Aspetti preiliminari del problema

Abbiamo degli acquiferi di vario tipo (fiumi, laghi, falde, sorgenti) e vogliamo prevedere l’andamento delle misurazioni di alcune grandezze fisiche associate all’acquifero oggetto di studio (livello idrometrico, portata della sorgente, livello della falda, ecc.); possiamo chiamare le misurazioni di queste grandezze “variabili target”. Oltre alle serie temporali delle variabili target disponiamo delle serie temporali di altre grandezze fisiche che ipotizziamo siano utili (insieme allo storico della variabile target) a predire i valori futuri delle variabili target, chiamiamo queste variabili indipendenti “predittori” o “feature predittive” o semplicemente “feature”.

Possiamo immaginare l’acquifero come un sistema dinamico complesso che riceve alcuni segnali di stimolo in input (siano essi di tipo inflow o outflow) e produce un insieme di risposte in output ovvero le variabili target con un certo ritardo (con un lag di risposta specifico per ogni variabile target rispetto ai diversi input).

Per un dato acquifero avremo la stessa grandezza fisica di interesse misurata da dispositivi collocati in diversi luoghi. Per comodità nel seguito parlerò della grandezza “livello della falda” ma i concetti esposti possono essere, mutatis mutandis, riferiti ad altre grandezze (portata di una sorgente, livello idrometrico, ecc.).

Tipicamente avremo la stessa grandezza (livello della falda) misurata da N diversi dispositivi dislocati in punti sparsi sul territorio oggetto di interesse; quindi avremo N diverse variabili target della stessa tipologia (associate allo stesso acquifero ma rilevate in punti distinti). Analogamente per i predittori, avremo la stessa grandezza fisica misurata in punti diversi come il livello delle precipitazioni misurato in P luoghi diversi da P pluviometri (e.g. mm di pioggia misurati da P pluviometri sparsi sul territorio).

Per esemplificare, consideriamo l’acquifero di Petrignano (dataset ACEA):

n. 2 variabili target:

* 1. P24, P25 (livello della falda)

n. 1 predittori essenzialmente legati all’inflow:

* 1. Precipitazioni in mm in Bastia Umbra, rainfall Bastia\_Umbra

n. 1 predittore legato all’outflow:

* 1. Emungimenti a fini potabili dall’impianto C1 in Petrignano, C10\_Petrignano

n. 2 predittori legati principalmente all’inflow ma anche all’outflow:

* 1. Temperature misurate in Bastia\_Umbra, Petrignano, Bastia\_Umbra, Petrignano  
     (la temperatura impatta l’inflow cambiando l’infiltrazione efficace (% delle precipitazioni che raggiunge la falda, e.g. neve, terreno ghiacciato, ecc.) ma può impattare anche l’outflow, il consumo di acqua in genere aumenta nel periodo estivo)
  2. livello idrografico del fiume Chiascio in Petrignano, Fiume\_Chiascio\_Petrignano: il bacino del fiume può alimentare la falda (inflow) ma bassi livelli del fiume posso essere correlati a periodi siccitosi ed indurre un maggiore utilizzo della falda in sostituzione delle acque fluviali (in periodi normali/non siccitosi le falde non vengono sfruttate per scopi irrigui, al contrario delle acque superficiali)

Le variabili sono numerose e presumibilmente esistono legami non lineari tra di esse (con lag che variano nel tempo, e.g. stagionalmente). La complessità è legata alla eterogeneità delle caratteristiche degli acquiferi, alla variabilità nel tempo per alcune di esse, e quindi alla variabilità nel spazio-temporale degli effetti dei predittori sulle variabili target.

**In primis, le serie temporali vanno studiate per capire le loro caratteristiche (e.g. stazionarietà, change point di media e varianza, presenza di trend, stagionalità, ecc.). Sicuramente troverete facilmente package R per svolgere questo tipo di analisi.**

Fatta questa analisi preliminare, dobbiamo cercare di **modellare le relazioni tra le variabili del problema**. Per cercare di attaccare il problema dobbiamo puntare a creare un modello approssimato delle relazioni tra le variabili. Consideriamo alcune semplificazioni.

1. Preprocessing delle time series pluviometriche: precipitazioni nevose.

Separare le time series della pioggia in precipiazioni piovose e precipitazioni nevose (trascurando gli effetti della temperatura sul terreno). Un primo esempio di semplificazione consiste nella valutazione del ruolo della temperatura sia sulle precipitazioni sia sulle caratteristiche del terreno attraverso cui fluisce l’acqua. Assumiamo in prima approssimazione che gli effetti della temperatura sulle precipitazioni siano significativi solo quando la temperatura è uguale o inferiore a -1°C, separando le time series della pioggia in due time series diverse (una per la pioggia e una pe le precipitazioni nevose). In sostanza trascuriamo gli effetti della temperatura sul terreno perchè assumiamo che il rallentamento nella filtrazione dell’acqua in un terreno gelato si manifesti solo per temperature <= -1°C e sia quindi concomitante con le precipitazioni nevose, quini il lag della time series della neve includerà anche il ritardo legato al terreno gelato.

1. Preprocessing delle time series pluviometriche: precipitazioni ininfluenti (infiltrazione efficace nulla).

Alcune precipitazioni di bassa intensità e breve durata producono un’infiltrazione efficace trascurabile (la percentuale di acqua piovana che arriva in falda e praticamente nulla); ogni zona dell’area di ricarica presenterà un proprio specifico valore minimo (threshold) delle precipitazioni cumulate superato il quale si potranno misurare effetti sulla falda; quindi per ognuna delle misurazione pluviometriche possiamo generare diverse serie temporali ognuna associata ad uno specifico valore di threshold. Ad esempio dalla serie originaria possiamo generare due time series, la prima che pone a zero le precipitazioni con un valore cumulato settimanale inferiore ad x mm di pioggia e la seconda che pone a zero le precipitazioni con un valore cumulato settimanale inferiore ad y mm di pioggia (e così via per altri threshold). lo scopo di generare nuove serie temporali pluviometriche “tagliate” è quello di eliminare dallo stimolo di inflow valori che rappresentano (per gli effetti di falda) solo rumore in quanto hanno un impatto trascurabile sui livelli di falda. Eliminando dallo stimolo il “rumore”, è possibile misurare più correttamente la correlazione tra stimolo e risposta ed identificare più precisamente i lag caratteristici (discorso analogo all’identificazione delle precipitazioni nevose).

1. Identificazione di Threshold e Lag ottimali da applicare alle feature predittive (associati a specifiche condizioni).

Poichè non conosciamo i threshold che possono variare sia stagionalmente sia in funzione delle condizioni di saturazione del terreno, dobbiamo analizzare l’andamento settimanale di questi valori di threshold. Se la disponibilità di dati lo consente, ha senso effettuare analisi distinte (e creare modelli predittivi distinti) per tre classi di anni molto diverse tra loro (anni siccitosi as, anni normali an, anni molto piovosi ap). Questa classificazione degli anni in tre gruppi ha senso dal punto di vista delle serie temporali ma andrebbe **corroborata da un’analisi di clustering delle time series** (vedi folder 19 e 20).

La determinazione del lag ottimale viene di solito determinata attraverso criteri di parsimoniosità (quanta informazione porta l’aggiunta di un ulteriore valore predittivo?) e si utilizzano solitamente criteri come l’AIC o il BIC. Tuttavia, come visto, non è detto che esista un lag ottimale per tutte le condizioni (tipo di predittore, tipo di anno, stagione, condizioni del terreno, livello della falda, ecc.) e la scelta di un lag “medio” da usare sempre è sicuramente più veloce ma in alcuni casi non ottimale. Quindi una volta ripulite le time series da falsi segnali (piogge cumulate basse e precipitazioni nevose) si devono analizzare queste serie per trovare threshold e lag ottimali specifici per le condizioni che ricorrono periodicamente come: anno as, an, ap, stagione di ricarica o di scarico dell’acquifero, livello della falda compreso tra determinati livelli “cruciali”.

Per livelli “cruciali” si intende i livelli di falda (se esistono) che (in condizioni di similarità di anno, stagione, livello delle precipitazioni cumulate, temperatura, ecc.) sono associati a regime shift come cambi nei valori ottimali del lag, cambio del valore medio della distribuzione della velocità di salita/discesa del livello della falda. Per evidenziare un regime shift (change point) nelle variabili target e nei predittori possono essere evidenziati con specifici package (vedi folder 01) oppure vedi <https://lindeloev.github.io/mcp/articles/packages.html.>

A livello empirico per avere un’idea di partenza di alcuni valori caratteristici (e.g. velocità di ricarica o scarico, lag l, threshold h) si possono cercare nelle time series periodi in cui alcune variabili predittive sono inattive (e.g. mancanza di precipitazioni) o sono stabili (valore della falda stabile), in questo senso i periodi all’inizio delle fasi di ricarica e di scarico dell’acquifero potrebbero fornire indicazioni utili perchè alcune variabili in gioco potrebbero aver esaurito i maggiori effetti sulla dinamica della falda, e.g. alla fine dell’estate/inizio autunno si dovrebbe, in genere, osservare una diminuzione dell’outflow e un periodo con precipitazioni assenti (o ridotte) per periodi sufficientemente lunghi (quindi gli effetti di breve periodo delle precedenti precipitazioni risulterebbero esauriti); le prime precipitazioni con infiltrazione efficace > 0, fornirebbero indicazioni su lag e tempi di ricarica della falda (in quelle condizioni: anno, stagione, livello della falda, assumendo che la temperatura abbia effetti solo se < 1°C).

Per ciasuna serie pluviometrica si possono valutare H valori di theshold h = 1, ..., H e L valori di lag l = 1, ..., L. Avremo quindi H x L serie temporali per ciascuna serie pluviometrica originaria. Per ciascuna settimana sceglieremo la prima coppia [h1, l1] e la seconda coppia [h2, l2] associate alle due time series modificate che risultano maggiormente correlate (coeff. di Spearman, ma vedi CANOVA o BNNPT anche folder 23) con la serie piezometrica della falda . Si dovrebbe priviligera una sequenza di h e di l che sia il più stabile possibile al variare del livello della falda; quindi si potrebbe costruire il pattern caratteristico di h e di l come una terza sequenza [h,l] ottenuta scegliendo di volta in volta ivalori tra h1 e h2 per h, l1 ed l2 per l che minimizzano la varianza della serie. Per scegliere h ed l.

Possiamo risparmiarci un po’ di lavoro considerando solo H serie e lasciando che la deterninazione del lag ottimale a programmi dedicati (criteri AIC, BIC, forse un’analisi lead-lag può essere utile (folder 22 e 22/optimal\_lag).

1. Assumiamo che l’outflow sia dominato dal fabbisogno idrico.

Supponiamo che su orizzonti temporali di breve termine (di interesse per il forecasting: giorni, settimane, mesi), l’outflow di un acquifero sia dominato dagli emungimenti (a fini potabili, industriali, irrigui, di allevamento, ecc.) poichè le tempistiche dell’outflow legato a fenomeni idrogeologici presenta tempistiche dilatate che non impattano in modo significativo la dinamica dell’acquifero di breve periodo (quest’ipotesi può risultare troppo approssimativa per acquiferi specifici). Quindi, se i predittori dell’outflow forniti dal committente (ACEA) descrivono in modo completo tutti gli emungimenti/outflow dell’acquifero, possiamo considerare risolto il problema di stima dell’outflow complessivo.

Rimangono da valutare i differenti effetti delle correlazioni spaziali tra le variabili. Potremmo considerare un modello che valuti esplicitamente le variazioni spazio-temporali e le correlazioni spazio-temporali, questo tipo di modelli esistono ma sono decisamente più complessi (se siete interessati posso fornirvi dei riferimenti). A causa della eterogeneità delle caratteristiche dell’acquifero possiamo aspettareci che alcuni piezometri presentino una dinamica delle misure più rapida/lenta a parità di sollecitazione (e.g. lag temporali diversi a parità di inflow o outflow).

1. Assumiamo che le correlazioni spaziali tra predittori producano effetti trascurabili.

Supponiamo che gli effetti di correlazione spaziale siano significativi solo tra predittori e variabili target (assumiamo che le correlazioni tra i predittori siano catturate dalle sole correlazioni temporali). Supponiamo inoltre che, queste correlazioni spaziali si manifestino producendo Trheshold h e Lag l diversi (a parità di stimolo).

Zone diverse dell’area di ricarica avranno coperture diverse (boschi, prati, zone urbanizzate, ecc.) con diversa porosità efficace, diversi valori di evotraspirazione, diverse pendenze (e quindi diversi effetti di ruscellamento), eccetera.

Facendo queste assunzioni semplificative cerchiamo di gestire l’eterogeneità dell’acquifero senza coinvolgere modelli che gestiscano le correlazioni spazio-temporali.

Ma perchè tutte queste complicazioni? Non possiamo semplicemente fare un’analisi di correlazione (Spearman) ed eliminare le feature ridondanti (per ridurre gli effetti delle collinearità o di relazioni nonlineari monotone) e poi fare l’analisi solo sulle feature selezionate? La risposta secondo me è dipende.

Dipende dalle caratteristiche dell’acquifero e dalle performance ottenibili con modelli semplici (e.g. assumendo che l’acquifero sia omogeneo). Si potrebbe adottare un approccio semplificato specie per acquiferi che non presentano caratteristiche fortemente eterogenee (dipenderà dal territorio e dall’estensione dell’area di ricarica). Consideriamo un caso semplificato di eterogeneità, un acquifero ipotetico per il quale si dispone di: 2 serie di misure piezometriche (due piezometri pz1 e pz2 collocati in parti opposte dell’area di ricarica, zone z1 e z2) e due serie pluviometriche (con i due pluviometri pv1, collocato vicino a pz1 e pv2 vicino a pz2). Supponiamo che le time series pv1 e pv2 siano fortemente/perfettamente correlate (e.g. i tempi di inizio aumento riduzione e termine delle precipitazioni sono uguali su tutta l’area di ricarica) ma che pv2 misuri livelli minori di pv1. Quindi i predittori sono sincronizzati ma pv1 presenta una infiltrazione efficace ie1 > 0 (supera il threshold h1) mentre la pioggia registrata da pv2 non è sufficientemente intensa e quindi ie2 = 0 (non supera il threshold h2). L’analisi di correlazione preliminare porterà a scegliere tra pv1 e pv2 (perfettamente correlate), supponiamo di scegliere pv1 (perchè ha una precipitazione cumulata maggiore nel periodo in esame).

Supponiamo infine che le caratteristiche idrogeologiche dell’acquifero portino l’infiltrazione efficace ie1 a spostarsi verticalmente fina alla falda in un tempo t1 e che successivamente il gradiente idraulico tra z1 e z2 (il livello piezometrico pz1 > pz2) generi uno spostamento orizzontale dell’acqua da z1 a z2 in un tempo t2 >> t1. In questo caso, pz1 presenterà una variazione significativa che risulterà correlata con la serie pv1 (opportunamente anticipata con un lag l1); al contrario il quantitiativo ie1 (rilevante nei pressi di pz1) si distriubuirà sulla restante parte della falda e la variazione del livello della falda in z2 (lontana da z1) potrebbe essere insignificante (oltre che con un lag decisamente maggiore).

Viceversa se avessimo preliminarmente scelto pv2 (eliminando pv1), avremmo osservato il (fuorviante) risultato che una modesta precipitazione in z2 ha generato effetti misurabili (e correlati) su pz1 e non ha prodotto effetti su pz2.

In generale, l’analisi preliminare di correlazione selezionerebbe solo uno tra due predittori fortemente correlati, ma i due predittori convogliano informazioni distinte a causa dell’eterogeneità dell’acquifero . Per questo sarebbe prudente sviluppare un modello per ciascuna variabile target considerando per tutti i predittori i possibili valori di threshold e lag.

Oltre a questo esempio limite, si devono tenere presente alcuni punti:

1. Gli algoritmi predittivi lavorano meglio in condizioni di maggiore omogeneità (da qui l’idea di ottimizzare le performance con parametri dei modelli per condizioni significative ricorrenti come as, an ap e stagioni di ricarica e scarico, piuttosto che adottare parametri “buoni per tutte le stagioni”).
2. Il coeff. Di Spearman rileva/misura relazioni (anche nonlineari) monotone (solo monotone) tra variabili.
3. La correlazione e la capacità predittiva non combaciano e nei modelli basati su alberi decisionali la ridondanza di dati può migliorare le performance predittive (anche se la collinearità nella Random Forest causa problemi con l’esatta valutazione della feature importance, nel gbm e nell’xgboost la collinearità dei predittori è meno problematica).

Possibili approcci risolutivi

In termini pratici, possiamo partirea applicando un approccio risolutivo parsimonioso e semplificato. In pratica assumiamo che non vi siano eterogeneità spaziali (“acquifero omogeneo”).

Approccio base per acquifero omogeneo.

Preprocessing.  
  
Missing value imputation.

Per intervalli ampi di missing value

<https://rdrr.io/cran/mice/man/ampute.html>Per maggiori dettagli sull’impotation vedi file IMP\_fimd.pdf (vedi folder 19)

oppure value imputation per clustering/similarità (vedi folder 19\_02, 19\_03)

Per intervalli brevi (1gg o 2gg), regressione lineare o media semplice .

Analisi di correlazione  
Pearson  
Spearman

CANOVA (folder 23)  
<https://rdrr.io/github/liyistat/canova/man/canova.html>

BNNPT (folder 23)  
<https://rdrr.io/github/liyistat/BNNPT/man/bnnpt.html>Modelli predittivi suggeriti: xgboost regression (vedi folder 14 e 16) e in misura minore Reg-SARIMA (vedi folder 13) o modelli standard vedi folder 02, 05.   
  
Nell’approccio di base ci affidiamo alla competenza del committente (ACEA) ed assumiamo che le feature predittive proposte siano idonee per la predizione della variabile target (e.g. livello piezometrico della falda). In particolare assumiamo che i contributi di ricarica dell’acquifero (inflow) siano esaustivamente rappresentati dalle misure pluviometriche fornite. Questa assunzione può rivelarsi ottimistica per alcuni acquiferi, ad esempio per l’acquifero di Petrignano viene fornita la misura delle precipitazioni registrate nel solo centro di Bastia Umbra. Per la misura dell’outflow si dovrebbe verificare che la misura fornita copra l’intero consumo di acqua captata per i diversi fini previsti (solo potabile e/o industriale e/o irriguo). In molti casi per le componenti dell’outflow non legate alla captazione da parte dei servizi idrici, si possono assumere tempi caratteristici dilatati che hanno un impatto lieve sulle dinamiche dell’acquifero nel breve periodo (giorni, settimane). Nel caso di specifici acquiferi tale assunzione potrebbe richiedere una verifica.  
In questo approcci di base le operazioni di feture engineering e selection potrebbero essere le seguenti:  
- missing value imputation;

- analisi di correlazione e rimozione delle variabili ridondanti (si scelgono/preservano le serie con varianza maggiore);

- se il modello lo richiede (e.g. regressione lineare, Reg-SARIMA), analisi di stazionalità (unit root test), trasformazione delle serie (e.g. log dei dati nel caso di distribuzioni skewed), scaling dei dati (e.g. normalizzazione), ecc.;

- separazione delle serie storiche pluviometriche in due serie distinte (pioggia o neve);  
- creazione di L serie storiche pluviometriche e di temperatura traslate in anticipo rispetto alla serie storica della falda con l’idea di valutare diversi possibili lag (l1 = 1 gg, l2 = 2 gg, ..., 1 settimana, .., 1 mese, ...).

- ulteriore modifica delle serie storiche traslate (anticipate) con creazione di H serie per ciascuna serie (quindi H x L serie, ognuna caratterizzata da threshold e lag [h, l].

Ogni variabile target avrà il proprio insieme di predittori (caratterizzati da threshold e lag) e se il clustering delle time series conferma la presenza di regime shift caratteristici/ricorrenti (anno as, an, ap e stagioni di ricarica o scarico), per ogni variabile target si potranno avere più modelli, ognuno con parametri ottimali per lo specifico regime.

Approccio base per acquifero eterogeneo.

Se le performance del modello base\_omogeneo non sono soddisfacenti o se si analizza un acquifero con una vasta area di ricarica e caratteristiche eterogenee, occorre evitare una rimozione delle feature (apparentemente ridondanti) in una fase preliminare (a valle dell’analisi di correlazione. Lo schema è analogo al modello con acquifero omogeneo ma è più oneroso perchè si ha un numero più elevato di predittori.

L’approccio per la selezione dei predittori miglori può essere simile a quello della stepwise forward regression o di tipo iterativo (vedi folder 25\_01).

Modelli predittivi suggeriti: xgboost (regression) vedi folder 14, 16 e modelli avanzati (reti neurali, MultiInputMultiOutput, spatio-temporal), vedi folder 04, 06, 16, 17, 18, 21.

Si possono fornire al modello ulteriori feature che vengono sfruttate nei modelli idrogeologici, oppure sfruttare un modello idrogeologico come riferimento per prevedere la dinamica della falda e poi implementare un modello (ad esempio basato su reti neurali) per prevedere l’andamento dell’errore associato al modello geologico (sfruttando gli stessi set di feature descritti spra), vedi folder 07, 08, 09, 10, 11, 12.

Approccio avanzato per acquifero eterogeneo, dati ACEA ritenuti non esaurienti, feature ispirate a modelli idrogeologici

(ed eventuale creazione di un modello idrogeologico a parametri concentrati).

In questo scenario dobbiamo completare le feature predittive fornite da ACEA con un insieme di altre feature (ricavabili da fonti dati pubblicamente accessibili). Prima di adottare questo approccio è necessario verificare che i predittori di inflow ed outlow siano non completamente rappresentativi. Si parte dall’outflow comparando gli emungimenti segnalati da ACEA con i fabbisogni idrici associati ai comuni serviti dall’acquifero. Poi si identifica l’area di ricarica dell’acquifero e si stimano alcuni parametri essenziali (copertura del territorio, evotraspirazione, pendenze del terreno, Curve number, porosità efficace, ecc.) per stimare l’infiltrazione efficace (e le sue fluttuazioni stagionali). Queste informazioni insieme ad altre associate ad altre permettono la calibrazione di modelli idrogeologici, vedi folder idrogeologia e sue sottocartelle.

Ecco un possibile algoritmo che definisce un approccio (empirico) per la valutazione di quali aree comunali contribuiscano ad alimentare un acquifero e con quali tempistiche si manifestino gli effetti di ricarica dell’acquifero.

Date una variabile dipendente (variabile target) e un insieme di variabili indipendenti (predittori) si vuole predire l’andamento della variabile target sfruttando le serie storiche dei predittori (oltre ai valori passati della variabile target). Supponiamo che nota la serie storica della variabile target fino all’istante t si voglia prevedere la variabile target all’istante t+X, dove X è un intervallo di tempo che definisce l’orizzonte temporale della previsione. L’algoritmo descritto di seguito è specifico per ciascun valore di X, quindi se X = 1gg, otterrò un certo modello predittivo, se X=1 settimana, otterò un diverso modello predittivo, ecc.

1. Individuo l’intersezione tra l’area dell’acquifero “A\_Acq” e tutte le aree comunali necessarie ad ottenere una “copertura” completa di A\_Acq i.e. individuo N comuni associati al processo di ricarica dell’acquifero.
2. Stimo la % di copertura di A\_Acq di ciascuna delle aree comunali (questo pu; essere fatto partendo da shapefile “.shp”oppure ricavato per alcuni acquiferi dalle fonti PDF individuate; l’ideale sarebbe determinarlo da .shp ed utilizzare le fonti PDF come verifica.
3. Separo ciascuna serie storica delle precipitazioni in due serie storiche distinte tramite il seguente split:
   1. la prima serie relativa alle precipitazioni;
   2. la seconda serie relativa alle nevicate;

se non trovo esplicitata questa distinzione, detemino i punti di spit identificando le unità di tempo (giorni) per le quali la temperatura è minore o uguale a -1 °C, giorni D\_snw; per tali giorni D\_snw considero 0.5 pioggia e 0.5 neve, mentre per tutti i giorni successivi con temperature che permangono <= -1 °C assumo 0.10 pioggia e 0.9 neve (sicuramente esistono modelli migliori di questo, non sono un esperto di meteorologia). Al primo giorno di pioggia a 0 °C considero il 100% di pioggia fino al successio giorno D\_snw.

1. Stimo la quantità di pioggia Q\_pgg associata a ciascuna area comunale e la sua evoluzione nel tempo ottenendo delle serie storiche Q\_pgg\_j\_t dove j = 1,..., N con N = numero dei comuni e t = t1, t2, ..., T tempi di misurazione delle precipitazioni nel periodo [t1, T].
2. Ordino i comuni per Q\_pgg in ordine di Q\_pgg decrescente; per ciascun comune j, costruisco le serie storiche traslate a partire dalla serie storica originaria Q\_pgg\_j\_t; le serie storiche traslate risulteranno sfasate in anticipo rispetto alla serie storica della variabile target per tener conto dei tempi caratteristici di filtrazione dell’acqua piovana. Se immaginiamo una sequenza di valori della variabile target V\_t1, V\_t2, ..., V\_tn, allora per ciascun comune j verranno costruite m serie storiche traslate Q\_pgg\_j\_t in modo che ciascuna serie storica risulti anticipata rispetto alla serie V\_tn (i.e. la causa “pioggia” anticipa l’effetto sull’acquifero):

serie Q\_pgg\_j\_t+k1 con una traslazione temporale di k1 unità temporali (e.g. k1 giorni, k1 settimane, ecc.)

serie Q\_pgg\_j\_t+k2 con una traslazione temporale di k2 unità temporali con k2 > k1

...  
serie Q\_pgg\_j\_t+km con una traslazione temporale di k2 unità temporali con k(m) > k(m-1).  
Oltre alla trslazione posso aumentare il numero di serie generate considerando diversi livelli di threshold h.

1. Una volta completata la costruzione dell Nxm serie storiche, si procede a creare N diversi modelli predittivi (xgboost di regressione, oppure basati su reti neurali) M\_c, c = 1, ..., N, (un modello predittivo per ciascun comune j), nel quale si adottano come predittori esclusivamente la serie Q\_pgg\_j\_t e le serie storiche traslate Q\_pgg\_j\_t+ki, i = 1, ..., m. Questi N modelli predittivi sono associati ad una specifica variabile target, quindi nel caso vi siano molteplici variabili target (come nel caso di predizione di diversi livelli piezometrici) si dovranno creare Nxm modelli per ciascuna variabile target. Normalmente questo non sarebbe necessario ma nei modelli che stiamo sviluppando non abbiamo incluso implicitamente nel modello la correlazione spaziale (tra predittori e tra variabili target). Quindi, se vogliamo lavorare con modelli base senza complicarli, possiamo valutare empiricamente quale feature risultino maggiormente predittive per una specifica variabile target. Ad esempio, supponiamo di dover predire l’andamento di due livelli piezometrici collocati in punti diversi dell’acquifero (variabili predittive Y e Z). Dai test potremmo scoprire per che il livello piezometrico Y (variabile target 1) le feature con maggiore valore predittivo siano la misurazione pluviometrica Q\_pgg\_2 (comune 2) con un anticipo (lag) di 5gg e la misurazione pluviometrica Q\_pgg\_5 (comune 5) con un anticipo di 3gg. Per il livello piezometrico Z (variabile target 2) le feature con maggiore valore predittivo potrebbero essere la misurazione pluviometrica Q\_pgg\_5 (comune 5) con un anticipo di 2gg e la misurazione pluviometrica Q\_pgg\_1 (comune 1) con un anticipo di 6gg. Questi effetti spaziali possono essere legati, ad esempio, alla eterogeneità del mezzo poroso che l’infiltrazione efficace attraversa per raggiungere la falda (o ad altre particolarità dell’acquifero). Quindi senza disporre di dati specifici sull’acquifero dobbiamo indagare e rivavare empiricamente la diversa importanza e rapidità di effetto delle precipitazioni (associate ad aree diverse) sulla ricarica dell’acquifero. In pratica cerchiamo di stimare i diverti tempi di ricarica associate alle diverse porzioni dell’area di ricarica dell’acquifero osservando il potere predittivo delle precipitazioni associate a queste porzioni di territorio.
2. L’eterogeneità delle caratteristiche idrogeologiche dell’acquifero pone anche un problema sulla definizione degli insiemi di training e di test degli algoritmi. Banalizziamo ed immaginiamo l’acquifero sia una bottiglia di vetro da 75cl. Se la bottiglia è quasi vuota (acquifero scarico a fine estate/inizio autunno) diciamo con soli 5cl di acua, l’aggiunta di 10cl (precipitazioni autunnali) farà salire il livello dell’acqua di x1. Se invece la bottiglia è quasi piena (acquifero all’inizio estate, dopo le precipitazioni primaverili) lo stesso quantitativo di 10cl di acqua farebbe salire il livello di x2 con x2 > x1 (ora sono vicino al collo della bottiglia, prima ero vicino al fondo). Il discorso vedrebbe x2 < xi se la “forma” dell’acquifero fosse quella di una bottiglia rovesciata. In pratica i tempi di ricarica e la risposta dinamica dell’acquifero alle precipitazioni risentono delle condizioni (livello) iniziale dell’acquifero. Glia lagoritmi predittivi lavorano meglio se applicati in condizioni “omogenee”, quindi se ci sono sufficienti dati, si dovrebbe cercare di selezionare i dati dell’insieme di training in modo che risultino coerenti con i dati dell’insieme di test. In pratica dobbiamo cercare di semplificare la vita all’algoritmo e dobbiamo cercare di addestrare l’algoritmo in condizioni simili a quelle che l’algoritmo incontrerà nella fase di test. Un modo per ottenere questo riultato è comparare tra loro periodi analoghi dal punto di vinsta delle condizioni/caratteristiche variabili dell’acquifero, in questo modo le feature predittive produrranno feature importance ed effetti comparabili tra la fase di training e di test. Questo è necesario per gli algoritmi basati su alberi decisionali perchè non possono interpretare fenomeni che non sono stati già osservato nella fae di training (con il deep learning si può in parte superare questo limite). Per esemplificare, dovremmo considerare un preprocessing delle serie temporali con questi obiettivi:
   1. svincolarsi dal calendario ed ancorare la scelta dei periodi di training e test sulla base delle caratteristiche stagionali e di ricarica dell’acquifero; si potrebbe considerare l’inizio del periodo come il punto di minimo della falda dell’acquifero (tipicamente a fine estate/inizio autunno); il periodo di carica terminerebbe alla fine della primavera, prima dell’inizio del picco di domanda estivo.
   2. Identificare anni con caratteristiche simili in termini di condizioni dell’acquifero; ad esempio in Umbria gli anni 2002, 2012, 2017 e 2019 sono stati anni siccitosi mentre il 2014 ed il 2015 sono stati anni di abbondanti precipitazioni (da verificare se vale anche per la Toscana). Potremmo decidere di valutare le performance del modello indipendentemente da queste condizioni dell’acquifero, oppure potremmo cercare di aumentare le performance del modello creando tre versioni con parametri ottimizza per i tre tipi di anno: anno siccitoso, anno normale, anno con precipitazioni abbondanti.
   3. Una volta scelti anni simili tra loro, possiamo verificare se un modello addestrato in un certo anno risulti predittivo in condizioni analoghe ma verificatesi in un altro periodo. Ad esempio, selazioniamo due anni con precipitazioni abbondanti, diciamo il 2014 ed il 2015 ed identifichiamo due date ad inizio autunno 2014 e 2015 con livelli di falda molto simili tra loro, diciamo d1\_14 e d1\_15. Identifichiamo una seconda coppia di date con livelli della falda simili al termine delle precipitazioni (probabilmente compresa tra la fine di aprile e la fine di maggio) supponiamo siano d2\_14 e d2\_15. Valutiamo gli intervalli di tempo Dt\_14r = [d1\_14, d2\_14] e Dt\_15r = [d2\_14, d2\_15] (r sta per ricarica) e scegliamo per il tarining quello con più varianza dei livelli piezometrici (un proxy della varietà di comportamento di precipitazioni e falda), se si hanno varianze molto simili scegliamo il più esteso tra i due; supponiamo che sia il Dt\_2015 quello scelto per la fase di training, cross validatio e hyperparameter optimization. Valutiamo le performance predittive dell’algoritmo per i dati del periodo Dt\_14. La stessa procedura verrà applicata nell fase di scarico dell’acquifero, quindi determinerò due intervalli Dt\_14s = [d3\_14, d4\_14] e Dt\_15s = [d3\_15, d4\_15]. Supponiamo che l’intervallo con maggiore vaianza della falda sia il Dt\_14s, quindi addestrerò l’algoritmo sul Dt\_14s e lo testerò su Dt\_15s. Alla fine avrò due algoritmi xgboost parametrizzati differentemente (uno per predire i livelli della falda durante il periodo di ricarica ed un secondo per i livelli del periodo di scarico). Un approccio analogo verrà seguito per gli anni siccitosi 2012, 2017 e 2019, e per gli anni “normali” 2006, .., 2011, 2013, 2016, 2018. In pratica avremo 6 modelli di regressione xgboost parametrizzati in modo differente per tipo di anno (siccitoso, normale, piovoso) e per stagione (ricarica o scarico dell’acquifero). Se i dati risultassero insufficienti per il training si dovrebbero concatenare intervalli di tempo analochi (ma di anni diversi) per creare un opportuno insieme di training, ad esempio potrei concatenare Dt\_11r, Dt\_13r e Dt\_16r per il training ed utilizzare il Dt\_18r per i test.  
      Se prendiamo ad esempio l’acquifero di Petrignano, abbiamo due variabili target (P24, P25) e quindi avremo un modello complessivo per ciascuna variabile target. Come variabili predittive, abbiamo le precipitazioni per il solo comune Di Bastia Umbra e abbiamo come ultimo predittore l’idrometria del fiume Chiascio. In termini operativi, possiamo assimilare l’idrometria ad una seconda precipitazione associata ad un secondo comune. Quindi alla fine avremo c= 2, due “comuni” (Basia Umbra e Fiume Chiascio). Per ottenere i due modelli finali M\_P24 ed M\_P25, parto dai modelli locali M\_1\_P24, M\_2\_P24 ed M\_1\_P25, M\_2\_P25.
3. Per ciascun modello M\_c, analizzo la feature importance ed escludo le serie con valore predittivo troppo basso (e.g. rimuovo/escludo le serie con coefficiente relativo inferore a 0.2, dove 1 è il coefficiente della feature con più alta importance).  
   Questa procedura dovrebbe aiutare ad stimare in modo approssimato i tempi caratteristici di filtrazione che variano da zona a zona nell’ipotesi (molto approssimata) che ogni area comunale sia associata ad un tempo di filtrazione medio omogeneo (e questa è una semplificazione accettabile) e stabile nel tempo (questa assunzione è meno accettabile perchè la filtrazione dipende dalle condizioni del terreno, e.g. saturazione, che variano nel tempo).  
   In definitiva per ciascun comune j avrò una lista di F\_j feature idonee sopravvissute alla scrematura (ovvero F\_j serie Q\_pgg\_j\_t+k con valore predittivo f >= 0.2, dove F\_j < m). Ad esempio nel caso di due soli comuni, per il comune 1 potremmo avere 3 serie idonee F\_1\_f1, F\_1\_f2, F\_1\_f3 (con potere predittivo f1 > f2 > f3 >= 0.2) mentre per il comune 2 potremmo avere 4 serie predittive F\_2\_f1, F\_2\_f2, F\_2\_f3, F\_2\_f4 (con potere predittivo f1 > f2 > f3 > f4 >= 0.2).
4. Ora costruisco un modello predittivo complessivo adottando la logica della stepwise regression (che è comoda anche se ha le sue criticità ed i suoi limiti) oppure con metodi più avanzati (vedi folder 25). Per farlo seleziono le feature escludendo quelle con basso valore predittivo. Prima di costruire il modello complessivo e procedere alla rimozione delle feature poco predittive, posso scegliere di escludere dal processo di cernita/selezione le feature che nei modelli locali M\_c hanno un alto valore predittivo. Ad esempio posso scegliere di preservare le feature con un coefficiente relativo (feature importance) superiore a 0.45 o 0.5, chiamiamo queste feature F\*, ad esempio F\*\_1\_f1. Per le restanti feature applicherò la procedure descritta di seguito. Prima di includere tutte le F\* preservate nello stesso modello possiamo svolgere un’analisi di correlazione tra la variabile target, le F\* e tra le stesse F\*. Nel caso due F\* risultino fortemente correlate tra loro e fortemente correlate con la variabile target, dobbiamo scegliere se considerarle ridondanti per il modello complessivo eliminando la feature con la feature importance f minore ottenendo così un modello più parsimonioso. Inoltre la collinearità dei predittori può causare difficoltà in alcuni modelli. In realtà i modelli basati su alberi decisionali sono sostanzialmente robusti/immuni rispetto alla collinearità dei predittori (specie gbm ed xgboost, nella RF perde attendibilità la feature importance).
5. Procedura di selezione delle feature per il modello complessivo.   
   Per ciascun comune j considero la feature F\_j a più alto potere predittivo f (tra quelle non rimosse) e costruisco un primo modello predittivo xgboost regression con queste le N feature F\_1, F\_2, ..., F\_j. Ad esempio, nel caso di due comuni (si veda punto7) il modello avrà le feature predittive F\_1\_f1 ed F\_2\_f1. Anche in questo caso procedo ad eliminare le feature con poter predittivo inferiore ad una cera soglia (e.g. coefficiente relativo < 0.3), ad esempio supponiamo che 1\_f1 = 0.6 e 2\_f1= 0.4, allora entrembe le feature sono idonee ed entrambe sopravvivono (se invece fosse 1\_f1 = 0.71 e 2\_f1= 0.29, solo la feature F\_1\_f1 sopravvivrebbe) .  
   Proseguo costruendo un secondo modello predittivo che avrà come feature F\_1\_f2 ed F\_1\_f2 e poi un terzo modello che avrà come feature predittive F\_1\_f3 per il comune 1 ed F\_2\_f3, F\_2\_f4 per il comune 2. La logica di partire con le feature a più alto potere predittivo porta ad aggregare le feature a più basso valore predittivo (in questo esempio nello stesso modello finiscono F\_1\_f3, F\_2\_f3, F\_2\_f4 e non solo F\_1\_f3 ed F\_2\_f3). Questa logica nel caso di numerosi comuni potrebbe portare ad una aggregazione eccessiva. Esempio: supponiamo che vi siano 10 comuni (j = 1, ..., 10) a copertura dell’area dell’acquifero, assumiamo che (in mancanza di informazioni specifiche sull’acquifero) che l’unione di questi 10 comuni approssimi l’area di ricarica principale dell’acquifero. Ora supponiamo che i 10 comuni abbiano un numero di feature idonee sopravvissute così distribuite:  
   comune 1: F\_1\_f1, F\_1\_f2  
   comune 2: F\_2\_f1, F\_2\_f2, F\_2\_f3  
   ...  
   comune 10: F\_10\_f1, ..., F\_10\_f10, F\_10\_f11  
   Se applicassimo lo schema dell’esempio precedente avremmo i seguenti modelli:  
   modello complessivo parte 1, M\_1: F\_1\_f1, F\_2\_f1, F\_3\_f1, ..., F\_10\_f1  
   modello complessivo parte 2, M\_2: F\_1\_f2, F\_2\_f2, F\_3\_f2, ..., F\_10\_f2  
   modello complessivo parte 3, M\_3: il terzo modello comprenderà tutte le rimanenti feature perchè il comune 1 aveva solo 2 feature.  
   Questo deve essere evitato, si devono costruire le parti del modello complessivo considerando solo i singoli modelli che non hanno esaurito le feature idonee. Per esemplificare, la procedura da seguire sarà la seguente:  
   modello complessivo parte 1, M\_p1: F\_1\_f1, F\_2\_f1, F\_3\_f1, ..., F\_10\_f1  
   modello complessivo parte 2, M\_p2: F\_1\_f2, F\_2\_f2, F\_3\_f2, ..., F\_10\_f2   
   modello complessivo parte 3, M\_p3: F\_2\_f3, F\_3\_f3, ..., F\_10\_f3 (escludo il modello del comune 1 che ha esaurito le feature)  
   modello complessivo parte 4, M\_p4: F\_3\_f4, F\_4\_f4, ..., F\_10\_f4 (escludo anche il modello del comune 2 che ha esaurito le feature)  
     
   ...  
   modello 10, M\_4: F\_9\_f10, F\_10\_f10, F\_10\_f11

La costruzione del modello complessivo può risultare laboriosa specie per acquiferi con molte variabili target e molti predittori, si dovrebbe quindi partire con un modello base (acquifero omogeneo) oppure testare una soluzione basata su xgboost regression o LSTM, CNN, o altra soluzione basata sulle reti neurali.

Nel caso si voglia sviluppare un modello idrogeologico, si dovrebbe preferire su un modello già esistente (sviluppato da ricercatori specializzati sull’argomento) per il quale è possibile ottenere (o stimare) i dati minimi necessari alla calibrazione del modello. Sarebbe preferibile partire con un modello esistente semplice che richieda un quantitativo di dati minimo.

Una volta determinati i dati caratteristici del territorio associato all’acquifero oggetto di studio, si possono sfruttare i dati predisposti per il modello idrogeologico anche per un modello predittivo (xgboost o modello basato su reti neurali). Nel caso di reti neurali, è noto che una qualsiasi funzione continua è approssimabile (con errore arbitrariamente piccolo) da una feedforward neural network a tre strati. Quindi se le feature sono rilevanti (e dovrebbero esserlo essendo incluse in modello idrogeologico) con un sufficiente quantitiativo di dati, è lecito aspettarsi che una opportuna rete neurale sia in grado approssimare la (incognita) funzione continua che mappa i dati in input (feature del modello idrogeologico) sul valore in output reale (e.g. variazione del livello della falda) o generato dal modello idrogeologico.

Una rassegna completa di modelli e codici R sfruttati in ambito idrogeologico si possono trovare ai seguentli link:

<https://cran.r-project.org/web/views/Hydrology.html>

<https://abouthydrology.blogspot.com/2012/08/r-resources-for-hydrologists.html>