Tecniche di analisi spaziale per la ricostruzione delle serie storiche di dati climatici

Enrico BONO, Goffredo LA LOGGIA, Leonardo V. NOTO

Dipartimento di Ingegneria Idraulica ed Applicazioni Ambientali, Università di Palermo, viale delle Scienze, 90100 Palermo, tel 091 6657726, *fax* 091 6657719, *e-mail* enrico@lifejackets.it, glal@idra.unipa.it, valerio@idra.unipa.it

Abstract
La disponibilità di dati pluviometrici affidabili è fondamentale per la maggior parte delle analisi ideologiche. In particolare, la stima della distribuzione spaziale della precipitazione è essenziale per stabilire il bilancio idrologico di un bacino e per la valutazione delle portate di piena. Sono state proposte diverse metodologie per la stima dei campi di precipitazione a partire da misure pluviometriche puntuali. In questo studio vengono messi a confronto vari algoritmi di interpolazione spaziale utilizzati per la stima dei campi di precipitazione in Sicilia. Gli algoritmi utilizzati comprendono modelli geostatistici, quali il *Kriging Ordinario*, il *Kriging Ordinario dei Residui*, l'*Inverse Distance Weighting* e diversi modelli regressivi, come la Regressione Lineare Semplice, la Regressione Multipla e la *Geographically Weighted Regression*. Alcuni di questi metodi fanno uso dell'elevazione per la stima spaziale dei campi di precipitazione, attraverso un DEM. Questi algoritmi sono applicati ai dati di precipitazione media mensile ed annuale in Sicilia, dati misurati in 248 stazioni pluviometriche su una superficie di circa 25.000 km^2. Le stime ottenute con i vari algoritmi vengono comparate tra loro utilizzando la validazione con un *dataset* indipendente non usano nella taratura dei modelli. I modelli spaziali utilizzati sono stati combinati con alcuni modelli temporali come i modelli Autoregressivi Periodici e i processi di generazione di Campi Casuali.

Abstract
The availability of good and reliable rainfall data are fundamental for the most of the hydrological analysis. Particularly the prediction of the spatial distribution of rainfall is essential for establishing the water balance and for the assessment of flood discharge in a basin. Many methods have been proposed to estimate rainfall fields starting from point measurements. This paper compares many of spatial interpolation algorithms in order to estimate the rainfall fields of Sicily. The used algorithms are geostatistical models as Ordinary Kriging, Residual Ordinary Kriging, Inverse Distance Weighting and several regressive model, as Simple Linear Regression, Multiple Regression and Geographically Weighted Regression. Some of these methods incorporate the elevation into the spatial prediction of rainfall fields by means of the use of a DEM. These algorithms are applied to annual and monthly average rainfall data of Sicily (Italy), measured at 248 rain gages on a surface of about 25,000 sqkm. Rainfall prediction produced with these algorithms are compared using validation on a independent dataset not used in models development. These spatial interpolation models are combined with many temporal methods as Periodic Autoregressive Model and Random Field Generation processes.

1 Introduzione

Le serie storiche di variabili meteoclimatiche sono alla base di numerosi modelli, utilizzati in diverse discipline. Per esempio i dati di precipitazione sono di primaria importanza per la corretta gestione della risorsa idrica o nello studio degli eventi di piena. Purtroppo è praticamente impossibile disporre di misure continue e complete. In genere vengono utilizzati solo record continui oppure si ignorano i dati mancanti. Con il primo approccio si perdono dati e i processi statistici sui dati rimanenti non possono che portare a risultati non veritieri. Con il secondo approccio, d'altra parte, si riduce il periodo degli eventi registrati, cosa che porta a sovrastimare la frequenza degli eventi straordinarî. D'altra parte utilizzare le serie storiche discontinue, senza alcuna ricostruzione dei dati mancanti, porta ad errori che presentano pattern spaziali e temporali (Stooksbury et al., 1999). L'alternativa valida agli approcci sopra citati consiste nello stimare i valori mancanti della variabile climatica. Vari autori si sono dedicati al problema dell'integrazione delle serie storiche, tra cui Tang, Eisched. Il primo autore si dedica alla ricostruzione dei dati pluviometrici mancanti in Malesia (Tang et al.,1996). Eisched (Eisched et al., 2000) ricostruisce le serie storiche di precipitazione e di temperature negli Stati Uniti Occidentali. In tutti gli studi si utilizzano vari algoritmi di interpolazione spaziale (Matheron, 1969), che permettono di determinare il dato mancante in una stazione in base ai valori registrati nelle stazioni limitrofe.

In questo studio si sono utilizzati diversi metodi di *interpolazione spaziale* per ricolmare le serie discontinue di dati pluviometrici registrati in 247 stazioni della Sicilia in 80 anni (***Errore. L'origine riferimento non è stata trovata.***). Per verificare e confrontare il comportamento dei vari metodi viene utilizzata la *validazione*.

Sulla base dati è stato effettuato un controllo a livello annuale, per evidenziare ed eventualmente eliminare dalla base dati valori anomali ed *outliers*. Questa ricerca di outliers è stata effettuata marcando come *dati dubbi* tutti i valori annuali che cadevano fuori da un opportuno intervallo costruito attorno alla media dei dati (Feng et al., 2004). Sono state utilizzate due metodologie:

1. per fissato anno, evidenziando anomalie nel *pattern* spaziale;
2. per fissata stazione, evidenziando anomalie temporali.

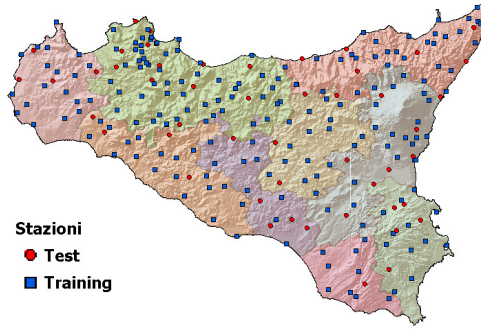

Figura 1 – Localizzazione delle 247 stazioni pluviometriche studiate

Le anomalie nello spazio e/o nel tempo sono state determinate con due criteri; il primo, più restrittivo, ha il seguente intervallo costruito sulla media dei dati:

$$(\mu - 3\sigma; \mu + 3\sigma) \qquad [1]$$

dove μ è la media (nello spazio o nel tempo) e σ è la deviazione standard. Per il secondo criterio, l'intervallo varia in base alla dimensione campionaria:

$$(\mu - K_n\sigma; \mu + K_n\sigma) \quad \text{con} \quad K_n = -0.9043 + 3.345\sqrt{\log(n)} - 0.4046\log(n) \qquad [2]$$

I dati evidenziati sono stati controllati direttamente sui documenti idrografici originali e quindi corretti o eliminati dalla base di dati.

2 Metodi di interpolazione spaziale per la ricostruzione del dato pluviometrico

L'interpolazione spaziale è una delle operazioni di analisi spaziale più diffuse. Data una grandezza che varia con continuità nello spazio, sia essa la quota del terreno, la precipitazione, la temperatura o la concentrazione di un gas inquinante nell'atmosfera, con l'interpolazione spaziale si cerca di fornire una stima di questa grandezza in punti in cui la stessa non è stata misurata. Questa stima viene fatta in tenendo presente i valori che la grandezza in esame assume in alcuni punti di campionamento. I processi di interpolazione si basano sulla legge di Tobler, secondo la quale "*tutti i siti sono correlabili, ma quelli più vicini sono più correlabili di quelli più lontani*". In quest'ottica

la migliore stima di una variabile in un determinato punto è più simile al valore misurato nel punto di campionamento più vicino.
Una delle possibili classificazioni degli interpolatori spaziali viene fatta differenziando i metodi che stimano la precipitazione utilizzando esclusivamente i dati pluviometrici registrati dai metodi che invece utilizzano anche altre informazioni, in primo luogo la quota del terreno.

2.1 Stima con utilizzo del solo dato di precipitazione

Tra i metodi afferenti al primo gruppo, troviamo l'interpolatore *IDW* (*Inverse Distance Weighting*), uno dei metodi più semplici. La stima P^* in ogni locazione $\underline{u}_0$ è la media dei valori P_i osservati in n punti vicini $\underline{u}_i$, pesati con l'inverso della distanza h_i tra $\underline{u}_i$ e $\underline{u}_0$. L'esponente β in genere è pari a 2.

$$P^*_{IDW}(\underline{u}_0) = \frac{1}{\sum_{i=1}^{n} \lambda_i} \sum_{i=1}^{n} [\lambda_i P_i] \quad \text{con} \quad \lambda_i = \frac{1}{h_i^{\beta}} \qquad [3]$$

L'idea alla base di questo metodo è che osservazioni P misurate in stazioni più vicine tra loro sono più simili rispetto alle precipitazioni misurate in punti più distanti tra loro; di conseguenza le osservazioni più vicine avranno un peso maggiore. Nei metodi geostatistici questa diversità tra i dati, e cioè il peso da attribuire ad ogni osservazione, viene misurata non dalla distanza euclidea, ma dalla semivarianza. Il primo metodo geostatistico usato è il *Kriging Ordinario* (Goovaerts, 1999).

2.2 Stima con utilizzo dell'elevazione

La *regressione lineare* è il primo dei metodi che utilizzano il dato morfologico nella stima delle piogge. Con questo metodo si suppone, infatti, che esista una relazione lineare tra la quota del terreno $Q(\underline{u})$ e la precipitazione, relazione espressa dalla [4].

$$P^*_{RL}(\underline{u}_0) = mQ(\underline{u}_0) + k \qquad [4]$$

I parametri m e k della retta interpolante vengono determinati effettuando una regressione lineare ai minimi quadrati ordinari tra i dati pluviometrici del *dataset* di *training* e le quote sul livello del mare delle stazioni di misura. Oltre alla regressione lineare $P - Q$ vengono anche studiate le relazioni $log(P) - Q$, $\sqrt{P} - Q$ e $log(P) - log(Q)$.
La regressione lineare ai minimi quadrati ordinari ha il difetto di tenere in egual considerazione tutti i punti, senza preoccuparsi di eventuali *outliers*, cioè misure affette da errori di misura o da errori sistematici. Con la *regressione robusta* invece, possiamo dare un peso diverso ai diversi punti del dataset; con un processo iterativo è possibile quindi trascurare gli *outliers* (*Figura 3*).
Un altro metodo regressivo è la *Geographically Weighted Regression*, nella quale i coefficienti m e k della retta di regressione non sono più unici per tutto il territorio, ma variano stazione per stazione (Brunsdon, 2001). La stima della precipitazione diventa quindi:

$$P^*_{GWR}(\underline{u}_0) = m(\underline{u}_0)Q(\underline{u}_0) + k(\underline{u}_0) \qquad [5]$$

L'estensione della regressione semplice è la *regressione multipla*, con la quale si cerca la relazione tra la variabile dipendente, in questo caso la precipitazione P, e altre variabili indipendenti. Come variabili indipendenti sono state usate la *quota del terreno* (indicata con DEM), estratta da un *digital terrain model* della Sicilia a risoluzione di 100x100m, e cinque sue trasformazioni (DEM, DEM^2, DEM^3, log(DEM), ln(DEM)); inoltre sono state inserite anche le variabili *pendenza del terreno* (SLP) e *esposizione* (ASP), anche queste con le stesse trasformazioni. Per completezza è stata testata la dipendenza di P ma anche di sue trasformazioni:

$$\begin{aligned}
&P^*_{MLR}(\underline{u}_0) = b_0 + (b_1 X_1(\underline{u}_0) + \ldots + b_n X_n(\underline{u}_0)) \\
&\log(P^*_{MLR}(\underline{u}_0)) = b_0 + (b_1 X_1(\underline{u}_0) + \ldots + b_n X_n(\underline{u}_0)) \\
&\ln(P^*_{MLR}(\underline{u}_0)) = b_0 + (b_1 X_1(\underline{u}_0) + \ldots + b_n X_n(\underline{u}_0)) \\
&\sqrt{P^*_{MLR}(\underline{u}_0)} = b_0 + (b_1 X_{1(\underline{u}_0)} + \ldots + b_n X_n(\underline{u}_0))
\end{aligned} \qquad [6]$$

La regressione utilizzata è del tipo *stepwise*: inizialmente si inseriscono nell'equazione tutte le variabili a disposizione, ed si elimina di volta in volta la variabile che produce il coefficiente di determinazione inferiore, fino a che l'eliminazione di un ulteriori parametri non produce variazioni significative nel coefficiente di determinazione.
L'ultimo metodo utilizzato, il *kriging dei residui*, è geostatistico. Questo modello permette di separare in ogni misura di precipitazione la variabilità intrinseca del fenomeno da un andamento globale che influenza le varie misure. Quindi ogni dato sarà composto da una parte deterministica più una seconda parte variabile con media nulla. La parte deterministica viene modellata con un modello regressivo, con la regressione lineare robusta e con la regressione geograficamente pesata; si calcolano quindi i residui nei punti di campionamento, cioè la differenza tra la precipitazione prevista in ogni punto di campionamento e la precipitazione stimata dal modello deterministico, e si effettua il kriging sui residui. In *Figura 2* vengono mostrate le mappe di precipitazione media annuale prodotte con i vari metodi.

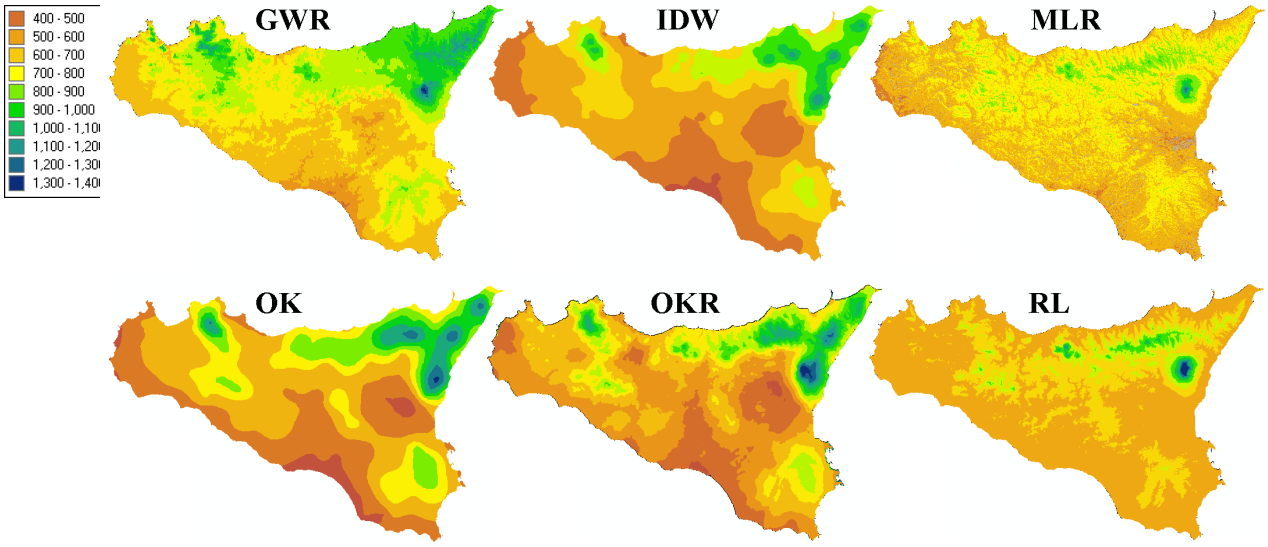

Figura 2 – Mappe di precipitazione annuale (GWR: Geographically Weighted Regression; IDW: Inverse Distance Weighting;MLR: Regressione Lineare Multipla; OK. Kriging Ordinario; OKR: Kriging Ordinario dei Residui; RL: Regressione Lineare Semplice)

3 Validazione

Per valutare quale metodo consenta di ottenere i migliori risultati si è utilizzata la *validazione*. Essa consiste nel dividere la base di dati in due gruppi: nei punti appartenenti al secondo gruppo (*dataset* di test) viene stimata la pioggia attraverso il modello costruito con i dati del primo sottoinsieme (*dataset* di *training*). La differenza tra singolo dato stimato e dato misurato costituisce l'errore; sugli errori si calcolano alcuni statistici che aiutano il confronto. In particolare sono stati utilizzati la media degli errori prodotti da ciascun metodo, l'*errore assoluto medio* MAE, l'*errore quadratico medio* MSE, *l'errore quadratico medio standardizzato* SMSE, la *distorsione* (*Bias*) o MBE (*Magnitude of Mean Errors*), la *deviazione standard* degli errori e la *varianza dell'errore* σ^2. Il confronto degli statistici permette di affermare che il metodo che produce gli errori minori nella stima della precipitazione è il *kriging dei residui* OKR (*Figura 4*, *Figura 5* e *Figura 6*).

4 Ricostruzione delle serie storiche

La ricostruzione delle serie storiche è fatta a scala annuale e a scala mensile. Per la stima dei valori annuali mancanti nelle serie storiche utilizziamo quindi il metodo OKR, cioè il *kriging dei residui* precedentemente sviluppato. Per la stima dei valori mensili utilizziamo invece diversi metodi. Il primo metodo, denominato *Soluzione 0*, consiste anch'esso nell'utilizzo dell'OKR, a scala mensile; viene cioè utilizzato un metodo di interpolazione spaziale. Inoltre l'analisi spaziale viene intrecciata con una analisi temporale: la stima annuale determinata con l'interpolazione spaziale viene

suddivisa in 12 parti per descrivere appunto le precipitazioni mensili. La suddivisione avviene secondo vari criteri temporali: in base ai valori medi mensili di precipitazione di ogni stazione (*Soluzione 1*); utilizzando un metodo autoregressivo periodico PAR del primo ordine (Soluzione 2) (Salas, 1993); generando serie sintetiche di precipitazioni mensili, facendo riferimento ai *Random Field* (*Soluzione 3*), con i quali si considera la generica serie Z(x) delle 12 precipitazioni mensili come una realizzazione casuale di un certo processo stocastico, che è appunto il *campo casuale* Z. Studiate le caratteristiche di Z possiamo generare tante serie sintetiche di media fissata, varianza nota e struttura di correlazione conosciuta. Ogni serie sintetica generata sarà una realizzazione casuale del processo gaussiano Z (Bras et al., 1994).
Anche per la valutazione dei vari metodi utilizzati per la stima dei valori mensili si utilizza la validazione. Il metodo migliore risulta essere la Soluzione 0, cioè il semplice kriging dei residui (*Figura 7* e *Figura 8*). Scelto il metodo da utilizzare per la stima dei valori mancanti nelle serie storiche, ne possiamo effettuare la ricostruzione. Il risultato finale è un database che contiene, per tutte le 247 stazioni studiate, le serie storiche di precipitazione mensile, per tutti gli anni dal 1921 fino al 2000. Ogni record possiede un *flag* che indica se il dato è interpolato o meno. Il database contiene varie altre tabelle, con le coordinate delle stazioni e i dati di precipitazione media mensile e media annuale, calcolati sia prima che dopo la ricostruzione delle serie storiche. Sebbene i dati immagazzinati su di una base di dati elettronica siano più facili da utilizzare per molte applicazioni, è stata prodotta una versione cartacea, nella quale i dati pluviometrici sono divisi per anno e le stazioni sono elencate in ordine geografico, seguendo gli standard tipografici degli annali idrologici pubblicati dall'Ufficio Idrografico Regionale.

5. Conclusioni e sviluppi futuri

Le serie storiche di dati pluviometrici presentano inevitabilmente discontinuità che si ripercuotono nei modelli idrologici nei quali vengono utilizzati. Attraverso l'uso delle tecniche di interpolazione spaziale integrata con l'analisi delle serie temporali è possibile stimare il dato meteoclimatico in una particolare posizione, a partire dai dati misurati in posizioni limitrofe: in particolare si possono determinare le stime dei dati mancanti nelle serie storiche. Il metodo che, attraverso la validazione, ha fornito i migliori risultati è stato il *kriging dei residui*, cioè la *Regressione Robusta ln(P)-Q* seguita dal *Kriging Ordinario dei Residui*. Gli errori e le deviazioni standard che si producono nella stima della precipitazioni ottenuta con il kriging dei residui risultano essere inferiori rispetto a quelli prodotti dagli altri metodi. Tra i possibili sviluppi di questo studio vi è l'ampliamento dei metodi di interpolazione utilizzati. Ad esempio è possibile effettuare la stessa analisi utilizzando interpolatori basati sulle reti neuronali come l'IRBFN (Lin et al., 2004). Inoltre lo stesso lavoro di ricostruzione delle serie storiche può essere effettuato su archivi di altre variabili meteoclimatiche, tra le quali, in primo luogo, le temperature.

Riferimenti bibliografici

Bras, R. L., Rodriguez-Iturbe I. (1994), *Random Functions and Hydrology*, Dover Publications, New York

Brunsdon, C., McClatchey, J., Unwin, D. J. (2001), "Spatial variations in the average rainfall-altitude relationship in Great Britain: an approach using geographically weighted regression", *International Journal of Climatology*, 21: 455-466

Eisched, J. K., Pasteris, P. A., Diaz, H. F., Plantico, M. S., Lott, N. J. (2000), "Creating a Serially Complete, National Daily Time Series of Temperature and Precipitation for the Western United States", *Journal of Applied Meteorology*, 39: 1580-1591

Feng, S., Hu, Q., Qian, W., (2004), "Quality control of daily meteorological data in China, 1951-2000: a new dataset", *International Journal of Climatology*, 24: 853-870

Goovaerts, P. (1999), "Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall", *Journal of Hydrology*, 228: 113-129

Lin, Gwo-Fong, Chen, Lu-Hsien (2004), "A spatial interpolation method based on radial basis function networks incorporating a semivariogram model", *Journal of Hydrology*, 288: 288-298.
Matheron, G. (1969), "Le Krigeage Universel", *Cahiers du Centre de Morphologie Mathematique,* Fontainebleau (France)
Stooksbury, D. E., Idso, C. D., Hubbard, K. G. (1999), "The effects of data gaps on the calculated monthly mean maximum and minimum temperatures in the continental United States: A spatial and temporal study", *Journal of Climate*, 12: 1524-1533
Tang, W. Y., Kassim A.H.M., Abubakar, S. H. (1996), "Comparative studies of various missing data treatment methods – Malaysian experience", *Atmospheric Research*, 42: 247-262

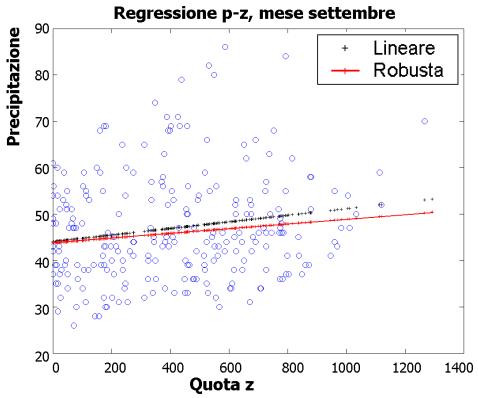

Figura 3 – Differenza regressione ai minimi quadrati ordinari e regressione robusta.

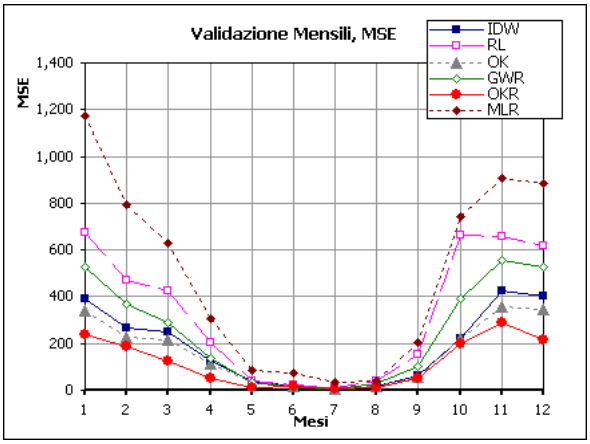

Figura 4 – MSE per la stima dei valori mensili

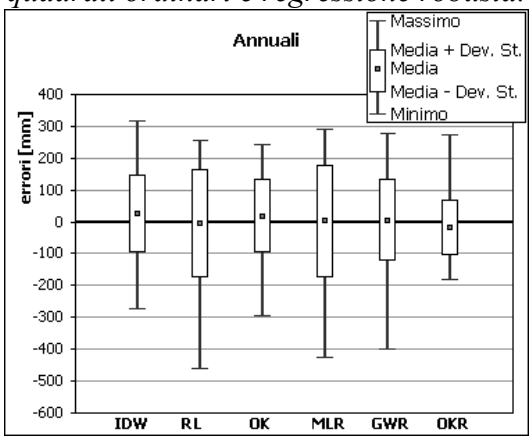

Figura 5 – Boxplot per la stima dei valori annuali

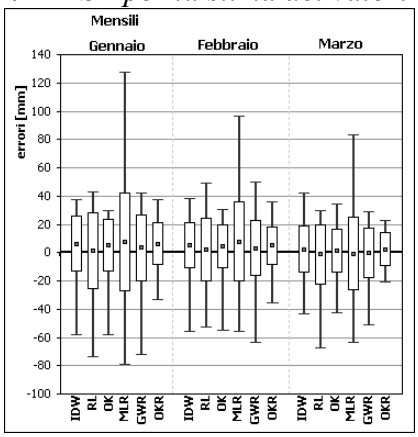

Figura 6 – Boxplot per la stima dei valori mensili con metodi spaziali

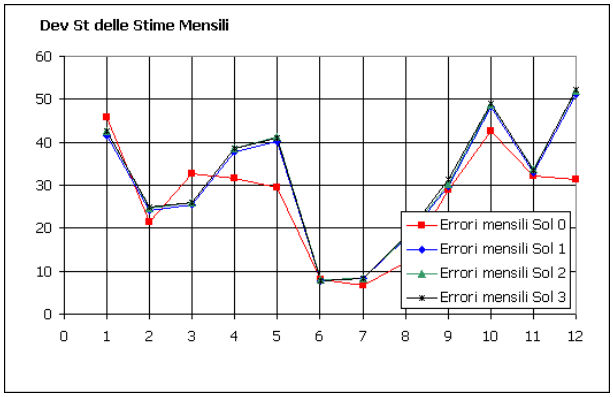

Figura 7 – Deviazione standard degli errori per il confronto dei metodi per la stima dei valori mensili

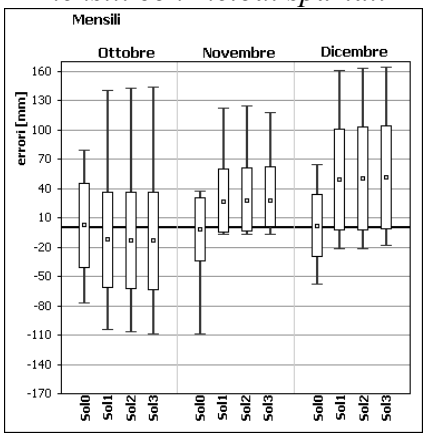

Figura 8 – Boxplot per il confronto dei metodi per le stime dei valori mensili