A Hybrid Machine Learning Approach for Identifying Flood Debris Drivers and Generation

Jasmine H. Bekkaye[1], Navid H. Jafari[1]

[1]Department of Civil & Environmental Engineering, Louisiana State University, Baton Rouge, LA, jbekka1@lsu.edu

LSU College of Engineering

LSU Department of Civil & Environmental Engineering

Background & Objectives

- Natural hazards generate tremendous amounts of debris waste that negatively impacts communities.
- Current disaster debris prediction models are inaccurate and have yet to utilize unsupervised algorithms, which could help guide flood debris models.
- The objective of this study is to demonstrate a hybrid unsupervised and supervised machine learning approach for understanding the relationships and drivers influencing flood debris quantities across a region using post-disaster waste data acquired in Beaumont, TX, after Hurricane Harvey.

Methodology

1. Aggregate post-disaster waste data to census block level.
2. Identify and characterize flood debris drivers.
3. Apply K-means clustering[1] to the waste data to create high, medium, and low debris clusters.
4. Conduct statistical testing to evaluate patterns between flood waste and drivers.
5. Build a Random Forest[2] (RF) classification model using 10-fold cross-validation and the clustered debris tonnage data.
6. Evaluate the model's performance for debris prediction.
7. Investigate variable importance in model construction.

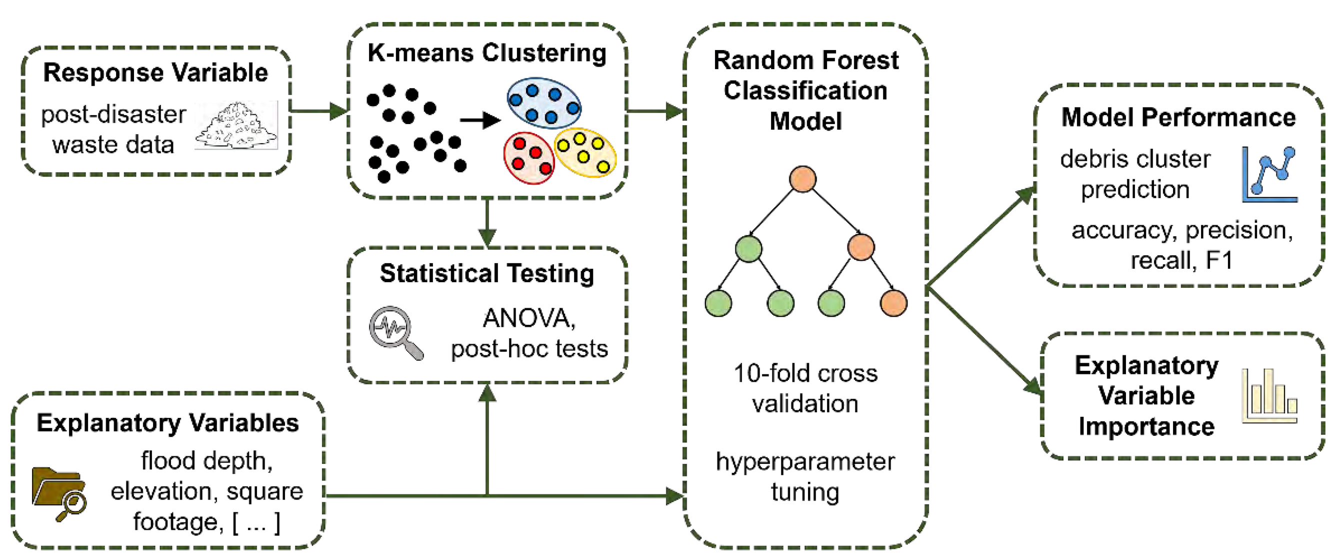


Figure 1. Hybrid machine learning framework for identifying flood debris drivers and generation.

K-Means Clustering

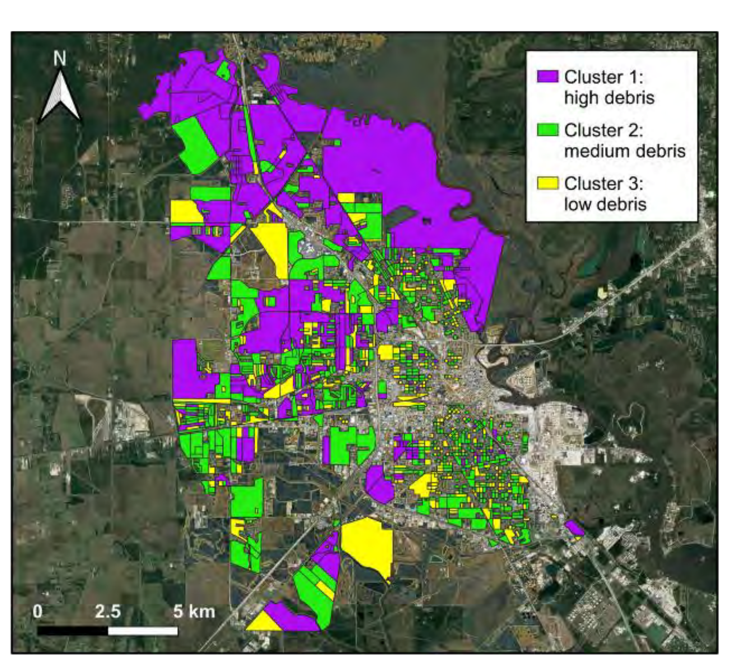


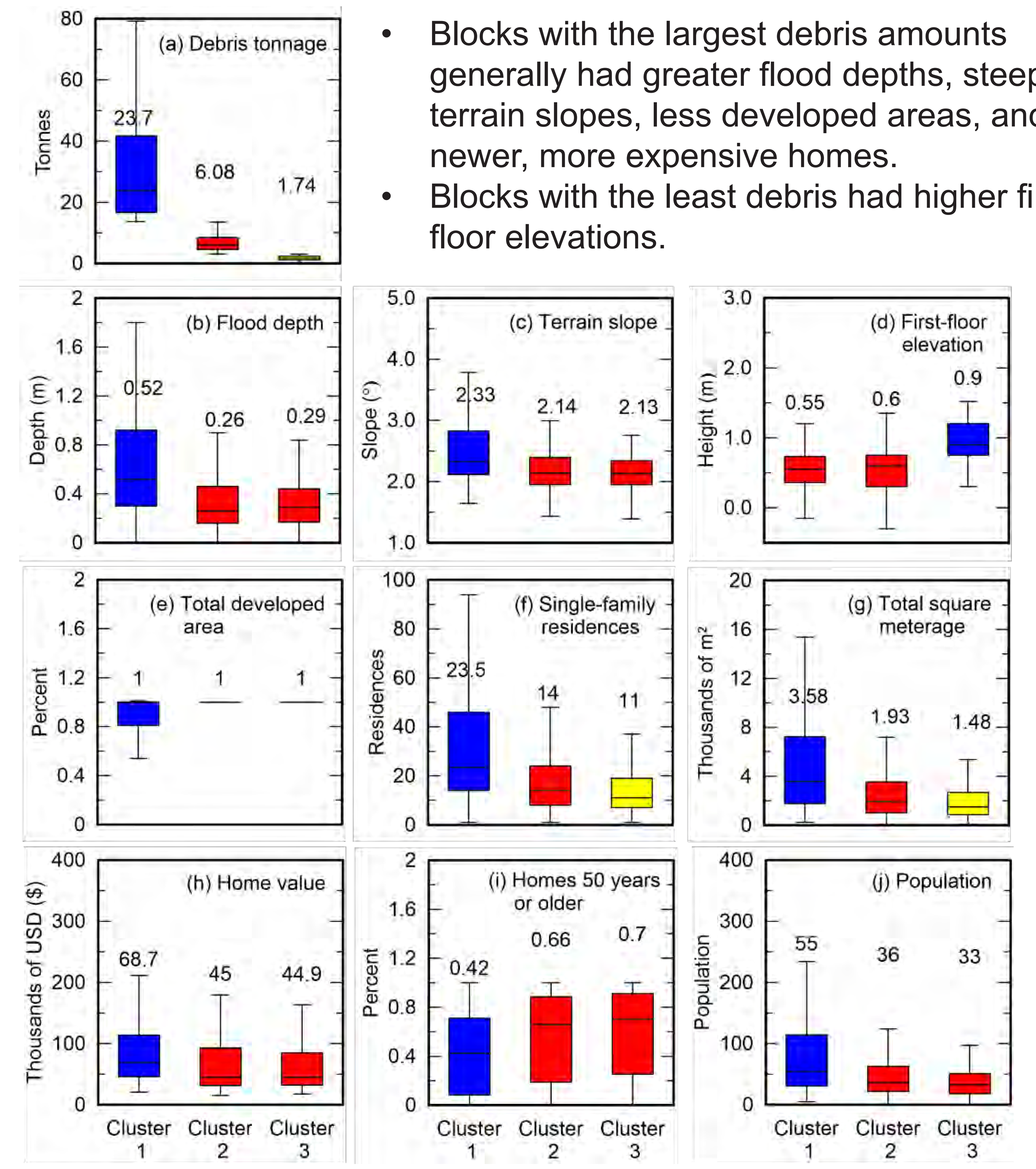
Figure 2. Spatial distribution of debris tonnage clusters.

Table 1. Summary statistics of tonnage clusters.

Cluster	n	Mean	SD	Median	Min	Max
1	182	47.4	91.3	23.7	13.7	882
2	497	6.74	2.76	6.08	3.08	13.6
3	306	1.71	0.74	1.74	0.07	3.03

- Most Cluster 1 blocks are in northern Beaumont adjacent to major rivers and floodplains.
- Census blocks in Clusters 2 and 3 are evenly dispersed and become more prominent towards the city center.

Statistical Testing of Clusters



- Blocks with the largest debris amounts generally had greater flood depths, steeper terrain slopes, less developed areas, and newer, more expensive homes.
- Blocks with the least debris had higher first-floor elevations.

Figure 3. Box plots illustrating debris tonnage and statistically significant explanatory variables. Variables with distinct distributions have a different color than other clusters and clusters with similar distributions are the same color.

Random Forest Model

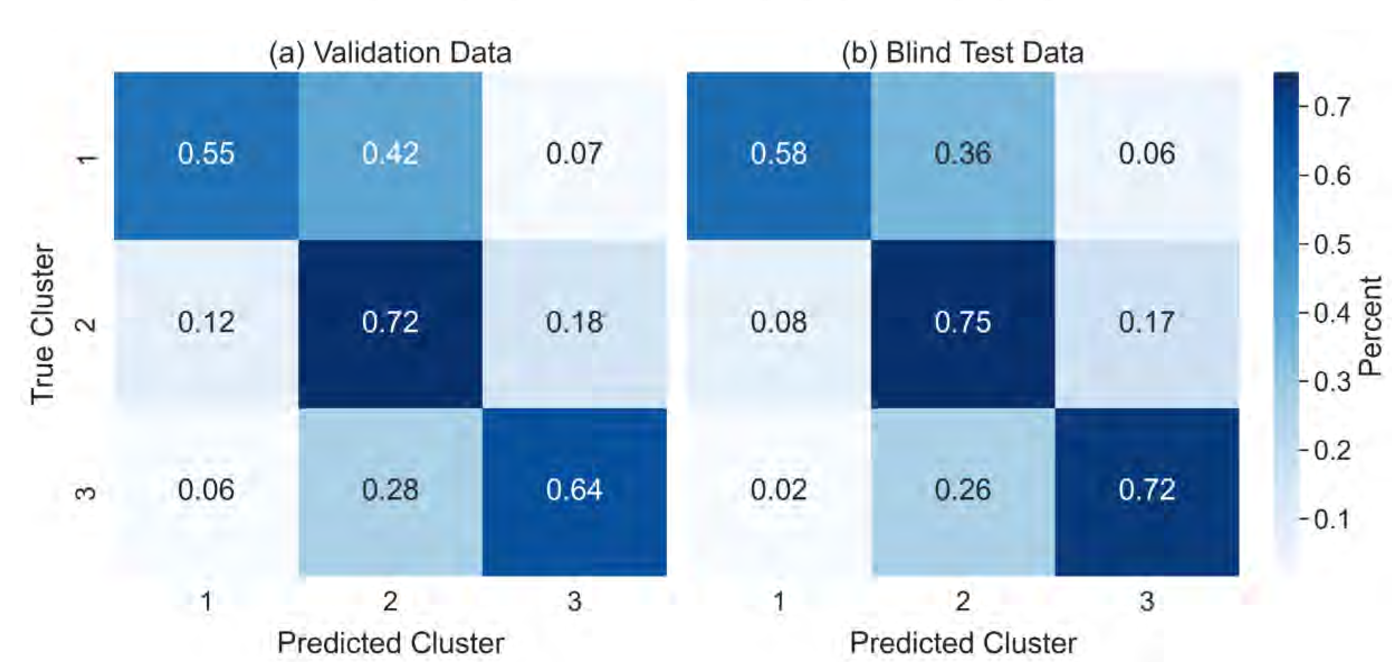


Figure 4. Confusion matrix for the RF model for the validation data (a) and the blind test data (b).

- Mean model accuracy: Validation data - 65.5%; Blind test data - 71.1%.
- Most accurately predicted Cluster 2 (success rate of 72 – 75%).
- Least accurately predicted Cluster 1 (success rate of 55 – 58%).
- Instances were frequently misclassified as Cluster 2, whereas instances were much less often misclassified as Clusters 1 and 3.

Acknowledgements

This work was supported by NSF through a Graduate Research Fellowship to Jasmine H. Bekkaye and through a RAPID grant (Award 1760718) that facilitated data collection.

Variable Importance

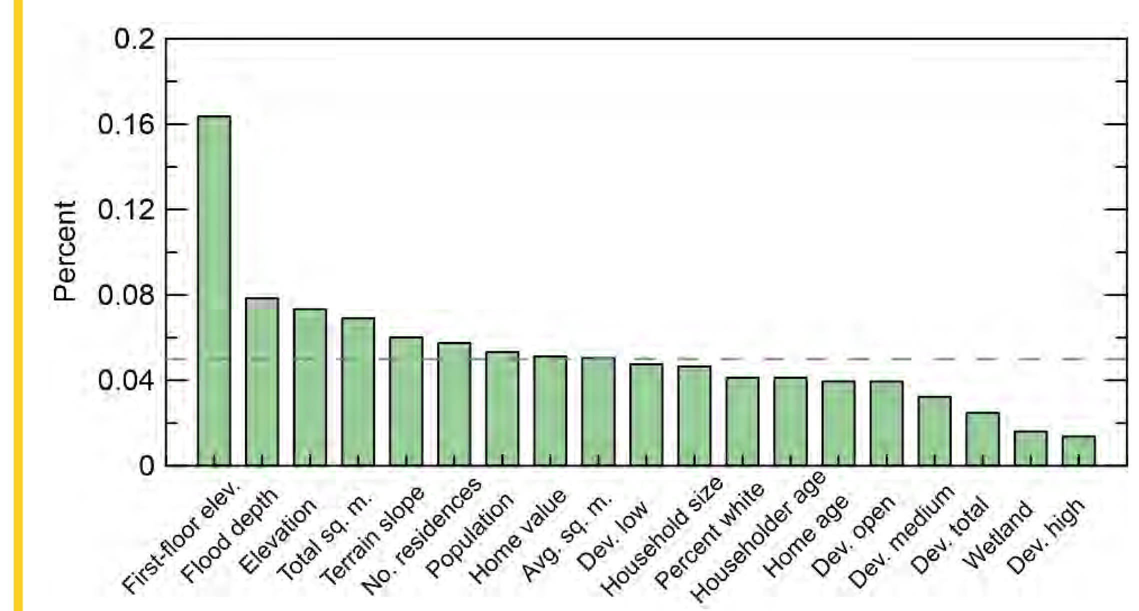


Figure 5. Relative variable importance in the RF model.

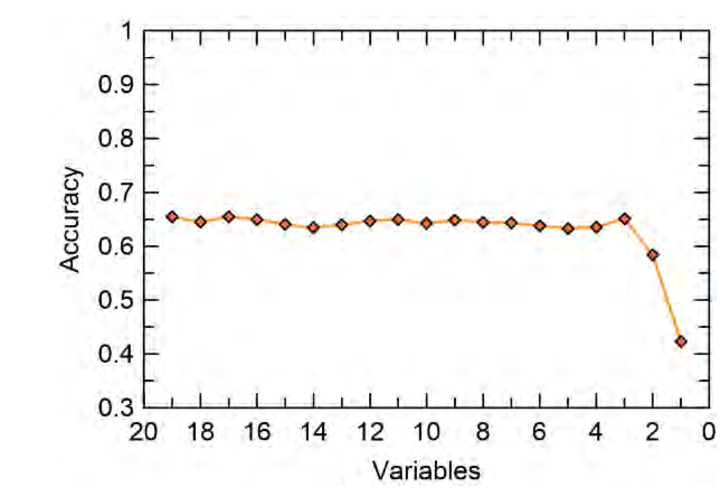


Figure 6. Removing variables in the RF model and the corresponding accuracy.

- Recursive feature elimination revealed that first-floor elevation, bare earth elevation, and total square meterage were the most important predictors of flood debris.
- All can be acquired pre-disaster, boosting debris prediction efforts.

Hazus Comparison

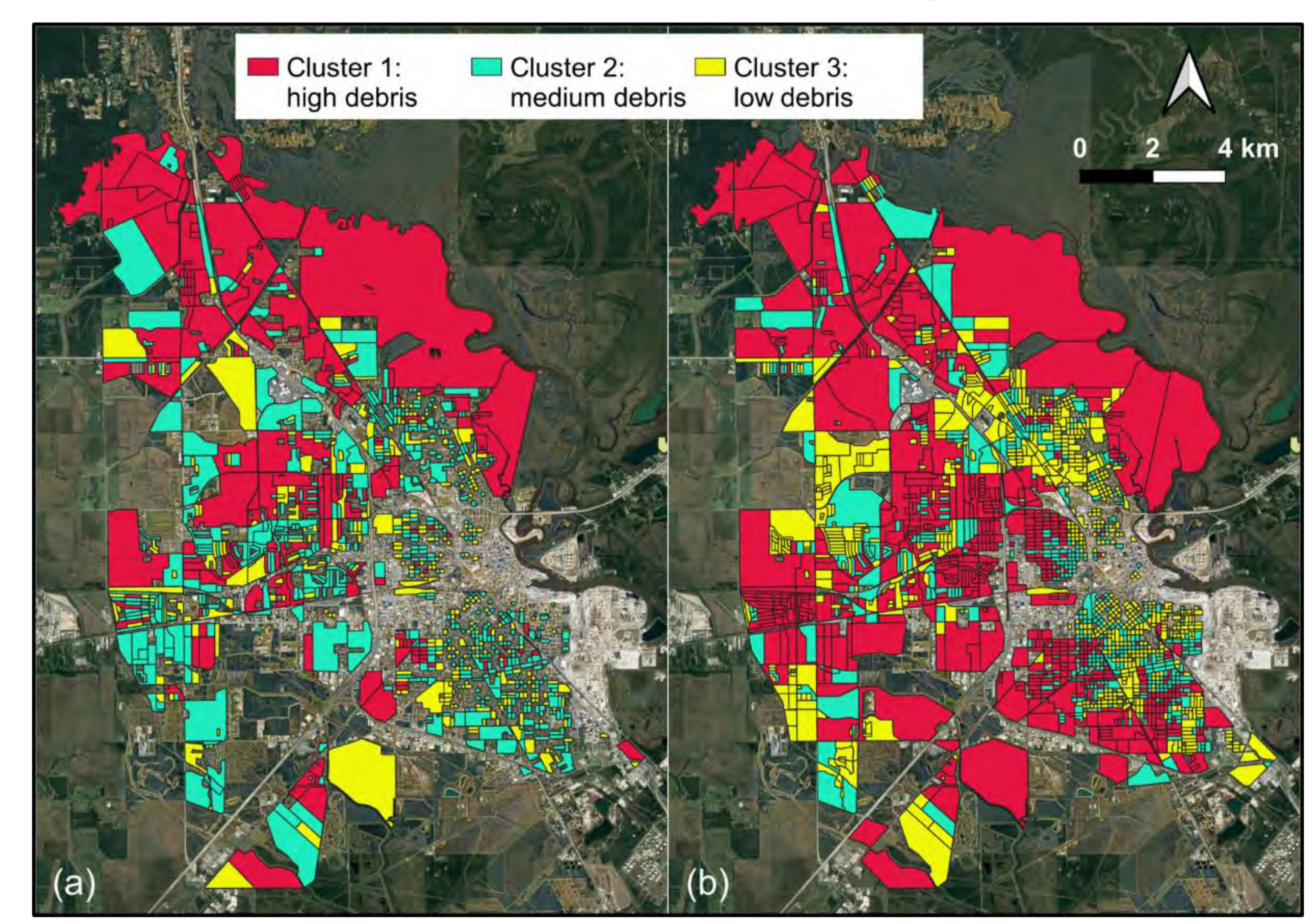


Figure 7. Spatial distribution of debris tonnage clusters from this study (a) and from Hazus estimates (b).

- Flood debris estimates generated using Hazus Flood Model.
- Hazus significantly overpredicts debris quantities (611 high debris census blocks compared to 182 high debris blocks).
- Maps of clustered flood debris predictions can aid in disaster debris removal and management operations.

Summary and Conclusions

- A hybrid machine learning approach can provide a good first step towards reducing uncertainty in predicting disaster debris.
- First-floor elevation emerged as a significant driver of flood debris generation but is not currently considered in prediction models.

References

[1] MacQueen, J. B. (1967). "Some methods for classification and analysis of multivariate observations." In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability,* (1), 81–297. California: University of California Press.

[2] Breiman, L. (2001). "Random Forests." *Machine Learning,* (45), 5–32.