Valutazione del danno alluvionale mediante strumenti di Statistica Multivariata e Machine Learning

Simone Montanari ([a]) ([b]), Christian Natale Gencarelli ([c]), Simone Sterlacchini ([c]) & Maurizio Vichi ([d])

([a]) Tesista triennale Dipartimento di Scienze Statistiche, Sapienza Università di Roma. E-mail: simonemontanar@gmail.com
([b]) Volontario per Information Management Team, Croce Rossa Italiana
([c]) Istituto di Geologia Ambientale e Geoingegneria del Consiglio Nazionale delle Ricerche (IGAG – CNR) Milano
([d]) Dipartimento di Scienze Statistiche, Sapienza Università di Roma

ABSTRACT

Importante rilievo nella preparazione e risposta a eventi alluvionali assume lo studio delle caratteristiche e della stima dei danni provocati.

Questo studio riporta i risultati preliminari dell'applicazione di metodi e modelli di statistica multivariata e machine learning.

Le informazioni raccolte a seguito dell'alluvione avvenuta nel gennaio 2014 nei comuni di Bastiglia e Bomporto sono state integrate con dati provenienti da modelli idraulici.

Tramite algoritmi di regressione si sono ottenute delle stime sui danni subiti dagli edifici privati e sono state analizzate le variazioni del danno in relazione a specifiche variabili.

Inoltre, con tecniche statistiche, ci si è concentrati sulle caratteristiche che incidono maggiormente nella clusterizzazione degli edifici colpiti.

I risultati ottenuti sostengono un maggior utilizzo di queste tecniche applicate in questi particolari contesti.

KEY WORDS: analisi multivariata, clustering, danni alluvionali, random forest, xgboost.

INTRODUZIONE

Secondo l'ultimo report pubblicato dal Centre for Research on the Epidemiology of Disasters, tra il 2001 e il 2020 sono stati registrati – in Europa – 999 disastri naturali, di cui 951 legati alla condizione climatica (di natura – quindi – meteorologica, idrologica e climatologica).

Questi eventi hanno portato circa 150.000 decessi e oltre 11 Milioni di persone danneggiate. I danni stimati ammontano a 217 Miliardi di Dollari.

Le alluvioni sono stati gli eventi più frequenti (41%), seguite da tempeste (27%), temperature estreme (23%) ed eventi secondari, quali incendi, frane, siccità (9%).

Relativamente alle persone danneggiate, più di 6.6 Milioni hanno subito danni a causa di eventi alluvionali che – sul piano economico – hanno contribuito al 50% delle perdite totali da eventi calamitosi di tipo climatico.

Secondo il rapporto 2020 di Legambiente, in Italia dal 2010 al 2020 sono stati registrati 946 fenomeni meteorologici estremi su 507 diversi Comuni, di cui 416 sono i casi di allagamento da piogge intense mentre 118 gli eventi causati da esondazioni fluviali.

È ormai da ritenersi consolidata la tendenza a una maggiore frequenza e intensità dei fenomeni meteorologici estremi, soprattutto in Italia in quanto al centro del Mar Mediterraneo, considerate uno degli "*hot spot*" del cambiamento climatico.

Secondo vari studi (Sofia Darmaraki et al. (2019) a esempio) l'area del Mediterraneo sarà infatti una delle più sensibili alle conseguenze del cambiamento climatico e vedrà, negli anni a venire, rapide e ingenti precipitazioni alternate a ondate di calore.

Questi rapporti hanno il fine di chiarire cause, caratteristiche e conseguenze degli eventi naturali che si susseguono nel tempo cercando di portare l'attenzione delle Amministrazioni a un livello di prevenzione piuttosto che di gestione dell'emergenza.

Il presente lavoro può considerarsi parte di questo obiettivo in quanto si sono definiti metodi per una più completa valutazione del danno provocato da un'alluvione in un contesto urbano.

DATI E TECNICHE

1. DESCRIZIONE DATASET

Questo studio prende in esame l'alluvione avvenuta nel gennaio 2014 nella provincia di Modena.

Il 19 gennaio 2014, a seguito delle ingenti piogge dei giorni precedenti, la rottura di un argine del fiume Secchia all'altezza della frazione di San Matteo causò l'alluvione di un'ampia zona di circa 52 km^2 comprendente i comuni e le frazioni di Bastiglia, Bomporto, Staggia, Villavara.

Secondo le analisi di Orlandini et al. (2015) queste zone sono state invase in meno di 30 ore da 37 milioni di m^3 d'acqua e sono rimaste inondate per circa 48 ore.

I danni stimati ammontano a 500 Milioni di Euro, di cui 16 solo per i danni a edifici privati.

A seguito dell'evento, ai cittadini colpiti sono state fatte compilare delle schede di danno in cui erano richieste

informazioni riguardo:

- Ubicazione dell'edificio danneggiato
- Caratteristiche dell'edificio danneggiato
- Danni in Euro registrati su beni mobili e immobili

Scopo di queste schede, il rimborso dei danni tramite fondi statali.

Questi dati sono stati integrati con:

- Informazioni geomorfologiche dell'area, ottenute da modelli idraulici
- Informazioni OMI (Osservatorio del Mercato Immobiliare), ottenute dall'Agenzia delle Entrate

In totale sono stati collezionati dati per 1918 abitazioni private. Di queste ne sono state analizzate 1366 a causa dei valori mancanti presenti per alcune caratteristiche (per la natura dei dati e l'obiettivo dello studio, si è deciso di non ricorrere ad alcuna interpolazione per ricavare i dati mancanti).

2. METODOLOGIA

Il lavoro si è focalizzato sulla regression lineare e sulla clusterizzazione dei dati che saranno presentate nelle successive Sezioni 2.1 e 2.2.

Inizialmente è stata condotta un'analisi dei valori anomali per individuare eventuali schede in contrasto con i valori medi complessivi.

Utilizzando la distanza di Mahalanobis e come soglia critica il quantile $\chi^2_{0.99}$, sono stati trovati 288 outliers.

Tutte le analisi successive sono state effettuate con e senza valori anomali per vederne le differenze.

Non sono state effettuate ulteriori analisi su questi valori in quanto al di fuori dell'obiettivo del presente lavoro.

2.1 Regressione non lineare

A partire dal lavoro di White del 1945, le analisi sulle perdite dovute alle alluvioni prendono in considerazione l'altezza dell'acqua come elemento principale da studiare e su cui, quindi, effettuare regressioni, trascurando altri fattori che potrebbero contribuire alla realizzazione dei danni.

Oggi, grazie all'avanzamento tecnologico, è possibile sfruttare tecniche più complesse rispetto a una regressione lineare, includendo anche molteplici variabili.

Esistono già alcuni studi che applicano metodologie simili ma è stato mostrato come i modelli sviluppati all'estero non si adattino al territorio italiano e quelli applicati in alcune zone d'Italia non restituiscano risultati soddisfacenti in altre aree (Amadio et al. (2016), Molinari et al. (2012)). I modelli pertanto soffrono di sitospecificità e solamente un'eventuale costruzione di un ricco database nazionale può permettere uno studio con risultati soddisfacenti su buona parte del territorio italiano.

In questo caso ci si è concentrati su due modelli di regressione non lineare basata su alberi decisionali: *random forest* e *xgboost*.

Gli *alberi decisionali* (*decision trees*) sono gli elementi alla base di un algoritmo random forest. Un albero decisionale è costituito da tre elementi: una radice, dei nodi e delle foglie.

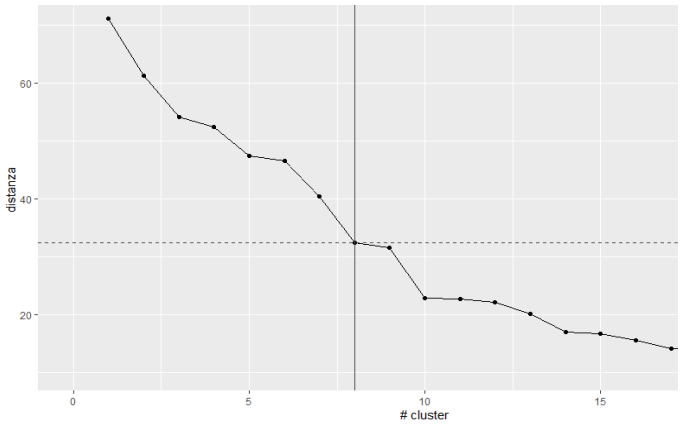

Fig. 1 – Screeplot

Un algoritmo decision trees divide il dataset di training in nodi, divisi a loro volta in altri nodi collegati ai precedenti tramite rami. Questo iterativamente fino ad arrivare a una foglia, ossia un nodo che non può essere ulteriormente diviso. Le foglie vengono utilizzate per predire il risultato tramite delle apposite metriche.

Il random forest è costituito da molteplici alberi decisionali (tecnica *ensemble*). Tramite *bagging* e una scelta casuale delle variabili "radici", viene presa la previsione di ogni albero della foresta e restituito un unico output finale, ottenendo una maggior accuratezza rispetto all'utilizzo del singolo albero e limitando il problema dell'overfitting.

Anche il modello xgboost (*Extreme Gradient Boosting Algorithm*) è un metodo ensemble ma utilizza la discesa del gradiente come algoritmo di ottimizzazione. L'idea principale è di correggere gli errori fatti precedentemente dal modello, quindi imparare e migliorare i risultati col passare delle iterazioni. Questo fino a trovare il minimo della funzione di perdita utilizzata.

Le analisi sono state quindi eseguite con entrambi i modelli, cercando di compararne i risultati tramite la lettura di *variable importance*, *trees rules* e *partial dependence plot*.

2.2 Cluster

Per scovare la presenza di pattern particolari che legano i dati a disposizione, si è deciso di applicare algoritmi di clustering.

L'analisi dei cluster consiste nel raggruppare le osservazioni in insiemi con l'obiettivo di avere dati simili all'interno dello stesso gruppo ma dissimili dalle unità appartenenti agli altri gruppi.

Esistono diverse tipologie di clustering e svariate metriche da poter utilizzare per raggiungere questo obiettivo, e la scelta è data dal tipo di dati a disposizione e dallo scopo di questa analisi.

Volendo semplicemente evidenziare l'esistenza di eventuali legami tra gruppi di dati senza assumere vincoli di alcun tipo, si è optato per un clustering gerarchico di tipo agglomerativo.

Ogni osservazione crea un proprio cluster e questi vengono man mano accoppiati, in base alla metrica scelta, fino alla creazione di un unico gruppo.

	Random Forest	XGboost
Train set	0.78	0.75
Test set	0.64	0.67

Tabella 1 – Valori di r^2 per train e test set nei rispettivi modelli

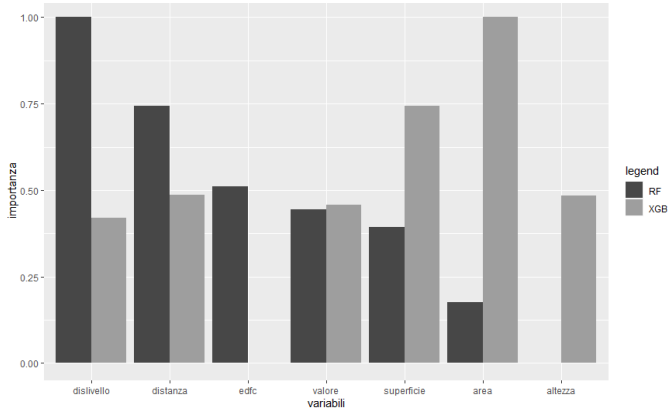

Fig. 2 – Variable importance per random forest e xgboost

Il numero di cluster da utilizzare viene scelto osservando lo screeplot riportato in Fig. 1. Sull'asse *y* si trova la distanza a cui sono legati i gruppi, sull'asse *x* è riportato il numero di cluster. Per scegliere quest-ultimo si vede il punto in cui è presente un gomito nella spezzata che si è formata, cercando di ottenere più cluster possibili senza complicarne la lettura.

3. RISULTATI

3.1 Statistica Descrittiva

Le variabili sono state analizzate da un punto di vista uni- e multi-variato, cercando forme particolari nelle distribuzioni o relazioni statisticamente rilevanti tra le *features*.

Per quanto riguarda la distribuzione, non si osservano caratteristiche particolari se non in due variabili (*altezza dell'acqua* e *distanza dal fiume*) per cui si nota una possibile mistura, cosa non interessante per questo lavoro.

Le correlazioni sono basse, anche per le variabili per cui si può presupporre un forte legame, come *altezza dell'acqua* e *danno subito* ($r^2 = -0.01$). Sono stati registrati valori compresi tra 0.6 e 0.8 per le coppie *superficie edificio – valore edificio* e *distanza dal fiume – altezza dell'acqua*.

3.2 Regressione non lineare

Il dataset originale è stato suddiviso in train set e test set. Gli algoritmi random forest e xgboost sono stati eseguiti sul train set e il risultato è stato applicato sul test set in modo da validarne l'accuratezza per nuovi dati.

Entrambi i modelli hanno subito il *tuning* oltre all'applicazione della *cross validation*, così da cercare di rendere il modello più robusto rispetto l'overfitting.

Sono stati testati diversi modelli cambiando i vari parametri.

I risultati migliori sono stati ottenuti ponendo l'85% dei dati nel train set e il rimanente 15% nel test set. I valori anomali individuati nella Sezione 2 non sono stati esclusi.

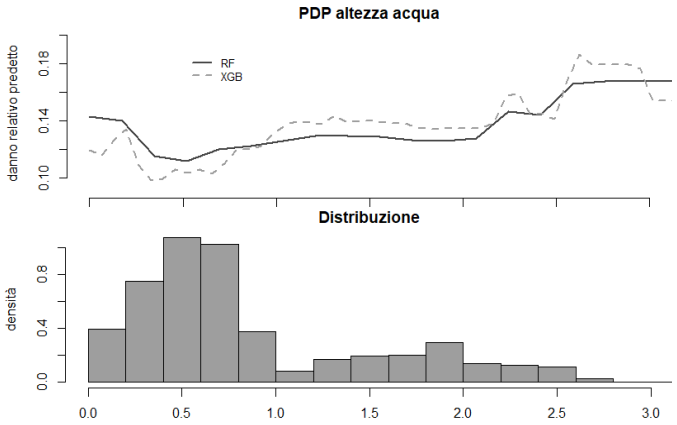

Fig. 3 – Partial dependence plot del danno relativo rispetto l'altezza dell'acqua

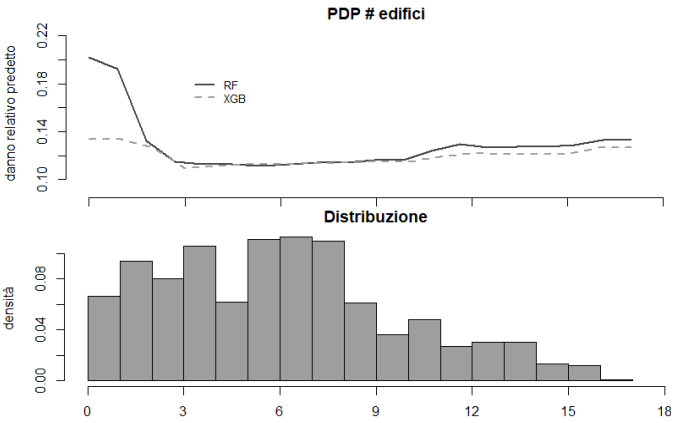

Fig. 4 – Partial dependence plot del danno relativo rispetto il numero di edifici

Le performance in termini di r^2 sono riportate in Tabella 1.

Nonostate i risultati ottenuti siano simili, c'è una notevole differenza guardando alla *variable importance* (Fig. 2), ossia all'impatto che assume ogni variabile sull'output del modello. Queste differenze derivano dalla diversa natura degli algoritmi e da questo grafico si nota il diverso utilizzo delle variabili.

Nonostante ciò, la figura resta importante per eventuali futuri approfondimenti.

I risultati ottenuti sono stati infine analizzati tramite i *partial dependence plot* (*pdp*), nei quali viene mostrato l'effetto marginale di una variabile sul risultato previsionale medio del modello di machine learning selezionato.

Interessante è la rappresentazione della variazione media dell'output al variare delle features *altezza dell'acqua* e *numero di edifici attorno a quello colpito*. Il grafico della distribuzione di densità della variabile aiuta nell'interpretabilità del *pdp*.

Per i dati collezionati si evidenzia come, in media, il danno tenda a crescere fino a un livello dell'acqua pari a 1 metro, andando poi incontro a una certa stabilità per poi riprendere la crescita nel range 2.0 – 2.5 metri (Fig. 3).

Per quanto riguarda gli edifici, si osserva come questi impattino positivamente nella riduzione del danno (Fig. 4) in quanto questo diminuisce all'aumentare del numero di edifici presenti in un raggio di 100 metri attorno a quello colpito.

Variabile	Varianza nei cluster (WSS)	Varianza tra cluster (GSS)	Pseudo F
Distanza fiume	286.00	1079.00	731.00
Dislivello	293.00	1072.00	711.00
Altezza acqua	548.00	817.00	289.00
Valore immobile	726.00	639.00	171.00
Superficie	822.00	543.00	128.00
Area	964.00	401.00	81.00
Edifici	1034.00	331.00	62.00
Danno relativo	1131.00	234.00	40.00

Tabella 2 – Valori della pseudo-f

3.3 Cluster

La cluster analysis è stata effettuata utilizzando un algoritmo gerarchico agglomerativo.

Come metrica di distanza si è fatto ricorso al metodo di Ward che si basa sulla minimizzazione della varianza intra-cluster, concentrandosi quindi sul rendere i gruppi il più omogenei possibili.

Analizzando lo scatterplot in Fig. 1 si è optato per un numero di cluster pari a 8 e si sono potute effettuare analisi interne ai gruppi, come il calcolo della silhouette e la pseudo-f sulle singole variabili.

Corretti i cluster grazie alla lettura della silhouette, dai risultati trovati con la pseudo-f si è notato come le variabili che hanno impattato maggiormente nella classificazione siano quelle relative alla posizione dei singoli edifici (distanza dal fiume, dislivello rispetto gli argini) piuttosto che le features urbanistiche (Tabella 2).

Questo dato risulta molto interessante anche nell'ottica di futuri approfondimenti.

4. CONCLUSIONI

Lo studio effettuato riprende e amplia principalmente quanto osservato da Carisi et al. nel 2018.

Come già sottolineato si tratta di analisi preliminari.

I modelli ad albero hanno restituito buone performance e la pseudo-f calcolata sulle variabili ha offerto spunti interessanti.

Quanto fatto dimostra come sia importante l'applicazione di nuove tecniche in questo campo, che risulta essere sempre più importante all'interno del più vasto ambito della prevenzione verso questo genere di fenomeni naturali.

Importante diventa anche il ruolo della raccolta dei dati in quanto il modo in cui ad oggi si reperiscono introduce fattori di errore almeno parzialmente evitabili.

Infine, ottenendo buoni risultati con questi modelli, assumono maggiore interesse eventuali studi condotti tramite l'utilizzo di algoritmi più avanzati e appartenenti al deep learning.

REFERENCES

Amadio, M., Mysiak, J., Carrera, L. et al. – Improving flood damage assessment models in Italy. Nat Hazards 82, 2075–2088 (2016). https://doi.org/10.1007/s11069-016-2286-0

Breiman, L. – Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Carisi, F., Schröter, K., Domeneghetti, A., Kreibich, H., and Castellarin, A. – Development and assessment of uni- and multivariable flood loss models for Emilia-Romagna (Italy), Nat. Hazards Earth Syst. Sci., 18, 2057–2079, https://doi.org/10.5194/nhess-18-2057-2018, 2018.

Centre for Research on the Epidemiology of Disasters – CRED Crunch 64 - Extreme weather events in Europe (2021)

Chinh, D.T.; Gain, A.K.; Dung, N.V.; Haase, D.; Kreibich, H. – Multi-Variate Analyses of Flood Loss in Can Tho City, Mekong Delta. Water 2016, 8, 6. https://doi.org/10.3390/w8010006

D'Alpaos, L., Brath, A., Fioravante, V., Gottardi, G., Mignosa, P., and Orlandini, S. – Relazione tecnico-scientifica sulle cause del collasso dell' argine del fiume Secchia avvenuto il giorno 19 gennaio 2014 presso la frazione San Matteo, Tech. rep., Bologna, Italy, 2014

Friedman, J. H. (2001) – Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5), 1189–1232. http://www.jstor.org/stable/2699986

Joe H. Ward Jr. (1963) – Hierarchical Grouping to Optimize an Objective Function, Journal of the American Statistical Association, 58:301, 236-244, DOI: 10.1080/01621459.1963.10500845

Legambiente – Rapporto 2020 dell'Osservatorio di Legambiente Cittàclima

Merz, B., Kreibich, H., and Lall, U.: Multi-variate flood damage assessment: a tree-based data-mining approach, Nat. Hazards Earth Syst. Sci., 13, 53–64, https://doi.org/10.5194/nhess-13-53-2013, 2013.

Molinari, D., Aronica, G., Ballio, F., Berni, N., and Pandolfo, C. – Le curve di danno quale strumento a supporto della direttiva alluvioni: criticità dei dati italiani, in: XXXIII Convegno Nazionale di Idraulica e Costruzioni Idrauliche – Brescia, 10–15 settembre 2012, Brescia, Italy, 2012

Morde, V. – XGBoost Algorithm: Long May She Reign!. https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

Nielsen, Frank. (2016) – Hierarchical Clustering. 10.1007/978-3-319-21903-5_8.

Nikulski, J. – The Ultimate Guide to AdaBoost, random forests and XGBoost. https://towardsdatascience.com/the-ultimate-guide-to-adaboost-random-forests-and-xgboost-7f9327061c4f

Orlandini, S., Moretti, G., and Albertson, J. D. – Evidence of an emerging levee failure mechanism causing disastrous floods in Italy, Water Resour. Res., 51, 7995–8011, https://doi.org/10.1002/2015WR017426, 2015.

White, G. – Human adjustment to floods, Department of Geography – University of Chicago, Chicago, USA, 1945