Giornale di Geologia Applicata 2008, 8 (1) 5-16 - doi: 10.1474/GGA.2008-08.1-01.0216

Modellazione e previsione nei sistemi idrogeologici mediante la tecnica E.P.R. (Evolutionary Polynomial Regression)

Davide Mancarella & Vincenzo Simeone

Politecnico di Bari – II Facoltà di Ingegneria – Taranto – Dipartimento Ingegneria per l'ambiente e lo Sviluppo Sostenibile (DIASS)
d.mancarella@poliba.it, v.simeone@poliba.it

Modelling and prediction in hydrogeological systems by means of E.P.R. (Evolutionary Polynomial Regression)

ABSTRACT: The hydrologic response of aquifers to meteorological conditions is often studied by means of time series analysis or physically based models. Modelling hydro-geological systems is however a challenging problem, often complicated by poor knowledge of basic assumptions such as hydraulic conductivity distribution, aquifer geometry and evapotraspiration rates. A better estimate of model parameters usually requires quite substantial investment and prolonged measurement campaigns. Collecting reliable data from different sources and public corporations is not always an easy task. Conversely, piezometric head and rainfall intensity data in long records are more easily available. Therefore, data-driven approaches (Ljung, 1999; Giustolisi, 2000) in groundwater hydrology and hydrogeology are attractive and potentially complementary in some cases. In the present work, the authors show some interesting results regarding the application of a recent symbolic regression technique, named evolutionary polynomial regression (Giustolisi and Savic, 2003, Mancarella, 2003, Giustolisi and Savic, 2006), in the conceptual modelling of a real hydrogeological system. The methodology's advantage lies in the model building procedure that is entirely based on time series data, and on the possibility of conceptualizing the physical insight into the process.

Key terms: Groundwater, Groundwater level forecast, Data-driven models, Conceptual models
Termini chiave: Acque sotterranee, Previsione dei livelli piezometrici, Modelli data-driven, Modelli concettuali

Riassunto

La risposta idrologica degli acquiferi alle precipitazioni può essere modellata con diversi approcci. E' frequente l'utilizzo di modelli fisicamente basati o il ricorso allo studio delle serie storiche con metodi statistici. L'utilizzo realistico di modelli fisicamente basati presuppone una buona conoscenza di base del sistema e un patrimonio di dati rilevante spesso di non facile reperibilità. La disponibilità di serie storiche idrologiche in periodi anche lunghi è invece più frequente. Pertanto gli approcci data-driven possono talvolta rappresentare una utile alternativa o uno strumento complementare importante nell'idrogeologia delle acque sotterranee. Nel presente lavoro gli autori si propongono di illustrare i risultati dell'applicazione di una recente tecnica di regressione simbolica denominata *Evolutionary Polynomial Regression* nella modellazione di un sistema idrogeologico reale, di cui sono disponibili le serie storiche pluviometriche e freatimetriche.

1. Introduzione

La modellazione dei sistemi idrogeologici si presenta, di norma, non semplice, per l'intrinseca complessità naturale degli stessi e per la forte non linearità dei processi che ne governano il comportamento e la risposta agli *input* esterni. Ciò rende difficile il controllo e la gestione di questi sistemi, anche quando si dispone di dati derivanti da un buon monitoraggio, in quanto il passaggio al livello decisionale necessita di un modello che sia in grado di determinare la risposta complessa del sistema acquifero correlando i dati di monitoraggio di *input* e *output*.

Esistono varie classi di modelli implementabili. Il modello adottato dovrà, tuttavia, essere in grado di interpretare la complessità dei fenomeni fisici che governano il comportamento dei sistemi idrogeologici, dall'altro evitare di introdurre un numero elevato di variabili e parametri, spesso difficilmente stimabili, anche se ciò può portare ad accettare ipotesi semplificative. In molti problemi ambientali, la grande scala spaziale non permette la conoscenza e la misura di tutti i parametri fisici coinvolti. In queste circostanze, la calibrazione di un modello, magari semplificato, ma più idoneo ad utilizzare la conoscenza rinveniente dai dati di monitoraggio, può divenire uno strumento complementare, ma essenziale per la gestione del sistema.

Nella presente nota vengono esposti i risultati della modellazione delle oscillazioni del livello di falda di un sistema idrogeologico reale mediante una recente tecnica di regressione simbolica, denominata *Evolutionary Polynomial Regression* (EPR) (Giustolisi e Savic, 2003; Giustolisi e Savic, 2006). Si tratta di uno strumento di

modellazione dei sistemi ambientali particolarmente versatile che presenta interessanti potenzialità di applicazione allo studio dei sistemi idrogeologici.

2. Classi di modelli

Il modello generale di un fenomeno fisico può essere rappresentato in maniera compatta da un'equazione matriciale generale del tipo:

$$Y = F(X, \Theta) \quad (1)$$

dove Y rappresenta l'uscita del modello, solitamente una o più grandezze di stato del sistema. La variabile di uscita può essere costituita da più variabili in forma di serie storiche. X è l'insieme delle variabili di ingresso, θ il vettore contenente i parametri del sistema, *F* il sistema di equazioni che mette in relazione queste grandezze. In generale sia θ che *F* possono essere tempo-varianti. In un modello dinamico si possono prevedere come dati di ingresso anche i valori storici delle variabili d'uscita.

Un modello che preveda la risposta di un acquifero alle precipitazioni, in termini di oscillazioni del livello di falda, può, in prima approssimazione, essere espresso attraverso una relazione matematica del tipo:

$$H = F(H, P, T, K, G) \quad (2)$$

in cui *H* è un vettore o una matrice contenente le altezze di falda in uno o più punti stabiliti, *P* è il vettore di precipitazioni nell'area di alimentazione della falda in uno o più punti stabiliti, *T* è il vettore delle temperature, *K* una matrice contenente le permeabilità degli elementi che costituiscono il sistema e *G* è una matrice che descrive la geometria del sistema. La definizione delle relazioni esistenti fra le diverse grandezze dell'equazione (2) e la relativa risoluzione può essere svolta utilizzando differenti approcci e servendosi di diverse classi di modelli.

2.1 Modelli fisicamente basati

I modelli così detti *fisicamente basati (white-box)* sono basati su equazioni differenziali e parametri che sono desunti dalla fisica del sistema e richiedono la conoscenza del dominio, delle condizioni iniziali ed al contorno con un buon grado di accuratezza, per essere usati a scopi previsionali. I parametri e le variabili hanno generalmente significato fisico ben preciso ed esprimono le proprietà e lo stato del sistema in termini espliciti. Con riferimento ai sistemi idrogeologici, un modello di questa famiglia è l'equazione di Richards (Richards, 1931) per la simulazione dei processi di infiltrazione nei mezzi insaturi. E' un modello certamente rigoroso che trova tuttavia una certa difficoltà di applicazione, specialmente in problemi di larga scala, a causa delle complesse funzioni e parametri che in esso compaiono e che vengono dedotte a loro volta da modelli costitutivi semiempirici di terreno disponibili in letteratura. Una ulteriore complicazione deriva dai fenomeni di isteresi (Dane & Wierenga, 1975) che, per essere tenuti in conto, richiedono non solo la condizione iniziale di umidità ma persino la "storia di carico" del contenuto idrico del suolo, dal momento che la curva di ritenzione in essiccamento è differente da quella in umidificazione. Il fenomeno dell'isteresi è spesso trascurato nella pratica a causa delle enormi incertezze connesse (Guymon, 1994).

In questo genere di modelli può spesso accadere che l'apparente maggior accuratezza e affidabilità del modello venga vanificata dalla insufficiente disponibilità di dati su parametri distribuiti, condizioni iniziali e al contorno, sul tensore di permeabilità che può essere spazio-variante e manifestare isteresi in condizioni insature, sui parametri dei modelli caratteristici, sulla geometria dei corpi acquiferi. L'applicazione di questi modelli risulta pertanto realisticamente rara, perlomeno per scopi previsionali.

2.2 Modelli data-driven

Un modello *data-driven (black-box),* contrariamente all'approccio fisicamente basato, è il frutto dell'identificazione di strutture matematiche flessibili, atte a rappresentare la dinamica del sistema, a partire da sequenze storiche di dati ingresso-uscita. Il processo di costruzione di modelli di sistemi complessi sulla base di dati misurati è denominato Identificazione dei Sistemi *(System Identification)* (Ljung, 1999). In questo caso, le equazioni che reggono la dinamica del sistema ed i relativi parametri non sono noti a priori e spesso non si prestano ad una spiegazione fisica immediata, tuttavia consentono di ottenere relazioni tra le variabili del sistema di grande utilità ed accuratezza. Si selezionano le variabili che costituiscono l'*input* e l'*output* del modello in relazione ai dati disponibili, secondo un approccio alquanto pragmatico, e ci si propone di costruire uno strumento che possa essere usato per prevedere i valori della variabile d'uscita del sistema. Le funzioni *F* da utilizzare e il valore dei parametri che appaiono nel modello non sono noti a priori ma ricercati mediante opportune tecniche numeriche. Questa fase di taratura, in questa famiglia di modelli, viene spesso denominata apprendimento *(training)*.

Una classe di modelli *black-box* ben nota è rappresentata dalle reti neurali *(Artificial Neural Networks,* ANN) che hanno trovato larga applicazione nei problemi di previsione e controllo dei sistemi in vari settori della scienza e della tecnica quali ad esempio l'ingegneria elettronica, l'informatica, e, più recentemente, l'idrologia ed idraulica (ASCE Task Commettee, 2000; Giustolisi, 2000). Una ulteriore classe di modelli *black-box* è rappresentata dalla *Genetic Programming* (GP) (Keijzer & Babovic, 2000). In particolare le ANN rappresentano strutture di interpolazione pura che mantengono al più basso livello l'informazione sulle relazioni funzionali *F* mentre la GP costruisce modelli con relazioni funzionali esplicite (regressione simbolica). Nelle reti neurali i modelli prodotti assumono la forma di sistemi di equazioni numeriche non lineari con i pesi calibrati sui dati.

2.3 Modelli concettuali

I modelli concettuali *(grey-box)* sono caratterizzati dall'avere una struttura matematica che deriva dalla concettualizzazione di un fenomeno fisico oppure dall'assunzione o semplificazione, attuata mediante opportune ipotesi, delle leggi che reggono la sua dinamica. Le equazioni del modello sono pertanto conosciute o assunte a priori mentre i parametri introdotti devono solitamente essere tarati attraverso analisi ingresso-uscita. Un modello afferente a questa famiglia, ben noto in idrogeologia, è quello di Horton (Horton, 1940). Modelli più complessi sono stati varie volte applicati alla modellazione di sistemi acquiferi (Wanakule & Anaya, 1993). L'informazione che si usa per la costruzione del modello deriva in parte dalla conoscenza fisica del fenomeno e dalla sua schematizzazione ed in parte dalle sequenze sperimentali di ingresso-uscita.

3. Evolutionary polynomial regression

Numerosi ricercatori, in questi ultimi anni, si sono riproposti di avvicinare i modelli *data-driven* ad una rappresentazione di tipo *grey-box* in cui fosse possibile introdurre la conoscenza fisica nelle fasi di identificazione, taratura e verifica del modello, pervenendo ad espressioni simboliche che, da un lato, nascessero in parte da una concettualizzazione, e dall'altro fossero in grado di evidenziare alcuni aspetti anche ignoti del fenomeno modellato (Keijzer & Babovic, 2000). Una frontiera promettente della ricerca è rappresentata dalla EPR, *Evolutionary Polynomial Regression* (Giustolisi & Savic, 2003; Mancarella, 2003; Giustolisi & Savic, 2006; Doglioni *et al.,* 2007; Mancarella *et al.,* 2007).

La EPR è una tecnica finalizzata alla costruzione di modelli simbolici a carattere polinomiale che sfrutta un semplice Algoritmo Genetico (GA) (Goldberg, 1989) per effettuare la ricerca nello spazio delle strutture possibili del modello e si estrinseca in due fasi:

(i) Identificazione della struttura del modello
(ii) Stima dei parametri.

La *Evolutionary Polynomial Regression* limita l'insieme di operatori utilizzati nella regressione simbolica ad un sottoinsieme costituito da addizione, moltiplicazione, potenza, logaritmo ed esponenziale. La struttura risultante ha una forma polinomiale in cui i termini monomi possono essere combinazioni più o meno complesse delle variabili di ingresso e uscita:

$$y = \sum_{j=1}^{m} a_j \cdot z_j + a_o \qquad (3)$$

dove

y è il valore previsto dell'uscita del modello;
a_j è il coefficiente tarabile del j^{mo} termine del polinomio;
a_0 è una costante opzionale (bias);
m è il numero di termini monomi figuranti nell'espressione, composti da variabili;
z_j è una variabile trasformata che è combinazione di funzioni delle variabili indipendenti scelte per la previsione, cioè gli ingressi $<x_1\ x_2\ \ldots\ x_k>$;
k è il numero delle variabili indipendenti scelte per la previsione.

Se lo scopo fosse quello di costruire un modello di previsione della risposta della falda alle precipitazioni mensili che tenga conto anche della temperatura e che usi i dati disponibili con due mesi di anticipo potremmo considerare come ingressi le variabili $<h_{t-2}\ h_{t-3}\ h_{t-4}\ h_{t-5}\ p_{t-2}\ p_{t-3}\ p_{t-4}\ p_{t-5}\ T_{t-2}\ T_{t-3}\ T_{t-4}\ T_{t-5}>$ assumendo così che spingendoci indietro nel tempo di 5 mesi, i dati possano veicolare informazione sufficiente a prevedere l'escursione di falda al mese t. Un modello ipoteticamente identificato da EPR, come maggiormente rappresentativo del sistema, potrebbe avere una struttura del tipo:

$$h_t = a_0 + a_1 \cdot h_{t-2} + a_2 \cdot p_{t-2} + a_3 \cdot h_{t-3} h_{t-4} p_{t-3} p_{t-4}^{0.5} - a_4 \cdot T_{t-2} T_{t-3} \qquad (4)$$

in cui $a_0 \ldots a_4$ sono i coefficienti che moltiplicano i diversi termini monomi. Se per un acquifero si ottenesse questo genere di relazione *input-output*, a valle della ricerca operata nello spazio delle soluzioni e della stima dei parametri, si potrebbero trarre alcune considerazioni: si potrebbe desumere che la precipitazione caduta 5 mesi prima p_{t-5} non influenza significativamente il livello di falda o, più correttamente, tale dato non veicola sufficiente informazione ai fini della previsione del livello di falda al mese t. Se EPR non ha selezionato questa variabile durante la costruzione del modello, allora nella storia di quel dato non c'è un contenuto di informazione significativo nella previsione a breve e medio termine (5 mesi in avanti). Analogamente si potrebbe affermare che, sulla base del modello ottenuto, la precipitazione caduta 4 mesi prima ha un peso relativo minore rispetto a quella avutasi 3 mesi prima. Il valore dei coefficienti stabilirà quale termine incide maggiormente nell'escursione di falda al mese t, mentre gli esponenti determinano il peso relativo delle variabili storiche all'interno dello stesso termine.

EPR può potenzialmente trovare applicazione nella modellazione dinamica dei sistemi, nella individuazione di relazioni tra parametri fisici, in problemi di classificazione nello stabilire l'appartenenza di un elemento ad un insieme.

3.1 La matematica dell'EPR

In questo paragrafo verranno brevemente descritte le basi numeriche del processo di induzione dei modelli dinamici costruiti da EPR a partire dalle serie storiche dei dati. Per ulteriori approfondimenti, i lettori interessati possono fare riferimento a pubblicazioni di dettaglio sugli algoritmi numerici implementati (Giustolisi e Savic, 2003; Mancarella, 2003; Giustolisi e Savic, 2006).

L'equazione (3), ai fini della costruzione del modello,

può essere scritta in forma matriciale nella maniera seguente:

$$\mathbf{Y}_{N\times 1}(\boldsymbol{\theta},\mathbf{Z}) = \left[\mathbf{I}_{N\times 1} \quad (\mathbf{Z}_j)_{N\times m}\right]\times\left[a_0 \quad a_1 \quad \mathrm{K} \quad a_m\right]^T = \mathbf{Z}_{N\times d}\times\boldsymbol{\theta}^T_{d\times 1} \tag{5}$$

in cui

$\mathbf{Y}_{N\times 1}(\boldsymbol{\theta},\mathbf{Z})$ è la stima ai minimi quadrati dell'uscita del modello (valori di target);

$\boldsymbol{\theta}_{d\times 1}$ è il vettore contenente $d = m+1$ parametri a_j del modello, con j =1…m, ed a_0;

$\mathbf{Z}_{N\times d}$ è una matrice costituita da **i**, vettore colonna unitario e da m vettori colonna di variabili trasformate $\mathbf{z}_j$, le quali, per un assegnato valore dell'indice j rappresentano un prodotto delle variabili *input* di previsione $\mathbf{x} = <\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_k>$ che possono essere raccolti in una matrice di ingressi $[\mathbf{x}_1\ \mathbf{x}_2\ \ldots\ \mathbf{x}_k]$;

N rappresenta il numero di dati disponibili, trattandosi, nel caso di una serie storica, del numero di misure effettuate su ogni variabile negli istanti passati. Con il simbolo × si denota il prodotto righe per colonne.

La EPR assume una struttura per il modello che abbia la forma dell'equazione (5) e ricerca la miglior forma della funzione F, ovvero la miglior combinazione delle variabili di *input* $\mathbf{X}_{s=1:k}$ effettuando mediante la tecnica dei minimi quadrati la stima dei parametri $\boldsymbol{\theta}$ per ciascuna combinazione di ingressi. La ricerca del miglior modello avente struttura (5) avviene minimizzando gli scarti quadratici medi (la cosiddetta funzione obiettivo) tra i valori previsti dalla stessa equazione (5) e i valori della variabile misurati. La ricerca della combinazione di *input* e dei relativi esponenti che minimizza la funzione obiettivo avviene mediante algoritmi genetici (Goldberg, 1989), una tecnica numerica di ottimizzazione che assicura una esplorazione globale dello spazio delle soluzioni possibili ed evita gli inconvenienti derivanti dalle metodologie classiche di ricerca basate su gradiente e matrice Hessiana (Kecman, 2000; Giustolisi & Savic, 2003). Questa fase prende il nome di ricerca evolutiva *(Evolutionary Search)*. Le strutture che manifestano maggior adattamento alla variabilità dei dati vengono mantenute, variate e ricalibrate finché non si ottengono i più bassi errori possibili di approssimazione.

Durante il processo iterativo di EPR, quando una struttura (5) viene identificata, i coefficienti *aj* possono essere stimati per mezzo della approssimazione ai minimi quadrati.

Le funzioni di costo adottate sono costruite in maniera tale da penalizzare la complessità delle formule risultanti o effettuare un controllo sui coefficienti *aj* (Giustolisi & Savic, 2003). Durante la ricerca delle soluzioni, i coefficienti *aj* per i quali si verifichi che il loro valore assoluto risulti inferiore alla loro deviazione standard, calcolata sulla base degli errori residui del modello, vengono azzerati. In seguito all'annullamento di questi termini gli altri coefficienti vengono ricalcolati, sempre con il metodo dei minimi quadrati. EPR costruisce più modelli, a diverso grado di complessità strutturale che possono poi essere oggetto di ulteriori valutazioni e giudizi.

3.2 Un test numerico per EPR

I dati delle serie storiche sono sempre caratterizzati da un certo rumore di fondo che rende talvolta difficile l'identificazione della dinamica del sistema. Prima di procedere ad un caso di studio reale si è ritenuto pertanto utile verificare l'attitudine di EPR a individuare la struttura matematica dei fenomeni per mezzo di un semplice test numerico. Si considerino tre variabili di ingresso $<\mathbf{X}_1\ \mathbf{X}_2\ \mathbf{X}_3>$ ed una grandezza **Y** che è funzione delle precedenti secondo la seguente legge:

$$\mathbf{Y} = a_0 + a_1\cdot\mathbf{Z}_1 + a_2\cdot\mathbf{Z}_2 + a_3\cdot\mathbf{Z}_3 = 10 + 1\cdot\mathbf{X}_1/\mathbf{X}_2 + 1\cdot\mathbf{X}_2/\mathbf{X}_3 + 1\cdot\mathbf{X}_3/\mathbf{X}_1 \tag{6}$$

Le variabili di ingresso sono state generate come tre vettori di numeri casuali tutti contenuti nell'intervallo [0.1-1], con distribuzione di probabilità normale. A queste serie di ingressi corrisponderà l'uscita **Y** secondo l'equazione (6).

Questa equazione rappresenta una legge che relaziona una certa variabile di uscita y ad altre tre grandezze che rappresentano variabili di ingresso per il modello. EPR può ricostruire la formula (6) a partire dalle sequenze di dati di ingresso e uscita.

A tale equazione si può aggiungere un rumore di fondo nei dati, generando quattro sequenze di numeri $\mathrm{n}(0, \sigma_l)$ aventi distribuzione di probabilità gaussiana a media nulla e deviazione standard σ_l (l = 1…4). Si ottengono così quattro differenti uscite Y_{Rl}:

$$\mathbf{Y}_{Rl} = 10 + 1\cdot\mathbf{X}_1/\mathbf{X}_2 + 1\cdot\mathbf{X}_2/\mathbf{X}_3 + 1\cdot\mathbf{X}_3/\mathbf{X}_1 + \mathrm{N}(0,\sigma_l) \tag{7}$$

con l =1…4

Ciascuna serie Y_{Rl} deriva da quella originale e risulta sporcata da questo rumore di fondo, avente deviazione standard con differenti livelli di ampiezza σ_l rispettivamente pari a 0, 5, 10, 20 % di σ_y, essendo σ_y la deviazione standard della variabile di uscita **Y** ottenuta dalla equazione (6).

Lo scopo del test è identificare la vera struttura di **Y** espressa nella formula (6) utilizzando i dati delle serie $\mathbf{Y}_{\mathrm{Rl}}$ dei dati ottenuti dalla sequenza originale ma corrotti dal rumore di fondo. I risultati ottenuti sono mostrati in Tabella 1. Per un livello nullo di σ_l, ossia in assenza di rumore di fondo, EPR ha identificato esattamente la struttura (6). Con l'incremento del livello di rumore nei dati, la formula risulta ancora straordinariamente vicina a quella originaria. Si notano scostamenti più significativi, sebbene piuttosto piccoli, solo quando la deviazione standard σ_l nella (7) raggiunge il 20 % della deviazione dell'uscita originale σ_y. In questo specifico caso, pertanto, l'equazione del fenomeno Y viene identificata correttamente nonostante sia presente un rumore di fondo pari al 20% della variabilità attorno alla media della grandezza Y.

Tabella 1 -. Formule identificate da EPR per differenti livelli del rumore di fondo sulla equazione (6).
Table 1 - Formulas identified by EPR at different background noise levels on equation (6)

Livello del rumore	Formule identificate da EPR
$\sigma_I = 0$	$Y_{EPR} = 10 + 1 \cdot X_1/X_2 + 1 \cdot X_2/X_3 + 1 \cdot X_3/X_1$
$\sigma_I = 0.1\ \sigma_Y$	$Y_{EPR} = 9.9884 + 1.0008 \cdot X_1/X_2 + 0.97882 \cdot X_2/X_3 + 0.99574 \cdot X_3/X_1$
$\sigma_I = 0.15\sigma_Y$	$Y_{EPR} = 9.9983 + 0.99823 \cdot X_1/X_2 + 1.0317 \cdot X_2/X_3 + 0.98168 \cdot X_3/X_1$
$\sigma_I = 0.2\sigma_Y$	$Y_{EPR} = 10.2326 + 0.78962 \cdot X_1/X_2 + 0.49317/X_3 + 1.0136 \cdot X_3/X_1$
	$Y_{EPR} = 9.8774 + 1.0047 \cdot X_1/X_2 + 0.95116 \cdot X_2/X_3 + 1.019 \cdot X_3/X_1$

4. Il caso di studio: L'acquifero superficiale di Brindisi

Ad Ovest di Brindisi (Figura 1) è presente una vasta area pianeggiante di forma irregolare che occupa la piccola depressione tettonica esistente in corrispondenza della "Soglia Messapica", individuata tra le Murge e la penisola salentina. Il substrato carbonatico mesozoico è sprofondato e si sono deposti dei litotipi ascrivibili al ciclo sedimentario Plio-Pleistocenico.

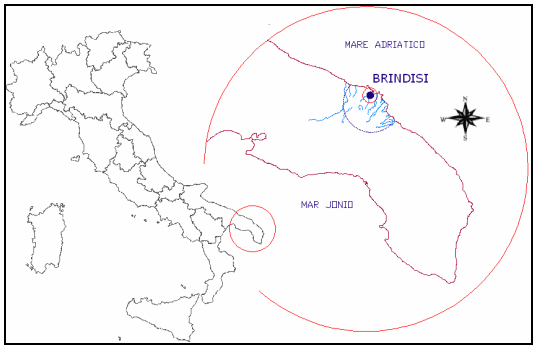

Figura 1 – Localizzazione dell'area di studio.
Figure 1 - Location of the study area.

La successione stratigrafia (Figura 2) si chiude con litotipi argillosi ascrivibili alla argille grigio azzurre subappenine sostenenti litotipi più permeabili costituiti da depositi quaternari stratificati composti da sabbie, calcareniti, conglomerati, livelli terrazzati e depositi alluvionali e colluviali in forma di depositi marini terrazzati (Pleistocene Superiore e Olocene). Detti litotipi ospitano un acquifero superficiale ben descritto da Ricchetti e Polemio (1996), caratterizzato da porosità primaria con valori di permeabilità variabili da 8×10^{-6} m/s a 1.4×10^{-4} m/s in cui la circolazione delle acque sotterranee è legata solo all'alimentazione diretta delle piogge. L'area è caratterizzata dalla presenza di un reticolo idrografico poco profondo, ma ben sviluppato con poche incisioni profonde, come quelle del Canale Cillarese, a nord-ovest del capoluogo, del Canale di Siedi, del Canale Foggia di Rau, del Fiume Grande, che, con la loro azione drenante, portano a giorno le acque della falda superficiale. Trattandosi di un acquifero permeabile per porosità con sola alimentazione diretta legata alle piogge si presta molto bene allo studio delle correlazioni fra le precipitazioni e le escursioni di falda.

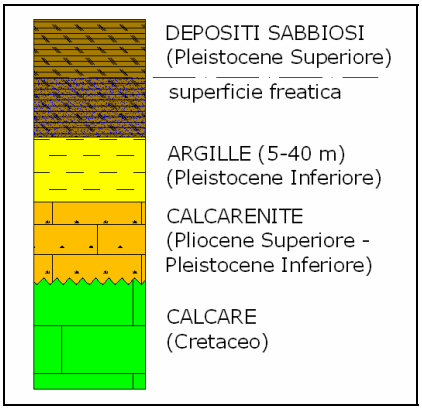

Figura 2 - Successione stratigrafica schematica dell'area dell'acquifero superficiale di Brindisi.
Figure 2 - Schematic stratigraphy in the area where the shallow aquifer of Brindisi is located.

4.1 I dati freatimetrici e pluviometrici

Lo studio è stato svolto utilizzando i dati del livello di falda misurati nella stazione freatimetrica del Servizio Idrografico e Mareografico presso la Casa Cantoniera (Via Appia km 717). Per detta stazione freatimetrica sono disponibili dati del livello di falda misurati 3 volte la settimana per un periodo di 44 anni dal 1953 al 1996, con alcune lacune di modesta entità. Si è scelto di svolgere l'analisi considerando come intervallo temporale di riferimento il mese. Come dati del livello di falda sono pertanto stati considerati i valori medi del livello di falda misurati nel mese considerato. Per quanto attiene i dati di pioggia e di temperatura si è fatto riferimento ai valori di precipitazione totale mensile misurati nella stazione termopluviometrica di Brindisi. Questa pur essendo posta più a valle della stazione di misura freatimetrica si è ritenuto fosse la più significativa in quanto di gran lunga più vicina delle altre al punto di misura della falda, soprattutto perché trattasi di un acquifero superficiale di modeste dimensioni. I dati usati in questo lavoro sono precipitazioni mensili e livelli piezometrici medi mensili. Il *set* di dati utilizzati in questo caso di studio è stato costituito dalle due serie storiche mensili (precipitazioni mensili e quote piezometriche) del periodo 1953-1996, ciascuna contenente 528 valori. In Tabella 2 sono riportate la caratteristiche delle stazioni.

Tabella 2 – Caratteristiche delle stazioni freatimetrica e pluviometrica.
Table 2 - Main features of the groundwater level and rain gouge stations.

Stazioni		Caratteristiche delle stazioni
Pluvio-metrica	Brindisi	Pluviometro registratore Quota: 28 m s.l.m. Altezza dell'apparecchio: 22 m
Freati-metrica	Casa Cantoniera SS 7, km 717	Latitudine: 5° 26' Longitudine: 40° 36' Quota del caposaldo di riferimento 35,92 s.l.m.

Sulla base dei dati termopluviometrici relativi al periodo di studio (1953-1996), l'area di studio è caratterizzata da un regime pluviometrico di tipo marittimo, con un minimo di piovosità nel mese di Luglio ed un massimo ricadente nel periodo di Novembre o Dicembre con una precipitazione totale media annua di poco più di 600 mm (Figura 3).

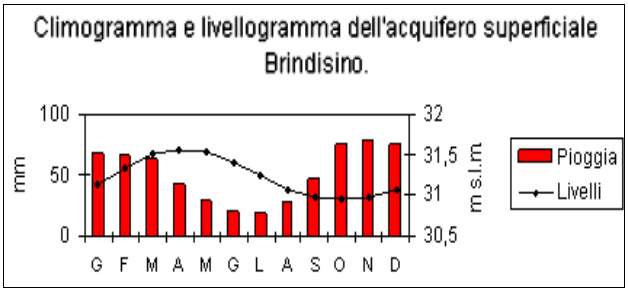

Figura 3 – Climogramma e livellogramma medio delle stazioni pluviometriche e fratimetriche studiate (da Maggio, 2003).
Figure 3 - Mean rainfall histogram and water level graph for the measurement stations (from Maggio, 2003).

Dall'analisi del livellogramma della falda (Figura 3) emerge che la ricarica della falda avviene essenzialmente nei primi tre mesi dell'anno, mentre le piogge dei primi tre mesi dell'anno non contribuiscono significativamente alla ricarica della falda, probabilmente perché inizialmente l'acqua di infiltrazione serve a ricostruire il contenuto d'acqua connesso alla capacità di campo. Vi è pertanto una sfasatura di circa 3 mesi fra il climogramma ed il livellogramma, in accordo con quanto evidenziato da Ricchetti e Polemio (1996). La ricarica appare significativa per piogge di bassa intensità e lunga durata, mentre piogge di elevata intensità e bassa durata non danno contributo significativo ai livelli di falda. L'analisi complessiva di lungo periodo evidenzia un *trend* generale di decrescita sia delle precipitazioni che dei livelli della falda (Figura 4).

Il *set* di dati utilizzati in questo caso di studio è costituito da due serie storiche (precipitazioni mensili e quote piezometriche), ciascuna contenente 528 valori. Per la realizzazione del modello EPR si è deciso di suddividere ciascuna serie storica in due sottoinsiemi: il primo *(training set),* costituito dalle prime 300 misurazioni da usarsi per la costruzione del modello, l'altro composto dai restanti 228 valori storici, da usare per la validazione del modello e per testarne la capacità di previsione in avanti. La serie storica dei dati di taratura *(training set)* risulta ininterrotta, mentre vi sono dei vuoti nelle rilevazioni dei livelli di falda, riscontrabili nei dati del *test set*. In particolare ve ne sono 5: due *gap* mensili, uno bimestrale, uno trimestrale ed uno di 7 mesi.

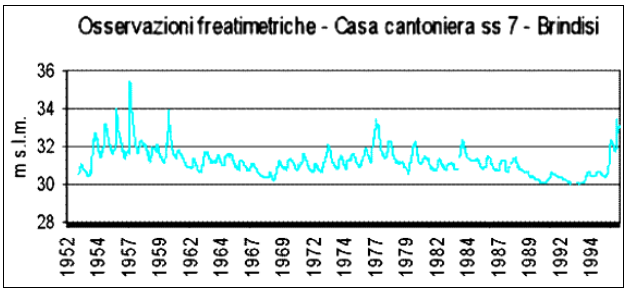

Figura 4 – Andamento delle osservazioni freatimetriche nel periodo studiato (da Maggio, 2003).
Figure 4 - Observed phreatimetric levels in the study period (from Maggio, 2003).

5. La modellazione dell'acquifero superficiale di Brindisi con EPR

L'acquifero superficiale di Brindisi è stato modellato con la tecnica EPR allo scopo di studiare e prevedere la dinamica della risposta della falda freatica alle precipitazioni. Come grandezza da modellare è stata scelta l'altezza di falda mentre le precipitazioni costituiscono la variabile di *input*.

Lo scopo è quello, da un lato, di verificare le capacità di previsione della tecnica descritta nel presente lavoro, e dall'altro di leggere e studiare la dinamica del sistema idrogeologico così come appare dalle serie storiche delle sue variabili di stato (quote freatimetriche) e forzanti (precipitazioni). Per la realizzazione del modello EPR si è deciso di suddividere ciascuna serie storica (quote piezometriche e precipitazioni mensili), contenente 528 valori in due sottoinsiemi: il primo *(training set),* costituito dalle prime 300 misurazioni da usarsi per la costruzione del modello, l'altro composto dai restanti 228 valori storici, da usare esclusivamente per la validazione del modello ossia per testarne la capacità di previsione. Il modello viene costruito e calibrato con i soli dati del *training set* mentre la sua capacità di previsione viene testata esclusivamente sui dati del *test set.*

5.1 Costruzione e validazione del modello

La serie storica dei dati di taratura *(training set)* risulta ininterrotta, mentre vi sono dei vuoti nelle rilevazioni dei livelli di falda, riscontrabili nei dati del *test set*. In particolare ve ne sono 5: due *gap* mensili, uno bimestrale, uno trimestrale ed uno di 7 mesi.

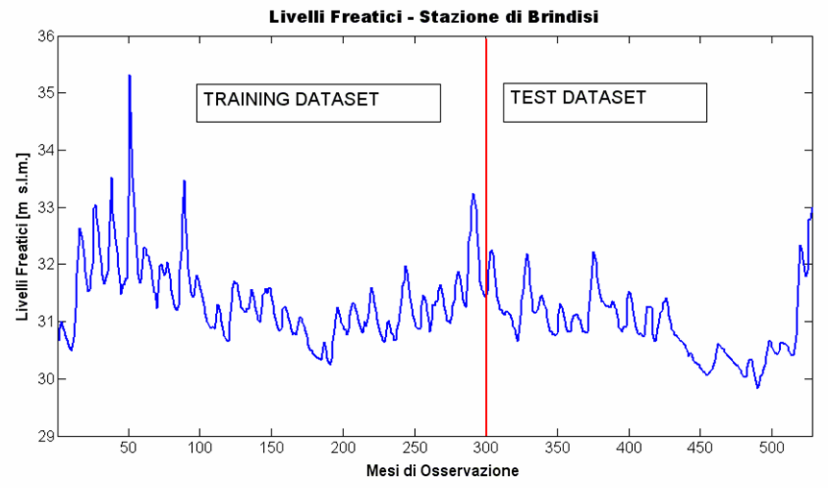

Figura 5 – Suddivisione dei livelli misurati in *training dataset* e *test dataset.*
Figure 5 - Data splitting into training set and test set.

Le serie storiche utilizzate per la modellazione EPR sono state dunque integrate con dati ricostruiti con le spline cubiche, ma non sono state sottoposte ad alcuna pre-elaborazione finalizzata alla rimozione dei *trend* e delle componenti cicliche pluriennali. La rimozione di tali componenti ha infatti senso soprattutto se si intende modellare separatamente la componente stocastica rispetto a quella deterministica. Inoltre il pre-processamento dei dati comporta sicuramente una perdita di informazioni.

Durante questo studio, sono stati costruiti numerosi modelli EPR confrontandone la *performance*. Sono state implementate, durante la ricerca evolutiva, le tecniche sopra descritte basate sul controllo dei coefficienti a_j ed una funzione di costo con penalizzazione delle strutture complesse.

I differenti modelli costruiti da EPR si differenziano per il loro diverso ordine di complessità e per il criterio di induzione adottato. Modelli più complessi possono manifestare un maggior adattamento ai dati del *training set* e del *test set* ma risultano di più difficile leggibilità.

In questa sede non si esporranno tutti i modelli risultanti ottenuti al variare del numero di termini polinomiali a_jZ_j ma ci si limiterà a descrivere solo i modelli più efficaci, sia in termini di capacità di generalizzazione nelle previsioni in avanti per differenti orizzonti temporali, sia in termini di leggibilità ed interpretabilità dei risultati.

5.2 Analisi dei principali risultati della modellazione

I migliori modelli sono risultati quelli a due e tre termini monomi, anche se si sono ottenute soluzioni interessanti anche per i modelli a quattro termini. Le previsioni effettuate sui dati del *test set* appaiono decisamente buone, sulla base di indicatori come la SSE ed il CoD (Doglioni *et al.,* 2007). La SSE (Somma degli errori quadratici medi) è un indicatore statistico che misura la deviazione totale dei valori di uscita del modello rispetto ai dati sperimentali. Per demarcare meglio questo indicatore si è pensato di rapportarlo con l'SSE del modello di persistenza $h_t=h_{t-1}$, indicando RSSE questo valore.

Il CoD (Coefficiente di Determinazione) è un buon indicatore di approssimazione ai dati sperimentali che misura la capacità del modello di spiegare la variabilità dei dati attorno alla media. Il suo valore rappresenta la porzione di varianza nei dati effettivamente spiegabile dal modello costruito. Il CoD è espresso mediante l'equazione (8):

$$\mathrm{CoD} = 1 - \frac{\mathrm{N}-1}{\mathrm{N}} \frac{\sum_{\mathrm{N}} \left(\hat{H} - H_{\exp}\right)^2}{\sum_{\mathrm{N}} \left(H_{\exp} - \mathrm{avg}\left(H_{\exp}\right)\right)^2} \qquad (8)$$

in cui N è il numero di dati nel *test set*, $\hat{H}$ è il vettore di tutte le previsioni della grandezza h, ed H_{exp} sono le corrispondenti osservazioni registrate nel *test set*.

Tra tutti, il miglior modello trovato da EPR è risultato il seguente:

$$h_t = 10.1 \cdot \sqrt{h_{t-1}} + \\ + 5.94 \cdot 10^{-6} \cdot p_{t-2} \cdot \left(h_{t-2}\right)^2 \cdot \sqrt{h_{t-1} \cdot p_{t-3} \cdot p_{t-4}} - 25.3 \qquad (9)$$

ottenuto con la ricerca evolutiva basata sul controllo dei coefficienti a_j dei termini polinomiali. Nella simbologia di questa equazione e delle successive, h_t ed h_{t-1} rappresentano rispettivamente il livello di falda al mese t e t-1, mentre p_{t-1}, p_{t-2}, p_{t-3}, p_{t-4}, p_{t-5} le precipitazioni nei cinque mesi precedenti a t.

In questa formula EPR figura un termine di persistenza non lineare in cui la quota di falda è isolata rispetto alle altre variabili di previsione. Questa caratteristica è stata riscontrata in numerosi altri modelli identificati da EPR, con minori capacità previsionali, che qui non si riportano per brevità.

Nella (9) compare anche un altro termine monomio non lineare, costituito da una combinazione di numerose variabili di ingresso, sul quale si riporteranno alcune considerazioni nel seguito.

Nella Figura 6 è mostrato il confronto tra l'*output* del modello (9) e i dati di taratura *(training set)*. L'immagine mostra un'eccellente aderenza ai dati sperimentali da parte del modello. Quest'ultimo descrive molto bene la variabilità dei dati sperimentali del set di taratura. Tuttavia occorre valutare la capacità di generalizzazione del modello ossia testarne la sua capacità previsionale con dati non adottati per calibrarlo, sui dati del *test set.*

Nella Figura 7 sono riportati i livelli di falda sperimentali del *test set* raffrontati con l'uscita del modello nella previsione rispettivamente di uno, due, quattro, sei mesi in avanti ed in simulazione. Nella figura sono riportati i valori degli indicatori di *performance* statistica già ricordati.

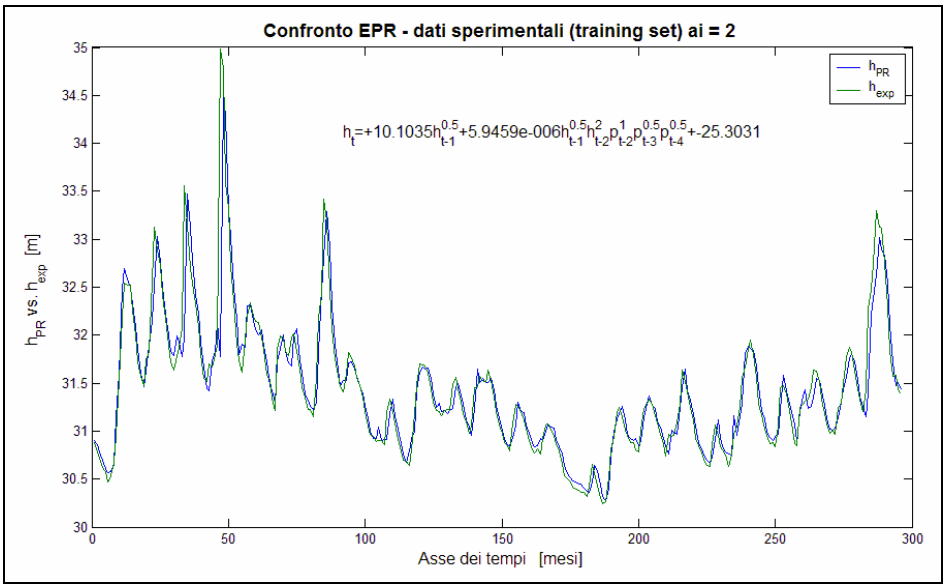

Figura 6 – Confronto tra l'uscita del modello EPR ed i dati sperimentali di taratura. h_{PR} rappresenta l'altezza del livello prevista mentre h_{EXP} è quella osservata.

Figure 6 - Comparison between experimental data from the training set and EPR forecast. h_{PR} represents predicted groundwater level while h_{EXP} is the observed.

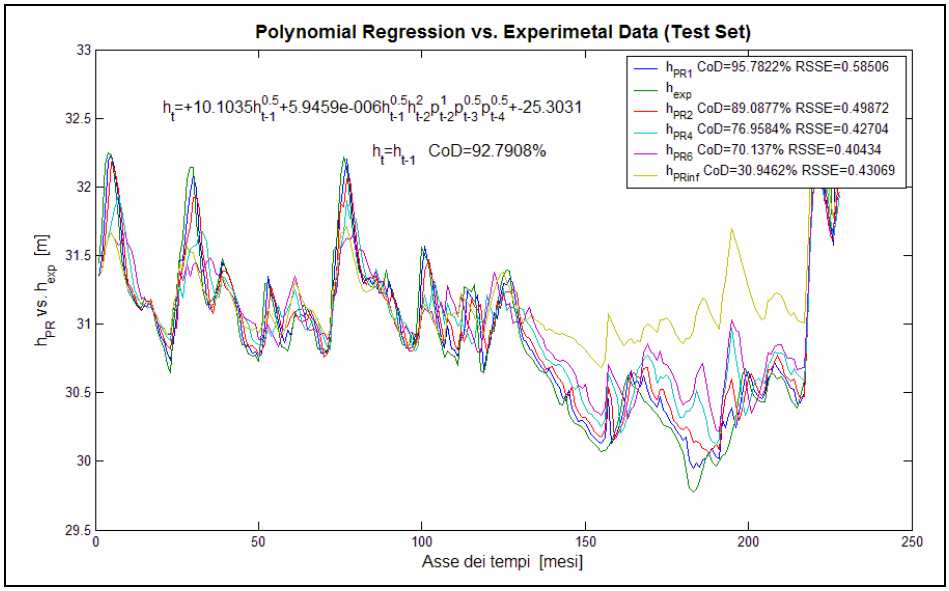

Figura 7 - Raffronto delle quote piezometriche previste con i dati del *test set.*

Figure 7 - Comparison between predicted and observed groundwater levels from the test set.

Il coefficiente di determinazione per l'adattamento ai dati nelle previsioni in avanti fino a sei mesi sono superiori al 70% e la somma degli errori quadratici medi è inferiore al 40% della SSE del modello di persistenza.

Ad una prima analisi della figura si evidenzia un errore grossolano di previsione del modello in corrispondenza dei mesi 182-188 e nei mesi 192-195 del *test set*. Questo errore di previsione di EPR si manifesta in corrispondenza dei mesi in cui i rilievi fratimetrici risultano assenti e per i quali la serie è stata localmente ricostruita mediante *spline* cubiche. Tale errore è tanto più evidente quanto più ci si spinge in avanti con la previsione, fino a presentare una discordanza totale. L'episodio riscontrato nel *test set* deriva probabilmente dal fatto che tali dati potrebbero anche essere completamente errati perché ricostruiti esclusivamente sulla base di una semplice interpolazione numerica sulla serie dei livelli di falda, senza tenere conto della forzante pioggia e della dinamica del sistema nel suo complesso. Di fatto potrebbero essere più affidabili i dati ottenuti attraverso la previsione EPR di quelli ottenuti attraverso la semplice interpolazione numerica, perché privi dell'informazione relativa alla forzante. La capacità di previsione di EPR è generalmente molto buona sulla base degli indicatori statistici sul *test set,* sebbene la qualità dai dati e la presenza di gap temporali può ridurne la efficacia. Un buon modello *data-driven* deve saper cogliere il massimo dell'informazione storica contenuta nei dati. A tal fine, si è ritenuto molto utile effettuare un'analisi di correlazione degli errori residui del modello (9). Questi ultimi sono ottenuti per sottrazione dei valori previsti con i dati sperimentali del *test set*. Se nei residui del modello è presente una informazione che possa essere correlata con le serie storiche dei dati di *input* o con i residui stessi, allora il modello non spiega completamente la dinamica espressa dai dati.

In Figura 8 sono riportati i valori del coefficiente di autocorrelazione ai vari *lag* (sfalsamenti temporali) dei residui. Questo correlogramma assume notevole importanza nel verificare che non vi sia informazione residua negli errori del modello.

Nella stessa figura vi è poi il diagramma della correlazione incrociata dei residui del modello con la serie storica delle precipitazioni, ai vari *lag*, il quale può rivelare eventuali relazioni tra l'uscita del modello ed il termine forzante, l'*input* pioggia. I correlogrammi mostrano un andamento attorno allo zero e contenuto in maniera più che soddisfacente nella fascia di confidenza del test al 95%.

Si può concludere pertanto che i residui del modello rappresentano soltanto un rumore di fondo o, perlomeno hanno un contenuto bassissimo di informazioni. Il modello ha colto esaurientemente la dinamica del fenomeno di risposta della falda alle precipitazioni in quanto non sussistono correlazioni tra la serie dei residui e quella dell'*input*.

Per confronto, si riporta un altro modello individuato da EPR con struttura più complessa, a tre termini monomi e bias. La miglior regressione simbolica a tre termini ottenuta è rappresentata dall'equazione (10):

$$h_t = 10.06 \cdot \sqrt{h_{t-1}} + 0.0074 \cdot p_{t-2} \cdot \sqrt{h_{t-2} \cdot p_{t-3}} + \\ +5.7 \cdot 10^{-8} \cdot \left[h_{t-1} \cdot h_{t-2}\right]^2 \cdot \sqrt{p_{t-1} \cdot p_{t-4} \cdot p_{t-5}} - 25.128 \qquad (10)$$

Anche in questo caso vi è un termine di persistenza non lineare, del tutto simile a quello riscontrato nella equazione (9) e questa volta due termini, entrambi legati alle precipitazioni, anch'essi non lineari. Il modello rappresentato dalla (10) è certamente più complesso e di difficile interpretazione fisica. Tuttavia in previsione a breve e medio termine, si dimostra più aderente ai dati del *test set,* come evidenziato nei risultati riportati nella Tabella 3.

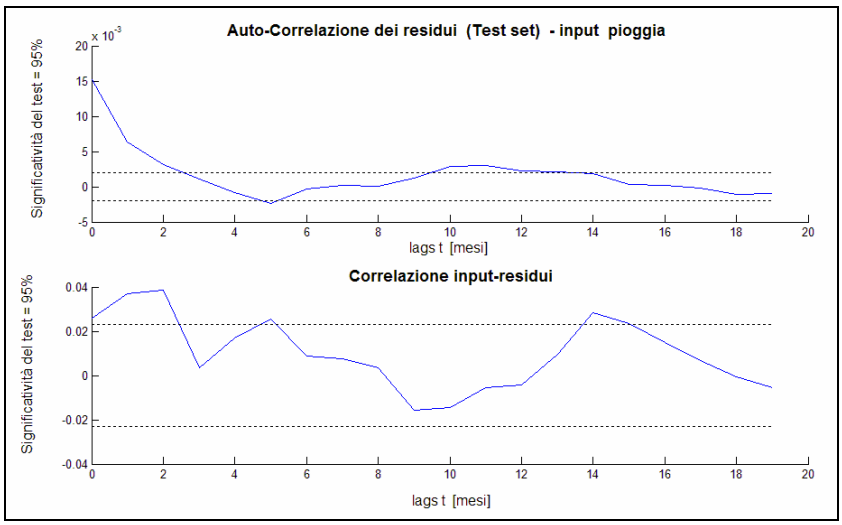

Figura 8 - Correlogrammi tra le serie storiche dei residui e degli input.
Figure 8 - Auto-correlation of residuals and cross-correlation between input and residuals

Tabella 3 - Adattamento dei modelli ai dati del *test set*.
Table 3 - Adaptation of identified model to test-set data.

CoD	modello a 2 termini ajZj	modello a 3 termini ajZj
pr. 1 mese	95.78 %	95.36 %
pr. 2 mesi	89.09 %	89.42 %
pr. 4 mesi	76.96 %	78.83 %
pr. 6 mesi	70,14 %	73.59 %
simulazione	30,95 %	50.55 %

Nella tabella sono riportati i valori del CoD. per vari orizzonti temporali di previsione sui dati sperimentali del *test set*. Nella previsione in avanti di un mese, il modello (9) presenta una accuratezza maggiore, mentre per orizzonti temporali più lontani, è il modello (10) a mostrarsi più efficace. Quest'ultimo pertanto può rappresentare un miglior strumento di previsione, anche nel caso in cui ci siano alcuni dati scarsamente affidabili o ricostruiti, ma risulta più simile ad un modello di tipo *black-box*.

Infine, merita una menzione particolare la regressione simbolica a due termini ottenuta da EPR applicando una tecnica di ricerca con forte penalizzazione di complessità:

$$h_t = 0.069 \cdot p_{t-2} + 10.47 \cdot \sqrt{h_{t-1}} - 27.38 \qquad (11)$$

Il modello è composto di due termini contenenti variabili isolate perché, evidentemente, la forte penalizzazione ha spinto la ricerca evolutiva a selezionare solo i termini particolarmente semplici e rappresentativi al tempo stesso.

Tale modello non risulta effettivamente utile nelle previsioni, dati i significativi errori riscontrabili. Tuttavia, tra tutti i modelli costruiti da EPR sulle serie storiche pluvio-freatimetriche di Brindisi, la (11) rappresenta il più parsimonioso ed i termini selezionati da EPR per la sua costruzione sono h_{t-1} e p_{t-2}, variabili fondamentali che sono presenti in tutti i migliori modelli.

5.3 Discussione dei risultati

La ricerca effettuata da EPR ha prodotto numerosi modelli a differenti livelli di complessità. I modelli con adattamento migliore ai dati sono risultati quelli con due, tre e quattro termini. I modelli ancora più complessi hanno manifestato un adattamento superiore ai dati del *training set* e peggiore al test set. Tali modelli soffrono pertanto di *overfitting*, ossia di sovradattamento ai dati di calibrazione che poi induce il modello ad avere una povera capacità di generalizzazione su nuovi dati. Alcune considerazioni di carattere interpretativo si possono invece riportare in merito ai modelli più semplici. Questi ultimi, essendo espressi da equazioni simboliche relativamente meno complesse, possono talvolta prestarsi ad interpretazioni fisiche, contribuendo a comprendere i meccanismi di risposta dell'acquifero alle precipitazioni.

In altri termini, allorquando la qualità e quantità dei dati risulti soddisfacente, è possibile leggere la dinamica del sistema idrogeologico così come appare dalle serie storiche delle sue variabili di stato (quote freatimetriche) e forzanti (precipitazioni).

La costruzione di queste formule è il risultato di una ricerca in uno spazio molto ampio di strutture del tipo (5), definite su un dominio rappresentato dall'insieme dei dati sperimentali delle serie storiche pluviometriche e freatimetriche. Le equazioni dedotte sono leggi di relazione tra queste variabili, interamente basate sui dati.

Con riferimento all'equazione (9), in essa figura un termine di persistenza non lineare in cui la quota di falda è isolata rispetto alle altre variabili di previsione. Questa caratteristica è stata riscontrata in numerosi altri modelli identificati da EPR, anche in quelli con minori capacità previsionali. In base all'equazione (9), si può pertanto ipotizzare che, in assenza di precipitazioni nei mesi precedenti, questo termine di persistenza, insieme al termine noto, rappresenti la legge di decadimento del livello di falda. Ipotizzando una quota piezometrica iniziale di 34 m s.l.m.m. e assenza di pioggia per svariati mesi, si ottiene per *h(t)* l'andamento mostrato in Figura 9.

La curva di decadimento dei livelli di falda presenta una lieve concavità rivolta verso l'alto: dapprima più ripida, tende successivamente, col passare dei mesi non piovosi, ad una riduzione della velocità di abbassamento della quota piezometrica. Ciò può essere spiegato col fatto che, riducendosi il volume idrico immagazzinato nell'acquifero per mancanza di nuovi apporti meteorici, si abbassano i gradienti idraulici, limitando così il drenaggio attraverso il reticolo idrografico, il deflusso verso il mare, nonché l'infiltrazione a favore dell'acquifero profondo. Tale effetto risulta accentuato soprattutto per lo spessore relativamente modesto della falda.

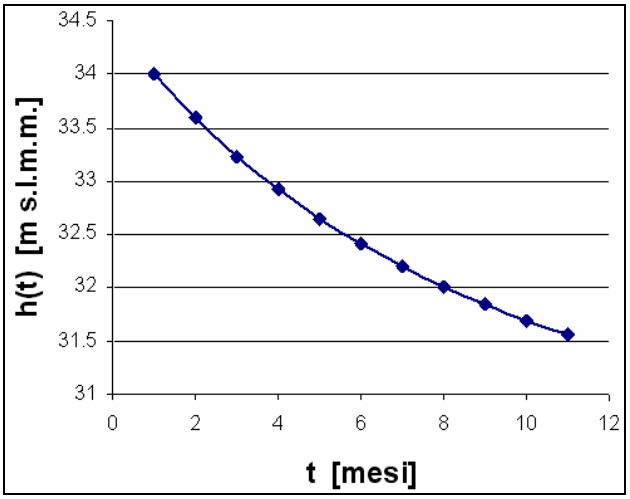

Figura 9 - Decadimento del livello di falda in assenza di precipitazioni.
Figure 9 - Conjectured groundwater level in absence of precipitation.

Sempre nell'equazione (9) compare un altro termine monomio piuttosto complesso, sul quale si possono fare alcune considerazioni. Questo termine monomio è costituito da una combinazione che è il prodotto di numerose variabili di ingresso, le precipitazioni di due, tre e quattro mesi prima, e le quote piezometriche della falda dei due ultimi mesi. L'afflusso meteorico, il termine forzante del sistema, è contemplato dal modello in questo secondo termine. In questo modello, è interessante osservare che, durante la ricerca evolutiva, EPR non ha selezionato come variabile di previsione in ingresso la precipitazione del mese precedente p_{t-1}: evidentemente il tempo di risposta dell'acquifero alle precipitazioni nella stazione freatimetrica è generalmente superiore al mese e più vicino ai due mesi. Infatti il termine p_{t-1} non è selezionato da EPR e pertanto non è portatore di un livello di informazione significativo ai fini della previsione di h_t. E' significativo il fatto che la pioggia totale mensile al tempo *t-2* compare nella (9) con un esponente maggiore rispetto agli altri mesi figuranti, assumendo un ruolo dominante, probabilmente perché incide maggiormente nella dinamica dell'escursione del livello di falda al tempo *t*. Tutto il secondo termine si annulla se non si sono avute contestualmente precipitazioni in tutti e tre i mesi. In questo caso solo la legge di decadimento determina la dinamica del sistema. Ciò è plausibile se si considera che, in assenza di precipitazioni, la riserva idrica di campo del suolo viene intaccata per prima, e per prima deve essere ripristinata quando sopraggiungono nuovi afflussi meteorici. Il modello risulta pertanto insensibile a precipitazioni anche importanti, che sopraggiungano dopo mesi siccitosi, mentre risponde con un incremento del livello di falda a periodi con piovosità modesta ma regolare. Il secondo termine monomio contiene anche il prodotto di h_{t-2} e h_{t-1} che smorza o accentua l'incidenza delle precipitazioni sull'uscita del modello. Se questi livelli sono più elevati, le precipitazioni dei mesi precedenti hanno un effetto maggiore nel causare l'innalzamento della superficie piezometrica. In effetti un livello freatico alto, nell'anno idrologico medio, può essere indirettamente rappresentativo di un certo grado di saturazione del sottosuolo. Una precipitazione che trova un suolo relativamente umido produce maggior infiltrazione efficace ai fini della ricarica.

Il modello espresso nell'equazione (10) presenta una struttura a tre termini più bias.

Anche in questo caso vi è un termine di persistenza non lineare, del tutto simile a quello riscontrato nella equazione (9) e questa volta non uno ma due termini legati alle precipitazioni, anch'essi non lineari. La precipitazione con un peso maggiore nella formula rimane p_{t-2}, con esponente uno, ma questa volta compaiono, con un ruolo più marginale, anche p_{t-1} e p_{t-5}. Il modello rappresentato dalla (10) è certamente di più difficile interpretazione fisica rispetto a (9). Tuttavia in previsione a medio e lungo termine, si è dimostrato più aderente ai dati del *test set,* come evidenziato nel paragrafo precedente. Di contro, quest'ultimo si presta più difficilmente ad una spiegazione fisica, e presenta una struttura più simile ad una formula di regressione pura, piuttosto che a carattere concettuale. Pertanto può rappresentare un miglior strumento di previsione ma più vicino ad un modello di tipo *black-box.*

In merito al modello (11), una formula di regressione simbolica a due termini ottenuta con una forte penalizzazione della complessità, si può evidenziare come quest'ultimo sia composto di due termini contenenti variabili isolate. Probabilmente, la forte penalizzazione ha spinto la ricerca evolutiva a selezionare solo i termini particolarmente semplici e rappresentativi al tempo stesso. Tale modello non risulta effettivamente utile nelle previsioni, dati i significativi errori riscontrabili. Tuttavia, tra tutti i modelli costruiti da EPR sulle serie storiche pluvio-freatimetriche di Brindisi, la (11) rappresenta il più parsimonioso. In esso i termini selezionati sono h_{t-1} e p_{t-2}. Si è riscontrato in tutti i modelli identificati da EPR che queste variabili sono sempre presenti e figurano con esponenti tali da avere un peso maggiore delle altre. Si potrebbe dedurre che queste variabili rappresentano di fatto le grandezze fondamentali nella dinamica delle escursioni freatiche dell'acquifero.

In conclusione, sulla base dei modelli costruiti con questa tecnica, l'acquifero risponde alle precipitazioni con circa due mesi di ritardo e proprio l'afflusso meteorico di due mesi prima sembra incidere nella variazione della superficie freatica in modo più significativo. Nei modelli più complessi, che presentano un'approssimazione migliore ai dati di verifica del *test set,* questi due valori sono comunque sempre presenti, ma compaiono anche gli altri valori storici, in particolare di precipitazione, soprattutto p_{t-3} e p_{t-4}. Le variabili p_{t-1} e p_{t-5} sono selezionate solo nei modelli più complessi: evidentemente il loro contributo nel produrre variazioni nella superficie freatica è ridotto oppure, semplicemente, il loro effetto sulla ricarica della falda è correlabile statisticamente con le altre variabili di precipitazione, e pertanto rappresentano portatori di informazione meno significativi rispetto a queste ultime.

6. Conclusioni

Nel presente lavoro l'acquifero superficiale di Brindisi è stato modellato facendo ricorso ad una nuova tecnica, denominata *Evolutionary Polynomial Regression*, allo scopo di studiare la risposta dinamica del sistema idrogeologico alle precipitazioni. I modelli, costruiti a partire dalle serie storiche disponibili, consistono in equazioni simboliche esplicite che relazionano le variabili idrologiche del sistema nel tempo e possono essere adoperati per scopi previsionali. Le equazioni generate possono variare nella complessità e capacità di descrivere

il fenomeno ma presentano termini simbolici che ne rendono possibile la leggibilità.

Questi modelli possono essere usati sia a scopi previsionali che conoscitivi. Nel primo caso assolvono il compito di prevedere l'andamento nel tempo di una certa grandezza di interesse, in questo caso le escursioni piezometriche, in relazione alle forzanti del sistema, le precipitazioni. Poiché le equazioni del modello sono a carattere simbolico, possono prestarsi ad interpretazioni fisiche. Pertanto possono rappresentare uno strumento di analisi dell'importanza relativa dei processi coinvolti e nel determinare quali grandezze hanno maggior rilevanza sulla dinamica del sistema. Alcune interazioni tra variabili coinvolte nel processo modellato potrebbero non essere colte se deboli in relazione al rumore di fondo. La qualità dei dati può rappresentare sempre una limitazione per questo tipo di approcci, specialmente nelle previsioni a più lungo termine, sebbene EPR abbia mostrato una buona capacità di generalizzazione sia attraverso un test numerico che sui dati del *test set.* La costruzione di queste formule è il risultato di una ricerca in uno spazio molto ampio di strutture possibili definite su un dominio rappresentato dall'insieme dei dati sperimentali delle serie storiche pluviometriche e freatimetriche. Le equazioni dedotte sono leggi di relazione tra queste variabili, interamente basate sui dati.

I risultati dell'applicazione di questa tecnica all'acquifero superficiale del territorio di Brindisi sono state molto incoraggianti. La correlazione dei residui del modello con i dati di ingresso risulta praticamente nulla. Gli errori residui sulle previsioni a diversi orizzonti temporali sono bassi ed alcuni termini si prestano ad interpretazioni fisiche.

La risposta dinamica del sistema appare fortemente condizionata dal tipo di distribuzione delle precipitazioni nel tempo più che dalle quantità, in accordo con quanto osservato dal punto di vista generale. La tecnica ha prodotto modelli matematici di meccanismi che sono effettivamente individuabili dal punto di vista qualitativo. Una precipitazione intensa che segue due mesi non piovosi non produce alcun effetto sull'innalzamento di falda, secondo i modelli selezionati da EPR per le serie storiche date. Alcuni termini ed alcune variabili sono ricorrenti nelle diverse equazioni, indipendentemente dal loro grado di complessità. Tali termini veicolano maggiori informazioni ai fini della previsione. Questo può dare indicazioni sui meccanismi di risposta alle precipitazioni e di ricarica. In particolare, il ritardo nel picco della risposta dell'acquifero nell'anno idrologico medio risulta superiore al mese evidenziando in maniera analitica come i termini che producono un incremento dei livelli freatici, ovviamente legati alle precipitazioni, risentono fortemente delle condizioni di saturazione nel sistema condizionando la risposta dell'acquifero.

Riferimenti bibliografici

ASCE Task Committee (2000). On Application of Neural Networks in Hydrology, II: hydrologic applications. Journal of Hydrologic Engineering, ASCE, 5(2), 124-137.

Babovic V. and Abbott M.B. (1997). The evolution of equations from hydraulic data, Part I: Theory. Journal of Hydraulic Research, Vol. 35, No 3, pp.1-14.

Dane J.H., Wierenga P.J. (1975). Effects of Hysteresis on Prediction of Infiltration, Redistribution and Drainage of Water in Layered Soils, Journal of Hydrology, vol. 25, pp 229-242.

Doglioni, A., Giustolisi, O., Savic, D. A., Webb B.W. (2007). An investigation on stream temperature analysis based on evolutionary computing, Hydrological Processes 21, DOI: 10.1002/hyp.6607

Giustolisi O. (2000). Simulating a Urban Drainage System by Non-Linear Time Invariant Dynamic Systems: Neural Networks, 4th International Conference on Hydroinformatics, Iowa City, Iowa State, USA 23-26 July.

Giustolisi, O. and Savic, D.A. (2003), Evolutionary Polynomial Regression (EPR): Development and Applications. Report 2003/01. School of Engineering, Computer Science and Mathematics, Centre forWater Systems, University of Exeter, 2003.

Giustolisi O. & Savic D.A. (2006) A Symbolic Data-Driven Technique Based on Evolutionary Polynomial Regression, J. of Hydroinformatics, 8(3), 227-222.

Goldberg, D.E., Genetic Algorithms in Search, Optimization and Machine Learning. Addison

Wesley, 1989.

Golub G.H. and Van Loan C.F. (1993). Matrix Computations. The Johns Hopkins University Press Ltd., London, UK.

Guymon G.L. (1994). Unsaturated Zone Hydrology, Prentice Hall, Englewood Cliffs, New Jersey.

Horton, R.E. (1940). An approach to a physical interpretation of infiltration capacity, Soil Science Society of American Proceedings, vol. 5, pp. 399-417.

Kecman, V. (2000). Learning and Soft Computing by SVM, NN and FLS, The MIT Press, Cambridge, MA.

Keijzer M. & Babovic V. (2000). Genetic Programming within a Framework of Computer-Aided Discovery of Scientific Knowledge – Proceedings of the Genetic and Evolutionary Computation Conference GECCO july 2000

Ljung, L., System Identification: Theory for the User 2e. Prentice-Hall Inc., UpperSaddle River, New Jersey, USA, 1999.

Maggio F. (2001). Analisi quali-quantitativa della risposta alle precipitazioni di acquiferi superficiali e profondi pugliesi, Tesi di laurea (Politecnico di Bari, Facoltà di Ingegneria di Taranto).

Mancarella D.(2003). Strategie evolutive nella realizzazione di modelli per la gestione delle risorse idriche sotterranee, Tesi di Laurea, Politecnico di Bari, Facoltà di Ingegneria di Taranto.

Mancarella D., Doglioni A, Simeone V., Giustolisi O. (2007). Inferring groundwater dynamics from time series data. International Conference ModelCARE 2007 - Calibration and Reliability in Groundwater Modelling – Copenhagen, September 9-13, 2007

Ricchetti E., Polemio M. (1996). L'acquifero superficiale del territorio di Brindisi: dati idrogeologici diretti e immagini radar da satellite, Memorie Della Società Geologica Italiana, Vol. LI, fascicolo II (1996), ISSN 0375-9857.

Richards L.A., 1931. Capillarity Conduction of Liquids in Porous Media, Physics, vol. 1, pp 318-333.

Wanakule, N. & Anaya, R. (1993). A Lumped Parameter Model for the Edwards Aquifer. Texas A&M University, College Station, Texas Water Resources Institute, Technical Report n. 163.