# Tornado Analysis

Stephen Merrill and Kate Gibson

11th February 2017

## Background

- The Storm Prediction Center (SPC) rates tornado severity on the Fujita Scale (F-scale)
- F-scale rating is an integer between 0 and 5, with 5 being most severe
- Rankings are determined subjectively
- Goal: Create a method to objectively rank tornado severity based on past rankings

# Data

- Information on 940 unique tornadoes in 2012 across US
- Date, time, state, property loss, crop loss, injuries, fatalities, length of impact, width of impact, beginning and ending longitude and latitude
- Reclassified state to region. Only included month out of date variables, did not include longitude and latitude

# Data

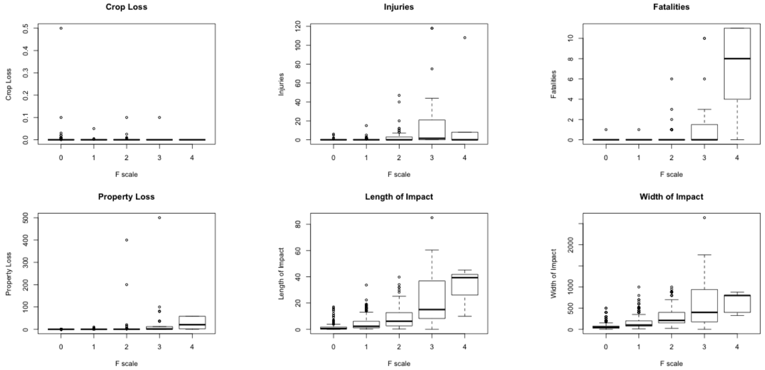


Figure: Distribution of our variables in relation to severity

# Spatial Distribution

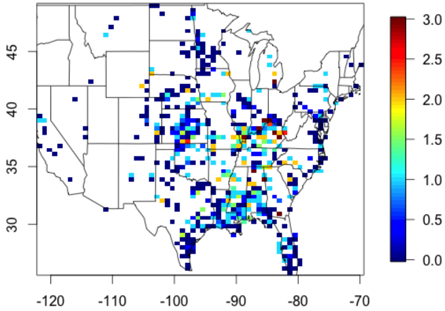


Figure: This plot shows the distribution of tornado severity in our data over the continental US

# Tree Model

- Divide predictor space into T non-overlapping regions $R_1, \ldots, R_T$
- For any $x_0 \in R_t$, make the same prediction

$$\hat{y}(x_0) = \sum_{t=1}^{T} \left( \arg\max_c \pi_c \right) \mathbb{1}(x_0 \in R_t)$$

Where $\pi_c$ is the proportion of $\{(x_i) : x_i \in R_t\}$ of class $c$

## Growing a Tree

Find regions that minimize:

$$Error(y, \hat{y}) + \lambda T$$

- $\lambda$ is a tuning parameter, T is the size of the tree - together they control overfitting
- $Error(y, \hat{y})$ = Gini Index = $\sum_{k=1}^{K} \hat{\pi}_{tk}(1 - \hat{\pi}_{tk})$
- $\hat{\pi}_{tk}$ is the proportion of observations in region $t$ of class $k$

# Random Forests

- Trees tend to be quite variable
- Use bootstrapped samples to create 500 different trees, each time only considering m=3 (chosen by cross-validation) variables
- This decorrelates the trees and gives us better predictive accuracy
- We then average over all the trees to get our predictions (take mode)

# Cross Validation for Random Forest

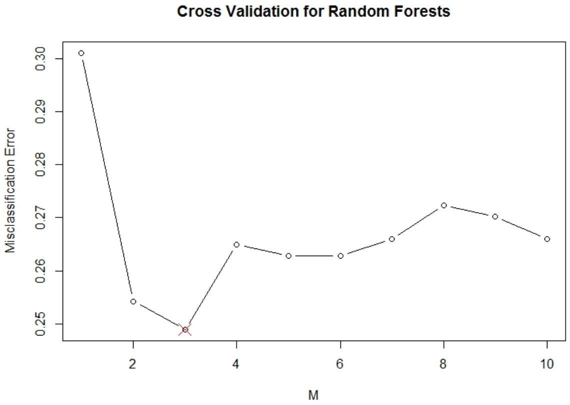


Figure: Considering 3 variables at each node minimizes error.

- Random forests help us complete our goals by allowing us to objectively predict classifications on a 0-5 integer scale for new tornadoes based on previous classifications
- Random forests don't make any assumptions, so we don't have to worry about linear/nonlinear effects

# One Tree in the Forest

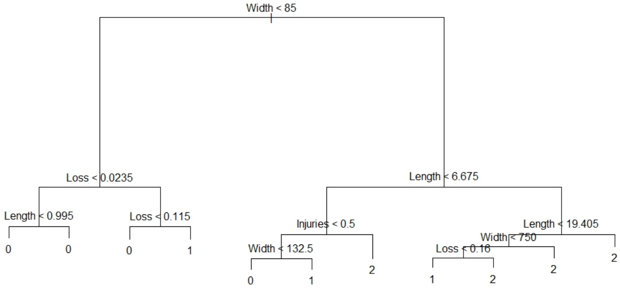


Figure: Width, Loss, Length and Injuries solely determine this tree

# Classifying a New Tornado

- Obtain a classification from each tree in the forest
- Choose final classification by taking the mode of the predictions

## Results

| Predicted | 0 | 1 | 2 | 3 | 4 | Classification Error |
|---|---|---|---|---|---|---|
| **Actual** | | | | | | |
| 0 | 528 | 46 | 2 | 0 | 0 | 0.0833 |
| 1 | 97 | 128 | 15 | 1 | 0 | 0.4688 |
| 2 | 8 | 37 | 46 | 4 | 0 | 0.5157 |
| 3 | 0 | 6 | 13 | 5 | 0 | 0.7916 |
| 4 | 0 | 0 | 1 | 3 | 0 | 1.0000 |

Table: Random Forest Confusion Matrix

- Overall Error: $232/940 = 0.2468$

# Results

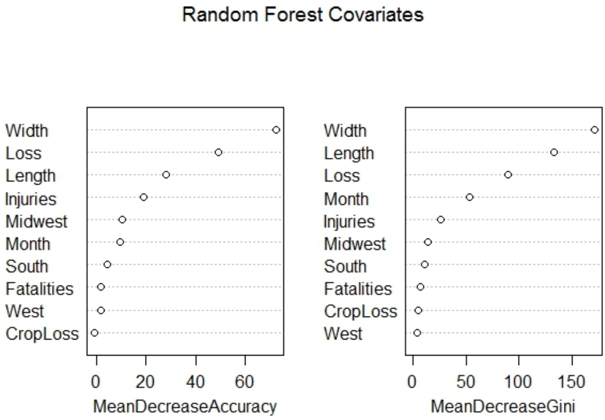


Figure: Effect if each covariate were to be left out of the tree

# Conclusion

- Using random forests allowed us to classify tornadoes objectively
- Poor classification of high F-scale tornadoes due to lack of data
- Only 0 scale tornadoes are classified reliably
- Width, length and loss are the most informative covariates

## Shortcomings and Further Studies

- Unable to make inference and quantify uncertainty of the effects the covariates have on F-scale rating
- Lack of wind speed data
- Better incorporate spatial effect, in the data there appears to be more of an effect than the results show
- Use spatial locations to classify tornadoes based on their likelihood to cause damage in populated areas