What is the best place to survive an apocalypse?

ELGUINDY Abdelrahman (21102968)
ELHUSSIENY Habiba (21105223)

May 2025

Project Report
DAC Master's Program
Sorbonne University

Contents

1 Introduction

In an increasingly unstable world marked by environmental crises, pandemics, and geopolitical conflicts, we begin to ask ourselves, If the world were to unravel, where would we want to be? With the rising of climate threats and increasing pressure on global systems, the idea of an "apocalypse" is becoming more familiar. In this project, we define an apocalyptic scenario as: a future characterized by increased environmental instability and a fracturing of global/national systems, leading to heightened local risks, resource scarcities, and potential for societal disruption.

Given this context, one critical question arises: *What is the best place to survive an apocalypse?*

To address this question, we propose the construction of a quantitative composite measure *the Survivability Index* designed to evaluate and compare the resilience of states and regions worldwide. This index integrates a variety of indicators, encompassing environmental, demographic, infrastructure, and health-related dimensions, in order to provide a comprehensive assessment of each state's ability to withstand and adapt to systemic crises.

This report presents the full data science pipeline used to build this index. It begins with a definition of our objectives and goals, followed by a discussion of data collection, cleaning, and the selection of relevant indicators (Section 3). Section 4 covers exploratory analysis and dimensionality reduction techniques. Section 5 outlines the methodology for index construction and presents the predictive model developed for feature importance and score estimation. Section 6 introduces the interactive visualization tool used to explore the results. Finally, Section 7 discusses the key insights and limitations of the project. Through this report, we aim to contribute a transparent and exploratory approach to evaluating regional survivability under hypothetical global collapse scenarios.

2 Objectives and Goals

The core motivation of this study is to anticipate the different states/regions that would offer the highest likelihood of survival in the face of systemic collapse or instability. To achieve this, our first objective is to define survivability in quantitative terms. This involves constructing a composite index based on measurable, real-world indicators such as disaster frequency, infrastructure quality, and healthcare capacity. A second goal is to conduct a global evaluation and ranking of states across the world based on their potential to resist disruptions.

To ensure the findings are accessible and useful, the project also aims to create a public-facing, interactive visualization platform. This tool will allow users, whether individuals, policymakers, or organizations, to explore regional survivability scores.

The outcomes of this project are intended to benefit a diverse audience. Primary beneficiaries include international organizations, environmental and humanitarian NGOs, and academic researchers focused on climate adaptation and disaster risk reduction.To inform them on where to prioritize efforts and where populations may be most vulnerable in the face of global crises. Additionally, there is also the general public, especially those interested in survival strategies, sustainability, and future planning. This project empowers individuals at multiple levels to make decisions about where to allocate resources, implement interventions, or even consider relocation in response to increasing global instability.

3 Data collection and preparation

The construction of the Survivability Index relies on a broad and diverse set of indicators that reflect the resilience and vulnerability of regions in the context of an apocalypse-like scenario. Our goal is to provide a multidimensional view of what constitutes survivability under systemic stress.

3.1 Data Sources

The data used in this project was aggregated from a diverse range of sources to construct a robust and comprehensive dataset. These sources include:

- **Pre-existing Structured Files**: A foundation of the dataset was built using CSV, TSV, and Excel files. These contained essential geographic and disaster-related information such as global city listings (worldcities.csv), historical tsunami, earthquake, and volcanic eruption records (volcano-events.tsv), and flood incidents (FloodArchive.xlsx).
- **Web Scraping**: Several public websites were scraped to collect region-specific data that was not available in bulk formats. Wikipedia was a key source for demographic (population), geographic (area, elevation), and climatic (average temperature) statistics at the state level. Additional specialized websites were used to collect indexes such as access to drinking water (atlasocio.com), the Food Security Index (impact.economist.com), the Global Peace Index (visionofhumanity.org), as well as the locations of volcanoes (civitatis.com) and nuclear power plants.
- **Application Programming Interfaces (APIs)**: To fill in missing or inconsistent data, APIs were leveraged. The Nominatim API provided geolocation data (latitude and longitude), and Open-Meteo was used to retrieve missing elevation values. Moreover, due to anti-scraping protections on certain sites, asynchronous programming (`asyncio`) was used to efficiently send and handle multiple concurrent requests.
- **Institutional Databases**: Several trusted institutional datasets were incorporated to cover broader global indicators. The World Bank provided data on the percentage of arable and agricultural land, healthcare infrastructure (number of physicians and hospital beds per 1,000 people), access to electricity, and causes of death. The Ecological Threat Index was sourced from visionofhumanity.org.
- **Geographic Matching via KDTree**: In several cases, disaster event datasets referenced geographic regions that no longer exist or had ambiguous names. To resolve this and accurately link each event (volcanic eruption, earthquake, tsunami) to a current administrative region, we implemented a `KDTree-based nearest neighbor search` (learned from the UE SAM). This allowed us to assign each event to its nearest state or province based on geospatial proximity, thus enabling us to compute reliable per-state counts of natural disaster events.

3.2 Variables Selection

To address the complexity of "survivability," we defined four fundamental pillars. These pillars represent distinct yet interconnected dimensions contributing to a region's ability to withstand

and adapt in a major crisis. Each pillar includes a carefully selected set of indicators grounded in theory and data availability at a global or sub-national level. This pillar structure was iteratively refined during the project development phase.

Pillar 1: Environmental Stability This pillar assesses the fundamental characteristics of a region's physical and climatic environment, favoring moderate and stable conditions. It considers indicators such as average annual temperature (in Celsius), elevation (in meter), and rainy days percentage and lastly the humidity percentage.

Pillar 2: Hazard Exposure This pillar quantifies a region's exposure to various types of major natural hazards (geological, climatic) and technological risks. Low exposure to these risks is considered beneficial. Indicators include distance to dangerous volcanoes and nuclear power plants (higher is better), frequency of earthquakes, tsunamis, and volcanic eruptions (lower is better), flood severity, and an ecological threat index (lower is better).

Pillar 3: Resource Availability & Infrastructure This pillar evaluates a region's capacity to meet the basic needs of its population in terms of natural resources and essential infrastructure. It includes indicators such as the percentage of arable and agricultural land, access to clean water (in percentage), average annual precipitation (in mm), access to electricity (in percentage), food security index and, the percentage of deaths due to communicable diseases (as a result for deficient basic sanitation and healthcare).

Pillar 4: Societal Stability & Health Systems This pillar focuses on aspects of social cohesion, demographic pressure, and the robustness of health systems. Indicators include population density, a peace index (lower is better), the number of physicians per 1000 people, and the number of hospital beds per 1000 people.

3.3 Data Cleaning

The collected raw data required significant cleaning to ensure consistency, comparability, and suitability for analysis. The main steps were:

- **Cleaning and Harmonization**: Unit conversions (e.g., feet to meters, square miles to km^2), translation of country/region names, standardization of data formats.
- **Manually Consolidated Data & Feature Engineering**: Certain data points required manual or semi-automated processing, particularly for custom features. For instance, the *minimum distance to the nearest volcano* and *minimum distance to the nearest nuclear power plant* were calculated using the Haversine formula based on geospatial coordinates. Similarly, *population density* was derived by dividing population by area.
- **Handling Missing Values**: Several strategies were employed:
  - Use of APIs to fill gaps (e.g., coordinates, elevation).
  - Missing values were imputed using a tiered approach: first, by computing the mean or median at the country level. If still unavailable, values were imputed using continental averages.

  These strategies were necessary due to the presence of `recording bias` in our data sources, meaning that some countries or regions systematically lacked data, not because the value

was zero, but because it was never collected or reported. This is particularly common in developing regions where data infrastructure is weaker or where certain indicators (like the number of hospital beds or clean water access) are not tracked at the sub-national level.

- **Variable Elimination**: Variables with high proportions of missing data and limited coverage—such as the *food security index*—were removed to avoid introducing bias or incorrect imputation.
- **Validation of Data Consistency**: Checks were performed to identify and correct inconsistencies, such as:
  - Ensuring percentage-based indicators did not exceed logical bounds (e.g., values between 0 and 100).
  - Flagging and correcting implausible values (e.g., negative population counts, extreme temperature values).
- **Outlier Treatment**: Potential outliers were identified using statistical techniques (e.g., Z-score) and visual inspection. These anomalies were mainly caused by errors during scraping (for example the population variable in which had values equal to zero). In these cases, country-level medians were used to impute corrected values. However, we were cautious of *group attribution bias*, where characteristics or statistics from the broader group (e.g., country-level) are incorrectly assumed to apply to individual states. While imputing with country-level medians helped with consistency, it also risked reinforcing false assumptions about regional uniformity. To mitigate this, we made sure that the imputed values were accurate. .

4 Data Analysis and Visualization

4.1 Exploratory Variable Relationships

This subsection aims to understand how the selected variables interact and behave statistically before conducting deeper analyses.

The first step involved computing *descriptive statistics* (mean, median, standard deviation, etc.) for all quantitative variables. This provided an initial overview of central tendencies and variability across regions.

Notably, this initial overview helped identify potential asymmetries and extreme values, which prompted a deeper investigation into the *distributions of individual variables* using histograms and kernel density plots. This distributional analysis revealed that several features exhibited substantial skewness(e.g. number of disaster events, population density, humidity percentage). Hence, for these features, a logarithmic transformation was applied. This reduces the impact of extreme values and makes distributions more symmetrical. Here's an illustrative example that compares the original (figure 1) and log-transformed distributions of population density (figure 2).

To investigate the relationships between variables, a correlation matrix was computed for the features (keeping in mind that they're all numerical), complemented by pairwise scatter plots to visually support these associations. For example, there is a very strong positive correlation between access to clean water and electricity, suggesting that regions with developed infrastructure

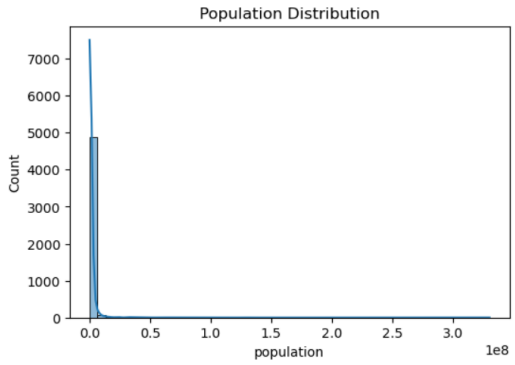

Figure 1: Original Population Density Distribution

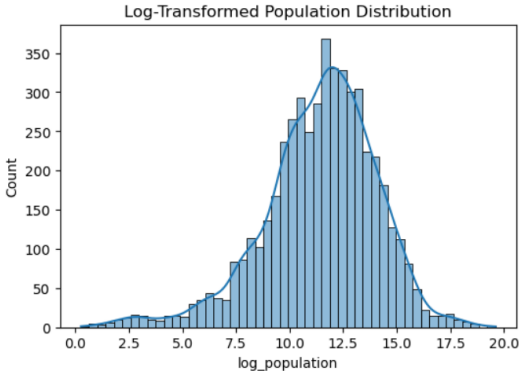

Figure 2: Log-Transformed Population Density Distribution

tend to excel across multiple basic services. As you can see in the figure 3, clean water access and electricity access both show strong negative correlations with the percentage of deaths from communicable diseases, indicating the critical role of infrastructure in reducing preventable mortality. Additionally, regions with higher earthquake frequency tend to be highly prone to tsunamis, reflecting a strong geophysical linkage. Regions with more physicians per capita are strongly associated with lower mortality from communicable and nutritional causes. Finally, regions that have a high peace index (which is bad) tend to have a higher ecological threat index.

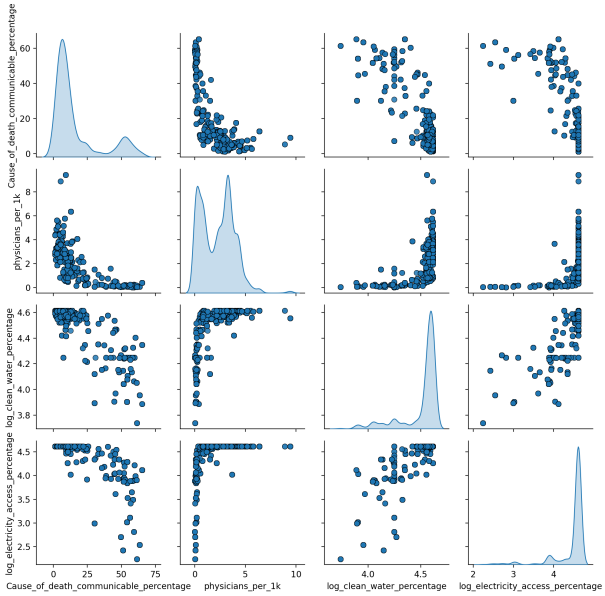

Figure 3: Example of pairwise scatter plots between certain variables

4.2 Comparative Analysis of Indicators Across Continents

To further understand regional variations in the factors contributing to survivability, we performed statistical comparisons of key indicators across different continents. Our initial exploration involved checking the assumptions for Analysis of Variance (ANOVA).

4.2.1 Rationale and Methodology

The primary goal was to determine if the average (or median, more appropriately for non-normally distributed data) values of our selected survivability indicators differed significantly from one continent to another. This helps identify broad geographical patterns and understand if certain continents inherently possess advantages or disadvantages in specific areas relevant to our survivability pillars.

Our process for each selected indicator (e.g., `log_electricity_access_percentage`, `peace_index`, `ecological_threat_index`, `avg_temp_celsius`) involved:

1. **Assumption Testing for ANOVA:**
   - **Normality within groups:** We used the Shapiro-Wilk test to assess if the indicator's values were normally distributed within each continent.
   - **Homogeneity of variances:** We used Levene's test to check if the variances of the indicator were similar across all continent groups.

   Our assumption checks consistently revealed that most indicators, when grouped by continent, did not meet the criteria for standard ANOVA (i.e., data within groups were often not normally distributed, and variances between continent groups were often unequal, with Levene's test p-values typically being very small, e.g., $p < 0.001$). *Source: User's provided ANOVA output.* Therefore, to draw robust conclusions, the non-parametric **Kruskal-Wallis H test** was identified as the more appropriate method for comparing the distributions (effectively, medians) of these indicators across continents. If the Kruskal-Wallis test indicated a significant overall difference, post-hoc tests (such as Dunn's test with Bonferroni correction) would be necessary to identify which specific pairs of continents differ.
2. **Visualization:** Box plots (or density plots used for visualizing in our presentation) were generated to visually compare the distributions of each indicator across continents.

4.2.2 Key Findings from Continental Comparisons

- **Electricity Access (`log_electricity_access_percentage`):** A highly significant difference in electricity access was observed across continents (overall test p-value < 0.001).
  - An interesting finding from the normality checks was for Europe: the Shapiro-Wilk test yielded a p-value of 1.0000 for `log_electricity_access_percentage`, accompanied by a "range zero" warning. This indicates that all European states/provinces in our dataset have virtually identical (100%) electricity access, resulting in no variance for this group on this specific (log-transformed) indicator. This perfect score sets Europe apart.
  - Post-hoc comparisons revealed that Africa has significantly lower median electricity access than all other continents.
  - Oceania also showed significantly lower access compared to Asia, Europe, North America, and South America, though its access was higher than Africa's.
  - Asia, Europe, North America, and South America did not show statistically significant differences amongst themselves in terms of already high electricity access (excluding the perfect score of Europe).

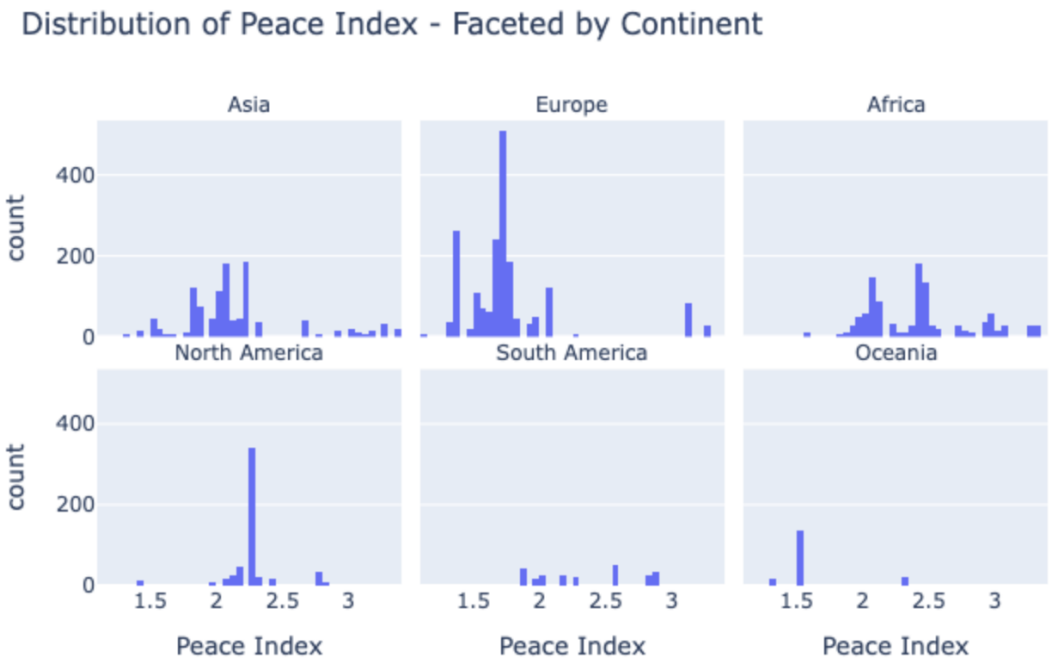

Figure 4: Enter Caption

This disparity in a fundamental infrastructure like electricity has profound implications for resource management and societal functioning in a crisis.

- **Peace Levels (`peace_index`):** Significant differences in peace levels were found across continents (overall test p-value < 0.001).
  - Post-hoc tests indicated that Europe and Oceania are, on average, substantially more peaceful (i.e., have better scores on the peace index) than Africa, Asia, North America, and South America.
  - Interestingly, Africa and South America did not show a statistically significant difference in their peace index scores from each other (pairwise p-value was not significant, e.g., $p \approx 0.2494$ from the user's Tukey output), suggesting broadly similar challenges regarding peace in these two continents based on our data.
  - North America also showed significant differences when compared to Europe/Oceania (less peaceful) and Asia (more peaceful than Asia in this comparison).

  Higher peace levels provide a stronger baseline for societal stability and coordinated responses during widespread disruptions.

- **Ecological Threat (`ecological_threat_index`):** The analysis revealed highly significant differences in ecological vulnerability patterns globally (overall test p-value < 0.001).
  - Europe exhibited a significantly lower average ecological threat index compared to all other continents, indicating greater baseline ecological stability and resilience.
  - Africa, conversely, presented with a significantly higher average ecological threat than Europe, North America, Oceania, and Asia, suggesting it might face the most severe

ecological challenges. Its difference with South America was also significant, though smaller.

- Asia and South America did not show a statistically significant difference from each other in their ecological threat profiles (pairwise p-value ≈ 0.4281), suggesting similar vulnerability levels despite their geographical diversity.
- North America and Oceania also showed no significant difference between themselves, both presenting with moderate threat levels that were higher than Europe but lower than Africa.

These patterns are critical, as ecological stability directly impacts resource availability and resilience to environmental shocks in an apocalypse scenario.

- **Average Temperature (`avg_temp_celsius`):** Statistically significant differences were also found for average temperatures across continents (overall test p-value < 0.001).
  - Europe stood out with significantly lower average temperatures compared to all other continents.
  - "Temperature equivalence zones" were observed: Africa and Oceania did not show a statistically significant difference in their average temperatures. Similarly, Asia and South America had statistically indistinguishable average temperature profiles.
  - North America occupied an intermediate position, differing significantly from most other groups.

  These thermal regimes directly influence agricultural potential, energy needs for heating/cooling, and human comfort, all vital aspects of survivability.

These continental comparisons, derived from statistical testing, provide valuable context for the state-level survivability scores. They highlight broad regional trends and disparities in fundamental conditions relevant to resilience in a crisis. While our index focuses on state-level differentiation, these wider geographical contexts clearly play a significant role. The consistent violation of ANOVA assumptions across most variables underscores the diverse and often non-normally distributed nature of global data, reinforcing the choice of robust statistical methods for comparison.

4.3 Dimensionality Reduction

To better understand the underlying structure of the dataset and reduce its complexity, dimensionality reduction techniques were applied. This step aimed to visualize patterns and reveal potential clusters. All numerical features were standardized using StandardScaler from scikit-learn. It centers the data to zero mean and scales it to unit variance. This step was necessary before applying dimensionality reduction techniques such as PCA and t-SNE.

The first technique used was t-SNE, applied directly on the full scaled dataset. Despite t-SNE being non-linear and focused primarily on preserving local structures, it revealed meaningful groupings: some states were found to cluster closely together, sharing similar survival-related characteristics. For instance, certain regions with both high ecological threat index and high peace index scores appeared in proximity, as it shows in figures 5 and **??**.

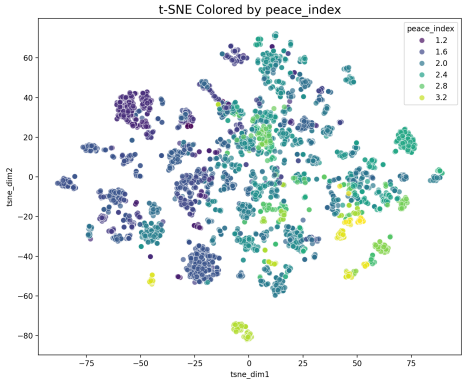

Figure 5: t-SNE by peace index

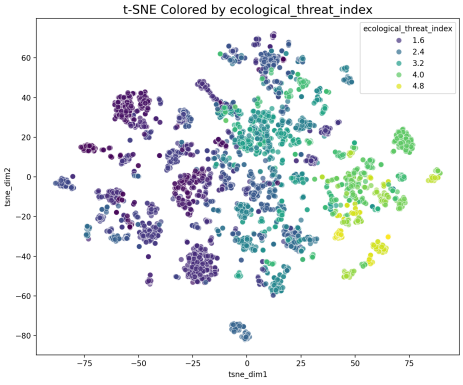

Figure 6: t-SNE by ecological threat index

To complement and stabilize the analysis, PCA was subsequently performed. The goal was to reduce dimensionality while preserving as much of the data's structure as possible. We retained the first eight principal components, which together explained approximately 80% of the total variance, thus ensuring that the majority of the dataset's information was maintained. To facilitate interpretation, the data was then projected onto the first two principal components (PC1 and PC2), which capture the dominant patterns in the data. The figure 7 helped identify the features most strongly associated with each principal axis. Our interpretation of these components is as follows :

- **PC1** distinguishes regions with robust healthcare and infrastructure from those facing serious health and ecological challenges: it shows variables like `Cause_of_death_communicable_perce` negatively associated with healthcare indicators like `clean_water_percentage`, but strong positive correlation between `physicians_per_1k` and `hosp_beds_per_1k` .
- **PC2** reflects climate-related variation, especially around precipitation: strong correlation between `rainy_days_percentage`, `log_avg_precip_mm` .

Next, we applied PCA by continent, projecting each state/province onto PC1–PC2 colored by continent to investigate whether regional patterns emerged. As shown in the figure 8, continental clustering was evident. For instance, most European states are positioned on the left side of the plot, corresponding to a high number of hospitals and clean water access — consistent with the loadings on PC1. This pattern suggests that continent-level groupings may reflect common socioeconomic and environmental characteristics (which is a bit expected).

Finally, we performed KMeans clustering on the PCA-transformed data (PC1 and PC2) . The aim was to identify natural groupings among the regions. We chose 6 clusters : Figure 9.

The KMeans clusters roughly aligned with continental boundaries. To better understand the nature of each cluster, we analyzed the most influential features within each group. The results were coherent and interpretable: for example, one cluster was defined by high healthcare capacity and infrastructure, while another showed high climate risks (tsunami, volcano, earthquake). Interestingly, we saw some regions from the United States grouping with parts of Southern Africa and the middle east. Looking closer, this particular cluster was characterized by "strong performance in clean water access and hospital bed availability, despite other differences.The African continent didn't just form one group. It split into at

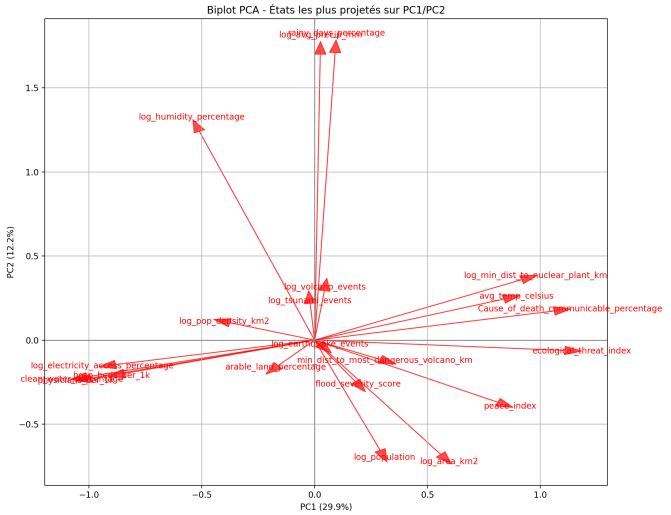

Figure 7: Biplot of Principal Components 1 and 2

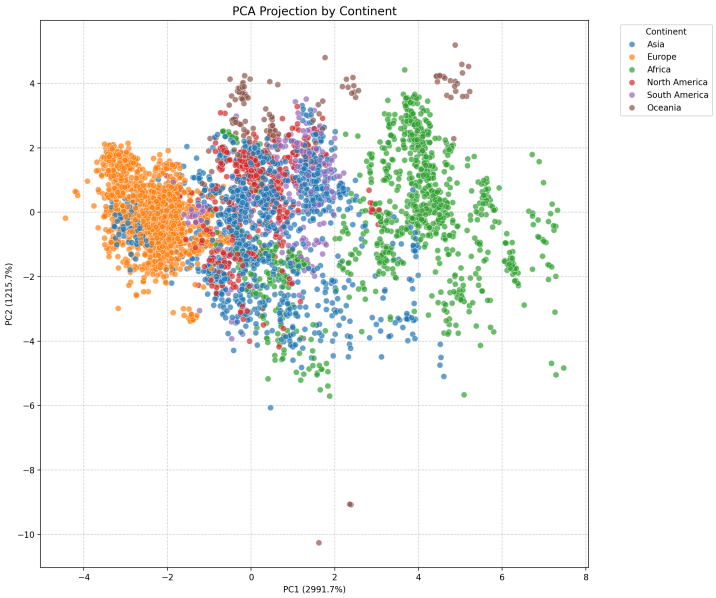

Figure 8: Projection of Data onto PC1 and PC2, by continent

least two major clusters, highlighting its internal diversity.One cluster, representing more central areas, was often characterized by "greater distances from nuclear plants (a positive), but higher ecological threat and more deaths from communicable diseases" Another distinct African cluster showed a different emphasis, with factors like the peace index playing a more dominant role in its unique profile, alongside environmental concerns. This clustering helps us see that 'survivability' isn't just about being on a particular continent; it's about a specific combination of many underlying factors.

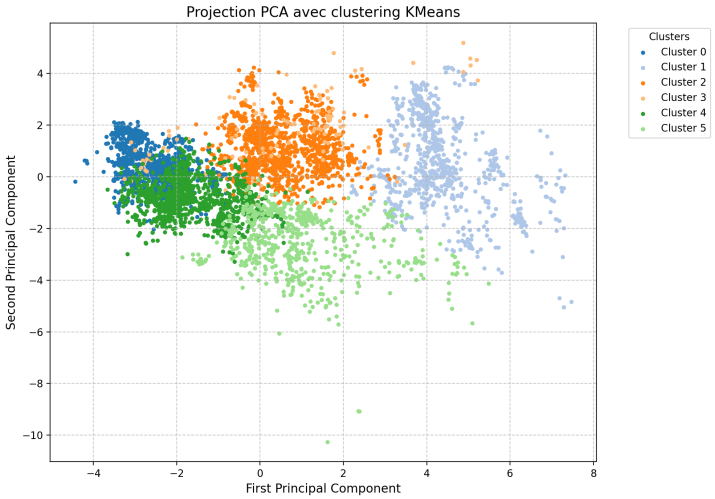

Figure 9: KMeans clustering on PC1 and PC2

5 Survivability Score and Predictive Model

After analyzing a wide range of variables, and gaining a comprehensive understanding of the key factors influencing survivability, we now formalize the **Survivability Score**. This score synthesizes information from the four previously defined pillars and provides a metric of a state's capacity to withstand and recover from apocalyptic scenarios. Once this score is computed, we can build a predictive model that estimates a state's survivability based on its indicators across the four pillars.

5.1 Survivability Score

As introduced in Section 3.2, we define four pillars, each comprising several indicators. For each indicator, we determine whether higher or lower values are more desirable, or whether an optimal range exists, based on domain knowledge and internet sources.

For instance, we set the optimal range for average temperature at **18–26 °C** based on the following sources:

- The World Health Organization recommends a minimum indoor temperature of 18 °C to protect vulnerable populations from health risks (NCBI).
- The human thermoneutral zone is typically centered around 21–25 °C, ensuring thermal comfort without additional energy expenditure (Wikipedia).
- Agronomic studies show that crop germination and growth are optimal between 20–30 °C (EOS Data Analytics).

For **elevation**, we define the optimal range as **50–1000 m**:

- This includes the global median elevation of human settlements, around 194 m (PNAS).
- It avoids elevations above 2000–2500 m, where altitude sickness and hypoxia risks become significant (Better Health Channel).

Regarding the **percentage of rainy days**, we choose an ideal range of **25–50%** :

- This range provides sufficient water without excessive cloud cover.

– It is inspired by the climate of Bilbao (44.7% rainy days), a representative oceanic climate (Weather and Climate, Wikipedia).

Finally, for **relative humidity**, we set the optimal range between **40–50%** (after applying a logarithmic transformation in the model):

– The CDC recommends maintaining indoor humidity between 40–50% to reduce allergens and viral transmission (Verywell Health).
– A slightly wider range of 40–60% is also supported by public health sources (40to60RH, Aprilaire).

Actually, the indicators collected span a wide range of scales and units—such as percentages, distances in kilometers, elevation in meters, and logarithmic transformations. In order to meaningfully combine them into scores and, ultimately, a unified **Survivability Score**, a normalization process is essential. The goal is to transform each indicator onto a common scale, typically 0 to 1, where 1 represents the most favorable situation for survivability and 0 the least favorable. Two main normalization methods were implemented, depending on the nature of the indicator and its relationship with survivability:

– **Min-Max Normalization (for "higher/lower is better" indicators)**: For indicators where a higher value is considered better (e.g., `min_dist_to_most_dangerous_volcano_km`), the normalization formula is:

$$\text{Normalized Score} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

For indicators where a lower value is considered better (e.g., `log_earthquake_events`), the formula is inverted:

$$\text{Normalized Score} = 1 - \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

Where X is the indicator's value for a given region, and $X_{\min}$ and $X_{\max}$ are the minimum and maximum values of that indicator observed across the entire dataset (after imputation and transformations).

– **Normalization for "Optimal Range" (for indicators where a moderate value is preferred)**: For certain indicators in Pillar 1 (Environmental Stability), such as `avg_temp_celsius`, `elevation_m`, and `rainy_days_percentage`, neither very high nor very low values are desirable. A moderate range is considered optimal. For these, a specific normalization function is applied:
  * Values falling within the defined ideal range receive a normalized score of 1.
  * For values below `ideal_min`, the score increases linearly from 0 (at the indicator's observed minimum, `X_min`) to 1 (at `ideal_min`). The formula is:

$$\text{Normalized Score} = \frac{X - \text{X}_{\min}}{\text{ideal}_{\min} - \text{X}_{\min}}$$

  * For values above `ideal_max`, the score decreases linearly from 1 (at `ideal_max`) to 0 (at the indicator's observed maximum, `X_max`). The formula is:

$$\text{Normalized Score} = 1 - \frac{X - \text{ideal}_{\max}}{\text{X}_{\max} - \text{ideal}_{\max}}$$

(Adjustments are made if $\text{X}_{\max} \leq \text{ideal}_{\max}$).

Once all relevant indicators are normalized to a 0-1 scale (where 1 is always preferable), the scores are aggregated.

Pillar Score Calculation : For each pillar, the score is calculated as the simple arithmetic mean of the normalized scores of all its constituent indicators. For example:

$$\text{Pillar 1 Score} = \frac{1}{N_1} \sum_{i=1}^{N_1} \text{Normalized Score}_i$$

where N_1 is the number of indicators in Pillar 1. This approach assigns equal weight to each normalized indicator within the same pillar.

Final Survivability Score Calculation : The final survivability score for each region is calculated as the simple arithmetic mean of the four pillar scores:

$$\text{Final Score} = \frac{\text{Pillar 1 Score} + \text{Pillar 2 Score} + \text{Pillar 3 Score} + \text{Pillar 4 Score}}{4}$$

This method, by default, assigns equal weight to each pillar in determining the overall score.

Insights from the Survivability Rankings : One of the most surprising results from our Survivability Index is the top-ranked region: the French Southern and Antarctic Lands. At first glance, this remote and sparsely inhabited territory may seem like an unlikely candidate for the safest place to survive a global systemic crisis. However, its exceptionally high score is largely driven by its performance in the Hazard Exposure pillar, where it ranks first due to its extreme geographical isolation and minimal exposure to natural or human-made threats, as you can see in the Figure 10. This highlights a key insight from our model — that remoteness can be a powerful asset in certain crisis scenarios.

That said, this ranking also reveals a limitation in our approach. The French Southern and Antarctic Lands benefit from imputed data or national-level French statistics for several pillars, including infrastructure and health systems. In reality, such data may not accurately represent conditions in this remote territory, where access to services and supplies is far more limited. This serves as a reminder that top rankings should always be interpreted with caution, especially when data granularity is insufficient.

State	Country	**Final survivability score (%)**	Environmental Stability	Hazard Exposure	Resource Availability & Infrastructure	Societal Stability & Health Systems
French Southern and Antarctic Lands	France	77	76	96	77	59
Gagauzia	Moldova	76	84	77	84	60
Dobrich	Bulgaria	76	80	77	76	71

Figure 10: Ranking of states by highest survivability score

In contrast, regions like Gagauzia in Moldova demonstrate a different profile of resilience. Gagauzia did not lead any single pillar, but consistently performed well across all four dimensions. This balance suggests a robust, multi-faceted form of survivability that may

be more reliable in practice. It also illustrates that high survivability scores can be achieved through different combinations of strengths. While some regions rank highly due to exceptional performance in one domain (like low hazard exposure), others, like Gagauzia, embody a steadier and perhaps more realistic resilience through balanced performance.

Overall, these results emphasize that there is no single equation for survivability. Regions can score highly by excelling in one area or by maintaining balance across all. At the same time, careful consideration of data quality and modeling choices is essential when interpreting the rankings and deriving actionable insights from them.

5.2 Predictive model

Once the normalized indicators were used to compute the final Survivability Score for each state, we used a model to extract important features and also built a predictive model to estimate this score from the raw normalized features. This serves two purposes:

- It allows interpretation and ranking of feature importance, helping us understand which variables most strongly influence survivability.
- It enables generalization: one can input custom values for the indicators of a new or hypothetical region, and obtain an estimated Survivability Score.

Feature Importance: To identify key indicators, we examined feature importance using a Random Forest model. This helped us detect which features most affect the Survivability Score.

Agricultural land percentage (18.3%) emerged as the most important predictor of survivability. This underscores the critical need for local food production, especially in scenarios where global supply chains break down. *Distance from dangerous volcanoes* (13.1%) ranks second, showing that proximity to major volcanic threats significantly impacts survivability. *Peace index* (11.8%) follows closely, highlighting the importance of pre-existing social and political stability. Societies with lower conflict are more likely to maintain order and manage crises effectively. *Ecological threat index* (8.3%) also contributes significantly, reinforcing the idea that existing environmental pressures can compound future risks.

Surprisingly, some variables commonly associated with survivability had very low importance: *Natural disaster frequency indicators* such as earthquakes, tsunamis, and volcanoes individually contributed less than 0.1%. Temperature and other climate-related variables had feature importance below 1%, suggesting that while climate matters, it may not be a key determinant of survival in extreme scenarios.

At the pillar level, we note that *Resource Availability & Social Stability* are the dominant factors in survivability. *Environmental Stability*, surprisingly, had the lowest overall contribution (0.51%), suggesting that human adaptability to various environmental conditions is high—provided other essential needs are met.

These insights reinforce the need to prioritize local resource availability and institutional strength over climatic perfection when planning for survival scenarios.

Model Evaluation: We then trained and evaluated two different regression models for prediction:

1. **Random Forest Regressor** (with 100 trees).

2. **XGBoost Regressor** (with 100 estimators, max depth of 5).

Both models were evaluated on training and test sets using standard regression metrics: R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE).

After analyzing the results, we noticed that both models perform very well, but XGBoost outperformed Random Forest slightly across all metrics on the test set. Notably, it proved particularly effective at accurately predicting lower survivability scores, which are crucial for identifying high-risk regions. The prediction performance of both models is visualized in Figure 11

Thus, we retain the XGBoost model as our final predictor. It can now be used to compute the Survivability Score for any new region by providing the normalized values of the relevant indicators.

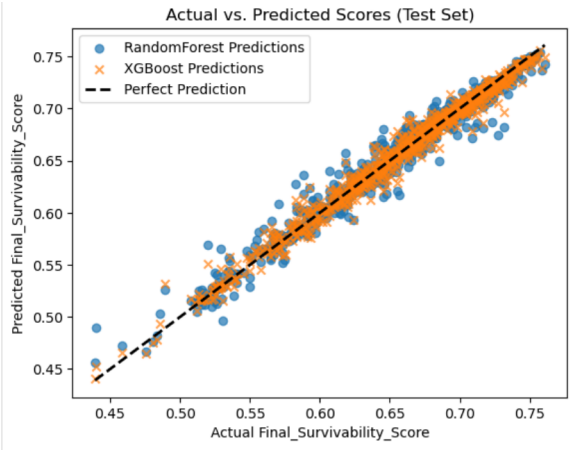

Figure 11: Random Forest vs XGBoost predictions

6 Interactive Application for Results Exploration

To enable easy exploration and a deeper understanding of the survivability index results, an interactive web application was developed in Python using the Streamlit library. This application loads the `survivability_results.csv` file (and potentially `df_states.csv` for supplementary raw data) and offers several analytical and visualization features.

The application allows users to navigate the data, compare regions, and explore the impact of different factors on survivability. The interface is structured with a sidebar for controls and filters, and a main area for displaying data tables and visualizations.

The application is structured around the following components:

Interactive Pillar Weight Adjustment ("What If" Scenario) : This feature allows users to modify the relative importance (weight) of each of the four pillars in the final score calculation. Numeric input fields are provided for entering percentage weights for each pillar. When one weight is changed, the other weights automatically adjust to maintain a total sum of 100%. The application dynamically recalculates the survivability score and the state rank based on these new weights. The main table and potentially the map update to reflect these changes, enabling direct sensitivity analysis.

Custom Prediction with XGBoost: In addition to modifying pillar weights, users can input custom values for any or all of the normalized indicators (e.g., agricultural land

%, ecological threat index, peace index, etc.). Once submitted, these values are passed into the trained `XGBoost` regression model to generate a predicted Survivability Score for a hypothetical or new region.

This tool not only allows users to explore global survivability rankings, but also offers a hands-on simulation environment to test hypothetical regions under crisis scenarios.

7 Conclusion and Future Perspectives

This report presented the development of a multidimensional composite index designed to assess the relative survivability of states worldwide, within the context of an "apocalypstic" scenario, characterized by increasing environmental instability and the fragmentation of global socio-economic systems. The Survivability Index is built upon four key pillars: Environmental Stability, Hazard Exposure, Resource Availability & Infrastructure, and Societal Stability & Health Systems.

The project followed a complete data science pipeline, beginning with extensive data acquisition from heterogeneous sources (open databases, APIs, and web scraping). Data was then cleaned, transformed (e.g., logarithmic transformations), and imputed where necessary. To better understand the relationships among variables and reduce redundancy, we performed correlation analysis across indicators. Furthermore, dimensionality reduction techniques such as Principal Component Analysis (PCA) were applied to capture the underlying structure of the data, and KMeans clustering was used to explore natural groupings among the regions based on their survivability profiles.

To facilitate the exploration and interpretation of the results, an interactive web application was developed. This application allows for the visualization of data in filterable and sortable tables and a cartographic representation of scores.

A major strength of this project lies in its multidimensional approach, which considers four distinct pillars. This structure enables allows for coverage of a broad spectrum of factors influencing survivability. Another key strength is the methodological flexibility of the framework: the modularity of the index allows for the addition, removal, or refinement of indicators.

Despite its strengths, the study has several limitations that must be considered when interpreting the results. One significant constraint is the quality and granularity of the data: many critical indicators, such as food security or access to clean water, are often only available at the national level, leading to potential inaccuracies when applied to states. Additionally, while missing data was addressed through imputation, some regions still suffer from incomplete information, which can affect score reliability. More importantly, The selection of indicators and their weights involves a degree of subjectivity. Furthermore, the distribution of scores across all regions reveals a notable clustering around moderate values — particularly near 70% survivability score. In fact, it may be an artifact of our equal-weighting scheme, which tends to pull composite scores toward the center. This observation provides an important critique of the index's construction, pointing to a possible over-smoothing of differences between regions. It also highlights a valuable avenue for future refinement, such as exploring alternative weighting schemes or introducing more dynamic scoring mechanisms to better capture regional disparities.

Several avenues for future improvement can be considered to enhance both the robustness and usability of the Survivability Index. First, improving data quality remains a priority, with efforts aimed at sourcing more granular and full data, as well as integrating information

on disaster intensity. Refining the weighting methodology is another promising direction ; exploring alternative weighting schemes or introducing more dynamic scoring mechanisms to better capture regional disparities.