**ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH**

**TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN**

🙤🙧🟍🙥🙦



**XÂY DỰNG MÔ HÌNH DỰ ĐOÁN**

**QUY MÔ SẠT LỞ ĐẤT**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện: | | |
| STT | Họ tên | MSSV |
| 1 | Nguyễn Thị Minh Phương | 19522065 |
| 2 | Chu Hà Thảo Ngân | 19521882 |
| 3 | Thái Minh Triết | 19522397 |

**TP. HỒ CHÍ MINH – 12/2021**

# GIỚI THIỆU

Trong đồ án lần này, chúng tôi đã áp dụng các kỹ thuật phân tích và trực quan dữ liệu để phân tích sự ảnh hưởng của các yếu tố tự nhiên lên các sự kiện sạt lở đất đã diễn ra, từ đó xây dựng mô hình máy học dự đoán quy mô (kích thước) sạt lở đất.

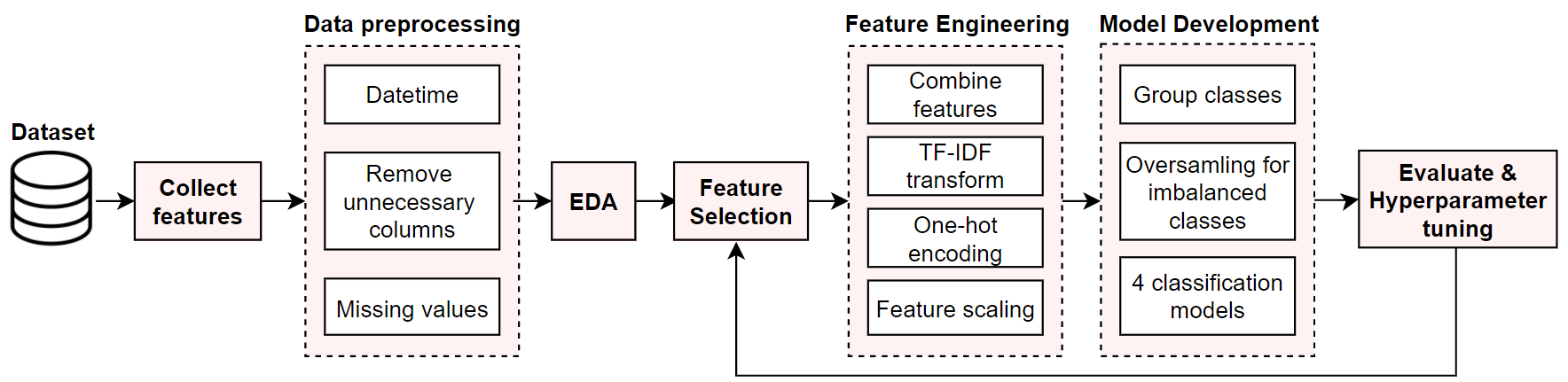
Chúng tôi đã thu thập thêm các thuộc tính về thời tiết, độ cao, mùa, mật độ dân số, kết cấu đất,... để bổ sung vào bộ dữ liệu gốc nhằm hỗ trợ cho quá trình phân tích. Sau đó sử dụng các thư viện *Scikit-Learn, Matplotlib và Seaborn* để tiền xử lý dữ liệu và phân tích thăm dò. Những insights sau khi EDA được kết hợp cùng với hai phương pháp thống kê *Chi-Square* và *ANOVA* để lựa chọn ra được các thuộc tính quan trọng, thực hiện Feature Engineering (*one-hot encoding, TF-IDF transform, MaxAbsScaler,...*) để xử lý trước khi đưa vào mô hình.

Đây là bài toán phân lớp trên dữ liệu mất cân bằng, nên chúng tôi đã áp dụng kỹ thuật *ADASYN* để oversampling cho những lớp thiểu số nhằm giải quyết vấn đề này. Thử nghiệm trên 4 mô hình phân lớp: *Logistic Regression, Support Vector Machine, Random Forest và Passive Aggressive Classifier* và sử dụng độ đo *Macro F1-score* để đánh giá cho bài toán mất cân bằng. Cuối cùng chúng tôi sử dụng *GridSearch* tinh chỉnh siêu tham số để có được mô hình tốt nhất.

Kết quả thu được với mô hình tốt nhất là *Logistic Regression* với accuracy bằng 0.717189 và macro F1-score bằng 0.522588.

# NỘI DUNG

Chúng tôi thực hiện đồ án theo quy trình phân tích dữ liệu cụ thể cho bài toán như sau:



*Hình 1. Quy trình phân tích dữ liệu*

## Giới thiệu các bộ dữ liệu

**2.1.1. Bộ dữ liệu gốc**

* Tên bộ dữ liệu: Global Landslide Catalog (GLC).
* Nguồn dữ liệu: <https://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.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả** | **Miền giá trị** |
| 1 | source\_name | object | Tên báo đưa tin | 9News, SkyNews,… |
| 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 | 1988-11-07 to 2017-09-28 |
| 5 | event\_time | float64 | Giờ diễn ra sạt lở đất |  |
| 6 | event\_title | object | Tiêu đề tin tức sạt lở đất | Landslide in Shazi,… |
| 7 | location\_description | object | Mô tả thông tin vị trí sạt lở | Tay Tra district,… |
| 8 | location\_accuracy | object | Khoảng cách chênh lệch giữa vị trí ghi nhận so với vị trí thực tế | unknown, exact, 5km, 10km, 25km, 100km… |
| 9 | event\_description | object | Mô tả sự kiện sạt lở đất | LANDSLIDES hit Surat … |
| 10 | landslide\_category | object | Loại sạt lở đất | landslide, mudslide, debris flow,… |
| 11 | landslide\_trigger | object | Nguyên nhân gây ra sạt lở đất | downpour, rain, continuousrain |
| 12 | landslide\_size | object | Mức độ sạt lở đất | large, small, medium,… |
| 13 | landslide\_setting | object | Môi trường xung quanh vị trí sạt lở đất | mine, above\_road, natural\_slope,... |
| 14 | fatality\_count | float64 | Số lượng người tử vong | 0 to 5000 |
| 15 | injury\_count | float64 | Số lượng người thương vong | 0 to 374 |
| 16 | storm\_name | object | Tên cơn bão xảy ra trước khi sạt lở | Agaton |
| 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 | glc, test, … |
| 20 | event\_import\_id | float64 | Mã cung cấp sự kiện sạt lở |  |
| 21 | country\_name | object | Tên quốc gia nơi xảy ra sự kiện | Vietnam,… |
| 22 | country\_code | object | Mã quốc gia | US, PH… |
| 23 | admin\_division\_name | object | Tên đơn vị hành chính | New York,… |
| 24 | admin\_division\_population | float64 | Dân số của đơn vị hành chính | 0 to 13M |
| 25 | gazeteer\_closest\_point | object | Vị trí trên bản đồ gần nơi xảy ra sạt lở nhất | Morongo Valley |
| 26 | gazeteer\_distance | float64 | Khoảng cách từ *"gazeteer\_closest\_point"* tới nơi xảy ra sạt lở | 3e-5 to 215.45 (km) |
| 27 | submitted\_date | object | Ngày nộp/hoàn thành sample trên dataset | 2014-04-01 to 2017-11-21 |
| 28 | created\_date | object | Ngày tạo sample trên dataset | 2017-11-20 to 2017-12-20 |
| 29 | last\_edited\_date | object | Ngày cuối cùng chỉnh sửa sample trên dataset | 2018-02-15 |
| 30 | latitude | float64 | Vĩ độ nơi xảy ra sạt lở | -46.77 to 72.62 |
| 31 | longitude | float64 | Kinh độ nơi xảy ra sạt lở | -179.98 to 179.99 |

**2.1.2. Dữ liệu thu thập thêm**

Kích thước các thuộc tính thu thập thêm: ***11033 x 36***, bao gồm 10 categorical features và 26 numerical features.

**2.1.2.1. Nguồn thu thập và tham khảo:**

* Thời tiết: <https://www.visualcrossing.com/weather-api/>
* Độ cao: <https://developers.airmap.com/docs/elevation-api/>
* Châu lục: <https://pypi.org/project/pycountry-convert/>
* Mùa: <https://www.nationalgeographic.org/encyclopedia/season/>
* Mật độ dân số: Bộ dữ liệu Gridded Population of the World Version 4.11 <https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-density-rev11/>
* Độ che phủ rừng: Bộ dữ liệu Hansen Global Forest Change v1.8 (2000-2020) <https://data.globalforestwatch.org/documents/134f92e59f344549947a3eade9d80783/explore/>
* Kết cấu của đất: Bộ dữ liệu OpenLandMap Soil Texture Class <https://developers.google.com/earth-engine/datasets/catalog/OpenLandMap_SOL_SOL_TEXTURE-CLASS_USDA-TT_M_v02>

**2.1.2.2. Phương pháp thu thập:**

Đối với dữ liệu thời tiết, chúng tôi sử dụng api từ website *VisualCrossing* để lấy chỉ số thời tiết tại vị trí và thời điểm xảy ra sạt lở. Một vài chỉ số có tỉ lệ giá trị bị khuyết rất lớn và ít quan trọng nên chúng tôi quyết định không thu thập chúng. Về độ cao so với mực nước biển, chúng tôi sử dụng api từ *websiteAirmap* để lấy dữ liệu độ cao tại vị trí xảy ra sạt lở.

Đối với dữ liệu về mật độ dân số, độ che phủ rừng và kết cấu của đất, chúng tôi sử dụng thư viện *ee* trong package *geemap* để truy xuất dữ liệu từ các dataset tương ứng trên nền tảng Google Earth Engine.

Thuộc tính châu lục được tạo thêm dựa vào thuộc tính cơ sở là *‘country’*. Thuộc tính mùa được tạo thêm dựa vào thuộc tính cơ sở là *‘event\_date’*, *‘latitude’* và *‘longitude’* theo các khái niệm về xuân phân, hạ chí, thu phân, đông chí ở hai nửa bán cầu.

*(Mô tả thuộc tính dữ liệu thu thập thêm tại* ***phụ lục A****)*

**2.1.3. Kết hợp các bộ dữ liệu**

* 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”, “longitude”, “event\_date”).
* Kích thước toàn bộ của bộ dữ liệu: 11033 x 67, bao gồm 32 categorical features và 35 numerical features.

## Tiền xử lý dữ liệu

Chúng tôi sử dụng bộ dữ liệu được kết hợp cuối cùng cho bước phân tích thăm dò và phát triển mô hình. Trước khi tiền xử lý, bộ dữ liệu có tới 54 thuộc tính chứa giá trị bị khuyết và tỉ lệ dữ liệu bị khuyết giá trị là 12.18% với 90051 giá trị.

*Chúng tôi tiến hành các bước tiền xử lý dữ liệu như sau:*

**2.2.1. Xử lý các thuộc tính dạng datetime**

Định dạng lại các thuộc tính dạng datetime: ‘*event\_date’*, ‘*created\_date’, ‘last\_edited\_date’, ‘submitted\_date’, ‘event\_time’* thành kiểu dữ liệu datetime trong Python.

**2.2.2. Loại bỏ các thuộc tính không cần thiết**

Chúng tôi tiến hành loại bỏ các thuộc tính không cần thiết ứng với từng lí do sau:

|  |  |
| --- | --- |
| **Loại bỏ thuộc tính** | **Lý do loại bỏ** |
| *'source\_link'* (2)*, 'photo\_link'* (17) | Vì là liên kết trang web |
| *'event\_id'* (3)*, 'event\_import\_id'* (20) | Vì là mã định danh |
| *'submitted\_date'* (27)*, 'created\_date'* (28)*, 'last\_edited\_date'* (29) | Chỉ chứa các thông tin tracking sample trên dataset |
| *'storm\_name'* (16)*, 'notes'* (18) | Số lượng sample là null trên 90% |
| *'event\_time'* (5) | Trong ‘event\_date’ đã chứa thông tin thời gian |
| *'event\_import\_source'* (19)*, 'source\_name'* (1) | Chứa những thông tin không quan trọng. |

**2.2.3. Xử lý các giá trị bị khuyết**

* *‘country\_name’* (21): sử dụng module *Nominatim* của thư viện *geopy* để định dạng toàn bộ tên quốc gia theo tiếng Anh cũng như xử lý tên quốc gia bị thiếu dựa trên toạ độ.
* *'admin\_division\_name'* (23) và ‘*location\_description’* (7): điền giá trị bị khuyết bằng tên quốc gia tương ứng.
* *‘gazeteer\_closest\_point’* (25): trong bộ dữ liệu, địa điểm này thường là tên thành phố hoặc tên thị trấn, tên ngôi làng… - nơi gần với sự kiện sạt lở xảy ra. Do đó, chúng tôi giải quyết địa điểm bị khuyết bằng tên đơn vị hành chính tương ứng tại nơi xảy ra sạt lở đó.
* Ở các thuộc tính chứa thông tin mô tả còn lại, thay thế giá trị bị khuyết là một chuỗi rỗng.
* *'admin\_division\_population'* (24): có mức độ tương quan vừa (~0.50) với mật độ dân số qua các năm, sử dụng phương pháp *KNN Imputation* để điền giá trị bị khuyết dựa trên mật độ dân số qua các năm.
* *‘fatality\_count’* (14), *‘injury\_count’* (15): giá trị trên 2 thuộc tính này là số nguyên và xuất hiện giá trị ngoại lai (outliers), do đó dùng giá trị *median* để thay thế cho những giá trị bị khuyết.

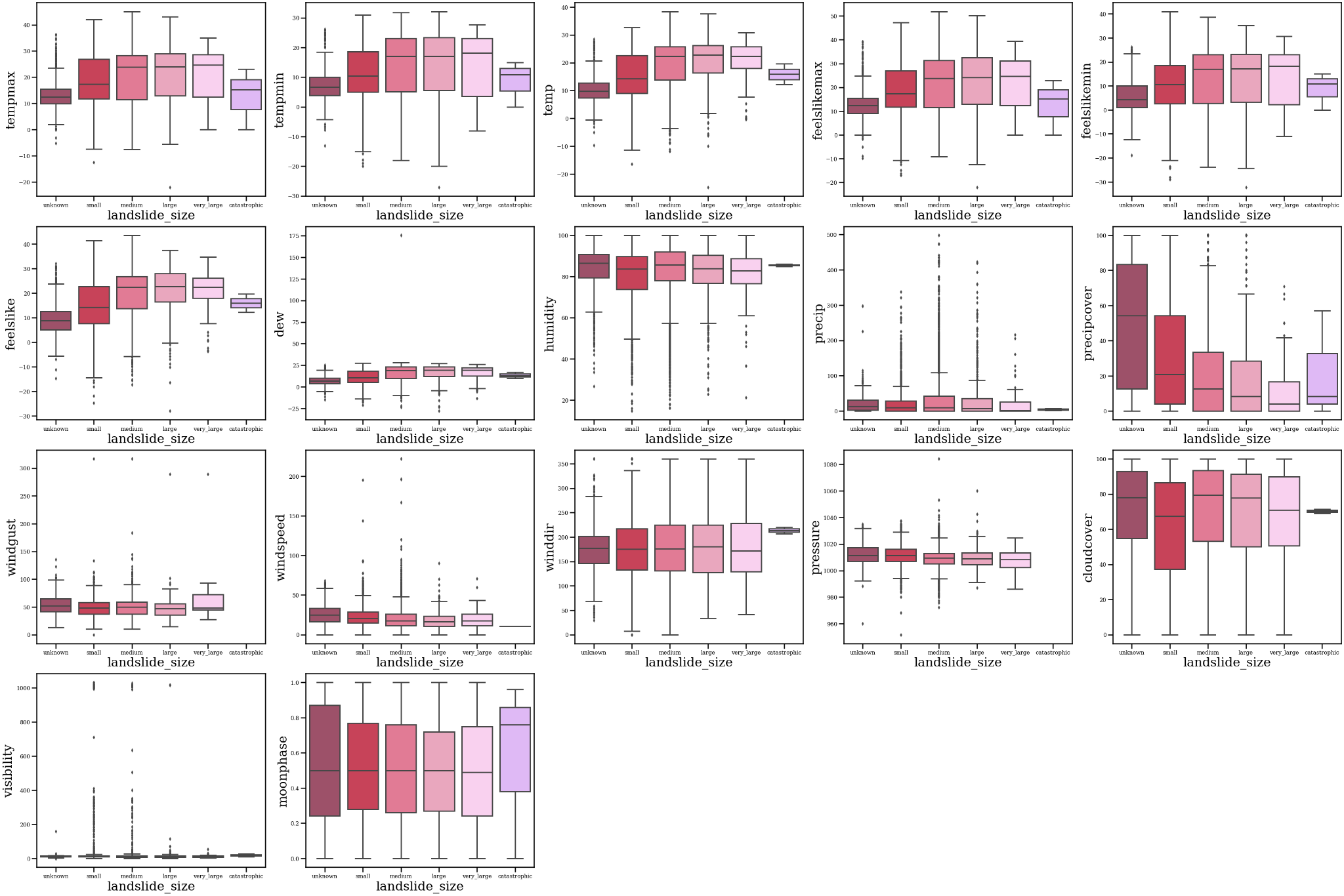
Với các thuộc tính là biến định lượng liên tục, chúng tôi plot phân bố giá trị của từng thuộc tính. Nếu thuộc tính tuân theo phân phối chuẩn, điền giá trị bị khuyết là giá trị *mean*. Nếu thuộc tính có nhiều outliers thì lựa chọn điền khuyết bằng giá trị *median* sẽ phù hợp hơn.

Cuối cùng, chúng tôi quyết định loại bỏ những sample có giá trị bị khuyết ở hầu hết các thuộc tính, cũng như vì chưa có cách điền khuyết phù hợp.

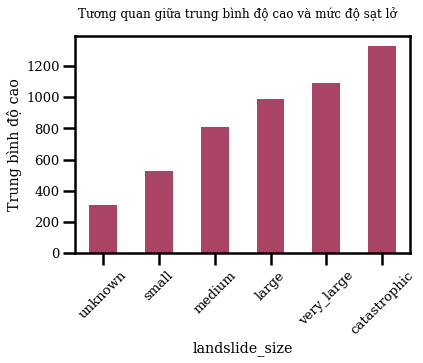
Kích thước của bộ dữ liệu sau khi xử lý các giá trị bị khuyết có được là ***9345 x 54*** so với bộ dữ liệu ban đầu là 11033 dòng và 67 cột.

## Phân tích thăm dò

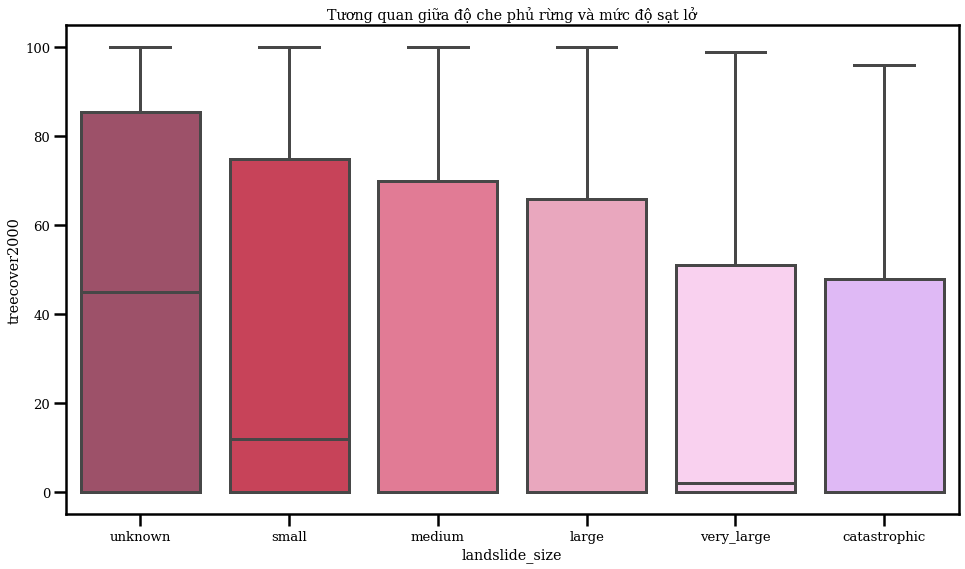
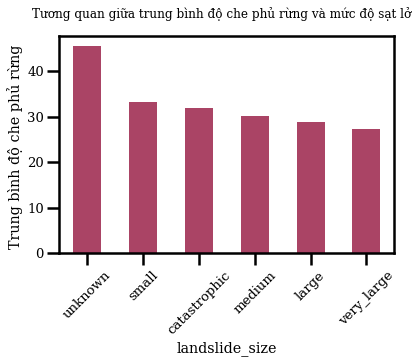
**2.3.1. Thống kê mô tả**



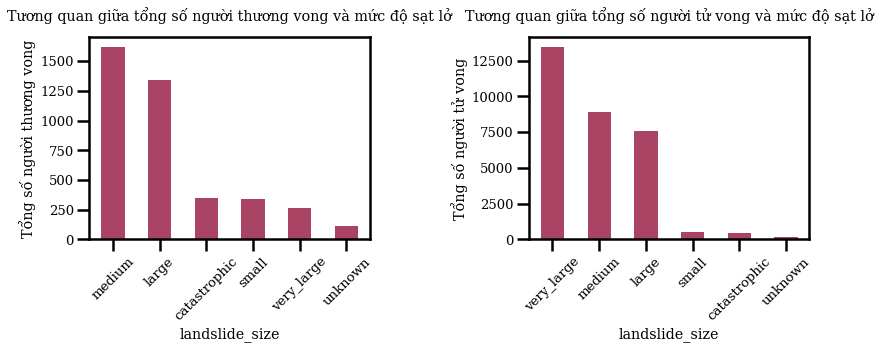
Boxplot các thuộc tính thời tiết đa số cho thấy có sự overlap lẫn nhau giữa các quy mô sạt lở, ngoài trừ một số thuộc tính cho thấy sự khác biệt như *‘temp’*, *‘feelslike’*, *‘precipcover’*. Các khu vực có chỉ số *‘temp’* và *‘feelslike’* càng cao, chỉ số *‘precipcover’* càng thấp thì có khả năng quy mô sạt lở càng nghiêm trọng.

****

Dựa vào boxplot và barplot tương quan giữa độ cao và quy mô sạt lở, chúng tôi nhận thấy có sự khác nhau về các độ cao địa hình giữa các quy mô sạt lở. Ở các khu vực có địa hình thấp dưới 800m đến dưới 1000m, quy mô sạt lở thường là vừa, nhỏ hoặc không xác định. Độ cao địa hình càng tăng thì quy mô sạt lở càng có xu hướng nghiêm trọng hơn, tuy nhiên chưa có sự chênh lệch đáng kể về độ cao ở quy mô sạt lở vừa, lớn, rất lớn và thảm khốc.

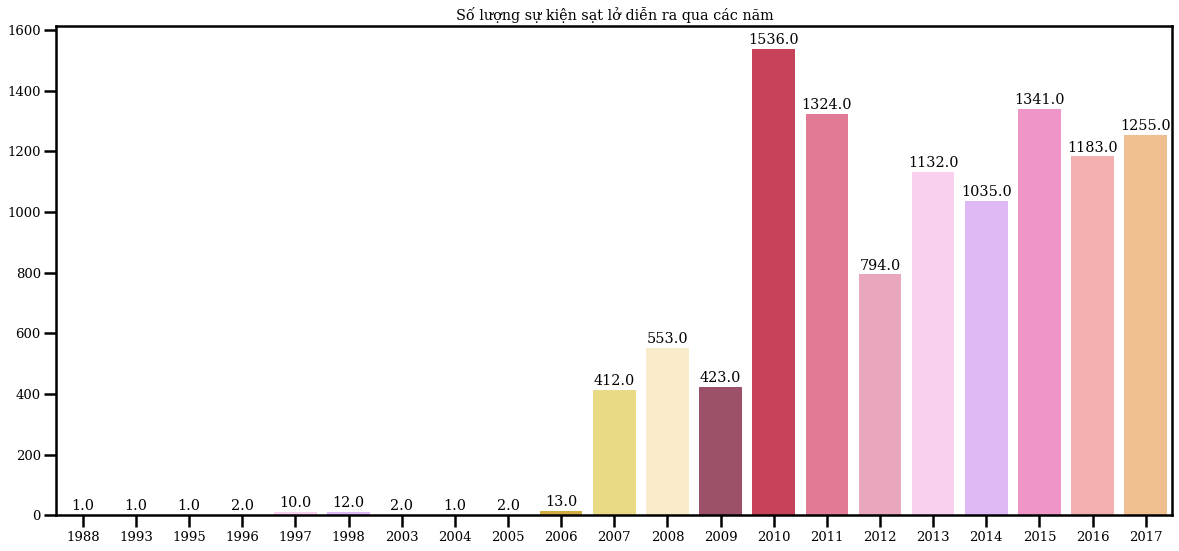
****

Nhìn chung quy mô sạt lở giảm dần ở các khu vực có độ che phủ rừng cao. Cho thấy được vai trò giữ đất của thảm thực vật, hạn chể nguy cơ xảy ra sạt lở nghiêm trọng.

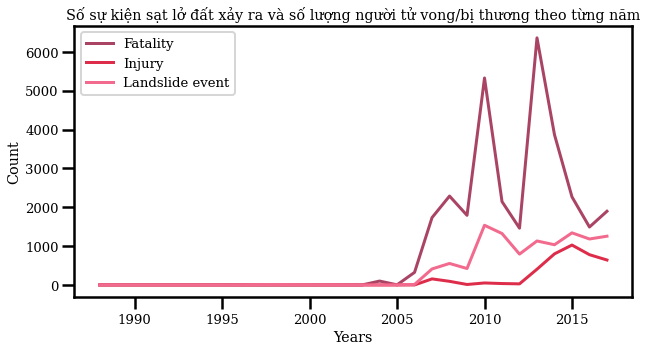
****

Các vụ sạt lở quy mô vừa và lớn khiến nhiều người thương-tử vong. Đặc biệt, sạt lở quy mô rất lớn gây thiệt hại về sinh mạng nhiều nhất trong khi số thương vong chỉ ở mức nhỏ.

**2.3.2. Phân tích trực quan dữ liệu thời gian**

****

Như đã đề cập ở mục 2.1.1, trung tâm Goddard Space Flight của NASA bắt đầu biên soạn bộ dữ liệu vào năm 2007, do đó từ năm 2007 trở đi mới đầy đủ thông tin các sự kiện sạt lở. Đặc biệt, vào tháng 7 năm 2013 có hơn 200 sự kiện sạt lở trên toàn thế giới. Trước đó, đã diễn ra thảm họa sạt lở tại khu vực thung lũng Kedarnath, Uttarakhand, Ấn Độ vào ngày 16/06/2013 khiến hơn 5000 người thiệt mạng, nguyên nhân do mưa lớn kéo dài [4]

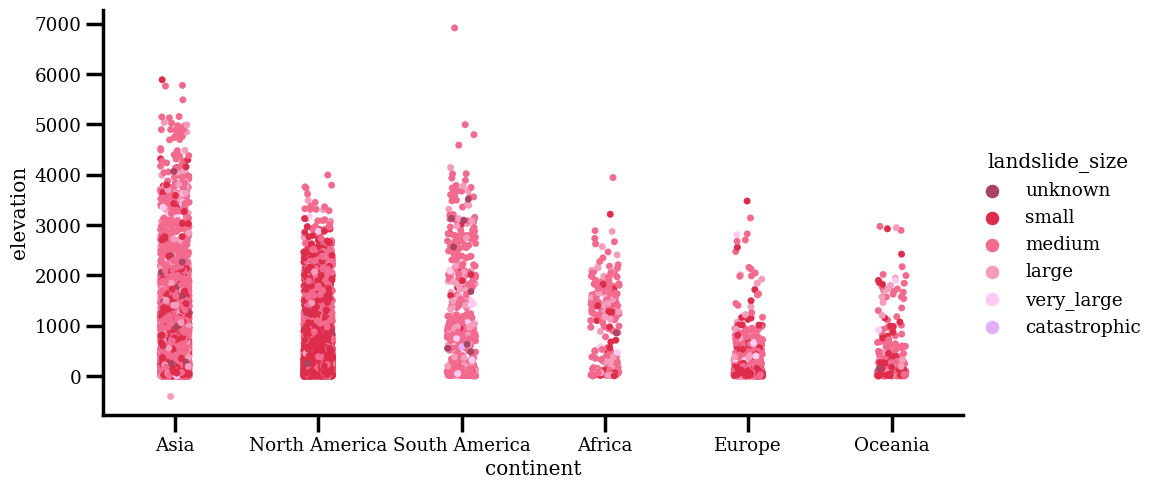


Năm 2010 và 2013 là năm có nhiều sự kiện sạt lở đất xảy ra, cũng như có nhiều số người tử vong.

**2.3.3. Phân tích tổng hợp thuộc tính**

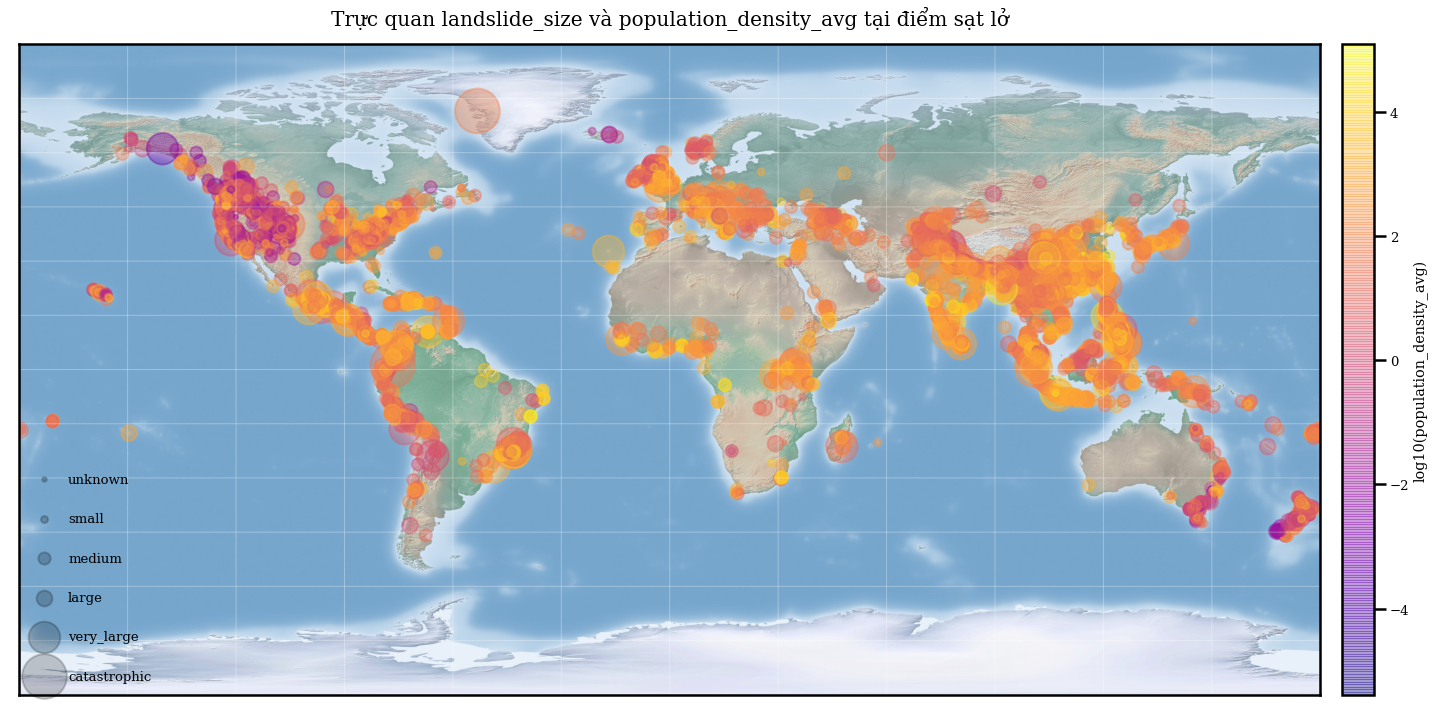
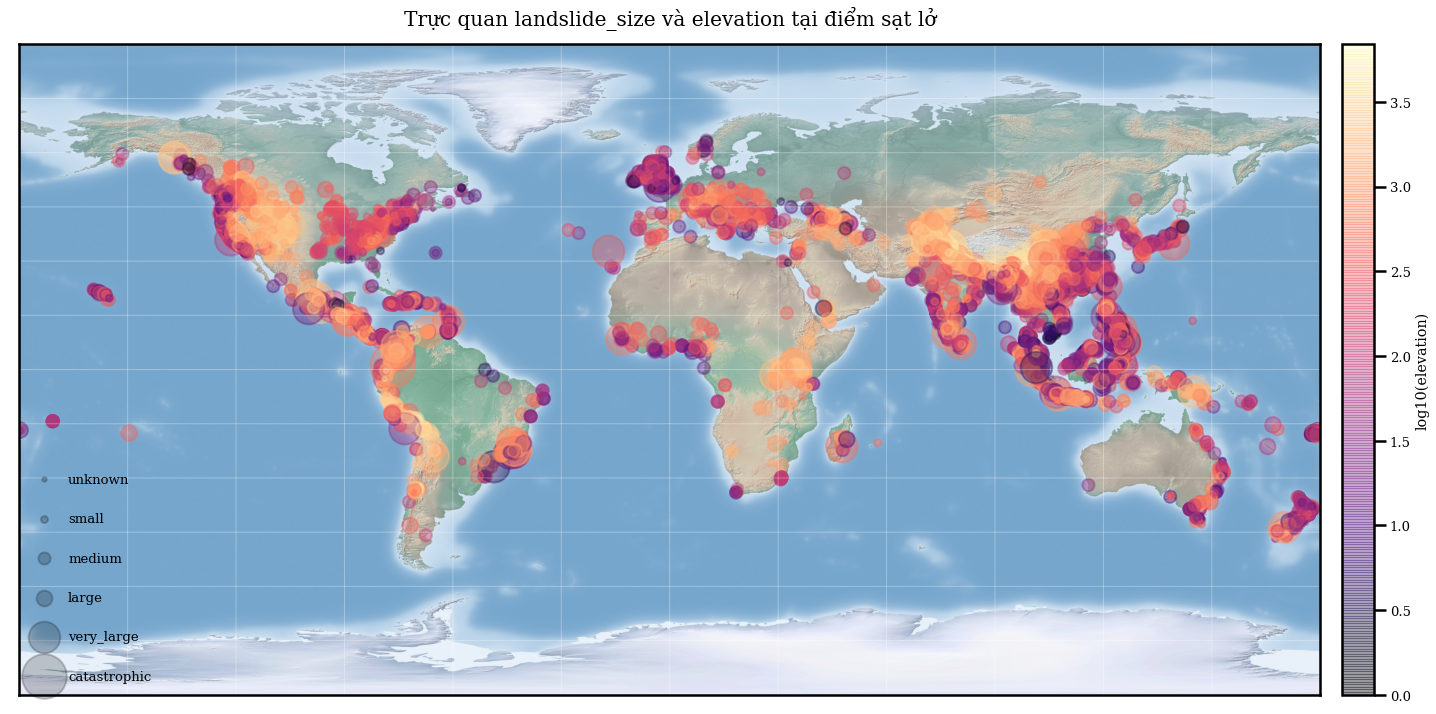


'United State' là quốc gia có nhiều sự kiện sạt lở đất diễn ra nhất, theo sau đó là 'India' và 'Philippines'. ‘India’ và ‘China’ là quốc gia có số lượng người tử vong do sạt lở đất nhiều nhất, ‘Guatemala’ và ‘Colombia’ có số lượng người bị thương do sạt lở đất nhiều nhất. Đặc biệt, ‘United States’ có số lượng sự kiện sạt lở đất nhiều nhất nhưng lại không nằm trong top các quốc gia có số lượng người bị thương/tử vong nhiều vì đa số các sự kiện sạt lở diễn ra ở bờ Tây nước Mỹ - nơi dân cư thưa thớt (bản đồ mục 2.3.4).



‘North America’ có số lượng sự kiện sạt lở nhiều thứ hai sau ‘Asia’ nhưng mức độ ít nghiêm trọng hơn, đa số diễn ra ở quy mô vừa (‘medium’), nhỏ (‘small’) hoặc không xác định (‘unknown’) ở địa hình thấp. Ở ‘Asia’ cho thấy quy mô sạt lở trải đều ở các độ cao địa hình, chủ yếu quy mô vừa và lớn. ‘South America’ và ‘Africa’ cho thấy nhiều điểm sạt lở có quy mô lớn trở lên và trải đều các độ cao địa hình.

**2.3.4. Phân tích trực quan trên bản đồ địa lý**

****

Dựa vào bản đồ kết hợp với kết quả ở mục 2.3.1, môt cách trực quan cho thấy được các điểm sạt lở ở khu vực Đông Nam Á, Đông Á, phía Đông Hoa Kỳ, Đông Úc, Vương quốc Anh, New Zealand xảy ra ở độ cao địa hình thấp, quy mô sạt lở phần lớn ở mức vừa và nhỏ. Xung quanh dãy Himalaya cho thấy quy mô sạt lở tương đối lớn.

Mật độ dân số cũng cho thấy ảnh hưởng nhất định đến mức độ sạt lở ở từng khu vực. Ở bờ Tây Hoa Kỳ, dân cư thưa thớt, sạt lở ít nghiêm trọng ở mức độ vừa và nhỏ. Khu vực Châu Á có mật độ dân số cao, quy mô sạt lở đa số từ mức độ vừa trở lên.

**2.3.5. Tổng kết phân tích thăm dò**

Sau đây là một số insights chúng tôi thu được từ bộ dữ liệu sau khi phân tích thăm dò:

* Kích thước của một sự kiện sạt lở đất hay gặp là *'medium'*, kích thước hiếm gặp nhất là *'very\_large'* và *‘catastrophic’*.
* Loại sạt lở đất hay gặp nhất là *'landslide'* và *'mudslide'*.
* Nguyên nhân gây ra sạt lở đất phần lớn là do mưa, nhiều nhất là *'downpour'*.
* Mùa hè là mùa thường xuyên xảy ra sạt lở đất nhất.
* North America và Asia là hai châu lục thường xuyên xảy ra sạt lở đất nhất.
* Tại United State, sạt lở đất diễn ra nhiều trong giai đoạn từ tháng 12 đến tháng 3, tại India là từ tháng 6 đến tháng 8, tại Vietnam là từ tháng 7 đến tháng 11.
* Các sự kiện sạt lở diễn ra vào mùa mưa bão, có bão tuyết.

Kết luận ảnh hưởng của các thuộc tính đến quy mô sạt lở:

* Nhìn chung, địa hình tại khu vực sạt lở càng cao, quy mô sạt lở càng nghiêm trọng. Do sự khác biệt về độ cao địa hình, nên mức độ sạt lở ở các quốc gia cũng có sự khác nhau.
* Những khu vực có chỉ số nhiệt độ ‘temp’ và ‘feelikes’ cao, chỉ số ‘precipcover’ thấp thì mức độ sạt lở càng lớn.
* Các khu vực độ che phủ thảm thực vật thưa thớt có nguy cơ xảy ra sạt lở với quy mô lớn hơn so với những khu vực có thảm thực vật dày.
* Có sự khác nhau chưa rõ rệt về quy mô sạt lở giữa các châu lục.

*(hình ảnh từ phân tích thăm dò tại* ***phụ lục B****)*

## Feature selection

Vì biến mục tiêu *‘landslide\_size’* là biến định tính, nên chúng tôi áp dụng hai phương pháp để lựa chọn đặc trưng là: *One Way ANOVA* và *Chi-Square Test*.

* + 1. **Lựa chọn biến dạng số (thuộc tính định lượng)**

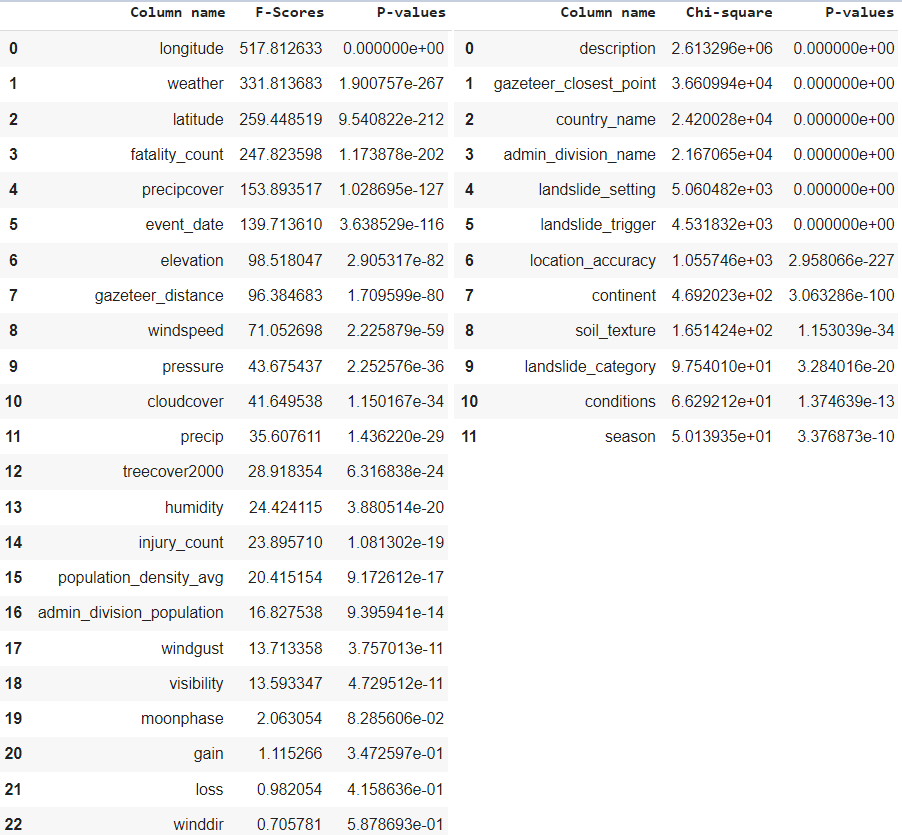
*One Way ANOVA* là phương pháp thống kê đo lường mối quan hệ tương quan giữa biến định lượng và biến định tính, bằng cách so sánh phương sai của biến định lượng theo từng lớp của biến định tính.

Khi áp dụng *One Way ANOVA* để tìm mối quan hệ tương quan giữa các thuộc tính định lượng so với biến mục tiêu *‘landslide\_size’*, kết quả trả về là 2 giá trị F-score và P-value cho mỗi thuộc tính. *F-score càng cao* và *P-value càng thấp* thì thuộc tính đó *càng quan trọng.*

**2.4.2. Lựa chọn biến phân loại (thuộc tính định tính)**

*Chi-Square Test* là phương pháp thống kê đo lường mối quan hệ tương quan giữa hai biến định lượng, bằng cách tính phân phối giữa các giá trị trong hai biến.

Khi áp dụng *Chi-Square Test* để tìm mối quan hệ tương quan giữa các thuộc tính định tính so với biến mục tiêu *‘landslide\_size’*, kết quả trả về là 2 giá trị Chi-square và P-value cho mỗi thuộc tính. *Chi-square càng cao* và *P-value càng thấp* thì thuộc tính đó *càng quan trọng.*

**

*Hình 2. Xếp hạng các thuộc tính theo mức độ quan trọng của*

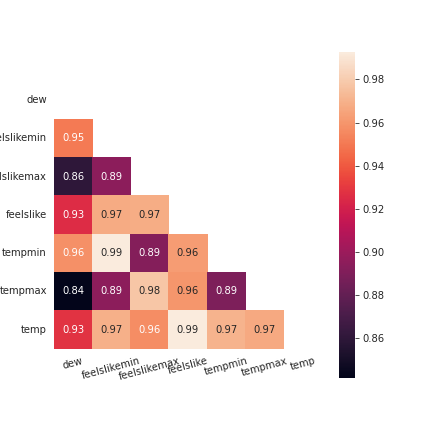
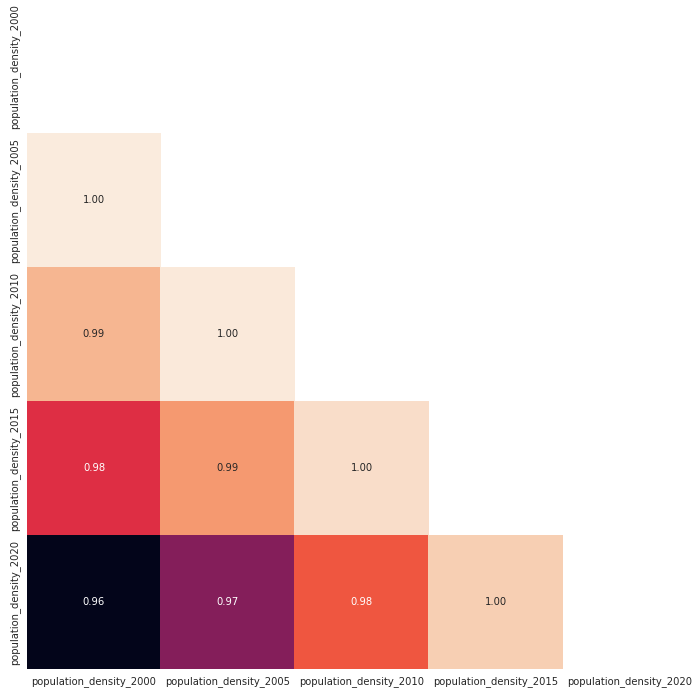
*biến định lượng (trái) và biến định tính (phải).*

Sau khi có được bảng xếp hạng mức độ quan trọng của các thuộc tính, chúng tôi chọn ngưỡng để lựa chọn thuộc tính đưa vào mô hình bằng cách sử dụng những thuộc tính quan trọng đã được tìm thấy sau khi phân tích thăm dò, thử nghiệm thêm và đánh giá nhiều lần để có thể chọn được một ngưỡng phù hợp nhất, chính là 16 biến định lượng và 8 biến định tính đầu tiên trong hai bảng trên.

## Feature engineering

**2.5.1. Kết hợp các nhóm thuộc tính có tương quan cao**

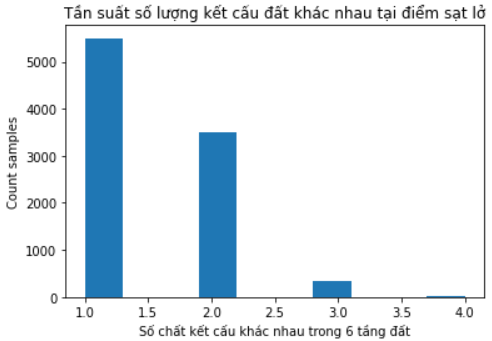
1. **Biến dạng số (thuộc tính định lượng)**

****

Dựa trên ma trận tương quan, ta thấy có 2 nhóm thuộc tính có tương quan rất mạnh lẫn nhau là:

* Nhóm thuộc tính về mật độ dân số: ‘population\_density\_2000’, đến ‘population\_density\_2020’.
* Nhóm thuộc tính về chỉ số thời tiết: 'dew', 'feelslikemin', 'feelslikemax', 'feelslike', 'tempmin', 'tempmax', 'temp'.

Nên chúng tôi kết hợp hai nhóm bằng cách lấy *mean(),* thu được hai thuộc tính mới thay thế là *‘population\_density\_avg’* và *‘dew\_temp\_avg’*.

1. **Biến dạng phân loại (thuộc tính định tính)**

Theo phân tích thăm dò 6 cột về kết cấu của đất theo độ sâu (*'soil\_texture\_0', 'soil\_texture\_10', 'soil\_texture\_30', 'soil\_texture\_60', 'soil\_texture\_100', 'soil\_texture\_200'*), chúng tôi nhận thấy đa số các sample chỉ có từ 1 đến 2 loại đất ở cả 6 tầng đất, nên chúng tôi sử dụng *mode()* để kết hợp 6 cột trở thành một cột duy nhất là *‘soil\_texture’*: chứa loại đất xuất hiện nhiều nhất trong 6 tầng.

**2.5.2. Kết hợp và mã hóa các cột chứa thông tin mô tả (text description)**

* Các cột có chứa thông tin mô tả bao gồm: 'event\_description', 'event\_title', và 'location\_description' .
* Xử lý: kết hợp thành cột duy nhất là ‘description’, sử dụng TF-IDF để mã hóa dữ liệu văn bản về vector dạng số theo unigram và bigram.

**2.5.3. Mã hóa One-hot cho các thuộc tính dạng biến phân loại**

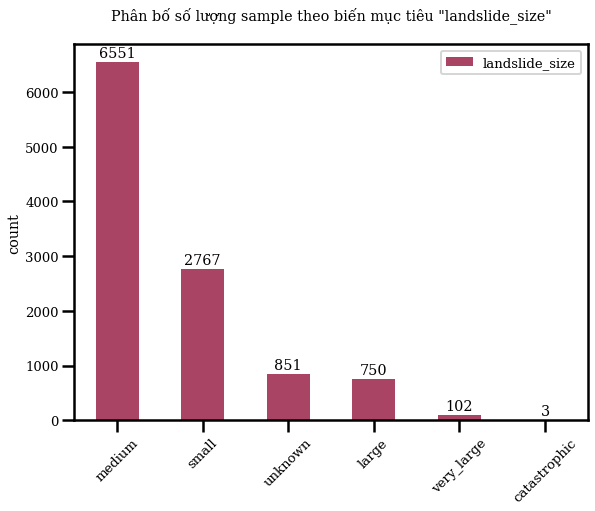
Các cột thuộc tính dạng biến phân loại được chuyển đổi về dạng 0, 1 bằng cách sử dụng hàm *pd.get\_dummies().*

**2.5.4. Feature scaling**

* Áp dụng *MaxAbsScaler()* để chuẩn hóa miền giá trị của tất cả các thuộc tính về phạm vi [-1, 1].
* Chúng tôi chọn *MaxAbsScaler()* do phương pháp này hoạt động tương tự như *MinMaxScaler()* và phù hợp với dữ liệu thưa (sparse data) thu được sau khi mã hóa TF-IDF dữ liệu dạng văn bản. Ngoài ra, nó không làm thay đổi hình dạng của dữ liệu.

## Phát triển mô hình

**2.6.1. Xử lý biến mục tiêu ‘*landslide\_size’***

Dựa trên số lượng các mẫu theo từng lớp của *‘landslide\_size’*, chúng tôi đã xử lý sơ bộ trước khi đưa vào mô hình dự đoán, bao gồm:

- Gộp các giá trị *‘catastrophic’* (thảm khốc) chung với *‘very\_large’*, vì số lượng mẫu có giá trị là *‘catastrophic’* và *‘very\_large’* rất ít (3 và 81 trên tổng 9345 điểm dữ liệu), và chúng có ý nghĩa tương tự nhau.

* Xử lý dữ liệu mất cân bằng

Đa số các sự kiện sạt lở có kích thước vừa đến nhỏ, chỉ có một số ít có kích thước lớn. Để mô hình dự đoán kích thước có thể hoạt động hiệu quả và không bị ảnh hưởng bởi sự mất cân bằng dữ liệu, chúng tôi áp dụng kỹ thuật *ADASYN* để oversampling cho những lớp có ít điểm dữ liệu hơn.

**2.6.2. Lựa chọn mô hình và độ đo đánh giá**

* 4 mô hình phân lớp: Logistic Regression, Support Vector Machine, Random Forest và PassiveAggressiveClassifier.
* Độ đo đánh giá: Macro F1-score và accuracy.

Trong đó chúng tôi tập trung vào *Macro F1-score* để xếp hạng các mô hình bởi vì đây là độ đo không nhạy cảm sự mất cân bằng của các lớp dữ liệu, phù hợp với bài toán mất cân bằng mà chúng tôi đang giải quyết.

**2.6.3. Tinh chỉnh mô hình**

Sau khi thử nghiệm trên 4 loại mô hình, chúng tôi chọn ra mô hình tốt nhất để tiếp tục tối đa hóa kết quả bằng cách sử dụng *GridSearchCV* để tinh chỉnh các hyperparameter cho mô hình tốt nhất.

## Kết quả và đánh giá

|  |  |  |  |
| --- | --- | --- | --- |
| **Phương pháp** | **Mô hình** | **Macro F1-score** | **Accuracy** |
| Trước khi xử lý mất cân bằng | Random Forest | 0.481892 | 0.713623 |
| PassiveAggressiveClassifier | **0.519439** | 0.715407 |
| Logistic Regression | 0.511298 | 0.718973 |
| SVM | 0.472977 | 0.717190 |
| Sau khi xử lý mất cân bằng | Random Forest | 0.498321 | 0.716476 |
| PassiveAggressiveClassifier | 0.520779 | 0.711484 |
| Logistic Regression | **0.521931** | 0.715050 |
| SVM | 0.477818 | 0.706847 |
| **GridSearch** | **Logistic Regression** | **0.522588** | **0.717189** |

# KẾT LUẬN

Sau khi quá trình thu thập, tiền xử lý và phân tích thăm dò, kết quả thu được cho thấy những thuộc tính chúng tôi thu thập thêm (*‘dew\_temp\_avg’, 'fatality\_count', 'injury\_count', 'precipcover', 'elevation', 'treecover2000', 'population\_density\_avg'*) có ảnh hưởng đến biến mục tiêu *‘landslide\_size’* và góp phần quan trọng trong việc phân tích cũng như phát triển mô hình.

Những thuộc tính được lựa chọn ra sau khi phân tích thăm dò được kết hợp với những thuộc tính quan trọng khi kiểm thử bằng *One Way ANOVA* và *Chi-Square*, chúng tôi đã lựa chọn ra được 24 thuộc tính quan trọng từ 67 thuộc tính ban đầu.

Việc áp dụng ADASYN để giải quyết vấn đề mất cân bằng dữ liệu đã giúp cho 4 mô hình đều đạt được độ chính xác cao hơn. Cùng với đó, chúng tôi sử dụng Grid Search để tinh chỉnh mô hình tốt nhất là *Logistic Regression* nhằm nâng cao hiệu suất mô hình, với độ đo *marco F1-score* = 0.522588 và *accuracy* = 0.717189 cao hơn kết quả của mô hình trước khi tinh chỉnh.

**TÀI LIỆU THAM KHẢO**

**[1]** Toward Data Science | Using the Chi-Squared test for feature selection with implementation

<https://towardsdatascience.com/using-the-chi-squared-test-for-feature-selection-with-implementation-b15a4dad93f1>

**[2]** Toward Data Science | ANOVA for Feature Selection in Machine Learning

<https://towardsdatascience.com/anova-for-feature-selection-in-machine-learning-d9305e228476>

**[3]** Geographic Data with Basemap | Python Data Science Handbook

**[4]** The Landslide Blog | The Kedarnath debris flow disaster in Uttarakhand

<https://blogs.agu.org/landslideblog/2013/06/21/the-kedarnath-debris-flow-disaster-in-uttarakhand/>

**PHỤ LỤC PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| **STT** | **Thành viên** | **Nhiệm vụ** |
| 1 | Nguyễn Thị Minh Phương | - Lựa chọn thuộc tính  - Phát triển mô hình  - Viết báo cáo và làm slide |
| 2 | Chu Hà Thảo Ngân | - Tiền xử lý dữ liệu  - Phân tích thăm dò  - Viết báo cáo và làm slide |
| 3 | Thái Minh Triết | - Thu thập dữ liệu  - Phân tích thăm dò  - Viết báo cáo và làm slide |

**PHỤ LỤC**

Link github của đồ án: <https://github.com/minhphuongzzz/DS105-final-project>

1. **Mô tả thuộc tính của bộ dữ liệu thu thập thêm**

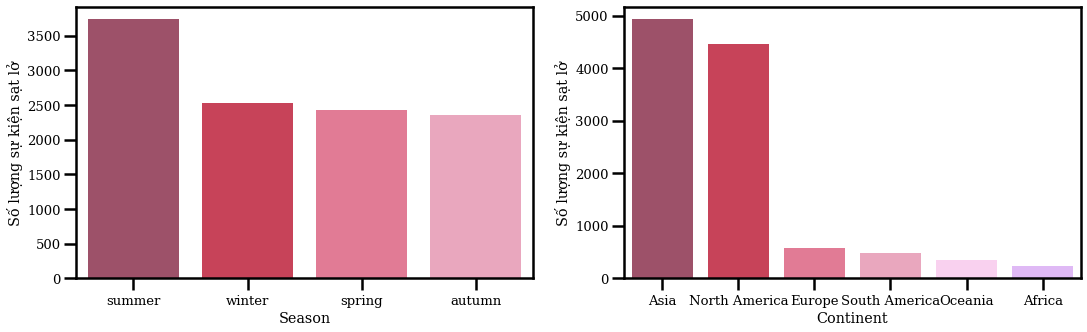
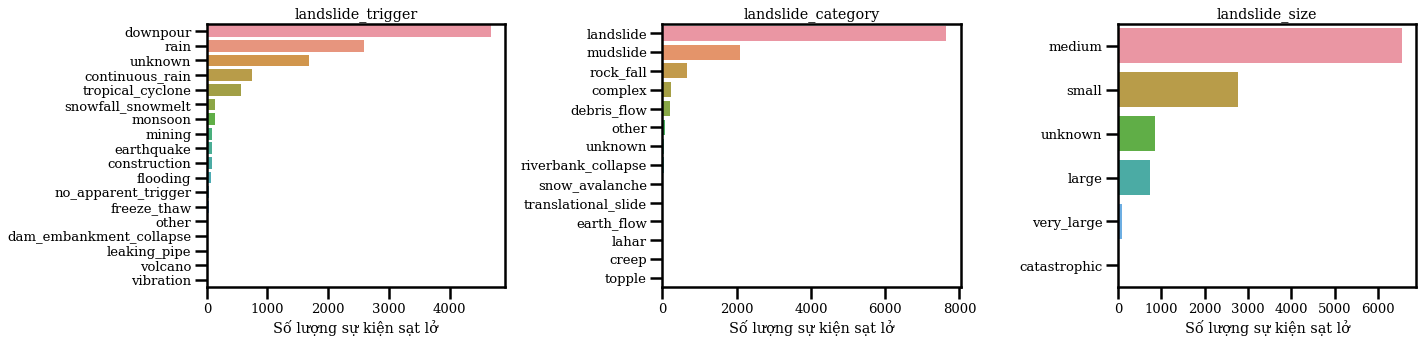
**1.** **Weather**

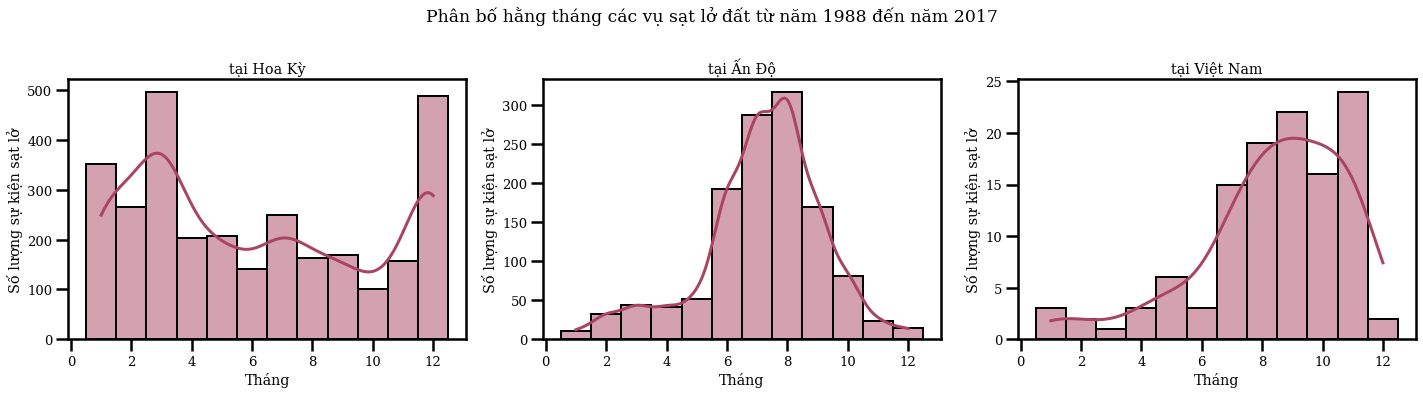
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả** | **Miền giá trị** |
| 32 | tempmax | float64 | Nhiệt độ cao nhất trong ngày tại địa điểm sạt lở | -22.1 to 45.1  (ºC) |
| 33 | tempmin | float64 | Nhiệt độ thấp nhất trong ngày tại địa điểm sạt lở | -27.1 to 32.1  (ºC) |
| 34 | temp | float64 | Nhiệt độ tại địa điểm sạt lở | -24.8 to 38.4  (ºC) |
| 35 | feelslikemax | float64 | Nhiệt độ cảm thấy cao nhất | -22.1 to51.9  (ºC) |
| 36 | feelslikemin | float64 | Nhiệt độ cảm thấy thấp nhất | -32.3 to 41.0  (ºC) |
| 37 | feelslike | float64 | Nhiệt độ cảm thấy | -28.0 to 43.6  (ºC) |
| 38 | dew | float64 | Điểm sương (nhiệt độ hóa sương) | -27.9 to 175.5  (ºC) |
| 39 | humidity | float64 | Độ ẩm tương đối | 14.69 to 100.0 (%) |
| 40 | precip | float64 | Lượng mưa tại thời điểm “datetime” | 0.0 to 498.0  (mm) |
| 41 | precipcover | float64 | Tỉ lệ số giờ trong ngày có lượng mưa khác 0 | 0.0 to 100.0  (%) |
| 42 | windgust | float64 | Gió giật | 0.0 to 316.8  (m/s) |
| 43 | windspeed | float64 | Tốc độ gió trung bình trên 1 phút | 0.0 to 222.1  (m/s) |
| 44 | winddir | float64 | Hướng gió | 0.0 to 360.0  (m/s) |
| 45 | pressure | float64 | Áp suất khí quyển theo mực nước biển hoặc khí áp | 951.8 to 1084.0  (Pa) |
| 46 | cloudcover | float64 | Tỷ lệ bầu trời bị che phủ bởi mây | 0.0 to 100.0  (%) |
| 47 | visibility | float64 | Tầm nhìn xa | 0.0 to 1030.1 (km) |
| 48 | moonphase | float64 | Tỉ lệ hình dạng của mặt trăng theo chu kỳ | 0.0 to 1.0 |
| 49 | conditions | object | Điều kiện thời tiết | Partially cloudy; Rain, Overcast; Snow;… |
| 50 | stations | object | Trạm thời tiết |  |

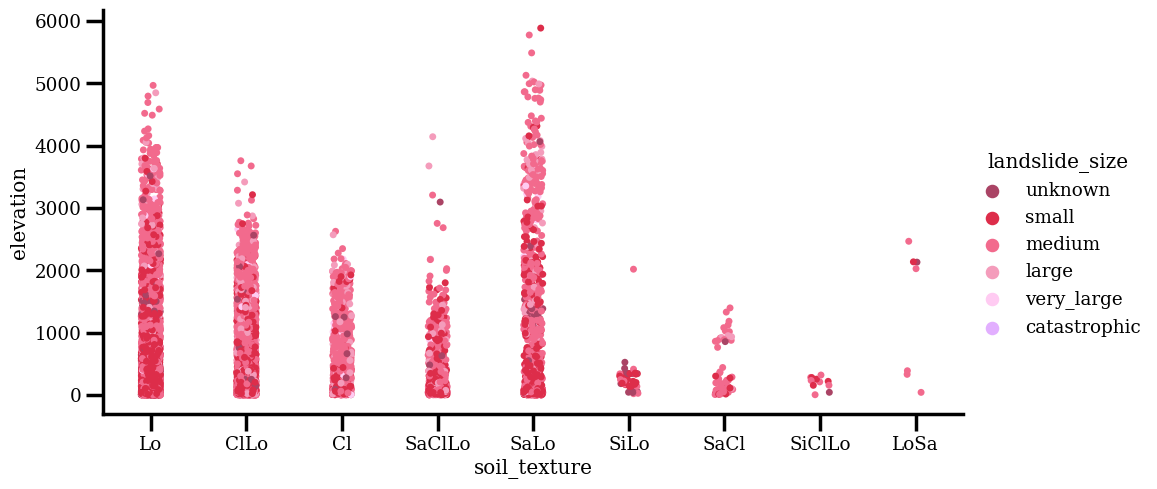
**2.** **Các features khác**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **STT** | **Tên thuộc tính** | **Kiểu dữ liệu** | **Mô tả** | **Miền giá trị** |
| 51 | elevation | int64 | Độ cao so với mực nước biển (mét) | -407.0 to 6916.0 (m) |
| 52 | continent | object | Châu lục | Asia, North America, South America, Africa, Europe, Oceania |
| 53 | season | object | Mùa | spring, summer, autumn, winter |
| 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) | 0 to 100 (%) |
| 55 | loss | int64 | Sự thay đổi từ forest sang non-forest trong khoảng thời gian từ năm 2000 - 2012 | 0 hoặc 1 |
| 56 | gain | int64 | Sự thay đổi toàn bộ từ non-forest sang forest trong khoảng thời gian từ năm 2000 - 2012 | 0 hoặc 1 |
| 57 | soil\_texture\_0 | object | Loại kết cấu đất ở bề mặt | * Lo * ClLo * SaClLo * SaLo * SaCl * SiLo * Cl * LoSa * SiClLo * Sa |
| 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 | 0.0 to 110884.89 (người/km2) |
| 64 | population\_density\_2005 | float64 | Mật độ dân số tại khu vực vào năm 2005 | 0.0 to 116474.90 (người/km2) |
| 65 | population\_density\_2010 | float64 | Mật độ dân số tại khu vực vào năm 2010 | 0.0 to 121829.19 (người/km2) |
| 66 | population\_density\_2015 | float64 | Mật độ dân số tại khu vực vào năm 2015 | 0.0 to 126877.76 (người/km2) |
| 67 | population\_density\_2020 | float64 | Mật độ dân số tại khu vực vào năm 2020 | 0.0 to 130805.85 (người/km2) |

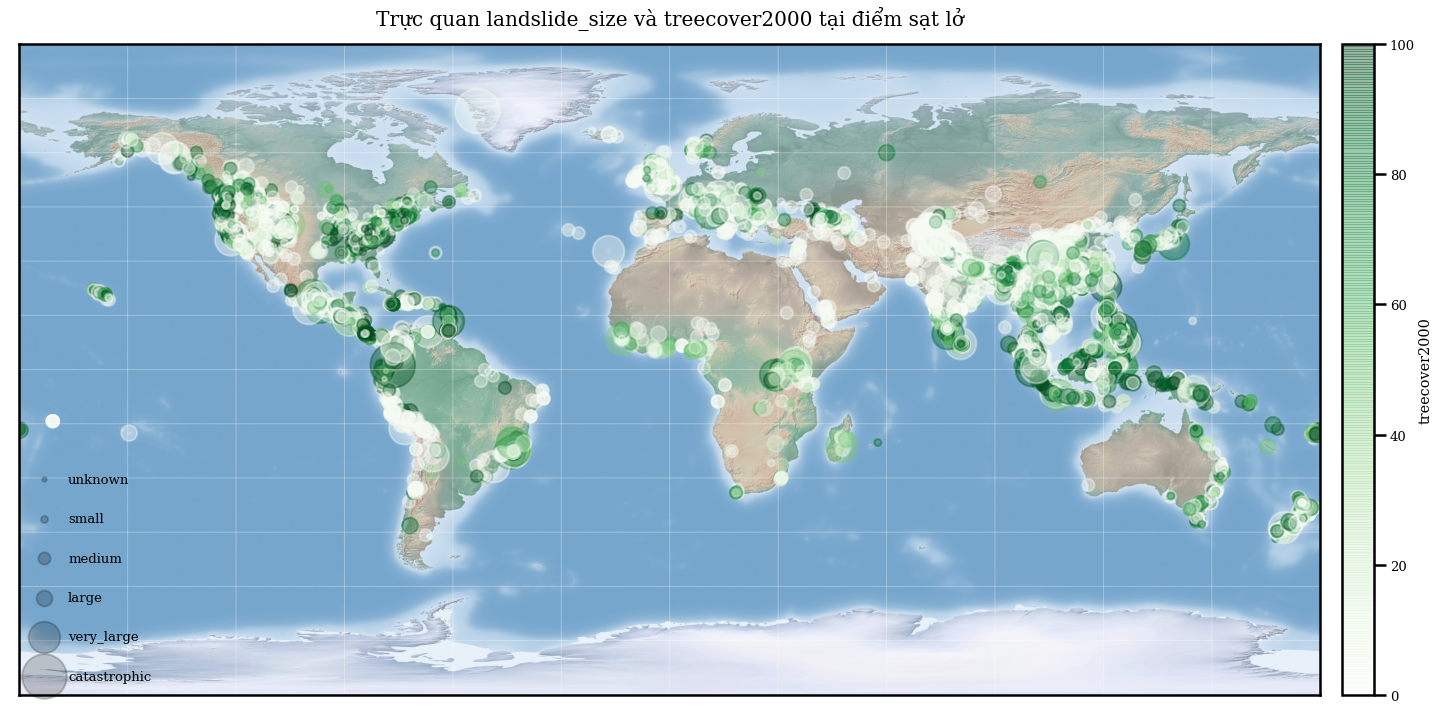
1. **Hình ảnh từ phân tích thăm dò**

*******Phân bố số lượng sample của thuộc tính: landslide\_trigger, landslide\_category, landslide\_size, season, continent*

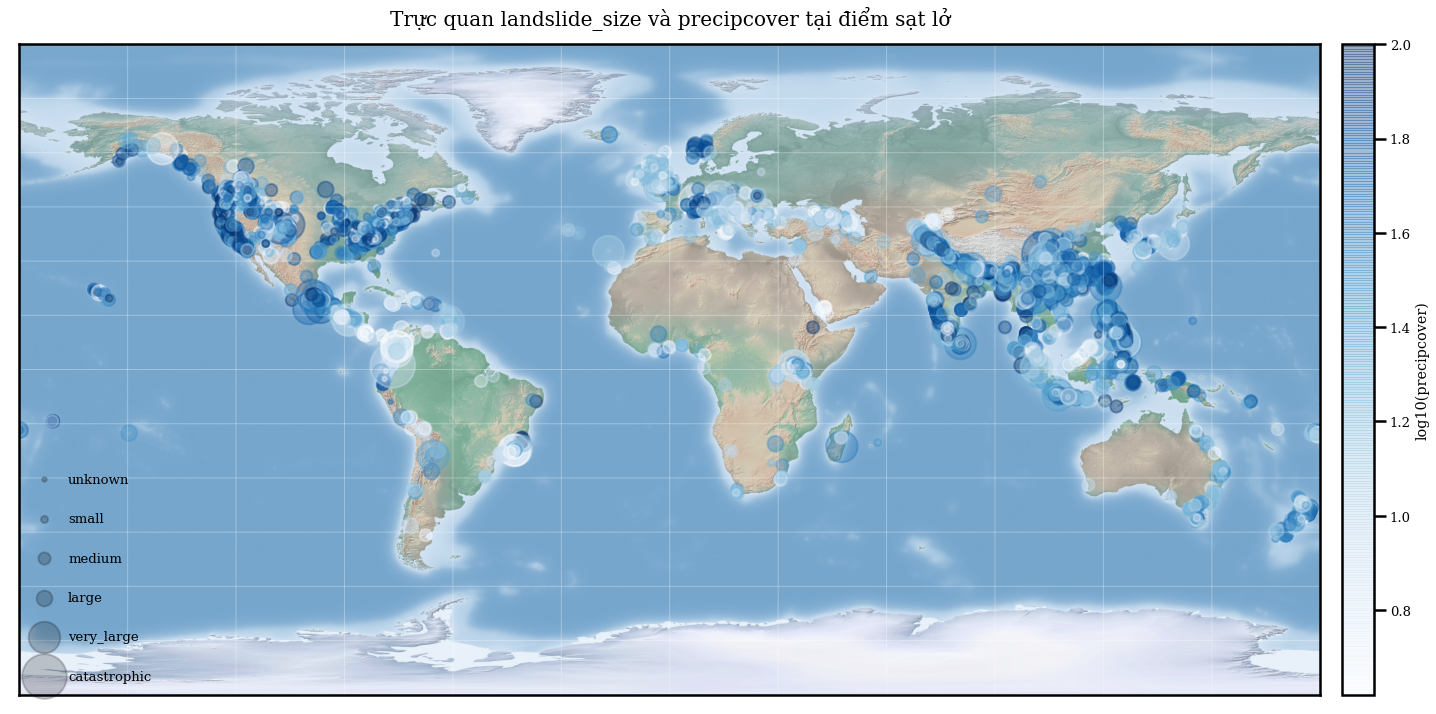
*******Phân bố hằng tháng các vụ sạt lở đất từ năm 1988 đến năm 2017 tại 3 quốc gia: United States, India, Vietnam. Những vụ sạt lở đất diễn ra vào mùa mưa bão, bão tuyết, mưa lớn kéo dài.*

*******Độ cao địa hình và quy mô sạt lở đất theo từng loại kết cấu đất*

*******Phân bố các địa điểm xảy ra sạt lở ở từng châu lục từ năm 1988 đến năm 2017*

****

*Trực quan mức độ sạt lở và độ che phủ rừng tại điểm sạt lở*

*******Trực quan mức độ sạt lở và lượng mưa tại điểm sạt lở*

*Trực quan mức độ sạt lở và nguyên nhân gây sạt lở********Wordcloud những từ xuất hiện nhiều ở các nhãn mức độ sạt lở đất (target) dựa trên 'event\_description'*