

UTILIZZO DI TECNICHE DI MACHINE LEARNING PER LA DEFINIZIONE DI MODELLI DI ALLERTA DA FRANA A SCALA REGIONALE

Gaetano Pecoraro, Michele Calvello

Dipartimento di Ingegneria Civile (DICIV), Università degli Studi di Salerno

gpecoraro@unisa.it, mcalvello@unisa.it

Guido Rianna, Alfredo Reder

Centro Euro-Mediterraneo sui Cambiamenti Climatici (CMCC)

guido.rianna@cmcc.it, alfredo.reder@cmcc.it

Massimo Esposito, Marco Pota

Istituto di Calcolo e Reti ad Alte Prestazioni del Consiglio Nazionale delle Ricerche (ICAR-CNR)

massimo.esposito@icar.cnr.it, marco.pota@icar.cnr.it

Sommario

I sistemi di allerta da frana a scala regionale sono tipicamente basati su correlazioni tra le frane avvenute nel passato e dati di monitoraggio pluviometrico al fine di individuare delle soglie pluviometriche d'innesco. Recentemente, la disponibilità di grandi dataset di forzanti atmosferiche offre la possibilità di includere ulteriori variabili incrementando l'affidabilità dei modelli. Tuttavia, condurre le analisi con le tecniche tradizionali può risultare piuttosto complesso e oneroso. Lo scopo di questo studio preliminare è dimostrare che le tecniche di machine learning possono essere utilizzate per analizzare dati di monitoraggio allo scopo di selezionare le variabili più significative per l'innesco di frane. I modelli sviluppati sono stati testati in una delle zone d'allerta definite dalla protezione civile per la gestione del rischio idrogeologico in Campania. Le variabili atmosferiche sono derivate dalla rianalisi atmosferica ERA5-Land (<http://dx.doi.org/10.24381/cds.adbb2d47>). I dati sugli eventi di frana sono ricavati da "FraneItalia", un catalogo georeferenziato di fenomeni franosi avvenuti in Italia sviluppato consultando fonti online dal 2010 in poi (<http://dx.doi.org/10.17632/zygb8jygrw.3>). I modelli sviluppati sono stati calibrati e validati al fine di definire delle combinazioni di variabili di pioggia e del contenuto d'acqua del suolo per la previsione delle condizioni d'innesco dei fenomeni franosi.

1. Introduzione

I modelli di allerta da frana a scala regionale sono tipicamente basati su soglie pluviometriche espresse in termini di pioggia cumulata o intensità media rispetto alla durata dell'evento meteorico, trascurando del tutto le condizioni antecedenti (Segoni et al., 2018). Ne consegue che ad ogni evento previsto di precipitazione di una certa intensità, indipendentemente dallo stato del terreno, il sistema rilascia un determinato livello di allerta, aumentando la probabilità di falsi allarmi. A tal proposito, studi recenti hanno dimostrato che l'affidabilità del modello di allerta potrebbe essere incrementata includendo altre variabili, tra cui il contenuto d'acqua del suolo (Reder e Rianna, 2021).

A tal proposito, in questa nota è proposto l'utilizzo di tecniche di machine learning al fine di selezionare le variabili più significative per l'innesco di fenomeni franosi pluvio-indotti. I dati sulle frane sono stati ricavati da un inventario disponibile online, mentre le variabili atmosferiche sono derivate dalla rianalisi atmosferica ERA5-Land. La procedura è stata testata nella zona di allerta Camp-3 in un arco temporale compreso tra il 2010 e il 2019.

2. Materiali e metodi

2.1 Area di studio e dati utilizzati

L'area di studio è la Camp-3, una delle otto zone d'allerta definite dalla Protezione Civile ai fini della gestione del rischio idrometeorologico in Campania (DPGR 299/2005). In questa zona, di estensione pari a circa 1619 km², sono compresi 109 comuni e i monti Lattari, Picentini e Partenio (Fig.1). Le condizioni orografiche e la vicinanza del mare favoriscono la formazione di tempeste a carattere convettivo (Fortelli et al., 2019; Scafetta e Mazzarella, 2021). La presenza sui versanti di depositi piroclastici di origine vulcanica su substrati carbonatici rende queste aree altamente suscettibili all'innescò di fenomeni franosi a cinematica veloce, tra cui: frane superficiali, debris flow, valanghe di detriti, flussi di detriti e flussi iperconcentrati (Cascini et al., 2008). Nell'area si sono registrati alcuni degli eventi franosi più catastrofici avvenuti in Europa, tra i quali rientrano i tragici eventi avvenuti sul massiccio del Pizzo d'Alvano tra il 4 e il 5 maggio 1998 (Haque et al., 2016).

Le informazioni sui fenomeni franosi occorsi nell'area di studio sono state ricavate da FraneItalia, un catalogo georeferenziato di fenomeni franosi avvenuti in Italia sviluppato consultando fonti online dal 2010 in poi (Calvello e Pecoraro, 2018). Gli eventi franosi sono classificati considerando due categorie di numerosità: eventi franosi singoli (SLE), per record che riportano una singola frana; eventi franosi areali (ALE), per record che si riferiscono a frane multiple provocate da un'unica causa innescante nella stessa Zona di Allerta Meteo. In Camp-3 sono stati registrati 120 eventi pluvio-indotti (72 SLE e 48 ALE) dal 2010 al 2019, la maggior parte dei quali (96 su 120) sono avvenuti tra ottobre e marzo.

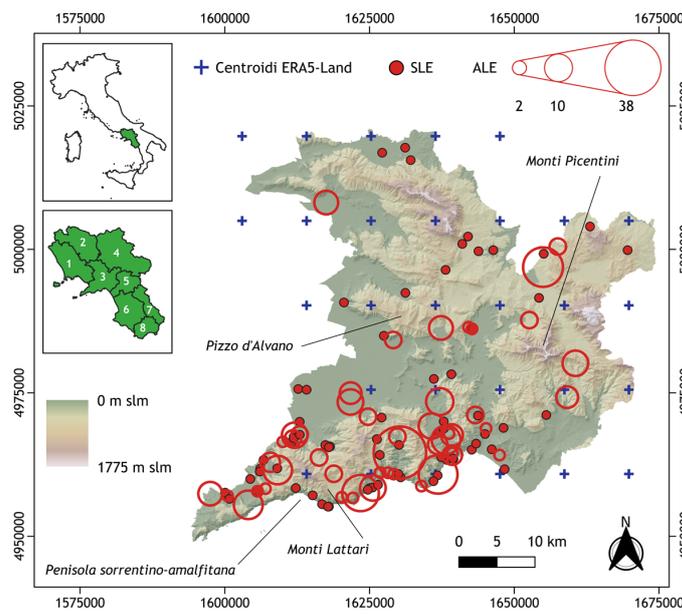


Fig 1. Mappa della zona d'allerta Camp-3 che raffigura i 120 eventi franosi considerati in questo studio e i centroidi delle celle ERA5-Land. Nei riquadri è mostrata la posizione dell'area di studio in Campania e in Italia.

I dati di pioggia e del contenuto d'acqua del suolo sono derivati dalla rianalisi atmosferica ERA5-Land (Munoz-Sabater et al., 2021), sviluppata dallo European Centre for Medium-Range Weather Forecasts (ECMWF). La rianalisi atmosferica fornisce un quadro consistente e completo dell'atmosfera combinando dati osservazionali dei satelliti e dei sensori terrestri con modelli meteorologici fisicamente basati. ERA5-Land rende disponibili circa 50 variabili atmosferiche a una risoluzione spaziale di 9 km e a una risoluzione temporale oraria. Data la rilevanza dei processi al suolo e di un'adeguata parametrizzazione dei processi, ERA5-Land rappresenta, a rigore, un "replaying" della

sola componente di suolo mentre utilizza per le variabili atmosferiche interpolazioni statistiche delle forzanti prodotte dalla rianalisi atmosferica ERA5 (con risoluzione spaziale di circa 31 km).

2.2 Metodologia

Questo studio parte dall'assunzione che a questa scala, ovvero considerando unitariamente l'intera area di studio di quasi 2000 km², gli eventi franosi pluvio-indotti possano essere correlati ad una combinazione di misure legate a due fattori: una condizione predisponente rappresentata dal valore del contenuto d'acqua negli strati superficiali del suolo e una condizione di innesco rappresentata dalle variabili di pioggia (Reder e Rianna, 2021).

I dati orari del dataset ERA5-Land sono stati pre-processati al fine di ottenere 13 variabili di input (*features*), calcolate con una discretizzazione temporale giornaliera coerente con l'informazione contenuta nel catalogo degli eventi franosi utilizzato come variabile dipendente (*outcome*), ovvero:

- massimi giornalieri delle medie geografiche dei valori di pioggia, cumulate su intervalli temporali di 1h, 3h, 6h, 9h, 12h, 18h, 24h, 36h, 48h e 72h (da F1 a F10);
- massimo giornaliero della deviazione standard geografica dei valori orari di pioggia (F11);
- massimo giornaliero della media geografica dei valori orari del contenuto d'acqua del suolo (F12);
- massimo giornaliero della deviazione standard geografica dei valori orari del contenuto d'acqua del suolo (F13).

Per trovare un modello che associ un outcome (frana/non frana, oppure probabilità di frana) ai dati noti di pioggia e contenuto d'acqua del suolo, è stato applicato un modello di machine learning per la classificazione basato sulla logica fuzzy chiamato Likelihood-Fuzzy Analysis (LFA), che consente: i) di descrivere le variabili tramite termini di variabili linguistiche (basso/medio/alto); ii) di associare ad ogni occorrenza di dati di input la probabilità della classe di interesse; iii) di presentare un modello interpretabile. Tale approccio mira ad ottimizzare una misura di performance scelta, tramite ottimizzazione degli insiemi fuzzy che rappresentano i termini linguistici di ogni variabile (Pota et al., 2018), e ottimizzazione del numero di termini per ogni variabile, dell'insieme di variabili costituenti il modello, e delle regole fuzzy (Pota et al., 2017). La misura di performance scelta in questo lavoro per l'ottimizzazione degli insiemi fuzzy è lo Squared Classification Error (SCE):

$$SCE = \frac{1}{KN} \sum_{k=1}^K \sum_{i=1}^N (P(k) - \delta_i^k)^2$$

dove $P(k)$ è la probabilità della k -esima classe calcolata dal modello, e δ_i^k è 1 se il campione i -esimo è associato alla k -esima classe, altrimenti 0. Lo SCE è stato utilizzato anche per la scelta del modello, dando precedenza alle performance piuttosto che all'interpretabilità, in buona parte garantita comunque per costruzione, come mostrato nell'esempio di Fig. 2 e nei risultati.

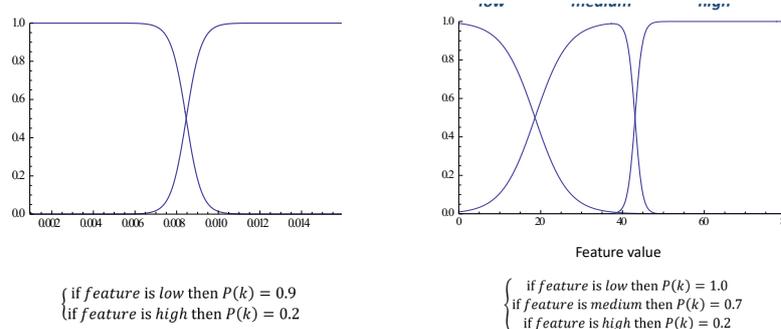


Fig 2. Esempi di logica fuzzy applicata definendo due (a) e tre (b) termini linguistici per una generica variabile (feature) del modello di machine learning.

3. Risultati e discussione

Le 13 variabili indipendenti definite rielaborando i dati orari della precipitazione e del contenuto d'acqua del suolo sono state correlate con la classe positiva dei 120 giorni con eventi franosi avvenuti in Camp-3 e la classe negativa dei rimanenti giorni dal 2010 al 2019, tramite modelli monovariati. In Figura 3 sono riportati i valori della funzione obiettivo (SCE, da minimizzare) per ciascuna variabile. A dispetto di un range di variazione dell'errore piuttosto limitato, emerge un trend monotono decrescente per intervalli delle precipitazioni compresi tra 1h e 18h (durata in corrispondenza della quale si registra l'errore minimo), con una lieve crescita fino a 72h e valori decisamente più elevati per la deviazione standard delle precipitazioni e per il contenuto d'acqua.

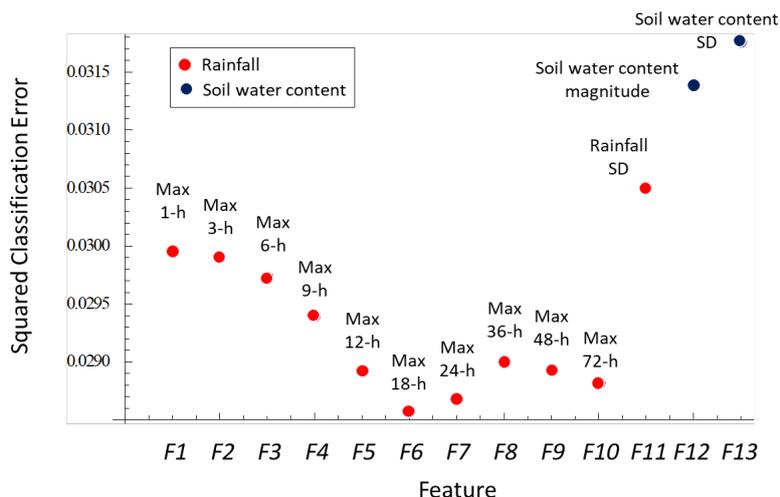


Fig 3. Andamento della funzione obiettivo per modelli definiti utilizzando, singolarmente, le 13 variabili calcolate con discretizzazione temporale giornaliera.

Tra i modelli monovariati, quello con la migliore capacità predittiva è dettagliato in Fig. 4. Dividendo la precipitazione cumulata a 18h in tre insiemi fuzzy (basso/medio/alto), si può notare che la massima probabilità di frana (37%) si ottiene per valori alti (superiori a circa 40 mm). È opportuno ricordare che tale valore si riferisce ad una precipitazione mediata sull'intera area di studio. Il grafico inoltre mostra un range intermedio in cui la probabilità di frana si attesta intorno al 23%, e un range di valori bassi della precipitazione in cui la probabilità di frana è minima (1%).

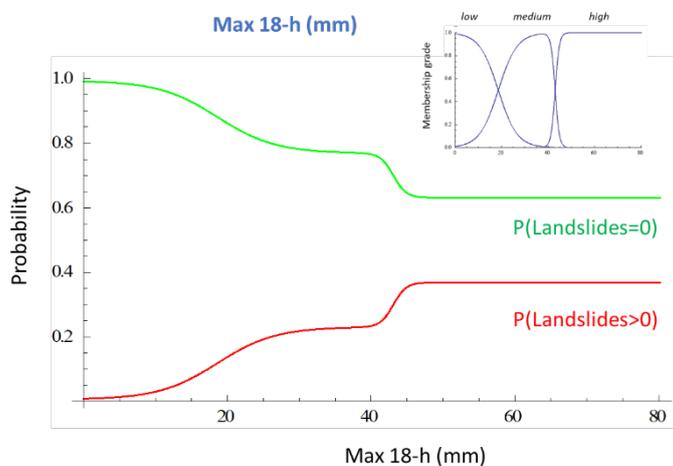


Fig 4. Probabilità di frana e non frana associate al modello basato sulla precipitazione cumulata a 18h.

Inoltre, è stato individuato un modello multivariato che considera la precipitazione cumulata a 18h e la deviazione standard del contenuto d'acqua (Fig. 5). Combinando le due variabili, si ottiene un modello a cui è associata una probabilità nulla di frane per valori bassi di entrambe le variabili indipendenti e soprattutto permette di individuare degli intervalli in cui la probabilità di frana sale al 46%. In definitiva, i risultati permettono di evidenziare l'apporto informativo del contenuto d'acqua.

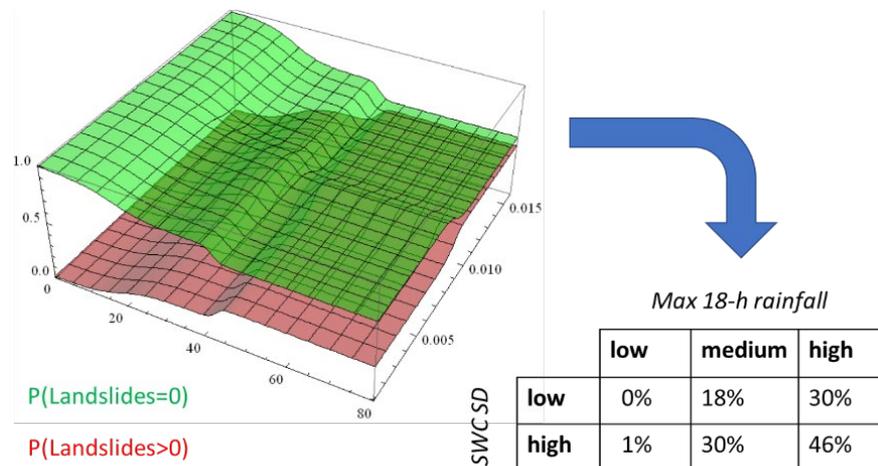


Fig 5. Probabilità di frana e non frana associate al modello che combina la precipitazione cumulata a 18h e la deviazione standard del contenuto d'acqua.

Infine, sono state confrontate le prestazioni di diversi modelli di machine learning, calcolate attraverso una cross-validazione a 10 fold. Gli indicatori di prestazione riportati in Tabella 1 rivelano che i modelli caratterizzati dalle prestazioni migliori sono quelli ottenuti tramite LFA e il metodo Naïve Bayes, in quanto entrambi raggiungono un F1 pari a 0.27. Tuttavia, il primo modello, a differenza del modello bayesiano, comprende solo le variabili più significative, ed è inoltre facilmente interpretabile, come visto sopra.

Tab 1. Indicatori di prestazione associati ai modelli di machine learning testati (i modelli ottenuti tramite selezione di variabili sono indicati con l'asterisco).

Modello	sensitività	precisione	F1
LFA *	0.54	0.18	0.27
Naïve Bayes	0.60	0.17	0.27
Bayes Network	0.72	0.15	0.24
Logistic	0.06	0.30	0.10
MLP	0.09	0.30	0.14
SVM	0.00	0.00	0.00
KNN	0.19	0.19	0.19
RIPPER *	0.06	0.20	0.10
C4.5 *	0.02	0.37	0.05
Random Forest	0.04	0.33	0.07

4. Conclusioni

L'utilizzo di tecniche di machine learning per la definizione di modelli che combinano, a scala regionale, variabili di pioggia e del contenuto d'acqua del suolo per la previsione di fenomeni franosi meteo-indotti è stato testato, in questo studio di carattere preliminare, nella zona di allerta Camp-3 in un intervallo temporale compreso tra il 2010 e il 2019. I modelli sviluppati hanno permesso di individuare alcune variabili significativamente correlate con gli eventi franosi considerati e hanno consentito di calcolare delle probabilità di avvenimento degli eventi piuttosto che delle semplici relazioni dicotomiche. Alcuni possibili sviluppi futuri dello studio potrebbero essere: confronti con altri modelli di allerta (ad es., modello utilizzato dalla regione Campania, soglie probabilistiche di pioggia utilizzate in letteratura); sviluppo di analisi prendendo in considerazione solo gli eventi areali più numerosi; calibrazione e validazione di modelli che impiegano altre variabili potenzialmente rilevanti.

Bibliografia

- Calvello M., Pecoraro G. (2018). FraneItalia: a catalog of recent Italian landslides. *Geoenvironmental Disasters*, 5(13).
- Cascini L., Ferlisi S., Vitolo E. (2008). Individual and societal risk owing to landslides in the Campania region (southern Italy). *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 2:3, 125-140.
- DPGR n. 299 30/06/2005, "Decreto del Presidente della Giunta Regionale della Campania: Il Sistema di Allertamento Regionale per il rischio Idrogeologico e Idraulico ai fini di protezione civile", Bollettino ufficiale della regione Campania, n. speciale 01/08/2005.
- Fortelli A., Scafetta N., Mazzarella A. (2019). Nowcasting and real-time monitoring of heavy rainfall events inducing flash-floods: an application to Phlegraean area (Central-Southern Italy). *Nat Hazards* 97, 861–889.
- Haque U., Blum P., da Silva P.F., Andersen P., Pilz J., Chalov S.R., Malet J.P., Auflič M.U., Andres N., Poyiadji, E., Lamas P.C., Zhang W., Peshevski I., Pétursson H.G., Kurt T., Dobrev N., García-Davalillo J.C., Halkia M., Ferri S., Gaprindashvili G., Engström J., Keellings D. (2016). Fatal landslides in Europe. *Landslides*, 13, 1545–1554.
- Muñoz-Sabater J., Dutra E., Agustí-Panareda A., Albergel C., Arduini G., Balsamo G., Boussetta S., Choulga M., Harrigan S., Hersbach H., Martens B., Miralles D.G., Piles M., Rodríguez-Fernández N.J., Zsoter E., Buontempo C., Thépaut J.-N. (2021). ERA5-Land: a state-of-the-art global reanalysis dataset for land applications. *Earth Syst. Sci. Data*, 13, 4349–4383.
- Pota M., Esposito M., De Pietro G. (2017). Designing rule-based fuzzy systems for classification in medicine. *Knowledge-Based Systems*, 124, 105-132.
- Pota M., Esposito M., De Pietro G. (2018). Likelihood-fuzzy analysis: From data, through statistics, to interpretable fuzzy classifiers. *International Journal of Approximate Reasoning*, 93, 88-102.
- Reder A., Rianna G. (2021). Exploring ERA5 reanalysis potentialities for supporting landslide investigations: a test case from Campania Region (Southern Italy). *Landslides*, 18, 1909–1924.
- Scafetta N., Mazzarella A. (2021). On the Rainfall Triggering of Phlegraean Fields Volcanic Tremors. *Water*, 13:154.
- Segoni S., Piciullo L., Gariano S.L. (2018). A review of the recent literature on rainfall thresholds for landslide occurrence. *Landslides*, 15:1483–1501.