Analyzing Houston Flooding Using Unsupervised Machine Learning

Etsuko Ishii¹,², Isaac Dykeman³, Zouhan Chen³, Devika Subramanian³

¹Department of Mathematical Engineering and Information Physics, the University of Tokyo, Japan
²Nakatani RIES: Research and International Experiences for Student, Nakatani Foundation, Japan
³Department of Computer Science, Rice University, U.S.A.

Flooding in Houston

Objective
This research aims at establishing a model which
• Predicts flooding at the home level in Houston based on integrating a range of readily available data sources: rainfall, 311 calls, flood plain designation, digital elevation models
• Uses unsupervised machine learning techniques to identify the key determiners of home flooding.

Conclusions
• Unsupervised machine learning identifies a cluster of over 50% of homes that have experienced multiple floods in Houston during 2011-2015 that are far from a bayou and that are not in FEMA designated flood plains.
• LIDAR analysis of these homes (with a 200mx200m tile around home) reveals that these regions are characterized by high density development and proximity to large tracts of land at low elevation.
• Further analysis of these homes at different scales (50mx50m to 500mx500m) is needed to identify key factors that cause flooding.

Unsupervised Machine Learning

Goal: identify groups of 311 call locations based on rainfall, bayou distance, floodplain, call frequency by K-means clustering
• Normalized input data (5000 x 8 features)
• Classified into three clusters (elbow method)

2nd K-means clustering: identify LIDAR signature of unusual cluster uncovered by 1st K-means clustering
• Extracted LIDAR DSM data for those 311 call locations
• Scaled to lie between given minimum and maximum value
• Classified into four clusters by K-means

Future Work
• 2nd K-means clustering by other features (LIDAR DEM, Digital Elevation Model)
  - road features
  - water pipe/ drainage/ sewage network data
  - land use features
• Establishing a fast and accurate flood prediction model at home level using diverse data sets rather than by running detailed physical simulations.

References