Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN

Ingegneria civile, chimica, ambientale e dei materiali

Ciclo 31

Settore Concorsuale: 08/A1

Settore Scientifico Disciplinare: ICAR/02

TITOLO TESI

MACHINE LEARNING PER L'IDROLOGIA: APPLICAZIONE DEL METODO RANDOM FORESTS PER LA PREVISIONE DEGLI EVENTI DI PIENA FLUVIALE CON UN APPROCCIO PROBABILISTICO

Presentata da: Paolo Leoni

Coordinatore Dottorato
Prof. Luca Vittuari

Supervisore
Prof. Alberto Montanari

Co-supervisore
Ing. Silvano Pecora

Esame finale anno 2019

SOMMARIO

CONTESTO DI RICERCA

Negli ultimi anni le osservazioni idrologiche e i modelli di previsione idrologica hanno avuto uno sviluppo considerevole, anche grazie al rapido incremento di nuove tecniche e tecnologie informatiche.

Il passaggio dalle reti di monitoraggio manuali a quelle automatiche in tempo reale ha permesso di raccogliere una quantità maggiore di dati in tempi estremamente brevi. L'aumento delle informazioni ha sicuramente migliorato la conoscenza del territorio, con una particolare attenzione dedicata ai processi che si verificano durante le precipitazioni più importanti ed il conseguente sviluppo degli eventi di piena.

In funzione degli obiettivi da raggiungere, sono disponibili diversi tipi di modelli idraulici e idrologici. La scelta della tipologia del modello è legata agli obiettivi dell'analisi: un modello semplice richiede tempi di sviluppo e di calcolo particolarmente brevi, con una minor precisione delle informazioni prodotte (*output*); al contrario, un modello più complesso, ad esempio un modello fisicamente basato, richiede tempi di sviluppo e di calcolo più lunghi, a vantaggio però della precisione delle informazioni prodotte.

Un modello idrologico per le previsioni di piena, sviluppato per scopi operativi, deve rispondere a tre requisiti principali: tempi di calcolo rapidi, informazioni previsionali il più affidabili possibile e robustezza a fronte di errori nei dati e di calcolo.

Negli ultimi anni è sensibilmente incrementata la necessità di avere a disposizione le previsioni degli eventi di piena in tempo reale per i principali corsi d'acqua; necessità motivata soprattutto dalla crescente numerosità degli eventi di piena di tipo "*flash flood*", ad oggi la più difficile, in termini previsionali, e pericolosa tipologia di piena.

I recenti eventi di piena in Emilia-Romagna (ad esempio, in Italia, Fiume Parma 2014, Fiume Trebbia 2015, ecc.) sono gli *input* principali dell'attività di ricerca presentata in questo elaborato.

Con l'intento di elaborare un modello di previsione degli eventi di piena che massimizzasse l'utilizzo dei dati osservati, in questo lavoro è stato inizialmente sviluppato un modello speditivo basato sul *Modello NASP – NAsh SPeditivo*, (Biondi & Versace, 2007) già in operatività nei Centri Funzionali, che utilizza solamente i dati raccolti dalla densa rete di monitoraggio. Per poter migliorare l'efficienza del modello è necessario considerare le condizioni iniziali del bacino, pertanto è stato necessario sviluppare anche un modulo di rifiuto basato sul metodo del *Curve Number* del *SCS* (Soil Conservation Service Engineering Division, 1972) che identifica le sole piogge efficaci come contributo del deflusso superficiale. Il modello speditivo sviluppato è utile per produrre rapidamente valide informazioni sui livelli idrometrici, consultabili durante gli eventi di precipitazione più intensa.

Nella fase successiva, è stata adottata una delle tecniche più efficaci di Machine Learning: il metodo Random Forests. Il modello di previsione, basato sulla tecnica del *Random Forests*, è stato sviluppato e applicato ai principali fiumi dell'Emilia-Romagna; i risultati ottenuti sono stati molto incoraggianti e sono stati confrontati con quelli ottenuti applicando la tecnica "*BMBP – Bayesian Multivariate Binary Predictor*" (Todini, 2008).

Successivamente, la tecnica del *Random Forests* è stata estesa per affrontare un secondo caso studio: lo sviluppo di un modello di previsione dei livelli idrometrici e delle portate per le sezioni principali del Fiume Po, che valuti anche la stima dell'incertezza predittiva.

In questo secondo caso, nel data set del modello convergono due diversi tipi di informazioni: i dati osservati di livello idrometrico o i valori di portata e quelli previsti dalla catena di modellistica previsionale *MIKE11 NAM/HD* già sviluppata e disponibile nel Sistema *FEWS* del bacino del Po.

STRUTTURA DELLA TESI

Il lavoro di tesi sintetizza l'attività di ricerca sviluppata presso l'Area Idrologia-Idrografia dell'Arpae Emilia-Romagna, sede di Parma.

L'attività dà seguito ad un percorso di conoscenza e studio degli eventi di piena e delle loro tipologie iniziato durante il periodo universitario e la tesi di laurea magistrale; a cui si aggiungono le tecniche di apprendimento automatico e l'approfondimento di una delle tecniche principali: il metodo *Random Forests*.

Il lavoro qui presentato è suddiviso in 4 capitoli, come indicato di seguito.

CAPITOLO 1

È un'introduzione all'opera di tesi, in cui sono state riportate le definizioni più importanti e le informazioni necessarie, al fine di comprendere le principali questioni che verranno trattate in seguito. Le tematiche principali trattate sono: le diverse tipologie degli eventi di piena, i diversi tipi di modelli idraulico-idrologici per la previsione di eventi di piena ed il sistema di monitoraggio e previsione FEWS sviluppato per supportare i centri di Protezione Civile.

CAPITOLO 2

Presentazione delle tecniche di intelligenza artificiale ed evoluzione del *Machine Learning* (apprendimento automatico) con particolare riferimento alle principali tecniche sviluppate nel corso degli anni; focus dettagliato sulle tecniche basate sugli alberi decisionali, con particolare attenzione al metodo tra i più recenti e performanti: il *Random Forests*. Illustrazione di tutti gli elementi che caratterizzano il metodo *Random Forests*, con particolare attenzione alla gestione dei dati in ingresso e degli output prodotti.

Lo scopo di questo capitolo è quello di fornire le conoscenze necessarie per comprendere i due successivi capitoli, dove verranno presentati i modelli previsionali sviluppati, ed i motivi che hanno guidato determinate scelte.

CAPITOLO 3

Presentazione dell'attività di ricerca condotta per sviluppare un modello previsionale per gli eventi di piena dei principali corsi d'acqua dell'Emilia-Romagna. Gli argomenti vanno dalla risoluzione del problema indiretto all'implementazione del Modello *NASP* fino allo sviluppo del modello di previsione basato sulla tecnica del *Random Forests*.

Per presentare il modello basato sul metodo *Random Forests*, sono stati descritti tutti i passaggi e le relative scelte condotte, come, ad esempio, quella di associare un valore di probabilità ai risultati ottenuti. Al termine, è stato riportato un confronto tra i risultati di entrambi i modelli e quelli ottenuti applicando la tecnica *BMBP*.

CAPITOLO 4

Il secondo caso ha riguardato lo sviluppo di un modello basato sul metodo *Random Forests*, per la previsione degli eventi di piena nelle principali sezioni del Fiume Po. Inoltre, vengono riportate le principali innovazioni rispetto al primo caso di studio, che rendono il processo più rapido e performante e con una rapida ed importante capacità. Al termine, gli incoraggianti risultati ottenuti sono stati presentati sia graficamente sia numericamente e successivamente confrontati con quelli ottenuti applicando il processore *MCP*.

SVILUPPO DEL LAVORO DI TESI

Un Sistema di Allerta a Previsioni è solo uno dei tanti anelli della catena di Protezione Civile. Un ruolo molto importante è ricoperto anche dal personale tecnico, altamente qualificato, dotato di una spiccata preparazione e lunga esperienza nel settore; tutte queste caratteristiche risultano necessarie nella ordinarietà, mentre, diventano fondamentali nelle situazioni critiche.

Un modello previsionale, per quanto rigoroso sia, ha sempre dei limiti operativi in parte noti a priori in fase di elaborazione in parte emersi solo in fase di utilizzo; in tal caso, l'esperienza e la preparazione di un utente tecnico sono decisivi per utilizzare al meglio il modello.

Spesso è proprio da chi utilizza quotidianamente questi strumenti che emergono le criticità e quindi le richieste di miglioramento dei prodotti già esistenti o l'elaborazione di nuovi; l'attività di ricerca qui proposta ne è un concreto esempio.

L'evento di piena di tipo "*flash flood*" che il 13 ottobre 2014 ha interessato la città di Parma, con l'esondazione del Torrente Baganza (affluente del Torrente Parma) ha evidenziato una marcata criticità previsionale in tutte le catene modellistiche operative disponibili: nonostante i notevoli quantitativi pluviometrici previsti dalla modellistica meteorologica, gli accumuli di pioggia osservati a fine evento hanno superato sensibilmente quelli previsti; con conseguente sottostima del colmo di piena da parte delle catene modellistiche idrologiche-idrauliche (AIPO, 2014).

È emersa quindi la necessità di trarre il maggior vantaggio possibile dai dati osservati, in particolare, di pioggia e livello idrometrico; la ricerca di un'informazione rapida e continuamente aggiornata, fino ad intervalli di 15 minuti, è risultata necessaria.

L'attività di ricerca qui presentata ha cercato di soddisfare una richiesta operativa volta non solo al miglioramento del Sistema di Allerta e Previsione, ma anche alle attività di Protezione Civile dove in fase di intervento due-tre ore di anticipo possono risultare vitali.

Dopo una profonda ricerca bibliografica basata sulle più recenti tecniche di elaborazione dati e sviluppo di modelli previsionali speditivi, si è deciso di sviluppare un modello afflussi-deflussi basato sul Modello speditivo NASP – Nash Speditivo (Biondi & Versace, 2007), già in uso presso i Centri Funzionali, e sul metodo CN-SCS; inizialmente solo per il bacino del torrente Parma, successivamente è stato esteso ai principali corsi d'acqua dell'Emilia-Romagna.

Come verrà illustrato successivamente, il modello tiene conto delle principali variabili: precipitazione, stato di saturazione del terreno, tempi di corrivazione e risposta, etc. In funzione delle precipitazioni osservate simula la portata prevista nelle prossime ore. Al termine del processo di calibrazione, il modello ha fornito risultati incoraggianti ma non pienamente soddisfacenti; pertanto, si è deciso di sperimentare, nel campo dell'idrologica e dell'idraulica, una delle principali tecniche di apprendimento automatico supervisionato, già consolidata in molti altri settori: il *Random Forests*.

Le tecniche di Machine Learning (apprendimento automatico) inizialmente applicate soprattutto in campo medico, negli ultimi anni hanno preso sempre più piede anche negli altri settori: economico, gestionale, informatico, etc. L'applicazione di queste tecniche nel campo dell'idrologia e dell'idraulica è sicuramente una novità nel panorama nazionale.

È stato quindi sviluppato un modello che utilizza gli stessi *input* e fornisce gli stessi *output* del Modello di Nash; al termine, sono state confrontate le performance di entrambi i modelli con quelli ottenuti applicando la tecnica "*BMBP – Bayesian Multivariate Binary Predictor*" (Todini, 2008).

Nell'ultima parte dell'attività di ricerca, riferendoci ai più recenti ed avanzati studi sulla valutazione dell'incertezza predittiva (Coccia e Todini, 2011), utilizzando in input sia i dati osservati, sia della sezione di riferimento sia di quelle a monte, e i dati di previsione forniti dal modello idrologico-idraulico MIKE11 è stato sviluppato un modello basato sulla tecnica del Random Forests. Inoltre, il modello è stato strutturato affinché in automatico valuti l'importanza dei dati osservati in *input* così da escludere informazioni superflue che rallenterebbero il processo di calcolo; sebbene già particolarmente rapido, dell'ordine di 2-4 minuti.

I risultati ottenuti, da entrambi gli sviluppi, sono stati particolarmente incoraggianti; ma soprattutto, l'applicazione di queste nuove tecniche nel campo dell'idrologia e dell'idraulica ha permesso di iniziare un percorso, a nostro avviso, particolarmente favorevole.

ELABORAZIONE IN AMBIENTE R

Negli ultimi anni, soprattutto nell'ambito della ricerca scientifica, l'ambiente di sviluppo open-source R ha visto una crescita esponenziale sia in termini di sviluppo sia nel numero degli utilizzatori. L'ambiente R, nato principalmente per l'analisi statistica dei dati, è dotato di un proprio linguaggio di programmazione basato sui più classici linguaggi Fortran e C; inoltre, è stata implementata una versione più intuitiva denominata *Rstudio* che presenta un'interfaccia grafica pratica e intuitiva anche per un utente meno esperto nell'ambito della programmazione.

Fin dagli inizi, le potenzialità dell'ambiente R hanno favorito la rapida crescita di una community di utenti che nel tempo ha stimolato e favorito numerosi miglioramenti ed altrettanti potenziamenti rendendo oggi l'ambiente R uno dei più utilizzati in ambito scientifico.

In ambito idrologico, l'ambiente R si stava progressivamente affermando già agli inizi di questo lavoro di tesi (2015); negli ultimi due anni nuovi applicativi e nuovi sviluppi hanno incrementato sensibilmente l'utilizzo di R. Come riporta Louise J. Slater, R ha avuto una "rapida crescita in ambito idrologico", a conferma che la modalità open-source spesso risulta vincente: dare la possibilità a numerosi utenti di utilizzare un prodotto e di collaborare attivamente allo sviluppo dello stesso, favorisce rapide crescite ed importanti miglioramenti (Slater et al., 2019).

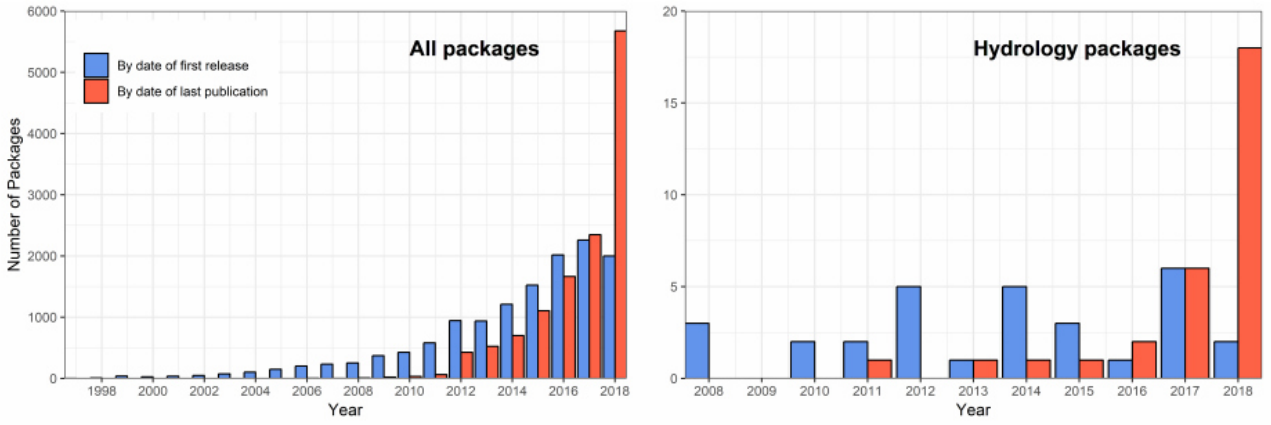


Figura 1 A sinistra, il numero di pacchetti disponibili su CRAN (1997-2018); a destra, quelli contenenti il termine "hydro" all'interno dei metadati del pacchetto (Louise J. Slater et al., 2019)

Inoltre, nell'ambito della ricerca si sentiva il bisogno di un ambiente di programmazione open-source e facilmente comprensibile anche dai meno esperti; la grande community sviluppatasi negli anni è una conferma oltre che una certezza per gli anni avvenire: uno dei principali obiettivi è proprio facilitare la diffusione dell'attività di ricerca con l'intento di creare reti di collaborazione virtuale.

Pertanto si è deciso fin da subito di utilizzare per l'attività di ricerca l'ambiente R, anche con un'ottica futura di condividere le elaborazioni e i risultati. In corso d'opera, al termine dell'elaborazione del primo modello è stato possibile implementarlo direttamente nel sistema FEWS; quest'ultimo è stato recentemente aggiornato dagli sviluppatori per favorire la compatibilità anche con i codici prodotti nell'ambiente R.

Tra i numerosi pacchetti disponibili vi è anche il pacchetto "*randomForest*" (Package 'random Forest', che è risultato essere uno degli elementi principali dei modelli sviluppati durante questa attività di ricerca; a cui

sono stati affiancati altri pacchetti e *tools*, relativi al *Random Forests*, per visualizzare i risultati e facilitarne la comprensione.

ABSTRACT

Lo sviluppo dei modelli di previsione degli eventi di piena richiede un gran numero di dati, una dettagliata conoscenza del territorio, dei processi fisici e delle loro interazioni; l'insieme di questi elementi definisce la complessità del modello.

Le informazioni necessarie per elaborare un modello previsionale, il più fedele possibile alla realtà, sono di difficile reperimento in molti casi pratici; inoltre, tanto maggiore è la mole delle informazioni da elaborare tanto maggiore sarà l'architettura informatica necessaria per processarle, con relativi costi di realizzazione e gestione non sempre supportabili.

In riferimento alle previsioni idrologiche ai fini di protezione civile, in caso di forti precipitazioni è necessario poter disporre di affidabili previsioni di livello idrometrico nel minor tempo possibile; soprattutto se il fine delle informazioni previsionali è il supporto di valutazioni tecniche e l'attività di protezione civile durante le emergenze.

In seguito alla grande piena del Fiume Po (2000) è stato sviluppato un sistema informatico denominato "*Flood Early Warning System (FEWS)*" che viene utilizzato e aggiornato periodicamente dall'Area Idrologia del Servizio Idro-Meteo-Clima di Arpae Emilia-Romagna (Italia) a supporto delle procedure e delle azioni di Protezione Civile.

Il Sistema si basa su avanzate catene modellistiche per la previsione degli eventi di piena dei principali corsi d'acqua del bacino del Fiume Po. Negli ultimi anni, gli eventi alluvionali, soprattutto di tipo "*Flash Flood*" sui corsi d'acqua cosiddetti "minori", sono sensibilmente aumentati. In particolare, molti degli ultimi eventi alluvionali sul suolo italiano sono stati caratterizzati da accumuli pluviometrici importanti e colmi di piena rapidi in sviluppo e consistenti in volumi.

I fenomeni di precipitazione più intensi sono di difficile previsione con alta risoluzione spaziale. Per superare questo limite previsionale, una valida soluzione è quella di utilizzare al meglio tutte le informazioni disponibili fornite dalla rete di monitoraggio.

In questo contesto, le tecniche di *Machine Learning* possono rivelarsi molto utili, massimizzando l'utilizzo delle informazioni in tempo reale. In particolare, il metodo *Random Forests* è tra le più recenti e performanti tecniche di Machine Learning.

Sfruttando le caratteristiche e le potenzialità del metodo *Random Forests* sono stati affrontati due casi di studio differenti che hanno portato all'elaborazione di due differenti modelli previsionali.

Il primo caso di studio si è concentrato sui principali fiumi della Regione Emilia-Romagna caratterizzati da tempi di risposta molto brevi. La scelta di questi fiumi non è stata casuale: negli ultimi anni si sono in detti bacini verificati diversi eventi di piena, in gran parte di tipo "*flash flood*".

Il secondo caso di studio riguarda le sezioni principali del Fiume Po, dove il tempo di propagazione dell'onda di piena è maggiore.

I dati in ingresso sono stati definiti in funzione degli obiettivi prefissati in entrambi i casi studio. Per il caso di studio (bacino del Fiume Po) ai dati osservati sono stati affiancati anche i dati di previsione provenienti dalla catena modellistica Mike11 NAM/HD.

Approfittando di una delle principali caratteristiche del metodo *Random Forests*, è stata stimata una probabilità di accadimento: questo aspetto è fondamentale sia nella fase tecnica che in fase decisionale per qualsiasi attività di intervento di protezione civile.

L'elaborazione dei dati e i dati sviluppati sono stati effettuati in ambiente R. Durante la fase di sviluppo, questo ambiente di programmazione ha consentito un'estrema versatilità; inoltre, il tempo di elaborazione si è rivelato estremamente ridotto.

Al termine della fase di validazione, gli incoraggianti risultati ottenuti hanno permesso di inserire il modello sviluppato nel primo caso studio all'interno dell'architettura operativa di FEWS.

1 EVENTI DI PIENA

1.1 INTRODUZIONE

In passato, i primi insediamenti urbani furono costruiti vicino ai corsi d'acqua in quanto necessitavano di disponibilità continua di risorsa idrica; da sempre considerata un bene vitale. L'acqua fu inizialmente utilizzata per fini di allevamento ed irrigui, ed infine, dalla prima rivoluzione industriale, anche per scopi produttivi.

Se la prossimità ai corsi d'acqua da un lato era fondamentale, dall'altro metteva però a serio rischio l'insediamento urbano: a seguito di eventi di precipitazione intensi il conseguente innalzamento del livello idrometrico causava inondazioni.

I primi interventi "di protezione" dalle inondazioni consistevano nel costruire o ricostruire gli insediamenti in luoghi più distanti o magari rialzati; mentre, i primi interventi "strutturali", come argini o vasche di espansione, appartengono ad un'epoca più recente.

Per quanto le tecniche di realizzazione delle opere strutturali siano in continua evoluzione e garantiscano una sicurezza maggiore, il rischio residuale di un'inondazione rimane uno dei rischi ambientali più rilevanti per la società moderna.

La raccolta dati e l'analisi degli eventi di piena sono precedenti ai primi interventi strutturali a protezione del territorio. Si pensi, ad esempio, al Nilometro degli antichi Egizi: pozzi o scale realizzati per misurare l'altezza delle piene del fiume Nilo, sulla base delle quali veniva stimato il raccolto. I primi punti di monitoraggio visivo del corso d'acqua venivano collocati sufficientemente a monte dell'insediamento urbano così, in caso di innalzamento idrometrico, vi era un lasso di tempo, ritenuto sufficiente, per mettere in sicurezza i territori a valle e ridurre i danni provocati dall'inondazione.

Utilizzando le prime osservazioni idrologiche, sono state realizzate le prime opere strutturali; negli ultimi anni, il supporto informatico ha permesso di potenziare sia l'attività di monitoraggio sia gli interventi strutturali dando vita ad un solido sistema di prevenzione.

Inoltre, negli ultimi decenni, alla prevenzione è stata affiancata la previsione: un valore aggiunto che permette di conoscere in anticipo gli eventi di piena così da valutare eventuali interventi a difesa delle persone e del territorio.

Oggi, attraverso le più recenti ed avanzate tecniche di prevenzione e previsione, è possibile conoscere e difendersi con sufficiente anticipo dagli eventi di piena e dalle inondazioni. Purtroppo, nonostante la prevenzione, gli interventi e la previsione, non si ha ancora e non si avrà mai totale sicurezza; pertanto è fondamentale continuare a sviluppare nuove tecniche per ridurre al minimo i potenziali danni prodotti dagli effetti di una piena.

1.2 EVENTI DI PIENA: DEFINIZIONE E TIPOLOGIE

Gli eventi di piena (in inglese, *"floods"*) vengono descritti dal *World Meteorological Organization – WMO* (WMO, 2011) come:

- l'aumento, solitamente repentino e di breve durata, del livello dell'acqua in un corso d'acqua, fino al raggiungimento di un picco tale da rappresentare una situazione di pericolosità; da cui il livello diminuisce gradualmente con dinamiche ad evoluzione generalmente più lenta;
- una portata insolitamente elevata alla sezione di chiusura o in un tratto di un canale/corso d'acqua;
- l'allagamento di zone costiere con acqua di mare a seguito altezze d'onda eccezionali.

Mentre il possibile effetto di un evento di piena, ovvero, l'inondazione (in inglese, "*flooding*") viene definita come la fuoriuscita di un volume d'acqua, dai normali confini del corso d'acqua, con l'occupazione di aree circostanti.

Il *WMO* identifica diverse tipologie di eventi di piena ognuna delle quali presenta delle proprie caratteristiche; di conseguenza anche le eventuali inondazioni ad esse associate possono avere elementi distintivi e quindi presentare criticità differenti.

Sul nostro territorio nazionale le piene fluviali sono sicuramente quelle più diffuse. Sono generate da eventi di precipitazione continui con intensità moderata-forte, si possono sviluppare nell'arco di pochi minuti come diverse ore o giorni. In relazione all'evento di precipitazione, le piene fluviali possono esser caratterizzate da un singolo picco di piena oppure da più picchi.

Precipitazioni intense e persistenti (durata di qualche ora), spesso localizzate, possono dare origine a piene improvvise, i cosiddetti "*flash floods*". Generalmente si verificano in seguito a eventi temporaleschi particolarmente intensi e persistenti, frutto di sistemi convettivi autorigeneranti che stazionano per diverse ore (fino a 5-7 ore) sulla stessa area. Lo sviluppo di una piena improvvisa è condizionato anche dalle caratteristiche del terreno, come permeabilità e pendenza, ma anche dal grado di saturazione del terreno o di compattezza: un terreno saturo, in seguito a eventi di precipitazione precedenti, favorirà lo sviluppo di piene a rapida evoluzione, tipica condizione autunnale-invernale-primaverile; così come un terreno compatto, in seguito ad un lungo periodo siccitoso, condizione comune in estate. Tra tutti gli eventi di piena, la tipologia "*flash flood*" è la più imprevedibile e devastante: la previsione di un sistema convettivo autorigenerante è estremamente complessa ed affetta da una forte incertezza, rendendo spesso inaffidabile la previsione meteorologica. L'attività di ricerca qui presentata si è concentrata soprattutto sulla risoluzione di tale problematica; ovvero, cercare di estendere l'orizzonte temporale di previsione in ambito meteorologico per le piene improvvise.

Se il corso d'acqua incontra alla foce delle avverse condizioni di moto ondoso (mare molto mosso, agitato, etc.), il normale deflusso viene ostacolato e si creano le condizioni per una piena estuarina o da rigurgito. In prossimità della foce il livello idrometrico sale rapidamente e rimane elevato fino a quando non si riduce il deflusso del corso d'acqua o l'intensità del moto ondoso. Tali condizioni possono causare anche degli allagamenti nelle aree circostanti.

In seguito ad abbondanti nevicate durante i periodi invernali o all'inizio della stagione primaverile si possono verificare repentini rialzi termici associati a venti sostenuti e precipitazioni; ciò può comportare lo sviluppo di piene da scioglimento nivale. Tale fenomeno si verifica quando si ha una rapida fusione del manto nevoso associata, eventualmente, a precipitazioni in forma liquida. È il caso dell'Emilia-Romagna nel dicembre 2009:

copiose nevicate hanno interessato tutta la regione a metà del mese, successivamente, tra la seconda e la terza decade di dicembre, forti venti caldi con modeste precipitazioni associate, hanno favorito in breve tempo la fusione di tutto il manto nevoso presente. La combinazione di due eventi relativamente modesti ha scaturito piene severe nei principali corsi d'acqua regionali. (Leoni & Montanari, 2014)

1.3 OSSERVAZIONI IDROLOGICHE E APPROCCI PREVISIONALI

La misura delle principali variabili meteorologiche (temperatura, umidità, pressione, intensità e direzione del vento, pluviometria, etc.), attraverso un'adeguata strumentazione, è una procedura consolidata con incertezze relativamente limitate.

Se la misura puntuale della precipitazione presenta un errore trascurabile; il calcolo della media areale delle precipitazioni osservate è affetta da un errore sensibilmente maggiore. Così come la stima quantitativa delle precipitazioni attraverso la valutazione condotta dai radar meteorologici o misure indirette da satellite (Todini, 2011) . Vari tentativi sono stati effettuati per combinare assieme le misure e rendere più affidabile il prodotto finale, ma il livello di errore rimane ancora significativo (Mazzetti and Todini, 2010).

Il livello idrometrico di un corso d'acqua è sicuramente la variabile più semplice da misurare: installando un'asta idrometrica è possibile "leggere" in tempo reale l'altezza del battente idrico. Attraverso le moderne stazioni di monitoraggio ad ultrasuoni, l'altezza idrometrica viene rilevata direttamente dallo strumento installato in punti specifici, ad esempio sui ponti.

Ma la sola misura dell'altezza idrometrica per fini di studio, progettazione e previsione non è sufficiente; si tratta infatti di una misura "puntuale" che spesso subisce notevoli variazioni già a pochi metri a monte o a valle del punto di monitoraggio. Variazioni dovute alla conformazione della sezione e dell'alveo fluviale: pendenza, restringimento della sezione oppure allargamento della stessa, etc. Inoltre, lo stesso punto può essere caratterizzato da movimenti di trasporto e di deposito del materiale solido presente in alveo; quindi, il letto del fiume può alzarsi ed abbassarsi sensibilmente, obbligandoci così ad una continua verifica del fondo alveo.

Attraverso l'utilizzo di un'adeguata strumentazione è possibile misurare la velocità della corrente lungo una serie di verticali che coprono l'estensione della sezione monitorata; con dette misure, nota l'altezza idrometrica, è possibile calcolare il valore di portata idrica, definita quale volume di fluido che transita nell'unità di tempo attraverso la sezione fluviale trasversale. Questa variabile, tra tutte, è la più importante in quanto permette di avere un quadro preciso ed affidabile per la sezione di interesse. Effettuando questa operazione più volte, con diverse altezze idrometriche, è possibile ricavare la scala di deflusso, ovvero: il legame tra altezza idrometrica e portata transitata, caratteristico per la sezione in esame; che rimarrà immutato fino a quando la sezione non subirà importanti variazioni.

Durante gli eventi di piena più severi le sezioni spesso subiscono profonde alterazioni; ciò comporta che i precedenti riferimenti non sono più validi: incisione dell'alveo, oppure, deposito detritico possono comportare una forte variazione del livello idrometrico misurato a parità di portata. Così come, un allargamento della sezione, dovuto magari alla rottura parziale delle sponde a monte della sezione. In tal caso, è necessario ripetere il rilievo con diverse condizioni idrometriche per realizzare la nuova scala di deflusso; grazie alla quale sarà possibile "aggiornare" le precedenti osservazioni idrometriche, relative alla scala precedente, attraverso i valori di portata.

Il monitoraggio e l'archiviazione delle informazioni raccolte oltre ad essere stati i primi passi nel campo dell'idrologia, così come in molti altri ambiti scientifici, continuano ad essere fondamentali per le valutazioni, per i progetti, per l'elaborazione dei modelli e quindi per i sistemi di prevenzione, previsione ed allertamento.

Grazie ai primi dati raccolti, in passato, è stato possibile realizzare le prime previsioni di livello idrico e portata fluviale: su base empirica, basandosi su una sezione strumentata più a monte era possibile intuire, più a valle, l'incremento del livello idrometrico; con precisione e tempistiche che dipendevano dal corso d'acqua considerato.

Le osservazioni e le prime previsioni empiriche hanno contribuito ad incrementare la conoscenza in ambito idrologico, a tal punto da conoscere, comprendere e riprodurre i principali processi fisici. Il primo tra questi è sicuramente il processo di trasformazione degli afflussi (precipitazione) in deflussi (portata).

La riproduzione matematica di questo processo è alla base dei principali modelli di previsione realizzati, dal più semplice di tipo concettuale al più complesso modello fisicamente basato.

1.4 MODELLISTICA IDROLOGICA E MODELLI DI PREVISIONE

Le osservazioni disponibili, raccolte nel tempo, e la conoscenza dei principali fenomeni idrologici ne hanno permesso la riproduzione matematica creando così i primi modelli idrologici.

È opportuno ricordare che un modello è la rappresentazione semplificata di un sistema reale (Moradkhani e Sorooshian, 2008) e, per quanto completo, non rappresenterà mai fedelmente la realtà. Di conseguenza, qualunque modello è affetto da semplificazioni, errori ed incertezza. Concettualmente, il miglior modello è quello che fornisce dei risultati più simili alla realtà, utilizzando il minor numero di parametri e una ridotta complessità modellistica.

Dal più semplice al più complesso, gli elementi che costituiscono un modello sono: dati di input, dati di output e parametri. I dati di input sono il punto di partenza per la creazione del modello e rappresentano l'insieme delle osservazioni disponibili; i dati di output, invece, sono i prodotti del modello, l'informazione che si vuole ottenere dal modello completato.

I parametri sono invece coefficienti, costanti o variabili nel tempo, che vengono introdotti nelle equazioni del modello per conferire flessibilità. I parametri possono essere ricavati da osservazioni di campo, oppure, dedotti attraverso degli studi preliminari. Input, parametri ed output sono tra loro legati da una serie di funzioni che descrivono i vari "sotto-processi" che insieme compongono il processo principale; che verrà riprodotto dal modello una volta completato.

Per esempio, un modello afflussi-deflussi riproduce il processo di "trasformazione" della precipitazione, su una determinata superficie, in portata, calcolata in una data sezione fluviale. I dati in input sono le precipitazioni, mentre, gli output sono i valori di portata generati dal modello. I parametri, in questo caso, sono tutte le informazioni disponibili ed implementabili nel modello che descrivono le caratteristiche del bacino: copertura vegetale, proprietà del suolo (impermeabilità, compattezza, etc.), topografia del bacino, tasso di saturazione del bacino, livello della falda sotterranea, etc. Solo implementando nel modello tutte queste informazioni e le relazioni tra loro si può realizzare una corretta modellazione del processo di trasformazione afflussi-deflussi; il più vicino a quello reale.

1.5 TIPOLOGIA DI MODELLI

Il continuo sviluppo delle tecniche di modellazione dei processi idrologici ha dato origine nel tempo a diverse tipologie di modelli, suddivise in differenti categorie; ognuna delle quali ha proprie caratteristiche e finalità.

Lo sviluppo o l'applicazione di un modello sono strettamente legati alle informazioni disponibili, necessarie sia in fase di realizzativa sia in fase operativa, e alle finalità che si vogliono raggiungere.

I modelli idrologici vengono classificati in funzione di: scala spaziale, scala temporale, tipologia di output e complessità del modello (Moradkhani e Sorooshian, 2008).

SCALA SPAZIALE

I modelli che rappresentano esplicitamente l'eterogeneità dei processi idrologici vengono definiti spazialmente distribuiti; mentre, quelli che "condensano" il bacino idrografico vengono definiti concentrati.

I modelli concentrati furono realizzati per la prima volta agli inizi della seconda metà del '900 (Madsen, 2003); si tratta di modelli speditivi, ma non per questo poco attendibili, che nel riprodurre il processo afflussi-deflussi sintetizzano le caratteristiche del bacino in uno o più parametri. Trovano largo impiego nei casi in cui, per varie ragioni, non si hanno sufficienti informazioni sulle caratteristiche del bacino idrografico. Il quale viene rappresentato in questi modelli con un numero definito di serbatoi in cascata, aventi una data capacità di accumulo dell'acqua presente al suolo. Per simulare il processo di trasformazione afflussi-deflussi, vengono definite una serie di equazioni matematiche associate che descrivono il processo di passaggio dell'acqua tra un serbatoio e l'altro (Houghton-carr, 1999). L'insieme di equazioni che compongono questi modelli rappresentano, in maniera semplificata, le fasi al suolo del ciclo idrologico mentre i parametri che costituiscono il modello cercano di riprodurre i valori medi delle caratteristiche dell'intero bacino (Madsen, 2003). Il più noto di questa tipologia di modelli è senz'altro il modello del serbatoio lineare o Modello di Nash; a cui verrà dedicata un'approfondita descrizione nei capitoli seguenti.

SCALA TEMPORALE

Generalmente i modelli idrologici vengono realizzati per operare in continuo cercando di riprodurre le condizioni idrologiche sia durante gli eventi di pioggia sia durante i periodi asciutti ("*dry period*"); ciò comporta che il modello deve anche cercare di riprodurre tutte le condizioni transitorie del bacino che influiscono direttamente al deflusso. I modelli appartenenti a questa classe vengono anche definiti dinamici, in riferimento alla loro peculiare caratteristica di modellare in continuo le condizioni del bacino idrografico.

Particolari situazioni, specie quelle più critiche, possono richiedere la realizzazione di modelli specifici che raccolgano e processino quante più informazioni possibili; è il caso ad esempio di un evento di piena prodotto da intense precipitazioni oppure dal cedimento di una diga, o ancora, la diffusione di un inquinante accidentalmente finito nel corso d'acqua. Questi modelli, operativi solo per un breve periodo temporale in funzione della loro struttura e della loro richiesta, vengono definiti a scala d'evento.

TIPOLOGIA DI OUTPUT

La classificazione dei modelli avviene anche in funzione della tipologia di output prodotti. In un modello deterministico le variabili, sia in input sia in output, assumono tutte valore puntuale; pertanto il modello

fornisce una stima esatta delle variabili in output. Ad ogni medesimo ingresso corrisponde sempre la medesima uscita.

In un modello stocastico almeno una tra variabili di stato e variabili in uscita è rappresentata specificando la sua distribuzione di probabilità. In questi modelli, le variabili in uscita sono sempre identificate con un margine di approssimazione, che dovrebbe esser specificato meditante la distribuzione di probabilità associata. È possibile che i modelli stocastici si compongano anche di una parte deterministica, impiegata per riprodurre una stima del valore medio delle variabili in uscita, alla quale si aggiunte una componente stocastica che permette di quantificare in termini di distribuzione di probabilità l'oscillazione attorno al valor medio. In funzione della struttura del modello, ad ogni medesimo ingresso corrispondono differenti uscite.

STRUTTURA DEL MODELLO

Le classificazioni viste fino ad ora possono esser considerate come "sotto-tipologie" delle tre principali tipologie di modelli: empirici, concettuali e fisicamente basati.

I primi modelli realizzati erano di tipo empirico, ovvero, basati esclusivamente sui dati raccolti; questa tipologia di modelli non considera i processi fisici che si verificano ma si basa esclusivamente sui dati e sulle relazioni tra *input* ed *output*. Dipendendo strettamente dai dati d'ingresso questa tipologia di modelli è stata definita "*black box*". Nella loro "semplicità" hanno subito una rapida evoluzione: prima, con i metodi statistici di regressione e correlazione per identificare una relazione tra input e output; poi, con tecniche più avanzate e performanti come quelle di apprendimento automatico.

Se, invece, l'intento è descrivere i processi delle componenti idrologiche bisogna ricorrere ad una tipologia più complessa rispetto a quella precedente, ovvero, ai modelli concettuali; dove il bacino idrografico viene schematizzato attraverso una serie di serbatoi in cascata. Per ogni serbatoio vengono considerati i processi che generano la ricarica, precipitazioni, infiltrazione e percolazione, e quelli che invece contribuiscono a svuotare il serbatoio, evaporazione, deflusso superficiale e sotterraneo. Le equazioni che costituiscono questa tipologia di modelli sono di tipo semi-fisico ("*grey-box models*") mentre i parametri sono generalmente definiti sia dai dati di monitoraggio sia dai processi di calibrazione (Ljung, 2002). La calibrazione di questi modelli richiede un vasto set di dati e numerosi tentativi per la definizione dei parametri; pertanto vi sono diversi gradi di complessità, in funzione degli elementi considerati e degli obiettivi da raggiungere. Modelli semplici che necessitano solamente di una rapida calibrazione sono caratterizzati da pochi parametri (esempio, Modello di Nash, solo due parametri: k ed n); modelli particolarmente complessi come lo "Stanford Watershed Model IV (SWM)", sviluppato da Crawford and Linsley nel 1966 ed ancora oggi il principale modello concettuale, conta ben 20 parametri.

La tipologia più complessa è sicuramente quella dei modelli fisicamente basati ("*white-box models*") che matematicamente cercano di rappresentare i processi reali utilizzando le variabili di stato che sono misurabili sia nel tempo che nello spazio. I processi idrologici che caratterizzano l'afflusso ed il deflusso sono rappresentati da una serie di equazioni alle differenze finite. In questo caso non sono necessari data set molto grandi per la calibrazione ma è richiesta una profonda conoscenza dei processi fisici che si intende modellare e la valutazione di un gran numero di parametri che descrivano le caratteristiche fisiche del bacino. Le informazioni necessarie per parametrizzare questi modelli sono sensibilmente maggiori rispetto alle altre tipologie già viste (topografia, dimensioni e caratteristiche della rete idrografica, etc.); inoltre, il dettaglio delle informazioni è necessariamente più spinto. Le performance di un ottimo modello fisicamente basato possono superare quelle degli altri due modelli; ciò implica, però, che in fase di realizzazione sono stati considerati tutti i processi e gli elementi influenti ai fini dell'elaborazione.

1.6 SCELTA DEI MODELLI

Il primo elemento che condiziona la scelta del modello è sicuramente l'obiettivo da raggiungere, ovvero, il fine di utilizzo del modello. Lo sviluppo, quindi l'esistenza, di diverse tipologie di modelli è indice che le necessità di utilizzo dei modelli idrologici-idraulici sono particolarmente diversificate.

Se il fine ultimo è quello di sviluppare/utilizzare un modello che sia in grado di riprodurre il più fedelmente possibile i processi fisici che si verificano in un bacino idrografico, allora, la scelta ricadrà su un modello fisicamente basato. In tal caso bisognerà però conoscere nel dettaglio le caratteristiche del bacino oltre che i principali processi che lo regolano; ciò comporta una robusta conoscenza del territorio attraverso osservazioni e campagne di monitoraggio più o meno continue. Il processo di raccolta delle informazioni, necessarie per la realizzazione del modello, richiede spesso molto tempo (anni o addirittura decadi). La conoscenza maturata andrà successivamente ingegnerizzata, cioè riprodotta numericamente, e poi verificata attraverso un data set di calibrazione. Questa seconda fase richiede tempi di sviluppo importanti, sebbene inferiori a quelli necessari nella prima fase (raccolta delle informazioni). Non meno importante è il supporto informatico nelle varie fasi: progettazione, calibrazione ed operatività; anche se i tempi di calcolo negli ultimi anni si sono notevolmente ridotti.

Se l'obiettivo è invece un modello attendibile che rappresenti con buona approssimazione i principali processi presenti nel bacino idrografico di cui conosciamo solo gli elementi principali, allora, un modello concettuale risulterà sicuramente più adeguato. In tal caso, i tempi necessari per la progettazione e calibrazione sono sensibilmente inferiori rispetto ad un modello fisicamente basato; a ciò si aggiunge il vantaggio di una rapida elaborazione che permette l'utilizzo del modello anche in tempo reale in virtù dell'immediatezza della risposta.

Agli antipodi, per struttura, dei modelli fisicamente basati ci sono i modelli empirici "*data-driven*"; inizialmente molto utili in condizioni di scarse informazioni sulle caratteristiche del bacino e con la sola disponibilità delle osservazioni. Negli ultimi anni hanno avuto una spiccata "evoluzione" con il supporto delle reti neurali (ASCE Task Committee on application of Artificial Neural Networks in Hydrology, 2000; ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000) (Giustolisi, 2000), prima, e dell'intelligenza artificiale, dopo. A tal punto che con un robusto set di dati di monitoraggio è possibile ottenere informazioni attendibili in tempi relativamente molto ristretti; sia in fase di progettazione sia in fase operativa.

Con lo scopo di raggiungere l'obiettivo prefissato all'inizio di questo lavoro di ricerca, realizzare un modello attendibile e speditivo per le previsioni idrometriche, l'attenzione è stata rivolta alle due categorie di modelli più speditivi: concettuali ed empirici.

1.7 SISTEMA DI PREVISIONE ED ALLERTAMENTO

La previsione idrologica è una valutazione dello stato futuro di alcuni fenomeni idrologici, come il livello idrometrico, la portata, il volume totale, l'area inondata, la media delle velocità della corrente in una sezione o in un tratto fluviale, etc.

Inizialmente, le previsioni idrologiche erano un contributo per la gestione degli eventi di piena; il continuo sviluppo che ha portato ad un sensibile incremento della loro affidabilità, ha reso le previsioni idrologiche indispensabili non solo per fini operativi durante gli eventi più importanti ma anche per fini di pianificazione,

sia durante i periodi asciutti sia in previsione degli eventi più importanti (da 2-3 giorni a poche ore dall'evento).

Già negli anni '90, i modelli idrologici erano considerati "il cuore di qualsiasi sistema di previsione degli eventi di piena" (Serban, 1991); oggi, i modelli idrologici, insieme alle osservazioni, sono il cardine dei sistemi di gestione e previsione.

L'immagine seguente illustra i passaggi fondamentali per l'elaborazione, lo sviluppo e la verifica di un modello previsionale.

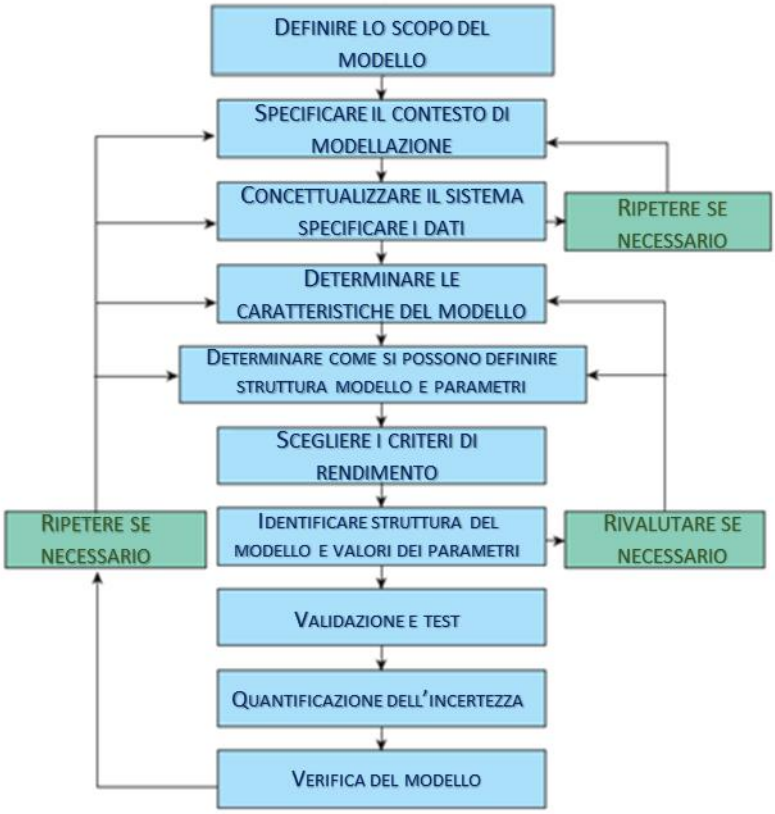


Figura 2: Schematizzazione del processo di sviluppo di un modello previsionale di piena.
World Meteorological Organization - WMO, 2011

Prima di procedere con la realizzazione di un modello è necessario definire le finalità; noto l'obiettivo si può procedere con la scelta e lo sviluppo del modello.

Nell'ambito dei modelli previsionali, un parametro di valutazione particolarmente importante da considerare è il tempo di preannuncio della previsione ("*forecast lead-time",* in inglese), ovvero, il tempo trascorso tra l'emissione della previsione ed il suo verificarsi. Il tempo di preannuncio è legato a diversi fattori: orografici, computazionali, finalità previsionali, etc. Per un bacino di dimensioni ridotte il tempo di preavviso sarà altrettanto ridotto; mentre, per grandi bacini il tempo di preavviso può raggiungere anche 7-10 giorni.

Il seguente grafico mostra le varie tempistiche previsionali: rispetto all'asse delle ascisse, nella parte superiore sono riportate le tempistiche meteorologiche; mentre, nella parte inferiore sono state schematizzate le tempistiche idrologiche.

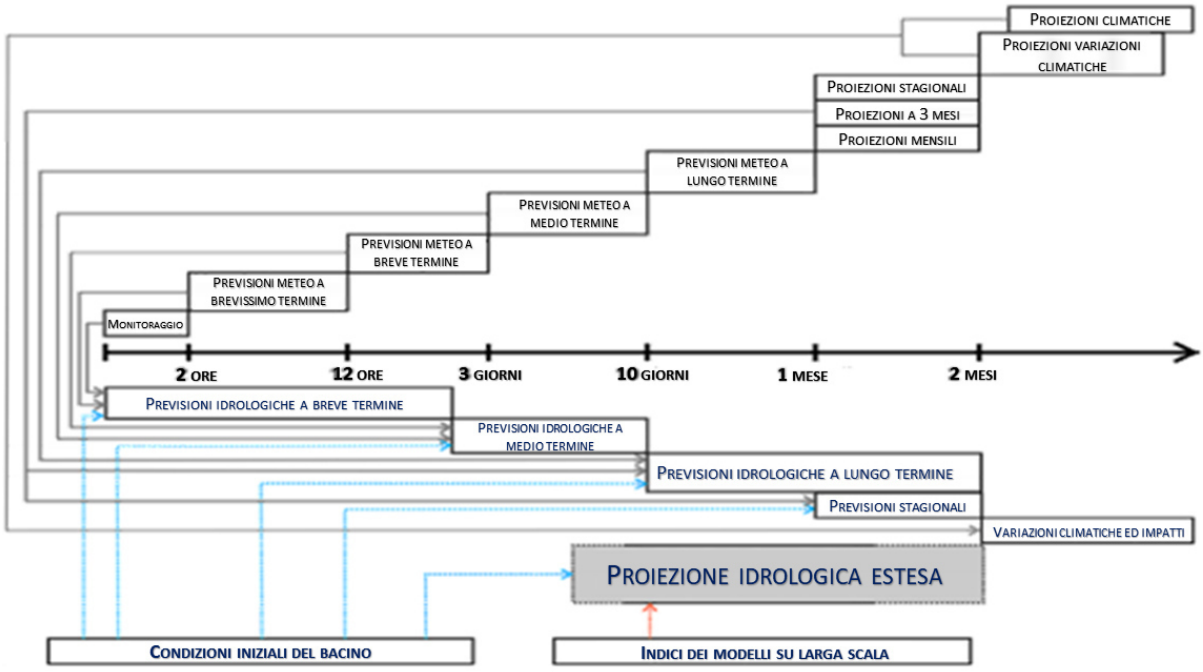


Figura 3 Rappresentazione degli intervalli previsionali in meteorologia (sopra l'ascissa) ed idrologia (sotto l'ascissa). World Meteorological Organization - WMO. (2006).

Generalmente, a livello mondiale, si definisce previsione a breve termine quella con un tempo di preannuncio di 48-72 ore; mentre, la previsione a medio termine raggiunge anche i 10 giorni (WMO, 2006).

In Italia, ad esempio, i tempi di preannuncio sono più brevi; le previsioni di piena per i fiumi appenninici hanno un tempo di preannuncio compreso in generale tra le 6 e le 12-18 ore. Le previsioni di piena con tempi di preannuncio inferiori alle tre ore, vengono identificate nel monitoraggio (*nowcasting*, in inglese*)*; ovvero, il peso delle informazioni raccolte dalla rete osservativa è dominante rispetto alle informazioni prodotte dai modelli di previsione meteorologica.

Le osservazioni raccolte dalla rete di monitoraggio sono gli input del modello di previsione idrologica-idraulica; queste sono affette da un tasso d'incertezza, che può essere ridotto ma non annullato. Il modello di previsione idrologica-idraulica è anch'esso affetto da un tasso di incertezza; pertanto per elaborare un modello di previsione è necessario considerare il problema dell'incertezza previsionale derivante dall'errore nei dati, dall'inadeguatezza del modello e dalla stima non ottimale dei parametri.

Tenendo anche conto dell'incertezza presente nei modelli previsionali, non esiste un unico modello sempre valido in ogni occasione ed in ogni ambito; quindi, prima di effettuare la scelta del modello, è necessario ricorrere ad un'attenta valutazione degli obiettivi da raggiungere, del tempo di calcolo, della conoscenza disponibile dei fenomeni idrologici e del territorio in esame.

Nel 1991 O'Connell suggerì di utilizzare un approccio modulare ai modelli concettuali, in cui il processo principale (formazione del deflusso) è composto da tanti sotto-processi, che possono esser considerati dei sotto-modelli del modello principale (O'Connell, 1991).

Il modello previsionale scelto, per essere utile, deve soddisfare gli obiettivi richiesti dalle parti interessate e dagli utenti finali delle previsioni; la complessità del modello deve esser coerente con l'effettiva "capacità di

portare informazioni" (Klemeš, 1988) in funzione dei dati disponibili per calibrare e successivamente fornire le informazioni previsionali (Goswami e O'Connor, 2007). Un modello particolarmente dettagliato che richiede numerose informazioni in ingresso e tempi di elaborazione molto lunghi potrà risultare anche preciso, ma se il tempo necessario per produrre l'informazione supera il tempo di preannuncio dell'evento, il modello sarà pressoché inutile.

Per fini previsionali previsione, considerando le numerose tipologie di modelli disponibili e considerando la marcata eterogeneità degli eventi pluvio-idro, è necessario realizzare un Sistema di Allerta e previsione degli eventi di piena, denominato "*Flood Forecasting Warning Systems – FFWS*"; contenente una serie di modelli o suite di modelli differenti tra loro, in funzione delle esigenze operative.

1.7.1 FEWS Po – Flood Early Warning System

Per realizzare un efficiente sistema di allertamento e previsione è necessario disporre innanzitutto di una robusta rete di monitoraggio in tempo reale delle principali variabili meteorologiche ed idrologiche; pertanto, sul bacino del fiume Po, anche in seguito alla "grande piena del 2000", è stata sviluppata una rete capillare di monitoraggio idro-meteorologico.

I dati raccolti in tempo reale sono il primo anello della catena previsionale, in quanto oltre a fornire un quadro istantaneo delle condizioni idro-meteo, servono ad alimentare il complesso sistema di modellistica previsionale sviluppato nel corso degli anni. I prodotti dei modelli di previsione e delle catene modellistiche previsionali svolgono un ruolo importante, che nei casi più critici diventa fondamentale, per la valutazione della situazione attuale, per la formulazione delle previsioni a breve e medio termine, per le attività di protezione civile, etc.

In concomitanza con la realizzazione della rete di monitoraggio in tempo reale è stato realizzato dall'area idrologia e idrografia dell'ARPAE Emilia-Romagna un sistema di allerta e monitoraggio per il bacino del fiume Po che prende il nome di "*Flood Early Warning System FEWS-Po*".

Il sistema consente la modellazione e la previsione delle piene del fiume Po e dei suoi principali affluenti; inoltre è integrato con i sistemi di previsione dei Centri Funzionali regionali di Protezione Civile.

FEWS-Po è una piattaforma informatica in grado di gestire le esecuzioni di diverse catene idrologico-idrauliche alimentate da campi di precipitazione forniti dalla rete di monitoraggio in tempo reale e da modelli meteorologici. La piattaforma permette l'acquisizione dei dati in tempo reale, la validazione dei parametri e la gestione-esecuzione dei modelli numerici relativi a tutti i processi fisici legati alle inondazioni dei fiumi. Al temine della fase di elaborazione, *FEWS-Po* consente la visualizzazione e la condivisione dei risultati; sia numerici sia grafici.

I principali risultati prodotti sono gli idrogrammi di piena riferiti alle sezioni idrometriche di interesse; ma anche rappresentazioni grafiche più avanzate, utili per facilitare la loro interpretazione. L'interfaccia del sistema è basata su interfaccia GIS e permette di visualizzare in tempo reale i dati georeferenziati derivanti da sensori in telemetria (stazioni meteorologiche ed idrometriche). L'informazione osservata puntualmente viene elaborata ed estesa spazialmente così da poter avere un quadro d'insieme più preciso e dettagliato oltre che confrontabile con le informazioni previsionali: mappe di precipitazione, temperatura, neve al suolo, etc.

Le informazioni che costituiscono l'input al sistema FEWS-Po consistono in:

- Precipitazioni osservate dalla rete di telerilevamento
- Precipitazioni previste fornite dai modelli di previsione meteorologica
- Corsa[1] operativa determinista del modello meteorologico, con un ciclo di assimilazione rapida, COSMO RUC (risoluzione 2,2 km)
- Corsa operativa deterministica del modello meteorologico COSMO-I2 (risoluzione 2,2 km)
- Corsa operativa deterministica del modello meteorologico COSMO-I5 (risoluzione 5 km)
- Corsa di insieme del sistema COSMO-LEPS (risoluzione di 7 km)
- Esecuzione determinista del modello ECMWF del Centro Europeo di Reading (risoluzione 16 km)

Utilizzando gli input sopra elencati e le catene modellistiche implementate, il sistema FEWS-Po fornisce in output una serie di previsioni generate dalla combinazione di tre diverse catene di modellistica previsionale, composte dai modelli idrologici HEC-HMS (Bennett, 1988), MIKE11-NAM (Nielsen e Hansen, 2018) e TOPKAPI (Todini e Ciarapica, 2002). Oltre ai modelli idrologici, sono stati implementati tre modelli idrodinamici per simulare l'andamento delle portate nel letto del fiume HEC-RAS (Chen, 1973), DHI-M11 e Deltares-Sobek (Deltares, 2014).

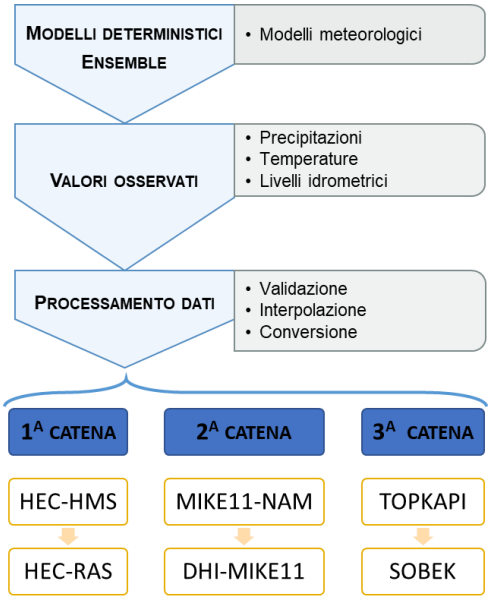


Figura 4 Struttura Sistema FEWS

[1] Corsa ("run" in inglese) è un termine tecnico utilizzato per identificare i risultati delle elaborazioni dei modelli

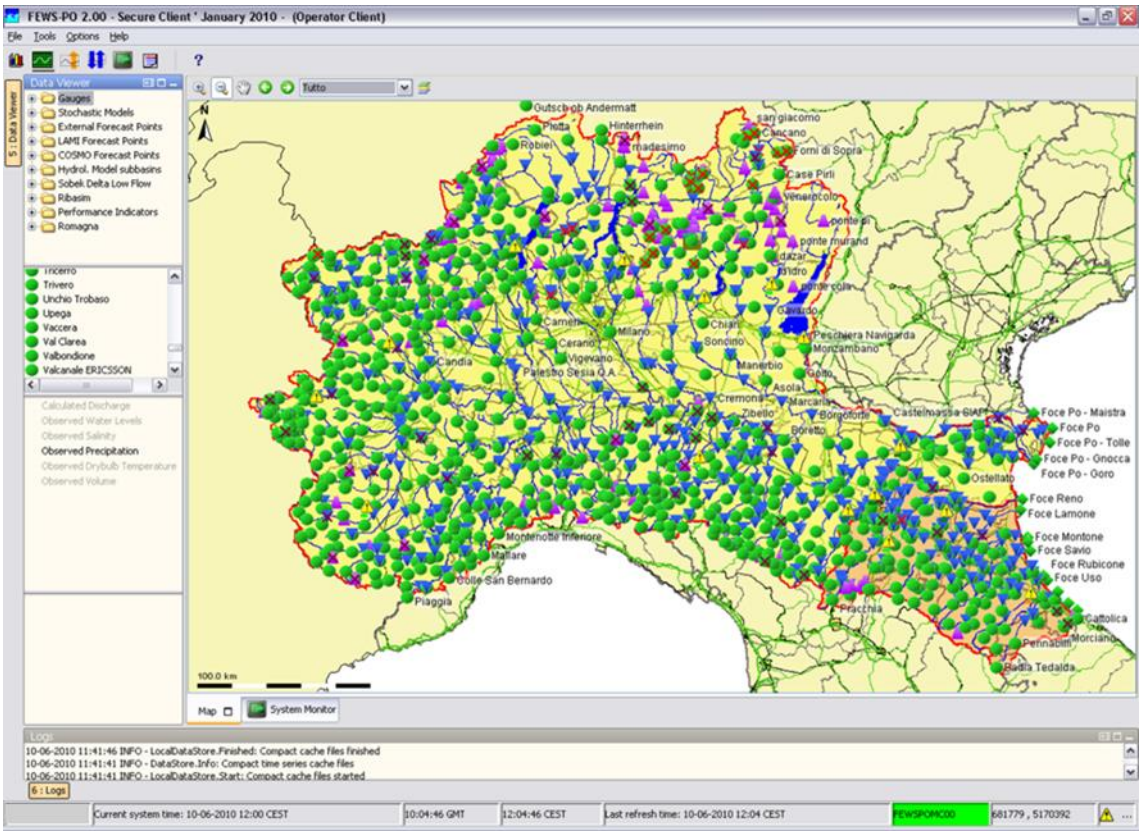


Figura 5 Interfaccia grafica del Sistema FEWS-Po

Il Sistema *FEWS-Po* per il Fiume Po, i suoi affluenti principali, il Fiume Reno e i corsi d'acqua della Romagna, elabora informazioni previsionali con una frequenza di 3-24 ore; nel dettaglio:

- 1 previsione determinista idrologica a brevissimo termine (+12/18 ore)
- 1 previsione deterministica idrologica a breve termine (+36/40 ore)
- 1 previsioni idrologiche deterministica a medio termine (+60/64 ore)
- 63 previsioni idrologiche a lungo termine per ensemble (+108/112 ore)

Nel medio-lungo periodo, l'incertezza delle previsioni di livello idrometrico, principalmente condizionata dall'incertezza delle previsioni meteorologiche, cresce con l'aumento della finestra temporale previsionale e la diminuzione della superficie del bacino. Per i bacini minori, come quelli dell'Emilia-Romagna, caratterizzati da tempi di risposta di poche ore, l'incertezza locale delle previsioni meteo influenza in modo significativo le previsioni di livello idrometrico, anche nel breve periodo.

In questi casi, i dati pluviometrici ed idrometrici acquisiti dalla rete in telemisura ad alta frequenza diventano fondamentali per valutare lo stato dell'evento, per studiare le condizioni del bacino e per elaborare possibili previsioni e/o interventi.

Il Sistema *FEWS-Po* è operativo, ai fini previsionali, 365 giorni all'anno, 24 ore al giorno. Un'altra caratteristica presente nel Sistema, molto importante per un sistema di allerta e previsione, è la possibilità di configurarlo in "stand alone"; condizione necessaria per la pianificazione, la simulazione in tempo differito o la ricostruzione di eventi. (Brian, et al., 2018).

1.8 LIMITI PREVISIONALI E CRITICITÀ RISCONTRATE

Qualsiasi modello previsionale, indipendentemente dalla propria tipologia e finalità, deve essere periodicamente aggiornato; l'aggiornamento motivato dal nuovo set di dati in ingresso o dalle nuove informazioni disponibili, generalmente, comporta una nuova calibrazione del modello e quindi un nuovo processo di validazione. Modelli più semplici, così come in fase di realizzazione, permettono calibrazioni più veloci e frequenti; modelli più complessi (ad esempio i fisicamente basati) richiedono tempi molto lunghi per la fase di calibrazione e validazione.

Per garantire nel tempo l'efficacia e l'efficienza del modello è necessario tener conto dei processi di aggiornamento fin dalla fase di elaborazione del modello.

Gli aggiornamenti periodici interessano non solo i modelli ma anche un sistema di allertamento e previsione, che periodicamente viene aggiornato con:

- nuovi dati di monitoraggio,
- aggiornamento dei modelli o delle catene modellistiche
- implementazione di nuovi modelli o nuove catene modellistiche
- implementazione nuove tecniche previsionali

Il Sistema FEWS-Po, grazie alla sua struttura versatile e modulabile, consente numerosi e frequenti aggiornamenti e, se necessario, implementazioni che periodicamente vengono condotti per soddisfare le finalità per cui è stato realizzato.

Gli aggiornamenti, sia dei modelli sia di un Sistema di allerta e previsione, sono nella maggior parte dei casi programmati ma talvolta rispondono anche a delle esigenze operative; ad esempio, un evento meteo-idro particolare (per tipologia, intensità, durata, etc.) non presente nei precedenti data set, oppure, la disponibilità di nuove importanti informazioni utili ad affinare le performance raggiunte precedentemente.

Questa attività di ricerca è nata proprio da una necessità operativa: negli ultimi anni si è registrato un sensibile incremento degli eventi di piena di tipo "*flash flood*"; per loro natura tanto imprevedibili quanto potenzialmente pericolosi. L'imprevedibilità meteorologica che caratterizza gli eventi di precipitazione che generano le piene improvvise ("*flash flood*") si ripercuote direttamente sulla previsione idrologica-idraulica. È bene ricordare che piccole variazioni chilometriche sono più che tollerabili per un modello meteorologico; ciò può comportare, però, marcate differenze di effetti al suolo: basti pensare all'evento di piena, una differenza di pochi chilometri, tra osservato e previsto, dei fenomeni di precipitazione più intensi si traduce in accumuli pluviometrici più severi per un dato bacino piuttosto che per uno attiguo, con conseguente sviluppo di piena.

Se ad oggi le previsioni meteo non riescono a superare questo limite, preziose informazioni ci giungono dai dati osservati; quindi, per massimizzare le informazioni in tempo reale si è cercato di identificare un modello che utilizzasse nel miglior modo possibile i dati osservati, fornendo una previsione rapida ed attendibile.

2 ARTIFICIAL INTELLIGENCE, MACHINE LEARNING E RANDOM FORESTS

La tecnica di Intelligenza Artificiale su cui è basato questo lavoro di ricerca è al momento attuale oggetto di attività di ricerca intensiva in ambito internazionale. In questo capitolo, viene presentata una revisione della letteratura sull'Intelligenza Artificiale: vengono prima riportati i principali metodi di apprendimento automatico e poi, nel dettaglio, il metodo *Random Forests* e le modalità attraverso le quali sono stati sviluppati i modelli dei due casi studio qui presentati.

2.1 INTRODUZIONE

"*L'Intelligenza Artificiale (Artificial Intelligence A.I., in inglese) è una disciplina appartenente all'informatica che studia le basi teoriche, la metodologia e le tecniche che consentono la progettazione di sistemi hardware e sistemi di programmi software in grado di fornire al computer prestazioni che, ad un osservatore comune, sembrerebbero appartenere esclusivamente all'intelligenza umana*" (Somalvico, 1997).

Pertanto, un risultato fornito da un "computer intelligente" può integrare un giudizio di un "umano intelligente". Uno dei primi esempi cardine dell'intelligenza artificiale (A.I.) è rappresentato dal "*Turing Test*", in cui viene mostrato come un computer intelligente fornisce risposte non distinguibili da quelle che darebbe un altro essere umano, se interrogato da un terzo essere umano (Turing, 1950).

I primi *bot*, o *robot*, dotati di Intelligenza Artificiale erano in grado di riprodurre le azioni su cui erano stati opportunamente addestrati, senza elaborazioni cognitive delle informazioni disponibili; successivamente, attraverso le tecniche di Machine Learning i bot sono stati programmati per apprendere ad eseguire azioni frutto sia dell'addestramento imposto sia della conoscenza acquisita in fase di addestramento. Per la prima volta, sono stati creati dei bot aventi la facoltà di elaborare cognitivamente le informazioni disponibili e capaci di fornire una "risposta intelligente" piuttosto che una "risposta semplice".

Uno degli ultimi obiettivi dell'Intelligenza Artificiale riguarda direttamente la logica e il "pensiero razionale", sfruttando a meglio le ben note regole dell'inferenza. Il bot programmato con questa strategia di apprendimento, basandosi sui dati disponibili, è in grado di generare, elaborare e riportare tutte le informazioni e fornire risultati corretti.

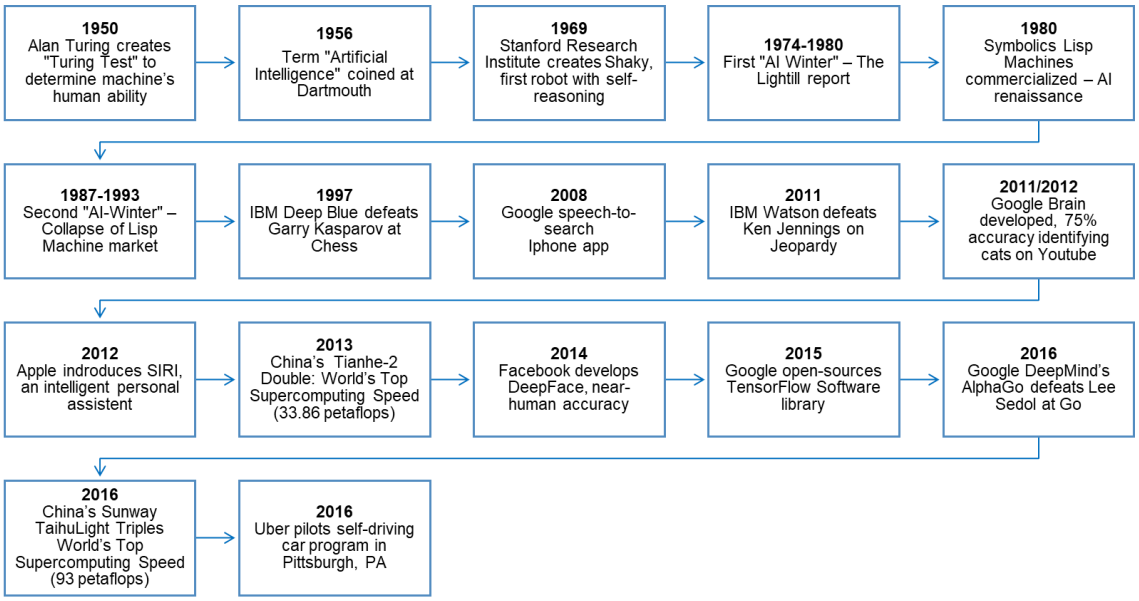


Figura 6 Evoluzione dell'Intelligenza Artificiale

2.2 MACHINE LEARNING

L'apprendimento automatico (*Machine Learning*, in inglese) è una dei settori più importanti dell'Intelligenza Artificiale (I.A.). Una macchina è considerata "intelligente" quando è in grado di apprendere dalla propria esperienza. Sulla base delle proprie esperienze storiche, la macchina è in grado di "pensare", "valutare" e quindi "scegliere" al fine di ottenere i migliori risultati; in sostanza, essa "guadagna esperienza" esattamente come gli umani. Tanto maggiore sarà l'esperienza acquisita, tanto maggiore sarà la capacità di fornire risultati più accurati.

La capacità di una macchina (*bot*) dotata di intelligenza artificiale può essere così illustrata: "*si dice che un programma per computer apprende dall'esperienza E in relazione ad alcune classi di compiti T ed alla misura di prestazione P, se le sue prestazioni nei compiti T, misurate da P, migliorano con l'esperienza E*" (Mitchell, 1997). Quindi, per realizzare un apprendimento automatico, è necessario definire l'esperienza E, il compito T e la misura di prestazione P.

Il bot per potersi migliorare ha la necessità apprendere/imparare attraverso dei feedback che valutino i risultati delle sue azioni. In riferimento alle metodologie di apprendimento, esistono tre categorie principali:

1 Apprendimento supervisionato: con feedback diretto
2 Apprendimento non supervisionato: nessun feedback esterno
3 Apprendimento per rinforzo: con feedback indiretto

2.2.1 APPRENDIMENTO SUPERVISIONATO

Lo scopo dell'apprendimento supervisionato è estrapolare una funzione da un set di dati denominato "set di addestramento" ("*training set*", in inglese). Questo set di dati è composto da valori in ingresso (*input*) e in uscita (*output*); generalmente, l'input è composto da due o più variabili; mentre, l'output è una sola variabile. L'algoritmo di apprendimento supervisionato esamina il "*training set*" fornitogli, con l'obiettivo di definire

una funzione in grado di attribuire l'output corretto a nuove istanze (nuovi dati), non presenti nel set di addestramento.

Nella risoluzione di un problema di apprendimento supervisionato è necessario considerare due parametri importanti dell'algoritmo di apprendimento utilizzato: la stabilità e l'adattabilità. Dati due diversi set di addestramento, la stabilità è la propensione dell'algoritmo a generare la stessa funzione o una funzione molto simile; mentre, l'adattabilità è la propensione dell'algoritmo a produrre diverse funzioni a partire da diversi set di dati di addestramento. Un algoritmo dotato di un'elevata stabilità genererà errori sistematici con probabilità maggiore; così come un algoritmo dotato di elevata adattabilità sarà condizionato sensibilmente dal data set di addestramento utilizzato.

Esistono due modalità di apprendimento supervisionato: classificazione e regressione. Nella classificazione, i dati sono raggruppati in categorie ben definite (ad esempio la classificazione della popolazione per anni, sesso, etc.); mentre, nella regressione i dati sono considerati singolarmente (valori numerici). Sia in fase di addestramento sia in fase operativa, un algoritmo strutturato sulla classificazione colloca le istanze in ingresso in ambienti (categorie) ben definite e distinte; a differenza di un algoritmo strutturato sulla regressione che colloca le istanze in ingresso in un ambiente libero.

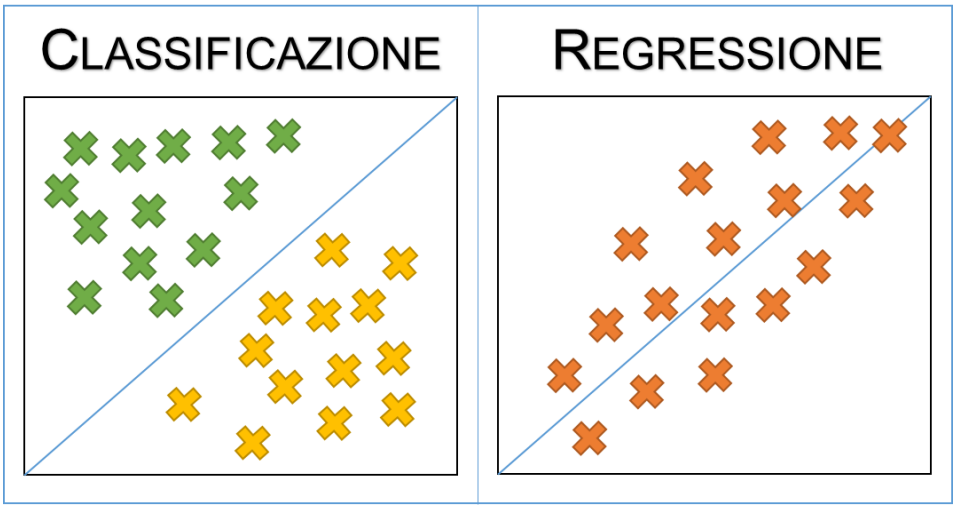


Figura 7 Rappresentazione schematica dalla tecnica di classificazione e regressione

La scelta della modalità di apprendimento supervisionato è guidata sia dal set di dati utilizzato sia dai risultati che si intendono ottenere: se l'intento è classificare una specie animale o vegetale, caratterizzata da connotati ben definiti, si utilizzerà un apprendimento basato sulla classificazione; mentre, se l'obiettivo è stimare delle variabili continue si utilizzerà un algoritmo basato sulla regressione.

2.2.2 APPRENDIMENTO NON SUPERVISIONATO

Questa tipologia di apprendimento è applicata quando il set di dati per l'apprendimento è composto dalle sole informazioni in ingresso (input); ovvero, sono assenti le informazioni in uscita (output). Per poter definire una funzione l'algoritmo ha la necessità di avere anche le informazioni in uscita; pertanto, in questo caso vengono utilizzate diverse tecniche. Una delle tecniche è l'analisi dei sotto-gruppi, detti anche *cluster*, del training set: consiste nel raggruppare dati simili in sottoinsiemi (*cluster*) con lo scopo di definire delle categorie. Il processo si arresta quando viene raggiunto un numero prefissato di cluster.

Un'altra tecnica è quella di ricercare in prima istanza delle regole di associazione per identificare la dipendenza tra le variabili disponibili; tale metodologia è molto diffusa nelle tecniche di "*data mining*" (estrazione dei dati).

2.2.3 APPRENDIMENTO PER RINFORZO

Questo tipo di apprendimento viene applicato quando non ci sono indicazioni precise sull'output desiderato. L'output è interpretato attraverso una ricompensa (rinforzo), che valuta quantitativamente la bontà della scelta fatta. La differenza con l'apprendimento supervisionato è nella verifica dell'output: nel primo caso l'output è ben noto e definito; mentre nell'apprendimento per rinforzo c'è solo una "valutazione" positiva o negativa dell'output.

Sono state sviluppate due tipologie di apprendimento per rinforzo, in funzione delle modalità di miglioramento delle performance dell'algoritmo: modalità passiva e modalità attiva. Considerando l'apprendimento mediante il rinforzo passivo, la strategia di apprendimento rimane invariata e l'obiettivo è quello di apprendere l'utilità dei vari stati e quindi la decisione migliore. Nell'apprendimento mediante rinforzo attivo, durante la fase esplorativa vengono fatte nuove scelte per valutarne l'efficacia. Per questi motivi, la strategia deve essere modificata in funzione delle informazioni ottenute. In primo luogo, viene preferita una fase di esplorazione iniziale e successivamente viene identificata una fase di definizione e una strategia di miglioramento. L'apprendimento di rinforzo viene applicato quando il set di dati di input e output non è molto coerente e/o affidabile, oppure, nel caso in cui viene identificata una strategia a lungo termine per raggiungere un obiettivo.

2.3 ALBERI DECISIONALI

Una delle tecniche di apprendimento più recenti si basa sugli alberi decisionali (*decision trees*, in inglese) generati da algoritmi di apprendimento. Questa tipologia di apprendimento è un valido strumento per guidare delle scelte che dipendono da diverse variabili; inoltre, permette la rappresentazione grafica degli alberi generati, permettendo così l'accesso a queste informazioni anche agli utenti meno esperti.

La struttura di un albero decisionale ricorda quella di un diagramma rovesciato, composto da:

- Nodo radice: è il primo nodo della struttura avente in uscita due rami, uno per ogni scelta che può esser condotta in funzione di una delle variabili coinvolte oppure al valore di una combinazione di più variabili.
- Nodi interni (nodi di test): collocati all'interno della struttura, si compongono di rami in ingresso e in uscita; ciascun nodo genitore è diviso in due nodi figlio.
- Foglie: caratterizzate solo da rami in entrata e collocate nella parte inferiore, rappresentano la parte terminale della struttura.

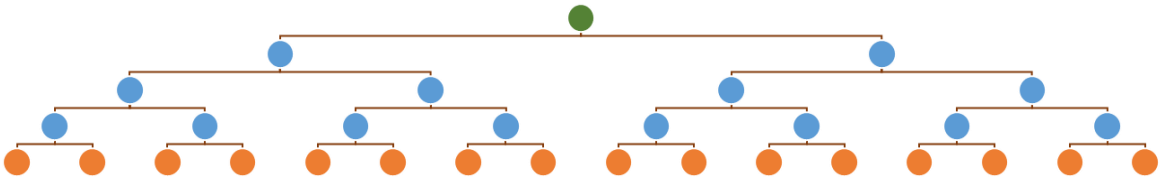


Figura 8 Struttura schematizzata di un albero decisionale, composto da: nodo radice (verde), nodi interni (azzurri), foglie (arancioni) e rami (linee marroni)

I dati di ingresso (*input*) di un albero decisionale possono essere un oggetto o un'istanza, caratterizzati da uno o più attributi; i dati di uscita (*output*) sono i valori di una variabile definita a priori o una classe.

Ad ogni ramo presente nello schema, sia quelli uscenti dal nodo radice sia quelli entrati ed uscenti dai nodi interni, corrisponde un test su uno o più attributi; i rami uscenti dai nodi rappresentano i possibili valori dell'attributo testato.

La scelta finale per tutte le istanze che vengono elaborate dall'algoritmo decisionale viene rappresentata dalle foglie.

La tecnica degli alberi decisionali rientra nella categoria dell'apprendimento supervisionato; pertanto si possono avere due tipologie di alberi: alberi di classificazione e alberi di regressione. L'operatività è la medesima già descritta nell'apprendimento supervisionato, gli alberi di classificazione (*classification trees*) vengono impiegati quando l'output, in virtù delle caratteristiche del dataset utilizzato, può assumere solo valori discreti predeterminati (o classi); mentre, gli alberi di regressione (*regression trees*) sono utilizzati quando l'output assume valori continui. La scelta di una tipologia rispetto all'altra è strettamente legata alla tipologia dei dati ed all'obiettivo da raggiungere. La parte cognitiva del problema è fondamentale: un'informazione povera e/o errata potrebbe condurre ad un errore di valutazione grave, favorito dalla creazione di alberi decisionali sbagliati.

Di seguito verranno illustrati i principali passaggi per l'elaborazione di un albero decisionale e le relative problematiche riscontrabili in fase di realizzazione; l'approccio e le possibili soluzioni verranno illustrate in seguito in riferimento ai casi studio affrontati.

2.3.1 Apprendimento (Learning)

La tecnica di apprendimento automatico attraverso l'albero decisionale rientra nella tipologia di apprendimento supervisionato; l'elaborazione di un albero decisionale, attraverso un algoritmo specifico, si basa sull'uso di uno o più data set consistenti, composti da informazioni in ingresso (input) ed informazioni in uscita (output). Essendo fortemente condizionati dai dati disponibili, queste tecniche di apprendimento sono strettamente "*data driven*".

Il primo set fornito all'algoritmo per la definizione dell'albero decisionale è il set di addestramento (*training set*), sulla base del quale l'algoritmo classifica le istanze presenti e genera l'albero decisionale. Questa è la fase più importante di tutto il processo di apprendimento automatico; la scelta attenta di un corretto "*training set*" è fondamentale per evitare impatti negativi sul processo di creazione dell'albero decisionale.

Successivamente, l'albero decisionale prodotto verrà testato con un nuovo set di dati, denominato "*test set*"; le istanze del test set sono prive della classe o variabile di output, in quanto questa verrà elaborata dall'algoritmo (output algoritmo) e successivamente confrontata con la classe o variabile reale per valutare

la bontà dell'albero decisionale prodotto. La qualità dell'albero decisionale dipende dalla capacità di classificare correttamente le istanze presenti del set di test.

In funzione dell'algoritmo utilizzato, con lo stesso data set, è possibile ottenere due o più alberi decisionali differenti tra loro; ciò è legato alla capacità dell'algoritmo di valutare i valori delle variabili presenti all'interno del training set. A parità di performance, è preferibile seguire il principio del "*rasoio di Occam*" ("*Occam's razor*"), ovvero, scegliere l'albero con la dimensione minore. La dimensione di un albero decisionale viene valutata in termini di: numero di nodi, numero di livelli (o profondità) e numero di attributi utilizzati; un albero con dimensioni ridotte ha una maggior capacità di generalizzazione, risulta meno dispendioso in termini di elaborazione e soprattutto permette una consultazione più rapida da parte dell'utente finale.

Sempre in riferimento alle dimensioni, un albero decisionale caratterizzato da un percorso diverso per ogni istanza, ovvero, da molti nodi e rami (condizione limite) sarà inefficace in presenza di nuove istanze differenti da quelle presenti nel set di addestramento. Al contrario, un albero decisionale caratterizzato da un numero estremamente ridotto di percorsi, generato però da un elevato numero di istanze tra loro differenti, fornirà risultati troppo generici e molto probabilmente non aggiungerà nulla di nuovo a ciò che è già noto.

L'obiettivo finale non è classificare correttamente le istanze presenti nel training set, ma fare in modo che l'albero decisionale classifichi correttamente le nuove istanze, (diverse da quelle presenti nel training set). Solo così l'albero decisionale tornerà utile, fornendo informazioni aggiuntive a quelle iniziali.

2.3.2 SET DI DATI PER L'ADDESTRAMENTO (*TRAINING SET*)

Il set di dati di addestramento è per l'algoritmo la fonte da cui trarre esperienza per addestrare l'albero decisionale. Generalmente un training set è una matrice di dati dove le righe rappresentano le istanze, ovvero le informazioni relative ad una specifica situazione; mentre, le colonne rappresentano gli attributi di riferimento. Nell'ultima colonna, aggiunta per raggiungere lo scopo (l'apprendimento), viene riportata la corretta informazione d'uscita (output), nota come classe; in questa fase utilizzata tra i valori di input per addestrare l'albero decisionale.

I set di dati sono caratterizzati da tre tipologie di attributi:

- Numerico: valore numerico discreto o continuo.
- Ordinale: non hanno valore numerico, ma possono essere ordinati secondo una logica precisa (come i giudizi: eccellente, buono, sufficiente, insufficiente); inoltre, non sono caratterizzati da una nozione di distanza.
- Nominale: non hanno valore numerico e non possono essere ordinati.

Istanza	Attributo A (numerico)	Attributo B (ordinale)	Attributo C (nominale)	Attributo n-esimo	Classe
…	…	…	…	…	…
101	1321	Rosso	Quadrato	…	A
102	1354	Giallo	Cerchio	…	D
103	4689	Blu	Rettangolo	…	F
…	…	…	…	…	…

Tabella 1 Esempio della struttura di un data set necessario per l'apprendimento, prima, e l'operatività, dopo, dell'albero decisionale.

In base al tipo di attributi, è necessario valutare la scelta dell'algoritmo e le sue operazioni.

La creazione di un albero decisionale con buone capacità di classificazione richiede un set completo di allenamento: le istanze devono rappresentare tutte le categorie di input presenti ed ogni attributo deve avere tanti valori quante sono le istanze del training set. Inoltre, è opportuno che tutte le classi siano incluse nel data set di apprendimento.

La condizione principale per raggiungere un valido risultato è che il set di allenamento sia coerente: per ogni classe assegnata, devono esserci istanze uguali con gli stessi valori; adottando questa strategia a medesime condizioni iniziali corrispondono le medesime condizioni finali, si evita quindi di "confondere" l'algoritmo. Per facilitare la fase di elaborazione, è utile evitare le istanze ridondanti: ripetere più volte quanto "è già stato visto" aumenterebbe solo i tempi di elaborazione senza apportare benefici in termini di risultati.

2.3.3 SCELTA DEGLI ATTRIBUTI

L'identificazione degli attributi dominanti in un processo permette di raggiungere il più rapidamente possibile la classificazione del maggior numero di istanze. La "classificazione ideale" si ha quando viene identificato "l'attributo ideale", ovvero, quell'attributo che da solo può classificare tutti le istanze; mentre, il caso opposto si verifica quando uno o più attributi lasciano invariate le proporzioni tra le classi in fase di classificazione. Pertanto, in fase di elaborazione è opportuno selezionare l'attributo, o gli attributi, che più si avvicinano alle condizioni dell'"attributo ideale".

In riferimento all'idea di entropia espressa da Shannon (Shannon, 1948), la quantità di informazioni ottenibili da un dato è in funzione della precedente disponibilità di informazioni attribuite a quel dato; ovvero, possiamo trarre informazioni utili da un attributo solo se questo attributo è stato osservato e valutato in passato.

Nelle tecniche di apprendimento basate sugli alberi decisionali, l'informazione ottenuta è la classe associata ad una data istanza. La conoscenza a priori è il numero di istanze appartenenti ad una particolare classe; il metodo più rapido e semplice per quantificare questa conoscenza all'interno del training set è calcolare il rapporto tra il numero di istanze che appartengono ad una determinata classe ed il numero totale di istanze presenti: un rapporto molto basso indica che quella classe è poco conosciuta all'interno del training set; mentre, un rapporto molto alto è indice di un'elevata conoscenza di quella classe. Come verrà mostrato nei successivi paragrafi, è importante comunicare all'utente finale le informazioni relative al training set: numero di dati, periodo di osservazione, numero di classi, istanze per ogni classe, dati mancanti, etc.

Un data set di addestramento incompleto, con uno o più attributi mancanti in varie istanze, influisce negativamente sull'elaborazione dell'albero decisionale. In questo caso, è consigliabile eliminare l'intera istanza in cui almeno un attributo è mancante.

2.3.4 RUMORE (*NOISE*)

La presenza di valori errati all'interno del set di addestramento rappresenta il "rumore" (*noise*, in inglese) del data set; questo può generare coppie di istanze uguali negli attributi ma con classificazione diversa. La presenza di un "rumore" molto forte porta l'algoritmo a generare un albero errato o "confuso"; che classificherà tutte o buona parte delle nuove istanze in modo errato. Per evitare di ricadere in questa situazione è fondamentale ridurre al minimo, o meglio eliminare il "rumore" all'interno del training set anche

a costo di escludere le istanze: una mancata classificazione per carenza di informazioni è un problema noto a priori, attraverso l'anagrafica del training set, che si può affrontare; l'errata classificazione, invece, risulta invisibile poiché annegata nei risultati finali.

L'eliminazione del "rumore" può essere condotto direttamente dall'utente esperto in fase preliminare, se il training set non è molto grande e/o gli errori sono noti ed in numero ridotto; mentre, per training set pesanti o per errori molto diffusi questa operazione risulta essere molto più difficile. Si dovrà ricorrere a dei "metodi di potatura" ("*pruning criteria*", in inglese) applicabili a priori o durante la fase di elaborazione dell'albero decisionale. Una delle tecniche più diffuse è la scelta per maggioranza: data una foglia (nodo terminale) in cui vengono classificate due istanze uguali ma con classi diverse, si predilige classificare l'istanza che ha la classe più presente tra tutte le istanze che hanno raggiunto quella foglia; rigettando l'altra istanza.

2.3.5 SOVRADATTAMENTO, SOTTOADATTAMENTO E POTATURA

La presenza degli attributi in numero maggiore del necessario in un training set è un'altra condizione particolarmente sfavorevole per la corretta elaborazione di un albero decisionale: l'algoritmo potrebbe elaborare un albero che ha definito relazioni tra attributi e classi inesistenti nella realtà. In tal caso si avrà una cattiva classificazione dei futuri set di dati, in quanto l'albero valuterà come importanti determinati attributi e relazioni tra essi; attributi e relazioni che in realtà sono inefficaci e quindi non utili per elaborare correttamente la classificazione.

Per risolvere il problema, molto comune, del sovradattamento (*overfitting*) esistono diverse tecniche alcune manuali altre automatiche; quest'ultime risultano essere particolarmente utili quando il training set è molto grande oppure quando le informazioni presenti non permettono all'utente esperto di identificare rapidamente i casi di *overfitting*.

È sempre consigliabile supervisionare il processo di "potatura" (*pruning*, in inglese) del training set, così da evitare che vengano eliminate preziose istanze e soprattutto di ricadere nella situazione opposta, ovvero, il sub-adattamento (*underfitting*). Ciò comporterebbe l'elaborazione di un albero decisionale troppo piccolo non all'altezza del processo di classificazione desiderato.

La potatura (*pruning*) è il metodo più diffuso ed efficace per risolvere il problema del sovradattamento che consiste nel ridurre il numero di istanze ridondanti nel training set; al termine del processo il risultato sarà una riduzione delle dimensioni dell'albero decisionale.

Esistono due tipologie di potatura in relazione al momento in cui viene effettuata questa operazione:

- Pre-potatura (*pre-pruning*): prima della creazione dell'albero, che potrebbe risultare troppo grande da generare e/o da gestire.

- Post-potatura (*post-pruning*): successiva alla creazione dell'albero; utile per eliminare le parti errate o non necessarie ai fini decisionali. Questa tecnica può essere eseguita procedendo in due direzioni:
  - dall'alto verso il basso ("top-down"), partendo dalla radice e seguendo i rami dell'albero fino all'identificazione dei sotto-alberi (nodi interni) da eliminare;
  - dal basso verso l'alto ("bottom-up"), partendo dalle foglie fino a raggiungere il punto (nodo intero) in cui è conveniente potare l'albero.

Indipendentemente dalla tecnica di potatura utilizzata, il risultato sarà sempre un peggioramento della capacità di classificare le istanze del set di addestramento; in quanto si sottraggono informazioni all'algoritmo decisionale, sebbene le restanti siano le più importanti. Tuttavia, l'obiettivo finale del processo di *pruning* è migliorare l'accuratezza nella classificazione di nuove istanze; pertanto, un peggioramento delle performance di classificazione in fase di addestramento non implica anche un peggioramento delle performance di classificazione in fase operativa. Anzi, l'obiettivo è quello di ridurre le istanze superflue dando maggior importanza alle istanze più rilevanti; soprattutto se appartenenti a particolari classi. Più semplicemente, il processo di *pruning* può servire per evidenziare determinate classi rispetto ad altre.

I vari metodi disponibili per eseguire il processo di potatura hanno caratteristiche e risultati differenti, di seguito vengono brevemente presentate le principali:

- Criterio di arresto: metodo pre-potatura dall'alto verso il basso; l'espansione dell'albero viene interrotta dall'algoritmo quando alcuni criteri vengono soddisfatti a priori (raggiungimento della profondità massima consentita dell'albero, il numero di classi nei nodi è inferiore al valore minimo consentito, etc.). Metodo molto semplice e rapido; ma con risultati spesso non soddisfacenti.

- Riduzione dell'errore: metodo di post-potatura dal basso verso l'alto. Si articola in quattro semplici passaggi: partendo dalla foglia il processo si arresta al primo nodo interno incontrato (primo nodo dal basso), il nodo viene sostituito da una foglia contenente la classe avente più istanze in quel nodo. Per verificare l'efficienza di questa sostituzione, viene effettuata la verifica dell'errore di classificazione sull'intero training set: se l'errore è stato ridotto, la modifica viene mantenuta ed il processo prosegue sulla foglia successiva; se l'errore è aumentato, la modifica viene annullata e il processo prosegue sulla foglia successiva.
- Potatura "*cost-complexity*": metodo di post-pruning dal basso verso l'alto; il processo viene avviato ad albero completo, gradualmente i nodi interni più bassi vengono sostituiti con le foglie, a cui vengono attribuite le classi con maggior istanze. Il processo termina quando viene raggiunto il nodo radice e definite tutte le classi.

- Potatura "pessimistica": metodo di pre-potatura dall'alto verso il basso; il processo confronta l'albero ottenuto con dei sotto-alberi generati dal medesimo training set. Attraverso delle verifiche di classificazione vengono scartati i sotto-alberi con il tasso d'errore più elevato. L'albero scelto sarà quello con la percentuale di errore di classificazione più bassa.

- Chi-quadro: potatura di tipo "bottom-up" in entrambe le modalità, dall'alto verso il basso o dal basso verso l'alto. La differenza tra la classe e ciascun attributo è valutata calcolando la distribuzione del chi-quadro; l'integrazione o la sottrazione di alcuni attributi definisce le prestazioni dell'albero decisionale.

2.3.6 VALIDAZIONE INCROCIATA (*CROSS-VALIDATION*)

Il processo di validazione è condotto a valle del processo di calibrazione: consiste nell'escludere una porzione di dati dall'intero data set che non verranno utilizzati in calibrazione (prima fase) ma in validazione (seconda fase); in sostanza, la validazione è la prima verifica delle performance del modello, sulla base della quale il modello viene migliorato o confermato per la fase operativa (terza fase).

Nel caso specifico degli alberi decisionali, la tecnica di "cross-validation" (validazione incrociata) viene applicata per verificare la capacità dell'albergo generato di prevedere eventi non conosciuti in fase di apprendimento. In relazione ai risultati ottenuti in questa fase, l'albero decisionale potrà essere confermato oppure rielaborato modificando i parametri di calibrazione o integrando tutto o parte del data set di validazione per rafforzare la fase di apprendimento. L'importanza e la necessità di utilizzo parziale o totale del data set a disposizione verrà discusso nella parte finale di questo elaborato.

La condizione più favorevole si ha in presenza di un vasto ed eterogeneo data set da cui è possibile creare i due sotto gruppi di dati: training set per l'addestramento e *validation set* per la validazione.

Per realizzare dei sotto gruppi del data set iniziale, una tecnica molto comune è denominata "*k-fold validation*": consiste nella suddivisione del set di addestramento in sottoinsiemi equidimensionali; per ogni sottoinsieme verrà generato un albero decisionale. La scelta dell'albero, da utilizzare in operatività, sarà determinata sulla base delle migliori prestazioni. Questa tecnica risulta vantaggiosa per data set molto grandi e omogenei tra loro.

Nei casi studio sviluppati in quest'attività di ricerca, è stata adottata un'altra tecnica di validazione, denominata "*leave-one-out cross-validation*" che consiste nell'eseguire numerosi test, estraendo ogni volta un'istanza dal training set; la scelta dell'albero ricadrà su quello avente la miglior performance. Per applicare questa metodologia è necessario disporre di un dataset sufficientemente grande e privo di istanze "eccezionali"; in tal caso, è opportuno che queste rimangano sempre all'interno del dataset.

2.3.7 ENSEMBLE LEARNING

Una delle tecniche più avanzate di apprendimento automatico è l'apprendimento d'insieme o "*Ensemble Learning*"; questa tecnica consiste nel creare numerosi alberi decisionali utilizzando il medesimo training set. Gli alberi generati sono combinati per ottenere un classificatore che dovrebbe risultare preciso almeno quanto il miglior albero generato. L'obiettivo di questa tecnica non è migliorare le performance decisionali, quanto ridurre la probabilità di classificare erroneamente una nuova istanza.

Per facilitare la comprensione di questa avanzata metodologia di classificazione, si può far riferimento al "teorema della giuria di Condorcet" (Condorcet, 1785): dato un gruppo di votanti, in cui l'esito del voto è binario, definiamo con (p) la probabilità che ha ogni votante di votare correttamente e con (M) la probabilità che il voto di maggioranza sia corretto:

$$Se\ p > 0.5\ \Rightarrow M > p$$

$$M\ \to 1\ se\ \begin{cases} p > 0.5 \\ Numero\ votanti \to \infty \end{cases}$$

Tenendo conto di due importanti assunzioni:

- I voti della giuria sono indipendenti
- La votazione è binaria, ovvero, ci sono solo due scelte nella fase di votazione

Il teorema afferma che in determinate condizioni, una risposta corretta può essere assicurata utilizzando un numero appropriato di votanti, la cui capacità di giudizio è leggermente migliore di una decisione casuale.

I sistemi decisionali in grado di classificare correttamente la maggior parte delle istanze, salvo una limitatissima frazione, vengono definiti "classificatori robusti" ("*strong learner*", in inglese); mentre, i sistemi che classificano leggermente meglio rispetto ad una supposizione casuale vengono definiti "classificatori deboli" ("*weak learner*", in inglese). In quest'ultima categoria rientrano ad esempio i ceppi decisionali ("*decision stump*", in inglese), ovvero, alberi decisionali composti solo dal nodo radice in cui ad ogni foglia è associata la classe più presente nel sottoinsieme corrispondente; una situazione limite, raramente adottata ma obbligata se è necessario classificare un data set particolarmente eterogeneo e privo di sufficienti informazioni.

2.3.7.1 CARATTERISTICHE DI APPRENDIMENTO

Un sistema di apprendimento ensemble è composto dai seguenti elementi:

- set di addestramento ("*training set*");
- algoritmo di induzione: sfrutta il set di addestramento per generare un albero decisionale;
- generatore di ensemble: che definisce la metodologia della diversificazione degli alberi;
- combinatore: valuta i risultati degli alberi decisionali in funzione della classificazione delle istanze e li combina per ottenere il risultato finale.

Il risultato finale sarà strettamente legato agli elementi sopra riportati e alle tre caratteristiche principali che dominano un sistema di apprendimento automatico:

- Dipendenza tra classificatori: l'influenza che ha l'addestramento di un albero sull'addestramento dell'albero successivo.
- Diversità: capacità di produrre alberi molto diversi tra loro; seppur a discapito della loro affidabilità. Se ben parametrizzata, genera ensemble molto efficaci.
- Dimensione dell'ensemble: numero di classificatori presenti nell'ensemble.

La dipendenza e la diversità sono strettamente legate tra loro: se l'ensemble include un numero ridotto di classificatori, all'aumentare della dipendenza, diminuisce la diversità e viceversa.

Pertanto, i metodi di apprendimento più efficaci sono in grado di generare un elevato numero di classificatori (elevata dimensione dell'ensemble) i quali risulteranno dipendenti tra loro ma grazie alla numerosità riusciranno a coprire un ampio *range* di possibilità, che renderà alcuni classificatori molto diversi da altri; un'elevata dimensione dell'ensemble permette di avere un *esemble* con un'elevata dipendenza e contemporaneamente un'elevata diversità.

L'unico limite per generare un *ensamble* molto grande è il tempo di calcolo necessario; limite superabile con un buon calcolatore ed un algoritmo ben ottimizzato.

Gli algoritmi di ensemble *learning* possono esser classificati in base al metodo di combinazione; le categorie principali sono due: metodi a votazione pesata e metodi di meta-apprendimento. I metodi a votazione pesata generano alberi indipendenti, ovvero, la produzione dell'albero successivo non è condizionata in alcun modo dall'albero precedente; la classificazione finale avviene per competizione in funzione delle performance di ogni singolo albero.

I metodi basati sul meta-apprendimento generano alberi dipendenti tra loro: ogni albero generato è condizionato dall'albero precedente; in questo caso, la classificazione finale avviene per cooperazione.

Di seguito verranno brevemente descritti i principali metodi a votazione pesata e di meta-apprendimento.

METODI A VOTAZIONE PESATA

Nei metodi a votazione pesata viene assegnato un peso ad ogni classificatore; il peso rappresenta l'influenza del risultato del singolo classificatore sul risultato finale.

a. Votazione a maggioranza: gli alberi vengono prodotti dopo aver suddiviso il training set (processo simile alla "*cross-validation*") e come classe finale viene preferita quella che ha ottenuto il maggior numero di voti. Questo metodo è conosciuto anche come voto di pluralità o metodo base di ensemble.

b. Somma di distribuzione: metodo simile al precedente, ma in questo caso la scelta della classe finale viene effettuata sulla base della probabilità associata ad ogni albero. Questo metodo è generalmente utilizzato nel caso in cui alle foglie dei singoli alberi non sono associate le classi ma dei valori di probabilità per ognuna delle classi presenti.

c. *Bagging* (*Bootstrap Aggregating*): anche in questo caso il training set viene suddiviso ulteriormente; vengono prodotti numerosi alberi, ognuno dei quali è caratterizzato da uno specifico training set casuale. La generazione, totalmente causale, di sub-training set prevede la possibilità di riconsiderare più volte la stessa istanza così da promuovere al meglio la diversità tra gli alberi generati. La scelta dell'output negli alberi di classificazione avviene per votazione a maggioranza; mentre, negli alberi di regressione avviene come media dei risultati ottenuti. Per sfruttare al meglio questa tecnica, con l'obiettivo di favorire la diversità tra gli alberi, è preferibile utilizzare un algoritmo di induzione molto suscettibile alle variazioni del training set.

d. Foreste casuali (*Random Forests*): è un'evoluzione della tecnica *bagging* che incrementa ulteriormente la differenza tra gli alberi prodotti, diminuendo sensibilmente la loro correlazione. Nella fase di produzione viene selezionato casualmente un sottoinsieme di attributi, preferendo tra questi il migliore da utilizzare nella tecnica di suddivisione. Con questa tecnica si riduce notevolmente il numero di alberi prodotti ma contemporaneamente aumenta la diversità a vantaggio della prestazione complessiva dell'ensemble. Il nome, foresta casuale, deriva dal fatto che vengono generati casualmente numerosi alberi; ma nonostante la produzione così massiccia l'algoritmo ha dei tempi computazionali particolarmente ridotti, poiché viene ridotta notevolmente o addirittura eliminata la fase più dispendiosa, ovvero, la scelta del miglior attributo. Rivedremo nel dettaglio questo metodo più avanti.

METODI DI META-APPRENDIMENTO

I metodi di meta-apprendimento si caratterizzano per la loro capacità di auto-apprendimento, ovvero, sono in grado di apprendere dagli alberi prodotti e dalle classificazioni che questi effettuano sul training set; in pratica, vengono prima generati una serie di classificatori che producono determinate classificazioni, sulla base delle quali vengono generati altri classificatori che a loro volta produrranno nuove classificazioni. Il processo si arresta quando vengono raggiunte determinate condizioni: risultati attesi, dimensioni dell'ensemble, etc.

- *Stacking*: questa tecnica fornisce un'elevata precisione nella generalizzazione; viene utilizzata per valutare l'affidabilità degli alberi prodotti, combinando a questi altri alberi prodotti mediate algoritmi differenti. Dapprima il training set viene diviso in due parti: una parte, per la creazione del meta-data set; l'altra parte, per la creazione dei classificatori di base. Il meta-data set contiene un'istanza per ogni istanza del data set originale dove però gli attributi originali sono sostituiti con le classificazioni prodotte da ciascun albero; mentre, l'output rimane la classe originale. Le nuove istanze vengono utilizzate per produrre un meta-classificatore che combina in un'unica previsione le diverse ipotesi. Le classificazioni del meta-classificatore rappresentano le effettive prestazioni degli algoritmi di induzione di base. Nel classificare una nuova istanza ogni albero di base produce una sua previsione; queste previsioni andranno a costituire una nuova istanza che verrà classificata dal meta-classificatore. Le prestazioni di questa tecnica sono paragonabili a quelle del miglior classificatore scelto tramite la tecnica di "*cross-validation*" (validazione incrociata).

- Albero arbitro (*arbiter tree*): questo metodo genera un albero aggiuntivo definito arbitro che viene utilizzato insieme agli altri alberi per scegliere la classe finale. Gli alberi di base classificano il training set, tramite il quale verrà indotto l'albero arbitro sfruttando la regola di selezione delle istanze dal training set; queste andranno a costituire il training set dell'albero arbitro in funzione della classificazione condotta dagli alberi di base. Attraverso un algoritmo di induzione verrà creato l'albero arbitro e generato il nuovo training set. In fase di classificazione di un'istanza si utilizza la regola di arbitraggio con votazione di maggioranza da parte degli alberi di base e dell'arbitro; eventuali pareggi vengono risolti a favore della classe scelta dall'arbitro.

- Alberi combinatori (*Combiner trees*): il metodo è molto simile al metodo *stracking* in quanto utilizza le previsioni effettuate dai classificatori di base per generare un training set per l'algoritmo di meta-apprendimento; il contenuto delle istanze viene determinato utilizzando la regola di combinazione. In fase di classificazione di una nuova istanza vengono dapprima generate le previsioni degli alberi di base; successivamente, dalle previsioni ottenute viene generata una nuova istanza che il combinatore provvede a classificare. La finalità principali di questa tecnica è unire le previsioni degli alberi di base imparando a conoscere le relazioni tra queste previsioni e quella corretta. Il nuovo training set generato viene utilizzato per produrre il combinatore; per ogni nuova istanza gli alberi di base produrranno la loro previsione generando una nuova istanza, sulla base della stessa regola di combinazione utilizzata per produrre il training set del combinatore. Quest'ultimo valuterà la classe finale della nuova istanza.

- *Grading* (classificazione): questa tecnica genera un nuovo classificatore per ogni albero base e lo usa come valutatore per determinare se le previsioni dell'albero di base sono corrette. Gli alberi di base sono generati secondo la tecnica della validazione incrociata e vengono utilizzati per classificare tutte le istanze del training set. Per ogni albero di base viene prodotto un training set, in cui la classe delle istanze è composta da una classe che indica la corretta o scorretta classificazione dell'istanza da parte dell'albero. Da tutti i training set generati viene prodotto un ulteriore albero che ha lo scopo di valutare-prevedere di quanto l'albero di base sbaglierà la classificazione. Per la classificazione di una nuova istanza si utilizzano gli alberi di base per identificare le possibili classi e gli alberi derivati per decidere quali alberi sono efficaci; la scelta ricade sugli alberi ritenuti affidabili per classificare quella istanza nel miglior modo possibile, in alternativa si procede con una tecnica di votazione. Per la sua struttura e per le relative performance, questa tecnica può esser ritenuta un'evoluzione della validazione incrociata; in quanto non sceglie un solo albero, ma valutando i casi singolarmente, definisce l'albero migliore per quella istanza.

- *Boosting*: questo metodo utilizza i set di addestramento con istanze pesate; nella fase di apprendimento di un'ipotesi, l'importanza di ogni istanza è proporzionale al suo peso. L'algoritmo d'induzione utilizzato deve essere in grado di gestire le istanze ponderate; se ciò non è possibile, le istanze possono essere replicate in proporzione al peso assegnato: la numerosità dell'instanza sostituisce il peso assegnatogli. I principali punti del processo sono:

    1. Elaborazione di un albero con un processo che considera prioritarie le istanze con peso maggiore.
    2. L'albero generato viene utilizzato per classificare il training set.
    3. Viene ridotto il peso delle istanze classificate correttamente e congiuntamente viene incrementato il peso delle istanze classificate erroneamente.
    4. Si ripetono i punti da 1 a 3 fino a quando non viene raggiunto un determinato numero di alberi.
    5. Ad ogni albero viene attribuito un peso proporzionale alle sue prestazioni sul training set.

Le nuove istanze verranno classificate attraverso un sistema di votazione per maggioranza da parte di tutti gli alberi presenti. L'obiettivo di questo metodo è produrre un elevato numero di alberi per coprire quante più istanze possibili. I difetti principali di questo metodo sono la suscettibilità ed il rumore: l'eventuale presenza di dati errati comporterà l'attribuzione da parte dell'algoritmo di pesi sempre maggiori; generando così un peggioramento delle prestazioni. Per ovviare a tale problematica i metodi di *boosting* più avanzati prevedono che le istanze classificate erroneamente vengano interpretate come contenenti dati errati; di conseguenza viene tolto il peso a queste istanze giudicate errate.

2.3.8 APPRENDIMENTO BASATO SULLA RILEVANZA

Un altro metodo per affrontare il problema del sovradattamento è identificabile nell'apprendimento basato sulla rilevanza (*Relevance Based Learning – Decision Tree Learning*). A differenza dei metodi di potatura (*pruning*) che cercano di ignorare gli attributi superflui, questo metodo cerca di identificare le dipendenze tra i vari attributi e di produrre un training set che contiene solo gli attributi ritenuti necessari all'apprendimento.

L'obiettivo è l'identificazione della determinazione consistente minima, ovvero, il più piccolo data set di attributi che da solo è grado di determinare la classe di un'istanza.

Per raggiungere tale scopo è necessario valutare ogni sottoinsieme di attributi a partire dai singoli fino ad arrivare all'intero insieme e verificare se il sottoinsieme corrente è consistente con il training set; il primo sottoinsieme consistente sarà anche quello minimo. Un approccio di questo tipo richiede però un notevole sforzo computazionale; se consideriamo una determinazione minima di dimensione p su un totale di attributi pari ad n, la determinazione minima verrà individuata solo dopo aver analizzato i sottoinsiemi di p. Il numero dei sottoinsiemi da analizzare è pari a:

$$\binom{n}{p} = O(n^p)$$

Si tratta quindi di un problema NP-completo, di conseguenza non utilizzabile quando n e p sono elevati. La verifica delle prestazioni mostra però come questa tecnica sia particolarmente efficace in presenza di training set con attributi irrilevanti, in quanto riesce a generare alberi molto più precisi nonostante un data set particolarmente ridotto. Risulta quindi efficace nel caso in cui la classe è determinata soltanto da un numero

ridotto di attributi; viceversa, è pressoché inefficace a causa della sua richiesta computazionale, quando si hanno numerosi attributi. Infine, non vi sono vantaggi se questo metodo viene applicato nei casi in cui tutti gli attributi concorrono alla determinazione della classe.

2.4 *RANDOM FORESTS*

Il metodo *Random Forests* è stato introdotto da Leo Breiman, ispirato al precedente lavoro di Amit e German. Questo metodo è un'estensione dell'idea di *bagging* di Breiman e un "*competitor*" del *boosting* (Breiman, 2001). Il metodo delle foreste casuali (*Random Forests*) può essere impostato per produrre in output una valutazione delle classi considerate (metodo di classificazione), o per produrre una valutazione numerica continua delle variabili considerate (metodo di regressione). Le informazioni di input, così come quelle di output, possono essere degli attributi (categorie) o dei valori numerici (variabili continue). Il contributo computazionale che fornisce il metodo *Random Forests* rispetto a simili metodi di classificazione è sicuramente quello più importante, a parità di performance riscontrate; questo grande vantaggio è legato alle caratteristiche vincenti di questo metodo:

- Capacità di gestire rapidamente sia il processo di classificazione sia il processo di regressione.
- Tempistiche di addestramento e di produzione dell'informazione richiesta particolarmente brevi.
- Numero dei parametri di settaggio ridotto.
- Strutturato per fornire una stima dell'errore di generalizzazione (classificazione).
- Particolarmente indicato per set di dati molto grandi.

Per sua natura il metodo *Random Forests* risulta molto utile anche per ricavare importanti informazioni dal data set utilizzato; informazioni facilmente accessibili e comprensibili anche da un utente non particolarmente esperto in materia.

Di seguito verranno presentate le principali caratteristiche che rendono questo metodo particolarmente versatile e vantaggioso. (Cutler et al., 2012; Louppe, 2014)

In fase di elaborazione l'algoritmo analizza il data set fornitogli e valuta l'importanza che le variabili hanno sul risultato finale; un'informazione preziosa poiché consente eventualmente di eliminare (manualmente o automaticamente) le variabili superflue, rendendo il processo di elaborazione ancora più veloce, ma soprattutto permette all'utilizzatore di conoscere quali sono le informazioni che dominano il processo di elaborazione della foresta casuale.

Nella fase di parametrizzazione dell'algoritmo, che genererà la foresta casuale, è possibile attribuire dei pesi alle classi presenti nel training set; questa funzione consente di "guidare" l'algoritmo in una valutazione differente di una o più variabili rispetto a quanto avrebbe fatto basandosi solamente sul training set. Ne risulterà una foresta casuale "forzata" che fornirà una valutazione differente rispetto ad una foresta casuale addestratasi autonomamente. Se utilizzata con la dovuta accortezza, questa funzione risulta molto importante nel caso in cui l'obiettivo è valutare una classe "rara" ma di particolare interesse all'interno del data set, senza che la valutazione di questa classe venga influenzata dalle altre classi aventi un numero di istanze sensibilmente maggiore.

Sia nella modalità classificazione sia nella modalità regressione, il metodo *Random Forests* consente di individuare eventuali valori mancanti all'interno del training set; ciò implica che non è più necessario il

controllo del training set in fase preliminare per verificare eventuali dati mancati. Congiuntamente, l'algoritmo individua i valori anomali all'interno del training set; i quali possono essere o scartati (errori) oppure considerati (valori eccezionali).

Il metodo *Random Forests* è basato sull'apprendimento supervisionato, ovvero, in fase di addestramento richiede sia le informazioni di input (causa) sia le informazioni di output (effetto); ma, se opportunatamente parametrizzato, l'algoritmo è in grado di imparare iterativamente dal training set fornitogli (apprendimento non supervisionato). I risultati ottenuti da un primo tentativo saranno le classi del training set nel tentativo successivo; il processo si arresta al raggiungimento di un obiettivo prefissato o di una determinata condizione, oltre la quale l'algoritmo non riesce a migliorare la propria foresta casuale. Questa modalità richiede tempi di elaborazione più elevati e un calo delle performance rispetto alla modalità supervisionata; ma in assenza degli "effetti" nel training set risulta particolarmente utile. Nei capitoli in cui verranno presentati i vari casi studio, si farà riferimento alle peculiarità appena descritte, che hanno contribuito positivamente nel raggiungimento degli obiettivi prefissati; ma soprattutto sono state le principali forzati che hanno motivato la scelta del metodo *Random Forests* rispetto agli altri metodi di apprendimento automatico.

2.4.1 L'ALGORITMO

In questo paragrafo vengono illustrate le nozioni di base relative all'algoritmo del *Random Forests*; necessarie per comprendere il costrutto della tecnica impletata, l'utilizzo delle variabili e la differenza sostanziale tra classificazione e regressione.

L'algoritmo *Random Forests* genera un insieme di alberi (*tree-based ensemble*) in cui ogni albero dipende da una serie di variabili casuali. Analiticamente, consideriamo un ambiente p-dimensionale dove il vettore casuale $X = (X_1, \dots, X_p)^T$ rappresenta le reali variabili in input o il predittore di valore reale X_i, mentre la variabile casuale Y è il valore reale di output; inoltre, assumiamo che $P_{XY}(X, Y)$ sia una distribuzione di probabilità. L'obiettivo è identificare una funzione di previsione $f(X)$ per predire Y. La funzione di predizione è determinata da una funzione di perdita $L(Y, f(X))$ e definita per minimizzare il valore atteso della perdita $E_{XY}(L(Y, f(X)))$ rispetto alla distribuzione che lega X e Y. Indicativamente, $L(Y, f(X))$ rappresenta quanto $f(X)$ si avvicina a Y, penalizzando i valori di $f(X)$ che risultano essere molto diversi da Y.

Generalmente l'errore (perdita) di L sono lo "*squared error loss"* (perdita quadratica) $L(Y, F(X)) = (Y - f(X))^2$ per la regressione e lo "*zero-one loss*" (perdita zero-uno) per la classificazione.

$$L(Y, f(X)) = I(Y \neq f(X)) = \begin{cases} 0 \; if \; Y = f(X) \\ 1 \; otherwise \end{cases}$$

Riducendo al minimo $E_{XY}(L(Y, f(X))$ per lo "*squadred error loss*", si ottiene la condizione attesa, nota come funzione di regressione:

$$f(x) = E(Y|X = x)$$

Nel caso della classificazione, se l'insieme di possibili valori di Y è denominato Υ, minimizzando $E_{XY}(L(Y, F(X)))$ per "*zero-one loss*" si ha la seguente funzione, nota anche come "*Bayes rule"* (regola di Bayes):

$$f(x) = argmax P(Y = y|X = x)$$

$$Y \in \mathcal{Y}$$

Il costrutto di esemble f è composto dai cosiddetti "*base learners*" (apprenditori di base) $h_1(x), \ldots, h_J(x)$; questi sono combinati insieme per generare un "ensemble predictor" (insieme dei preditori) $f(x)$.

Nella tecnica di regressione, i "*base learners*" sono mediati tra loro:

$$f(x) = \frac{1}{J}\sum_{J=1}^{J} h_j(x)$$

Mentre, nella tecnica di classificazione, $f(x)$ è la classe che è stata prevista con una frequenza maggiore; scelta per votazione (*voting)*.

$$f(x) = argmax_{y\epsilon\mathcal{Y}} \sum_{J=1}^{J} I(y = h_j(x))$$

Nel metodo *Random Forests*, il j_{esimo} apprenditore di base è un albero decisionale denominato $h_j(X, \Theta_j)$; dove Θ_j è una raccolta di variabili casuali e le Θ_j sono indipendenti per $j = 1 \ldots J$. Ovvero, ogni albero che fa parte dell'insieme è caratterizzato da variabili indipendenti, ciò implica che tutti gli alberi sono tra loro indipendenti.

2.4.2 ALBERI DI CLASSIFICAZIONE E REGRESSIONE

Gli alberi che compongono le foreste casuali sono basati sul processo ricorsivo di partizionamento binario. (Breiman, 1984; Hastie, Trevor, Tibshirani, Robert, Friedman, 2009; Izenman, 2008; Zhang and Singer, 2010)

Gli alberi generati dividono lo spazio predittore attraverso una sequenza di partizioni binarie, definite "*split*", sulle singole variabili. Il primo nodo che compone l'albero, definito nodo radice ("*root node*"), comprende l'intero spazio predittore. I nodi non succeduti da suddivisioni vengono definiti "nodi terminali" (*terminal nodes*") e costituiscono la partizione finale dello spazio predittore. Ogni nodo terminale si divide in due nodi discendenti, uno a sinistra ed uno a destra, in base al valore di una delle variabili predittive.

Per una variabile predittiva continua, la suddivisione è determinata da un punto di divisione ("*split-point*"); i punti in cui il predittore è più piccolo dello *split-point* ricadranno sulla condizione di sinistra, gli altri in quella di destra.

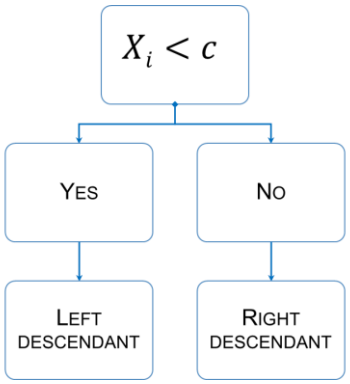


Figura 9 Rappresentazione schematica della suddivisione di una variabile predittiva continua

Una variabile predittrice categoriale X_i prende i valori da un insieme finito di categorie, definito come $S_i = \{S_{i,1}, \ldots, S_{i,m}\}$. Una suddivisione già definita colloca un sottoinsieme di queste categorie $S \subset S_i$ a sinistra mentre la porzione restante a destra.

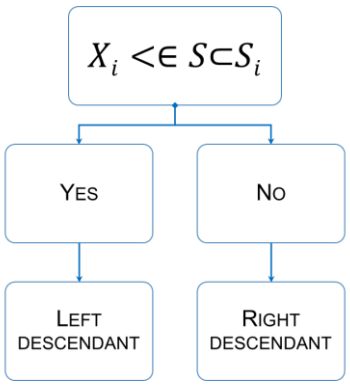


Figura 10 Rappresentazione schematica della suddivisione di una variabile predittiva categoriale

La suddivisione caratteristica, che ogni albero utilizza per suddividere un nodo in due porzioni, è scelta considerando ogni possibile divisione su ogni variabile predittrice e scegliendo quella migliore (*best*) secondo alcuni criteri. Nel caso della regressione, se i valori di risposta sul nodo sono $y_1, \ldots, y_n$ un tipico criterio di suddivisione al nodo ("*splitting*") è la media dei residui al quadrato ("*mean squared residual*"):

$$Q = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

Dove il valore previsto del nodo è la media dei valori di risposta:

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Mentre, nella classificazione in cui le K-classi sono così identificate $Classi = 1, \ldots, K$ generalmente viene utilizzato l'indice di Gini ("*Gini index*") come criterio di suddivisione:

$$Q = \sum_{k \neq k'}^{k} \widehat{p_k}\ \widehat{p_{k'}}$$

Dove $\widehat{p_k}$ è la proporzione delle osservazioni di classe k nel nodo:

$$\widehat{p_k} = \frac{1}{n}\sum_{i=n}^{n} I(y_i = k)$$

Il criterio di suddivisione fornisce una misura della "bontà di adattamento" ("*goodness of fit*"), nel caso della regressione, o di "purezza" ("*purity*") per un nodo nella classificazione; con valori elevati che rappresentano un adattamento inadeguato (regressione) o un nodo impuro (classificazione).

Una suddivisione crea due nodi discendenti, uno a sinistra (left) e l'altro a destra (right); i nodi discendenti sono Q_L e Q_R. Le loro dimensioni del campione sono rispettivamente n_L e n_R.

La scelta della loro divisione è guidata dalla ricerca nel minimizzare il valore:

$$Q_{split} = n_L Q_L + n_R Q_R$$

Per una variabile predittrice continua, identificare la miglior divisione possibile comporta l'ordinamento dei valori del predittore e la considerazione delle suddivisioni tra ogni coppia distinta di valori consecutivi. Generalmente, si considera il punto medio dell'intervallo, anche se sarebbe sufficiente e corretto considerare qualunque valore contenuto nell'intervallo. I valori di Q_L e Q_R ed anche di Q_{split} vengono calcolati per ciascuno di questi possibili punti di divisione, solitamente utilizzando un altgoritmo a rapido aggiornamento. Per una variabile predittiva categoriale Q_L, Q_R e Q_{split} vengono calcolati in tutti i possibili modi per scegliere un sottoinsieme di categorie da assegnare a ciascun nodo discendente. Identificata una divisione, i dati vengono suddivisi nei due nodi discendenti e ciascuno di questi nodi viene considerato allo stesso modo del nodo originale.

La procedura si ripete in modo ricorsivo finché non viene soddisfatto un criterio di arresto che, ad esempio, può essere quando tutti i nodi suddivisi contengono meno di un numero fisso di casi. Gli ultimi nodi vengono definiti "terminali".

Per tutte le osservazioni, si ottiene un valore previsto nei nodi terminali; i valori previsti saranno una media della risposta per i problemi di regressione o il calcolo della classe più frequente per i problemi di classificazione.

Per la previsione di un nuovo punto, il suo insieme di valori predittori viene utilizzato per "scorrere" il nuovo punto lungo l'albero fino a quando non ricade in un nodo terminale e la previsione per il nodo terminale viene utilizzata come previsione per il nuovo punto.

Un processo ricorsivo di questo tipo potrebbe portare alla creazione di alberi molto grandi e, quindi, foreste particolarmente estese ("*overgrowing problem*", "*overfitting problem*"); il metodo Random Forests, per come è definito, controlla automaticamente le dimensioni dell'albero e delle foreste: attraverso una parametrizzazione del numero di alberi preliminare. Pertanto, non sono necessarie tecniche di "potatura" ("*pruning*").

2.4.3 DEFINIZIONE RANDOM FORESTS

Il metodo *Random Forests* genera alberi $h_j(X, \Theta_j)$ come apprenditori di base; il dataset di apprendimento necessario per generare la foresta è un insieme $\mathrm{D} = \{(x_1, y_1), \dots, (x_N, y_N)\}$, dove $x_i = (x_{i,1}, \dots, x_{i,p})^T$ indicano i p-predittori mentre y_i rappresentano le risposte associate ai p-predittori; x_i e y_i rappresentano le istanze di apprendimento e sono legati tra loro da una particolare relazione θ_j di Θ_j, da cui risulta l'albero $\widehat{h_j}(x, \theta_j, \mathrm{D})$.

La formulazione sopra riportata è quella originale di Breiman, padre del metodo *Random Forests*, in operatività la componente casuale θ_j non è considerata esplicitamente ma è implicitamente utilizzata per produrre la casualità attraverso due specifiche modalità: la prima, simile al *bagging*, permette una creazione indipendente dell'albero (questa prima randomizzazione costituisce una parte di θ_j); la seconda, nella suddivisione del nodo, identifica in maniera indipendente su ogni nodo la divisione migliore su un sottoinsieme selezionato casualmente di variabili predittrici m anziché su tutti i predittori p (questa seconda randomizzazione costituisce la parte rimanente di θ_j).

Il confronto finale tra gli alberi generati viene condotto con la tecnica del "voto non pesato" ("*unweighted voting*") nella classificazione; mentre, la tecnica della "media non pesata" ("*unweighted averaging*") nella regressione.

2.4.4 OUT OF BAG DATA (OOB)

L'errore di previsione del metodo *Random Forests* viene quantificato attraverso la tecnica "*OOB Out-Of-Bag error*", letteralmente "errore fuori borsa"; "*OOB-error*" è l'errore di predizione medio su ciascun campione di apprendimento x_i emerso nel classificare il dataset di addestramento. Le informazioni relative a "*OOB-error*" sono molto utili, oltre che per stimare gli errori di classificazione, anche per valutare l'importanza delle variabili.

Per stimare l'errore di classificazione, prima di tutto è necessario considerare le dimensioni degli alberi: alberi molto gradi sono rappresentativi di una classificazione che ha utilizzato tutte le osservazioni disponibili; quindi, i risultati di verifica saranno eccessivamente positivi ma relegati al solo training set. Inoltre, per la stima dell'errore di classificazione si utilizzano solo le istanze che sono state classificate come OOB.

Per il metodo di regressione con ottimizzazione "*squared error loss*" (perdita quadratica), l'errore di generalizzazione è stimato utilizzando l'errore quadratico medio dei valori *OOB* (MSE):

$$MSE_{OOB} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \widehat{f_{OOB}}(x_i))^2$$

dove $\widehat{f_{OOB}}(x_i)$ è la previsione per l'istanza i-esima

Per il metodo di classificazione con ottimizzazione "*zero-one loss*" (perdita zero-uno), l'errore di classificazione è stimato utilizzando "*Out-Of-Bag Error Rate*" (tasso di errore di classificazione):

$$E_{OOB} = \frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \widehat{f_{OOB}}(x_i))$$

Utilizzando il tasso di errore, nella modalità di classificazione, è possibile ottenere un tasso di errore di classe per ognuna delle classi considerate ed una "*confusion matrix*" dove vengono incrociate le "risposte" y_i e le istanze erroneamente associate $\widehat{f_{OOB}}(x_i)$.

La tecnica di valutazione dell'errore di classificazione, uno dei principali "punti di forza" del metodo Random Forests, verrà nuovamente trattata nei capitoli relativi ai casi studio sviluppati in questa tesi.

2.4.5 MIGLIORAMENTO E POTENZIAMENTO DELL'ALGORITMO *RANDOM FORESTS*

È possibile cercare di migliorare, entro un certo limite, le performance dell'algoritmo *Random Forests* ottimizzando i tre parametri principali in funzione degli obiettivi prefissati; i parametri su cui intervenire per poter migliorare le performance del modello sono:

- *m*, il numero di variabili predittive selezionate casualmente su ciascun nodo
- *J*, il numero di alberi che compongono la foresta casuale
- *tree size*, dimensione dell'albero calcolata in funzione del numero di "split" presenti nel nodo più vicino al nodo radice; oppure, calcolata in funzione del massimo numero di nodi terminali.

Tra questi, il parametro avete un peso maggiore è *m* (numero di variabili predittive). Nella modalità "classificazione", il valore di default è $m = \sqrt{M}$ dove *M* è il numero totale di predittori; mentre, nella modalità "regressione" $m = \frac{N}{3}$ dove *N* è la dimensione del campione. Generalmente, l'algoritmo Random Forests non è molto simile alla variazione di *m* (Díaz-Uriarte e Alvarez de Andrés, 2006); come verrà illustrato nei successivi capitolo, nei casi studio affrontati il numero di variabili particolarmente influenti nell'elaborazione della foresta casuale sono in numero sensibilmente ridotto rispetto all'insieme delle variabili predittive (*m*).

Numerosi metodi ensemble generano un elevato numero di alberi (parametro *J*) che compongono la foresta casuale; questa tecnica permette di ridurre inizialmente l'errore di classificazione ma oltre un certo valore di *J*, quindi per foreste molto grandi composte da alberi ridondanti, si ha un'inversione dell'errore di classificazione, ovvero, aumenta con l'aumento di *J*. Nel metodo *Random Forests* ciò non si verifica: per valori molto piccoli di *J* spesso si può riscontrare una forte variabilità nelle performance di classificazione; ma oltre un certo valore di *J* l'errore di classificazione si stabilizza e rimane invariato all'aumentare di *J*. Questa peculiarità apporta due importanti vantaggi nella fase di addestramento-calibrazione: il primo, è possibile verificare il valore di *J-minimo* oltre il quale l'errore di classificazione rimane costante, così da poter ridurre al minimo i tempi di elaborazione; il secondo, si può saltare il passaggio precedente e definire l'algoritmo sul valore suggerito da Breiman (Breiman, 2001) pari a *J*=500, un numero sufficientemente grande per garantire la stabilizzazione dell'errore di classificazione e contemporaneamente non troppo oneroso in termini computazionali. Nel terzo capitolo, verrà presentato un test condotto nel primo caso studio per l'ottimizzazione del parametro *J* ed i motivi associati alla scelta effettuata per il valore finale di *J*.

2.4.6 DATASET SBILANCIATI E CLASSI PESATE

Un data set di addestramento (*training set*) caratterizzato da numerose istanze attribuite ad un numero ristretto di classi e/o dall'assenza di alcune classi viene definito sbilanciato. Questo è uno dei problemi principali dei classificatori che si ripercuote direttamente sulle performance del classificatore; in tali condizioni, il classificatore riuscirà a classificare correttamente solo le classi con numerose istanze, mentre, valuterà con un elevato tasso di errore le classi a cui sono associate poche istanze. Nel caso limite in cui una classe non presentasse neanche una sola istanza associata (classe-assente), il classificatore non sarà in grado di classificare le future istanze appartenenti alla classe-assente nella fase di training set. L'assenza di una o più classi può essere risolta con pre-elaborazioni del training set, esistono diverse tecniche che risolvono solo parzialmente il problema; altrimenti, si deve ricorrere ad un metodo di apprendimento non supervisionato.

In riferimento ai data set sbilanciati, il *Random Forests* presenta un metodo efficace per pesare le classi da classificare: attribuendo dei pesi alle classi del training set, "diminuendo l'importanza" delle classi più numerose ed "aumentando" quella delle classi meno numerose, è possibile "bilanciare" il set di addestramento.

Questo metodo è applicabile anche in presenza di un training set bilanciato dove è necessario in fase di classificazione dare una maggior importanza solo ad alcune classi; ad esempio per far fronte a quei casi in cui è più grave concludere che tutto andrà bene piuttosto che concludere in modo errato che non tutto andrà bene.

2.4.7 IMPORTANZA DELLE VARIABILI PREDITTIVE

In fase di addestramento l'algoritmo valuta l'importanza delle variabili predittive presenti, ciò permette di eliminare le variabili poco influenti o del tutto inutili, a vantaggio dei tempi di processamento, e dar maggior considerazione a quelle più importanti, soprattutto in fase decisionale. Alcuni metodi di classificazione utilizzano delle tecniche di valutazione per selezionare le variabili principali prima del processo di classificazione; ciò comporta il rischio di trascurare alcune variabili non importanti in fase di elaborazione-costruzione ma rilevanti nella fase di perfezionamento della classificazione.

Il metodo *Random Forests* utilizza una tecnica tanto particolare quanto semplice per misurare l'importanza delle variabili. Consideriamo, tra le tante variabili presenti nel training set, una variabile *k*; per ogni albero viene eseguita la seguente procedura di valutazione. Inizialmente si valutano le performance ottenute con la variabile *k* originaria; successivamente, viene modificata utilizzando casualmente i valori "fuori borsa" (*out-of-bag data*); il nuovo risultato ottenuto viene confrontato con quello precedente. Il confronto riguarda non solo le performance della variabile *k* ma anche delle altre variabili; quindi è possibile valutare oltre all'importanza della singola variabile *k* anche come questa influisce sulla valutazione delle altre variabili. In questo modo è possibile capire se una o più variabili oltre ad essere importanti influiscono a tal punto da condizionare le capacità predittive dell'algoritmo utilizzato.

In fase operativa, come verrà mostrato nei successivi capitoli, sapere quali sono le variabili dominanti aiuta l'operatore nella comprensione dei risultati forniti dal metodo *Random Forests* e quindi nella valutazione finale. Ad esempio, apprendere che la variabile *Z* abbia un notevole peso in termini di elaborazione del *Random Forests*, consente all'operatore di valutare i risultati forniti dal modello in funzione dell'affidabilità dei valori della variabile *Z*; valori, che per ipotesi, risultano essere affidabili fino ad un certo limite (a vantaggio dell'elaborazione del *Random Forests*) e poi totalmente inaffidabili (a svantaggio dell'elaborazione del *Random Forests*).

L'importanza delle variabili viene valutata attraverso due indici "*Mean Decrease Accuracy*" (diminuzione media dell'accuratezza) e "*Mean Decrease Gini*" (diminuzione media dell'indice di Gini): il primo, associa un valore ad ogni variabile in funzione dell'influenza che questa ha sulla precisione dei risultati ottenuti; il secondo, associa un valore ad ogni variabile in funzione dell'influenza che questa ha sulla realizzazione dell'albero decisionale.

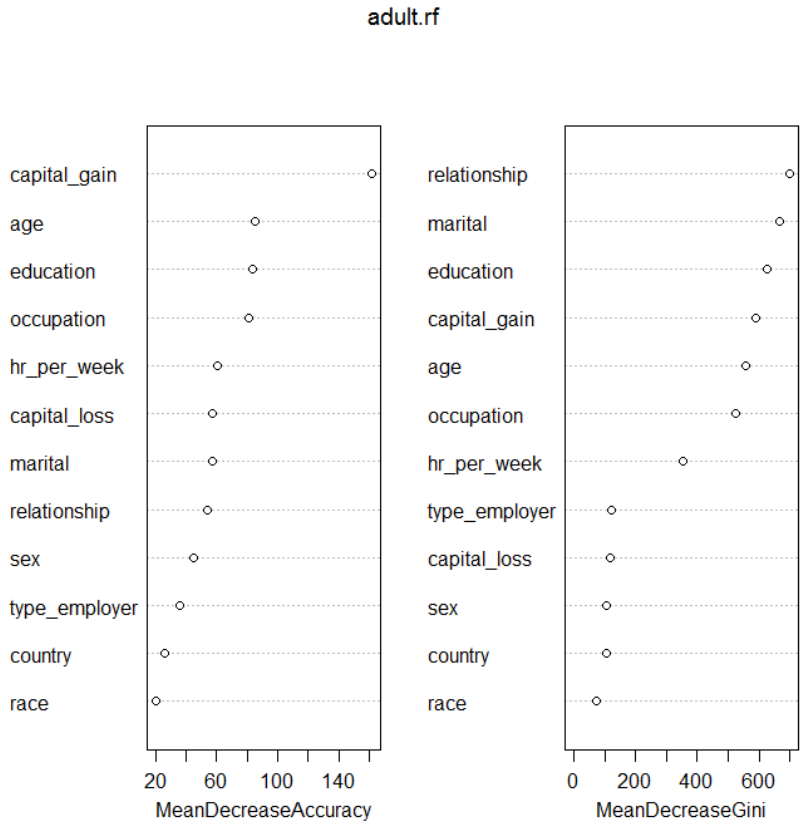


Figura 11 Esempio grafico dei due indici di valutazione delle variabili predittive

Generalmente, per la valutazione delle variabili, si utilizza il "*Mean Decrease Gini*". Questi due indici verranno ripresi nel dettaglio nei casi studio presentati.

2.4.8 PROSSIMITÀ TRA DUE OSSERVAZIONI

La prossimità ("*proximity*", in inglese) tra due osservazioni è il rapporto tra di esse all'interno dei nodi terminali: se queste si troveranno nello stesso nodo terminale, la loro prossimità sarà 1; mentre, se non sono all'interno dello stesso nodo terminale la prossimità è 0.

Generalmente, la distanza tra due osservazioni nello spazio euclideo condiziona la distanza tra queste anche all'interno dell'albero: due osservazioni vicine, nello spazio euclideo, risulteranno vicine "di foglia" anche al termine del processo di classificazione. Questa tecnica è utile in fase di training per identificare eventuali osservazioni che sono sensibilmente lontane dai dati presenti nel training set, a tal punto da generare nodi terminali (foglie) con una singola osservazione, o comunque, con un numero di osservazioni particolarmente

ridotto; in fase operativa, risulta importante per classificare quei valori che non erano presenti in fase di training.

Per quest'ultimo motivo è importante avere un training set eterogeno ed affidabile così da comporre un albero con tante foglie tra loro vicine; in tal caso, le future osservazioni verranno classificate in nodi appositi, invece che esser collocate in uno spazio predittivo privo di osservazioni di addestramento che impone al classificatore di collocare queste istanze in una foglia o in quella a fianco, con un conseguente calo delle performance del modello.

2.5 APPROCCIO PROBABILISTICO

Nel paragrafo 2.4.2 sono state presentate le due modalità di elaborazione del metodo *Random Forests*: classificazione e regressione. Utilizzando le potenzialità del metodo di classificazione, per ogni istanza è possibile identificare una probabilità di corretta classificazione associata alle classi presenti nel training set; ovvero, attraverso una corretta formulazione dell'algoritmo del *Random Forests*, per ogni nuova istanza viene fornito un output composto da un valore percentuale per ogni classe presente nella fase di training set. Pertanto la nuova istanza potrà ricadere totalmente (probabilità 100%) in una delle classi presenti, oppure, ricadere parzialmente in due o più istanze; la somma delle percentuali risulterà sempre 100%.

Ad esempio, consideriamo un training set contenente 100 istanze e 3 classi (ognuna delle quali rappresenta una soglia di livello idrometrico); al termine della classificazione, verrà prodotto un albero decisionale avente in ogni nodo terminale (foglia) le tre classi considerate e le relative probabilità di corretta classificazione.

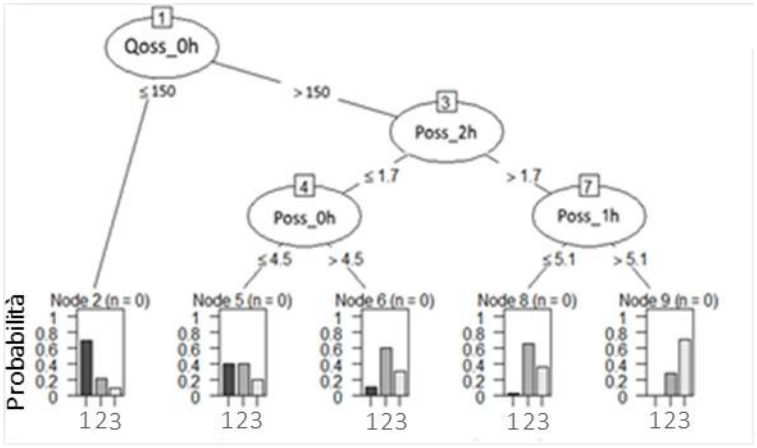


Figura 12 Rappresentazione semplificata di un albero decisionale con output probabilistico

In fase operativa, si richiede di classificare l'istanza 101; la quale, in funzione dei valori ad essa associati, ricadrà nel nodo terminale n°8, dove sono state identificate le seguenti probabilità di classificazione:

- Classe 1: 5%
- Classe 2: 60%
- Classe 3: 35%

Pertanto, l'istanza 101 non verrà classificata in una sola classe ma su tutte e tre le classi; la percentuale maggiore è riferita alla classe 2, ma una probabilità, seppur minore, è presente anche in classe 3 e in classe 1, prossima a zero ma non nulla.

Un'informazione di questo tipo permette all'utilizzatore di effettuare eventuali valutazioni sulla base dei risultati forniti dal modello.

Se consideriamo le classi come soglie di allertamento, per ipotesi, la Classe 3 racchiude un livello di rischio sensibilmente maggiore rispetto alla Classe 2; a tal punto da giustificare degli interventi sul campo anche con una probabilità medio-bassa (35%).

Nei successivi capitoli (3 e 4) è stato ulteriormente approfondito questo concetto, in relazione ai casi studio affrontati e alle modalità con cui questa tecnica è stata ottimizzata per il raggiungimento degli obiettivi prefissati.

3 PREVISIONE DELLE PIENE FLUVIALI NEI BACINI DELL'EMILIA-ROMAGNA

3.1 INTRODUZIONE

L'orografia dell'Emilia-Romagna è caratterizzata da una struttura ben definita: l'area pianeggiante, parte della Pianura Padana, si estende verso nord-est; mentre, l'area montuosa-collinare si estende verso sud-ovest. Le due aree sono divise dalla storica Via Emilia, costruita dal console romano Marco Emilio Lepido nel II sec. a.C.

I depositi alluvionali trasportati dal fiume Po nel corso dei millenni compongono l'area pianeggiante nota come Val Padana; una parte di questa è afferente all'Emilia-Romagna e rappresenta l'intera area di pianura della regione.

La regione è divisa in Emilia, da Piacenza fino alla provincia di Bologna (Ferrarese incluso), e Romagna; nell'area emiliana sono presenti le vette più elevate, tra cui la vetta più alta della regione, Monte Cimone (2165 m s.l.m.).

La rete idrografica regionale è molto sviluppata e comprende i seguenti corsi d'acqua:

- il Fiume Po, il principale corso d'acqua italiano;
- 7 vie navigabili di secondo ordine: Fiume Enza, F. Panaro, F. Parma, F. Reno, F. Secchia, F. Taro e F. Trebbia
- 27 corsi d'acqua di terzo ordine
- oltre 40 corsi d'acqua di quarto ordine

I corsi d'acqua sono divisi in due gruppi principali: i corsi d'acqua occidentali, che sfociano nel Fiume Po, e i corsi d'acqua orientali, che sfociano tutti nel Mar Adriatico.

Tutti i fiumi della regione eccetto il Po hanno carattere torrentizio con eventi di piena particolarmente severi in autunno e primavera, talvolta anche in inverno, e lunghi periodi siccitosi, in Estate. Durante gli eventi di piena, nella parte alta dei corsi d'acqua si verificano frequenti fenomeni di erosione, sia del fondo sia spondale, che provocano l'innesco di frane a causa dell'elevata presenza di un substrato facilmente erodibile; le piene fluviali, quindi, sono caratterizzate da un consistente trasporto solido.

I bacini appenninici emiliani registrano precipitazioni più intense e frequenti rispetto ai bacini romagnoli, a causa della maggior altitudine e vicinanza alla dorsale appenninica più esposta alle perturbazioni provenienti dal Mar Tirreno. I bacini dove si verificano le precipitazioni maggiori sono quelli del Fiume Trebbia e del Fiume Taro (tra Piacenza e Parma); procedendo verso est si ha una diminuzione delle precipitazioni, sia in intensità sia in volumi.

Gli accumuli maggiori e le precipitazioni più intense si verificano in prossimità dello spartiacque appenninico, tra Toscana, Emilia-Romagna e Liguria; dove le condizioni orografiche tendono ad amplificare i fronti perturbati provenienti dal Mar Ligure o contribuiscono alla formazione di celle temporalesche autorigeneranti. Procedendo verso la pianura le precipitazioni sono via via più modeste.

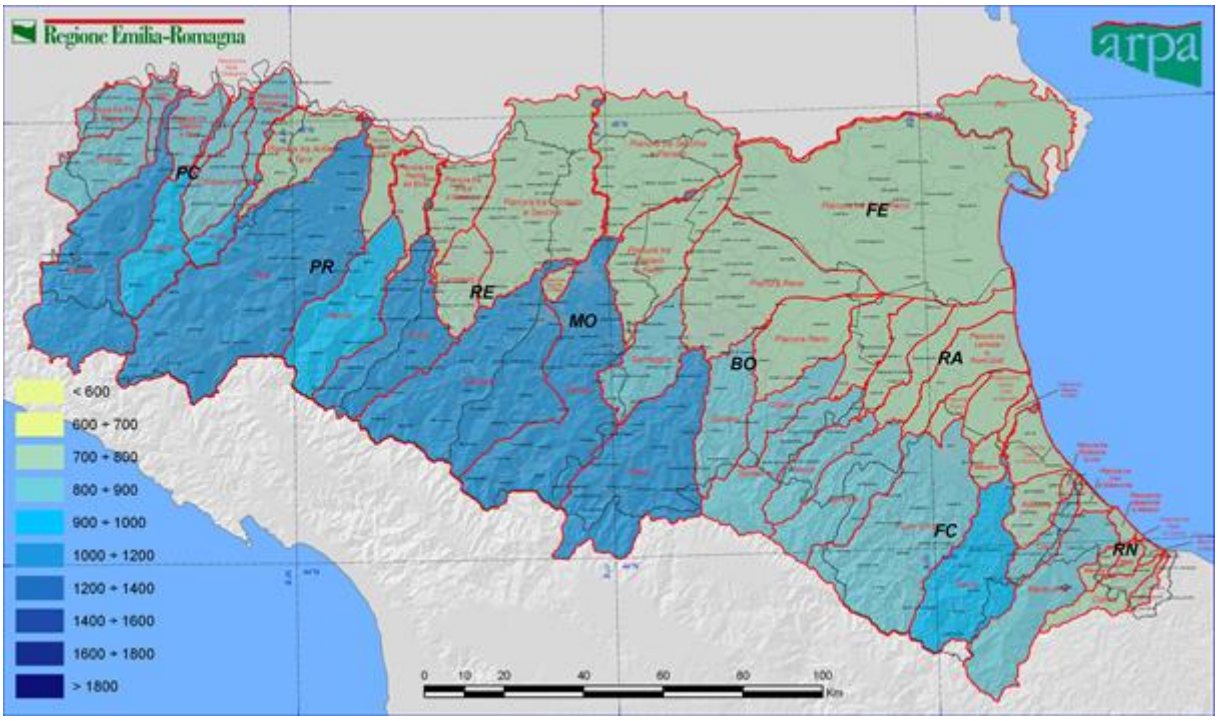


Figura 13 Precipitazioni medie annuali (periodo 1991 – 2006)
Fonte: ARPAE Emilia-Romagna

I bacini, generalmente stretti ed allungati, presentano tempi di risposta particolarmente brevi nelle aree collinari (2-4 ore, talvolta anche meno); mentre, procedendo verso la prima pianura i tempi salgono fino a 5-8 ore. Tempistiche così ridotte obbligano a produrre informazioni previsionali nel minor tempo possibile, così da permettere eventuali interventi nel breve arco temporale a disposizione; pertanto è necessario ricorrere alle più recenti e rapide tecniche previsionali per fronteggiare le situazioni più critiche.

Per il caso studio presentato in questo capitolo è stato utilizzato il bacino del fiume Parma (fiume di secondo ordine) come "bacino pilota"; la scelta è stata dettata dalle necessità operative, a seguito del grande evento di piena verificatosi nell'ottobre 2014, con l'obiettivo di massimizzare l'utilizzo dei dati osservati per generare un'affidabile informazione previsionale, nel minor tempo possibile.

Inoltre, il bacino del Parma è tra i principali dell'Emilia-Romagna e le sue caratteristiche sono molto simili agli altri bacini emiliani; mentre, i bacini romagnoli hanno caratteristiche più simili all'affluente del Fiume Parma, il Torrente Baganza.

3.1.1 Il bacino del Fiume Parma

Il bacino del Fiume Parma si estende nella provincia di Parma per circa 815 km^2; di cui il 60% in area di montagna e il restante in area di pianura.

Il principale corso d'acqua, il fiume Parma, nasce dal Lago Santo e dai laghi Gemio e Scuro sul crinale del Monte Orsaro e del Monte Sillara. Il fiume scorre per una lunghezza totale di circa 100 km lungo la direttrice nord-est.

Nella città di Parma riceve il contributo del torrente Baganza, il suo principale affluente; il tratto a valle della città fino alla confluenza con il fiume Po è interamente arginato.

Un'alta piovosità caratterizza il regime pluviale, di tipo appenninico sub-costiero, solo nelle aree vicine al crinale appenninico; mentre nella parte alta della pianura (200-300 m di quota) le precipitazioni sono modeste. Nel bacino idrografico, gli afflussi medi variano da 800 mm/anno (area pianeggiante) a 2000 mm/anno (crinale).

Il regime idrologico è di tipo torrentizio con eventi di piena durante l'autunno e la primavera, e intensi eventi di siccità nella stagione estiva. Le portate più alte si verificano solitamente nella stagione autunnale. La forma allungata e stretta del bacino fa sì che durante gli eventi di piena più intensi, i picchi siano stretti ed elevati; sia per il fiume Parma sia per il torrente Baganza. Per contenere i volumi di piena durante gli eventi principali, riducendo il rischio delle inondazioni, a monte della città di Parma è stata realizzata una cassa di espansione.

Le due sezioni idrometriche di riferimento sono quella di Ponte Verdi per il Fiume Parma e di Ponte Nuovo per il Torrente Baganza.

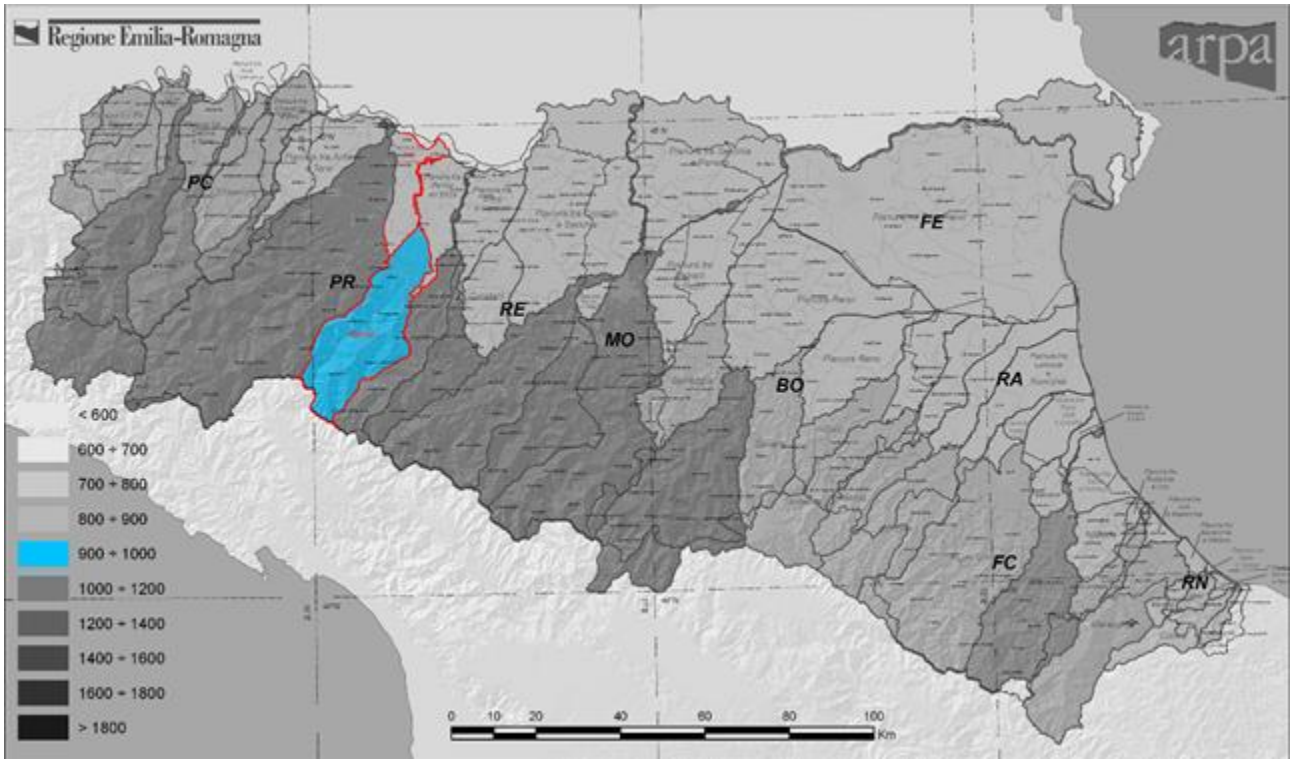


Figura 14 Inquadramento geografico del bacino del Fiume Parma
Fonte: ARPAE Emilia-Romagna

3.2 DATI IDRO-METEOROLOGICI

In seguito alla grande piena del Fiume Po (anno 2000) è stata sviluppata una rete capillare di monitoraggio idro-meteorologico, necessaria anche per contrastare la complessità morfologica e idrologica della regione. In precedenza, numerose stazioni meccaniche erano già presenti sul territorio; ma il passaggio a stazioni di monitoraggio automatiche ha permesso una maggior frequenza di acquisizione dei dati congiuntamente ad una maggior accuratezza delle informazioni raccolte.

I dati di monitoraggio, raccolti in telerilevamento in tempo reale, vengono rapidamente inoltrati ai vari centri funzionali. Nell'immediato, le funzioni principali della rete osservativa sono il monitoraggio ed il supporto alla fase previsionale; i dati raccolti in tempo reale dalle stazioni vengono utilizzati per varie finalità: valutare le condizioni meteo-idro attuali, implementare le catene modellistiche per le previsioni a breve e medio termine, supportare le attività di protezione civile, etc.

Ma i dati raccolti sono anche utilizzati per il supporto di studi meteorologici ed agro-meteorologici, per la valutazione del bilancio idrico, per la pianificazione, gestione e tutela delle risorse idriche, per la protezione e la difesa del territorio e della fascia costiera, per gli studi quantitativi a supporto della gestione della qualità delle acque superficiali (deflussi minimi vitali), ecc.

La rete di monitoraggio in tempo reale dell'Emilia-Romagna è costituita da oltre 480 stazioni, a cui si aggiungono più di 100 stazioni automatiche con registrazioni di dati e stazioni meccaniche.

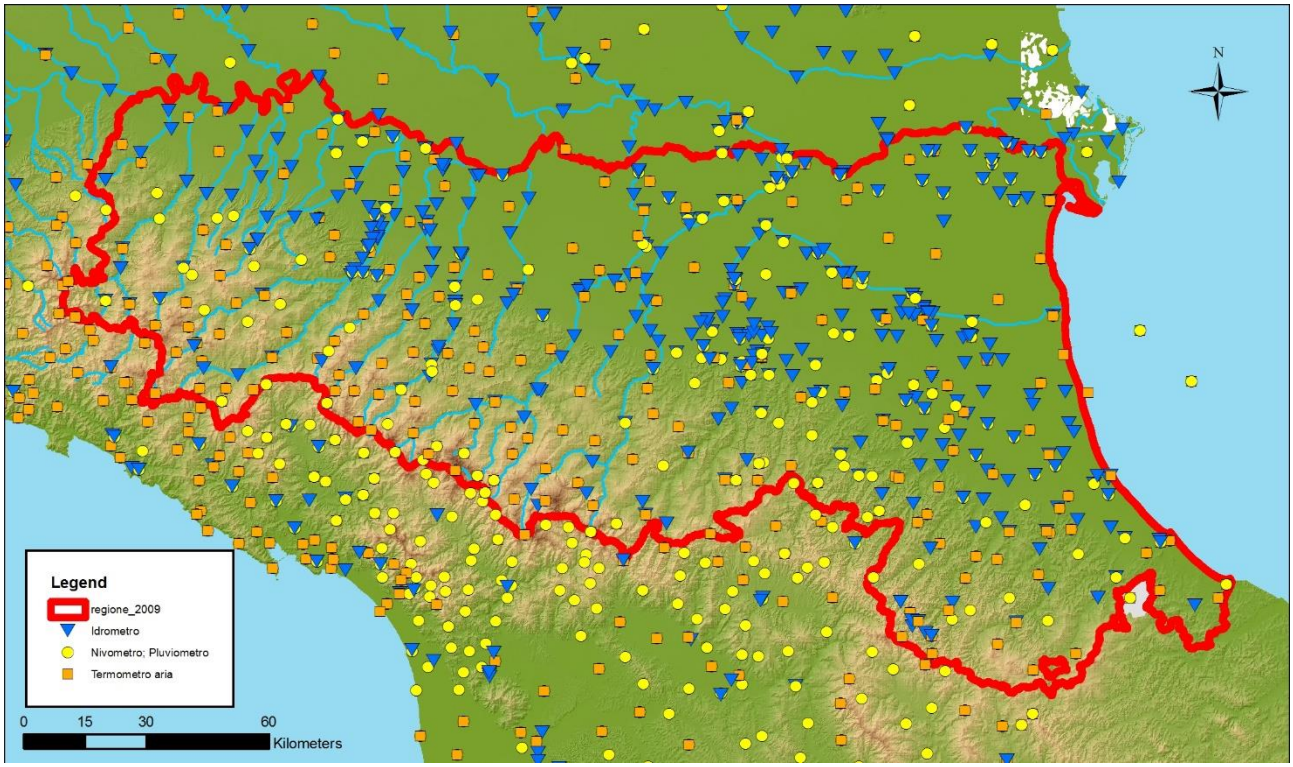


Figura 15 Punti di monitoraggio della rete dell'Emilia-Romagna
Fonte: ARPAE Emilia-Romagna

3.3 SISTEMA MODELLISTICO

Un sistema di monitoraggio, allerta e previsione denominato *FEWS* (*Flood Early Warning System*) è stato implementato nell'area idrologia e idrografia del Servizio Idro-Meteo-Clima (SIMC) di ARPAE Emilia-

Romagna: il Sistema è stato sviluppato non solo per il territorio regionale dell'Emilia-Romagna ma anche per l'intero bacino del Fiume Po (*FEWS-Po*).

Questo sistema è nato con l'intento principale di modellare e prevedere le inondazioni fluviali ed è integrato con i sistemi di previsione dei Centri Funzionali regionali.

FEWS-Po è una piattaforma informatica in grado di gestire le esecuzioni di diverse catene idrologico-idrauliche alimentate da campi di precipitazione forniti dalla rete di monitoraggio in tempo reale e da modelli meteorologici; collegato ai sistemi di acquisizione dati in tempo reale, consente la validazione dei parametri e la gestione ed esecuzione di modelli numerici relativi a tutti i processi fisici legati alle piene fluviali. Il Sistema consente, infine, la condivisione dei risultati sia all'interno della piattaforma sia online.

I risultati prodotti sono generalmente idrogrammi di piena riferiti alle sezioni idrometriche di interesse; ma anche rappresentazioni grafiche più avanzate, utili per facilitare la comprensione anche all'utente meno esperto.

L'interfaccia di sistema è basata su un'interfaccia GIS, quindi, il Sistema consente la visualizzazione in tempo reale di dati georeferenziati, derivanti da sensori in telemetria (pluviometri, termometri, idrometri). Tutti i dati, sia puntuali sia spaziali, vengono memorizzati nel sistema; le mappe di pioggia e temperatura sono utilizzate come input dai modelli idrologici.

L'input del sistema *FEWS-PO*, per fini idrologici-idraulici, è costituito da:

- precipitazioni osservate:
- precipitazioni previste fornite dai modelli di previsione meteorologica:
  - Corsa[2] operativa deterministica del modello meteorologico, con un ciclo di assimilazione rapida, COSMO RUC (risoluzione di 2,2 km)
  - Corsa operativa deterministica del modello meteorologico COSMO-I2 (risoluzione di 2,2 km)
  - Corsa operativa deterministica del modello meteorologico COSMO-I5 (risoluzione di 5 km)
  - Corsa di insieme del sistema COSMO-LEPS (risoluzione di 7 km)
  - Esecuzione deterministica del modello ECMWF del Centro europeo di lettura (risoluzione 16 km)

Gli output prodotti dal Sistema FEWS-Po sono una serie di previsioni generate dalla combinazione di tre diverse catene modellistiche, composte dai modelli idrologici *HEC-HMS*, *Mike 11-NAM* e *TOPKAPI*; a questi modelli idrologici, i modelli idraulici *HEC-RAS*, *Mike 11-HD* e *Sobek* sono associati per la simulazione della propagazione della portata nel letto del fiume.

[2] vedi nota paragrafo 1.7.1

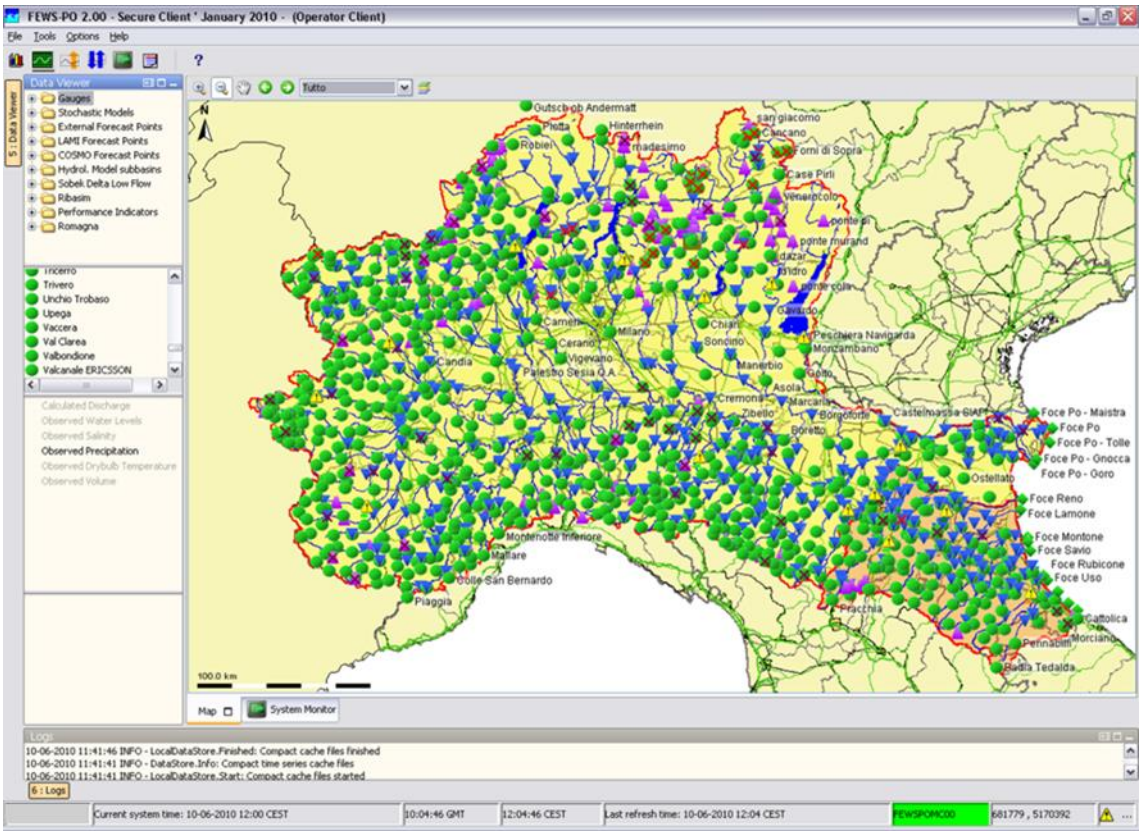


Figura 16 Interfaccia grafica del Sistema *FEWS-Po*
Fonte: ARPAE Emilia-Romagna

Nel dettaglio, per il fiume Po e i relativi affluenti e per il fiume Reno ed i corsi d'acqua della Romagna, il Sistema FEWS genera informazioni previsionali con una frequenza variabile tra le 3 e le 24 ore. I prodotti che periodicamente vengono forniti sono:

- 1 previsione deterministica idrologica a brevissimo termine (+ 12-18 ore)
- 1 previsione deterministica idrologica a breve termine (+ 36-40 ore)
- previsioni idrologiche deterministiche a medio termine (+ 60-64 ore)
- 63 previsioni idrologiche ensemble a lungo termine (+ 108-112 ore)

Nel medio-lungo periodo, l'incertezza della previsione del livello idrometrico, principalmente condizionata dall'incertezza delle previsioni meteorologiche, cresce con l'aumentare dell'arco temporale previsionale e con la diminuzione della superficie del bacino.

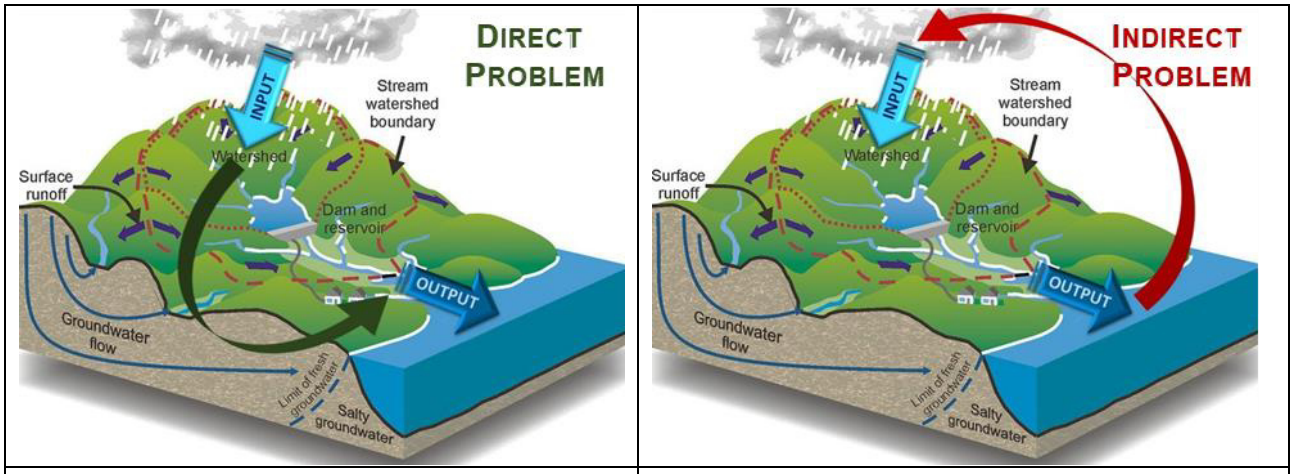
Anche nel breve periodo, per i bacini minori come quelli dell'Emilia-Romagna caratterizzati da tempi di risposta di poche ore, l'incertezza locale delle previsioni meteorologiche influenza in modo significativo le previsioni di livello idrometrico.

In questi casi, i dati pluviometrici e idrometrici acquisiti da una rete di telemisura ad alta frequenza diventano fondamentali per valutare lo stato dell'evento, per studiare le condizioni del bacino e per elaborare possibili previsioni e/o interventi.

Il sistema FEWS-Po è operativo a scopo previsionale 365 giorni all'anno, 24 ore al giorno. Può anche essere utilizzato in una configurazione "stand alone" per la pianificazione, la simulazione o la ricostruzione di eventi. (Brian, et al., 2018).

3.4 CASO DI STUDIO

Per il raggiungimento dell'obiettivo iniziale (sviluppare un modello basato sui dati osservati per poter disporre di previsioni idrometriche valide, immediate e continuamente aggiornate soprattutto durante gli eventi più estremi) si è cercato di risolvere il problema inverso dell'idrologia: identificare il volume di pioggia che ha generato un già definito livello idrometrico.



Problema diretto: dalle precipitazioni osservate alla simulazione dei livelli idrometrici.	**Problema indiretto**: dalle portate osservate alla simulazione delle precipitazioni.

Figura 17 Rappresentazione concettuale del problema diretto (sinistra) ed inverso (destra)

Uno degli obiettivi più frequenti in idrologia è identificare il livello idrometrico prodotto da una precipitazione osservata; nei casi più semplici si cerca di parametrizzare tutti i processi che legano l'input (afflusso) all'output (deflusso), mentre, nei casi più complessi si procede ad un'analisi più approfondita che termina in una serie di complesse equazioni matematiche. Partendo da uno ietogramma osservato si cerca di riprodurre un'idrogramma associato e le sue principali caratteristiche: la fase di crescita, il colmo e la fase di scarico; al termine, per un dato ietogramma, il modello genererà come output un unico e solo idrogramma di piena.

Invece, se l'obiettivo è identificare lo ietogramma che ha prodotto un dato livello idrometrico, come il colmo di piena, (problema inverso) è bene ricordarsi che quest'ultimo può essere generato da diversi ietogrammi e quindi la risoluzione di tale problema è sicuramente più complessa rispetto a quella del problema diretto.

In merito a questo specifico caso studio, si è cercato di sviluppare un modello che data una sezione idrometrica a cui sono associate le tre soglie di allertamento di Protezione Civile, basandosi sulle precipitazioni osservate, fosse in grado in corso di evento di identificare le precipitazioni che avrebbero portato nell'immediato al superamento delle soglie di allertamento; prediligendo l'elaborazione di un'informazione tempestiva e con il minor numero di mancati allarmi (condizione più gravosa per le finalità di questo caso studio).

Un'approfondita ricerca bibliografica in riferimento alla risoluzione del problema inverso per fini previsionali e di allertamento ha preceduto un'analisi dettagliata delle caratteristiche idrologiche dei principali bacini dell'Emilia-Romagna. Basandosi sulle consistenti informazioni e sulle lunghe serie storiche di osservazioni disponibili è stato possibile implementare un modello speditivo basato sul Modello di Nash, il NASP – NAsh Speditivo (Biondi & Versace, 2007), già in uso nei Centri Funzionali; e successivamente implementare un

modello previsionale con una delle più avanzate tecniche di apprendimento automatico, il metodo Random Forests.

Di seguito verranno illustrati i passaggi che hanno portato all'elaborazione del Modello di Nash e successivamente, con maggior dettaglio, i passaggi compiuti per l'elaborazione del modello previsionale basato sul *Random Forests*.

3.4.1 ANALISI DEGLI EVENTI DI PIENA ED IMPLEMENTAZIONE DI UN MODELLO "CLASSICO" SPEDITIVO

Nel 2004, a seguito della Direttiva del Presidente del Consiglio dei Ministri Dir. PCM 27/02/2004, per le principali sezioni idrometriche sono state identificate tre soglie idrometriche di allertamento; le soglie sono caratteristiche della sezione a cui si riferiscono ed ad ognuna di esse sono stati associati degli interventi di protezione civile.

- Sotto la soglia 1: nessun evento di inondazione. Livelli idrometrici regolari.
- Sopra la soglia 1: evento di inondazione non significativo. Livelli idrometrici corrispondenti al riempimento del letto del fiume poco profondo e generalmente al di sotto del livello naturale del terreno.
- Sopra la soglia 2: evento di piena con erosione limitata e fenomeni di trasporto. Livelli idrometrici corrispondenti all'occupazione delle aree alluvionali o all'espansione del corso d'acqua, con il coinvolgimento delle banche. Il livello idrometrico può superare il livello naturale delle aree circostanti.
- Sopra la soglia 3: evento alluvionale significativo con fenomeni di erosione e trasporto diffusi. Livelli idrometrici corrispondenti all'occupazione dell'intera sezione fluviale, prossimi ai valori massimi del margine di sicurezza della sponda del fiume (bordo libero).

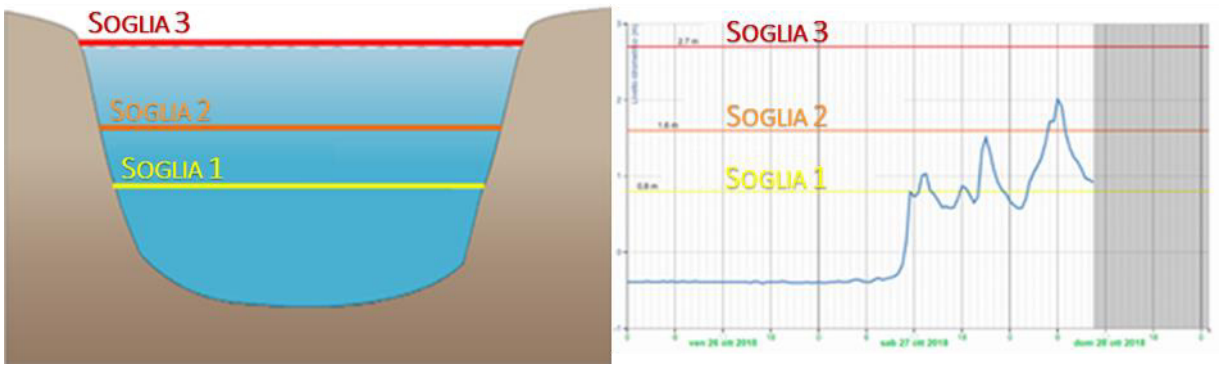


Figura 18 Sinistra: rappresentazione grafica delle soglie di allertamento
Destra: idrogramma e soglie di allertamento.
Fonte: Arpae Emilia-Romagna

Ad ogni soglia idrometrica, espressa in metri, è possibile associare un valore di portata utilizzando la scala di deflusso; per ogni sezione, sono stati identificati gli eventi piena con un colmo pari o superiore alla rispettiva soglia 1.

Selezionati gli eventi di piena, è necessario identificare l'evento pluviometrico che ha dato origine all'evento di piena; quindi, definire l'intero evento pluvio-idrometrico.

Come verrà mostrato successivamente, solo una parte delle precipitazioni contribuisce al deflusso; con tempi differenti in funzione delle caratteristiche del bacino e del grado di saturazione. Generalmente, ogni sezione ha un tempo di risposta caratteristico, ovvero, il tempo che intercorre tra lo scroscio di pioggia più intenso ed il passaggio del colmo di piena, da esso generato, nella sezione di interesse; il tempo di risposta è un parametro fondamentale in quanto, una volta definito, permette di conoscere indicativamente l'arco temporale che si ha a disposizione tra lo scroscio e il passaggio dell'onda di piena. Per massimizzare gli effetti, tutti gli interventi operativi dovranno esser compiuti all'interno della finestra temporale identificata.

In riferimento al bacino pilota del fiume Parma, con l'obiettivo di identificare il tempo di risposta si è assunto inizialmente che questo sia coincidente con il tempo di corrivazione calcolato con l'equazione di Giandotti:

$$T_c = \frac{4\sqrt{A} + 1.5 * L}{0.8\sqrt{H}}$$

T_c= tempo di corrivazione [h]
A= estensione bacino [km^2]
L= lunghezza canale principale [km]
H= altitudine media bacino [m]

Applicando l'equazione per il fiume Parma, alla sezione principale Ponte Verdi e per il torrente Baganza, alla sezione principale Ponte Nuovo, si ottengono i seguenti tempi di corrivazione:

Fiume Parma – Ponte Verdi	
A= 600 km^2 L= 62.83 km H= 646 m	$\boldsymbol{T_c = 9.5\ h}$

Torrente Baganza– Ponte Nuovo	
A= 177 km2 L= 58.51 km H= 685.24 m	$\boldsymbol{T_c = 6.9\ h}$

Si è fatto quindi riferimento a queste tempistiche per la selezione degli eventi pluvio-idrometrici e la ricostruzione degli eventi di piena; nell'immagine seguente è stato riprodotto concettualmente un evento pluvio-idrometrico.

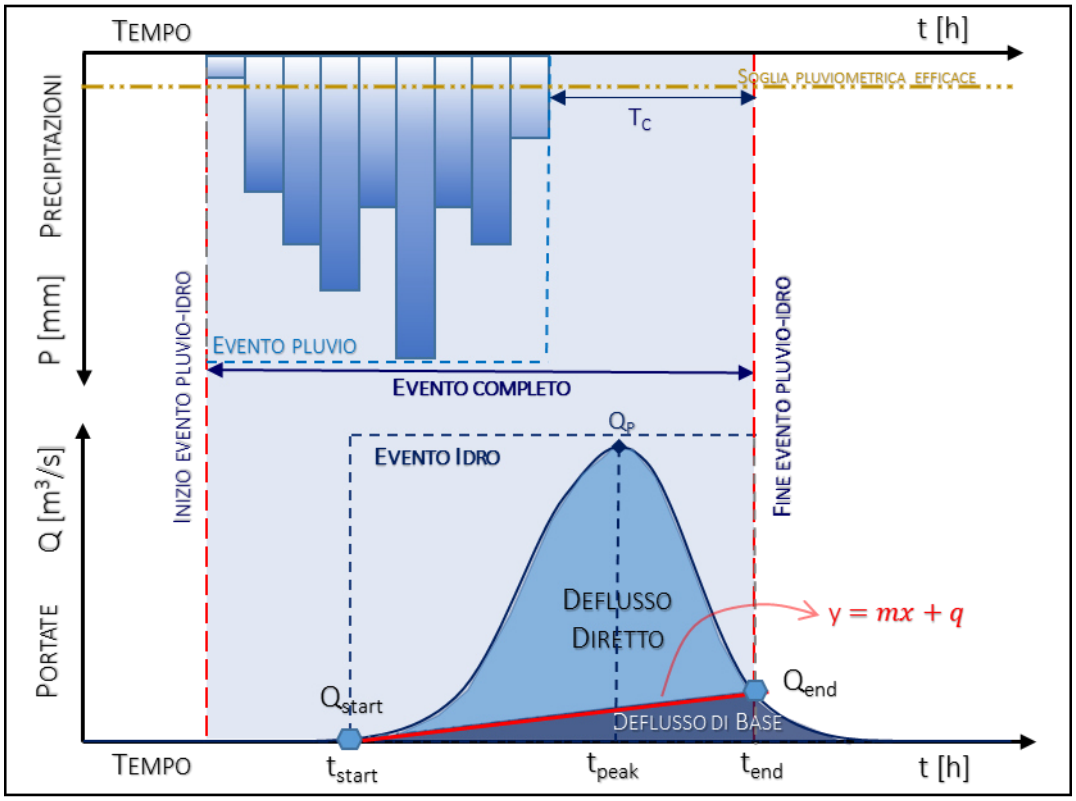


Figura 19 Rappresentazione concettuale di un evento pluvio-idrometrico: identificazione dell'evento pluviometrico e separazione dell'idrogramma di piena dalla portata di base.
Q_p=colmpo di piena; T_c=tempo di corrivazione.

L'inizio dell'evento pluvio-idrometrico coincide con l'inizio delle precipitazioni, mentre la fine dell'evento corrisponde alla fine delle precipitazioni più il tempo di corrivazione della sezione considerata. Le precipitazioni con intensità inferiore a 1 mm/h non sono state considerate in quest'analisi: il loro contributo è trascurabile rispetto al deflusso totale.

Per ogni evento identificato, il volume di piena è stato separato dalla portata di base utilizzando la seguente equazione lineare:

$$y = mx + q$$

$$m = \frac{Q_{finale} - Q_{iniziale}}{T_{finale} - T_{iniziale}}$$

$$q = \frac{(t_{finale} * Q_{iniziale}) - (t_{inziale} * Q_{finale})}{t_{finale} - t_{iniziale}}$$

Per ogni passo temporale (orario, in questa prima analisi) è stato possibile identificare la portata di piena:

$$Q_{piena}{}_{t_i} = Q_{piena,obs}{}_{t_i} - Q_{piena,base}{}_{t_i}$$

Nota la portata, in funzione del tempo è possibile calcolare il volume di piena per ogni passo temporale:

$$V_{piena} = \sum_{i=t_{inizio}}^{t_{fine}} Q_{piena,i} * \Delta t \qquad dove \qquad \Delta t = 1h$$

L'analisi di un evento di piena richiede la conoscenza sia del massimo valore di portata transitato (intensità dell'evento) sia dei volumi transitati (consistenza dell'evento); la conoscenza di questi due elementi consente di avere una buona visione d'insieme dell'evento già dall'idrogramma.

Ad esempio, dati due idrogrammi della stessa sezione, relativi a due eventi di piena differenti, caratterizzati dalle medesime condizioni iniziali del bacino e dal medesimo valore di picco (Q_{max}), ma con una marcata differenza di volumi, si può dedurre che l'evento avente volumi inferiori è stato generato da precipitazioni di durata minore ma con maggior intensità; mentre, l'idrogramma caratterizzato da un volume di piena maggiore, molto probabilmente sarà stato generato da precipitazioni più modeste ma continue e di maggior durata.

Partendo dalle portate e dai volumi di piena è possibile procedere all'analisi-ricostruzione degli eventi di piena.

Per questo caso studio, la ricostruzione degli eventi di piena è stata condotta utilizzando il Modello Nash accoppiato ad un modello di rifiuto SCS-CN (Soil Conservation Service Engineering Division, 1972), necessario per identificare i volumi di pioggia che effettivamente contribuiscono alla formazione dell'evento di piena.

Analizzando i principali eventi di piena, per ogni sezione, sono stati definiti i parametri caratteristici del modello di Nash: *n* (parametro di forma) e *k* (parametro di scala). (Bahremand, 2009)

Per ricavare i due parametri, è necessario prima calcolare il massimo valore di portata dell'evento ($Q_{piena,\ max}$) ed il relativo volume ($V_{piena,\ max}$), in funzione del *time step* considerato; quindi, il tempo di picco ($t_{piena,\ picco}$), ovvero, l'istante in cui è stato registrato il massimo volume rispetto l'inizio dell'evento.

Colmo di piena $$Q_{piena,max_{t_p}} = Q_{piena,obs_{t_p}} - Q_{base,t_p}$$

Volume colmo di piena $$V_{piena,max_{t_p}} = Q_{piena,max_{t_p}} * \Delta t \quad \text{con } \Delta t = 1h$$

Tempo di picco $$t_{piena,picco} = t_{Q_{piena,max_{t_p}}} - t_{inizio}$$

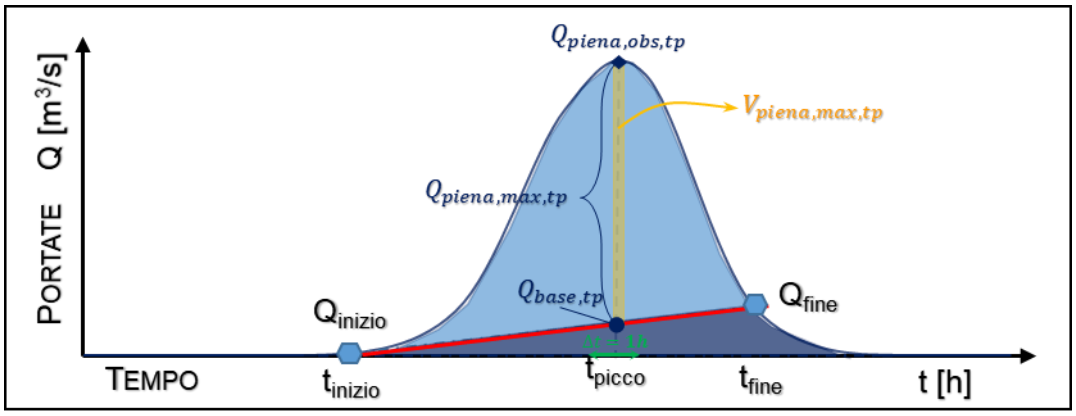


Figura 20 Rappresentazione grafica di un evento di piena e separazione del volume di piena dal volume di base

In riferimento all'intero evento di piena, il volume totale è dato dalla sommatoria dei volumi discretizzati:

$$V_{piena,tot} = \sum_{i=t_{inzio}}^{t_{fine}} V_{piena,i}$$

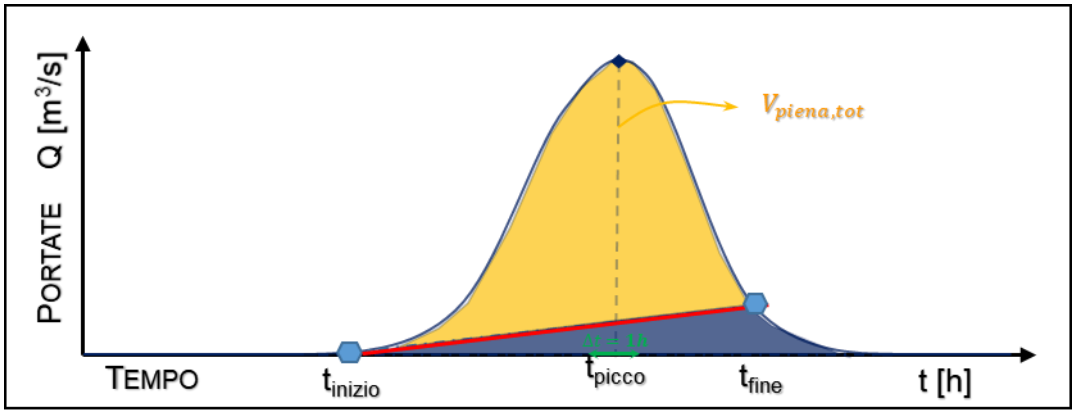


Figura 21 Rappresentazione grafica del volume di piena (area gialla) separato dal volume di base (area blu)

Dal rapporto tra il massimo volume di piena ($V_{piena,\ max}$) e il volume totale dell'evento di piena ($V_{piena,tot}$), è possibile calcolare la portata adimensionale (q_{piena}); necessaria per ricavare, successivamente, I tre parametri β, n, k.

$$q_{piena} = \frac{V_{piena,max}}{V_{piena,tot}}$$

In riferimento a (Bhunya et al., 2005, 2004) i parametri β, n, k si ottengono dalle seguenti relazioni:

$$\beta = q_{piena} * t_{piena,picco}$$

$$n(\beta) = \begin{cases} se\ 0.01 < \beta < 0.35 \rightarrow n = 5.53 * \beta^{1.75} + 1.04 \\ se\ \beta \geq 0.35\ \rightarrow\ n = 6.29 * \beta^{1.998} + 1.157 \end{cases}$$

$$k = \frac{t_{picco}}{n-1}$$

Noti ora *n* e *k*, è possibile ricavare il tempo di risposta ($t_{risposta}$), come prodotto tra i due parametri:

$$t_{risposta} = n * k$$

Le equazioni sopra riportate sono state applicate per ogni singolo evento di piena, così da definire, in funzione della precipitazione totale, i valori *n*, *k* e $t_{risposta}$.

Noti i valori di *n*, *k* e $Q_{piena,\ max}$ in funzione delle differenti finestre temporali (dall'inizio delle precipitazioni fino un'ora prima –ultimo *time step*- del picco di piena), utilizzando l'equazione proposta da Rigon (Rigon et al., 2011) è stato calcolato il tempo di picco di Nash:

$$t_{picco,Nash} = \frac{t_{step}}{1 - \left[e^{(-\frac{t_{step}}{k})}\right]^{\frac{1}{n-1}}}$$

Il tempo di preannuncio è dato dalla differenza tra il tempo di picco di Nash e il tempo di picco di piena:

$$t_{\text{preannuncio}} = t_{picco,Nash} - t_{picco,piena}$$

Quindi, è possibile calcolare il seguente integrale, noto come *S-Curve* o *S-Hydrograph* (Dooge, 1973):

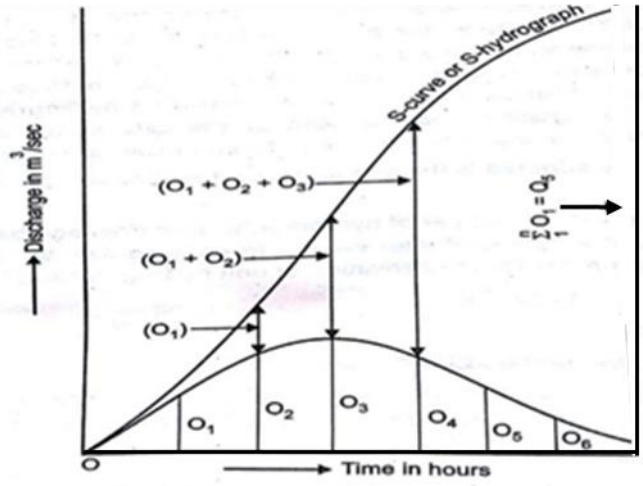


Figura 22 Rappresentazione grafica dell'integrale S-Curve

In riferimento a: $t_{picco,Nash}$ $$S(t_p) = \Gamma(\frac{t_{picco,Nash}}{k}; n)$$

In riferimento alla finestra temporale considerata: $$S(t_p - t) \equiv S\left[\frac{t_p - t}{k}\right] = \Gamma(\frac{t_{preannuncio}}{k}; n)$$

Il passo seguente, consiste nel calcolare l'intensità massima della pioggia netta ($i_{max,calc}$), nel passo temporale di riferimento; il valore viene calcolato per tutti i passi temporali considerati.

$$i_{max,calc} = \frac{Q_{piena,max}}{A * \left[\frac{S(t_p) - S(t_p - t)}{3.6}\right]}$$

La massima intensità di precipitazione netta ($i_{max,calc}$) calcolata nel passo temporale di riferimento (in questo caso un'ora, quindi l'afflusso massimo orario) è anche definita "critica", in quanto generatrice del massimo valore di portata dell'evento di piena ($Q_{piena,\ max}$).

Successivamente, si confrontano i valori di massima intensità osservata e massima intensità calcolata per definire il minimo tempo di preannuncio:

$$\min(t_{preannuncio}) = i_{max,calc}(t) \geq i_{max,obs}(t)$$

La precipitazione totale critica è calcolata come il prodotto tra l'intensità massima e il tempo di preannuncio e rappresenta il volume d'afflusso necessario per generare il volume di piena considerato:

$$P_{tot} = i_{max,calc} \times t_{preannuncio}$$

Per ogni evento considerato, avendo ora calcolato P_{tot}, *n* e *k* è stato possibile identificare una relazione lineare che lega la pioggia totale (P_{tot}) al tempo di risposta ($t_{risposta} = n * k$) e successivamente una relazione che lega n e k, caratteristica per ogni sezione.

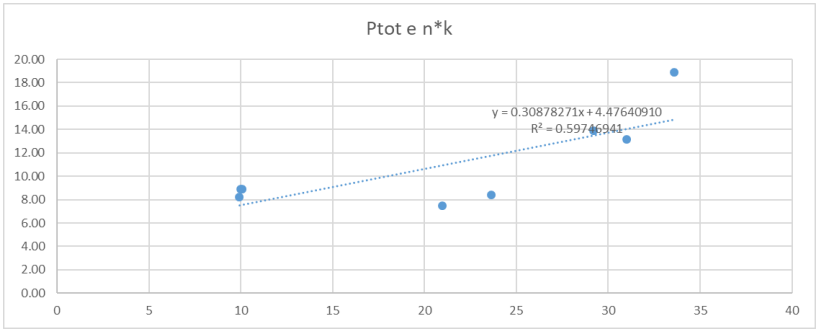


Figura 23 Relazione lineare tra Precipitazione totale e tempo di risposta

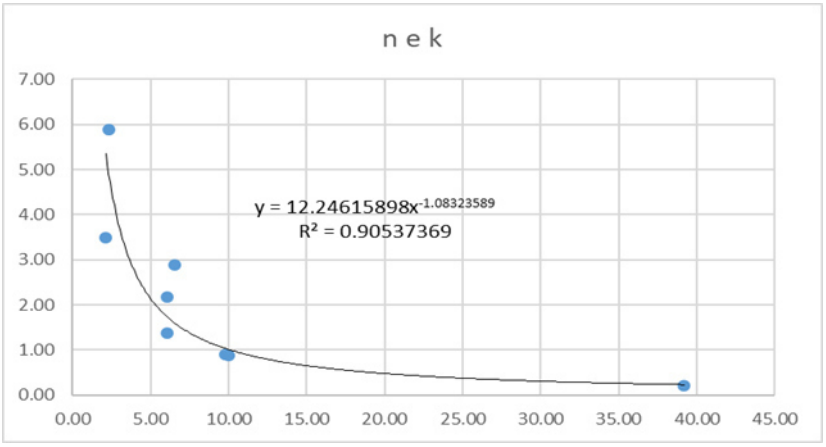


Figura 24 Relazione tra *n* (ascissa) e *k* (ordinata)

Per ogni passo temporale in funzione della finestra mobile considerata, identificata con il tempo di risposta caratteristica di ogni sezione, è stata calcolato l'afflusso minimo; il volume di afflusso superiore a questa soglia, contribuisce direttamente al deflusso, ovvero, rappresenta la pioggia netta. Ciò consente, indirettamente, di valutare "in continuo" le condizioni di saturazione del bacino.

In corso di evento, quindi, è possibile identificare e separare dalla precipitazione, la pioggia netta; la quale contribuirà al deflusso superficiale noti i parametri *n* e *k* della sezione considerata. Utilizzando queste informazioni, in relazione alla precipitazione osservata, è possibile simulare il valore di portata atteso entro le prossime t_{ore}; dove *t* è un arco temporale pari al tempo di preannuncio.

Il Modello di Nash, inizialmente calibrato ed applicato al bacino del Fiume Parma, è stato successivamente calibrato ed applicato a tutte le principali sezioni dei fiumi dell'Emilia-Romagna. Come vedremo più avanti, le informazioni ottenute (caratteristiche dei bacini idrografici considerati) in questa prima fase hanno rappresentato lo "studio preliminare" per la seconda parte del caso studio relativa all'applicazione del metodo *Random Forests*; i risultati ottenuti dal Modello di Nash verranno presentati a confronto con quelli ottenuti con il metodo *Random Forests*.

3.4.2 DATA SET DI CALIBRAZIONE

Tutte le principali sezioni dei fiumi dell'Emilia-Romagna dispongono di serie storiche, di portata e precipitazione, dal 2005; per questo caso studio è stato considerato un periodo di dieci anni, 2005-2015. Il data set di calibrazione è il medesimo sia per il modello speditivo di Nash sia per il modello basato sul metodo *Random Forests*; condizione necessaria per procedere con il confronto dei risultati e la verifica delle performance.

Il modello speditivo Nash è caratterizzato da un processo di calibrazione manuale, come illustrato nel paragrafo precedente, e il data set utilizzato ha lo scopo di verificare le performance del modello prima dell'eventuale fase operativa (processo di validazione).

Il modello basato sul metodo *Random Forests*, invece, è caratterizzato da un processo di calibrazione automatica: punto forte delle tecniche di autoapprendimento, come visto nel capitolo 2. L'unica operazione compiuta dall'operatore è la "preparazione strutturale" del data set necessaria, prima, per la calibrazione del modello e, poi, per l'operatività. Per ogni passo orario (*time step*) viene generata in automatico un'istanza

(riga di dati) contenente le piogge orarie osservate spazializzate e le portate osservate nelle N-ore precedenti; il parametro N rappresenta la finestra temporale caratteristica di ogni sezione e pari al tempo di risposta del bacino chiuso alla data sezione (come presentato nel precedente paragrafo). Inoltre, solo in fase di calibrazione, ad ogni istanza del data set è stata aggiunta un'ulteriore informazione: la massima condizione di allerta idrometrica osservata nelle N-ore successive al "tempo zero" t_0.

In fase di calibrazione, per ogni passo temporale, l'istanza sarà composta da "cause" (precipitazioni e portate) ed "effetto" (massima soglia idrometrica); il *Random Forests* apprenderà "causa-effetto" attraverso questo *training set* mentre in fase operativa, essendo presenti nel data set solo le "cause", produrrà in previsione gli "effetti".

3.4.3 IMPLEMENTAZIONE DEL METODO *RANDOM FORESTS*

Di seguito verranno illustrati i principali passaggi necessari per l'implementazione del metodo *Random Forests* all'interno del modello previsionale; inoltre, verranno riportate le criticità riscontrate in fase di elaborazione e le scelte adottate per superarle.

3.4.3.1 PREPAZIONE DEL DATA SET DI ADDESTRAMENTO ("*TRAINING SET*")

Il data set di addestramento (*training set*) è sicuramente l'elemento principale del metodo *Random Forests*; poiché essendo basato su un algoritmo ad apprendimento automatico, la qualità del *training set* influisce direttamente ed incisivamente nelle performance del modello. Nel capitolo 2 sono stati rapidamente presentati alcuni metodi per sopperire, in parte, a data set non eterogenei o caratterizzati da dati mancati; in questo caso studio, tutte le serie di dati utilizzate sono state preventivamente validate e non presentano valori particolarmente alterati o mancanti. Condizione necessaria per massimizzare l'utilizzo dei dati disponibili.

Il primo passo da compiere per l'implementazione del metodo *Random Forests* è la preparazione del *training set*: la serie di dati disponibile à stata commutata in matrice dove ogni riga rappresenta un'istanza.

Esempio di un'istanza di addestramento: l'istanza relativa al tempo t_0 è composta dalle osservazioni di precipitazione e portata precedenti N_{ore} a t_0 e dalla massima soglia idrometrica osservata nelle N_{ore} successive a t_0.

TIME	DATI OSSERVATI: PRECIPITAZIONE E PORTATA							DATI OSSERVATI
t_0	P_{-Nh}	P_{-5h}	P_{-4h}	P_{-3h}	P_{-2h}	P_{-1h}	P_{-0h}	MASSIMA SOGLIA IDROMETRICA
	Q_{-Nh}	Q_{-5h}	Q_{-4h}	Q_{-3h}	Q_{-2h}	Q_{-1h}	Q_{-0h}	

Tabella 2 Esempio di un'istanza del *training set*

Esempio di un'istanza di validazione o operativa: l'istanza relativa al tempo t_0 è composta dalle osservazioni di precipitazione e portata precedenti N_{ore} a t_0. La massima soglia idrometrica nelle N_{ore} successive a t_0 è l'informazione generata dal modello (output del Modello).

TIME	DATI OSSERVATI: PRECIPITAZIONE E PORTATA	DATI ELABORATI DAL *RANDOM FORESTS*

t_0	P_{-Nh}	P_{-5h}	P_{-4h}	P_{-3h}	P_{-2h}	P_{-1h}	P_{-0h}	MASSIMA SOGLIA IDROMETRICA
	Q_{-Nh}	Q_{-5h}	Q_{-4h}	Q_{-3h}	Q_{-2h}	Q_{-1h}	Q_{-0h}	

Tabella 3 Esempio di un'istanza di validazione o operativa

A titolo di esempio, in riferimento alla sezione principale del bacino pilota (Fiume Parma a Ponte Verdi), dall'analisi preliminare condotta per il Modello di Nash speditivo è emerso che il tempo di risposta è di 6 ore; quindi, ogni istanza del *training set* del modello *Random Forests* dovrà contenere sette coppie (precipitazione e portata) di valori e l'ultima colonna rappresenterà la massima soglia idrometrica raggiunta nelle 6 ore successive a t_0.

Data 19-20/01/2009	Precipitazione							Portata							Max Soglia +6h
	P_{oss0h}	P_{oss1h}	P_{oss2h}	P_{oss3h}	P_{oss4h}	P_{oss5h}	P_{oss6h}	Q_{oss0h}	Q_{oss1h}	Q_{oss2h}	Q_{oss3h}	Q_{oss4h}	Q_{oss5h}	Q_{oss6h}	
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
22:00	1.27	1.54	1.25	1.63	1.06	0.85	0.57	38.20	35.59	31.83	29.43	29.43	30.62	29.43	**0**
23:00	1.30	1.27	1.54	1.25	1.63	1.06	0.85	40.88	38.20	35.59	31.83	29.43	29.43	30.62	**0**
00:00	3.55	1.30	1.27	1.54	1.25	1.63	1.06	46.49	40.88	38.20	35.59	31.83	29.43	29.43	**0**
01:00	2.30	3.55	1.30	1.27	1.54	1.25	1.63	52.40	46.49	40.88	38.20	35.59	31.83	29.43	**0**
02:00	1.72	2.30	3.55	1.30	1.27	1.54	1.25	65.13	52.40	46.49	40.88	38.20	35.59	31.83	**1**
03:00	4.19	1.72	2.30	3.55	1.30	1.27	1.54	88.30	65.13	52.40	46.49	40.88	38.20	35.59	**2**
04:00	2.82	4.19	1.72	2.30	3.55	1.30	1.27	98.00	88.30	65.13	52.40	46.49	40.88	38.20	**2**
05:00	2.21	2.82	4.19	1.72	2.30	3.55	1.30	104.03	98.00	88.30	65.13	52.40	46.49	40.88	**2**
06:00	3.04	2.21	2.82	4.19	1.72	2.30	3.55	155.33	104.03	98.00	88.30	65.13	52.40	46.49	**2**
07:00	3.11	3.04	2.21	2.82	4.19	1.72	2.30	160.20	155.33	104.03	98.00	88.30	65.13	52.40	**2**
08:00	4.59	3.11	3.04	2.21	2.82	4.19	1.72	218.01	160.20	155.33	104.03	98.00	88.30	65.13	**3**
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…

Tabella 4 Porzione del training set utilizzato per la sezione Ponte Verdi del fiume Parma

L'elaborazione del *training set* è stata automatizzata all'interno dell'ambiente R; per poter esser successivamente utilizzata dal pacchetto "*randomForest*" in R.

3.4.3.2 NUMERO DI ALBERI GENERATI PER LA FORESTA CASUALE

Nel paragrafo 2.4.5 sono stati presentati i tre parametri principali per settare e migliorare, se possibile, il metodo Random Forests:

- *m*, il numero di variabili predittive selezionate casualmente su ciascun nodo
- *J*, il numero di alberi che compongono la foresta casuale
- *tree size*, dimensione dell'albero calcolata in funzione del numero di "split" presenti nel nodo più vicino al nodo radice; oppure, calcolata in funzione del massimo numero di nodi terminali.

In riferimento a questo specifico caso studio e seguendo quanto suggerito in bibliografia, sono stati effettuati dei test di verifica, volti a migliorare le performance del modello, solo sul numero di alberi (parametro *J*); come verrà presentato successivamente, il parametro *J* è generalmente il più influente, mentre *m* e "*tree size*" rivestono un ruolo importante in presenza di un'elevata numerosità di variabili e di dataset fortemente eterogenei che comportano la produzione di molte "foglie" (nodi terminali).

Il metodo Random Forests in ambiente R è parametrizzato di default con un numero di alberi pari a 500 ($J = 500$); ma con l'obiettivo di verificare l'impatto sulle performance e sul tempo di calcolo, sono stati effettuati alcuni test variando il numero di alberi.

PRIMO TEST [J = 500]

Il primo test realizzato con il parametro $J = 500$ (numero di alberi = 500), presenta una stima del tasso di errore generale (*OOB estimate of error rate*) di 0.60%; ovvero, classifica correttamente il 99,40% dei valori.

Numero di alberi *Number of trees*	500	Stima del tasso di errore *OOB estimate of error rate*	0.60%

	Confusion matrix				
	Soglia 0	Soglia 1	Soglia 2	Soglia 3	Errore di classificazione (%) *Class.error (%)*
Soglia 0	75659	74	2	0	0.0010
Soglia 1	204	275	39	0	0.4691
Soglia 2	41	75	150	2	0.4402
Soglia 3	5	0	15	16	0.5555

Tabella 5 *Confusion matrix* ottenuta con J = 500

Facendo riferimento alla "*Confusion matrix*", generabile attraverso il pacchetto "*randomForest*" in ambiente R:

	Classificazioni corrette (in verde nella *confusion matrix*)	Classificazioni errate
Soglia 0	75659	76
Soglia 1	275	243
Soglia 2	150	118
Soglia 3	16	20
TOTALE	76100 (99.40%)	457 (0.60%)

Tabella 6 Valutazione classificazione con J=500

In fase di training, generalmente il numero di istanze e il tasso di errore di classificazione sono inversamente proporzionali; nel nostro caso, lo score peggiore è risultato essere quello della soglia 3, affetta dal minor numero di istanze.

È necessario precisare che la "*Confusion matrix*" è relativa alla sola fase di *training*; pertanto, l'addestramento nella fase iniziale, i primi alberi, può presentare delle carenze (errori), ma con l'aumentare delle dimensioni della foresta (più alberi) queste carenze possono esser superate.

Il grafico sotto riportato (figura 13) mostra la "storia" dell'addestramento in funzione delle dimensioni della foresta casuale: con i primi 20-25 alberi, il tasso di errore di classificazione risulta relativamente elevato per tutte le soglie, indice che la foresta è ancora troppo piccola; superati i 50 alberi, la soglia 1 e 2 raggiungono un tasso di errore che risulterà costante con il crescere della foresta. La soglia 3, più variabile rispetto alle soglie 1 e 2, mostra un apparente miglioramento tra i 70 e i 90 alberi; mentre, tende a stabilizzarsi su un valore di errore leggermente più elevato oltre i 100 alberi.

Dal grafico si può notare che oltre i 100 alberi non vi sono miglioramenti apprezzabili per quanto riguarda le performance di classificazione di soglia 1 e soglia 2. Per soglia 3, invece, si possono dedurre due importanti considerazioni: il tasso di errore non raggiunge un valore costante, ma rimane sempre "nervoso", ciò è dovuto alle poche informazioni disponibili e spesso eterogenee.

La soglia 0 (linea rossa tratteggiata) è prossima allo 0% fin dai primi alberi generati e rimane sempre al di sotto dell'errore generale (OOB), linea nera continua; facilmente intuibile dall'algoritmo in quanto caratterizzata da istanze pari a zero (precipitazioni) e valori di portata bassi e costanti nel tempo.

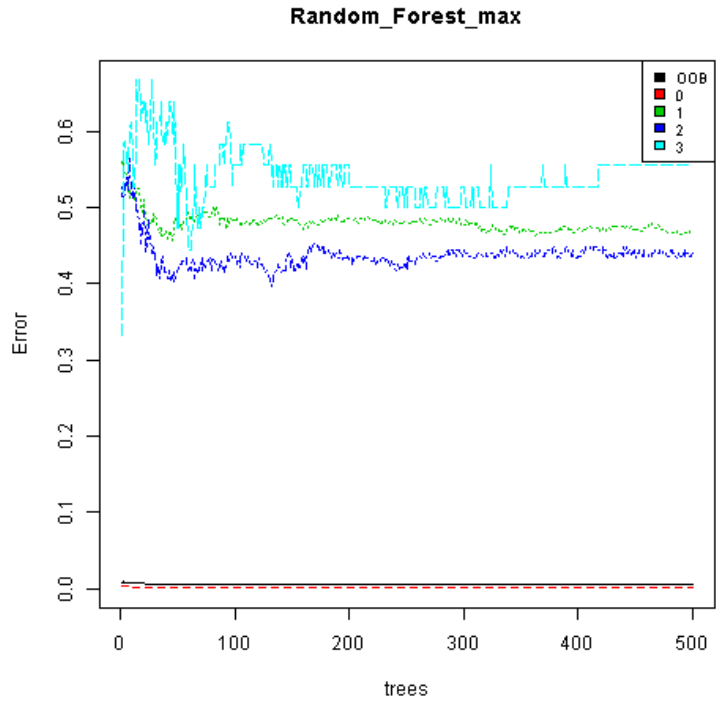


Figura 25 Grafico dell'andamento dei tassi di errore di classificazione [J= 500]

È opportuno ricordare che in questa fase si verificano e valutano solo le performance relative all'addestramento del *Random Forests*, che non sono rappresentative della bontà del modello ma aiutano a comprendere la qualità del *training set* utilizzato e l'apprendimento maturato della foresta casuale. Tanto migliori saranno le performance in questa fase, tanto migliori saranno le performance finali del modello; ma, come verrà illustrato successivamente, non è implicito che a performance non soddisfacenti in questa fase corrispondano scarse performance del modello.

SECONDO TEST [J =1000]

Il secondo test realizzato con il parametro $J = 1000$ (numero di alberi = 1000), presenta una stima del tasso di errore generale (*OOB estimate of error rate*) di 0.59%; ovvero, classifica correttamente il 99,41% dei valori.

Numero di alberi *Number of trees*	1000	Stima del tasso di errore *OOB estimate of error rate*	0.60%

	Confusion matrix				
	Soglia 0	Soglia 1	Soglia 2	Soglia 3	Errore di classificazione (%) *Class.error (%)*
Soglia 0	75669	64	2	0	0.0008
Soglia 1	202	274	42	0	0.4710
Soglia 2	41	75	150	2	0.4402
Soglia 3	5	1	12	18	0.5000

Tabella 7 *Confusion matrix* ottenuta con J = 1000

Facendo riferimento alla "*Confusion matrix*", generabile attraverso il pacchetto "*randomForest*" in ambiente R:

	Classificazioni corrette (in verde nella *confusion matrix*)	Classificazioni errate
Soglia 0	75669	66
Soglia 1	274	244
Soglia 2	150	118
Soglia 3	18	18
TOTALE	76111 (99.41%)	446 (0.59%)

Tabella 8 Valutazione classificazione con J=1000

In questo test è stato raddoppiato il numero di alberi, da 500 a 1000; dai risultati ottenuti in fase di verifica del *training* non emergono apprezzabili scostamenti rispetto al primo test: il lieve miglioramento può esser valutato come "rumore" piuttosto che un effettivo miglioramento generato dall'incremento del numero di alberi.

A supporto di questa valutazione, il grafico relativo agli andamenti dei tassi d'errore di classificazione di questo test mostra andamenti e valori molto simili a quelli emersi nel test precedente.

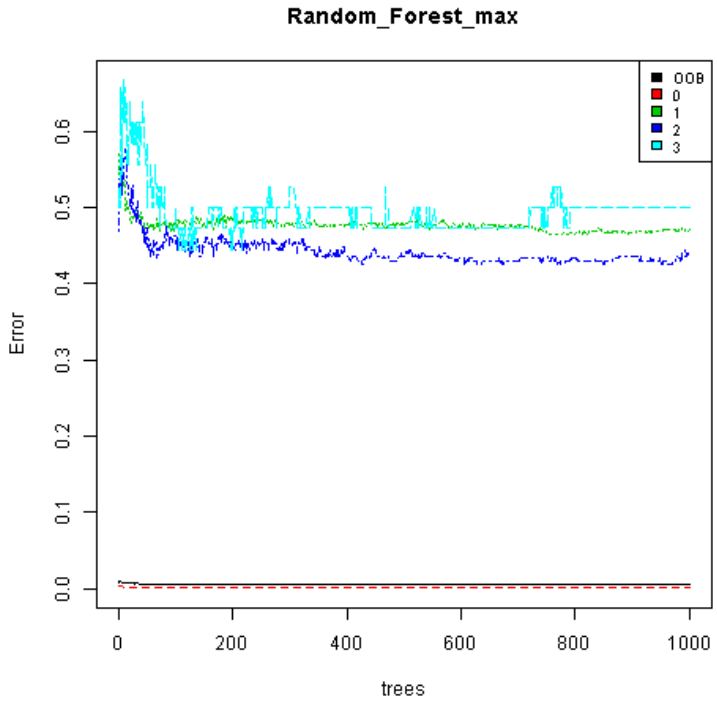


Figura 26 Grafico dell'andamento dei tassi di errore di classificazione. [J=1000]

In questo secondo test un miglioramento, seppur minimo, è comunque riscontrabile; pertanto si è ritenuto opportuno effettuare un altro test, incrementando ancora la numerosità degli alberi.

TERZO TEST [J =1250]

Il terzo test realizzato con il parametro $J = 1250$ (numero di alberi = 1250), presenta una stima del tasso di errore generale (*OOB estimate of error rate*) di 0.59%; ovvero, classifica correttamente il 99,41% dei valori.

Numero di alberi *Number of trees*	1250	Stima del tasso di errore *OOB estimate of error rate*	0.59%

	Confusion matrix				
	Soglia 0	Soglia 1	Soglia 2	Soglia 3	Errore di classificazione (%) *Class.error (%)*
Soglia 0	75666	65	4	0	0.0009
Soglia 1	201	277	40	0	0.4652
Soglia 2	43	75	148	2	0.4478
Soglia 3	6	0	12	18	0.5000

Tabella 9 *Confusion matrix* ottenuta con J = 1250

Incrementando ulteriormente il numero di alberi, da 1000 a 1250, le performance non hanno subito apprezzabili miglioramenti; anzi, c'è stato un leggero peggioramento della capacità di classificazione. Inoltre, come verrà presentato in seguito, l'incremento del numero di alberi comporta un incremento dei tempi di calcolo.

Non avendo riscontrato alcun beneficio nei test di incremento del numero di alberi; è stato condotto un ulteriore test diminuendo il numero di alberi di partenza.

QUARTO TEST [J =250]

Il quarto test realizzato con il parametro $J = 250$ (numero di alberi = 250), presenta una stima del tasso di errore generale (*OOB estimate of error rate*) di 0.58%; ovvero, classifica correttamente il 99,42% dei valori.

Numero di alberi *Number of trees*	250	Stima del tasso di errore *OOB estimate of error rate*	0.58%

	Confusion matrix				
	Soglia 0	Soglia 1	Soglia 2	Soglia 3	Errore di classificazione (%) *Class.error (%)*
Soglia 0	75667	63	5	0	0.0008
Soglia 1	200	277	41	0	0.4653
Soglia 2	42	74	150	2	0.4403
Soglia 3	6	1	10	19	0.4722

Tabella 10 *Confusion matrix* ottenuta con J = 250

Facendo riferimento alla "*Confusion matrix*", generabile attraverso il pacchetto "*randomForest*" in ambiente R:

	Classificazioni corrette (in verde nella *confusion matrix*)	Classificazioni errate
Soglia 0	75667	68
Soglia 1	277	241
Soglia 2	150	118
Soglia 3	19	17
TOTALE	76113 (99.42%)	444 (0.58%)

Tabella 11 Valutazione classificazione con J=250

Nel quarto test è stato dimezzato il numero degli alberi iniziali, da 500 a 250; le performance in fase di addestramento non hanno subito apprezzabili cambiamenti.

Come mostra il grafico relativo agli andamenti del tasso di errore di classificazione, dopo una prima fase di "stabilizzazione" (entro i primi 100 alberi), soglia 1 e soglia 2 raggiungo un valore di errore stabile; mentre persistono le condizioni di "rumore" nella soglia 3.

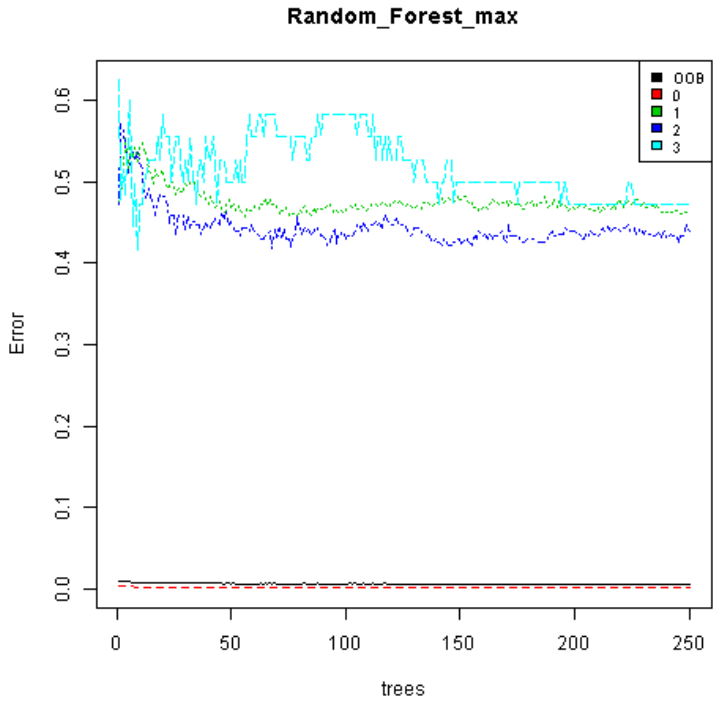


Figura 27 Grafico dell'andamento dei tassi di errore di classificazione. [J=250]

Ridurre ulteriormente il numero di alberi, porterebbe ad ottenere una foresta caratterizzata da un maggior "rumore" della soglia 3; a cui si assocerebbe una marcata volatilità delle performance.

Come si può apprendere da letteratura (Cutler et al., 2012), considerando di aver raggiunto buona stabilità nella classificazione, apprezzabili miglioramenti delle performance di training sono dell'ordine del 5-10% sulla classificazione generale e 10-20% su quella delle singole classi (nel nostro caso, le soglie).

Nei test eseguiti, le percentuali di miglioramento ottenute sono ben lontane da quelle suggerite; se escludiamo il quarto test, 250 alberi, poiché troppo vicino alla condizione di "instabilità" della soglia 3, la scelta del numero di alberi superiore a 500 è pressoché indifferente.

Se la differenza in termini di performance è trascurabile; per valutare la scelta migliore, bisogna però considerare i tempi di calcolo necessari per processare il modello in funzione del numero di alberi generati.

3.4.3.3 TREES-TIME-ERRORS TEST

Per verificare le relazioni tra numero di alberi, tempo di calcolo e performance in fase di training è stato eseguito un test specifico: considerando sempre il training set del bacino del Fiume Parma, è stato eseguito un *loop* del metodo *Random Forests* con incremento del numero di alberi; partendo da un minimo di 10 alberi, ad ogni ciclo è stato aggiunto un albero fino al raggiungimento di una foresta composta da 1250 alberi. Le performance e i tempi di calcolo sono stati memorizzati. La relazione tra numero di alberi generati, tempo di calcolo e numero di errate classificazioni (errori) è rappresentata nel grafico seguente:

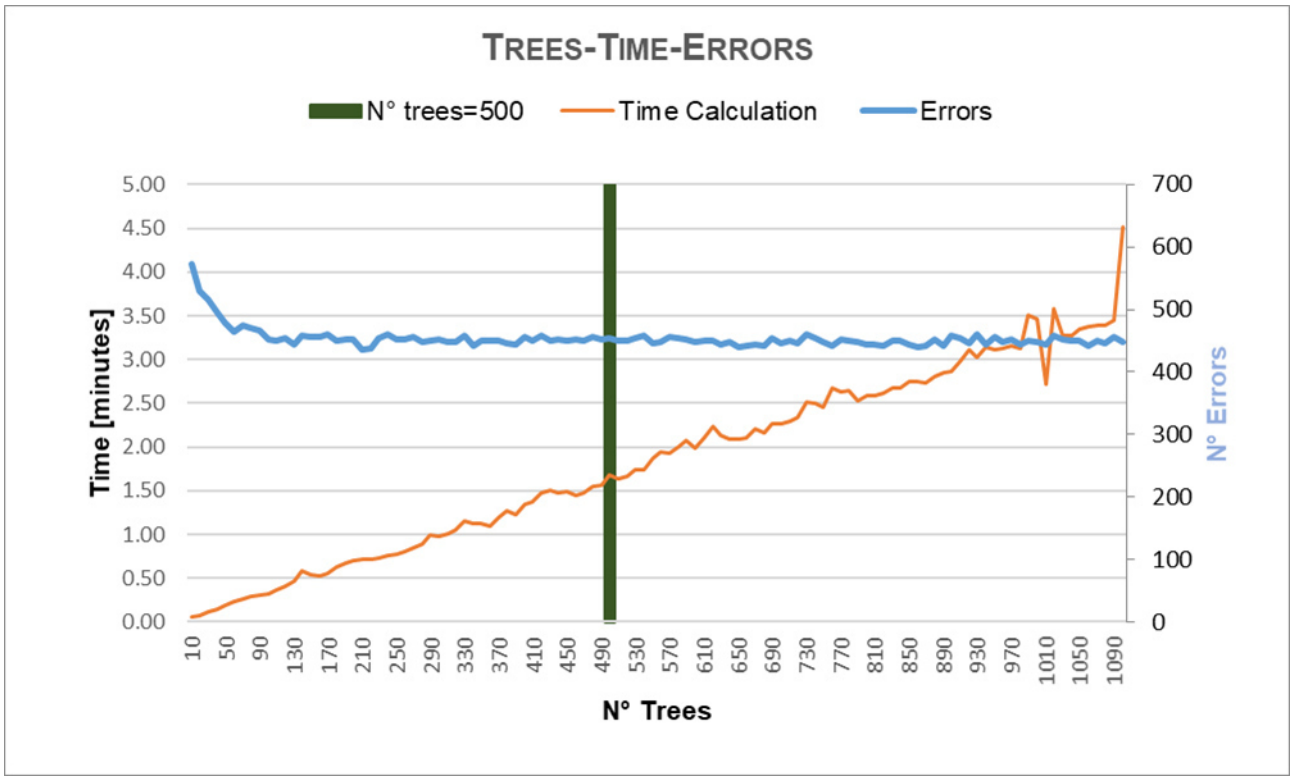


Figura 28 Rappresentazione grafica della relazione: numero di alberi – tempo di calcolo – numero di errori di classificazione

Si evince che il tempo di calcolo (linea arancione), fino a 1100 alberi (circa), cresce linearmente con il numero degli alberi; poi, ad ogni albero aggiunto, il tempo di calcolo cresce esponenzialmente. Il numero totale di errori di classificazione (linea azzurra) si attesa tra 430 e 450, già con i primi 80-100 alberi; a conferma di quanto visto nei quattro test precedentemente descritti.

Pertanto, la scelta finale è legata principalmente al tempo di calcolo. Considerando il valore di default, foresta con 500 alberi, il tempo di calcolo necessario è di poco superiore ai 90 secondi (un minuto e mezzo); ogni 250 alberi aggiunti, il tempo di calcolo aumenta di circa 60 secondi. Per generare una foresta casuale con 1000 alberi sono richiesti circa tre minuti e mezzo; valore più che accettabile in fase operativa.

Per una singola sezione idrometrica, quindi per un solo modello basato sul metodo *Random Forests*, elaborare dei risultati in un minuto e mezzo o tre minuti e mezzo è sostanzialmente indifferente; ma se l'obiettivo è estendere il modello a n-sezioni idrometriche, il tempo totale di elaborazione cresce sensibilmente, senza ottenere alcun miglioramento in termini di performance.

Pertanto, completati e valutati questi test, si è ritenuto opportuno per avere la massima efficienza modellistica, adottare il valore di default, ovvero, 500 alberi; tale valore è anche suggerito da Cutler (Cutler et al., 2012), a valle di numerose verifiche.

3.4.3.4 IMPORTANZA DELLE VARIABILI

Il metodo *Random Forests* consente eventualmente di attribuire dei pesi alle variabili presenti nel *training set*; generalmente, i pesi vengono attribuiti in funzione dell'importanza della variabile, della numerosità della variabile e della qualità dei dati della variabile.

In tal caso, tali modifiche, rientrano nella fase di *setting* del metodo *Random Forests* e si ripercuotono direttamente sulle performance del modello.

Per questo caso studio, a seguito di alcuni tentativi iniziali, si è ritenuto non necessario attribuire dei pesi alle variabili (Soglia 0, Soglia 1, Soglia 2 e Soglia 3) ma di considerare tutte le istanze ugualmente importanti; nonostante la consistente numerosità delle istanze di Soglia 0.

Il pacchetto in R del metodo *Random Forests*, indipendentemente dai pesi eventualmente attribuiti alle variabili, consente di visualizzare (numericamente e graficamente) l'importanza delle variabili utilizzate dedotta dall'algoritmo. Questa funzione caratteristica risulta essere particolarmente utile ed importante, in quanto:

- Consente di verificare l'effettiva importanza di una o più variabili, rispetto alle altre, in fase di *training* dell'albero e, quindi, anche in fase operativa.
- Valutare, a posteriori, in fase operativa se la presenza di dati mancati o alterati di una o più determinate variabili influisce sensibilmente sull'elaborazione del risultato finale.

In riferimento al bacino pilota del fiume Parma, rappresentato graficamente i valori attribuiti dall'algoritmo *Random Forests* si ottengono le seguenti informazioni.

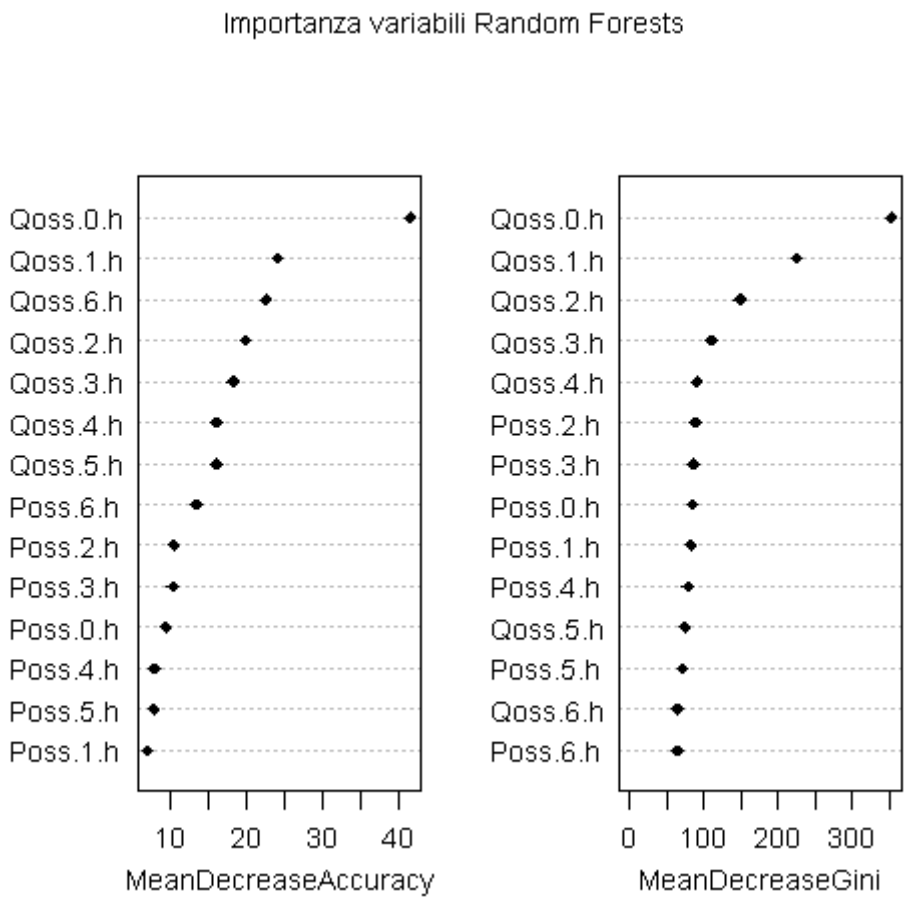


Figura 29 Rappresentazione grafica dell'importanza delle variabili definite dall'algoritmo in fase di *training*

Il grafico di sinistra rappresenta l'importanza attribuita dall'algoritmo ad ogni variabile per migliorare l'accuratezza della classificazione; mentre, il grafico di destra rappresenta l'importanza attribuita ad ogni variabile per l'elaborazione della foresta causale.

Queste informazioni possono essere utilizzate per eseguire un nuovo *training* del modello, intervenendo sulle variabili ed i relativi dati, a cui corrisponderà un nuovo albero e delle nuove valutazioni; oppure, in fase operativa, a supporto dell'operatore.

In quest'ultimo caso, in riferimento al grafico riportato sulla destra, l'operatore potrà dedurre che l'albero finale è stato generando considerando principalmente i valori di portata osservata nelle ultime quattro ore (dall'ora più recente a quella più remota); le altre variabili, hanno un peso minore e svolgono un ruolo di "supporto".

Pur non conoscendo la fisicità del bacino, il metodo *Random Forests* ha ritenuto più importanti i valori di portata nelle ultime ore: descrittori delle condizioni di saturazione del bacino; successivamente i valori di precipitazione osservata nelle quattro ore passate: che per questo bacino rappresentano il tempo di trasformazione dell'afflusso in deflusso in condizioni di prossima saturazione. I valori di precipitazione delle quattro ore precedenti presentano un indice molto simile, a conferma che la pioggia caduta su questo bacino in questo arco temporale contribuisce al deflusso entro le sei ore (n-ore) successive.

Esaminando il grafico di sinistra, si può dedurre invece che in fase operativa l'assenza o l'alterazione del dato relativo alla portata osservata nelle ultime ore influisce maggiormente rispetto all'assenza o all'alterazione del dato relativo alla precipitazione osservata. Pertanto, è preferibile non utilizzare il modello in assenza dell'ultimo valore di portata o viziato da un'alterazione, in quanto si avrebbero dei risultati molto probabilmente errati; mentre, l'assenza della pioggia osservata nell'ultima ora è sicuramente meno impattante in fase di elaborazione dell'informazione da parte del modello.

Al di là del risultato atteso, queste informazioni forniscono un valido feedback della qualità del metodo Random Forests: utilizzando solo i dati a disposizione ("causa ed effetto"), l'algoritmo è riuscito ad individuare le principali variabili che dominano un evento di piena e a dar loro la corretta importanza.

La fase di training può definirsi conclusa, l'albero viene memorizzato sia graficamente sia "numericamente" così da poter esser richiamato per valutare il *validation set* (dataset di valutazione) o l'*operational set* (dataset operativo).

3.4.3.5 Generazione Random Forests

Al termine del processo di calibrazione, viene generato e memorizzato l'albero finale; l'immagazzinamento dell'albero finale consente il suo riutilizzo in fase di validazione e/o operativa. Le nuove istanze, diverse da quelle utilizzate in fase di calibrazione, verranno classificate in funzione delle caratteristiche (nodi, rami e foglie) dell'albero memorizzato. Inoltre, è possibile riprodurre un'immagine dell'albero finale che consente la "classificazione manuale" delle nuove istanze, senza alcun supporto informatico.

Per questo caso studio, le foglie dell'albero finale forniscono una "previsione di probabilità" delle soglie idrometriche relative alla sezione per cui è stato realizzato l'albero. Ogni "foglia" fornirà quattro informazioni, valide per la futura finestra temporale caratteristica di ogni sezione:

- Probabilità di non superamento della soglia 1
- Probabilità di superamento della soglia 1
- Probabilità di superamento della soglia 2
- Probabilità di superamento della soglia 3

Al termine di questo elaborato è stato riportato il *plot* dell'albero finale generato per il Fiume Parma, sezione Ponte Verdi.

In fase operativa, la possibilità di poter salvare e/o stampare l'albero finale permette di poter utilizzare l'elaborazione del *Random Forests* senza l'ausilio informatico; in condizioni operative di emergenza o elevata criticità il supporto cartaceo rimane ancora oggi la soluzione più sicura.

3.4.3.6 Operatività Random Forests

Le nuove istanze verranno classificate sulla base dell'albero ultimo memorizzato e dovranno presentare la medesima struttura (numerosità, ordine, formato, etc.) delle istanze utilizzate in fase di calibrazione.

A differenza del data set di calibrazione, in fase di validazione-operativa l'istanza sarà priva dell'ultima colonna relativa alla massima soglia di allerta raggiunta nelle N-ore successive; in quanto output del modello.

Figura 30 Rappresentazione grafica del processo di elaborazione delle informazioni del *Random Forests*

Il *plot* successivo riporta i risultati ottenuti con il modello Random Forests in relazione alle precipitazioni osservate e alla portata osservata.

Il grafico in alto (ietogramma) rappresenta le precipitazioni orarie registrate e spazializzate sul bacino del fiume Parma chiuso alla sezione di Ponte Verdi. Il grafico in basso rappresenta l'*output probabilistico* del modello Random Forests confrontato con la portata osservata (linea blu). Gli istogrammi rappresentano la probabilità di accadimento, della condizione assente, e la probabilità di superamento, delle tre soglie.

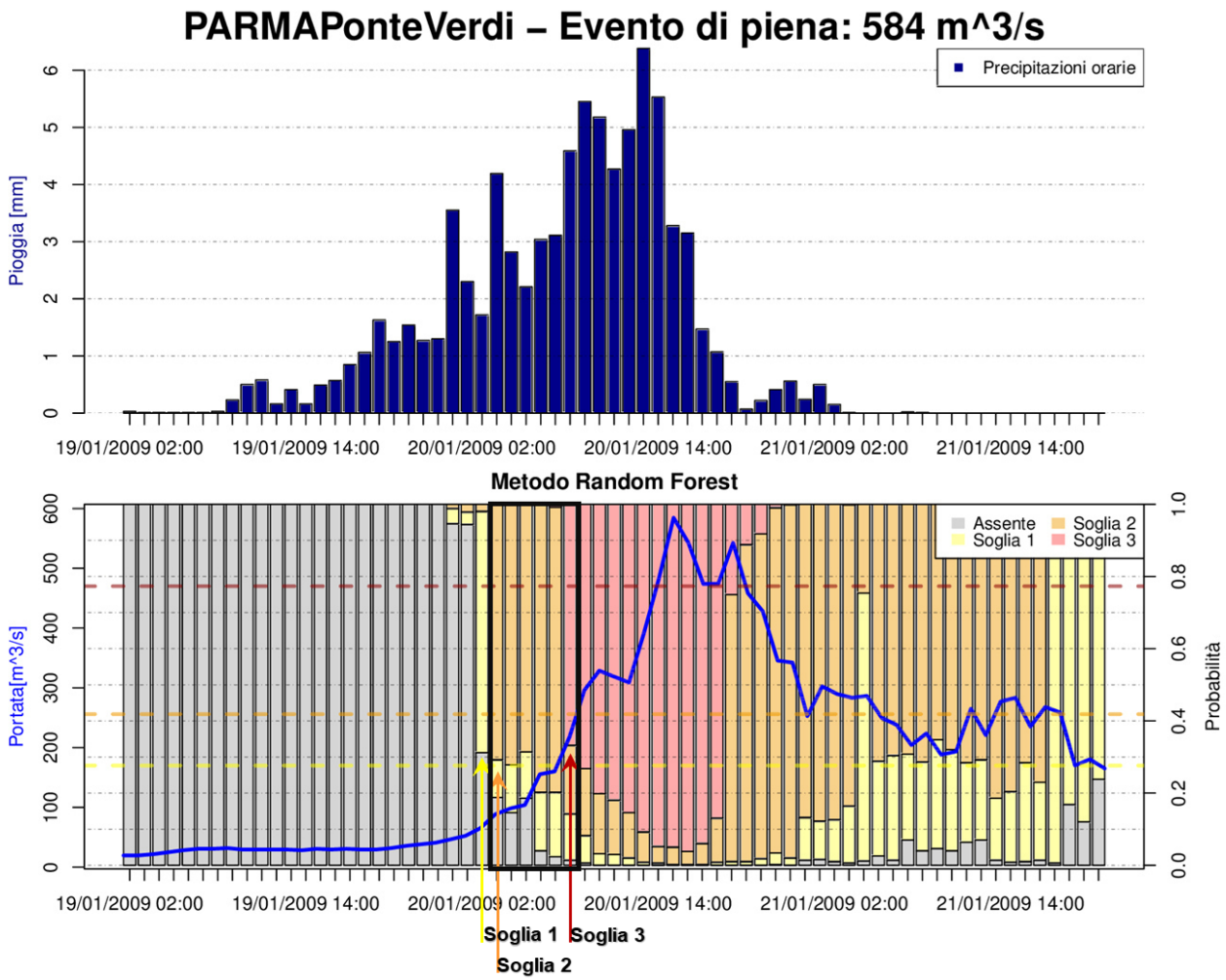


Figura 31 Rappresentazione grafica automatizzata in ambiente R per ogni evento di piena identificato

A titolo di esempio, è stato riportato un evento di piena generato da precipitazioni non particolarmente intenso ma durature nel tempo (oltre 36 ore di pioggia). Le prime piogge sono state registrate nel primo

pomeriggio del 19 gennaio 2009, con una portata di base particolarmente ridotta (indice di un bacino asciutto); una recrudescenza dei fenomeni è stata registrata tra le 00:00 e le 02:00 del 20 gennaio: in previsione, l'intensità delle precipitazioni ha prodotto un sensibile innalzamento della probabilità di superamento della soglia 1 entro le prossime sei ore (finestra temporale caratteristica per la sezione Ponte Verdi, del Fiume Parma); le precipitazioni sono state tali da attribuire al superamento di soglia 1 nelle prossime ore la probabilità più elevata. Dai valori di portata osservati (linea blu) si può notare che entro le successive sei ore dalla previsione, il superamento di soglia 1 è effettivamente avvenuto. La medesima analisi può esser condotta per il superamento di soglia 2 e soglia 3.

3.4.4 RISULTATI DEI MODELLI

Per questo caso studio sono state analizzate le principali sezioni idrometriche dei fiumi dell'Emilia-Romagna; per tutte le sezioni analizzate sono stati applicati sia il modello di Nash sia il metodo Random Forests.

Tutti gli eventi analizzati, sono stati riprodotti graficamente, come mostrato nel paragrafo precedente.

Con l'obiettivo di confrontare e valutare le diverse prestazioni, i risultati forniti dai due diversi modelli sono stati resi comparabili, scegliendo per la fase di confronto:

- Modello Nash: la soglia di allerta a cui è riferita la massima portata simulata nelle N-ore successive (finestra temporale caratteristica della sezione considerata).
- Modello *Random Forests*: la soglia di allerta con il valore più elevato di probabilità di superamento.

La verifica della performance è stata condotta utilizzando le tabelle di contingenza e gli indici di valutazione più comuni.

In riferimento alla tabella di contingenza, sono state identificate le seguenti condizioni:

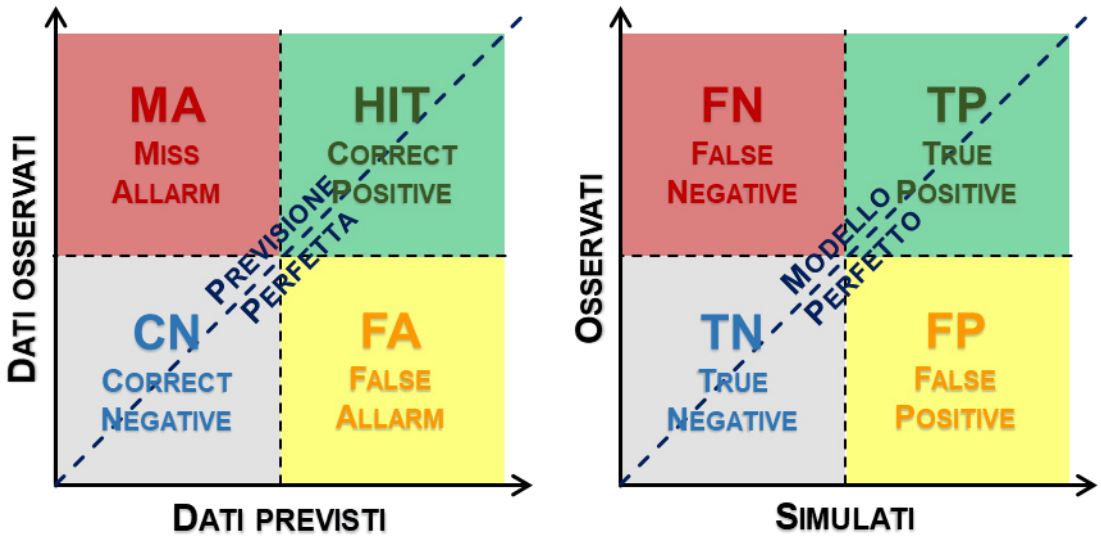


Figura 32 Rappresentazione grafica della verifica delle performance previsionali e modellistiche
A sinistra: rappresentazione grafica di una tabella di contingenza per un modello di previsione
A destra: rappresentazione grafica della tabella di contingenza per la valutazione degli effetti di un modello di previsione in ambito di

Di seguito, le condizioni di valutazione delle performance del modello:

- $Q_{obscond-1h}$: condizione di allerta relativa alla portata osservata a -1h
- $Q_{obscond0h}$: condizione di allerta relativa alla portata osservata a 0h
- $Q_{simcond-1h}$: condizione di allerta simulata a -1h
- $Q_{simcond0h}$: condizione di allerta simulata a 0h
- $\max Q_{simcond+(1h\rightarrow N)h}$: massima condizione di allerta simulata tra +1h e Nh
- $\max Q_{obscond+(1\rightarrow N)h}$: massima condizione di allerta osservata tra +1h e Nh

Soglia 1
$\boldsymbol{HIT}: Q_{obscond-1h} < 1 \;\&\; Q_{obscond0h} \geq 1 \;\&\; \max Q_{simcond+(1\rightarrow N)h} \geq 1$
$\boldsymbol{MA}: Q_{obscond-1h} < 1 \;\&\; Q_{obscond0h} \geq 1 \;\&\; \max Q_{simcond+(1\rightarrow N)h} < 1$
$\boldsymbol{FA0}: Q_{simcond-1h} < 1 \;\&\; Q_{simcond0h} \geq 1 \;\&\; \max Q_{obscond+(1\rightarrow N)h} = 0$
$\boldsymbol{CN}: Q_{simcond0h} < 1 \;\&\; \max Q_{obscond+(1\rightarrow N)h} < 1$
$\boldsymbol{NA}: altri\ dati$

Tabella 12 Condizioni di valutazione performance del modello per Soglia 1

Soglia 2
$\boldsymbol{HIT}: Q_{obscond-1h} < 2 \;\&\; Q_{obscond0h} \geq 2 \;\&\; \max Q_{simcond+(1\rightarrow N)h} \geq 2$
$\boldsymbol{MA}: Q_{obscond-1h} < 2 \;\&\; Q_{obscond0h} \geq 2 \;\&\; \max Q_{simcond+(1\rightarrow N)h} < 2$
$\boldsymbol{FA1}: Q_{simcond-1h} < 2 \;\&\; Q_{simcond0h} \geq 2 \;\&\; \max Q_{obscond+(1\rightarrow N)h} = 1$
$\boldsymbol{FA0}: Q_{simcond-1h} < 2 \;\&\; Q_{simcond0h} \geq 2 \;\&\; \max Q_{obscond+(1\rightarrow N)h} = 0$
$\boldsymbol{CN}: Q_{simcond0h} < 2 \;\&\; \max Q_{obscond+(1\rightarrow N)h} < 2$
$\boldsymbol{NA}: altri\ dati$

Tabella 13 Tabella 12 Condizioni di valutazione performance del modello per Soglia 2

Soglia 3
$\boldsymbol{HIT}: Q_{obscond-1h} < 3 \;\&\; Q_{obscond0h} \geq 3 \;\&\; \max Q_{simcond+(1\rightarrow N)h} \geq 3$
$\boldsymbol{MA}: Q_{obscond-1h} < 3 \;\&\; Q_{obscond0h} \geq 3 \;\&\; \max Q_{simcond+(1\rightarrow N)h} < 3$
$\boldsymbol{FA2}: Q_{simcond-1h} < 3 \;\&\; Q_{simcond0h} \geq 3 \;\&\; \max Q_{obscond+(1\rightarrow N)h} = 2$
$\boldsymbol{FA1}: Q_{simcond-1h} < 3 \;\&\; Q_{simcond0h} \geq 3 \;\&\; \max Q_{obscond+(1\rightarrow N)h} = 1$
$\boldsymbol{FA0}: Q_{simcond-1h} < 3 \;\&\; Q_{simcond0h} \geq 3 \;\&\; \max Q_{obscond+(1\rightarrow N)h} = 0$
$\boldsymbol{CN}: Q_{simcond0h} < 3 \;\&\; \max Q_{obscond+(1\rightarrow N)h} < 3$
$\boldsymbol{NA}: altri\ dati$

Tabella 14 Tabella 12 Condizioni di valutazione performance del modello per Soglia 3

L'obiettivo cardine di questo caso studio era elaborare un modello affidabile avente un numero di mancati allarmi prossimo o uguale a zero e un numero più basso possibile di falsi allarmi.

Nelle successive tabelle sono riportati i risultati ottenuti in fase di calibrazione, su un periodo continuo di dieci anni (2005-2015), sia con il modello di Nash sia con il modello *Random Forests*.

BAGANZA	SOGLIA	MODELLO DI NASH				MODELLO RANDOM FORESTS				
		TP	FN	FP	TN	TP	FN	FP	TN	NA

Ponte Nuovo	Soglia 1	65	26	268	50272	91	0	0	51433	3784
	Soglia 2	32	16	191	53950	48	0	0	54612	648
	Soglia 3	19	5	123	54662	24	0	0	55017	267

PARMA		MODELLO DI NASH				MODELLO RANDOM FORESTS				
	SOGLIA	TP	FN	FP	TN	TP	FN	FP	TN	NA
Berceto	Soglia 1	3	3	30	80657	6	0	0	80748	22
	Soglia 2	1	0	7	80752	1	0	0	80772	3
	Soglia 3	0	1	1	80768	1	0	0	80773	2
Marzolara	Soglia 1	7	0	68	49493	7	0	0	79760	29
	Soglia 2	2	0	18	79728	2	0	0	79784	10
	Soglia 3	0	1	4	79792	1	0	0	79791	4
Ponte Verdi	Soglia 1	58	19	156	74905	77	0	0	75709	758
	Soglia 2	28	10	71	75943	38	0	0	76232	274
	Soglia 3	4	1	12	76479	5	0	0	76507	32

TREBBIA		MODELLO DI NASH				MODELLO RANDOM FORESTS				
	SOGLIA	TP	FN	FP	TN	TP	FN	FP	TN	NA
Bobbio	Soglia 1	8	0	55	96001	8	0	0	96169	60
	Soglia 2	2	0	12	96186	2	0	0	96221	14
	Soglia 3	0	1	1	96233	1	0	0	96229	7
Cabanne	Soglia 1	39	18	89	93682	57	0	0	93986	442
	Soglia 2	7	9	24	94310	16	0	0	94377	92
	Soglia 3	2	1	3	94464	3	0	0	94469	13
Rivergaro	Soglia 1	27	2	125	94059	29	0	0	94724	403
	Soglia 2	2	0	9	95098	2	0	0	95135	19
	Soglia 3	0	1	1	95160	1	0	0	95147	8
Salsominore	Soglia 1	77	0	469	89469	77	0	0	92252	1020
	Soglia 2	30	1	251	91735	31	0	0	92904	414
	Soglia 3	6	3	26	93174	9	0	0	93253	87
Valsigiara	Soglia 1	23	2	53	93494	25	0	0	93783	211
	Soglia 2	1	3	10	93941	4	0	0	93988	27
	Soglia 3	0	2	4	93995	2	0	0	94004	13

TIDONE		MODELLO DI NASH				MODELLO RANDOM FORESTS				
	SOGLIA	TP	FN	FP	TN	TP	FN	FP	TN	NA
Rottofreno	Soglia 1	7	0	79	31929	7	0	0	32154	64
	Soglia 2	2	0	10	32204	2	0	0	32208	15
	Soglia 3	0	1	5	32223	1	0	0	32217	7

NURE		MODELLO DI NASH				MODELLO RANDOM FORESTS				
	SOGLIA	TP	FN	FP	TN	TP	FN	FP	TN	NA
Ferriere	Soglia 1	13	2	69	81988	15	0	0	82220	69
	Soglia 2	1	1	14	82234	2	0	0	82294	8

	Soglia 3	0	1	1	82296	1	0	0	82300	3
Farini	Soglia 1	29	0	271	87481	29	0	0	88413	229
	Soglia 2	5	0	26	88562	5	0	0	88644	22
	Soglia 3	1	0	9	88640	1	0	0	88665	5

CHIAVENNA	SOGLIA	MODELLO DI NASH				MODELLO RANDOM FORESTS				
		TP	FN	FP	TN	TP	FN	FP	TN	NA
Saliceto	Soglia 1	8	12	57	91118	20	0	0	91246	199
	Soglia 2	0	2	6	91448	2	0	0	91447	16
	Soglia 3	0	1	2	91472	1	0	0	91457	7
Montanaro	Soglia 1	51	1	192	73099	52	0	0	74000	524
	Soglia 2	6	2	44	74400	8	0	0	74507	61
	Soglia 3	0	1	1	74581	1	0	0	74569	6

TARO	SOGLIA	MODELLO DI NASH				MODELLO RANDOM FORESTS				
		TP	FN	FP	TN	TP	FN	FP	TN	NA
Borgotaro	Soglia 1	21	2	40	91850	23	0	0	91997	175
	Soglia 2	6	5	20	92058	11	0	0	92106	78
	Soglia 3	2	0	5	92174	2	0	0	92181	12
Castellina di S.	Soglia 1	37	3	213	93241	40	0	0	94042	373
	Soglia 2	16	6	114	93865	22	0	0	94255	178
	Soglia 3	4	2	37	94325	6	0	0	94403	46
Fidenza Siap	Soglia 1	102	37	197	62439	139	0	0	63289	2200
	Soglia 2	5	3	17	65551	8	0	0	65585	41
	Soglia 3	0	1	2	65627	1	0	0	65628	5
Ponte Lamberti	Soglia 1	3	4	32	76384	7	0	0	76489	48
	Soglia 2	3	4	9	76479	7	0	0	76501	36
	Soglia 3	0	2	2	76534	2	0	0	76532	10
Ponte Taro	Soglia 1	18	12	74	81549	30	0	0	81775	542
	Soglia 2	6	7	16	82141	13	0	0	82142	192
	Soglia 3	0	1	1	82365	1	0	0	82331	15
Pradella	Soglia 1	23	2	56	86716	25	0	0	86972	232
	Soglia 2	3	7	16	87083	10	0	0	87145	75
	Soglia 3	0	1	6	87219	1	0	0	87223	6
S. Secondo	Soglia 1	6	18	26	88072	24	0	0	88166	474
	Soglia 2	6	12	26	88234	18	0	0	88299	347
	Soglia 3	0	1	3	88668	1	0	0	88645	18
Toccalmatto	Soglia 1	46	58	158	84613	104	0	0	85193	989
	Soglia 2	2	19	33	86055	21	0	0	86132	150
	Soglia 3	0	1	11	86264	1	0	0	86298	5
Tornolo	Soglia 1	13	20	32	93552	33	0	0	93649	212
	Soglia 2	2	5	11	93817	7	0	0	93848	39
	Soglia 3	0	1	3	93889	1	0	0	93889	4

ENZA	SOGLIA	MODELLO DI NASH				MODELLO RANDOM FORESTS				
		TP	FN	FP	TN	TP	FN	FP	TN	NA
Compiano	Soglia 1	11	1	261	89248	12	0	0	90413	81
	Soglia 2	2	0	23	90465	2	0	0	90492	12
	Soglia 3	0	1	6	90498	1	0	0	90500	5
Selvanizza	Soglia 1	19	1	103	90591	20	0	0	90997	122
	Soglia 2	4	2	12	91062	6	0	0	91108	25
	Soglia 3	0	1	4	91115	1	0	0	91135	3
Sorbolo	Soglia 1	105	37	285	75794	142	0	0	78448	4855
	Soglia 2	50	33	194	79760	83	0	0	81099	2263
	Soglia 3	24	39	96	81297	63	0	0	81899	1483
Vetto	Soglia 1	70	5	195	79459	75	0	0	80350	671
	Soglia 2	23	0	76	80637	23	0	0	80940	133
	Soglia 3	1	0	45	80885	1	0	0	81090	5

CROSTOLO	SOGLIA	MODELLO DI NASH				MODELLO RANDOM FORESTS				
		TP	FN	FP	TN	TP	FN	FP	TN	NA
Cadelbosco	Soglia 1	2	16	28	88977	18	0	0	89037	222
	Soglia 2	1	2	5	89245	3	0	0	89241	33
	Soglia 3	1	0	1	89286	1	0	0	89268	8
Puianello	Soglia 1	13	6	68	83823	19	0	0	83971	73
	Soglia 2	0	2	4	84054	2	0	0	84053	8
	Soglia 3	0	1	1	84063	1	0	0	84058	4

SECCHIA	SOGLIA	MODELLO DI NASH				MODELLO RANDOM FORESTS				
		TP	FN	FP	TN	TP	FN	FP	TN	NA
Ca de Caroli	Soglia 1	32	9	237	86956	41	0	0	87683	385
	Soglia 2	15	6	118	87661	21	0	0	87906	182
	Soglia 3	0	1	2	88114	1	0	0	88102	6
Ponte Cavola	Soglia 1	13	0	139	93149	13	0	0	93775	117
	Soglia 2	7	0	30	93732	7	0	0	93859	39
	Soglia 3	2	1	11	93843	3	0	0	93887	15
Ponte Dolo	Soglia 1	27	3	111	59774	30	0	0	60142	180
	Soglia 2	5	3	20	60240	8	0	0	60304	40
	Soglia 3	1	0	5	60330	1	0	0	60347	4
Ponte Lugo	Soglia 1	48	2	162	92954	50	0	0	93789	623
	Soglia 2	7	1	40	94176	8	0	0	94394	60

	Soglia 3	0	1	2	94464	1	0	0	94455	6
Rossenna	Soglia 1	38	6	170	85943	44	0	0	86420	345
	Soglia 2	11	2	62	86537	13	0	0	86705	91
	Soglia 3	0	1	3	86807	1	0	0	86803	5
Gatta	Soglia 1	58	3	259	85549	61	0	0	87126	722
	Soglia 2	2	0	11	87849	2	0	0	87896	11
	Soglia 3	0	1	4	87886	1	0	0	87904	4
Rubiera SS9	Soglia 1	99	26	277	63452	125	0	0	65474	3095
	Soglia 2	29	5	133	67502	34	0	0	68135	525
	Soglia 3	7	4	32	68484	11	0	0	68586	97
Ponte Alto	Soglia 1	85	10	199	85903	95	0	0	88326	5004
	Soglia 2	40	6	137	90550	46	0	0	91764	1615
	Soglia 3	1	1	8	93340	2	0	0	93369	54
Ponte Bacchello	Soglia 1	54	23	194	87994	77	0	0	90249	5115
	Soglia 2	31	12	121	92409	43	0	0	93696	1702
	Soglia 3	1	2	14	95289	3	0	0	95331	107
Pioppa	Soglia 1	23	33	180	89728	56	0	0	91592	3976
	Soglia 2	8	15	60	94015	23	0	0	94527	1074
	Soglia 3	0	1	8	95573	1	0	0	95594	29

PANARO	SOGLIA	MODELLO DI NASH				MODELLO RANDOM FORESTS				
		TP	FN	FP	TN	TP	FN	FP	TN	NA
Fiumalbo	Soglia 1	38	14	97	84190	52	0	0	84628	565
	Soglia 2	4	4	18	85149	8	0	0	85212	25
	Soglia 3	0	2	2	85236	2	0	0	85237	6
Pievepelago	Soglia 1	17	8	55	92957	25	0	0	93186	137
	Soglia 2	3	3	25	93200	6	0	0	93324	18
	Soglia 3	0	1	2	93342	1	0	0	93344	3
Ponte Val Sasso	Soglia 1	50	1	217	72667	51	0	0	73832	580
	Soglia 2	12	0	35	74160	12	0	0	74385	66
	Soglia 3	0	1	1	74465	1	0	0	74457	5
Fanano	Soglia 1	40	19	32	93711	59	0	0	93821	311
	Soglia 2	6	1	16	94104	7	0	0	94162	22
	Soglia 3	0	1	12	94145	1	0	0	94187	3
Ponte Samone	Soglia 1	34	0	45	94552	34	0	0	94749	249
	Soglia 2	2	2	19	94931	4	0	0	94999	29
	Soglia 3	1	0	6	95015	1	0	0	95025	6
Spilamberto	Soglia 1	137	37	185	80456	174	0	0	81174	1735
	Soglia 2	3	8	24	82913	11	0	0	82966	106
	Soglia 3	0	3	6	83055	3	0	0	83049	31
Bomporto	Soglia 1	31	13	223	91846	44	0	0	93302	1970
	Soglia 2	13	14	63	93961	27	0	0	94351	938
	Soglia 3	2	0	6	95281	2	0	0	95273	41

	SOGLIA	MODELLO DI NASH	MODELLO RANDOM FORESTS

RENO		TP	FN	FP	TN	TP	FN	FP	TN	NA
Pracchia	Soglia 1	32	7	55	77496	39	0	0	77753	233
	Soglia 2	7	2	31	77849	9	0	0	77979	37
	Soglia 3	1	1	5	77996	2	0	0	78019	4
Silla	Soglia 1	72	6	208	49650	78	0	0	50716	987
	Soglia 2	18	2	40	51440	20	0	0	51658	103
	Soglia 3	0	1	7	51748	1	0	0	51777	3
Vergato	Soglia 1	110	0	249	65230	110	0	0	66951	2197
	Soglia 2	33	1	48	68585	34	0	0	68822	402
	Soglia 3	0	1	5	69241	1	0	0	69249	8
Calcara	Soglia 1	10	2	34	61837	12	0	0	61912	75
	Soglia 2	1	1	5	61980	2	0	0	61983	14
	Soglia 3	0	1	1	62001	1	0	0	61991	7
Forcelli	Soglia 1	36	16	125	48529	52	0	0	49645	1764
	Soglia 2	28	3	93	49891	31	0	0	50606	824
	Soglia 3	0	1	4	51435	1	0	0	51444	16
Lavinio di Sopra	Soglia 1	52	22	276	66536	74	0	0	67865	636
	Soglia 2	7	4	38	68414	11	0	0	68520	44
	Soglia 3	0	2	11	68544	2	0	0	68565	8
Castenaso	Soglia 1	15	5	133	75113	20	0	0	75477	253
	Soglia 2	2	0	35	75644	2	0	0	75729	19
	Soglia 3	1	0	30	75671	1	0	0	75740	9
Sesto Imolese	Soglia 1	1	3	35	79605	4	0	0	79725	56
	Soglia 2	1	2	19	79710	3	0	0	79752	30
	Soglia 3	0	1	11	79762	1	0	0	79775	9
Mordano	Soglia 1	4	11	70	93488	15	0	0	93683	249
	Soglia 2	0	2	4	93930	2	0	0	93921	24
	Soglia 3	0	1	3	93951	1	0	0	93936	10
Casola V.Senio	Soglia 1	11	0	77	66582	11	0	0	66823	80
	Soglia 2	2	0	18	66854	2	0	0	66902	10
	Soglia 3	1	0	2	66905	1	0	0	66909	4
Castelbolognese	Soglia 1	9	16	69	59184	25	0	0	59366	378
	Soglia 2	2	3	16	59671	5	0	0	59703	61
	Soglia 3	1	0	4	59764	1	0	0	59759	9
Casalecchio T.V.	Soglia 1	95	7	275	85877	102	0	0	87771	2397
	Soglia 2	4	7	20	90022	11	0	0	90082	177
	Soglia 3	0	1	3	90272	1	0	0	90259	10

LAMONE	SOGLIA	MODELLO DI NASH				MODELLO RANDOM FORESTS				
		TP	FN	FP	TN	TP	FN	FP	TN	NA
Marradi	Soglia 1	79	8	326	79626	87	0	0	81206	884
	Soglia 2	13	4	105	81755	17	0	0	82073	87
	Soglia 3	0	2	7	82162	2	0	0	82170	6
Strada Casale	Soglia 1	30	0	198	82800	30	0	0	83963	307
	Soglia 2	4	0	42	84069	4	0	0	84276	20
	Soglia 3	0	1	1	84301	1	0	0	84294	5

Sarna	Soglia 1	12	1	185	71813	13	0	0	72688	103
	Soglia 2	2	0	53	72598	2	0	0	72787	15
	Soglia 3	0	1	9	72788	1	0	0	72797	6
Reda	Soglia 1	12	5	151	84743	17	0	0	85316	366
	Soglia 2	0	2	22	85594	2	0	0	85666	31
	Soglia 3	0	1	5	85702	1	0	0	85685	13

MONTONE		MODELLO DI NASH				MODELLO RANDOM FORESTS				
	SOGLIA	TP	FN	FP	TN	TP	FN	FP	TN	NA
Castrocaro	Soglia 1	41	6	333	89773	47	0	0	91756	670
	Soglia 2	6	2	117	91862	8	0	0	92394	71
	Soglia 3	1	0	12	92439	1	0	0	92466	6
Ponte Vico	Soglia 1	20	5	120	81752	25	0	0	82630	727
	Soglia 2	7	1	43	82844	8	0	0	83120	254
	Soglia 3	2	1	4	83230	3	0	0	83281	98
Predappio	Soglia 1	55	10	358	81305	65	0	0	84089	1169
	Soglia 2	22	1	305	83304	23	0	0	85090	210
	Soglia 3	1	0	115	84852	1	0	0	85316	6

RONCO		MODELLO DI NASH				MODELLO RANDOM FORESTS				
	SOGLIA	TP	FN	FP	TN	TP	FN	FP	TN	NA
Coccolia	Soglia 1	19	18	117	90374	37	0	0	90990	922
	Soglia 2	3	4	27	91688	7	0	0	91768	174
	Soglia 3	1	1	4	91904	2	0	0	91895	52

SAVIO		MODELLO DI NASH				MODELLO RANDOM FORESTS				
	SOGLIA	TP	FN	FP	TN	TP	FN	FP	TN	NA
San Carlo	Soglia 1	44	7	198	85523	51	0	0	86412	727
	Soglia 2	9	5	70	86717	14	0	0	86966	210
	Soglia 3	1	1	5	87172	2	0	0	87169	19
Borello	Soglia 1	11	3	149	59050	14	0	0	59904	196
	Soglia 2	4	0	33	59901	4	0	0	60066	44
	Soglia 3	1	0	8	60063	1	0	0	60107	6

MARECCHIA		MODELLO DI NASH				MODELLO RANDOM FORESTS				
	SOGLIA	TP	FN	FP	TN	TP	FN	FP	TN	NA
Rimini	Soglia 1	14	9	79	50170	23	0	0	50392	279
	Soglia 2	3	7	28	50529	10	0	0	50584	100
	Soglia 3	1	0	4	50686	1	0	0	50685	8

RUBICONE		MODELLO DI NASH				MODELLO RANDOM FORESTS				
	SOGLIA	TP	FN	FP	TN	TP	FN	FP	TN	NA

Savignano	Soglia 1	6	0	51	31073	6	0	0	31222	79
	Soglia 2	3	0	8	31258	3	0	0	31274	30
	Soglia 3	0	1	1	31312	1	0	0	31300	6

PISCIATELLO		MODELLO DI NASH				MODELLO RANDOM FORESTS				
	SOGLIA	TP	FN	FP	TN	TP	FN	FP	TN	NA
Calisese	Soglia 1	27	28	103	49329	55	0	0	49604	392
	Soglia 2	7	6	26	49878	13	0	0	49932	106
	Soglia 3	1	0	1	50051	1	0	0	50045	5

USO		MODELLO DI NASH				MODELLO RANDOM FORESTS				
	SOGLIA	TP	FN	FP	TN	TP	FN	FP	TN	NA
Santarcangelo	Soglia 1	22	9	108	68596	31	0	0	68939	348
	Soglia 2	1	9	12	69204	10	0	0	69226	82
	Soglia 3	0	4	5	69279	4	0	0	69283	31

SOGLIA	MODELLO DI NASH				MODELLO RANDOM FORESTS				
	TP	FN	FP	TN	TP	FN	FP	TN	NA
Soglia 1	2567	718	10613	5630790	3285	0	0	5717451	3285
Soglia 2	635	301	3459	5749277	936	0	0	5765535	936
Soglia 3	92	116	807	5776048	208	0	0	5778344	208
Totale	3294	1135	14879	17156115	4429	0	0	17261330	4429

Confrontando i risultati ottenuti in fase di calibrazione si evince una performance sensibilmente più elevata del modello *Random Forests* rispetto al modello speditivo di *Nash*; a conferma dell'elevata capacità di apprendimento in fase di addestramento del metodo *Random Forests*, sono l'assenza di FN (Falsi Negativi, o mancati allarmi) e di (Falsi Positivi, o falsi allarmi).

L'albero generato è stato prodotto con le istanze del *training set* in cui erano presenti sia le cause (valori osservati) sia gli effetti (superamento soglie idrometriche); in fase di verifica, utilizzando le medesime istanze, prive però degli effetti, l'algoritmo ha egregiamente identificato gli effetti prodotti (output modello) dalle relative istanze.

Questa è una delle principali peculiarità dei metodi più avanzati di intelligenza artificiale: saper riconoscere in operatività ciò che è stato già visto almeno una volta.

Al termine di questo elaborato, verrà affrontato e discusso l'utilizzo di questo modello in fase operativa.

3.4.5 PERFORMANCE DEI MODELLI

I dati utilizzati in questo caso studio per le comporre le istanze del *training set* sono stati precedentemente controllati e validati; pertanto, la qualità dei dati è verificata.

Trattandosi di data set particolarmente eterogenei, caratterizzati da un'elevata numerosità di valori nulli (precipitazioni) o variazioni trascurabili (portate nei periodi estivi), affidarsi al valore di accuratezza complessiva del processo di classificazione non è sufficiente per avere un riscontro attendibile delle performance del modello; è opportuno ricorrere a tecniche più selettive e mirate alla valutazione del raggiungimento degli obiettivi prefissati.

Esistono diversi indici per valutare le prestazioni dei modelli; è importante conoscere lo scopo del modello per scegliere gli indici di valutazione corretti su cui basare la verifica delle performance.

Ad esempio, nel caso in cui gli effetti di una previsione "estrema" influenzano significativamente le attività, l'economia o la sicurezza delle persone, è importante valutare l'affidabilità del modello in situazioni "estreme"; attribuendo eventualmente un peso maggiore per i mancati allarmi.

In questo caso, è meglio dotarsi di un classificatore che fornisce un'accuratezza di predizione elevata rispetto alla classe di minoranza, pur mantenendo un'accuratezza ragionevole per la classe di maggioranza.

Un'altra valutazione, più avanzata che richiede la conoscenza anche di altri elementi, è basata sul rapporto costi/benefici; ad esempio, quantificare il numero di falsi allarmi sostenibili rispetto al numero dei mancati allarmi. (Chen et al., 2004).

Per valutare le performance dei modelli realizzati in questo caso studio, si è fatto riferimento a cinque dei principali indici di valutazione delle prestazioni dei modelli; validi sia per i modelli classici sia per i modelli basati sulle tecniche di intelligenza artificiale. (Shaikhina et al., 2017):

1. "*Probability of detection POD*": valuta la frazione degli eventi osservati che è stata effettivamente prevista; è un indice sensibile ai mancati allarmi, ma ignora i falsi allarmi.

$$POD = \frac{HITS}{HITS + MISS}$$

2. *"False Alarm Ratio FAR*": valuta la frazione degli eventi previsti ma non verificatisi; è un indice sensibile ai falsi allarmi, ma ignora i mancati allarmi.

$$FAR = \frac{FALSE}{HITS + FALSE}$$

3. "*Critical Success Index CSI*": valuta il rapporto degli eventi previsti e verificatisi rispetto a tutti gli eventi osservati; risulta sensibile agli "*hits*" ma penalizza i falsi allarmi e i mancati allarmi.

$$CSI = \frac{HITS}{HITS + FALSE + MISS}$$

4. "*Success Ratio SR*": valuta la frazione effettivamente osservata di tutti gli eventi previsti; ignora la presenza dei mancati allarmi.

$$SR = \frac{HITS}{HITS + FALSE}$$

5. "*BIAS*": valuta la relazione tra gli eventi previsti e gli eventi osservati.

$$BIAS = \frac{HITS + FALSE}{HITS + MISS}$$

I risultati degli indici ottenuti per le principali sezioni considerate sono riportati di seguito:

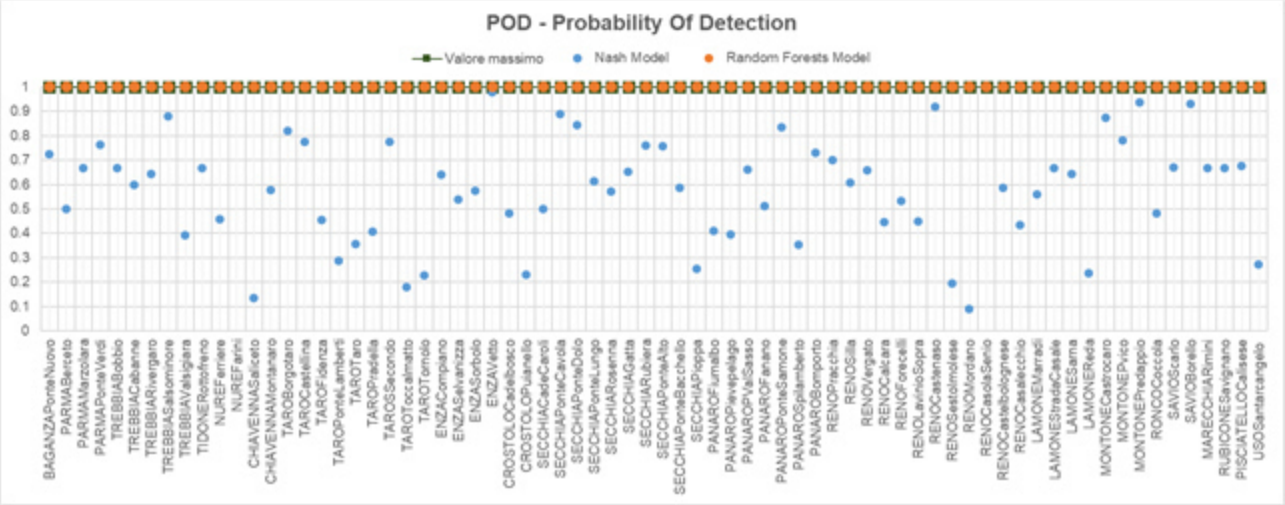


Figura 33 Risultati indice POD

L'indice POD (*Probability Of Detection*) per il modello di Nash risulta essere particolarmente variabile; mentre, per il modello Random Forests, in fase di calibrazione, l'indice raggiunge il valore massimo su tutte le sezioni analizzate.

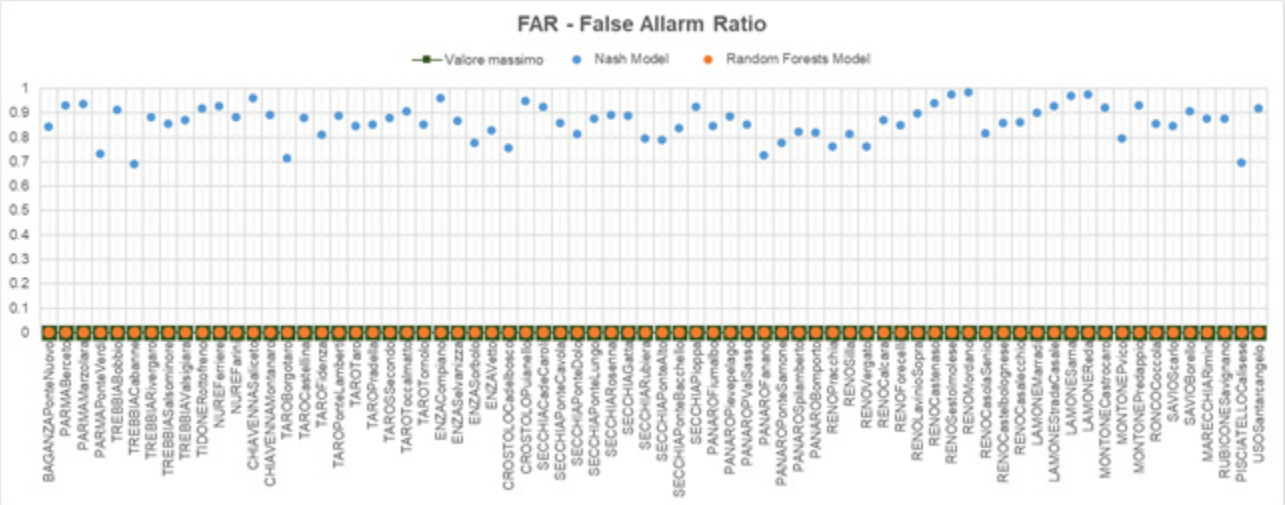


Figura 34 Risultati indice FAR

L'assenza di "falsi allarmi" nei risultati del metodo *Random Forests* fornisce un valore dell'indice FAR (*False Allarm Ratio*) pari a zero; mentre, l'elevata numerosità di falsi allarmi riscontrata nel modello di Nash è confermata dall'elevato valore dell'indice FAR, compreso tra 0.7 e 0.95.

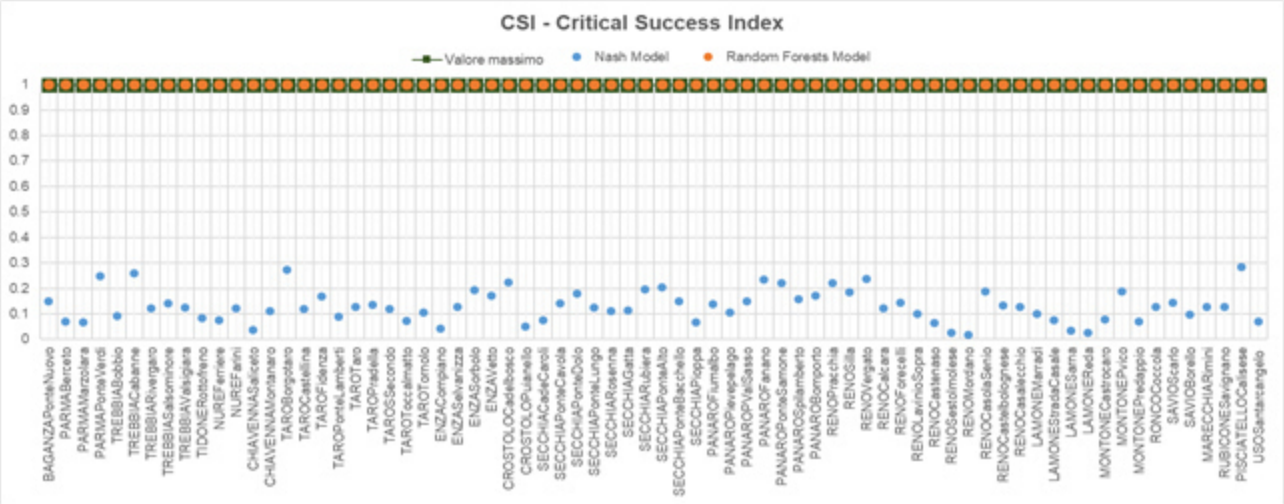


Figura 35 Risultati indice CSI

L'indice CSI (*Critical Success Index*) pari al valore massimo in tutte le sezioni per il metodo Random Forests conferma, in fase di calibrazione, l'assenza di mancati allarmi e/o falsi allarmi; mentre, denota la presenza di falsi e mancati allarmi nelle performance del modello di Nash.

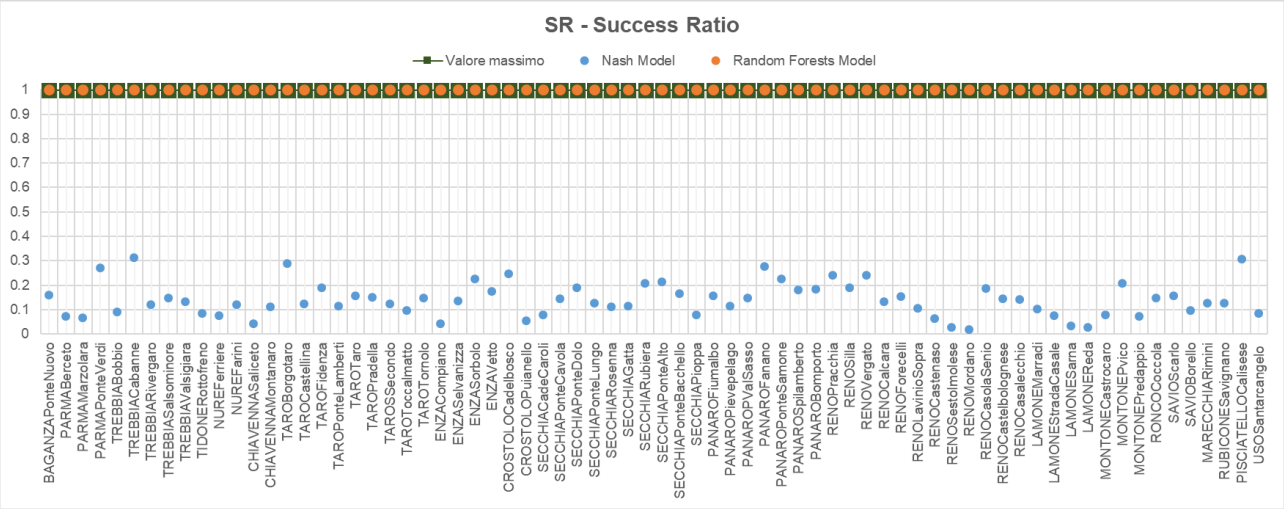


Figura 36 Risultati indice SR

Rispetto all'indice CSI, l'indice SR (*Success Ratio*) non considera i mancati allarmi; i valori riscontrati con il modello di Nash, come per l'indice CSI, risultano essere in diversi casi non soddisfacenti. Per il metodo Random Forests, in relazione ai risultati ottenuti, l'indice SR è ridondante rispetto all'indice CSI.

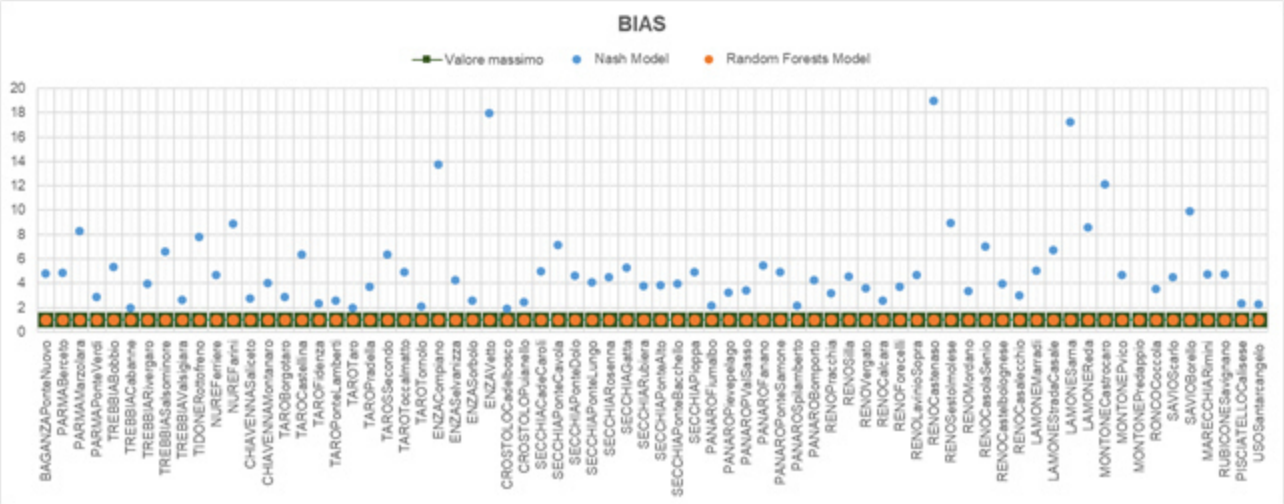


Figura 37 Risultati indice BIAS

L'indice BIAS, che fornisce una valutazione complessiva delle performance del modello, risulta pari a 1 per il metodo Random Forests; mentre, per il modello di Nash, in alcuni casi è soddisfacente in altri pienamente insufficiente.

Lo scopo di questo studio era identificare un modello che potesse prevedere gli eventi di piena con un adeguato tempo di preavviso, necessario per poter disporre le principali attività di protezione civile, con una bassa numerosità di falsi allarmi e, in particolare, di mancati allarmi.

I risultati ottenuti con il modello *Random Forests* in fase di calibrazione sono stati particolarmente soddisfacenti; in quanto l'algoritmo *Random Forests* è in grado di memorizzare un'elevata numerosità di dati che consente in fase operativa di identificare gli effetti prodotti da osservazioni simili a quelle già viste in passato.

La probabilità associata al superamento delle soglie idrometriche, fornita dal modello, è un valore aggiunto per i tecnici che dovranno valutare gli interventi e le attività in fase d'evento; in molti casi, soprattutto per i superamenti della soglia 3 (soglia di allarme), i danni scaturiti da un mancato allarme presentano dei costi economici sensibilmente superiori rispetto a quelli scaturiti da numerosi falsi allarmi. Pertanto, potrebbe essere opportuno intervenire anche con percentuali di superamento di soglia particolarmente bassi; per questo motivo, l'output del modello è sempre caratterizzato dalla probabilità di superamento delle quattro condizioni (assente, soglia 1, soglia 2 e soglia 3) indipendentemente dal loro valore. Solo per la fase di verifica qui presentata è stata selezionata per ogni istanza di output la condizione avente la percentuale maggiore.

Nella parte conclusiva di questo elaborato verranno riportate alcune riflessioni e considerazioni in merito al metodo *Random Forests*, con particolare attenzione alla fase operativa; anticipiamo qui che le performance ottenute in fase di calibrazione non devono trarre in inganno: il metodo è sicuramente performante, ma ben lontano dal "modello perfetto"; come verrà illustrato successivamente, l'algoritmo è in grado di riconoscere ciò che ha già "imparato" dal passato, ma potrebbe risultare inefficace nel valutare nuove istanze estreme ben lontane da quanto osservato in fase di *training* (calibrazione). In alcuni casi è possibile "forzare" l'addestramento con delle istanze sintetiche; in altri, questa operazione potrebbe risultare comunque non sufficiente.

3.4.6 ULTERIORI CONFRONTI

Per verificare la bontà del modello *Random Forests*, utilizzando il medesimo set di dati (2005-2015) sulle medesime sezioni è stata applicata la tecnica "*BMBP – Bayesian Multivariate Binary Predictor*" (Todini, 2008); i risultati ottenuti sono stati confrontati rapidamente confrontando i due indici più rappresentativi per le finalità definite: *POD* e *FAR*.

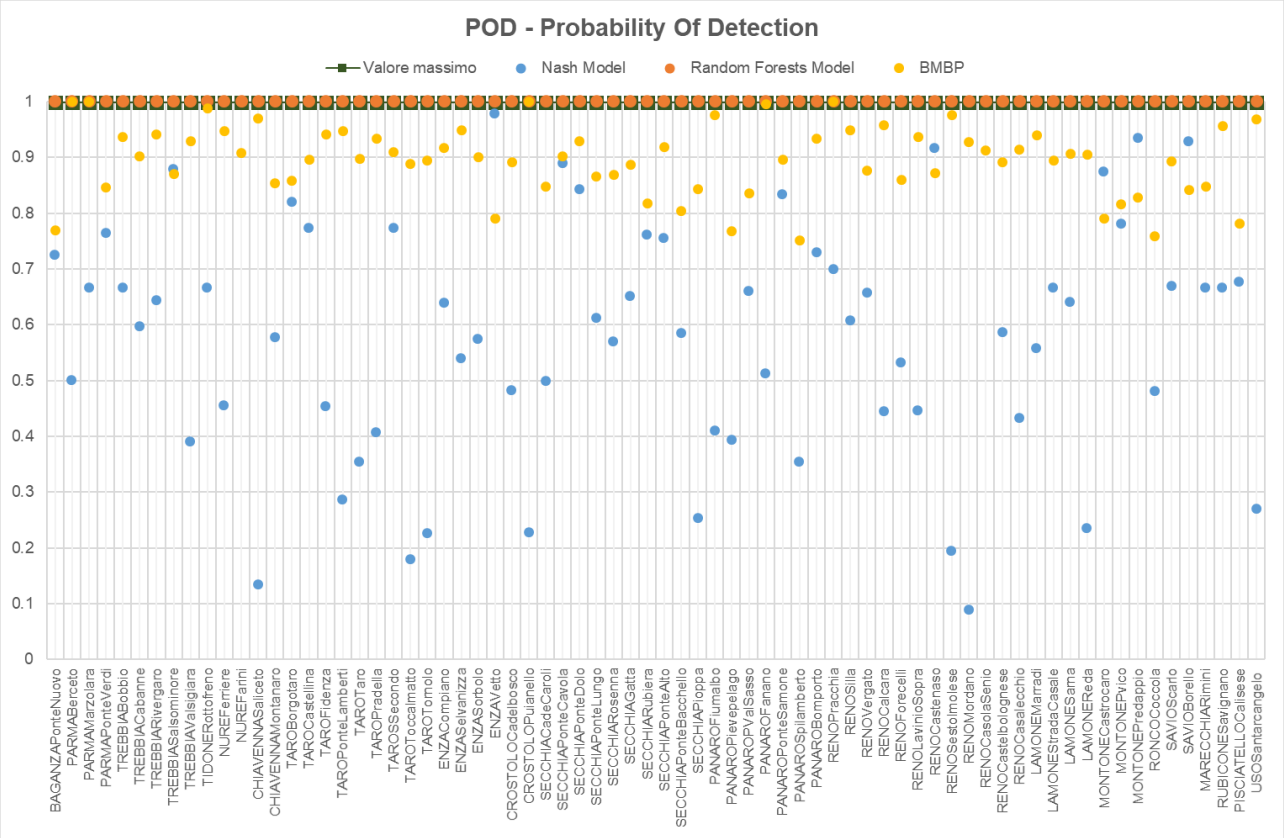


Figura 38 Risultati confronto indice POD

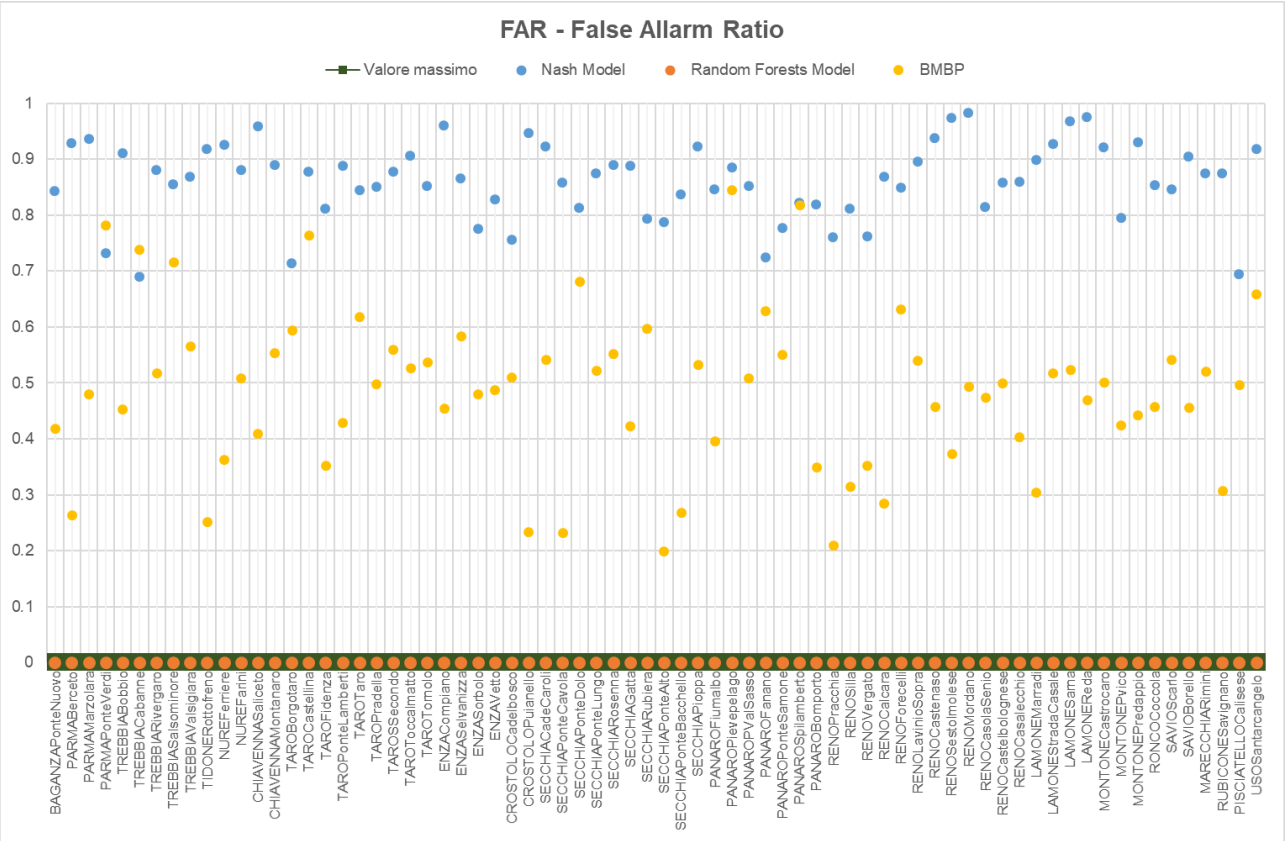


Figura 39 Risultati confronto indice FAR

La tecnica "*BMBP – Bayesian Multivariate Binary Predictor*" presenta dei risultati migliori rispetto al Modello di Nash e confrontabili con il metodo Random Forests; sebbene quest'ultimo risulti più performante soprattutto nel caso dei "falsi allarmi".

4 PREVISIONE DELLE PIENE FLUVIALI NEL BACINO DEL FIUME PO

4.1 ULTERIORI CONFRONTI

Il Fiume Po è il più importante corso d'acqua italiano, sia per lunghezza del canale principale (650 km) sia per estensione del bacino idrografico superficiale (oltre 71.000 km^2). Il bacino si estende a sud delle Alpi ed a nord dell'Appennino Settentrionale e presenta caratteristiche idrologiche e geomorfologiche eterogenee.

Nella parte di sinistra idraulica dell'asta principale, l'estensione del bacino è di circa 43000 km^2, di cui 16000 km^2 sono regolati dai grandi laghi della Lombardia i quali occupano un'area di 890 km^2. L'area in destra idraulica presenta un'estensione di circa 27000 km^2. La porzione alpina ricoperta da ghiacciai perenni si estende per circa 600 km^2.

Afferiscono al bacino del Po le regioni italiane: Liguria, Piemonte, Valle d'Aosta, Lombardia, Trentino-Alto Adige, Veneto, Emilia-Romagna e Toscana; a queste si aggiunge una porzione della Svizzera e diverse valli francesi. Complessivamente sul bacino del Po si contano 3210 comuni.

Oltre ad essere la principale area geografica italiana per estensione, il bacino del Fiume Po è anche l'area economica più importante del territorio nazionale: il Prodotto Interno Lordo (PIL) è pari a circa il 40% di quello nazionale. Ciò è dovuto alla presenza di importanti industrie, attività agricole e zootecniche. Per questi motivi, il bacino del Fiume Po è tra i più monitorati, analizzati e controllati in Europa (Leoni, 2014).

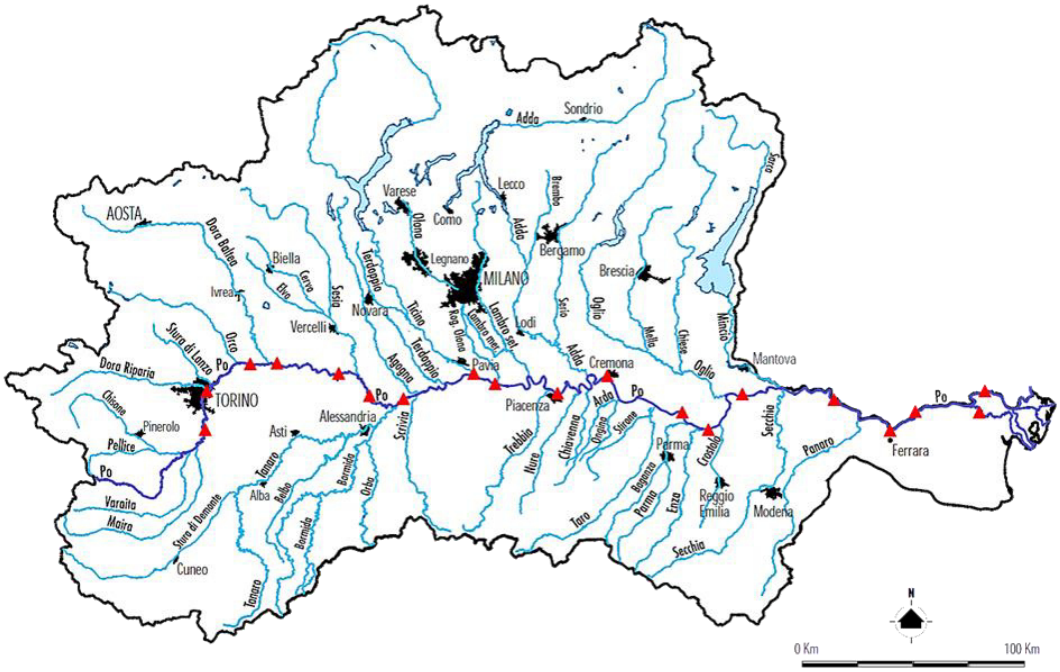


Figura 40 Bacino del fiume Po: asta principale e suoi affluenti. Fonte: Autorità di Bacino del Fiume Po

4.1.1 IL BACINO DEL FIUME PO

Il principale fiume italiano nasce a Pian del Re, ai piedi del Monte Monviso a 2022 m sul livello del mare e scorre da ovest verso est fino a sfociare nel Mare Adriatico con un ampio delta.

Nel primo tratto, tra Moncalieri (Piemonte) e Valenza (Piemonte), il fiume scorre tra le colline torinesi e il Monferrato: è delimitato dai rilievi morfologici a destra e dai grandi accumuli alluvionali delle conoidi a sinistra, generati dagli affluenti di sinistra. A Isola S. Antonio (Piemonte), il Fiume Po raccoglie le acque del Fiume Tanaro: fin qui, l'asta fluviale ha una lunghezza di circa 270 km, con un bacino idrografico di 25320 km^2.

Dalla confluenza del fiume Tanaro al delta, per circa 375 km il Fiume Po è contenuto da continue strutture arginali artificiali. Il regime di flusso è influenzato dalle condizioni idrologiche e dalla sistemazione idraulica degli affluenti, nonché dai lavori di difesa e sistemazione direttamente effettuati sul fiume stesso.

I principali affluenti del Fiume Po sono i corsi d'acqua piemontesi (Dora Baltea, Sesia e Tanaro) e lombardi (Ticino, Adda e Olona), mentre il contributo degli affluenti emiliani è modesto.

Nella prima sezione, tra il Fiume Tanaro e il Fiume Ticino, il Po è caratterizzato da una pendenza dell'ordine di 0,35 ‰. La confluenza del Fiume Po con il Fiume Ticino genera una trasformazione del regime idrico che acquisisce connotati tipicamente fluviali, condizionato dall'approvvigionamento idrico regolamentato, con un notevole contributo glaciale e un'assenza di trasporto solido. La pendenza media diminuisce a 0,18 ‰, quindi gradualmente e progressivamente raggiunge un valore di 0,14 ‰ all'altezza di Revere-Ostiglia (Lombardia, MN). In questa porzione dell'asta fluviale, il livello idrometrico varia di circa 10 metri tra condizioni normali e condizioni di piena.

Le arginature principali sono continue su entrambi i lati già dalla città di Torino e si estendono fino alla foce, per un totale di 1100 km, presentando un percorso molto irregolare; realizzate a più riprese nel corso degli

anni sono tra loro molto diverse e separate da una distanza variabile tra 1 km ed oltre 4 km. L'elevata distanza tra gli argini di sponda destra e quelli di sponda sinistra genera un'ampia area lungo il corso d'acqua che svolge essenzialmente la funzione di laminazione durante gli eventi di piena principali (piane alluvionali chiuse o golene).

Dalla valle di Revere-Ostiglia al delta, l'alveo viene incanalato per alcuni tratti tra gli argini, con una distanza inferiore ai 500 m. Inoltre, non riceve alcun affluente, ad eccezione del fiume Panaro.

Fino alla fine del secolo scorso, a partire dal Ponte della Becca (regione Piemonte, PV) il sistema arginale non era completamente chiuso e il Po e i suoi affluenti inondavano la pianura circostante con le acque di piena. I continui allagamenti della parte terminale, in prossimità del delta, rendevano quest'area più simile ad un lago piuttosto che al tratto finale di un corso d'acqua naturale.

Attualmente, le banche del Po sono state estese ai suoi numerosi affluenti, per una lunghezza complessiva di oltre 1500 km; questa condizione raggiunta è particolarmente critica e comporta una difficile e delicata gestione, soprattutto in concomitanza dell'arrivo delle piene degli affluenti alla foce e del passaggio della piena del Po.

Il regime fluviale del Po è misto tra il tipo alpino (eventi alluvionali in tarda primavera ed estate, con lunghi periodi siccitosi in inverno) e il tipo appenninico (eventi di piena primaverili ed autunnali con prolungata siccità estiva).

Le piene fluviali sono generalmente concentrate nei mesi autunnali a causa delle piogge intense e persistenti di matrice atlantica, con la quota neve ancora molto elevata; occasionalmente possono verificarsi anche in inverno e con maggior probabilità in primavera, in concomitanza con la fusione nivale.

Nel corso degli anni sono state registrate inondazioni massicce e talvolta devastanti: l'evento alluvionale più severo nella storia recente è sicuramente quello del 1951 (colmo di piena superiore ai 10000 m^3/s e multipla rottura arginale con conseguente inondazione del Polesine); mentre il più importante, in termini di volume e durata, è quello del 2000.

4.2 DATI IDRO-METEOROLOGICI E SISTEMA MODELLISTICO

A tutela degli aspetti economici e sociali da sempre di primo ordine, il bacino del fiume Po è monitorato dalla fine del XIX secolo: la sezione di chiusura idrometrica di Pontelagoscuro (FE) ha la più lunga serie storica di monitoraggio di tutte le sezioni idrometriche italiane.

Attualmente, il bacino è monitorato in tempo reale da una densa di rete di strumentazione ed è dotato delle più avanzate tecniche di previsione.

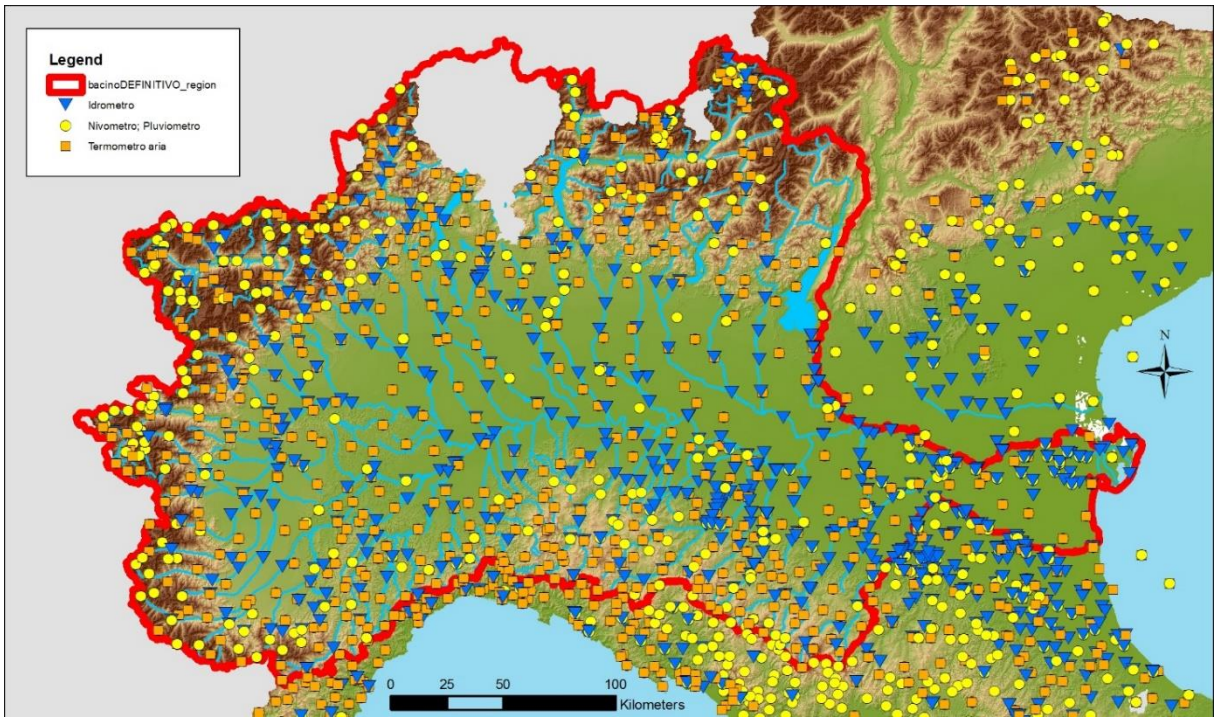


Figura 41 Punti di monitoraggio della rete del bacino del fiume Po

Le principali caratteristiche del sistema di previsione degli eventi di piena *FEWS-Po* sono state introdotte nel paragrafo 3.3. Per il bacino del Po, è stato inoltre sviluppato un sistema denominato *DEWS* ("*Drought Early Warning System*"). Questo sistema viene utilizzato durante i periodi di siccità per l'analisi e la valutazione degli eventi di magra.

DEWS-Po è caratterizzato da una struttura molto simile a *FEWS-Po*; oltre ad utilizzare le informazioni provenienti dalla rete di monitoraggio è costituito da catene modellistiche specifiche per gli eventi di siccità: le previsioni idro-meteorologiche coprono un arco temporale più lungo, fino a 15 giorni, concentrandosi su aspetti e caratteristiche tipiche dei periodi di magra fluviale.

È stato possibile sviluppare questo secondo caso di studio per la stima dell'incertezza predittiva in funzione dei dati osservati e previsti, soprattutto in condizioni di piena, grazie alle numerose informazioni raccolte negli anni dalla rete di monitoraggio ed a quelle prodotte, ed archiviate, dai modelli meteorologici ed idrologici. La finestra temporale considerata per questo caso studio comprende il periodo tra il 2000 e il 2014.

Le informazioni utilizzate per lo sviluppo del modello, basato sul metodo *Random Forests*, sono:

- Dati osservati: altezze idrometriche e portate osservate in continuo dalla rete di monitoraggio dal 2000 al 2014
- Dati previsionali: portate previste dalla catena idrologica-idraulica *MIKE11 NAM/HD*; disponibili per evento dal 2000 al 2005, in continuo dal 2005 al 2009 e dal 2013 al 2014.

Il modello è stato sviluppato per le principali sezioni idrometriche del Fiume Po: Piacenza, Cremona, Boretto, Borgoforte e Pontelagoscuro.

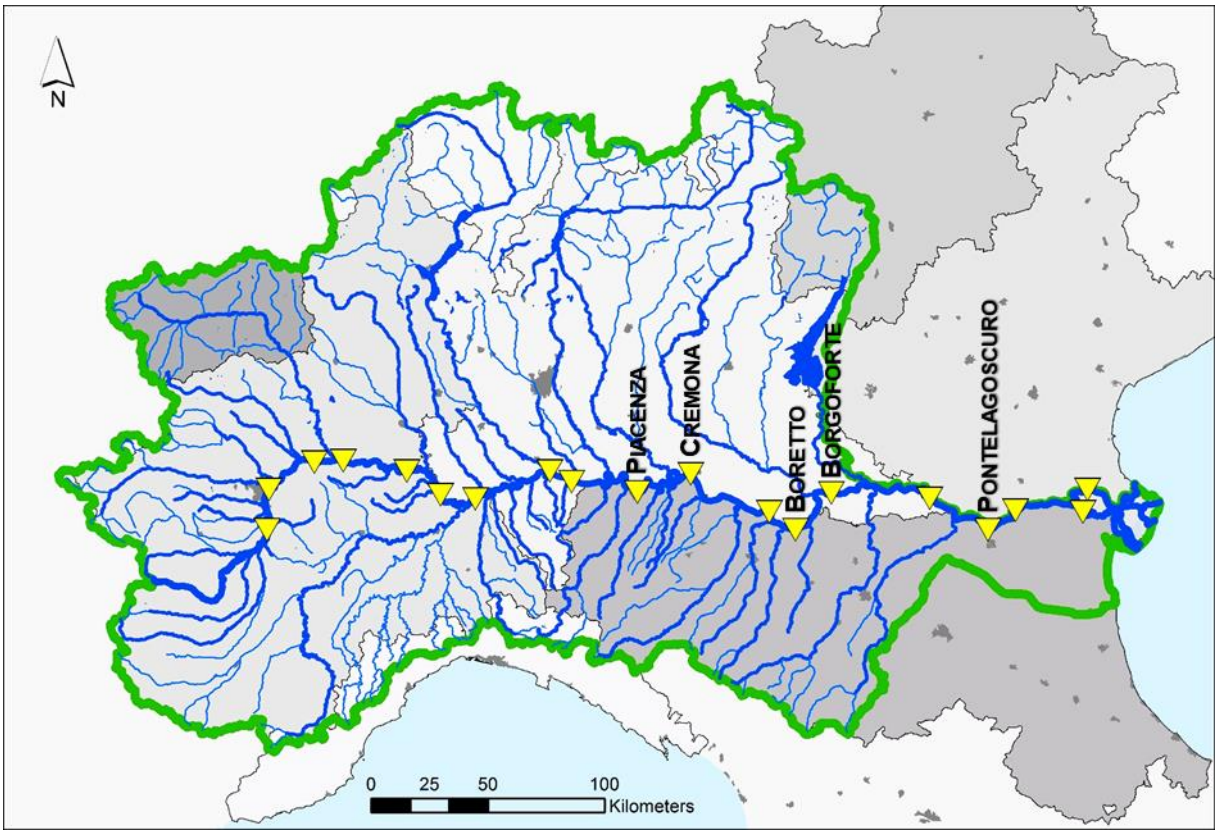


Figura 42 Sezioni idrometriche strumentate presenti lungo l'asta principale del fiume Po

4.2.1 EVENTI DI PIENA DEL FIUME PO

Le piene del fiume Po sono più prolungate (fino a 72 ore) rispetto a quelle dei fiumi dell'Emilia-Romagna (massimo 12 ore); i tempi di traslazione dell'onda di piena variano in funzione della provenienza dei contributi maggiori: considerando la sezione di chiusura di Pontelagoscuro, le onde di piena più veloci sono generate dai contributi dei fiumi lombardi, mentre più lente sono le piene di origine piemontese; i contributi emiliani risultano essere generalmente modesti e non destano particolare preoccupazione se considerati singolarmente.

Tempi di traslazione del colmo di piena relativamente lunghi permettono di valutare i possibili effetti con maggiore attenzione e consentono interventi preventivi; d'altra parte, i volumi che transitano nel canale principale sono elevati e possono ledere le strutture arginali presenti. Ciò comporterebbe l'inondazione delle aree circostanti con conseguenti danni ingenti.

Generalmente, gli eventi di piena del fiume Po sono caratterizzati da un unico picco di piena (esempio: novembre 2016) ma in alcuni casi questi possono essere caratterizzati da numerosi picchi di piena (ad esempio, in ottobre e novembre 2014); in tal caso: l'elaborazione previsionale risulta essere più complessa, la sollecitazione arginale è sensibilmente più marcata e con essa la pericolosità associata all'evento.

Ad esempio, l'evento di piena verificatosi nel novembre 2016 è stato caratterizzato da un solo colmo di piena: generatosi in Piemonte a seguito delle intense e persistenti precipitazioni, il colmo ha raggiunto la sezione di chiusura del bacino (Pontelagoscuro) tre giorni dopo. Di seguito sono stati riportati i tempi di traslazione del

colmo di piena per le sezioni di riferimento e le relative soglie idrometriche di allertamento definite dalla Protezione Civile (già trattate nel paragrafo 3.4.1).

Sezioni idrometriche	H_{max} [m]	Q_{max} [m^3/s]	Data Ora	Traslazione	Soglie idrometriche di allertamento		
					L1	L2	L3
Isola S. Antonio	8.55	9900	26/11 02:30	-	5.5	6.5	8.0
Ponte Becca	5.75	7900	26/11 15:30	+13h	3.5	4.5	5.5
Spessa Po	6.68	8100	26/11 20:00	+17h 30’	4.5	5.5	6.5
Piacenza	7.54	7500	27/11 04:20	+1d 1h 50’	5.0	6.0	7.0
Cremona	3.33	6900	27/11 14:20	+1d 11h 50’	2.2	3.2	4.2
Boretto	6.33	5800	28/11 12:00	+2d 9h 30’	4.5	5.5	6.5
Borgoforte	6.76	5900	28/11 20:10	+2d 17h 40’	5.0	6.0	7.0
Sermide	8.18	6000	29/11 07:00	+3d 4h 30’	7.0	8.0	9.0
Pontelagoscuro	1.62	5700	29/11 10:00	+3d 7h 30’	0.5	1.3	2.3

Tabella 15 Evento di piena novembre 2016: durata evento e traslazione del colmo. Fonte: ARPAE Emilia-Romagna

Un importante evento di piena multi-picco si è verificato nel novembre 2014: ripetute perturbazioni atlantiche hanno interessato il Nord Italia generando un evento alluvionale con tre colmi di piena distinti. L'evento di piena è durato per quasi un mese e ha compromesso le aree circostanti e la struttura degli argini: in diversi punti il Fiume Po ha rotto gli argini inondando le aree circostanti, soprattutto nella provincia di Reggio Emilia; numerosi anche i sormontamenti degli argini minori.

Il grafico mostra l'effettiva intensità dell'evento nelle principali sezioni del fiume Po.

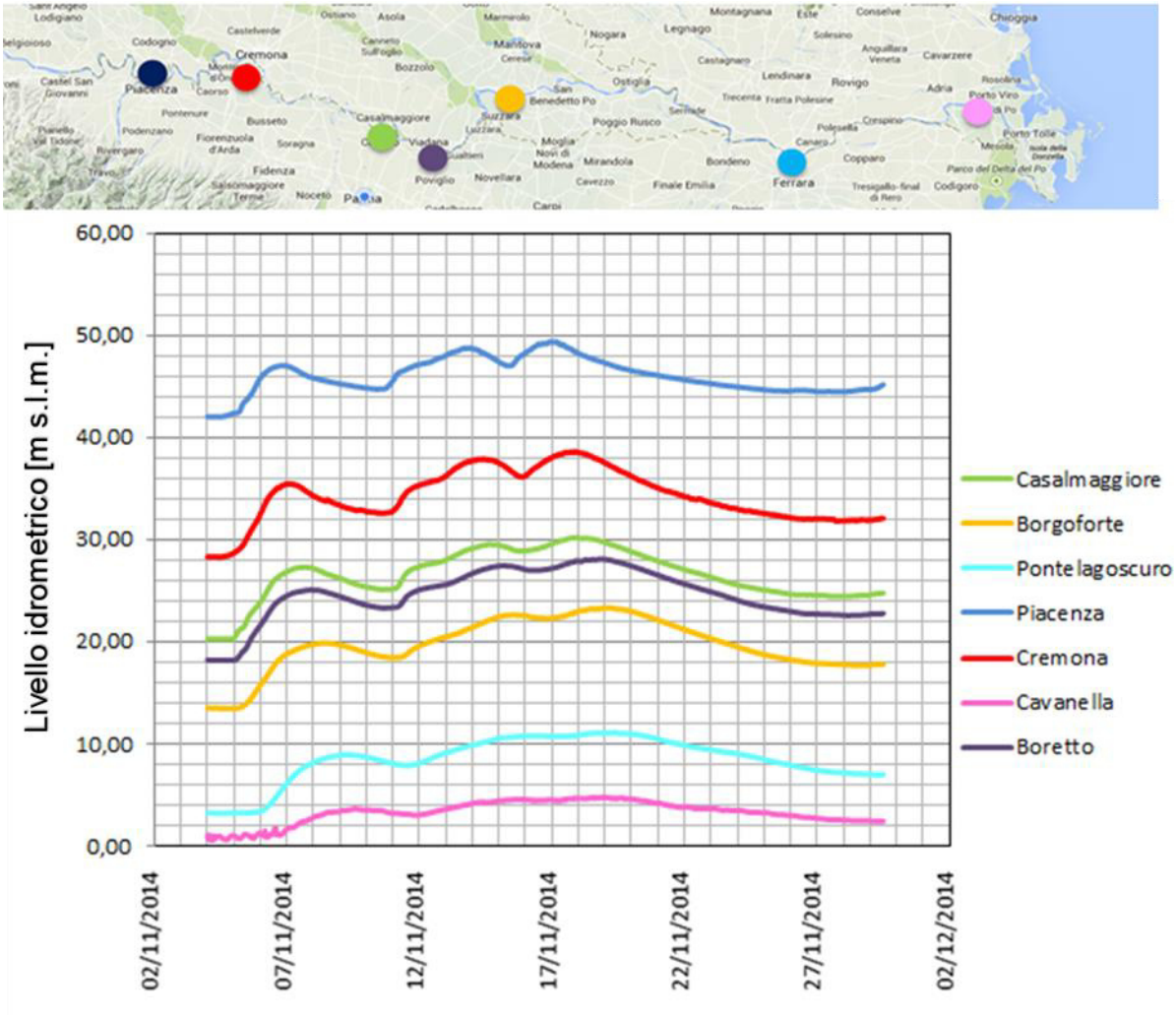


Figura 43 Sezioni idrometriche di riferimento e relativi livelli idrometrici registrati durante l'evento di piena di novembre 2014. Fonte: ARPAE Emilia-Romagna

4.3 *Random Forests* per gli eventi di piena del Fiume Po

Le numerose informazioni raccolte nel corso degli anni e gli studi condotti sul bacino del Po hanno permesso di acquisire numerose e preziose informazioni. L'oggetto di questo caso di studio è l'elaborazione di un modello previsionale probabilistico basato sulla tecnica del *Random Forests*, calibrato sulle principali sezioni dell'asta principale.

A differenza del caso di studio relativo ai corsi d'acqua dell'Emilia-Romagna; in questo caso, l'algoritmo utilizza oltre ai dati osservati anche i dati di previsione forniti dalla catena modellistica idrologico-idraulica "*MIKE 11 NAM/HD*".

Di seguito verranno illustrati i principali passaggi che hanno portato alla realizzazione del modello previsionale.

4.3.1 Training set

Per tutte le sezioni idrometriche considerate, il *training set* disponibile per l'addestramento è relativo al periodo 2000-2014 e si compone di:

a) Livelli idrometrici (o portate) orari osservati

b) Livelli idrometrici (o portate) previsti dalla catena modellistica idrologico-idraulica (*MIKE11 NAM/HD*) sulla base dei dati meteorologici del modello di previsione meteorologica (*COSMO*).

c) Massimo valore idrometrico (o massima portata) osservato nelle successive 12, 24 o 36 ore, rispetto all'orario di riferimento; valore espresso in classi definite in continuo.

4.3.1.1 PORTATE ORARIE (LIVELLO IDROMETRICI) OSSERVATE

Considerando le caratteristiche idrauliche del fiume Po per comporre il *training set* sono state utilizzate le portate orarie osservate (eventualmente, è possibile impostare il modello anche con i livelli idrometrici) nelle sei ore precedenti all'orario di riferimento: arco temporale sufficiente per definire la progressione idrometrica, condizione necessaria per l'apprendimento dell'algoritmo *Random Forests*.

Inoltre, in riferimento ai valori osservati, nel *training set* è stata aggiunta un ulteriore informazione relativa all'andamento dei valori di portata rispetto all'ora precedente:

$$se\ Q_{oss_{0h}} > Q_{oss_{-1h}} \rightarrow 1\ (andamento\ crescente)$$

$$se\ invece\ Q_{oss_{0h}} < Q_{oss_{-1h}} \rightarrow 2\ (andamento\ decrescente)$$

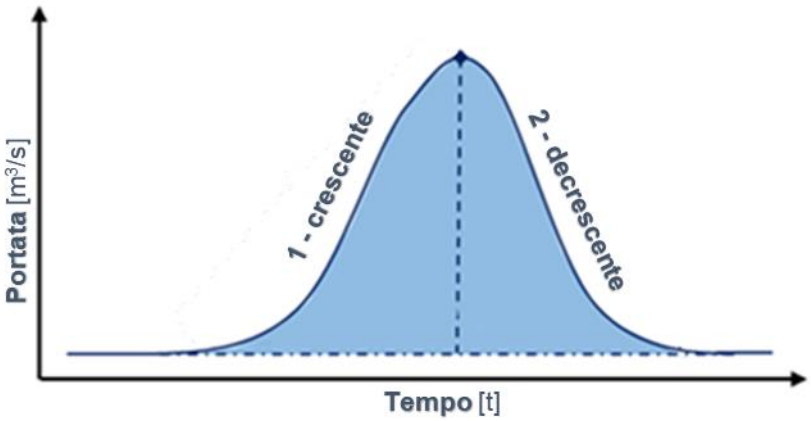


Figura 44 Rappresentazione grafica delle fasi principali dell'idrogramma di piena.

L'identificazione dell'andamento delle portate (o del livello idrometrico), in crescita o in decrescita, e l'inserimento di questa informazione sia nel *training set* sia nel data set operativo, è un valore aggiunto che consente all'algoritmo di escludere le condizioni opposte, evitando una condizione di "confusione" in fase di apprendimento ma soprattutto operativa. Pertanto le istanze "crescenti" (valore 1) verranno valutate con solo con le istanze crescenti presenti nel *training set*; medesima modalità per le istanze "decrescenti" (valore 2).

4.3.1.2 PORTATE ORARIE (LIVELLO IDROMETRICI) PREVISTE

Sono stati considerati i valori di portata prevista dalla catena modellistica *MIKE 11 NAM/HD*; elaborati giornalmente nel sistema *FEWS-Po* alle ore 06:00. Per la calibrazione del modello, sono state scelte le seguenti finestre temporali: +12 ore, +24 ore e +36 ore.

La scelta è stata guidata dalle esigenze operative di valutazioni e interventi a breve, medio e lungo termine; tempistiche relative al bacino del Fiume Po. In funzione della finestra temporale identificata, per la calibrazione del modello, è stato identificato il massimo valore di portata osservato; denominato nel set di allenamento come "classe".

4.3.1.3 DEFINIZIONE DELLE CLASSI DI RIFERIMENTO

Ad ogni istanza (riga) del *training set* è stata aggiunto il valore della classe di appartenenza. Le classi vengono definite in funzione dell'ultimo valore di portata osservato (t0); ogni classe rappresenta un *range* di portata. La classe aggiunta al *training set* è relativa al massimo valore di portata osservata nelle N-ore successive (12, 24 o 36 ore); espresso nella relativa classe di appartenenza.

Per ogni istanza vengono generate diciannove classi crescenti e diciannove classi decrescenti in funzione dell'ultimo valore di portata osservato; il numero delle classi generate è stato definito arbitrariamente.

Per definire i valori di portata delle classi sono stati utilizzati due diversi coefficienti: il primo, un coefficiente moltiplicativo, per le classi crescenti e il secondo, un coefficiente di riduzione per le classi decrescenti. Il primo valore dell'insieme crescente è una funzione del valore osservato moltiplicato per un coefficiente definito del 2,5%. Il primo valore dell'insieme decrescente è una funzione del valore osservato ridotta di un coefficiente definito del 2,5%.

I seguenti valori delle classi si basano sui valori precedenti, incrementati o ridotti in funzione del coefficiente indicato. Per l'identificazione delle classi è stata utilizzata la lettera **T** (T="*Thresholds*", in inglese). Le due "classi estreme", T_{max} e T_{min}, sono definite in base alle seguenti condizioni:

T_{MAX}

$$Se\ T_{UP18} + 2.5\% > Q_{MAX\_oss}\ nella\ finestra\ temporale\ considerata \rightarrow \boldsymbol{T_{MAX}} = T_{UP18} + 2.5\%$$

$$Invece\ se\ \ T_{UP18} + 2.5\% < Q_{MAX\_oss}\ nella\ finestra\ temporale\ considerata \rightarrow \boldsymbol{T_{MAX}} = Q_{MAX\_oss}$$

Dove: T_{UP18}= 18ª soglia crescente

Q_{MAX_oss} = Massimo valore di portata osservato nelle N-ore successive

T_{MIN}

$$SE\ T_{DW18} - 2.5\% < Q_{MIN\_oss}\ nella\ finestra\ temporale\ considerata\ \ \rightarrow \boldsymbol{T_{MIN}} = T_{DW18} - 2.5\%$$

$$Invece\ Se\ \ T_{DW18} - 2.5\% > Q_{MIN\_oss}\ nella\ finestra\ temporale\ considerata\ \ \rightarrow \boldsymbol{T_{MIN}} = Q_{MIN\_oss}$$

Dove: T_{DW18}= 18ª soglia decrescente

Q_{MIN_oss} = Minimo valore di portata osservato nelle N-ore successive

Le classi comprese tra T_{MAX} e T_{MIN} sono automaticamente definite in funzione dell'ultimo valore osservato, secondo i criteri definiti e riportati nella tabella seguente:

Insieme di soglie crescenti	$T_{MAX}=T_{UP18}$+2.5% *oppure* Q_{MAX_oss}
	$T_{UP18}=T_{UP17}$+2.5%
	$T_{UP17}=T_{UP16}$+2.5%
	$T_{UP16}=T_{UP15}$+2.5%
	$T_{UP...}=T_{UP...}$+2.5%
	$T_{UP4}=T_{UP3}$+2.5%
	$T_{UP3}=T_{UP2}$+2.5%
	$T_{UP2}=T_{UP1}$+2.5%
	$T_{UP1}=T_{OBS}$+2.5%
Ultimo valore di portata osservato	T_OSS
Insieme di soglie decrescenti	$T_{DW1}=T_{OBS}$-2.5%
	$T_{DW2}=T_{DW1}$-2.5%
	$T_{DW3}=T_{DW2}$-2.5%
	$T_{DW4}=T_{DW3}$-2.5%
	$T_{DW...}=T_{DW...}$-2.5%
	$T_{DW16}=T_{DW15}$-2.5%
	$T_{DW17}=T_{DW16}$-2.5%
	$T_{DW18}=T_{DW17}$-2.5%
	$T_{MIN}=T_{DW18}$-2.5% *oppure* Q_{MIN_oss}

Tabella 16 Definizione delle soglie in funzione dell'ultimo valore di portata osservato

In riferimento ad una sola istanza, il processo di definizione delle classi è rappresentato schematicamente nell'immagine sottostante:

- Linea azzurra: rappresenta i valori di portata osservata oltre le sei ore precedenti all'orario di riferimento; non più presenti nell'ultima istanza generata e quindi non oggetto di valutazione da parte dell'algoritmo.
- Linea blu: rappresenta i valori di portata osservata entro le sei ore precedenti all'orario di riferimento; costituiscono l'informazione osservativa dell'istanza del training set.
- Rettangoli: ogni rettangolo rappresenta una classe definita come riportato nella tabella 2
- Cerchio rosso: rappresenta il massimo valore di portata previsto per le N-ore successive dalla catena modellistica *MIKE11 NAM/HD*
- Quadrato verde: rappresenta il massimo valore di portata osservato nelle N-ore successive

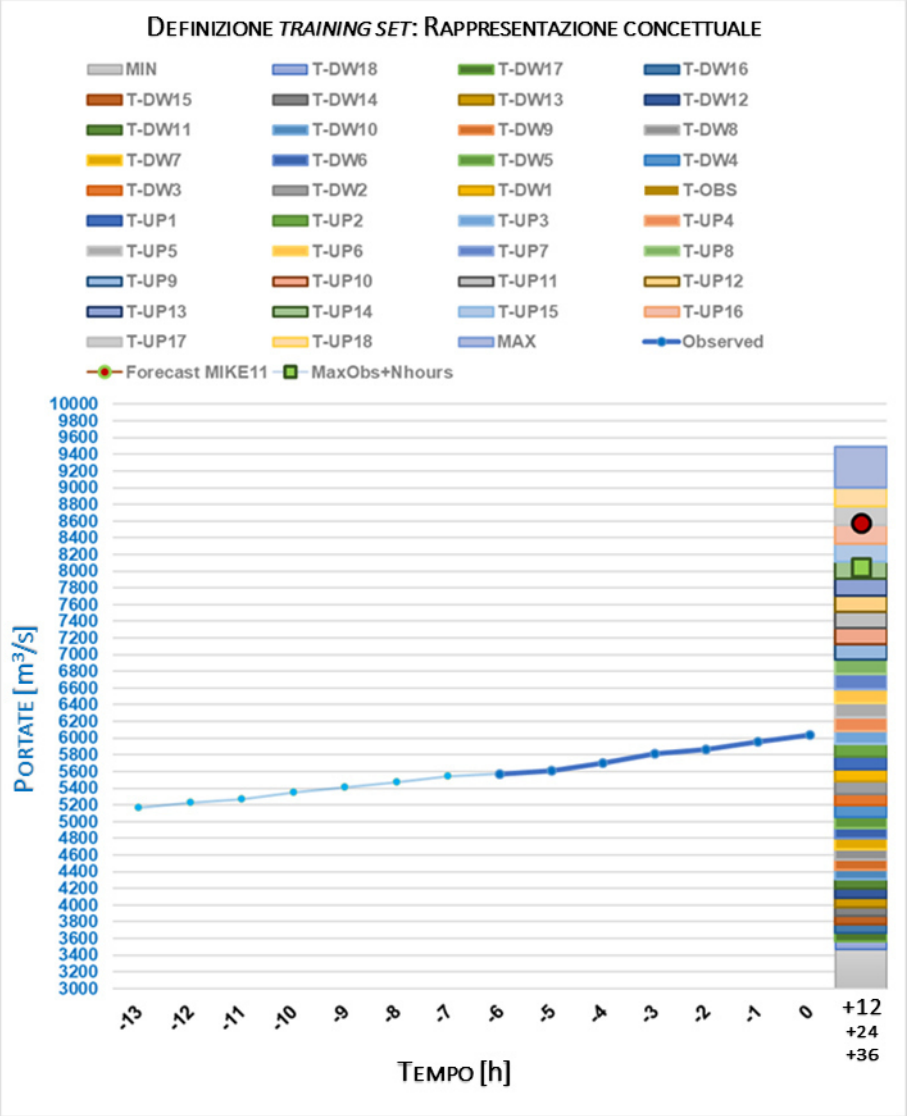


Figura 45 Rappresentazione grafica concettuale della definizione delle classi e degli elementi costituenti il training set.

Nell'istanza di calibrazione, il massimo valore osservato nelle N-ore successive viene sostituito dalla relativa classe; quindi in fase di validazione-operativa, il modello fornirà in previsione la classe, ovvero, un *range* di valori di portata.

Ad esempio: il massimo valore osservato nelle N-ore successive all'orario di riferimento è pari a 8050 m^3/s, che ricade nella classe T_{UP14}; nell'istanza di calibrazione, come condizione di verifica, verrà riportata la classe T_{UP14}

Massimo valore di portata osservato nelle N-ore successive: $Q_{MAXoss_{+Nh}} = 8050 \frac{m^3}{s}$

Classe in cui rientra $Q_{MAXoss_{+Nh}}$: $T_{UP14} = 7950 \div 8150 \ \frac{m^3}{s}$

$$Q_{MAXobs_{+Nh}} \rightarrow Class_{T_{UP14}}$$

Ovvero, considerando la medesima istanza in operatività (quindi privata dell'informazione relativa alla massima portata osservata nelle N-ore successive) il modello in previsione, se corretto, dovrà attribuire la massima probabilità di accadimento a $Class_{T_{UP14}}$.

4.3.1.4 SEZIONI IDROMETRICHE PRECEDENTI

Ad eccezione di Piacenza, nei *training set* delle altre sezioni idrometriche sono stati inseriti anche i valori di portata osservata nelle ore precedenti nelle sezioni a monte rispetto a quella in esame. L'aggiunta di queste informazioni consente di formulare una previsione più robusta in riferimento ai volumi e durata dell'evento.

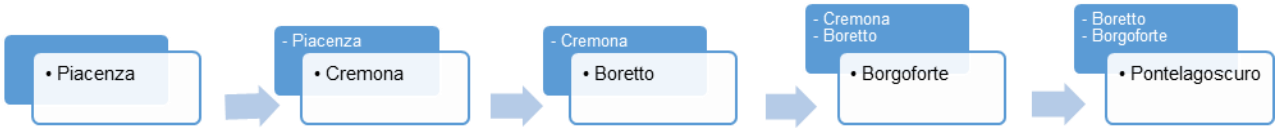


Figura 46 Sezioni idrometriche considerate: sfondo bianco per la sezione di riferimento; sfondo blu per le sezioni precedenti contenute nel *training set* della sezione di riferimento.

Ogni singola istanza del training set sarà caratterizzata dalla seguente struttura:

Sezione considerata									Sezione precedente						Classi
Q_{oss0h}	$Q_{fkstmax+Nh}$	Asc. Disc.	Q -1h	Q -2h	Q -3h	Q -4h	Q -5h	Q -6h	Q -1h	Q -2h	Q -3h	Q -4h	Q -5h	Q -6h	$Class_T$
Portata osservata	Massimo valore di portata previsto nelle N-Ore successive dalla catena modellistica MIKE11 NAM/HD	Andamento idrometrico crescita/decrescita	Portate osservate nelle ore precedenti nella sezione in esame						Portate osservate nelle ore precedenti nella sezione precedente a quella in esame						Classe associata alla massima portata osservata nelle N-Ore successive [solo nel *training set*]

Tabella 17 Struttura delle istanze del *training set*

4.3.2 IMPLEMENTAZIONE DEL METODO *RANDOM FORESTS*

Il modello basato sulla tecnica del *Random Forests*, sviluppato per questo caso di studio, è stato calibrato sulle principali sezioni idrometriche del fiume Po: Piacenza, Cremona, Boretto, Borgoforte e Pontelagoscuro.

Rispetto al modello realizzato per i principali corsi d'acqua dell'Emilia-Romagna, sono stati apportati alcuni miglioramenti necessari per adattare il modello alle nuove finalità richieste e congiuntamente per sfruttare al meglio le potenzialità della tecnica del *Random Forests*.

- Le soglie di riferimento, identificate con il termine "classi", vengono ricalcolate per ogni intervallo temporale (ogni istanza) in funzione dell'ultimo valore di portata osservato. (punto C del paragrafo precedente)

- In funzione dell'ultimo valore di portata osservato, viene prodotto un sottoinsieme (*training subset*) del *training set* iniziale; vengono così ridotti ulteriormente i tempi di elaborazione senza però alterare le performance del modello.

In riferimento ai criteri di identificazione del sottoinsieme di addestramento (*training subset*), è stata identificata una condizione tale per cui tutte le istanze presenti nel *training set* originario aventi un valore di portata osservato inferiore alla metà dell'ultimo valore di portata osservato (Q_{UNDER}) non sono incluse nel nuovo sottoinsieme:

Condizione di non inclusione nel *training subset*:

$$Q_{oss} < Q_{UNDER} \text{ dove } Q_{UNDER} = \frac{Q_{OSS_{0h}}}{2}$$

Pertanto, il *training subset* generato conterrà una serie di istanze aventi tutte un valore di portata osservato maggiore di Q_{UNDER}.

Congiuntamente, ad ogni istanza appartenente al *training subset* viene associata una classe, in funzione di quanto presentato nel punto C del paragrafo precedente; essendo le classi definite in funzione dell'ultimo valore di portata osservato, per le istanze del *training subset* la classe attribuita risulterà variabile ad ogni passo temporale.

Ad esempio, si ipotizzi in fase di calibrazione che l'ultima istanza in ingresso è caratterizzata da un valore di portata osservato Q_{oss} = 6500 m^3/s, in funzione della quale vengono definite tutte le classi; il massimo valore di portata osservato nelle N-ore successive è $Q_{MAXoss+Nh}$ = 7500 m^3/s, a cui corrisponde la classe T_{UP5}.

Congiuntamente, una generica istanza del *training subset* è caratterizzata da un valore di portata osservato Q_{oss} = 5500 m^3/s, anche in questo caso le classi vengono definite in funzione di Q_{oss}; il massimo valore di portata osservato nelle N-ore successive è $Q_{MAXoss+Nh}$ = 7500 m^3/s, come l'esempio precedente. La classe corrispondente, in questo secondo esempio, è però differente rispetto al primo e risulta essere T_{UP12}; in quanto generata in funzione di un valore osservato più basso. Tale procedimento viene condotto per tutte le istanze presenti nel training subset.

Queste implementazioni consentono all'algoritmo di elaborare un subset ridotto rispetto a quello iniziale e di poter valutare l'ultima istanza sulla base di un training subset elaborato in funzione di essa; il vantaggio di un *training subset* ridotto diventa importante per gli eventi più importanti (sopra la soglia 1 di allerta di Protezione Civile), dove la numerosità del *training subset* si riduce fino ad un ordine di grandezza rispetto al *training set*.

4.3.2.1 RANDOM FORESTS "ON THE FLY"

In riferimento ai paragrafi di presentazione del metodo *Random Forests* (Capitolo 2), l'algoritmo per poter classificare correttamente una nuova istanza (istanza di validazione o operativa) deve necessariamente aver memorizzato un albero decisionale, elaborato su uno specifico *training set*; più semplicemente, l'algoritmo prima deve apprendere e solo successivamente è in grado di valutare le nuove informazioni in ingresso.

Le nuove tecniche di ottimizzazione, implementate in questo caso studio, modificano "in continuo" il *training set*: ad ogni nuova istanza viene prodotto un *training subset* (sottoinsieme del *training set*) e tutte le istanze presenti vengono "riclassificate"; pertanto l'applicazione "classica" del metodo *Random Forests*, sviluppata nel primo caso di studio, risulta essere non compatibile.

Per consentire la corretta implementazione delle nuove tecniche è stato necessario trasformare il metodo Random Forests da statico a dinamico, "*on fly*": successivamente alla creazione del *training subset* e alla "riclassificazione" di tutte le istanze, viene generato l'albero decisionale che apprende dal *training subset* e valuta la nuova istanza. A differenza di quanto visto nel primo caso studio, non verrà prodotto e memorizzato un albero decisionale per ogni sezione idrometrica; ma verrà prodotto un albero decisionale ad ogni nuova istanza per ogni sezione idrometrica. In questo modo si ottiene un ulteriore vantaggio: il *training set* sarà corredato di tutte le nuove istanze disponibili, di conseguenza l'algoritmo calibrerà in continuo l'albero decisionale sfruttando tutte le informazioni disponibili e raccolte fino a quell'istante. Tutto ciò con tempi di calcolo particolarmente contenuti: dall'ingresso della nuova istanza alla sua valutazione, da parte dell'algoritmo *Random Forests*, occorrono circa 120 secondi.

4.3.3 ELABORAZIONE DEL MODELLO PREVISIONALE

Nel paragrafo precedente sono state presentate le nuove tecniche adottate, di seguito verranno illustrate le principali fasi di elaborazione del modello; il processo che consente di ottenere l'informazione di *output* può essere suddiviso in tre fasi cardine:

- Fase 1: elaborazione del training subset
- Fase 2: elaborazione dell'albero decisionale
- Fase 3: valutazione dell'ultima istanza ed output probabilistico

FASE 1: elaborazione del training subset

Il modello riceve in input l'ultimo dato osservato relativo alla portata (o al livello idrometrico); viene quindi elaborata un'istanza avente la stessa struttura delle istanze contenute nel *training set* relativo alla medesima sezione: portata osservata, massima portata prevista dalla catena modellistica , condizione di ascesa o discesa rispetto agli ultimi valori presenti nel training set, livelli o portate osservate alle sezioni precedenti (ad esclusione della sezione di Piacenza). In questo caso, l'istanza sarà priva dell'informazione relativa alle classi (massima portata nelle N-ore successive); in quanto questa sarà l'output del modello.

- Ultimo valore di livello o portata osservato:

Sezione considerata
Q_{oss0h}

Tabella 18 Ultimo valore osservato

- Creazione dell'istanza:

Sezione considerata								Sezioni a monte				
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	…	Q_{oss-5h}	Q_{oss-6h}	Q_{oss0h}	Q_{oss-1h}	…	Q_{oss-5h}	Q_{oss-6h}

Tabella 19 Esempio di un'istanza del *training set*

- Creazione delle classi in funzione dell'ultimo valore osservato (paragrafo 4.3.1.3)

- Elaborazione del training subset in funzione dell'ultimo valore di portata osservato (paragrafo 4.3.2):

Sezione considerata								Sezioni a monte					Classe
…	…	…	…	…	…	…	…	…	…	…	…	…	…
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	…	Q_{oss-5h}	Q_{oss-6h}	Q_{oss0h}	Q_{oss-1h}	…	Q_{oss-5h}	Q_{oss-6h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	…	Q_{oss-5h}	Q_{oss-6h}	Q_{oss0h}	Q_{oss-1h}	…	Q_{oss-5h}	Q_{oss-6h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	…	Q_{oss-5h}	Q_{oss-6h}	Q_{oss0h}	Q_{oss-1h}	…	Q_{oss-5h}	Q_{oss-6h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	…	Q_{oss-5h}	Q_{oss-6h}	Q_{oss0h}	Q_{oss-1h}	…	Q_{oss-5h}	Q_{oss-6h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	…	Q_{oss-5h}	Q_{oss-6h}	Q_{oss0h}	Q_{oss-1h}	…	Q_{oss-5h}	Q_{oss-6h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	…	Q_{oss-5h}	Q_{oss-6h}	Q_{oss0h}	Q_{oss-1h}	…	Q_{oss-5h}	Q_{oss-6h}	T_{UP}/T_{DW}
…	…	…	…	…	…	…	…	…	…	…	…	…	…

Tabella 20 Esempio di una porzione del *training subset*

FASE 2: elaborazione dell'albero decisionale

L'algoritmo *Random Forests* basandosi sul *training subset* elabora una foresta casuale composta da 500 alberi; come riportato nel capitolo 3, una numerosità di 500 alberi risulta essere idonea alla maggior parte dei casi, ciò è stato verificato rapidamente anche in questo caso di studio (la verifica non è stata qui riportata poiché non aggiunge nulla di nuovo a quanto già illustrato precedentemente).

Sebbene i tempi di calcolo siano particolarmente ridotti, il modello è stato ottimizzato con l'intento di ridurli ulteriormente; senza però alterarne le performance. Per valutare nel miglior modo possibile l'ultima istanza in ingresso, il modello prevede un processamento multiplo (*loops*) dell'algoritmo *Random Forests*: l'algoritmo è settato per produrre un numero di alberi decisionali pari al numero di variabili osservate presenti nel *training subset*, il quale viene ridotto prima al minimo e successivamente incrementato ad ogni *loop*; inoltre, per ogni *loop*, ovvero, per ogni albero decisionale generato viene condotta in automatico una valutazione speditiva, a verifica della bontà del processo di calibrazione sul *training subset* utilizzato. Se i risultati di calibrazione ottenuti nell'ultimo *loop* sono peggiori rispetto a quelli del *loop* precedente, il processo di elaborazione si interrompe e l'algoritmo *Random Forests* procede con la valutazione dell'ultima istanza in ingresso, utilizzando l'albero decisionale elaborato nel *loop* precedente.

Il numero massimo di *loops* è pari al numero di valori di portata osservata, oltre all'ultimo valore osservato, presenti nel training subset; ovvero, sei. Di seguito verrà riportato schematicamente il funzionamento operativo dell'algoritmo Random Forests e le variabili utilizzate ad ogni loop:

- *1° loop:* ultimo valore (Q_{oss-1h}) di portata osservato nella sezione considerata ed in quelle a monte (se presenti), massimo valore di portata previsto dalla catena modellistica MIKE11 NAM/HD nelle prossime N-ore, variazione rispetto al valore precedente e classe attribuita all'istanza.

Sezione considerata				Sezioni a monte			Classe
…	…	…	…	…	…	…	…
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-6h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-6h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-6h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-6h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-6h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-6h}	T_{UP}/T_{DW}
…	…	…	…	…	…	…	…

Tabella 21 Esempio di una porzione del *training subset* utilizzato nel 1° loop

Al termine del processo di *training*, l'algoritmo fornisce un valore percentuale relativo alla capacità di calibrazione del modello, denominato "*cap_calibr_1*"; sia l'albero decisionale sia il valore "*cap_calibr_1*" vengono memorizzati dal modello.

- *2° loop:* ultimo (Q_{oss-1h}) e penultimo (Q_{oss-2h}) valore di portata osservati nella sezione considerata ed in quelle a monte (se presenti), massimo valore di portata previsto dalla catena modellistica MIKE11 NAM/HD nelle prossime N-ore, variazione rispetto al valore precedente e classe attribuita all'istanza.

Sezione considerata					Sezioni a monte			Classe
…	…	…	…	…	…	…	…	…
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-2h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-2h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-2h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-2h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-2h}	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	Q_{oss-1h}	Q_{oss-2h}	Q_{oss0h}	Q_{oss-1h}	Q_{oss-2h}	T_{UP}/T_{DW}
…	…	…	…	…	…	…	…	…

Tabella 22 Esempio di una porzione del *training subset* utilizzato nel 2° loop

Al termine del processo di *training*, l'algoritmo fornisce un valore percentuale relativo alla corretta capacità di calibrazione del modello, denominato "*cap_calibr_2*"; questo valore viene confrontato con il precedente "*cap_calibr_1*". Se il precedente valore (*cap_calibr_1*) è migliore di quello attuale

(*cap_calibr_2*), il processo si arresta e l'albero decisionale elaborato nel 2° *loop* non viene immagazzinato. L'istanza verrà classificata utilizzando l'albero decisionale elaborato nel 1° *loop*. Se invece il valore attuale è migliore di quello precedente, l'albero decisionale elaborato in questo *loop* sovrascrive il precedente, cap_calibr_2 sostituisce cap_calibr_1 ed il processamento procede con il 3° *loop*.

☒ Se $cap\_calibr\_1 > cap\_calibr\_2$ → processo di elaborazione interrotto e valutazione istanza con albero decisionale generato nel 1° *loop*.

☑ Se $cap\_calibr\_1 < cap\_calibr\_2$ → processo di elaborazione continua; l'albero decisionale sovrascrive quello generato nel precedente *loop* e cap_calibr_2 sostituisce cap_calibr_1.

- *3° loop:* ultimo ($Q_{oss\text{-}1h}$), penultimo ($Q_{oss\text{-}2h}$) e terzultimo ($Q_{oss\text{-}3h}$) valore di portata osservati nella sezione considerata ed in quelle a monte (se presenti), massimo valore di portata previsto dalla catena modellistica MIKE11 NAM/HD nelle prossime N-ore, variazione rispetto al valore precedente e classe attribuita all'istanza.

Sezione considerata						Sezioni a monte				Classe
...	...	...	...	...	...	...	...	...	...	...
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	Q_{oss0h}	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	Q_{oss0h}	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	Q_{oss0h}	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	Q_{oss0h}	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	Q_{oss0h}	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	T_{UP}/T_{DW}
Q_{oss0h}	$Q_{fkstMAX+Nh}$	Asc./Disc.	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	Q_{oss0h}	$Q_{oss\text{-}1h}$	$Q_{oss\text{-}2h}$	$Q_{oss\text{-}3h}$	T_{UP}/T_{DW}
...	...	...	...	...	...	...	...	...	...	...

Tabella 23 Esempio di una porzione del *training subset* utilizzato nel 3° loop

Al termine del processo di *training*, l'algoritmo fornisce un valore percentuale relativo alla corretta capacità di calibrazione del modello, denominato "*cap_calibr_3*"; questo valore viene confrontato con il precedente "*cap_calibr_2*". Se il precedente valore (*cap_calibr_2*) è migliore di quello attuale (*cap_calibr_3*), il processo si arresta e l'albero decisionale elaborato nel 3° *loop* non viene immagazzinato. L'istanza verrà classificata utilizzando l'albero decisionale elaborato nel 2° loop. Se invece il valore attuale è migliore di quello precedente, l'albero decisionale elaborato in questo *loop* sovrascrive il precedente, cap_calibr_3 sostituisce cap_calibr_2 ed il processamento procede con il 4° *loop*.

☒ Se $cap\_calibr\_2 > cap\_calibr\_3$ → processo di elaborazione interrotto e valutazione istanza con albero decisionale generato nel 2° *loop*.

☑ Se $cap\_calibr\_2 < cap\_calibr\_3$ → processo di elaborazione continua; l'albero decisionale sovrascrive quello generato nel precedente *loop* e cap_calibr_3 sostituisce cap_calibr_2.

- *4° loop: [medesima modalità dei loop precedenti]*
- *5° loop: [medesima modalità dei loop precedenti]*
- *6° loop: [medesima modalità dei loop precedenti]*

Dai test condotti è emerso che i sei *loops* potenzialmente elaborabili dal modello sono più che sufficienti per classificare nel le istanze in ingresso; generalmente dopo il 3° o 4° loop si ha un peggioramento, seppur lieve, delle performance di calibrazione e solo raramente questo peggioramento si riscontra al 5° *loop*.

Considerata la consistente numerosità dei dati, dal 2000 al 2014, quest'implementazione comporta un'importante riduzione dei tempi nella verifica di tutto il set di dati disponibile; mediamente, per ogni istanza i tempi di elaborazione diminuiscono del 50-60%. In fase operativa, le tempistiche di elaborazione si riducono mediamente da 150-180 secondi a 90-120 secondi.

FASE 3: valutazione dell'ultima istanza ed output probabilistico

Successivamente, l'algoritmo *Random Forests* sulla base delle informazioni dedotte (albero finale) dal *training subset* valuta l'ultima istanza in ingresso; al termine della valutazione, associa ad essa un valore percentuale per ogni classe definita.

Come riportato nel paragrafo 4.3.1.3, in riferimento all'ingegnerizzazione del modello, questo elabora 38 classi nella Fase 1, in funzione dell'ultimo valore di portata osservato; e in questa fase ad ognuna di queste classi associa un valore percentuale di accadimento.

L'output del modello, può esser così riprodotto graficamente: il cerchio verde rappresenta l'ultimo valore di portata osservato, il *range* dei valori di portata delle classi sono rappresentati dagli istogrammi di colore arancione; mentre, l'output del modello è rappresentato dall'areale blu.

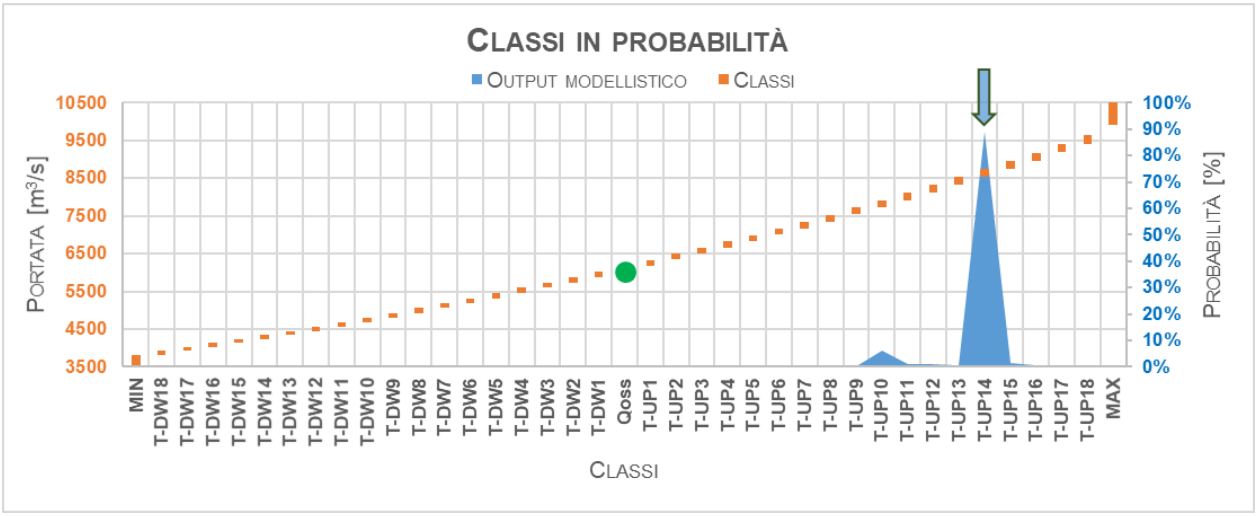


Figura 47 Rappresentazione grafica dell'output del modello *Random Forests* per le principali sezioni del Fiume Po.

Nell'esempio rappresentato è stato attribuito il maggior valore percentuale, pari a 88.95% (ordinata di destra), alla classe T_{up14}; ovvero, l'algoritmo *Random Forests*, sulla base del training subset utilizzato ed in riferimento all'ultimo valore osservato, suggerisce con elevata probabilità che il valore di portata raggiunto nelle prossime N-ore sarà compreso tra 8527 m^3/s e 8741 m^3/s. Valori percentuali sensibilmente inferiori sono stati attribuiti alle classi comprese tra T_{up10} e T_{up16}; la somma di tutti i valori percentuali è pari al 100%.

Oltre a fornire l'informazione completa, tutte le classi con le rispettive percentuali, il modello identifica la classe con la percentuale più alta, eventualmente suggerendola all'utente finale come informazione "ultima".

Quest'ultima elaborazione è più utile per fini di studio (analisi performance modello, etc.) più che per la fase operativa, dove spesso viene richiesta la possibilità di conoscere l'informazione completa con la quale valutare eventuali decisioni e/o interventi.

Per le finalità di protezione civile, percentuali previsionali apparentemente non significative, poiché molto basse, delle condizioni più gravose possono invece esser sufficienti per attivare delle azioni di pianificazione o di intervento sul territorio.

4.3.4 RISULTATI E PERFORMANCE DEL MODELLO

L'informazione numerica previsionale prodotta dal modello, ad ogni "*time step*", può esser rappresentata come in figura 8; per valutare le performance del modello, necessarie per le finalità di questo elaborato, è necessario considerare solo la classe avente il valore percentuale più elevato rispetto alle altre. Pertanto, per ogni "*time step*" è stata identificata la classe "dominante", avente la probabilità maggiore rispetto alle altre, a cui corrisponde un determinato "*range*" di portate; come illustrato nei paragrafi precedenti. I risultati sono stati rappresentati come segue:

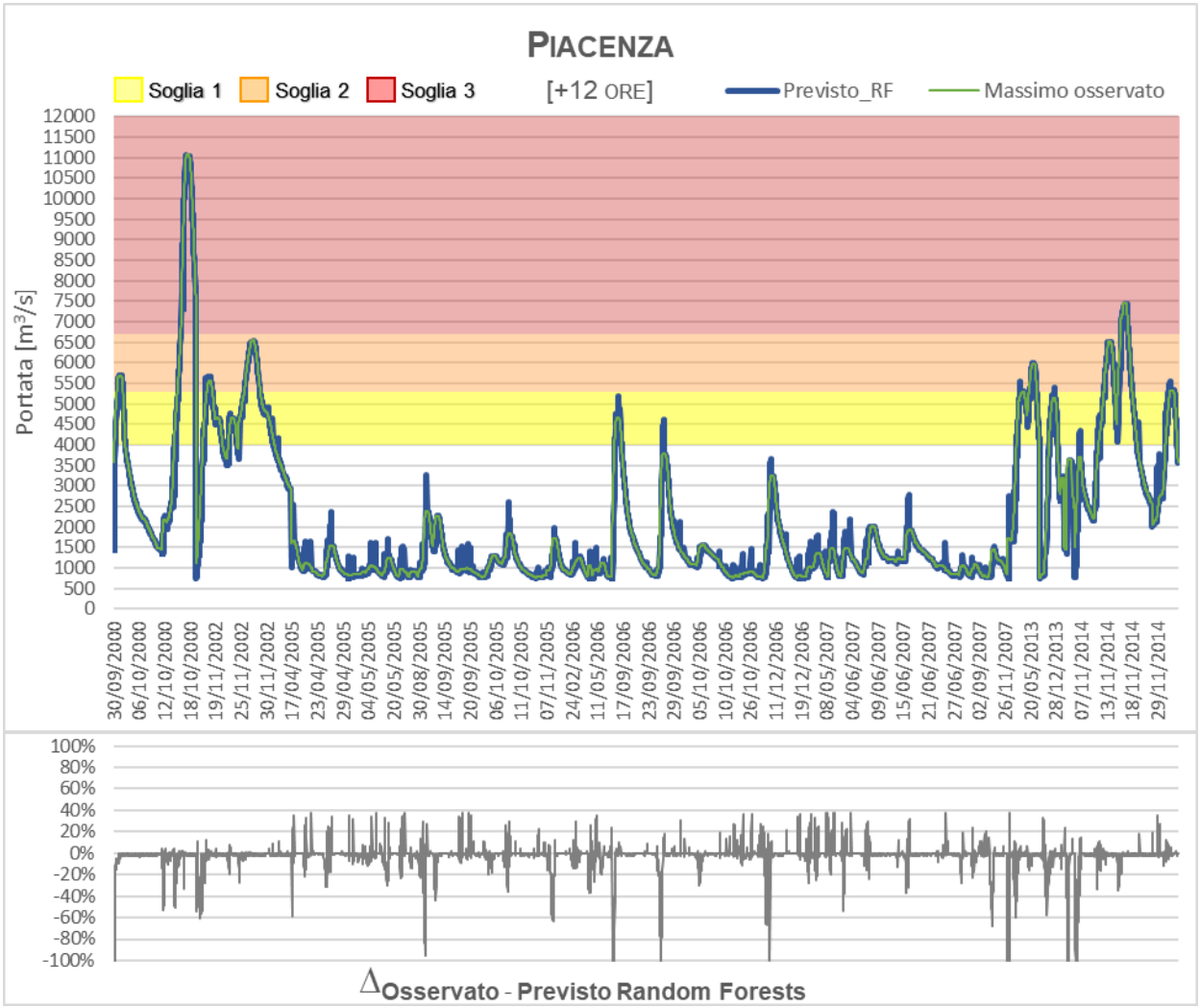


Figura 48 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 12 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*.

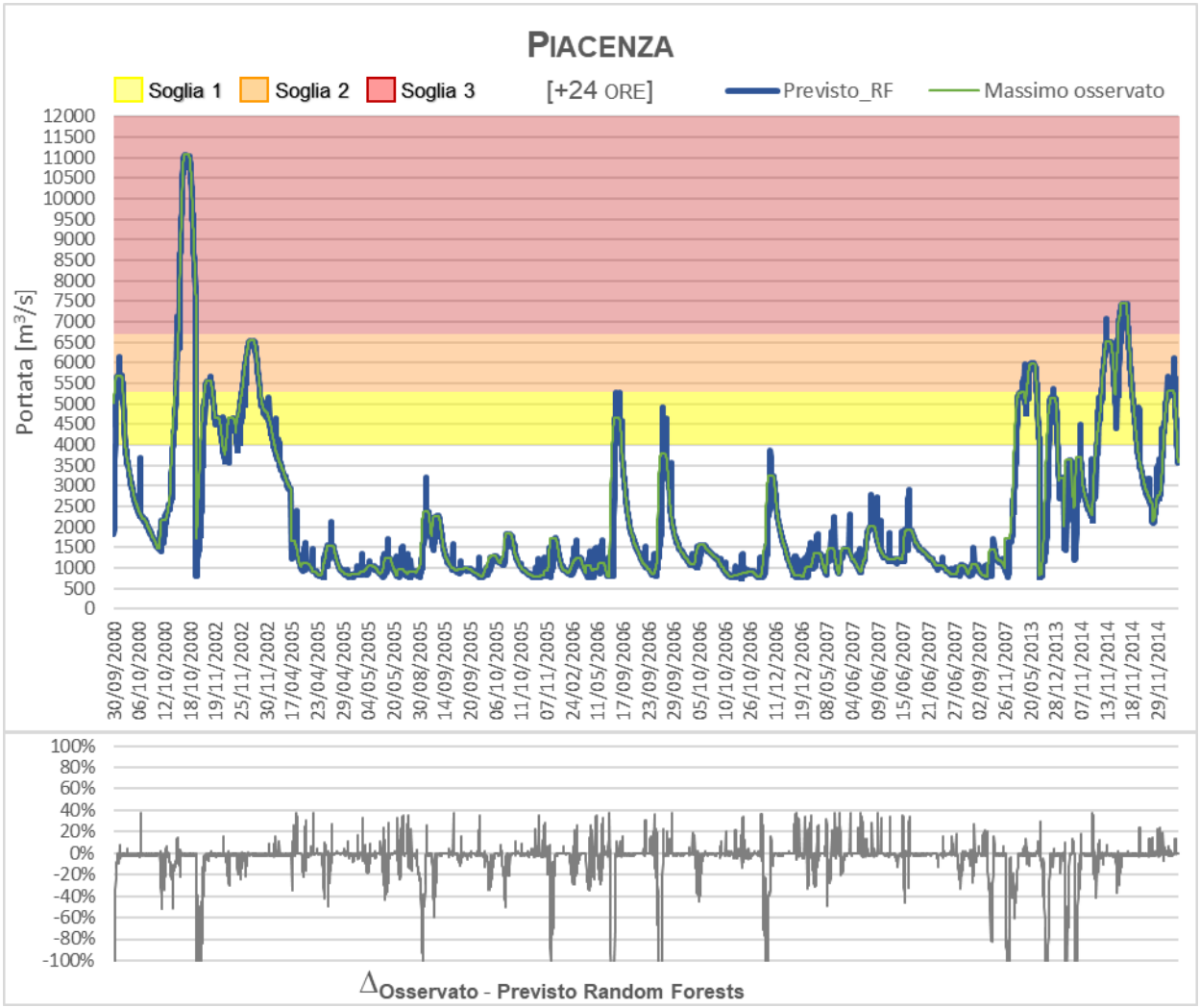


Figura 49 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 24 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

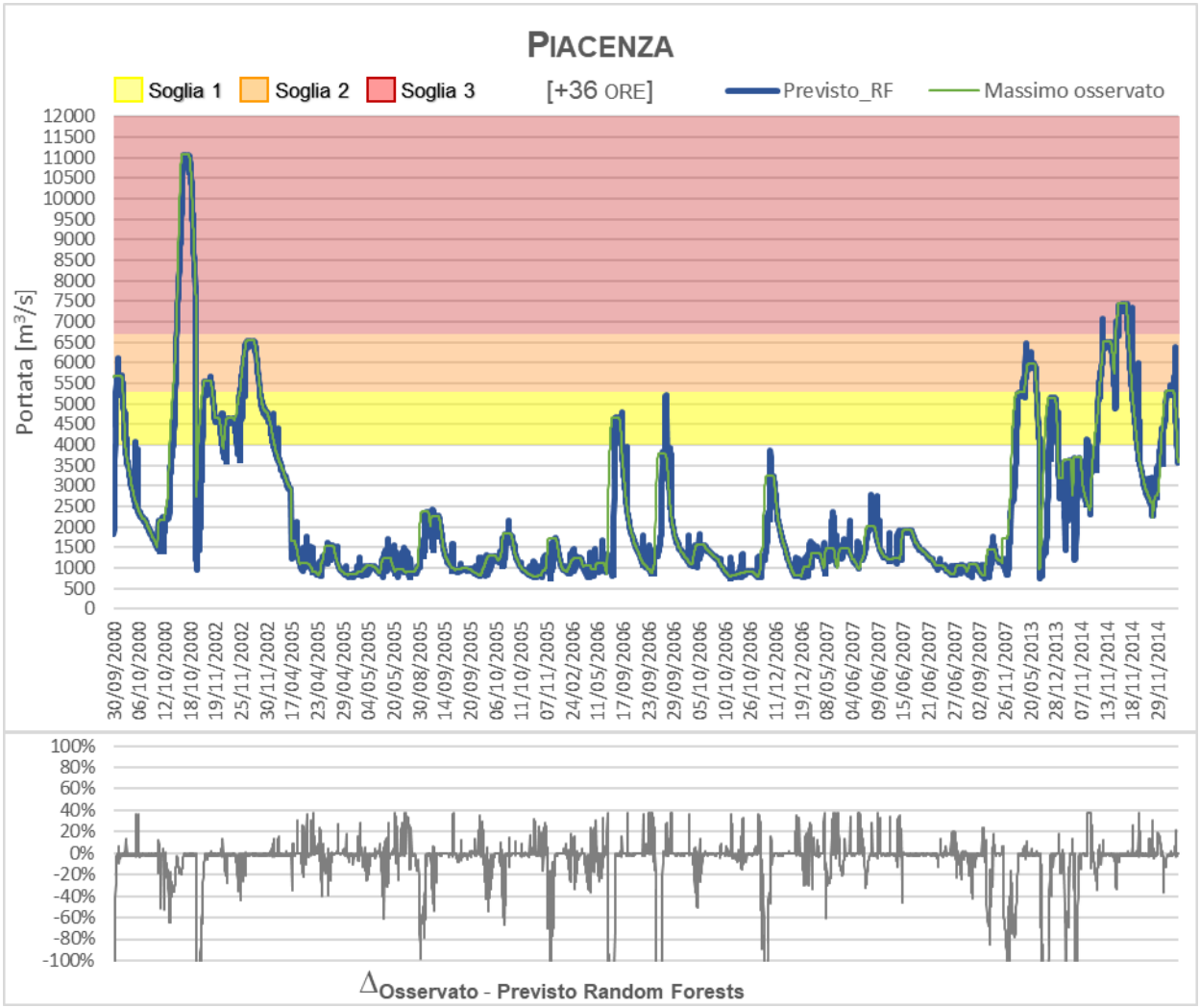


Figura 50 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 36 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

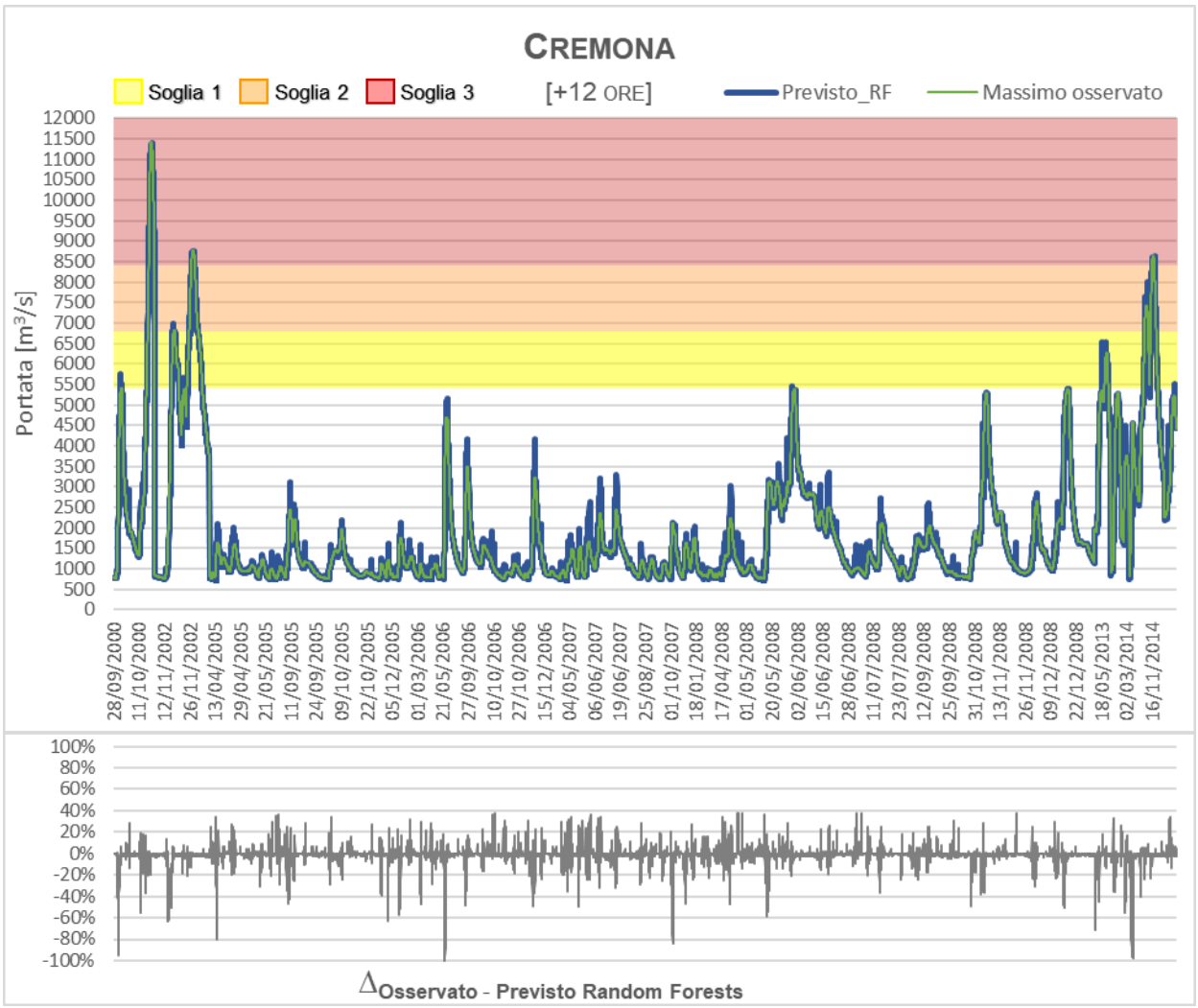


Figura 51 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 12 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

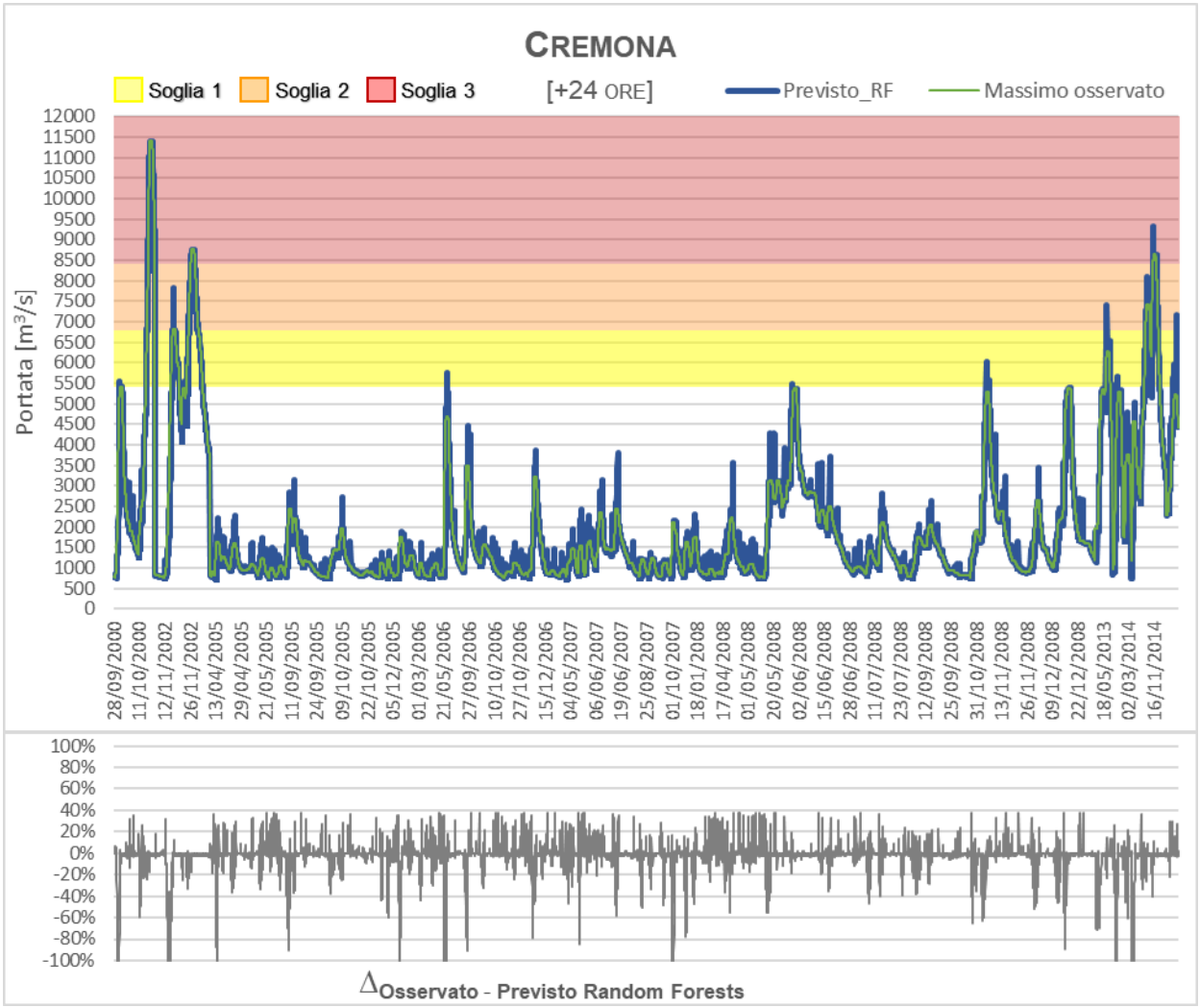


Figura 52 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 24 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

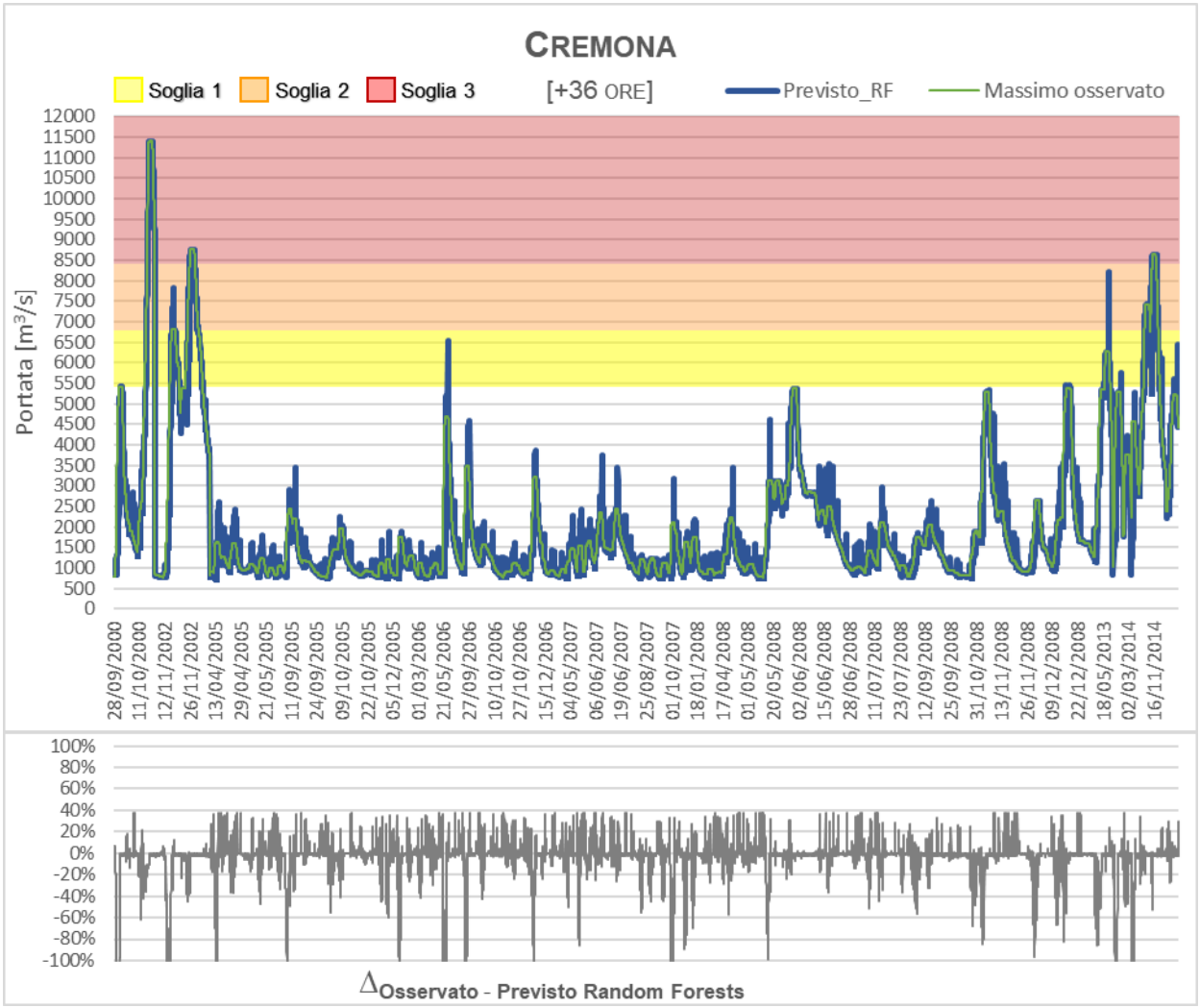


Figura 53 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 36 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

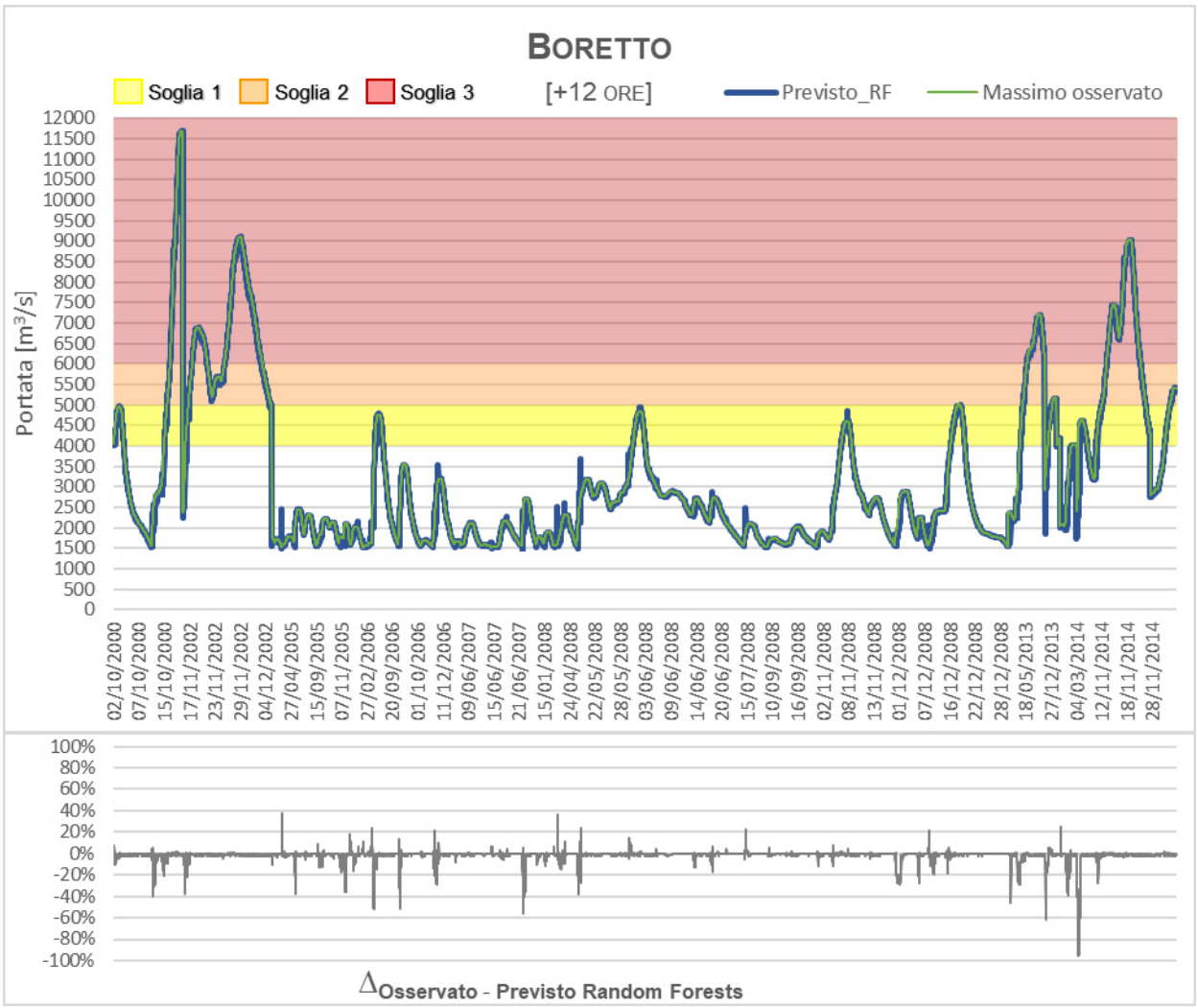


Figura 54 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 12 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

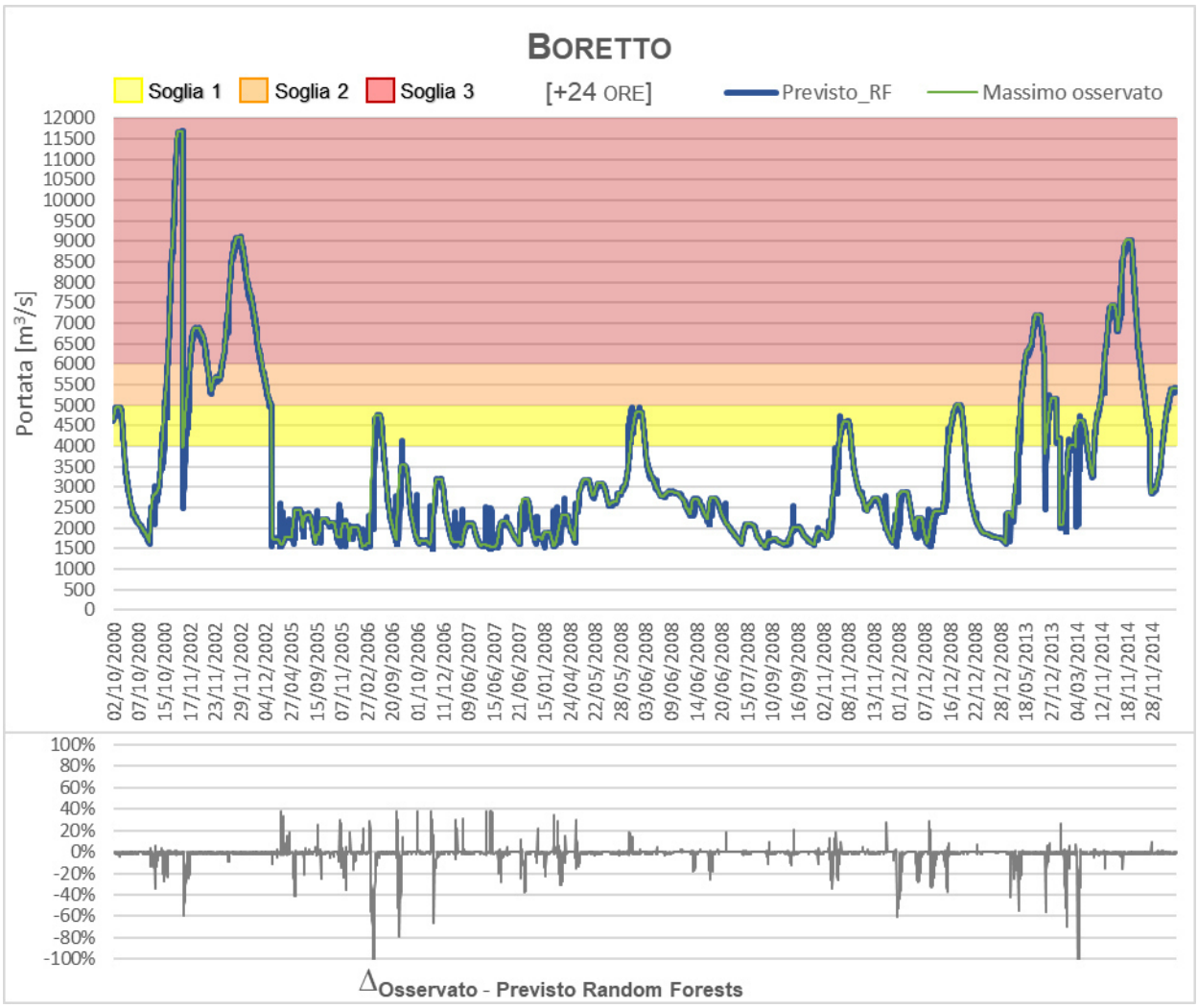


Figura 55 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 24 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

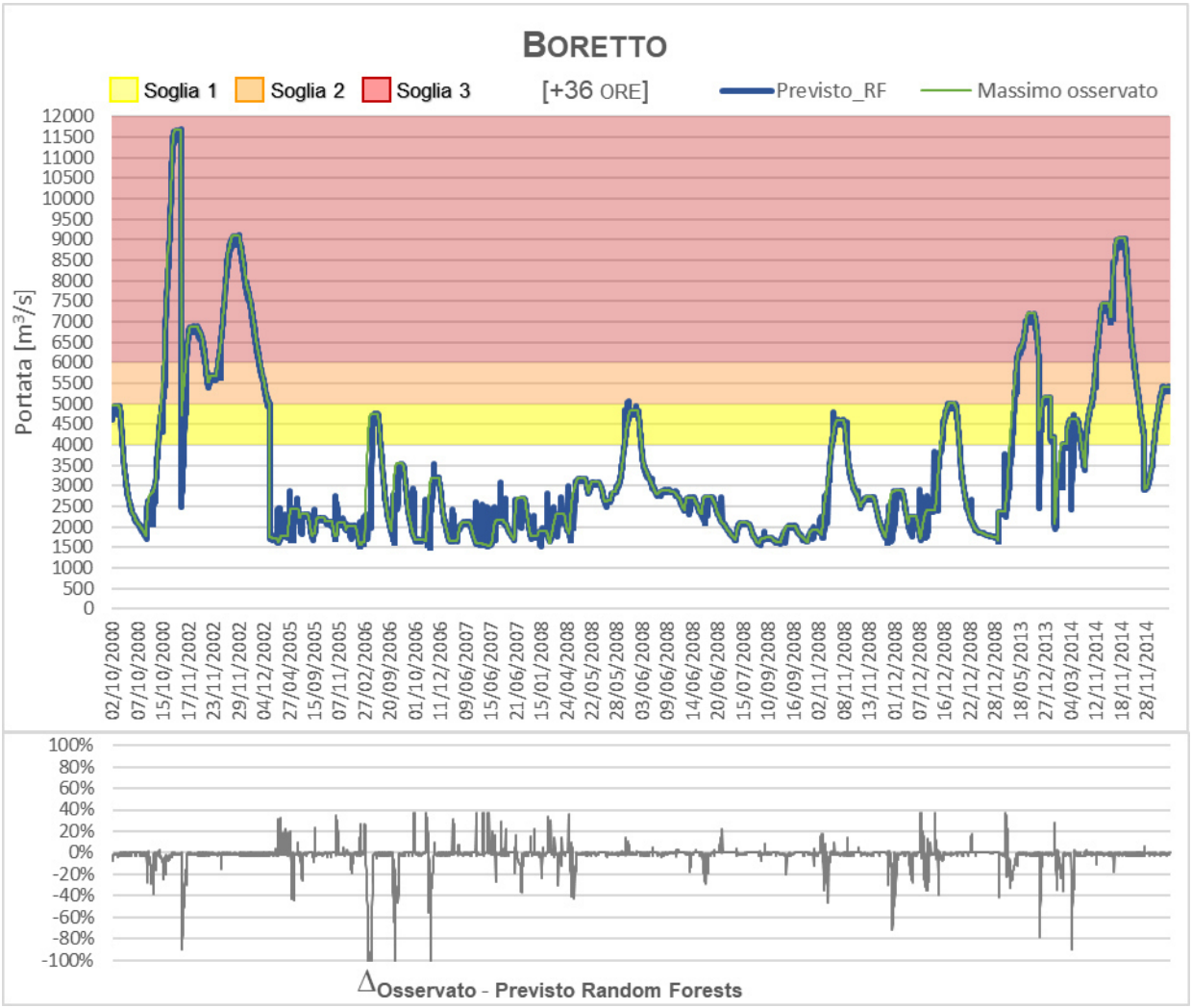


Figura 56 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 36 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

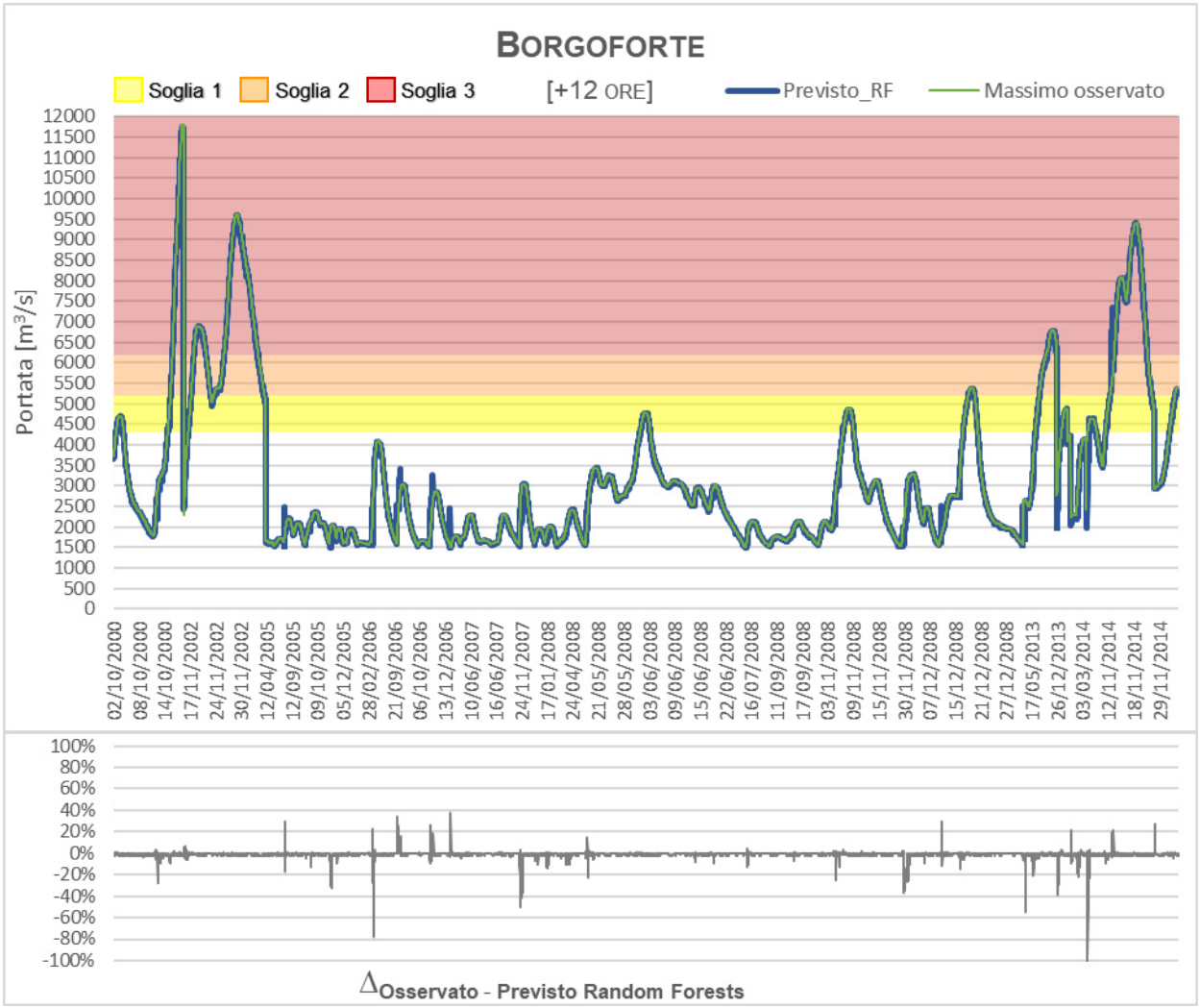


Figura 57 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 12 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

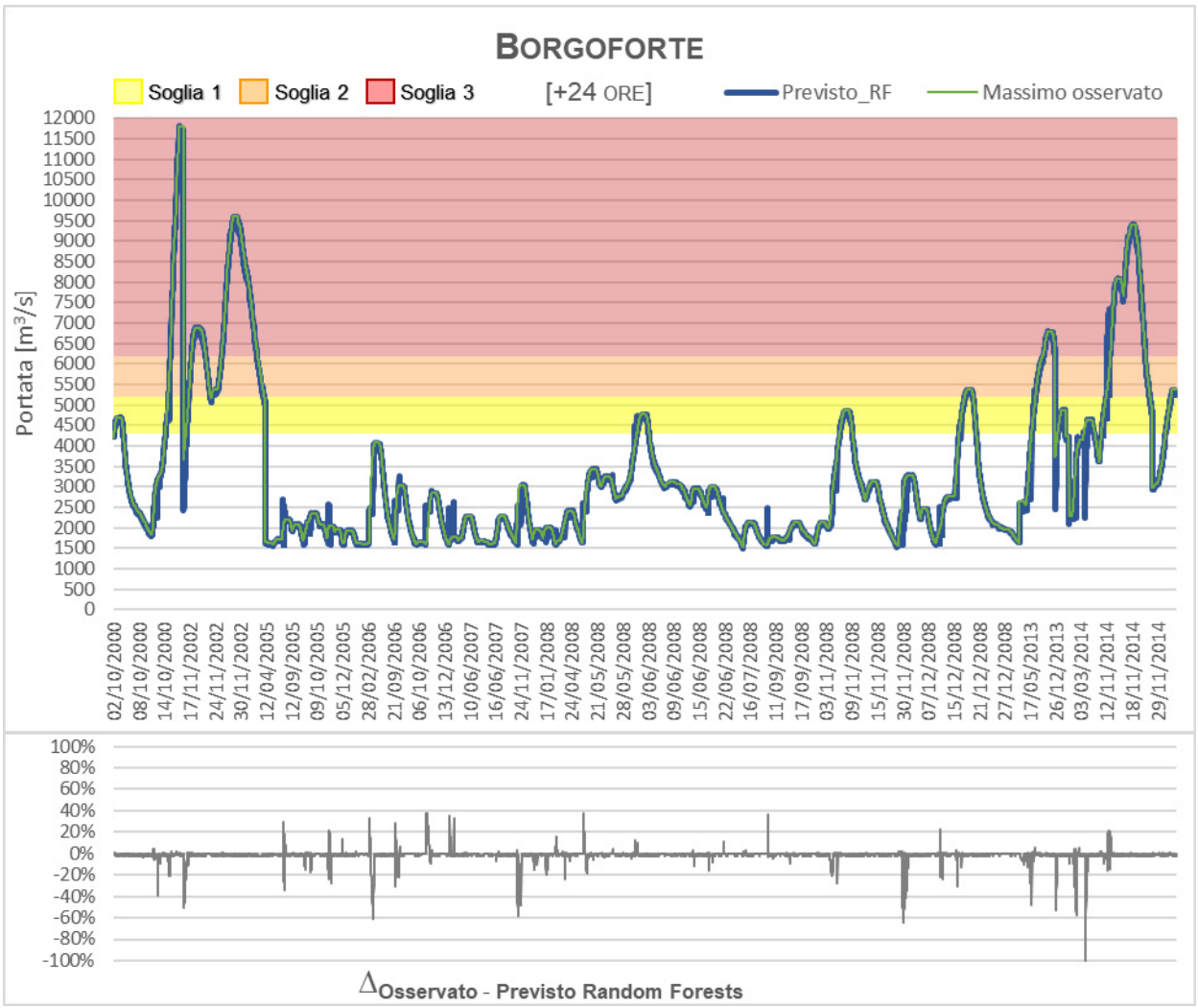


Figura 58 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 24 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

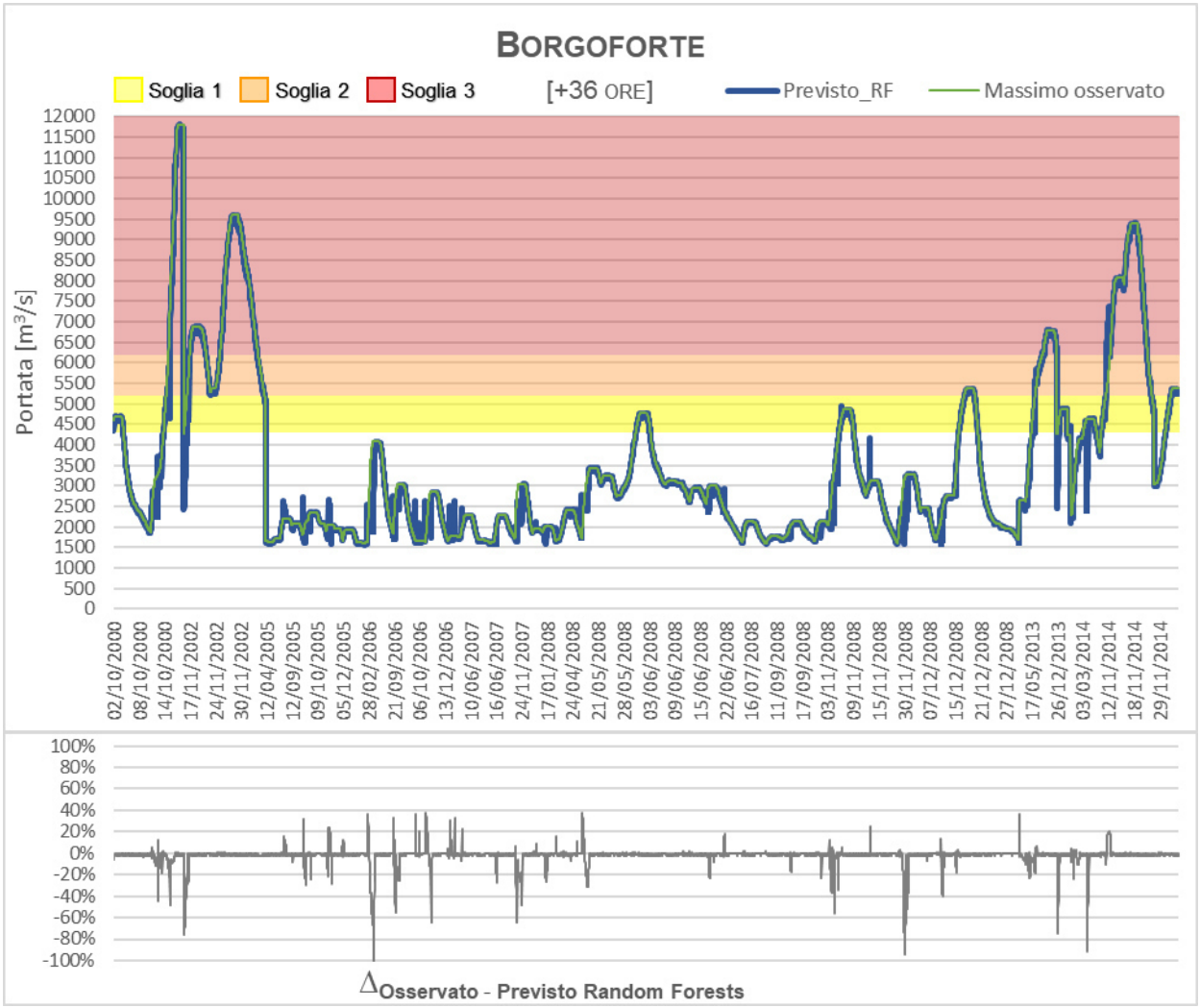


Figura 59 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 36 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

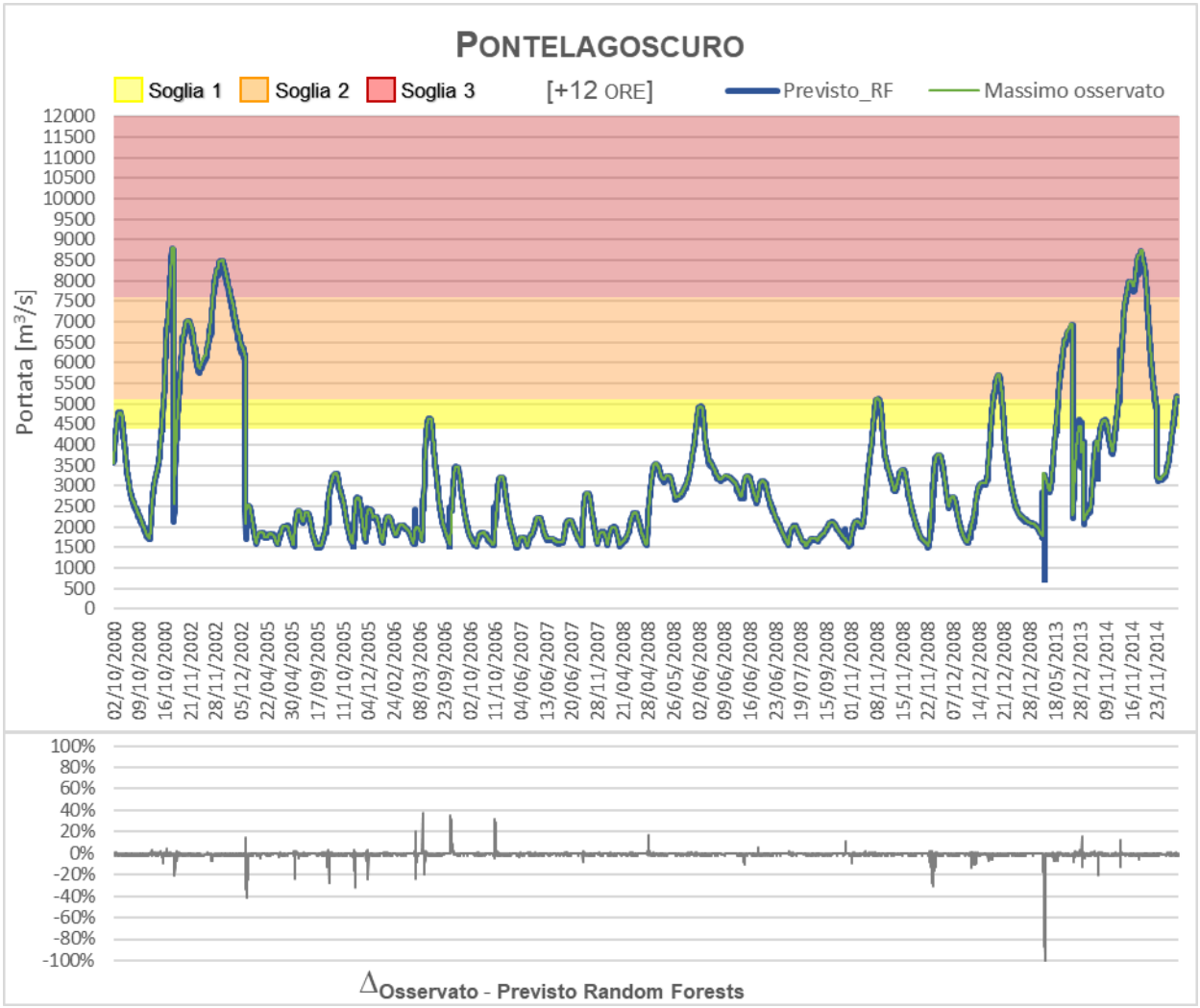


Figura 60 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 12 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

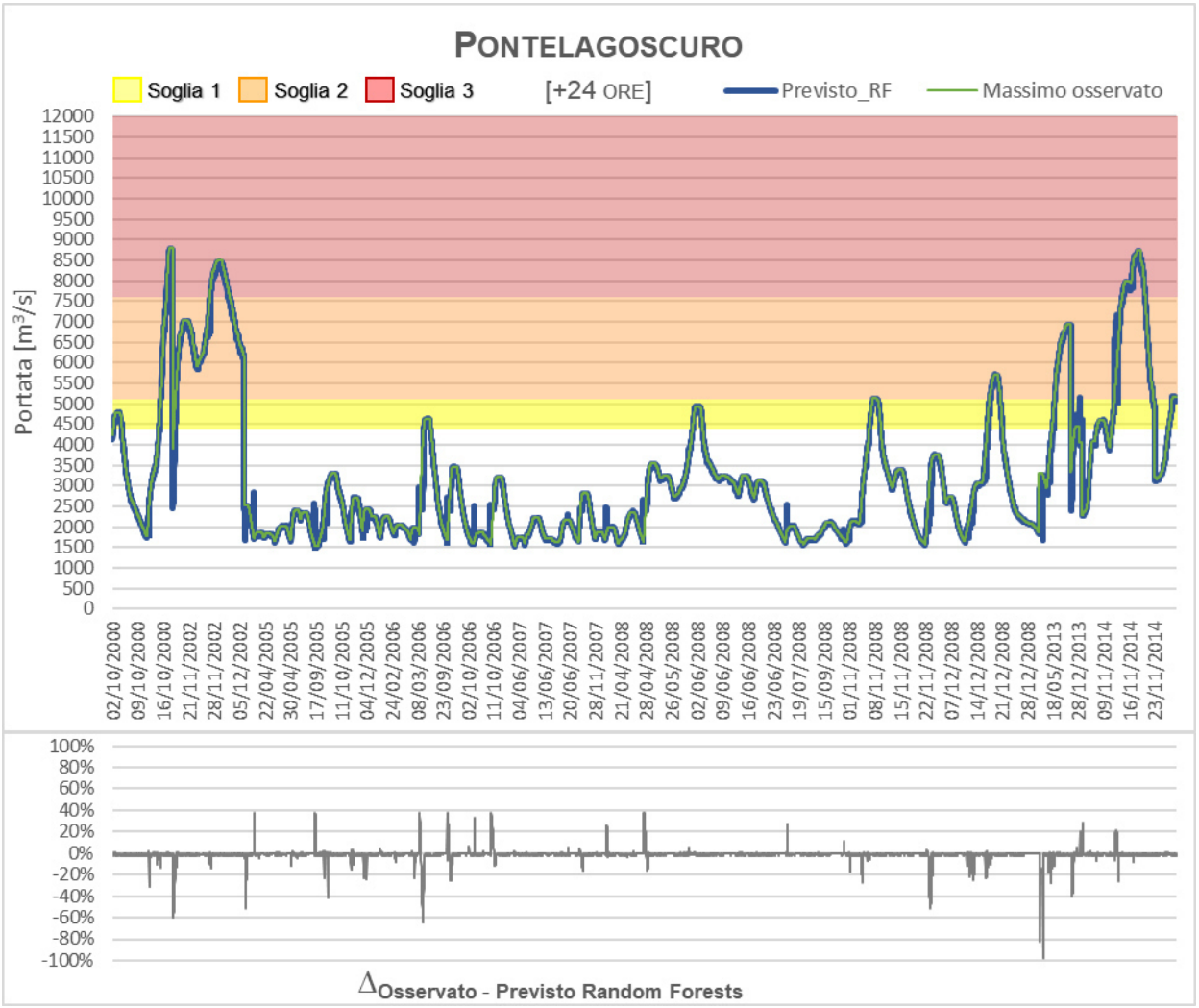


Figura 61 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 24 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

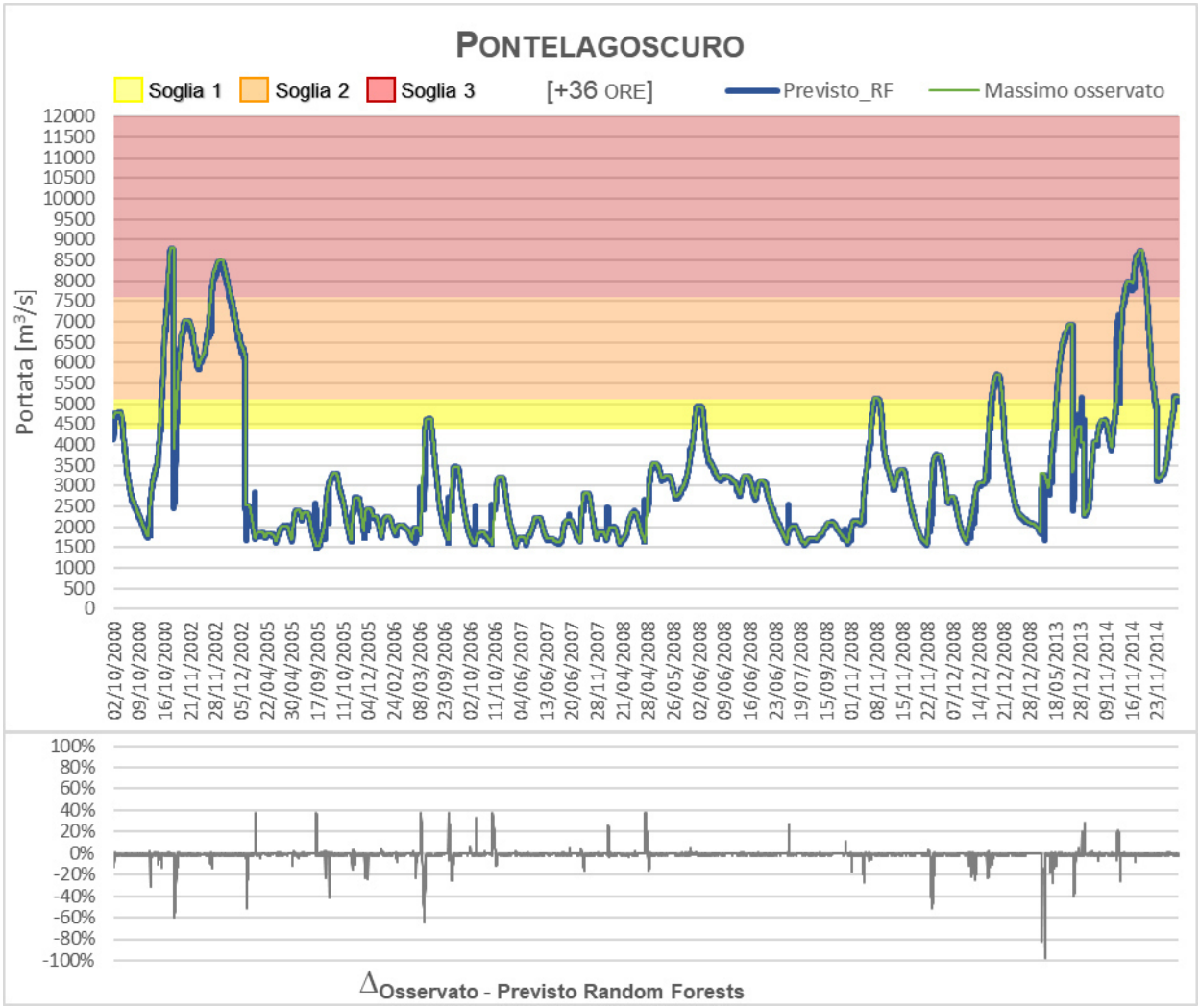


Figura 62 Grafico superiore: rappresentazione grafica dell'output del modello *Random Forests* confrontato con il massimo valore osservato nelle 36 ore successive. Grafico inferiore: delta percentuale tra massimo valore osservato e valore previsto dal modello *Random Forests*

Da una prima analisi speditiva, si hanno due importanti conferme:

- L'errore previsionale è in funzione dell'arco temporale previsionale; tanto maggiore è l'arco di previsione tanto maggiore è l'errore commesso dal modello in previsione.

- Nelle prime sezioni, Piacenza e Cremona, il tasso di errore è sensibilmente maggiore rispetto alle altre tre sezioni; ciò è attribuibile sia alla mancanza di informazioni aggiuntive relative alle sezioni precedenti (Piacenza) sia al contributo -talvolta importante- dei tributari lombardi (sezione di Cremona). Le sezioni più a valle, non ricevendo importanti contributi, sono strettamente correlate a quella a monte; pertanto le performance sono migliori.

4.3.4.1 Analisi Performance

Un modello che elabora in continuo tutte le informazioni disponibili, come quello sviluppato in questo caso di studio, richiede un'attenzione particolare in fase di valutazione delle performance; per ogni "*step*

temporale", vengono generati nuovi *output* in sostituzione di quelli precedenti. Talvolta, in funzione delle ultime informazioni acquisite, le differenze tra gli ultimi risultati e quelli precedenti possono essere consistenti.

In questo caso specifico, per effettuare un'analisi completa e più corretta possibile delle performance del modello, l'informazione probabilistica previsionale dovrebbe essere memorizzata ad ogni elaborazione (in altre parole, ad ogni "*time step*") così da permettere la valutazione di tutte le 38 classi definite.

Per le elaborazioni che seguono, come per quelle grafiche, è stata considerata solamente la classe avente la percentuale di probabilità dominante.

Ripercorrendo quanto presentato nel Capitolo 3, la valutazione delle performance è stata condotta utilizzando i cinque indici di riferimento:

1. "*Probability of detection POD*": valuta la frazione degli eventi osservati che è stata effettivamente prevista; è un indice sensibile ai "successi previsionali", ma ignora le sovrastime previsionali.

$$POD = \frac{HITS}{HITS + MISS}$$

POD			
	RF +12h	**RF +24h**	**RF +36h**
Piacenza	0.931	0.904	0.876
Cremona	0.925	0.889	0.870
Boretto	0.971	0.964	0.951
Borgoforte	0.978	0.978	0.973
Pontelagoscuro	0.955	0.962	0.946

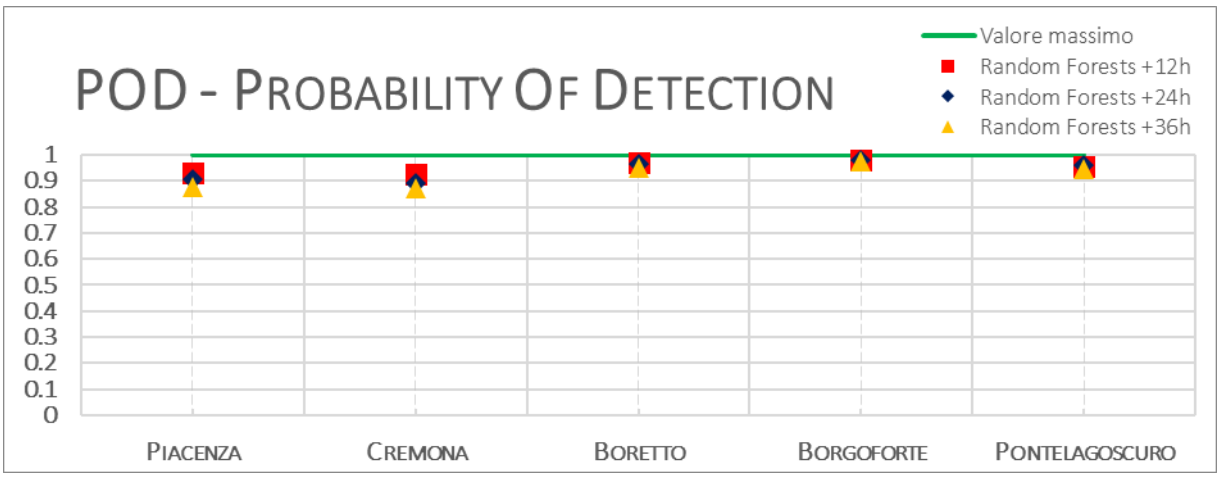


Figura 63 Rappresentazione numerica e grafica dell'analisi delle performance del modello

2. *"False Alarm Ratio FAR*": valuta la frazione degli eventi previsti ma non verificatisi; è un indice sensibile alle sovrastime previsionali, ma ignora le sottostime previsionali.

$$FAR = \frac{FALSE}{HITS + FALSE}$$

FAR			
	RF +12h	RF +24h	RF +36h
Piacenza	0.019	0.017	0.017
Cremona	0.019	0.014	0.015
Boretto	0.001	0.002	0.003
Borgoforte	0.007	0.006	0.006
Pontelagoscuro	0.001	0.004	0.006

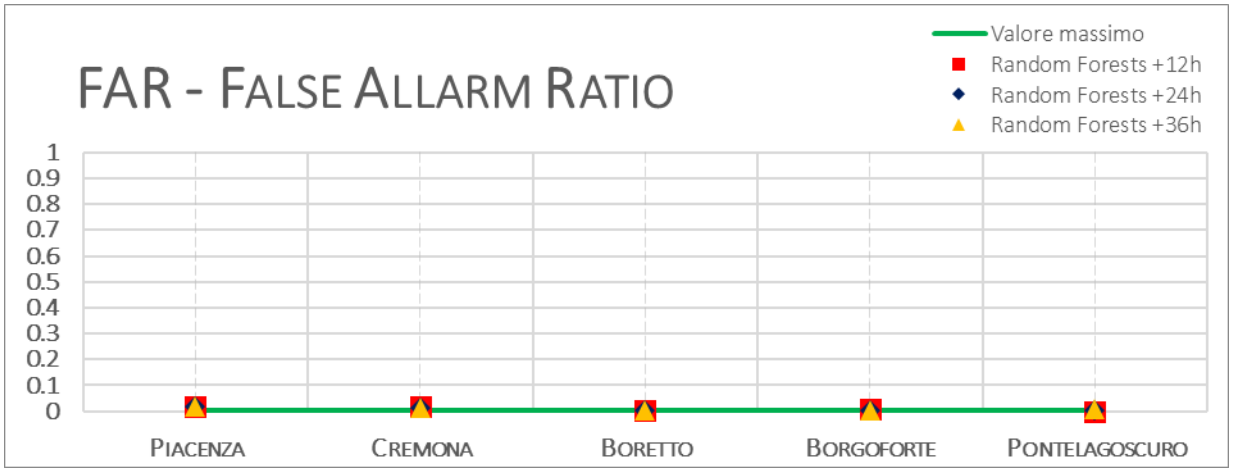


Figura 64 Rappresentazione numerica e grafica dell'analisi delle performance del modello

3. "*Critical Success Index CSI*": valuta il rapporto degli eventi previsti e verificatisi rispetto a tutti gli eventi osservati; risulta sensibile agli "*hits*" ma considera in maniera ridotta gli errori previsionali, sovrastime e sottostime.

$$CSI = \frac{HITS}{HITS + FALSE + MISS}$$

CSI			
	RF +12h	RF +24h	RF +36h
Piacenza	0.915	0.890	0.863
Cremona	0.908	0.878	0.859
Boretto	0.969	0.962	0.949
Borgoforte	0.970	0.973	0.968
Pontelagoscuro	0.954	0.958	0.940

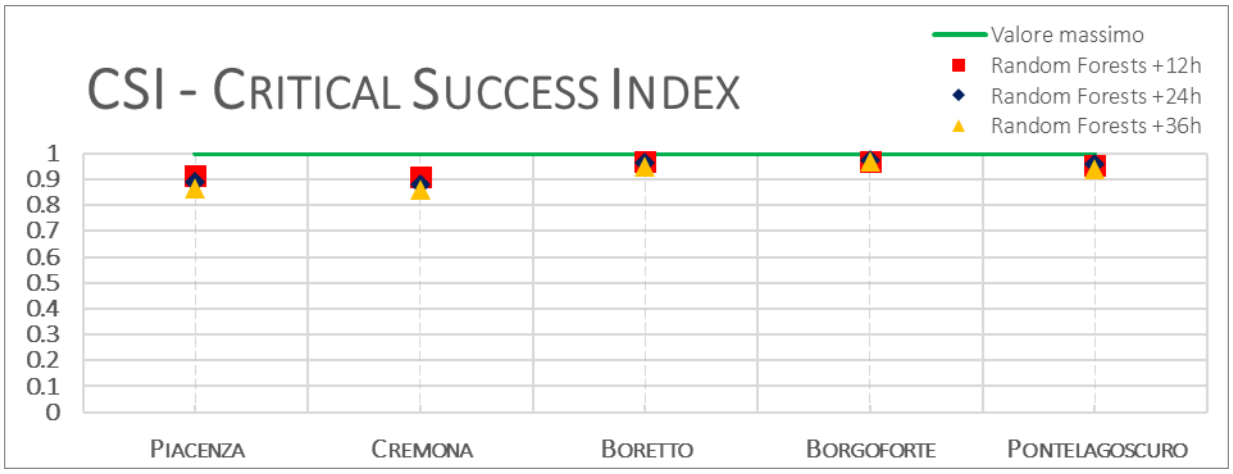


Figura 65 Rappresentazione numerica e grafica dell'analisi delle performance del modello

4. "*Success Ratio SR*": valuta la frazione effettivamente osservata di tutti gli eventi previsti; ignora le sottostime previsionali.

$$SR = \frac{HITS}{HITS + FALSE}$$

SR			
	RF +12h	**RF +24h**	**RF +36h**
Piacenza	0.981	0.983	0.983
Cremona	0.981	0.986	0.985
Boretto	0.999	0.998	0.997
Borgoforte	0.993	0.994	0.994
Pontelagoscuro	0.999	0.996	0.994

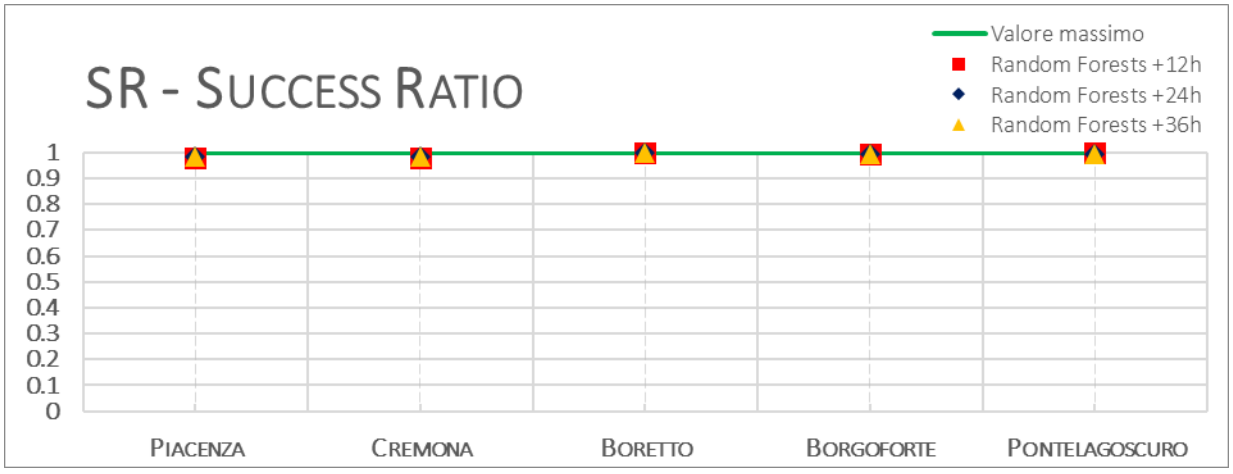


Figura 66 Rappresentazione numerica e grafica dell'analisi delle performance del modello

5. "*BIAS*": valuta la relazione tra gli eventi previsti e gli eventi osservati.

$$BIAS = \frac{HITS + FALSE}{HITS + MISS}$$

BIAS			
	RF +12h	RF +24h	RF +36h
Piacenza	0.948	0.920	0.892
Cremona	0.942	0.902	0.884
Boretto	0.972	0.966	0.954
Borgoforte	0.985	0.984	0.979
Pontelagoscuro	0.956	0.966	0.952

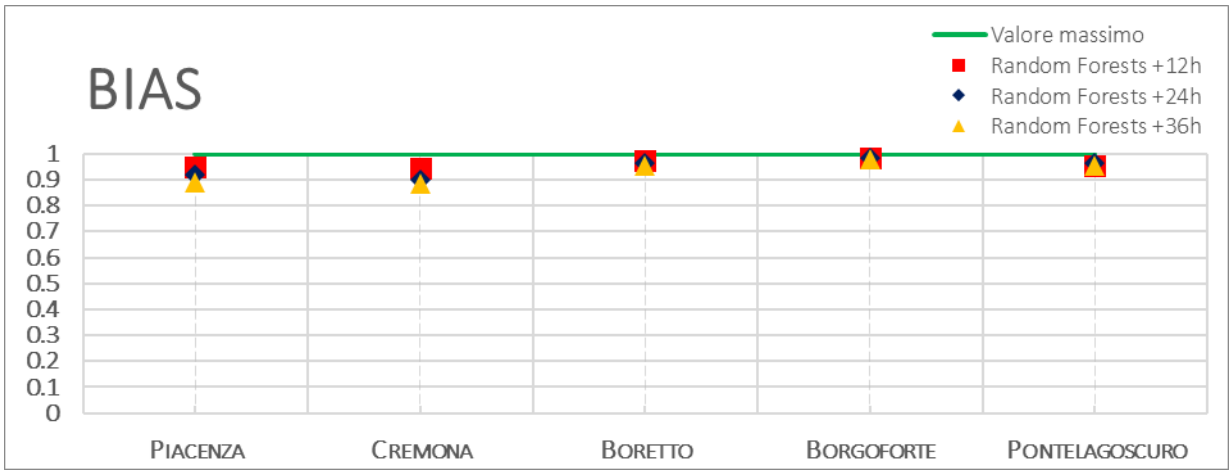


Figura 67 Rappresentazione numerica e grafica dell'analisi delle performance del modello

I risultati ottenuti in questo caso di studio confermano l'elevata capacità del metodo *Random Forests* di "memorizzare" le informazioni ingresso; su cui basarsi successivamente in fase di operativa. L'importanza di un *training set* corposo ed eterogeneo permette di ampliare il dominio di valutazione; così come l'assenza o la scarsità di informazioni comporta un peggioramento delle performance modellistiche, seppur contenuto in questo caso (Piacenza e Cremona).

Dal dettaglio dei pesi attribuiti alle variabili dall'algoritmo in fase di *training*, come presentato nel paragrafo 3.4.3.4, emerge che la variabile più rilevante sia in fase di elaborazione della foresta casuale (grafico di destra) sia in fase di perfezionamento della valutazione (grafico di sinistra) è quella relativa al massimo valore previsto dalla catena modellistica MIKE11 NAM/HD; in tutti e tre i casi: +12 ore (Q_{max12h_Df}), +24 ore ($Q_{max24h\_\_Df}$) e +36 ore (Q_{max36h_Df}).

In fase di elaborazione della foresta casuale, la variabile previsionale Q_{max_Df} è seguita dalle informazioni di portata osservata alle sezioni precedenti a quella di riferimento: Boretto e Borgoforte, nel caso di Pontelagoscuro (fig. 29); mentre, per affinare la valutazione dell'ultima istanza, utilizzando una foresta già elaborata, l'algoritmo considera predominante le ultime informazioni relative alla portata osservata alla sezione di riferimento.

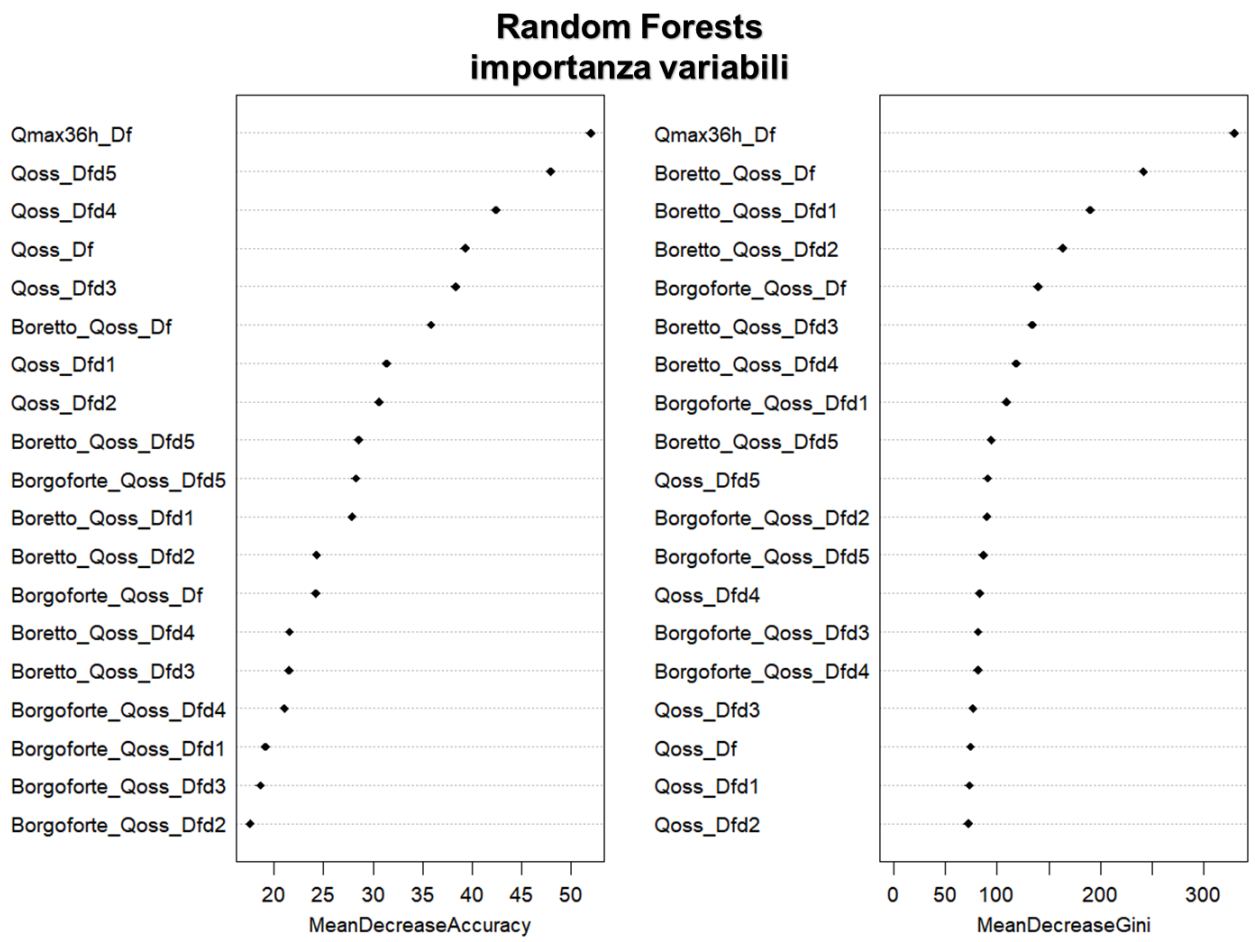


Figura 68 Rappresentazione grafica dell'importanza delle variabili definite dall'algoritmo in fase di training.
Sezione: Pontelagoscuro – Valutazione variabili modello a +36 ore

Il modello sviluppato in questo caso di studio, così come quello sviluppato nel caso di studio precedente, non ha alcuna informazione indotta relativa alla fisicità del bacino; ma l'algoritmo è in grado di dedurre attraverso il *training set* fornitogli che per formulare un'ipotesi in probabilità sulle future classi (*range* di portata) nelle ore successive ha bisogno della variabile previsionale della catena modellistica Mike11 NAM/HD. Mentre, per migliorare l'accuratezza dell'ipotesi previsionale in probabilità sulle future classi ha necessità di "conoscere" i valori osservati alla sezione di riferimento e successivamente anche i valori osservati relativi alle sezioni precedenti.

L'assenza all'interno del *training set* dell'informazione previsionale comporterebbe un sensibile peggioramento delle performance del modello in quanto verrebbe a mancare l'unica informazione relativa "al futuro"; il modello risulterebbe, quindi, fortemente condizionato dalle sole informazioni relative ai valori osservati.

4.3.5 ULTERIORI CONFRONTI E CONSIDERAZIONI

I risultati di questo caso di studio relativi alla sezione di Pontelagoscuro (sezione di chiusura del bacino) sono stati confrontati con quelli ottenuti dalla catena MIKE11-MCP. L'*MCP* (*Model Conditional Processor*) è un

post-processore Bayesiano per la stima dell'incertezza predittiva introdotto dal Prof. Ezio Todini nel 2008 e sviluppato presso l'Università di Bologna (Coccia e Todini, 2011). Il metodo è basato sull'applicazione del teorema di Bayes per la stima della distribuzione di probabilità della variabile osservata condizionata alle previsioni dei modelli. I dati, per poter essere utilizzati, vengono trasformati nel campo Gaussiano applicando la Trasformazione Normale Quantile (TNQ), in cui le distribuzioni marginali delle variabili sono Normali Standard e la loro distribuzione congiunta può anche essere rappresentata a tratti mediane due o tre distribuzioni Gaussiane Multivariate Troncate. Applicando il teorema di Bayes si ricava la stima della distribuzione di probabilità predittiva, pari al rapporto tra la distribuzione congiunta di tutte le variabili e la distribuzione congiunta delle sole variabili previste. Il risultato ottenuto è una distribuzione Normale in campo Gaussiano che viene ritrasformata nello spazio reale dei dati utilizzati attraverso una trasformazione TNQ inversa.

In fase di verifica dei risultati, il modello Random Forests è stato ingegnerizzato per fornire in output anche il valore di portata previsto al termine di una finestra temporale definita: +12 ore, +24 ore e +36 ore; tale informazione è utile in questa fase per valutare la capacità previsionale in termini di "precisione temporale" del modello.

Anche se non è lo scopo per cui viene sviluppato l'MCP sviluppato per derivare la distribuzione di probabilità predittiva, il confronto tra i due metodi è stato condotto su più eventi in termini di valore atteso condizionato. Di seguito sono riportati tre dei più recenti eventi a Pontelagoscuro, in riferimento al set di dati utilizzato in cui è stata considerata la finestra temporale più lunga, +36 ore.

Si tenga tuttavia presente che i risultati della catena MIKE11-MCP dipendono fortemente dalla qualità dei risultati del modello condizionante (MIKE11) in quanto l'MCP è un post-processore statistico che permette esclusivamente di derivare la distribuzione di probabilità predittiva condizionatamente al modello (o ai modelli) previsionali utilizzati.

Evento 20 maggio-23 giugno 2008

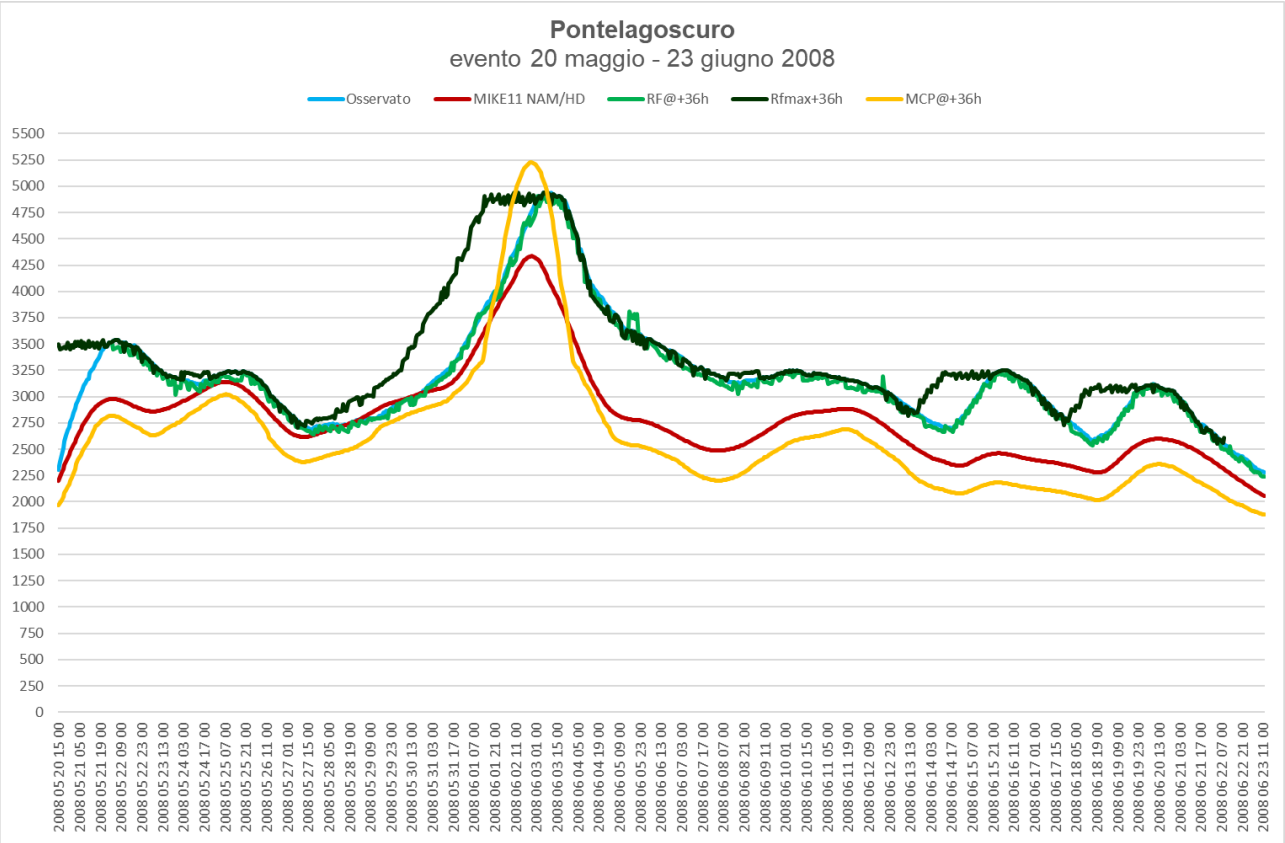


Figura 69 Evento di piena 20 maggio – 23 giugno 2008: confronto grafico tra osservato, Mike11 NAM/HD, *Random Forests* e MCP

Evento 5-18 novembre 2008

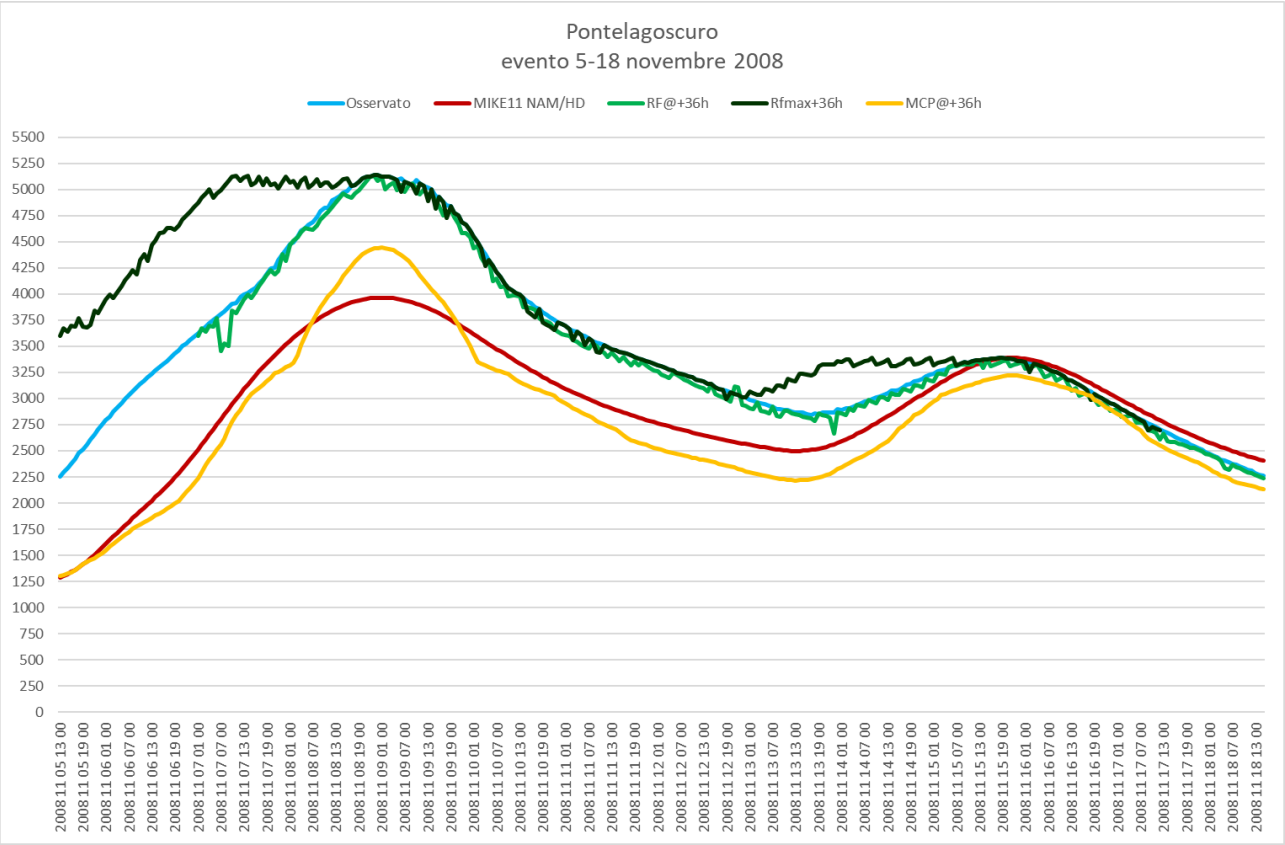


Figura 70 Evento di piena 5 novembre – 18 novembre 2008: confronto grafico tra osservato, Mike11 NAM/HD, *Random Forests* e MCP

Evento 13-25 dicembre 2008

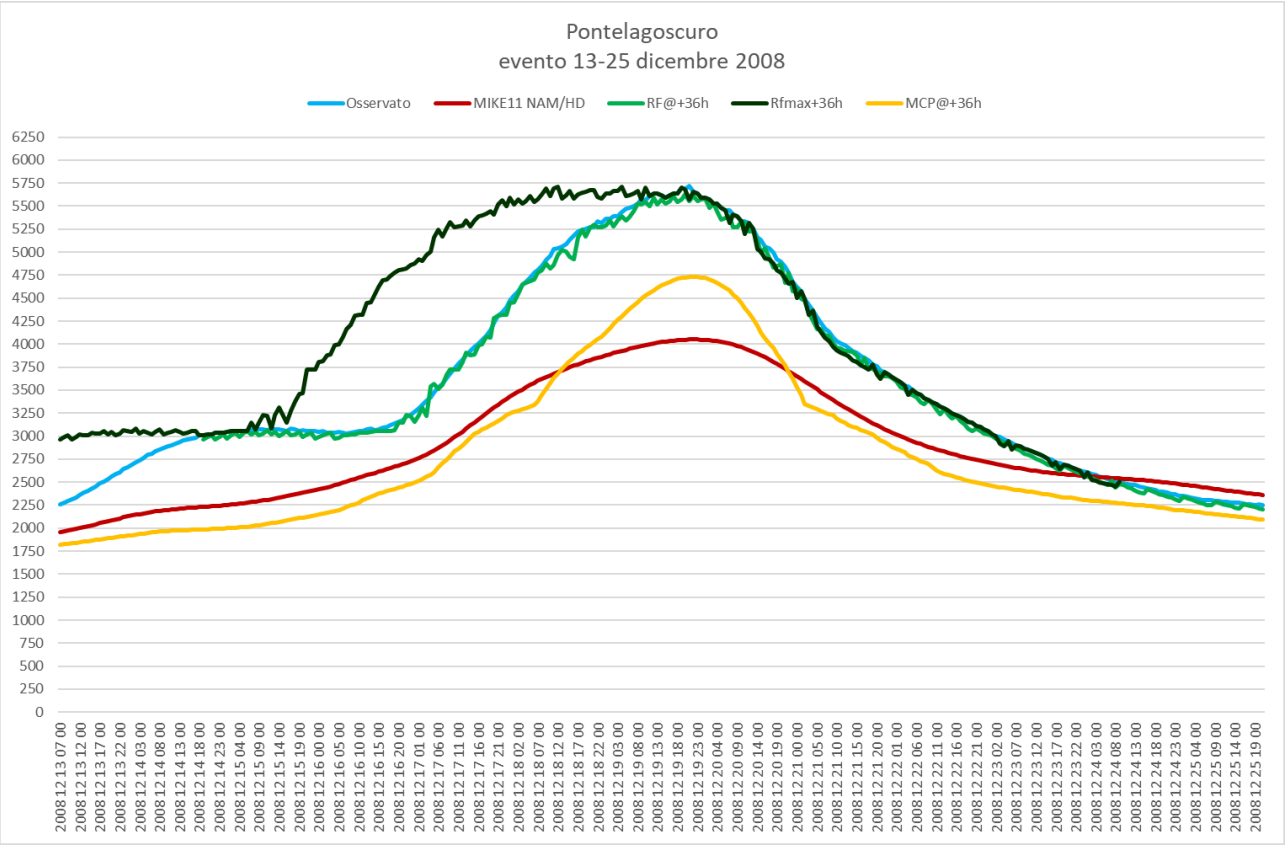


Figura 71 Evento di piena 13 dicembre – 25 dicembre 2008: confronto grafico tra osservato, Mike11 NAM/HD, *Random Forests* e MCP

Da questi e dagli altri confronti condotti, emerge una buona capacità del *Random Forests* di "classificare" gli eventi simili a quelli presenti nel *training set*; l'informazione previsionale "puntuale" al termine della finestra temporale considerata (+36 ore), linea verde chiaro, segue bene l'andamento del valore osservato, seppur affetta da un "rumore" occasionalmente più marcato. Più stabile, invece, l'informazione previsionale riferita al massimo valore previsto per le prossime 36 ore (linea verde scuro); per ragioni operative di protezione civile, quest'ultima informazione è preferita in quanto consente di conoscere il massimo valore atteso nelle prossime ore sul quale valutare i possibili interventi.

CONCLUSIONI

Agli inizi della mia attività di ricerca, ovvero durante la fase dedicata principalmente all'approfondimento bibliografico, ho potuto dedurre che l'evoluzione tecnico–scientifica talvolta è mossa da nuove conoscenze e talvolta da nuove tecniche; solo occasionalmente nuove tecniche e nuove conoscenze concorrono congiuntamente all'evoluzione tecnico–scientifica, e quando ciò accade più che un "passo in avanti" si verifica un "balzo in avanti". Talvolta, le conoscenze e/o le tecniche note e collaudate da tempo in una determinata disciplina vengono importante in un'altra disciplina e ciò spesso comporta una positiva sintesi delle conoscenze.

Nassim Nicholas Taleb in "*Giocati dal caso*" (Taleb, 2005), uno dei suoi saggi che avuto il piacere di leggere in quest'ultimo periodo, sostiene come il progresso scientifico proceda a salti, a differenza di quanto sostenuto fino al XX secolo quando gran parte della comunità scientifica riteneva che questo fosse "continuo". Graficamente, l'evoluzione tecnico-scientifica è sensibilmente più vicino ad una linea tratteggiata incrementale, piuttosto che una retta lineare o una curva crescente.

I "gradini" della linea tratteggiata sono quasi sempre generati dalla necessità; solo raramente, molto raramente, dalla casualità: in condizioni di "quiete" si tende a non variare la situazione attuale e gli elementi che la caratterizzano; mentre, in seguito ad un evento -generalmente negativo- si compie "uno sforzo" che altera la condizione attuale e produce un miglioramento, un'evoluzione, un "gradino" verso l'alto (salvo casi eccezionali).

Se vogliamo, una visione più "limitata" del principio "causa-effetto", caratterizzata però da un coefficiente dispersivo, più o meno grande; difficilmente una "causa" sortisce un "effetto" di ugual misura, soprattutto se la "causa" è di origine naturale e l'"effetto" è un prodotto antropogenico.

Nel campo dell'idrologia e dell'idraulica, gli egizi con il "Nilometro" segnavano il livello delle piene e lo rapportavano all'andamento del raccolto. Così, prevedevano l'andamento del raccolto basandosi sulle piene stagionali del Nilo; il Nilo, come gran parte dei corsi d'acqua, è caratterizzato da piene periodiche con intensità variabile: talvolta devastanti, ma spesso "vitali" per la fertilizzazione dei campi. Grazie all'esperienza e alle osservazioni, gli egizi sapevano che il Nilo non tutti gli anni era distruttivo e che il suo contributo era fondamentale per il loro raccolto; così come sapevano (o meglio erano convinti di sapere, riprenderò il concetto più avanti) che c'era uno stretto legame tra livello segnato sul "Nilometro" e andamento del raccolto.

In epoca più recente, XVIII secolo in Pianura Padana, un territorio a noi più vicino, gli abitanti di allora dedussero che le piene più modeste del fiume Po potevano esser contenute grazie a dei manufatti in terra, gli argini. Ciò comportò notevoli vantaggi: la maggior vicinanza al corso d'acqua, l'impiego di terre più fertili e quindi raccolti più ricchi, maggior sicurezza (forse…), etc. È bene ricordare che in passato le piene del fiume Po allagavano tutta la parte centro-orientale del bacino padano e che a seguito degli eventi di piena più intensi il corso d'acqua subiva grosse alterazioni e talvolta rimaneva fuori alveo per diversi mesi, anche anni.

Al tempo degli egizi, l'unica cosa che si poteva fare era: osservare e cercare di capire; probabilmente, al tempo, gli egizi erano già convinti di aver capito, sicuramente facevano il loro meglio mettendo in pratica il massimo delle loro conoscenze.

Nel XVIII secolo, le conoscenze e le tecniche hanno permesso di compiere un passo in più, non solo osservare e capire, ma anche intervenire per proteggersi.

Dai primi argini, localizzati e spesso non sufficienti, si è passati ad una "cintura" arginale che accompagna il fiume Po per tutto il tratto di pianura; congiuntamente sono state realizzate altre opere strutturali di difesa, come le casse d'espansione o le aree golenali.

Nell'ultimo secolo, grazie ai numerosi passi -e qualche balzo- in avanti compiuto dall'informatica e dall'elettronica è stato possibile raccogliere, incrementare e mettere a sistema le conoscenze e le tecniche in materia di idraulica ed idrologia, provenienti anche da diverse aree geografiche.

A seguito del grande evento di piena del fiume Po nel 2000, si è deciso di mettere a sistema tutte le informazioni disponibili, di tipo sia osservativo sia previsionale; il sistema *FEWS-Po*, nato agli inizi del XXI secolo e in continuo sviluppo, è sicuramente un "gradino" dell'evoluzione osservativa-previsionale, per quanto riguarda il principale bacino italiano.

Il Sistema è nato dalla necessità di difendersi ulteriormente contro gli eventi pluviometrici ed idrometrici più intensi; soprattutto dopo il raggiungimento dei limiti strutturali delle arginature, ci si è resi conto che per migliorare le condizioni di sicurezza gli interventi strutturali mirati non erano più sufficienti, pertanto bisognava ricorrere ad estesi interventi non strutturali.

Causa (evento di piena), necessità (mitigazione del rischio attraverso il miglioramento e la condivisione del sistema di monitoraggio e previsione per il bacino del fiume Po) ed effetto (Sistema FEWS-Po).

Le prime implementazioni del Sistema FEWS-Po, che comprende anche l'Emilia-Romagna, sono state i punti di monitoraggio e i modelli previsionali disponibili; successivamente, grazie all'aumento delle conoscenze e allo sviluppo di nuove tecniche, il Sistema è stato, ed è tutt'ora, periodicamente aggiornato.

Nel corso degli anni, il costante incremento delle conoscenze, l'applicazione di nuove tecniche ma soprattutto gli eventi più severi hanno migliorato le modalità di osservazione e le capacità previsionali degli eventi pluvio-idro su tutto il bacino del fiume Po, Emilia-Romagna compresa.

Ma gli eventi di "piena lampo" ("*flash flood"*) verificatisi recentemente sui principali affluenti del fiume Po, ma non solo, hanno messo a dura prova l'intero sistema, sia sotto il profilo previsionale sia sotto quello operativo; precipitazioni intense, estremamente localizzate e di breve durata, difficilmente prevedibili dai modelli meteorologici previsionali che generano in 2-3 ore colmi di piena eccezionali sono ad oggi una delle sfide più grandi per l'intero sistema di protezione civile.

Un evento pluviometrico particolarmente intenso che si sviluppa rapidamente e che genera repentini innalzamenti idrometrici è estremamente difficile da prevedere con sufficiente accuratezza temporale e spaziale; ciò che però si può fare, è osservare. In fase di evento è possibile massimizzare l'utilizzo delle informazioni provenienti dai sensori pluviometrici ed idrometrici installati sul territorio; così da poter valutare in tempo reale l'entità dell'evento e i possibili sviluppi, al fine di permettere agli organi di protezione civile di intervenire nel minor tempo possibile e nel miglior modo possibile.

Questo lavoro di ricerca parte proprio da qui, da un'esigenza prettamente operativa, ovvero: in riferimento all'evento pluvio-idrometrico che si sta osservando, cercare di capire cosa succederà nell'immediato futuro -entro le 12 ore- in una data sezione idrometrica dei fiumi dell'Emilia-Romagna; al fine di supportare gli interventi di protezione civile in fase di evento.

L'implementazione all'interno sistema di monitoraggio e previsione per l'Emilia-Romagna di un modello speditivo utile in fase di *nowcasting* per prevedere nell'immediato futuro i possibili scenari idrometrici, attraverso le sole informazioni osservate, sarebbe stato sicuramente un valore aggiunto ad un sistema già solido ma in continua evoluzione.

Il primo approccio, con i dati di precipitazione e portata a disposizione (2005-2015), è stato quello di elaborare un modello speditivo basato sul Modello NASP (Biondi & Versace, 2007), già in operatività nei Centri Funzionali; la calibrazione ha interessato le principali sezioni dei corsi d'acqua dell'Emilia-Romagna. Successivamente, si è fatto ricorso ad una delle più note tecniche di apprendimento automatico, il metodo *Random Forests*, per sviluppare un modello particolarmente speditivo con la capacità di associare una probabilità di superamento nelle prossime ore delle soglie di allertamento di Protezione Civile, basandosi sugli ultimi dati osservati.

Le tecniche di apprendimento automatico (*machine learning*) sono un "ramo" dell'intelligenza artificiale; le più avanzate sono capaci di identificare i legami causa-effetto analizzando enormi moli di dati. Sviluppate inizialmente in campo medico a supporto dei processi diagnostici, sono state rapidamente applicate ed estese ad altri settori: automobilistico, informatico, pubblicitario, etc.

A differenza dei modelli classici (esempio, modello di Nash), che necessitano di un'attenta calibrazione per superare il test di validazione, prima, e quello operativo, poi; un modello basato sull'apprendimento automatico prevede solo un settaggio iniziale dell'algoritmo necessario per definire le modalità con cui questo apprenderà dal data set di addestramento (*training set*). Quando si elaborano modelli basati sulle tecniche di apprendimento automatico, è più corretto utilizzare il termine addestramento (*training*) piuttosto che calibrazione, come per i metodi classici; infatti, per ottenere output differenti non è sufficiente modificare uno o due parametri del modello, ma è necessario ridefinire le modalità con cui l'algoritmo apprende dal *training set*, ad esempio: modificando il valore del numero di alberi generati, il numero di variabili considerate, etc.

Modificare il settaggio di un algoritmo di apprendimento automatico non è semplice, al più si possono variare i pochi parametri che lo compongono; in ogni caso, ciò influisce direttamente sull'addestramento piuttosto che sui risultati. Se quest'ultimi risultano essere non soddisfacenti, molto probabilmente il problema risiede nei dati riportati nel *training set* o nella sua errata definizione e/o identificazione.

Tecniche di apprendimento come il *Random Forests* sono particolarmente performanti anche con *training set* molto grandi ed eterogenei; l'importante è che questi siano coerenti, ovvero: l'algoritmo, in fase di apprendimento, valuta il *training set* secondo il principio di "causa-effetto"; la presenza all'interno del *training set* di dati "incoerenti" potrebbe alterare la fase di apprendimento dell'algoritmo con conseguente errore di valutazione delle istanze in ingresso.

Pertanto, in presenza di dati incoerenti è consigliabile aggiungere dei pesi di valutazione a favore delle informazioni più attendibili e a sfavore dei dati meno attendibili; cosicché l'algoritmo, a parità di informazione, prediligerà le istanze aventi un peso maggiore. Se la numerosità di dei dati incoerenti, o ritenuti tali, risulta elevata è bene rimuovere tali istanze ed utilizzare solamente quelle ritenute attendibili.

Nel primo caso di studio affrontato, Capitolo 3, i dati disponibili per il periodo 2005-2015 erano già stati accuratamente validati; pertanto privi di errori strumentali e/o di ricezione. La buona validità dei dati è riscontrabile dalle performance ottenute in fase di calibrazione; a conferma che l'algoritmo è in grado di riconoscere e valutare correttamente le informazioni su cui è stato addestrato.

Al termine della fase di addestramento (*training*), è possibile verificare il peso che l'algoritmo ha associato alle variabili presenti nel *training set* per "addestrarsi"; pur non considerando alcuna informazione relativa alla fisicità dei bacini o alle principali nozioni di idraulica, l'algoritmo ha valutato come "prioritarie" le informazioni relative agli ultimi valori di portata osservati (generalmente, ultime 3-5 ore; nei bacini più grandi 6-9 ore), collocando a valle di queste le informazioni relative alle altezze di pioggia.

Ad esempio, in fase di calibrazione del modello di Nash per il bacino del fiume Parma alla sezione di chiusura Ponte Verdi è stato stimato un tempo di risposta del bacino di circa 3-4 ore; il tempo di risposta è il tempo che intercorre tra il volume di pioggia osservato e l'incremento idrometrico da esso generato alla sezione di riferimento. Sensibilmente più breve del tempo di corrivazione, oltre nove ore e mezzo applicando la formula di Giandotti. Lo stesso arco temporale ritenuto rilevante ai fini dell'apprendimento dall'algoritmo *Random Forests*. Come anticipato, la fisicità del bacino non è indotta nelle tecniche di apprendimento automatico, ma è -in parte- dedotta dall'algoritmo al termine della fase di apprendimento.

Nel secondo caso di studio, relativo alle principali sezioni del bacino del fiume Po, il training set è stato integrato con le informazioni idrometriche osservate nelle sezioni precedenti e con le informazioni previsionali prodotte dalla catena modellistica Mike11 NAM/HD. A scopo di ricerca, l'obiettivo era di elaborare un modello, basato sulla tecnica del Random Forests, che associasse una probabilità di superamento ai possibili livelli idrometrici o valori di portata previsti e valutarne la bontà confrontando i risultati ottenuti con i modelli e le tecniche già sviluppate. Anche in questo caso, il modello non considera in ingresso la fisicità del bacino; al termine della fase di *training* fornisce però l'importanza associata alle variabili presenti nel training set. Confrontando la nota fisicità del bacino e la valutazione delle variabili da parte dell'algoritmo, si può apprezzare come quest'ultimo abbia correttamente "dedotto" che gli ultimi valori osservati hanno un peso maggiore per la valutazione probabilistica previsionale, rispetto alle informazioni previsionali provenienti dalla catena modellistica Mike11 NAM/HD; congiuntamente, emerge una parziale difficoltà nel valutare i possibili scenari futuri per le prime sezioni (Piacenza e Cremona) a causa dell'assenza delle informazioni provenienti dalla sezioni più a monte (Piacenza) e all'apporto dei tributari lombardi, non considerato direttamente nel *training set* (Cremona). Performance maggiormente soddisfacenti si hanno soprattutto nella sezione di chiusura Pontelagoscuro, il cui training set è composto sia dalle ultime informazioni osservate relative alla sezione stessa sia dalle informazioni osservate alle sezioni di monte: permettere all'algoritmo di "conoscere" ciò che è già successo nelle sezioni di monte, gli consente di valutare con maggiore accuratezza cosa succederà nella sezione d'interesse posta più a valle; soprattutto, nel caso in cui il fenomeno di traslazione dell'onda di piena prevale su quello di laminazione.

Se si fa riferimento ai principali indici di valutazione delle performance, POD (Probability of Detection) e FAR (False Allarme Ratio), nel secondo caso di studio i valori, per quanto soddisfacenti, sono risultati inferiori rispetto a quelli ottenuti nel primo caso di studio; ciò è attribuibile ad un *training set* non sufficientemente eterogeneo e consistente: il numero degli eventi di piena del fiume Po osservati è inferiore rispetto agli eventi di piena osservati negli affluenti emiliano-romagnoli nel periodo 2005-2015.

Le performance ottenute nei due casi di studio sono state confrontate con due delle più recenti ed avanzate tecniche per la stima dell'incertezza predittiva; sviluppate presso l'Università di Bologna ed oggetto di numerose pubblicazioni di settore. La prima tecnica "*BMBP – Bayesian Multivariate Binary Predictor*" (Todini, 2008) non necessita delle ipotesi del modello probabilistico è stata applicata ad entrambi i casi di studio; mentre, la seconda, "MCP - *Model Conditional Processor"* (Todini, 2008) è un post-processore bayesiano per la stima dell'incertezza predittiva.

CONSIDERAZIONI SU RISULTATI E PERFORMANCE

Sia nel primo caso di studio sia nel secondo, i valori degli indici di valutazione delle performance sono risultati molto elevati; soprattutto nel primo caso di studio dove *POD* è risultato essere uguale a uno e FAR pari a zero. Ovvero, basandosi solo sull'informazione derivata dagli indici emerge che il modello è in grado di prevedere correttamente tutti gli eventi, senza generare neanche un mancato allarme.

Una tale considerazione della qualità dei modelli comporterebbe sia un errore di valutazione sia un grave errore in fase operativa: in quanto, l'utente finale, si aspetterebbe dai modelli una previsione "perfetta", in ogni caso.

Per evitare di cadere in questo grossolano errore è necessario, in fase di elaborazione del modello, tener conto delle principali peculiarità delle tecniche di apprendimento automatico: queste forniscono elevate performance in fase di *training*; le stesse performance sono riscontrabili in fase operativa se i valori delle variabili del dataset operativo sono simili a quelli contenuti nel *training set*.

In caso contrario le performance potrebbero decadere rapidamente; pertanto è necessario supportare modelli di questo tipo con altri modelli o tecniche, come di seguito, oppure "rinforzare" l'apprendimento del modello con serie sintetiche.

In ogni caso, è opportuno informare colui che beneficerà dell'utilizzo del modello, sul periodo di *training* utilizzato e sulle relative performance; ad esempio, nel primo caso di studio: numero di eventi superiori a soglia 3 contenuti nel training set, numero di eventi superiori a soglia 2 contenuti nel training set, etc.

Se si fa riferimento ai soli risultati ottenuti, senza considerare le principali caratteristiche delle tecniche di apprendimento automatico, si rischia di valutare i modelli sviluppati in questi due casi di studio come "impeccabili".

DAL TRAINING SET ALL'OPERATIONAL SET

L'elevata capacità del metodo Random Forests di apprendere da un "training set" coerente, eterogeneo e consistente è già stata ampiamente descritta; ma per verificare il comportamento dell'algoritmo in fase operativa sono stati condotti alcuni test utilizzando sia gli eventi reali, ma non presenti nel *training set*, sia gli eventi "sintetici", ricavati incrementando i valori di afflusso e deflusso di alcuni dei principali eventi reali.

Considerando le peculiarità del Random Forests e le varie prove condotte durante il periodo di ricerca, com'era lecito aspettarsi e come suggeriscono le principali fonti bibliografiche consultate, in operatività la tecnica di apprendimento automatico utilizzata è tanto più vincente quanto più l'osservato è simile allo storico. In "real-time", se l'evento pluvio-idrometrico che l'algoritmo "sta osservando" è simile ad uno o più eventi da esso già osservati e valutati in fase di *training*; allora, l'informazione previsionale generata dall'algoritmo risulterà attendibile.

In breve, l'algoritmo è sufficientemente "preparato" per valutare correttamente la tipologia di evento in corso.

La situazione cambia radicalmente nel caso in cui l'algoritmo è costretto a valutare un evento in "*real time*" con caratteristiche particolarmente differenti rispetto a quelli già osservati in fase di *training*.

In riferimento al primo caso di studio, di seguito, verranno presentati tre casi emblematici relativi al maggior evento pluvio-idrometrico verificatosi sul bacino del fiume Parma; il comportamento del modello verrà valutato su:

- Caso A: evento pluvio-idrometrico estremo non presente nel *training set*
- Caso B: evento pluvio-idrometrico estremo presente nel *training set*
- Caso C: evento pluvio-idrometrico estremo sintetico non presente nel *training set*

Il 13 ottobre 2014 intense e persistenti precipitazioni interessarono anche il bacino del fiume Parma con conseguente innalzamento repentino dei livelli idrometrici.

Da un'analisi dettagliata dell'evento è possibile apprezzarne la sua eccezionalità:

- L'intensità massima oraria dell'evento, calcolata sull'intero bacino, è stata di poco superiore al massimo valore fino ad allora osservato (15,28 mm/h, 13 ottobre 2014 ore 13:00, contro i 15.17 mm/h, 20 ottobre 2013).
- L'intensità tri-oraria massima dell'evento, calcolata sull'intero bacino, è stata sensibilmente superiore al massimo valore fino ad allora osservato; non solo, durante l'evento del 2014, il massimo valore fino ad allora osservato è stato superato ben quattro volte. Inoltre, l'intensità tri-oraria registrata durante l'evento del 2014, occupa oggi i primi otto valori; a conferma che l'eccezionalità dell'evento è riscontrabile in tutta, o quasi, la sua durata (dalle ore 12:00 alle ore 16:00).

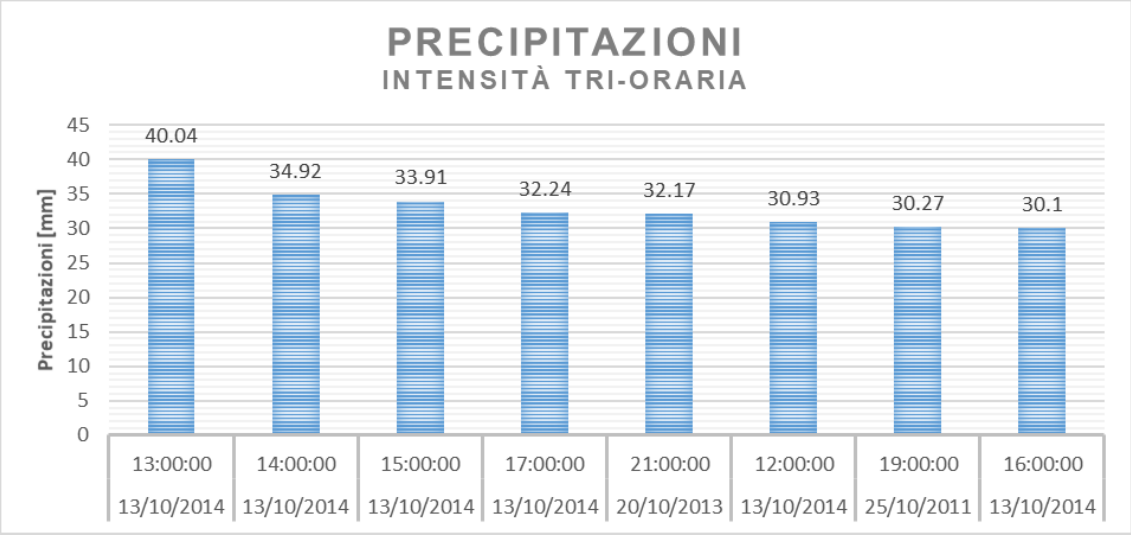


Figura 72 Massima intensità tri-oraria delle precipitazioini osservate nel periodo 2005-2015

- L'evento si è verificato dopo un lungo periodo siccitoso, i valori idrometrici alle sezioni principali erano ai minimi annuali e talvolta vicini ai minimi osservati; di conseguenza tutto il bacino era completamente asciutto.
- Il nubifragio che ha colpito il bacino ha dato origine ad un colmo di piena quasi "istantaneo" ("*flash flood*"); sono state sufficienti le prime tre ore di nubifragio per generare un rapido incremento dei livelli idrometrici.

- Tra le 15:00 e le 16:00, un'ora, il valore di portata alla sezione di Ponte Verdi è passato da 123.00 m^3/s a 584 m^3/s; se si fa riferimento alle soglie di Protezione Civile, in meno di un'ora, si è passati da "assente" a "livello 3 – criticità elevata". Nell'ora successiva, dalle 16:00 alle 17:00, il valore di portata è aumentato fino a 939 m^3/s (colmo di piena e valore mai misurato prima).
- I due incrementi orari menzionati non solo sono stati superiori rispetto al massimo incremento orario fino ad allora osservato; ma tra le 15:00 e le 16:00 l'incremento è stato pari quasi al doppio del massimo incremento orario osservato fino al 2014.

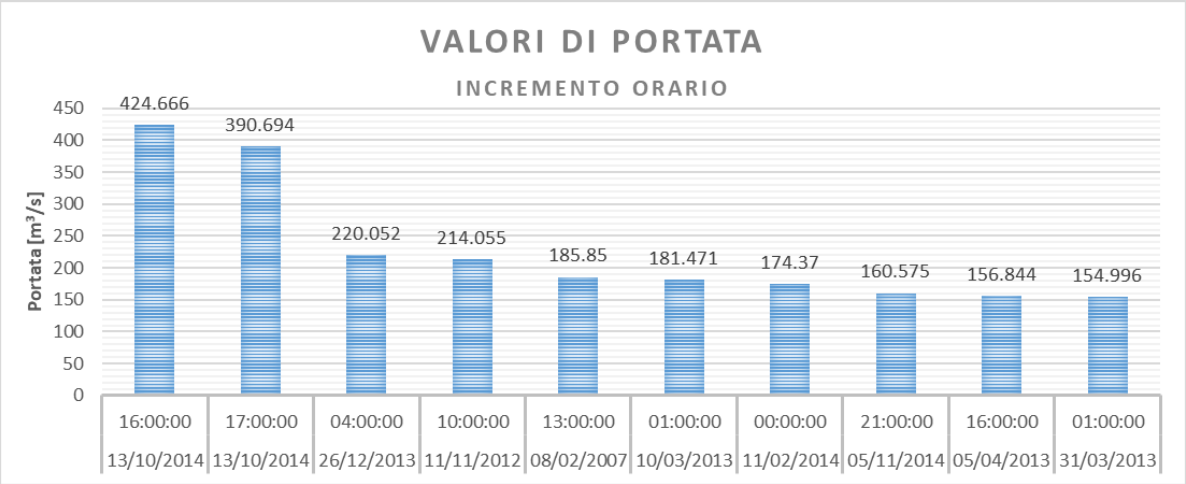


Figura 73 Massimi valori di portata osservati nel periodo 2005-2015

Ricordando quanto riportato nel capitolo 3, relativo al primo caso di studio: il bacino del fiume Parma alla sezione principale, Ponte Verdi, ha un tempo di risposta associato di circa sei ore; pertanto, il set di dati utilizzato in fase di *training*, e quindi anche in operatività, è composto dalle ultime sei ore di precipitazione e dagli ultimi sei valori di portata osservata. Inoltre, come presentato nel Capitolo 3, l'algoritmo in fase di valutazione considera prima gli ultimi valori di portata e successivamente i valori di precipitazione.

CASO A: evento pluvio-idrometrico estremo non presente nel *training set*

Se nel *training set* omettiamo questo evento e successivamente interroghiamo l'algoritmo per valutare l'evento, con la foresta casuale elaborata senza di esso, si ottiene il seguente risultato:

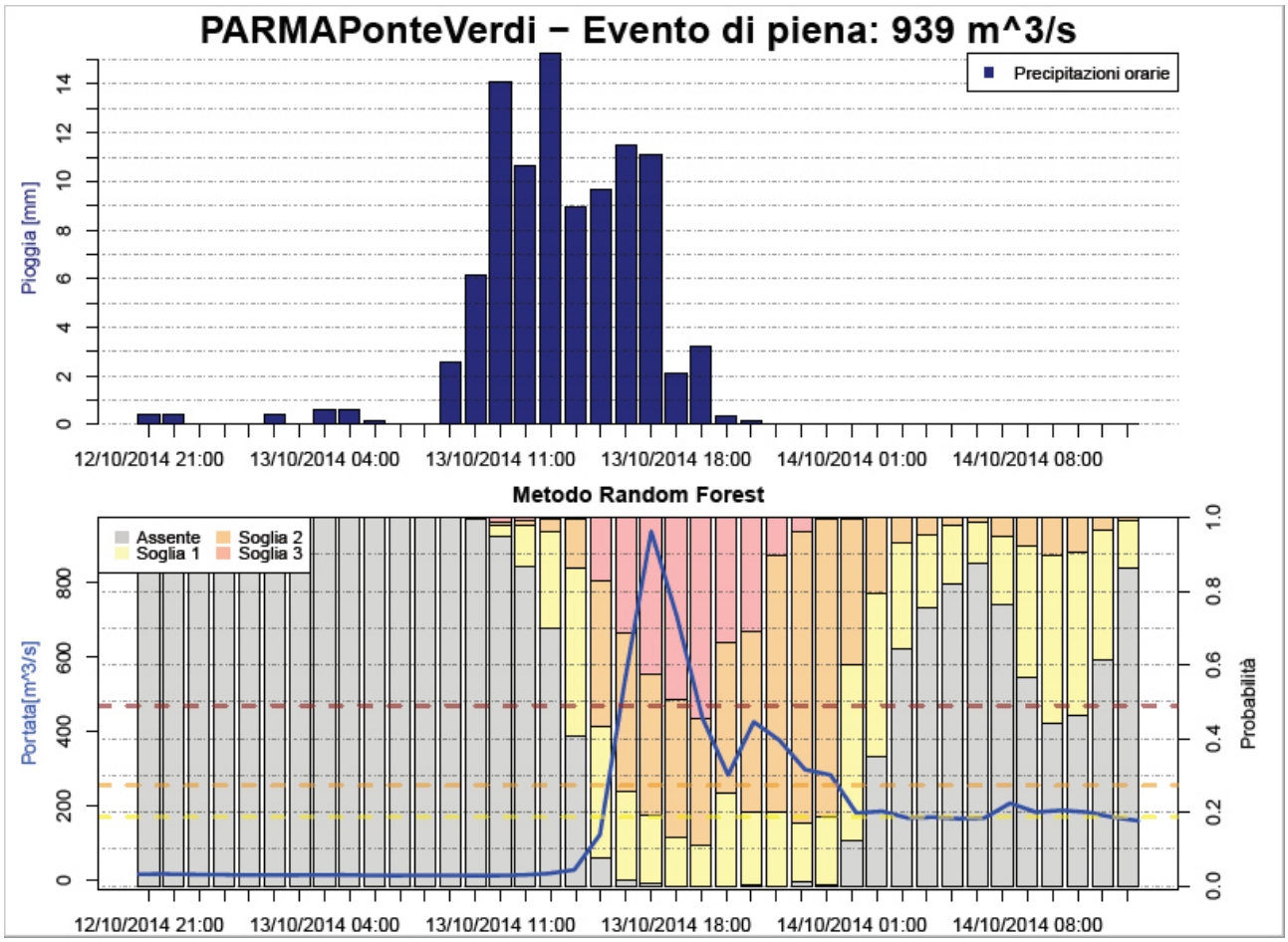


Figura 74 Rappresentazione grafica dell'output del modello in riferimento all'evento 2014, non presente nel *training set*

La probabilità di superamento della soglia 1 e 2 è maggiore delle altre solo ad un'ora dall'effettivo superamento; mentre, la probabilità di superamento della soglia 3 rimane bassa anche durante il suo effettivo superamento.

Una performance decisamente inferiore rispetto a quelle presentate nel capitolo 3.

Il motivo di tale deficit previsionale è identificabile nell'addestramento dell'algoritmo che è stato condotto su un *training set* dove i valori di portata molto bassi sono stati seguiti da incrementi sensibilmente più modesti rispetto a quelli del 2014; così come le precipitazioni orarie, intensità così elevate erano uniche all'interno degli eventi presenti nel *training set* e non si ripetevano per più ore come per l'evento del 2014.

In sostanza, l'algoritmo non avendo mai visto incrementi orari di portata così repentini associati a precipitazioni così intense, non ha valutato correttamente l'evento; sottostimandolo.

Mentre, se focalizziamo l'attenzione sulla fase di esaurimento dell'onda di piena, il modello prevede correttamente e con le giuste tempistiche il rientro in "Soglia 2" e "Soglia 1"; ciò è attribuibile al fatto che la fase di esaurimento dell'evento idrometrico è stata molto simile a quelle già osservate e presenti nel *training set*.

CASO B: evento pluvio-idrometrico estremo presente nel *training set*

Inserendo nel *training set* l'evento estremo del 2014, come fatto nel Capitolo 3, e successivamente interrogando l'algoritmo per valutare l'evento, con la foresta casuale elaborata anche con esso, si ottiene il seguente risultato:

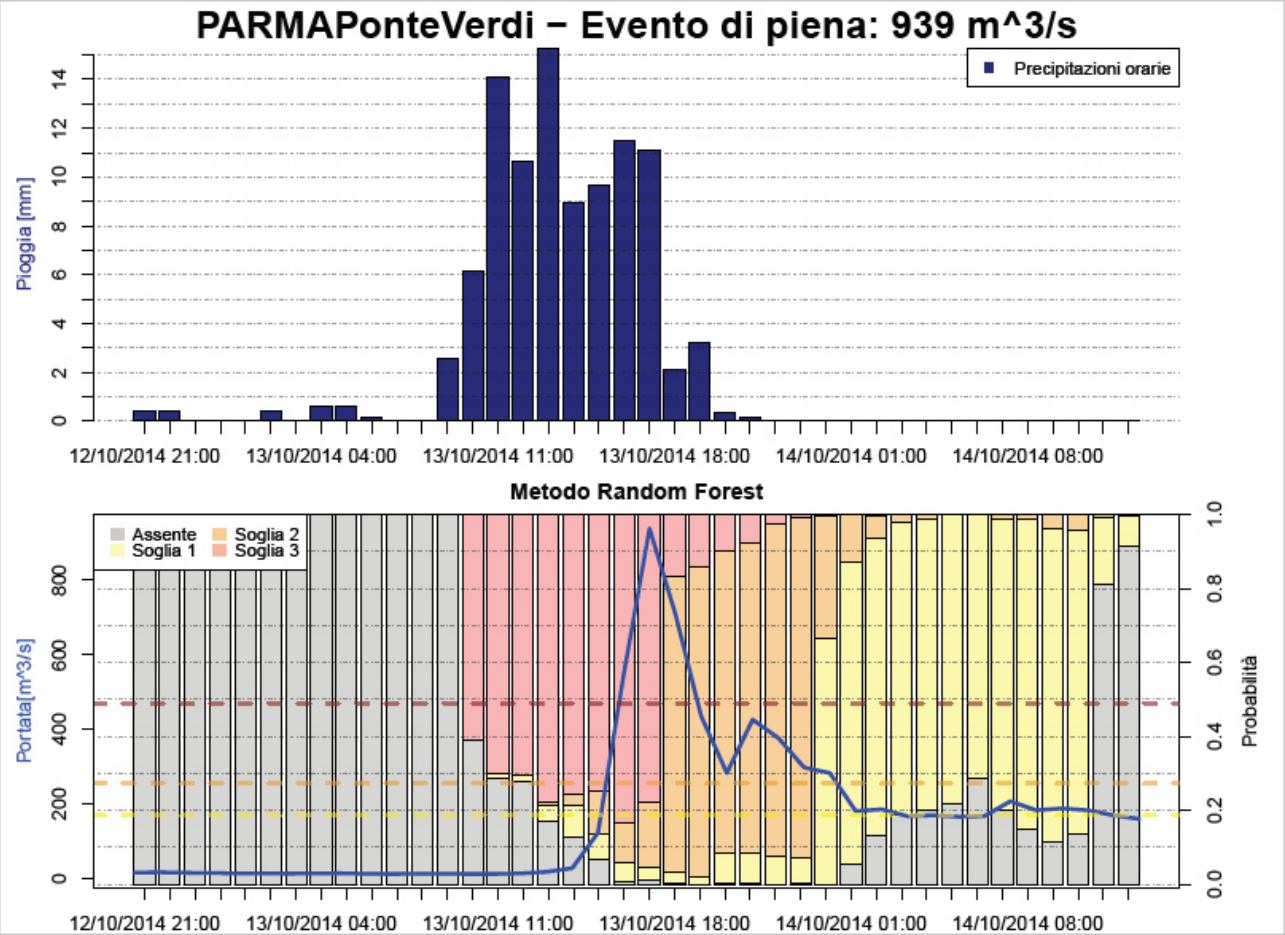


Figura 75 Rappresentazione grafica dell'output del modello in riferimento all'evento 2014, presente nel training set

In questo caso, già dopo le prime due ore di pioggia, la probabilità si superamento della soglia 3 cresce sensibilmente e rimane elevata anche nelle ore successive; una corretta valutazione così anticipata è possibile in quanto l'algoritmo ha appreso da un *training set* dove era presente almeno un caso con valori di portata molto bassi e pioggia intensa per più di due ore consecutive.

CASO C: evento pluvio-idrometrico estremo sintetico non presente nel *training set*

Ipotizziamo di vover valutare un evento non presente nel *training set* ed ancora più estremo in termini di precipitazioni, stesso ietogramma del 2014 incrementato del 25%, e portata, stesso idrogramma incrementato del 25%; il *training set* è corredato con tutti i dati a disposizione 2005-2015 (evento estremo del 2014, compreso).

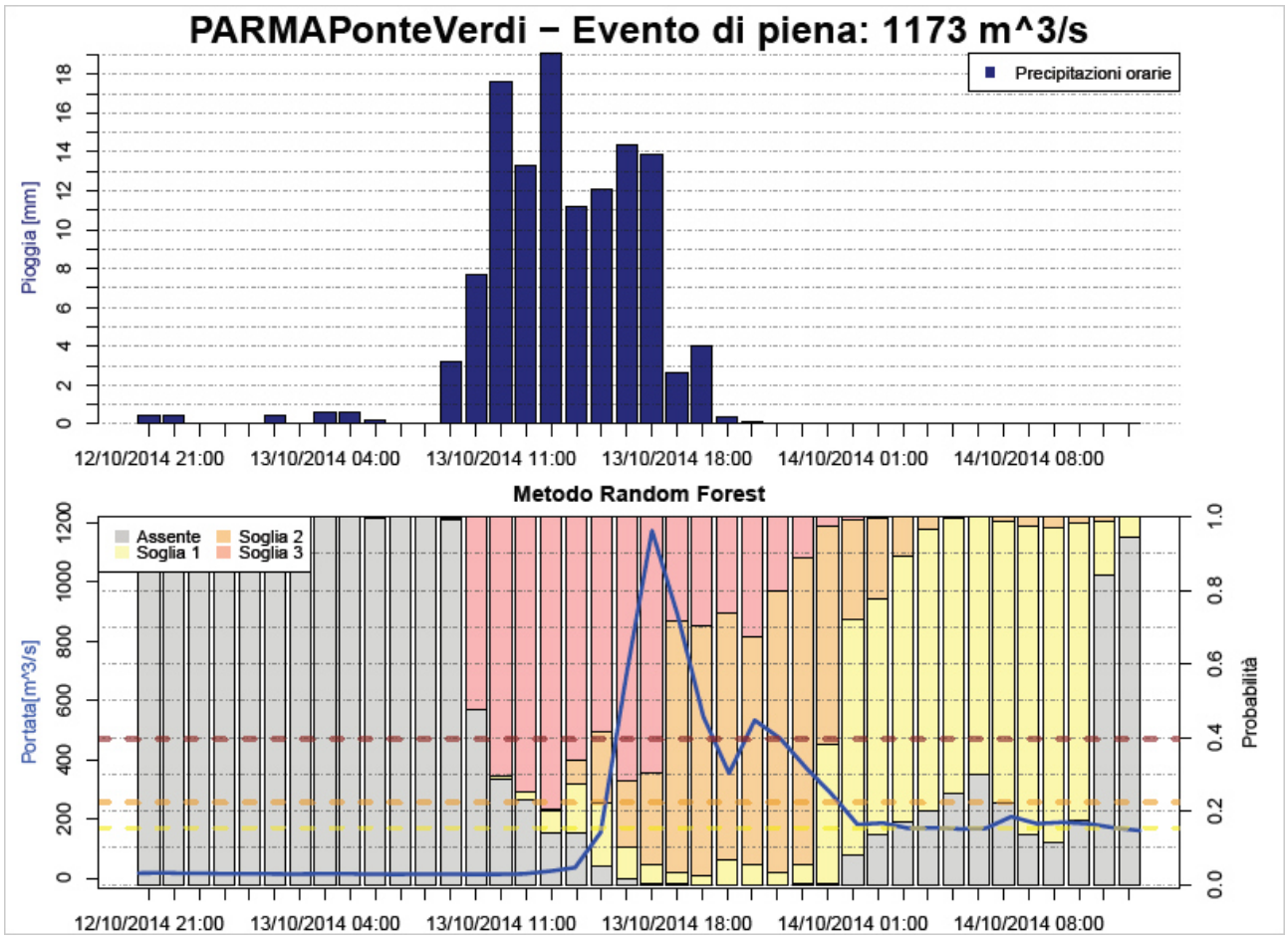


Figura 76 Rappresentazione grafica dell'output del modello in riferimento ad un evento più estremo rispetto a quello del 2014, non presente nel training set

In questo caso "sintetico", nonostante si tratti di un caso ancora più estremo rispetto all'evento del 2014, nei risultati la probabilità di superamento della soglia 3 prevale sulle altre con sufficiente anticipo rispetto all'effettivo superamento; questo perché l'algoritmo "ha già visto" una situazione simile, dove: nonostante valori di portata molto bassi a causa delle intense precipitazioni si è avuta una rapidissima crescita del valore di portata.

L'algoritmo ha potuto apprendere la condizione "limite inferiore": rapidi incrementi idrometrici pur con portate iniziali molto basse; con tale condizione di bacino, in presenza di afflussi pari o superiori a quelli del 2014 questo è in grado di valutare correttamente il superamento di soglia 3.

Pertanto è corretto sostenere che l'algoritmo ha un'elevata capacità di apprendimento ed un'elevata capacità di valutazione degli eventi simili o "compresi" a quelli riportati nel *training set*; d'altra parte è necessario identificare preventivamente i limiti della foresta casuale elaborata con un determinato *training set* (ad esempio, se questo è privo di eventi in soglia 3, oppure, la numerosità non è sufficiente per ottenere un'elevata performance in fase operativa.

SVILUPPI FUTURI

Per evitare scarse performance previsionali in fase operativa a seguito di eventi eccezionali, come quello dell'ottobre 2014 per il bacino del fiume Parma, è consigliabile realizzare preventivamente delle serie

sintetiche "limite" da inserire nel *training set*, come precipitazioni intense e persistenti su un bacino inizialmente asciutto. Come già riportato, l'apprendimento da parte dell'algoritmo di "situazioni limite" consente di ottenere buone performance previsionali anche per eventi più intensi.

In alternativa, è consigliabile adottare un approccio più cautelativo riducendo i valori delle soglie di allertamento.

Inoltre, durante gli eventi di precipitazione più intensi è consigliabile affiancare il modello *Random Forests* con un modello statistico che valuti l'intensità delle precipitazioni osservate prima ancora che l'afflusso si trasformi in deflusso; così da guadagnare minuti preziosi, spesso vitali, per le valutazioni e/o gli interventi di protezione civile.

L'algoritmo sia in fase di *training* sia in fase operativa, valuta l'informazione in funzione della "variazione" rispetto all'osservazione precedente; tanto maggiore è frequente l'informazione fornita all'algoritmo tanto più rapida sarà la valutazione da parte dello stesso.

Gli eccezionali incrementi orari di portata durante l'evento dell'ottobre 2014 sono apprezzabili anche a passi temporali più ridotti, ad esempio, 15 minuti. Pertanto, l'algoritmo parametrizzato con informazioni orarie fornisce una singola informazione previsionale ogni ora; mentre, l'algoritmo parametrizzato con informazioni al quarto d'ora fornisce quattro informazioni previsionali ogni ora. Un aggiornamento previsionale più frequente permette di valutare con un maggiore anticipo gli incrementi di precipitazione o portata; guadagnando così minuti preziosi in condizioni critiche di valutazione ed intervento.

Il modello Random Forests calibrato per le principali sezioni dei fiumi dell'Emilia-Romagna ed inserito all'interno del Sistema FEWS-Po, considerando il tempo di *polling* delle stazioni pluvio-idrometriche, è stato calibrato con un passo temporale di 15 minuti e la valutazione delle nuove informazioni in ingresso viene effettuata ogni 15 minuti; così da ottimizzare il più possibile la fase di valutazione dell'informazione osservata.

Infine, in riferimento al secondo caso di studio, per aumentare la capacità di apprendimento e di conseguenza migliorare anche la capacità di valutazione dell'algoritmo, il modello è stato modificato per consentire l'aggiunta "in continuo" dell'ultima istanza osservata all'interno del *training set* ed il successivo ricalcolo dell'albero decisionale; con questa modalità si avrà sempre il *training set* aggiornato all'ultima osservazione e un albero nuovo "*on the fly*" ad ogni aggiornamento elaborato con tutte le informazioni disponibili fino a quell'istante. Come mostrato, tale implementazione è stata presentata all'interno del secondo caso di studio (Capitolo 4), a breve sarà disponibile anche nel Sistema FEWS-Po nel modello relativo ai fiumi dell'Emilia-Romagna.

Il continuo aggiornamento delle informazioni disponibili con conseguente ricalcolo delle informazioni richieste è una delle peculiarità vincenti delle tecniche di apprendimento automatico: l'aggiornamento del *training set* è completamente automatico e l'algoritmo può contare su un data set di addestramento completo e aggiornato in "*real time*".

Riprendendo il concetto base di un saggio più recente di Taleb, "*Antifragile – Prosperare nel disordine*" (Taleb, 2014), l'aggiornamento ed il ricalcolo in continuo rendono queste tecniche "antifragili". A differenza di una tecnica robusta, valida nella maggior parte dei casi ma inefficace nell'eccezionalità; la tecnica "antifragile" subisce l'evento eccezionale -errata valutazione- ma istantaneamente ne apprende le caratteristiche, così da non subirlo una seconda volta, e soprattutto l'apprendimento immediato permette di ridurre il divario di valutazione nei confronti di un nuovo evento, ancora più eccezionale del precedente. Un modello di questo

tipo non è di certo infallibile, ma resiste, migliora e si evolve in seguito al susseguirsi degli eventi. Questa è una delle peculiarità vincenti delle tecniche di apprendimento dell'intelligenza artificiale.

BIBLIOGRAFIA

AIPO (Agenzia Interregionale per il fiume Po), 2014. Relazione preliminare sugli eventi di Parma e Baganza del 13-14 ottobre 2014

ASCE Task Committee on application of Artificial Neural Networks in Hydrology, 2000. Artificial neural networks in hydrology. I: preliminary concepts. J. Hydrol. Eng. https://doi.org/10.5121/ijsc.2012.3203

ASCE Task Committee on Application of Artificial Neural Networks in Hydrology, 2000. Artificial Neural Networks in Hydrology II: Hydrologic Applications. J. Hydrol. Eng.

Bahremand, A. &. M. R., 2009. Mathematical computation of Nash model parameters for hydrograph prediction. s.l., 3rd International Conference on Approximation Methods and Numerical Modelling in Environment and Natural Resources, (p. 1-5). , pp. 1-5.

Bennett, T., 1988. Development and application of a continuous soil moisture accounting algorithm for the hydrologic engineering center hydrologic modeling system (HEC-HMS). M.S. Thesis. Dept. of Civil and Environmental Engineering, University of Califor.

Bhunya, P.K., Ghosh, N.C., Mishra, S.K., Ojha, C.S., Berndtsson, R., 2005. Hybrid Model for Derivation of Synthetic Unit Hydrograph. J. Hydrol. Eng. https://doi.org/10.1061/(asce)1084-0699(2005)10:6(458)

Bhunya, P.K., Mishra, S.K., Ojha, C.S.P., Berndtsson, R., 2004. Parameter Estimation of Beta Distribution for Unit Hydrograph Derivation. J. Hydrol. Eng. https://doi.org/10.1061/(asce)1084-0699(2004)9:4(325)

Biondi, D., Versace, P., 2007. Peak flow estimation under parameter uncertainty in a real time flood warning system for ungauged basins.

Breiman, L., 2001. Manual on setting up, using, and understanding random forests v3. 1. Tech. Report, http//oz.berkeley.edu/users/breiman, Stat. Dep. Univ. Calif. Berkeley, …. https://doi.org/10.2776/85168

Breiman, L., 1984. Classification and regression trees Regression trees. Encycl. Ecol. https://doi.org/10.1007/s00038-011-0315-z

Brian, M., Dell'Aquila, V., Leoni, P. & Pecora, S., 2018. La modellistica idrologica per le emergenze. Ecoscienza, pp. 28-29.

Chen, C., Liaw, A., Breiman, L., 2004. Using random forest to learn imbalanced data, Berkeley Department of Statistics Tech Reports. https://doi.org/ley.edu/sites/default/files/tech-reports/666.pdf

Chen, Y., 1973. Mathematical Modeling of Water and Sediment Routing in Natural Channels. Ph.D. Dissertation, Department of Civil Engineering, Colorado State University, Ft. Collings, CO: s.n.

Coccia, G., 2011. Analysis and developments of uncertainty processors for real time flood forecasting.. s.l.:Alma Mater Studiorum, Università di Bologna. doi:10.6092/unibo/amsdottorato/3423.

Coccia, G., Todini, E., 2011. Recent developments in predictive uncertainty assessment based on the model conditional processor approach. Hydrol. Earth Syst. Sci. https://doi.org/10.5194/hess-15-3253-2011

Condorcet, J.D.A., 1785. Essai sur l'application de l'analyse a la probabilite des decisions rendues a la pluralite des voix, Condorcet Selected Writings.

Cutler, A., Cutler, D.R., Stevens, J.R., 2012. Random forests, in: Ensemble Machine Learning: Methods and Applications. https://doi.org/10.1007/9781441993267_5

Deltares, S., 2014. SOBEK User Manual. Delft, the Netherlands.: s.n.

Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics. https://doi.org/10.1186/1471-2105-7-3

Dooge, J.C.I., 1973. Linear Theory of Hydrologic System, US Department of Agriculture. https://doi.org/10.1016/0003-6870(73)90259-7

Giustolisi, O., 2000. Simulating a Urban Drainage System by Non-Linear Time Invariant Dynamic Systems: Neural Networks.. Iowa City, Iowa State, USA 23-26 July, s.n.

Goswami, M., O'Connor, K.M., 2007. Comparative assessment of six automatic optimization techniques for calibration of a conceptual rainfall-runoff model. Hydrol. Sci. J. https://doi.org/10.1623/hysj.52.3.432

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning The Elements of Statistical LearningData Mining, Inference, and Prediction, Second Edition, Book. https://doi.org/10.1007/978-0-387-84858-7

Houghton-carr, H.A., 1999. Assessment criteria for simple conceptual daily rainfall-runoff models. Hydrol. Sci. J. https://doi.org/10.1080/02626669909492220

Izenman, A.J., 2008. Modern Multivariate Statistical Techniques. https://doi.org/10.1007/978-0-387-78189-1

Klemeš, V., 1988. A hydrological perspective. J. Hydrol. https://doi.org/10.1016/0022-1694(88)90179-5

Leoni, P. Montanari, A., 2014. Assessment of dynamic of flood in the Secchia River (Italy). IAHS2014, Bologna, Italia

Leoni, P., 2014. Analisi, simulazione e classificazione delle principali piene del fiume Po. Tesi Magistrale. Alma Mater Studiorum, Università di Bologna.

Ljung, L., 2002. Black-box models from input-output measurements. https://doi.org/10.1109/imtc.2001.928802

Louppe, G., 2014. Understanding Random Forests. Cornell Univ. Libr. https://doi.org/10.13140/2.1.1570.5928

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Adv. Water Resour. https://doi.org/10.1016/S0309-1708(02)00092-1

Mazzetti, C., Todini, E., 2010. Combining weather radar and raingauge data for hydrologic applications, in: Flood Risk Management: Research and Practice. https://doi.org/10.1201/9780203883020.ch159

Mitchell, T. M., 1997. Machine Learning. McGraw-Hill Science/Engineering/Math. 2-3.

Moradkhani, H., Sorooshian, S., 2008. General Review of Rainfall-Runoff Modeling: Model Calibration, Data Assimilation, and Uncertainty Analysis, in: Hydrological Modelling and the Water Cycle. https://doi.org/10.1007/978-3-540-77843-1_1

Nielsen, S.A., Hansen, E., 2018. Numerical simulation of the rainfall-runoff process on a daily basis. Hydrol. Res. https://doi.org/10.2166/nh.1973.0013

O'Connell, P., 1991. A historical perspective in recent advances in the Modeling of Hydrologic Systems. A historical perspective in recent advances in the Modeling of Hydrologic Systems. Bowles D.S. and O'Connell P.E. (Eds.), Kluwer: Dordrecht, pp. 3 – 30. 1-12.

Rigon, R., D'Odorico, P., Bertoldi, G., 2011. The geomorphic structure of the runoff peak. Hydrol. Earth Syst. Sci. https://doi.org/10.5194/hess-15-1853-2011

Shaikhina, T., Lowe, D., Daga, S., Briggs, D., Higgins, R., Khovanova, N., 2017. Decision tree and random forest models for outcome prediction in antibody incompatible kidney transplantation. Biomed. Signal Process. Control. https://doi.org/10.1016/j.bspc.2017.01.012

Shannon, C.E., 1948. A Mathematical Theory of Communication (Part I). Bell Syst. Tech. J. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x

Slater, L.J., Thirel, G., Harrigan, S., Delaigue, O., Hurley, A., Khouakhi, A., Prodoscimi, I., Vitolo, C., Smith, K., 2019. Using R in hydrology: a review of recent developments and future directions. Hydrol. Earth Syst. Sci. Discuss. https://doi.org/10.5194/hess-2019-50

Soil Conservation Service Engineering Division, 1972. Section 4: Hydrology, in: National Engineering Handbook.

Somalvico, M., 1997. Intelligenza Artificiale. Roma: Enciclopedia italiana di scienze lettere ed arti (p. 735-738).

Taleb, N., 2005. Giocati dal caso. Il ruolo della fortuna nella finanza e nella vita. Il Saggiatore.

Taleb, N., 2014. Antifragile. Il Saggiatore.

Todini, E., Ciarapica, L., 2002. The TOPKAPI Model, in: Mathematical Models of Large Watershed Hydrology.

Todini, E., 2005. Rainfall-Runoff Models for Real Time Flood Forecasting. In: Encyclopedia of Hydrological Sciences, J. Wiley & Sons.. p. Chapter HSA131.

Todini, E., 2011. Bayesian Conditioning of a Random Field to Point Measurements. https://doi.org/10.1007/978-94-010-0810-5_35

Todini, E., 2008. A model conditional processor to assess predictive uncertainty in flood forecasting. Int. J. River Basin Manag. https://doi.org/10.1080/15715124.2008.9635342

Todini, G., 2008. A new snowfall detection algorithm for high latitude regions based on a combination of active and passive sensors. Bologna: Ph. D. Dissertation thesis, Alma Mater Studiorum, Università di Bologna. DOI 10.6092/unibo/amsdottorato/982 0.

Turing, A., 1950. Computing machinery and intelligence-AM Turing. Mind Q. Rev. Psychol. Philos.

WMO, 2011. Manual on flood forecasting and warning, World Meteorological Organization. https://doi.org/10.1016/j.transproceed.2010.08.027

WMO, 2006. TECHNICAL REGULATIONS: Volume III - Hydrology, WMO-No.49.

Zhang, H., Singer, B.H., 2010. Recursive Partitioning and Applications.

RINGRAZIAMENTI

Al termine di questo intenso percorso, desidero ringraziare il Prof. Alberto Montanari per l'opportunità concessami e per i preziosi contributi.

Inoltre, vorrei esprimere i miei più sinceri ringraziamenti all'Ing. Silvano Pecora, Dirigente dell'Area Idrologia-Idrografia di Arpae Emilia Romagna, per avermi concesso la possibilità di condurre l'attività di ricerca presso l'Area; ma soprattutto, per tutti i consigli e gli stimoli durante questi anni.