**BÁO CÁO TIẾN ĐỘ ĐỒ ÁN MÔN HỌC**

**THÔNG TIN NHÓM**

|  |  |  |  |
| --- | --- | --- | --- |
| **STT** | **MSSV** | **Họ và tên** | **Vai trò** |
| 1 | 19522065 | Nguyễn Thị Minh Phương | Nhóm trưởng |
| 2 | 19521882 | Chu Hà Thảo Ngân |  |
| 3 | 19522397 | Thái Minh Triết |  |

**TÊN ĐỀ TÀI:** PHÂN TÍCH VÀ DỰ ĐOÁN ẢNH HƯỞNG CỦA CÁC YẾU TỐ TỰ NHIÊN ĐẾN TÌNH TRẠNG SẠT LỞ ĐẤT.

**I. Nguồn dữ liệu**

**1. Bộ dữ liệu chính**

- Tên bộ dữ liệu: *Global Landslide Catalog (GLC).*

- Nguồn dữ liệu: [https://data.nasa.gov/Earth-Science/Global-Landslide-Catalog-Export/dd9e-wu2v](%20https://data.nasa.gov/Earth-Science/Global-Landslide-Catalog-Export/dd9e-wu2v)

- Kích thước: 11033 x 31, bao gồm 22 categorical features và 9 numerical features.

- Mô tả bộ dữ liệu: bộ dữ liệu *GLC* được biên soạn năm 2007 tại *NASA Goddard Space Flight Center*. Bộ dữ liệu chứa thông tin liên quan đến các sự kiện sạt lở đất trên toàn thế giới khoảng từ năm 1988 đến năm 2017.

**2. Dữ liệu thu thập thêm**

**2.1. Thời tiết**

- Nguồn dữ liệu: <https://www.visualcrossing.com/weather-api>

- Kích thước: 11033 x 19, bao gồm 2 categorical features và 17 numerical features.

- Mô tả nguồn dữ liệu: dữ liệu được thu thập thêm từ Weather API của *VisualCrossing*. Bộ dữ liệu có chứa các thông tin liên quan đến thời tiết tại thời điểm và địa điểm xảy ra sự kiện sạt lở đất trong bộ *GLC*.

**2.2. Elevation**

- Nguồn dữ liệu: <https://developers.airmap.com/docs/elevation-api>

- Kích thước: 11033 x 1, là numerical feature.

- Mô tả nguồn dữ liệu: dữ liệu được thu thập từ Eleavation API của trang web Airmap. Chứa thông tin về độ cao so với mực nước biển cho hầu hết các vị trí địa lý trên Trái Đất. Độ đo được sử dụng là mét. Độ phân giải không gian: 1 arc-second (khoảng 30 mét).

**2.3. Continent**

- Nguồn dữ liệu: *package pycountry\_convert.*

- Kích thước: 11033 x 1, là categorical feature.

- Mô tả nguồn dữ liệu: package sử dụng dữ liệu từ Wikipedia, cho phép thực hiện việc chuyển đổi giữa tên quốc gia (chuẩn ISO) sang mã quốc gia và châu lục. Các giá trị của châu lục thu được từ bộ dữ liệu gồm có:

* **Asia:** Châu Á
* **North America:** Bắc Mĩ
* **South America:** Nam Mĩ
* **Europe:** Châu Âu
* **Africa:** Châu Phi
* **Oceania:** Châu Đại Dương

**2.4. Season**

- Nguồn tham khảo: <https://www.nationalgeographic.org/encyclopedia/season/>

- Kích thước: 11033 x 1, là categorical feature.

- Mô tả nguồn dữ liệu: dữ liệu về mùa tại sự kiện xảy ra sạt lở đất được thu thập dựa vào thời điểm và vị trí địa lý nó so với đường xích đạo.

|  |  |  |
| --- | --- | --- |
| **Vị trí** | **Thời gian** | **Season** |
| Bắc bán cầu | 21/03 – 20/06 | Spring |
| 21/06 – 22/09 | Summer |
| 23/09 – 20/12 | Autumn |
| 21/12 – 20/03 năm sau | Winter |
| Nam bán cầu | 21/03 – 20/06 | Autumn |
| 21/06 – 22/09 | Winter |
| 23/09 – 20/12 | Spring |
| 21/12 – 20/03 năm sau | Summer |

**2.5. Mật độ dân số**

- Nguồn thu thập: Bộ dữ liệu **Gridded Population of the World Version 4.11**

* <https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev11>
* Google Earth Engine: <https://developers.google.com/earth-engine/datasets/catalog/CIESIN_GPWv411_GPW_UNWPP-Adjusted_Population_Density>

- Kích thước: 11033 x 5, là numerical feature.

- Mô tả nguồn dữ liệu: dữ liệu về mật độ dân số (người/km2) của thế giới trong các năm 2000, 2005, 2010, 2015 và 2020 được thu thập và ước tính bởi *Center for International Earth Science Information Network, Columbia University* và *NASA Socioeconomic Data and Applications Center*, dữ liệu được hiệu chỉnh để phù hợp với dự báo triển vọng dân số thế giới năm 2015 (The World Population Prospects: 2015 Revision) của Liên Hợp Quốc.

- Cách thức thu thập: dữ liệu được thu thập sử dụng thư viện *ee* trong package *geemap* trong python để truy xuất đến bộ dữ liệu **GPWv411:UN-Adjusted Population Density** trong Google Earth Engine, sau đó lấy dữ liệu về mật độ dân số ở từng năm theo kinh độ và vĩ độ.

**2.6. Độ che phủ rừng**

- Nguồn thu thập: bộ dữ liệu **Hansen Global Forest Change v1.8 (2000-2020)**

* <https://data.globalforestwatch.org/documents/134f92e59f344549947a3eade9d80783/explore>
* Google Earth Engine: <https://developers.google.com/earth-engine/datasets/catalog/UMD_hansen_global_forest_change_2020_v1_8>

- Kích thước: 11033 x 3, là numerical feature.

- Mô tả nguồn dữ liệu: dữ liệu về độ che phủ rừng toàn cầu năm 2000 và sự thay đổi (tăng và giảm) về độ che phủ rừng từ năm 2000 đến năm 2020. Dữ liệu được tổng hợp và số hóa từ hình ảnh vệ tinh Landsat 7.

- Cách thức thu thập: dữ liệu được thu thập sử dụng thư viện *ee* trong package *geemap* trong python để truy xuất đến bộ dữ liệu **Hansen Global Forest Change v1.8 (2000-2020)** trong Google Earth Engine, sau đó lấy dữ liệu về độ che phủ rừng theo kinh độ và vĩ độ.

**2.7. Kết cấu của đất (Soil Texture)**

- Nguồn thu thập: Bộ dữ liệu **OpenLandMap Soil Texture Class (USDA System)**

* Google Earth Engine: <https://developers.google.com/earth-engine/datasets/catalog/OpenLandMap_SOL_SOL_TEXTURE-CLASS_USDA-TT_M_v02>

- Kích thước: 11033 x 6, là categorical feature.

- Mô tả nguồn dữ liệu: dữ liệu về 12 loại kết cấu đất (soil texture) theo hệ thống USDA ở 6 độ sâu: 0 cm, 10 cm, 30 cm, 60 cm, 100 cm và 200 cm.

- Cách thức thu thập: dữ liệu được thu thập sử dụng thư viện *ee* trong package *geemap* trong python để truy xuất đến bộ dữ liệu **OpenLandMap Soil Texture Class (USDA System)** trong Google Earth Engine, sau đó lấy dữ liệu về loại kết cấu đất ở từng độ sâu theo kinh độ và vĩ độ.

**3. Bộ dữ liệu kết hợp cuối cùng**

- Bộ dữ liệu cuối cùng được kết hợp từ các bộ dữ liệu thành phần trên, dựa trên khóa là vị trí và thời điểm xảy ra sự kiện sạt lở đất (“latitude”, “longtitude”, “event\_date”).

**-** Kích thước toàn bộ của bộ dữ liệu: 11033 x 67, bao gồm 32 categorical features và 35 numerical features.

**II. Thông tin mô tả bộ dữ liệu**

**1. *Global Landslide Catalog (GLC)***

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Name** | **Datatype** | **Description** |
| 1 | source\_name | object | Tên báo đưa tin |
| 2 | source\_link | object | Liên kết dẫn đến tin |
| 3 | event\_id | int64 | Mã sự kiện sạt lở đất |
| 4 | event\_date | object | Giờ/ngày/tháng/năm diễn ra sạt lở đất |
| 5 | event\_time | float64 | Giờ diễn ra sạt lở đất |
| 6 | event\_title | object | Tiêu đề tin tức sạt lở đất |
| 7 | location\_description | object | Mô tả thông tin vị trí sạt lở |
| 8 | location\_accuracy | object | Khoảng cách chênh lệch giữa vị trí ghi nhận so với vị trí thực tế |
| 9 | event\_description | object | Mô tả sự kiện sạt lở đất |
| 10 | landslide\_category | object | Loại sạt lở đất |
| 11 | landslide\_trigger | object | Nguyên nhân gây ra sạt lở đất |
| 12 | landslide\_size | object | Mức độ sạt lở đất |
| 13 | landslide\_setting | object | Môi trường xung quanh vị trí sạt lở đất |
| 14 | fatality\_count | float64 | Số lượng người tử vong |
| 15 | injury\_count | float64 | Số lượng người thương vong |
| 16 | storm\_name | object | Tên cơn bão xảy ra trước khi sạt lở |
| 17 | photo\_link | object | Đường dẫn tới hình ảnh khu vực bị sạt lở |
| 18 | notes | object | Ghi chú |
| 19 | event\_import\_source | object | Nguồn cung cấp sự kiện sạt lở |
| 20 | event\_import\_id | float64 | Mã cung cấp sự kiện sạt lở |
| 21 | country\_name | object | Tên quốc gia nơi xảy ra sự kiện |
| 22 | country\_code | object | Mã quốc gia |
| 23 | admin\_division\_name | object | Tên đơn vị hành chính |
| 24 | admin\_division\_population | float64 | Dân số của đơn vị hành chính |
| 25 | gazeteer\_closest\_point | object | Vị trí trên bản đồ gần nơi xảy ra sạt lở nhất |
| 26 | gazeteer\_distance | float64 | Khoảng cách từ "gazeteer\_closest\_point" tới nơi xảy ra sạt lở |
| 27 | submitted\_date | object | Ngày nộp/hoàn thành sample trên dataset |
| 28 | created\_date | object | Ngày tạo sample trên dataset |
| 29 | last\_edited\_date | object | Ngày cuối cùng chỉnh sửa sample trên dataset |
| 30 | latitude | float64 | Vĩ độ nơi xảy ra sạt lở |
| 31 | longitude | float64 | Kinh độ nơi xảy ra sạt lở |

**2. Weather**

|  |  |  |  |
| --- | --- | --- | --- |
| **Index** | **Name** | **Datatype** | **Description** |
| 32 | tempmax | float64 | Nhiệt độ cao nhất trong ngày tại địa điểm sạt lở |
| 33 | tempmin | float64 | Nhiệt độ thấp nhất trong ngày tại địa điểm sạt lở |
| 34 | temp | float64 | Nhiệt độ tại địa điểm sạt lở |
| 35 | feelslikemax | float64 | Nhiệt độ cảm thấy cao nhất (ºC) |
| 36 | feelslikemin | float64 | Nhiệt độ cảm thấy thấp nhất (ºC) |
| 37 | feelslike | float64 | Nhiệt độ cảm thấy (ºC) |
| 38 | dew | float64 | Điểm sương (nhiệt độ hóa sương) |
| 39 | humidity | float64 | Độ ẩm tương đối (tính theo %) |
| 40 | precip | float64 | Lượng mưa tại thời điểm “datetime” |
| 41 | precipcover | float64 | Tỉ lệ số giờ trong ngày có lượng mưa khác 0 |
| 42 | windgust | float64 | Gió giật |
| 43 | windspeed | float64 | Tốc độ gió trung bình trên 1 phút |
| 44 | winddir | float64 | Hướng gió |
| 45 | pressure | float64 | Áp suất khí quyển theo mực nước biển hoặc khí áp (tính theo đơn vị millibars hay hectopascals) |
| 46 | cloudcover | float64 | Tỉ lệ bầu trời bị che phủ bởi mây (%) |
| 47 | visibility | float64 | Tầm nhìn xa (km) |
| 48 | moonphase | float64 | Tỉ lệ hình dạng của mặt trăng theo chu kỳ |
| 49 | conditions | object | Điều kiện thời tiết |
| 50 | stations | object | Trạm thời tiết |

**3. Các features khác**

|  |  |  |  |
| --- | --- | --- | --- |
| 51 | elevation | int64 | Độ cao so với mực nước biển (mét) |
| 52 | continent | object | Châu lục |
| 53 | season | object | Mùa |
| 54 | treecover2000 | int64 | Phần trăm độ phủ tán cây tại khu vực trong năm 2000 (áp dụng với thảm thực vật cao trên 5 mét) |
| 55 | loss | int64 | Sự thay đổi từ forest sang non-forest trong khoảng thời gian ghi nhận. |
| 56 | gain | int64 | Sự thay đổi toàn bộ từ non-forest sang forest trong khoảng thời gian từ năm 2000 - 2012 |
| 57 | soil\_texture\_0 | object | Loại kết cấu đất ở bề mặt |
| 58 | soil\_texture\_10 | object | Loại kết cấu đất ở độ sâu 10 cm |
| 59 | soil\_texture\_30 | object | Loại kết cấu đất ở độ sâu 30 cm |
| 60 | soil\_texture\_60 | object | Loại kết cấu đất ở độ sâu 60 cm |
| 61 | soil\_texture\_100 | object | Loại kết cấu đất ở độ sâu 100 cm |
| 62 | soil\_texture\_200 | object | Loại kết cấu đất ở độ sâu 200 cm |
| 63 | population\_density\_2000 | float64 | Mật độ dân số tại khu vực vào năm 2000 (người/km2) |
| 64 | population\_density\_2005 | float64 | Mật độ dân số tại khu vực vào năm 2005 (người/km2) |
| 65 | population\_density\_2010 | float64 | Mật độ dân số tại khu vực vào năm 2010 (người/km2) |
| 66 | population\_density\_2015 | float64 | Mật độ dân số tại khu vực vào năm 2015 (người/km2) |
| 67 | population\_density\_2020 | float64 | Mật độ dân số tại khu vực vào năm 2020 (người/km2) |

**III. Exploratory Data Analysis**

**1. Thống kê mô tả bộ dữ liệu**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| **event\_id** | 11033 | 5598.953141 | 3249.228647 | 1 | 2785 | 5563 | 8435 | 11221 |
| **event\_time** | 0 | NaN | NaN | NaN | NaN | NaN | NaN | NaN |
| **fatality\_count** | 9648 | 3.219424 | 59.886178 | 0 | 0 | 0 | 1 | 5000 |
| **injury\_count** | 5359 | 0.751819 | 8.458955 | 0 | 0 | 0 | 0 | 374 |
| **event\_import\_id** | 9471 | 4798.56307 | 2789.125559 | -111.1673 | 2386.5 | 4773 | 7189.5 | 9669 |
| **admin\_division\_population** | 9471 | 157760.0458 | 829734.5446 | 0 | 1963 | 7365 | 34021 | 12691840 |
| **gazeteer\_distance** | 9471 | 11.873689 | 15.598228 | 0.00003 | 2.363845 | 6.25487 | 15.81561 | 215.4489 |
| **longitude** | 11033 | 2.520441 | 100.908393 | -179.980766 | -107.8717 | 19.6946 | 93.948 | 179.9914 |
| **latitude** | 11033 | 25.881887 | 20.415054 | -46.7748 | 13.9176 | 30.5345 | 40.866259 | 72.6275 |
| **tempmax** | 11004 | 18.857534 | 10.49672 | -22.1 | 11.2 | 21.1 | 27.9 | 45.1 |
| **tempmin** | 11004 | 12.979198 | 9.323941 | -27.1 | 4.8 | 13.2 | 22.1 | 32.1 |
| **temp** | 9776 | 17.725378 | 8.279178 | -24.8 | 10.8 | 19.5 | 25.1 | 38.4 |
| **feelslikemax** | 11001 | 19.957177 | 12.40151 | -22.1 | 11.2 | 21.1 | 30 | 51.9 |
| **feelslikemin** | 11001 | 12.307445 | 10.235743 | -32.3 | 2.3 | 13.2 | 22.1 | 41 |
| **feelslike** | 9773 | 17.842014 | 9.817092 | -28 | 10.4 | 19.5 | 25.5 | 43.6 |
| **dew** | 9740 | 14.202977 | 8.887815 | -27.9 | 7.3 | 14.9 | 22.6 | 175.5 |
| **humidity** | 9740 | 81.949669 | 13.122496 | 14.69 | 76.9175 | 85.02 | 90.95 | 100 |
| **precip** | 9776 | 30.272759 | 52.780018 | 0 | 0.3 | 10.4 | 35.9725 | 498 |
| **precipcover** | 11004 | 26.585559 | 30.337672 | 0 | 0 | 13.64 | 45.83 | 100 |
| **windgust** | 3826 | 50.90149 | 18.46986 | 0 | 38.025 | 50 | 59.4 | 316.8 |
| **windspeed** | 9752 | 20.844924 | 12.993564 | 0 | 12.6 | 18.4 | 27.6 | 222.1 |
| **winddir** | 9269 | 176.340533 | 64.694292 | 0 | 132.7 | 175.9 | 219.4 | 360 |
| **pressure** | 8312 | 1010.184312 | 7.630745 | 951.8 | 1005.9 | 1010.1 | 1014.3 | 1084 |
| **cloudcover** | 9776 | 66.812848 | 30.089074 | 0 | 48.6 | 76 | 91.5 | 100 |
| **visibility** | 9698 | 18.535482 | 74.701551 | 0 | 7.2 | 10.8 | 14.9 | 1030.1 |
| **moonphase** | 11033 | 0.505495 | 0.304086 | 0 | 0.26 | 0.5 | 0.76 | 1 |
| **elevation** | 11033 | 714.618599 | 844.603545 | -407 | 83 | 347 | 1120 | 6916 |

**2. Kiểm tra giá trị khuyết**

|  |  |
| --- | --- |
| **Tên thuộc tính** | **Số lượng missing value** |
| **event\_time** | 11033 |
| **notes** | 10716 |
| **storm\_name** | 10456 |
| **photo\_link** | 9537 |
| **windgust** | 7207 |
| **injury\_count** | 5674 |
| **pressure** | 2721 |
| **winddir** | 1764 |
| **admin\_division\_name** | 1637 |
| **country\_code** | 1564 |
| **gazeteer\_closest\_point** | 1563 |
| **event\_import\_source** | 1563 |
| **gazeteer\_distance** | 1562 |
| **event\_import\_id** | 1562 |
| **country\_name** | 1562 |
| **admin\_division\_population** | 1562 |
| **fatality\_count** | 1385 |
| **visibility** | 1335 |
| **humidity** | 1293 |
| **dew** | 1293 |
| **windspeed** | 1281 |
| **feelslike** | 1260 |
| **temp** | 1257 |
| **conditions** | 1257 |
| **precip** | 1257 |
| **cloudcover** | 1257 |
| **stations** | 1228 |
| **event\_description** | 862 |
| **source\_link** | 846 |
| **location\_description** | 102 |
| **landslide\_setting** | 69 |
| **feelslikemax** | 32 |
| **feelslikemin** | 32 |
| **precipcover** | 29 |
| **tempmax** | 29 |
| **tempmin** | 29 |
| **landslide\_trigger** | 23 |
| **submitted\_date** | 10 |
| **landslide\_size** | 9 |
| **location\_accuracy** | 2 |
| **landslide\_category** | 1 |
| **created\_date** | 1 |
| **elevation** | 0 |
| **moonphase** | 0 |
| **continent** | 0 |
| **source\_name** | 0 |
| **latitude** | 0 |
| **longitude** | 0 |
| **last\_edited\_date** | 0 |
| **event\_title** | 0 |
| **event\_date** | 0 |
| **event\_id** | 0 |
| **season** | 0 |

Tỉ lệ dữ liệu bị khuyết giá trị: 15.03 %

**3. EDA dữ liệu *Global Landslide Catalog***

* **Target: 'landslide\_trigger' (nguyên nhân gây sạt lở)**

Chart

Description automatically generated

* Số lượng label ban đầu: 18
* Gom một số label về chung 1 label (ví dụ gom các loại mưa về chung label ‘*rain’*)
* Số lượng label sau khi xử lý: 12
* Số lượng label ‘rain’ mất cân bằng => sử dụng phương pháp giải quyết mất cân bằng như SMOTE, ADASYN
* **Target: 'landslide\_size’ (mức độ sạt lở)**

Chart, funnel chart

Description automatically generated

* Số lượng label ban đầu: 6
* Gom label ‘*very\_large*’ chung với label ‘*large’*, bỏ label ‘*catastrophic’*
* Số lượng label sau khi xử lý: 4
* **Loại sạt lở đất (landslide\_category)**

**Chart, bar chart

Description automatically generated**

* **Số lượng những sự kiện sạt lở diễn ra hằng năm**

Chart, bar chart, histogram

Description automatically generated

* Số lượng sự kiện sạt lở bắt đầu tăng từ năm 2007
* Hoặc từ năm 2017 NASA mới bắt đầu thu thập các thông tin sạt lở.
* **Ma trận tương quan giữa thương-tử vong và dân số địa phương**

Giữa thương-tử vong và dân số địa phương (đơn vị hành chính) không có tương quan. Khu vực dân số tuy đông nhưng số thương-tử vong do sạt lở cũng thấp.

Chart, waterfall chart, treemap chart

Description automatically generated

* **Ma trận tương quan các thuộc tính**

A picture containing text

Description automatically generated

* Chart

  Description automatically generated**Regplot thương-tử vong và dân số địa phương**

Chart

Description automatically generated

**Chart

Description automatically generated**

* Chart, bar chart

  Description automatically generated**Top 10 quốc gia xảy ra sạt lở nhiều nhất**
* **Phân bố hàng tháng các lần xảy ra sạt lở đất ở một số quốc gia**

Tháng diễn ra sạt lở nhiều nhất:

* Tại Mỹ: từ tháng 12 - tháng 3
* Tại Ấn Độ: tháng 6 - tháng 8
* Tại Việt Nam: tháng 7 - tháng 11
* Diễn ra vào mùa mưa bão, bão tuyết

Chart, histogram

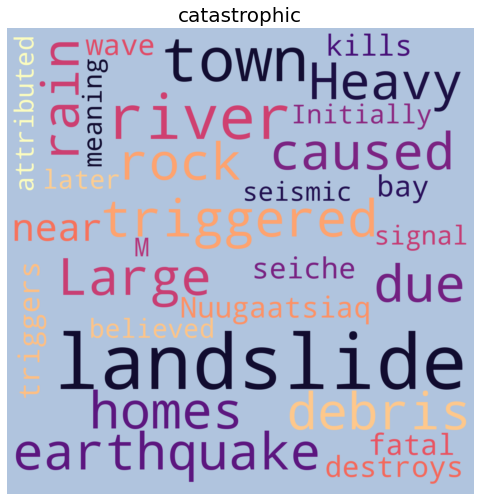
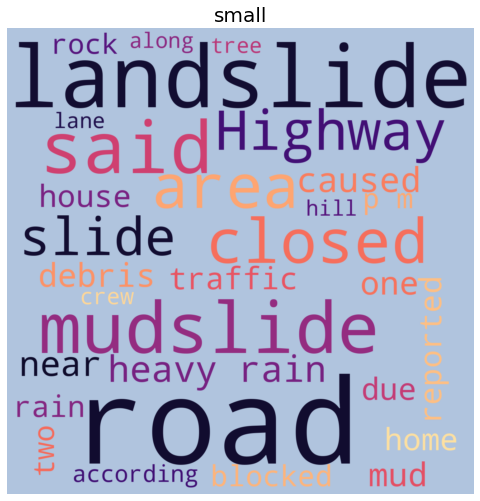
Description automatically generated

Chart, histogram

Description automatically generated

* Chart, histogram

  Description automatically generated**Wordcloud mô tả sự kiện sạt lở theo thuộc tính landslide\_size**



**4. EDA dữ liệu thời tiết**

* **Ảnh có chứa văn bản, bảng trắng

  Mô tả được tạo tự độngBoxplot dữ liệu thời tiết so với thuộc tính landslide\_size**

Ảnh có chứa văn bản, bảng trắng

Mô tả được tạo tự động

**5. Trực quan hóa bản đồ**

* **Trực quan hóa các địa điểm ghi nhận sạt lở trên thế giới**
* **Trực quan hóa các địa điểm ghi nhận sạt lở ở khu vực Châu Á**

****