ANALISI GEOSTATISTICHE

Introduzione

Lo scopo del lavoro richiesto è l'elaborazione di dati per la generazione di due indici geostatistici, rispettivamente un indice globale e un indice locale, relativamente a dati di eventi sismici registrati dall'Istituto Nazionale di Geofisica e Vulcanologia (INGV). Tali dati riguardano in particolare un determinato periodo di tempo (anno solare) e un determinato territorio, in questo caso rappresentato da una regione italiana. I dati che sono stati scelti e utilizzati per questa attività sono quelli ricavati da movimenti tellurici aventi magnitudo superiore a 2 registrati nel corso dell'anno 2016 nella regione Marche. La scelta è stata determinata dal fatto che, nel periodo indicato, il territorio preso in esame è stato soggetto a notevole attività sismica; dunque, è stato ritenuto di particolare interesse per l'analisi dei dati.

Cenni di teoria: Analisi geostatistiche e indici geostatistici

Gli indici geostatistici sono indicatori numerici che permettono di determinare il tipo di distribuzione spaziale di un dataset di features vettoriali, come punti o poligoni, e inoltre di calcolare il grado di autocorrelazione spaziale presente tra di essi. Essi vengono calcolati sui software GIS attraverso algoritmi di geometria computazionale che consentono di eseguire operazioni geometriche in maniera automatica. Il calcolo di questi indici permette di classificare la distribuzione dei dati nelle seguenti tipologie:

- **Dispersa**: i dati sono molto sparsi sulla zona d'interesse e non sono presenti particolari raggruppamenti (clusters);
- **Random**: i dati sono disposti in modo casuale e sono presenti sia raggruppamenti che dati sparsi;
- **Clusterizzata**: vi sono raggruppamenti di dati molto vicini tra di loro.

Per tutti gli indici, l'attendibilità del risultato viene verificata attraverso un test statistico per il quale l'ipotesi nulla è sempre H_0: "la distribuzione dei dati è di tipo "random", ovvero si testa se il dataset è risultato di un processo casuale e dunque non vi è alcuna autocorrelazione spaziale. Vi sono inoltre due approcci fondamentali per effettuare le analisi geostatistiche:

- **Point Pattern Analysis (PPA)**: si prendono in considerazione solo le posizioni degli elementi vettoriali;
- **Spatial Autocorrelation Analysis (SA)**: si considerano, oltre alle posizioni, anche i valori degli attributi associati;

In entrambi i casi lo studio della distribuzione spaziale di features può essere svolto a livello **globale** o **locale**; pertanto, si possono distinguere indici globali ed indici locali. L'analisi globale caratterizza il comportamento complessivo della distribuzione, verificando la presenza di cluster di dati, e produce un report HTML con i risultati del calcolo e del test statistico. È bene ricordare che questo tipo di analisi funziona bene se si effettua l'ipotesi di consistenza del comportamento su tutta l'area di interesse. Fra gli indici globali si annoverano:

- **Nearest Neighbour Index (NNI);**
- **Indice Globale di Moran.**

L'analisi locale studia il comportamento di ciascun elemento rispetto ai circostanti, identificando e classificando i cluster di dati. Il risultato in ambiente GIS di queste elaborazioni è una mappa che fornisce una rappresentazione cartografica dell'eterogeneità spaziale. Alcuni indici locali utilizzati sono:

- **Indice di Getis e Ord Gi*;**
- **Indice di Anselin.**

Nel caso in esame, gli indici che vengono richiesti sono l'indice Globale di Moran e l'indice locale di Getis e Ord Gi*.

Indice Globale di Moran (Global Moran I)

L'indice di Moran è un indicatore **globale** che fornisce una misura del grado di autocorrelazione spaziale della distribuzione dei dati. Esso è definito come:

$$I = \frac{n}{S_0} * \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2}$$

[Equazione 1]

- z_i = *differenza tra il valore dell'attributo associato all'elemento i-esimo ed il valor medio dell'attributo.* $z_i = Z_i - \bar{Z}$
- z_j = *differenza tra il valore dell'attributo associato all'elemento j-esimo ed il valor medio dell'attributo.* $z_j = Z_j - \bar{Z}$
- n = *numero totale di elementi vettoriali;*
- $w_{i,j}$ = *peso spaziale tra elementi i e j: esso può essere calcolato in modi diversi, ad esempio l'inverso della distanza tra i e j;*
- S_0 = *Somma dei pesi spaziali;*

I valori che l'indice I così definito può assumere appartengono all'intervallo [-1; +1]. Un valore di I compreso fra -1 e 0 è indice di autocorrelazione negativa, corrispondente ad una distribuzione estremamente dispersa, mentre un valore di I compreso fra 0 e 1 è indice di autocorrelazione positiva, quindi corrispondente ad una distribuzione con clusters. Quando il valore di I è pari a 0 ciò indica invece una distribuzione randomica, e quindi con correlazione spaziale nulla. Dopo aver eseguito il calcolo dell'indice I, il software effettua anche un test statistico di significatività sul risultato, utilizzando l'ipotesi H_0 definita precedentemente. Il test viene eseguito calcolando il valore normale standardizzato, denominato z-score [Equazione 2]. Per valori positivi statisticamente significativi di z-score, maggiore è il valore di z-score e più elevata è la probabilità che il dataset contenga dei cluster. Per valori negativi statisticamente significativi di z-score, minore è il valore di z-score e più elevata è la probabilità che il dataset sia disperso.

$$z = \frac{I - E[I]}{\sqrt{V[I]}}$$

[Equazione 2]

- E[I] = *valore atteso*;
- V[I] = *varianza.*

Indice di Getis e Ord

L'indice di Getis e Ord è un indicatore locale che viene utilizzato per individuare la posizione dei clusters e distinguere, in base ai valori di un determinato attributo associato a ciascun punto, se i clusters sono raggruppamenti di valori "alti" o "bassi". Per ogni elemento del dataset, questo indicatore viene calcolato attraverso la seguente equazione:

$$Gi^* = \frac{\sum_{j=1}^{n} w_{i,j}x_j - \bar{X}\sum_{j=1}^{n} w_{i,j}}{S\sqrt{\frac{n\sum_{j=1}^{n} w_{i,j-}^2 (\sum_{j=1}^{n} w_{i,j})^2}{n-1}}}$$

[Equazione 3]

Dove:

- $S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \bar{X}^2}$
- x_j = *valore dell'attributo nel punto j;*
- $\bar{X}$ = *valore medio dell'attributo;*
- $w_{i,j}$ = *peso spaziale fra i punti i e j.*

Successivamente al calcolo di Gi* viene effettuato il test di significatività su questo risultato. In questo caso, siccome è dimostrato che Gi* è distribuita come una normale standard, non è necessaria l'operazione di normalizzazione e standardizzazione in quanto il risultato rappresenta già di per sé uno z-score. Una volta eseguito il test, il software GIS nel quale si esegue questa elaborazione assegna ciascun punto od un poligono ad un cluster con un livello di probabilità pari a 90%, 95% e 99% che il risultato sia corretto. I cluster possono essere di due tipi:

- **Hotspot**: raggruppamento di valori simili ed **elevati** dell'attributo;
- **Coldspot**: raggruppamento di valori simili e **bassi** di un attributo.

Per valori positivi statisticamente significativi di z-score, maggiore è il valore dell'indice e più elevata è la probabilità che il punto appartenga a un cluster "**hot spot**", allo stesso modo per valori negativi statisticamente significativi di z-score minore è il valore dell'indice e più elevata è la probabilità che il punto appartenga a un cluster di tipo "**cold spot**". Infine, è inoltre possibile che il dato venga classificato come non appartenente ad alcun cluster.

Al termine dell'elaborazione, viene fornita una rappresentazione cartografica dell'indice di Getis e Ord, nella quale ogni punto o poligono viene rappresentato con il colore rosso se esso viene classificato come appartenente ad un hotspot o con il colore blu se è invece appartenente ad un coldspot, e i 3 gradi di significatività secondo i quali è avvenuta l'assegnazione (90%, 95% e 99%) vengono rappresentati con 3 diverse gradazioni del colore, dal più chiaro al più scuro.

Svolgimento dell'esercizio

Per l'elaborazione richiesta sono stati utilizzati i dati presenti nell'archivio dell'Istituto Nazionale di Geofisica e Vulcanologia. In particolare, filtrando opportunamente tali dati attraverso gli strumenti del portale, vengono ricavate informazioni riguardanti gli eventi di interesse e precisamente quelli avvenuti in Italia e con magnitudo compresa tra 2 e 10. Ogni riga di informazione presenta diversi attributi, tra cui coordinate geografiche (latitudine e longitudine), intensità (magnitudo), profondità in chilometri, data e ora dell'evento. Inoltre, per poter analizzare ed elaborare separatamente i due semestri dell'anno preso in esame, vengono scaricati i dati specificando nel filtro i due periodi di interesse, andando a generare due dataset.

Per lo svolgimento dell'esercizio è stato utilizzato il software ArcGIS Pro. In esso i due insiemi di dati vengono importati in formato tabellare e convertiti in layer di punti. I dati sono forniti in coordinate geografiche, ed in particolare il loro sistema di coordinate è il World Geodetic System 84 (WGS84, codificato in GIS come EPSG:4326). Per questa ragione questi layer di punti sono stati successivamente proiettati nel sistema di riferimento cartografico Universal Transverse Mercator nel fuso 32N, basato sull'ellissoide WGS84 (Codifica GIS: EPSG 32632), poiché, per effettuare le elaborazioni geostatistiche richieste, è necessario che le distanze siano espresse in metri. Dopo aver eseguito questa operazione, sono stati infine isolati i dati relativi al solo territorio della regione italiana scelta per l'esercizio, le Marche, effettuando una clip con lo shapefile corrispondente ai confini regionali, reperito dall'archivio di ISTAT.

Al fine della comprensione del fenomeno, risulta essere di particolare interesse visualizzare in mappa i punti relativi all'attributo "magnitudo", differenziando i punti relativi ai dati registrati nel primo semestre da quelli del secondo semestre. Ogni punto è stato rappresentato in mappa con colori e dimensione diversi a seconda del valore della magnitudo. Si osserva dalle mappe a pagina 3 che vi è una netta differenza nella distribuzione dei dati sul territorio regionale. In *Figura 1* si può notare come i dati siano distribuiti sul territorio in abbastanza sparsa, con l'eccezione di alcuni raggruppamenti di dati localizzati a Sud e Ovest della regione, in corrispondenza della catena montuosa degli Appennini. I valori di magnitudo registrati risultano essere compresi tra 2 e 3,5. Per quanto riguarda il secondo semestre, in *Figura 2*, è evidente che vi sia una notevole disomogeneità nella disposizione dei dati, in quanto è presente un ampio raggruppamento di punti localizzato a Sud-Ovest della regione. In questa zona sono anche stati registrati i terremoti di magnitudo più elevata, fino ad arrivare ad un massimo di 5,9, il quale è stato registrato nel comune di Visso (MC) il giorno 26 ottobre 2016.Tutto questo è riconducibile al fatto che nel suddetto anno, nei mesi di agosto e ottobre, le regioni del centro Italia sono state interessate da considerevole attività sismica.

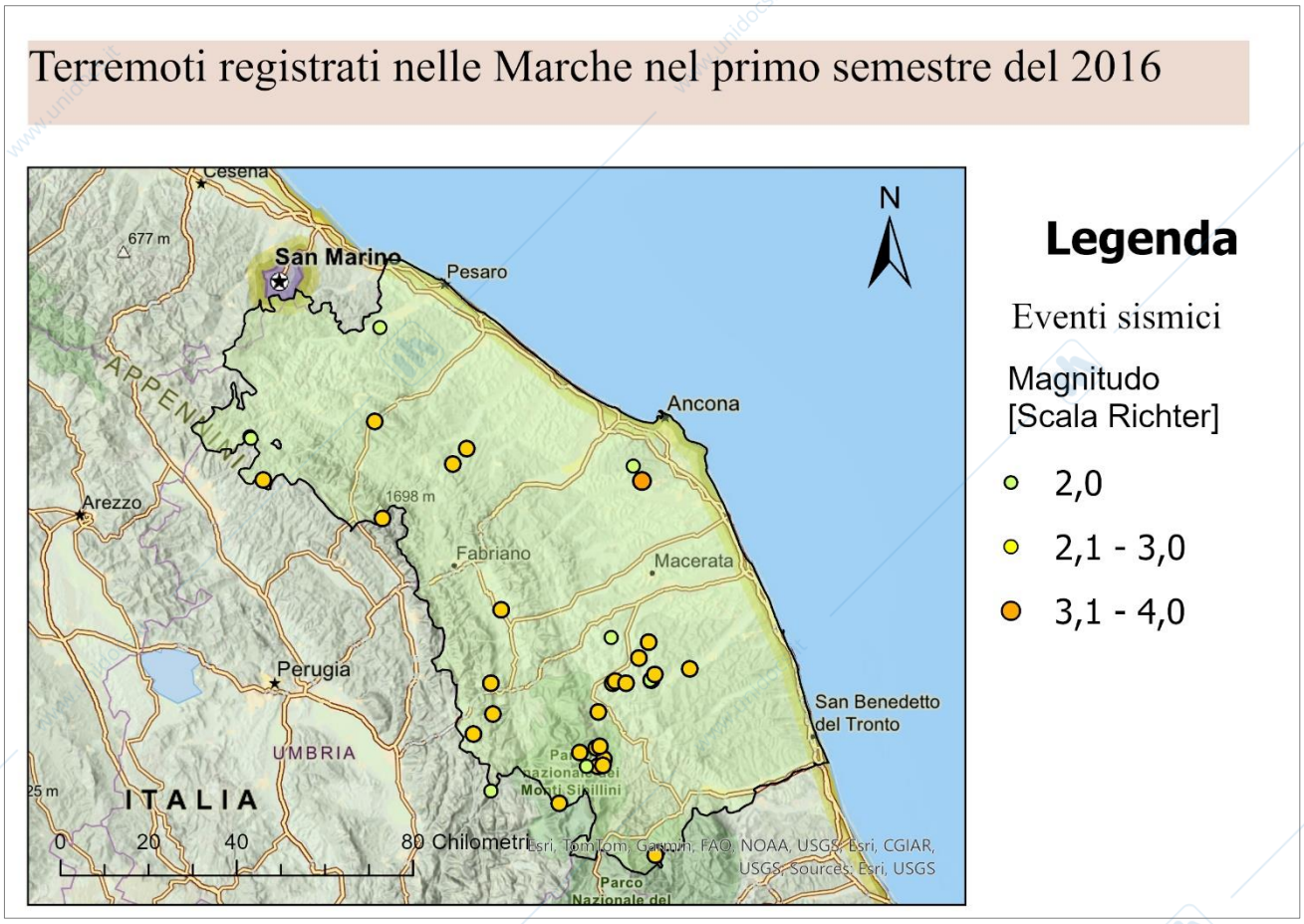


Figura 1: Mappa raffigurante i terremoti avvenuti nel primo semestre del 2016 nel territorio della regione Marche.

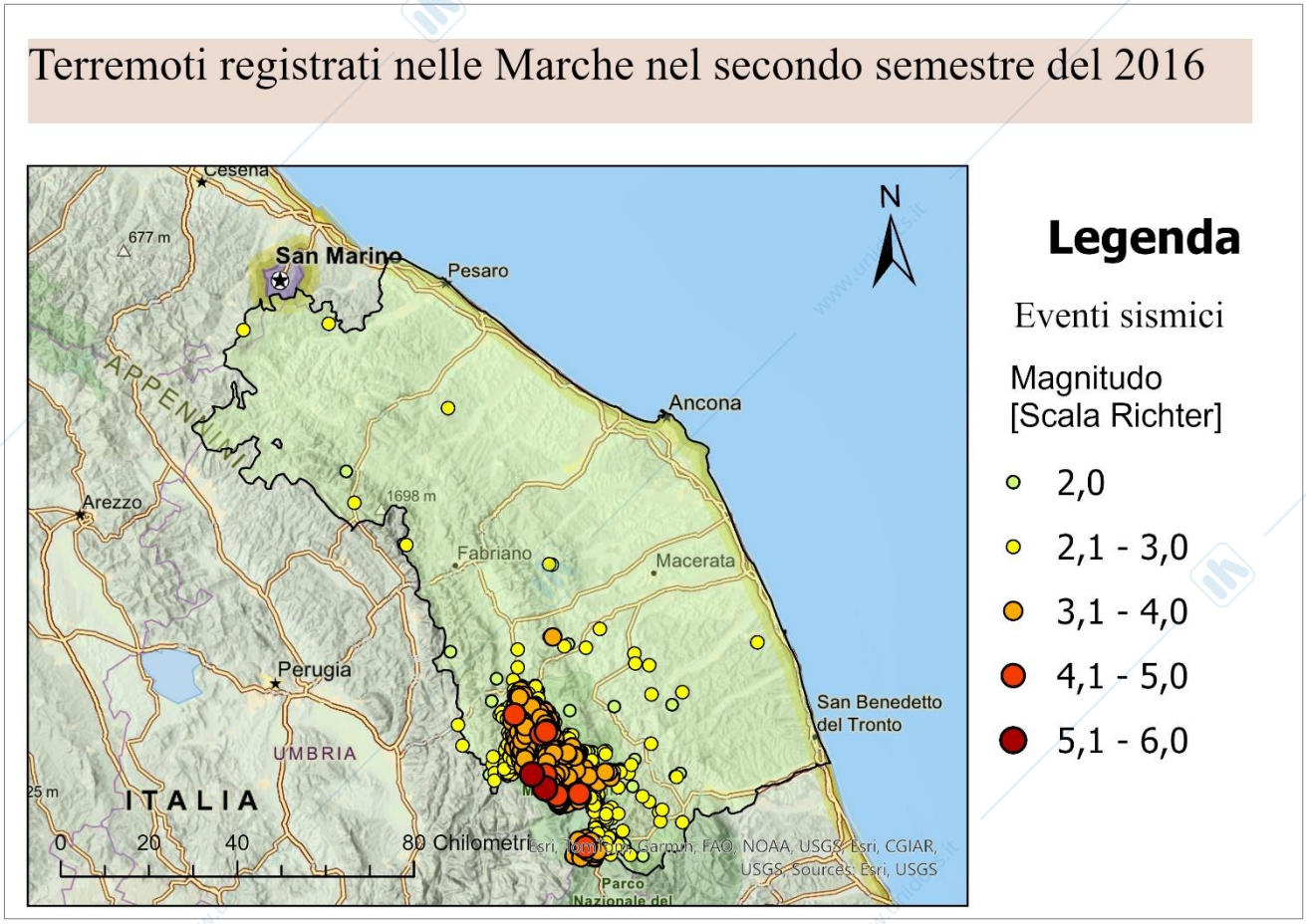


Figura 2: Mappa raffigurante i terremoti avvenuti durante il secondo semestre del 2016 nel territorio della regione Marche.

Calcolo dell'indice globale di Moran

La prima operazione che è stata effettuata è il calcolo dell'indice globale di Moran, separatamente per il primo e il secondo semestre, per individuare la tipologia di distribuzione spaziale dei dati nei due scenari. In ArcGIS Pro, l'output di questa operazione ha fornito due report in formato html che contengono il valore numerico dell'indice di Moran nonché i risultati del test statistico. Essi sono riportati in *Figura 3, Tabella 1* e in *Figura 4, Tabella 2.*

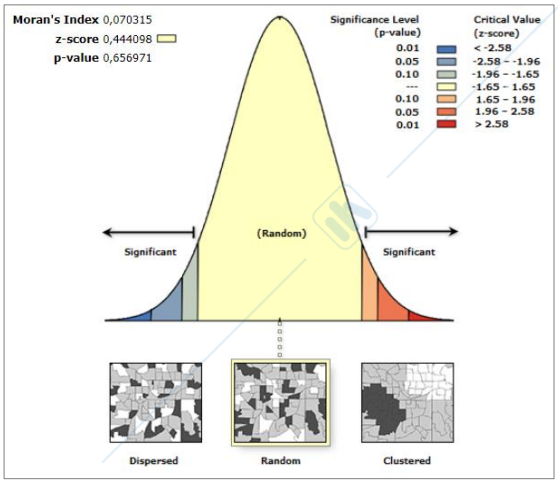


Figura 3: Report dell'indice globale di Moran relativo ai dati del primo semestre del 2016.

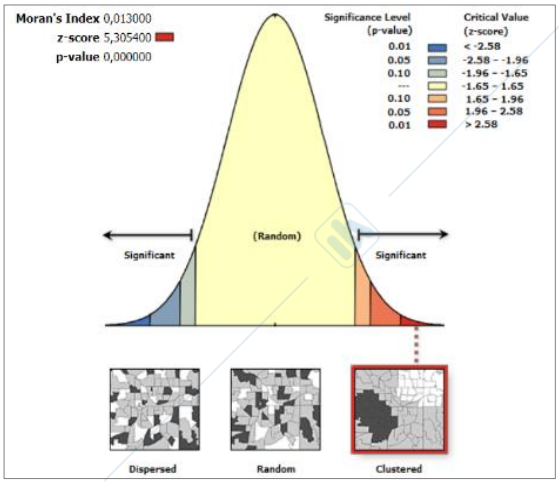


Figura 4: Report dell'indice globale di Moran relativo ai dati del secondo semestre del 2016.

Tabella 1: Risultati dell'indice di Moran relativi al primo semestre.

Indice di Moran	0,070315
Indice Atteso	-0,027778
Varianza	0,048788
z-score	0,444098
p-value	0,656971

Tabella 2: Risultati dell'indice di Moran relativi al secondo semestre.

Indice di Moran	0,013000
Indice Atteso	-0,000258
Varianza	0,000006
z-score	5,305400
p-value	0,0000001

Si osserva in *Figura 3* e in *Tabella 1* che il calcolo dell'indice globale di Moran ha fornito un valore di I pari a 0,07 e un valore di z-score pari a 0,444, portando a classificare la distribuzione dei dati sismici relativi al primo semestre come **randomica**. Per quanto riguarda il test statistico, il valore del p-value è risultato pari a 0,66, e ciò indica che la probabilità che questo pattern di dati possa essere il risultato di un processo casuale, è circa del 66%.

Per quanto riguarda i dati del secondo semestre, il report mostrato in *Figura 4* e in *Tabella 2* ha evidenziato un valore di I pari a 0,013, un valore di z-score pari a 5,305, classificando il pattern assunto dai dati spaziali come una distribuzione **clusterizzata**. Il valore del p-value, infatti, è tendente a 0, il che indica che in questo caso la distribuzione dei dati è altamente improbabile che sia frutto di un processo casuale, e che quindi l'ipotesi H_0 può essere rifiutata con una probabilità di errore minore dell'1%. Si può affermare dunque, alla luce di questi risultati, che solo nel secondo semestre si è verificata una situazione nella quale i dati registrati hanno elevata autocorrelazione spaziale. Ciò risulta essere coerente ed in linea con quanto è stato osservato qualitativamente a pagina 2.

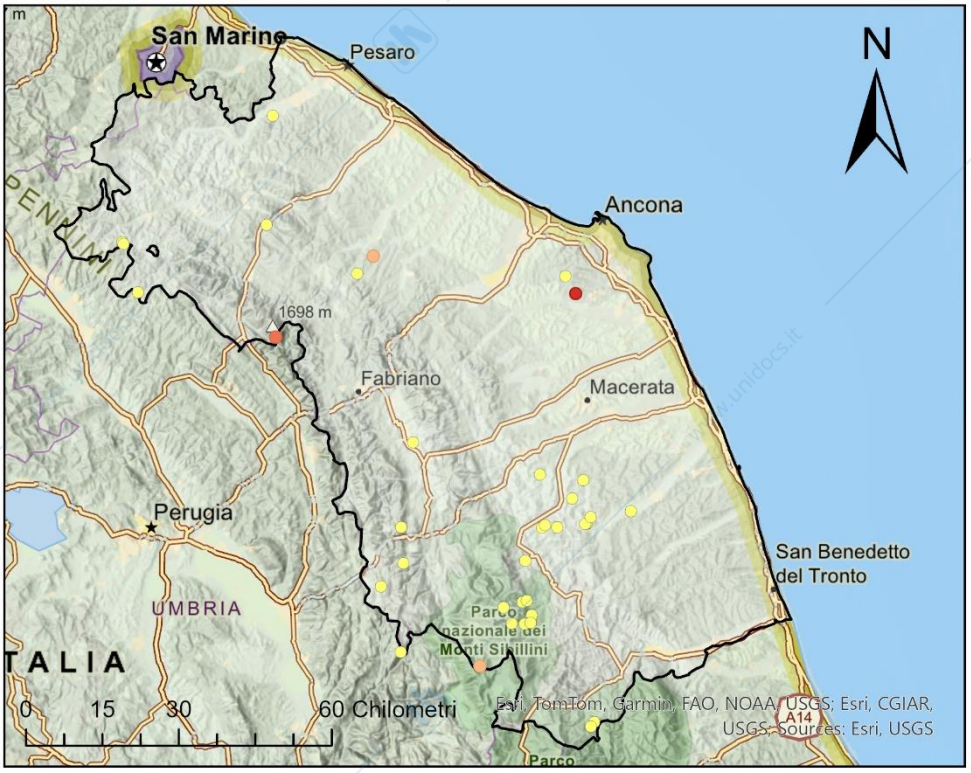
Calcolo dell'indice di Getis e Ord Gi*

Dopo aver classificato le distribuzioni dei dati con l'indice di Moran, si è proceduto con il calcolo dell'indice di Getis e Ord Gi* utilizzando il tool Hot-spot Analysis di ArcGIS Pro, per ricavare l'informazione su dove i dati sismici, caratterizzati da valori alti o bassi di magnitudo, sono raggruppati spazialmente in cluster. Questa elaborazione, eseguita anch'essa separatamente per il primo e secondo semestre, ha fornito in output le rappresentazioni cartografiche riportate in *Figura 5* a pagina 5. Nella prima mappa, corrispondente al calcolo dell'indice per i dati del primo semestre, si osserva un numero esiguo di cluster di dati, i quali risultano essere solo di tipo hotspot. Ciò risulta coerente con il risultato ottenuto dal calcolo dell'indice di Moran, Per quanto riguarda il secondo semestre invece, i cui risultati sono rappresentati nella seconda mappa, si possono osservare parecchi hotspot, evidenziati con il colore rosso. Sono presenti anche altrettanti punti considerati non statisticamente significativi, mentre anche in questo scenario non vi sono stati rilevati coldspot.

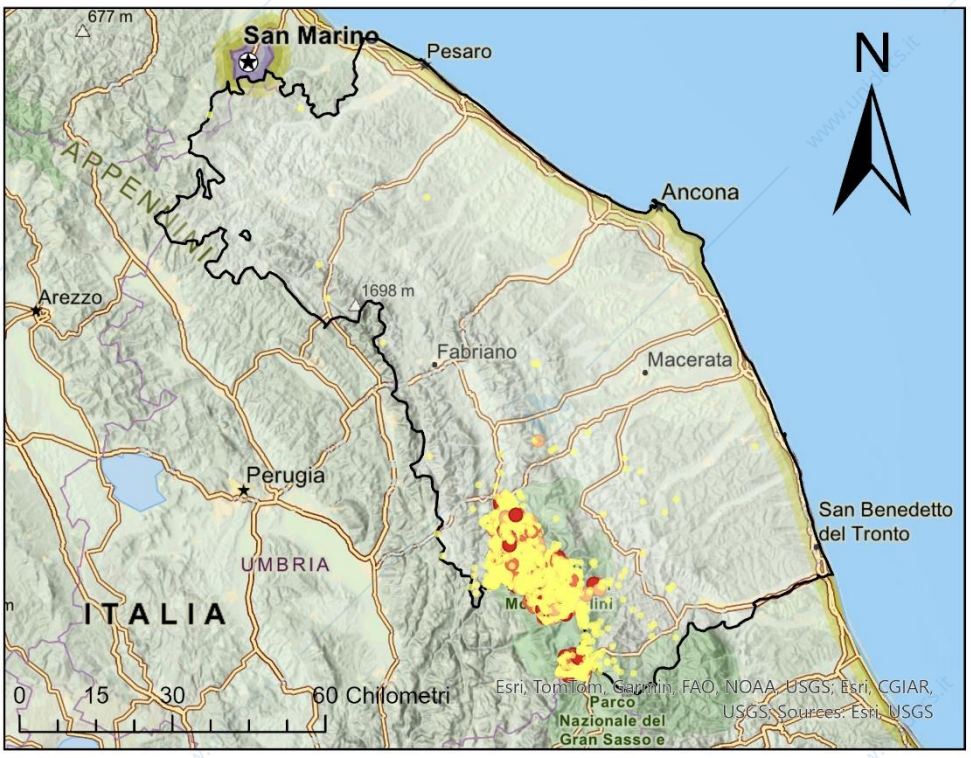
In conclusione, viene riportata la schermata di lavoro in ArcGIS Pro a pagina 6 in *Figura 6.*

Indice di Getis e Ord Gi*

Primo semestre 2016



Secondo semestre 2016



Legenda

Hot Spots e Cold Spots

Valore di Gi*

- Cold Spot with 99% Confidence
- Cold Spot with 95% Confidence
- Cold Spot with 90% Confidence
- Not Significant
- Hot Spot with 90% Confidence
- Hot Spot with 95% Confidence
- Hot Spot with 99% Confidence

Figura 5: Rappresentazione cartografica dell'indice di Getis e Ord Gi*.

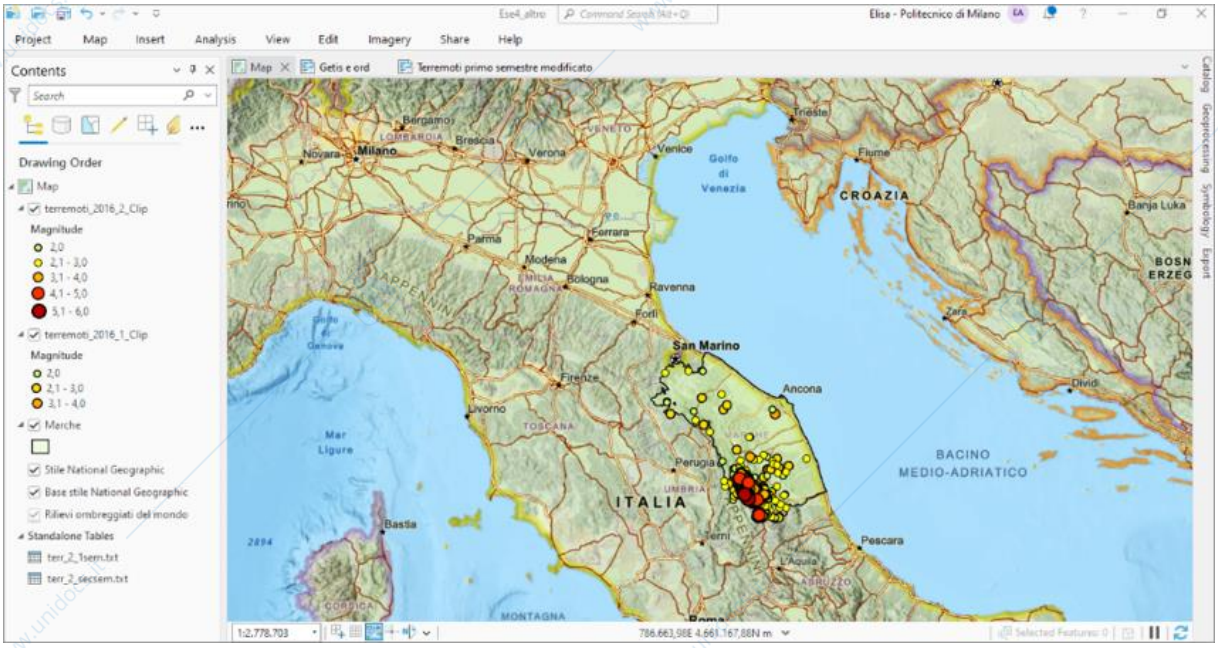


Figura 6: schermata di lavoro in ArcGIS Pro.

Sitografia

- Sito web dell'Istituto Nazionale di Geofisica e Vulcanologia **(INGV):** **https://terremoti.ingv.it/**
- Istituto Nazionale di Statistica, Archivio Cartografia Nazionale, **https://www.istat.it/it/archivio/222527**