Exploratory Data Analysis - Tornado

Stat536 - Dr. Heaton

Rachel Messick & Jacob Mortensen

March 7, 2015

Goals of Analysis

The reach of tornados can be devastating to a community's economic and physical wellness. Because of these potentially significant ramifications, the Storm Prediction Center (SPC) at the National Ocean and Atmospheric Administration (NOAA) is interested in tracking the trends in tornado severity over space and time. However, accurately mapping these trends requires a viable metric for measuring the damages associated with a tornado. The Fujita scale (F-scale) is the current standard for measuring such damages.

The purpose of this analysis is to build an objective classification system consistent with previous classifications, such that the SPC can better understand how the severity of tornados might be varying over space and time.

Data

An immediately obvious feature of the data is that the response is categorical with 5 levels. This will play an important role in determining what method of analysis we use. Initially, there are several variables that appear to have clear relationships with F-scale. In Figure 1 we can see that number of injuries, number of fatalities, the length and width of the tornado path all have clear relationships with the F-scale that a tornado is assigned. In each case, as injuries, fatalities, length and width increase, we see that F-scale also increases.

Many of the variables included in the data set are related to when the tornado occurred. While this

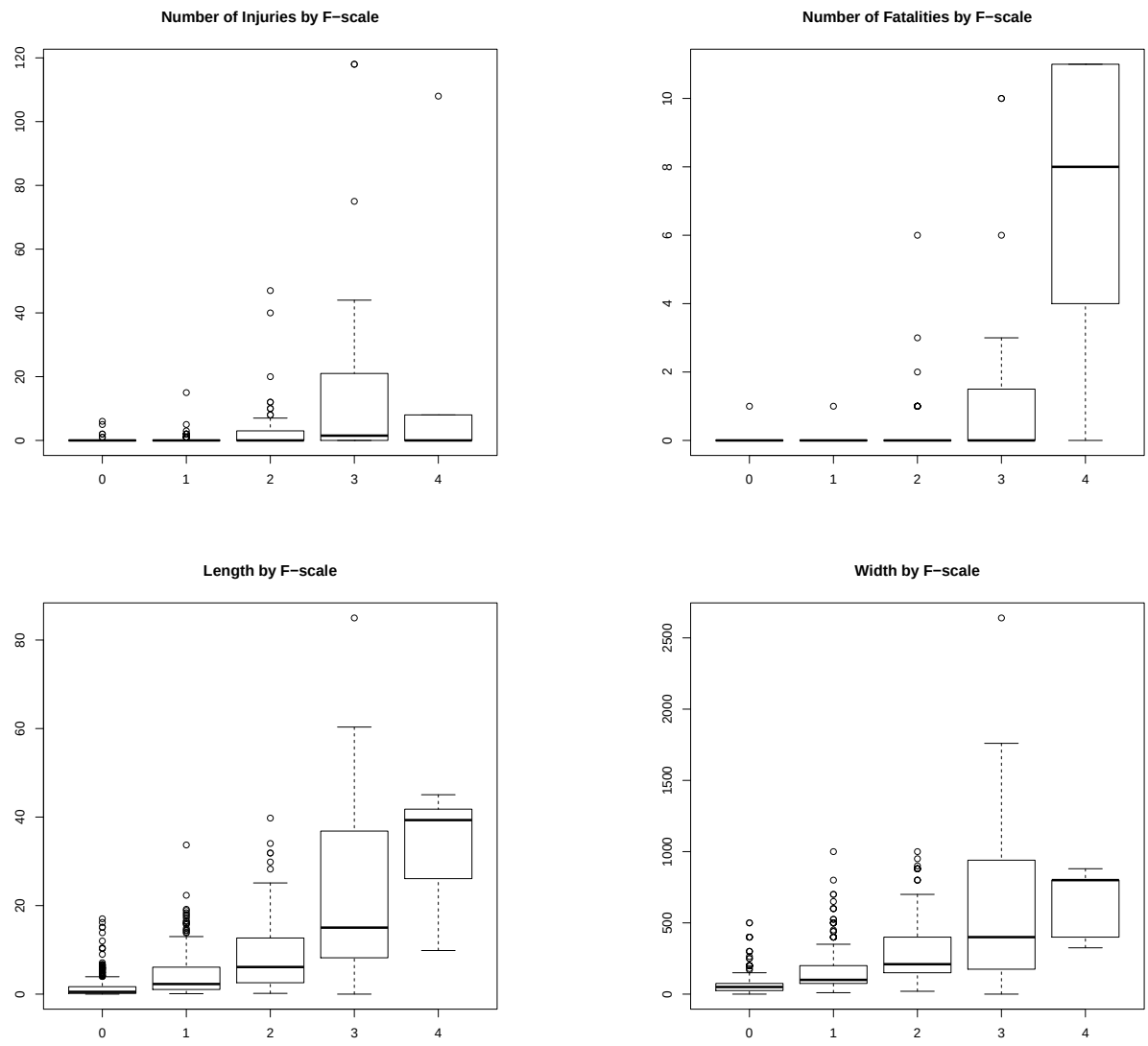

Figure 1: Relationship between F-scale and various other variables

information may not improve our ability to classify the severity of a tornado, it could prove valuable by helping us to identify repeated observations.

Method

Because the response is a non-binary classification, neither linear regression nor logistic/probit regression is suitable. One possible method might involve something similar to logistic regression, but instead of the responses following a Bernoulli distribution, they follow a sort of multinomial distribution such that each observation has a set of probabilities of landing in each of the possible classes. Alternatively, we know

from literature that one method for handling classification is classification trees. We do not, however, have sufficient knowledge of the topic to carry out an analysis.