Preparing a Flood Risk Index for the State of Assam, India

Zeyu Chen
zc2796@nyu.edu

Bo Long
bl2665@nyu.edu

Qiuyi Wei
qw2316@nyu.edu

Abstract

To address the recurrent flooding in Assam, India, the government is seeking to identify the most susceptible areas to bolster its preparedness and response tactics. In our project, we utilized feature engineering, the Random Forest Classification model, and K-means clustering to categorize each Revenue Circle (an administrative division) in Assam according to the degree of flood damage experienced.

1 Introduction

The state of Assam in India is affected by floods every year. About 40% of the state is prone to floods and experiences widespread losses and damages to lives, livelihoods, and infrastructure annually. Past allocations for disaster relief funding have been heavily influenced by political factors [1]. Therefore, there is an urgent need for a tender allocation model that is based on the actual extent of flood damage or flood vulnerability.

Collaborated with Civic Data Lab, we received a dataset consisted of 64 flood related variables in a monthly granularity from May 2021 to Aug 2023 for each revenue circle. We developed two classification models to categorize the 180 revenue circles of Assam based on their vulnerability to floods. These models aim to assist in creating targeted interventions for improved flood preparedness, classifying each revenue circle with a fragility label to assess the risk and preparedness levels of each revenue circle.

2 Problem objective and algorithm

2.1 Problem objective

Guided by our mentor, we set our goal to classify each of the 180 Revenue Circles into six distinct categories based on their susceptibility to flood damage. These categories are: Very Low, Low, Medium, High, Very High, and Extreme. The model inputs include monthly flood-related data for each Revenue Circle (RC, administrative district), such as precipitation and elevation. The output is a specific vulnerability label assigned to each Revenue Circle.

2.2 Algorithm

Upon discovering significant intercorrelations among the variables, we opted to implement Principal Component Analysis (PCA) to address these correlations and diminish the dimensionality of our dataset. Subsequently, PCA factors were generated. We then employed Random Forest classification and K-means clustering techniques on these PCA-derived factors to evaluate flood vulnability for each RC in Assam state.

Preprint. Under review.

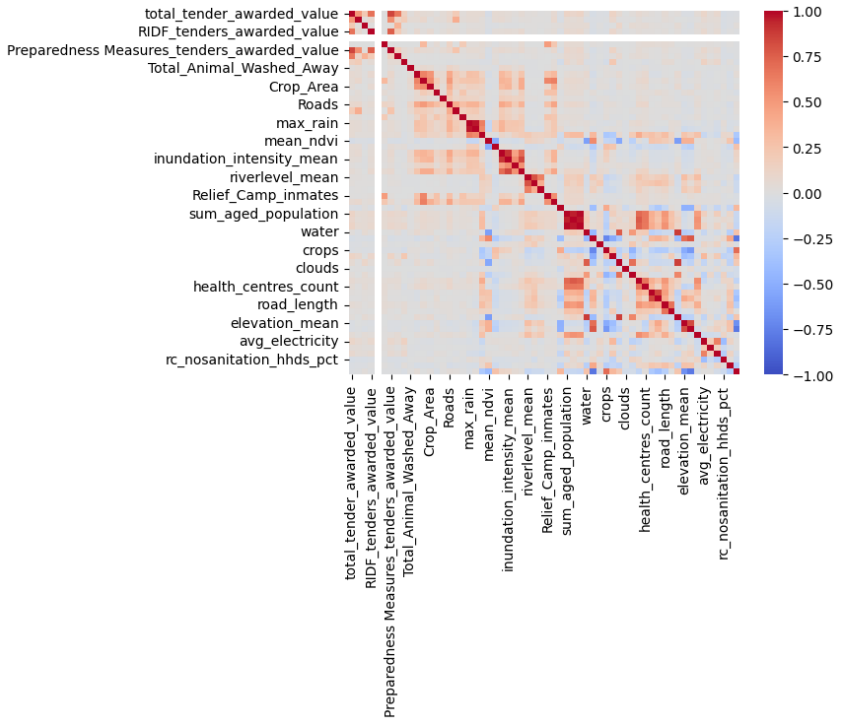

Figure 1: Correlation heat map

2.2.1 Exploratory data analysis

As we delved into the dataset, the most striking observation was the strong correlation present among the variables as shown in the correlation heat map, which can cause co-linearity problem.

2.2.2 Principal component analysis

Guided by mentorship and outcomes from the correlation analysis, we pinpointed the most suitable variable groups for Principal Component Analysis. Here's the process:

- Step 1: Identify the 20 most correlated variables for each of the 64 variables in the dataset.
- Step 2: Keep only variable pairs with a correlation above 0.4, and group these highly correlated variables.
- Step 3: Normalize the data, perform singular value decomposition on each identified variable group, and derive the corresponding PCA factors.

The generated PCA factors are treated as eleven indexes representing different domains with respect to flood: damage index, government investment index, inundation index, rain index, river index, infrastructure population index, road index, terrain index, land index, drainage index, electricity index. In our following steps, we are going to utilize these indices as the dataset for developing classification models.

2.2.3 Random forest classification

In harnessing the robustness of the Random Forest classification model, we began our analysis of flood damage by applying it to the PCA factors. We take the damage index as the target variable and use the remaining indices as independent variables. The strength of Random Forest lies in its ability to handle a large number of input variables and its effectiveness in classifying complex datasets, making it particularly suitable for our analysis involving diverse factors. This model excels in reducing overfitting, providing more reliable predictions compared to simpler models.

Consultations with our mentors highlighted that variables like total government awards and cumulative rainfall have a sustained influence on the effects of flooding. This realization underscored the need to incorporate their cumulative values into our analysis, ensuring a more comprehensive understanding of the long-term impacts of these factors. Such insights were pivotal in shaping our approach, allowing for a more nuanced and effective assessment of flood vulnerability across the Revenue Circles. Consequently, we conducted PCA on both the original dataset and a modified dataset that incorporated cumulative values for certain variables, allowing us to compare the performance of the models. 6

2.2.4 K-means clustering

Understanding that K-means clustering primarily groups data points based on spatial distances, we recognized its independence from concerns about collinearity among features. This characteristic of K-means clustering made it an ideal method for our analysis, particularly after the application of Principal Component Analysis (PCA). Consequently, we applied K-means clustering to both the row dataset and the PCA-derived factors, focusing on generating a comprehensive flood index.

However, it's important to note that K-means clustering inherently does not provide a natural ranking of the clusters; it groups entities based on similarities in features. So we have to assign label for each clustered group.

By this method, we were able to not only cluster the Revenue Circles based on their flood vulnerability, hazard, and resilience profiles but also rank these clusters in terms of their relative susceptibility to flood damage. This two-pronged approach of clustering and ranking offered a nuanced understanding of the flood risk landscape across Assam, facilitating targeted and effective intervention strategies. 5

3 Evaluation

In this section, we conduct a detailed evaluation of our methodology, including Random Forest classification and K-means clustering, and their performance in the analysis of flood vulnerability. We first focus on data processing, particularly on the significance of calculating cumulative values for key variables, and then move on to the specific application and assessment of these methods.

3.1 Data

3.1.1 DATA OVERVIEW

We are given a dataset consist of 64 flood related variables in a monthly granularity from May 2021 to Aug 2023 for each revenue circle, categorized into six different types:

- **Flood Proneness Variables:** elevation, slope, drainage density, etc.
- **Socio-economic Vulnerability Variables:** aged and young population, etc.
- **Demographic Variables:** total population, sex ratio, etc.
- **Environmental Vulnerability Variables:** number of roadways, schools, hospitals, etc.
- **Government Responses Variables:** number of relief camps, monetary assistance, etc.
- **Damages and Losses variables:** lives lost, roads and bridges damaged, etc.

3.1.2 Data processing

In our study, we focused on calculating cumulative values for certain key variables to better understand their long-term impact on flood effects. For this purpose, we employed a function named `rolling_cumulate`, which performs a rolling cumulative calculation over a specified time window (defaulted to 3 months) for selected columns. The primary variables we concentrated on for cumulative values were total rainfall (`sum_rain`) and total tender awarded value (`total_tender_awarded_value`). This method helped us capture the accumulative changes of these variables over time, thereby providing a richer data perspective for analyzing flood vulnerability.

3.2 Methodology

3.2.1 Random forest classification

The evaluation of the Random Forest classification model's performance, as indicated by the provided metrics, reveals a nuanced understanding of its predictive capabilities.

Table 1: Classification performance metrics

Class	Precision	Recall	F1-Score	Support
Very Low	0.855662	0.974705	0.911313	1186
Low	0.142857	0.044118	0.067416	68
Median	0.200000	0.069767	0.103448	43
High	0.250000	0.037037	0.064516	54
Very High	0.444444	0.063492	0.111111	63
Extreme	0.166667	0.230769	0.193548	26
			Accuracy	0.815278

Table 2: Classification performance metrics with Cumulative Variables

Class	Precision	Recall	F1-Score	Support
Very Low	0.852149	0.986509	0.914420	1186
Low	0.285714	0.058824	0.097561	68
Median	0.333333	0.046512	0.081633	43
High	0.500000	0.055556	0.100000	54
Very High	0.200000	0.031746	0.054795	63
Extreme	0.161290	0.192308	0.175439	26
			Accuracy	0.823611

- Precision: gauges the accuracy of positive predictions, shows notable improvement across most classes when cumulative variables are introduced, suggesting enhanced accuracy in correctly identifying positive instances. This is particularly evident in classes with lower instances, reflecting the model's refined ability to discriminate.
- Recall: indicates the proportion of actual positives correctly identified, also improves in the 'Very Low' and 'Extreme' classes with the addition of cumulative variables. This suggests a better capture of all relevant instances in these categories.
- F1-score: a critical metric that harmonizes precision and recall. It generally shows an upward trend, especially for the 'Very Low' class, signaling a more balanced performance.

In summary, while the classifier appears to have a high overall accuracy, this metric is skewed by the high performance on the very low class, which has a disproportionately large number of instances. The model struggles with all other classes, as evidenced by low F1-scores. This is an indication of an imbalanced dataset, likely attributable to over 80% of the months lacking flood events for each revenue circle.
Notably, models utilizing cumulative variables demonstrate superior performance. Overall accuracy of the model sees a modest improvement, rising from 81.5278% to 82.3611%, underscoring the positive impact of including cumulative variables in the analysis. This reinforces the approach's validity, recommending its continued application in our analysis.

3.2.2 K-means clustering

In light of the insights gained from the Random Forest classification and its emphasis on the importance of certain variables in flood damage prediction, we tailored our approach to K-means clustering. Recognizing that K-means is fundamentally a distance-based algorithm, we decided to apply it in two distinct ways:

1. Clustering with Raw Dataset: First, we applied K-means clustering directly to the raw dataset. This step was taken to explore the natural groupings within the data based on the original variables, and to generate a flood index. Despite the presence of collinearity in the raw dataset, K-means clustering's

focus on spatial relationships between data points allows it to effectively identify clusters without being significantly hindered by collinearity issues.

2. Clustering with PCA-derived Indices: Secondly, we applied K-means clustering to the PCA-derived indices. The PCA process effectively condensed the dataset into factors that represent key domains related to flood, such as damage index, rainfall index, and inundation index. By clustering on these PCA factors, we aimed to generate a flood index that developed by employing all the relevant factors from PCA, including the critical damage index. The K-means algorithm was instrumental in clustering the Revenue Circles based on similarities in these factors, effectively grouping them according to their shared characteristics in flood vulnerability, flood resilience, and flood hazard.

To generate a ranking for clustered groups, we aggregated key indices derived from PCA – such as the government investment index, damage index, inundation index, and rain index – to create a composite score. This aggregated score was then used to rank the K-means groups.

Based on the analysis of clustering outcomes from both the raw dataset and the PCA-derived data, we opted to proceed exclusively with the PCA-based method. This decision was influenced by the observation that the raw dataset clustering was adversely affected by data imbalance, leading to suboptimal centroid positioning. In contrast, the PCA-based clustering did not exhibit this issue and effectively encapsulated key characteristics such as damage, inundation, and rainfall for each Revenue Circle.

Instead of assigning a precise ranking to each Revenue Circle (RC) from 1 to 180, we employed K-means clustering to categorize them into 6 groups, ranging from 1 to 6. This approach is deemed appropriate as it is challenging to distinctly prioritize RCs when they exhibit similar levels of damage and characteristics in other aspects.

3.3 Results

The resulting classification from K-means clustering is justifiable as the flood index aptly mirrors the extent of damage, inundation levels, and rainfall in specific areas. The flood index from K-means clustering also coincide with the result from random forest classification. Furthermore, it reveals that previous tender allocations were not solely based on damage levels, as some of the most damaged areas sometimes don't receive tenders. In June 2022, the southern region of Assam, which has a high population density 4, experienced considerable flood damage and inundation. This area, despite its high flood index, notably did not receive adequate tender allocations.

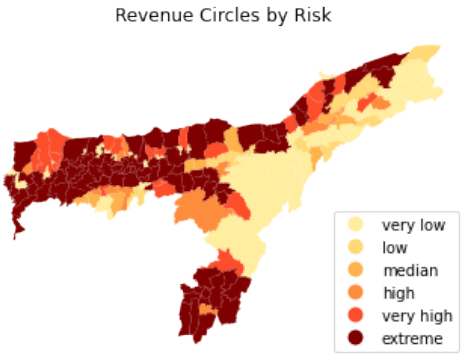

Figure 2: Random forest classification result

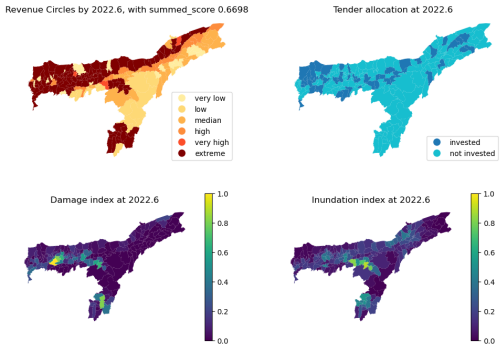

Figure 3: K-means clustering result

4 Conclusions and future step

The frequent and devastating floods in Assam, India, necessitate a shift from politically-influenced disaster relief allocations to a more data-driven approach. We developed two classification models to assess the flood vulnerability of Assam's 180 revenue circles represents a critical step towards this goal. Our Random Forest classifier shows high accuracy and make acceptable flood vulnerability

map, but the model accuracy is skewed by its performance on the predominant 'very low' class. The K-means results effectively mirror real-world damage, inundation, and rainfall, highlighting discrepancies in historical tender distributions. Nonetheless, establishing a method for validation remains a crucial next step. Additionally, a future enhancement could involve creating a flood probability model 7. This model, when combined with the flood vulnerability data obtained from the Random Forest analysis, could yield a new flood index. This new index could then be used to cross-verify the flood index obtained from the K-means clustering.

5 Lessons learned

The dataset utilized for this project, in contrast to the more refined datasets typically encountered in academic settings, presented several challenges. Comprising 5040 rows and 64 columns, its relatively small size and high dimensionality raised concerns about potential overfitting. Additionally, the dataset was markedly imbalanced, a common issue in real-world data. This imbalance was particularly evident as most revenue circles did not experience flood impact in the majority of months, leading to a skewed distribution of outcomes. This resulted in the Random Forest classification model displaying a significant imbalance in its results, with an exceedingly high accuracy for the 'Very Low' label and comparatively lower accuracy for other labels. To address these challenges, we invested considerable effort and experimented with numerous feature engineering techniques, although only a few proved to be effective. This process involved a rigorous trial-and-error approach, where we meticulously tested different methods to determine which ones could successfully enhance the model's performance given the dataset's complexities.

Despite these challenges, a notable improvement in the model's performance was achieved by integrating cumulative values for certain variables. This strategy proved effective in mitigating some of the issues posed by the dataset's limitations, particularly in addressing the imbalanced nature of the data. The key takeaway from this experience is the importance of adapting data preprocessing and feature engineering techniques to suit the specific challenges of a dataset. By thoughtfully adjusting our approach to accommodate the dataset's characteristics, we were able to enhance the model's accuracy and derive more meaningful insights from the analysis.

6 Student contributions

Bo Long: Conducted principal component analysis and random forest classification. Wrote the corresponding parts of the report.

Qiuyi Wei: Conducted EDA, regression, visualization, and the K-mean clustering. Wrote the corresponding parts of the report.

Zeyu Chen: Conducted data preprocessing and cleaning, feature engineering, visualization, and the random forest classification. Wrote the corresponding parts of the report.

References

[1] Cole, S., Healy, A. & Werker, E. (2012) Do voters demand responsive governments? Evidence from Indian disaster relief. In *Journal of Development Economics*, **97**(2):167-181. Elsevier. `https://doi.org/10.1016/j.jdeveco.2011.05.005`.

[2] Badhe, Y.P., Medhe, R.S. & Shelar, T. (2019) Site Suitability Analysis for Water Conservation Using AHP and GIS Techniques: A Case Study of Upper Sina River Catchment, Ahmednagar (India). *Hydrospatial Analysis* **3**(2):49-59.

[3] Das, S. (2019) Comparison among influencing factor, frequency ratio, and analytical hierarchy process techniques for groundwater potential zonation in Vaitarna basin, Maharashtra, India. *Groundwater for Sustainable Development* **8**, 617-629. DOI: `https://doi.org/10.1016/j.gsd.2019.03.003`

[4] Mukhopadhaya, S. (2016) GIS-based site suitability analysis: case study for professional college in Dehradun. *Journal of Civil Engineering and Environmental Technology* **3**(1):60-64.

[5] Anbazhagan, S., Ramasamy, S.M. & Gupta, S.D. (2005) Remote sensing and GIS for artificial recharge study, runoff estimation and planning in Ayyar basin, Tamil Nadu, India. *Environmental Geology* **48**(2):158-170. DOI: `https://doi.org/10.1007/s00254-005-1284-4`

Acknowledgements

We would like to express our deep gratitude to our mentors for their invaluable assistance throughout our project. Special thanks to Dr. Brian McFee from New York University, whose expertise and insights have greatly contributed to our research. We are also immensely grateful to Sai Krishna Dammalapati, Meenu Francis, and Jeeno George from Civic Data Lab for their continuous support and guidance. Their dedication and commitment have been a significant driving force behind the success of this project.

Appendix

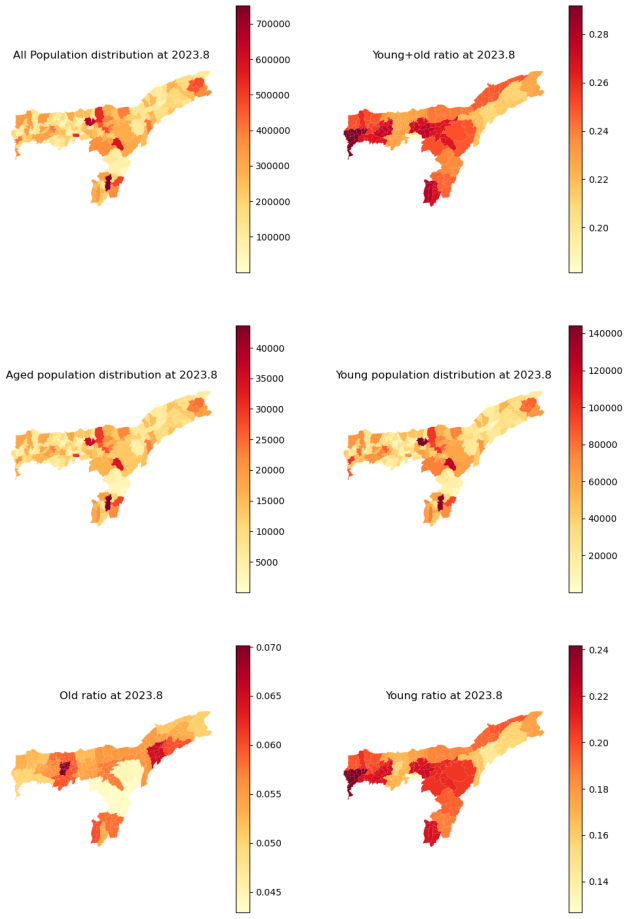

Figure 4: Population distribution for Assam State

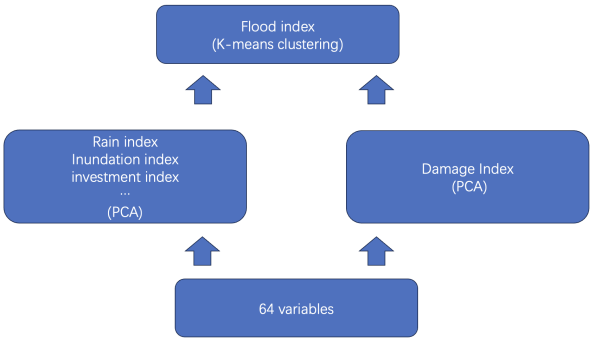

Figure 5: K-means model structure

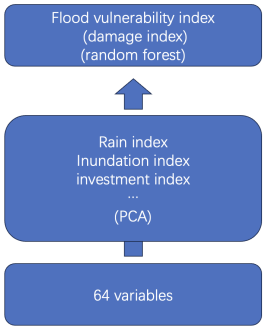

Figure 6: Random forest model structure

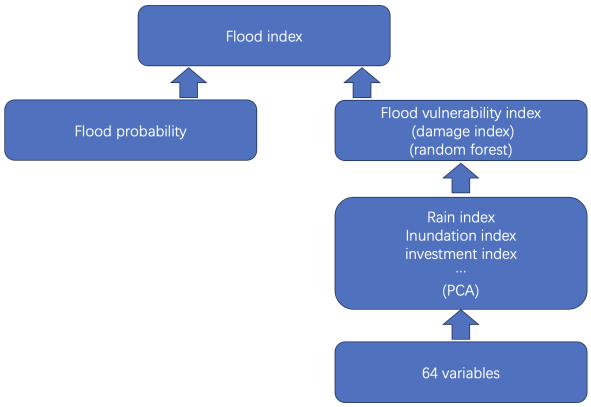

Figure 7: Possible future step