Progetto II

Simone Manti, Matricola: 566908

9/12/2021

Introduzione

La seguente analisi è rivolta ad enti idrogeologici dell'Emilia Romagna, precisamente della provincia di Bologna. Il nostro obiettivo è aiutare questi enti nella costruzione di strutture che possano evitare pericoli causati da frane e fenomeni idraulici: l'idea è quella di dividere in alcuni gruppi i comuni di Bologna, in modo da capire dove e cosa va costruito in ogni gruppo, in base alle proprie esigenze.

Presentazione del problema

I dati esaminati risalgono al 2017 e provengono da una tabella fornita dall'Istat (dettagli in fondo alla relazione).

I parametri presi in considerazione sono i seguenti:

F1 Area pericolosità idraulica bassa

F2 Area pericolosità idraulica media

F3 Area pericolosità idraulica elevata

F4 Area pericolosità da frana pai[1] moderata

F5 Area pericolosità da frana pai media

F6 Area pericolosità da frana pai elevata

F7 Area pericolosità da frana pai molto elevata

Tutti i fattori sono stati misurati in kmq. Per maggiore chiarezza, mostriamo il summary della tabella:

	F1	F2	F3	F4	F5	F6	F7
Min.	0.10000	0.10000	0.040	0.000000	0.0000000	0.000000	0.0000000
1st Qu.	3.54000	3.54000	1.375	0.000000	0.0000000	0.000000	0.0000000
Median	21.28000	21.45000	2.950	0.000000	0.0000000	2.360000	0.0000000
Mean	33.60891	33.69164	9.032	1.373273	0.3372727	8.256182	0.6605455
3rd Qu.	44.27500	44.27500	9.530	1.950000	0.5150000	14.910000	0.8100000
Max.	159.11000	159.11000	50.200	13.750000	2.8800000	39.990000	4.6700000

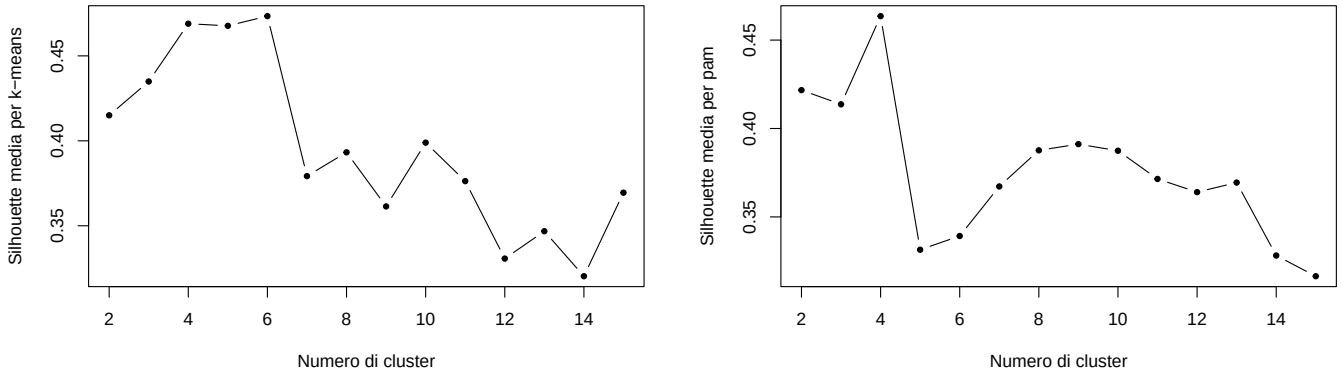
Clustering

La nostra analisi procederà come segue:

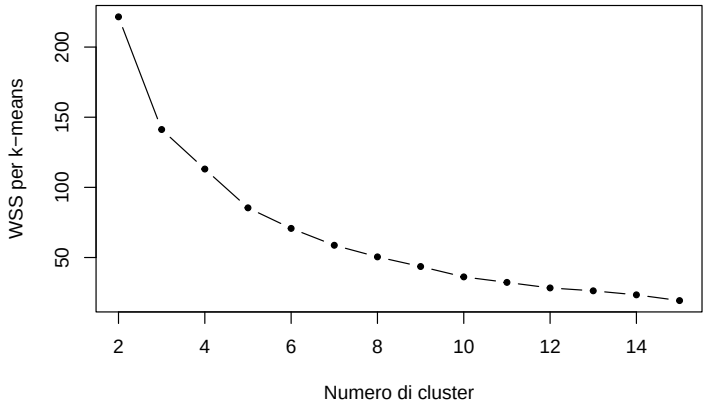
[1] Piani di assetto idrogeologico

1. Considereremo e confronteremo due metodi a punti prototipo: k-means e pam (con la distanza euclidea).
2. Considereremo e confronteremo tre metodi gerarchici, i quali differiscono solo dal tipo di distanza tra cluster: l'average linkage method, il single linkage method e il complete linkage method (con la distanza euclidea tra punti).
3. Sceglieremo il migliore tra i migliori metodi di 1 e 2. Una volta scelto il metodo che si comporta meglio ci concentreremo su quello, cercando di interpretare la suddivisione trovata.
4. Concluderemo riassumendo i risultati principali ottenuti.

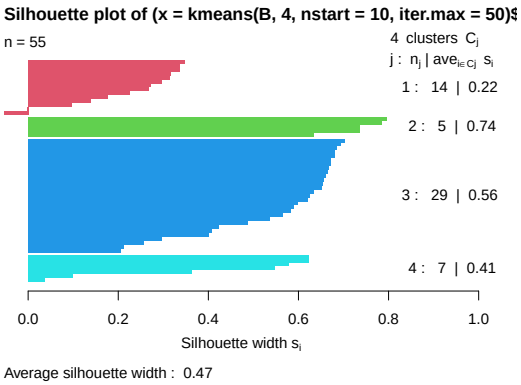
Dopo aver opportunamente standardizzato i dati, cominciamo la nostra analisi studiando il comportamento della silhouette media per i metodi k-means e pam.



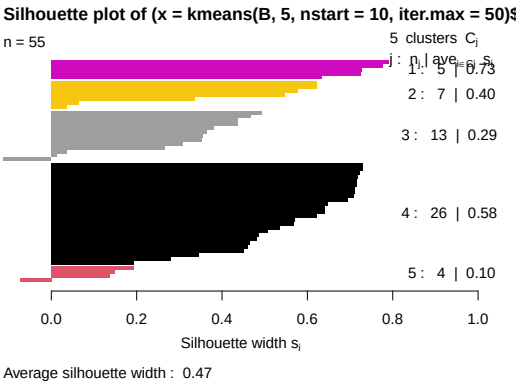
Dai grafici notiamo che i valori rilevanti della silhouette media per k-means si ottengono in 4,5,6,7 mentre per pam si ottengono in 3 e 4. Per rafforzare la nostra analisi per quanto riguarda k-means, studiamo anche il grafico relativo alla WSS, per vedere se esiste k per cui passando da k a k+1 la WSS non crolla in modo sostanziale (ovvero è presente un "gomito").



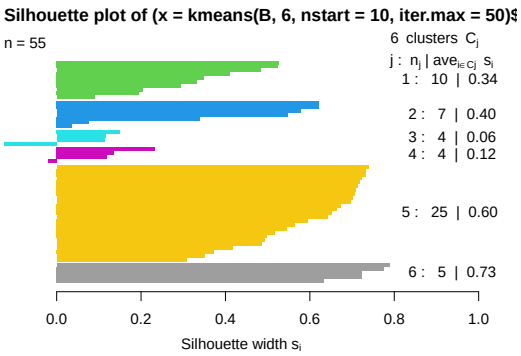
Purtroppo il grafico precedente non ci dà informazioni aggiuntive, riportiamo allora le silhouette per i valori elencati prima sia per k-means che per pam.



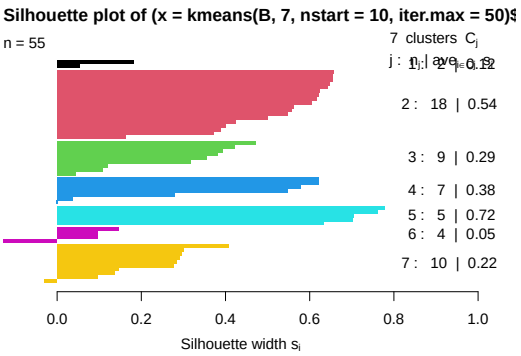
Average silhouette width : 0.47



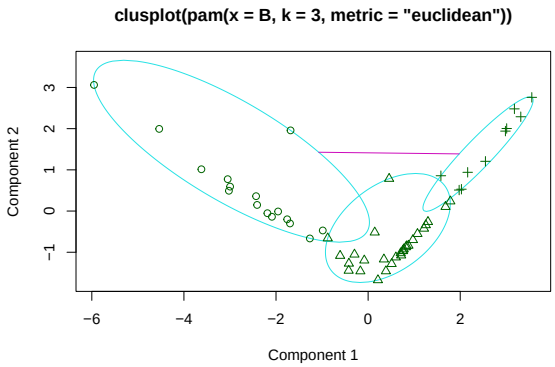
Average silhouette width : 0.47



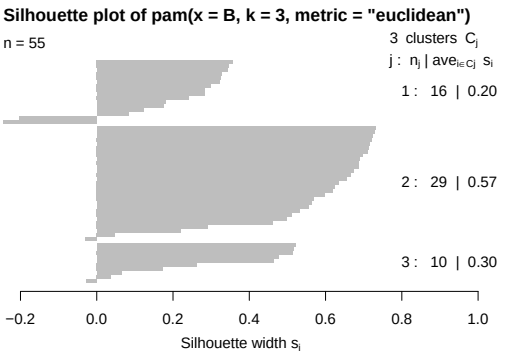
Average silhouette width : 0.46



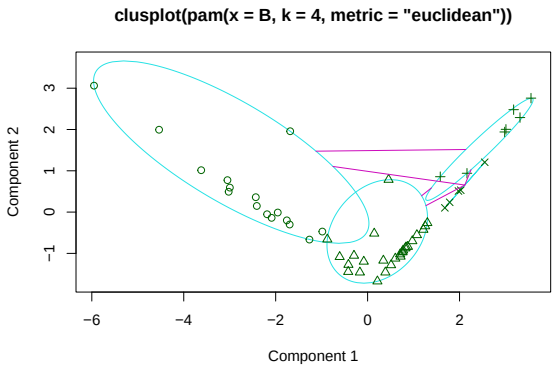
Average silhouette width : 0.38



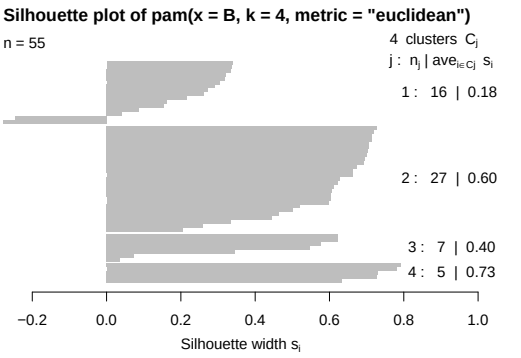
These two components explain 79.4 % of the point variability.



Average silhouette width : 0.41



These two components explain 79.4 % of the point variability.

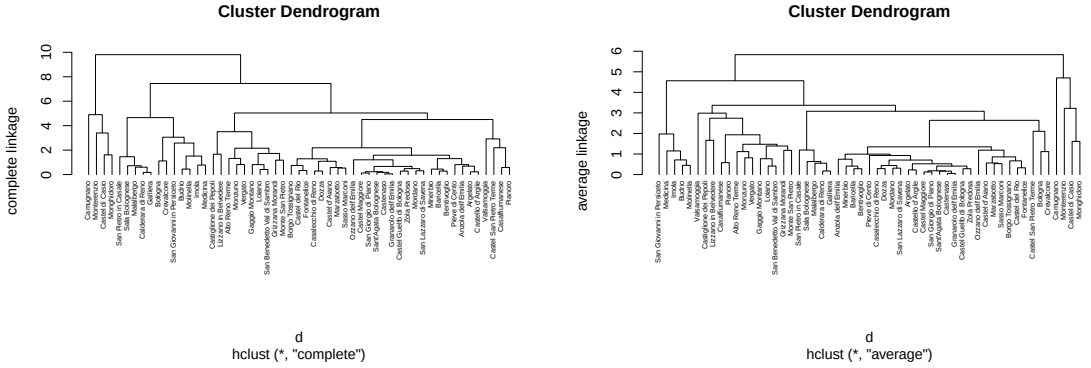


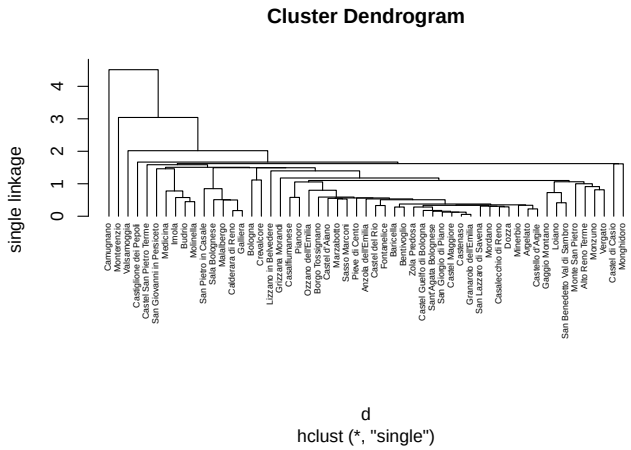
Average silhouette width : 0.46

Scegliamo di utilizzare k-means con k = 4: infatti in questo caso si ha una buona silhouette media (0.47)

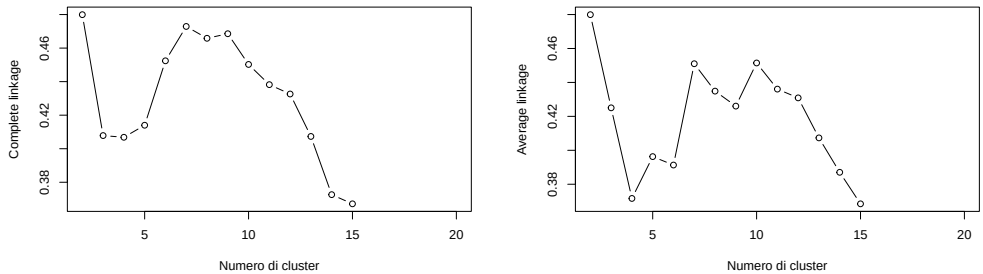
e si ha silhouette negativa solo per due dati. Per la precisione, dai grafici delle silhouette si restringe la ricerca del parametro ottimale ai valori 4 e 5 (perchè per 6 e 7 sono presenti valori più disomogenei delle silhouette) e si preferisce il primo in quanto a parità di silhouette media "sbagliamo di meno" (i.e. i valori con silhouette negativa sono più grandi) e si hanno valori più omogenei della silhouette. Osserviamo anche che le suddivisioni ottenute da k-means con k = 4 e pam con k=4 (k=3 si esclude perchè abbiamo più dati con silhouette negativa) sono molto simili: questo conferma la bontà dell'analisi. Tra i due scegliamo il metodo k-means.

Consideriamo adesso i tre metodi gerarchici già introdotti prima. Per prima cosa, rappresentiamo i tre dendogrammi.





Già dai dendogrammi riusciamo a escludere il single linkage: infatti, si osserva immediatamente che potando questo albero nelle fasi iniziali si ottengono cluster composti da un singolo elemento. Questo è sintomo di una scelta sbagliata della distanza tra cluster. Confrontiamo invece le silhouette medie degli altri due metodi gerarchici.



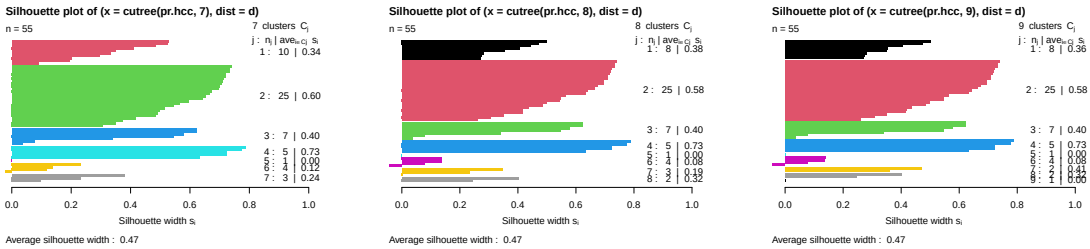
Da questi grafici emerge che i valori rilevanti della silhouette media per complete linkage si ottengono in 6,7,8,9,10 mentre per average linkage si ottengono in 7,8,9,10,11. Rappresentiamo in due tabelle le numerosità dei cluster per i casi sopraelencati.

1	2	3	4	5	6	7	8	9	10
10	29	7	5	1	3	NA	NA	NA	NA
10	25	7	5	1	4	3	NA	NA	NA
8	25	7	5	1	4	3	2	NA	NA
8	25	7	5	1	4	2	2	1	NA
8	25	2	5	5	1	4	2	2	1

1	2	3	4	5	6	7	8	9	10	11
13	28	5	5	1	2	1	NA	NA	NA	NA
12	28	5	5	1	2	1	1	NA	NA	NA
10	28	5	5	1	2	2	1	1	NA	NA
10	25	3	5	5	1	2	2	1	1	NA
10	25	2	5	5	1	2	1	2	1	1

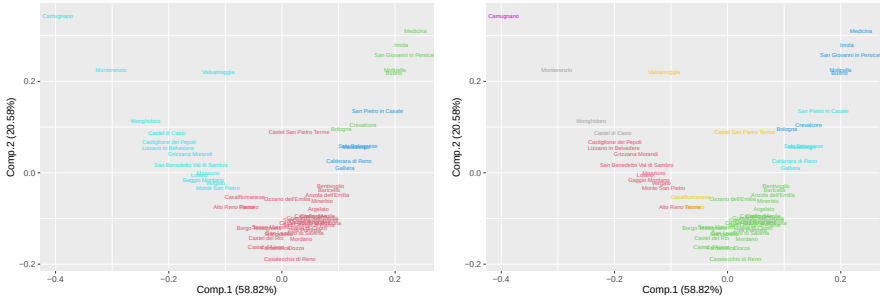
Notiamo che nell'average linkage si ha lo stesso problema del single linkage già discusso in precedenza, ovvero si hanno numerosi cluster fatti da un singolo elemento e questo è sintomo di una scelta sbagliata della distanza tra cluster. Dunque, tra i metodi gerarchici il migliore è il complete linkage.

Studiamo anche in questo caso le silhouette per i valori con silhouette media più alta, ovvero per 7,8,9.



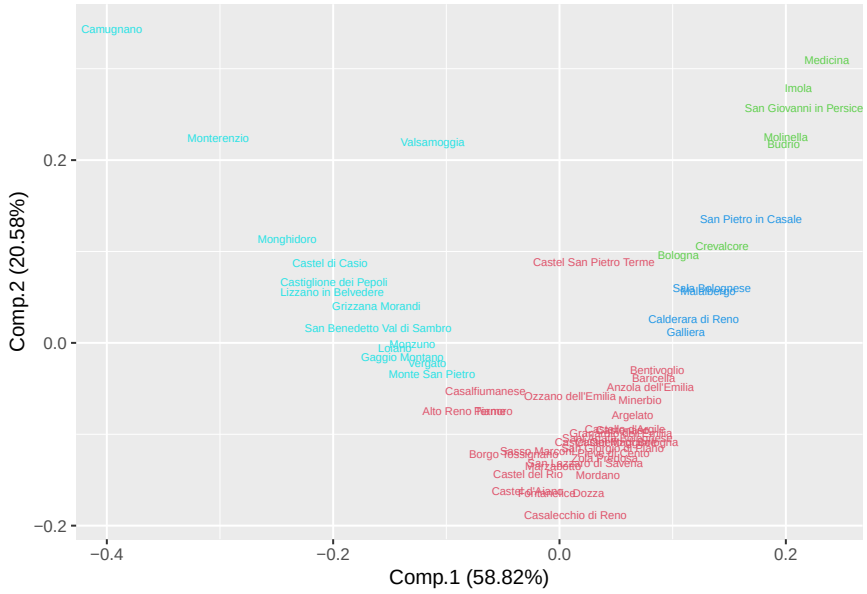
Scegliamo la divisione in 7 cluster: infatti, in questo caso a parità di silhouette media abbiamo una disposizione delle silhouette più uniforme e "sbagliamo di meno".

A questo punto confrontiamo i due modelli vincitori ottenuti proiettando le osservazioni sul piano principale e studiando la disposizione dei cluster per entrambi.



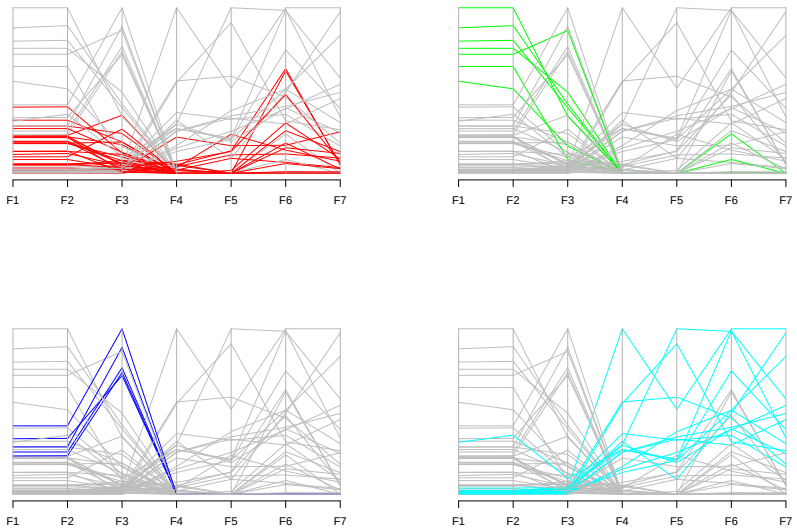
Tra i due scegliamo come modello finale il k-means con k=4, in quanto il complete linkage crea un cluster di 3 elementi e un cluster di 4 elementi che sembrano avere poco significato (le proiezioni delle osservazioni sono molto lontane tra di loro). Tuttavia le due suddivisioni in cluster sono molto simili: questo riconferma

Da ora in poi ci concentreremo sui cluster dati dal metodo a punti prototipo k-means con k = 4. Cominciamo la nostra interpretazione proiettando le osservazioni sul piano principale (i.e. sul piano generato dalle prime due componenti principali) e studiando la disposizione dei cluster. Osserviamo (grazie ai grafici dati da pam) che le prime due componenti principali riescono a descrivere circa il 79% di varianza spiegata.



Osserviamo che eccetto Bologna, Crevalcore e Castel San Pietro Terme gli altri comuni sembrano ben raggruppati e i cluster sembrano sufficientemente distanziati.

Passiamo adesso all'intepretazione dei cluster: per farlo, vediamo i grafici a coordinate parallele.



Possiamo fare alcune considerazioni:

- Il cluster rosso può essere interpretato come **comuni ad alta pericolosità da frana**. Difatti questi comuni presentano valori abbastanza bassi per tutti i fattori eccetto per F6.
- Il cluster ciano può essere interpretato come **comuni ad altissima pericolosità da frana**. In effetti,

gli elementi di questo cluster presentano valori alti negli ultimi 4 fattori, soprattutto in F7.

- Il cluster verde può essere interpretato come **comuni a medio-bassa pericolosità idraulica**. Infatti, tutti i comuni appartenenti a questo cluster presentano valori alti rispetto a F1 e valori leggermente più bassi rispetto a F2 e F3. Rispetto agli ultimi 4 fattori presentano valori trascurabili.
- Il cluster blu può essere interpretato come **comuni ad elevata pericolosità idraulica**, in quanto presentano valori elevati rispetto a F3.

Proviamo infine a capire perchè il metodo k-means sbaglia su Bologna, Crevalcore e Castel San Pietro Terme. Riportiamo nella seguente tabella i valori originali (i.e. non standardizzati) di ognuna delle tre osservazioni precedenti.

	F1	F2	F3	F4	F5	F6	F7
Bologna	88.49	81.28	8.26	0.01	0	9.52	0.00
Castel San Pietro Terme	63.63	63.95	6.16	0.29	0	24.72	0.26
Crevalcore	102.75	102.68	4.19	0.00	0	0.00	0.00

Sia Bologna che Crevalcore sono stati raggruppati come "comuni a medio-bassa pericolosità idraulica". tuttavia presentano valori decisamente sopra la media per F2 e valori quasi nella media per F3: effettivamente, entrambi si comportano come comuni a medio-alta pericolosità idraulica. Per quanto riguarda Castel San Pietro Terme osserviamo che presenta un valore molto sopra la media per F6 (dunque effettivamente è un comune ad alta pericolosità da frana) tuttavia presenta valori ancora più alti per F1 e F2: anche questo comune è a medio-alta pericolosità idraulica (come Bologna e Crevalcore) ed è anche un comune ad alta pericolosità da frana.

Conclusione

Grazie all'analisi svolta siamo in grado di stabilire quali strutture sono necessarie per prevenire problemi idrogeologici. Precisamente, l'analisi precedente ci permette di dividere i comuni di Bologna in 4 gruppi: comuni ad alta pericolosità da frana, comuni ad altissima pericolosità da frana, comuni a medio-bassa pericolosità idraulica e comuni ad elevata pericolosità idraulica. L'analisi è stata svolta con il metodo a punti prototipo k-means, dopo averlo confrontato al metodo pam e a tre metodi di tipo gerarchico: il metodo complete linkage, il metodo average linkage e il metodo single linkage.

Inoltre, durante l'analisi abbiamo notato un comportamento "anomalo" (i.e. il metodo k-means sbagliava in questi casi) di 3 comuni: Bologna, Crevalcore e Castel San Pietro Terme. Dopo aver intepretato i 4 cluster siamo stati in grado di capirne il motivo e abbiamo studiato i 3 comuni separatamente e ad hoc.

Concludendo, la nostra analisi riesce sufficientemente bene a suddividere i comuni di Bologna: per ogni cluster, quindi, possiamo intervenire nel modo più pertinente possibile, ovvero abbiamo soddisfatto l'obiettivo prefissato all'inizio.

Dataset utilizzato

Il dataset utilizzato è reperibile dal seguente link http://dati.istat.it//Index.aspx?QueryId=55347.