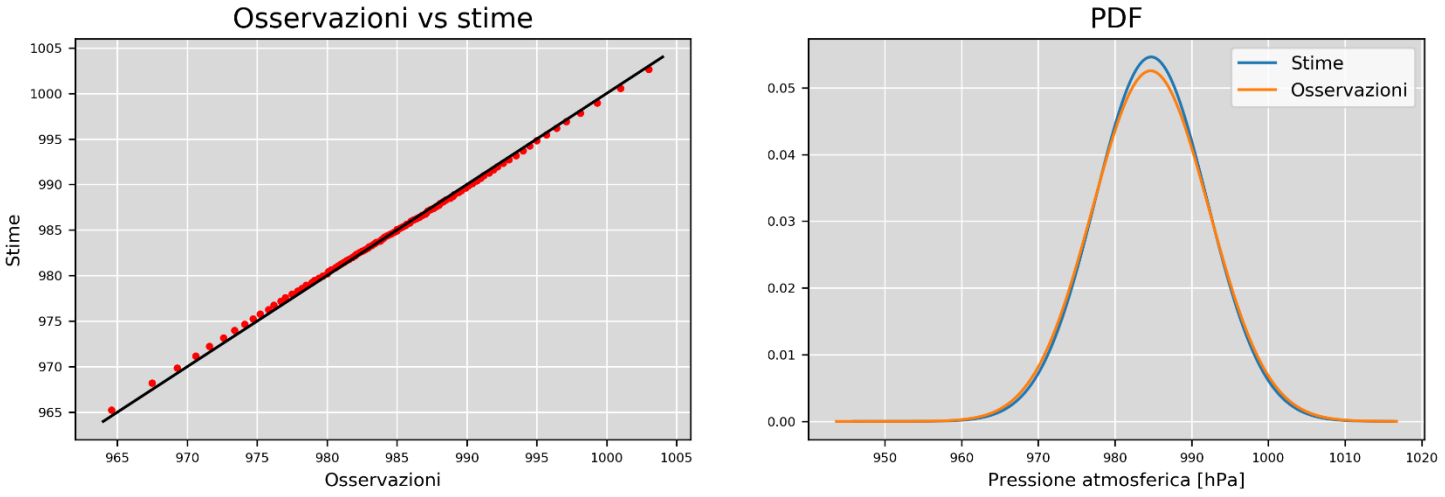

Fig. 38 – Kriging non esatto con semivariogramma lineare.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

Dal grafico a sinistra si evince che l'interpolazione tramite kriging esatto con semivariogramma lineare tende a sovrastimare leggermente i valori più bassi della serie originale, mentre i valori più alti vengono leggermente sottostimati. Nel complesso, comunque, tutti i punti si trovano nelle immediate vicinanze della diagonale del grafico, mostrando che la qualità dell'interpolazione è generalmente molto buona. Il grafico a destra non aggiunge informazioni sostanziali. Per quanto riguarda il kriging effettuato con semivariogramma esponenziale, si nota che le performance globali sembrano migliori: i punti del grafico a sinistra appaiono più vicini alla diagonale e le PDF nel grafico a destra sembrano leggermente più vicine rispetto al grafico analogo illustrato precedentemente. Di

seguito vengono proposti gli ultimi due grafici relativi al kriging non esatto con semivariogramma lineare: le differenze rispetto agli altri due metodi sono minime, anche se si nota un leggero peggioramento.

4.2.2. Risultati RBF-Network

Per quanto riguarda le tre RBFN addestrate, esse mostrano comportamenti diversi a seconda della funzione di attivazione utilizzata. Tali performance non sono prevedibili a priori, quindi verranno estrapolate dall'analisi dei grafici, così come già fatto per il kriging. Di seguito si illustra la tabella coi valori delle serie interpolate tramite le tre RBFN.

	Serie originale	RBFN con kernel multiquadric	RBFN con kernel lineare	RBFN con kernel gaussiano
Numerosità	561.835	735.840	735.840	735.840
Media	984,66 hPa	984,71 hPa	984,72 hPa	984,74 hPa
Dev. St.	7,59 hPa	7,43 hPa	7,36 hPa	7,15 hPa
Min	943,7 hPa	919,3 hPa	943,7 hPa	943,7 hPa
q25%	980,3 hPa	980,7 hPa	980,7 hPa	980,9 hPa
q50%	984,8 hPa	984,9 hPa	984,9 hPa	984,9 hPa
q75%	989,1 hPa	989,0 hPa	989,0 hPa	988,9 hPa
Max	1016,7 hPa	1026,4 hPa	1016,7 hPa	1016,7 hPa

Tab. 15 - Statistiche descrittive delle serie interpolate.

A differenza del kriging, le RBFN mostrano evidenti comportamenti diversi: è sufficiente osservare i numeri in tab.15 per notare alcune differenze. In primo luogo, il valore medio è diverso per i tre metodi, anche se la differenza è solo alla seconda cifra decimale, e differisce da quella della media originale (che risulta minore delle altre tre). Anche la deviazione standard è diversa, in particolare è leggermente minore di quella originale, quindi ci si aspetta che i grafici siano più lisci. Inoltre, mentre le interpolazioni tramite kernel lineare e gaussiano rientrano nei limiti della serie originale, i valori massimi e minimi della RBFN con kernel multiquadric superano di molto quelli della serie storica. I quartili, comunque, rimangono invariati per le tre serie.

Si procederà ora a illustrare l'esempio di interpolazione della prima e dell'ultima settimana di osservazioni, in modo da poter confrontare visivamente le performance dei tre metodi utilizzati.

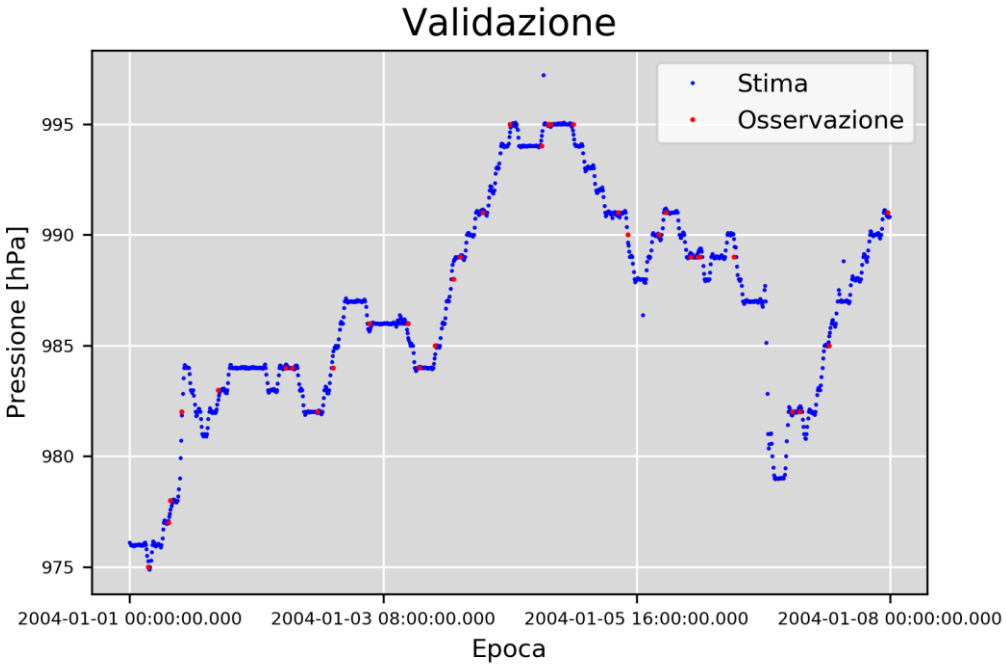

Fig. 39 - Osservazioni interpolate con RBFN, kernel multiquadric; prima settimana di osservazioni.

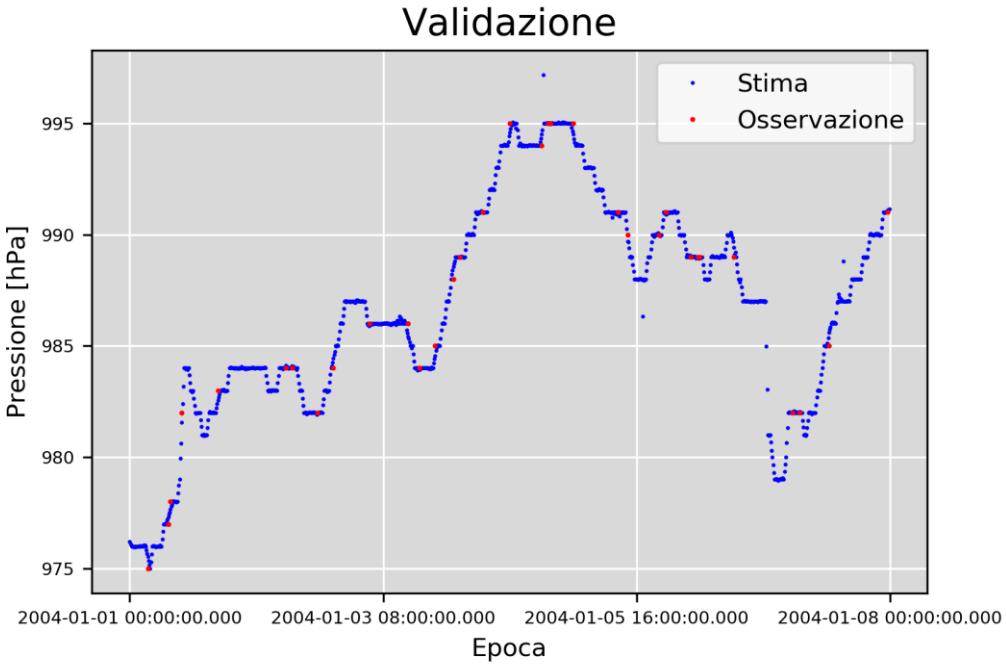

Fig. 40 - Osservazioni interpolate con RBFN, kernel lineare.

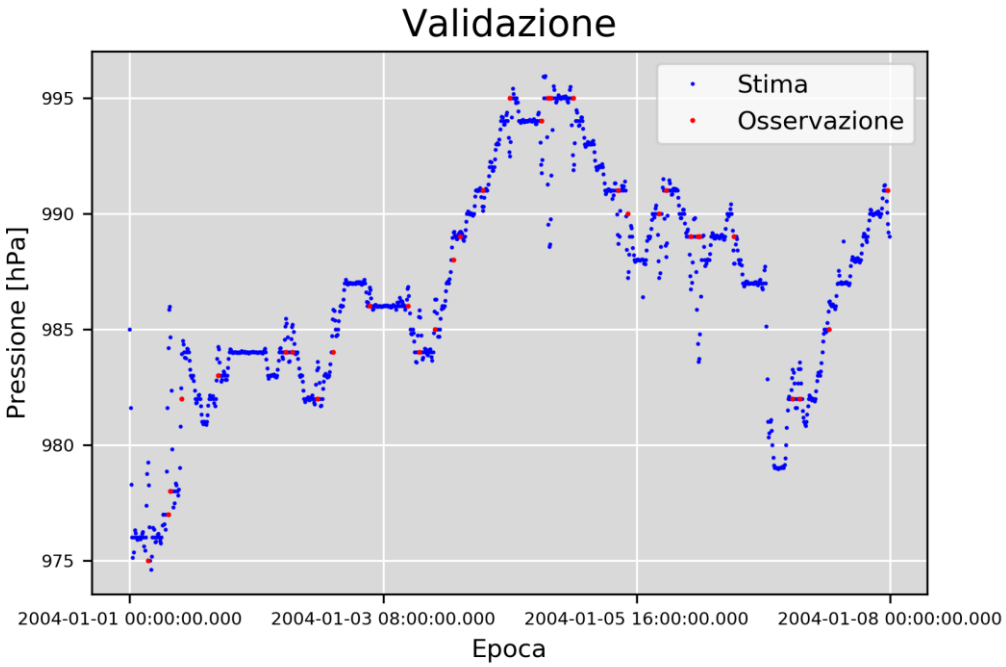

Fig. 41 - Osservazioni interpolate con RBFN, kernel gaussiano.

Mentre i primi due grafici risultano molto simili tra loro e anche rispetto all'interpolazione esatta tramite kriging lineare, la RBFN con kernel gaussiano mostra un comportamento insolito. I punti sono, infatti, decisamente più sparsi rispetto a tutti gli altri modelli mostrati finora. Questo comportamento anomalo è probabilmente causato dalla tendenza del kernel gaussiano a ridurre molto velocemente a zero il valore della stima. Analizzando i valori grezzi forniti da questo modello per l'intero primo mese di osservazioni, infatti, si nota come il periodo vuoto piuttosto lungo dopo il 10 ottobre vada proprio a zero, laddove invece gli altri metodi operano un'interpolazione praticamente lineare.

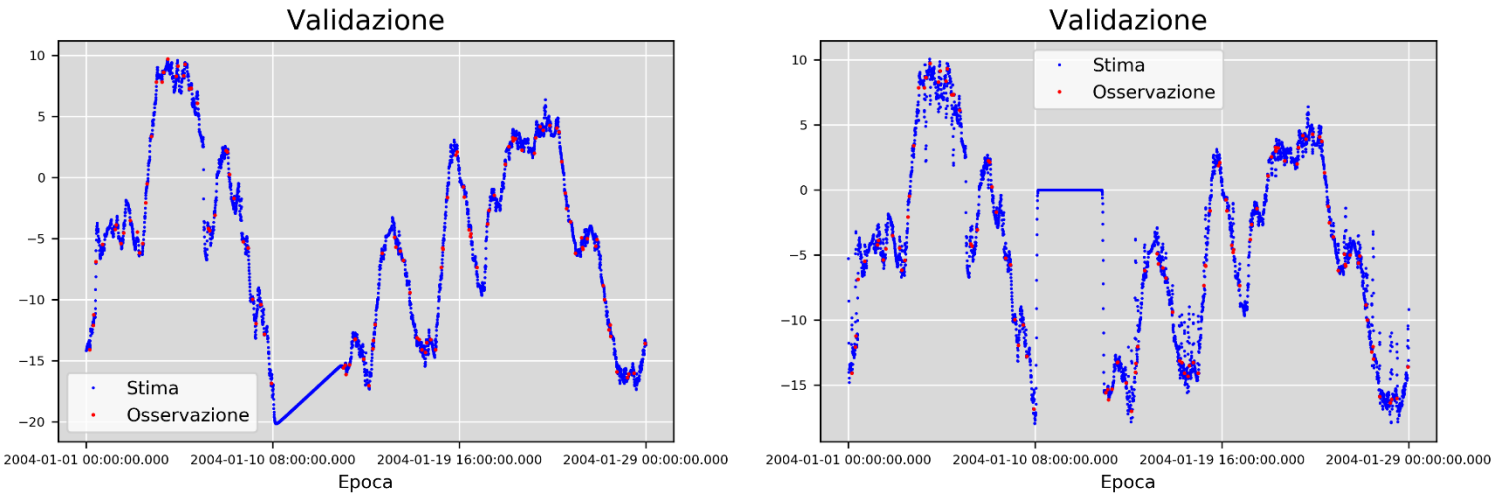

Fig. 42 - Output grezzo delle RBFN con kernel multiquadric a sinistra, con kernel gaussiano a destra.

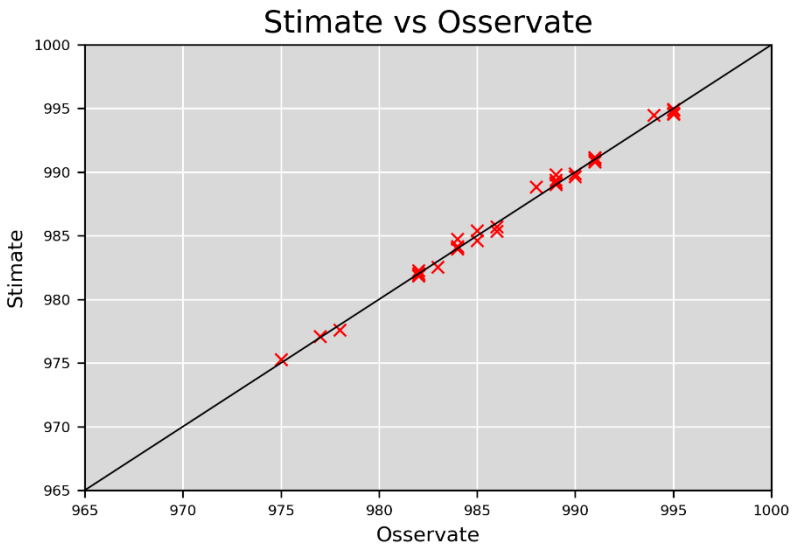

Fig. 44 - Osservazioni contro stime, RBFN con kernel multiquadric; prima settimana di osservazioni.

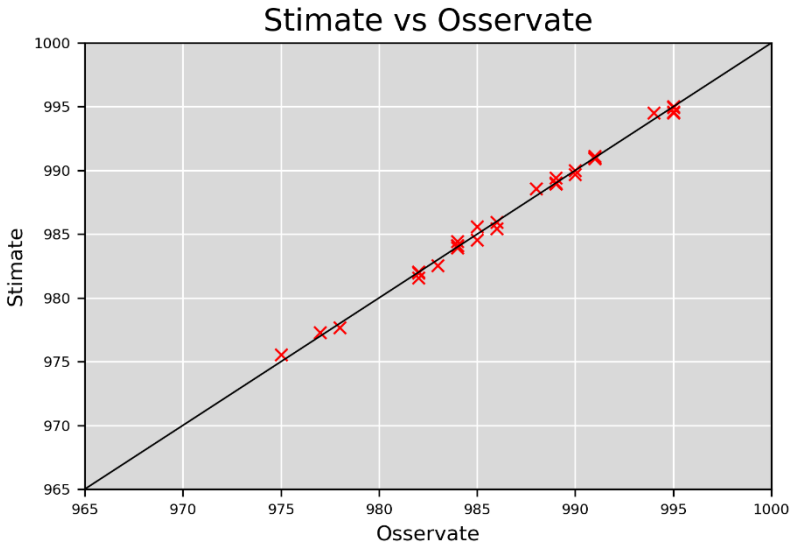

Fig. 43 - Osservazioni contro stime, RBFN con kernel lineare.

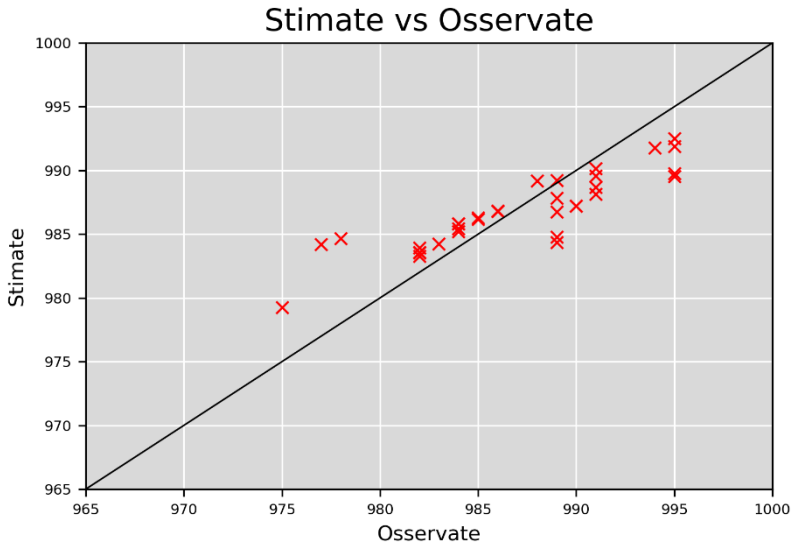

Fig. 45 - Osservazioni contro stime, RBFN con kernel gaussiano.

Il comportamento anomalo della RBFN con kernel gaussiano è visibile anche nel grafico in fig.45: pur essendo tendenzialmente vicini alla diagonale, i punti risultano comunque più sparsi rispetto agli altri metodi. Anche nell'esempio dell'ultima settimana di osservazioni si nota un comportamento anomalo, anche se più sistematico, della RBFN con attivazione gaussiana. Gli altri due metodi, invece, hanno performance del tutto analoghe a quelle del kriging.

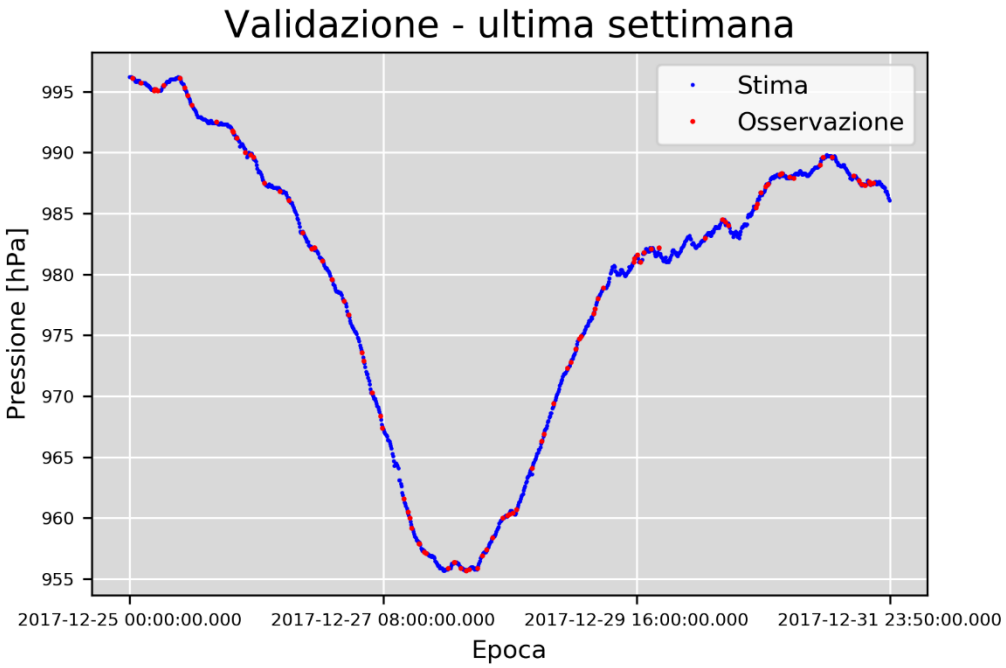

Fig. 46 - Osservazioni interpolate con RBFN, kernel multiquadric; ultima settimana di osservazioni.

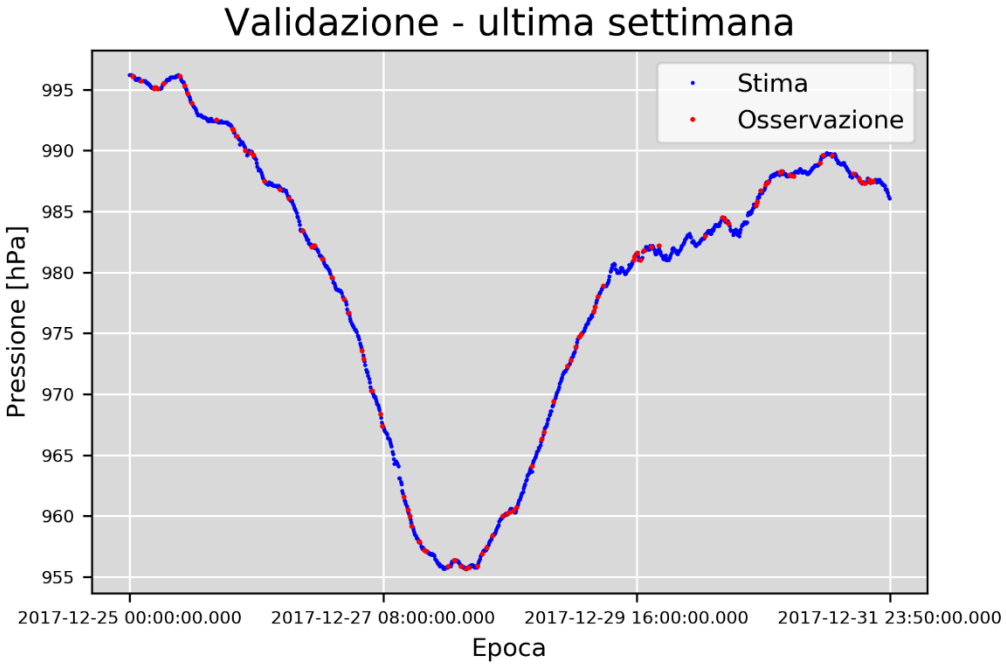

Fig. 47 - Osservazioni interpolate con RBFN, kernel lineare.

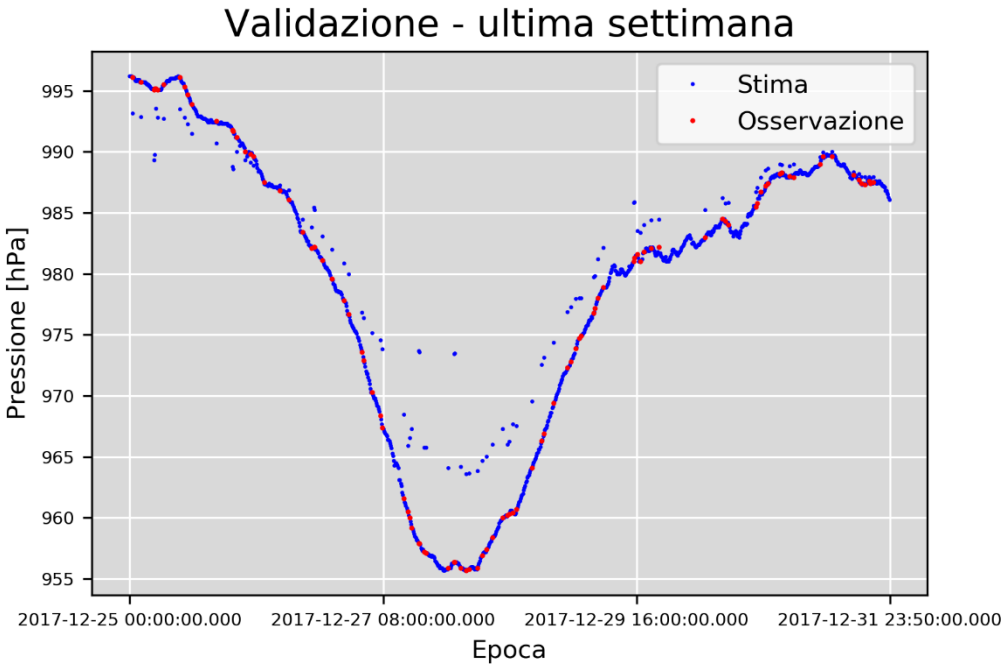

Fig. 48 - Osservazioni interpolate con RBFN, kernel gaussiano.

Osservando il confronto tra i valori stimati tramite RBFN con attivazione gaussiana e quelli reali, si nota che l'interpolazione è sistematicamente errata: nello specifico, i valori fino a 990 hPa vengono sovrastimati, mentre quelli al di sopra di tale soglia vengono sottostimati.

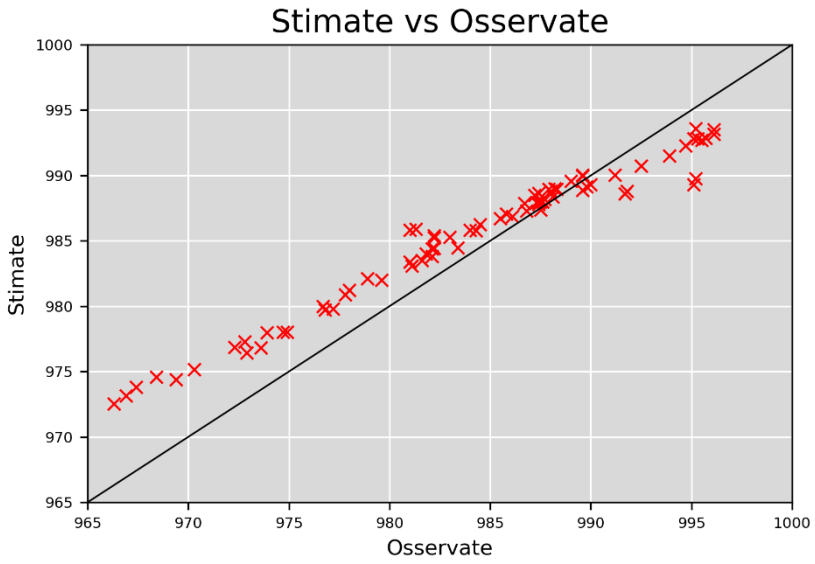

Fig. 49 - Osservazioni contro stime, interpolazione tramite RBFN ad attivazione gaussiana.

Anche le performance delle tre RBFN addestrate è stata valutata in termini di percentili e funzioni densità di probabilità. Nella tabella qui sotto si trovano i percentili estremi delle diverse distribuzioni generate.

Quantile	Serie originale	RBFN con kernel multiquadric	RBFN con kernel lineare	RBFN con kernel gaussiano
1%	964,6 hPa	964,8 hPa	965,1 hPa	965,6 hPa
99%	1003,0 hPa	1002,9 hPa	1002,8 hPa	1002,4 hPa

Tab. 16 - 1° e 99-esimo quantile per tutte le serie esaminate.

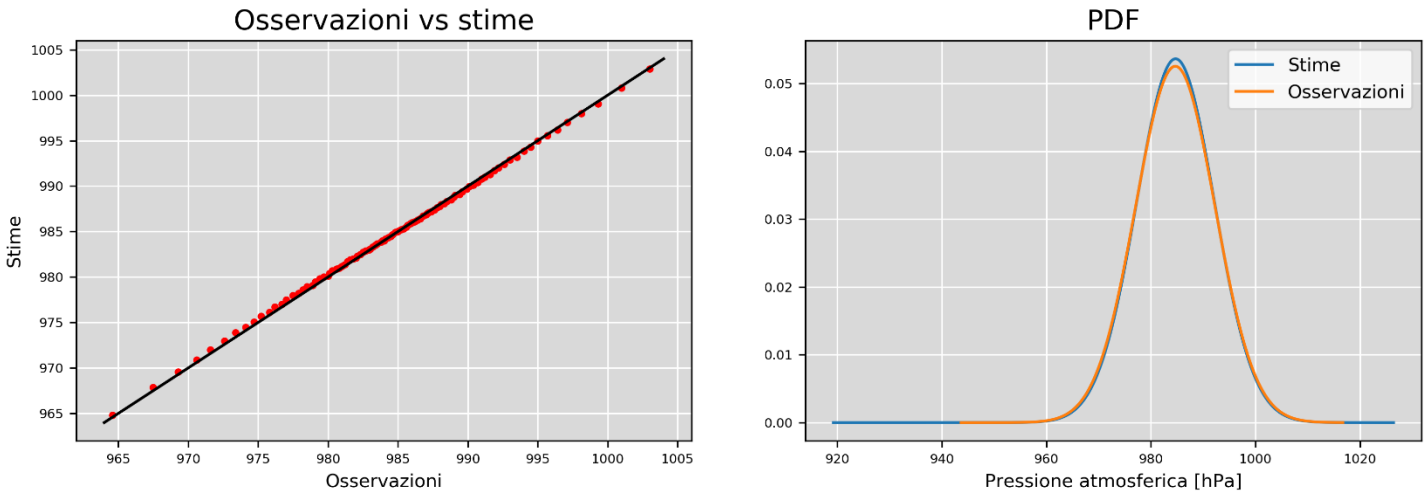

Fig. 50 - RBFN con kernel multiquadric.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

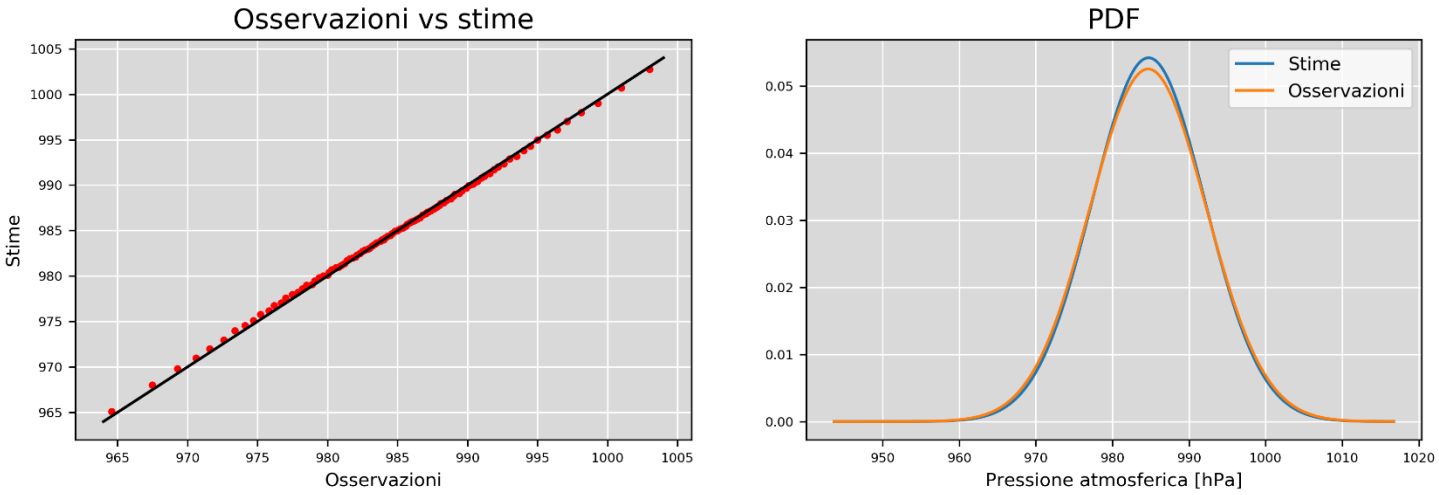

Fig. 51 - RBFN con kernel lineare.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

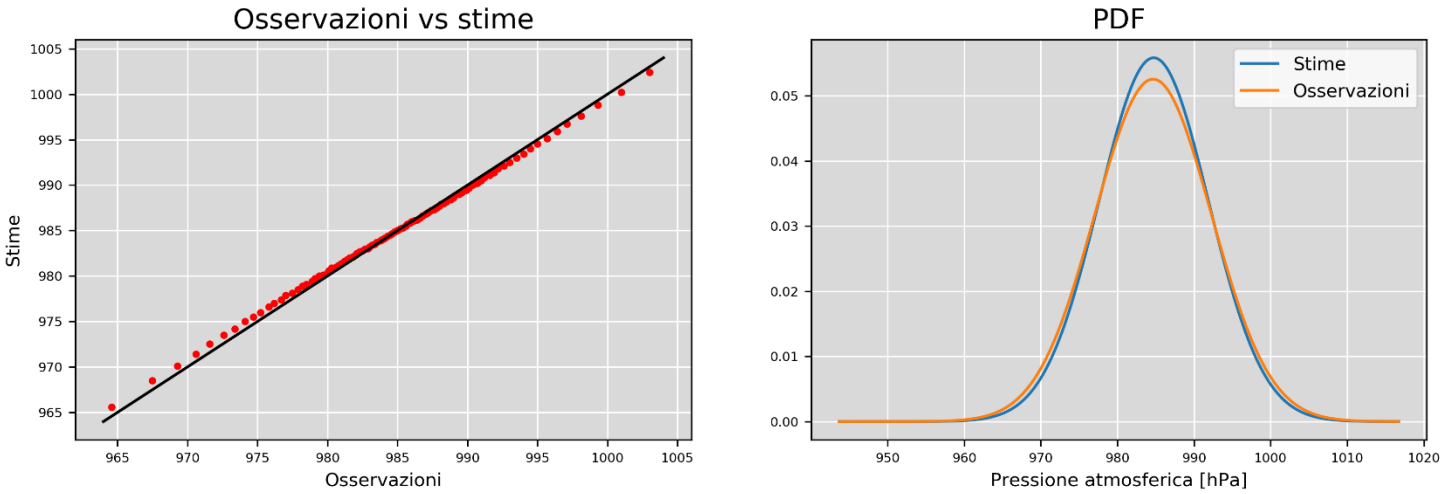

Fig. 52 - RBFN con kernel gaussiano.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

Nei grafici appena illustrati si notano diverse cose interessanti. In primo luogo nella PDF della serie generata dalla RBFN con kernel multiquadric ha una coda sinistra molto allungata dovuta al fatto che tale serie ha una stima minima decisamente inferiore al corrispettivo osservato. A parte ciò, com'era lecito aspettarsi dai grafici delle stime della prima e ultima settimana, gli algoritmi con kernel multiquadric e lineare hanno performance molto simili. L'ultimo metodo della triade, invece, risulta peggiore: entrambi i grafici si discostano dalla situazione reale molto più degli altri.

4.2.3. Ulteriori considerazioni

In conclusione, sono stati calcolati i coefficienti di determinazione medi per ogni metodo e i tempi di calcolo richiesti da ciascun algoritmo su una serie temporale di 735.840 epoche. I risultati sono mostrati in tab.17.

	Kriging esatto lineare	Kriging esatto esponenziale	Kriging non esatto lineare	RBFN multiquadric	RBFN lineare	RBFN gaussiana
R^2 medio	0,997	0,997	0,994	0,996	0,997	0,955
Minuti	4,85	4,96	59,11	0,80	0,65	1,78

Tab. 17 - Coefficienti di determinazione e tempi di calcolo di ciascun algoritmo testato.

La RBFN ad attivazione gaussiana ha un coefficiente di determinazione “nettamente” inferiore rispetto agli altri algoritmi a causa, probabilmente, del comportamento anomalo che il kernel gaussiano genera nella predizione di intervalli di valori di lunghezza maggiore rispetto alla mediana. I grafici in fig.42 nel paragrafo precedente evidenziano questo comportamento. Per approfondire la questione, è interessante osservare come variano i diversi kernel in funzione del parametro “*a*” usato per la loro calibrazione: questo parametro corrisponde alla distanza media tra due valori noti.
Per comodità sono riportate le formule dei kernel:

- Multiquadric: $f(r) = \sqrt{1 + (ar)^2}$
- Lineare: $f(r) = r$
- Gaussiana: $f(r) = e^{-(ar)^2}$

Con $r = \|t - t'\|$ e $a = E[r]$. È immediato notare che il kernel lineare non dipende da alcun parametro, poiché essendo la relazione lineare, tale valore viene “assorbito” dai pesi della rete in modo non dissimile da ciò che accade con il kriging esatto con semivariogramma lineare. I grafici in fig.53, invece, illustrano la dipendenza delle altre due funzioni dal parametro “*a*”. Risulta subito evidente che, al netto della successiva calibrazione della rete, il grado di dipendenza tra due valori decresce molto più rapidamente se descritta da un kernel gaussiano piuttosto che con da uno multiquadric. Questo implica che più sono distanti i buchi della serie, peggiore sarà la predizione, specialmente con l’utilizzo di una funzione di attivazione gaussiana.

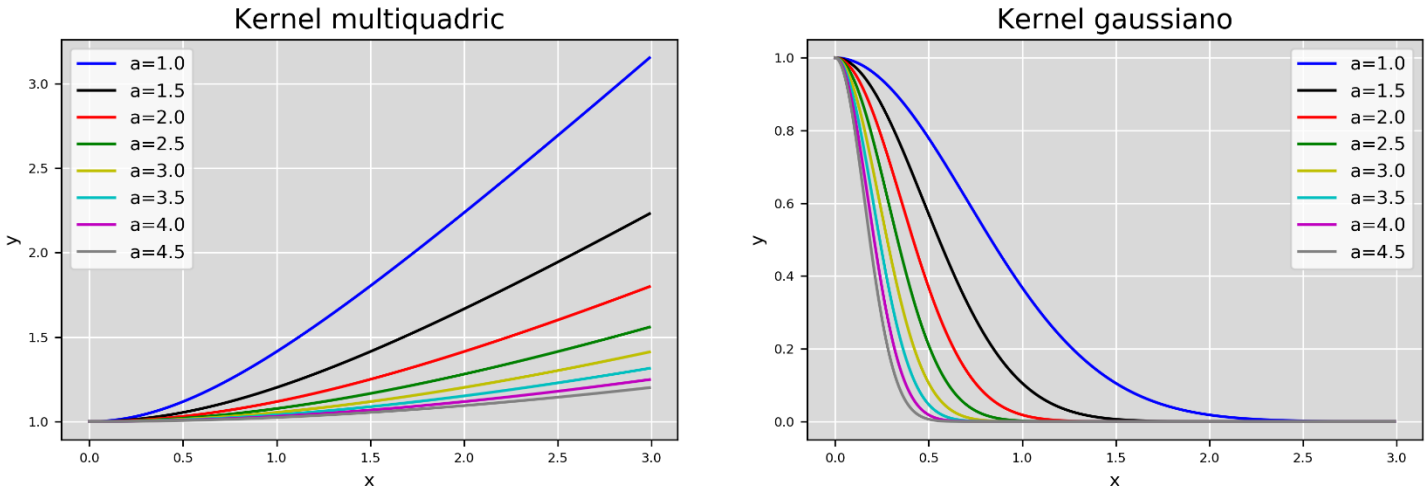

Fig. 53 - A sinistra: funzione multiquadric al variare del parametro "a".
A destra: funzione gaussiana al variare del parametro “a”.
Notare che i grafici hanno scale diverse: le differenze, in questo senso, sono “assorbite” dai parametri della rete durante l’addestramento.

Escludendo questo modello, tutti gli altri metodi mostrano performance molto alte in termini di qualità dell'interpolazione. Dal punto di vista dei tempi di calcolo, però, il kriging non esatto impiega circa un'ora di tempo a causa dell'altissimo numero di predizioni da effettuare. Tra gli algoritmi rimasti, quello con le prestazioni migliori risulta la RBFN con kernel lineare: il coefficiente di determinazione è pari agli altri, ma i tempi di calcolo sono radicalmente inferiori a qualsiasi interpolazione tramite kriging e alla RBFN con kernel multiquadric (che pure presenta un problema coi valori estremi).

4.3. Temperatura

4.3.1. Risultati kriging

La tabella sottostante illustra le principali statistiche descrittive delle serie di temperatura interpolate utilizzando il kriging con diversi semivariogrammi. Per comodità, la prima colonna contiene le statistiche della serie storica, già mostrate nel capitolo 2.

	Serie originale	Kriging con variogramma lineare - senza noise	Kriging con variogramma esponenziale - senza noise	Kriging con variogramma esponenziale – con noise
Numerosità	469.869	788.400	788.400	788.400
Media	13,10 °C	13,08 °C	13,08 °C	13,17 °C
Dev. St.	8,25 °C	8,31 °C	8,31 °C	8,18 °C
Min	-12,7 °C	-12,7 °C	-12,7 °C	-11,1 °C
q25%	6,5 °C	6,3 °C	6,3 °C	6,5 °C
q50%	12,8 °C	12,9 °C	12,9 °C	13,0 °C
q75%	19,4 °C	19,5 °C	19,5 °C	19,5 °C
Max	36,2 °C	36,2 °C	36,2 °C	35,9 °C

Tab. 18 - Principali statistiche descrittive delle serie interpolate.

A differenza del caso della pressione atmosferica, in questa tabella emerge una sottile differenza: in particolare si nota che il kriging non esatto ha valori estremi più moderati rispetto alle altre serie analizzate: ciò è coerente con lo smussamento operato dall'interpolazione non esatta, che non tocca necessariamente i valori noti della serie e tende a smorzare i valori estremi. Come nel paragrafo precedente, verranno mostrati a titolo d'esempio le interpolazioni effettuate con i tre metodi nella prima e nell'ultima settimana di osservazioni.

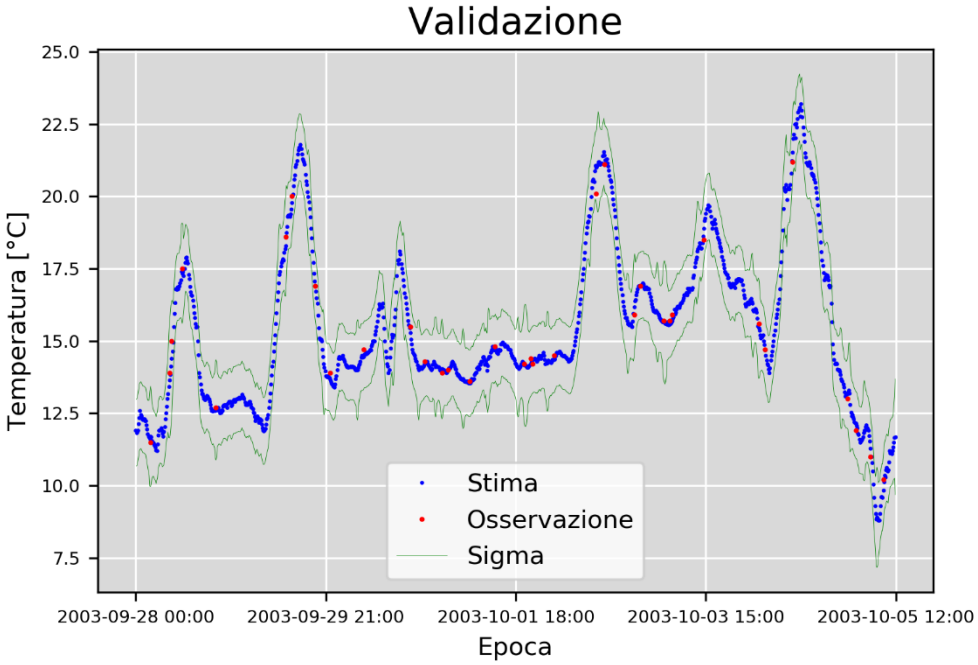

Fig. 54 - Interpolazione tramite kriging esatto con semivariogramma lineare; prima settimana di osservazioni.

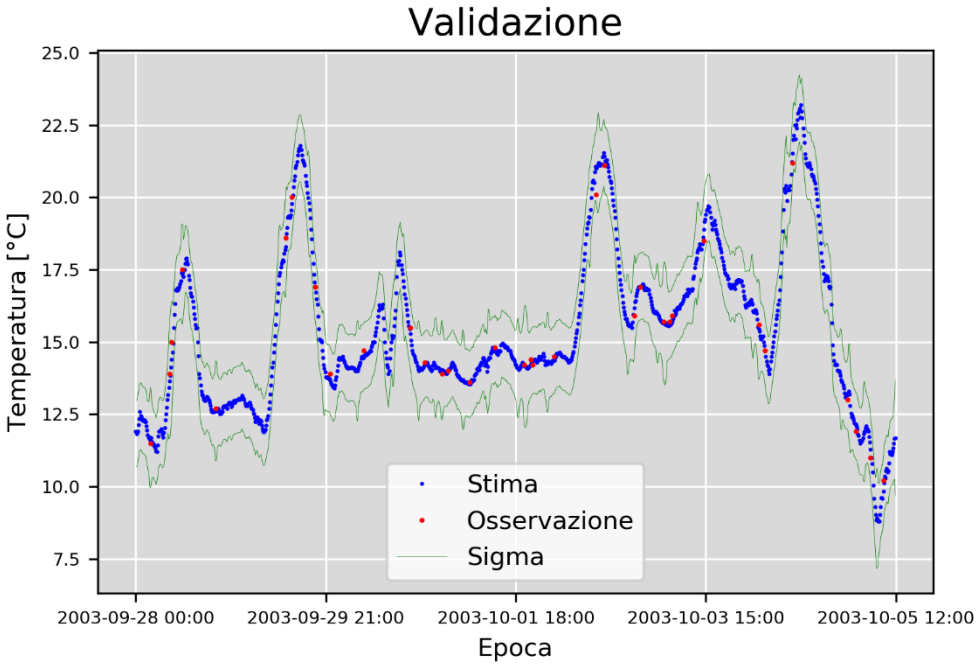

Fig. 55 - Interpolazione tramite kriging esatto con semivariogramma esponenziale.

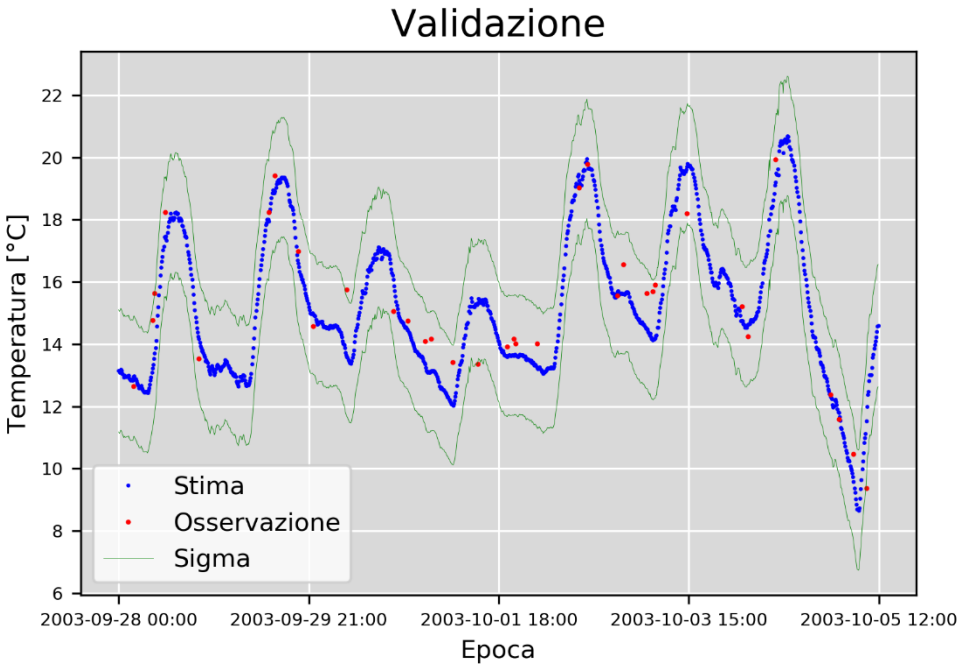

Fig. 56 - Interpolazione tramite kriging non esatto con semivariogramma lineare.

Come nel caso della variabile meteorologica precedente, l'interpolazione non esatta fornisce una stima più morbida a scapito dell'accuratezza: se si osserva l'intorno del 10 ottobre alle ore 18:00, infatti, si nota che l'interpolazione non esatta genera un picco che invece non dovrebbe esistere: nell'intorno di quel valore le osservazioni sono pressoché costanti, come indicato nei due grafici sopra. Una stima preliminare della bontà del fit è fornita tramite lo strumento grafico utilizzato anche in precedenza.

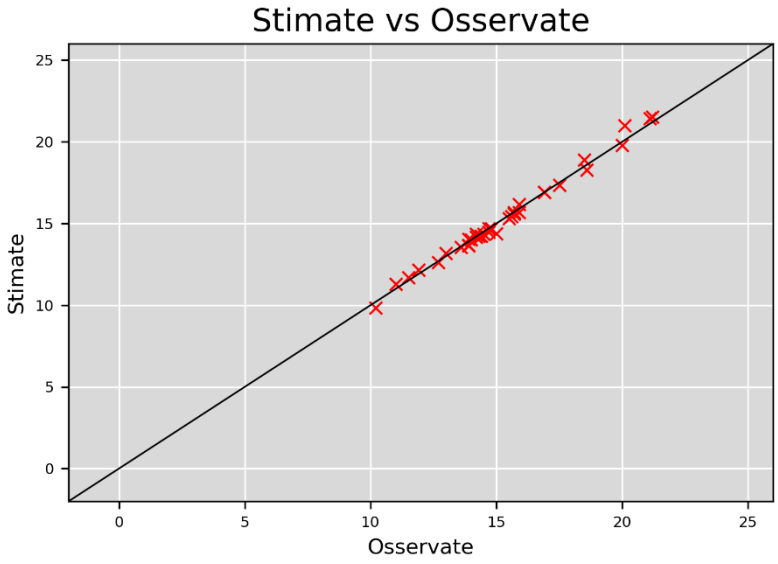

Fig. 57 - Osservazioni contro stime, kriging esatto con semivariogramma lineare; prima settimana di osservazioni.

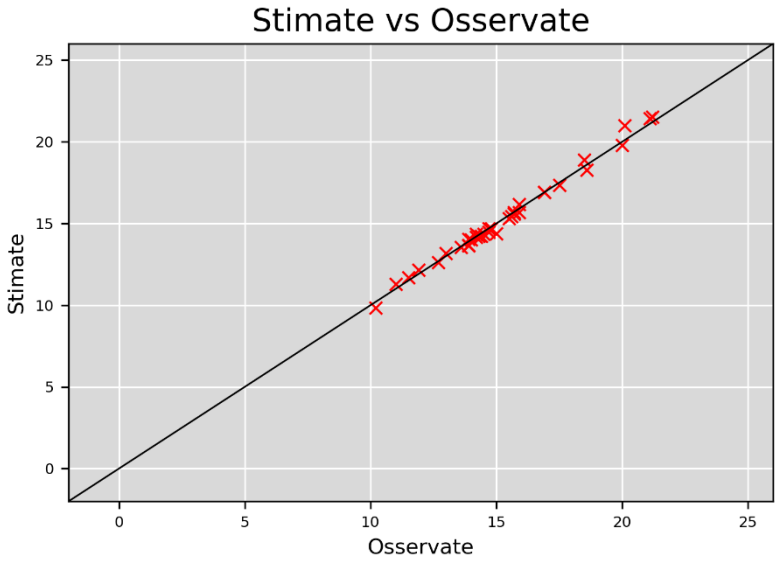

Fig. 58 - Osservazioni contro stime, kriging esatto con semivariogramma esponenziale.

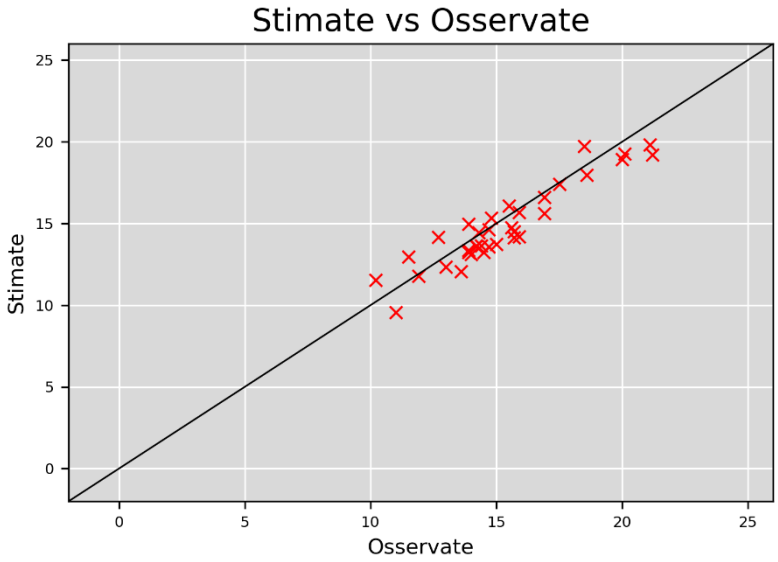

Fig. 59 - Osservazioni contro stime, kriging non esatto con semivariogramma lineare.

Analogamente a quanto già visto con la pressione atmosferica, il kriging non esatto genera un'interpolazione peggiore rispetto agli altri due metodi. Questo è valido anche per l'interpolazione effettuata sull'ultima settimana della serie temporale esaminata.

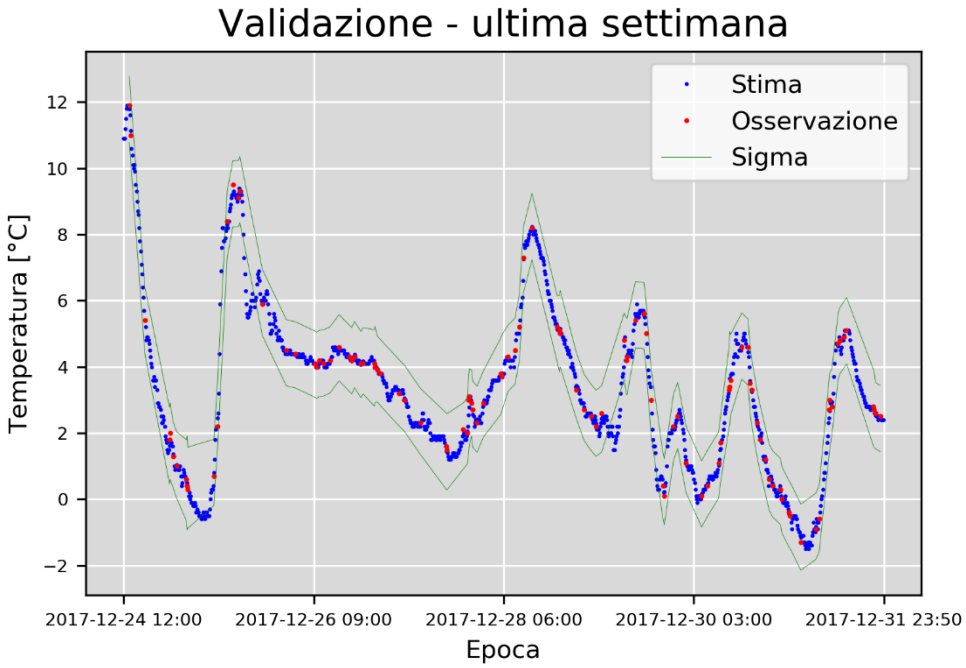

Fig. 60 - Interpolazione tramite kriging esatto con semivariogramma lineare; ultima settimana di osservazioni.

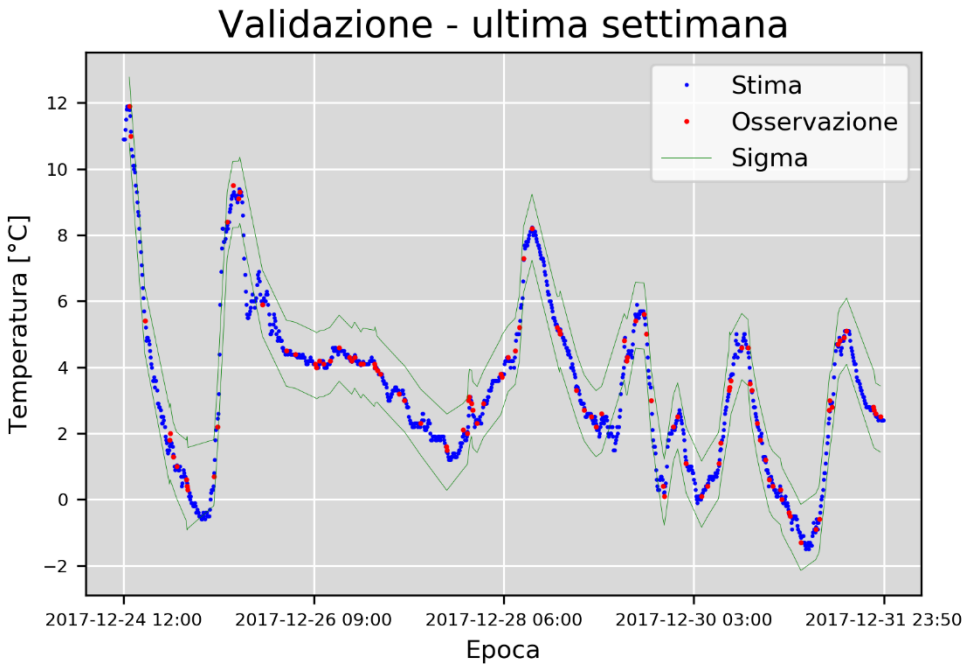

Fig. 61 - Interpolazione tramite kriging esatto con semivariogramma esponenziale.

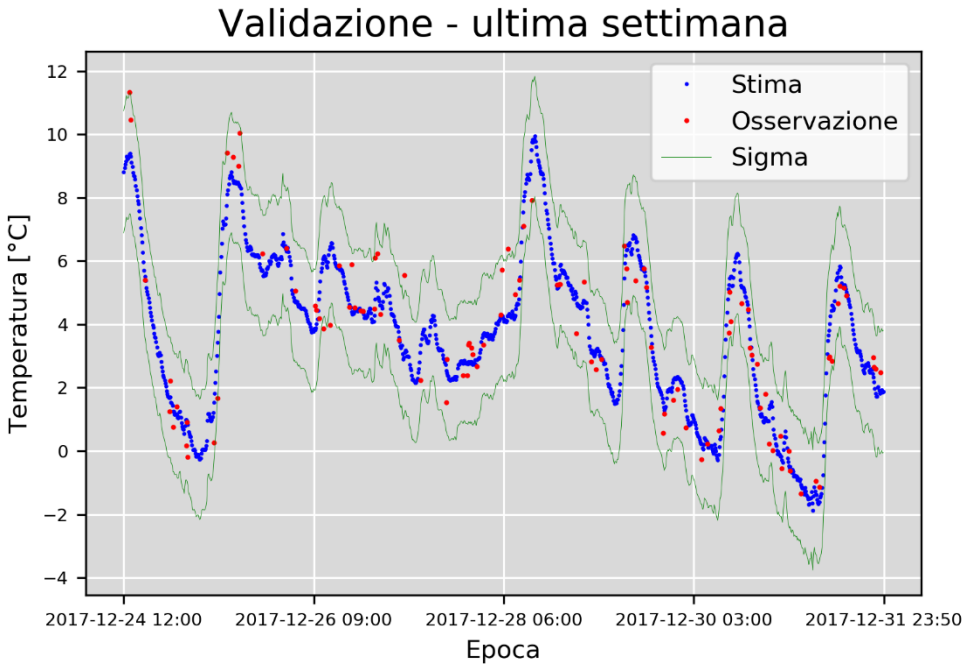

Fig. 62 - Interpolazione tramite kriging non esatto con semivariogramma lineare.

Anche in questo valgono le considerazioni esposte sopra, come pure nel paragrafo precedente, riguardo la bontà del fitting dei tre metodi: il kriging non esatto con semivariogramma lineare risulta avere le performance peggiori, come confermato dal grafico in fig.62.

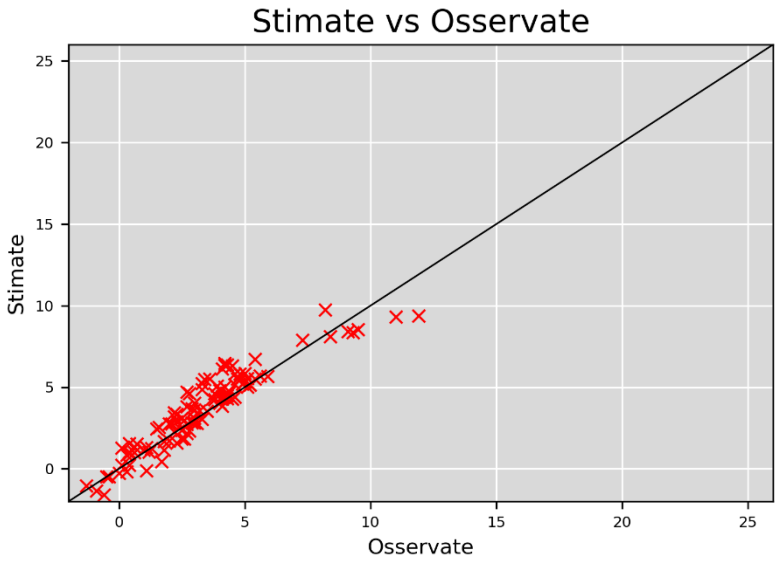

Fig. 63 - Osservazioni contro stime, kriging non esatto con semivariogramma lineare; ultima settimana di osservazioni.

La verifica dei percentili non fornisce, tuttavia, ulteriori informazioni rilevanti: tutti i metodi sembrano determinare una distribuzione delle stime praticamente identica a quella della serie storica. In tabella vengono riportati i risultati numerici, mentre più in basso si trovano tutti gli altri grafici utilizzati per la valutazione delle performance.

Quantile	Serie originale	Kr. lineare	Kr. esponenziale	Kr. non esatto
1%	-2,5 °C	-2,6 °C	-2,6 °C	-2,1 °C
99%	30,7 °C	30,4 °C	30,4 °C	30,3 °C

Tab. 19 - 1° e 99-esimo quantile per tutte le serie esaminate.

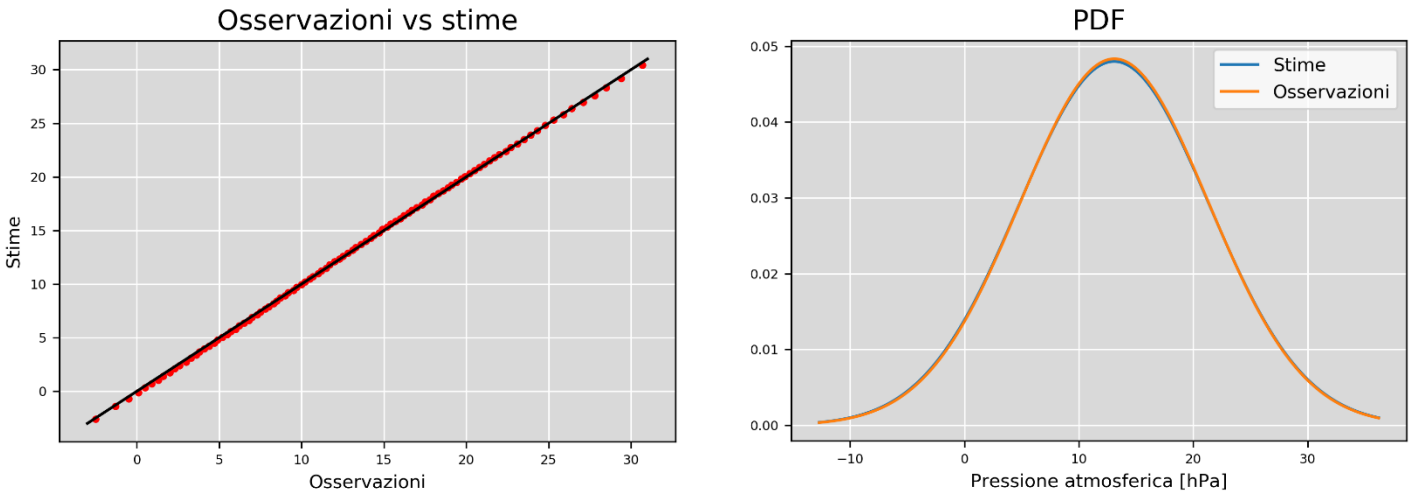

Fig. 64 - Kriging esatto con semivariogramma lineare.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

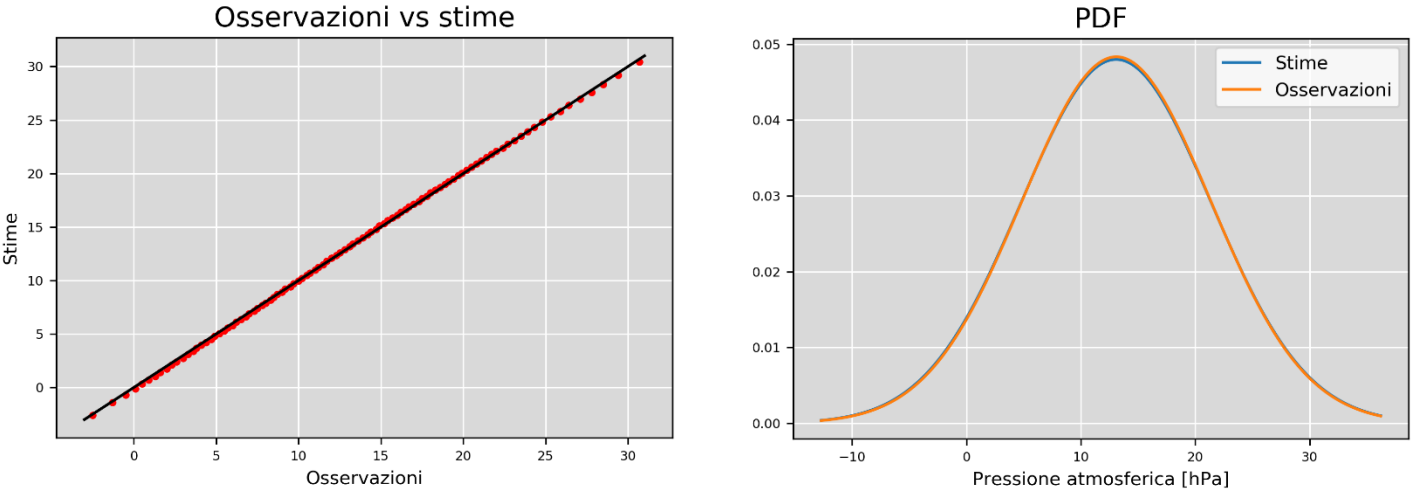

Fig. 65 - Kriging esatto con semivariogramma esponenziale.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

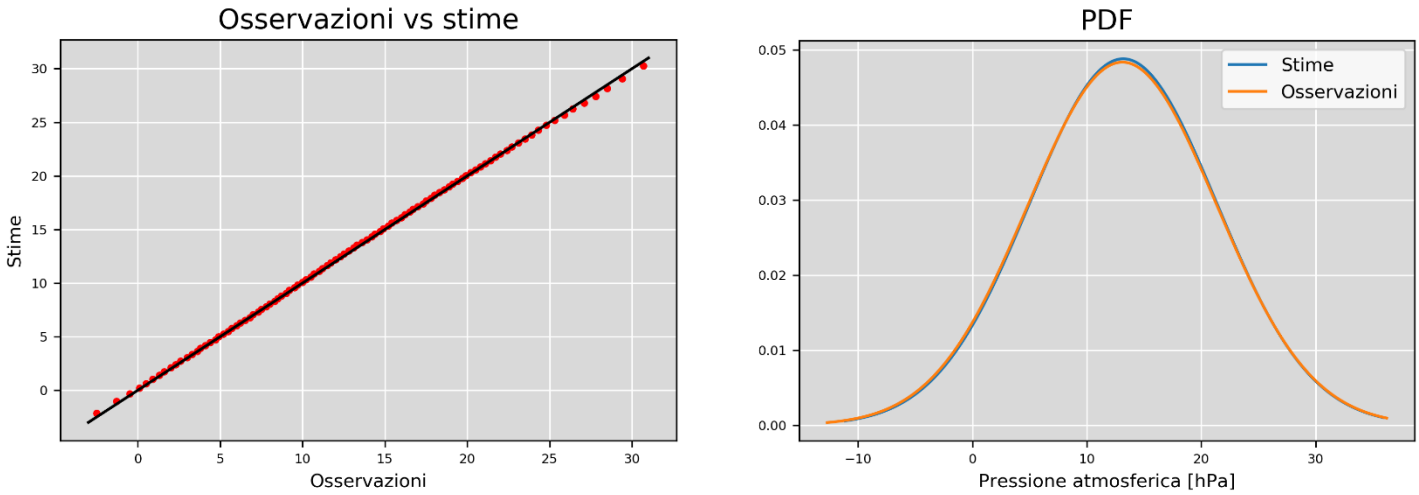

Fig. 66 - Kriging non esatto con semivariogramma lineare.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

L'analisi di questi grafici, purtroppo, risulta inconcludente, poiché non è possibile ricavarne informazioni interessanti sui diversi metodi: le differenze tra i diversi grafici sono impercettibili. L'unica informazione di interesse che emerge è che tutti gli algoritmi sembrano ugualmente validi, dato che le distribuzioni dei quantili delle serie interpolate sembrano molto simili a quelle della serie storica.

4.3.2. Risultati RBF-Network

I numeri in tab.20 mostrano le principali statistiche descrittive delle serie generate dai tre tipi di RBFN, più quelle della serie originale.

	Serie originale	RBFN con kernel multiquadric	RBFN con kernel lineare	RBFN con kernel gaussiano
Numerosità	469.869	788.400	788.400	788.400
Media	13,10 °C	13,08 °C	13,08 °C	13,09 °C
Dev. St.	8,25 °C	8,31 °C	8,31 °C	8,27 °C
Min	-12,7 °C	-12,7 °C	-12,7 °C	-13,2 °C
q25%	6,5 °C	6,3 °C	6,3 °C	6,4 °C
q50%	12,8 °C	12,9 °C	12,9 °C	12,9 °C
q75%	19,4 °C	19,5 °C	19,5 °C	19,5 °C
Max	36,2 °C	36,2 °C	36,2 °C	36,2 °C

Tab. 20 - Principali statistiche descrittive delle serie interpolate.

La tabella stabilisce che esistono alcune differenze tra le diverse statistiche, ma sono minime: non si va oltre il mezzo grado di differenza in nessuna colonna. Si prosegue, quindi, nell'analisi dei risultati mostrando come performano le tre RBFN nell'interpolazione delle osservazioni della prima e dell'ultima settimana.

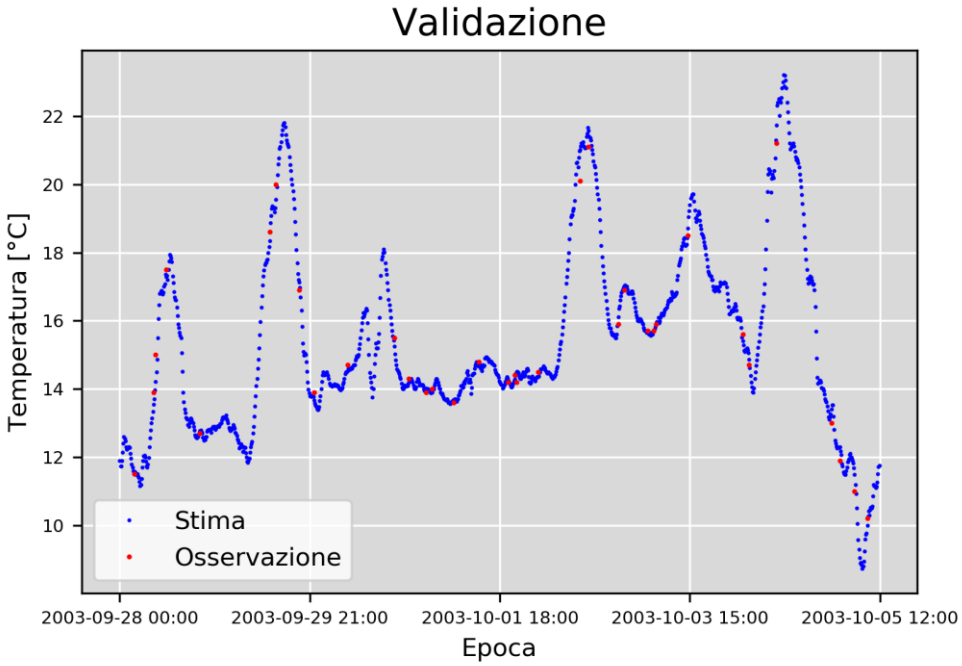

Fig. 67 - Osservazioni interpolate con RBFN, kernel multiquadric; prima settimana di osservazioni.

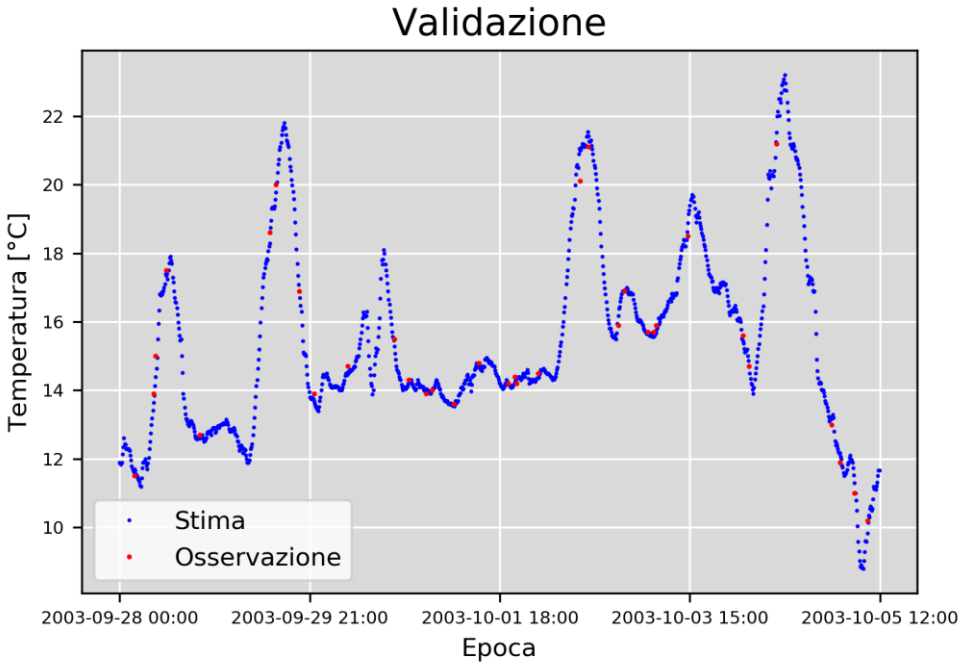

Fig. 68 - Osservazioni interpolate con RBFN, kernel lineare.

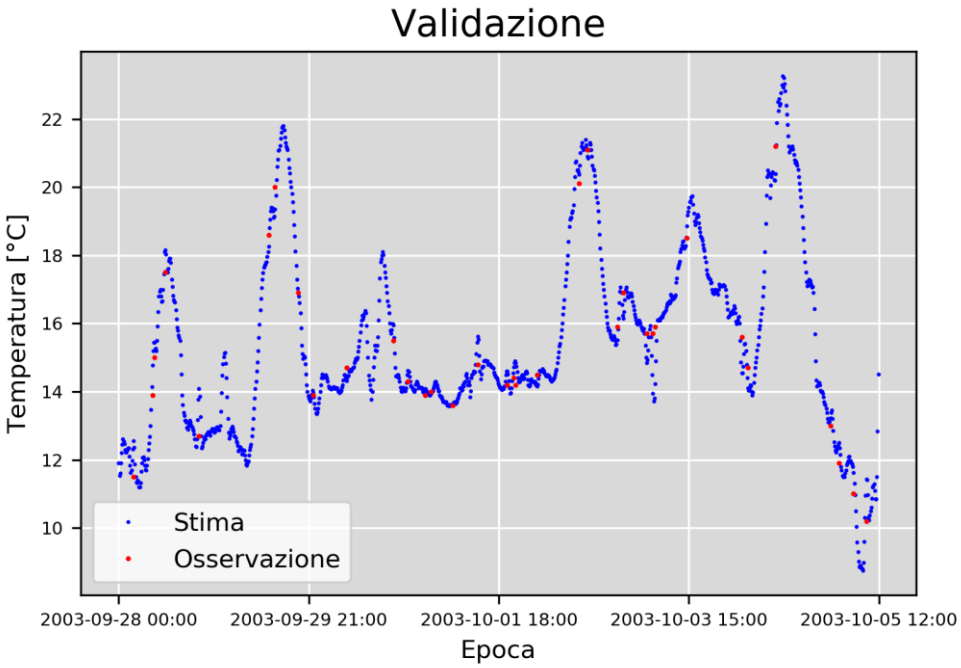

Fig. 69 - Osservazioni interpolate con RBFN, kernel gaussiano.

Anche se l'effetto è, in questo caso, meno marcato, si può notare come il grafico dell'interpolazione con kernel gaussiano dia una soluzione più sparsa rispetto agli altri due metodi. Il motivo è probabilmente, da ricercarsi nell'anomalo comportamento di questo tipo di kernel che tende ad appiattire i valori sullo zero. Questo trend, tuttavia, non è visibili nei grafici mostrati di seguito, tuttavia è interessante notare che le performance della RBFN ad attivazione gaussiana sono comunque peggiori delle altre reti.

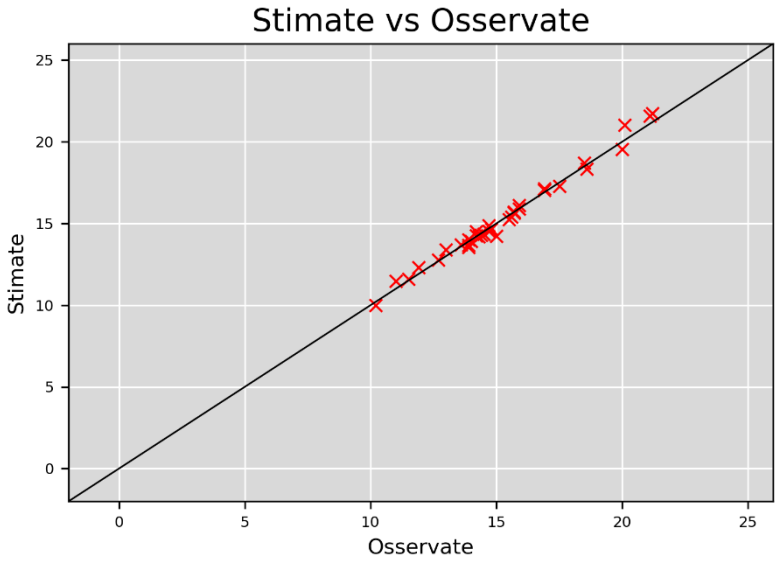

Fig. 70 - Osservazioni contro stime, RBFN con kernel multiquadric; prima settimana di osservazioni.

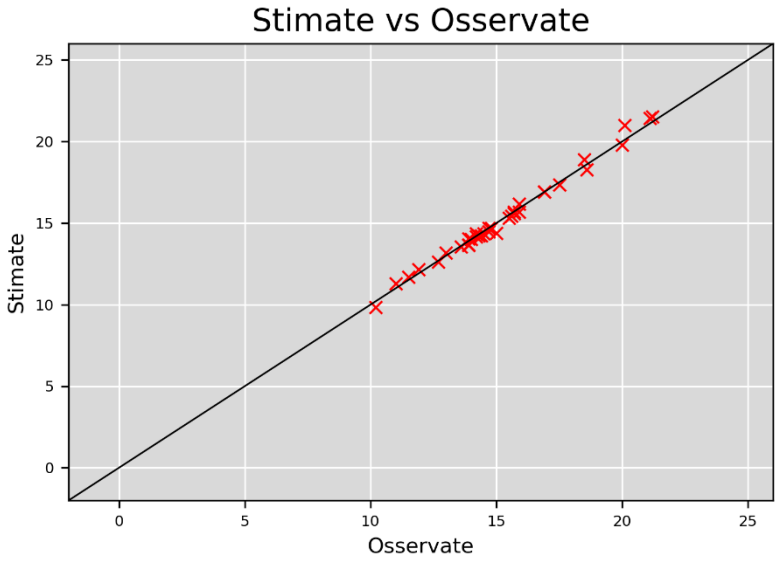

Fig. 71 - Osservazioni contro stime, RBFN con kernel lineare.

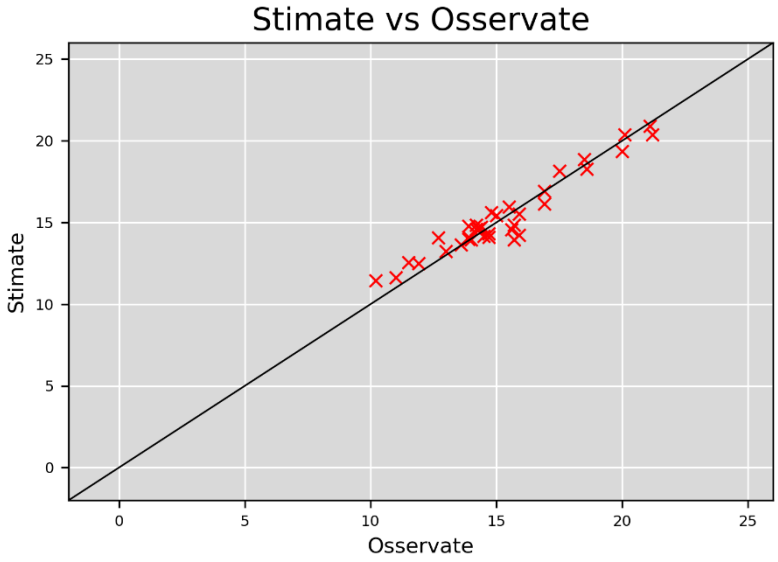

Fig. 72 - Osservazioni contro stime, RBFN con kernel gaussiano.

I risultati relativi all'ultima settimana di osservazioni non forniscono nessuna informazione aggiuntiva, poiché il comportamento delle reti è del tutto analogo a quello tenuto nei grafici appena esposti. Vengono tuttavia inseriti per completezza.

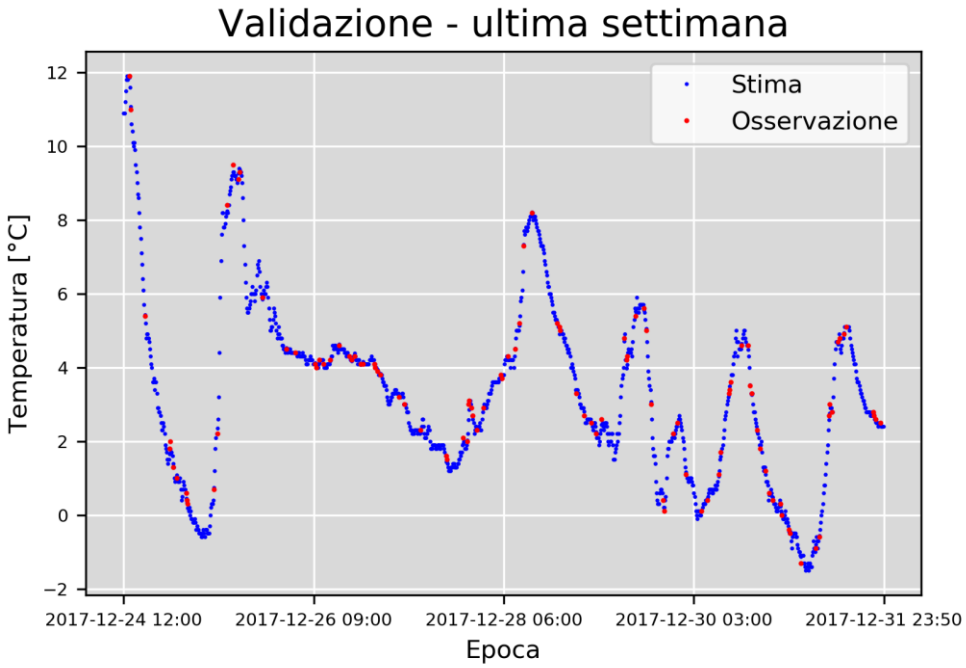

Fig. 73 - Osservazioni contro stime, RBFN con kernel multiquadric; ultima settimana di osservazioni.

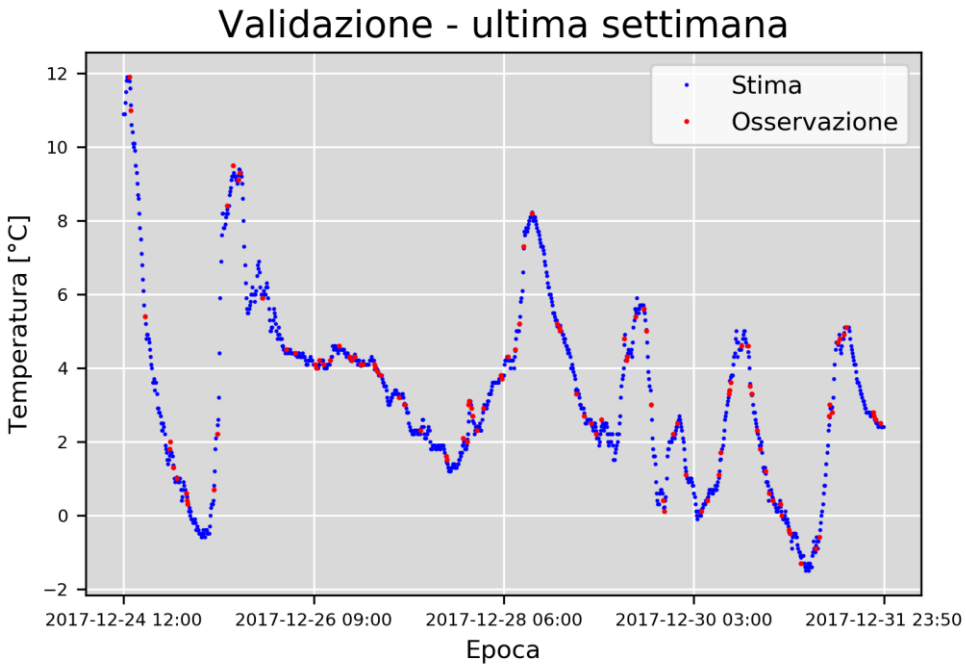

Fig. 74 - Osservazioni contro stime, RBFN con kernel lineare.

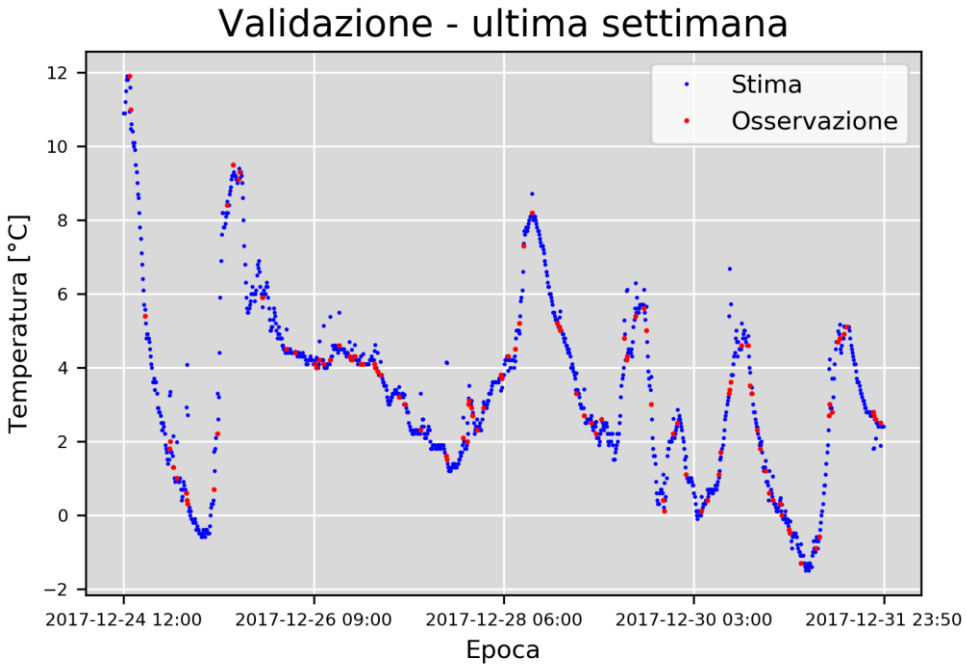

Fig. 75 - Osservazioni contro stime, RBFN con kernel gaussiano.

Piuttosto che mostrare il grafico delle osservate contro le stime per l'ultima settimana, in questo caso risulta più interessante lo stesso tipo di grafico ma applicato ai residui delle interpolazioni dell'ultimo mese. Osservare l'output grezzo del modello, in questo caso, fornisce informazioni più interessanti che si perdono nel passaggio all'output reale, ovvero aggiungendo la media istantanea all'output grezzo.

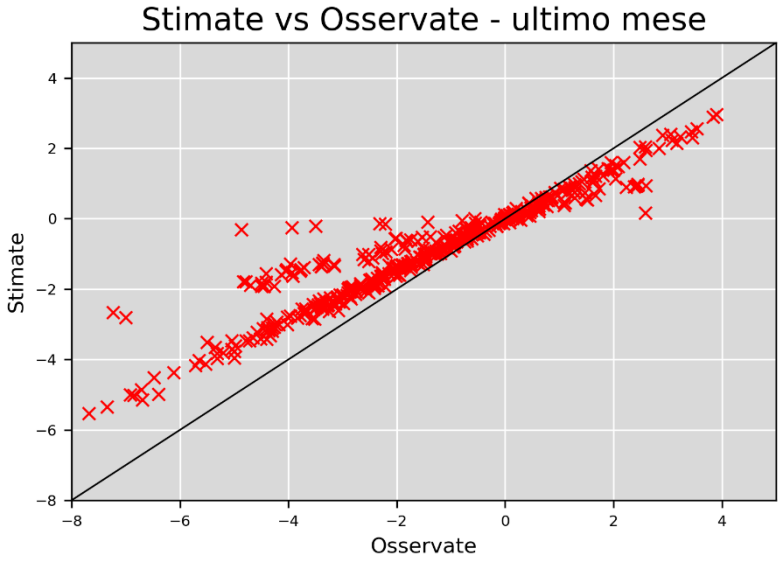

Fig. 76 - Osservate contro stime dei residui, ultimo mese; interpolazione tramite RBFN con attivazione gaussiana.

È interessante notare che l'errore compiuto sembra sistematico, come nel caso della pressione atmosferica: si distinguono bene due direttrici che determinano la sovrastima dei valori più bassi e la sottostima dei valori più alti. In alcuni casi sembra quasi che non ci sia correlazione tra i valori osservati e quelli stimati.

Vengono ora illustrate le rimanenti statistiche utilizzate per la valutazione della bontà dei fitting: primo e ultimo percentile, grafici dei percentili della serie storica contro quelli delle serie interpolate e grafici delle funzioni densità di probabilità.

Quantile	Serie originale	RBFN con kernel multiquadric	RBFN con kernel lineare	RBFN con kernel gaussiano
1%	-2,5 °C	-2,6 °C	-2,6 °C	-2,5 °C
99%	30,7 °C	30,4 °C	30,4 °C	30,4 °C

Tab. 21 - 1° e 99-esimo quantile per tutte le serie esaminate.

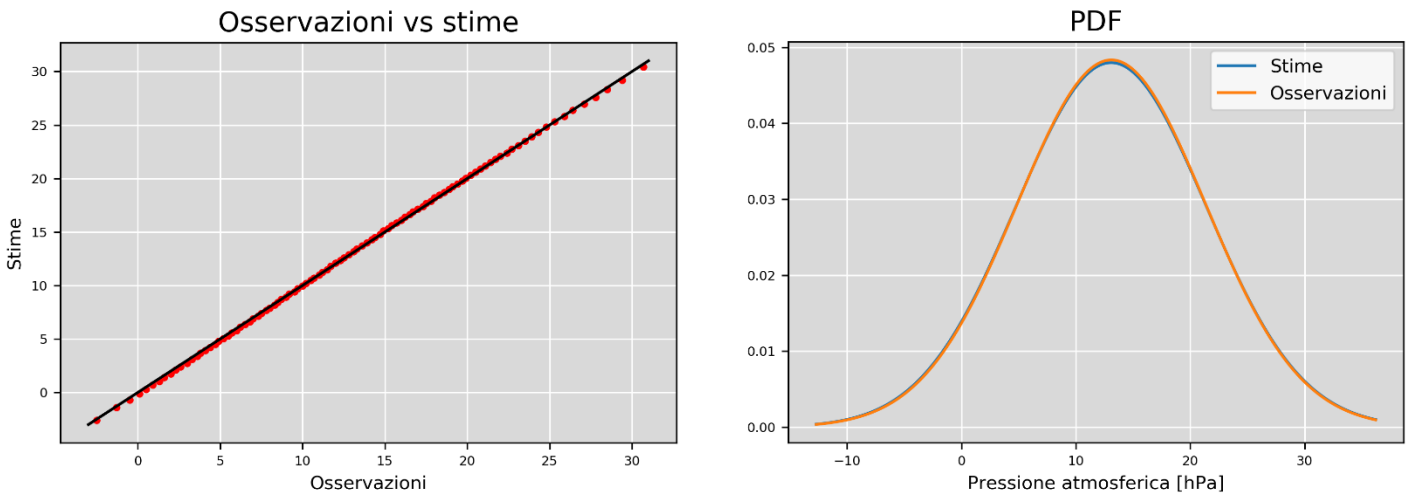

Fig. 77 - RBFN con kernel multiquadric.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

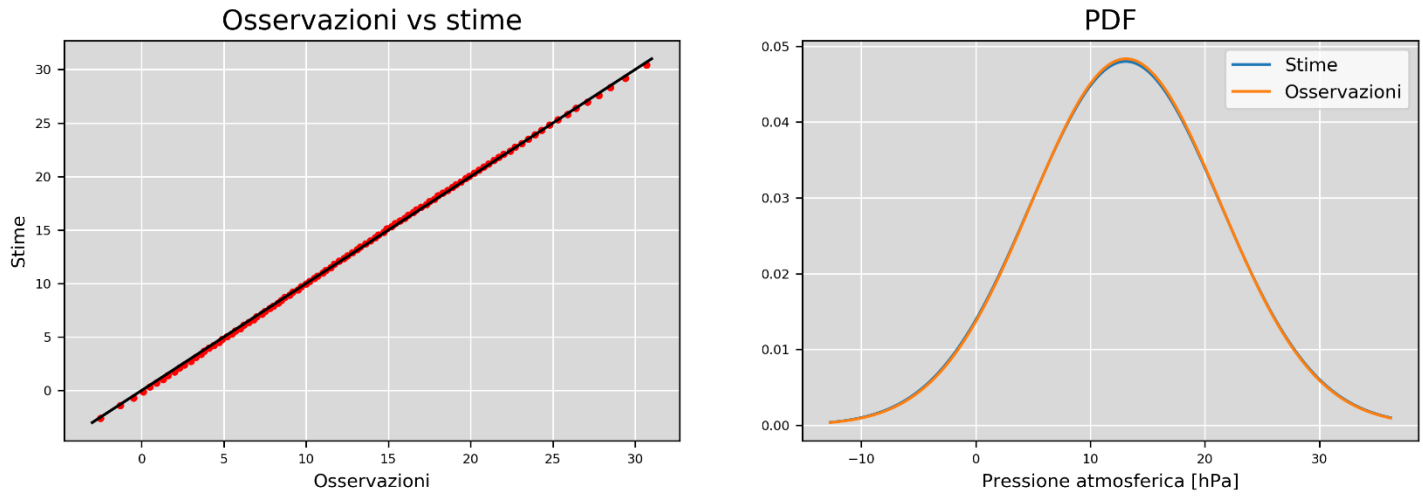

Fig. 78 - RBFN con kernel lineare.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

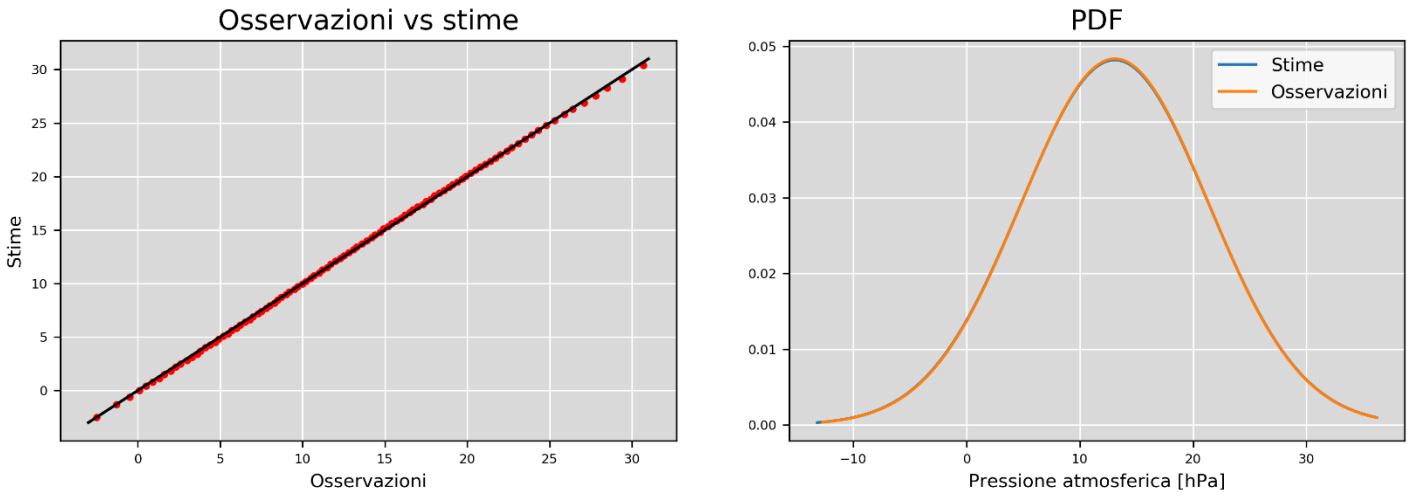

Fig. 79 - RBFN con kernel gaussiano.
A sinistra: confronto tra i quantili interpolati e quelli originali; a sinistra: funzioni densità di probabilità delle stesse serie.

Anche in questo caso, come per il kriging, non è possibile estrapolare alcuna informazione utile dai grafici appena mostrati, se non l'ottima aderenza delle distribuzioni interpolate con quella storica.

4.3.3. Ulteriori considerazioni

L'ultimo confronto è stato effettuato osservando i valori medi del coefficiente di determinazione e dei tempi di calcolo richiesti da ciascun algoritmo.

	Kriging esatto lineare	Kriging esatto esponenziale	Kriging non esatto lineare	RBFN multiquadric	RBFN lineare	RBFN gaussiana
R^2 medio	0,991	0,991	0,897	0,989	0,990	0,952
Minuti	8,59	8,89	54,04	0,73	0,68	1,61

Tab. 22 - Coefficienti di determinazione e tempi di calcolo di ciascun algoritmo testato.

Nel caso della temperatura, l'algoritmo con le performance peggiori è il kriging non esatto con semivariogramma lineare: ha contemporaneamente il coefficiente di determinazione minore e il maggiore tempo di calcolo. La RBFN con attivazione gaussiana ha un valore di R^2 comunque inferiore agli altri metodi, quindi verrà scartato. Tra gli algoritmi rimanenti, i due kriging esatti hanno tempi di calcolo di circa 8 minuti, superati in termini di prestazioni dalle due RBFN che hanno tempi di calibrazione e predizione inferiori al minuto. Tra questi ultimi due, quello con il coefficiente di determinazione più alto è, come nel caso della pressione atmosferica, la RBFN con kernel lineare.

5. Conclusioni

Scopo di questo progetto di tesi è l'identificazione, tra 2 diversi principi di interpolazione e diverse parametrizzazioni, di un algoritmo che garantisca un minimo errore di predizione per l'interpolazione di serie storiche di variabili meteorologiche. I dati analizzati derivano da sensori di pressione atmosferica e temperatura di proprietà di ARPA Lombardia. Il lavoro di tesi si inserisce, infatti, nel progetto LAMPO: una collaborazione tra Politecnico di Milano, ARPA Lombardia, lo spin-off del Politecnico GReD e l'Università di Padova.

L'algoritmo di predizione verrà utilizzato per rendere le diverse serie temporali, ove possibile, omogenee rispetto all'intervallo di campionamento, riempendo eventuali 'no-data'. La scelta dell'algoritmo ha richiesto quindi una preliminare classificazione delle serie temporali oggetto dell'interpolazione per caratterizzare i diversi tipi di 'no-data' in base alla lunghezza del periodo in cui le osservazioni mancano e alla distribuzione dei tali periodi all'interno della serie temporale. Si è deciso di confrontare il kriging, un algoritmo di predizione classico della geostatistica, con un algoritmo che deriva dal machine learning, come le reti neurali artificiali.

La scelta degli algoritmi è stata dettata dalla mia curiosità rispetto alle due tecniche, apprese durante il mio corso di studi, che sebbene diverse per la modellizzazione del dato, in un caso stocastica, nell'altro deterministica, e nell'utilizzo nella prima di un principio di minimizzazione della varianza dell'errore di predizione, assente nella seconda, basata sulla possibilità di "addestrare un algoritmo" sulla base di un sufficiente numero di problemi di interpolazione, presentano una apparente similarità. Questa è relativa alla determinazione di pesi da dare alle osservazioni prima di combinarle per ottenere il valore predetto. Nel primo caso attraverso la stima empirica del variogramma dei dati, nel secondo nella scelta di una funzione di peso nella combinazione non lineare delle osservazioni. Nello specifico, il kriging è stato utilizzato per compiere interpolazioni esatte e non esatte (ovvero assumendo che i dati fossero affetti da rumore bianco non correlato con le osservazioni) con semivariogrammi lineari o esponenziali; le reti neurali artificiali scelte, invece, appartengono alla classe delle Radial Basis Function Network e sono state utilizzate per compiere interpolazioni esatte, verificando le prestazioni di kernel lineari, multiquadric e gaussiani.

Dall'analisi dei risultati è emerso che il modello con le prestazioni migliori in termini di coefficiente di determinazione e tempi di calcolo è la RBF-Network con funzioni di attivazione lineari, per entrambe le variabili meteorologiche analizzate. A parità di bontà dell'interpolazione, infatti, il la stima empirica del variogramma nel kriging richiede tempi di calcolo maggiori. Le RBF-Network,

create come approssimatori universali, sono in grado in questo caso di risolvere il problema in maniera rapida e sufficientemente accurata.

I risultati ottenuti saranno utilizzati per l'interpolazione di serie storiche di altre variabili meteorologiche così come di serie temporali di osservazioni di contenuto di acqua precipitabile in atmosfera ottenuto dall'uso innovativo di stazioni GNSS a basso costo nell'ambito del progetto LAMPO. Il progetto, infatti, prevede l'utilizzo di una rete neurale per effettuare predizioni di piogge intense nell'area test del Bacino del Seveso in Lombardia. Basato sulla disponibilità di serie temporali di osservazioni metereologiche (dal 2010 al 2017), su una analisi approfondita delle condizioni meteorologiche cha hanno portato alla generazione di piogge intense e localizzate e sull'uso innovativo di osservazioni diffuse e continue della quantità di acqua precipitabile nella troposfera tramite sensori GNSS a basso costo, il training di tale rete neurale richiede una lunga attività di preprocessing dei dati per renderli omogenei e privi di 'buchi'.

6. Bibliografia e sitografia

1. P. Benevides, João Catalão, Giovanni Nico, Pedro M. A. Miranda, "Evaluation of rainfall forecasts combining GNSS precipitable water vapor with ground and remote sensing meteorological variables in a neural network approach," Proc. SPIE 10786, Remote Sensing of Clouds and the Atmosphere XXIII, 1078607 (9 October 2018).
2. Maria Cleofe´ Valverde Ramirez, Haroldo Fraga de Campos Velho, Nelson Jesus Ferreira, "Artificial neural network technique for rainfall forecasting applied to the Sao Paulo region", Journal of Hydrology 301 (2005) pagg. 146–162.
3. Costa, Jean-Pierre, Luc Pronzato, and Eric Thierry. "A comparison between Kriging and radial basis function networks for nonlinear prediction." NSIP. 1999.
4. Rusu, Cristian, and Virginia Rusu. "Radial basis functions versus geostatistics in spatial interpolations." *IFIP International Conference on Artificial Intelligence in Theory and Practice*. Springer, Boston, MA, 2006.
5. www.arpalombardia.it