

# Analisi dei Terremoti e delle faglie nel mondo (1970-2025)

## Introduzione

In questo lavoro di gruppo, abbiamo deciso di analizzare un dataset preso USGS.gov, che contiene varie informazioni sui terremoti avvenuti in tutti il mondo dal 1970 al 2025. Il dataset è stato selezionato in modo da avere solo terremoti di magnitudo superiore a 5.0 della scala Richter. Ci siamo limitati a questa magnitudo in quanto il sito consente di scaricare fino a 20.000 eventi la volta, quindi per coprire tutta la serie storica è stato necessario unificare tutte le richieste creando un unico dataset con 87798. Inoltre, dal sito globalquakemodel.org il quale contiene la mappa di tutte le faglie attive nel mondo. In questo modo abbiamo potuto confrontare le posizioni dei terremoti rispetto alle faglie. I dati sono stati analizzati tenendo conto sia della loro posizione geografica e di altre caratteristiche presenti nel dataset e sia della loro frequenza nel corso del tempo. Queste analisi sono state effettuate sia attraverso l'utilizzo di mappe, sia attraverso l'utilizzo di grafici al fine di descrivere al meglio tutte le variabili quantitative e qualitative presenti.

## Dataset Faglie

Il dataset contiene la mappa di tutte le faglie attive presenti nel mondo, ed è possibile importarlo tramite il seguente codice:

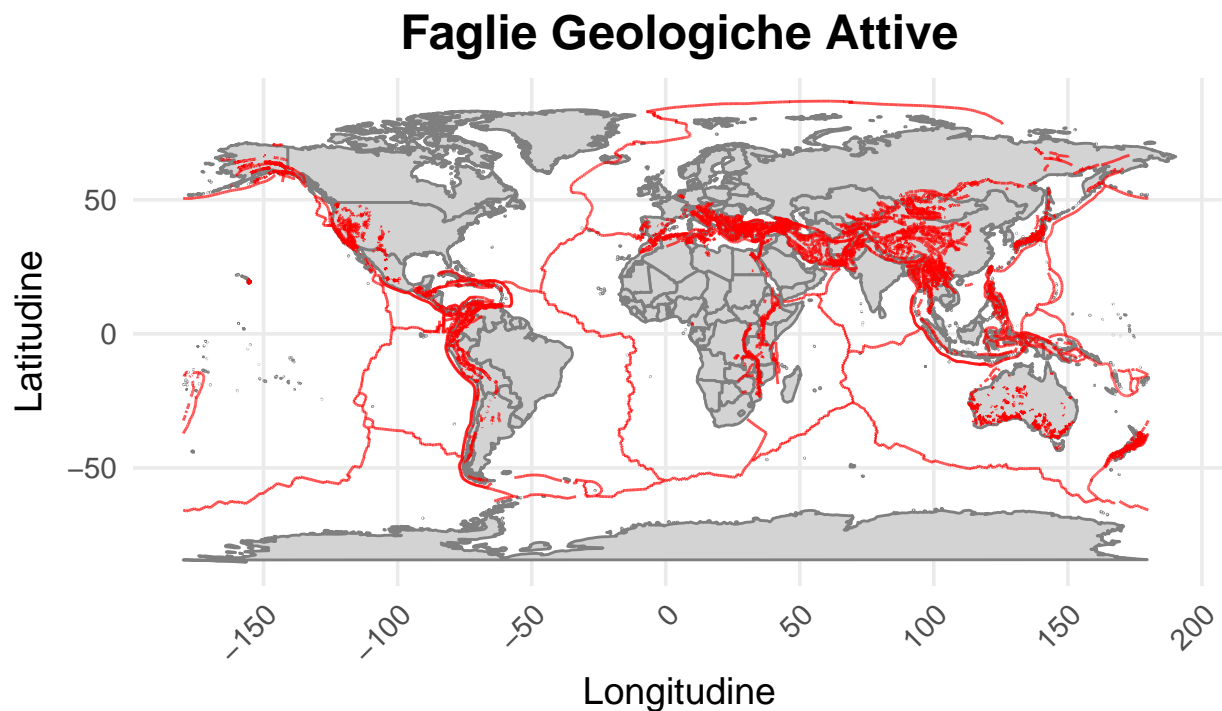
```
library(readr)
library(sf)
library(ggplot2)
library(dplyr)
library(maps)
library(tidyr)
library(RColorBrewer)
library(stringr)
library(rnaturalearth)
library(rnaturalearthdata)
library(ggrepel)
library(lubridate)
```

Per semplicità si è deciso di presentare tutte le librerie utilizzate nel lavoro nel precedente blocco di codice.

```
faglie <- st_read("gem_active_faults.shp")

## Reading layer 'gem_active_faults' from data source
##   'C:\Users\Roberto\Desktop\Progetto in R. L-D-R\gem_active_faults.shp'
##   using driver 'ESRI Shapefile'
## Simple feature collection with 16195 features and 26 fields
## Geometry type: LINESTRING
## Dimension:      XY
## Bounding box:   xmin: -180 ymin: -66.163 xmax: 180 ymax: 86.805
## CRS:           NA
```

```
ggplot() +
  # Aggiungi la mappa del mondo
  borders("world", colour = "gray50", fill = "lightgray") +
  # Aggiungi le faglie
  geom_sf(data = faglie, color = "red", size = 1, alpha = 0.7) +
  labs(title = "Faglie Geologiche Attive", x = "Longitudine", y = "Latitudine",
        color = "Magnitudo") +
  theme_minimal(base_size = 14) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1),
    plot.title = element_text(hjust = 0.5, size = 18, face = "bold"),
    legend.position = "right")
```



Nel grafico, viene mostrata la mappa nel mondo in grigio, e in rosso, le faglie geologiche attive. Come si può vedere le faglie si distribuiscono principalmente lungo i confini delle placche tettoniche con una forte concentrazione nella parte sud-ovest dell'Asia a formare la catena dell'Himalaya. La presenza di faglie geologiche è di particolare interesse in quanto nei pressi di queste zone si osserva una maggiore probabilità che vi siano terremoti. Questo fenomeno verrà visualizzato in seguito, attraverso il dataset "Earthquake" in cui è presente la lista dei terremoti registrati dal 1970-2025.

### Dataset Earthquake:

I dati sono stati presi dal sito USGS.gov filtrati con una magnitudo superiore a 5.0. Essendo la frequenza dei terremoti correlata alla magnitudo il sito consentiva solo di scaricare 20000 eventi per volta, producendo troppe richieste di dati all'ente fornitore. Inoltre, il dataset sull'intero fenomeno era piuttosto frammentato si è deciso quindi di unire le informazioni distribuite su diversi anni in un unico blocco per avere una visione globale del fenomeno.

In R, i dati sono stati importati tramite il seguente codice:

```
terremoti <- read.csv("Earthquake_1970-2025.csv")
```

## Analisi Premilinare.

Il nostro dataset è composto da 22 variabili di varia natura per un totale di quasi 88000 osservazioni. Per semplicità, mostriamo un breve estratto delle variabili più significative usate nell'analisi del fenomeno.

```
library(pander)
pander(head(terremoti %>% select(latitude, longitude, depth, mag, nst, type, time)))
```

latitude	longitude	depth	mag	nst	type	time
-20.18	-70.62	35	5	37	earthquake	2024-12-31T23:13:20.048Z
-6.668	150.6	10	5.1	61	earthquake	2024-12-31T20:09:41.043Z
-4.05	151.6	14.52	5	59	earthquake	2024-12-31T17:09:39.374Z
-17.66	168.2	66.61	5.1	121	earthquake	2024-12-30T05:49:02.808Z
-29.93	-72	10	5.5	127	earthquake	2024-12-30T05:41:06.678Z
-29.92	-72.06	10	5.5	178	earthquake	2024-12-30T05:40:49.261Z

Nella seguente tabella, le variabili indicano:

- **latitudine e longitudine:** rappresentano le coordinate geografiche di dove è avvenuto il terremoto.
- **depth:** La profondità, misurata in km rispetto alla superficie terrestre.
- **mag:** la magnitudo del terremoto in scala Richter.
- **nst:** numero di stazioni che hanno registrato il terremoto.
- **type:** l'origine del terremoto inteso come causa scatenante. (Terremoto terrestre, esplosione nucleare, collasso di miniere, frane ed esplosioni.)
- **time:** data e ora della rilevazione del fenomeno.

Con la funzione summary mostriamo le caratteristiche delle variabili quantitative menzionate sopra.

```
pander(summary(terremoti %>% select(latitude, longitude)))
```

latitude	longitude
Min. :-77.080	Min. :-180.00
1st Qu.: -19.215	1st Qu.: -72.48
Median : -3.692	Median : 102.12
Mean : 1.003	Mean : 40.61
3rd Qu.: 23.936	3rd Qu.: 143.03
Max. : 87.386	Max. : 180.00

```
pander(summary(terremoti %>% select(depth, mag, nst)))
```

depth	mag	nst
Min. : -3.00	Min. :5.000	Min. : 0.0
1st Qu.: 10.00	1st Qu.:5.100	1st Qu.: 69.0

depth	mag	nst
Median : 33.00	Median :5.200	Median :122.0
Mean : 67.16	Mean :5.363	Mean :162.1
3rd Qu.: 57.00	3rd Qu.:5.500	3rd Qu.:219.0
Max. :700.00	Max. :9.100	Max. :934.0
NA	NA	NA's :58768

Ciò che è possibile notare è che per la variabile nst ci sono diversi dati mancanti (NA). Per la variabile depth, si può notare che la mediana e la media non coincidono, lasciando intendere in questa prima fase, la presenza di asimmetria nei dati. Maggiori conferme si hanno guardando la differenza tra il terzo quartile(3rd Qu) e il massimo(Max). Il terzo quartile della variabile mag, ci suggerisce che il 75% dei dati non supera il 5.5 di magnitudo.

## Analisi Quantitativa

Come prima analisi, ci siamo concentrati sulle variabili quantitative cercando possibili correlazioni fra di esse. A tal fine abbiamo utilizzando la funzione “cor” del pacchetto stats, che di default utilizza il coefficiente di correlazione di Pearson. Per una maggiore interpretazione del valore dei coefficienti, si è deciso di utilizzare una heat map generata tramite il seguente codice.

```
library(dplyr)
library(tidyr)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

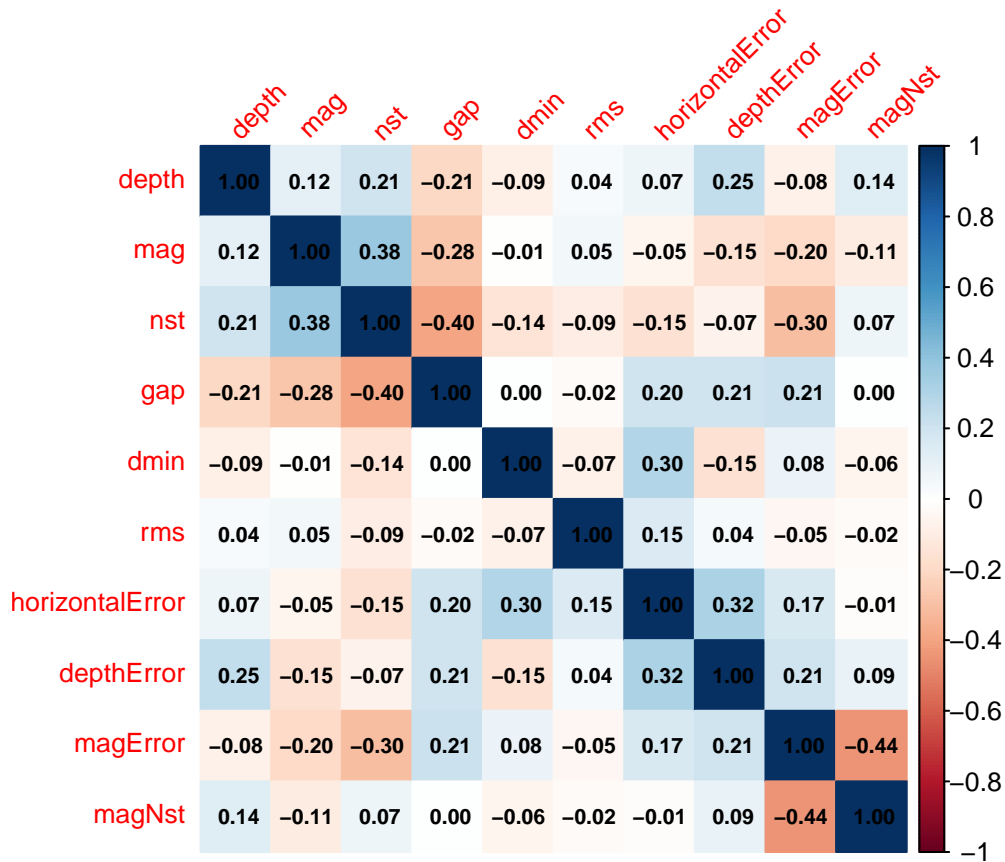
```
library(ggcorrplot)
##seleziono le variabili per l'analisi
df_corr= terremoti%>%
  dplyr::select(depth, mag, nst, gap, dmin, rms, horizontalError, depthError, magError, magNst)
df_na=df_corr %>%
  summarise(across(c(depth, mag, nst, gap, dmin, rms, horizontalError, depthError, magError, magNst),
    ~sum(is.na(.))))

#### seleziono le variabili da pulire da NA
df_corr_clean <- df_corr %>%
  drop_na(depth, mag, nst, gap, dmin, rms, horizontalError, depthError, magError, magNst)

#calcolo la matrice di correlazione
df_f=cor(df_corr_clean)

library(corrplot)

corrplot(df_f,
  method = "color",      # Usa colori per distinguere meglio i livelli di correlazione
  type = "full",         # Mostra solo la parte inferiore della matrice
  tl.cex = 0.8,          # Riduce la dimensione delle etichette per evitare sovrapposizioni
  tl.srt = 45,           # Ruota le etichette per migliorarne la leggibilità
  addCoef.col = "black", # Mostra i valori della correlazione in nero
  number.cex = 0.7)      # Riduce la grandezza dei numeri dentro la matrice
```



Dalla heat map, si evince vi siano numerose correlazioni fra le variabili di cui, risultano di particolare interesse le seguenti:

- **Mag e nst:** Queste due variabili appartengono rispettivamente al livello di magnitudo e al numero di stazioni sismiche che hanno registrato il terremoto. Sono correlate positivamente in quanto un evento con alta magnitudo sarà rilevato da più stazioni sismiche presenti nel mondo.
- **nst e gap:** Con un maggior numero di stazioni (nst elevato), la copertura attorno all'epicentro è più uniforme, riducendo il gap. Al contrario, con meno stazioni la copertura è più concentrata, generando un gap più ampio.
- **Mag Error e magNst:** Queste due variabili appartengono rispettivamente al livello di magnitudo e al numero di stazioni sismiche che hanno partecipato alla determinazione della magnitudo. Per questo motivo osserviamo un coefficiente di correlazione negativo.
- **Horizontal Error e dmin:** L'errore orizzontale potrebbe essere correlato alla distanza minima (dmin) perché terremoti più lontani dalla rete di rilevamento (maggiore dmin) tendono ad avere una localizzazione meno precisa, aumentando l'errore orizzontale. Questo è dovuto alla geometria della rete sismica e alla riduzione della precisione con la distanza.
- **Horizontal Error e depth error:** L'errore orizzontale e l'errore di profondità sono spesso correlati perché entrambi dipendono dalla qualità e dalla distribuzione dei dati sismici. Una scarsa copertura strumentale o una geometria sfavorevole possono influenzare simultaneamente la precisione della localizzazione sia in orizzontale che in profondità.