Priority Risk Identification and Scope Measurement

Abhijeet Talaulikar, Sajid Hussain Rafi Ahamed,
Siddharth Susarla, Yuthika Shekhar, Hailey Thanki

Group 06, DSC 483
Goergen Institute for Data Science
University of Rochester

1. Introduction

Figure 1: Risk Areas

This project aimed to determine the survey variables that are the best predictors of priority risks. The United Nations (UN) is an intergovernmental organization whose stated purposes are to maintain international peace and security, develop friendly relations among nations, achieve international cooperation, and be a center for harmonizing the actions of governments. It consists of 193 member states. The UN has 4 main pillars: Peace and Security, Human Rights, Rule of Law and Development. The UN Secretary General, General Assembly and Security Council have the most control over the UN.

The Secretary General chairs the Regional Monthly Review (RMR) and it is a body within the UN, that conducts bi-annual surveys. The RMR was conceived under the Human Rights Up Front policy and was developed to ensure that the UN system has a shared understanding of situations and takes early and coordinated action for the prevention of consequences of existing risks. The risk areas are broken down into 13 categories as shown in Fig 1.

2. Dataset Description

Figure 2: Countries that were survey participants

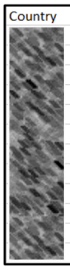
The raw survey data frame contains information about 53 surveys conducted within a span of 3 years starting 2019 until 2021. These surveys were taken from participants belonging to 53 different countries of the world by 62 UN entities. A total of 667 individuals participated in these surveys. Fig. 2 shows the countries the surveys were conducted in. Fig 2 is a snippet of the raw data frame before performing transformations.

Following are the columns of the raw data frame:
1. Topic - represents the alias associated with the country the participant is from.
2. Date - represents the day the survey was conducted.
3. Topic ID - it is a unique key that represents each row of the data frame.
4. Participant ID - it is a unique key that represents each participant that took part in the surveys.
5. Question - this column represents all the questions asked in the surveys to each participant throughout the years.
6. Value - it contains the answer to a particular question filled in by a survey participant.
7. Scope - this column represents whether the variables in the row are predictor variables or outcome variables.

Topic	Date	Topic_ID	Participant_ID	Question	Value	Scope
Country 1	Wednesday, 01 December, 2021	RMR_Country1_Survey_211201_coded-2022-05-17637720503256747952RMR_CT_25_7	6.38E+17	(q126) 37. how do you assess national capacity to cope with the identified risks?	Moderate	Predictor
Country 2	Tuesday, 19 November, 2019	RMR_Country2_Survey_191119_coded-2022-03-18637090568908120276RMR_CT_37_8	6.37E+17	(q160) 43. how do you assess local un system capacity to engage and help address such risks?	Low	Predictor
Country 3	Thursday, 25 February, 2021	RMR_Country3_Survey_210225_coded-2022-05-17637491571930721337RMR_CT_29	6.37E+17	(Q121) 10.14. Are there indications that the risk related to internal security could increase over the next 6 months to an extent that it should be a priority in the upcoming RMR discussion, also bearing in mind the implications of the COVID-19 pandemic?	Yes: Strong	Outcome

Figure 3: Raw Data

The questions in the surveys are broadly classified into 6 categories:
1. Change in UN engagement by extent and urgency - For e.g. ”From your perspective, would the scenario you identify require a change in UN engagement over the next 6 months?”
2. Priority risks - (Multi-select) For e.g. ”Please select the risk areas you would like to identify as a priority for the RMR discussion.”
3. Identity and demographics - For e.g. ” Please select your entity from the drop-down list. If you are submitting for the Resident Coordinator or equivalent, please ensure to select ’UN Resident Coordinator’.”
4. National coping capacity - For e.g. ”How do you assess national capacity to cope with the identified risks?”
5. Risk increase potential - For e.g. ”Are there indications that the risk related to environment climate could increase over the next 6 months to an extent that it should be a priority in the upcoming RMR discussion, also bearing in mind the implications of the COVID-19 pandemic?”
6. UN response capacity - For e.g. ”How do you assess local un system capacity to engage and help address such risks?”

Country	Survey_ID	Participant	outcome	q109	q111	q112	q114	q115	q117	q118	q120	q121	q123	q124	q126	q127	q129
	Country47	p6368927(	Democratic space | In		Moderate					Yes: Moderate				Yes: Minor	Moderate		
	Country47	p6368931	Democratic space | Ju		High					Yes: Minor							
	Country47	p6368936	Democrati	Yes: Minor	Don't know					Yes: Minor							
	Country47	p6368938	Justice & r	Yes: Moderate						Yes: Moderate							
	Country47	p6368938	Democratic space | In		Moderate	Yes: Minor				Yes: Moderate				Yes: Minor	Moderate		
	Country47	p6368940	Economic	Yes: Moderate					Moderate								
	Country47	p6368940	Democrati	Yes: Mode	Moderate					Yes: Moderate							
	Country47	p6368940	Democrati	Yes: Stron	Low												
	Country47	p6368973	Displacem	Don't know			Moderate			Don't know							
	Country49	p6368689	Economic	Yes: Minor					Moderate								
	Country49	p6368728	Democratic space | Di		Moderate	Yes: Stron	Moderate		Moderate	Yes: Moderate							
	Country49	p6368728	Democratic space | Di		Very high		High		High								

Participant responses

Question responses

Figure 4: Transformed Data

Each survey has about 42 questions that each participant from various countries was asked. Some transformations were performed on the data frame and was un-pivotted (Fig. 4). A new column called Survey ID was created, which is one of the unique identifiers in the data frame. This attribute represents a unique key, which is a combination of the Country ID and date. Then, the data frame is pivoted in a way that every row represents each participant's answers and most of the columns represent the question asked and an individual's answer to the question. The key uniquely identifies each row of the data frame. The question ID corresponding to this is extracted from the text. Each survey and participant ID combination is considered as a row. Each row is interpreted as an opinion of a participant in a specific survey about a country. The country aliases are also mapped to the country names. The pivoted data frame contains about 860 rows and 48 columns.

There is missing data in the data frame due to two reasons. One of the reasons can be interpreted as non-response bias, which represents the survey participants that may have answered the questions with "Don't know" or skipped the question. The other reason for missing values is that the question wasn't asked, since the surveys are always evolving and the older surveys do not have some of the questions from newer surveys. The imputation for both of these cases was handled differently.

The predictor variables in this scenario are the answers to the various survey questions, and the outcome column is the combination of different risk factors on UN action urgency.

3. Exploratory Analysis

3.1. Data Sparsity

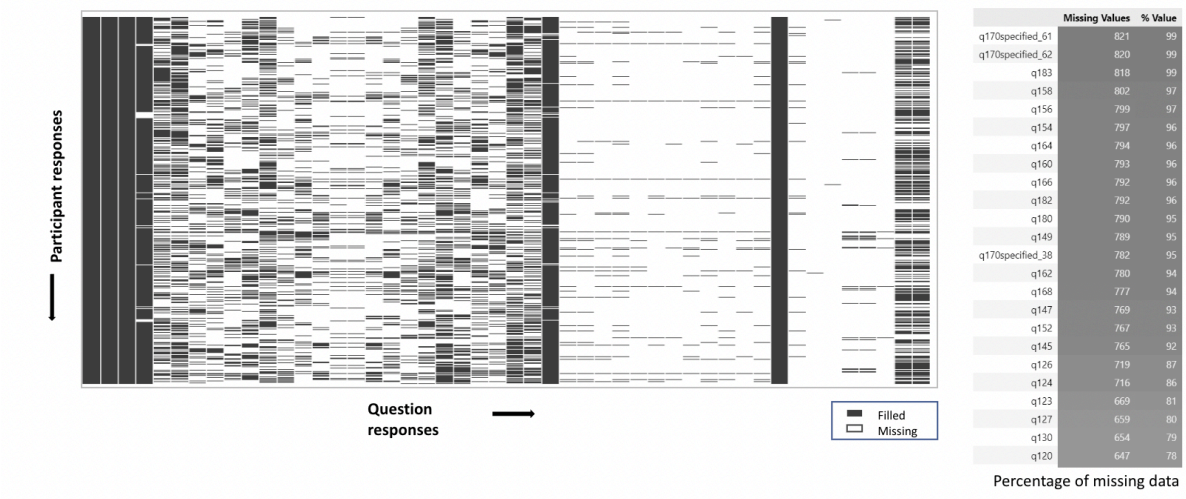


Figure 5: Data Sparsity

Sparse data is a variable in which the cells do not contain actual data within data analysis. Sparse data is empty or has a zero value. This data is different from missing data because sparse data shows up as empty or zero, while missing data doesn't show what some or any of the values are.

There are two types of sparsity:
• Controlled sparsity: when there is a range of values for multiple dimensions that have no value.
• Random sparsity: when there is sparse data scattered randomly throughout the datasets

The visualization of sparse data can be seen in Fig. 5. The darker cells represent non-zero and non-null values. The lighter cells represent the missing values in the dataset. It can be observed that there is random sparsity in the dataset.

3.2. Distribution of data over the years

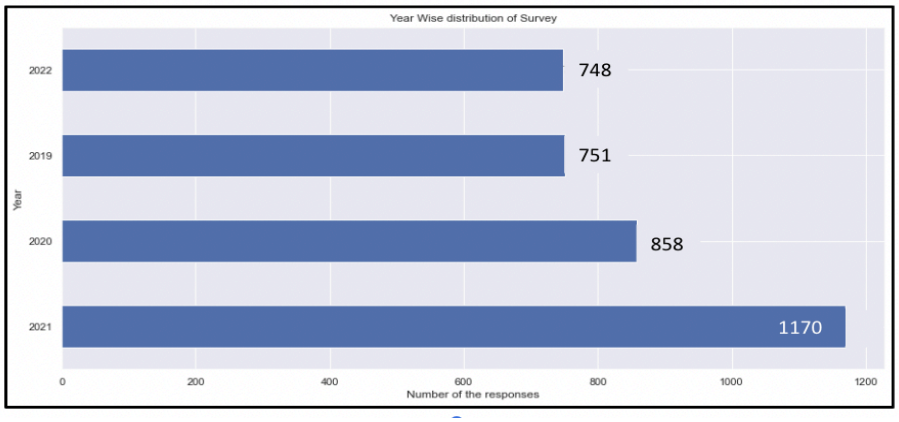


Figure 6: Data Distribution over the Years

Fig. 6 shows the number of surveys conducted in all the countries over the years. We had information about the surveys starting in the year 2019 until the year 2022. It can be observed that most of the rows contain information about the surveys conducted in 2021. The least amount of information is available for the surveys conducted recently in 2022. However, the differences in the amount of data available throughout the years are not significantly different. Typically, countries go through a single survey in a year. However, there may be another survey conducted in a country based on the risks evaluated by the analysis team of the UN.

3.3. Number of surveys conducted in a particular nation over the years

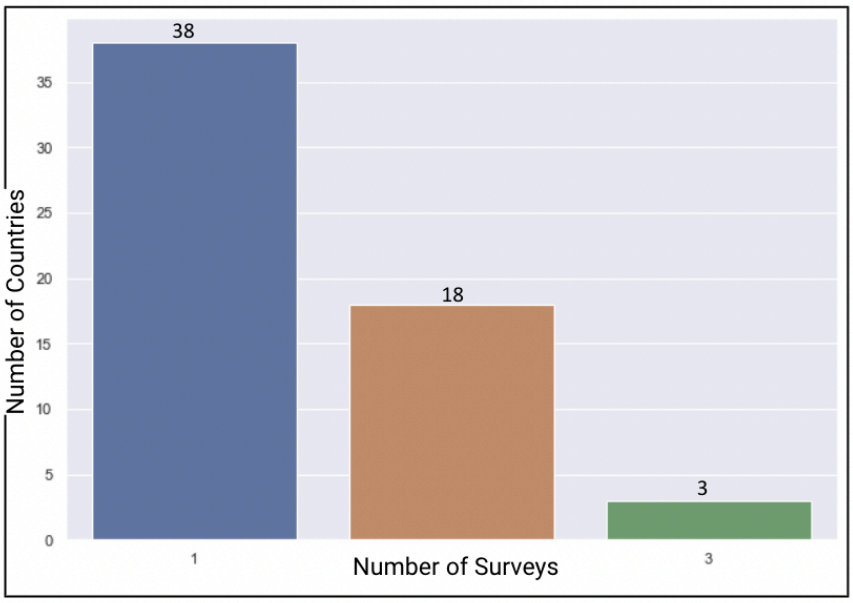


Fig. 7 shows the number of countries where 1, 2 or 3 surveys were conducted over the years. According to the bar chart, only 1 survey was conducted in about 37 countries. 2 surveys were conducted in about 18 countries. 3 surveys were conducted in about 3 countries. So, in most of the countries, only one survey was conducted and 2-3 surveys were conducted in fewer countries compared to the number of countries where only 1 survey was conducted. About 85% of the countries had a single survey from 2019 until 2022.

Figure 7: Number of Surveys Conducted over the Years

3.4. Distribution of participants in countries

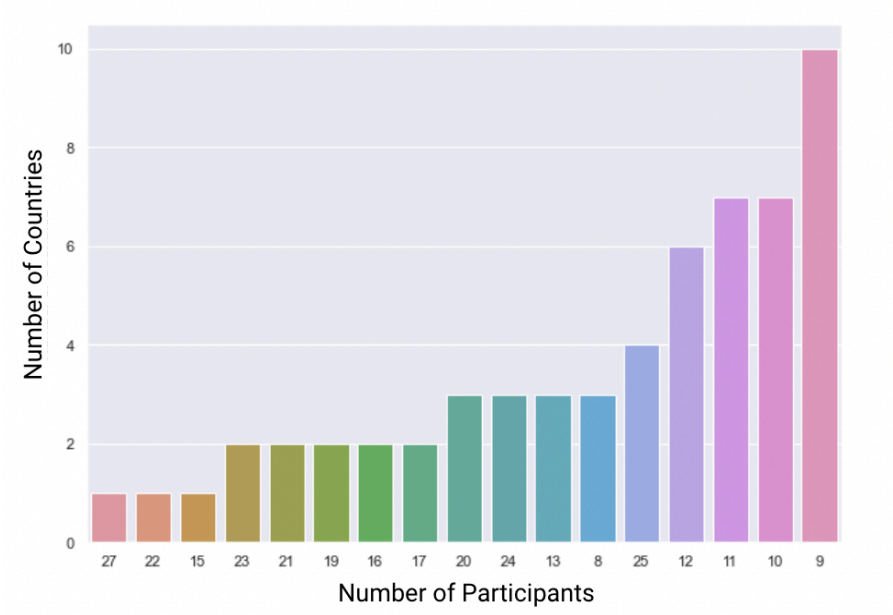


Fig. 8 shows the number of countries where a certain number of participants participated in the surveys conducted from 2019 until 2022. For e.g. from the plot, it can be observed that there were 10 countries where only 9 participants participated in the surveys in those 3 years. As can be observed, 27 countries had only a single participant over the course of these 3 years. Most countries had no more than 0-4 participants. About 24 countries had 9-10 participants take the survey.

Figure 8: Number of Participants from Countries

3.5. Distribution of Risk Factors

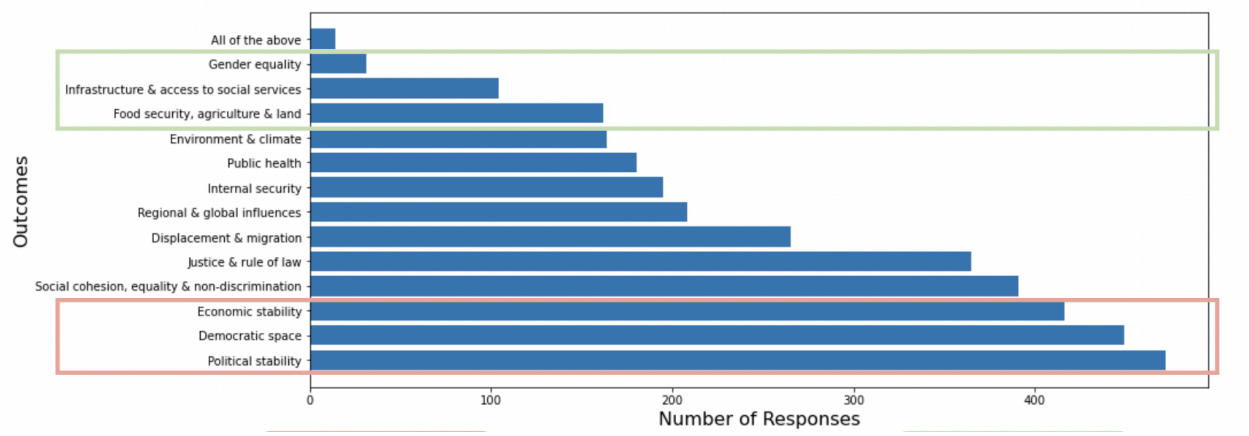


Figure 9: Number of responses and the frequency of risk factors

The most commonly occurring risk factors are Political stability, Democratic Space and Economic Stability. The least commonly occurring risk factors are Infrastructure, Gender Equality and Food Security. Fig 9 shows the number of times each risk factor occurred in the survey responses.

3.6. Most voted risk areas by RMR survey year

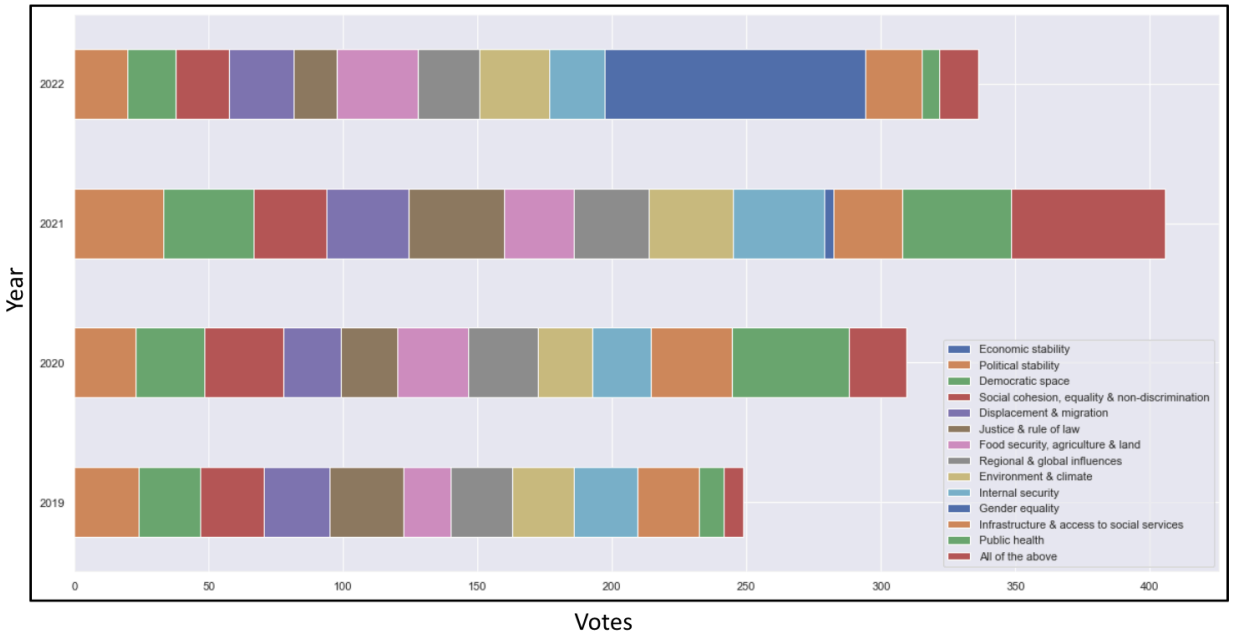


Figure 10: Most voted risk areas by RMR survey year

Fig. 10 shows the breakdown of the risk factor categories by year. The percentage of votes for each risk factor seemed to remain consistent over the years, except for economic stability in 2022 and public health in 2020-21. This phenomenon can be attributed to the Covid-19 pandemic.

3.7 Distribution of UN Extent Urgency

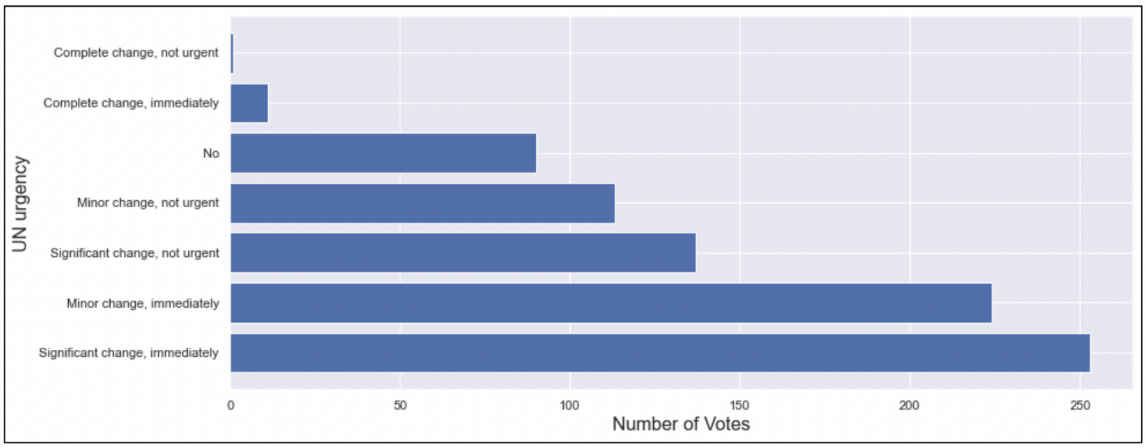


Figure 11: Distribution of UN Extent Urgency

Fig. 11 shows the frequency of votes on the nature of action that needs to be taken by the UN in order to mitigate the Risks. These responses are ordinal ranging from ”not urgent” to ”immediately” Majority of the responses (about 250 responses) said that there are significant changes which need to be acted on immediately.

3.8. UN Urgency aggregated based on Region

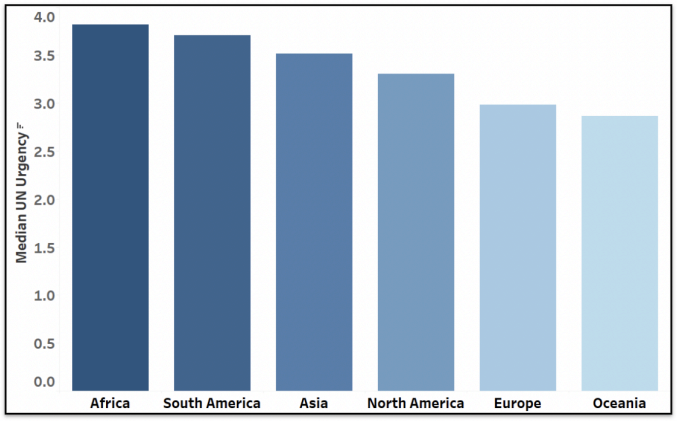


Figure 12: UN Urgency aggregated based on Country

Fig. 12 shows the median of the scale of the urgency of action required in the opinion of the survey participants from each country/continent. Africa has the highest median urgency across all risk factors, and also the highest number of countries in the Data. Oceania has the fewest number of countries in the data which appear to have a low urgency.

3.9. Voting patterns of UN Entities for Extent and Urgency of UN Action

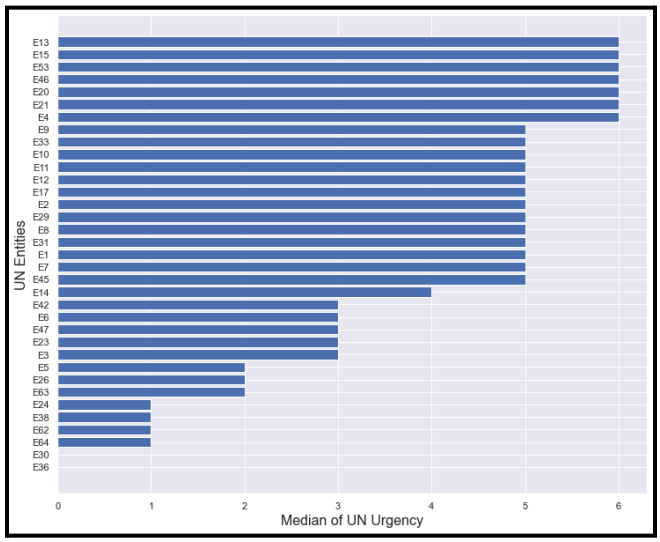


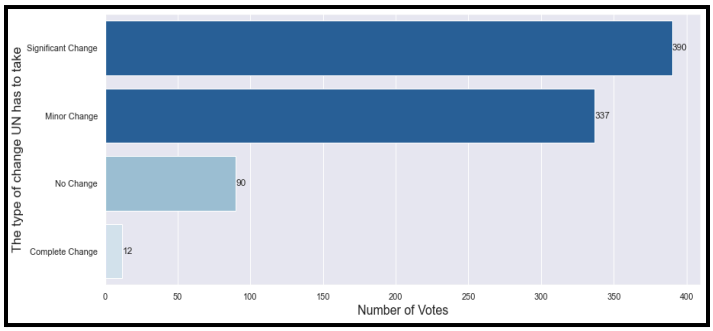
Fig. 13 represents the voting patterns of UN entities for the extent and urgency of UN action.

The median is insensitive to very strong positive and negative responses.

There are about 60 UN Entities which are referred to here by their aliases ranging from E1 to E64. Some examples of these UN entities are UNICEF, OCHA, World Food Program, DPPA etc.

The median is calculated over all the countries where the survey was conducted. Figure 13: Voting patterns of UN Entities

3.10. Distribution of UN Extent & Urgency



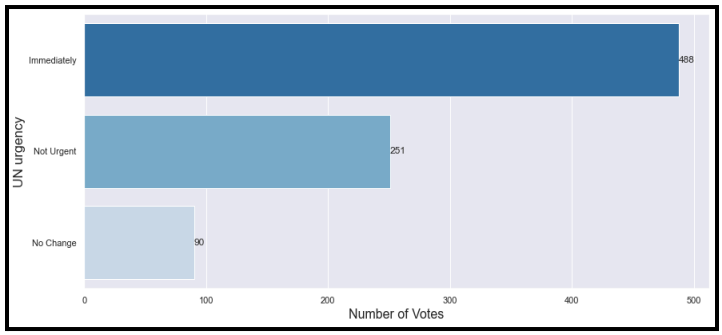


Figure 14 (a): Distribution of UN Extent & Urgency

Fig. 14 (a) shows how the responses regarding the urgency of UN involvement in mitigating the risk factors of survey participants were distributed. Moreover, it also presents the extent to which the UN's involvement is necessary. We

observed the UN's Extent and Urgency at a much granular level and dealt with them separately. About 350 participants voted for significant change which UN needs to take and over 480 participants voted for immediate UN Urgency.

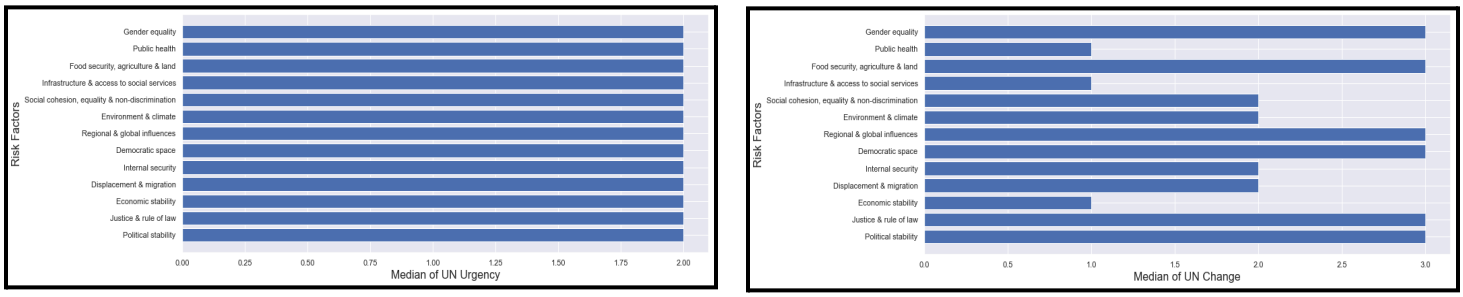


Figure 14 (b): Distribution of UN Extent & Urgency

Interestingly, Economic Stability, which was the most common risk factor, was voted as requiring minor change. We find the median urgency to be quite similar across all risk factors. Fig. 14 (b) represents the median of the ratings given by survey participants with regard to the amount of change required to be made as a consequence of UN involvement. It also shows the median of the ratings given by survey participants with regard to the urgency with which the UN should get involved in order to mitigate the impending risk the region might face.

3.11. Most Voted Risk Areas

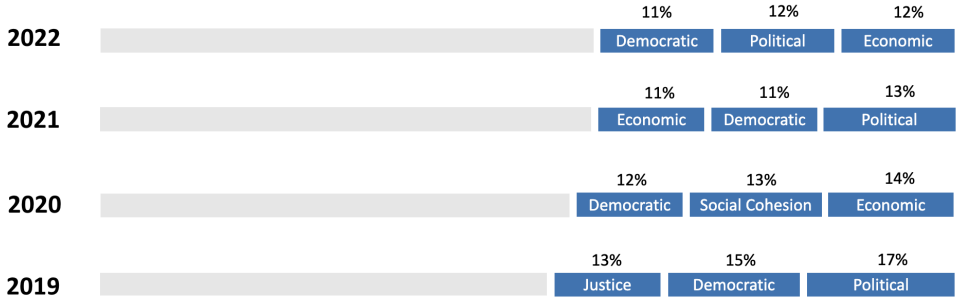


Figure 15: Most voted risk areas

Fig. 15 shows the risk areas that were the most voted. It can be observed that over the years, Democratic Stability, Political Stability and Economic Stability consistently remained the risk areas that were the most voted.

4. Model Development

10 qualitative responses were presented to participants across surveys. Standardized the weight of the responses using a Net Promoter Score. As mentioned earlier, the responses to the survey questions were ordinal, namely: Yes: Very Strong | Yes: Very High | Yes: Strong, High | Yes: Moderate, Moderate | Yes: Minor, Low | No | None, Don't Know.
These responses were first mapped to an integer score in order to fit an ML model. These responses are given by survey participants for questions related to risks. The mapping is shown in Table 1.

Response	Score
Yes: Very Strong, Very High	5
Yes: Strong, High	4
Yes: Moderate, Moderate	3
Yes: Minor, Low	2
No	1
None, Don't Know	0

Table 1: Mapping of Responses for Risk-Related Questions

Response	Score
Very High	1
High	2
Moderate	3
Low	4
None , Don't Know	0

Table 2: Mapping of Responses for Coping Question

There are two types of questions based on coping capacity-National and UN System. 6 qualitative responses were presented to participants across surveys, namely Very High | High | Moderate | Low | None, Don't Know. These responses correspond to the questions regarding coping capacity. We standardized the weight of the responses using a Net Promoter Score as shown in Table 2.

The response scores corresponding to the questions related to a particular risk factor summed up. The scores are combined based on mapping given by the UN for coping capacity. Then, the mean of the scores is calculated by grouping the countries in which the survey was conducted and the survey ID variables. This would give us the score of each risk factor the country is facing and to what extent according to the survey participants. This brings attention to the risk factor that is believed to be the one requiring the earliest action by the UN.

The following steps were followed for getting the data ready for an ML model:

1. We consider the last two surveys conducted in a country to perform these calculations since the clustering is a reflection of the current situation in countries, we take only the latest 2 surveys which are usually 6 months apart.
2. The outcome column is split into multiple rows for performing further steps.
3. We then created a dictionary to capture the urgency for every risk factor for each country.
4. Converted the previously created dictionary to a single vector for each country. This vector captures the response urgency of every risk factor.

We fit this data into a K-Means Clustering algorithm. K-means clustering is a method of vector quantization, originally from signal processing, that aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster.

4.1. K-Means

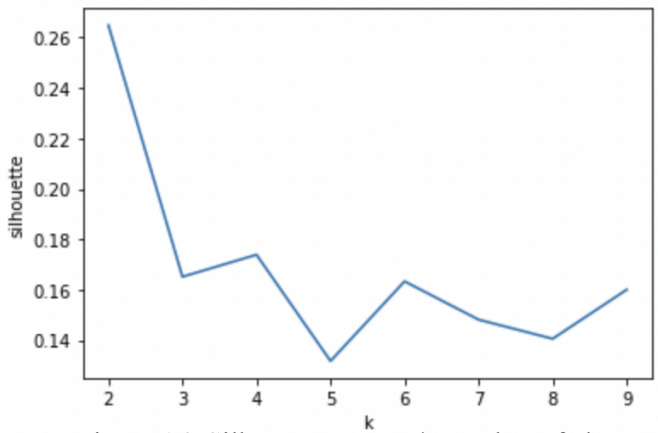


Figure 16: Silhouette score v/s number of clusters

For each problem, there might be a certain number of clusters that would be the most appropriate. In order to determine this, we performed a step in order to determine the optimal number of clusters for K-Means clustering based on the Silhouette score. Since we had very little data, we expected and chose a very low number of clusters. Fig. 16 shows the Silhouette score for a certain number of clusters. As can be observed in the graph, the optimal number of clusters was 2.

We perform a PCA analysis in order to reduce the dimensionality of the data. Then, we performed the K-means clustering based on the results obtained for the number of clusters required for maximizing the Silhouette score.

4.2. FP-Growth

Frequent Pattern Growth Algorithm is the method of finding frequent patterns without candidate generation. It constructs an FP Tree rather than using the generate and test strategy of Apriori. The focus of the FP Growth algorithm is on fragmenting the paths of the items and mining frequent patterns.

Using the FP-Growth algorithm we identified the pattern within the priority risks and impending risks for all the countries where the surveys were conducted.

5. Performance and Results

We determined the perception of risk increases across RMR risk areas and surveys. Fig. 17 shows the top 5 risk predictors based on the sum of scores are political stability, economic stability, democratic space, displacement & migration and environment & climate.

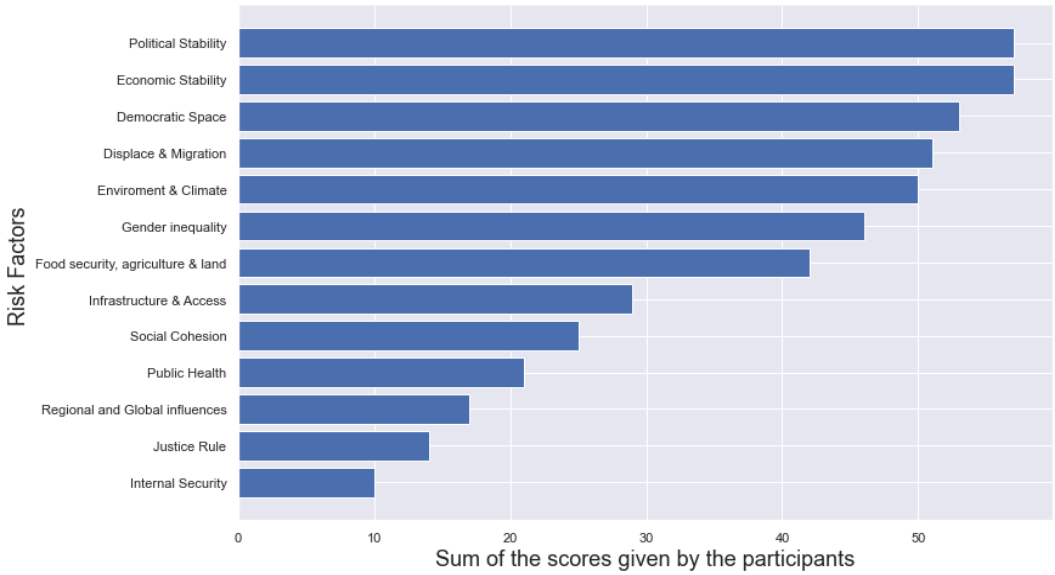


Figure 17: Risk predictors based on the sum of scores

Next, we determined the consensus of risk increases across RMR risk areas and surveys. Fig. 18 shows the top 5 risk predictors based on the total number of votes are political stability, economic stability, democratic space, displacement & migration and environment & climate.

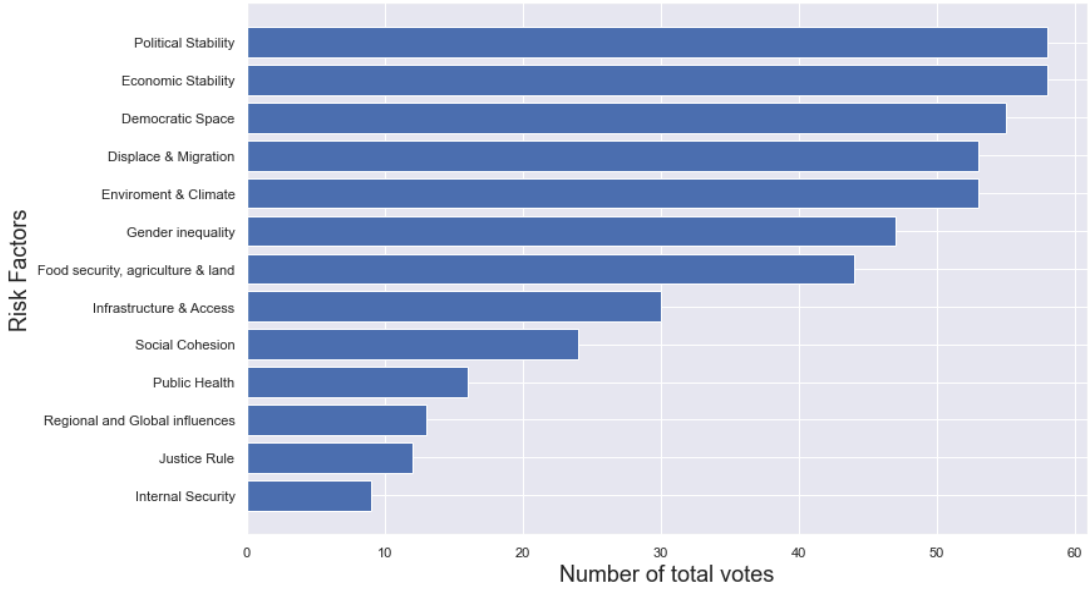


Figure 18: Risk predictors based on the number of votes

Figure 19: Countries grouped by Risk Perceptions

We grouped countries based on participant perceptions of the 13 risks. The goal was to enable the UN to standardize its actions based on similar risks. 13 features were utilized to determine prevalence of risk factors. Fig. 19 shows the clustering based on the risk factor classification that was determined. The features of these two clusters are as follows:

Cluster C1:

- Economic stability, political stability and public health are a concern.
- Food security is perceived as a risk.

Cluster C2:

- Overall risk, no distinguishing characteristics.

Figure 20: Countries grouped by Risk Urgency

Next, the countries were grouped based on participant perceptions of urgency in UN intervention. The goal was to enable the UN to allocate resources to countries in effective timeframes. 7 qualitative features were identified for determining UN action urgency. Fig. 20 shows the clustering based on the risk urgency classification that was determined. The features of these two clusters are as follows:

Cluster C1:

- Countries in this cluster believe that UN intervention needs a complete change, and it needs to happen immediately.

Cluster C2:

- Countries in this cluster believe the UN needs to enact some minor changes immediately, but significant changes are not urgent.

Some interesting patterns we observed using the FP Growth algorithm. Fig. 21 shows the risk that would potentially follow a certain priority risk in the opinion of the survey participants. For e.g. when economic stability, political stability, democratic space or justice are an identified risk, 30% of the participants thought the UN engagement needed a significant change over coming months.

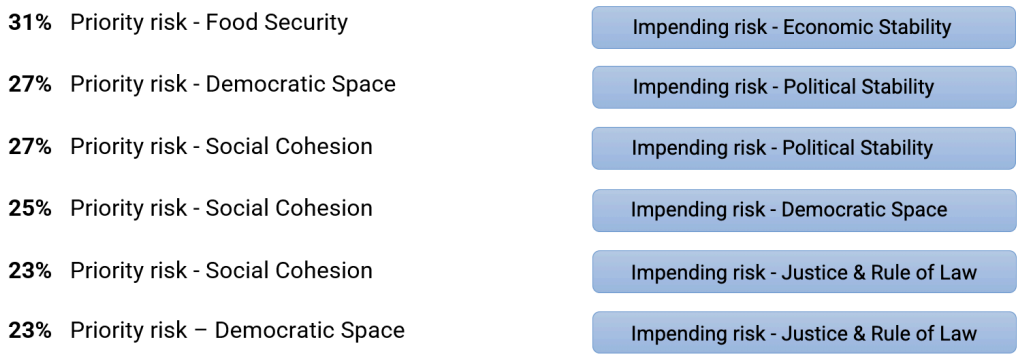


Figure 21: FP-Growth, Priority/Impending risks

The goal was to predict priority risks using answers to other questions. We discovered that there was insignificant correlation between the outcome variables and the predictor variables, as shown in Fig. 22. The metric used to determine the performance of models was the F1 score. Since the outcome is a multi-select variable, we looked at both precision and recall metrics.

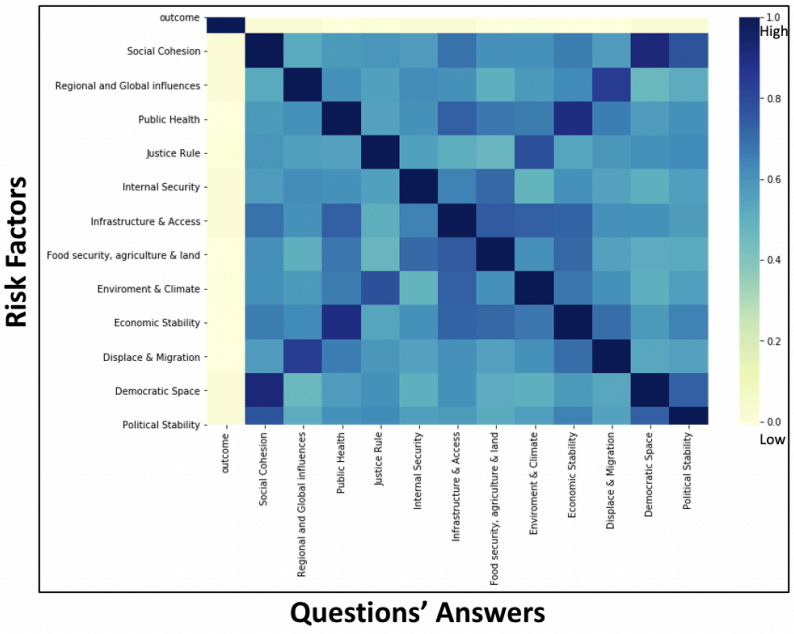


Figure 22: Heatmap (predictors/outcome)

Model Performance:
Random Forest = 50%
Naïve-Bayes = 10%

The reason behind the poor model performance is that there was a severe class imbalance. For e.g. for a certain risk factor "Infrastructure and access to social services", there were about 45 instances and for "Internal security", about 114 instances, which are low numbers compared to other risk factor categories.

6. Conclusion and Next Steps

1. We presented our results and insights to the UN using the dashboard created on Stream lit. All our analysis and insights were divided into different categories and presented as different tabs on the Stream lit dashboard.

2. There were notable patterns between priority and impending risks. The absence of social cohesion indicated a higher chance of economic/political stability.

3. Most questions were asked conditionally based on other answers. This led to a 1-1 mapping with the priority risk and hence were the best predictors.

4. We created scripts that automated the process of generating the dashboard that helps UN understand the survey results better.

5. It helped UN get a holistic picture of risks that countries under the RMR scanner are facing and their urgency.

References

[1] https://en.wikipedia.org/wiki/K-means_clustering
[2] United Nations Charter". www.un.org. 17 June 2015. Archived from the original on 18 March 2022. Retrieved 20 March 2022.
[3] https://sisudata.com/glossary/what-is-sparse-data
[4] https://www.softwaretestinghelp.com/fp-growth-algorithm-data-mining/