Spatiotemporal Risk Modeling of Traffic Accidents: A Case Study of Dallas, Texas, USA

Benjamin Acker & May Yuan, PhD

CaGIS AUtoCarto & UCGIS Symposium, 2018

UT DALLAS

Research Motivation

- Road injury is the 10th leading cause of death in the world – World Health Organization
- Geography alone cannot predict traffic accidents
- Yet geography can show how site characteristics amplify or dampen the frequency and severity of accidents
- This research is part of a larger collaborative project between University of Texas at Dallas, Dallas, Southern Methodist University, and the Dallas Fire-Rescue Department and is funded by National Institute of Standards and Technology (NIST)

UT DALLAS

Research Objective

- Develop a spatiotemporal model to understand how both time and site characteristics influence accident likelihood
  - Spatial resolution: 100m segment
  - Temporal resolution: 1 hour
- Types of features:
  - Time of Day/Week
  - Road features
  - Space Syntax
  - Weather
  - 2nd order effects of other accidents
- Focused on Dallas, TX, but model should be robust enough to apply to other urban areas

UT DALLAS

Previous Research

- Geographers have extensively studied the influence of site characteristics on traffic accidents with diverse methods, ranging from simple visualization techniques, to hot spot analysis with kernel density estimation (KDE) and local indicators of spatial autocorrelation (LISA), to cluster analysis using k-means and K-function, and to spatial regression analysis (Yao et al., 2015).
- Space syntax analysis has been used to model traffic flow (Jiang and Liu 2009, Serra et al 2015, and Omer et al 2017). It has also been used in the study of traffic accidents (Wang et al 2013, Obeidat et al 2017, and Omer et al 2017).
- This research synthesizes site characteristics, space syntax analysis, and temporal characteristics into a single model

Network Kernel Density Estimation (KDE)

- 2D KDE is a flawed way of smoothing traffic accidents, though it has been done (Anderson 2009, and Hashimoto et al. 2016).
- Two Network KDE algorithms developed by Okabe et al. (2009).
- Used for spatiotemporal analysis by Kaygisiz et al. (2015) and Romano and Jiang (2017).
- Too computationally intensive for entire city of Dallas

Space Syntax Analysis

- Proposed by Hillier et al. (1976) as a way of quantifying the topology of a network
- Traditionally done with Axial Analysis
- Angular Segment Analysis proposed as a way of quantifying the geometry of a network (Hillier and Iida, 2005)
- Angular Segment Analysis with centerlines instead of axial lines is computationally simpler (Turner, 2007)
- Choice
  - Count of times shortest path between two nodes uses a particular segment
  - In essence, how often will an omnipotent traveler choose a segment
- Integration
  - Average of the distance from a particular segment to all other segments
  - In essence, how centrally located is a segment

UT DALLAS

Accident Interaction

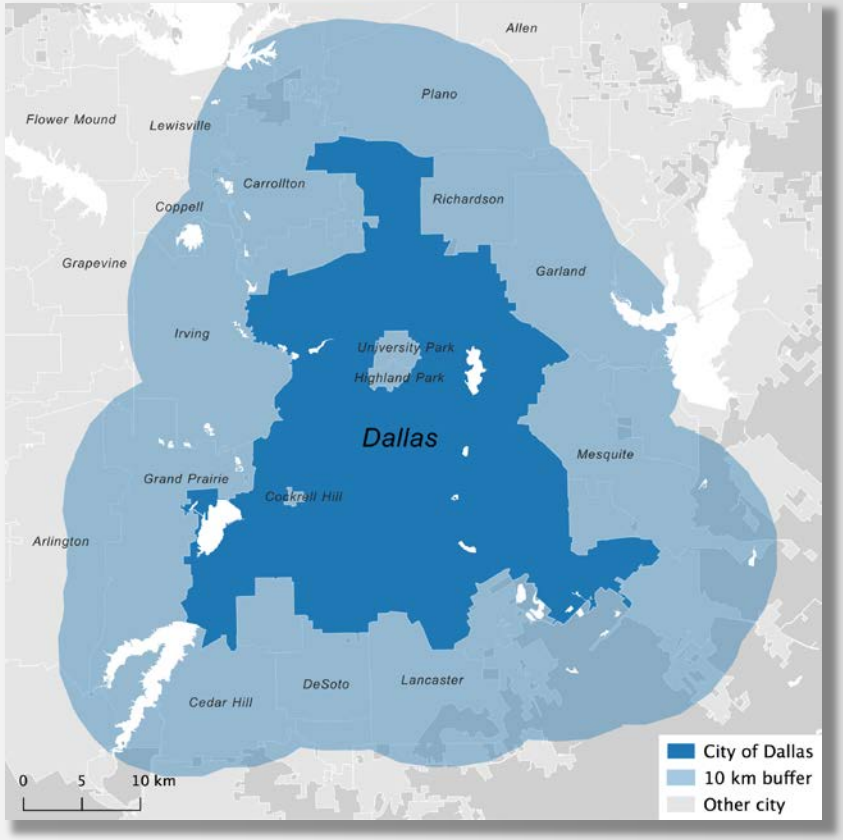
- O'Sullivan and Perry (2013) delineate the difference between first-order structure and second-order structure in spatial pattern modeling
  - Impetus for looking at the second-order structure, "cascading effects", of traffic accidents
- Inspired by "near repeats" in criminology, where a previous crime increases the likelihood of a new crime nearby in space and time (Townsley et al. 2003)
- As in criminology, a directional constraint in the temporal dimension is necessary
- Any point within 200 meters of an accident that occurred in the preceding hour
- Euclidean distance chosen over network distance, as line-of-sight is more important than travel distance

Statistical Learning: Literature Review

- James et al. (2013) and Hastie et al. (2009) describe various statistical learning approaches to predictive modeling
  - They show that logistic regression models and random forest decision tree models can be used not only for binary discrete classification, but also for assessment of the probability of binary discrete classification
- King & Zeng (2001) show how to use logistic regression models for rare-event data
  - For case-control designs, simply modify the intercept coefficient

Study Area

- City of Dallas
- 68,480 accidents over the course of 2015 and 2016
- Space syntax analysis is vulnerable to edge effects, so roads within a 10 km buffer of Dallas were used for space syntax



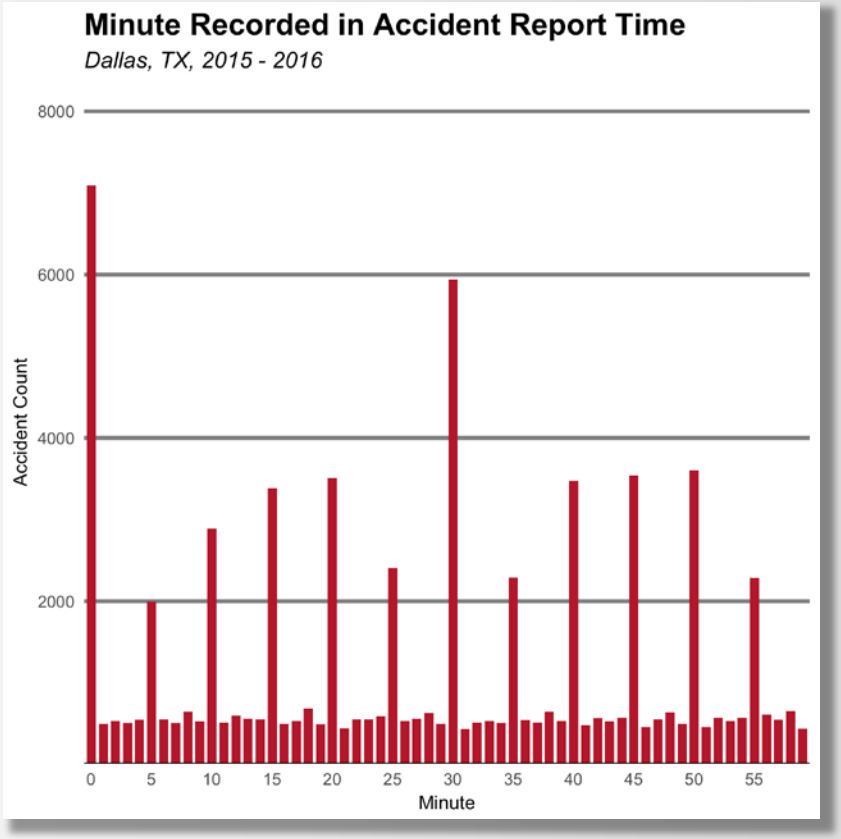
Data

- Traffic accident reports
  - Texas Department of Transportation (TXDOT)
- Road centerlines
  - Texas Department of Transportation (TXDOT)
- Weather (Daily Summaries)
  - NOAA National Centers for Environmental Information
- Ancillary cartographic data
  - City of Dallas, US Census Bureau, Texas Natural Resources Information System, Natural Earth, USGS

UT DALLAS

Temporal Resolution of Accident Data

- Accident reports do not appear to have a consistent temporal resolution greater than 1 hour
- An argument could be made for 30 minute resolution, but this research opted to use 1-hour intervals



Statistical Learning Approache

- Objective:
  - Model the probability of an accident occurring at any 100m road segment at any hour in Dallas, TX
- General characteristics:
  - Consider each 100m segment at each hour in 2015-2016 as an observation
  - Dependent variable is binary value, representing whether an accident occurred at that space-time location
  - Case-control design - create sample of non-accidents, as the full dataset for 2015-2016 would consist of ~1.5 billion observations
  - Sample included all 68,480 accidents and 245,616 non-accident street segments
  - Training and testing data split by separating odd days from even days

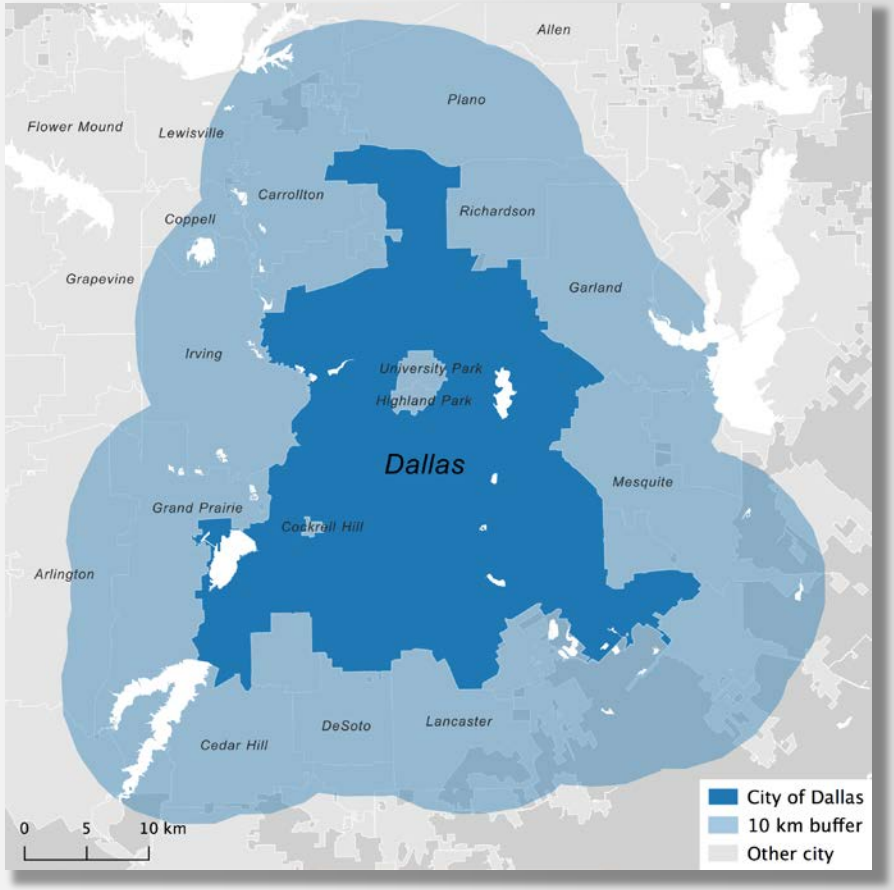
UT DALLAS

Parallel Computing

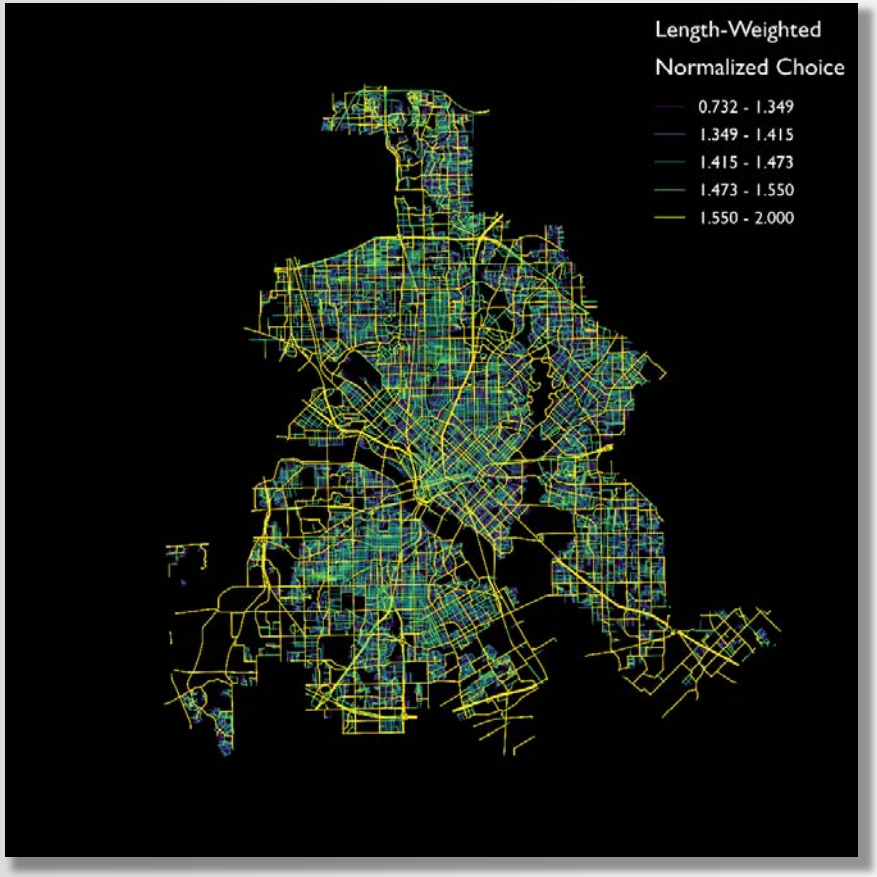
- Computer specifications:
  - 128 GB RAM
  - Intel Xeon CPU E5-2687W v3 @ 3.10GHz
- 15 Threads using R libraries DoParallel and ForEach
  - Calculating accident near-repeats
  - Building non-accident sample
  - Building random forests
- Microsoft Open R
  - GLM estimation and prediction

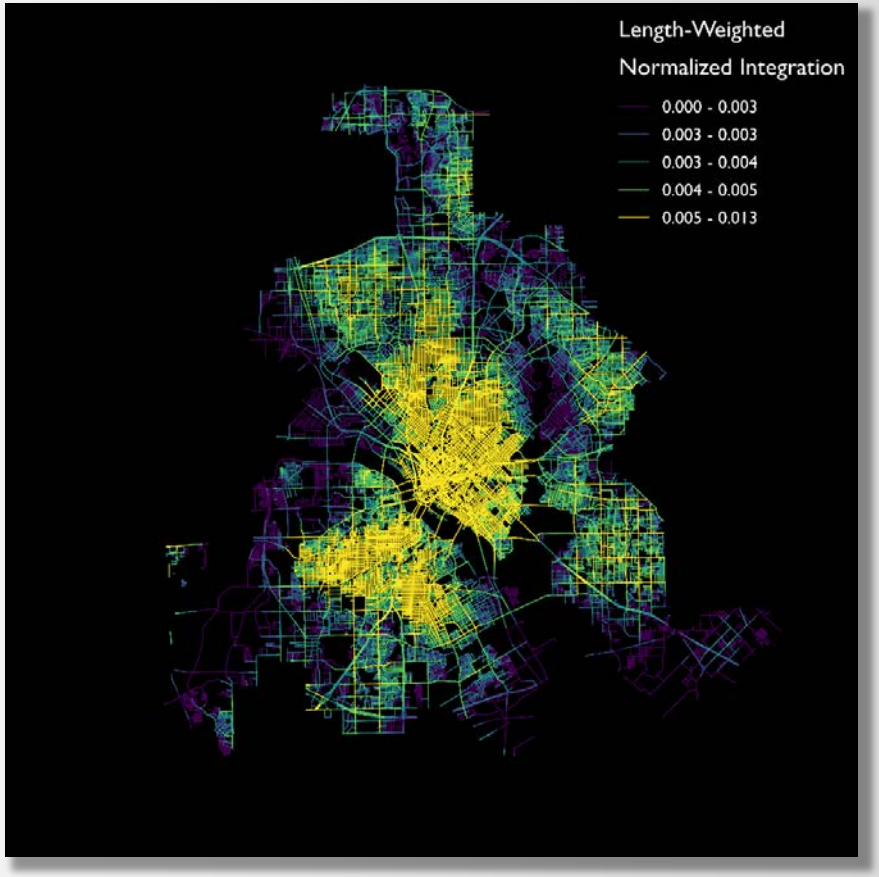
Space Syntax in Dallas, TX

- Angular Segment Analysis
- Length-Weighted
- Range of 10 kilometers
- Study area included a 10 km buffer around Dallas in order to avoid edge effects



Space Syntax Outputs for Dallas, TX

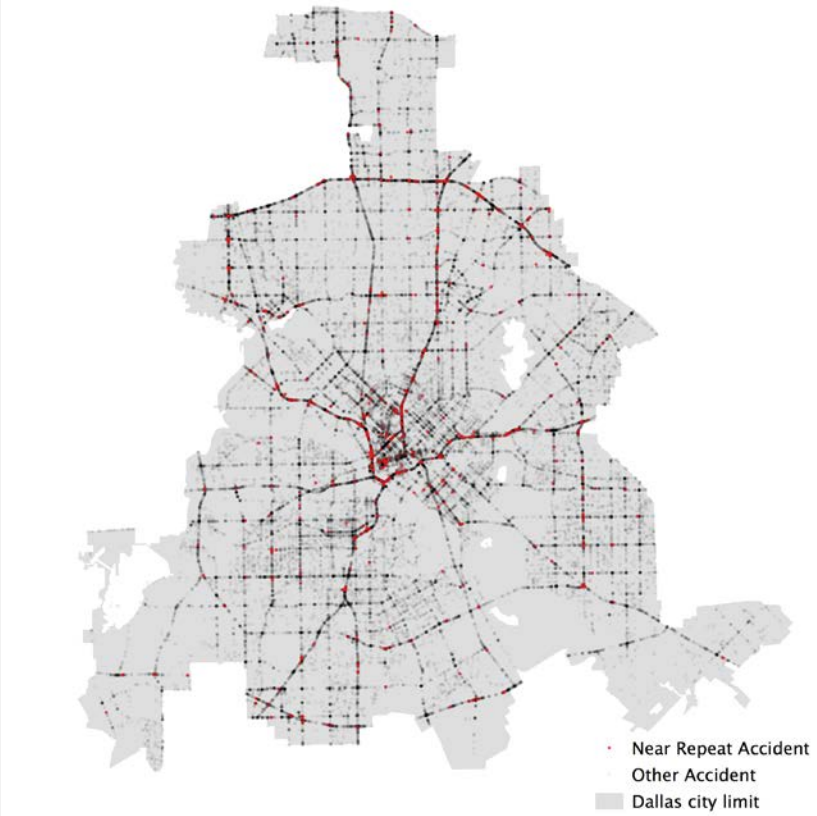




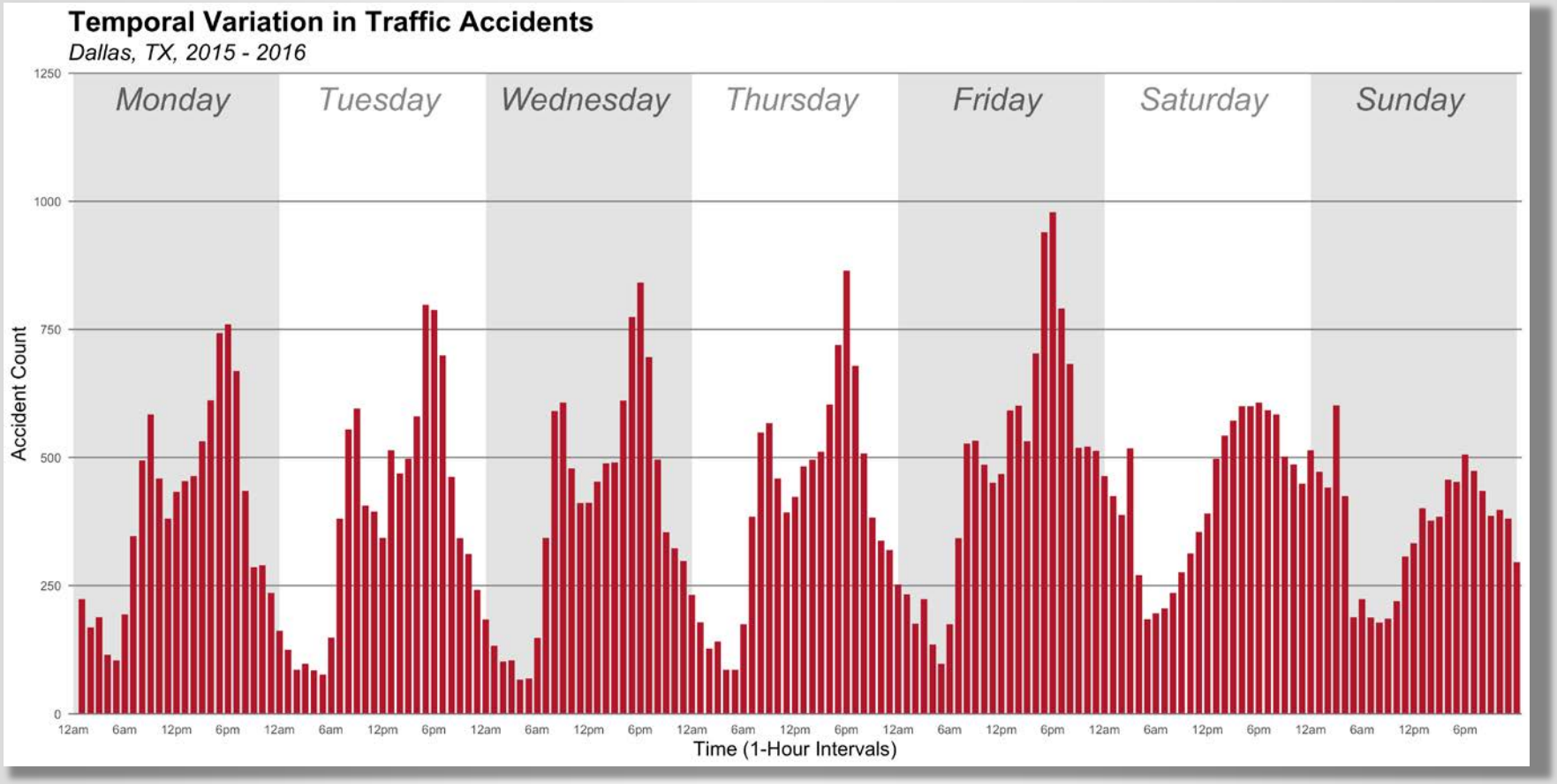
UT DALLAS

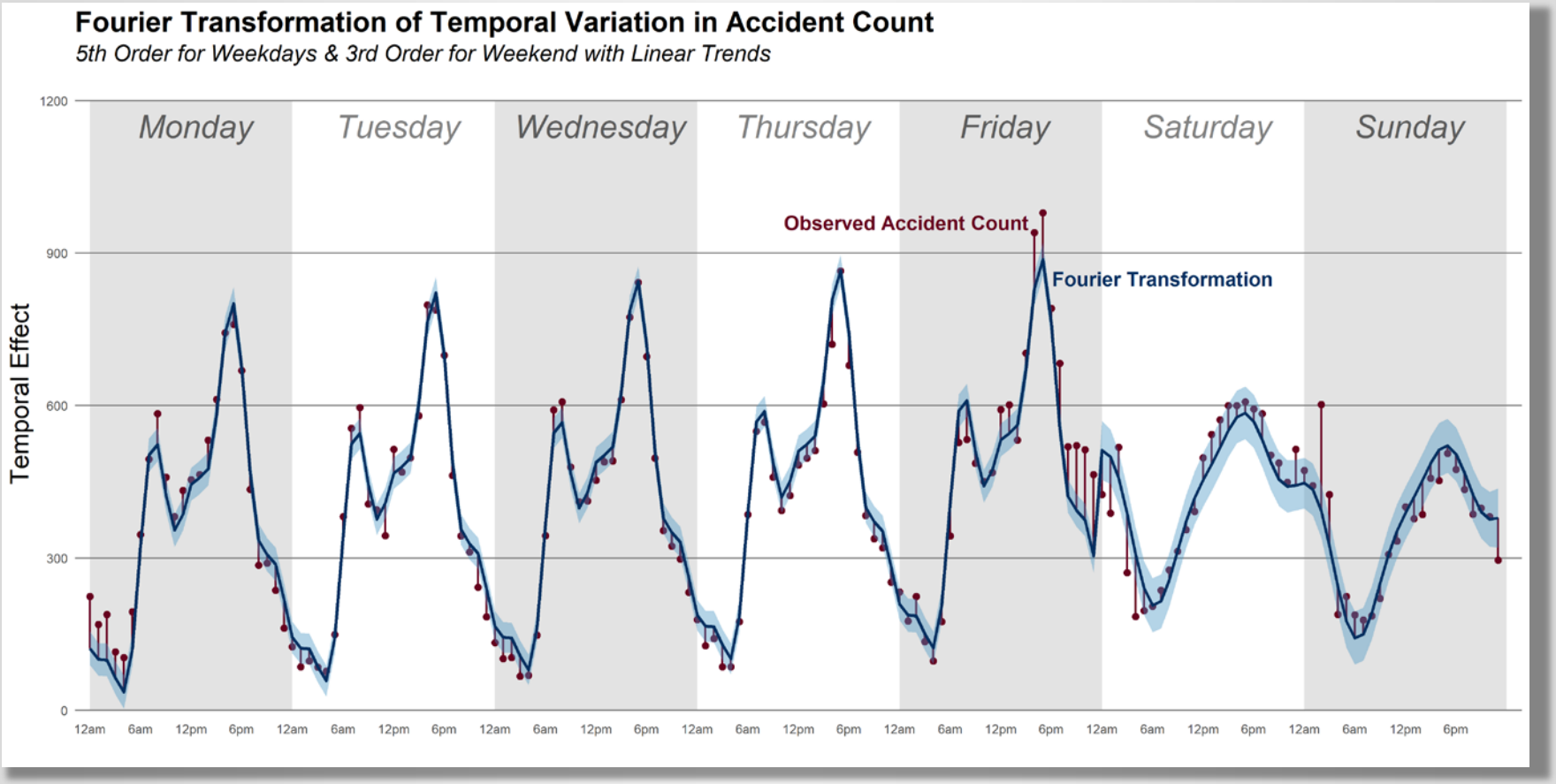
Near Repeat/Cascading Effects

- Near Repeat
  - Any point within 200 meters of an accident that occurred in the preceding hour
  - Euclidean distance chosen over network distance, as line-of-sight is more important than travel distance
- 600 near repeat accidents in Dallas between 2015 and 2016
- Largely concentrated on highways



UT DALLAS


Temporal Variation in Traffic Accidents
Dallas, TX, 2015 - 2016
Monday
Tuesday
Wednesday
Thursday
Friday
Saturday
Sunday
Accident Count
1250
1000
750
500
250
0
12am
6am
12pm
6pm
Time (1-Hour Intervals)



Weekday R^2_{adj} = 0.95

Weekend R^2_{adj} = 0.77

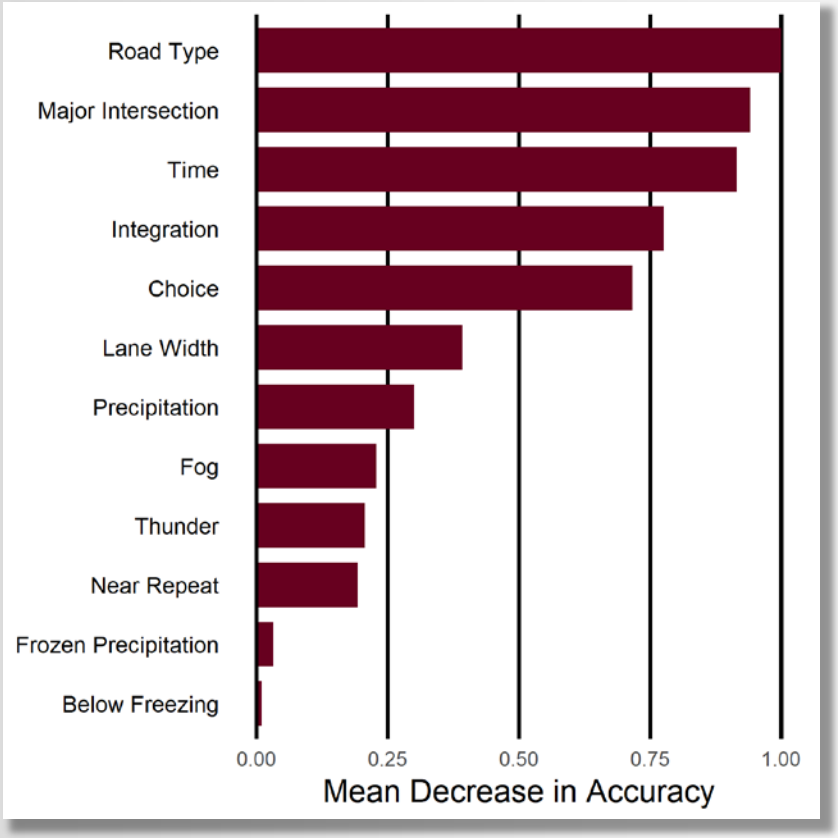
Logistic Regression Variable Importance

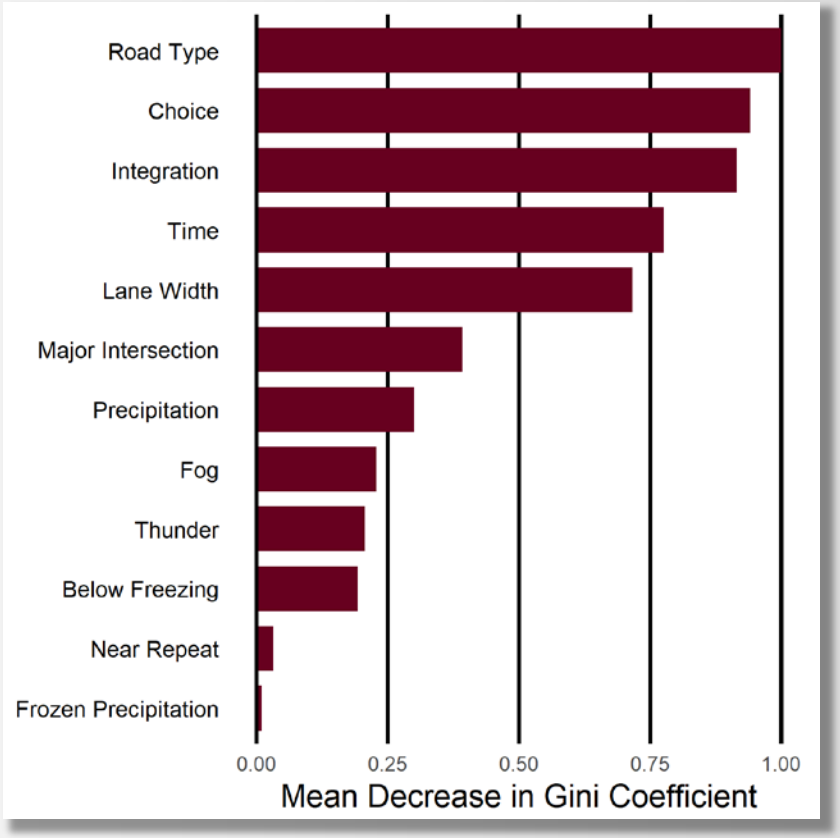
Variable	ΔAUC
Near Repeat	**0.02%**
Frozen Precip.	0.00%
Thunder	0.00%
Fog	0.00%
Below 0°C	0.00%
Precipitation	0.00%

Variable	ΔAUC
Time	**1.18%**
Lane Width	0.00%
Major Inters.	**1.49%**
Road Type	**4.71%**
Integration	**0.33%**
Choice	**0.16%**

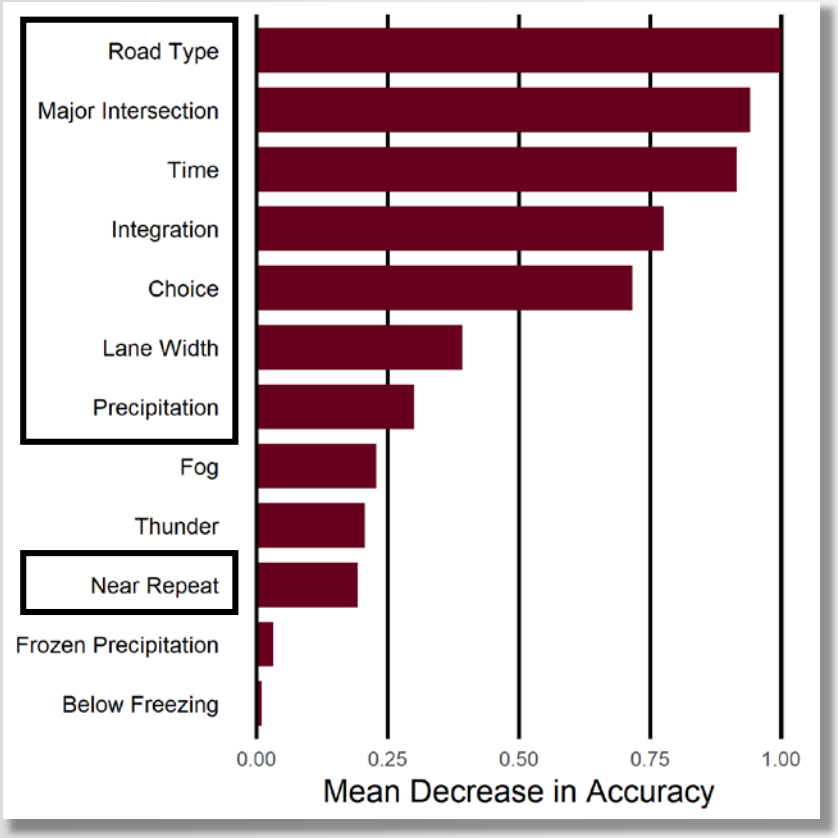
UT DALLAS

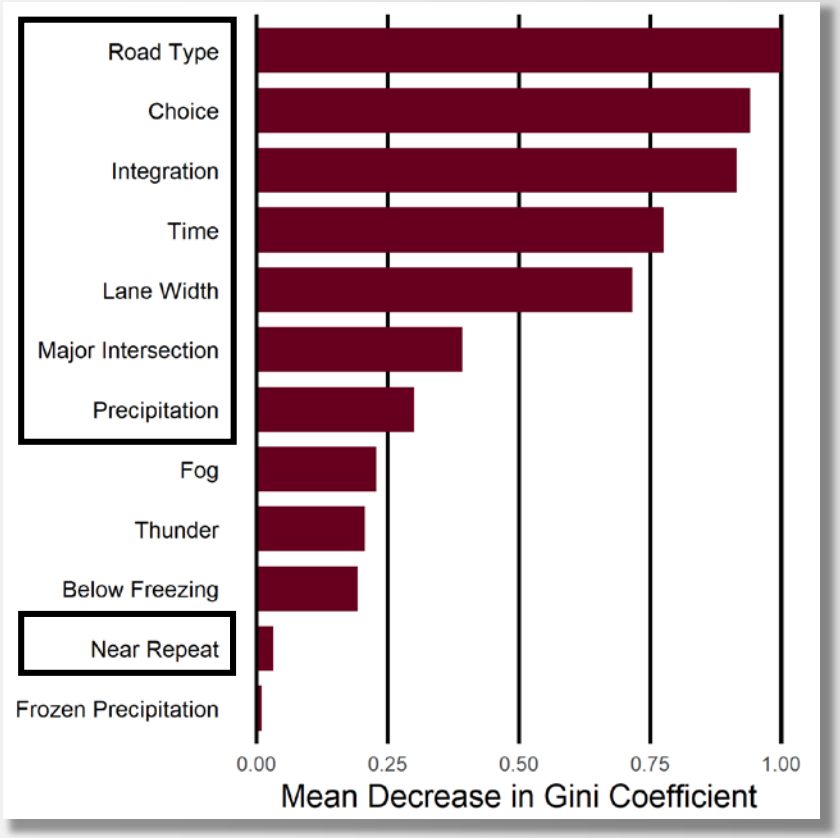
RF Variable Importance





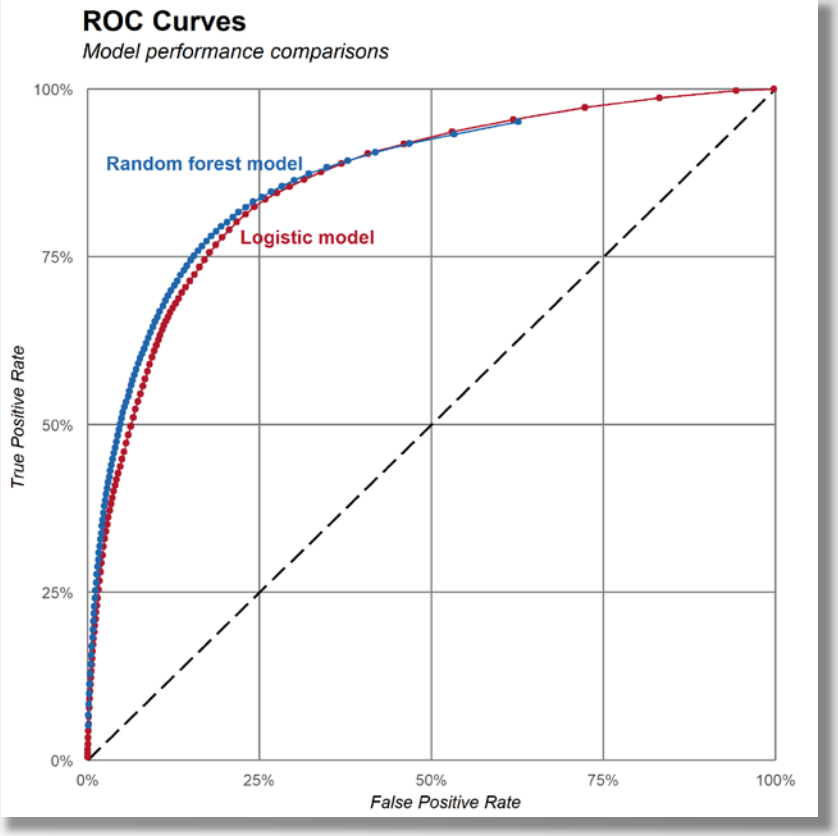
RF Variable Importance





Model Comparison

Metric	Logistic Regression	Random Forest
RMSE	0.3350	0.3272
Accuracy	84.11%	85.42%
Sensitivity	52.31%	51.83%
Specificity	93.04%	94.86%



Logistic Regression Coefficient Estimates

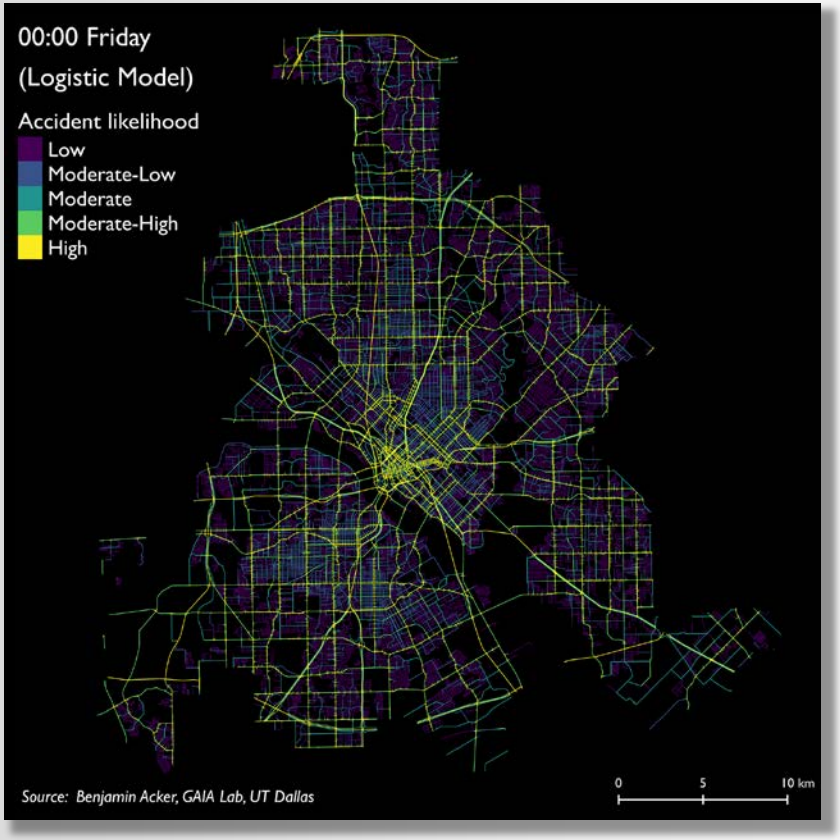
Variable	Estimate	P-value
Choice	1.356	0.000
Choice x Time	0.001	0.002
Integration	49.04	0.000
Integration x Time	0.122	0.000
Major Intersection (T)	0.918	0.000
Intersection x Time	0.001	0.000

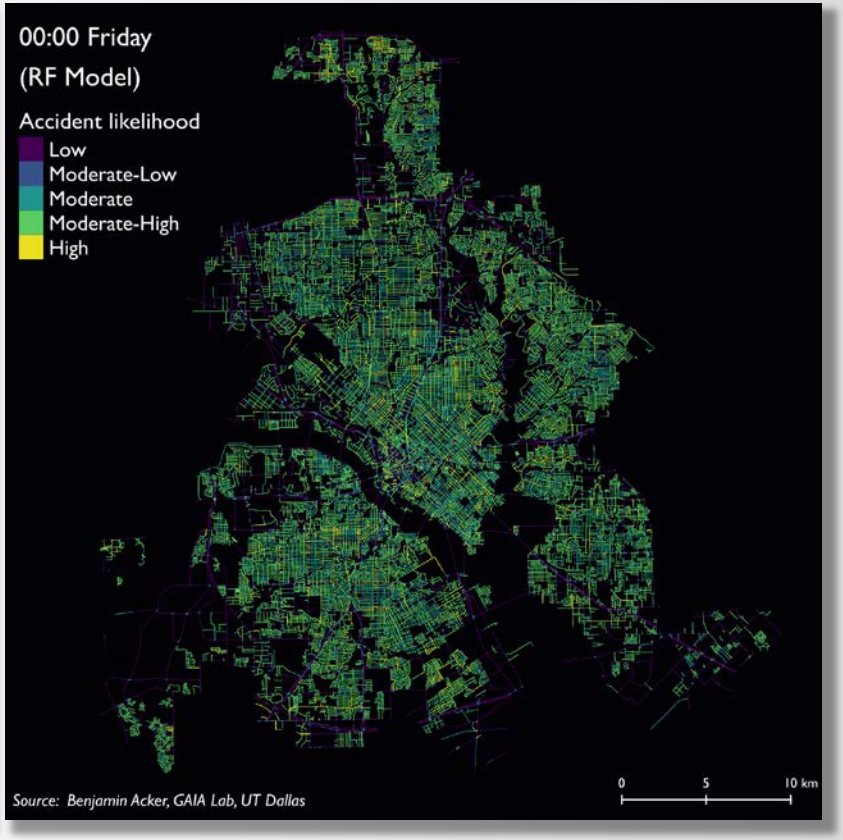
Variable	Estimate	P-value
Highway	2.405	0.000
Other Arterial Road	0.281	0.000
Local Road	-0.694	0.000
Time	0.000	0.997
Near Repeat (T)	1.287	0.000

Accident Likelihood Classification

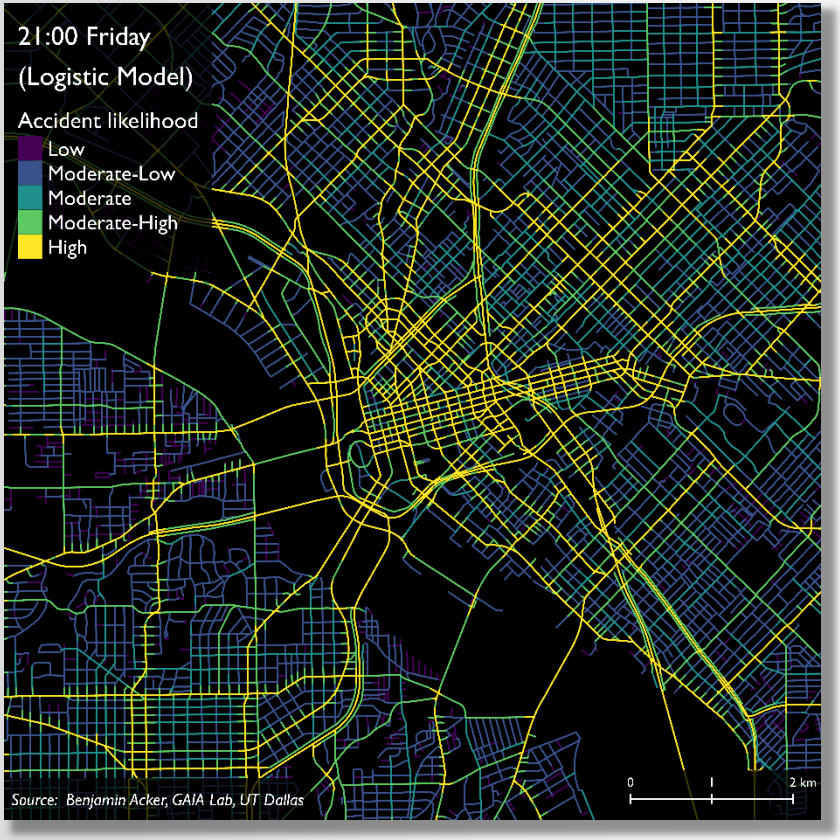
Classification	Logistic (%)	Random Forest (%)
Low	0.0002 - 0.0007	0.0000 – 0.0041
Moderate-Low	0.0007 - 0.0012	0.0041 – 0.0053
Moderate	0.0012 - 0.0018	0.0053 – 0.0068
Moderate-High	0.0018 - 0.0044	0.0068 – 0.0087
High	0.0044 - 2.9527	0.0087 – 0.0192

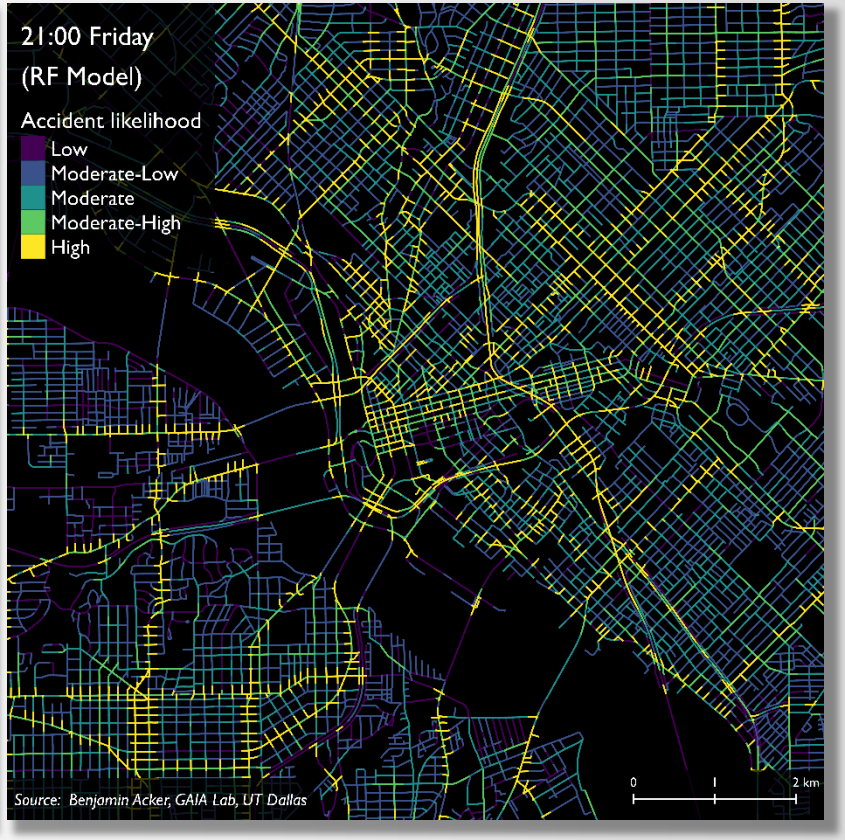
Model Output

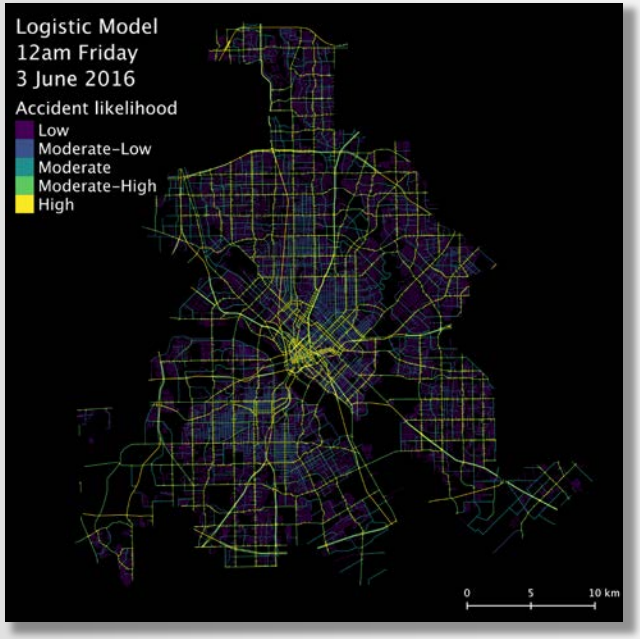


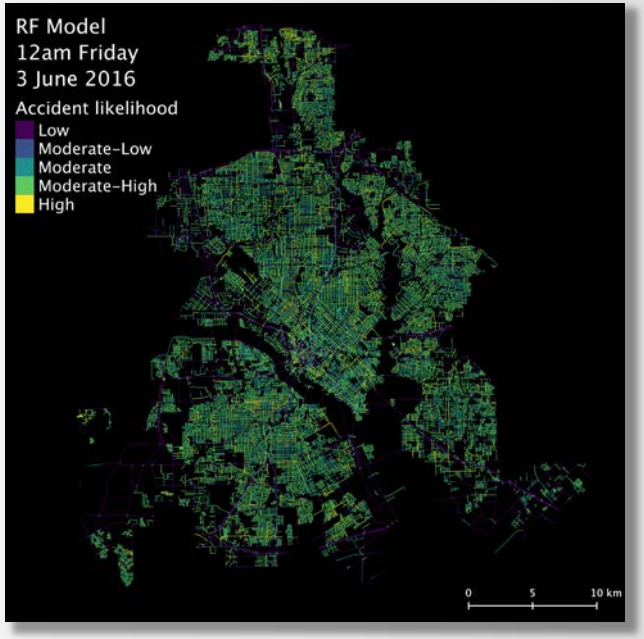


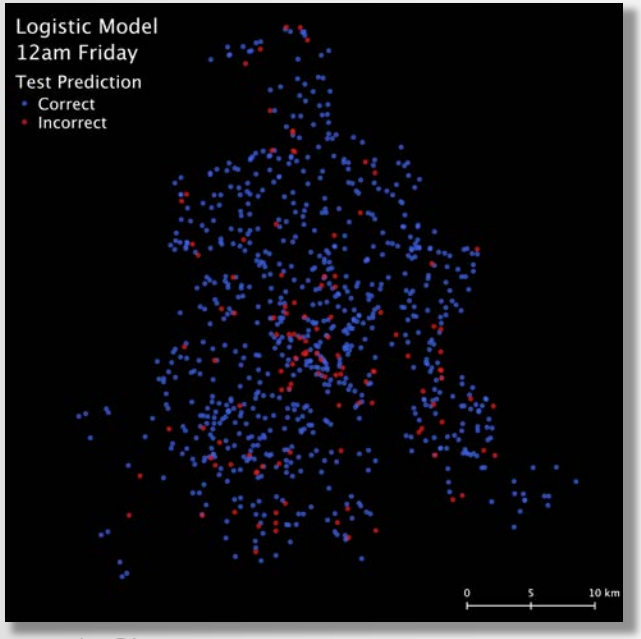
Model Output: Downtown 9pm

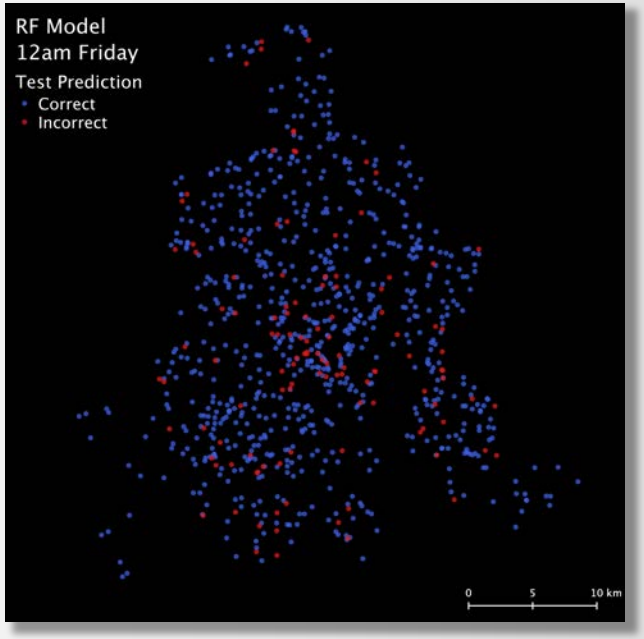




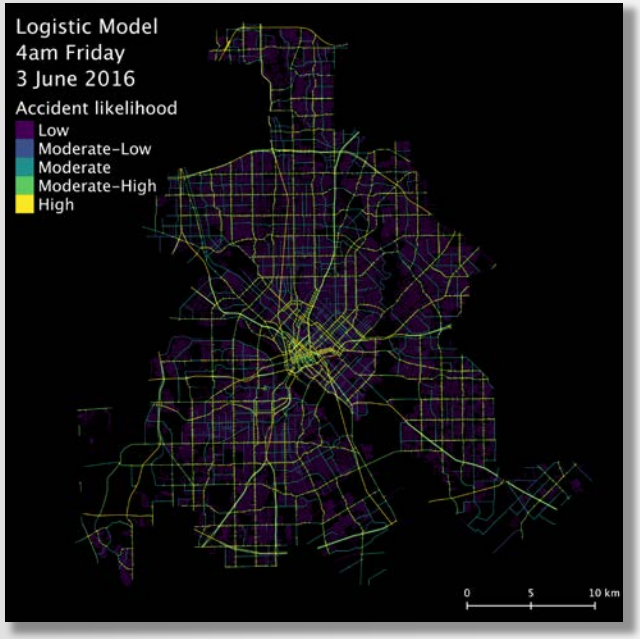

Logistic Model
12am Friday
3 June 2016
Accident likelihood
Low
Moderate-Low
Moderate
Moderate-High
High
0
5
10 km

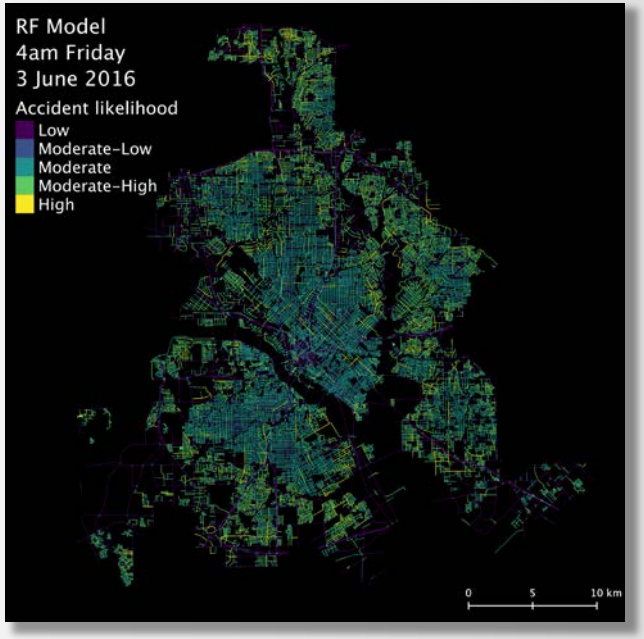

RF Model
12am Friday
3 June 2016
Accident likelihood
Low
Moderate-Low
Moderate
Moderate-High
High
0
5
10 km

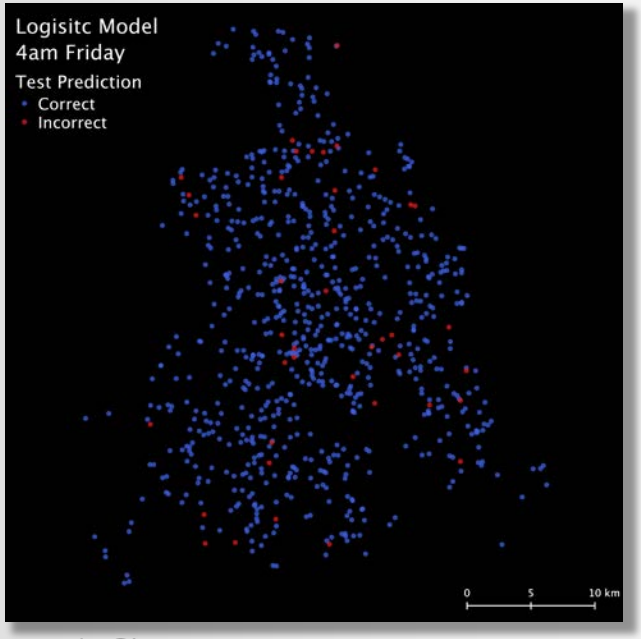

Logistic Model
12am Friday
Test Prediction
Correct
Incorrect
0
5
10 km

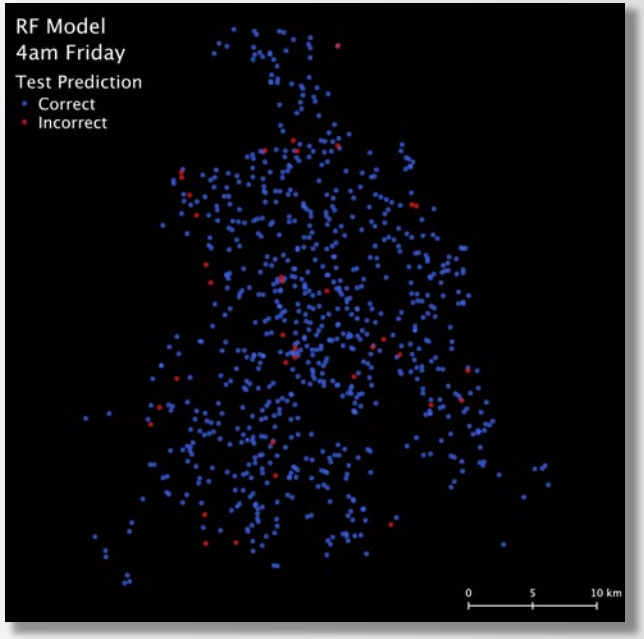

RF Model
12am Friday
Test Prediction
Correct
Incorrect
0
5
10 km

UT DALLAS

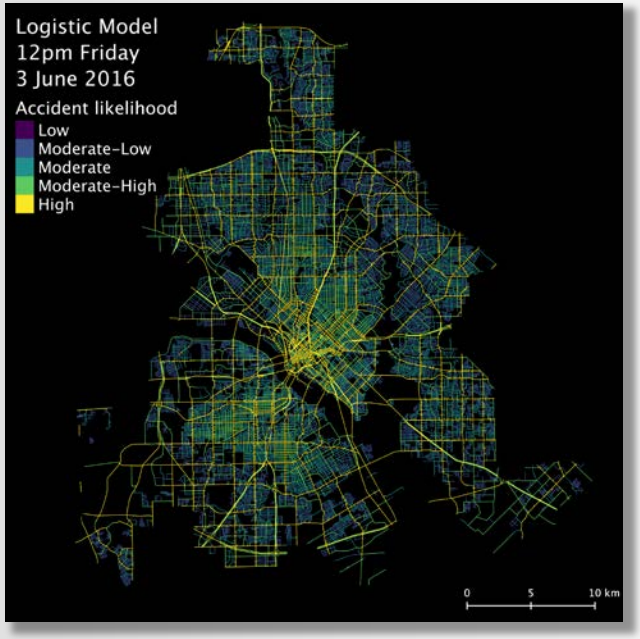

Logistic Model
4am Friday
3 June 2016
Accident likelihood
Low
Moderate-Low
Moderate
Moderate-High
High
0
5
10 km

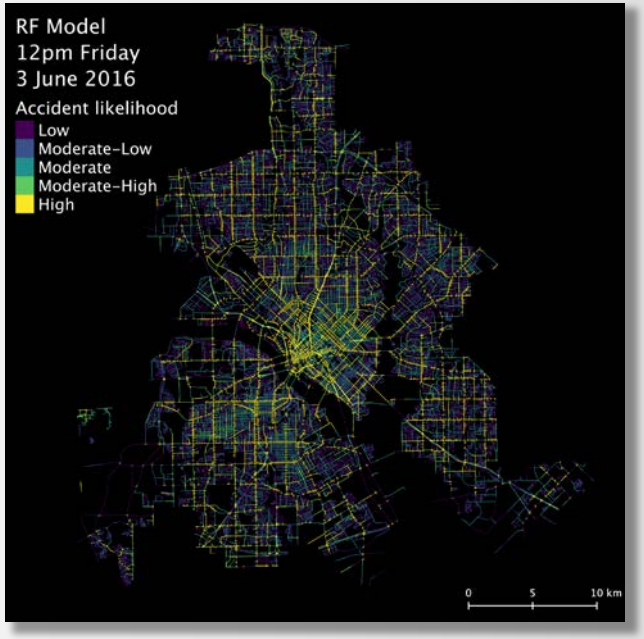

RF Model
4am Friday
3 June 2016
Accident likelihood
Low
Moderate-Low
Moderate
Moderate-High
High
0
5
10 km

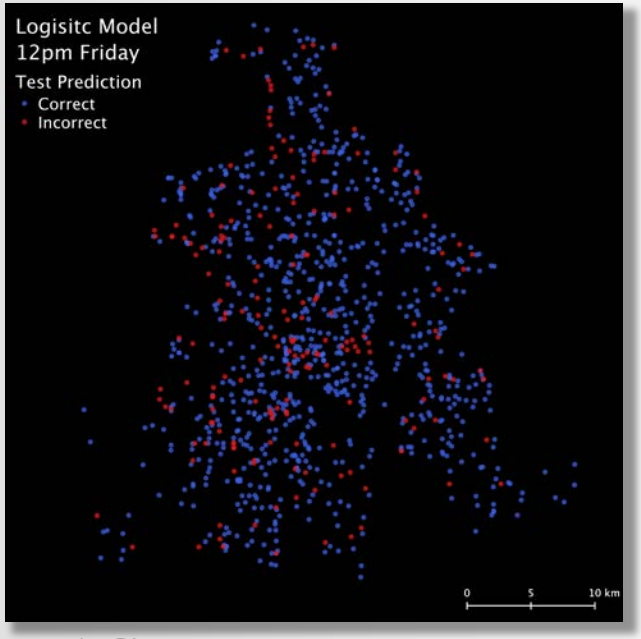

Logisitc Model
4am Friday
Test Prediction
Correct
Incorrect
0
5
10 km

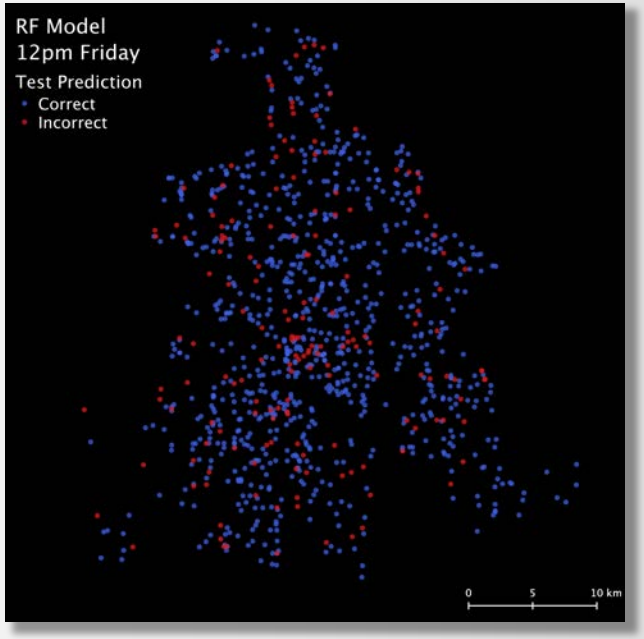

RF Model
4am Friday
Test Prediction
Correct
Incorrect
0
5
10 km

UT DALLAS

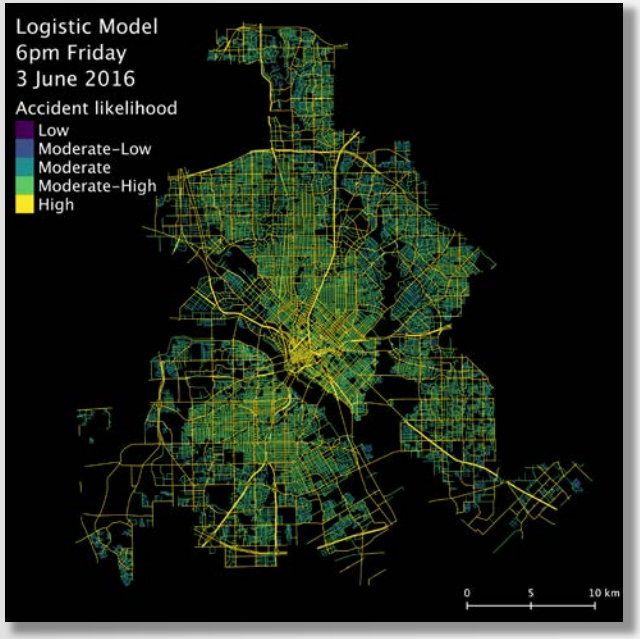

Logistic Model
12pm Friday
3 June 2016
Accident likelihood
Low
Moderate-Low
Moderate
Moderate-High
High
0
5
10 km

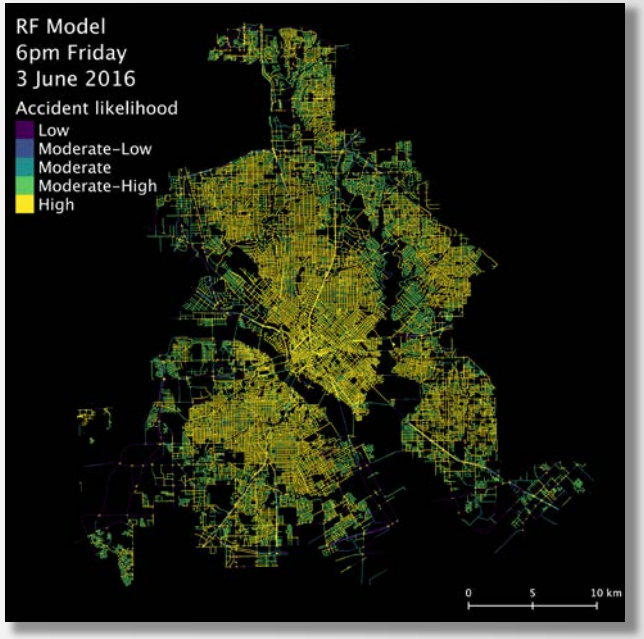

RF Model
12pm Friday
3 June 2016
Accident likelihood
Low
Moderate-Low
Moderate
Moderate-High
High
0
5
10 km

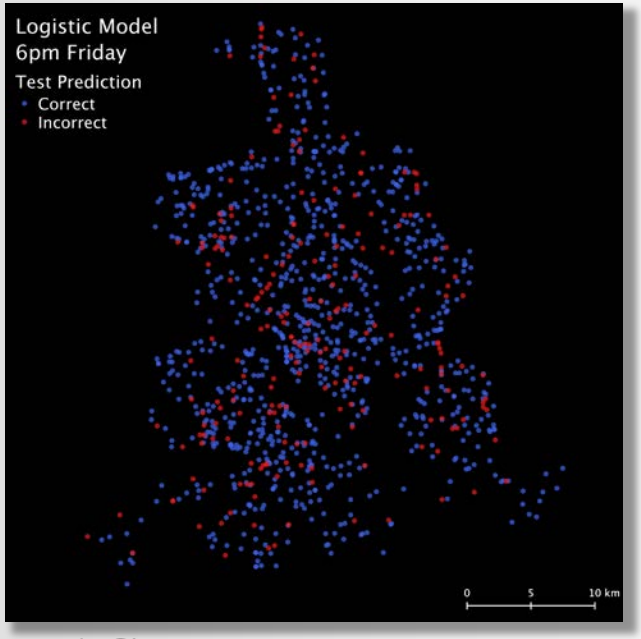

Logisitc Model
12pm Friday
Test Prediction
Correct
Incorrect
0
5
10 km

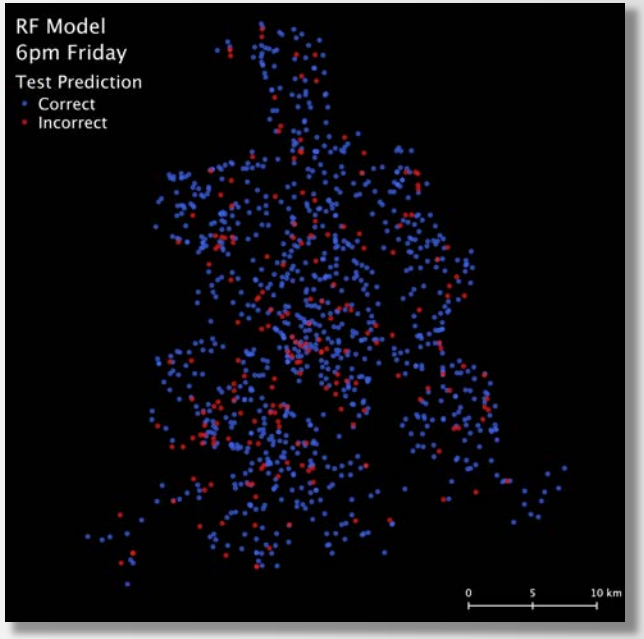

RF Model
12pm Friday
Test Prediction
Correct
Incorrect
0
5
10 km

UT DALLAS


Logistic Model
6pm Friday
3 June 2016
Accident likelihood
Low
Moderate-Low
Moderate
Moderate-High
High
0
5
10 km

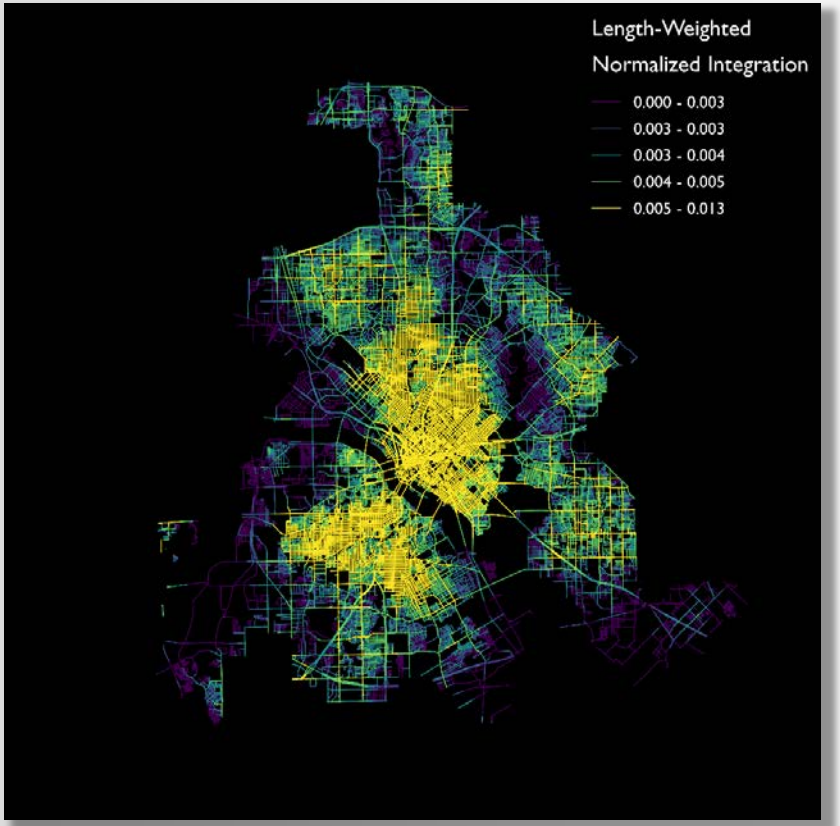

RF Model
6pm Friday
3 June 2016
Accident likelihood
Low
Moderate-Low
Moderate
Moderate-High
High
0
5
10 km

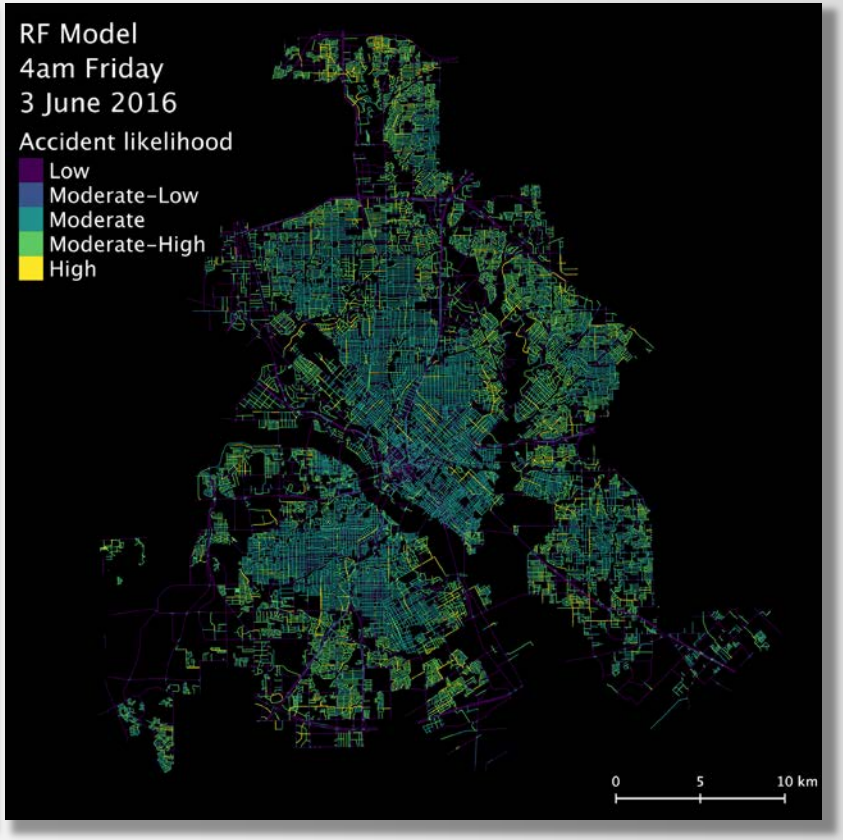

Logistic Model
6pm Friday
Test Prediction
Correct
Incorrect
0
5
10 km


RF Model
6pm Friday
Test Prediction
Correct
Incorrect
0
5
10 km

UT DALLAS

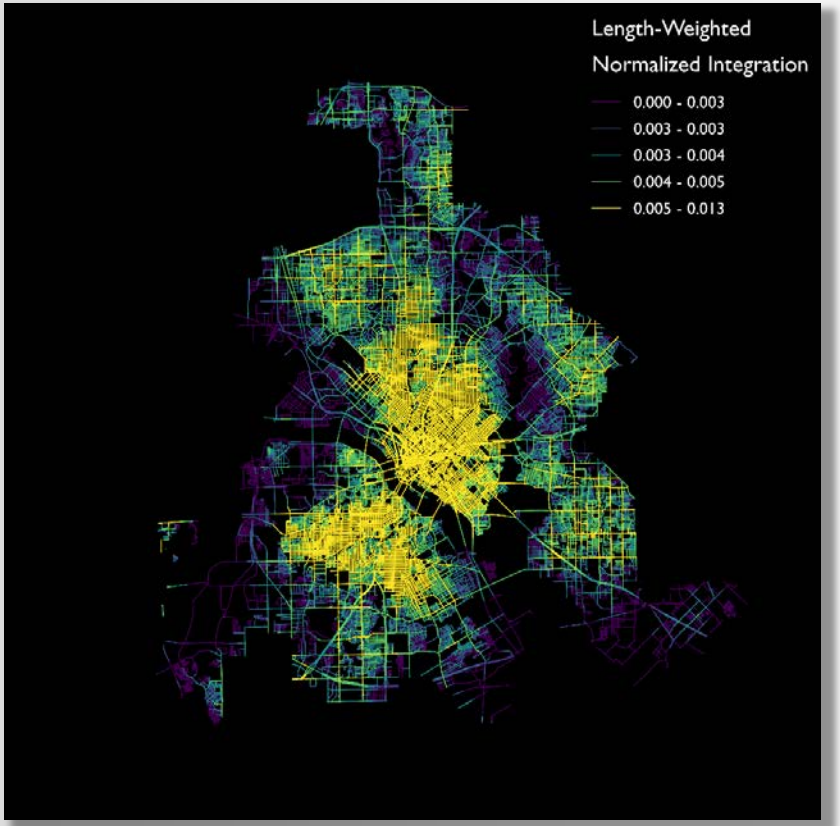
RF Identifies Interaction between Time and Space Syntax

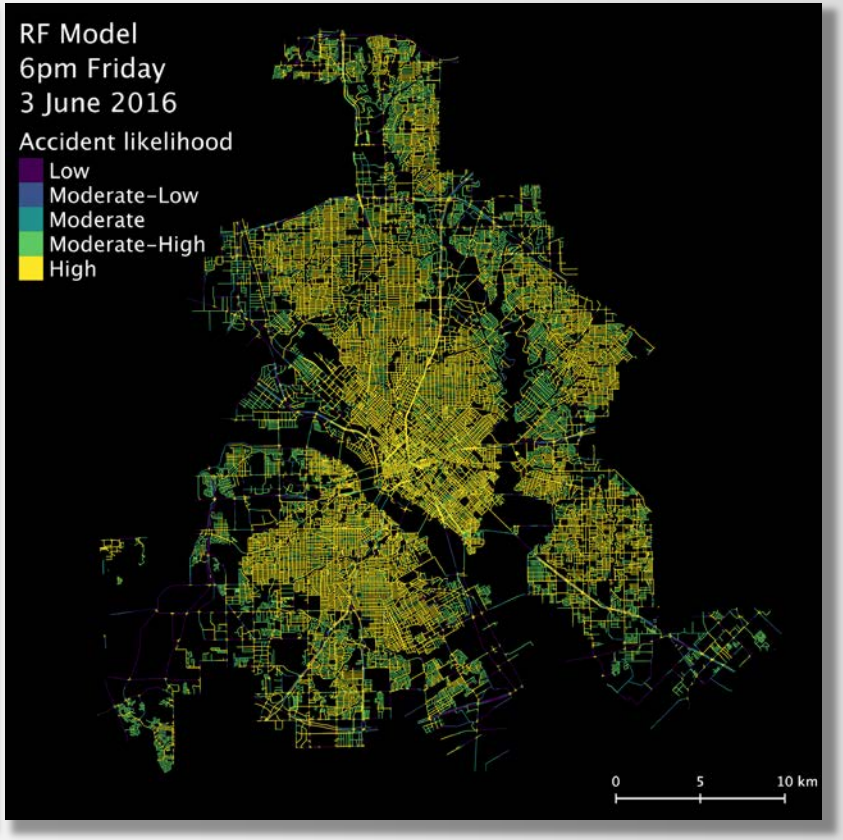


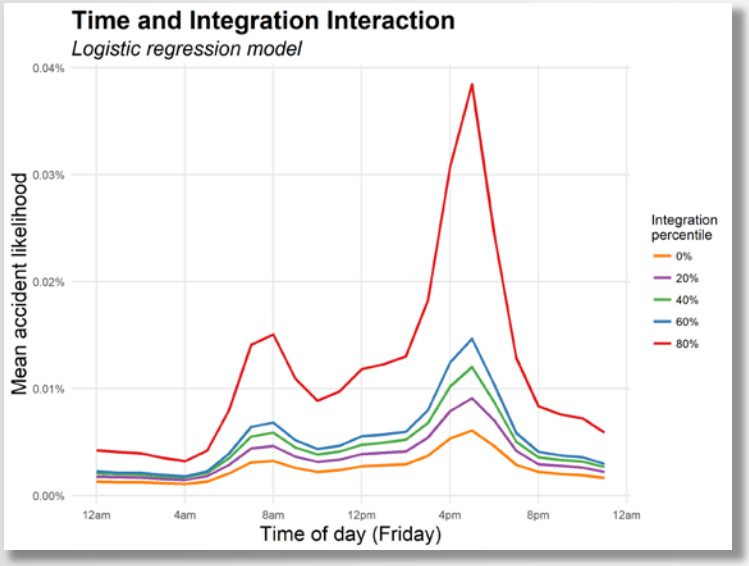


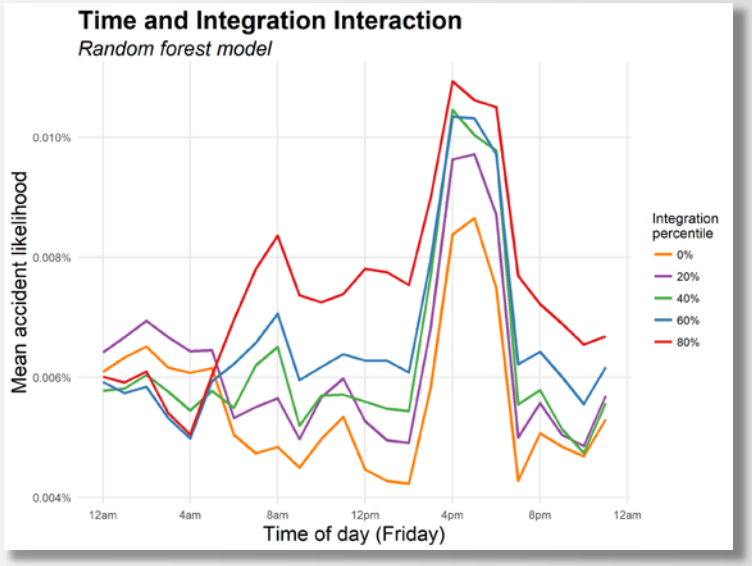
UT DALLAS

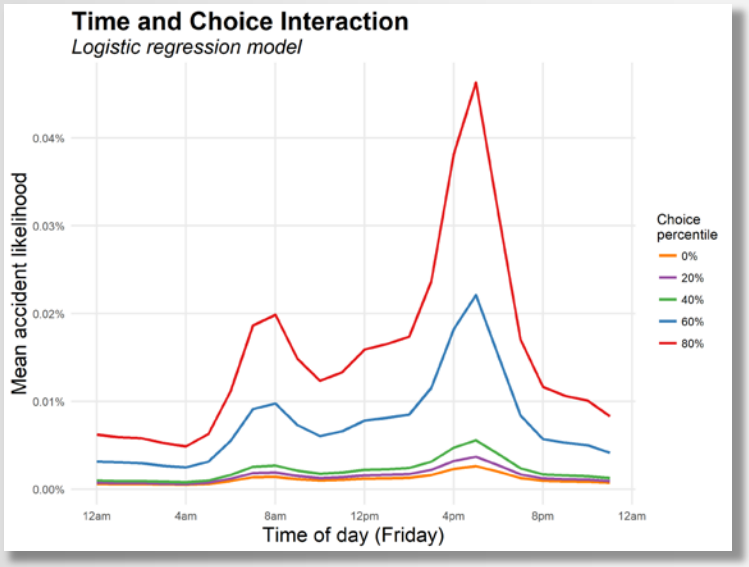
RF Identifies Interaction between Time and Space Syntax

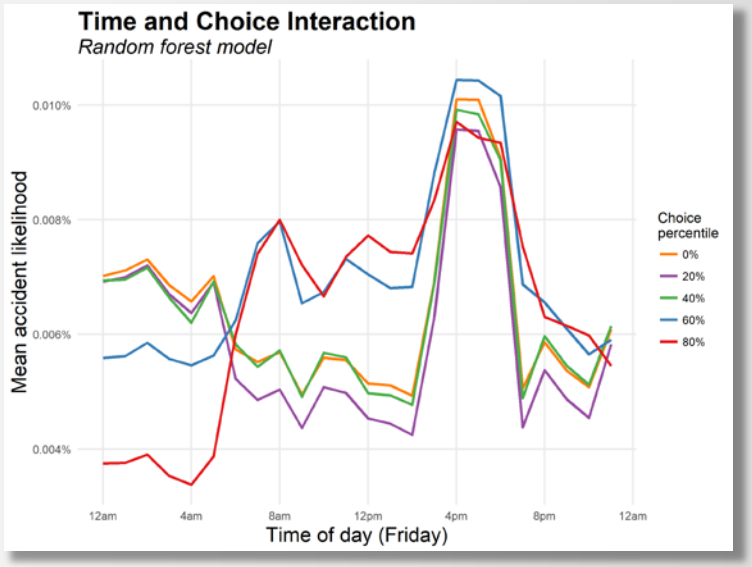





Time and Integration Interaction
Logistic regression model
Mean accident likelihood
0.04%
0.03%
0.02%
0.01%
0.00%
12am
4am
8am
12pm
4pm
8pm
12am
Time of day (Friday)
Integration percentile
0%
20%
40%
60%
80%


Time and Integration Interaction
Random forest model
Mean accident likelihood
0.010%
0.008%
0.006%
0.004%
12am
4am
8am
12pm
4pm
8pm
12am
Time of day (Friday)
Integration percentile
0%
20%
40%
60%
80%


Time and Choice Interaction
Logistic regression model
Mean accident likelihood
0.04%
0.03%
0.02%
0.01%
0.00%
12am
4am
8am
12pm
4pm
8pm
12am
Time of day (Friday)
Choice percentile
0%
20%
40%
60%
80%


Time and Choice Interaction
Random forest model
Mean accident likelihood
0.010%
0.008%
0.006%
0.004%
12am
4am
8am
12pm
4pm
8pm
12am
Time of day (Friday)
Choice percentile
0%
20%
40%
60%
80%

Conclusions

- Together, site characteristics, space syntax variables, and time are excellent predictors of traffic accidents
- The interaction between time and site characteristics is particularly interesting
  - Time and integration could represent movement between suburbs and city center
  - Time and choice could represent density of traffic on large roads
  - Time and intersection could represent danger of making a turn
- Second-order structure, specifically cascading effects/near repeats, does appear to have a positive increase in traffic accident likelihood

UT DALLAS

Limitations

- Model predicts likelihood of accident occurring on road segment, not the risk for individual drivers
  - The latter would require underlying population data, i.e. spatio-temporal traffic counts
  - Site characteristics might be less important than driver characteristics, e.g. intoxication, sleep deprivation, texting, etc…
- Model relies on consistent reporting of accidents, but not all communities are necessarily equally likely to call the police after an accident
  - Carrollton Police Dept. reported that they had received an unusually low number of crime reports in predominately Hispanic district during early 2017
  - They worry a fear of deportation is the root cause
    - Dallas Morning News (April 2017)

Future Work

- Explore which model's explanation of time/site interaction is more representative of reality
- Construct spatiotemporal model of traffic counts
  - Would allow modeling of the risk for individual drivers on each road segment
  - Would enable use of network space-time Ripley's k-function to discern the second-order structure of accidents, i.e. cascading effects
- Use traffic accident events to model pattern of life
- Collect weather data with higher spatial and temporal resolution

UT DALLAS

Questions?