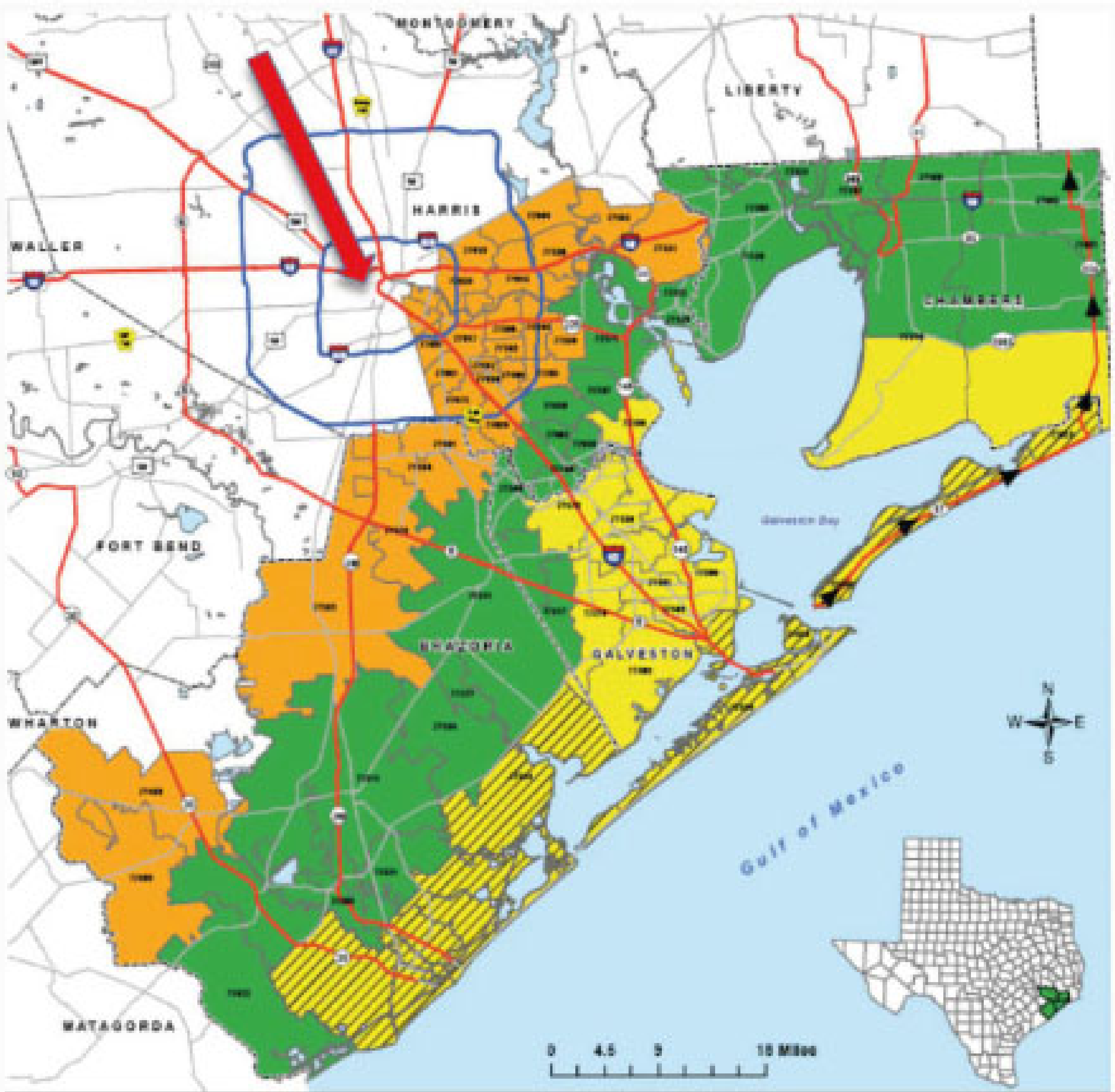
Using big data and machine learning to build spatially fine-grained models of hurricane wind and flood damage risk

Devika Subramanian and Isaac Dykeman and Etsuko Ishii

Project collaborators: Leonardo Duenas-Osorio, Bob Stein, Rick Wilson (Rice)

Motivation

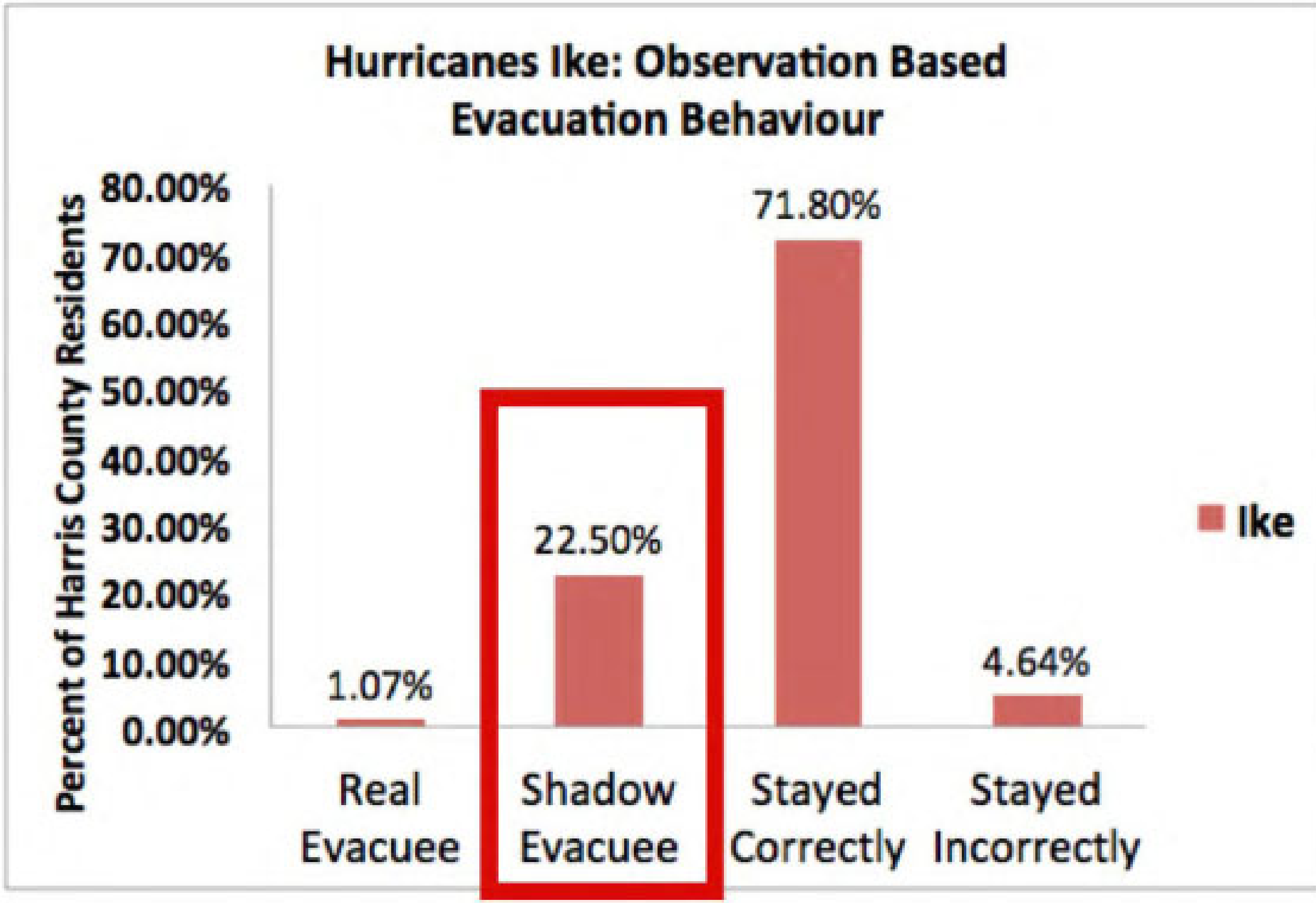


- No information about hurricane induced wind or flood risk is made available to people in the white zone.
- The white zone is where the bulk of the Harris County population resides.

Evacuation zones based on storm surge risks alone!

 5/9/18

Shadow Evacuation



- Real evacuee = at risk and decided to evacuate

  Shadow evacuee = not at risk and decided to evacuate

  Stayed correctly = not at risk and decided to shelter in place

  Stayed incorrectly = at risk and decided to shelter in place

Shadow evacuation math

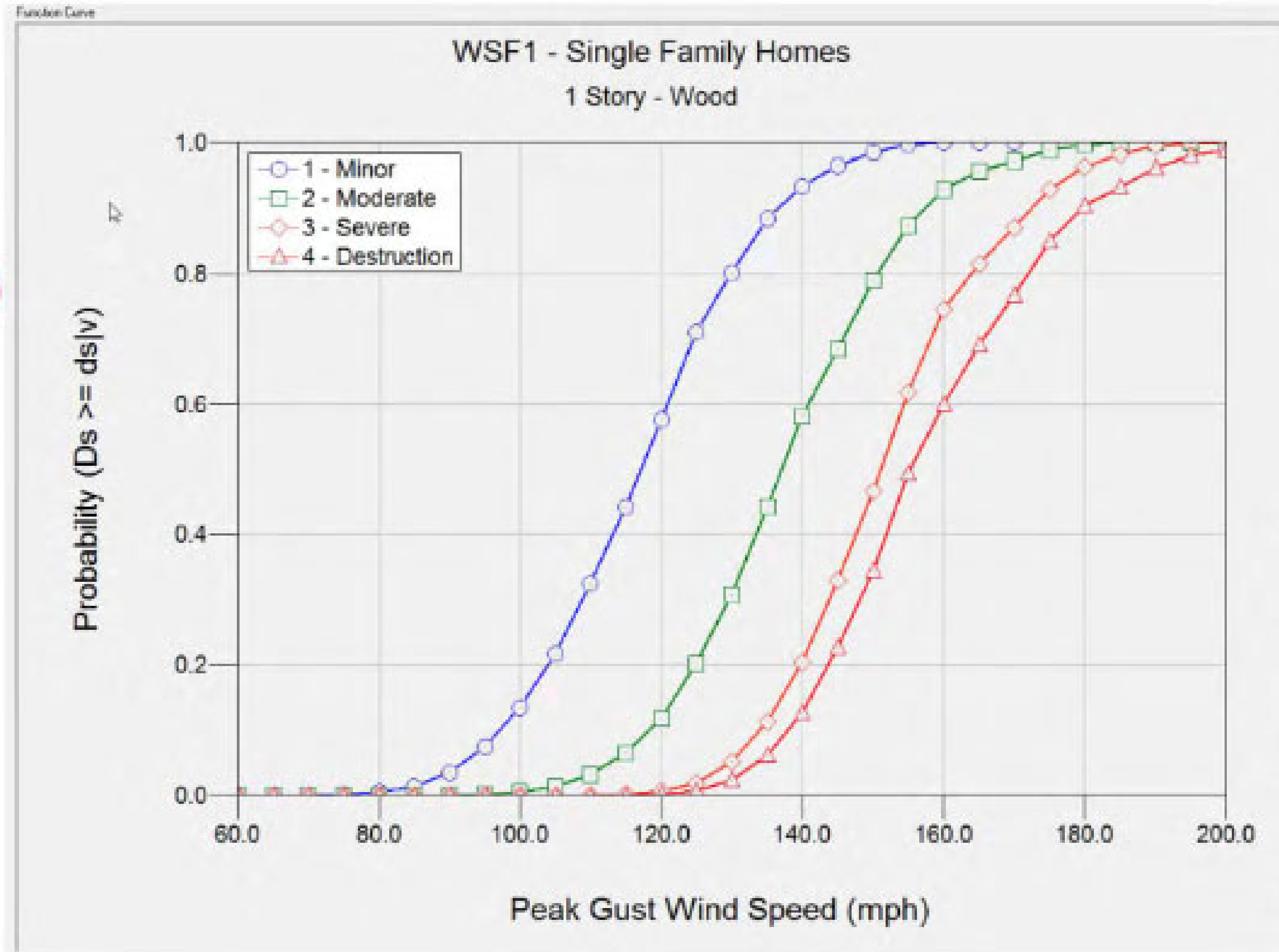
- The Houston transportation network handles 750,000 commuters everyday.
- But, between 100,000 - 500,000 in Galveston and other coastal areas in storm surge risk zones need to evacuate through Houston.
- 100,000 - 500,000+ shadow evacuators in Houston can really stress the transportation system.

Approaches for predicting wind risk

- Physics-based models: Build engineering models that estimate risk to a structure due to wind, rain (HAZUS-MH4) (based on structural engineering, physical modeling and expert opinion).
- Data-driven models: Learn models that predict risk based on large-scale observational data. (empirical, data-driven approach)

Fragility curve models (HAZUS-MH4)

Frame (4 types)
Floors (2 types)
Roof shape (2 types)
Nailing pattern (3 types)
Roof-deck attachment (2)
Garage type (3 types)
Terrain (5 types)
Wind speed



Pinelli 2004, Vickery et.al. 2004

http://etd.lsu.edu/docs/available/etd-04152009-092337/unrestricted/Friedland_dis.pdf

How well does this model perform?

- Hurricane Ike struck Galveston/Houston in

Natural Hazards Review / Volume 15 Issue 3 - August 2014

Technical Papers

DETAILS FIGURES REFERENCES RELATED

Building and Validating Geographically Refined Hurricane Wind Risk Models for Residential Structures

Devika Subramanian; Josue Salazar; Leonardo Duenas-Osorio, A.M.ASCE; and Robert Stein

FULL TEXT DOWNLOAD TOOLS SHARE

Abstract

Accurate estimation of risk to residential structures from hurricane winds is critical for emergency planning and post-event recovery. Models based on fragility curves are widely used for assessing wind damage risk at the county and census tract levels. Large-scale evaluation of the predictive accuracy of these models has

Authors

Devika Subramanian
Professor, Dept. of Computer Science, Rice Univ., 6100 Main St. MS 132, Houston TX 77005 (corresponding author). E-mail: devika@rice.edu

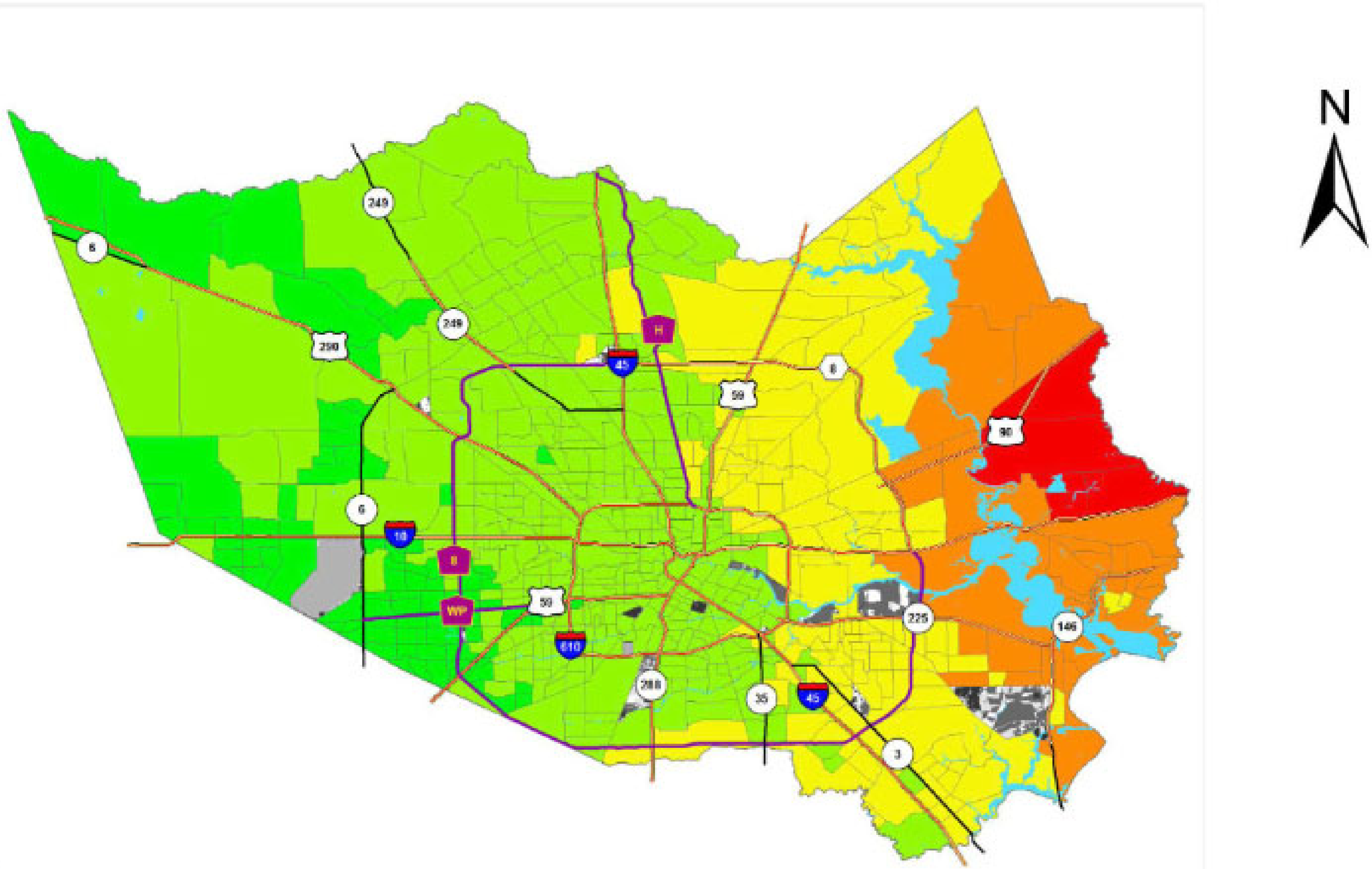
Josue Salazar
Graduate Student, Dept. of Computer Science, Rice Univ., 6100 Main St. MS 132, Houston TX 77005. E-mail: salazar.josue@rice.edu

Leonardo Duenas-Osorio, A.M.ASCE
Associate Professor, Dept. of Civil Engineering, Rice Univ., Houston TX 77005.

Robert Stein
Lena Gohlman Fox Professor, Dept. of Political Science, Rice Univ., Houston TX 77005.

HAZUS-MH4 wind risk model for Hurricane Ike

Census
Tract
Level

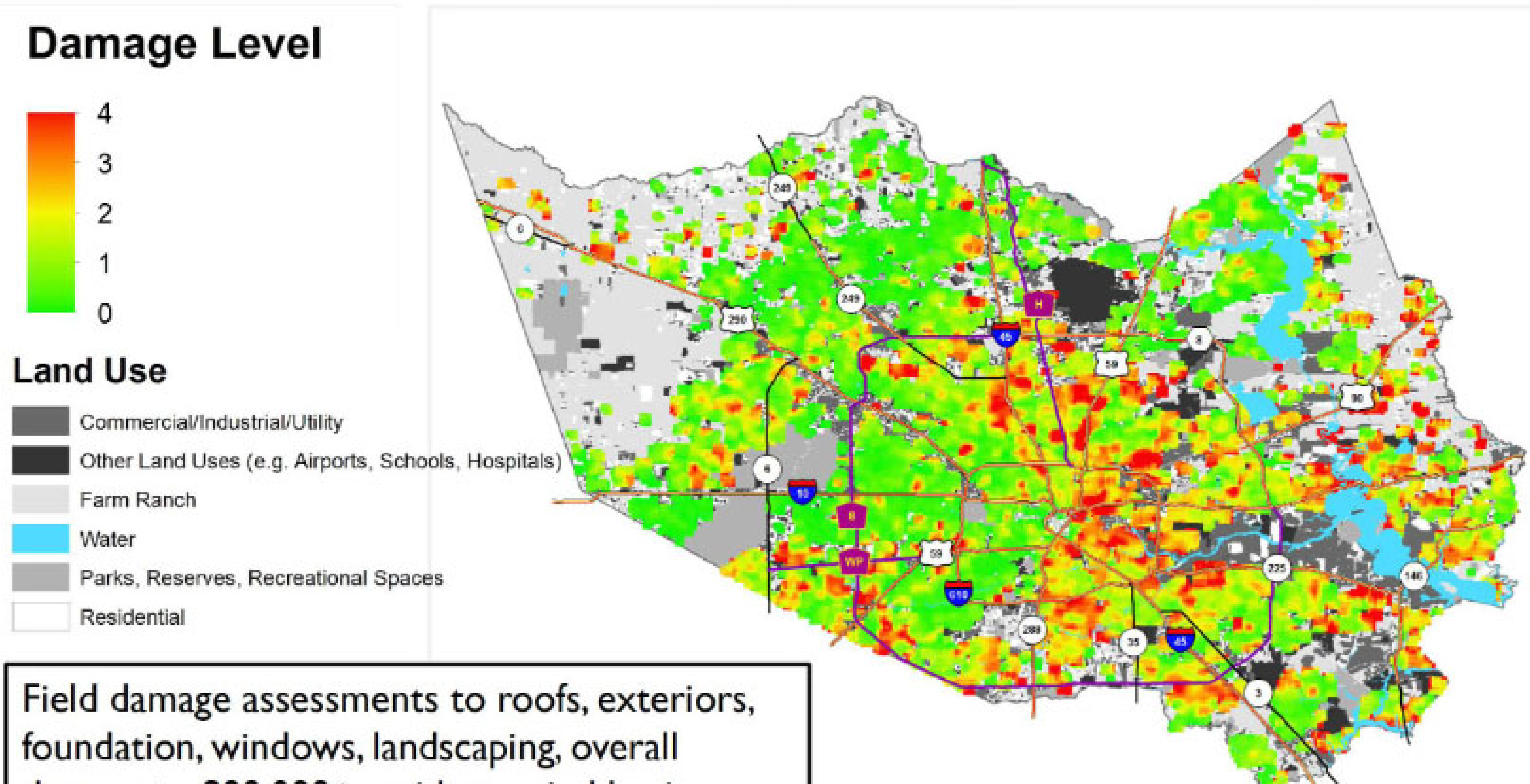


Damage %

- [0,10]
- (10,20]
- (20,30]
- (30,40]
- (40,100]
- No Data

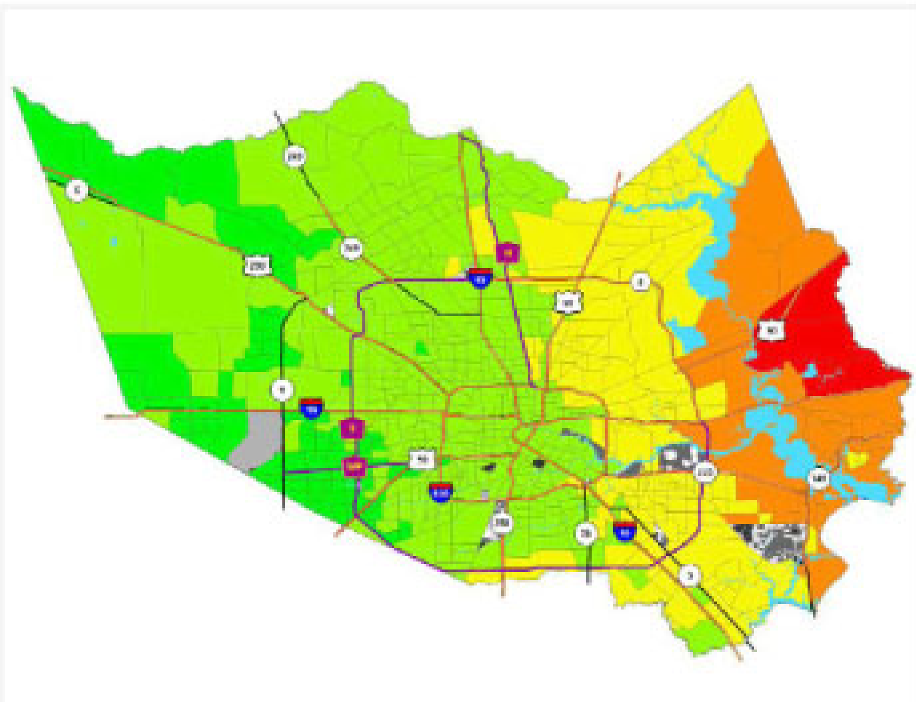
Observed wind damage from Hurricane Ike

Source: Harris County Housing Authority, 2008

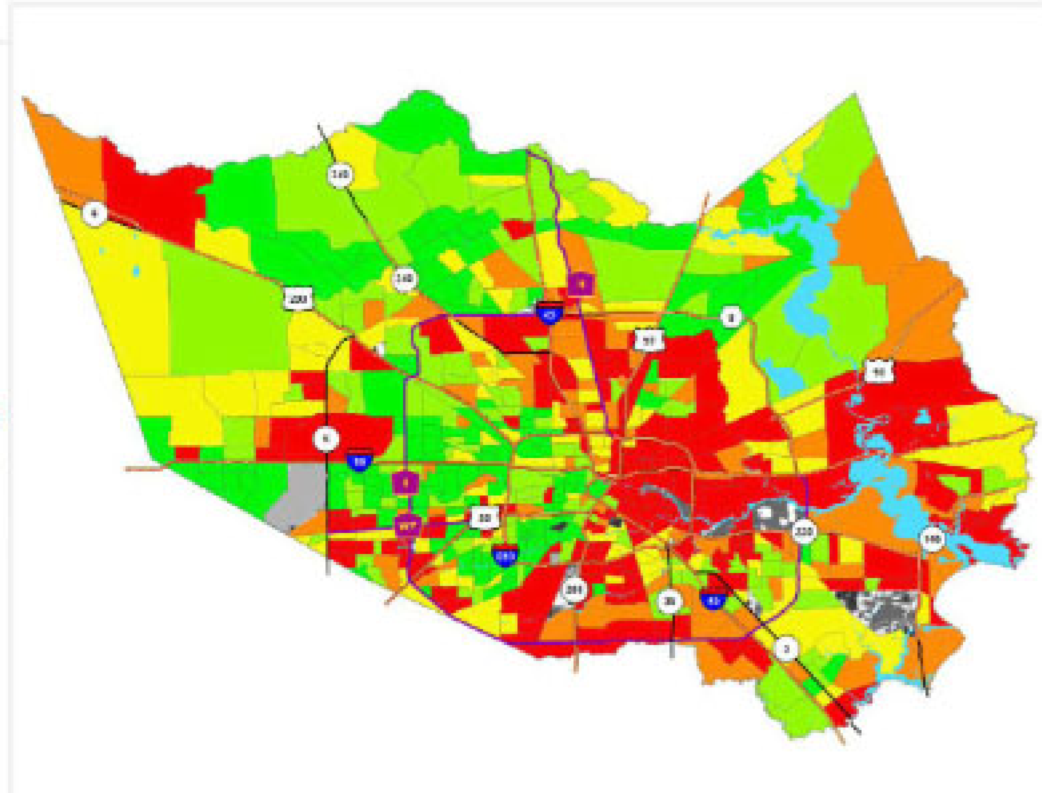


Field damage assessments to roofs, exteriors, foundation, windows, landscaping, overall damage to 800,000+ residences in Harris County

A side by side comparison at the census tract level



Predicted



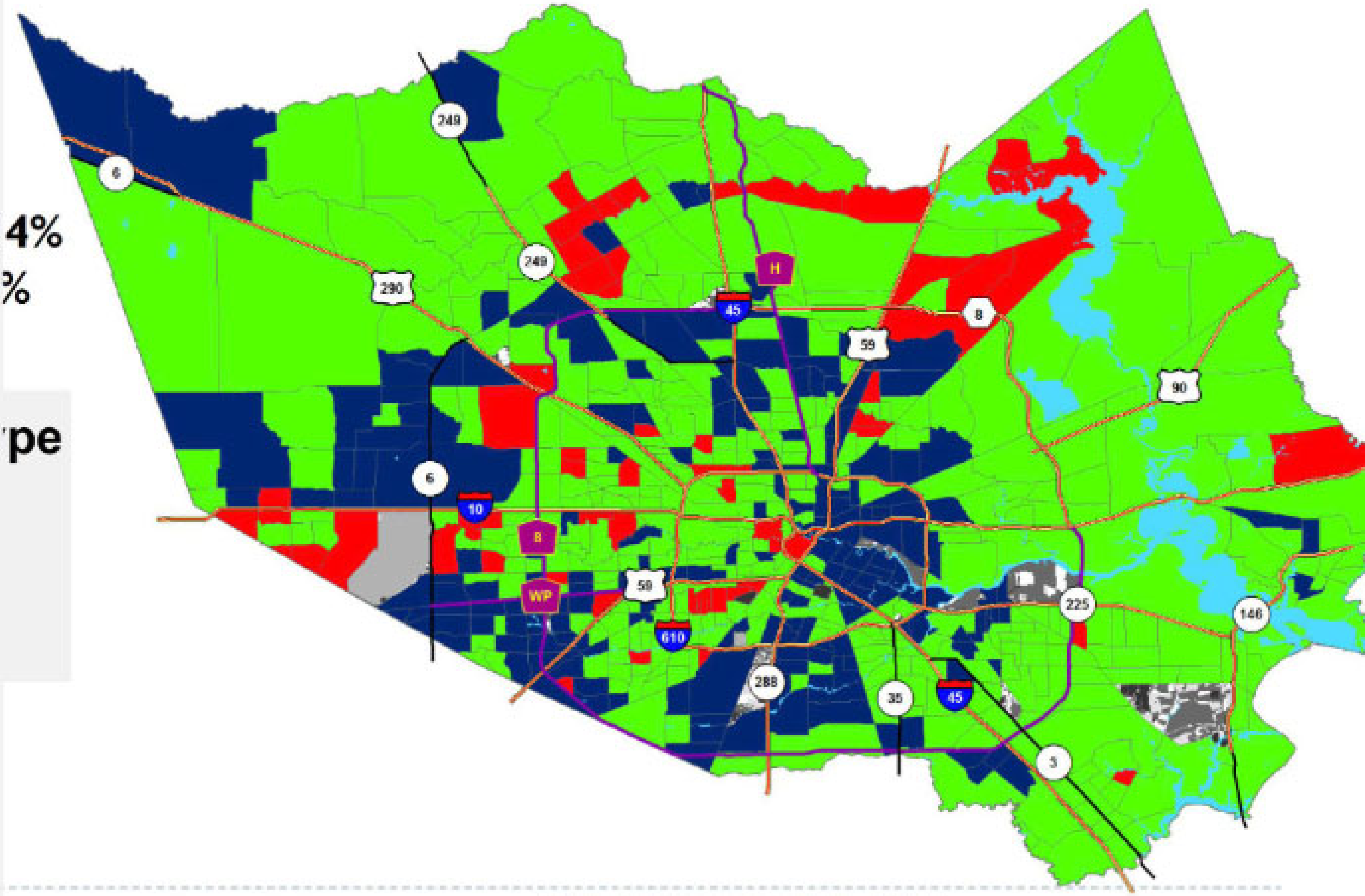
Actual

Prediction errors at census tract level (HAZUS-MH4)

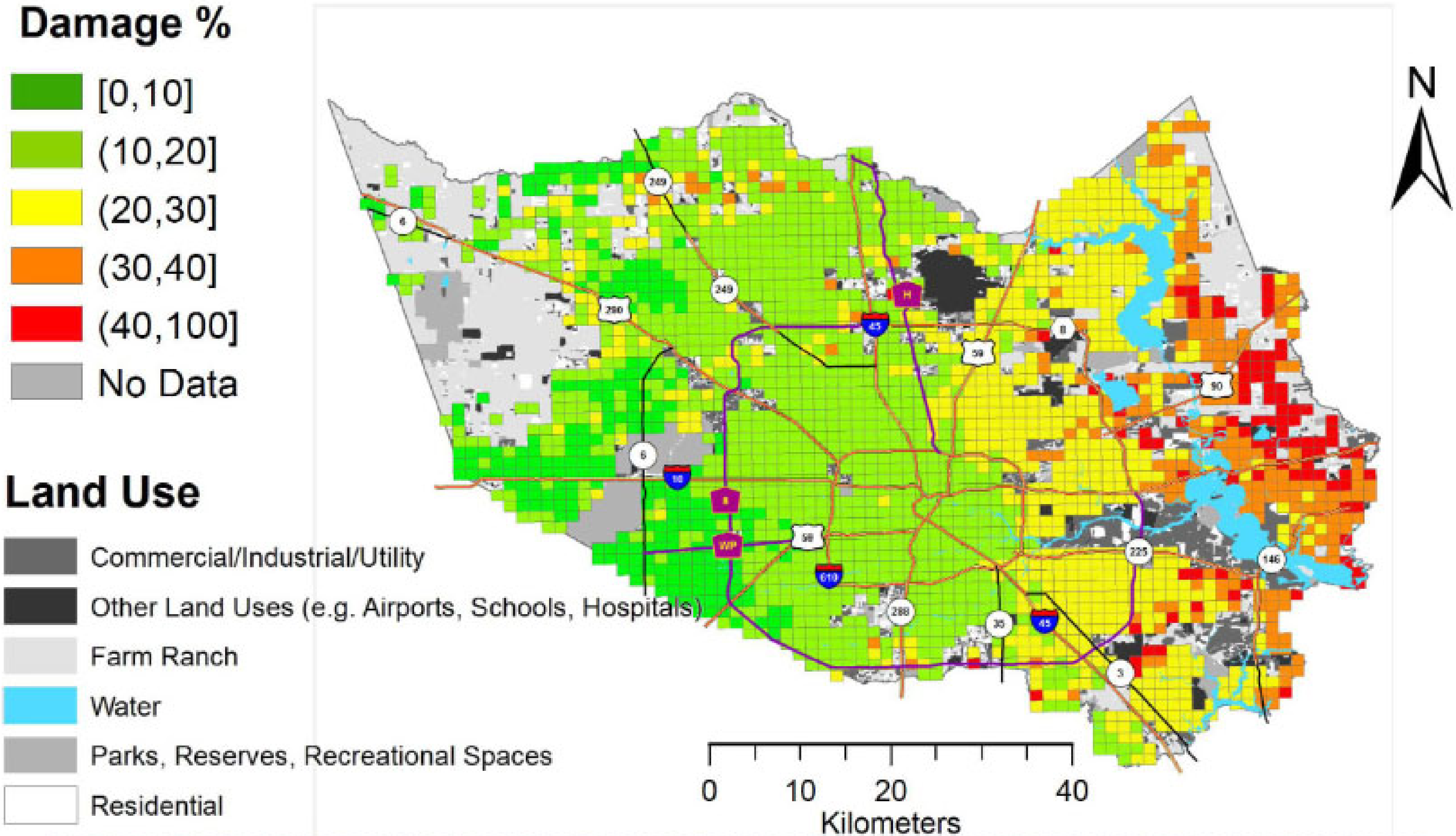
Correct = 59.24%
Under = 31.28%
Over = 9.48%

Prediction Type

- Under
- Correct
- Over

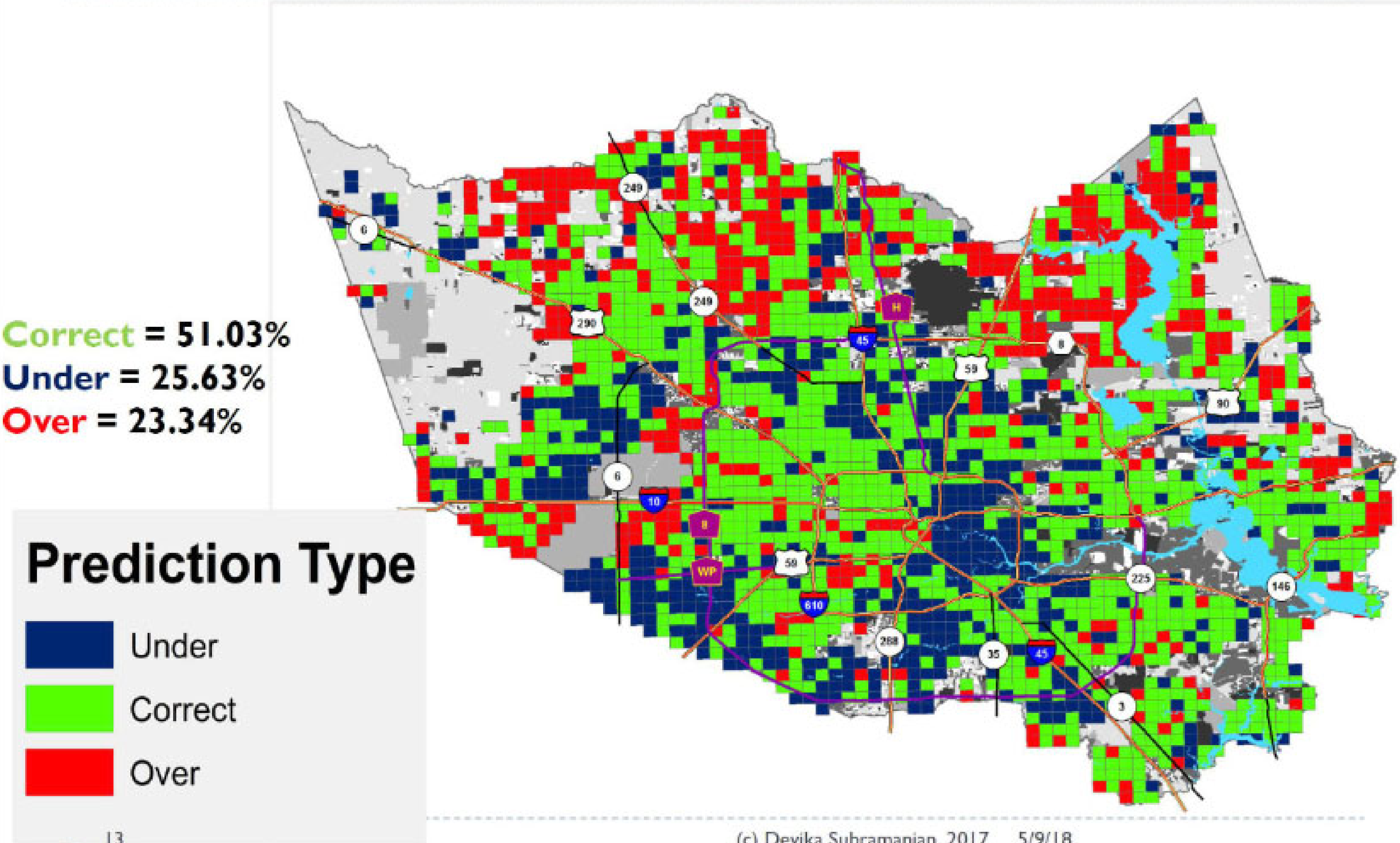


Predictions at the 1km square block level



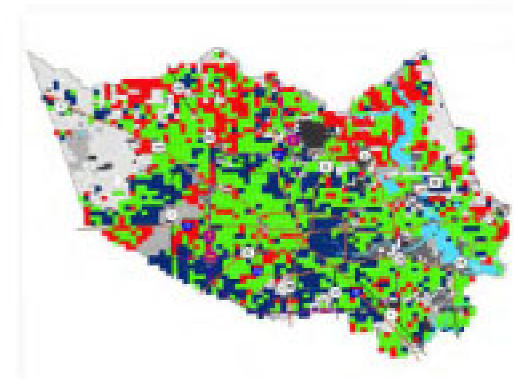
Prediction errors at the 1 km square block level

Correct = 51.03%
Under = 25.63%
Over = 23.34%

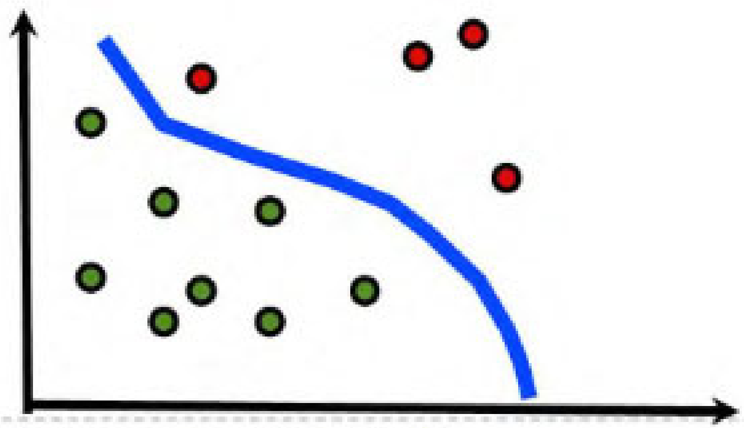


Error analysis using machine learning

What are the characteristics of the tracts/blocks where the fragility curve based model makes errors?



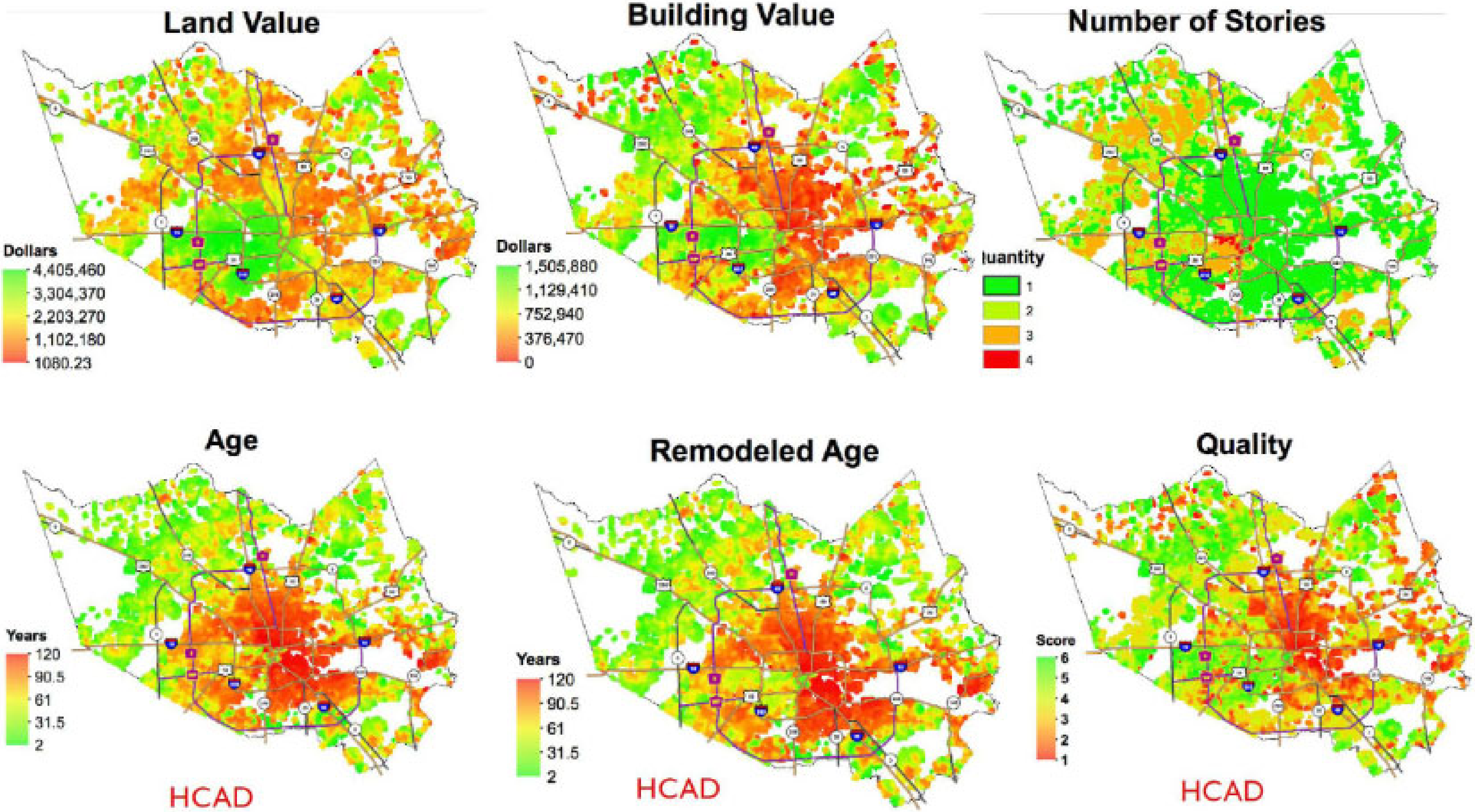
Training data



Model features

- Features averaged at census tract/block level
  - Terrain: Roughness length, HGAC landcover, tree100m, highway1km
  - Building properties: land value, building value, extra features value, exterior wall, floors, age, remodeled age, construction quality, garage type
  - Construction code enforcement: in an incorporated area, incorporation year, built after incorporation
  - Hazard characteristics: Max wind, wind swath, wind direction, wind duration, wind steadiness

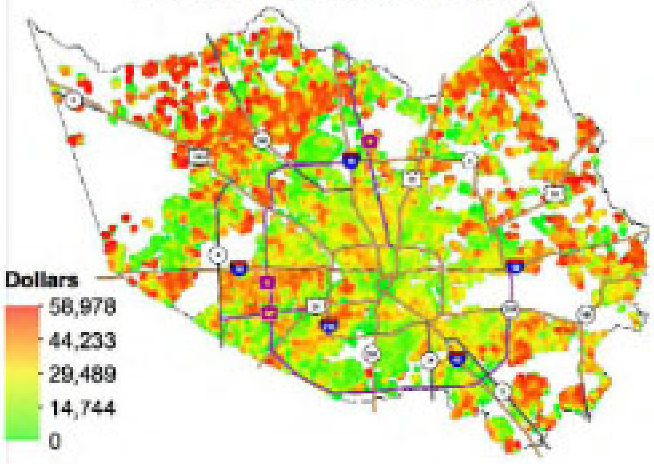
Explanatory variables



Explanatory variables

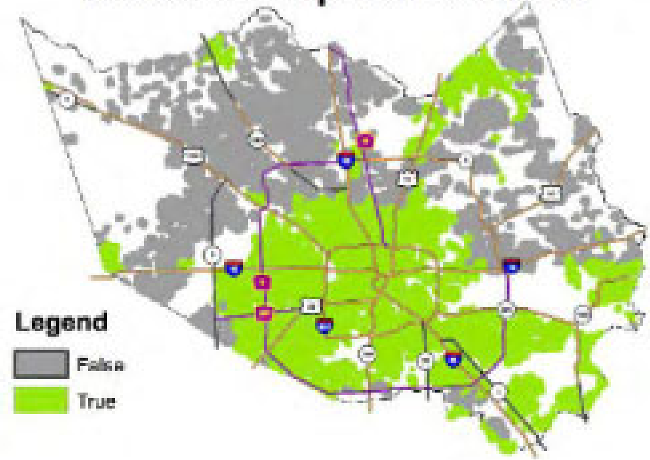
HCAD

Extra Features Value



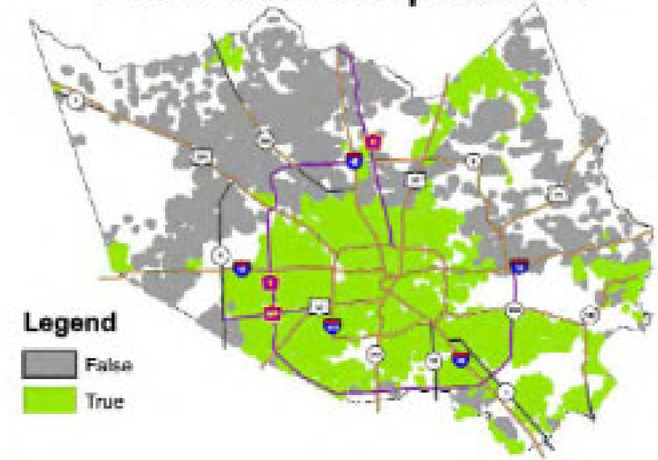
County

Inside Incorporated Area?

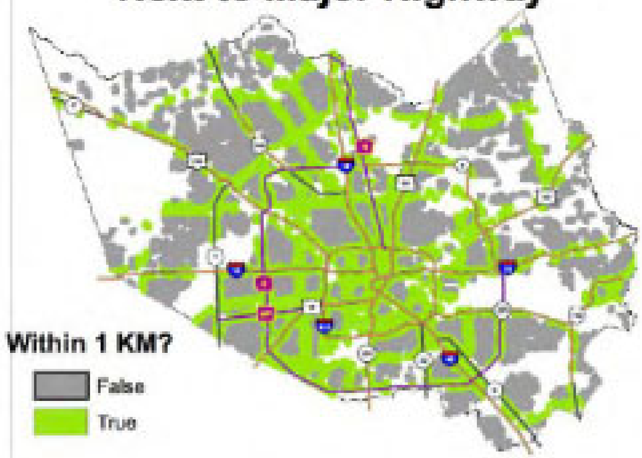


County

Built After Incorporation?

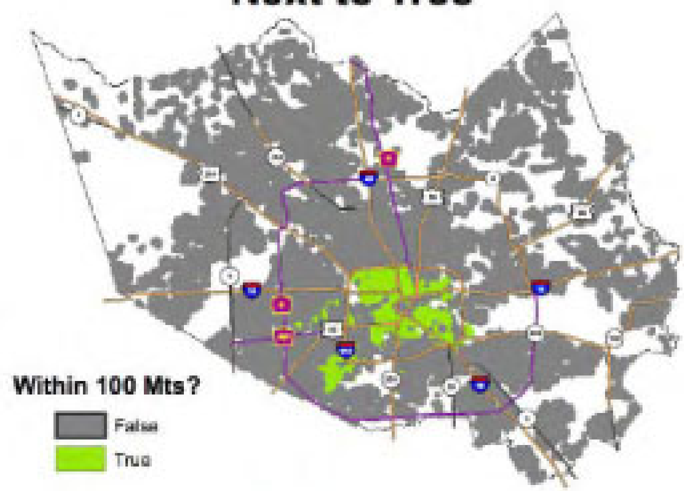


Next to Major Highway



County

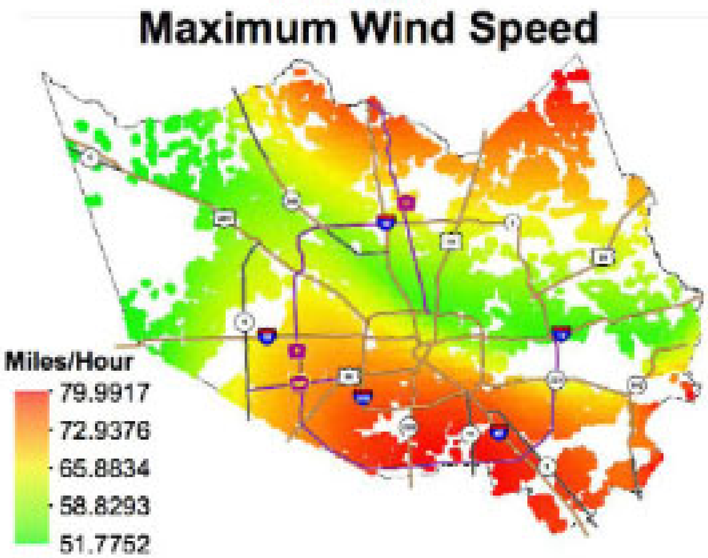
Next to Tree



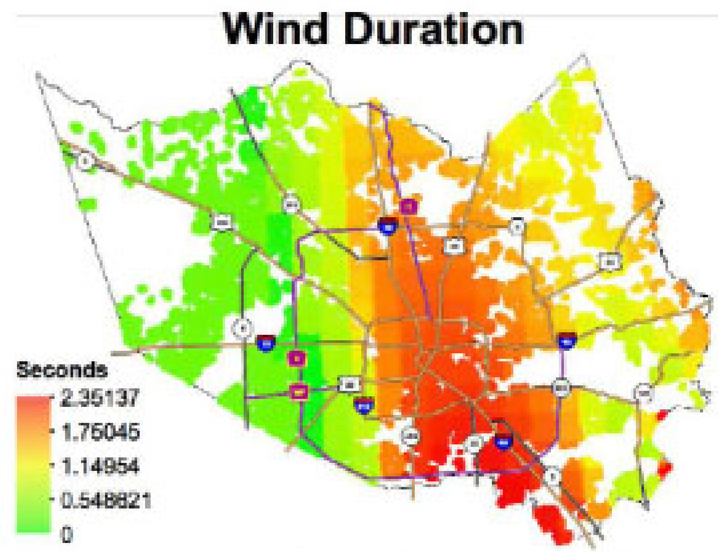
County

Explanatory variables

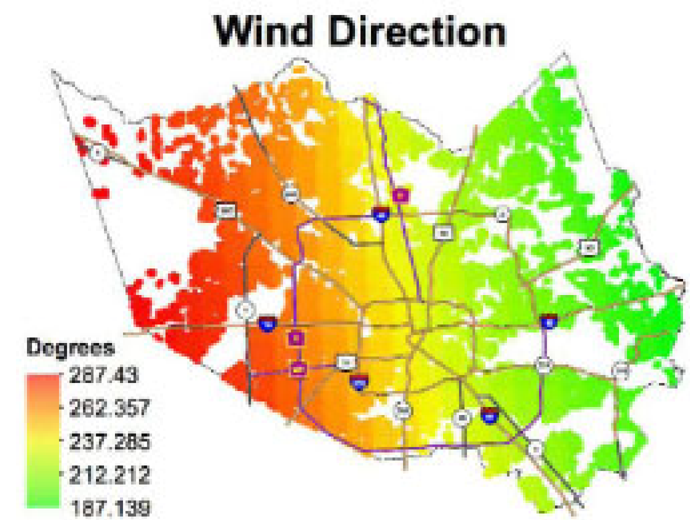
NOAA

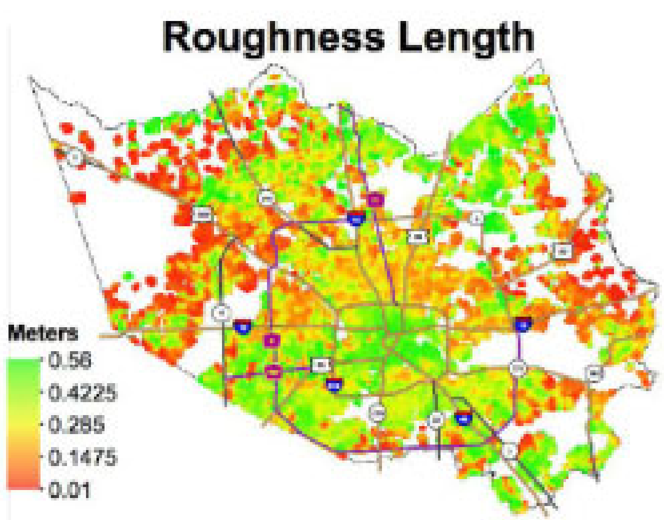


NOAA

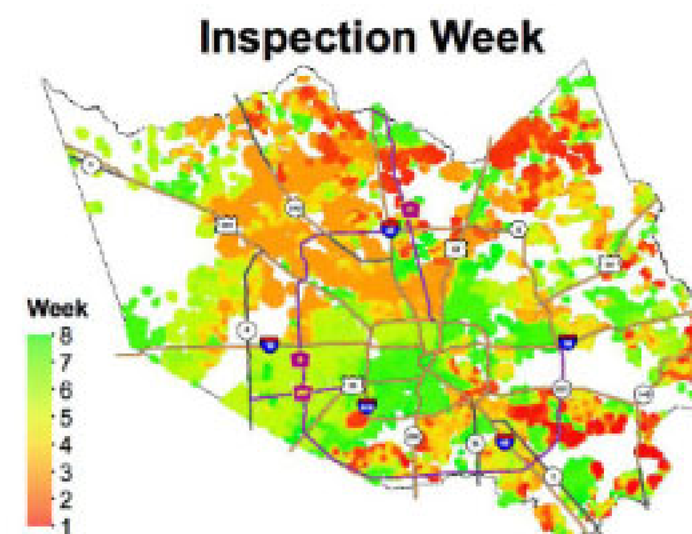


NOAA





HGAC

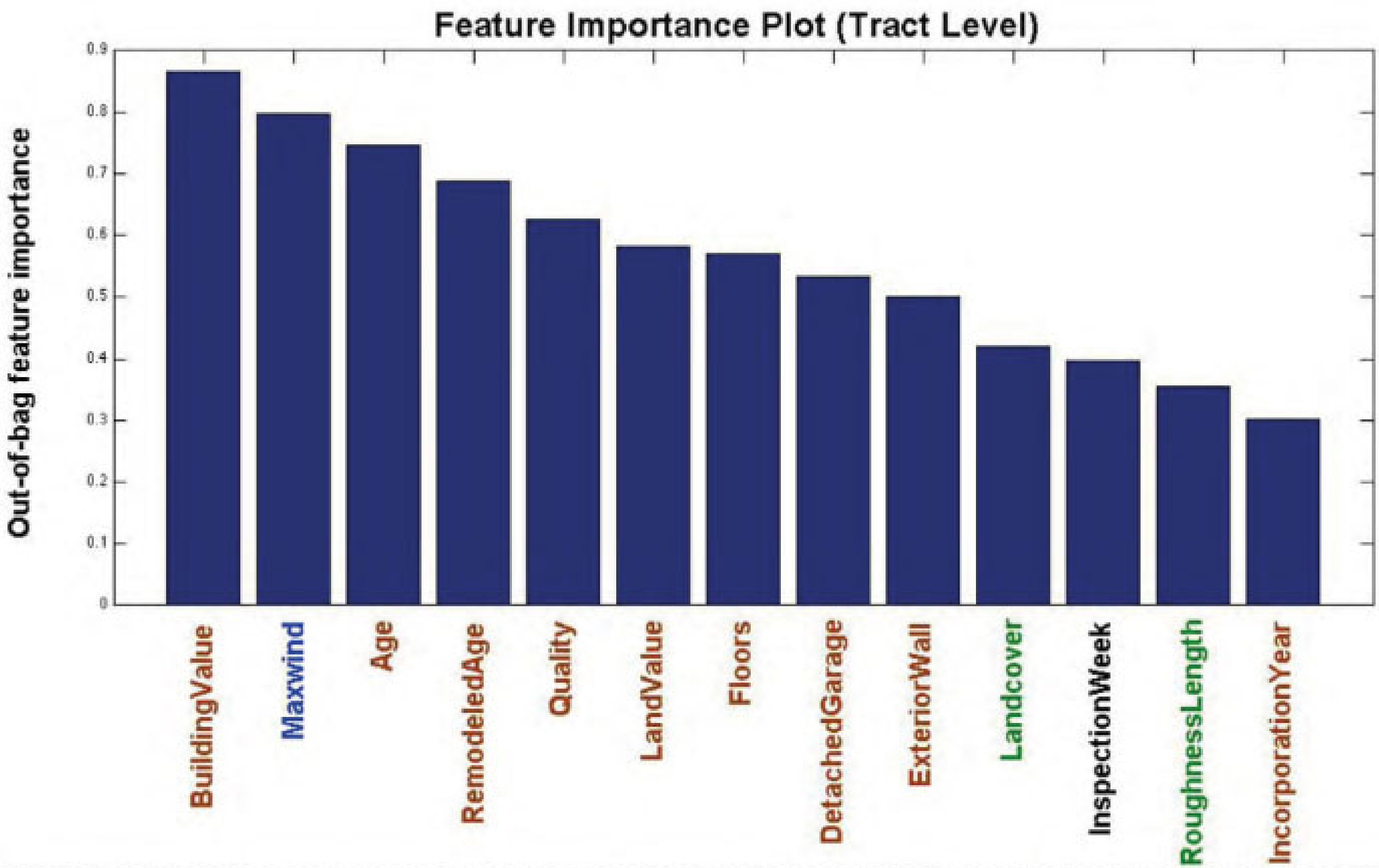


HCHA

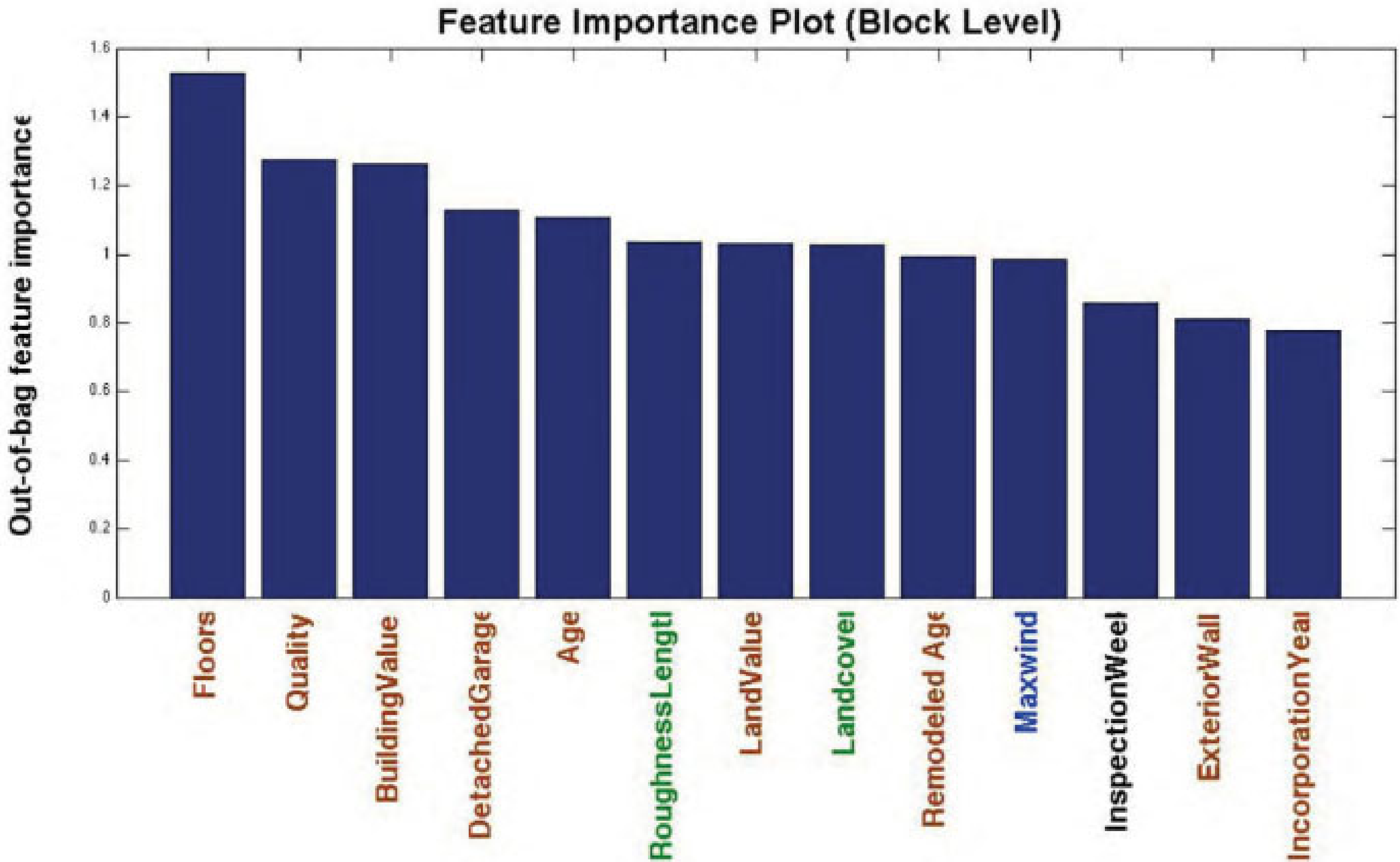
Structure of the model

- Classify tracts/blocks that are correctly predicted or incorrectly predicted (over/under)
- Best classifier is random forest classifier
  - 30 decision trees, 10 fold cross validated setting, for tract level.
    - 70% accuracy for tract level model
  - 100 decision trees, 10 fold cross validated setting, for block level.
    - 67% accuracy for block level model
- Visualize the random forest to yield a variable importance plot.

Error analysis of HAZUS-MH4 at tract level



Error analysis of HAZUS-MH4 at block level

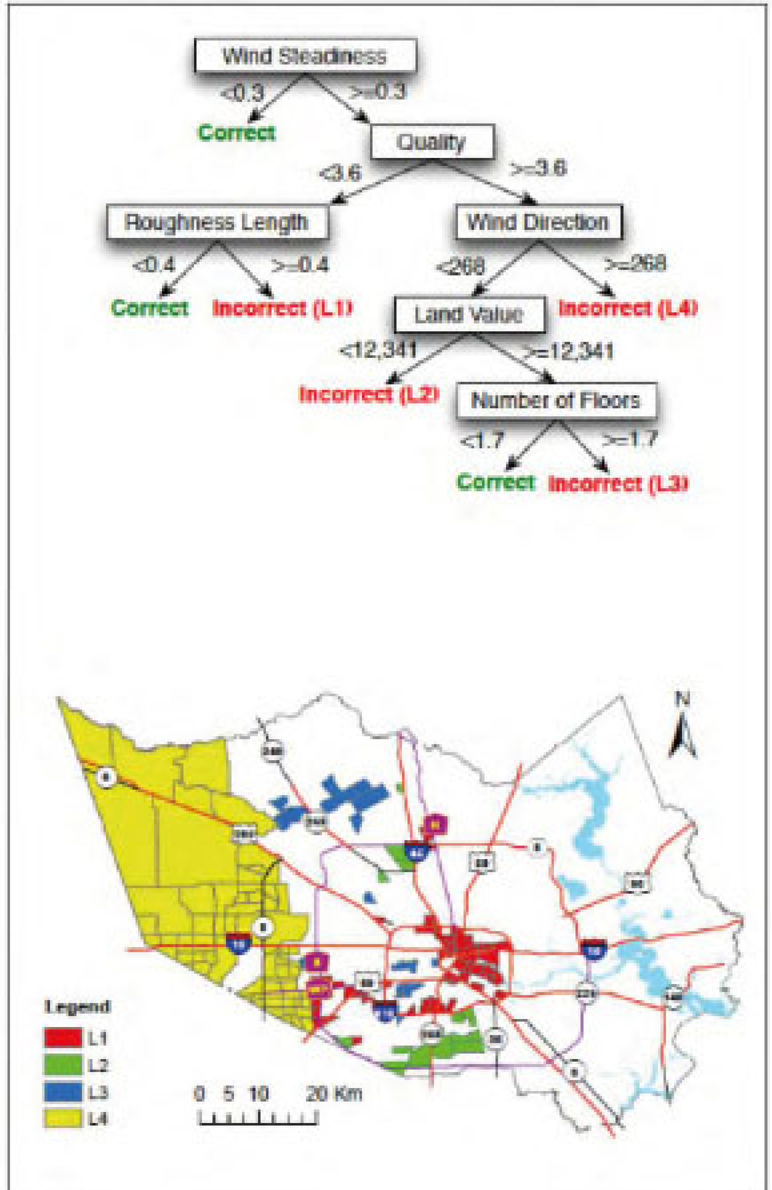


Interpreting results

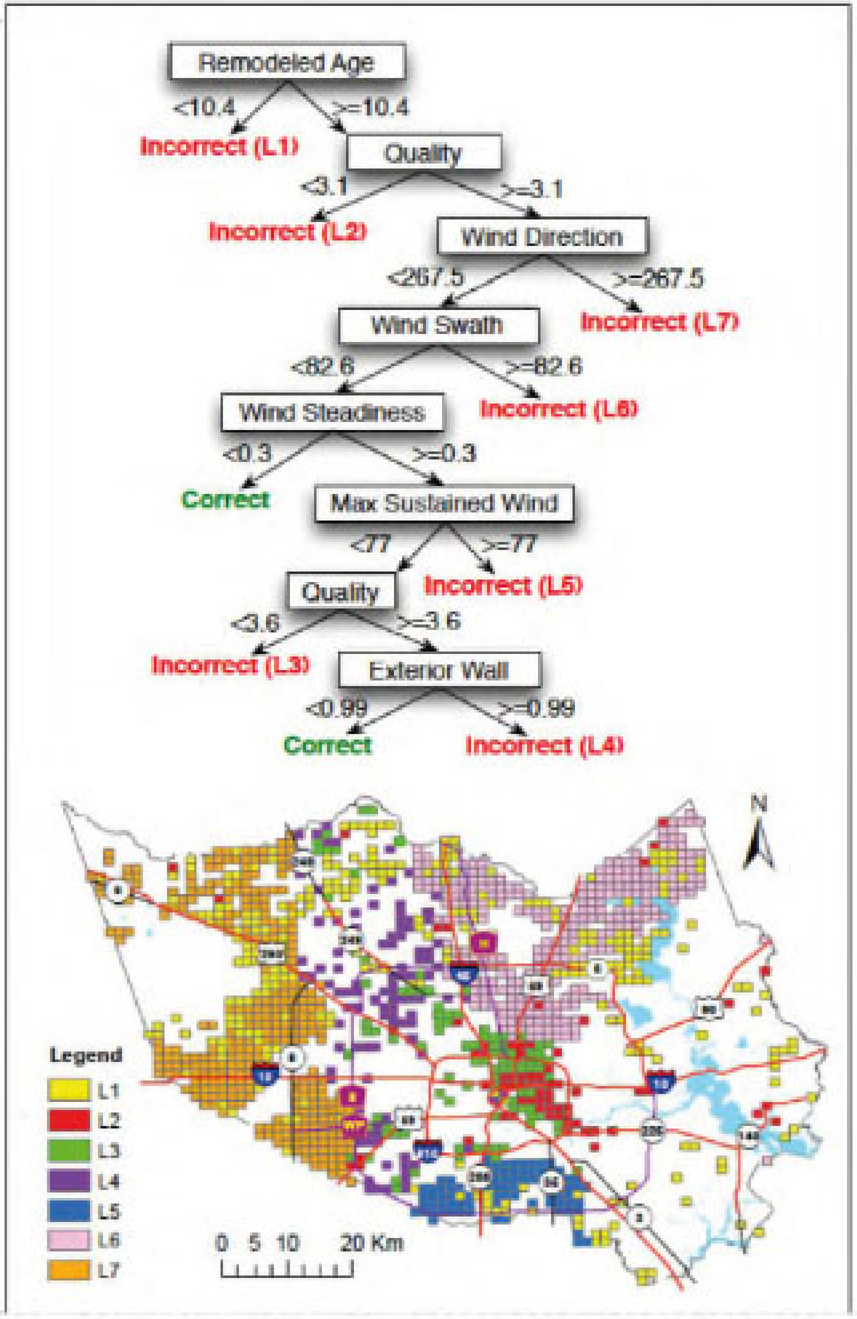
- Variable importance plots tell us which factors are critical in designing the next generation of fragility curves models.

Old parameters	New parameters
Frame	Building value
Floors	Land value
Roof shape	Quality of construction
Roof-deck attachment	Age
Nailing pattern	Remodeled Age
Garage type	Incorporation Year
Terrain	Exterior wall
Wind speed	

Detailed explanations of errors (tract level)



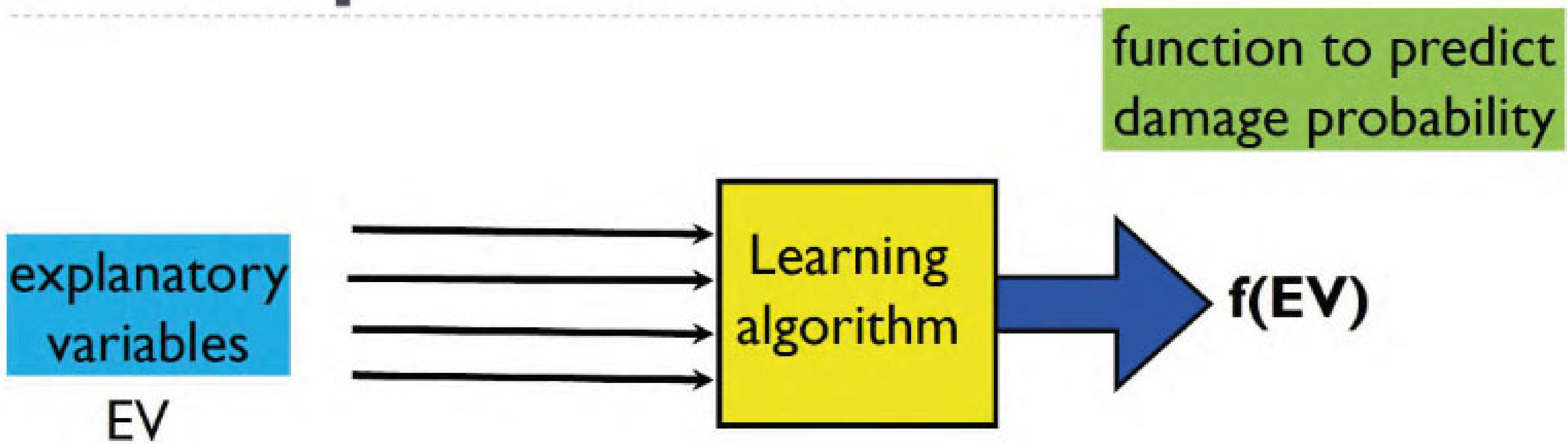
Detailed error explanations at block level

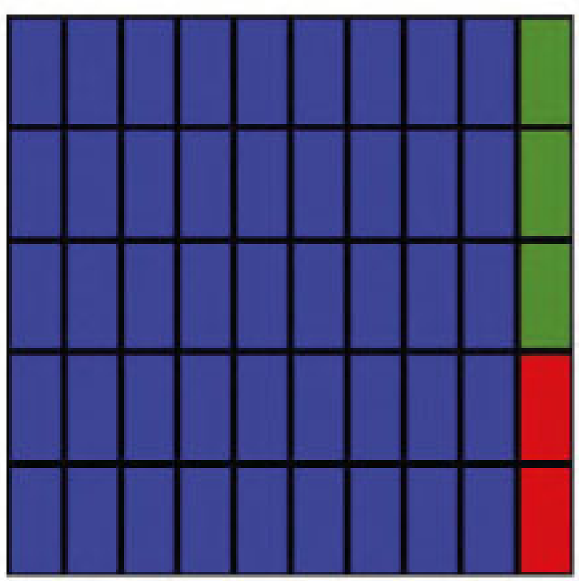


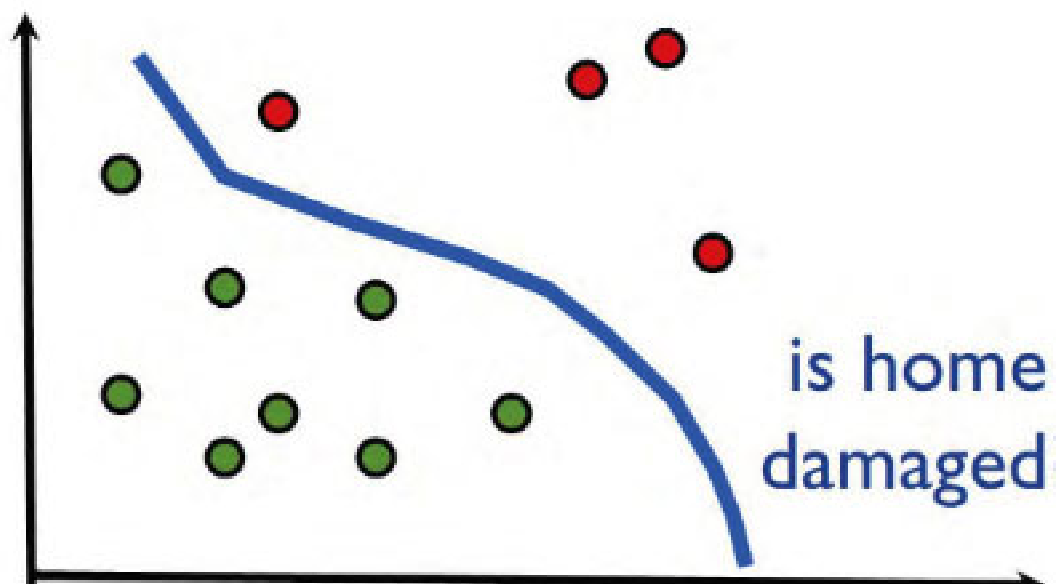
Models built directly from data

- More parameters and more complex functional forms are needed to estimate damage probabilities.
- What should the prediction target be?
  - Risk at home level?
  - Risk at block level?
  - Risk at tract level?
- How should model be evaluated?
  - At home level (classification task; AUC)
  - At block level (regression task; mean squared error)
  - At tract level (regression task; mean squared error)

Attempt 1


function to predict damage probability
explanatory variables
EV
Learning algorithm
f(EV)




is home damaged?

Explanatory variables (features)

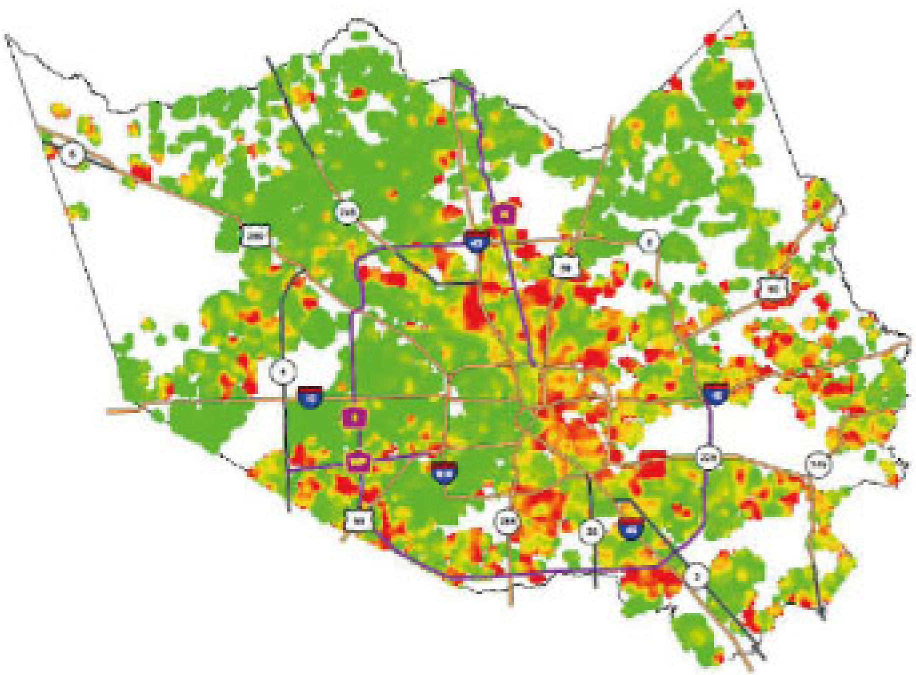
- Terrain properties at residence level: roughness length, HGAC land cover, tree within 100 meters, within 1 km of a major highway, distance to water bodies
- Wind properties at residence level: sustained wind speed, wind direction, maximum wind speed, wind duration, wind steadiness (NOAA).
- Residence properties: land value, building value, extra-features-value, quality, age, remodeled age, number of stories, week inspected after Ike, detached/attached garage, type of exterior wall (brick/stucco), incorporation year, built after incorporation year?, in an incorporated area?

Learning protocol and model evaluation

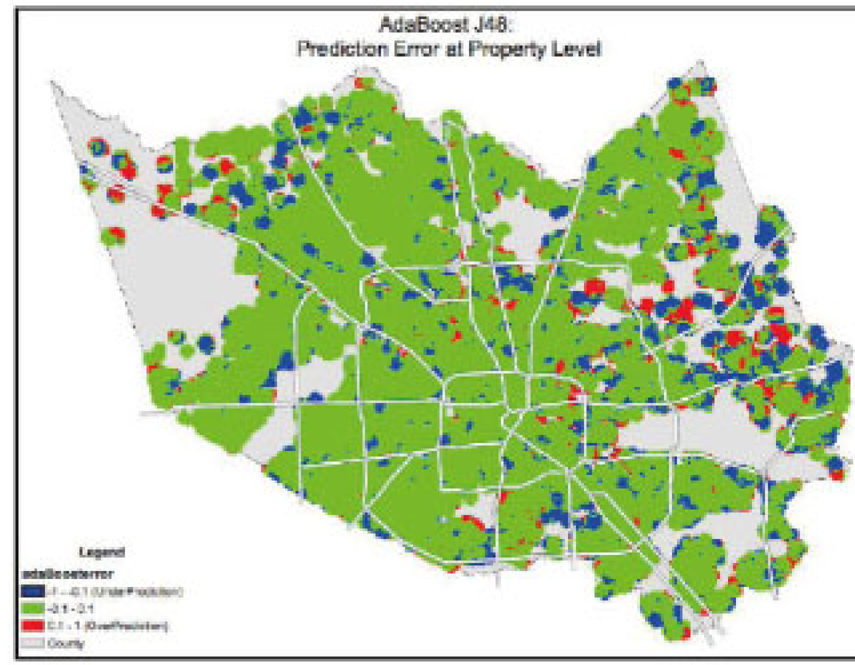
- Model structure:
  - Gradient boosted decision trees to predict damage at residence level.
  - Individual home damage predictions aggregated to kilometer square block level.
- Model evaluation: 10 fold cross validation
  - Evaluate accuracy at residence level
  - Evaluate errors (under, over, correct) at block level.

Prediction at home level

Prediction



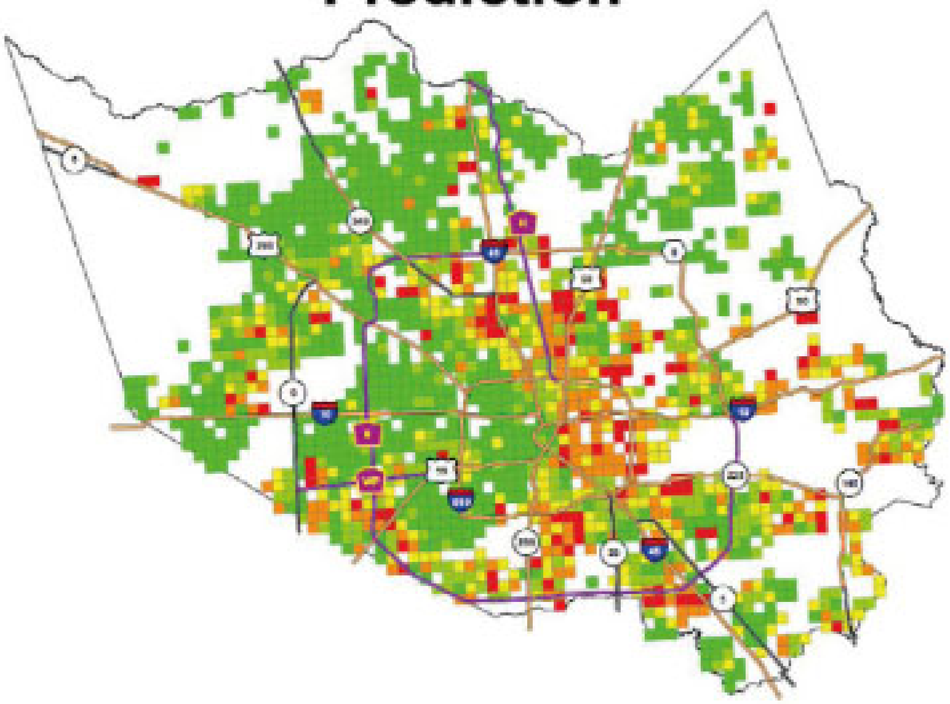
Error



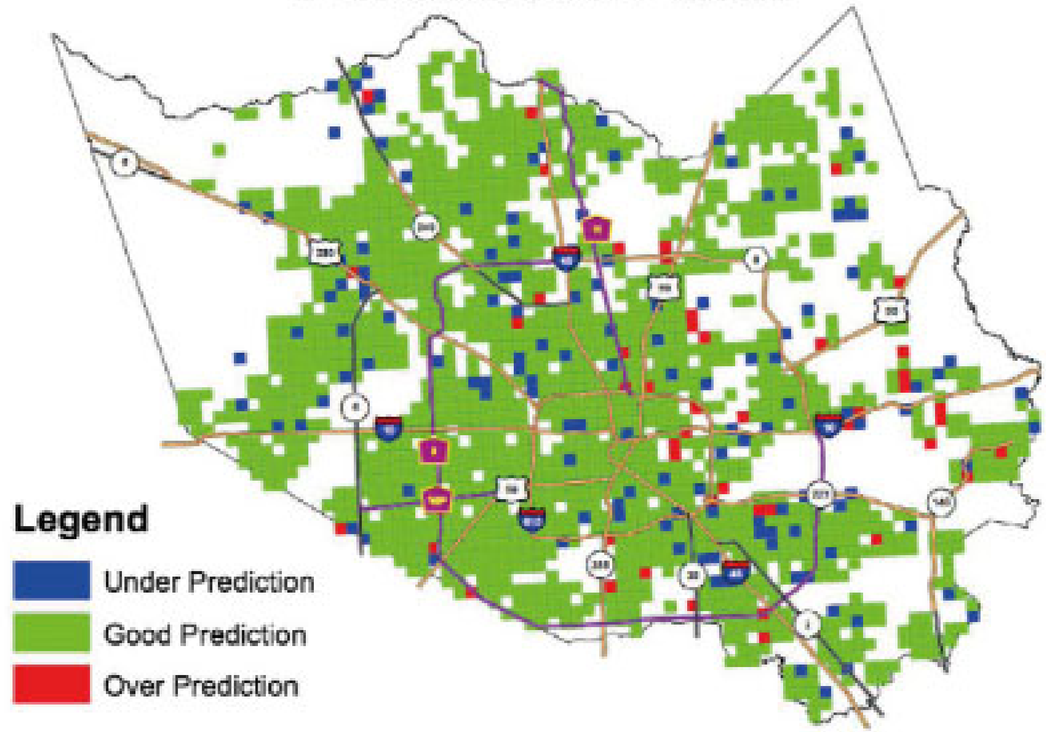
False positives : 8.6% , False negatives: 12.1%, Accuracy: 79.25%

Performance at block level

Prediction



Prediction Error

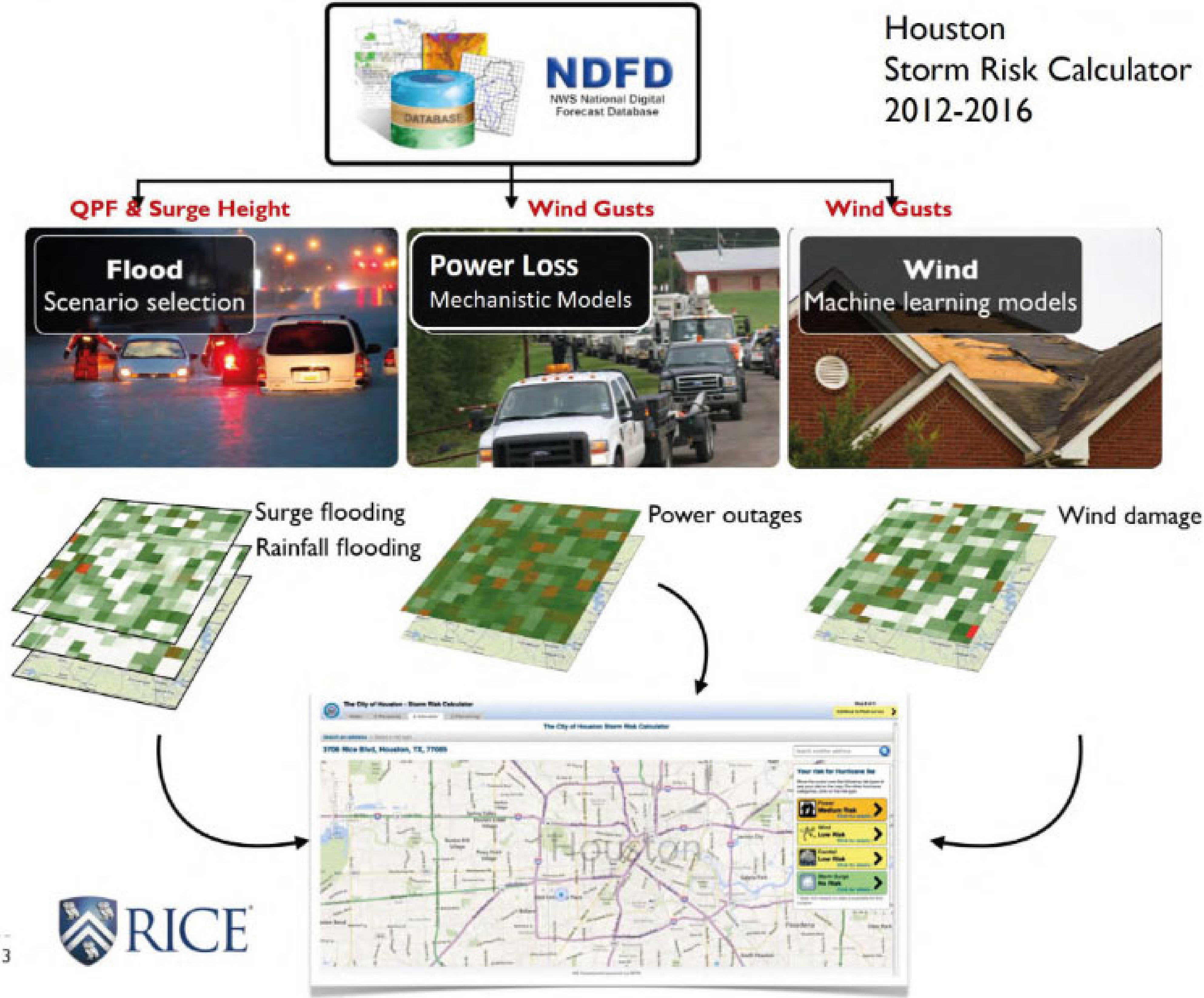


Accuracy : 88.8%, false positives: 2.82%, false negatives:8.3%

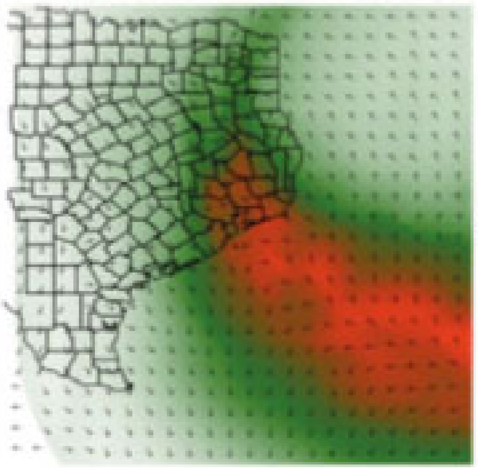
Generalizability?

- Overfitted to the Ike scenario.
- Need more sophisticated regularization strategy to control for overfitting.
- Models hard to debug!
- Unclear how to extract features to improve model performance!
  - Physics not helpful, alas.

Houston Storm Risk Calculator 2012-2016

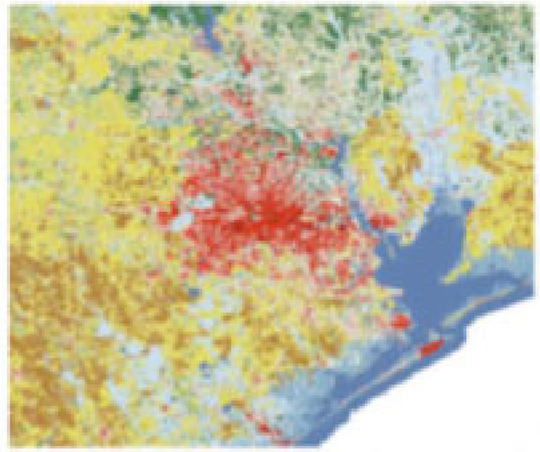


Integrating data sources



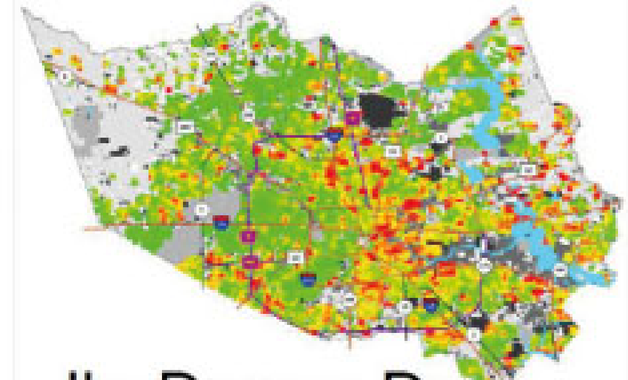
Wind data
(NOAA)

+



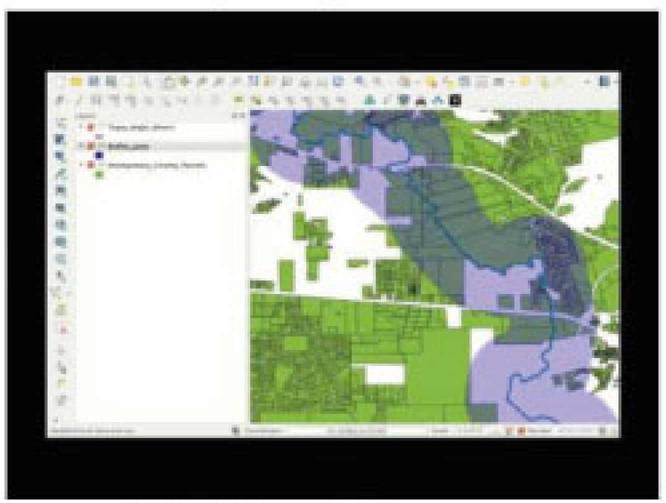
Landcover data
(MRLC)

+



Ike Damage Data
(HCHA)

+



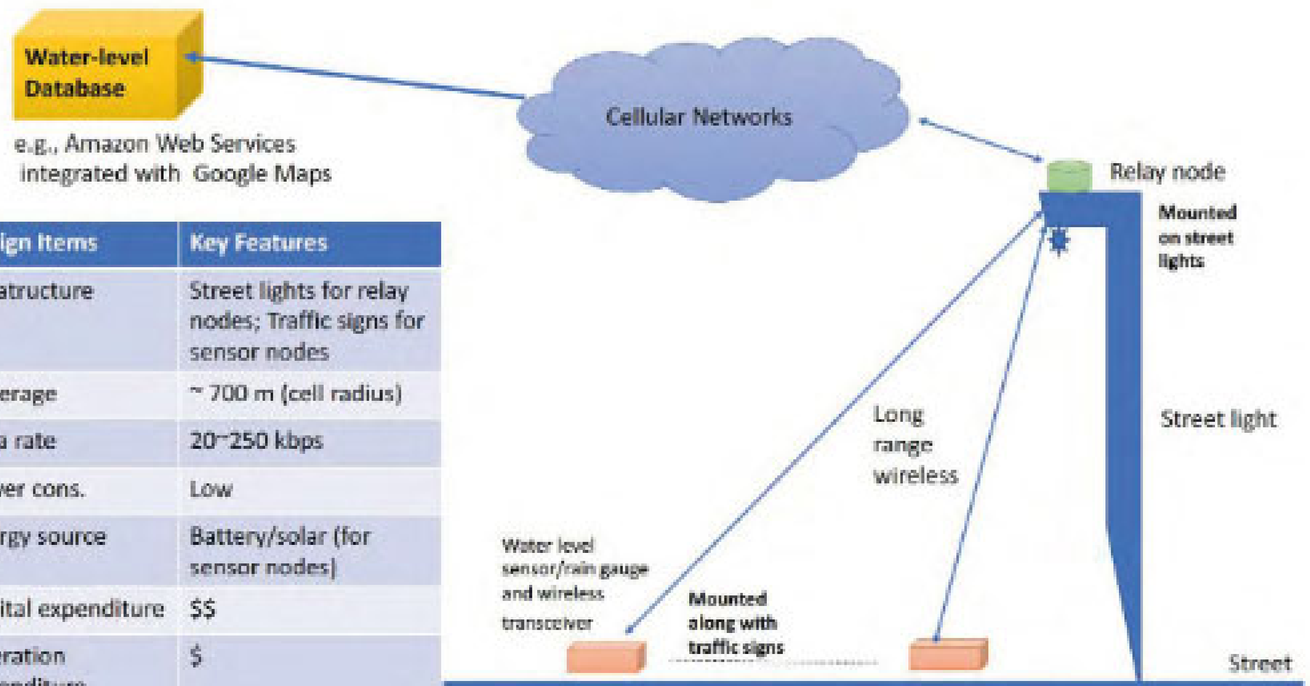
HCAD data
(Harris County)

+

LIDAR data
(Rice GIS)

Prediction at home level

New sensor networks

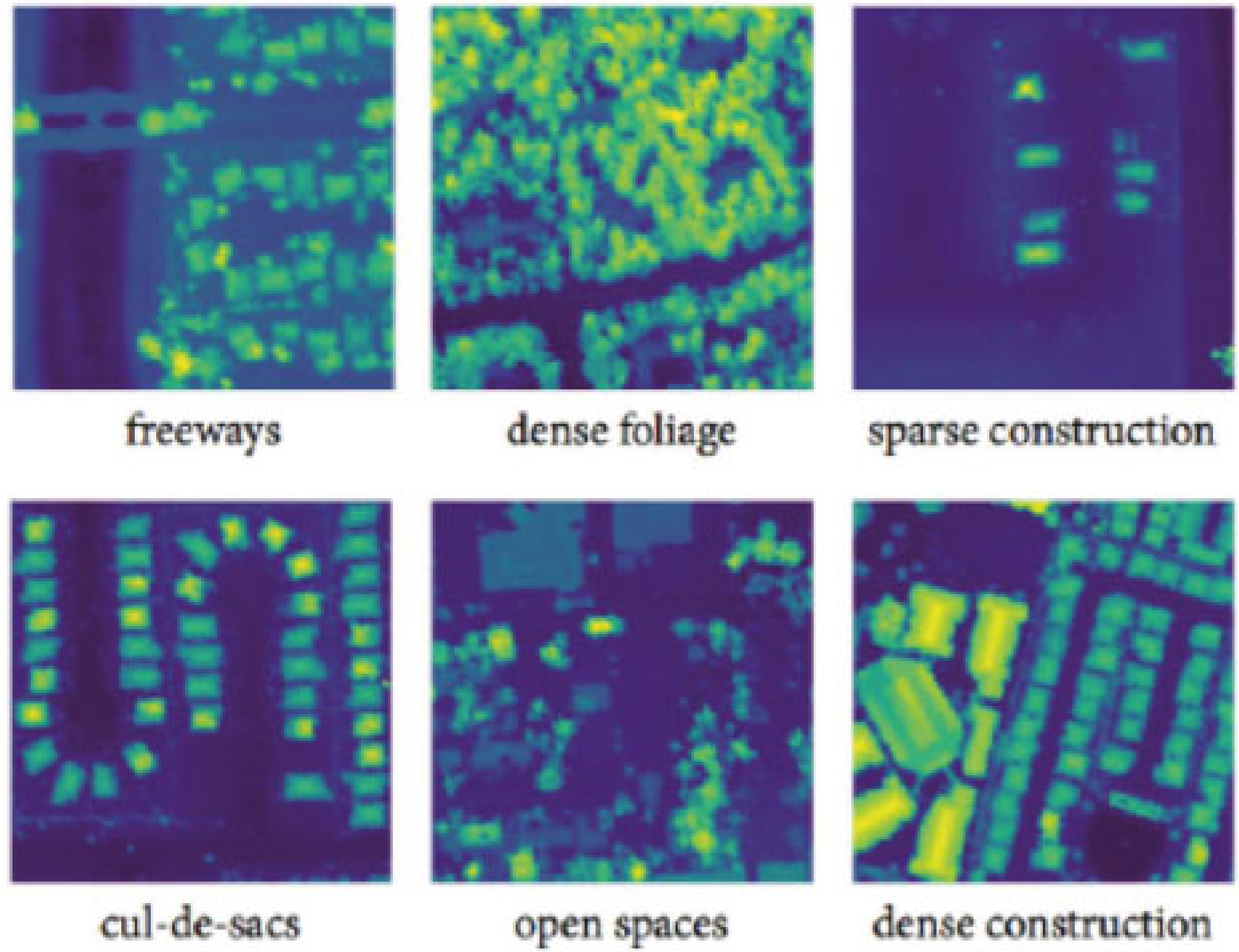


Design Items	Key Features
Infratructure	Street lights for relay nodes; Traffic signs for sensor nodes
Coverage	~ 700 m (cell radius)
Data rate	20~250 kbps
Power cons.	Low
Energy source	Battery/solar (for sensor nodes)
Capital expenditure	$$
Operation expenditure	$

Gary Woods and Frank Li and team
At Rice OEDK

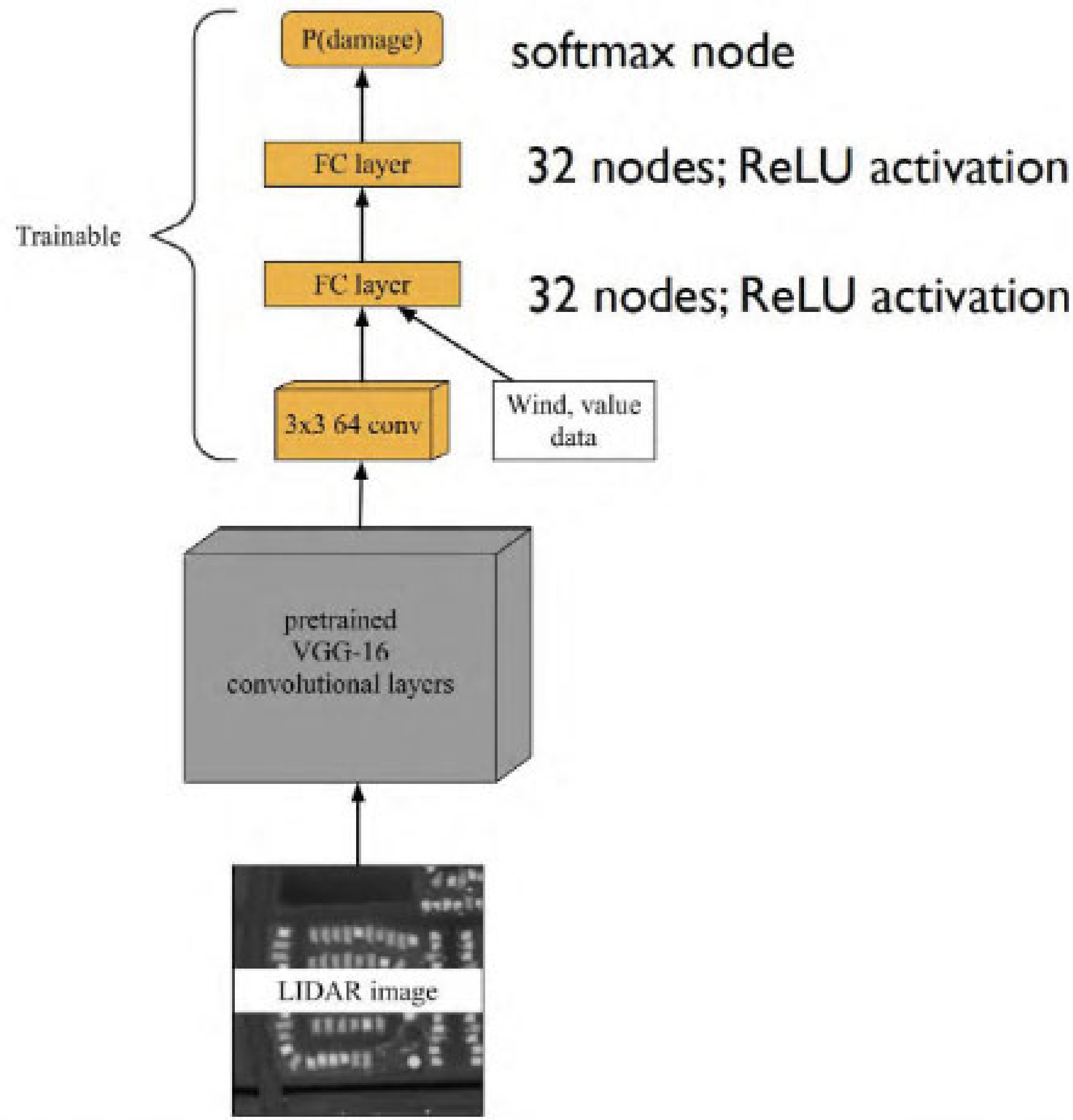
Challenges

- Data at varying spatial resolutions
  - LIDAR at 60 cm horizontal resolution
  - Damage information at home level
  - Landcover data at 30 m resolution
  - HCAD data at parcel level
- Class imbalance (damaged homes about 20% of data set)
- Have to learn relevant features and decision function jointly.

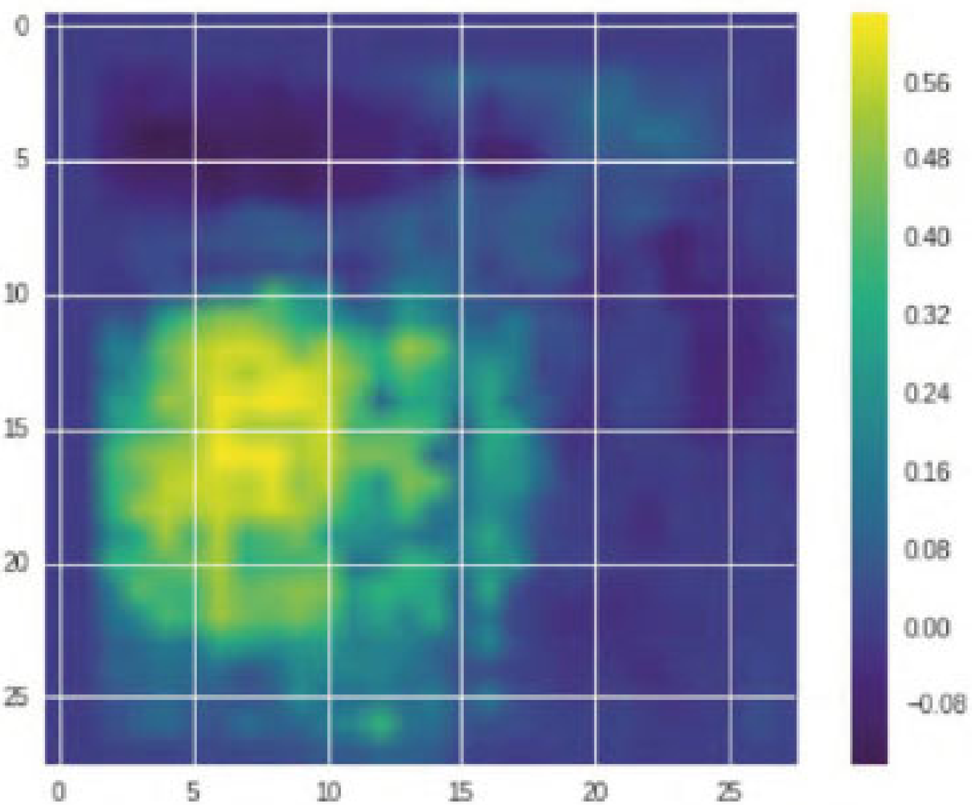


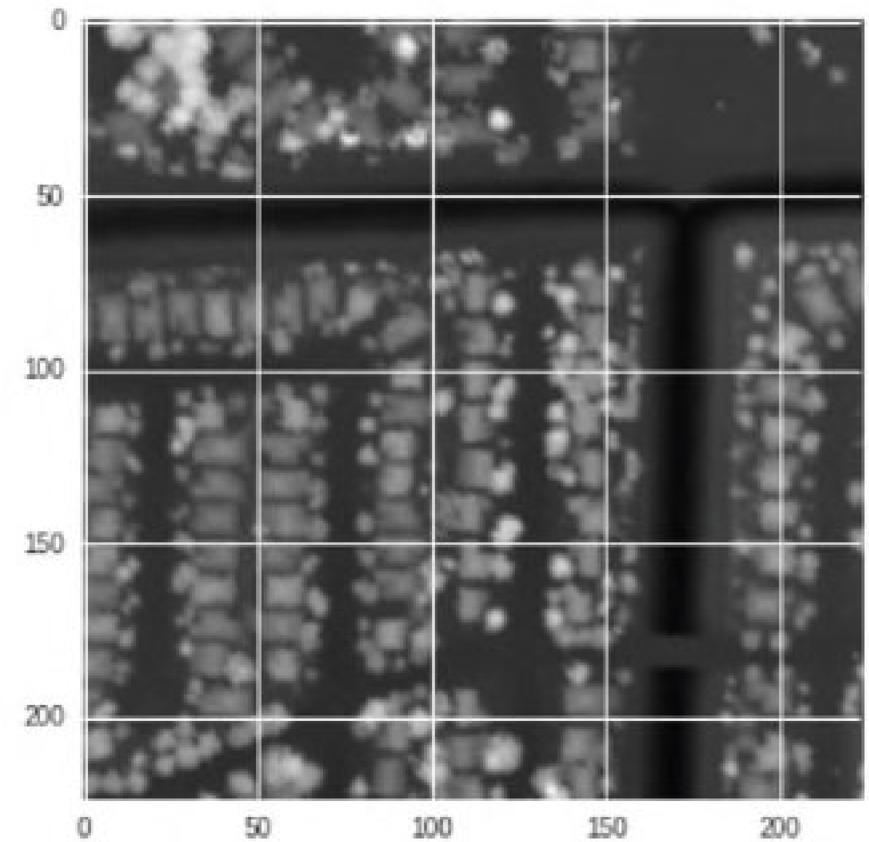
A deep architecture for predicting wind risk

- AUC in an 80/20 train/test configuration is 0.70 at the individual home level.
- Simple spatially autoregressive model with coarser terrain features in the same setting did not perform as well.



Saliency maps





Ongoing work

- Tuning architectures; training protocols; training parameters
- Rigorous comparison against spatially autoregressive models
- Interpreting learned model and extracting relevant aspects of built environment, wind and home characteristics that account for wind damage
  - Inferring the physics from learned distributions

Publications

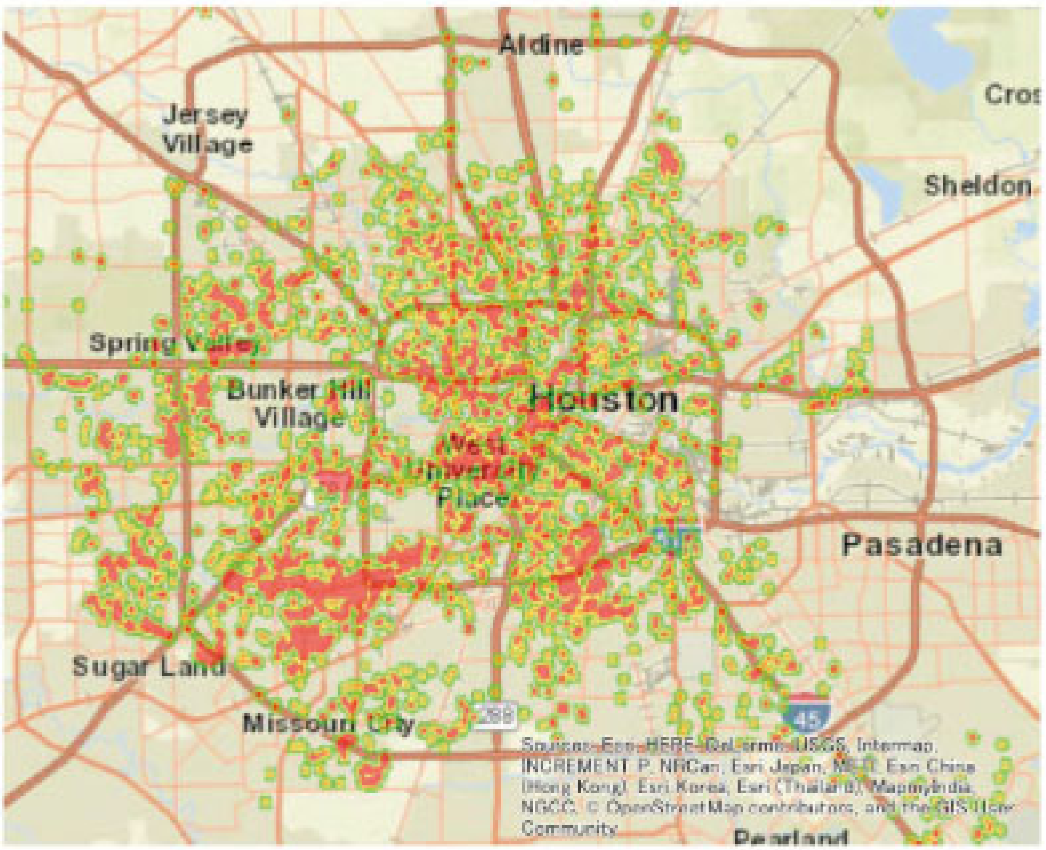
- Incorporating high-resolution LIDAR imagery to predict wind damage risk using deep convolutional networks, in preparation (with I. Dykeman)
- Building and validating geographically refined hurricane wind risk models for residential structures, Natural Hazards Review, 2014 (with J. Salazar, J., L. Duenas-Osorio, and R. Stein)
- How Risk Perceptions Influence Evacuations from Hurricanes and Compliance with Government Directives, Policy Studies Journal 2013, 41(2):319-342, (with R. Stein, B. Buzcu-Guven, L. Duenas-Osorio, and D. Kahle)
- Engineering-based hurricane risk estimates and comparison to perceived risks in storm-prone areas, Natural Hazards Review, 13(1), 4556; 2012 (with L. Duenas-Osorio, B. Buzcu-Guven, R. Stein)
- Who evacuates when disaster approaches: the role of risk, information and location, (with R. Stein and L. Duenas-Osorio), Social Science Quarterly, 91(3); 2010.
- Interface network models for complex urban infrastructure systems, Journal of Infrastructure Systems, 17(4); 2011 (with J. Winkler, L. Dueas-Osorio, R. Stein)

Analyzing flooding in Houston with unsupervised learning

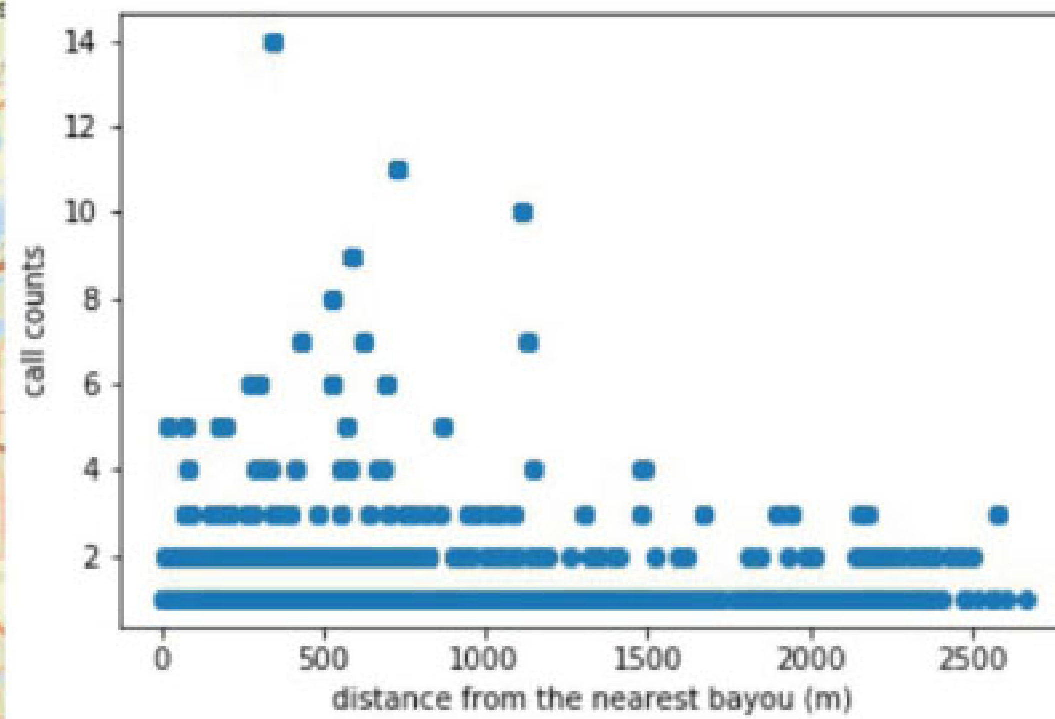
- Predict flooding at the home level in Houston based on integrating a range of readily available data sources: rainfall, 311 calls, flood plain designation, digital elevation models

- Use unsupervised machine learning techniques to identify the key determiners of home flooding

311 call density map 2011-2015

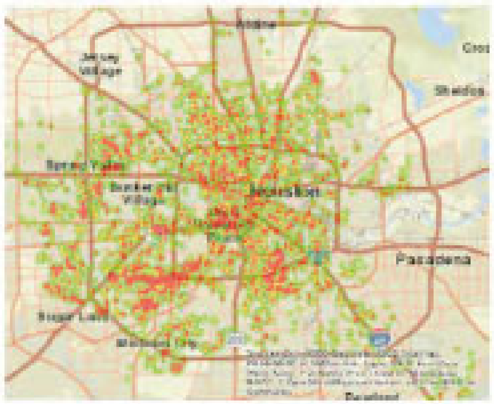
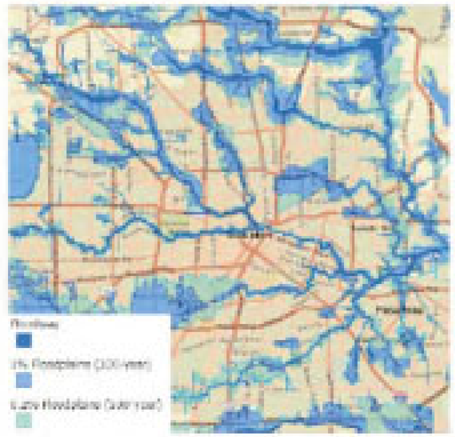


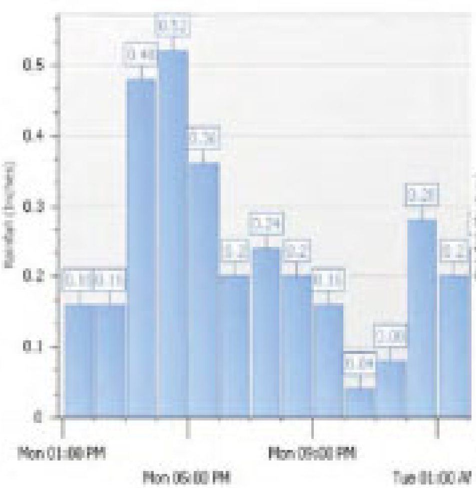
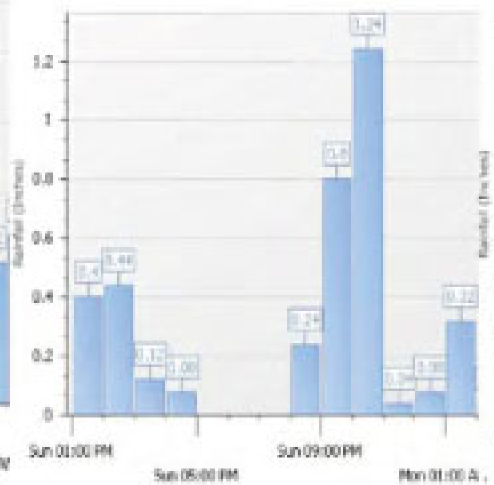
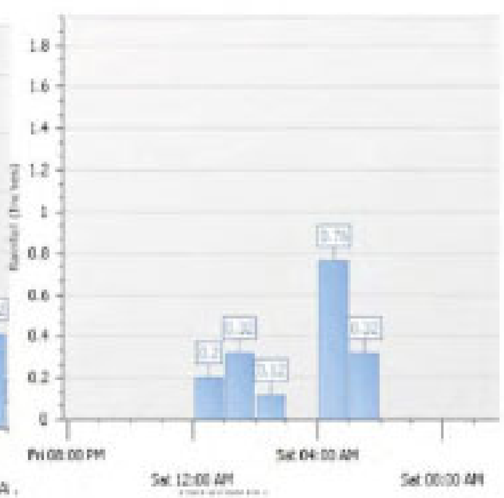
For homes not in a designated flood plain, scatterplot of distance from the nearest bayou and 311 call count in 2013-2015.



51.2% of homes not in designated flood plain experienced more than one flood.

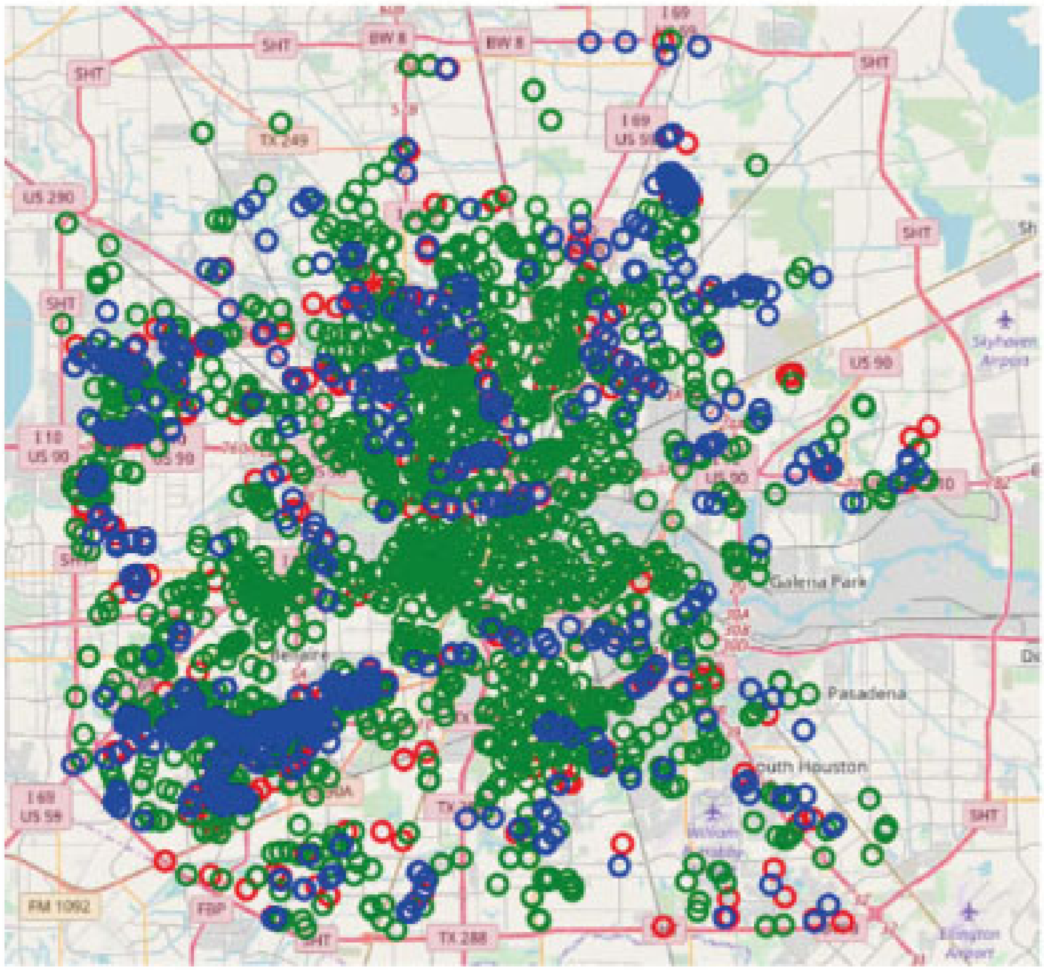
Unsupervised machine learning

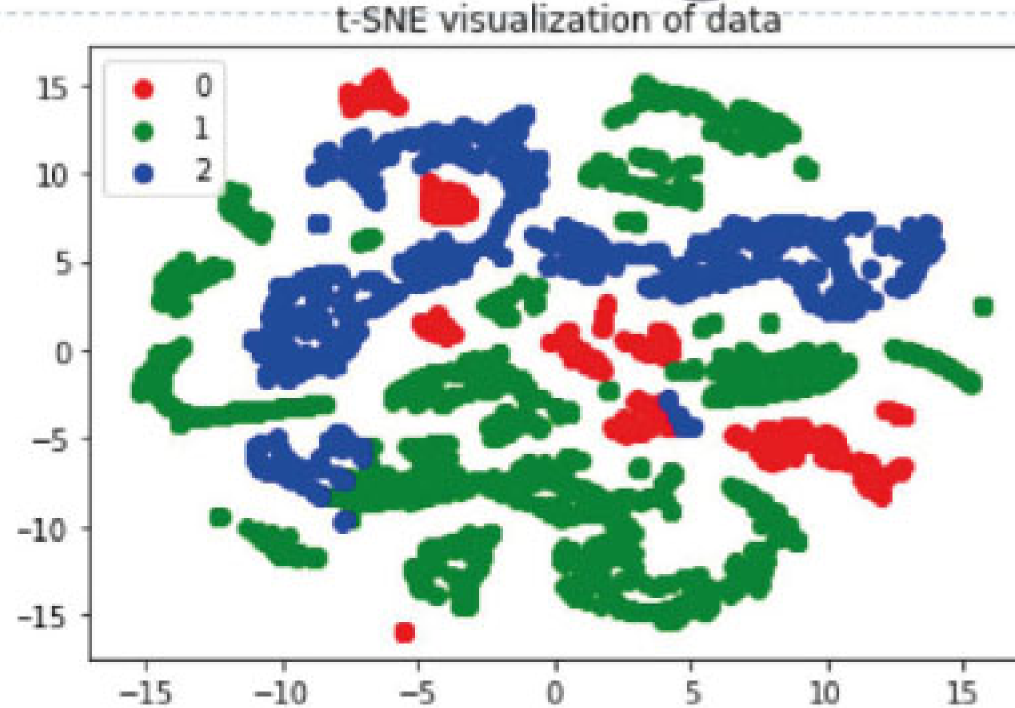



Cluster 311 calls
Based on
- rainfall
- bayou distance
- floodplain
- call frequency

Results of k-means clustering

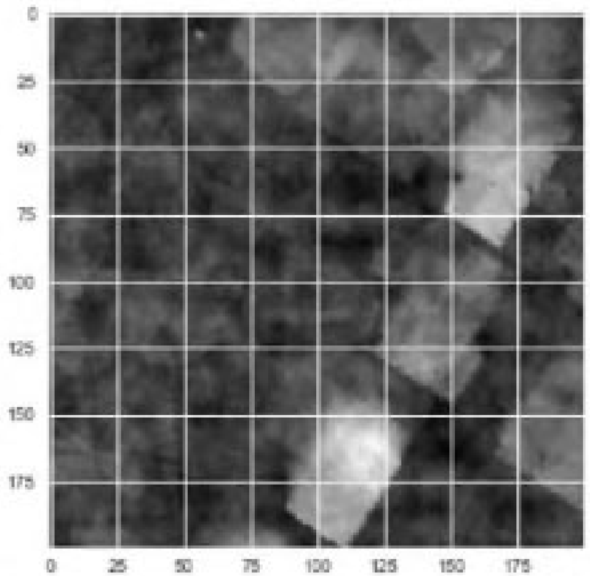




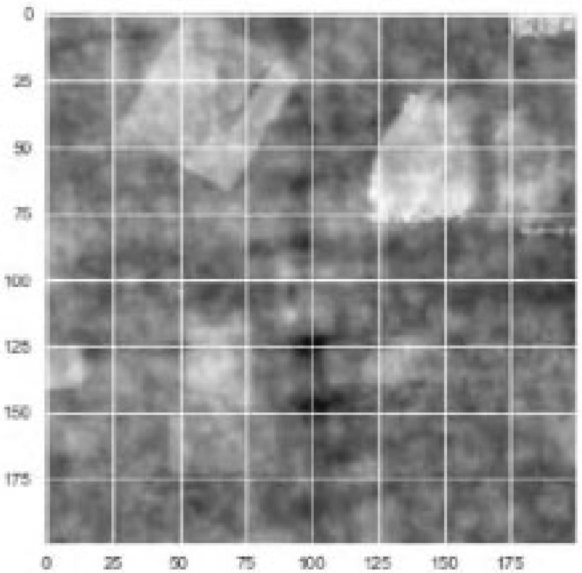
	bayou distance(m)	call counts	flood plain	percentage
A	42.6	2.35	0.400	15.0%
B	**768**	**2.78**	**0.109**	**54.5%**
C	140	1.58	0.368	30.4%

- Cluster A: very close to bayou, in flood plain, more than 2 floods on average
- **Cluster B: far from bayou, not in flood plain, more than 2 floods on average**
- Cluster C: close to bayou, in flood plain, more than 1 flood on average

LIDAR signature of cluster B



High density development and proximity to large tracts of land at low elevation



Major intersections in high density developments

Data-driven analysis as a useful stand-in for physics-based models

- We can build a fast and accurate flood prediction model at street level by merging diverse data sets rather than by running detailed physics simulations of water flow requiring knowledge of hydrology, soil permeability, drainage networks, rainfall, channel structures, and accurate terrain models.