Understanding Landslide Hazards: Analysis, Insights, and Implications for Disaster Management and Resilience

Report

marvelous-echidna

Introduction

Landslides represent a significant natural hazard that, despite their devastating impact on global communities, often remain under the radar. Driven by a need to understand the patterns of these occurrences and the factors that contribute to their severity, our research addresses critical questions regarding regional susceptibility, the human consequences of these disasters, and potential trends linked to broader environmental changes, such as climate change. Our initial analysis has uncovered a geographical pattern in the frequency and impact of landslides, laying the groundwork for more in-depth exploration.

This study delves into the relationship between the size of landslides and the fatalities they cause, and assesses the effectiveness of a logistic regression model in predicting the size of landslides. Our results show that larger landslides tend to be associated with higher fatality rates, and the model achieved a ROC AUC score of 0.689, indicating moderate predictive capability. Additionally, a map illustrating the geographic distribution of landslides highlights regions that are particularly at risk, providing valuable information for disaster preparedness and mitigation efforts. Moreover, our models that predict the number of casualties offer crucial insights for emergency response planning, particularly in areas like Ithaca that are prone to such events. These findings emphasize the necessity of improving prediction models and strengthening regional disaster response strategies to better manage landslide risks amid changing environmental conditions.

Data description

Our analysis utilizes NASA's Global Landslide Catalog, which compiles landslide incidents worldwide from 2007 to 2019. The inception of this data set was motivated by the need for a unified source of data to enhance landslide hazard modeling and risk assessment. The data set meticulously catalogs events gathered from news reports, scientific articles, and eyewitness accounts and aims to support the scientific community in understanding and mitigating landslide risks.

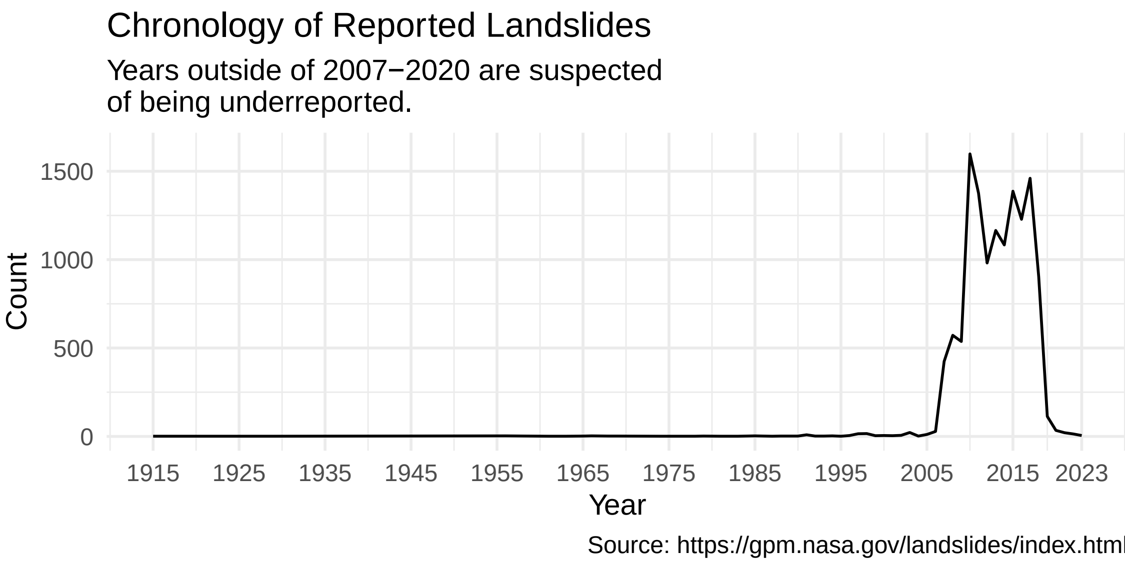
Overview and Funding: Each record in this dataset details a single landslide event, providing information on the date, time, location, trigger, type, and extent of the area affected, as well as the human repercussions in terms of injuries and fatalities. These attributes afford researchers like us a comprehensive view of each incident's context and impact. The catalog was developed by NASA, aligning with its mission to study Earth surface changes and their implications for human populations. This project reflects the agency's commitment to leveraging its technological and scientific resources to foster a safer, more informed world.

Data Collection: The data collection process for this study integrated a variety of sources including news reports, eyewitness accounts, and scientific analyses. This amalgamation enriched the dataset, albeit introducing complexities due to the varying degrees of visibility of each reported event. Factors such as the accessibility of the affected areas, the level of media attention, and the availability of local reporting mechanisms significantly influenced which events were recorded and how comprehensively. These elements introduce a layer of variability that must be considered when assessing the accuracy and thoroughness of the dataset.

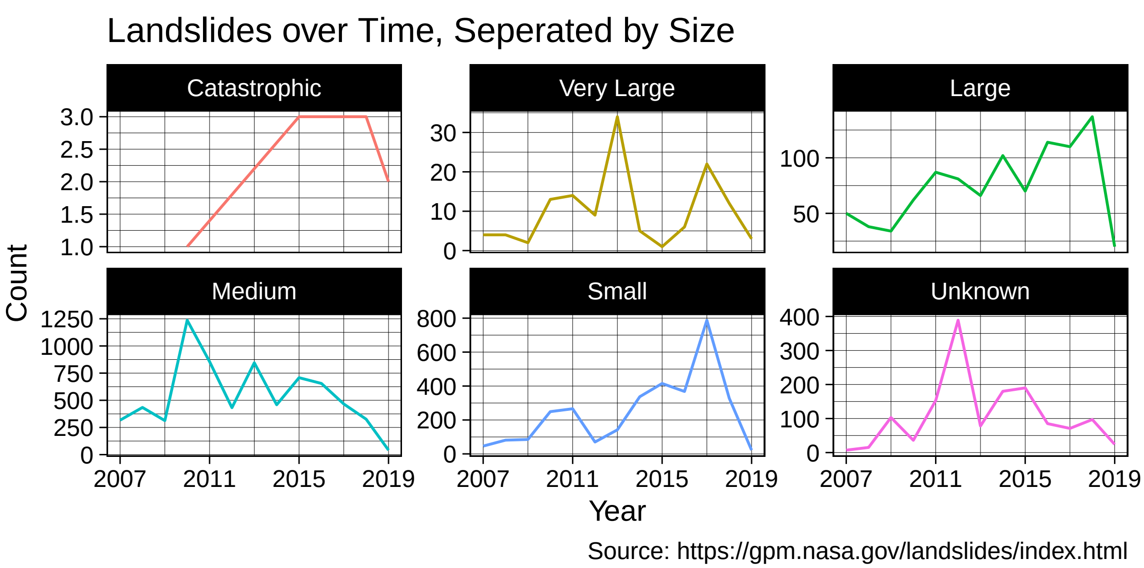
Preprocessing Steps: Before we could analyze the data, it went through a thorough preprocessing phase to enhance its uniformity and reliability. This included refining the accuracy of geographical coordinates and restructuring date and time information into a more usable format. We also standardized these formats across the dataset to facilitate temporal analyses and integrated categorical standards to classify the characteristics of each landslide. These critical steps were necessary to prepare the dataset for detailed analysis and to enable its

Ethical Considerations: It's important to note that the dataset includes sensitive information regarding human casualties. While the individuals involved did not directly participate in the data collection, the use of this data is intended to serve the public good by enhancing our understanding of landslides and facilitate improved disaster response strategies.

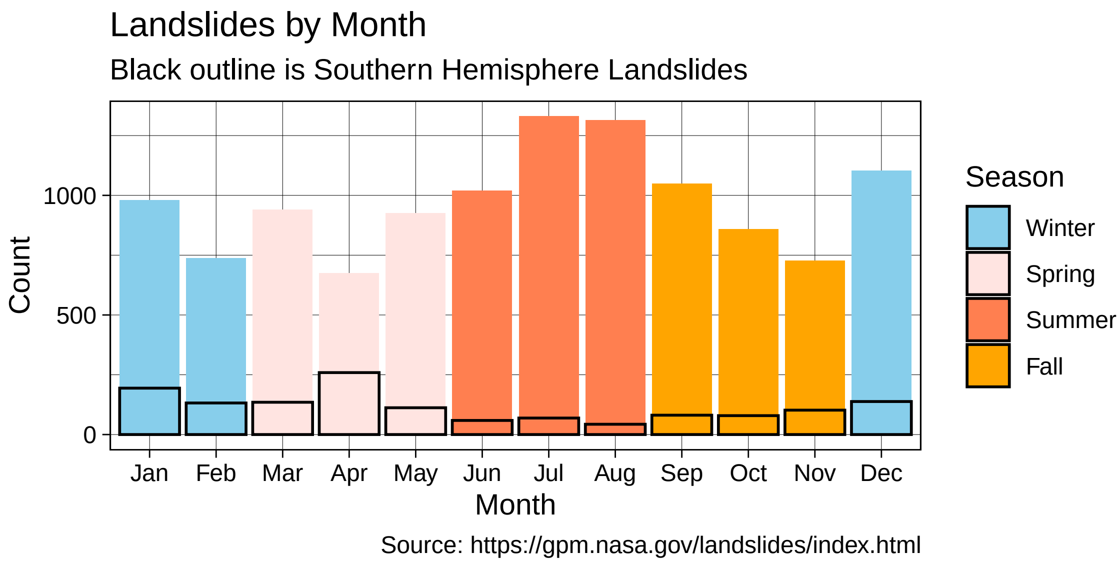
Data analysis



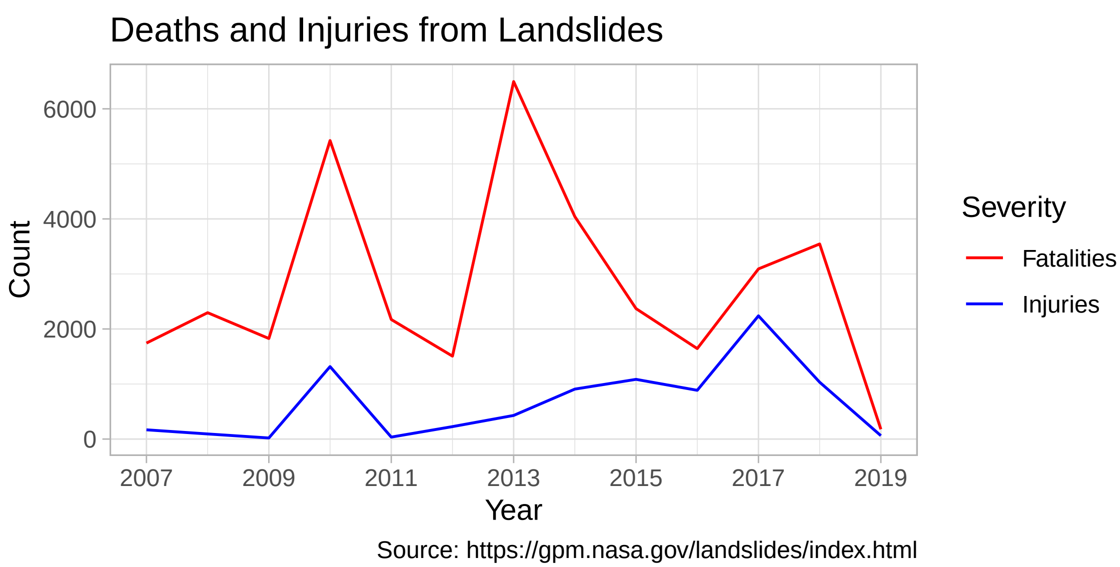
In our analyses involving time, we only included data between 2007 and 2019. This is because there is an unrealistically few number of landslides recorded outside of this range, and this is most likely due to a lack of reporting rather than a lack of landslides occurring.



From the graphs we can see that the number of small and large landslides has increased over time, while the number of medium landslides has decreased. This could potentially be due to a change in how landslides are classified, and not necessarily a change in actual landslide events. The number of landslides each year appears to fluctuate greatly and does not follow any apparent pattern.



Landslides appear to be seasonal, reaching maximum in the winter and summer seasons and minimum in the fall and spring seasons. This indicates that there are seasonal factors, mainly weather, that contribute to the risk of landslides

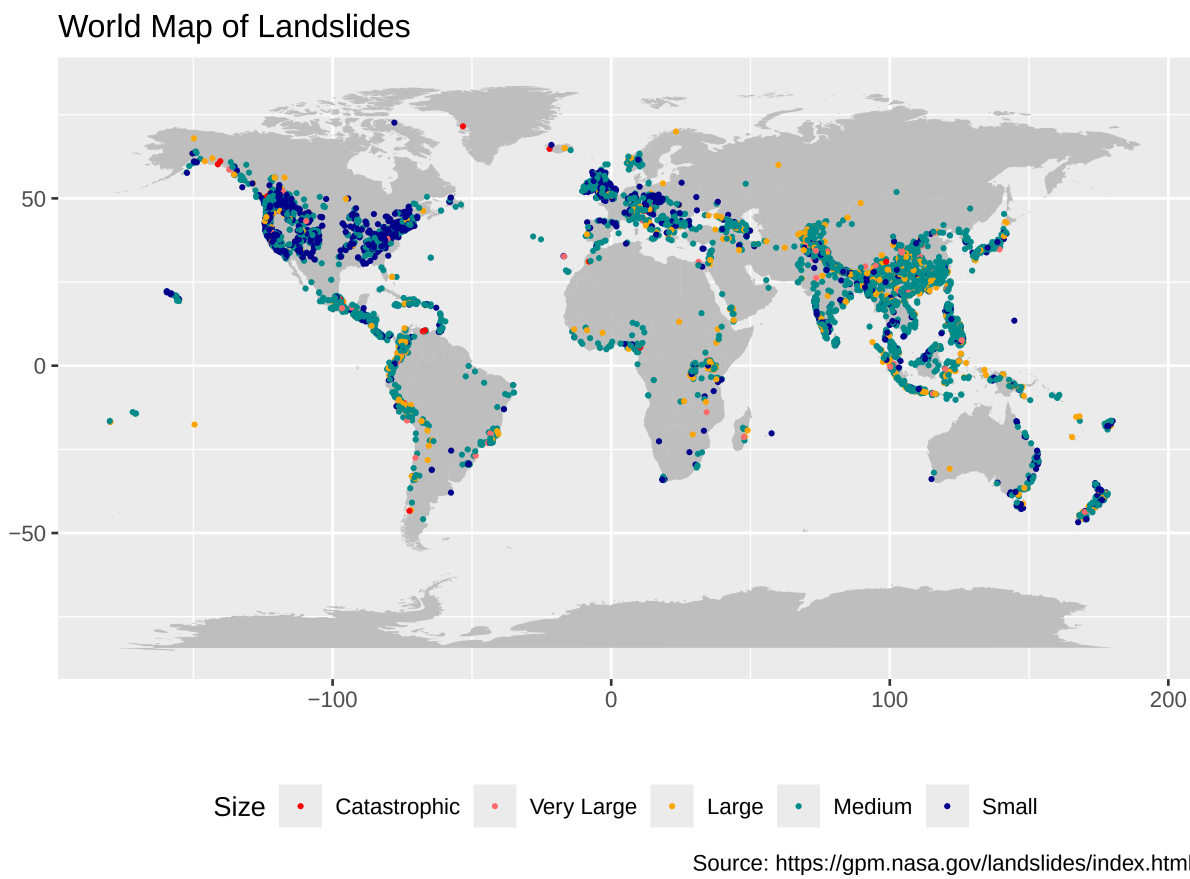


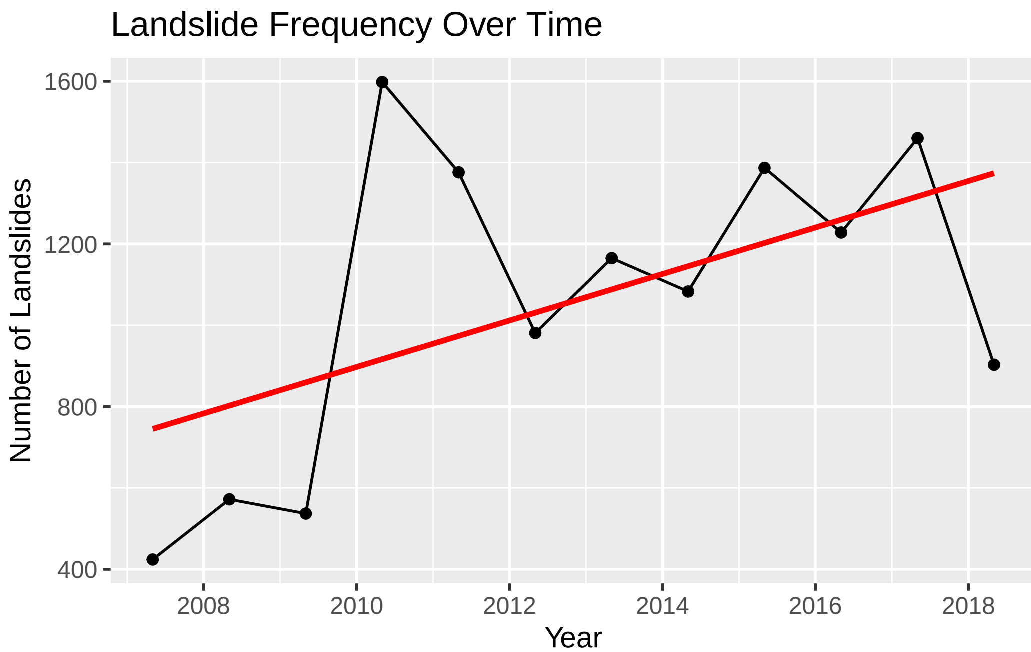
Deaths and injuries from landslides fluctuates greatly over time, with no apparent pattern.

```
# A tibble: 1 x 12
  r.squared adj.r.squared sigma statistic   p.value    df  logLik     AIC    BIC
      <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>   <dbl>  <dbl>
1    0.0692        0.0686  59.3      112. 3.71e-159     7 -58104. 116227. 1.16e5
# i 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

```
# A tibble: 5 x 2
    .pred landslide_size
    <dbl> <chr>
1  21.4   catastrophic
2 160.    very_large
3   9.43  large
4   0.276 medium
5  -0.179 small
```

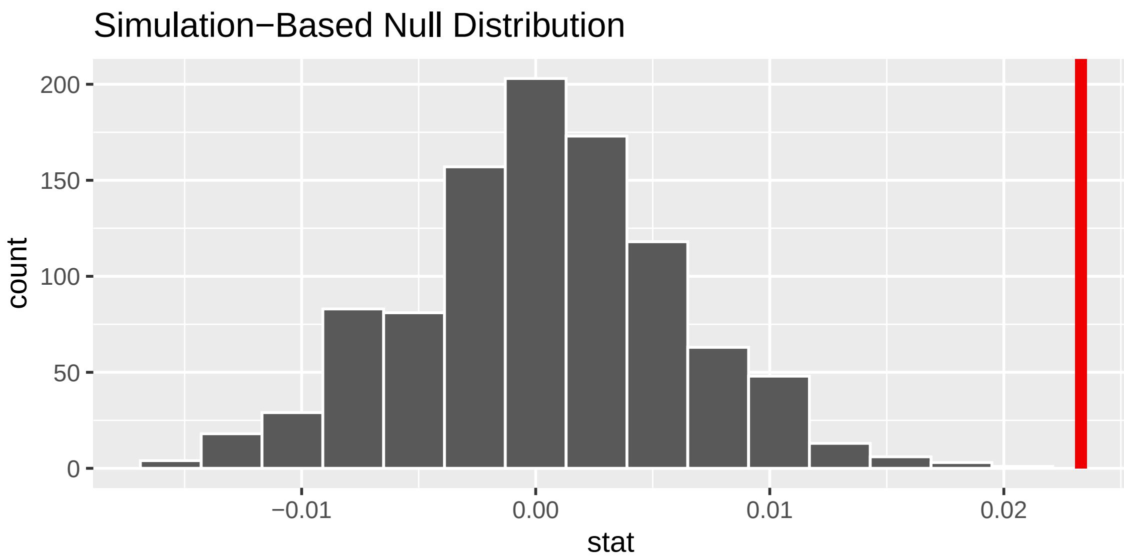
From our linear regression model, we can predict how many casualties we could expect to have if a landslide were to occur in Ithaca this year. The model predicts that we would have very few (if any) casualties if a small or medium landslide were to occur, approximately 10 casualties if a large landslide were to occur, and almost 160 casualties if a very large landslide were to occur. The model predicts that there would only be 21 casualties for a catastrophic landslide, however this is unrealistic and is likely due to the fact that the database has very few catastrophic landslides for our model to use.






```
# A tibble: 10 x 2
   admin_division_name   avg_fatality
   <chr>                        <dbl>
 1 Tolima                       4201.
 2 Badakhshān                    719
 3 Friuli-Venezia Giulia         667
 4 Sulawesi Tengah               226.
 5 Western Area                  197.
 6 Putumayo                      164.
 7 Junín                         154.
 8 Albay                         142.
 9 Vargas                        124
10 Kābul                         114
```

Evaluation of significance



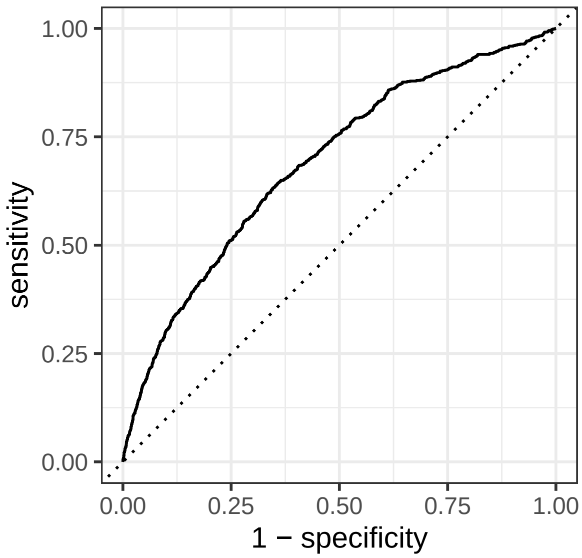

```
# A tibble: 1 x 1
  p_value
    <dbl>
1       0
```

$$H_0 : \hat{p}_{big}^{summer} = \hat{p}_{big}^{other}$$

$$H_A : \hat{p}_{big}^{summer} > \hat{p}_{big}^{other}$$

To find if the proportion of large, very large, and catastrophic landslides is different in the summer than it is in other months, we conducted a hypothesis test. We generated a null distribution of 1000 samples by permuting whether the event was in summer or not (in their respective hemispheres). We calculate a p value of 0, which surpasses a significance level of 0.05. Therefore, reject the null hypothesis that the proportion of big landslides is not different in summer than in other months in favor of the alternative hypothesis that big landslides are more likely in summer months.

```
# A tibble: 1 x 3
  .metric .estimator .estimate
  <chr>   <chr>          <dbl>
1 roc_auc binary         0.689
```



We utilized a logistic regression model to predict the severity of a landslide. This could be used to estimate the risk of landslides in a particular area or for a particular time. As shown in the ROC curve, our model does not perform very well, but performs somewhat okay with a AUC value of around 0.689.

Interpretation and conclusions

Based on the data analysis and visualizations presented, I can make the following key observations:

1. The landslide size categories show distinct patterns over the years analyzed (2007-2019). The "catastrophic" category exhibits high peaks in certain years, indicating the occurrence of severe landslide events. The "very_large" and "large" categories also display peaks, though less extreme than "catastrophic." The number of catastrophic landslides is also less than those of the other landslide sizes. The "large" category in particular seems to have been rising throughout the years, which might imply that the number of large landslides is getting more common. The "medium" and "small" categories seem to fluctuate more regularly with such pronounced spikes as well, and also having more occurrences of landslides at that level in general compared to the other categories. The "medium" category has been decreasing (however still more than the large category), but the "small" category has been rising as well, indicating that smaller landslides are happening more often as well.
2. The fatality and injury counts associated with landslides representation shows that there is a correlation between the spikes in fatality counts and the years when "catastrophic", "very large" and "medium" landslide events occurred, as seen in the size analysis graphs.

This suggests that those categories of landslides might have resulted in the increased in the spike of fatalities. The graph also tells us that the fatalities count exceeds that of the injury count, which implies that fatalities are more common than injuries occurring.

3. The world map displays the geographic distribution of landslide events, color-coded by their size category. This visualization highlights the regions most prone to landslides of varying severities during the analyzed period.

4. Our ML model predict whether a landslide is "big" or "little" based on various predictor variables. Given the ROC AUC score of 0.689 indicates that the model's performance is better than random chance, but there is still room for improvement. Further analysis, such as comparing this score with other models or assessing additional performance metrics, may provide a more comprehensive understanding of the model's predictive capabilities.

5. The linear regression model seeks to predict how many casualties we could expect to have if a landslide were to occur in Ithaca this year. The model predicts that we would have very few (if any) casualties if a small or medium landslide were to occur, approximately 10 casualties if a large landslide were to occur, and almost 160 casualties if a very large landslide were to occur. The model predicts that there would only be 21 casualties for a catastrophic landslide, however this is unrealistic and is likely due to the fact that the database has very few catastrophic landslides for our model to use. Using that hypothesis we create the null hypothesis that there is no significant association between the categorical variable "landslide size", the variables longitude, and latitude and fatality counts. Because we got a low p-value, we are able to reject the null hypothesis in favor of the alternative hypothesis. Furthermore, with the regression model, we see that very_large landslides are definitiely more common.

6. To try figuring whether certain geographic regions or administrative divisions are more prone to experiencing deadlier landslides compared to others we created a chart, we were able to see that Tolima had the highest amount of fatality from landslides, which means that location and landslide sizes also play a role in the fatality rates of the people there, which goes to show that geography ranges between the landslides as well. This is also shown by the previous graphs, especially the map, as well.

7. The Landslide Frequency Over Time graph seeks to answer our other hypothesis where The frequency of landslides has increased over the time period covered in the dataset. It is shown through our least squares line that there is a steady increase in landslide frequency over time.

In the context of real-life applications, this analysis can inform disaster preparedness and mitigation efforts. Regions identified as hotspots for severe landslides may benefit from targeted risk assessment, early warning systems, and infrastructure reinforcement. The correlation between catastrophic landslides and high fatality rates underscores the importance of effective emergency response and evacuation plans in vulnerable areas.

Furthermore, understanding the temporal patterns and potential triggers (e.g., precipitation, seismic activity) of landslides could aid in predictive modeling and risk management strategies. Continuous monitoring and data collection would further strengthen these efforts, enabling more robust analysis and informed decision-making.

The analysis of landslide frequency over time provides valuable information for urban planners and policymakers to assess the changing risk landscape and incorporate resilience measures into urban development projects. This includes zoning regulations, land use planning, and infrastructure design that take into account landslide-prone areas and potential impacts. Communicating the findings of these analysis to the public and engaging communities in disaster preparedness and response efforts are crucial steps in enhancing public safety and resilience. Providing accessible information on landslide risks, mitigation measures, and emergency protocols empowers individuals and communities to take proactive steps to protect themselves and their properties.

Limitations

While the data and visualizations provide valuable insights, it's important to note that the analysis covers a relatively short time frame (2007-2019). Additionally, the geographic coverage of the data may be skewed toward certain regions, either due to reporting biases or the actual distribution of landslide events. This could lead to an over-representation or under-representation of specific areas, affecting the conclusions regarding regional differences in landslide severity. Longer-term trends and patterns may not be fully captured. Additionally, the accuracy and completeness of the data could be influenced by factors such as reporting methods and data collection practices in different regions. Underreporting or missing data from certain areas or events could lead to an incomplete picture of the global landslide situation.

While the conclusions drawn from this analysis are insightful, it's essential to approach them with appropriate caution and contextualization. Factors such as data quality, completeness, and potential biases should be carefully considered and addressed through further research and cross-validation with other relevant data sources. The analyses also do not account for socioeconomic and demographic factors that could influence the fatality counts and the impact of landslides. For example, population density, urbanization levels, infrastructure quality, and access to emergency services could vary significantly across different regions, affecting the vulnerability and resilience of communities to landslide events.

Acknowledgments

Kirschbaum, D.B., Adler, R., Hong, Y. *et al.* A global landslide catalog for hazard applications: method, results, and limitations. *Nat Hazards* **52**, 561–575 (2010). https://doi.org/10.1007/s11069-009-9401-4